Institut des ACTUAIRES

Dauphine | PSL UNIVERSITÉ PARIS

Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires le 30 janvier 2025

Par : Léopoldine BAUMARD
Titre : Prédiction de la sinistralité en assurance annulation

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité ci-dessus

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Europ Assistance Holding
Signature :

europ assistance Holding
S.A.S au capital de 26 385 968 €
Siège Social : 2 rue Pillet-Will
75009 PARIS
01 58 34 23 00
SIRET 632.016 382 00178

Directeur de Mémoire en entreprise :
Nom : Raphaël Flambard
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)

Secrétariat :

Bibliothèque :

Signature du responsable entreprise

Signature du candidat

LB

Résumé

L'assurance annulation est un des risques majeurs de l'assurance voyage. Cette garantie offre une protection financière en cas d'événements imprévus empêchant le voyage. Depuis la pandémie du Covid-19, le volume d'activité s'est intensifié, entraînant une croissance rapide du marché de l'assurance voyage. Par conséquent, les compagnies d'assurance doivent ajuster leur suivi de rentabilité.

L'expansion de ce marché nécessite un pilotage précis de la sinistralité des garanties annulation. Dans ce contexte, ce mémoire cherche à prédire la sinistralité de la garantie annulation à partir des données mensuelles d'un partenaire commercial de Europ Assistance.

En premier lieu, il convient de comprendre les spécificités des polices annulation, principalement vendues en B2B2C par l'intermédiaire de partenaires comme des agences de voyage en ligne.

Ensuite, un Modèle Linéaire Généralisé (GLM) est utilisé pour prédire la sinistralité de la garantie annulation. Les variables utilisées dans ce modèle sont des variables dites de souscription, disponibles dès l'achat d'une police d'assurance. Le GLM, apprécié pour sa transparence et son interprétabilité, permet de prédire l'*Incurred in Percentage of Trip Cost*, charge de sinistres en fonction du coût du voyage, pour chaque contrat. Ce modèle performant permet d'intégrer les résultats dans les outils de pilotage et de suivi de rentabilité.

Enfin, un modèle de *Gradient Boosting* (GBM) est mis en place pour prédire la charge de sinistres sur les mois de départ à venir, à l'aide d'informations sur la sinistralité déjà développée. L'étude de ce modèle, conçu à l'aide de variables de souscription et de sinistralité, montre que la combinaison de ce type de variables n'est pas pertinente. La charge de sinistres prédite pour les mois de départ futurs dépend de la sinistralité développée au moment où est réalisée la prédiction.

Mots-clés : *Assurance annulation*, *Assurance voyage*, *couverture temporaire*, *Gradient Boosting*, *Prédiction*, *Sinistralité*.

Abstract

Cancellation insurance is one of the major risks in travel insurance. This coverage offers financial protection in the event of unforeseen events preventing travel. Since the Covid-19 pandemic, the volume of business has intensified, leading to rapid growth in the travel insurance market. As a result, insurance companies are having to adjust their profitability monitoring.

The expansion of this market requires precise management of the claims experience for cancellation cover. Against this backdrop, this dissertation seeks to predict the claims experience of cancellation cover based on monthly data from a Europ Assistance business partner.

First, it is necessary to understand the specific features of cancellation policies, which are mainly sold on a B2B2C basis via partners such as online travel agencies.

Then, a Generalized Linear Model (GLM) is used to predict the claim frequency for trip cancellation coverage. The variables used in this model are underwriting variables, available at the time of purchasing an insurance policy. The GLM, valued for its transparency and interpretability, is used to predict the *Incurred in Percentage of Trip Cost*, which represents the claim costs as a percentage of the trip cost, for each contract. This efficient model allows the results to be integrated into tools for monitoring and managing profitability.

Finally, a *Gradient Boosting Model* (GBM) is implemented to predict the claim costs for upcoming departure months using information on already-developed claims. The study of this model, built using underwriting and claims-related variables, shows that the combination of such variables is not relevant. The predicted claim costs for future departure months depend on the developed claims at the time the prediction is made.

Keywords : *Cancellation Insurance*, *GLM*, *Travel Insurrance*, *Gradient Boosting*, *Prediction*.

Note de Synthèse

Contexte de l'étude

L'objectif de cette étude est de concevoir un modèle permettant de prédire la charge de sinistres en assurance annulation. Dans un contexte de croissance du marché de l'assurance voyage, il devient nécessaire d'utiliser des outils de pilotage performants. Cette prédiction vise donc à améliorer la tarification et le suivi de la rentabilité d'un partenaire commercial spécifique de Europ Assistance. À ce titre, l'objectif de cette étude est de répondre à la question, "comment prédire la sinistralité des garanties annulation pour un compte spécifique ?". Pour répondre à la question, deux méthodes distinctes de modélisation sont employées, pour estimer la sinistralité en cours de développement des contrats, vendus par le biais du partenaire commercial de Europ Assistance. Ce partenaire commercial distribue ses produits dans cinq pays de l'Union Européenne avec la France et l'Allemagne comme principaux marchés.

Au sein de cette étude, plusieurs notions de vocabulaire se dégagent. Elles permettent d'étudier le comportement des voyageurs en fonction des caractéristiques de leur contrat.

- *Booking window* correspond à l'intervalle de temps entre la date de souscription de la police d'assurance et la date de départ ;
- *Exposition time* indique la durée d'exposition au risque de l'assureur. Elle correspond à la durée de la *booking window* ;
- *Trip duration* indique la durée du voyage ;
- *Cancellation window* indique le temps entre la date de départ et la date d'annulation du voyage.

Ces différentes notions deviennent des variables de modélisation. Elles permettent de discriminer les données afin de réaliser des prédictions de la charge de sinistres. La figure 1 représente la situation de l'étude.

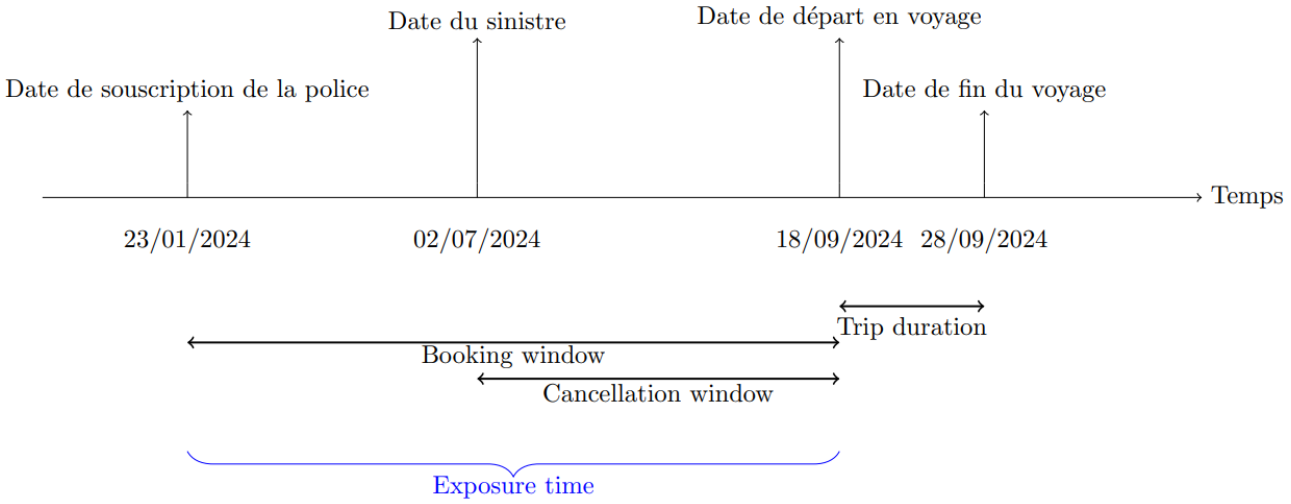

FIGURE 1 : Illustration d'un schéma simplifié d'annulation

Il convient de noter que le partenaire commercial peut mettre en place un barème d'annulation à destination du voyageur. Dans ce cas, lorsque le voyageur annule son voyage, un pourcentage de pénalité est appliqué sur le coût du voyage. Cette pénalité est inversement proportionnelle à la *cancellation window*. Il s'agit de la perte financière du voyageur. Lorsque celui-ci est assuré, l'assureur prend en charge la pénalité instaurée par le partenaire commercial. La charge de sinistre dépend alors de la pénalité mise en place par le partenaire commercial.

Afin d'estimer la sinistralité de ce compte, il convient de prédire la charge de sinistres. Celle-ci est prédite en pourcentage du coût du voyage. L'*Incurred in Percentage of Trip Cost*, notée IPTC, est un taux de sinistres en fonction du montant total du coût du voyage. Il s'agit d'exprimer la charge de sinistres en fonction du montant d'exposition au risque. L'IPTC correspond à la prime pure en pourcentage du coût du voyage et indique le pourcentage moyen de remboursement réalisé par Europ Assistance. Cette variable est donnée par l'expression

$$IPTC = \frac{\text{Charge de sinistre}}{\textit{Trip cost}}, \tag{1}$$

et suit les tendances de la fréquence des sinistres du portefeuille. L'analyse de sa dérive permet de piloter de manière efficiente la profitabilité du produit. Elle prend en compte les pénalités mises en place par le partenaire commercial. Par conséquent, c'est une variable stable qui permet des comparaisons spatio-temporelles de la charge de sinistres. Il convient de pas confondre l'IPTC avec le coût moyen qui correspond au pourcentage du coût du voyage moyen remboursé par EA en cas de sinistre.

Dans cette étude, les deux méthodes de modélisation mises en œuvre distinguent deux types de variables. Les variables de souscription (disponibles lors de la souscription d'une police d'assurance) et les variables de sinistralité (porteuses d'information sur la sinistralité passée). Ces modèles sont conçus dans un principe de continuité avec les hypothèses de modélisation appliquées par Europ Assistance.

Prédiction de la charge de sinistres avec des variables de souscription

Un premier modèle est consacré à la construction d'un modèle linéaire généralisé (GLM). Ce modèle a pour objectif d'analyser l'exposition au risque de la compagnie au moment de la souscription par l'assuré d'un contrat d'assurance. Il est conçu uniquement à partir des variables disponibles à la

souscription d'un contrat d'assurance. Ces variables discriminantes concernent le mois de départ, la durée du voyage, l'intervalle de temps entre la réservation et le départ, appelée *booking window* ainsi que le coût du voyage.

À titre d'exemple, le mois de départ permet de discriminer les contrats dont la sinistralité est stabilisée versus ceux dont la sinistralité est en cours de développement. Cette distinction permet d'utiliser des modèles de régression afin d'estimer cette sinistralité. La méthodologie de projection est la suivante. Le graphique 2 permet de comprendre comment est prise en compte la sinistralité. L'unité de l'axe temporel du graphique est le mois de départ. Ce schéma permet d'illustrer le raisonnement appliqué lorsque la date de départ a lieu plus de 3 mois avant la date de vision. Dans cet exemple, la date de départ des voyageurs 1 et 2 (septembre 2023) a lieu 3 mois avant la date de vision (31 décembre 2023). Pour ces voyages, la sinistralité réelle est prise en compte et la charge de sinistres associée à ces contrats est considérée comme stable. Ainsi, il n'est pas nécessaire de faire des projections pour les voyageurs 1 et 2. Dans ce cas, les données observées sont prises en compte dans la modélisation. Pour les contrats dont le mois de départ se situe dans l'intervalle de temps jusqu'à 3 mois avant la date de vision, un *pro rata* est appliqué entre la sinistralité réelle et projetée. Tandis que pour le voyageur 3 dont le départ se situe après la date de vision sur la frise chronologique, il faut donc appliquer la sinistralité moyenne constatée sur un historique de 12 mois de départ, une fois les sinistres stabilisés, soit les mois de novembre 2022 à octobre 2023 ($m-3$ par rapport à la date de départ). Cette méthodologie est décrite selon les modèles développés dans ce mémoire.

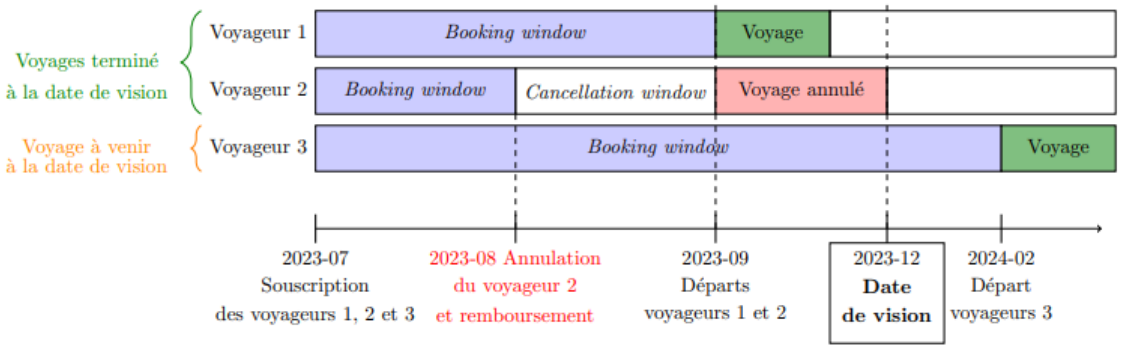

FIGURE 2 : Schéma de la méthodologie de projection

Une analyse statistiques des données permet d'abord de comprendre comment ces facteurs influencent la probabilité de survenance d'un sinistre. Les conclusions de cette analyse ont permis de construire un modèle GLM pertinent. Ce modèle, par son approche paramétrique, produit des résultats satisfaisants en termes de performance prédictive, pour les contrats dont la sinistralité n'est pas encore développée. Ainsi, il rend possible la mise en place d'un modèle dit de "tarification" adéquat qui s'incorpore dans les outils déjà utilisés pour le suivi de ce compte.

L'enjeu est alors de tarifier le contrat d'assurance annulation à l'aide d'une régression GLM avec une distribution Tweedie. Cette loi statistique permet de modéliser une prime pure grâce à la présence d'une masse en zéro et des valeurs positives sur le reste de la distribution. Le modèle permet de prédire l'IPTC à partir du mois de départ, du pays d'achat de la police d'assurance, de la durée du voyage, de la *booking window* et de la tranche du coût du voyage. Pour chaque contrat, la prime pure en pourcentage du coût du voyage est calculée en utilisant les coefficients estimés du modèle GLM. Cette prédiction pour chaque ligne du contrat permet d'obtenir, par agrégation, la charge de sinistres estimée pour chaque mois de départ.

Les résultats obtenus permettent de comparer les charges de sinistres prédites par le modèle GLM,

par le modèle actuel et la charge réellement observée.

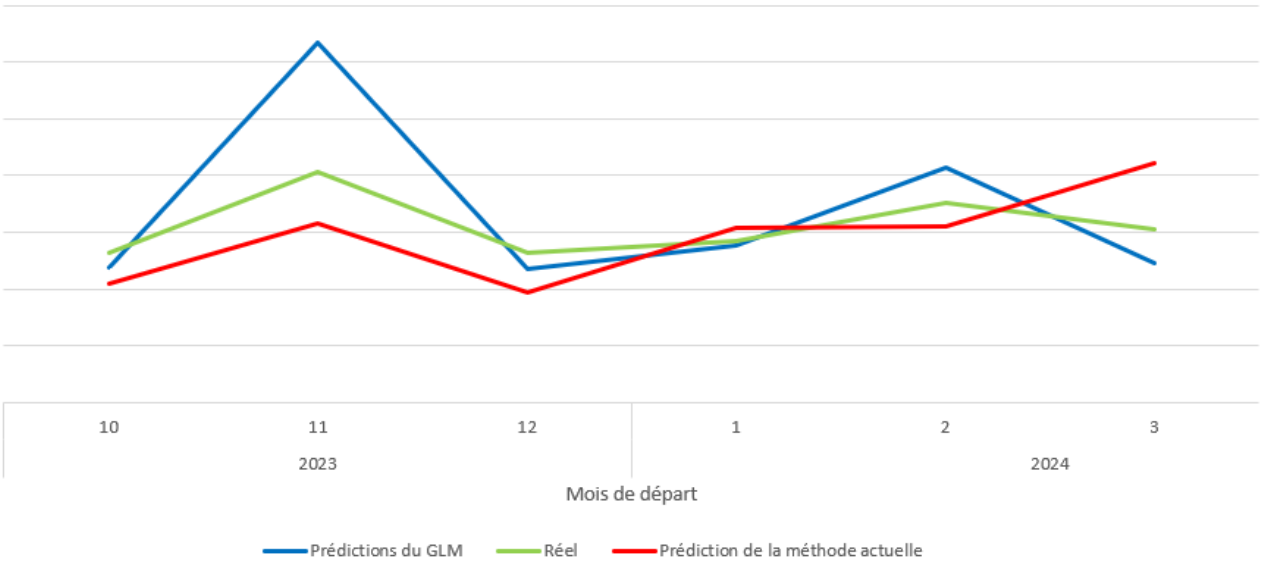

FIGURE 3 : Comparaison des prédictions de charge de sinistres en fonction des mois de départ, à la date de vision décembre 2023

Le graphique 3 permet de comparer les charges de sinistres prédite par la méthode actuelle et celle prédite par la méthode GLM. Il met en avant que le modèle GLM (courbe bleue) suit mieux les tendances imposées par les données réelles (courbe verte) que la méthode actuelle (courbe rouge). Le modèle actuel réalise des prédictions grâce à des calculs de moyennes des variables intéressées sur les douze derniers mois glissants. L'objectif est alors de réaliser de meilleures prédictions.

L'analyse graphique permet de mettre en avant que les prédictions du GLM et du modèle actuel sont relativement proches. Néanmoins, le modèle GLM possède la qualité de pouvoir facilement s'adapter à un changement de comportement des voyageurs, lorsque la structure de ses données d'entrée se modifient. Il faut remarquer aussi que le GLM capte mieux les tendances que le modèle actuel. À titre d'exemple, l'analyse de sinistralité selon les pays de souscription montre que celle-ci est très variable (à l'instar de l'écart entre l'Allemagne et la Suisse). Dans le cas où, la répartition du portefeuille se déséquilibre entre les pays, le modèle actuel s'adapte très lentement à cette modification, puisqu'il réalise seulement un calcul de moyenne. Alors que le modèle GLM prend en compte cette modification dès la souscription du contrat.

Prédiction de la charge de sinistres avec des variables de sinistralité

Un second type de modèle est étudié en prenant en compte des variables de sinistralité observée. Pour ce faire, un modèle plus sophistiqué est utilisé, le *Gradient Boosting*, noté GBM. L'objectif de ce modèle est de prendre en compte des variables porteuses d'informations sur la sinistralité passée. Il s'agit alors de détecter le niveau de stabilisation des sinistres et d'adapter les prédictions en fonction de cette observation et des montants déjà payés.

Les variables porteuses d'information sur la sinistralité permettent de suivre le développement de la charge de sinistres. La variable cible devient alors l'IPTC ultime, c'est à dire, l'IPTC une fois que la charge de sinistres est entièrement développée. En considérant des souscriptions récentes, qui ont donc déjà eu lieu, certains contrats ont une sinistralité développée tandis que pour d'autres, elle n'est

pas (entièrement) développée, comme le montre le graphique 4. Ce graphique montre au 31 décembre 2023 la charge de sinistres observée grâce à la courbe bleue, variable notée dans la modélisation IPTC observée (`IPTC_obs`). La courbe rouge indique cette charge à la vision du 31 juillet 2024, variable notée dans la modélisation IPTC ultime (`IPTC_ult`) . L'écart entre les deux reflète le développement de la sinistralité sur les mois de départ de janvier à juillet 2024.

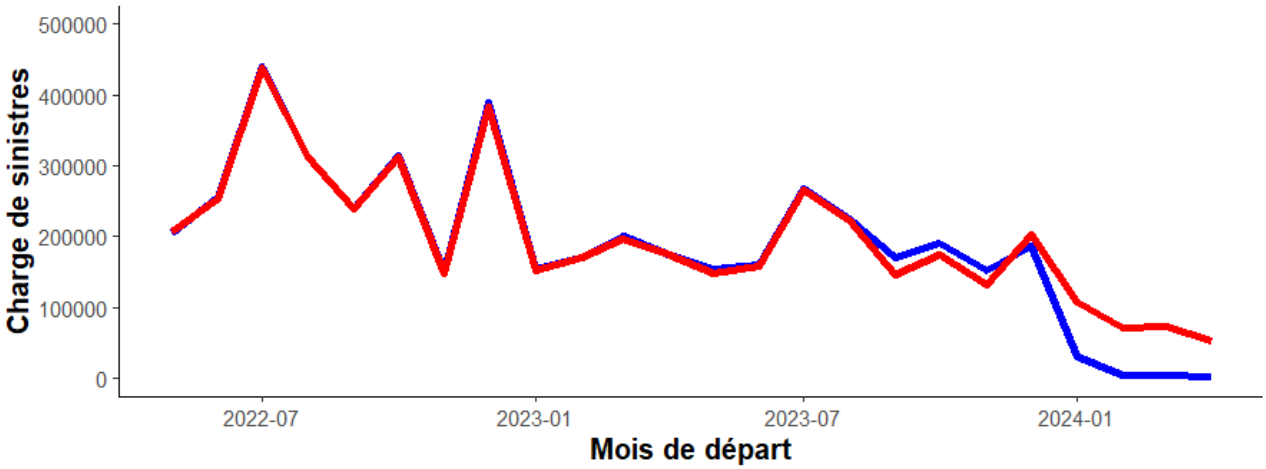

FIGURE 4 : Comparaison de la charge de sinistres observée au 31 décembre 2023 et développée au 31 juillet 2024

Dans ce contexte, il convient de différencier la sinistralité stabilisée de la sinistralité à estimer. La décomposition de la sinistralité d'après ce critère se réalise grâce aux données de mois de départ. Cette variable permet de discriminer de la meilleure façon le développement de la charge de sinistres.

L'algorithme mis en place est un algorithme de *Gradient Boosting Machine* (GBM). Le GBM fonctionne en construisant à chaque étape un arbre de décision qui cherche à minimiser l'erreur de prédiction du modèle précédent grâce aux résidus. Il s'agit pour chaque modèle intermédiaire, appelé modèle faible, de prédire la variable cible (IPTC ultime) en ajustant les résidus du modèle précédent. Cette mise à jour est pondérée par un facteur de taux d'apprentissage η comme indiqué dans l'équation 2. Soit $\mathbf{x}$, le vecteur d'informations disponibles dans le jeu de données. Pour l'arbre m, la relation est donnée par

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \eta \cdot h_m(\mathbf{x}). \tag{2}$$

Où $f_m(\mathbf{x})$ est le modèle final après l'ajout de l'arbre m.
$f_{m-1}(\mathbf{x})$ est le modèle à l'itération précédente.
η est le taux d'apprentissage.
$h_m(\mathbf{x})$ est l'arbre à l'étape m.

Le taux d'apprentissage η contrôle la contribution de chaque arbre au modèle final. Lorsqu'il est trop élevé, il peut conduire à un surapprentissage du modèle, aussi appelée *overfitting*. Tandis que lorsqu'il est trop faible, le modèle peut nécessiter un trop grand nombre d'itérations pour converger. La détermination des hyperparamètres de ce modèle est reéalisée par grille de recherche.

Le modèle final, après M itérations, est la somme pondérée des modèles faibles construits au cours

des itérations. Il est donné par l'expression

$$f_M(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{m=1}^{M} \eta \cdot h_m(\mathbf{x}). \tag{3}$$

Dans cette étude, la méthode `XGBoost` est utilisée en raison de ses performances élevées et de sa flexibilité. C'est une version optimisée, régulière et performante du GBM classique. Ainsi, la fonction h_m est la fonction qui permet l'apprentissage du modèle. Elle définit un arbre de régression entraîné sur les résidus obtenus par le modèle faible précédent. Elle dépend, par conséquent, du modèle utilisé et est donc spécifique au modèle `XGBoost`. Elle se définit par l'expression qui cherche à minimiser la fonction de perte $\mathcal{L}$, avec une régularisation quadratique sur h_m, multiplié par λ, le paramètre de pénalité. La régularisation est destinée à limiter la complexité de la fonction h_m. Il faut noter que le terme γT permet de gérer l'élagage de l'arbre et est expliqué ultérieurement.

$$h_m = \arg \min_{h \in \mathcal{F}(\mathbb{R} \to \mathbb{R})} \sum_{i=1}^{n} \left(\mathcal{L}\left(r_i, h(x_i)\right) + \frac{1}{2} \lambda h(x_i)^2 \right) + \gamma T. \tag{4}$$

Ce modèle nécessite un traitement des données important comme l'encodage et le traitement des données manquantes.

Cependant, il apparaît que le comportement du modèle n'est pas celui espéré puisqu'il ne permet pas de distinguer l'état de développement du sinistre. En outre, la charge prédite dépend de manière trop importante de la charge de sinistres observée à la date de vision. Ainsi, la qualité des prédictions ne s'améliore pas avec l'introduction de ces informations. Les données de sinistralité semblent plutôt apporter de la confusion en demandant au modèle de prédire des valeurs très différentes de la variable cible. En outre, le caractère communément appelé "boîte noire" de ce modèle de *machine learning* est un frein à l'utilisation de ce modèle par Europ Assistance. Cette partie de l'étude permet néanmoins d'explorer cette nouvelle approche de modélisation pour Europ Assistance. Elle constitue alors une première ébauche destinée à être approfondie en fonction des besoins de l'entreprise.

Puisqu'à ce stade le modèle GBM mis en œuvre dans cette étude ne permet pas de détecter la stabilisation de la charge de sinistre via une variable discriminante (distance entre le mois de départ et la date d'extraction des sinistres), une solution alternative est alors choisie pour intégrer le processus actuel de suivi de rentabilité. En combinant la modélisation GLM avec la modélisation actuelle, qui sépare manuellement les sinistres selon l'état de développement de leur charge, le modèle obtenu est pertinent. Dans cette modélisation, Europ Assistance a choisi de retenir l'utilisation du GLM pour prédire la sinistralité des contrats dont la sinistralité n'est pas développée. Tandis que pour les contrats ayant une sinistralité développée, il s'agit d'appliquer la méthode actuelle.

Conclusion de l'étude

Pour améliorer le suivi de la rentabilité de ce partenaire commercial, deux types de modèles ont été conçus. Le modèle GLM, en se combinant à la méthode actuelle, permet d'améliorer significativement les estimations réalisées mensuellement, en se basant sur les informations de souscription. L'ajout des informations de sinistralité, présentées dans le second type de modèle, ne permet pas au modèle de détecter la stabilisation de la sinistralité et n'apporte pas d'amélioration des résultats de prédiction. Afin que cette étude propose une solution concrète pour le suivi de la tarification, la charge de sinistres est prédite par un modèle GLM. Ainsi, la rentabilité de ce contrat d'assurance annulation est pilotée en respectant les contraintes imposées par la modélisation actuellement utilisée. Elle peut alors être utilisée pour suivre la rentabilité des autres partenaires commerciaux de Europ Assistance.

Synthesis note

Context of the study

The objective of this study is to design a model to predict claim costs for cancellation insurance. In a growing travel insurance market, efficient management tools are essential. This prediction aims to enhance pricing and profitability monitoring for a specific commercial partner of Europ Assistance. As such, the study seeks to answer the question, "How can cancellation insurance claims for a specific account be predicted?"

To address this question, two distinct modeling methods are employed to estimate claims still under development for contracts sold through Europ Assistance's commercial partner. This partner sells its products in five European Union countries, with the French and German markets being the most significant.

Several key concepts emerge in this study, facilitating the analysis of traveler behavior based on contract characteristics:

- Booking window: The time interval between the insurance policy subscription date and the departure date.
- Exposure time: The insurer's risk exposure duration, equivalent to the booking window duration.
- Trip duration: The length of the journey.
- Cancellation window: The time between the trip departure date and its cancellation date.

These concepts serve as modeling variables, enabling the data to be segmented for claim cost predictions. Figure 5 illustrates the study's context.

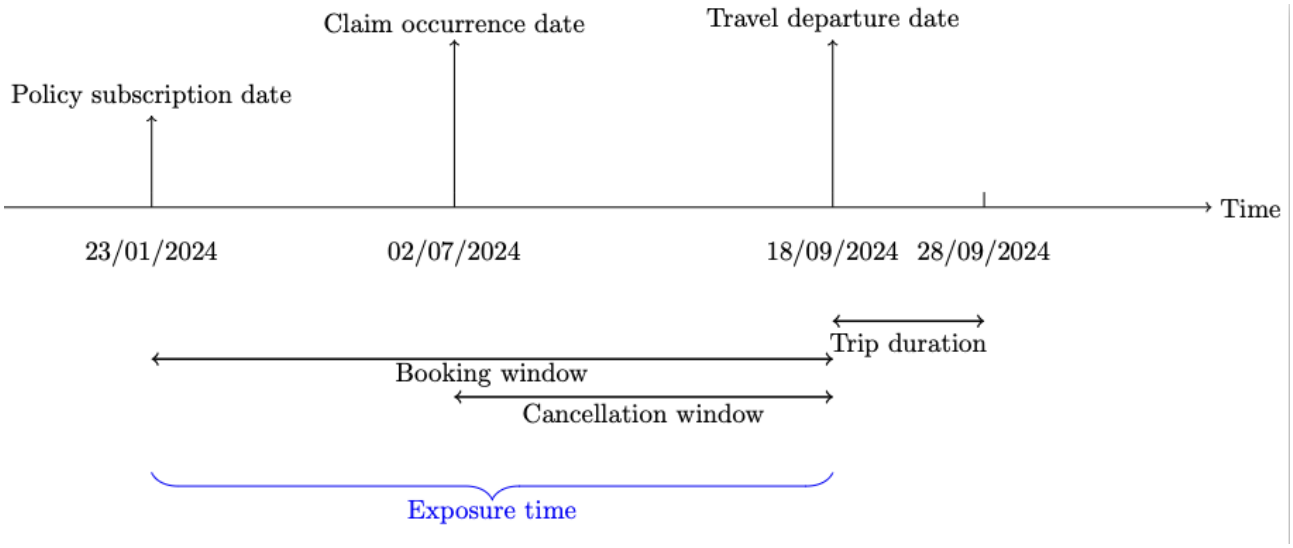

Figure 5: Illustration of a simplified annulation scheme

It should be noted that the commercial partner may implement a cancellation fee schedule for travelers. If a traveler cancels their trip, a percentage penalty is applied to the trip cost, inversely proportional to the cancellation window. This represents the traveler's financial loss. When the traveler is insured, the insurer covers the penalty imposed by the partner. Consequently, the claim cost depends on the penalty scheme established by the commercial partner.

To estimate claims for this account, the study predicts claim costs as a percentage of the trip cost. The Incurred in Percentage of Trip Cost (IPTC) is a claims ratio relative to the total trip cost. It expresses claim costs as a percentage of risk exposure and represents the pure premium percentage of the trip cost, indicating Europ Assistance's average reimbursement rate. The IPTC is defined as:

$$IPTC = \frac{\text{Claim Cost}}{\text{Trip Cost}}, \tag{5}$$

and it follows the trends of portfolio claim frequency. Analyzing its variations helps manage the product's profitability effectively. This variable accounts for penalties imposed by the commercial partner, making it stable and suitable for spatiotemporal comparisons. It is important not to confuse IPTC with the average cost, which represents the average percentage of the trip cost reimbursed by Europ Assistance in case of a claim.

In this study, the two modeling methods differentiate between two types of variables: subscription variables (available when the insurance policy is purchased) and claims variables (providing information on past claims). These models are designed in continuity with the modeling assumptions applied by Europ Assistance.

Claim cost prediction using subscription variables

The first model focuses on constructing a Generalized Linear Model (GLM). This model aims to analyze the company's risk exposure at the time of policy subscription by the insured. It is based solely on variables available at the time of insurance policy subscription and incorporates several discriminating variables, such as the month of departure, trip duration, the interval between booking and departure (referred to as the booking window), and trip cost.

For instance, the month of departure differentiates between contracts with stabilized claims versus those with claims still under development. This distinction allows for the use of regression models to estimate these claims. The projection methodology is outlined below.

Figure 6 illustrates how claims are accounted for. The time axis unit is the month of departure. This diagram demonstrates the reasoning applied when the departure date occurs more than three months before the observation date. In this example, the departure dates for travelers 1 and 2 (September 2023) occur three months before the observation date (December 31, 2023). For these trips, actual claims are considered stable, and no projections are required. These observed data are included in the modeling process.

For contracts with departure dates within three months of the observation date, a *pro rata* is applied between actual and projected claims. For traveler 3, whose departure occurs after the observation date on the timeline, average claims observed over a historical period of 12 months (November 2022 to October 2023) are used. This methodology aligns with the models developed in this study.

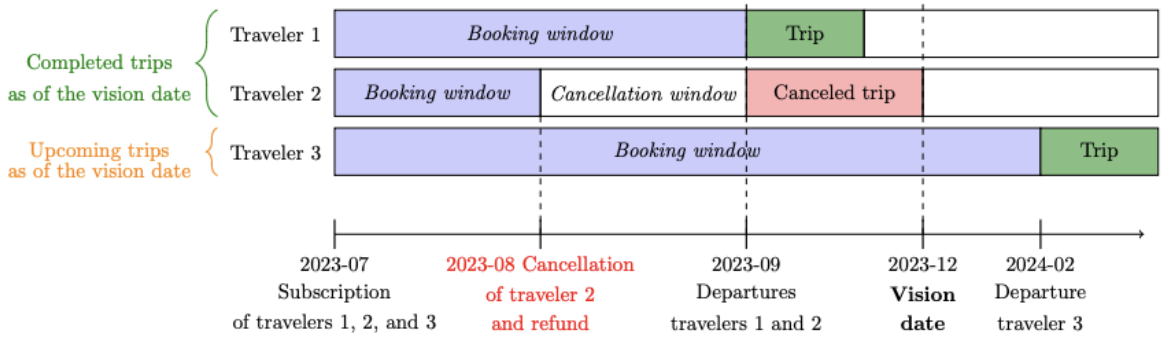

Figure 6: Projection methodology diagram

A statistical analysis of the data helps understand how these factors influence the probability of a claim. The conclusions of this analysis led to the construction of a relevant GLM. This parametric approach delivers satisfactory predictive performance for contracts with undeveloped claims, enabling the implementation of an adequate "pricing" model that integrates into existing tools used to monitor this account.

The goal is to price the cancellation insurance contract using a GLM regression with a Tweedie distribution. This statistical distribution models pure premium with a zero mass and positive values for the remainder of the distribution. The model predicts IPTC using the month of departure, the country of purchase, trip duration, booking window, and trip cost range. For each contract, the pure premium as a percentage of the trip cost is calculated using the estimated coefficients from the GLM. This prediction for each contract line allows for aggregated claim cost estimates for each departure month.

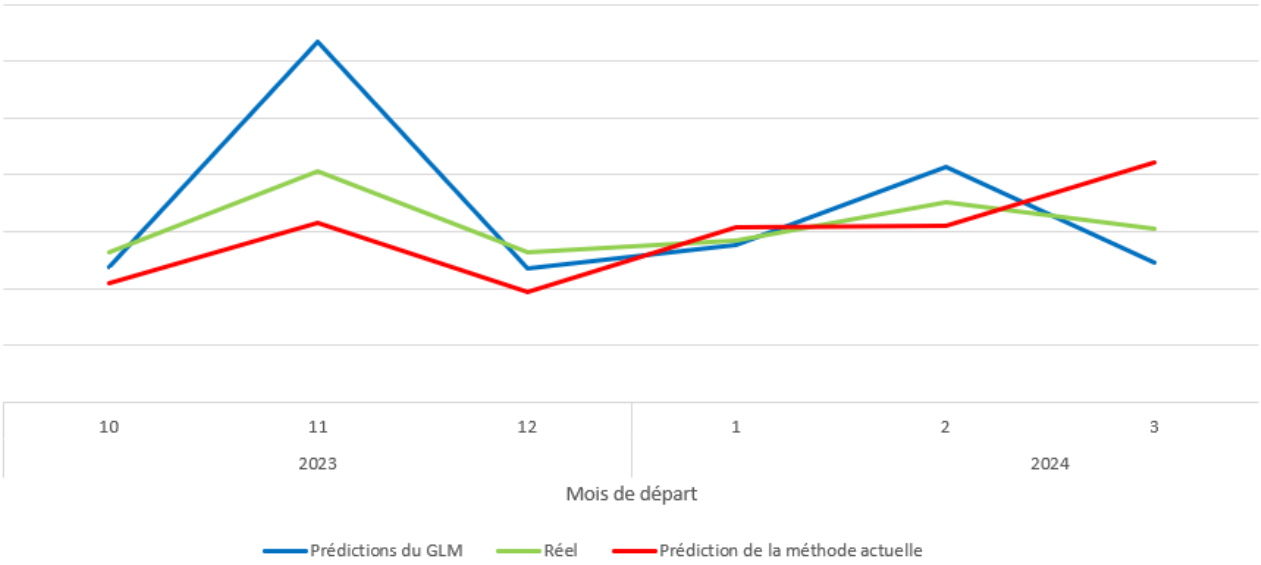

Figure 7: Comparison of predicted claim costs by departure onth as of December 2023

Figure 7 compares claim costs predicted by the current method and the GLM method. The GLM model (blue curve) better aligns with real data trends (green curve) than the current method (red curve).

The graphical analysis highlights that GLM and the current model predictions are relatively close. However, the GLM has the advantage of adapting to changes in traveler behavior when the structure of input data changes. For example, claims analysis by the country of subscription shows significant variability (e.g., differences between Germany and Switzerland). If the portfolio distribution shifts between countries, the current model's simple averaging adapts slowly, whereas the GLM accounts for this change at the time of subscription.

Claim cost prediction using claims variables

The second model type incorporates observed claims variables. A more sophisticated gradient boosting model is used to account for past claims information. This approach aims to detect claim stabilization levels and adjust predictions accordingly.

The variables with claims information monitor the development of claim costs. The target variable becomes the ultimate IPTC, representing the IPTC once claims are fully developed. Figure 8 illustrates the observed claims as of December 31, 2023 (blue curve) and their development as of July 31, 2024 (red curve). The difference reflects the claims development for January to July 2024 departures.

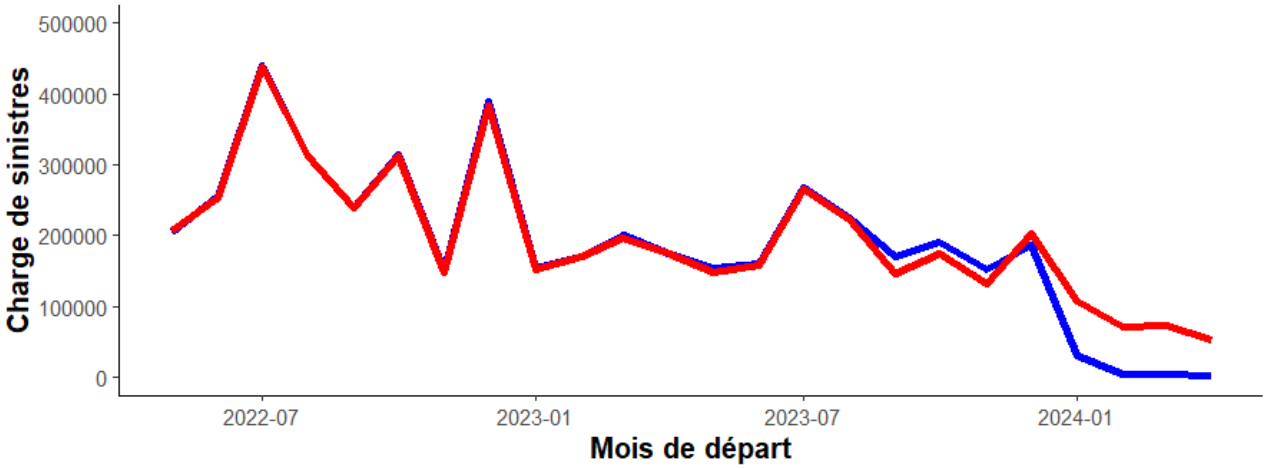

Figure 8: Comparison of observed claims as of December 31, 2023, and fully developed claims as of July 31, 2024

In this context, stabilized claims are distinguished from claims requiring estimation. This distinction uses the departure month variable to best differentiate claim development.

The gradient boosting algorithm (GBM) builds at each step a decision tree that minimizes the previous model's prediction error. This iterative process updates the model with residuals and is weighted by a learning rate η as shown in Equation 6. The final model after M iterations is the weighted sum of intermediate models

$$f_M(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{m=1}^{M} \eta \cdot h_m(\mathbf{x}). \tag{6}$$

In this study, the `XGBoost` method is used due to its high performance and flexibility. It is an optimized, regularized, and efficient version of the classic GBM. Thus, the function h_m represents the function that enables the model's learning process. It defines a regression tree trained on the residuals obtained from the previous weak model. Consequently, it depends on the specific model being used and is therefore specific to the `XGBoost` model. This function is defined by the expression that seeks to minimize the loss function $\mathcal{L}$, with quadratic regularization on h_m, multiplied by λ, the penalty parameter. The regularization aims to limit the complexity of the function h_m. Note that the term γT handles tree pruning and is explained later.

$$h_m = \arg \min_{h \in \mathcal{F}(\mathbb{R} \to \mathbb{R})} \sum_{i=1}^{n} \left(\mathcal{L}\left(r_i, h(x_i)\right) + \frac{1}{2} \lambda h(x_i)^2 \right) + \gamma T. \tag{7}$$

This model requires extensive data processing, such as encoding and handling missing data.

However, it appears that the model's behavior is not as expected, as it does not allow for distinguishing the development state of claims. Furthermore, the predicted costs rely too heavily on the observed claims at the prediction date. As a result, the prediction quality does not improve with the inclusion of this information. Claims data seem to introduce confusion by requiring the model to predict values significantly different from the target variable. Moreover, the "black-box" nature commonly associated with this machine learning model hinders its adoption by Europ Assistance. Nevertheless, this part

of the study provides an opportunity to explore this new modeling approach for Europ Assistance. It serves as an initial draft to be further refined based on the company's needs.

Since, at this stage, the GBM model implemented in this study does not detect the stabilization of claim costs,

Europ Assistance has chosen to retain the use of the GLM to predict claims for policies with undeveloped claims experience. For policies with developed claims experience, the current method will continue to be applied.

Conclusion of the study

To improve profitability monitoring for this commercial partner, two types of models were developed. The GLM model, combined with the current method, significantly improves monthly estimations using subscription information. Adding claims information through the second model does not enhance prediction quality due to its inability to detect claims stabilization.

As a concrete solution for pricing monitoring, claim costs are predicted using the GLM model. Thus, the profitability of this cancellation insurance contract is managed while respecting the constraints imposed by the current modeling process. This approach can also be applied to monitor the profitability of other Europ Assistance commercial partners.

Remerciements

Tout d'abord, je tiens à remercier Raphaël Flambard pour sa bienveillance et son accompagnement durant ces six mois de stage. Sa pédagogie, sa patience et sa passion de `R` m'ont permis de progresser tout au long de cette période. Un grand merci à toute l'équipe *Underwritting & Pricing* pour leur accueil chaleureux.

Je tiens à remercier chaleureusement tous ceux qui ont participé à la relecture de ce mémoire et en particulier Asmae, Martine, Benjamin et Tuan.

Par ailleurs, je tiens à remercier vivement toutes les personnes que j'ai croisé au cours de mes précédents stages et en particulier aux ACM pour leur chaleureux accueil en toutes circonstances. Les échanges bénéfiques m'ont permis de m'épanouir dans le domaine de l'actuariat.

Mes remerciements s'adressent à Martine Carré-Tallon et Quentin Guibert pour leur confiance et leur disponibilité tout au long de mon parcours à l'Université Paris-Dauphine. En outre, mes remerciements s'adressent aux différentes équipes pédagogiques qui ont rendu possible ce parcours atypique en transmettant leur passion avec bienveillance. À ce titre, Sylvia Dobyinsky, responsable du parcours MIASHS à Nanterre, Kate Boillot Patterson et Adrien Suru. Je tiens également à remercier Stéphanie Incorvaia, Judith Ntsame et Corinne Virique pour leur aide précieuse.

En outre, mes remerciements vont à mes parents, ma famille et mes amis pour leur soutien indéfectibles tout au long de ces sept années. Un grand merci à tous ceux qui ont fait semblants de s'intéresser à ce mémoire en ouvrant au moins une fois le PDF quotidien sur vos conversation *WhatsApp*. Une dédicace particulière au BDS pour notre stage aux ACM qui fut un moteur pour les futurs actuaires que nous serons. Enfin, ces remerciements ne peuvent se conclure sans citer un groupe dont le nom est trop long pour le soutien tant scolaire que moral qu'il a été pendant cette année.

Table des matières

Introduction

L'assurance annulation est l'un des risques majeurs de l'assurance voyage en termes de volumes de primes. Comme le souligne Erick Morazin, anciennement *Senior Vice-President Global Travel* chez AXA Partners (2023), après la pandémie du Covid-19 "la sécurité sanitaire, la sûreté du voyageur et de sa famille sont devenues des préoccupations essentielles" permettant un essor fulgurant du marché de l'assurance voyage et en particulier de la garantie annulation. Dans ce contexte, la garantie annulation est devenue un élément crucial pour les voyageurs, offrant une protection financière en cas d'événements imprévus les empêchant de partir en voyage. En conséquence, les compagnies d'assurance ont pu développer des produits adaptés aux nouveaux besoins des voyageurs et de leurs partenaires commerciaux dans le cadre d'un schéma de distribution de type B2B2C. Ainsi, l'assurance annulation est devenue un pilier essentiel de l'industrie de l'assurance voyage, afin de se prémunir du risque lié aux événements géopolitiques, climatiques ou encore sanitaires incertains.

L'expansion de ce marché génère des besoins spécifiques en matière de gestion de la rentabilité, notamment en ce qui concerne la sinistralité des garanties annulation. Ces garanties représentent une part importante du volume des primes en assurance voyage, nécessitant un pilotage efficace. C'est dans ce contexte que cette étude s'inscrit, avec pour objectif de répondre à la question suivante : **comment prédire la sinistralité des garanties annulation pour un compte spécifique ?** En utilisant les données mensuelles de la garantie annulation fournies par un partenaire commercial de Europ Assistance, ce mémoire vise à approfondir la compréhension du risque annulation, un sujet encore peu exploré par les actuaires.

Afin de maîtriser les enjeux autour de l'assurance annulation, ce mémoire commence par une présentation de Europ Assistance et un lexique auquel peut se référer le lecteur pour connaître la signification d'un sigle donné.

La modélisation de la garantie annulation est particulière par bien des aspects. Dans un premier temps, il s'agit de comprendre les spécificités des polices annulation. Cette garantie est principalement vendue en schéma B2B2C : l'assureur vend son produit d'assurance par l'intermédiaire d'un partenaire (agence de voyages en ligne, croisiériste, ...) à un voyageur. La garantie annulation est dite temporaire, ce qui signifie que la durée d'exposition au risque de l'assureur varie en fonction de l'intervalle de temps entre la date de souscription et la date de départ de la police. Ensuite, il s'agit de comprendre comment s'interprète et se calcule la sinistralité pour un compte d'assurance annulation. Le calcul et l'estimation de celle-ci a pour objectif de piloter de manière optimale la profitabilité du compte associé. Afin de réaliser cet objectif, des données mensuelles provenant du centre de gestion des sinistres de Europ Assistance sont utilisées.

Dans un deuxième temps, un modèle de tarification est conçu, en utilisant des variables dites de souscription. Ces données sont disponibles à la souscription d'un contrat d'assurance. Afin de prédire la sinistralité de la garantie annulation associée à ce compte, un Modèle Linéaire Généralisé, noté GLM (*Generalized Linear Model*), est utilisé. En utilisant une régression de Tweedie, cette modélisation permet de réaliser des calculs de prime pure pour chaque police d'assurance grâce aux données de

souscription. Dans cette étude, la prime pure s'obtient en pourcentage du coût du voyage. La variable cible est notée IPTC pour *Incurred in Percentage of Trip Cost* et représente la charge de sinistres en fonction du coût du voyage. Cette méthode de modélisation est fiable et pertinente. En outre, les prédictions réalisées sont vérifiées grâce à une méthodologie de *backtesting* avec les données de Europ Assistance. Les prédictions de primes pures sont ensuite intégrées dans les outils de pilotage afin de faciliter les prises de décisions stratégiques (tarification, participation aux bénéfices, ...) concernant le compte de ce partenaire commercial.

Enfin, dans un dernier temps, un modèle permettant de suivre la rentabilité du compte est mis en place en complément du modèle développé dans le deuxième chapitre. Le *Gradient Boostig Model*, noté GBM, a pour objectif, à partir des données de souscription et de sinistralité, de prédire le pourcentage de sinistres ultimes en fonction du coût du voyage total par mois. Dans ce chapitre, les données disponibles sont à la fois des variables de souscription et des variables porteuses d'information sur la sinistralité. Ainsi, la *cancellation window**, l'IPTC observé sont des informations utilisées pour prédire la charge de sinistres à l'ultime. Afin de distinguer les contrats ayant une sinistralité en cours de développement ou développée, il s'agit d'utiliser une variable discriminante. Dans le cadre de l'assurance annulation, la variable donnant l'information du mois de départ permet de statuer sur l'état de développement du sinistre. Il est à noter qu'une étude menée précédemment en interne permet de considérer que les sinistres sont développés deux mois après leur départ. À ce titre, une variable indiquant la stabilité du sinistre est prise en compte par le modèle. L'objectif de cette modélisation est de pouvoir incorporer le modèle créé à l'outil de pilotage déjà mis en place. Les enjeux sont pluriels, le modèle doit être plus performant que l'actuel et doit pouvoir s'insérer de manière aisée dans le *dashboard* utilisé au quotidien par les équipes. Le modèle GBM est adapté pour saisir des liens complexes entre les différentes variables explicatives.

*Intervalle de temps entre la date d'annulation et la date de départ.

Présentation de Europ Assistance

Présentation générale de Europ Assistance

Europ Assistance (2019b), noté EA dans la suite de cette étude, a été fondée en 1963 à Paris par Pierre Desnos. L'inventeur de Europ Assistance a su anticiper le besoin croissant de services d'assistance pour les voyageurs, à une époque où les voyages internationaux deviennent de plus en plus courants. À l'origine, l'entreprise est spécialisée dans la protection des voyageurs à l'étranger, en offrant des services d'assistance en cas d'urgence médicale, de perte de documents ou d'autres incidents imprévus. Rapidement, Europ Assistance élargit son champ d'action pour inclure l'assistance routière et l'assurance voyage, devenant ainsi un pionnier dans plusieurs domaines clés de l'assistance. La réputation de EA est telle qu'elle est évoquée dans la culture populaire, comme dans le film Les Bronzés, réalisé par Patrice Leconte Leconte (1978), où Michel Blanc déclare : "Tu parles, je vais te montrer mes bras, j'ai été bouffé par les moustiques (...) je vais me faire rapatrier par Europ Assistance, ça va pas faire un pli tu vas voir".

En six décennies, Europ Assistance a évolué d'une *start-up* innovante à un leader mondial dans le domaine de l'assistance. Aujourd'hui, l'entreprise est présente dans 41 pays à travers le monde, grâce à un réseau de 41 centres d'assistance, qui opèrent 24 heures sur 24 pour répondre aux besoins de leurs clients. Europ Assistance emploie près de 8 000 personnes, lui permettant de répondre efficacement aux demandes des clients, qu'il s'agisse de services médicaux, de soutien en cas de panne automobile ou de gestion de situations d'urgence. Le réseau de prestataires de Europ Assistance est composé d'environ 750 000 prestataires de services, parmi lesquels des professionnels de santé, des mécaniciens et des experts en assistance. Ce réseau global lui permet de fournir une réponse rapide et efficace à ses clients, peu importe l'endroit où ils se trouvent.

Depuis 2001, Europ Assistance est une filiale détenue à 100% par le groupe Generali, l'un des leaders mondiaux de l'assurance. Cette relation a permis à Europ Assistance de bénéficier des ressources financières et de l'expertise de Generali, tout en lui permettant de rester fidèle à ses valeurs. Generali, qui opère dans plus de 50 pays et gère environ 618 milliards d'euros d'actifs, a intégré Europ Assistance dans son "Generali *Care Hub*", une entité dédiée à la fourniture de services de soins et d'assistance. Cette intégration a renforcé les capacités de Europ Assistance, en lui permettant de se positionner comme un acteur clé dans les services de soins et d'assistance au niveau mondial. Une version simplifiée de l'organisation de la structure est présentée dans la figure 9, réalisée par Europ Assistance SA (2023), dans son rapport sur la solvabilité et la situation financière (SFCR) 2023.

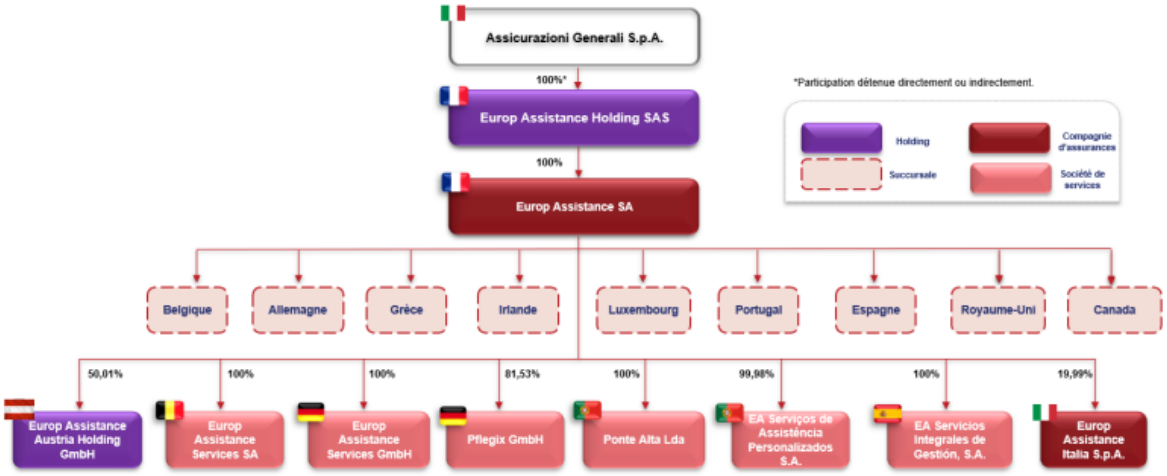

FIGURE 9 : Organigramme de Europ Assistance

Europ Assistance est une compagnie d'assurance présente dans neuf pays à l'aide de ses succursales, lui conférant alors un réseau mondial pour répondre aux besoins de ses clients.

En 2014, comme en témoigne L'ARGUS DE L'ASSURANCE (2023), Antoine Parisi a pris les rênes de l'entreprise en tant que Directeur Général. À son arrivée, il a initié le plan stratégique "*WeConnect*" en 2015, avec des objectifs ambitieux pour Europ Assistance. Le plan visait à transformer l'entreprise d'une collection d'entités indépendantes en un groupe cohérent, en mettant en avant les synergies internes et en alignant les stratégies à travers l'organisation. Il a également cherché à renforcer les liens avec le groupe Generali, en soutenant les initiatives stratégiques du groupe tout en différenciant Europ Assistance de ses concurrents sur le marché. Un autre objectif clé du plan était de relancer la croissance, en augmentant le chiffre d'affaires de l'entreprise, qui est passé de 1,3 milliard d'euros en 2014 à un objectif de 2 milliards d'euros d'ici 2020, avec une attention particulière portée à la croissance des primes d'assurance dans le secteur du voyage.

L'évolution des primes acquises nettes de réassurance (en millions d'euros) entre 2018 et 2023, dans le graphique 10 montre que EUROP ASSISTANCE SA (2022) se positionne en leader sur le marché grâce au volume de primes généré depuis 2018.

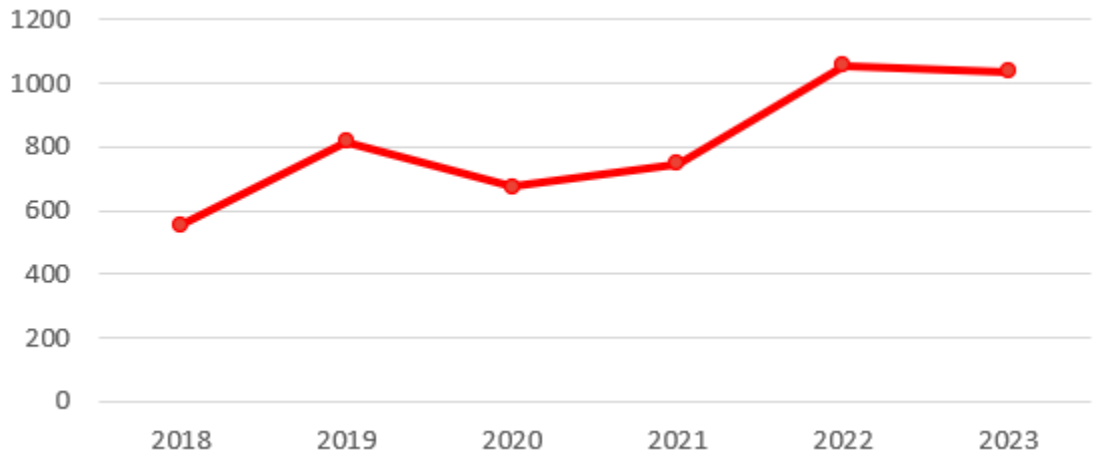

FIGURE 10 : Augmentation des primes acquises depuis 2018

À titre d'exemple, un élément explicatif est le renforcement de la position de Europ Assistance sur le segment de marché des locations saisonnières. La hausse du volume de primes à partir de 2021 est due en majeure partie à la reprise de l'activité après l'épidémie du Covid-19. D'après le rapport SFCR de Europ Assistance SA (2023), une baisse des primes acquises nettes de réassurance est principalement liée à l'arrêt du contrat de réassurance accepté en quote-part avec les États-Unis. En outre, la hausse des prix appliquée pour compenser l'inflation combinée et une hausse du volume des ventes ont joué un rôle majeur dans cette atténuation.

En 2022, Europ Assistance SA (2022) a enregistré un chiffre d'affaires de 2,4 milliards d'euros, renforçant ainsi sa position de leader sur le marché de l'assistance. L'entreprise réalise environ 13 millions d'interventions chaque année, soit plus de 35 000 interventions quotidiennes, illustrant son rôle majeur dans la gestion des situations d'urgence. Elle assure ainsi la protection et le bien-être de plus de 300 millions de personnes à travers le monde, soulignant la confiance que ses clients lui accordent.

L'innovation a également été au cœur de la stratégie "*WeConnect*". Europ Assistance s'est engagée à développer ses lignes de métier traditionnelles, telles que l'assistance routière et l'assurance voyage, tout en explorant de nouveaux domaines comme l'assistance à domicile et les soins aux personnes âgées. Pour ce faire, l'entreprise a capitalisé sur ses capacités en matière de gestion de réseaux, de plateformes et de technologies. Grâce à cette stratégie, Europ Assistance a fait des progrès significatifs. L'entreprise a lancé de nouveaux services d'assistance à domicile et de soins aux seniors, répondant ainsi à la demande croissante de services de soutien pour les personnes âgées, un segment en pleine expansion. En outre, Europ Assistance a réalisé des acquisitions ciblées, notamment aux États-Unis, où elle est devenue le troisième plus grand assureur voyage du pays. Cette expansion a renforcé la position de l'entreprise sur le marché mondial de l'assurance voyage.

Les produits d'assurance voyage de Europ Assistance sont principalement vendus par l'intermédiaire de partenaires commerciaux. Ces partenaires comprennent des émetteurs de cartes de crédit, des compagnies de transport telles que les compagnies aériennes, ferroviaires et de croisière, ainsi que des agences de voyage en ligne et des sociétés de location de vacances. Ces partenaires jouent un rôle majeur en permettant à Europ Assistance de proposer ses produits d'assurance voyage à un large public, via les canaux de vente des partenaires commerciaux. Les produits vendus sont généralement des contrats à court terme, qui couvrent spécifiquement le voyage acheté par le client final sur le canal de vente du partenaire. En plus de ces canaux, Europ Assistance vend également ses produits directement aux consommateurs. Bien que cette méthode représente une part plus petite des ventes globales, elle permet de proposer des polices annuelles qui couvrent tous les voyages effectués par le client au cours de l'année. Ces polices sont souvent associées à d'autres produits, comme des couvertures d'assistance routière, offrant ainsi une protection complète pour les voyageurs fréquents.

Avec plus de 8 500 collaborateurs répartis dans 44 pays, Europ Assistance dispose d'une capacité de réponse rapide et efficace aux besoins de ses clients, où qu'ils se trouvent. L'entreprise est présente dans plus de 200 pays et territoires, soutenue par un réseau international de 425 000 prestataires partenaires, incluant des hôpitaux, des garages, des transporteurs et autres services de secours. Europ Assistance a investi plus de 150 millions d'euros au cours des cinq dernières années dans l'innovation technologique, afin d'améliorer ses services numériques. Cet investissement s'est traduit par le développement d'applications mobiles et de plateformes d'assistance en ligne, renforçant ainsi l'accessibilité et l'efficacité de ses services. L'entreprise maintient un taux de satisfaction client de 92%, ce qui reflète son engagement constant envers la qualité de service et la satisfaction de ses clients. En 2022, Europ Assistance a géré plus de 1,5 million de dossiers médicaux à travers le monde, démontrant son expertise et sa capacité à intervenir dans des situations médicales critiques. La même année, l'entreprise a pris en charge plus de 4,8 millions d'assistances routières, un chiffre en constante augmentation en raison de la croissance mondiale du parc automobile.

Grâce à son expertise, son réseau mondial et son engagement envers l'innovation, Europ Assistance continue de se positionner comme un acteur clé dans le domaine de l'assistance globale, répondant aux besoins de ses clients dans un monde en constante évolution. Europ Assistance structure ses activités autour de trois "business lines" principales : *Travel*, *Auto* et *Personal Lines*.

Travel

L'assurance voyage est un domaine clé pour Europ Assistance, qui continue de croître en réponse à l'augmentation du tourisme mondial. En 2023, une étude de l'Ipsos (2024) a révélé que près de 7 Français sur 10 prévoyaient de partir en vacances malgré un contexte économique et international tendu, soulignant ainsi l'importance des assurances voyage. Europ Assistance propose une gamme complète de couvertures, conçues pour protéger les voyageurs contre une multitude de risques.

Les couvertures d'annulation et d'interruption de voyage sont essentielles pour rembourser les dépenses non remboursables si un voyage est annulé ou interrompu en raison de circonstances imprévues, comme une maladie grave, un accident ou un décès. D'autres événements couverts peuvent inclure des catastrophes naturelles, des actes de terrorisme, ou d'autres incidents imprévus.

En matière de santé, l'assurance voyage de Europ Assistance couvre les frais médicaux en cas de maladie ou de blessure survenant pendant le voyage. Cela inclut les frais d'hospitalisation, le transport médical d'urgence et le rapatriement si nécessaire. Ces services sont pertinents pour les voyageurs qui se rendent dans des destinations où les soins médicaux peuvent être coûteux ou difficiles d'accès.

L'assurance couvre également les bagages, en offrant une compensation pour les bagages perdus, volés ou endommagés. En cas de retard de ces derniers, Europ Assistance offre une aide pour les localiser et rembourse les articles essentiels que le voyageur a dû acheter en attendant leur retour.

Les retards de voyage et les correspondances manquées sont une autre préoccupation pour les voyageurs. À ce titre, Europ Assistance offre une couverture pour les dépenses supplémentaires telles que les repas, l'hébergement et les frais de transport engagés en raison de ces retards.

Enfin, la couverture de la responsabilité civile personnelle protège les voyageurs contre les responsabilités légales découlant de blessures corporelles accidentelles ou de dommages matériels causés à des tiers pendant le voyage, y compris les frais de défense juridique.

En plus des couvertures d'assurance, Europ Assistance propose une gamme de services d'assistance pour soutenir les voyageurs avant et pendant leur voyage. Ces services incluent des informations de voyage, la transmission de messages d'urgence, des services de traduction et une aide pour les documents de voyage perdus. Ces avantages font de l'assurance voyage de Europ Assistance un produit indispensable pour les voyageurs fréquents, les familles, les voyageurs d'affaires et les aventuriers qui recherchent une protection complète et une assistance 24/7.

Auto

L'assurance automobile est un autre pilier de Europ Assistance, offrant une protection pour les conducteurs confrontés à des problèmes liés à leur véhicule, en cas de panne, d'accident ou de vol.

L'assistance en cas de panne est une caractéristique centrale de l'assurance auto de Europ Assistance. Lorsqu'un véhicule tombe en panne, un technicien est envoyé sur place pour effectuer des réparations mineures, permettant ainsi au véhicule de reprendre la route le plus rapidement possible. Si les réparations ne peuvent pas être effectuées sur place, Europ Assistance organise le remorquage du véhicule vers le garage le plus proche ou un lieu choisi par l'assuré, en prenant en charge les frais associés.

En cas d'accident, Europ Assistance offre une assistance complète, incluant l'aide aux démarches administratives et la prise en charge des premiers frais. Cela comprend la coordination avec les services de secours, l'aide à la rédaction de rapports d'accident et le soutien pour les démarches auprès des compagnies d'assurance. Si le véhicule est immobilisé pour une durée prolongée, un véhicule de remplacement peut être fourni, permettant à l'assuré de poursuivre ses activités quotidiennes sans interruption. Pour les conducteurs dont le véhicule est immobilisé pour une durée prolongée, Europ Assistance propose une assistance voyage, qui comprend l'organisation et la prise en charge des frais de transport pour permettre à l'assuré de poursuivre son voyage ou de rentrer chez lui. Cela inclut la réservation de billets de train, de bus ou d'avion, ainsi que l'organisation de l'hébergement temporaire si nécessaire.

L'assurance auto de Europ Assistance propose également une aide précieuse en cas de vol de véhicule. Cela inclut l'assistance pour signaler le vol aux autorités, l'organisation d'un transport alternatif et l'aide pour les démarches administratives nécessaires. Si le véhicule ne peut pas être rapidement récupéré ou réparé, Europ Assistance organise le rapatriement du véhicule au domicile de l'assuré ou à un garage de son choix.

Cette assurance est particulièrement avantageuse pour plusieurs catégories de clients. Les conducteurs quotidiens bénéficient d'un soutien immédiat en cas de panne ou d'accident, les voyageurs fréquents peuvent compter sur une couverture complète lors de leurs déplacements, les propriétaires de véhicules anciens trouvent une tranquillité d'esprit face aux pannes fréquentes et les familles peuvent garantir la sécurité et la mobilité de tous les membres en cas de problème avec le véhicule principal.

Personal Lines

L'assurance "*Personal Lines*" de Europ Assistance est une offre diversifiée conçue pour protéger les individus et les familles contre une variété de risques quotidiens. Cette assurance regroupe plusieurs types de couvertures, offrant ainsi une protection globale et une tranquillité d'esprit aux assurés dans leur vie de tous les jours.

La couverture habitation est l'un des éléments clés de l'assurance "*Personal Lines*". Elle protège les propriétaires et les locataires contre les dommages causés à leur domicile par des événements tels que les incendies, les cambriolages, les tempêtes et autres catastrophes naturelles. Cette couverture comprend généralement la réparation ou le remplacement des biens endommagés, ainsi que l'hébergement temporaire si la résidence devient inhabitable, offrant ainsi une protection essentielle en cas de sinistre majeur.

La responsabilité civile est une autre composante importante de cette assurance. Elle protège les individus contre les réclamations résultant de dommages corporels ou matériels causés à des tiers. Que ce soit un accident domestique, un invité blessé, ou encore un incident impliquant un animal de compagnie, cette couverture prend en charge les frais médicaux, les réparations et les coûts juridiques associés, offrant ainsi une protection contre les imprévus du quotidien.

L'assurance santé individuelle est une autre composante importante de l'offre "*Personal Lines*". Elle permet aux assurés de couvrir les frais médicaux, les hospitalisations et les traitements spécialisés. Europ Assistance propose également des services d'assistance médicale, tels que des consultations à distance avec des professionnels de santé et la coordination des soins médicaux, garantissant ainsi un accès rapide et facile aux soins nécessaires.

Enfin, Europ Assistance propose des couvertures additionnelles, telles que l'assurance contre le vol d'identité et la protection juridique. L'assurance contre le vol d'identité aide à couvrir les coûts de restauration d'identité et les pertes financières associées à la fraude, tandis que la protection juridique

offre une assistance en cas de litiges légaux, couvrant les frais d'avocats et les dépenses judiciaires.

Conclusion

En conclusion, Europ Assistance offre une gamme complète de services d'assistance et de couvertures d'assurance qui répondent aux besoins variés de ses clients à travers le monde. Grâce à son réseau étendu de prestataires et à son engagement en faveur de l'innovation et de l'excellence du service, Europ Assistance permet à ses clients de vivre leur vie et de voyager avec confiance et sérénité, sachant qu'ils sont bien protégés contre les aléas de la vie.

Lexique

Le lexique permet de se référer aux termes utilisés dans le corps du mémoire et d'y retrouver les définitions des expressions.

AIC : *Akaike Information Criterion*
ACPR : Autorité de Contrôle Prudentiel et de Résolution
BW : *Booking window*, intervalle de temps entre la date de souscription de la police d'assurance et la date de départ.
B2B2C : *Business to Business to Consumer*, schéma de distribution en assurance annulation
CFAR : *Cancel For Any Reason*, type d'assurance annulation qui permet d'annuler son voyage sans justificatif.
CW : *Cancellation window*, intervalle de temps entre la date d'annulation de la police d'assurance et la date de départ .
COR : *Combined Operating Ratio*, ratio combiné indiquant les coûts totaux (y compris frais de gestion) sur les revenus des primes sur une période donnée.
Date de vision : Date d'extraction des données, date à laquelle les données sont visualisées.
EA : Europ Assistance
IA : Intelligence artificielle
IPTC : *Incurred in Percentage of Trip Cost*, pourcentage du coût du voyage.
GBM : *Gradient Boosting Machine*, méthode d'apprentissage supervisée.
GTO : *Gross Turn Over* chiffre d'affaires brut de frais, primes brutes nettes de frais.
GLM : *Generalized Linear Model*, Modèle Linéaire Généralisée, méthode de modélisation de la charge de sinistres.
LR : *Loss Ratio*, charge de sinistres rapportée aux primes acquises sur une période donnée.
OCW : *Occurrence window*, intervalle de temps entre la date de survenance du sinistre et la date de départ.
OPW : *Opening window*, intervalle de temps entre la date de déclaration du sinistre et la date de départ.
PBI : *PowerBI* est utilisé pour visualiser les données de profitabilité.
SFCR : rapport sur leur solvabilité et leur situation financière.
TC : *Trip Cost*, coût du voyage.
T&C's : *Terms and Conditions*, conditions générales d'utilisation de la police d'assurance.

Chapitre 1

Contexte et spécificités de l'assurance voyage annulation

Pour le lecteur familier des notions de l'assurance voyage, il convient de se rendre au point 1.2 car le point 1.1 énonce les principales notions de l'assurance voyage. Les particularités propres au partenaire commercial étudié dans ce mémoire sont rappelées brièvement au début du chapitre 2.

Ce chapitre établit les fondations nécessaires pour comprendre et découvrir les enjeux de l'assurance annulation, afin d'aborder sereinement les modélisations proposées dans les chapitres suivants.

1.1 Caractéristiques de la garantie annulation

Afin de saisir les enjeux de la garantie annulation, cette section analyse le contexte économique, les différents types d'assurance voyage puis les spécificités de l'assurance annulation. Enfin, une description du partenaire commercial, sur lequel est réalisée cette étude, est proposée.

1.1.1 Contexte économique

La pandémie du Covid-19 a donné lieu à de nombreuses restrictions sanitaires dans le monde entier. La mise en place de confinements successifs, la fermeture des frontières et les limitations de déplacements nationaux et internationaux ont considérablement réduit les possibilités de voyage lors de cette période.

Ces restrictions sanitaires n'ont eu pour conséquences que de différer dans le temps les voyages et renforcer le désir des voyageurs de partir. Depuis la fin de cette période de crise, un besoin croissant de voyager est constaté. À ce titre, l'essor du télétravail, permettant de travailler depuis n'importe quel endroit, a également accentué cette tendance. Comme le montre la figure 1.1, la proportion des individus désireux de partir en voyage est en nette croissance depuis l'année 2021 qui représente l'année de fin des restrictions sanitaires.

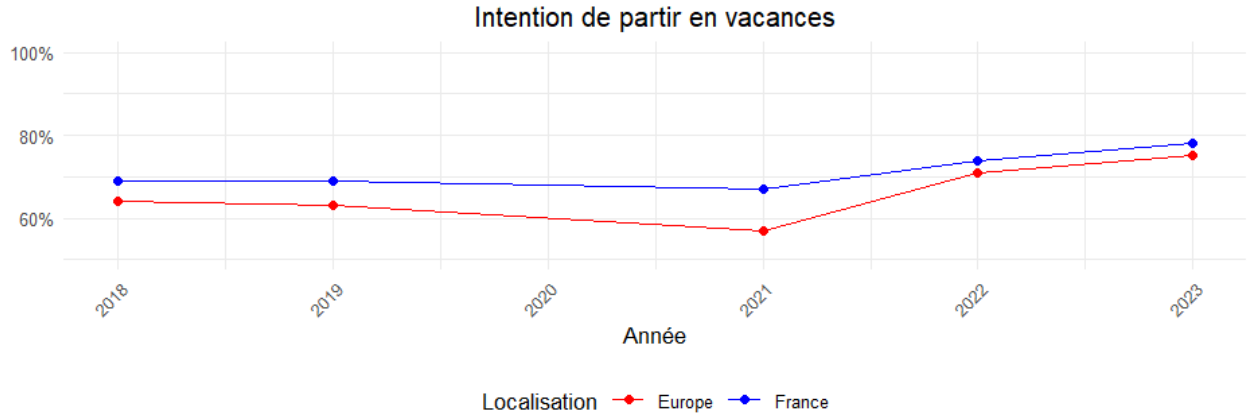

FIGURE 1.1 : Évolution de l'intention de départ en vacances

Ce graphique a été réalisé à partir des données des baromètres annuels réalisés par EUROP ASSISTANCE (2018), depuis 2018.

Les préoccupations des voyageurs se concentrent sur leur sécurité sanitaire et physique, dans un contexte d'incertitudes tant écologiques que pandémiques ou géopolitiques. Ainsi, cette hausse de l'intention de partir en vacances s'accompagne de la prise de conscience d'une nécessité de se prémunir contre le risque d'imprévus avant ou au cours du voyage.

L'étude réalisée par l'IPSOS (2024) dans le 23ème Baromètre annuel des Vacances pour Europ Assistance démontre que les Français sont sensibles aux risques climatiques. En effet, 54% d'entre eux déclarent que le risque de canicule a un impact sur le choix de destination et leur envie de voyager. En outre, cette étude met en exergue une sensibilité accrue aux risques géopolitiques montrant que plus de 58% des Français déclarent que leur envie de voyager est affectée par les conflits armés.

L'émergence de ces nouveaux risques et le climat anxiogène provoqué par les tensions géopolitiques incitent les voyageurs à se prémunir contre les risques et en particulier contre la survenance d'un événement les obligeant à annuler leur voyage. En outre, l'augmentation de l'allocation du budget dédiée au voyage depuis la pandémie du Covid et l'inflation dans les pays européens incitent les voyageurs à se prémunir contre les risques affectant leur voyage.

Une étude du *Customer Lab* d'ALLIANZ PARTNERS (2023) montre une augmentation importante des intentions d'achat de garanties d'assurance voyage. Elles ont plus que doublé entre 2019 et 2022, passant de 21% en 2019 à 55% en 2022 pour les voyages internationaux. C'est un cercle vertueux puisque les chiffres de l'*International Trade Administration* montrent une croissance significative de 41,7% du trafic aérien international en mars 2023. Par conséquent, l'assurance voyage devient une solution pour pérenniser et/ou sécuriser son voyage. Ceci explique la forte croissance du marché de l'assurance voyage depuis 2020.

En 2024, une étude réalisée par NEXT MOVE STRATEGY CONSULTING (2023) évalue le marché de l'assurance voyage à 20 milliards de dollars. Il devrait atteindre 60 milliards de dollars d'ici 2030, avec un taux de croissance annuel moyen de 14% entre 2024 et 2030. D'après cette étude, l'Europe détient la plus grande part de marché en 2022. Il convient de noter que cette information est importante car le partenaire commercial étudié dans ce mémoire propose majoritairement des voyages en France et en Europe.

En France, les principaux acteurs dudit marché sont Europ Assistance, Allianz Partners avec Mondial Assistance, Allianz Travel, Chapka Assurance (Aon), AXA ou Mutuaide, filiale de Groupama,

qui établissent des partenariats avec des partenaires commerciaux tels que des agences de location, de voyages, des compagnies aériennes, etc. Ces partenariats permettent aux acteurs du marché de pérenniser leur portefeuille de produits. La pandémie du Covid-19 a donné lieu à de nombreuses innovations dans le secteur permettant de rassurer les voyageurs lors de la réservation de leur voyage.

Les innovations dans ce secteur se traduisent aussi par l'émergence d'*assurtech* à l'instar de Yupwego, Meetch ou Heymondo d'origine espagnole, comme le montre Perrin (2024) dans son article. Ces *start-ups* de l'assurance se spécialisent sur des segments très spécifiques de l'assurance voyage et se distinguent, la plupart du temps, par le recours aux méthodes de *pricing* via l'Intelligence Artificielle (IA).

La pluralité des acteurs et des produits existant au sein de ce secteur implique de passer quelques instants à analyser les différents types d'assurance voyage.

1.1.2 Les différents types d'assurance voyage

Pour comprendre l'assurance voyage, il est essentiel de comprendre les différentes offres qui existent sur le marché et leur mode de fonctionnement, comme le souligne France Assureurs (2024). Par conséquent, il faut ainsi distinguer les polices d'assurance voyage de court terme et de long terme.

L'assurance voyage permet de couvrir les voyageurs individuels, les familles, les professionnels en déplacement et les groupes. Elle répond ainsi aux diverses exigences et préférences des différents segments de l'industrie du voyage. Les options proposées incluent des polices pour un seul voyage, des polices multi-voyages, et des polices annuelles, ainsi que des couvertures spécialisées pour des voyages spécifiques comme les sports d'aventure, les croisières, ou les voyages d'affaires.

Les polices d'assurance voyage offrent plusieurs types de couverture pour répondre aux divers besoins des voyageurs. Parmi les principales couvertures, l'annulation de voyage prend une place prépondérante. L'étude porte exclusivement sur cette garantie qui est détaillée dans la partie suivante. Néanmoins, il est pertinent d'explorer l'ensemble des garanties offertes par les assureurs afin de comprendre la structure du marché de l'assurance voyage.

En cas de maladie ou de blessure à l'étranger, la couverture médicale d'urgence prend en charge les frais médicaux, y compris l'hospitalisation, les consultations et les médicaments. Elle peut également inclure le rapatriement médical si nécessaire. La couverture des bagages et des effets personnels offre une indemnisation en cas de perte, de vol ou de dommage aux bagages et aux effets personnels durant leur voyage. Si un voyage est retardé ou interrompu en raison de circonstances imprévues telles que des catastrophes naturelles ou des grèves, l'assistance compense les dépenses supplémentaires encourues (hébergement, repas ...). Par ailleurs, la responsabilité civile protège les voyageurs en cas de dommages matériels ou corporels causés à des tiers pendant leur voyage. Elle couvre ainsi les frais juridiques et les indemnités potentielles ce qui réduit le risque financier des assurés. En cas de problèmes juridiques à l'étranger, l'assistance juridique et administrative offre une aide pour les démarches administratives et les consultations juridiques. Pour les voyageurs pratiquant des sports extrêmes ou des activités à risque, il existe des options de couverture spécifique. Ces polices spécialisées couvrent les incidents liés à ces activités, assurant une protection adéquate pour des voyages d'aventure.

Ces garanties existent sous des temporalités différentes de long et court terme. Un *focus* sur les assurances annulation est réalisé, montrant qu'elles représentent la majeure partie des souscriptions en assurance voyage. Les services d'assurance voyage à long terme sont principalement associés aux cartes bancaires, mais leur couverture est souvent limitée par rapport à une assurance annulation complémentaire proposée par une compagnie d'assurance.

Tout d'abord, seules les cartes bancaires haut de gamme, comme *Visa Premier*, *Gold Mastercard*

ou *American Express Gold*, offrent ces prestations d'assurance. Par conséquent, les voyageurs qui ne possèdent pas de telles cartes n'ont pas accès à cette couverture. De plus, pour bénéficier d'un remboursement, il est nécessaire que la prestation annulée ait été payée avec la carte bancaire concernée, ce qui réduit la flexibilité des voyageurs. Enfin, les plafonds d'indemnisation des assurances de carte bancaire sont souvent inférieurs à ceux d'une assurance annulation spécifique. Comme l'indique Nicolas Sinz dans son entretien sur France TV Info (2023), « la garantie des frais médicaux risque d'être insuffisante par rapport aux frais de santé aux États-Unis. Donc il faut se renseigner au regard de ce qu'est le voyage, la destination et les personnes qui partent avec vous. »

Les polices d'assurance voyages à court terme couvrent un événement sur une durée spécifique. Ce produit a donc une durée de vie courte. Les agences de voyage sont des acteurs majeurs de la distribution des assurances annulation. Ce schéma de distribution est appelé *Business to Business to Consumer* (B2B2C). Elles offrent souvent des polices d'assurance en partenariat avec des assureurs, à l'instar du partenaire commercial de Europ Assistance dont la sinistralité est étudiée dans ce mémoire.

Le modèle de distribution en B2B2C est un schéma, expliqué par le site Adobe (2023) , qui met en relation deux partenaires commerciaux (assureurs et agences de voyage, compagnies aériennes ou plate-forme de réservation en ligne) pour offrir des assurances voyage intégrées. Ce mode de distribution permet aux assureurs d'augmenter leur portefeuille de clients en créant des offres spécifiques au partenaire commercial. Cette collaboration entre partenaire commercial et assureur permet à l'assureur d'avoir accès aux données de souscription et aux habitudes de réservation des clients de ce partenaire. Ces données lui permettent alors de mieux appréhender la compréhension du risque de son portefeuille et d'affiner ses méthodes de *pricing*. La rémunération de cette collaboration se fait à partir de commissions, comptabilisées le plus souvent comme des frais d'acquisition dans le plan comptable assurance ou par un mécanisme de partage des profits. Ce mécanisme n'a pas de conséquence sur le *pricing* du produit. La prise en compte de cette rémunération est abordée ultérieurement dans cette étude.

Il convient de chercher maintenant à comprendre les caractéristiques de la garantie annulation, produit présenté dans cette étude.

1.1.3 Présentation de l'assurance annulation

Histoire de l'assurance voyage

Tout d'abord, il est intéressant de reprendre les grandes notions de l'histoire de l'assurance pour comprendre l'intérêt et l'émergence de l'assurance annulation.

Le besoin de se prémunir contre la survenance d'un aléa trouve ses origines chez les Babyloniens au IIème millénaire avant J.-C. avec le « prêt à la grosse aventure », expliqué avec soin dans la page Wikipédia (2024b) dédiée à ce sujet. Il s'agit d'un prêt à un taux d'intérêt élevé, non remboursable en cas de vol des marchandises ou de naufrage du navire. Ces contrats étaient principalement destinés à protéger les marchandises transportées par mer, moyen de commerce majeur à cette époque.

Au XIXème siècle, l'essor des voyages touristiques en bateau et en train a donné naissance à de nouveaux besoins en termes de couverture du risque. L'industrialisation et l'expansion des réseaux de transport ont rendu les voyages plus fréquents et accessibles à une plus grande partie de la population. Par conséquent, les premières assurances voyage ont vu le jour pour couvrir les risques inhérents à ces modes de transport. Ces assurances étaient principalement axées sur la protection contre les accidents, les pertes matérielles et les imprévus qui pouvaient survenir lors des déplacements, comme les retards ou les incidents techniques.

L'une des premières compagnies à proposer une forme d'assurance annulation moderne était la *Travelers Insurance Company* aux États-Unis, fondée en 1864, selon Travelers Insurance (2024).

Cette compagnie a introduit des contrats contenant des clauses permettant le remboursement des frais de voyage en cas d'annulation pour des raisons spécifiques, telles que des problèmes de santé graves ou des décès dans la famille. Cette innovation a marqué un tournant dans l'industrie des assurances, en apportant une sécurité financière supplémentaire aux voyageurs. Les clauses d'annulation permettaient aux individus d'obtenir une compensation financière des coûts engagés, réduisant ainsi les pertes économiques liées à des circonstances imprévues.

Au fil du temps, avec l'essor du tourisme de masse et la diversification des modes de transport, les assurances annulation ont évolué pour couvrir un large éventail de situations imprévues. L'avènement de l'aviation commerciale, la popularisation des croisières et l'augmentation des voyages internationaux ont multiplié les types de risques auxquels les voyageurs sont exposés. Aujourd'hui, les assurances annulation offrent des protections étendues, incluant les épidémies, les catastrophes naturelles, les conflits géopolitiques et bien d'autres aléas. Elles permettent aux voyageurs de planifier leurs déplacements avec une tranquillité d'esprit accrue, sachant qu'ils peuvent être indemnisés en cas d'imprévu majeur.

Cette évolution continue reflète l'importance croissante accordée à la sécurité et à la gestion des risques dans le secteur du voyage, faisant de l'assurance annulation un élément incontournable de la planification de tout voyage. C'est dans ce contexte que Pierre Desnos fonde, en 1963, Europ Assistance afin de protéger les individus voyageant à l'étranger et de leur permettre de se prémunir contre les risques pouvant affecter leur voyage.

Présentation de l'assurance annulation

En France, l'assurance annulation de voyages permet de se prémunir contre le risque d'une annulation sur présentation d'un justificatif pour le décès d'un proche, une maladie, un accident, un licenciement économique, un dommage matériel grave lié à l'habitation du voyageur, la survenance d'un événement naturel sur le lieu de la destination du voyage, la survenance d'un événement aléatoire correspondant aux caractéristiques descriptives du contrat de la garantie. Les motifs d'annulation peuvent varier selon les contrats proposés par les différentes compagnies d'assurance. Il convient de différencier les annulations de voyages des annulations d'événements et de billetterie, un domaine dans lequel l'assurtech Meetch se positionne comme un leader sur le marché. Selon le site Forbes France (2024), "plus de 2 millions de clients étaient assurés en 2022" par cette assurtech.

L'objectif des compagnies d'assurance comme Europ Assistance est d'indemniser les bénéficiaires de la police d'assurance annulation des frais engagés pour un voyage. Cette annulation doit être la conséquence de la réalisation d'un événement garanti dont la survenance a lieu avant le début du voyage. Le montant de remboursement du sinistre est limité par un tableau de garanties et le montant du coût du voyage. Ainsi, il n'est pas possible de bénéficier d'un remboursement d'un montant supérieur au coût du voyage, noté dans la modélisation *Trip Cost*, ayant pour abréviation TC. Le coût du voyage est une valeur importante puisque la prime payée se calcule en pourcentage de celui-ci. Il représente l'exposition au risque de l'assureur en termes de coût.

Pour illustrer ce propos, il faut prendre l'exemple de l'assurance souscrite auprès de Europ Assistance ou par le biais d'un partenaire commercial. L'assurance prend effet lorsque les événements cités dans les conditions générales du contrat surviennent. Les événements concernés sont énumérés dans les *Terms and Conditions*, notés *T&Cs* correspondant aux conditions générales. Ce document est fourni au bénéficiaire de la police d'assurance, avant la signature du contrat, explicitant l'étendue de la garantie, les modalités de déclaration de sinistres et les clauses légales régissant le produit.

Les conditions générales s'appuient sur le code des assurances et s'appliquent à tous les assurés indépendamment de leur profil. À ce titre, un exemple d'événements mentionnés dans les *T&Cs* est présenté dans la figure 1.2.

Extrait des *T&Cs* de Europ Assistance

- Maladie Grave, Blessure Grave ou décès :
  - d'un Assuré ;
  - d'un Compagnon de voyage ;
  - d'un Membre de la Famille ;
  - de la personne en charge de veiller sur les personnes mineures ou personnes majeures handicapées dont Vous êtes le responsable légal ou le tuteur légal ;
- Décès d'un Membre de la Famille au 3ème Degré.
- Dommage Important au Domicile ou aux Locaux Professionnels d'un Assuré.
- Perte d'emploi salarié ou de fonction non-salariée d'un Assuré.
- Commencement d'un emploi au sein d'une nouvelle entreprise dans laquelle l'Assuré n'a pas été engagé durant les six mois précédant la conclusion du nouveau contrat de travail. Les différents contrats conclus avec des entreprises d'intérim seront considérés comme des contrats conclus avec les entreprises dans lesquelles l'intérimaire exerce son activité.
- Convocation ou assignation d'un Assuré d'avoir à comparaître en tant que partie, témoin, membre d'un jury devant une juridiction judiciaire ou une autorité publique.
- Convocation d'un Assuré pour assister au sein d'un bureau de vote électoral.
- Vol de documents empêchant l'Assuré de commencer ou de continuer le Voyage.
- Casse ou Accident du véhicule appartenant à l'Assuré, l'empêchant de commencer ou de continuer le Voyage.
- Arrivée d'un enfant dans le cadre de son adoption par un Assuré.
- Echec imprévisible et injustifié d'une demande de visa par un assuré.
- Un Acte Terroriste commis, dans les 14 jours précédant la Date de départ, dans un rayon de 100 km autour de la destination que Vous avez prévu de visiter pendant Votre Voyage. Toutefois, la couverture ne s'applique que si :
  - le pays dans lequel se trouve cette destination n'a pas subi d'acte terroriste dans les 30 jours précédant la date mentionnée dans le certificat d'assurance, et
  - aucune recommandation déconseillant de se rendre dans cette destination n'était émise par une autorité gouvernementale de Votre Pays d'Origine au moment où Vous avez souscrit la Police.

Si l'événement garanti se rapporte à l'un des Assurés, les autres Assurés pourront être couverts pour ce même événement garanti.

FIGURE 1.2 : Exemple de *T&Cs*

Pour obtenir un remboursement des pertes subies, la présentation d'un justificatif (certificat de décès, certificat médical, extrait de naissance, livret de famille) est requise ainsi que la facture de

l'achat du voyage. Par conséquent, l'événement doit être imprévu et justifié afin de pouvoir bénéficier de la couverture annulation jusqu'au jour du départ.

Les polices d'assurance annulation CFAR

Lorsqu'un justificatif est requis pour annuler un voyage après la souscription d'une garantie annulation, il s'agit d'une assurance pour causes justifiées. Il existe trois catégories de couverture. Le premier type couvre les périls dénommés. Ils sont cités dans les *T&Cs* du contrat d'assurance. Le deuxième type de couverture couvre toutes causes justifiées et enfin le dernier type de couverture d'annulation est appelé *Cancel For Any Reason* (CFAR). Pour cette couverture, les assurés peuvent annuler leur voyage pour n'importe quelle raison. De la même manière que les assurances avec justificatif, les dépenses couvertes sont encadrées. Elles ne peuvent pas dépasser un certain montant fixe, qui est indiqué dans les *T&Cs* signées par le bénéficiaire de la police d'assurance.

Dans le cas d'une police d'assurance CFAR, l'assureur n'exige pas de justificatif et applique ainsi une franchise ; un reste à charge pour l'assuré de 25% en moyenne sur le coût du voyage. L'assurance CFAR est plus onéreuse qu'une assurance avec justificatif. Dans la plupart des cas, le produit CFAR est proposé en complément d'un produit d'assurance classique. De ce fait, le prix total de cette police se situe autour de 10% du coût du voyage. Le CFAR peut encourager des comportements entravant l'efficience du marché. Par exemple, des voyageurs peuvent acheter un voyage sans intention ferme de partir, en comptant sur l'assurance pour récupérer une partie des coûts.

Ce produit est très peu présent en Europe car le risque supporté par l'assureur est difficilement quantifiable. Les enjeux de gestion des risques, de viabilité financière et de prévention des abus rendent ce type de produit complexe à implémenter dans le cadre réglementaire européen. Par conséquent, ce type de produit présente une compatibilité limitée avec le cadre réglementaire assurantiel européen. Le risque d'annulation n'est désormais plus cantonné à la survenance d'un événement aléatoire. L'assuré a la possibilité d'annuler son voyage s'il ne souhaite plus partir ou s'il éprouve une appréhension à voyager.

En dehors de l'Europe, durant la pandémie de Covid-19, l'offre CFAR était le seul moyen de pouvoir annuler son voyage et obtenir un remboursement lorsque la raison était la peur de tomber malade. Par conséquent, les pertes subies par les compagnies proposant du CFAR ont été importantes. À titre d'exemple, la compagnie australienne *CoverMore* a supprimé son offre CFAR mise en place en 2018, après la pandémie du Covid-19 à la suite des importantes pertes subies, comme l'explique un article paru dans le journal Travel Weekly (2024).

Cependant, comme expliqué lors de la présentation du contexte économique, les voyageurs expriment de plus en plus un besoin de se couvrir contre les imprévus justifiés ou non. Depuis la pandémie, les annulations de voyage pour crainte de contamination, d'aléas climatiques et autres motifs ont pris une ampleur qui n'avait pas été anticipée. En Europe, il existe des solutions alternatives non assurantielles afin que les entreprises du secteur touristique bénéficient des avantages du CFAR. Elles mettent en place des tarifs flexibles d'annulation ou de modification, disponible jusqu'à une certaine date avant le départ. Compte-tenu du délai, ces entreprises peuvent revendre à moindre coût le voyage et limiter leurs pertes sur le produit vendu.

Une perspective économique

Il convient de s'intéresser au marché de l'assurance annulation comme un jeu d'équilibre entre l'offre (l'assureur) et la demande (le voyagiste). Lorsqu'un voyage est annulé, plusieurs dynamiques entrent en jeu et affectent à la fois l'assureur et le voyagiste. Il est important de noter qu'une couverture annulation

n'exclut pas la mise en place de mécanismes financiers supplémentaires (pénalités, franchises) pour le voyageur lors de l'annulation de son séjour. Ces notions sont traitées ultérieurement.

Lorsque le partenaire commercial de Europ Assistance subit un sinistre d'annulation, il peut réaffecter le voyage à un autre client en le remettant en vente sur le marché, à un prix réduit. Cette décote du prix permet d'attirer de nouveaux acheteurs dans un intervalle de temps réduit. Elle entraîne une diminution du profit pour le voyagiste. Néanmoins, le voyagiste ne subit qu'une perte partielle de la valeur du voyage, alors que l'assureur, en remboursant intégralement le montant dû à l'assuré, subit une perte sèche totale.

D'un point de vue économique, cette situation illustre un problème d'offre et de demande sur le marché de l'assurance annulation, comme représenté dans la figure 1.3. Dans ce graphique, l'axe vertical indique le prix, tandis que l'axe horizontal représente la quantité de voyages assurés. La courbe de demande (D), décroissante, reflète le comportement du voyagiste, tandis que la courbe d'offre (S), croissante, représente celui de l'assureur. L'équilibre entre ces courbes détermine le prix du marché et le volume des contrats d'assurance vendus. Les zones vertes et bleues indiquent respectivement les surplus initiaux du voyagiste et de l'assureur.

Lorsqu'un assureur propose une assurance annulation, il s'engage à couvrir le risque de devoir rembourser un voyage annulé. Cependant, en cas de sinistre, l'assureur subit une perte totale, correspondant au montant qu'il doit rembourser, comme le montre la zone rouge du graphique. Contrairement au voyagiste, l'assureur ne peut pas récupérer ses pertes une fois le sinistre déclaré. L'annulation d'un voyage entraîne pour l'assureur l'obligation de rembourser l'intégralité du montant couvert par l'assurance, ce qui constitue une perte sèche nette. Cette perte est complète et immédiate, car l'assureur ne dispose d'aucun moyen de réutiliser ou de revendre la police d'assurance annulée. Ainsi, la zone rouge sur le graphique illustre la perte sèche totale de l'assureur.

Contrairement à l'assureur, le partenaire commercial peut remettre le bien en vente avec une décote du prix initial, ce qui vient diminuer partiellement son surplus. La zone verte du graphique représente donc le surplus économique du voyagiste, qui, bien que réduit en raison de la nécessité de revendre le voyage à un prix inférieur, ne disparaît pas entièrement. Généralement, la décote du prix de vente du voyage permet une revente rapide du voyage initialement annulé.

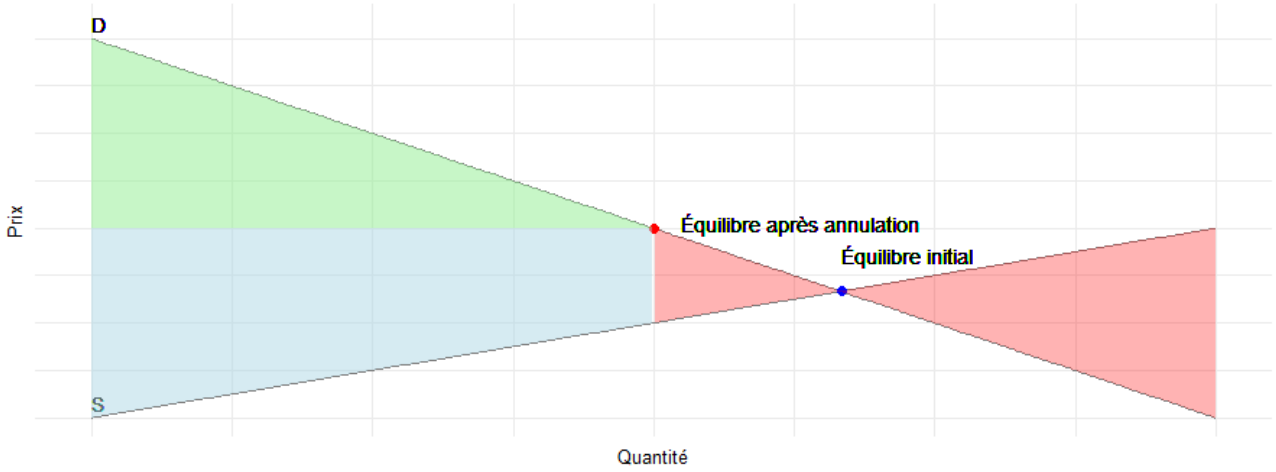

FIGURE 1.3 : Perte sèche et surplus en assurance annulation

Ce phénomène souligne pour l'assureur l'importance d'une évaluation pertinente du risque et de la tarification afin de ne pas mettre en péril la rentabilité du portefeuille. En outre, cette étude met en lumière l'avantage pour le voyagiste de mettre en œuvre un partenariat avec un assureur, puisque, en cas d'annulation, une partie des pertes peut être récupérée grâce à la revente du voyage sur le marché.

L'analyse montre comment les pertes se répartissent différemment entre l'assureur et le voyagiste en cas d'annulation et illustre les asymétries et les rigidités dans la réallocation des services en assurance voyage pour l'assureur.

1.1.4 Vocabulaire spécifique à l'assurance annulation

Dans le cadre des assurances annulation temporaires, particulièrement celles achetées en complément d'un voyage via un schéma de distribution B2B2C, la gestion du risque est étroitement liée à la temporalité. Deux concepts clés émergent dans cette analyse : la *booking window* et la *cancellation window*. Ces notions se rapportent respectivement à la durée d'exposition au risque et à l'intervalle entre la souscription et l'annulation. Leur définition, ainsi que leur importance dans la gestion des polices d'assurance, sont examinées ci-après. De manière générale, tout intervalle de temps mentionné avec le terme *window* fait référence à un écart par rapport à la date de départ du voyage.

Booking window (BW)

La *booking window* correspond à l'intervalle de temps entre la date de souscription de la police d'assurance et la date de départ. Sa durée a des conséquences sur les risques associés à la police. Plus elle est longue, plus la probabilité qu'un événement couvert par cette garantie survienne est élevée. Certains assureurs peuvent ajuster leur tarification en fonction de cette durée et augmenter le tarif des polices pour les longues BW supportant un risque plus important que les courtes BW. *A contrario*, des BW trop courtes (24 ou 48h avant la date de départ) peuvent être interdites, afin de limiter le risque d'annulation. Remarquons qu'il s'agit d'exemples existant sur le marché de l'assurance annulation afin de mieux comprendre les enjeux qui le régissent. La description du produit annulation étudiée dans ce mémoire est réalisée dans une partie suivante. Ces différentes limitations peuvent être mises en place afin de limiter les phénomènes d'aléa moral.

Il est pertinent d'illustrer ce propos par un exemple. Dans cette situation, il est possible de souscrire

une assurance après la date d'achat du voyage (ce raisonnement s'applique également lorsque les deux dates coïncident, bien que l'exemple soit alors moins explicite). Après avoir réservé son périple, le voyageur prend conscience qu'un problème majeur est fortement probable avant le départ, tel que des difficultés à obtenir son visa. Il souscrit alors une assurance annulation. Dans ce cas, l'assuré modifie son comportement car il se sait assuré. Il devient moins vigilant face à un risque de non-obtention de son visa, puisqu'en cas d'annulation de son voyage, il est indemnisé. Sans assurance, il aurait probablement évité la prise d'un tel risque, sachant qu'il risquait une perte sèche de la valeur du montant du voyage. Il s'agit alors d'un problème d'aléa moral, concept expliqué ci-après.

Cancellation window (CW)

La *cancellation window* indique le temps entre la date de départ et la date d'annulation du voyage. Pour rappel, un sinistre en assurance annulation se définit par la survenance d'un événement garanti par les conditions générales du contrat sous présentation d'un justificatif par l'assuré. Il ne peut avoir lieu qu'avant la date de départ en voyage. Par conséquent, il n'est pas possible d'annuler son voyage après la date de départ. Cependant, il arrive que la déclaration soit réalisée par l'assuré après la date de départ en voyage. Dans ce cas, ces sinistres possèdent une CW négative (cas où la date de départ est antérieure à la date de déclaration du sinistre). Afin de lever toute ambiguïté liée à la *cancellation window*, deux variables ont été introduites :

- L'*occurrence window* est l'intervalle de temps entre la date de départ et la date de survenance du sinistre.
- L'*opening window* indique l'écart entre la date de départ du voyage et la date de déclaration du sinistre.

Pour la garantie étudiée, la date de souscription de la police est la même que la date d'achat du voyage. Dans ce cas, si le voyageur souhaite souscrire à une police d'assurance annulation, il doit obligatoirement acheter la police lors de l'achat du voyage. Après la date de départ, il n'est plus possible d'annuler son voyage. Ainsi, la durée d'exposition au risque de l'assureur (*exposure time*) est la période entre la date de départ et la date de souscription de la police. Le schéma de la figure 1.4 permet de saisir les aspects techniques du fonctionnement de l'assurance annulation.

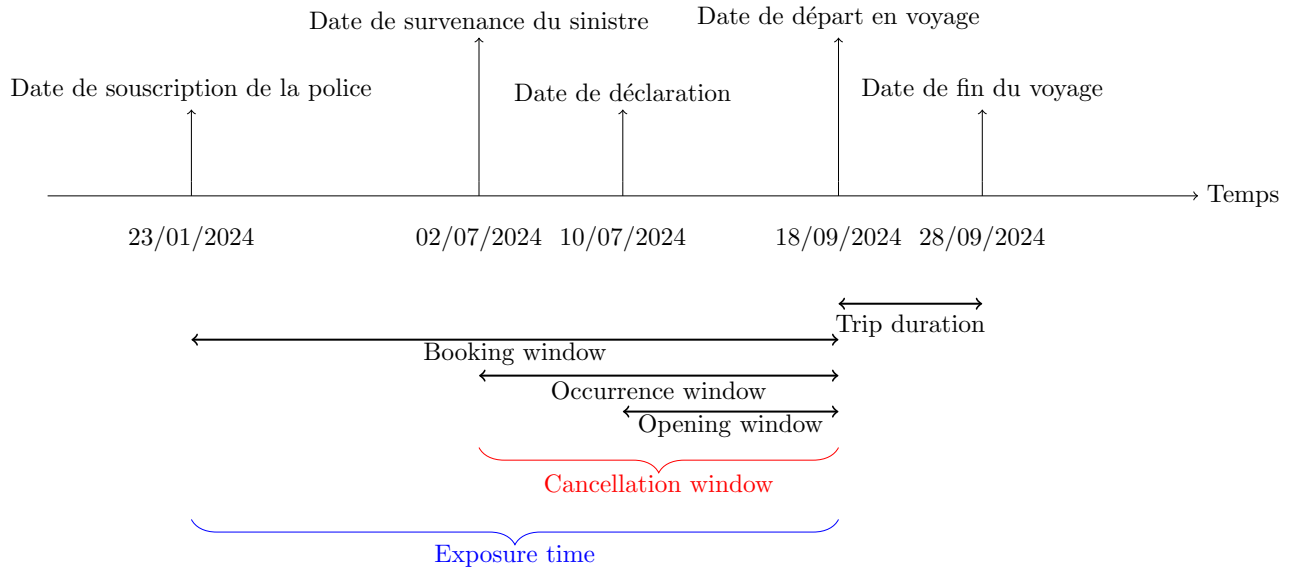

FIGURE 1.4 : Illustration d'un schéma d'annulation

Dans la suite de l'étude, dans une optique de simplification et de reflet de la réalité, il est pertinent de considérer que la date de survenance et la date de déclaration du sinistre coïncident. Par conséquent, les périodes *opening window* et *occurrence window* sont identiques et nommées *cancellation window*. Les intervalles de temps à retenir dans cette étude sont les suivants :

- *Booking window* ;
- *Cancellation window*.

Afin de faciliter la lecture du schéma précédent, la figure 1.5 représente la situation de l'étude.

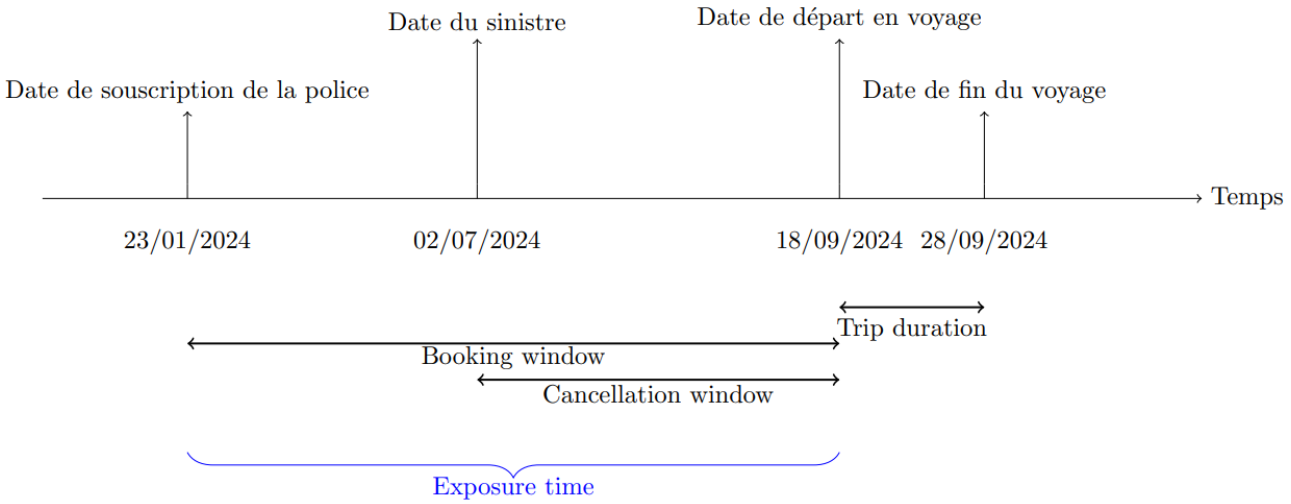

FIGURE 1.5 : Illustration d'un schéma simplifié d'annulation

Barème d'annulation

En règle générale, le partenaire commercial met en place un barème d'annulation dont le voyageur prend connaissance lorsqu'il achète son voyage. Ce barème contient le montant des pénalités appliquées par le voyagiste à son client lors de l'annulation du voyage. Les pénalités s'appliquent à tout voyageur annulant son voyage, qu'il ait souscrit ou non à une assurance annulation. En d'autres termes, le montant de pénalités correspond au montant que le voyagiste ne rembourse pas au client lorsque celui-ci annule son voyage. Par conséquent, le montant de pénalité appliqué est croissant au fur et à mesure que la date du départ approche.

Lorsque ce client est assuré, le montant de cette pénalité est prise en charge par l'assureur. Comme la pénalité est décroissante avec la durée de la CW, le montant versé par l'assureur à l'assuré l'est aussi. La pénalité croit au fur et à mesure que la CW diminue (que l'annulation s'approche de la date départ). Par conséquent, la *cancellation window* a des conséquences sur l'évaluation de la sinistralité de ce produit.

Grossièrement, il faut comprendre que la pénalité est entièrement payée par l'assureur lorsque le voyageur a souscrit à une assurance annulation. Comme le montant des pénalités augmente dans le temps, le montant des sinistres potentiels remboursés par l'assureur augmente lui aussi dans le temps. Il est à noter que certains produits d'assurance annulation imposent des conditions de couverture relatives à la CW, comme l'impossibilité d'annuler son voyage 48 heures ou 24 heures avant la date de départ. Cette condition n'est pas imposée dans le produit vendu par Europ Assistance. Par conséquent, le lecteur n'est pas surpris de constater des sinistres sur cette période.

Dans le cadre de cette étude, le partenaire commercial de Europ Assistance a mis en place un barème d'annulation. Pour rappel, celui-ci correspond au montant restant à charge au voyageur assuré à la suite d'une annulation et remboursé par Europ Assistance. Il faut donc comprendre le montant de pénalité est remboursé par l'assureur lorsque celui-ci est assuré.

Dans cette étude, il existe quatre paliers de pénalités qui varient en fonction du pays d'achat de la police. La pénalité contient un montant minimum, plancher permettant au partenaire de couvrir ses frais de dossier. Ce montant minimum de pénalité permet d'éviter un comportement opportuniste. Le schéma dans la figure 1.6 du fonctionnement du barème d'annulation permet d'illustrer les propos afin de faciliter la compréhension du sujet.

Il convient d'illustrer ces propos par l'exemple d'une assurance annulation souscrite par un assuré dont le montant total est de 1000€. Soit un barème d'annulation suivant dont les dates et les montants sont donnés à titre d'illustration.

- Lorsque l'assuré annule son départ plus d'un mois avant son départ, le partenaire commercial (en vert dans le schéma) lui rembourse le montant intégral de son voyage diminué d'un montant plancher, fixé à 200€, dans cet exemple ;
- Si l'annulation a lieu entre 40 jours et 21 jours inclus avant le départ, la pénalité est de 30% du coût du voyage ;
- Si l'annulation a lieu entre 20 jours et 8 jours inclus avant le départ, la pénalité est de 60% du coût du voyage ;
- Après 7 jours, le partenaire commercial applique une pénalité de 100% du prix du voyage et l'assureur (en rouge dans le schéma) prend en charge la totalité du montant engagé par l'assuré.

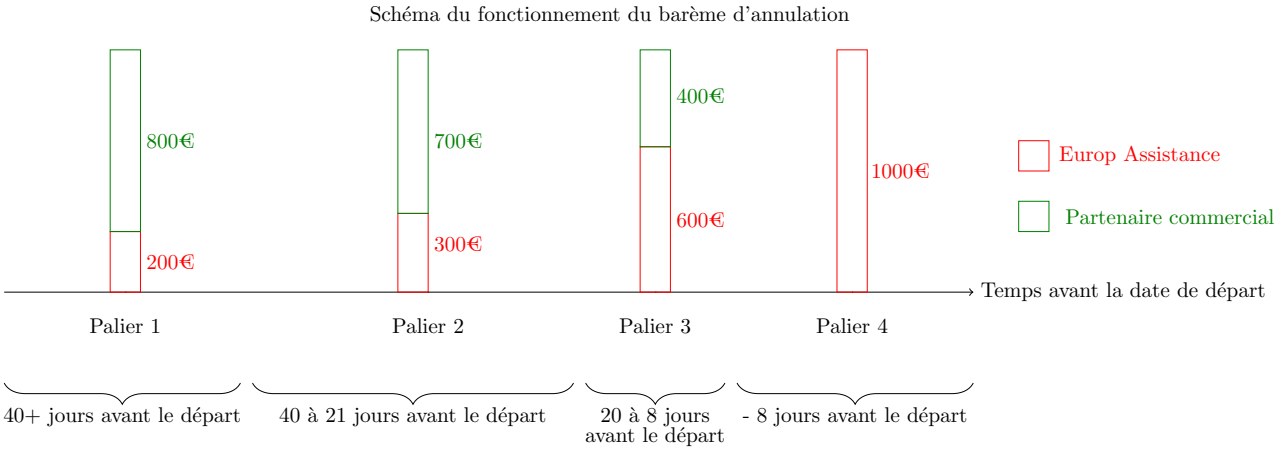

FIGURE 1.6 : Illustration d'un barème d'annulation

Pour conclure sur la mise en place de pénalités par le partenaire, celle-ci a deux conséquences majeures pour l'assureur.

- La première conséquence s'évalue en termes de volume, compte-tenu du barème d'annulation appliqué par le voyagiste, les voyageurs ont tendance à souscrire à une assurance annulation. Cela augmente le nombre de polices vendues et par conséquent, le nombre de sinistres.
- La seconde conséquence s'évalue en termes de coût des sinistres. En effet, plus le sinistre a lieu tardivement, c'est-à-dire, à proximité de la date de départ, plus l'assureur doit prendre en charge un pourcentage important du coût du voyage non remboursé par le partenaire commercial.

Anti-sélection, aléa moral et assurance annulation

Au cours de cette partie, le lecteur s'est familiarisé avec des outils que l'assureur utilise pour se prémunir contre les comportements d'aléa moral et d'anti-sélection, deux comportements résultant d'une asymétrie d'information.

L'aléa moral se réfère au comportement d'un assuré qui, après avoir souscrit une assurance, modifie son comportement en prenant plus de risques qu'il ne l'aurait fait en l'absence de couverture, puisqu'en cas de sinistre, il ne supporte pas entièrement la charge financière. Ce concept économique a été théorisé par Kenneth ARROW (1950). Il illustre une cause d'inefficience du marché. Ce type de comportement contribue à l'augmentation de la charge de sinistres de l'assureur. Mécaniquement, celui-ci répercute cette augmentation sur les primes des assurés, rendant le marché inefficace. Pour se prémunir contre ce comportement, des mesures sont mises en œuvre telles que les franchises obligeant les assurés à partager une partie de la perte financière.

Avant de comprendre le fonctionnement de ce mécanisme en assurance annulation, l'illustration de ce phénomène dans un domaine connu de tous s'impose. En assurance automobile, un agent économique conduisant une voiture de location assurée a tendance à prêter une attention mineure à celle-ci puisqu'il se sait assuré. En assurance annulation, l'aléa moral se manifeste par exemple, par des annulations fréquentes de voyage. Le voyageur sait que ses pertes financières sont couvertes par un remboursement de son assurance annulation.

D'autre part, l'anti-sélection, aussi appelée sélection adverse, a été mise en évidence par AKERLOF (1970), Prix Nobel d'économie en 2001 par son célèbre article « *The Market for Lemons* ». Dans cet

article, Akerlof démontre comment une asymétrie d'information entre acheteurs et vendeurs aboutit à une inefficience du marché. Dans le domaine de l'assurance, l'anti-sélection survient avant la souscription de la police d'assurance. Puisque les assurés connaissent mieux leur risque que l'assureur, ils ont un intérêt à souscrire une assurance lorsqu'ils anticipent un risque élevé. Ceux qui anticipent un faible risque peuvent souscrire une assurance en cas d'aversion forte au risque.

En assurance annulation, un voyageur, sachant qu'il est plus susceptible d'annuler un voyage à cause d'un problème de santé ou d'un visa défaillant, a tendance à souscrire une couverture. Sans présence de ce risque supplémentaire d'annulation, l'individu n'aurait probablement pas souscrit à l'assurance. Ce comportement entraîne une augmentation des profils de risque élevés dans le portefeuille de l'assureur, augmentant la charge de sinistres. L'augmentation des coûts de l'assureur implique une augmentation des primes afin de maintenir une rentabilité constante. De ce fait, les assurés ayant un profil de risque faible quittent le marché, révélant une inefficacité de celui-ci.

Afin de se prémunir contre l'anti-sélection, en assurance annulation, il existe des clauses d'exclusion de certaines causes d'annulation. En outre, l'impossibilité de souscrire une assurance annulation après l'achat de son voyage (ou après un délai très court) permet d'éviter que les assurés ne profitent de leur connaissance anticipée d'un risque accru.

Finalement, en situation d'asymétrie d'information, la charge de sinistres de l'assureur augmente donnant lieu à une augmentation des primes payées par les assurés. Cette situation entraîne un déséquilibre du marché qui devient inefficient. Dans ce contexte, les assureurs mettent en œuvre des moyens afin de limiter ces comportements. Europ Assistance met en place une franchise, la possibilité d'annuler son voyage sur présentation d'un justificatif, des clauses d'exclusions ainsi que l'impossibilité d'acheter sa police d'assurance après l'achat du voyage.

1.1.5 Description du partenaire commercial

L'analyse des données du partenaire commercial de Europ Assistance permet de dégager plusieurs tendances sur les intervalles de temps mis en exergue précédemment. Les données datent de 2023 avec une profondeur d'historique de 5 ans permettant de souligner des tendances dans le comportement des voyageurs souscrivant à une police annulation par le biais du partenaire commercial. Les polices d'assurance annulation sont majoritaires comme le montre le tableau 1.1. Il faut rappeler que les chiffres indiqués ci-dessous sont anonymisés à l'aide de coefficients.

Informations	**A**	**B**
Nombre de polices	1 091 192	118 118
Nombre de polices annulation	1 087 470	93 536
Montant de sinistres (€)	9 765 237	1 009 434
Montant de sinistres annulation (€)	9 421 234	841 204
Proportion de sinistres annulation (%)	99,66	79,19
Proportion montant de sinistres annulation (%)	96,48	83,33

TABLE 1.1 : Répartition des polices annulation

Le partenaire commercial étudié est un partenaire français proposant des voyages principalement en France et en Europe. Ce partenaire est scindé en deux entités notées A et B. Europ Assistance analyse la profitabilité de ce partenaire au global, sans distinguer les deux entités. Ainsi, les analyses des principales caractéristiques sont, dans ce chapitre, présentées d'un point de vue macro. Les assurés sont les clients de ce partenaire ayant souscrit à un produit d'assurance annulation si leur domicile se situe en France (FR), en Belgique (BE), aux Pays Bas (NL), en Allemagne (DE) ou en Suisse (CH). De

plus, comme indiqué précédemment, pour ce produit, les demandes de remboursement pour annulation doivent être accompagnées d'un justificatif. Les annulations pour toutes raisons (CFAR) ne sont pas prises en compte par Europ Assistance.

Tout d'abord, il est important d'analyser la durée d'exposition au risque de l'assureur représentée par la *booking window*. Parmi les individus souscrivant à ce produit d'assurance annulation, la majorité des souscriptions ont une BW dite longue (supérieure à un mois). Au total, chaque année, moins de 10% des réservations se font moins d'une semaine avant la date de départ. La figure 1.7, ci-dessous, met en avant un pic de BW entre 1 et 2 mois avant le départ entre 2019 et 2023.

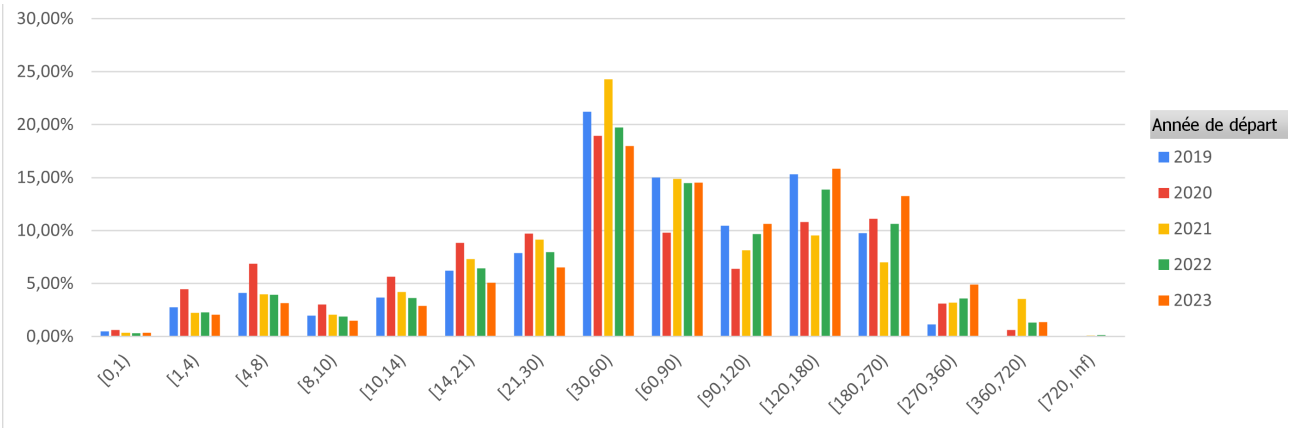

FIGURE 1.7 : Répartition des *booking window*

Cette étude préliminaire souligne que les voyageurs réservent leur voyage majoritairement au moins un mois avant leur départ. Il faut rappeler que dans cette étude, la date d'achat du voyage et de souscription de la police d'assurance sont identiques et que les voyageurs peuvent souscrire à une assurance avec une BW nulle (jour d'achat du voyage coïncide avec celle du départ).

La figure 1.8 ci-après souligne que la part d'annulation entre 0 et 7 jours (rectangles rouges, jaunes et verts) est majoritaire pour tous les intervalles de BW excepté ceux supérieurs à 180 jours (6 mois). De manière logique, une petite proportion de voyageurs ayant une longue BW annule son voyage avec une longue CW tandis que majoritairement, les annulations se font à l'approche de la date de départ.

Il est intéressant de chercher à comprendre si la BW influe sur l'utilisation de ce produit, c'est-à-dire sur la date d'annulation du voyage. Pour cela, une brève analyse de la répartition de la *cancellation window* permet de noter que la BW semble avoir un impact relativement faible sur l'intervalle d'annulation avec la date de départ, la *cancellation window*.

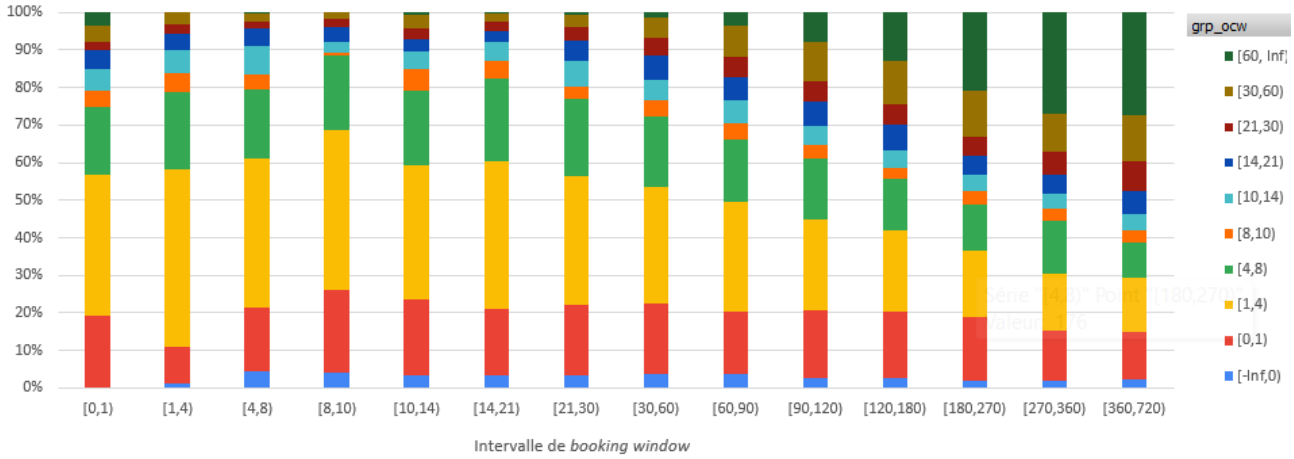

FIGURE 1.8 : Répartition des *cancellation window* en fonction des *booking window*

Dans ce graphique, la variable `grp_ocw` *indique les différents groupe de cancellation window.*

Le mémoire réalisé par Hervé GNONLONFOUN (2023) au sein de Europ Assistance a permis de mettre en lumière que, quelle que soit la BW, les voyageurs ont tendance à annuler leur voyage entre 0 et 7 jours avant leur départ. Plus la date de départ se rapproche, plus la probabilité de survenance d'un sinistre est importante et la sinistralité augmente ce qui réduit la profitabilité du contrat du partenaire commercial. Il est donc important de comprendre et de prédire cette sinistralité de manière précise. Pour cela, une définition de la sinistralité et son application à l'assurance annulation sont présentées dans la partie suivante.

1.2 Sinistralité en assurance annulation

Il convient de rappeler que l'objectif de ce mémoire est de construire un modèle de prédiction de la sinistralité en assurance annulation afin de suivre la rentabilité d'un compte d'un partenaire commercial. À ce titre il s'agit de comprendre comment modéliser la charge de sinistres afin de pouvoir mieux appréhender la rentabilité de ce partenaire. Par conséquent, il s'agit de modéliser la charge de sinistres à partir d'une variable spécifique et de réaliser un meilleur suivi de la rentabilité que celui obtenu par la méthode actuelle. Cette méthode actuellement mise en place est, par ailleurs, expliquée dans la partie 1.3. En d'autres termes, il s'agit de mettre en œuvre un modèle de suivi de rentabilité et non un modèle de provisionnement classique. Par conséquent, l'existence de Variables discriminantes dans la base de données permettent de mettre en relation la charge de sinistres à prédire avec des variables porteuses d'information sur le développement des sinistres. Ce type de modélisation permet de créer des modèles plus précis en termes d'estimation de la charge de sinistres.

Tout d'abord, une définition de la sinistralité en assurance annulation est présentée. Enfin, la variable à estimer pour apprécier la sinistralité et la profitabilité du compte étudié est explorée.

1.2.1 Définition et formalisation de la sinistralité

L'objectif de ce mémoire est d'utiliser des méthodes d'apprentissage supervisé pour prédire la sinistralité du produit annulation de ce partenaire. À ce titre, il convient de définir et d'établir les enjeux de la prédiction de la sinistralité.

La sinistralité en assurance non-vie peut se définir classiquement comme les sinistres payés nets

de recours, augmentés des frais de gestion des sinistres et des charges de provisions techniques nettes de recours. Elle se calcule brute ou nette de réassurance. Il existe plusieurs manières de quantifier la sinistralité.

En assurance annulation, la composante la plus importante est la fréquence, ratio du nombre de sinistres et du total de polices vendues sur une période donnée. Néanmoins, cette approche ne prend pas en compte les barèmes d'annulation mis en place par le partenaire commercial d'Europ Assistance. Cette négligence peut conduire à des erreurs d'interprétation et de pilotage de la profitabilité du compte de ce partenaire.

La sinistralité s'évalue donc en termes de fréquence et de coût moyen des sinistres survenus au sein d'une population assurée sur une période donnée, donnant la charge totale supportée par Europ Assistance. Ce calcul permet à Europ Assistance d'évaluer les risques associés à son contrat et de réviser la tarification en cas de mauvais résultat. Par conséquent, c'est un indicateur clé qui permet d'estimer la profitabilité du compte du partenaire commercial. Il est à noter que les différents indicateurs de rentabilité sont définis ci-après. Il est à noter que la sinistralité continue de se développer jusqu'à ce que la date de départ soit dépassée pour tous les assurés : c'est dans ce contexte que l'on cherche à prédire la sinistralité du compte du partenaire commercial de Europ Assistance.

L'estimation de la sinistralité nécessite celle de la fréquence des sinistres et du coût moyen. La fréquence représente la probabilité de survenance d'un sinistre. Elle est généralement calculée à partir des données historiques comme le ratio du nombre de sinistres et du nombre de polices. Le coût moyen indique le montant moyen d'indemnisation versé par l'assureur pour chaque sinistre.

Ainsi, le modèle coût moyen/fréquence permet de quantifier la charge de sinistres en décomposant les coûts en deux facteurs. Le premier facteur permet d'estimer combien de sinistres surviennent (fréquence). Le second facteur estime combien coûte en moyenne chaque sinistre (coût moyen). Ce modèle permet de calculer le coût total des sinistres de manière simple en multipliant le nombre total de polices par le coût moyen et la fréquence.

L'utilisation de ce modèle repose sur des hypothèses qui ne sont pas toujours conformes à la réalité, venant apporter des limites à ce modèle. Ce modèle repose sur l'indépendance des sinistres entre eux. Autrement dit, la réalisation d'un sinistre n'affecte pas la réalisation d'un autre sinistre. Cette hypothèse fondamentale dans ce modèle n'est pas viable en assurance annulation comme l'a montré la pandémie du Covid-19. En outre, le portefeuille doit être homogène pour généraliser les résultats obtenus. Cette hypothèse implique que les polices d'assurance soient suffisamment similaires en termes de risques couverts, de caractéristiques des assurés et de conditions de souscription.

Ce modèle est actuellement utilisé en supposant l'indépendance des événements. La plupart du temps, il n'y a pas de survenance de catastrophes majeures provoquant des sinistres dits "graves". Ces catastrophes, si elles se produisent, peuvent être gérées par des contrats de réassurance. De plus, l'homogénéité est assurée grâce à un portefeuille étudié suffisamment large, permettant ainsi une certaine fiabilité des résultats obtenus.

Pour débuter, il s'agit d'illustrer ce modèle par un exemple simple. En assurance automobile, le coût moyen des sinistres s'analyse comme la moyenne des montants des indemnités versées par l'assureur pour chaque sinistre. Il représente le coût moyen supporté par l'assureur pour chaque sinistre survenu. Par exemple, si l'assureur indemnise pour un montant total de 5 000 000€ correspondant à 1 000 sinistres, le coût moyen par sinistre est de 5 000€. Dans ce cas, le coût moyen s'évalue en absolu, en montant de charge de sinistres versés par l'assureur aux assurés. Il ne dépend donc pas directement de la valeur de la voiture, contrairement à l'assurance annulation où la prime est directement proportionnelle au coût. Cette méthodologie ne peut donc pas s'appliquer à l'assurance annulation.

Contrairement aux autres garanties en assurance Non-Vie, les primes d'assurance annulation s'ex-

priment, la plupart du temps, en fonction du coût du voyage (*Trip Cost*, noté TC) et elles sont proportionnelles à ce coût. Pour chaque voyage, le montant de la prime et du sinistre, s'il a lieu, sont liés au montant du voyage.

Dans un portefeuille d'assurance annulation, les montants de voyages sont multiples et varient d'une police à l'autre. Les voyages proposés par ce partenaire sont divers et variés avec des durées de voyage allant de courts séjours à des vacances de longue durée et de *standing* différent. Cette diversité implique une variation du coût des sinistres. Le coût moyen peut alors varier sans conséquence sur la sinistralité. Pour ce partenaire, la répartition des montants des voyages, comme le montre la figure 1.9, est la suivante. Ce graphique illustre que la majorité des voyages ont un coût relativement faible, autour de 500€. Une décroissance du nombre de voyages s'observe à mesure que le coût du voyage augmente. En outre, une infime partie du nombre de voyages ont un coût supérieur à 2000 €. Il faut remarquer la présence d'une queue à droite illustrant l'existence d'un petit nombre de voyages avec des coûts élevés. La fréquence de ce type de voyage est faible en comparaison aux voyages de coût inférieur à 2000 €.

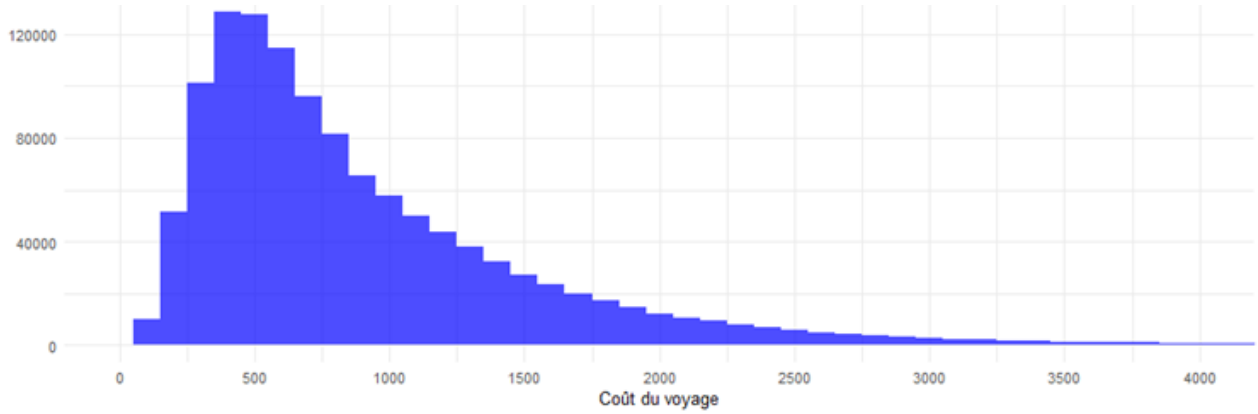

FIGURE 1.9 : Distribution des *Trip Cost*

Par conséquent, la sinistralité en assurance annulation ne peut pas s'apprécier en termes de coût moyen. Utiliser la notion de coût moyen pour évaluer la sinistralité annulation peut conduire à des estimations imprécises et des erreurs d'interprétation. Par exemple, la dérive de la sinistralité peut ne pas être remarquée compte-tenu de la structure du modèle. Une augmentation des coûts moyens des sinistres payés ne signifie pas automatiquement que la rentabilité de l'entreprise est compromise.

À cela s'ajoute que la sinistralité en assurance annulation est influencée par la temporalité. D'une part, les sinistres ne sont pas distribués de manière uniforme dans le temps, car la saisonnalité (hiver/été) influe sur leur distribution. D'autre part, plus la date de départ du voyage approche, plus la probabilité de survenance d'un sinistre augmente. Ce phénomène se traduit par des coûts d'annulation plus élevés en raison des pénalités accrues imposées par le partenaire commercial.

Ainsi, une dérive du coût moyen en montant n'indique pas directement un problème de profitabilité. La sinistralité ne peut donc pas s'évaluer uniquement en termes de montants payés. Il s'agit de supprimer l'échelle du coût moyen. Dans le modèle coût moyen/fréquence, le second facteur (coût moyen) est normalisé. L'utilisation d'un taux pour le coût moyen, défini dans le paragraphe suivant, permet de pallier ces problèmes.

1.2.2 Choix de la variable cible

L'*Incurred in Percentage of Trip Cost*, notée IPTC, apparaît comme la variable la plus adaptée pour représenter de manière optimale le risque associé à cette garantie. Cette variable à estimer est un taux de sinistres en fonction du montant total du coût du voyage. Il s'agit d'exprimer la charge de sinistres en fonction du montant d'exposition au risque. L'IPTC correspond à la prime pure en pourcentage du coût du voyage et indique le pourcentage moyen de remboursement réalisés par Europ Assistance. Cette variable est donnée par l'expression

$$IPTC = \frac{\text{Charge de sinistre}}{Trip\ cost}, \tag{1.1}$$

et suit les tendances de la fréquence des sinistres du portefeuille. Il ne faut pas confondre l'IPTC avec le coût moyen qui correspond au pourcentage du coût du voyage moyen remboursé par EA en cas de sinistre. Elle est basée sur les données historiques des voyageurs dont le départ a eu lieu durant les 12 derniers mois glissants.

Cette variable est donc significative et l'analyse de sa dérive permet de piloter de manière efficiente la profitabilité du produit. Elle prend en compte les pénalités mises en place par le partenaire commercial, comme le montrent les différents pics de la figure 1.10. Par conséquent, c'est une variable stable qui permet des comparaisons spatio-temporelles.

La distribution illustrée dans la figure 1.10 montre des valeurs supérieures à 1. Cela s'explique par le fait que Europ Assistance rembourse le pourcentage du prix du voyage associé ainsi que les dépenses supplémentaires en activité, restauration ou autres dépenses annexes, réservées lors de l'achat du voyage et non prises en compte dans le calcul de la prime d'assurance. Il faut rappeler que le partenaire étudié vend ses voyages dans différents pays et n'applique pas les mêmes pénalités dans ces pays.

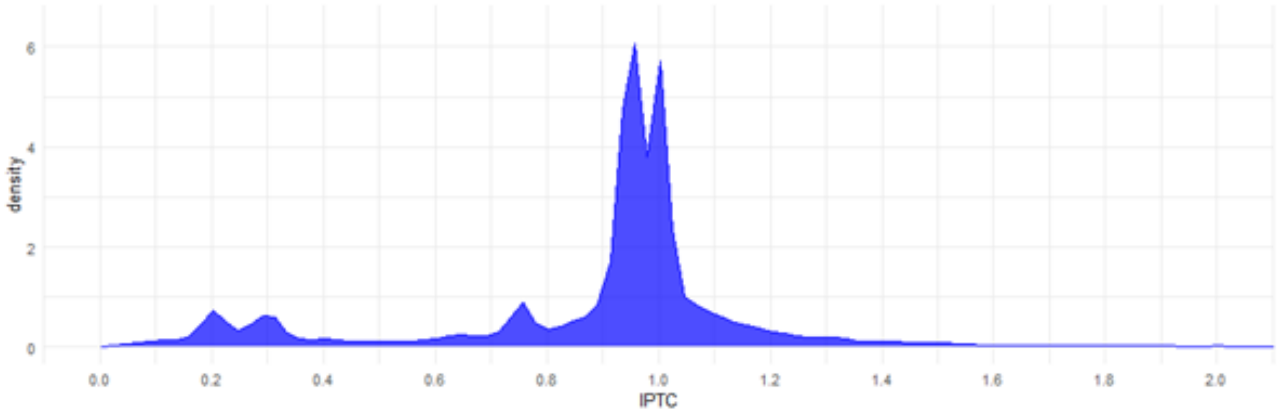

FIGURE 1.10 : Répartition de l'IPTC

Les deux pics sur la gauche du graphique reflètent des paliers de pénalités. À titre d'exemple, le pic d'IPTC à 20% indique que Europ Assistance rembourse 20% du coût du voyage aux bénéficiaires de la police. Le partenaire commercial a donc pris en charge 80% du prix du voyage. Cette faible pénalité est signe d'une longue *cancellation window*. Le bénéficiaire de la police a annulé son voyage longtemps avant la date de départ de celui-ci. Dans ce cas, il s'agit d'une longue *booking window* qui implique une longue *cancellation window*. Il convient de noter que l'adverbe "longtemps" fait référence à une durée d'au moins un mois avant le départ. Le partenaire commercial vendant ses produits dans différents pays d'Europe, les barèmes d'annulation ne sont pas identiques. Cette différence explique la présence de pics très rapprochés, par exemple les pics à 20% et 30% d'IPTC.

Les deux pics d'IPTC proches de 1 viennent appuyer le constat fait dans le mémoire de Hervé Gnonlonfoun (2023). Celui-ci stipule que les voyageurs ont tendance à annuler tardivement leur voyage, quelle que soit la *booking window*. Compte-tenu des pénalités qui s'appliquent, plus le voyageur tarde à annuler son voyage, plus les montants pris en charge par Europ Assistance sont conséquents. Les deux pics que met en avant le graphique indiquent que Europ Assistance rembourse 100% ou presque du coût du voyage aux assurés.

Ce graphique met en exergue la relation entre le pourcentage de la charge de sinistres à régler, la *booking window*, la *cancellation window* et le montant du voyage. Avant de construire un modèle permettant de prédire cette sinistralité, il s'agit de présenter la méthode actuellement utilisée et de comprendre les motivations de l'existence de ce travail.

1.3 Présentation de la méthode de prédiction en vigueur

Après une description du profil de risque du partenaire étudié, il convient de comprendre le fonctionnement de la méthode d'estimation actuelle et ses limites.

1.3.1 Contexte et cadre théorique

L'objectif de la méthode en vigueur est de pouvoir suivre la profitabilité du compte du partenaire de Europ Assistance. Projeter la sinistralité permet, à mi-année, d'estimer le montant de sinistres réglés à la fin de l'exercice et de connaître la rentabilité du produit vendu.

Cette profitabilité s'évalue à partir de deux ratios, indicateurs phares de la rentabilité d'un produit en assurance. Le *Loss Ratio* (LR) se calcule comme le ratio des prestations payées et des primes acquises. Il est défini par l'expression

$$LR = \frac{\text{Charge de sinistre}}{\text{Primes acquises}}. \tag{1.2}$$

Pour rappel, dans cette étude, les primes sont dites acquises une fois la date de départ dépassée. Le LR ne prend pas en compte les commissions versées au partenaire.

Le second ratio utilisé est le *Combined Operating Ratio* (COR), le ratio combiné. Il permet de mesurer la performance opérationnelle de l'assureur sur ce produit. Il se calcule comme le ratio des coûts totaux sur les primes acquises. Il est défini par l'expression

$$COR = \frac{\text{Charge de sinistre} + \text{Frais et autres charges techniques}}{\text{Primes acquises}}. \tag{1.3}$$

Le terme "frais et autres charges techniques" intègre, en particulier, les charges de prestations (sinistres payés), les frais d'acquisition et d'administration ainsi que les autres charges techniques telles qu'elles sont définies par l'ACPR. Ainsi, ce ratio prend en compte les commissions versées au partenaire.

- Lorsque $COR < 100\%$: le produit vendu est rentable. L'assureur gagne plus de primes qu'il ne dépense en charges techniques.
- Lorsque COR = 100% : le seuil de rentabilité est atteint pour ce produit. Les revenus des primes couvrent exactement les charges techniques. L'assureur ne réalise ni profit ni perte.
- Lorsque $COR > 100\%$: le produit vendu n'est pas rentable. L'assureur dépense plus en charges techniques qu'il ne gagne en prime.

L'objectif de cette étude est de pouvoir donner, dès la souscription des polices, le COR ultime qui est observé après développement complet de la sinistralité, une fois les primes acquises.

La sinistralité s'apprécie en termes d'exposition en euros. Les primes dépendent du coût total du voyage. Le chiffre d'affaires de Europ Assistance s'apprécie grâce au *Gross Turn Over* (GTO) qui indique le volume des primes brutes de frais. En réalité, pour connaître le réel chiffre d'affaires, il faut prendre en compte le montant de commissions à verser au partenaire qui n'est pas à négliger.

Concrètement, il s'agit lorsqu'un exercice est encore très peu développé, par exemple avec seulement quatre mois de souscription disponibles, de pouvoir estimer les indicateurs de rentabilité. Sans estimation les indicateurs de rentabilité, LR et COR, ne donnent pas une vision ultime. Cette estimation permet alors de suivre la rentabilité du compte, d'avoir un positionnement stratégique ou encore de mener des négociations avec le partenaire commercial.

Afin d'obtenir les chiffres à la fin de l'exercice comptable, une projection est réalisée à partir des données dites "stabilisées". La charge de sinistre d'un contrat est dite stabilisée lorsque la date de départ est dépassée de 3 mois. À titre d'exemple, pour une date de vision au 31 décembre 2023 (date d'extraction des données), les voyages ayant un mois de départ compris entre les mois d'octobre 2023 et décembre 2023 sont dits instables. Les données des contrats dont la date de départ est antérieure à la date d'octobre 2023 sont quant à elles considérées comme stables. Par conséquent, la notion de sinistre stabilisé s'apprécie en fonction de la distance entre le mois de départ du voyage et la date de vision.

Le graphique 1.11 permet de comprendre comment est prise en compte la sinistralité. L'unité de l'axe temporel du graphique est le mois de départ. Ce schéma permet d'illustrer le raisonnement appliqué lorsque la date de départ a lieu plus de 3 mois avant la date de vision. Dans cet exemple, la date de départ des voyageurs 1 et 2 (septembre 2023) a lieu 3 mois avant la date de vision (31 décembre 2023). Pour ces voyages, la sinistralité réelle est prise en compte et la charge de sinistres associée à ces contrats est considérée comme stable. Ainsi, il n'est pas nécessaire de faire des projections pour les voyageurs 1 et 2. Dans ce cas, les données observées sont prises en compte dans la modélisation. Pour les contrats dont le mois de départ se situe dans l'intervalle de temps entre $m - 3_{\text{date de vision}}$ et la date de vision, un *pro rata* est appliqué entre la sinistralité réelle et projetée. Dans cet exemple, il s'agit des voyages dont le mois de départ se situe entre octobre 2023 et décembre 2023. Le schéma 1.12 illustre ce raisonnement et permet de mieux comprendre ce propos. Tandis que pour le voyageur 3 dont le départ se situe après la date de vision sur la frise chronologique, il faut donc appliquer la sinistralité moyenne constatée sur un historique de 12 mois de départ, une fois les sinistres stabilisés, soit les mois de novembre 2022 à octobre 2023 ($m - 3$ par rapport à la date de départ). Cette méthodologie est décrite ci-après.

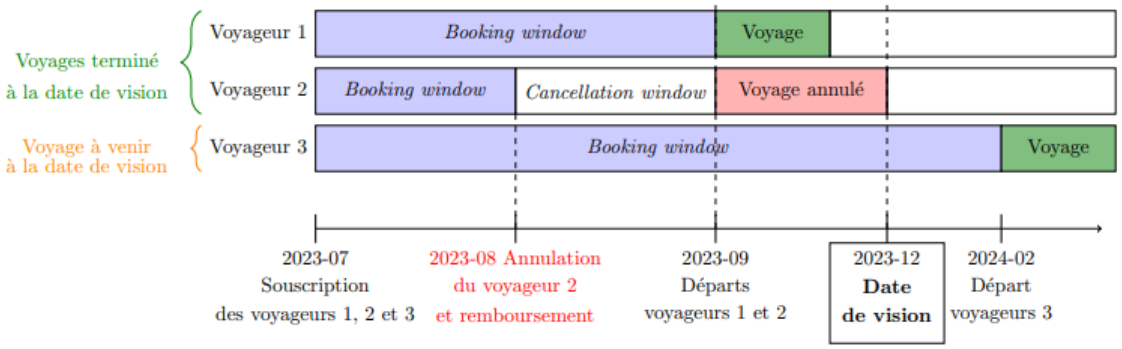

FIGURE 1.11 : Schéma de la méthodologie de projection

Dans cette étude, toutes les dates de vision sont données sous le format "année-mois" et correspondent à la fin du mois concerné.

1.3.2 Méthodologie de gestion des données

Dans l'objectif de réaliser une estimation de la sinistralité, il faut au préalable réaliser un travail sur la base de données.

Dans un premier temps, les données sont récupérées auprès du gestionnaire de sinistres de Europ Assistance. Ces données brutes sont réparties en trois bases de données.

Les données de vente des polices sont rassemblées en une première base. Ensuite, les données des sinistres sont rassemblées en une seconde base. Puis, une jointure est réalisée entre les deux bases grâce aux identifiants des assurés. Enfin, afin de pouvoir traiter les données, une base de données agrégées est créée, diminuant le nombre de lignes présentes dans la base initiale. Cette base est appelée base de *pricing*. Elle regroupe à la fois les données de polices et de sinistres. Cette agrégation implique une perte d'informations individuelles car les assurés sont regroupés par catégorie. Les catégories correspondent aux croisements des modalités des différentes variables comme des intervalles de BW, de durée de voyage, ... Ainsi, chaque catégorie définit un profil de risque particulier. Il s'agit alors d'un modèle de *pricing* collectif.

Afin de comprendre la structure des données utilisées, il convient d'analyser les retraitements effectués. La base de données de vente contient les données de ventes de polices non retraitées. Une première étape du traitement consiste à appliquer les taux de change pour les achats de polices qui ne sont pas effectués en euros. Ensuite, les calculs de variables sont effectués comme la BW ou encore la durée du voyage (*travel duration*).

La base de sinistres contient les données personnelles de l'assuré, les paiements déjà réalisés ou provisionnés et le type de couverture choisi par l'assuré. À cette étape, quelques retraitements sont déjà effectués. Par exemple, lorsque la date de souscription de la police est manquante, elle est retraitée à l'aide d'un calcul propre à Europ Assistance prenant en compte la BW moyenne.

Ces deux bases sont jointes en une base contenant les données de polices et de sinistres. Cette base est appelée base de *pricing*. Elle est ajustée de manière à pouvoir modifier la date de vision de la base, date à laquelle les polices souscrites et les sinistres déclarés sont observés. En d'autres termes, la date de vision correspond à la date d'extraction des données. Ainsi, la date de souscription de la police doit être antérieure à la date de vision choisie. De manière générale, la date de vision choisie est la date de fin de mois du mois précédent. Un exemple s'impose pour éclaircir la situation.

En considérant que l'étude est réalisée en début d'année 2024, il convient de chercher à connaître le nombre de polices achetées, la quantité de sinistres déclarés, leur charge déclarée à ce jour ou à venir ou encore leur charge totale, les montants à payer ... en somme, des critères de suivi du portefeuille, comme pour tout assureur Non-Vie.

À la date de réalisation de l'étude, la date de vision choisie est le 31 décembre 2023. Toutes les polices dont la date de souscription est antérieure à la date du 31 décembre 2023 sont inscrites dans la base de données. Néanmoins, il existe de nombreux sinistres qui ne sont pas (encore) déclarés. L'objectif du modèle, détaillé ci-après, est d'estimer la charge de sinistres à venir afin de pouvoir piloter la profitabilité du compte de ce partenaire. En janvier 2025, une fois l'année 2024 écoulée, il s'agit de vérifier que les prédictions de l'IPTC réalisées par le modèle étaient fiables et pertinentes. Pour cela, la date de vision est modifiée au 31 juillet 2024 afin de comparer les projections et les chiffres réels. Finalement, il s'agit d'appliquer une méthodologie de *backtesting*.

Cette méthode est couramment utilisée en finance et en assurance afin de contrôler la pertinence

des projections réalisées par un modèle. Il s'agit d'appliquer le modèle aux données historiques pour tester l'exactitude et l'efficacité de la modélisation mise en œuvre. Le *backtesting* aide ainsi à ajuster les paramètres du modèle pour améliorer leur précision et leur robustesse et à valider les hypothèses sous-jacentes. Néanmoins, cette méthode est limitée par le risque de surajustement aux données historiques et l'incapacité à prévoir les conditions futures du marché. Il est nécessaire d'avoir une approche prudente et complémentaire avec d'autres techniques de validation.

1.3.3 Analyse de la modélisation

A partir des données de la période précédente (fin de mois précédent), il s'agit de mettre à jour les données à la date de vision souhaitée. Par exemple, en août 2024, il s'agit de mettre à jour les données avec les nouvelles entrées du mois du juillet 2024. Après les traitements des données, la base finalisée est exportée puis utilisée sur *PowerBI* (PBI) qui permet une visualisation dynamique et esthétique des données.

Le traitement de la base de données a, entre autres, pour objectif de réduire le nombre de lignes afin de pouvoir utiliser la base de manière efficace. Pour cela, il faut utiliser une méthodologie de groupement et résumé des données, couramment employée pour agréger des données. Elle consiste à utiliser `group by` et `summarize` en langage `R`. Cette approche permet de regrouper les données de la base en fonction des variables clés et calculer les statistiques (somme, moyenne, maximum, ...) résumées pour chaque groupe de données.

Dans cette étude, les données sont regroupées par date de vision, date de départ, BW, date de souscription et pays de souscription. Pour chaque groupe de données, les statistiques calculées sont les sommes des *Ttrip cost* (montants totaux des voyages), des primes payées, des nombres et des montants de sinistres. La totalité des sommes engagées par Europ Assistance se décompose comme la somme des provisions et des montants déjà payés.

Dans une seconde étape, il s'agit de calculer les variables intermédiaires et cibles. Les variables cibles correspondent aux projections réalisées. En règle générale, dans cette méthode, les projections sont calculées comme une moyenne des valeurs de la variable souhaitée sur 12 mois de départ précédents.

Cependant, trois temporalités différentes permettent de calculer les variables cibles. Ces différentes temporalités permettent d'obtenir plusieurs méthodes de projection. Elles sont calculées comme la moyenne de leur valeur sur les deux derniers mois, six derniers mois ou douze derniers mois glissants. Dans un premier temps, les variables cibles concernent les primes puis les commissions versées au partenaire créant ainsi la prime nette. Ensuite, les variables cibles concernent l'estimation des sinistres. L'IPTC, calculé comme le ratio de la charge de sinistres et du total du coût du voyage, est donné pour chaque catégorie de profil de risque, noté i, par l'expression

$$IPTC_i = \frac{\text{Charge de sinistre}_i}{Trip\ cost_i}. \tag{1.4}$$

Il s'agit alors d'exprimer la sinistralité moyenne comme un pourcentage moyen remboursé par contrat. Puis, ce taux est appliqué au coût du voyage total à cette date de vision. Le nombre de sinistres est estimé comme le nombre de polices à la date de vision auquel on applique la fréquence estimée à cette date.

La réalisation de la projection de sinistralité sur les dernières données disponibles de ventes permet alors de donner une estimation de la profitabilité dès la souscription. Dans cette méthode, les sinistres déjà survenus sont pris en compte, tandis que pour les voyages dont l'écart entre la date de départ et la date de vision est inférieure à 3 mois, une sinistralité estimée est appliquée à partir des 12 derniers mois observés sur les autres voyageurs. Il est à noter que lorsque l'écart est positif, la date de départ

en voyage a lieu dans l'intervalle de temps entre $m - 3_{\text{date de vision}}$ et la date de vision. Lorsque cet écart est négatif, la date de départ a lieu après la date de vision (date d'extraction des données). Ce raisonnement est illustré dans le schéma 1.12.

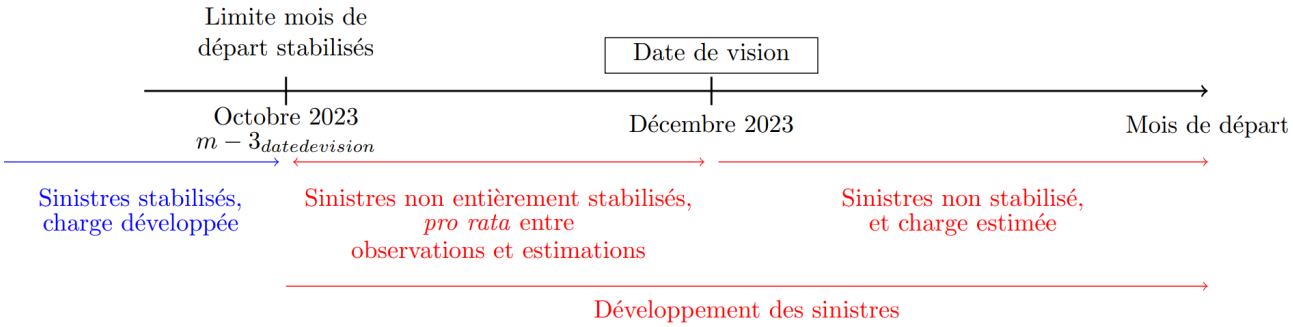

FIGURE 1.12 : Développement de la charge de sinistres

Cette méthodologie offre une estimation sans informations manquantes puisque la profitabilité est ré-estimée chaque mois permettant d'ajuster la prévision de charge de sinistres au fur et à mesure de l'exercice comptable.

Une fois ces calculs intermédiaires réalisés, une base de données finalisée est extraite et insérée dans un document PBI permettant de piloter de manière dynamique et esthétique les données de sinistralité et profitabilité de ce compte. Finalement, la profitabilité s'apprécie grâce à plusieurs indicateurs, détaillés ci-dessous.

Le nombre de polices par mois de souscription permet d'apprécier les tendances par période et la collecte de primes en nombre de polices vendues. La figure 1.13 montre une forte hausse de souscription pendant l'année 2022 et des niveaux de souscriptions stables sur les deux autres années. Le contexte inflationniste pourrait expliquer en partie la baisse du nombre de polices vendues en 2024.

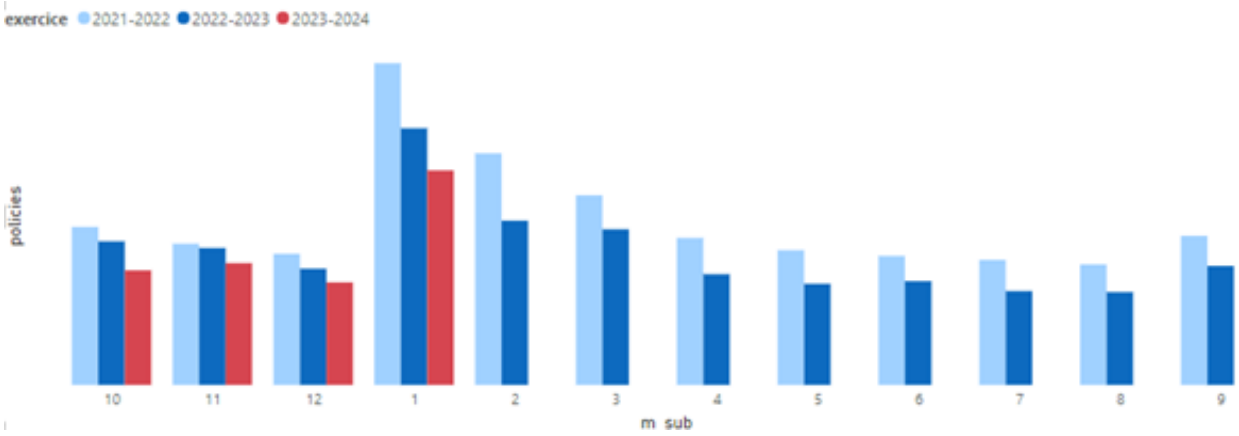

FIGURE 1.13 : Répartition des polices annulation par mois de départ

L'absence d'echelle sur les graphiques suivants est expliquée par l'anonymisation des chiffres de cette étude.

La figure 1.14 présente la répartition du GTO (montant total des primes brutes) par pays de souscription, offrant ainsi une vue d'ensemble des revenus issus des primes selon leur origine géographique. L'Allemagne et la France sont les pays collectant le plus de primes pour ce partenaire commercial. Cette analyse permet de comprendre quels marchés géographiques génèrent le plus de revenus pour

l'assureur. Les primes collectées doivent ensuite être rapportées aux sinistres afin de comprendre la sinistralité et la rentabilité de ces marchés.

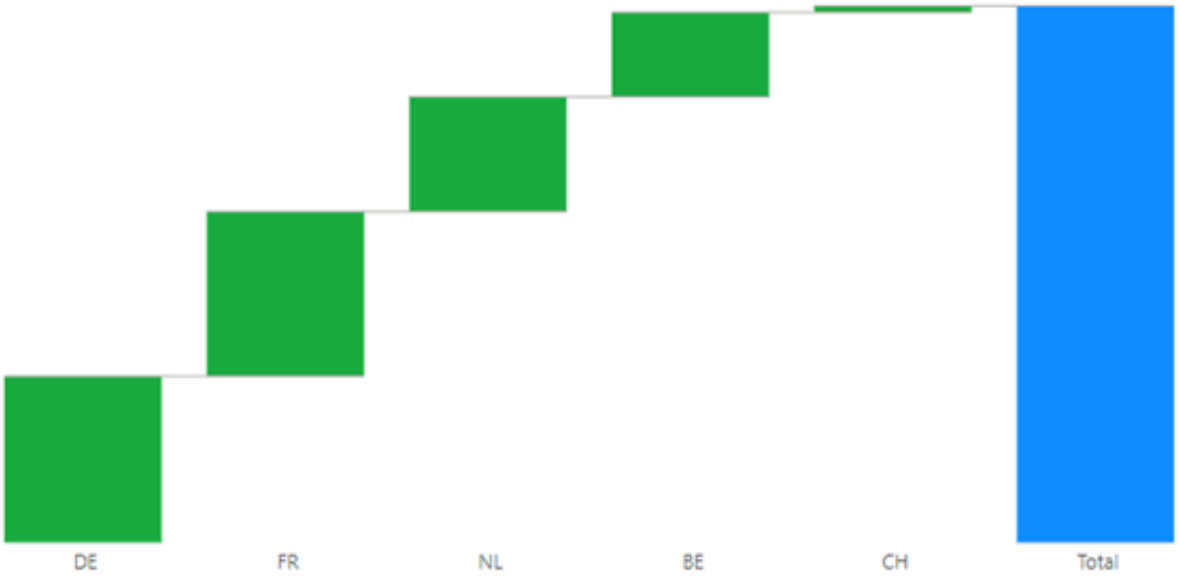

FIGURE 1.14 : Répartition du GTO annulation par pays

La répartition des sinistres par pays de ventes des polices, figure 1.14, souligne une différence marquée entre la France et l'Allemagne. Ces deux pays ont un profil de collecte de primes similaire, néanmoins, les sinistres sont plus importants en Allemagne qu'en France. Cette différence de sinistralité a des conséquences sur la tarification des contrats et le suivi de rentabilité des portefeuilles par pays.

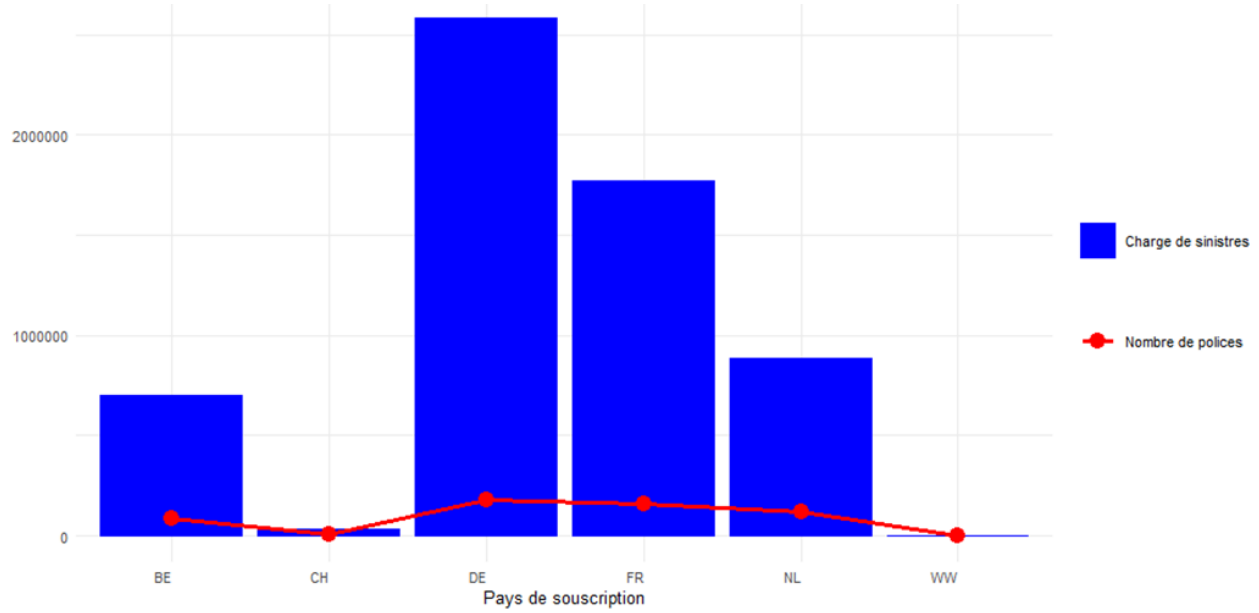

FIGURE 1.15 : Exposition par pays de souscription de la police

Enfin, la rentabilité s'analyse grâce à une comparaison des *loss ratios*. Le LR observé est comparé au LR projeté. Celui-ci est estimé grâce à une moyenne des douze derniers mois glissants des composantes

de ce ratio. Cette méthode a l'inconvénient de répercuter pendant une année les événements exceptionnels dans les prédictions réalisées, particulièrement pendant les années Covid. Un raisonnement identique est appliqué pour le calcul du COR projeté.

Dans la figure 1.16, les rectangles bleus correspondent aux LR observés au 30 juin 2024 et les rectangles jaunes aux LR projetés à cette date. L'objectif de cette méthode est qu'après développement, l'observé (en bleu) atteigne le niveau des prévisions réalisées en juin (en jaune). Au 30 juin 2024, les rectangles jaunes pour les dates allant de juin 2023 à juin 2024 indiquent alors la sinistralité en cours de développement. Le modèle estime que la différence entre le rectangle bleu et le rectangle jaune indique la sinistralité devant encore se développer, telle qu'estimée par le modèle.

Il existe un niveau de sinistralité maximum qui permet d'atteindre l'équilibre du contrat (COR 100%) et un niveau de COR cible, non indiqués sur ce graphique pour des raisons de confidentialité. La position du LR par rapport au COR cible permet alors de comprendre la rentabilité de ce compte. La profitabilité est estimée chaque mois pour ajuster au fur et à mesure les prévisions. Ces projections permettent de visualiser la pertinence de la tarification du produit et d'envisager une révision quand cela est nécessaire. L'analyse approfondie de ce graphique 1.16 contenant les données réelles et projetées montre que le LR projeté, en jaune sur le graphique, est plus élevée que le LR réel car la charge de sinistres n'est pas totalement développée. Néanmoins, après comparaison avec le LR final qui n'apparaît pas sur ce graphique, il apparaît que les projections se sont avérées trop optimistes. Le rectangle bleu a dépassé le rectangle jaune, correspondant à une situation où les charges de sinistres sont sous-estimées. Ainsi, ces projections réalisées manquent de fiabilité et de précision. L'estimation de l'IPTC semble trop simpliste pour capter les différents effets et impacts des variables influençant la charge de sinistres. Dans ce contexte de croissance du marché de l'assurance annulation, il convient d'utiliser d'autres modèles mathématiques adaptés afin de prédire au mieux la sinistralité de ce compte et de construire un outil de pilotage fiable et pertinent.

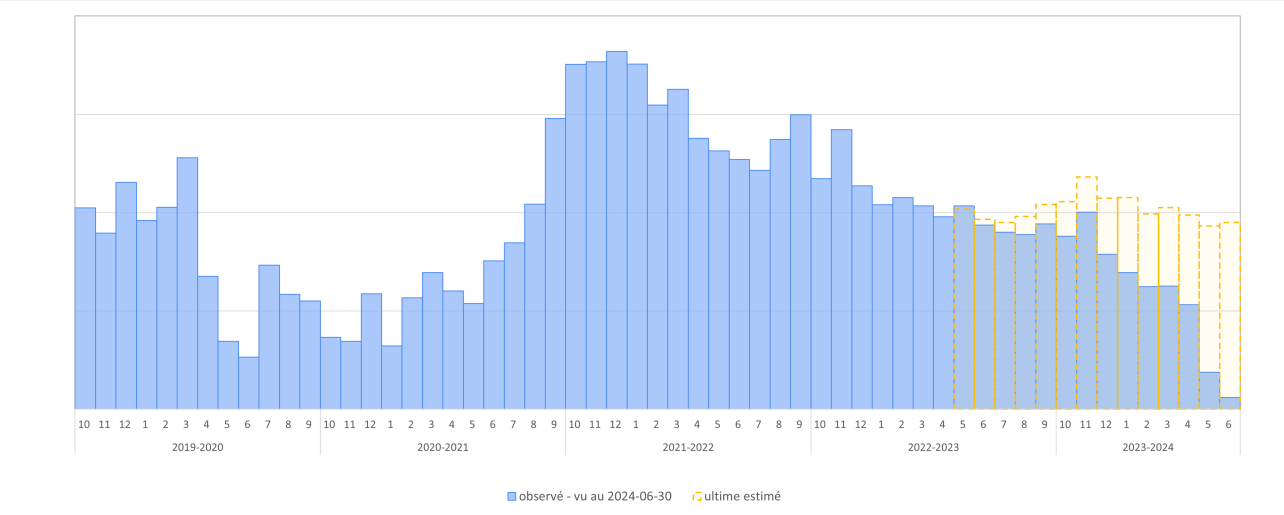

FIGURE 1.16 : *Loss ratio projeté et réel*

Conclusion

Pour conclure ce chapitre, le secteur de l'assurance voyage est en plein essor. Comme les garanties annulation représentent plus de 80% des polices vendues, Europ Assistance se doit de développer un modèle capable de prédire efficacement la sinistralité de ce produit.

Ce modèle doit permettre à l'assurer de pouvoir piloter correctement la profitabilité du compte du partenaire. En outre, il doit permettre de substituer au modèle actuellement utilisé dont les prédictions sont réalisées à partir d'un calcul de moyenne sur les douze derniers mois de départ glissants. Dans une logique de continuité, la prédiction de la sinistralité doit permettre d'utiliser le *dashboard* actuellement utilisé par Europ Assistance dans le suivi de la rentabilité des comptes. Afin de mener ce suivi, le principal indicateur est le *loss ratio*.

Le partenaire commerciale vend ses produits dans cinq pays européens : l'Allemagne , la France, les Pays-Bas, la Belgique et la Suisse. Ces ventes se réalisent à travers deux entités distinctes. L'Allemagne et la France constituent les principaux marchés de ce partenaire commercial.

Les principales notions de l'assurance annulation, comme la *booking window* ou la durée du voyage, permettent de discriminer les données. Chaque variable discriminante, brièvement abordée dans ce chapitre, est présentée avant chaque modélisation. La mise en œuvre d'un modèle tentant de répondre à cette problématique en prédisant l'IPTC est décrite dans le chapitre suivant.

Chapitre 2

Modèle de tarification en assurance annulation

La tarification d'un produit d'assurance permet d'évaluer et de quantifier le risque ainsi que la prime pure associée à la police d'assurance. Les modèles de tarification en assurance annulation utilisent des données historiques sur les sinistres, des caractéristiques sur les voyages réalisés ainsi que des informations sur les comportements des voyageurs. Avant de procéder à la description du modèle de tarification, il convient de rappeler des éléments de contexte de l'étude pour le lecteur n'ayant pas lu le premier chapitre ou peu familier des notions d'assurance.

Rappels théoriques sur la prime d'assurance

La mise en œuvre d'un modèle de tarification pour un contrat d'assurance annulation nécessite de rappeler les fondements théoriques de la prime payée par l'assuré, présentés par Etienne Marceau (2013), dans son ouvrage.

Propriétés recherchées pour les primes

Pour tarifier un contrat d'assurance, il est nécessaire de déterminer une prime suffisante pour couvrir les risques associés à la souscription ainsi que la rentabilité de l'assureur et l'utilité de l'assuré. De cette manière, les bons principes de primes, c'est à dire les principes de prime utilisable dans la pratique par l'assureur, respectent un certain nombre de propriétés. Soit $\mathcal{P}$ la fonction de prime qui associe à toute variable aléatoire $L \in \mathcal{L}$ de perte la valeur $\mathcal{P}(L)$ correspondant à la prime, où $\mathcal{L}$ représente l'ensemble des variables aléatoires réelles positives.

Les différentes propriétés recherchées pour la fonction $\mathcal{P}$ sont présentées ci-dessous.

Marge de sécurité Pour assurer sa rentabilité, l'assureur utilise une marge de sécurité. Elle permet à l'assureur de fixer la valeur de la prime telle qu'en moyenne celle-ci couvre la valeur des sinistres à indemniser. Elle est telle que

$$\forall L \in \mathcal{L}, \mathcal{P}(L) \geq \mathbb{E}[L]. \tag{2.1}$$

Exclusion de marge injustifiée Lorsqu'un risque n'est pas aléatoire, l'assureur n'ajoute aucune valeur en acceptant ce risque. Dans ce cas, la prime doit donc correspondre à la perte certaine. Cela

implique que

$$\forall L \in \mathcal{L} \text{ tel que } \exists x \in \mathbb{R}, L = x \text{ p.s.}, \mathcal{P}(L) = x. \tag{2.2}$$

Majoration Si une perte est plafonnée par une certaine valeur, il est alors nécessaire que la prime associée soit inférieure ou égale à ce plafond,

$$\forall L \in \mathcal{L} \text{ tel que } \exists M \in \mathbb{R}, L \leq M \text{ p.s.}, \mathcal{P}(L) \leq M. \tag{2.3}$$

Invariance d'échelle Si une perte est multipliée par un facteur positif, la prime associée doit être ajustée proportionnellement. Cela se traduit mathématiquement par la relation

$$\forall L \in \mathcal{L}, \forall \lambda \in \mathbb{R}_+^*, \mathcal{P}(\lambda L) = \lambda \mathcal{P}(L). \tag{2.4}$$

Sous-additivité Lorsqu'on considère la somme de deux pertes, la prime totale ne doit pas dépasser la somme des primes individuelles correspondantes. Ce principe peut être exprimé par

$$\forall (L_1, L_2) \in \mathcal{L}^2, \mathcal{P}(L_1 + L_2) \leq \mathcal{P}(L_1) + \mathcal{P}(L_2). \tag{2.5}$$

Additivité Lorsque deux pertes sont indépendantes, la prime associée à leur somme doit être égale à la somme des primes individuelles. Ce principe peut être formalisé par l'expression

$$\forall (L_1, L_2) \in \mathcal{L}^2 \text{ tels que } L_1 \perp\!\!\!\perp L_2, \mathcal{P}(L_1 + L_2) = \mathcal{P}(L_1) + \mathcal{P}(L_2). \tag{2.6}$$

Invariance par translation Cette propriété garantit que la prime réagit de manière prévisible et proportionnelle à des ajustements fixes des pertes. Elle assure que l'ajout d'une valeur fixe aux pertes se reflète directement dans la prime, sans affecter la relation entre la prime et la composante aléatoire des pertes. Elle est telle que

$$\forall L \in \mathcal{L}, \forall a \in \mathbb{R}, \mathcal{P}(a + L) = a + \mathcal{P}(L). \tag{2.7}$$

Principes de prime

Lors de la tarification des produits d'assurance, plusieurs méthodes permettent de réaliser le calcul de la prime. Elles permettent de varier la stratégie de tarification en fonction du produit d'assurance. Chaque méthode repose sur un principe particulier et respecte les propositions énoncées précédemment. Ces principes sont énoncés ci-après.

Principe de la prime pure Lorsque la tarification est réalisée à partir de la prime pure, elle correspond à l'espérance du coût du sinistre. Par conséquent, elle est généralement inférieure au coût du sinistre grâce au principe de mutualisation. Ce principe repose sur la mise en commun des risques d'un grand nombre d'assurés permettant de répartir les coûts des sinistres sur l'ensemble du portefeuille de l'assureur. Cette prime est définie par la relation

$$\forall L \in \mathcal{L}, \mathcal{P}(L) = \Pi^{\text{pure}}(L) = \mathbb{E}[L]. \tag{2.8}$$

Principe de la valeur espérée Ce principe est la méthode la plus simple à mettre en œuvre et la plus utilisée. Il s'agit d'ajouter à la prime pure, un pourcentage supplémentaire qui permet à l'assureur de se protéger contre les incertitudes. En pratique, Europe Assistance utilise ce principe de la valeur espérée. La marge de sécurité correspond à λ. La relation est alors

$$\forall \lambda \in \mathbb{R}_+^*, \forall L \in \mathcal{L}, \mathcal{P}(L) = (1+\lambda)\mathbb{E}[L]. \tag{2.9}$$

Principe de la variance Lorsque le risque est très volatile, une stratégie consiste à ajuster la prime pure d'une proportion de la variance, tel que

$$\forall \lambda \in \mathbb{R}_+^*, \forall L \in \mathcal{L}, \mathcal{P}(L) = \mathbb{E}[L] + \lambda \mathbb{V}[L]. \tag{2.10}$$

Principe de l'écart-type Ce principe est similaire au principe précédent. L'utilisation de l'écart-type au lieu de la variance permet de modérer l'impact des valeurs extrêmes, tout en restant sensible à la volatilité des données. L'écart-type modère l'effet des valeurs extrêmes en comparaison à la variance. La relation devient alors

$$\forall \lambda \in \mathbb{R}_+^*, \forall L \in \mathcal{L}, \mathcal{P}(L) = \mathbb{E}[L] + \lambda \sqrt{\mathbb{V}[L]}. \tag{2.11}$$

Synthèse des éléments théoriques

Afin de comprendre les raisons de l'utilisation massive de la prime pure en assurance, le tableau 2.1 permet de synthétiser l'ensemble des propriétés vérifiées par les différents principes. À ce titre, il faut noter que la prime pure est le seul principe à vérifier les propriétés citées précédemment.

Propriétés des primes	**Prime pure**	**Valeur espérée**	**Variance**	**Écart type**
Marge de sécurité	Oui	Oui	Oui	Oui
Exclusion de marge injustifiée	Oui	Non	Oui	Oui
Majoration	Oui	Non	Non	Non
Invariance d'échelle	Oui	Oui	Oui	Oui
Sous-additivité	Oui	Oui	Oui	Non
Additivité	Oui	Oui	Oui	Non
Translation	Oui	Non	Oui	Oui

TABLE 2.1 : Récapitulatif des équations des propriétés pour chaque principe de primes

D'autres primes et mesures de risques existent mais ne présentent pas d'intérêt pour la suite de cette étude. Cette étude utilise le principe de prime pure dans une logique de continuité de tarification du contrat et simplicité à mettre en œuvre pour Europ Assistance.

La prime commerciale

La prime réellement payée par le voyageur correspond à la prime commerciale qui contient la prime pure, les chargements (frais de l'assureur y compris les commissions) et les taxes associées à cette police

d'assurance, comme le rappelle l'Institut des Actuaires (2017). Le schéma, en figure 2.1, rappelle brièvement ce principe qui s'écrit mathématiquement comme le montre l'équation

$$\text{Prime Commerciale} = \text{Prime Pure} \times (1 + \text{Marge de sécurité}) + \text{Chargements}. \tag{2.12}$$

Les chargements couvrent les frais généraux, tels que les frais administratifs ou de gestion des sinistres et les commissions versées aux intermédiaires. Le chargement de sécurité permet de protéger l'assureur des imprévus et des fluctuations du marché ou de saisonnalité.

Enfin, le pourcentage de taxes appliqué à la police, taxe spéciale sur les conventions d'assurance, est inscrit dans le code général des impôts et peut être modifié lors du projet de loi de finances voté par le parlement.

À cela s'ajoutent les frais et les partages de bénéfices réalisés par l'assurance que l'assureur doit au voyagiste.

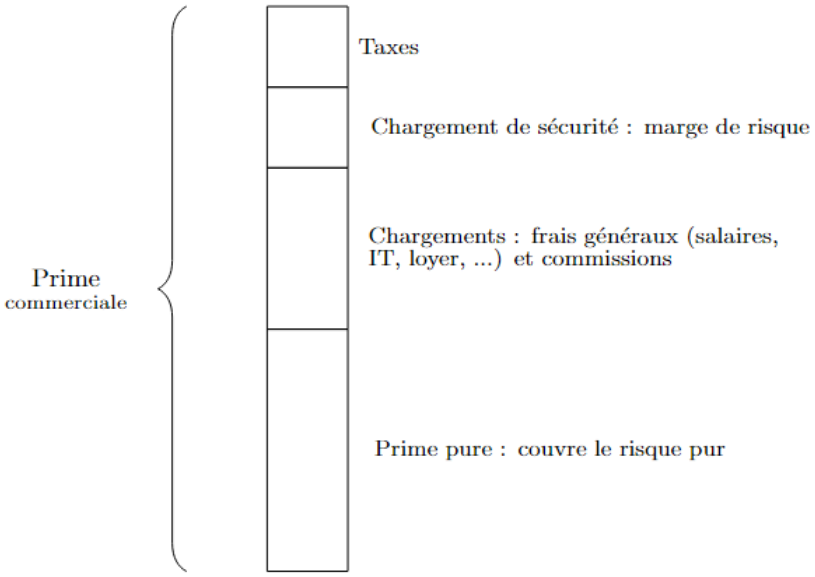

Figure 2.1 : Constitution de la prime commerciale

Il faut noter qu'en assurance annulation, les commissions versées au voyagiste par l'assureur sont relativement importantes. Le montant de ces commissions varie selon le type de contrat. En cas de schéma de distribution de type B2B2C, où le voyagiste joue un rôle d'intermédiaire entre l'assureur et l'assuré, ces commissions doivent être soigneusement négociées et intégrées dans la tarification globale afin de garantir une répartition équitable des coûts et des revenus entre toutes les parties impliquées.

Pour déterminer la prime pure d'un contrat, il faut estimer sa sinistralité, à l'aide des données disponibles.

Rappels des spécifités en assurance annulation

En assurance annulation, les montants de sinistres sont grandement corrélés au *trip cost*. Ainsi, la prime s'exprime en pourcentage de celui-ci. Les sinistres s'évaluent donc relativement au montant du *trip cost*. Cette variable est notée IPTC et est présentée dans le chapitre 1.

Deux méthodes permettent d'estimer la sinistralité. Une première méthode consiste à estimer le coût moyen des sinistres et la fréquence en affectant une loi de probabilité à chaque composante du modèle (coût moyen et fréquence). Dans ce cas, la prime pure s'obtient en multipliant la fréquence des sinistres par la gravité des sinistres, tel que Prime pure = Fréquence × Coût moyen.

Une seconde méthode consiste à estimer directement la prime pure sans utiliser un modèle de prédiction "coût moyen/fréquence". Dans cette étude, l'IPTC estimé est directement appliqué au coût du voyage et permet de prédire contrat par contrat la charge de sinistres associée.

Pour estimer cette charge de sinistres, deux types de variables sont disponibles :

- les variables de souscription : elles correspondent aux données disponibles après la souscription à une police annulation. Il s'agit de la date de départ, de la date de fin du voyage, de la date de souscription (correspondant à la date d'achat du voyage), de la durée du voyage et du pays d'achat de la police.

- les variables de sinistralité : elles permettent de prendre en compte le développement du sinistre, lorsqu'il a lieu, au fur et à mesure que la date de départ approche. Elles viennent en complément des variables de souscription. Ces variables correspondent à la date d'annulation du voyage, à la charge de sinistres en pourcentage du coût du voyage (IPTC) à la date de calcul, à l'IPTC ultime et à la *cancellation window*.

En outre, inclure la sinistralité observée aujourd'hui parmi les variables explicatives risque de masquer les effets des autres variables sur l'IPTC. Il faut noter que cette variable n'est pas une variable de souscription. Ainsi, seule la sinistralité serait prise en compte, ce qui n'est pas pertinent. Par conséquent, ces variables de sinistralité ne sont pas utilisées dans le GLM permettant de prédire l'IPTC. L'IPTC est prédit pour chaque contrat et représente la prime pure payée par l'assuré. Il s'agit d'estimer la charge de sinistres potentielle pour chaque contrat en fonction du coût du voyage.

Par conséquent, il s'agit de réaliser un GLM uniquement avec les variables de souscription. Les résultats sont comparés au modèle déjà mis en place par Europ Assistance. L'intégration de la sinistralité dans le modèle de prédiction est étudiée dans le chapitre 3 avec un modèle permettant de prendre en compte le développement de la sinistralité.

2.1 Aspects théoriques de la modélisation GLM

Une méthode classique de tarification en assurance est l'utilisation des modèles linéaires généralisés, notés GLM et présentés notamment par Cornillon et al. (2019) dans leur ouvrage. Cette méthode fiable dont les résultats s'interprètent facilement permet d'analyser les premiers résultats de la modélisation de la prime pure de ce produit.

2.1.1 Aspects théoriques de la modélisation GLM

Les modèles linéaires généralisés permettent de modéliser une relation non linéaire entre une variable dépendante et plusieurs variables indépendantes. Ces modèles sont une généralisation des modèles linéaires et procurent l'avantage de pouvoir modéliser une relation non linéaire entre les variables ainsi que de ne pas imposer la normalité de la variable cible, comme le soulignent Frédéric Planchet, Antoine Miseray (2023). Dans cette étude, il s'agit d'étudier le lien entre l'IPTC, variable cible à prédire et des variables de souscription constituant les variables explicatives. Les variables dites de

souscription sont les données disponibles lorsque le voyageur souscrit une police d'assurance. Cette relation, pour chaque police d'assurance, est notée

$$g(\mathbb{E}[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{2.13}$$

où y_i est l'IPTC associée au contrat i, g, une fonction de lien décrite ci-dessous et *où* β_0 *est l'intercept,* β_1, β_2, ..., β_p sont les coefficients de régression associés à chaque variable explicative x_{i1}, x_{i2}, ..., x_{ip}.

Ces modèles sont une généralisation des modèles linéaires puisqu'ils permettent la modélisation des données dont la distribution ne suit pas une loi normale. En assurance, ils sont utilisés en tarification pour calculer la prime associée à un contrat.

Dans ce cas, le GLM est utilisé pour prédire la sinistralité à partir des données de souscription disponibles. Il est nécessaire de rappeler qu'à partir des données historiques de l'assureur, c'est-à-dire les données à disposition pour créer le modèle adéquat, le GLM estime une valeur pour chaque coefficient de régression β_i associé à chaque variable explicative (le vecteur $\mathbb{X}$ de variables explicatives). Ainsi, les données de chaque variable X doivent être disponibles pour chaque contrat dont la sinistralité doit être estimée. Ils reposent sur trois composantes qui définissent les conditions d'application de ces modèles.

La première composante du GLM est la variable aléatoire associée à la distribution des données de la variable cible dont $(y_i)_i$ sont des réalisations. La loi de densité doit être inclue dans la famille exponentielle, c'est-à-dire qu'elle s'écrit de la forme

$$f_{\alpha_i}(y_i) = \exp\left(\frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi)\right), \tag{2.14}$$

où $\alpha_i \in \mathbb{R}$, $\phi \geq 0$ est un paramètre de dispersion et a, b et c sont des fonctions à valeurs dans $\mathbb{R}$.

La plupart des lois de distribution classiques, telles que les lois normales, gamma ou poisson, font partie de la famille exponentielle. Le choix de cette composante dépend de la nature de la variable cible. Dans ce cas, il s'agit d'un pourcentage de charge de sinistres en fonction du coût du voyage. Ainsi, compte-tenu de la structure de la loi de probabilité, les données sont modélisées par une loi de Tweedie décrite ultérieurement.

La deuxième composante est une composante déterministe, notée $\eta(x_i; \theta)$ est une famille paramétrique de fonctions. Dans cette étude, la formulation classique est choisie en supposant qu'elle s'exprime sous forme d'une combinaison linéaire des prédicteurs, c'est à dire

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}. \tag{2.15}$$

Il s'agit de déterminer le choix de la combinaison linéaire des variables explicatives qui permet d'expliquer les variations de la variable cible. Des transformations de variables peuvent être appliquées aux variables explicatives afin d'améliorer la performance du modèle. Pour cela, des méthodes de sélection de variables sont utilisées et expliquées dans une prochaine partie.

La troisième composante du GLM est la fonction de lien, notée g, comme le montre l'équation 2.13. Cette fonction strictement monotone et dérivable, permet de relier linéairement l'espérance de la variable cible et une combinaison linéaire des prédicteurs. Elle est choisie de manière à être bijective.

Les hypothèses et conditions d'existence des GLM sont explicités dans l'ouvrage de Cornillon et al. (2019).

Dans cette étude, la composante choisie est une fonction de Tweedie décrite dans la partie suivante.

2.1.2 Spécifités de la modélisation de Tweedie

En 1984, le physicien et statisticien Maurice Tweedie à l'université de Liverpool a introduit les distributions Tweedie, notées $T_p(\mu, \sigma^2)$. Celles-ci sont caractérisées par une masse positive en zéro et positive et continue sur le reste de l'ensemble de définition. Elles sont des cas particuliers des modèles de dispersion exponentiels et sont caractérisées par une relation spécifique entre la moyenne et la variance de la variable aléatoire. Autrement dit, la variance est une fonction de la moyenne.

Ces distributions, expliquées dans l'article de Delong et al. (2021), possèdent deux paramètres. Le paramètre de puissance, noté p, contrôle la relation entre la moyenne et la variance. Il influence la forme de la distribution. Tandis que le paramètre de dispersion σ contrôle l'échelle de la variance par rapport à la moyenne. Le changement de σ modifie l'échelle de la variance, mais la relation exponentielle entre la moyenne et la variance reste déterminée par le paramètre de puissance p.

Dans ce contexte, la fonction génératrice des cumulantes* permet de définir les moments de la distribution. Les modèles de dispersion exponentiels (ED) font partie de la famille exponentielle†. Ils se concentrent sur les distributions pour lesquelles la variance est spécifiquement une fonction de la moyenne.

Ces distributions sont souvent utilisées pour traiter des données hétéroscédastiques, c'est-à-dire des données où la variance des erreurs varie en fonction des valeurs des variables. Par exemple, elles permettent de modéliser la variance comme une fonction de la moyenne. Dans ce cas, les distributions de Tweedie sont des cas particuliers des modèles de dispersion exponentiels, où la variance est proportionnelle à une puissance de la moyenne.

Les distributions de Tweedie sont des lois composées de Poisson-Gamma, lorsque $p \in]1, 2[$. Elles peuvent être vues comme des distributions de Poisson où le nombre d'événements suit une distribution de Poisson. Chaque événement a une taille suivant une distribution Gamma. La relation moyenne-variance est telle que si $Y \sim T_p(\mu, \sigma^2)$, alors la moyenne $\mu = \mathbb{E}(Y)$.

La relation suivante est obtenue $\text{Var}(Y) = \sigma^2 \mu^p$, où $p \in \mathbb{R}$ est appelé le paramètre de puissance Tweedie, hyperparamètre du modèle à déterminer, σ^2 est le paramètre de dispersion positif et μ est la moyenne.

La distribution de probabilité P_{θ,σ^2} sur les *ensembles mesurables* A, est donnée par

$$P_{\theta,\sigma^2}(Y \in A) = \int_A \exp\left(\frac{\theta \cdot z - \kappa_p(\theta)}{\sigma^2}\right) \cdot \nu_\Lambda(dz), \tag{2.16}$$

pour une mesure σ-finie ν_Λ. Cette représentation utilise le paramètre canonique θ d'un modèle de dispersion exponentiels et d'une fonction génératrice des cumulants.

Les distributions de Tweedie sont étudiées en utilisant le cadre général des modèles de dispersion exponentiels, avec une dépendance particulière sur le paramètre canonique θ et la fonction cumulante pour caractériser la distribution. Ainsi, dans les distributions de Tweedie :

- La forme de la densité de probabilité de ces distributions est écrite en termes du paramètre canonique θ qui est commun aux modèles de dispersion exponentiel ;

*Une fonction cumulante $\kappa(\theta)$ est définie, par Wikipédia (2023), comme le logarithme de la fonction génératrice des moments $\varphi(\theta)$ d'une variable aléatoire X. Elle se note :

$$\kappa(\theta) = \log(\varphi(\theta)) = \log(\mathbb{E}[e^{\theta X}]).$$

Elle permet d'obtenir les cumulants d'une loi de probabilité, paramètres utilisés pour décrire la forme de la distribution (moyenne, variance ou encore le kurtosis).

†Démonstration en annexe.

- κ_p est utilisée pour définir et travailler avec les moments et les cumulants de la distribution. Son expression dépend de p et est donnée, selon le paramètre canonique, par

$$\kappa_p(\theta) = \begin{cases} \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha}, & \text{pour } p \neq 1,2 \\ -\log(-\theta), & \text{pour } p = 2 \\ e^{\theta}, & \text{pour } p = 1 \end{cases} \tag{2.17}$$

avec $\alpha = \frac{p-2}{p-1}$ ou $p = \frac{\alpha-2}{\alpha-1}$.

Il est nécessaire de rappeler que les dérivées successives de la fonction cumulante $\kappa(\theta)$ fournissent les informations sur les moments de la distribution de X.

La première dérivée de $\kappa(\theta)$ par rapport à θ, pour $p = 1$ permet de déterminer l'espérance de X. En effet,

$$\kappa'(\theta) = \frac{d}{d\theta}\kappa(\theta) = \mathbb{E}[X].$$

La deuxième dérivée de $\kappa(\theta)$ par rapport à θ, pour $p = 2$ permet de déterminer la variance de X. De même,

$$\kappa''(\theta) = \frac{d^2}{d\theta^2}\kappa(\theta) = \mathrm{Var}(X).$$

Ainsi, les dérivées successives de la fonction cumulante κ permettent de déterminer les cumulants de la distribution. Ils sont utilisés pour décrire la forme de la distribution (moyenne, variance, asymétrie, kurtosis, etc.).

Les trois paramètres de la loi de Tweedie p, μ et σ^2 (paramètre de dispersion) peuvent être écrits en fonction des paramètres des lois de Poisson et Gamma.

Le paramètre de puissance p détermine la nature exacte de la distribution de Tweedie et influence directement la forme de la fonction cumulante κ. Ainsi, lorsque sa valeur est modifiée, la forme de la distribution de Tweedie est différente. Chaque distribution possède une fonction cumulante κ et une relation moyenne-variance particulière.

- Pour $p = 0$: on obtient une distribution normale, où $\mathrm{Var}(Y) = \sigma^2$;
- Pour $p = 1$, on obtient une distribution de Poisson, où $\kappa(\theta) = e^{\theta}$ et $\mathrm{Var}(Y) = \sigma^2$;
- Pour $p = 2$, on obtient une distribution Gamma, où $\kappa(\theta) = -\log(-\theta)$ et $\mathrm{Var}(Y) = \sigma^2\mu^2$;
- Pour $1 < p < 2$: distribution Tweedie. Dans ce cas, la distribution est continue lorsque la variable réponse est positive ($Y > 0$) et avec une masse pour $Y = 0$;
- Pour $p = 3$: on obtient une distribution inverse gaussienne.

Pour les valeurs de $p \neq 1, 2$, la fonction $\kappa_p(\theta)$ prend des formes plus complexes en fonction de $\alpha = \frac{p-2}{p-1}$.

L'estimation de cet hyperparamètre est expliquée dans la section suivante qui permet de visualiser les hypothèses et données de modélisation. La complexité de la fonction de la densité de Tweedie génère des difficultés dans le calcul de la vraisemblance. Selon Akaike (1974), l'*Akaike Information Criterion* (AIC) est un critère utilisé pour la sélection de modèles, basé sur l'équilibre entre la qualité de l'ajustement du modèle et sa complexité. Il est calculé à partir de la vraisemblance maximisée du modèle choisi. Par conséquent, comme l'explique Wood (2006), pour les modèles de régression de type Tweedie, ce critère peut-être complexe à calculer. Afin de pallier ce problème, la déviance résiduelle est utilisée comme alternative pour évaluer la qualité d'ajustement du modèle. Une autre approche consiste à utiliser la validation croisée pour comparer les performances des modèles.

2.2 Mise en œuvre de la modélisation GLM

Afin de mettre en place la modélisation étudiée précédemment, il convient d'analyser les variables explicatives servant à prédire la variable cible.

2.2.1 Présentation des hypothèses de modélisation

Il convient de rappeler que le partenaire commercial étudié dans ce mémoire est un partenaire français vendant des voyages en France et à l'étranger. Les principaux marchés sont les marchés allemands et français. Enfin, il faut noter que les chiffres sont anonymisés à l'aide d'un coefficient multiplicateur afin de conserver la confidentialité de EA.

Ce voyagiste possède deux entités distinctes notées respectivement A et B. L'entité A générant la majorité du chiffre d'affaires du partenaire, les études sont orientées de manière à prendre cette information en compte.

Profondeur de l'historique

Les données prises en compte dans la modélisation se situent sur les périodes de avril 2022 aux dernières données disponibles.

Un des principaux défis de cette modélisation réside dans la qualité et la disponibilité des données historiques. En effet, l'épidémie de Covid-19 a considérablement perturbé les comportements de voyage des individus, rendant certaines parties de l'historique de données difficilement exploitables. Cette période de pandémie a introduit des anomalies et des variations exceptionnelles qui ne reflètent pas les tendances habituelles. En particulier, les données de 2020 et 2021, marquées par les confinements, les restrictions de voyage ont été exclues de l'étude puisqu'elles ne représentent plus la réalité des comportements des voyageurs. Les vagues successives de l'épidémie présentent des comportements atypiques qui peuvent biaiser les résultats si elles ne sont pas traitées avec soin. À ce titre, l'année 2022 présente des caractéristiques particulières mises en lumière dans la figure 2.2. La vague de Covid-19, Omicron, a rendu cette année très sinistrée en comparaison au nombre de polices souscrites.

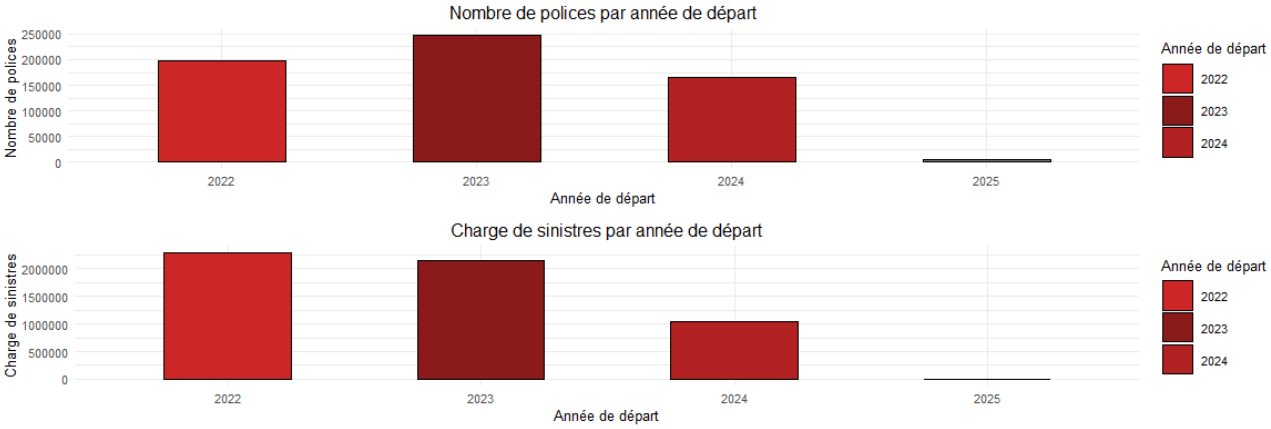

FIGURE 2.2 : Exposition du portefeuille par année de départ

À ce titre, il convient de rappeler que la garantie annulation proposée dans ce contrat ne fonctionne que sous conditions, sur présentation d'un justificatif. Ainsi, il n'est pas possible d'annuler son voyage par peur d'être contaminé. Afin de s'assurer que les prédictions réalisées sont fiables, l'ensemble

de la base de données est divisé en deux sous-ensembles distincts de données : l'ensemble d'entraînement des données et l'ensemble de test. Comme l'explique Zuer (2024) dans son article, l'ensemble d'entraînement correspond à l'ensemble qui permet de former le modèle. Il permet à l'algorithme d'apprentissage de s'ajuster aux données disponibles. Tandis que l'ensemble de test est destiné à mesurer la performance du modèle sur des données inconnues de celui-ci. Cet ensemble simule les conditions réelles et évalue la capacité de généralisation du modèle. Il permet de calculer l'erreur réalisée par le modèle. Dans cette étude, 70% des données sont tirées aléatoirement entre avril 2022 et avril 2024 pour l'ensemble d'entraînement et 30% pour l'ensemble de test. Cette séparation des deux ensembles est nécessaire pour éviter d'évaluer la qualité du modèle final en utilisant les mêmes données que celles qui ont servi à son entraînement. Dans ce cas, l'erreur réalisée par le modèle est minimale, alors qu'elle est plus élevée sur des données inédites.

Traitement des données manquantes

Avant de débuter la modélisation GLM, il convient de procéder à une explication des variables explicatives et un suivi des retraitements effectués.

- **Prise en compte de toute charge positive :** Il faut noter que la base de données contient une colonne de charge de sinistres. Lorsque celle-ci contient un montant positif, cela signifie que EA a réalisé un versement vers un assuré. Comme l'objectif est de prédire la charge de sinistres remboursée par EA, il convient de prendre en compte toute charge positive dans la modélisation. Bien que certaines polices sinistrées puissent contenir des erreurs de données, tout montant décaissé doit être inclus dans le modèle de prédiction.

  Les données de la base de données sont des données dites « ligne à ligne », chaque ligne concerne une police. Cette police est observable à différentes dates de vision. Dans ce mémoire, la date de vision correspond à la date à laquelle les données sont observées. Ainsi, l'exemple utilisé précédemment est repris. Une police est souscrite le 23 janvier lors de l'achat du voyage et la date d'annulation est le 2 juillet, soit 2 mois et 16 jours avant la date de départ, comme le montre la figure 1.5 expliquée dans le chapitre 1.

  Dans ce cas, à la date de vision du 30 juin 2024, il n'y a pas de développement de sinistres, tandis qu'à la date du 31 juillet 2024, la charge de sinistres a pris une certaine valeur. Cette charge de sinistres peut augmenter dans le temps selon le traitement de celui-ci. Cela implique que la valeur de la charge de sinistres à une date de vision ne peut être inférieure à la valeur à une date de vision suivante.

  Il faut noter que les conditions générales peuvent varier selon le pays d'achat de la police. Cela implique que les pénalités appliquées par le voyagiste ne sont pas identiques en fonction du pays d'achat de la police. Dans certains pays, le voyagiste prend en charge la totalité du coût lorsque la *cancellation window* est supérieure à un certain seuil. Tandis que dans d'autres pays, le voyagiste ne prend en charge que 80% du coût du voyage. Ainsi, lorsque le voyageur a souscrit une police d'assurance annulation, EA intervient et rembourse 20% du coût du voyage restant, sous réserve des conditions d'application de la garantie. Cette différence de condition d'application des pénalités selon le pays influe sur la distribution de la variable à expliquer. En effet, la distribution de l'IPTC souligne des pics de charge de sinistres aux niveaux des pourcentages de pénalités appliqués par le voyagiste.

- **Date de souscription manquante :** En outre, lorsque la date de souscription est manquante, elle est remplacée par une date déterminée à l'aide de la *booking window* moyenne, notée BW. Il convient de se souvenir que la BW correspond à l'intervalle de temps entre la date de souscription de la police d'assurance et la date de départ en voyage.

- ***Trip cost* manquant :** Un des retraitements majeurs réalisés dans la base de données est la correction des montants de *trip costs* manquants. La méthodologie de cette correction est d'estimer le montant du voyage manquant grâce au montant de sinistres remboursés par l'assureur à l'assuré. Il est à noter que cette méthode ne permet d'estimer que les valeurs manquantes en cas de sinistres puisque le coût du sinistre permet d'estimer le montant du *trip cost*. Dans ce cas, le *trip cost* moyen est utilisé. L'inconvénient est que cette méthode ne permet pas de capter les différences d'impacts de changement de pourcentage de couverture sur l'aléa moral et l'antisélection. Comme cet impact n'est pas quantifiable facilement, le choix est fait de prendre cette modélisation simple qui permet tout de même d'améliorer la base de données initiale.

  Cette correction augmente le montant du *trip cost*. Ainsi, elle a pour conséquence une baisse des pourcentages d'IPTC attendus par rapport à ceux estimés dans la méthode actuelle. Ce retraitement est réalisé sur les lignes de données concernant les polices sinistrées. Il s'agit d'utiliser la valeur de la CW et du montant de charge de sinistres, à partir de la relation

$$\begin{aligned} \text{Charge de sinistres} &= \text{pourcentage de pénalité} \times \textit{Trip cost} \\ &\Leftrightarrow \\ \textit{Trip cost} &= \frac{\text{charge de sinistres}}{\text{pourcentage de pénalité.}} \end{aligned}$$

  La valeur de la *cancellation window* (CW) permet de connaître le pourcentage de pénalité à appliquer sur la police dont le *trip cost* est manquant. Il convient de rappeler que la CW correspond à l'intervalle de temps entre la date d'annulation du voyage (réalisation du sinistre) et la date de départ.

  Ainsi, une estimation des *trip costs* manquants est obtenue, à l'exception des polices sinistrées ayant une CW trop élevée et achetée dans un pays qui n'applique pas de pénalité avant un certain seuil de CW. À ce titre, la structure des données s'observe dans le tableau 2.2. Pour illustrer

***Trip costs* estimés**	***Trip costs* initiaux**	**Charge de sinistres**
50 752 425	50 752 425	149 149
220 355	0	214 175

TABLE 2.2 : Tableau des données de *trip costs* estimés

  ce propos, il faut reprendre l'exemple précédent. Dans ce cas, la CW est égale à 2 mois et 16 jours. Cette durée est catégorisée dans les longues CW et pour certains pays d'achat de polices, le voyagiste rembourse entièrement le voyage. Dans d'autres pays, cette valeur de la CW implique une pénalité avec un montant fixe, indépendant du *trip cost*. Dans les deux cas cités, il n'est pas possible de déterminer le montant du *trip cost* manquant et associé à la police sinistrée, car il est indépendant de la pénalité appliquée. De même, il n'est pas possible de déterminer les montants de *trip costs* pour les polices non sinistrées.

  À titre d'exemple, dans la figure 2.3, une structure de ce type est observée pour une des entités du partenaire.

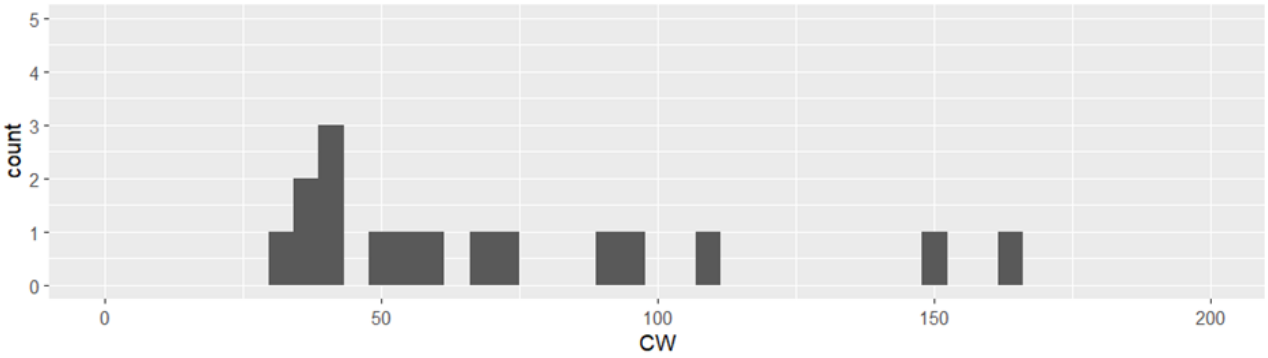

FIGURE 2.3 : Répartition des données manquantes pour les *trip costs* pour les longues *cancellation windows*

Afin de faciliter la lecture des données, les *trip costs*, les BW et les durées de voyage sont groupés en fonction des quantiles de leur distribution afin que chaque catégorie soit équirépartie en termes d'exposition.

- **Police manquante :** À ces retraitements s'ajoute la correction des données qui ne possèdent pas de police. Dans ce cas, la ligne de police est nulle alors qu'une charge de sinistres est indiquée dans la colonne associée. Comme souligné précédemment, il faut prendre en compte cette charge puisqu'elle a réellement été versée. Il s'agit en réalité d'une police annulée, concernant seulement 0,11% des données.
- ***Booking window* négative :** Enfin, la gestion des *booking windows* négatives impliquant la souscription d'une police d'assurance après la date de départ de celui-ci, se fait en rapportant à 0. Cela concerne un nombre très faible de polices avec une faible exposition en termes de charge de sinistres.

2.2.2 Présentation des variables de souscription

Afin de déterminer la valeur de la prime pure, c'est-à-dire de réaliser la tarification du produit d'assurance, des variables disponibles dès la souscription au contrat d'assurance sont utilisées. Les données sont observées à une date de vision (date d'extraction de celles-ci)du 31 juillet 2024. Il faut rappeler que pour le partenaire A, 96% de la charge de sinistres est créée par des sinistres annulation et pour le partenaire B, cela concerne 83% de la charge de sinistres. Au total, pour le partenaire commercial de EA, 95% de la charge de sinistres est due aux sinistres annulation. Or, la charge de sinistres en assurance annulation dépend du coût du voyage, puisque l'assureur rembourse le coût du voyage annulé à l'assuré. Il semble donc pertinent d'utiliser cette variable comme variable à prédire, autrement dit la variable cible.

Compte tenu des retraitements décrits précédemment, il faut noter que pour le partenaire B, les données de *trip cost* manquantes correspondent à un montant de charges plus important que pour les polices ayant un *trip cost* affecté dès le départ. Dès lors, l'hypothèse que la charge de sinistres perd en corrélation avec le *trip cost* est posée.

En présence d'un sinistre, pour le partenaire commercial, le *trip cost* et l'IPTC sont corrélés à 65%. La valeur de cette corrélation s'explique par :

- la présence de valeur d'IPTC supérieures à 1 car EA rembourse des prestations complémentaires en addition au coût du voyage ;

- un nombre important de données manquantes pour les *trip costs*.

Dans un second temps, les variables suivantes sont sélectionnées :

- Mois de départ ;
- Mois de souscription ;
- Groupe de BW ;
- Groupe de *trip cost* ;
- Groupe de durée du voyage ;
- Pays d'achat de la police.

Une analyse de ces variables est réalisée dans l'objectif de comprendre leur structure et leurs relations entre elles.

Année et mois de départ

Les mois de départ sont observés à partir d'avril 2022 jusqu'aux données dernièrement disponibles, soit juillet 2024. Le mois de départ prend des valeurs de 1 à 12 correspondant au mois de départ dans l'année. Le graphique de la figure 2.4 reflète une augmentation des volumes de souscription et par conséquent des sinistres en 2022, expliquée en partie par la vague Omicron. Un retour aux comportements habituels des consommateurs en 2023 est observé. Ainsi, le début de l'année 2022 est anormalement sinistrée.

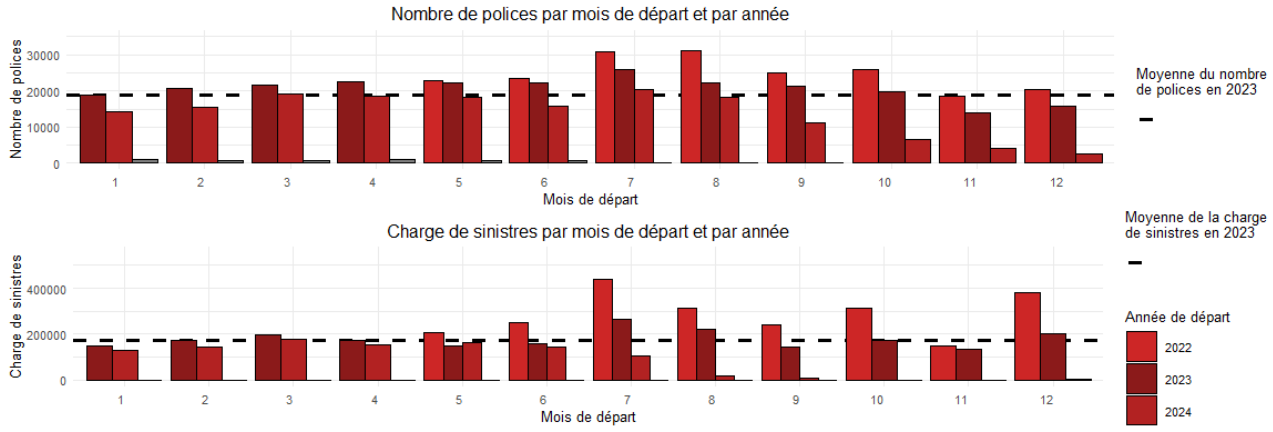

FIGURE 2.4 : Exposition selon l'année et le mois de départ

Les mois de départ les plus sinistrés sont les mois correspondant aux vacances d'été et aux fêtes de fin d'année. En comparant ces données au nombre de polices par mois, il faut remarquer que les mois d'été sont des mois contentant beaucoup de polices. Tandis que les autres mois de l'année sont relativement stables à l'exception de quelques mois comme le mois de novembre.

L'écart de sinistralité entre les différents mois de départ témoigne de l'importance de prendre cette variable en compte dans la modélisation.

Dans le modèle final, l'année de départ n'est pas une variable explicative car celui-ci utilise les données des 12 derniers mois glissants pour prédire les valeurs de la charge de sinistres. Ce raisonnement est expliqué en détail dans la partie concernant les résultats du modèle.

Mois de souscription

Le mois de souscription prend des valeurs de 1 à 12 correspondant au mois de souscription de la police d'assurance. Il est nécessaire de rappeler que celui-ci est identique au mois d'achat du voyage. Les données de la base concernent les voyages dont la date de départ se situe entre avril 2022 et juillet 2024. Ainsi, comme le montre la figure 2.5, ces données ont des dates de souscription qui s'étendent de 2020 à 2024, définissant ainsi un spectre large de BW.

Comme expliqué précédemment, le nombre de souscriptions et de sinistres en janvier 2022 est anormalement élevé à cause de la vague Omicron de la pandémie du Covid-19. En outre, les voyageurs ont pris conscience et appris à utiliser à bon escient la garantie annulation comme le montre la figure suivante. Les polices d'assurance souscrites en 2022 sont particulièrement sinistrées en comparaison aux autres années.

Néanmoins, une distinction entre les mois de souscription apparaît sur cette figure témoignant de l'importance de considérer cette variable dans le modèle final.

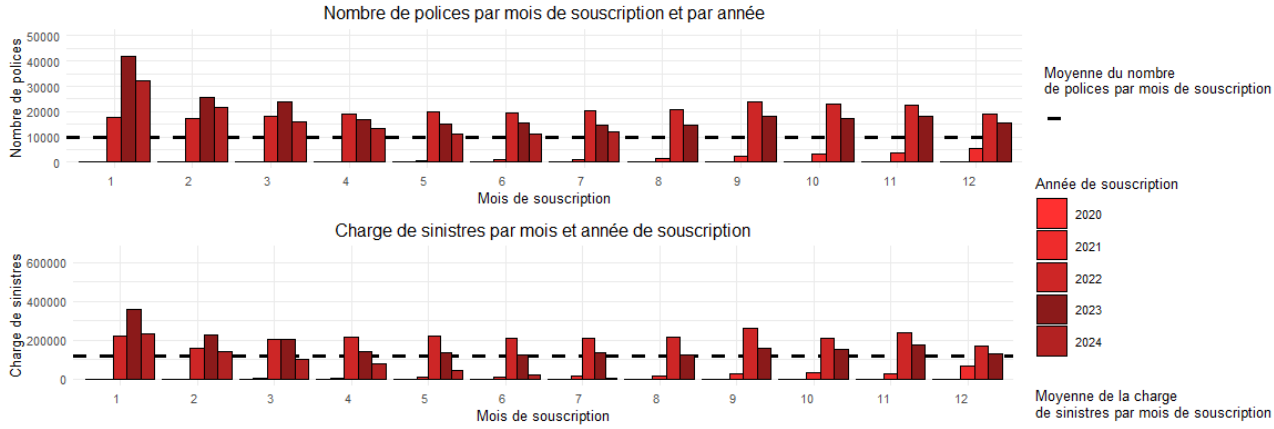

FIGURE 2.5 : Exposition selon le mois de souscription

Groupe de BW

Les *booking windows*, intervalle entre la date de souscription de la police d'assurance et la date de départ, permettent d'étudier le comportement des voyageurs qui souhaitent s'assurer. La figure 2.6 met en avant que les assurés achètent leur police majoritairement plus d'un mois avant la date de départ en voyage. Sans surprise, ces BW sont aussi les plus sinistrées, car plus cet intervalle est long, plus la probabilité qu'un événement survienne est important.
Un nombre non négligeable de polices souscrites 7 jours ou moins avant le départ et un nombre relativement faible de sinistres sur cet intervalle de temps est noté.

Enfin, le nombre de polices souscrites en 2022 dépassant de manière significative le nombre de polices souscrites en 2023, l'hypothèse que les voyageurs expriment un besoin de couverture plus important dans un contexte de crise sanitaire est posée.

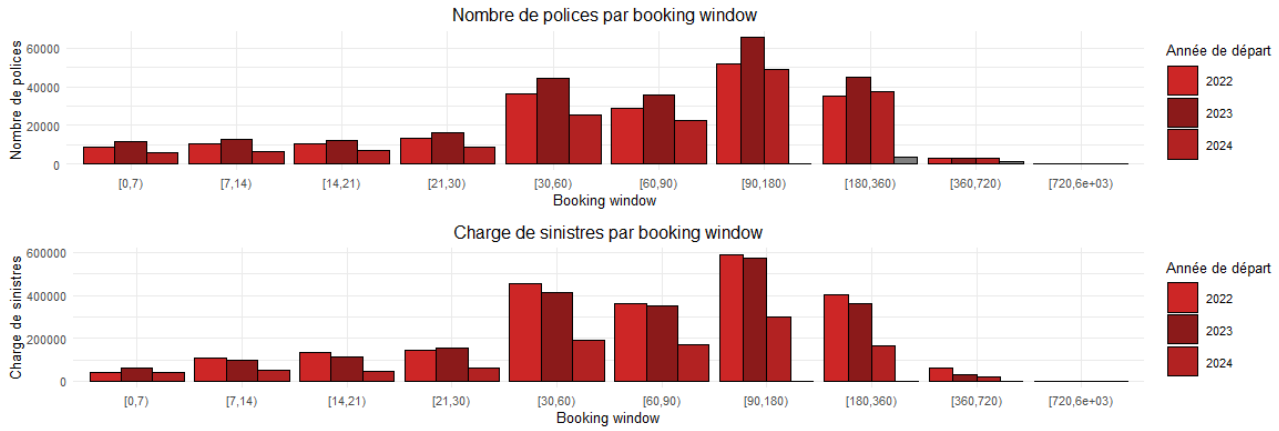

FIGURE 2.6 : Exposition selon la *booking window*

Afin d'analyser précisément les différences d'exposition en termes de polices et de sinistres entre 2022 et 2023, la figure 2.7 met en lumière le ratio pour ces deux grandeurs entre les deux années. Ainsi, pour une *booking window* de 14 à 21 jours, l'année 2022 est plus de 2,5 fois plus sinistrés que l'année 2023. La longueur de la *booking window* semble être une variable importante pour estimer la sinistralité en assurance annulation.

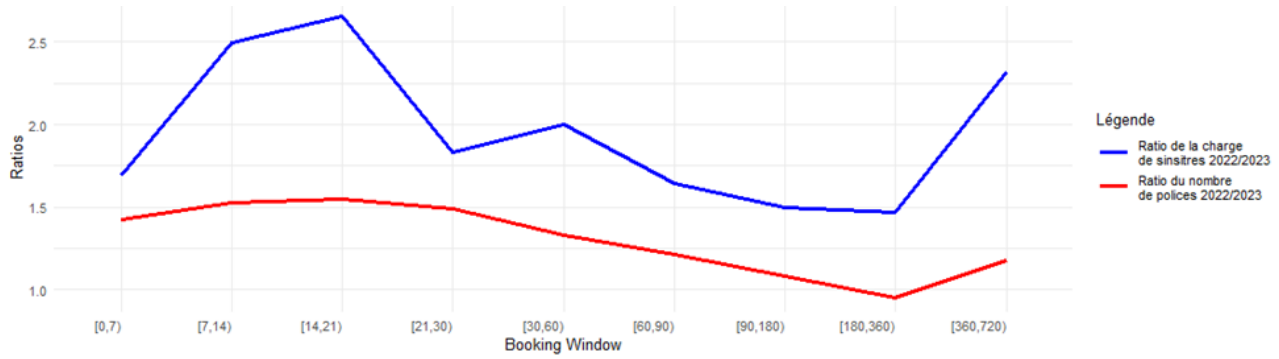

FIGURE 2.7 : Ratio des grandeurs d'exposition entre 2022 et 2023

Groupe de *trip cost*

Les primes de la police d'assurance sont calculées en fonction d'une tranche de *trip cost*. Il est donc pertinent d'analyser la répartition de ces tranches de coût du voyage. Afin de préserver la confidentialité des données, les détails numériques de cette grille ne sont pas dévoilés dans ce mémoire. Il faut noter que le *trip cost* contient un certain nombre de données manquantes qui sont estimées lorsqu'il y a présence d'un sinistre. Un premier aperçu du graphique mettant en lumière le nombre de polices pour chaque catégorie de coût du voyage montre que trois cases de la grille sont particulièrement représentées. Il y a peu de polices pour des voyages à coûts faibles. Il est intéressant de noter que Europ Assistance ne possède que les données des voyageurs ayant souscrits une police d'assurance. Le profil du partenaire commercial est tel qu'il faut comprendre qu'en majorité, les voyageurs réservant un voyage à faible coût ne se couvrent pas contre un risque d'annulation. Il faut remarquer aussi que plus le coût du voyage augmente, plus la fréquence d'apparition des polices dans une case de la grille diminue.

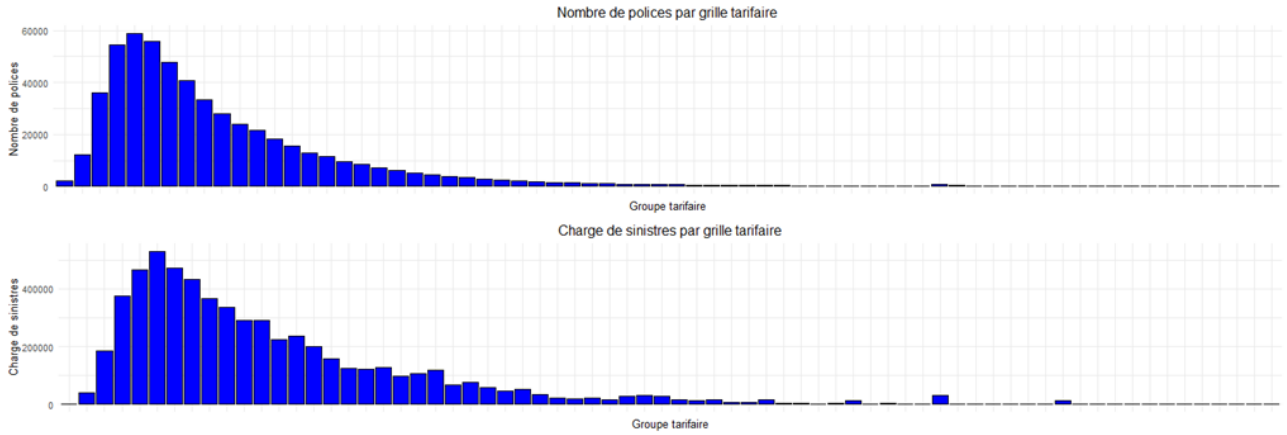

FIGURE 2.8 : Exposition selon la tranche de *trip cost*

Il est intéressant de comparer le graphique de la fréquence des polices par catégorie de la grille avec la fréquence des sinistres. Cette comparaison montre que la charge de sinistres diminue moins vite que le nombre de polices par catégorie. Quelques pics de charge de sinistres se dénotent pour des coûts de voyage plus élevés car le montant de remboursement dépend de ce coût du voyage, mettant en exergue un effet prix. Néanmoins, il faut noter que les grilles les plus sinistrées sont celles dont le montant du coût du voyage est faible. Lorsque le nombre de polices est important, la charge de sinistres augmente par un effet volume.

Le groupe tarifaire de la police est donc une variable à prendre en compte. Dans cette étude, les *trip costs* sont regroupés de manière à obtenir une exposition identique dans chaque groupement, comme le montre la figure 2.9. Ainsi, il y a une répartition homogène des polices en fonction du coût du voyage. Les deux groupes aux extrémités (moins et plus onéreux), ont un nombre légèrement plus élevé d'assurés.

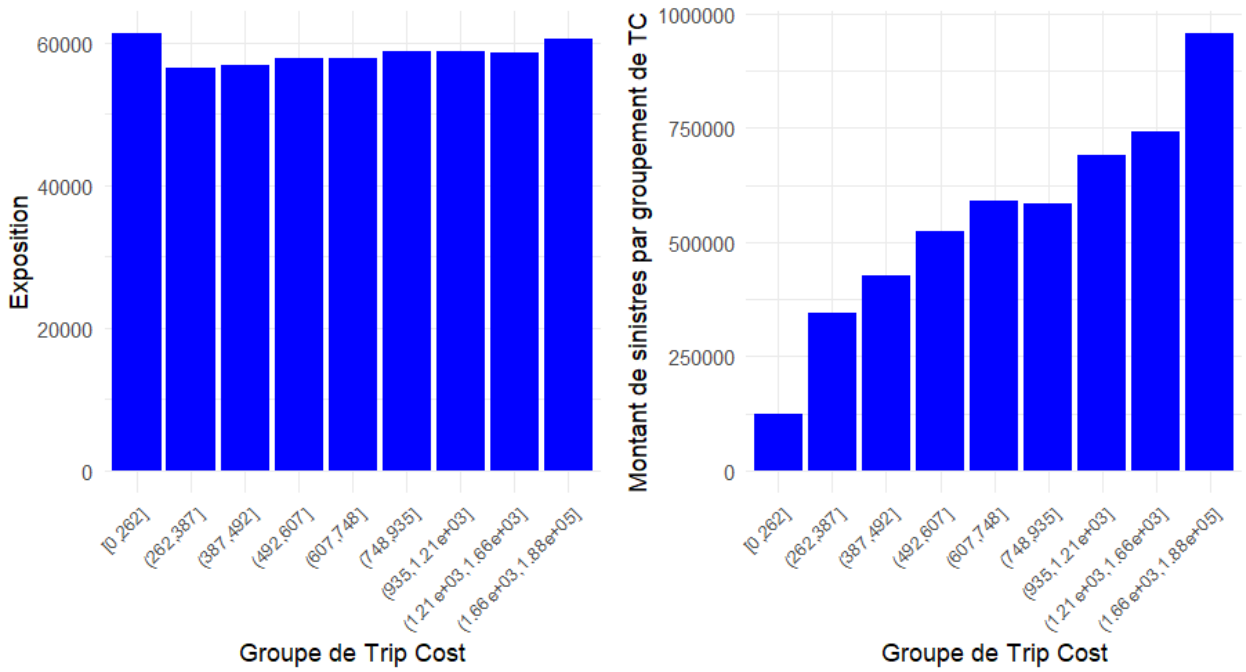

FIGURE 2.9 : Exposition des intervalles de *trip cost*

Dans cette figure, une analyse graphique indique une nette augmentation du montant des sinistres à mesure que le coût du voyage augmente. Cette observation est attendue puisque les tranches sont exprimées en montant et les voyages les plus onéreux sont associés aux sinistres les plus coûteux. Autrement dit, le graphique compare les volumes de charges de sinistres en valeur absolue et non en valeur relative*. Ainsi, ce graphique n'est pas un outil utilisé pour comparer la rentabilité des segments. Cela est possible uniquement avec une mesure relative comme la fréquence ou l'IPTC. Ainsi, il est normal d'observer que les groupes de *trip cost* les plus sinistrés sont les groupements concernant les voyages les plus onéreux.

Ces deux graphiques permettent de souligner l'importance de cette variable dans la modélisation mise en place. Bien que l'exposition soit à peu près homogène entre les différentes catégories de la grille de *trip cost*, les montants des sinistres augmentent de manière significative pour les voyages les plus chers. Autrement dit, bien qu'il n'y ait pas un volume important sur les tranches de *trip cost* onéreuses, ces tranches sont les plus risquées pour l'assureur à cause du montant du coût du voyage à rembourser en cas de sinistre.

Groupe de durée du voyage (td)

L'étude de la durée du voyage, grâce au graphique 2.10, montre que la majorité des polices d'assurance annulation sont souscrites pour des voyages qui ont lieu entre 4 et 8 jours. Néanmoins, l'intervalle de durée de voyage le plus sinistré est celui concernant les voyages de 8 à 14 jours. Ce résultat semble cohérent car ces voyages sont plus onéreux que ceux cités précédemment, incitant alors les voyageurs à se prémunir du risque d'annulation. Ce graphique montre aussi que les voyages dits "courts" entre 2 et 4 jours possèdent une faible charge de sinistres.

*Le lecteur peut se référer à l'explication donnée dans la partie 1.2.2 du premier chapitre sur l'intérêt d'utiliser l'IPTC comme valeur cible.

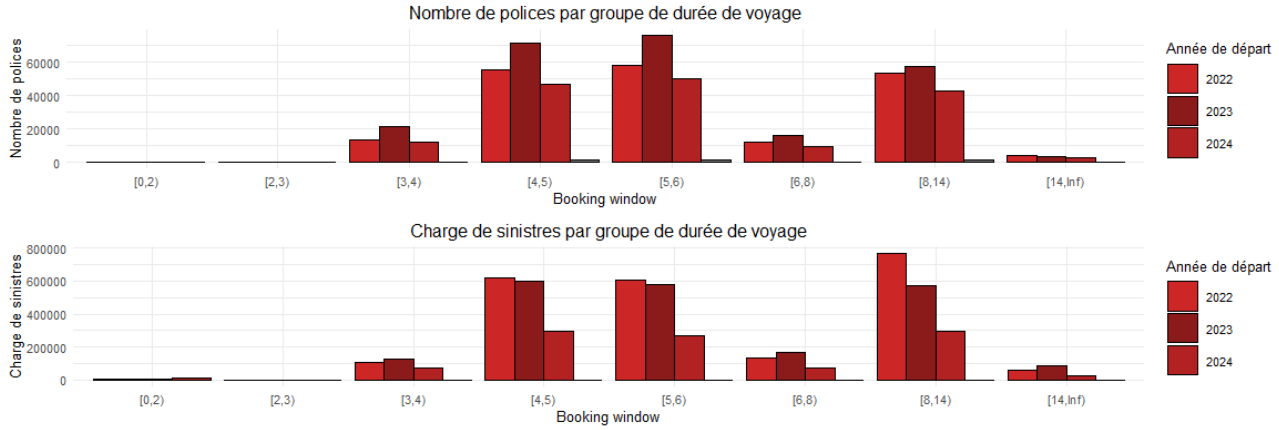

FIGURE 2.10 : Exposition des intervalles de durée de voyage

La différence de sinistralité selon les intervalles de durée de voyage incite à considérer cette variable dans la modélisation de la charge de sinistres.

Pays d'achat de la police

Le partenaire commercial de EA vend majoritairement ses polices d'assurance en Allemagne et en France, comme le met en avant la figure 2.11. Le pays le plus sinistré est l'Allemagne. La disparité de la sinistralité entre les pays d'achat des polices implique de devoir analyser les données en les segmentant par pays.

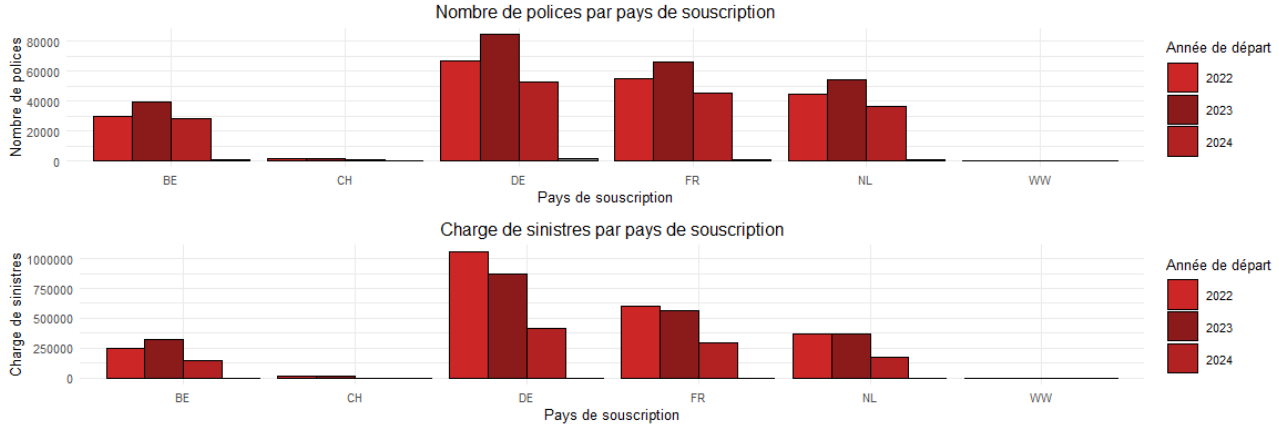

FIGURE 2.11 : Évolution de l'exposition selon le pays d'achat de la police

Il convient de remarquer que le portefeuille allemand a été particulièrement sinistré en 2022 et semble être moins rentable que les autres pays de souscription.

Ainsi, toutes les variables présentées ci-dessus sont les variables dont les informations sont disponibles lorsqu'un voyageur souscrit une police d'assurance. Afin qu'elles soient utilisées dans une modélisation GLM, il faut analyser et quantifier les relations entre ces différentes variables qualitatives.

2.2.3 Évaluation des corrélations entre les variables dans le modèle de tarification

Afin de connaître le lien entre ces variables, il convient de mettre en œuvre plusieurs tests permettant de déterminer et quantifier les liens entre elles.

V de Cramer

Dans un premier temps, le V de Cramer permet de mesurer l'intensité des relations entre deux variables. Il est utilisé pour analyser la structure des données. À ce titre, il est nécessaire de rappeler les bases théoriques du V de Cramer, en s'appuyant sur les notes du cours dispensé par Grasland (2000).

D'après la page Wikipédia (2024a) dédiée au V de Cramer, cette mesure statistique a été introduite par Harald Cramer en 1946 et permet de mesurer l'association entre deux variables qualitatives. Il prend des valeurs comprises entre 0 (pas d'association entre les variables) et 1 (association totale entre les variables). En outre, il permet une alternative au coefficient de corrélation pour analyser les relations entre différentes variables catégorielles.

Le V de Cramer utilise un tableau de contingence indiquant la fréquence pour chaque facteur. Afin d'obtenir un résultat généralisable à toute taille de tableau de contingence, il faut calculer la valeur de la statistique du Chi-2 et la normaliser en tenant compte de la taille de l'échantillon et des dimensions du tableau (nombre de modalités prises par chaque facteur). Le tableau de contingence pour les variables X possédant k modalités et Y avec p modalités est donnée dans le tableau 2.2.3.

	Y_1	$\cdots$	Y_j	$\cdots$	Y_p	**Total**
X_1	N_{11}	$\cdots$	N_{1j}	$\cdots$	N_{1p}	$N_{1.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_i	N_{i1}	$\cdots$	N_{ij}	$\cdots$	N_{ip}	$N_{i.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_k	N_{k1}	$\cdots$	N_{kj}	$\cdots$	N_{kp}	$N_{k.}$
Total	$N_{.1}$	$\cdots$	$N_{.j}$	$\cdots$	$N_{.p}$	$N_{..}$

Calcul de la statistique de Chi-2 (χ^2) : La statistique de test du χ^2 évalue le degré du lien entre deux variables grâce à une métrique appelée l'écart à l'indépendance. Après avoir réalisé le tableau de contingence, il s'agit de calculer les écarts à l'indépendance pour chaque case du tableau de contingence. L'écart à l'indépendance est la différence entre l'effectif observé et l'effectif théorique divisé par l'effectif total du tableau de contingence. L'effectif théorique est l'effectif qui serait observé s'il y avait indépendance entre les deux modalités étudiées de X et Y, si les modalités étaient attribuées de manière totalement indépendante.

- Calcul des effectifs théoriques : $N^*_{ij} = \frac{N_{i.} \times N_{.j}}{N_{..}}$. L'écart à l'indépendance permet de comprendre la forme de cette éventuelle relation entre ces deux modalités étudiées.
- Calcul des écarts à l'indépendance : $e_{ij} = (N_{ij} - N^*_{ij})$

  L'écart entre l'effectif théorique et l'effectif observé peut résulter de fluctuations aléatoires ou d'un lien entre les deux modalités. Cet écart représente l'écart entre l'effectif réellement observé et l'effectif en cas d'indépendance. Ainsi, plus cet écart est important, plus il est probable qu'il existe une dépendance entre ces modalités. Lorsque l'écart est positif, cela signifie que l'effectif observé est supérieur à l'effectif théorique, indiquant une sur-représentation de la population

dans cette catégorie. À l'inverse, si l'écart est négatif, l'effectif observé est inférieur à l'effectif théorique, indiquant une sous-représentation de la population dans cette catégorie. Le test de χ^2 permet de connaître la raison de l'écart entre effectif théorique et observé. Il se calcule de la manière suivante pour chaque case du tableau de contingence. Il faut alors calculer les χ^2 dits "locaux".

- Calcul des χ^2 locaux : $\chi^2_{ij} = \frac{(N_{ij}-N^*_{ij})^2}{N^*_{ij}}$. Comme pour l'écart à l'indépendance, plus la valeur du χ^2 local est importante, plus les deux modalités semblent avoir un lien significatif entre elles. La statistique du χ^2 se détermine en sommant tous les χ^2 locaux et en le divisant par le degré de liberté. Le degré de liberté est déterminé par la multiplication suivante, $(k-1) \times (p-1)$, où k et p correspondent aux nombres de modalités prises par chaque variable.

Calcul du V de Cramer : Après avoir calculé la statistique du χ^2, il est possible de procéder au calcul du V de Cramer. Cette mesure se détermine grâce à la relation 2.18.

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, p-1)}} \tag{2.18}$$

Il faut rappeler que le test du χ^2 est utilisé pour tester l'hypothèse d'indépendance entre deux variables catégorielles en mesurant l'écart entre les effectifs observés et les effectifs théoriques s'il y avait indépendance entre ces variables. Lorsque la valeur de la statistique est élevée, il est probable qu'il existe une dépendance significative entre les variables. Dans ce cas, les différences observées ne sont pas dues aux fluctuations aléatoires. Cette métrique témoigne ainsi de l'existence d'une relation entre les variables étudiées.

Le V de Cramer permet de mesurer l'intensité de l'association entre deux variables catégorielles. Il est calculé à partir de la valeur du χ^2 et est normalisé pour prendre en compte la taille de l'échantillon. Par conséquent, il ne dépend pas de sa taille. Cette mesure permet donc de quantifier le lien mis en avant par la statistique du χ^2. Dans cette étude, la matrice du V de Cramer correspond à la matrice présentée dans la figure 2.12.

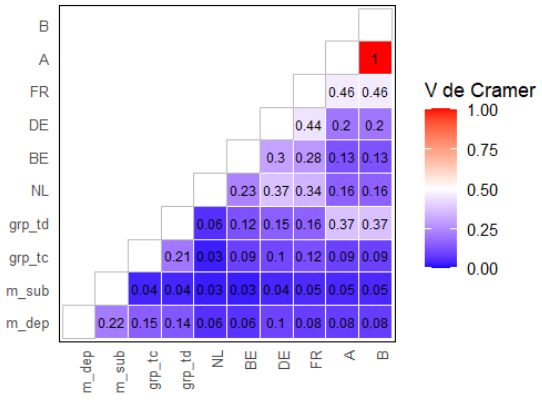

FIGURE 2.12 : Matrice de Cramer

Dans un premier temps, il faut remarquer que les variables qualitatives ne sont pas corrélées entre elles. Ainsi, il semble qu'elles peuvent être conservées dans le modèle. Le tableau met en avant quelques liens entre nos variables comme le mois de souscription et le mois de départ ou les pays d'achat de polices France et Allemagne. Comme 62.42% des polices de ce partenaire commercial sont achetés dans

ces deux pays, la corrélation entre ces deux variables est plus forte qu'avec les autres pays d'achat de polices.

Il est important de noter que les variables `A` et `B` sont totalement corrélées puisqu'elles prennent la valeur 0 ou 1 en fonction de l'entité dans laquelle a été achetée la police. Il convient de conserver cette décomposition pour le V de Cramer afin d'observer l'intensité du lien pour chaque variable selon les entités.

L'analyse descriptive des données se poursuit avec une étude de la multicolinéarité.

Etude de la multicolinéarité

Présentation de la multicolinéarité Afin de tester la multicolinéarité de nos variables explicatives, les valeurs VIF sont calculées pour chaque variable du modèle. Comme l'explique l'article écrit par Anderson (2024c), la *Variance Inflation Factor* (VIF) est une mesure utilisée pour détecter la multicolinéarité entre les variables explicatives dans un modèle de régression linéaire. Cette situation se produit lorsqu'une variable explicative dans un modèle de régression peut être prédite de manière linéaire à partir des autres variables explicatives. En somme, une variable explicative s'exprime comme combinaison linéaire d'autres variables explicatives. Dans ce cas, ces variables apportent des informations redondantes au modèle. Elles se chevauchent dans leur capacité à expliquer la variance de la variable dépendante. Cette colinéarité affecte les coefficients de régression en les rendant instables.

La VIF d'une variable explicative X_i est définie comme

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}, \tag{2.19}$$

où R_i^2 est le coefficient de détermination obtenu par la régression de la variable X_i sur les autres variables explicatives du modèle.

Concrètement, pour chaque variable explicative X_i du modèle, une régression est effectuée en utilisant X_i comme variable dépendante et les autres variables comme variables explicatives. Le coefficient de détermination R_i^2 de cette régression mesure la proportion de la variance de X_i expliquée par les autres variables. Plus R_i^2 est élevé, plus X_i est colinéaire avec les autres variables. Ainsi, la valeur de la VIF est élevée.

Par conséquent, les valeurs de la VIF s'interprètent de la manière suivante :

- VIF proche de 1 : il n'y a pas de multicolinéarité entre la variable X_i et les autres variables explicatives ;
- VIF comprise entre 1 et 5 : il y a une multicolinéarité modérée ;
- VIF supérieure à 5 : il y a une forte multicolinéarité.

Conséquences de la multicolinéarité Lorsqu'il y a multicolinéarité entre des variables explicatives du modèle, les coefficients de régression sont instables. En d'autres termes, les valeurs de ces coefficients peuvent changer de manière significative lorsque les données varient. En conséquence, les erreurs standards des coefficients augmentent réduisant la précision des estimations.

En cas de valeurs de VIF importantes, certaines variables peuvent être supprimées ou transformées afin réduire la multicolinéarité du modèle.

La mesure notée GVIF permet de généraliser la mesure du VIF lorsque les variables possèdent plusieurs degrés de liberté. Dans cette étude, les variables catégorielles ont plusieurs modalités entraînant plusieurs degrés de liberté. Cette mesure permet de déterminer le niveau de multicolinéarité entre chaque variable.

La transformation $\text{GVIF}^{1/(2*\text{Df})}$ permet d'ajuster la mesure du GVIF pour rendre la comparaison entre les variables plus limpide, en prenant en compte les degrés de liberté de la variable. Cette mesure s'interprète d'une manière identique à la VIF.

Ainsi, la table 2.3 indique que la tranche tarifaire et le *trip cost* sont des variables colinéaires. La variable du *trip cost* est retirée du modèle puisque l'information est comprise dans la variable regroupant la tranche tarifaire.

Les résultats du modèle montrent les valeurs du VIF pour chaque variable du modèle. Le VIF mesure la quantité de multicolinéarité dans un modèle de régression. Des valeurs élevées de VIF (> 5) indiquent une forte multicolinéarité.

Variables	**GVIF**	**Df**	$\mathbf{GVIF^{1/(2*Df)}}$
`m_dep`	1,442	11	1,017
`pol_client`	1,512	1	1,230
`TC_ret`	5,252	1	2,292
`grp_td`	2,018	7	1,051
`grp_bw`	1,253	9	1,013
`grp_tc`	6,653	9	1,111
`NL`	23,335	1	4,831
`BE`	18,398	1	4,289
`DE`	29,461	1	5,428
`FR`	28,017	1	5,293

TABLE 2.3 : Analyse de la colinéarité des variables

Dans cette étude, les variable `grp_tc` et `TC_ret` ont une VIF élevée indiquant une forte multicolinéarité. Ces deux variables apportent une information redondante. En outre, les variables concernant les pays ont une VIF relativement élevée.

En retirant la variable `TC_ret`, les nouvelles valeurs des VIF sont répertoriées dans la table 2.4. Celle-ci met en lumière les variables permettant de réaliser une régression de Tweedie. Les variables concernant les pays sont conservées car la métrique $\text{GVIF}^{1/(2*\text{Df})}$ dépasse légèrement la valeur du seuil fixé à 5. Néanmoins, l'information apportée par ces variables est pertinente. En couplant cette analyse avec le V de Cramer, il convient de conserver ces variables afin de prendre en compte les différences de comportements entre les pays de souscription.

Récapitulatif des variables explicatives

Les différentes études menées ont permis de réaliser un tri des variables explicatives et de les sélectionner pour la modélisation GLM. Le tableau 2.5 permet de synthétiser les variables retenues pour la modélisation de Tweedie.

Variables	GVIF	Df	$GVIF^{1/(2*Df)}$
m_dep	1,440	11	1,017
pol_client	1,511	1	1,229
grp_td	1,994	7	1,051
grp_bw	1,252	9	1,013
grp_tc	1,734	9	1,031
NL	23,335	1	4,831
BE	18,398	1	4,289
DE	29,461	1	5,428
FR	28,017	1	5,293

TABLE 2.4 : Analyse de la colinéarité des variables

Variables	Nom de modélisation	Variables explicatives
Entité	pol_client	Oui
Mois de départ	m_dep	Oui
Mois de souscription	m_sub	Oui
Année de départ	y_dep	Non
Groupe de durée de voyage	grp_td	Oui
Groupe de *booking window*	grp_bw	Oui
Groupe de *trip cost*	grp_tc	Oui
Trip cost	TC_ret	Non
Pays-Bas	NL	Oui
Belgique	BE	Oui
France	FR	Oui
Allemagne	DE	Oui

TABLE 2.5 : Récapitulatif des variables explicatives

2.3 Analyse des résultats de la modélisation

2.3.1 Analyse des résultats de la modélisation de Tweedie

Il est important de valider ce modèle avec un ensemble de données de test indépendant et de comparer ses performances avec d'autres modèles potentiels. Les diagnostics des résidus doivent être examinés pour s'assurer que les hypothèses du modèle sont respectées et qu'il n'y a pas de problèmes de surajustement ou de points influents.

Détermination des hyperparamètres

Les hyperparamètres principaux de la modélisation Tweedie sont le paramètre de puissance p, le paramètre de dispersion σ^2 et le choix de la fonction de lien g. La bonne sélection et estimation de ces paramètres permet d'obtenir un modèle Tweedie efficace et précis.

Les distributions Tweedie sont régulièrement utilisées en assurance grâce à la présence d'une masse en zéro et d'une distribution positive et continue sur l'ensemble de définition. En outre, elle permet de capter des queues lourdes et possède une fonction variance hétérogène représentant mieux la réalité des écarts entre les sinistres. La valeur du paramètre p rend cette distribution flexible, ce qui permet

de s'adapter à un large panel de données.

En assurance, l'utilisation de cette distribution permet de prendre en compte directement la prime pure sans utiliser un modèle de coût moyen/fréquence en modélisant le montant total de sinistres (le montant de pertes total). Il s'agit dans ce cas de faire un modèle unique grâce à un GLM avec une fonction de lien log et dont la composante déterministe est une distribution de Tweedie.

La figure 2.13 montre une grande hétérogénéité dans les variances pour différentes moyennes, avec certaines variances proches de zéro et d'autres beaucoup plus élevées, une distribution Tweedie pourrait s'appliquer, car elle capture à la fois les valeurs nulles ou proches de zéro et les variances élevées observées dans d'autres sous-groupes.

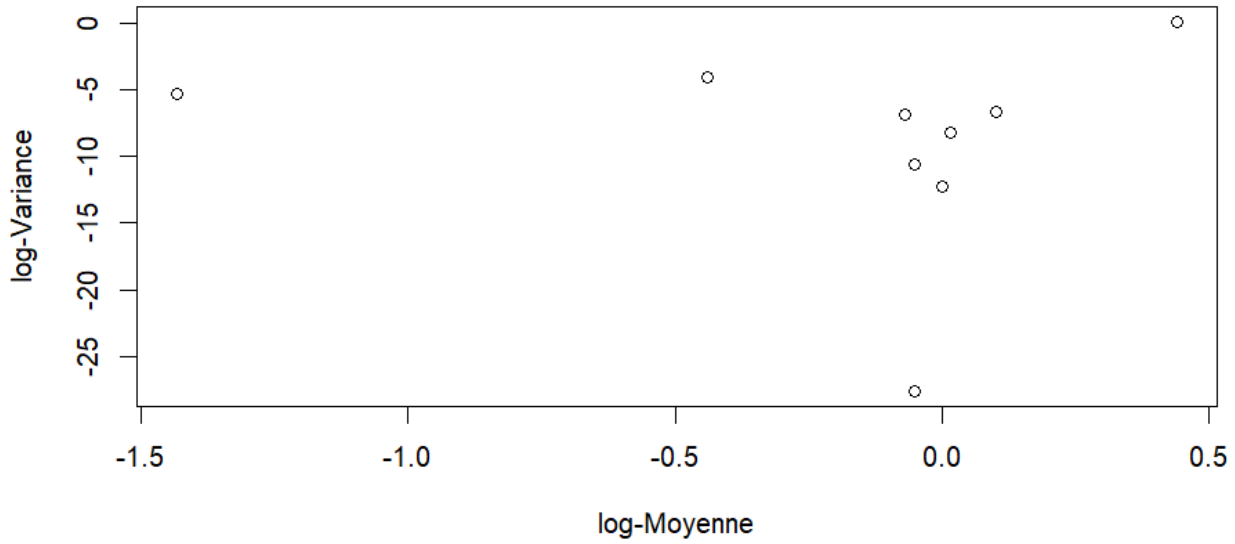

FIGURE 2.13 : Relation entre la log-moyenne et la log-variance de l'IPTC

Dans cette étude, la distribution de la variable `IPTC_obs` est fortement asymétrique, avec une concentration élevée de valeurs proches de zéro et quelques valeurs beaucoup plus élevées, comme le montre la figure 2.14. Le graphique met en avant un nombre significatif de valeurs très élevées, indiquant une longue queue à droite.

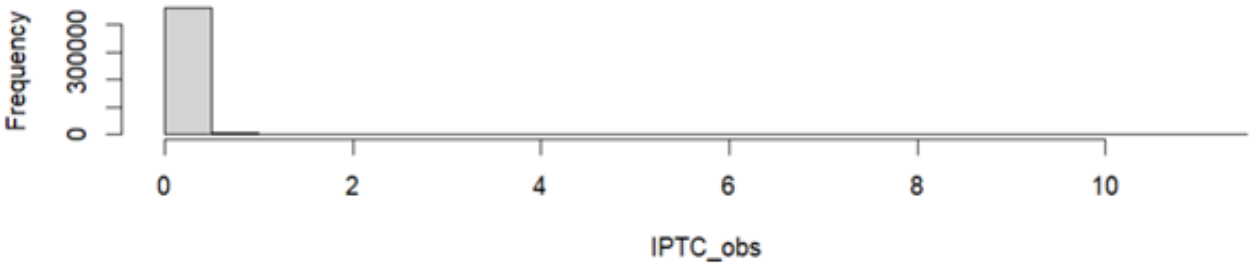

FIGURE 2.14 : Distribution de l'IPTC

Afin de déterminer la valeur optimale des paramètres, il convient de se servir de la fonction `tweedie.profile` pour estimer le paramètre de puissance p optimal pour un modèle de régression Tweedie.

La log-vraisemblance d'une observation y donnée s'écrit, avec $c(y, \sigma^2, p)$ est une constante de normalisation qui dépend de y, σ^2 et p,

$$\log L(y; \mu, \sigma^2, p) = \frac{1}{\sigma^2}\left[y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right] - c(y, \sigma^2, p). \tag{2.20}$$

La fonction `tweedie.profile` du *package* `Tweedie` de `R` et détaillée par Dunn (2022a) choisit la valeur de p qui maximise la log-vraisemblance totale du modèle. En considérant,

- y_i est l'observation i^{me} ;
- μ_i est la valeur prédite pour l'observation i^{me} ;
- σ^2 est le paramètre de dispersion ;
- p est le paramètre de puissance ;
- Γ est la fonction Gamma.

La log-vraisemblance est donnée par l'expression

$$\log L(\theta; y) = \sum_{i=1}^{n}\left[\frac{y_i\mu_i^{1-p}}{(1-p)\sigma^2} - \frac{\mu_i^{2-p}}{(2-p)\sigma^2} + \frac{y_i^{2-p}}{(2-p)\sigma^2} - \log\Gamma(y_i+1)\right], \tag{2.21}$$

Dans cette équation 2.21, le premier terme représente la contribution de l'observation y_i en relation avec la moyenne prédite μ_i. Le deuxième terme pénalise la log-vraisemblance pour la distance entre l'observation y_i et la moyenne μ_i. Le troisième terme ajuste la log-vraisemblance en fonction de la dispersion des observations. Enfin, le dernier terme normalise les observations en utilisant la fonction Gamma Γ.

La log-vraisemblance Tweedie est complexe en raison de la nature flexible de sa distribution. Cette caractéristique peut impliquer des problèmes de convergence de l'algorithme vers un maximum global de la log-vraisemblance. L'algorithme peut réaliser des boucles et rester bloqué sur des *maxima* locaux uniquement voire même diverger.

Il convient de rappeler que le paramètre p contrôle la forme de la distribution Tweedie. Il détermine la relation entre la moyenne et la variance des données et est compris entre 1 et 2 comme indiqué précédemment. Afin de déterminer la valeur de ce paramètre, la fonction `tweedie.profile` du *package* `Tweedie` suit un procédé en plusieurs étapes.
Tout d'abord, une grille de valeurs potentielles pour p est définie. Elles sont comprises entre $1,1$ et $1,9$. Pour chaque valeur de p dans cette séquence, la fonction ajuste un modèle de régression Tweedie et calcule la log-vraisemblance correspondante.
Ensuite, la fonction sélectionne la valeur de p qui maximise la log-vraisemblance. Cette valeur devient la valeur optimale de p, retenue pour la modélisation de Tweedie. Enfin, afin de faciliter la lecture, la fonction retourne les valeurs de p testées ainsi que leur log-vraisemblance correspondante à l'aide d'un graphique. L'analyse de celui-ci permet de déterminer la valeur optimale de p. Après plusieurs recherches, il apparaît que la détermination du paramètre p optimal avec la log-vraisemblance n'est pas optimal, comme illustré dans la figure 2.15. La valeur $p = 1.3$ est retenue comme l'argument qui maximise la log-vraisemblance du modèle.

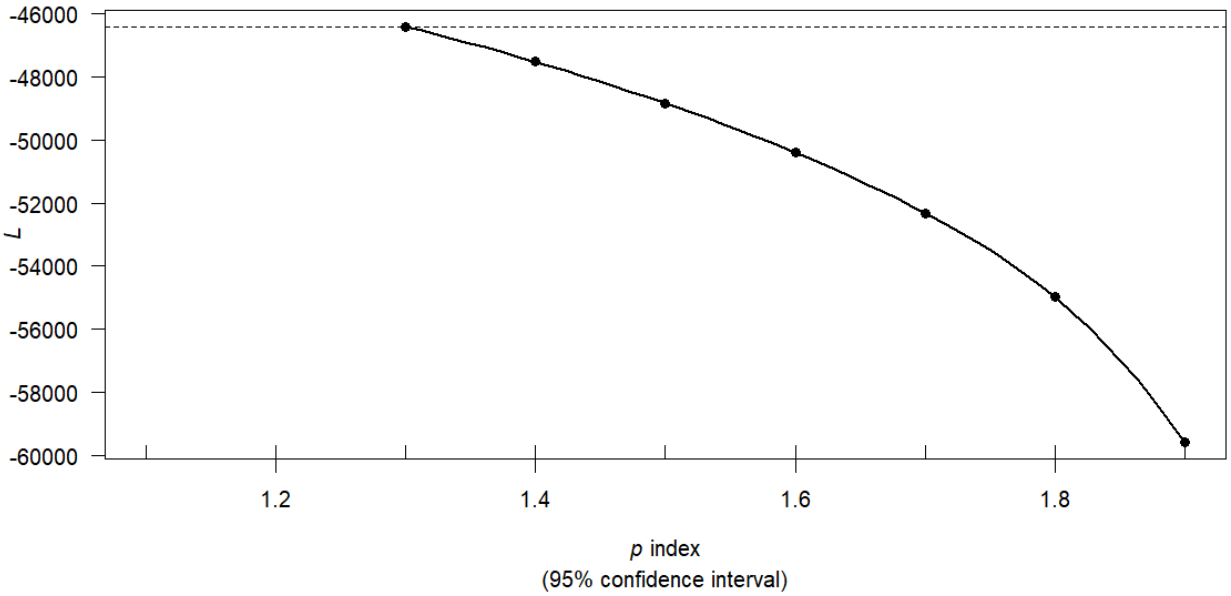

FIGURE 2.15 : Détermination du paramètre de puissance optimal p

Afin de conforter le choix de cette valeur, le graphique 2.16 permet de mettre en relation les valeurs de l'AIC en fonction des valeurs des différents paramètres de puissance p. La fonction `AICTweedie` du *package* `Tweedie` permet de calculer l'AIC dans le cadre d'un modèle de régression de Tweedie. Pour rappel, ce critère doit être minimisé. Pour la valeur $p = 1,3$, l'AIC n'est pas minimal, néanmoins sa valeur reste acceptable.

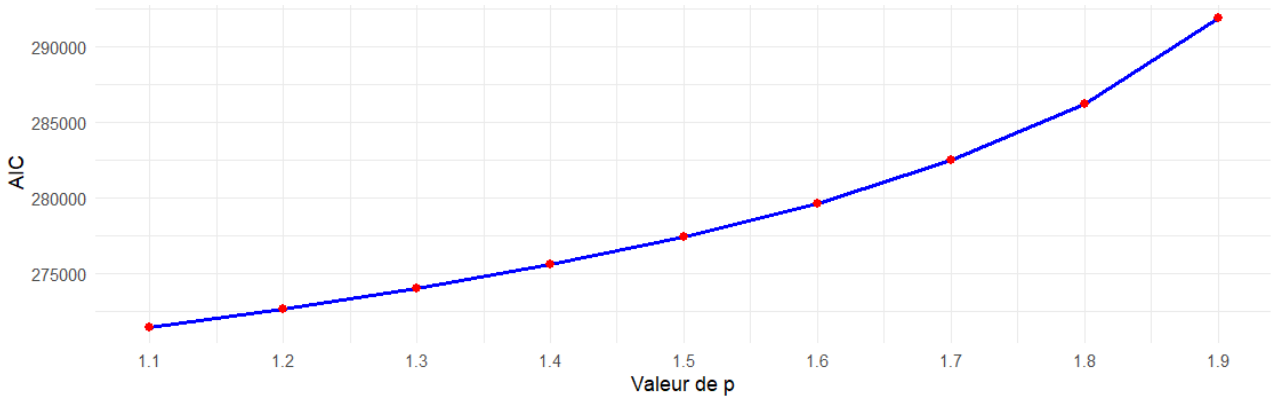

FIGURE 2.16 : AIC en fonction de la valeur du paramètre de puissance p

En conclusion, dans cette étude, la valeur retenue est $p = 1,3$ pour l'hyperparamètre de puissance.

Après avoir exposé la théorie sous-jacente, il convient d'étudier dès à présent la mise en œuvre concrète du modèle avec une description du modèle utilisé.

Description et analyse du modèle retenu

Dans cette étude, une régression de Tweedie avec un lien logarithmique est utilisée pour prédire la variable Y.

Soit $Y = \texttt{IPTC\_obs}$, alors l'espérance $\mathbb{E}[Y]$ est notée μ. En considérant,

- $\mu = \mathbb{E}[Y]$, l'espérance de la variable dépendante Y, `IPTC_obs` ;
- $\log(\mu)$, la fonction de lien logarithmique appliquée à l'espérance μ ;
- β_0, l'intercept du modèle ;
- $\beta_1, \beta_2, \dots \beta_{10}$, les coefficients associés aux variables explicatives respectives (`m_dep`, `grp_tc`, ... , `m_sub`, etc.). Ils sont estimés à partir des données historiques ;

la fonction de lien logarithmique est définie par

$$\begin{aligned}\log(\mu) = \beta_0 &+ \beta_1 \times \texttt{m\_dep} + \beta_2 \times \texttt{grp\_tc} \\ &+ \beta_3 \times \texttt{pol\_client} + \beta_4 \times \texttt{grp\_td} \\ &+ \beta_5 \times \texttt{grp\_bw} + \beta_6 \times \texttt{NL} \\ &+ \beta_7 \times \texttt{BE} + \beta_8 \times \texttt{DE} + \beta_9 \times \texttt{FR} + \beta_{10} \times \texttt{m\_sub}.\end{aligned} \tag{2.22}$$

Ce modèle suppose que la relation entre les prédicteurs et la variable réponse est multiplicative. Ainsi, l'effet de chaque prédicteur sur la variable réponse est exponentiel. Autrement dit, chaque unité d'augmentation d'un prédicteur a pour effet de multiplier `IPTC_obs` par $\exp(\beta)$, où β est le coefficient de régression associé à ce prédicteur.

Si un coefficient est positif, cela signifie que l'augmentation de la variable prédictive entraîne une augmentation exponentielle de IPTC_obs. Tandis que pour un coefficient négatif, il s'agit d'une diminution exponentielle de `IPTC_obs`.

Les variables catégorielles s'interprètent en fonction de la modalité de référence, non incluse dans le modèle. Lorsque le coefficient est positif, cette catégorie a une `IPTC_obs` plus élevée que la catégorie de référence qui a pour valeur l'intercept, β_0.

La modélisation de Tweedie pour un paramètre de puissance $p = 1,5$ et une fonction de lien log, g, s'écrit

$$\begin{aligned}\log(\mathbb{E}[\widehat{\texttt{IPTC\_obs}}]) = \widehat{\beta_0} &+ \widehat{\beta_1} \times m_{\texttt{dep}} \\ &+ \widehat{\beta_2} \times \texttt{grp\_tc} + \widehat{\beta_3} \times pol\_client \\ &+ \widehat{\beta_4} \times \texttt{grp\_td} + \widehat{\beta_5} \times \texttt{grp\_bw} \\ &+ \widehat{\beta_6} \times \texttt{NL} + \widehat{\beta_7} \times BE + \widehat{\beta_8} \times DE \\ &+ \widehat{\beta_9} \times FR + \widehat{\beta_{10}} \times m_{\texttt{sub.}}\end{aligned} \tag{2.23}$$

Les coefficients résultants de la modélisation sont fournis dans le tableau 3.3.2 présenté en annexe.

Les résultats de cette régression montrent que plusieurs variables, notamment certaines catégories de `m_dep`, `m_sub`, `grp_td`, `grp_bw` et `grp_tc`, ainsi que les pays `NL`, `BE` et `FR`, ont des effets significatifs sur la variable cible. Les différentes valeurs des coefficients dévoilent des impacts variés selon les variables explicatives sur l'IPTC.

Analyse des coefficients de régression du modèle GLM Une brève analyse des valeurs des coefficients significatifs du modèle avec la figure 2.17 permet de mettre en avant les catégories influant à la hausse ou à la baisse sur l'IPTC prédit par rapport à la valeur de l'intercept.

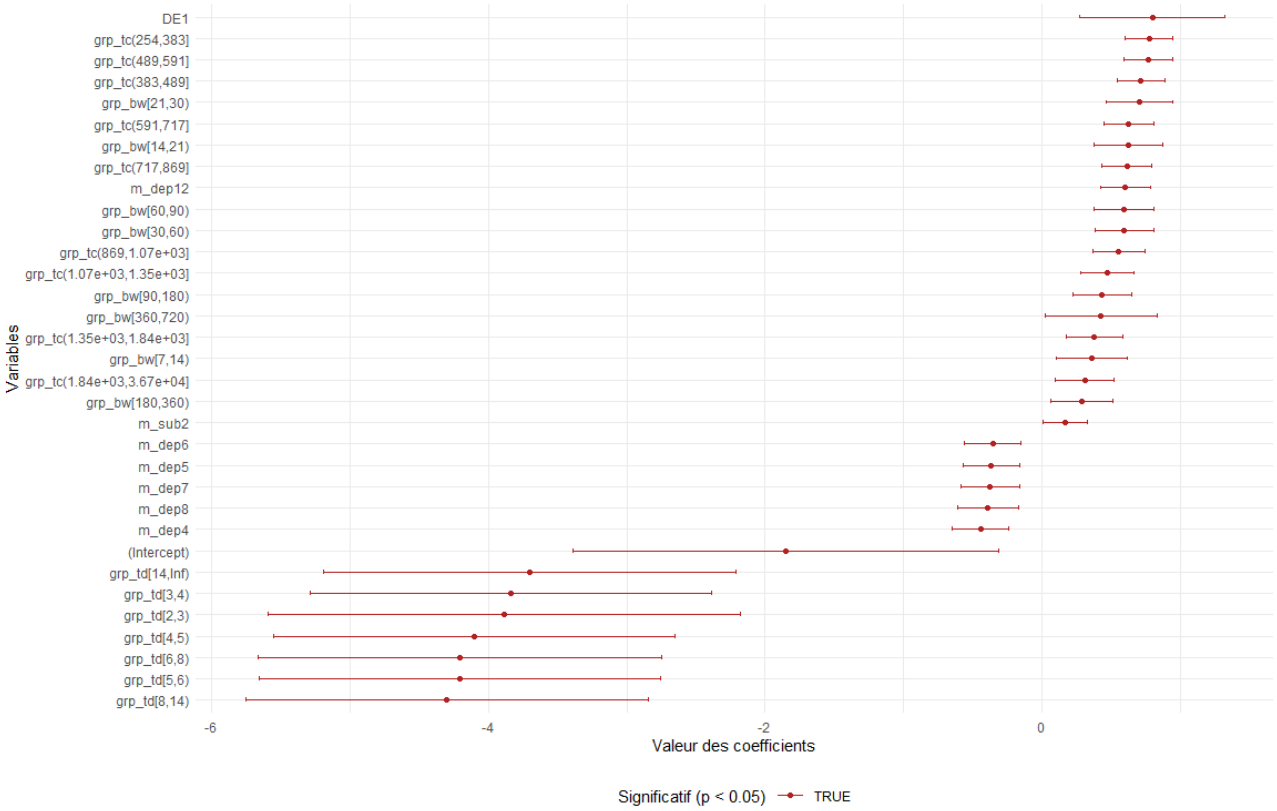

FIGURE 2.17 : Coefficients des variables significatives avec intervalles de confiance à 95%

Dans cette figure, les coefficients significatifs sont indiqués par la couleur rouge. Ils correspondent aux coefficients de régression ont une influence statistiquement significative sur l'IPTC. Tandis que les coefficients non significatifs, non présents sur la figure, n'ont pas une influence statistiquement significative. Si la variable est significative, plus le coefficient est éloigné de zéro, plus la variable explicative a un impact sur l'IPTC. Les valeurs des coefficients sont comprises entre 1 et -6, engendrant de longs intervalles de confiance pour certaines variables. Celles-ci concernent des intervalles de voyage ayant des expositions très volatiles. L'analyse de ces coefficients est complémentaire à d'autres analyses comme celle de la performance du modèle à l'aide des erreurs de modélisation.

Évaluation des performances du modèle retenu : Analyse de la MSE, MAE et RMSE Afin d'évaluer la performance d'un modèle, plusieurs mesures sont utilisées, comme le montre le tableau 2.6.

La *Mean Absolute Error* (MAE) est la moyenne des différences absolues entre les valeurs prédites et les valeurs réelles. Elle permet de comprendre l'erreur moyenne sans prendre en compte la direction de l'erreur qui correspond à la sous ou surestimation. Une MAE de 0,0265 indique que, en moyenne, les prédictions du modèle s'écartent des valeurs réelles de 0,0265 unités.

La *Mean Squared Error* (MSE) est la moyenne des carrés des erreurs entre les valeurs prédites et les valeurs réelles. Cette mesure pénalise davantage les grandes erreurs que les petites erreurs, car elle

élève chaque différence au carré. Une MSE de 0,0146 signifie que la moyenne des carrés des erreurs est de 0,0146 unités carrée.

La *Root Mean Squared Error* (RMSE) est la racine carrée de la MSE. Une RMSE de 0,1206 indique que les prédictions du modèle s'écartent des valeurs réelles de 0,1206 unités en moyenne. Cette mesure est utile car elle est exprimée dans la même unité que la variable cible, l'IPTC.

Mesure	**Valeur**
Mean Absolute Error (MAE)	0,0265
Mean Squared Error (MSE)	0,0146
Root Mean Squared Error (RMSE)	0,1206

TABLE 2.6 : Mesure des erreurs de modélisation

Ces trois valeurs sont relativement faibles comparativement aux valeurs de Y comprises entre 0 et 1. Ces mesures indiquent que le modèle fait des prédictions précises.

La performance du modèle doit être validée sur un ensemble de données de test indépendant pour s'assurer qu'il ne surajuste pas les données d'entraînement. Les diagnostics des résidus (comme les graphiques de résidus vs valeurs ajustées et les *QQ-plots*) doivent être examinés pour vérifier que les hypothèses du modèle sont respectées.

Évaluation des performances du modèle retenu : courbe *lift* Pour analyser la performance du modèle, la courbe *lift* du modèle est utilisée. Cette métrique est utilisé pour analyser la performance dans une logique de continuité avec les outils déjà utilisés dans le processus interne et par le logiciel Akur8. Le *lift* est le rapport entre le taux de positivité (*cumulative rate*) dans chaque décile et le taux de positivité global. Dans cette courbe, l'axe des abscisses représente les segments de la population triés par ordre décroissant de la probabilité prédite par le modèle. Chaque décile contient 10% des observations. Tandis que l'axe des ordonnées représente l'augmentation de la proportion d'événements capturés par le modèle par rapport à une sélection aléatoire.

$$\text{Lift} = \frac{\textit{cumulative rate}}{\textit{cumulative percentage}}$$

- Un *lift* de 1 signifie que le modèle n'apporte aucune amélioration par rapport à une sélection aléatoire ;
- Un *lift* supérieur à 1 signifie que le modèle est meilleur que le hasard ;
- Un *lift* inférieur à 1 signale que le modèle est pire que le hasard.

À titre d'exemple, le graphique se lit de la manière suivante : "le modèle prédit les observations positives avec une efficacité 100 fois meilleure qu'une sélection aléatoire." La courbe, présentée dans la figure 2.18, montre une augmentation progressive du *lift* en fonction des déciles. Le modèle est donc capable de discriminer les événements de manière efficace. Les observations dans les derniers déciles ont des valeurs de *lift* plus élevées. Il prédit mieux les événements pour ces segments, comme le montre le pic important du *lift* au dernier décile. Le modèle semble efficace pour identifier les observations avec les plus fortes probabilités.

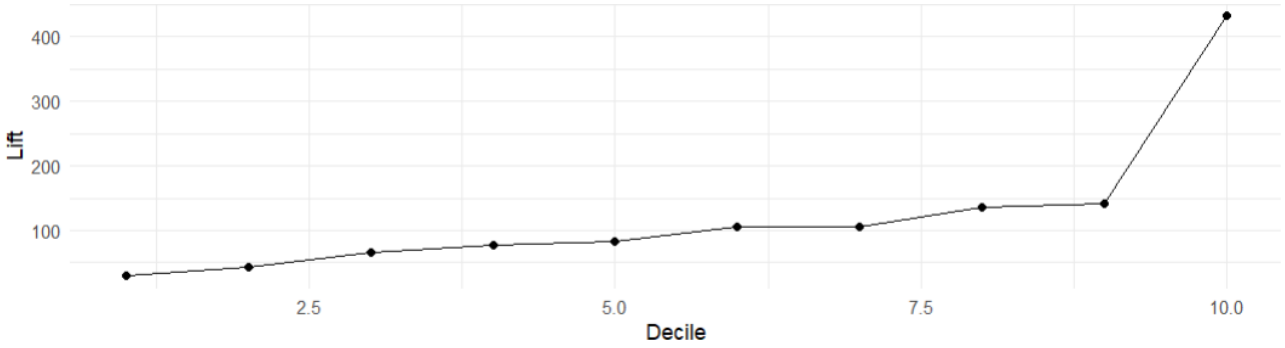

FIGURE 2.18 : Courbe *lift*

La ligne de référence, *lift = 1*, représente une sélection aléatoire. Lorsque la courbe *lift* est au-dessus de cette ligne, le modèle est plus performant qu'une sélection aléatoire pour identifier les événements.

Ainsi, la courbe indique que le modèle de régression Tweedie a une bonne capacité de classement et qu'il est efficace pour identifier les observations les plus probables de l'événement étudié. Néanmoins, il convient de noter l'éventualité d'un sur-apprentissage car les données sont peu nombreuses pour les grandes valeurs d'IPTC.

Sélection des variables explicatives La sélection de variables dans un modèle contribue à une meilleure performance du modèle puisque cette méthode permet d'exclure les variables dont l'effet est négligeable. Plusieurs méthodes de sélection de variables existent, dans cette étude, la méthode statistique est utilisée à l'aide des tests Anova et ANOVA :

- Le test ANOVA permet de comprendre l'apport des variables explicatives au modèle. Les résultats sont répertoriés dans le tableau 2.7.

Variable	***Df***	***Deviance***	***Resid. Df***	***Resid. Dev***
NULL			272 429	56 561
`m_dep`	11	1027,73	272 418	55 533
m_sub	11	28,91	272 407	55 504
`grp_tc`	9	402,69	272 398	55 101
`pol_client`	1	8,60	272 397	55 093
`grp_td`	7	191,46	272 390	54 901
`grp_bw`	9	191,26	272 381	54 710
`NL`	1	216,02	272 380	54 494
BE	1	21,33	272 379	54 473
DE	1	195,74	272 378	54 277
FR	1	6,42	272 377	54 270

TABLE 2.7 : Résultats du test ANOVA (analyse de la déviance)

  L'analyse de ce tableau montre la diminution de la déviance résiduelle lorsqu'une variable est ajoutée au modèle. La déviance totale du modèle sans prédicteur est de 56 561 et ne fait que décroître avec l'ajout des variables. Il faut noter que plus la réduction de la déviance est grande, plus la variable ajoutée apporte d'information au modèle.

- Le test Anova (Type II) présente les tests de chi-deux de la vraisemblance (*LR Chisq*) pour chaque variable. Pour rappel, le calcul de cette statistique de test est présenté dans la section

abordant le calcul du V de Cramer. Ce test permet d'évaluer l'importance statistique de chaque variable grâce à la valeur de la probabilité $p - value$, comme le montre le tableau 2.8.

Variable	***LR Chisq***	***Df***	***Pr(> Chisq)***
`m_dep`	208,225	11	$< 2.2.10^{-16}$ ***
m_sub	10,465	11	0,489 150
`grp_tc`	115,015	9	$< 2.2.10^{-16}$ ***
`pol_client`	0,019	1	0,890 875
`grp_td`	95,282	7	$< 2.2.10^{-16}$ ***
`grp_bw`	70,370	9	$< 1.2.10^{-14}$ ***
`NL`	0,680	1	0,409 718
`BE`	4,032	1	0,044 636 *
`DE`	10,801	1	0,001 015 **
`FR`	2,430	1	0,119 070

TABLE 2.8 : Résultats du test Anova (Type II)

L'analyse de ce tableau montre que les variables de pays sont peu significatives, hormis l'Allemagne et la Belgique. En outre, les variables `pol_client` et les `m_sub` ne le sont pas non plus.

Finalement, les résultats montrent que les variables `m_dep`, `grp_td`, `grp_bw`, `grp_tc`, BE et FR sont statistiquement significatives. Toutes les variables contribuent à la diminution de la déviance, d'après le test ANOVA, figure 2.8. Les autres variables ne semblent pas significatives d'après le test Anova, dans le tableau 2.8.

Néanmoins, une analyse complémentaire à ce travail montre l'importance de conserver les variables distinguant les entités du partenaire (`pol_client`) et les mois de souscription (`m_sub`) dans le modèle puisqu'elles traduisent des différences de comportement. En outre, ces variables contiennent des modalités très significatives tandis que d'autres le sont moins. Pour les variables concernant les pays de souscription, le raisonnement à appliquer est identique. Il s'agit de conserver les particularités propres au pays de souscription.

Dans ce contexte, même si le modèle n'est pas optimal dans le sens mathématique du terme, toutes les variables présentées sont conservées afin de tenir compte des différences induites par ces variables.

Analyse des résidus du modèle

Pour valider le modèle, il convient de procéder à l'analyse des résidus pour vérifier les hypothèses d'un GLM. Le graphique de la figure 2.19 présente les résidus de déviance par rapport aux valeurs prédites pour la régression Tweedie.

Résidus de déviance Dans ce cas, les résidus de déviance sont analysés. Ils mesurent la contribution d'une observation individuelle à la déviance totale du modèle. Ils correspondent à la différence entre la log-vraisemblance d'un modèle saturé (qui ajuste parfaitement les données) et la log-vraisemblance du modèle ajusté et se calculent comme le montre l'équation ci-après. En considérant que $L(y_i; \hat{\mu}_i)$ est la log-vraisemblance de l'observation y_i sous le modèle avec la valeur prédite $\hat{\mu}_i$ et $L(y_i; y_i)$ est la log-vraisemblance d'un modèle saturé (la valeur observée est égale à la valeur prédite). Pour une observation y_i avec une valeur prédite $\hat{\mu}_i$, le résidu de déviance d_i est donné par

$$d_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2\left(L(y_i; y_i) - L(y_i; \hat{\mu}_i)\right)}\,. \tag{2.24}$$

Les résidus de déviance sont des valeurs normalisées. Ils peuvent être positifs ou négatifs, indiquant si l'observation est au-dessus ou en dessous de la valeur prédite.

Dans ce modèle, il n'y a pas de tendance dans les données pour les faibles prédictions. Les valeurs élevées ont de faibles résidus pouvant éventuellement témoigner d'un surapprentissage des données pour des valeurs élevées d'IPTC. La distance de Cook, analysée ci-après permet d'affiner cette interprétation. Il faut également remarquer que la répartition n'est pas symétrique autour de zéro.

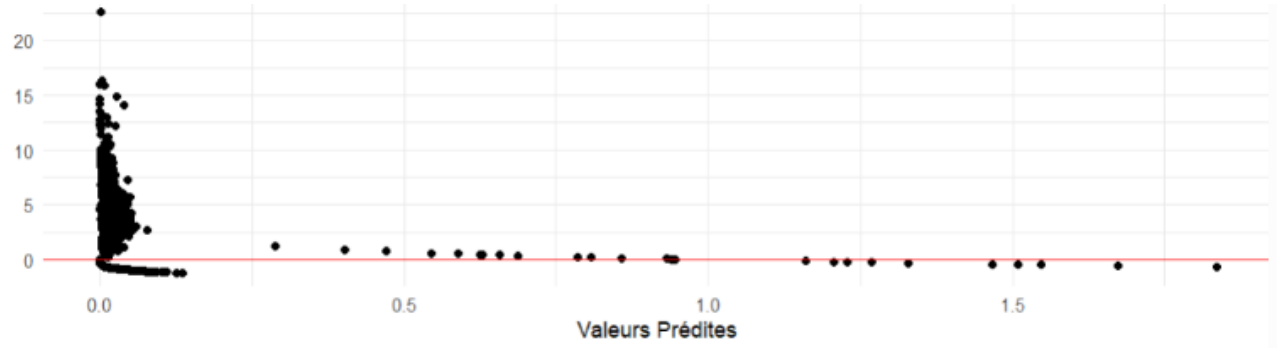

FIGURE 2.19 : Analyse des résidus de déviance

Une analyse du graphique permet d'observer que la majorité des résidus est concentrée autour de la valeur prédite de zéro puisque la variable cible prend majoritairement des valeurs proches de zéro. Les résidus montrent une large dispersion pour les petites valeurs prédites. Il y a une tendance des résidus à diminuer et se stabiliser autour de zéro pour des valeurs prédites plus élevées. La variabilité des résidus semble plus élevée pour les valeurs prédites faibles, signe d'une hétéroscédasticité, comme le montre le test de Breusch-Pagan, expliqué sur le site de ANDERSON (2024a).

Ce test est utilisé pour détecter la présence d'hétéroscédasticité dans les résidus d'un modèle de régression. Dans cette étude, la statistique de Breusch-Pagan est de 848,18. Cette valeur élevée indique une forte présence d'hétéroscédasticité. En outre, la faible valeur de la *p-value* ($< 10^{-14}$) permet de rejeter l'hypothèse nulle d'homoscédasticité et de conclure sur l'hétéroscédasticité des résidus.

Ainsi, il est nécessaire d'ajuster le modèle avec des erreurs robustes de White.

Erreurs robustes de White Dans l'analyse de régression, une hypothèse clé réside dans l'homoscédasticité des résidus, c'est-à-dire qu'ils ont une variance constante à travers toutes les observations. Cependant, cette hypothèse peut être violée dans les données réelles, où la variance des erreurs peut varier. L'hétéroscédasticité peut biaiser les tests statistiques et mener à des conclusions incorrectes. Lorsqu'il y a hétéroscédasticité, la variance des erreurs σ n'est pas constante, c'est à dire,

$$\text{Var}(\epsilon_i) = \sigma_i^2 \neq \sigma^2. \tag{2.25}$$

Les erreurs robustes de White est une méthode pour corriger ce problème et expliquée dans l'article de ANDERSON (2024b).

Même en présence d'hétéroscédasticité, les estimateurs des coefficients de régression restent non-biaisés. Cependant, leurs erreurs standards estimées peuvent être incorrectes, ce qui affecte la validité des tests de significativité et des intervalles de confiance.

Ainsi, les erreurs robustes de White ajustent la matrice de variance-covariance des coefficients pour prendre en compte l'hétéroscédasticité. Cette méthode ne repose pas sur l'hypothèse d'homoscédasticité et permet des inférences statistiquement valides même lorsque l'hétéroscédasticité est présente.

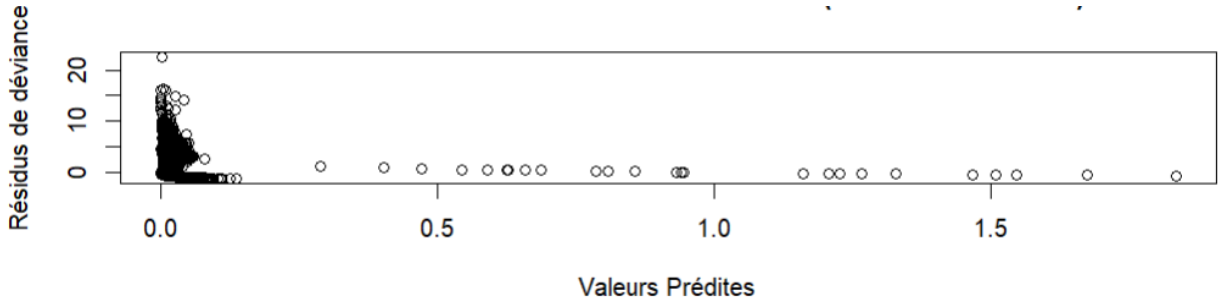

FIGURE 2.20 : Analyse des résidus avec prise en compte des erreurs robustes

La matrice de variance-covariance robuste de White est définie selon l'équation 2.26.

$$\mathbf{V}_{\text{robuste}} = (\mathbf{X}^\top \mathbf{X})^{-1} \left(\mathbf{X}^\top \mathbf{\Omega} \mathbf{X} \right) (\mathbf{X}^\top \mathbf{X})^{-1} \tag{2.26}$$

où $\mathbf{X}$ est la matrice des variables explicatives et $\mathbf{\Omega}$ est une matrice diagonale où chaque élément diagonal ω_i est le carré des résidus pour chaque observation $\omega_i = \hat{\epsilon}_i^2$.

En pratique, les erreurs robustes de White sont calculées en ajustant le modèle de régression classique. Puis, les erreurs standards des coefficients sont recalculées en utilisant la formule ci-dessus.

L'ajustement du modèle avec des erreurs robustes de White permet d'obtenir des inférences statistiques robustes même en présence d'hétéroscédasticité. Ainsi, les erreurs standards des coefficients sont correctes, rendant les tests de significativité et les intervalles de confiance plus fiables.

Après ajustement du modèle, les résultats obtenus sont montrés dans la figure 2.20.

Le graphique quantile-quantile, appelé *QQ plot*, figure 2.21, compare les quantiles théoriques et empiriques de la distribution des résidus. Une brève analyse de ce graphique montre que les résidus ne sont qu'asymptotiquement gaussien.

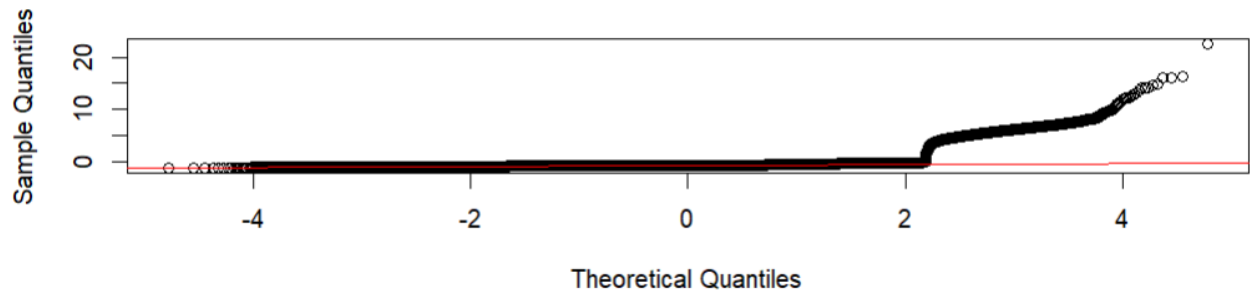

FIGURE 2.21 : QQ plot des résidus de déviance avec erreurs robustes

La distance de Cook, présentée dans le graphique 2.22, met en exergue une absence d'individus dits *outliers* en considérant un seuil classiques de cette distance. Ce seuil correspond à une distance égale à 1 comme indiqué en pointillés sur le graphique. Cette notion est expliquée dans l'article de QUÉBEC CENTRE FOR BIODIVERSITY SCIENCE (2023). La distance de Cook est une mesure qui permet d'identifier les points influents dans un modèle GLM. Il est à noter qu'un tel point se définit comme une observation dont l'exclusion de l'ensemble de données entraîne une modification significative du modèle de régression. Dans ce contexte, la distance de Cook permet de quantifier l'influence de chaque point d'observation.

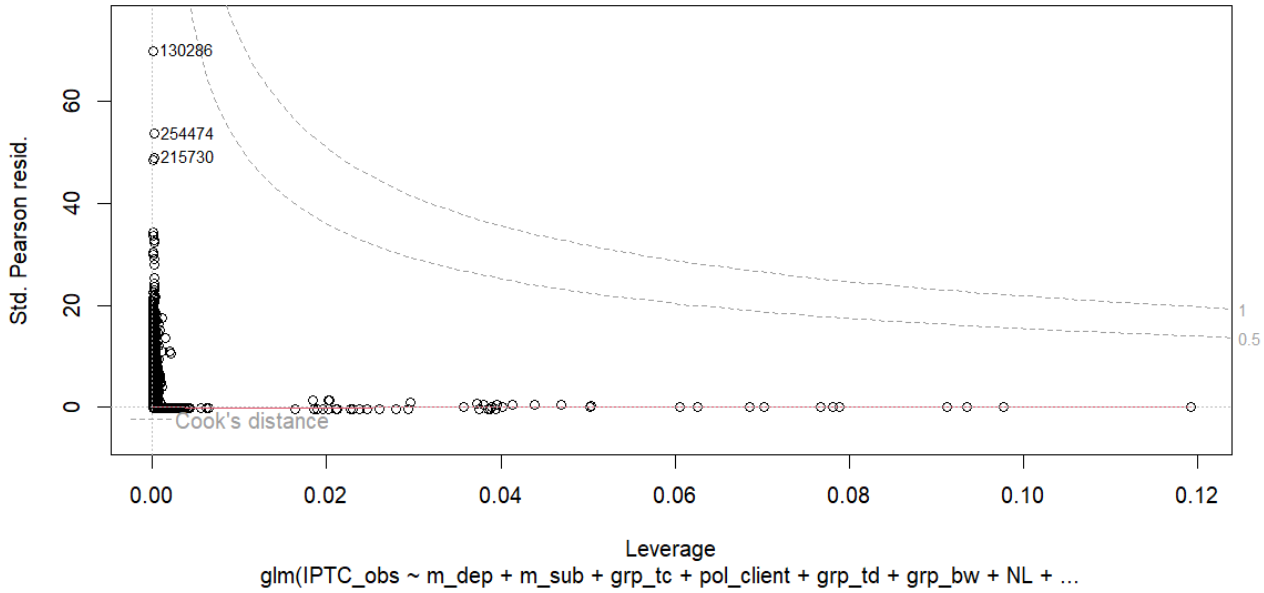

FIGURE 2.22 : Analyse du modèle grâce à la distance de Cook

2.3.2 Confrontation des résultats GLM aux autres modèles

D'un point de vue opérationnel, il s'agit alors d'observer les résultats du modèle, c'est à dire les prédictions de la charge de sinistres du modèle de tarification. Pour tester la performance de ce modèle en comparaison au modèle actuel, il est nécessaire de mettre en œuvre une méthodologie de *backtesting*. À ce titre, un raisonnement particulier pour la récupération des données est mis en place et décrypté ci-dessous, à l'aide du schéma 2.23.

Une base de données à vision au 31 décembre 2023, notée 202312 est utilisée pour réaliser les prédictions. Il faut comprendre que le terme "date de vision" réfère à la date d'extraction de la base de données. Il s'agit alors des données vues à cette date. La base extraite permet ainsi de constater le développement des sinistres à cette date de vision choisie. Pour ce compte, une étude antérieure a montré que les sinistres sont quasiment entièrement développés au bout de trois mois, par la suite, cette période de développement des sinistres est notée $m-3$ par rapport à la date de vision choisie. Ainsi, le modèle GLM est utilisé avec une base de données ayant trois mois de recul par rapport à la date de vision du 31 décembre 2023. Soit dans cet exemple, des données stabilisées au 30 septembre 2023, notée 202309. Le schéma suivant synthétise le mécanisme pour le lecteur ayant eu du mal à suivre le raisonnement.

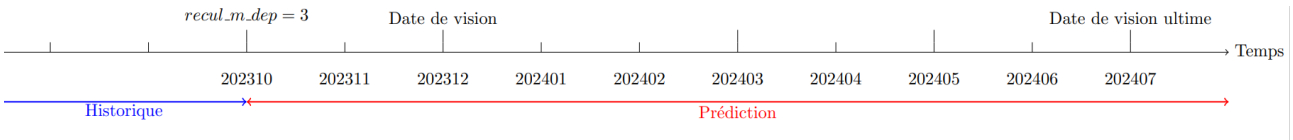

FIGURE 2.23 : Schéma explicatif de la date de vision

Présentation des données des sorties du GLM

Tout vient à point à qui sait attendre, il est désormais temps de visualiser les prédictions réalisées par le modèle.

Au préalable, il faut garder en mémoire que les données sont prises en compte à partir de mai 2022 inclus (2022-05) pour limiter les effets du Covid-19 dans l'apprentissage des données. Les données sont exprimées en mois de départ. L'entraînement du modèle est réalisé sur les douze derniers mois glissants précédents ($m-2$ à $m-13$ de la date de vision). Les prédictions sont réalisées sur les mois instables ($m-2$ la date de vision jusqu'à la date des dernières données disponibles).

Il convient de réaliser l'entraînement du GLM sur les douze derniers mois glissants dans une logique de continuité avec la méthode utilisée actuellement. Cette méthode contient une limite importante qu'il convient de mémoriser lors de l'analyse des résultats graphiques. L'utilisation des douze derniers mois glissants donne une place importante aux données extrêmes en cas de choc, à l'instar de l'épidémie du Covid-19, dans les prédictions d'IPTC.

Les intervalles de temps sont à comprendre à l'aide du graphique 2.23. Les résultats du modèle GLM sont obtenus de la manière suivante :

- Construction du modèle avec une date de vision donnée : à titre d'exemple, 202312, sur des départs jusqu'à $m-3$ de la date de vision (dans cet exemple, il s'agit des données du 2022-05 au 2023-10).
- A partir du modèle GLM construit dans ce chapitre, prédiction de la sinistralité sur la base à vision 202407 sur les mois de départ jusqu'à 2024-04. (Il est nécessaire de filtrer les mois de départ de 2023-10 à 2024-04).

Le modèle construit a vocation à être utilisé pour prédire les données dites instables à une date de vision qui est la date actuelle de visualisation des données. À titre d'exemple, lors de l'écriture de ce mémoire, en août 2024, il s'agit de prédire la sinistralité de ce compte pour les mois à partir de juin 2024.

La figure 2.24 présente les IPTC moyens prédits par le GLM mis en place.

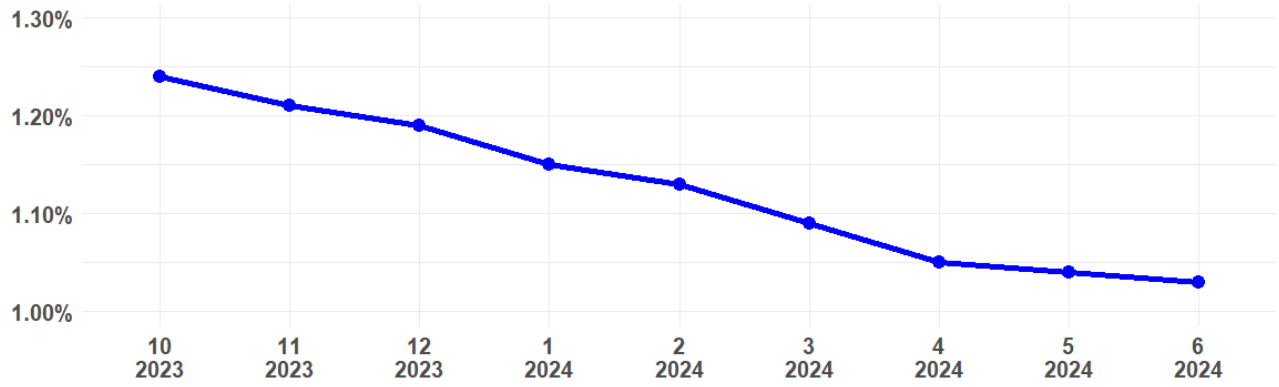

FIGURE 2.24 : IPTC moyen par mois de départ - Méthode GLM

Les résultats de la prédiction sont intéressants dans la mesure où l'IPTC moyen prédit par mois de départ reflète les tendances observées. L'IPTC prédit est sensiblement plus élevé pour les mois hivernaux que pour les mois d'été.

Comparaison aux modèles du logiciel Akur8

AKUR8 (2024) est une solution logicielle spécialisée dans la tarification des assurances, ayant pour objectif d'automatiser le processus de modélisation. Ainsi, il permet de générer les coefficients de régression et la sélection de modèle de manière automatisée rendant plus rapide le processus de tarification d'un produit.

Comparer les résultats permet de vérifier que le GLM construit offre un niveau de précision comparable à celui d'un logiciel spécialisé. Comme Akur8 intègre des techniques d'optimisation des hyperparamètres, ce modèle permet de vérifier l'absence de zones d'amélioration du modèle GLM. Akur8 et le GLM utilisent des régressions de Tweedie qui permettent alors de comparer les résultats des deux modèles. La comparaison aux modèles Akur8 est réalisée en deux parties. Dans un premier temps, il s'agit de sélectionner un modèle pertinent parmi les modèles proposés par le logiciel. Puis, il convient dans un second temps, d'analyser les résultats de ce modèle choisi aux résultats fournis par le GLM.

Processus de sélection de modèles Akur8 Akur8 produit plusieurs modèles nécessitant de les sélectionner afin de choisir le plus pertinent pour comparer les résultats du GLM. Ce logiciel présente aussi plusieurs métriques pour sélectionner le modèle le plus performant. Les métriques permettant de sélectionner le meilleur modèle sont le coefficient de Gini, la déviance moyenne et le RMSE (Root Mean Squared Error) et sont présentées ci-après.

Le coefficient de Gini indique le pouvoir prédictif du modèle. Plus celui-ci est élevé, plus le modèle est performant. Les modèles entre 7 et 10 variables possèdent des coefficients de Gini relativement élevés, autour de 30%.

La déviance moyenne mesure l'ajustement d'un modèle statistique. Une déviance plus faible indique un meilleur ajustement du modèle aux données. Le graphique de la déviance moyenne montre que les modèles avec un plus grand nombre de variables ont une déviance moyenne plus faible, mais les différences ne sont pas significatives après 7 variables.

Le RMSE mesure la différence entre les valeurs observées et les valeurs prédites par le modèle. Une valeur de RMSE plus faible indique un meilleur ajustement du modèle aux données. Les modèles avec plus de variables semblent avoir un RMSE similaire, avec peu de différence entre 11 et 12 variables.

Afin de s'inscrire dans une logique de continuité avec la modélisation GLM, les modèles sont produits pour un paramètre de puissance, $p = 1,3$. Compte-tenu de la valeur de cet hyperparamètre, pour choisir le meilleur modèle, un compromis entre ces métriques doit être trouvé. Les modèles 11 et 12 ont des coefficients de Gini relativement élevés et des RMSE similaires et relativement faibles, tandis que la déviance moyenne ne varie pas significativement après 7 variables. En tenant compte de ces observations, le modèle avec 9 variables est préféré pour éviter un surajustement. Il réalise des performances comparables aux autres modèles avec un nombre légèrement inférieur de variables.

Les variables utilisées sont répertoriées dans la figure 2.25. Le *spread* indique la contribution de cette variable au modèle. En vert foncé, le *spread* est représenté sans tenir compte de l'impact entre les variables alors qu'en vert clair, il est représenté en tenant compte de 5% données extrêmes. Les données extrêmes sont des valeurs d'IPTC anormalement élevées qui proviennent de cas spécifiques.

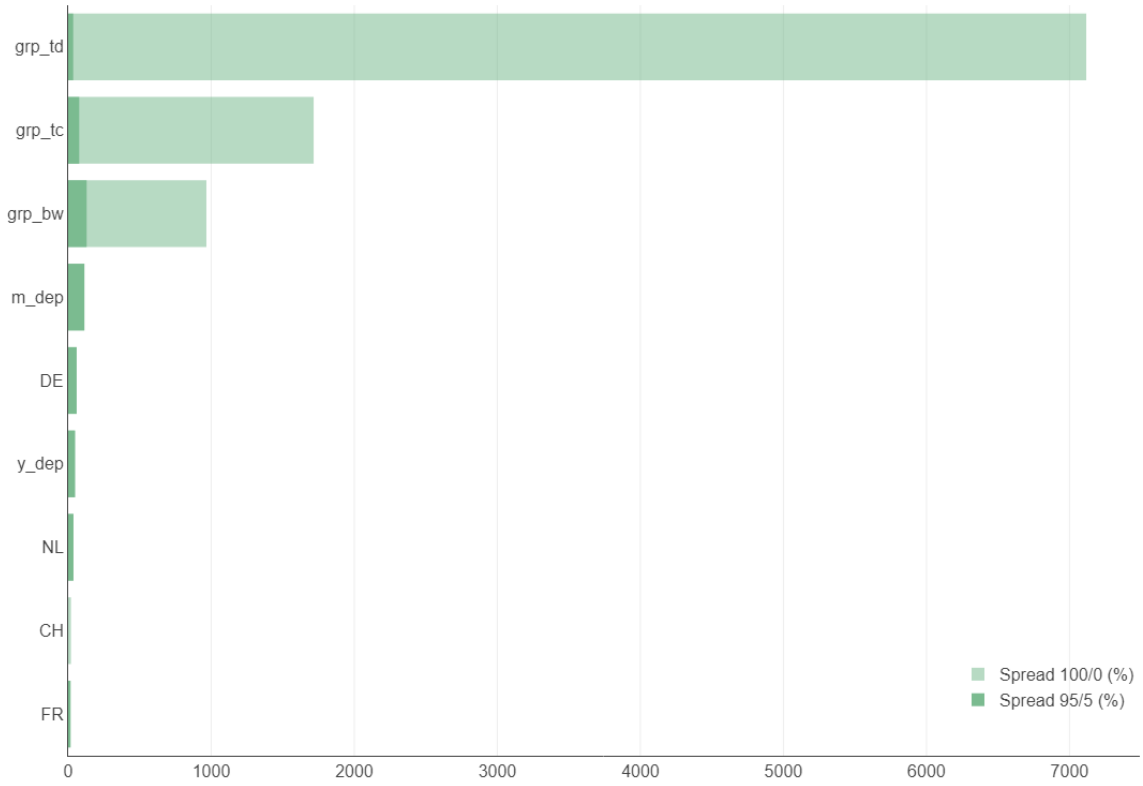

FIGURE 2.25 : Description du modèle proposé par Akur8

Cette représentation met en avant l'importance de la durée du voyage dans ce modèle ainsi que celui du coût du voyage. Il est intéressant de noter que l'année de départ ne constitue pas une variable primordiale pour le modèle. Cette analyse valide l'hypothèse de réaliser les prédictions sur une base contenant les données des douze derniers mois glissants.

L'analyse des modèles proposés par Akur8 et la sélection d'un modèle approprié permet de valider les hypothèses du GLM construit précédemment. Si les hypothèses techniques sont vérifiées, il convient dès à présent de vérifier que les sorties du modèle Akur8 correspondent à celles du modèle GLM ainsi

que ce qui est observé dans la réalité.

Analyse des résultats obtenus En comparant les résultats des modèles obtenus, il s'agit de vérifier la robustesse du modèle GLM. Si les coefficients et les prévisions sont similaires, cela renforce la confiance dans la validité du modèle GLM.

Afin de maintenir une ligne directrice avec le modèle GLM, la courbe *lift* est utilisée comme critère de sélection de modèle. L'écart entre l'IPTC observé et prédit est relativement faible, comme le montre la figure 2.26.

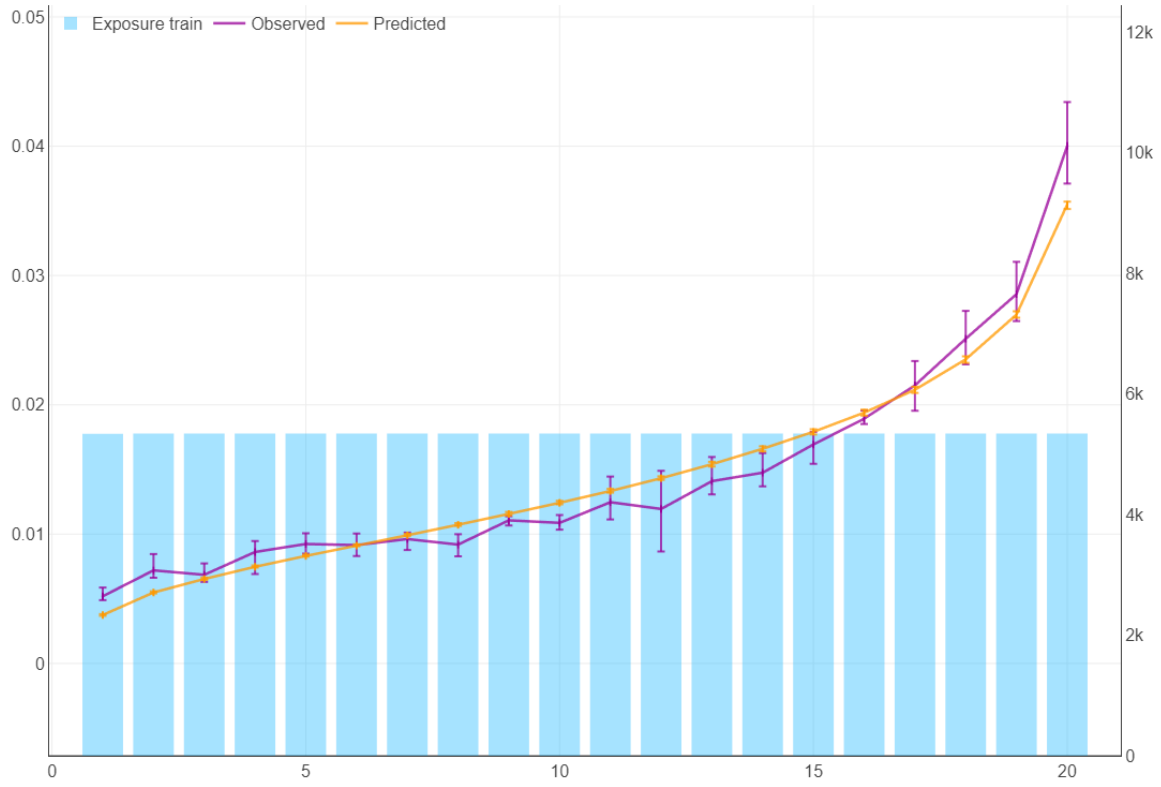

FIGURE 2.26 : *Lift curve* du modèle proposé par Akur8

Ce graphique compare les sinistres observés et prédits sur l'ensemble des segments de l'exposition. Les barres bleues du graphique représentent l'exposition au risque dans chaque segment. Cette exposition correspond aux montants assurés.

Les courbes violettes et orange représentent respectivement les sinistres observés et prédits. Lorsque les deux courbes sont proches, le modèle GLM du logiciel Akur8 est bien calibré. Il prédit avec précision les sinistres attendus. L'analyse de cette figure montre que le modèle capture correctement la relation entre les variables explicatives et le pourcentage de charge de sinistres dans la majorité des segments.

Il faut noter que les segments situés vers la droite du graphique montrent une augmentation rapide des sinistres observés et prédits, ce qui correspond à des segments à risque plus élevé. Le modèle proposé par Akur8 tend à capturer correctement cette tendance. Ce propos est illustré par la convergence des courbes observée et prédite dans ces segments.

Les prédictions réalisées par Akur8, illustrées dans le graphique 2.27, sont du même ordre que celles du modèle GLM. L'axe des ordonnées est un compteur pour les valeurs d'IPTC situées en abscisses. La proximité des valeurs prédites avec celles observées réellement montre que ce modèle est pertinent.

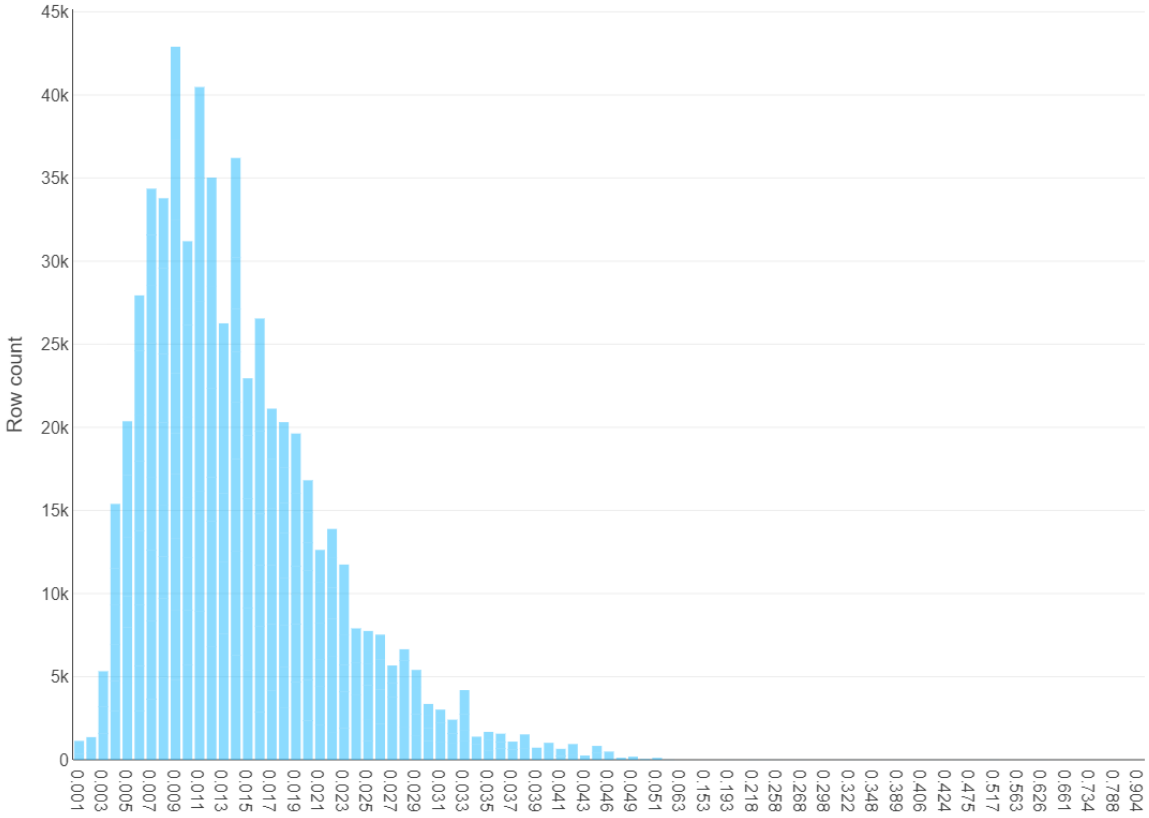

FIGURE 2.27 : Description du modèle proposé par Akur8

Ce graphique met en avant que la majorité des prédictions réalisées par le modèle Akur8 sont comprises entre $0,9\%$ et $1,5\%$ de la charge de sinistres. Il existe des valeurs extrêmes reflétant la réalité de la situation.

Par conséquent, les comparaisons entre le modèle Akur8 et le modèle GLM créé précédemment permettent de conclure positivement sur la viabilité et l'intérêt du modèle créé. Ce modèle prend son sens dans la mesure où il permet de traiter les données et de modéliser l'IPTC sur une seule interface et un seul code `R`.

Comparaison avec les données réelles

L'objectif final de la modélisation GLM est d'améliorer les performances de prédiction de la charge de sinistres par rapport au modèle actuel. Pour rappel, la charge de sinistres estimée par la méthode actuelle est calculée à partir d'une moyenne de la charge sur les douze derniers mois de départ glissants. Dans le graphique 2.30, la courbe verte représente la variable cible. Il s'agit de l'IPTC observée à la date d'extraction des données, la référence à reproduire le plus fidèlement possible. Cette courbe sert

de *benchmark* pour évaluer la précision et l'efficacité du nouveau modèle.

Un exemple concret de cette perturbation est l'important écart de prédiction observé pour le mois d'avril. Cet écart s'explique en grande partie par une charge exceptionnelle en avril 2022, directement liée à la vague Omicron. Cette vague, survenue à un moment où les restrictions étaient encore présentes mais où l'activité reprenait progressivement, a engendré un pic de sinistres non anticipé par les modèles basés sur les tendances pré-pandémiques. Le graphique 2.4 illustre clairement cette charge de sinistres répartie par mois de départ en mettant en avant la différence de la charge pour le mois d'avril. Ces variations mois par mois doivent être prises en compte dans l'analyse pour ajuster les prédictions futures.

Les données réelles d'IPTC à la date de vision juin 2024 sont représentées dans la figure 2.28.

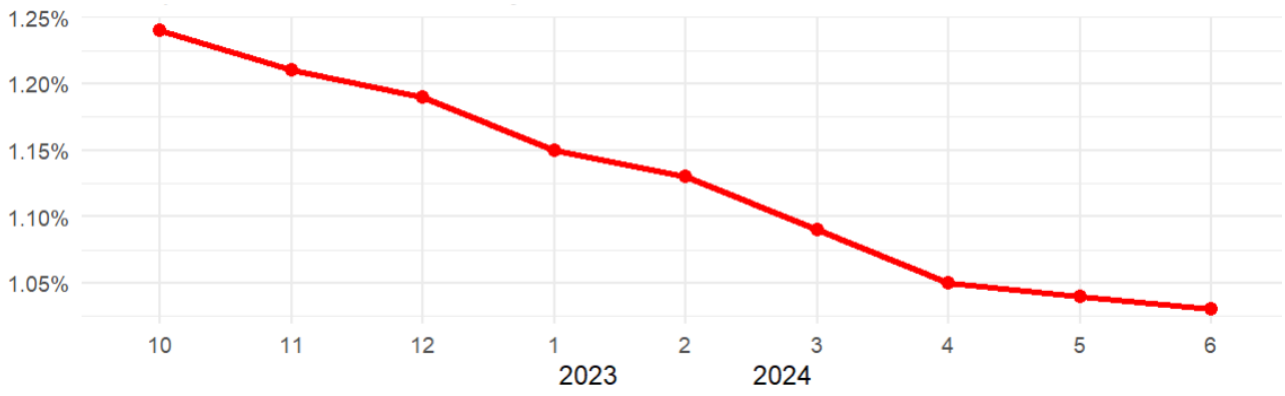

FIGURE 2.28 : Données réelles des IPTC en fonction des mois de départ à date de vision juin 2024

Le courbe du graphique ci-dessus décroît puisque la charge de sinistres n'est pas totalement développée pour les mois d'avril à juin 2024. À cette date de vision, la charge de sinistres est considérée comme stable pour les mois précédant le mois de mars 2024. Pour les autres mois, l'IPTC observé en juin 2024 est plus faible car la charge de sinistres est en cours de développement.

La figure 2.29 montre les prédictions de charges de sinistres permettant de calculer la prime pure par contrat, pour les mois d'octobre 2023 à avril 2024.

Le montant de prime pure pour chaque contrat s'écrit alors

$$\text{Prime Pure}_i = IPTC_{\text{prédit}_i} \times Trip\ cost_i. \tag{2.27}$$

La charge de sinistres totale s'exprime comme la somme des primes pures individuelles sur une période donnée,

$$\text{Charge de sinistres totale} = \sum_{i=1} \text{prime pure}_i. \tag{2.28}$$

La charge de sinistres est utilisée comme métrique pour comparer les résultats des différents modèles (actuel et GLM). Dans cette étude, la charge de sinistres totale est comparée pour chaque mois de départ.

Le graphique 2.29 compare la charge de sinistres prédite par le GLM et la charge de sinistres observée. Il permet de conclure que les résultats obtenus sont probants. Le modèle GLM a tendance à sur ou sous-évaluer les valeurs réelles à différents moments. Il faut noter que les données d'entraînement du modèle contiennent encore des données affectées par les contraintes sanitaires dues à l'épidémie du Covid-19. Globalement, le modèle retrace les tendances observées dans la réalité, ce qui en fait tout de même un modèle performant.

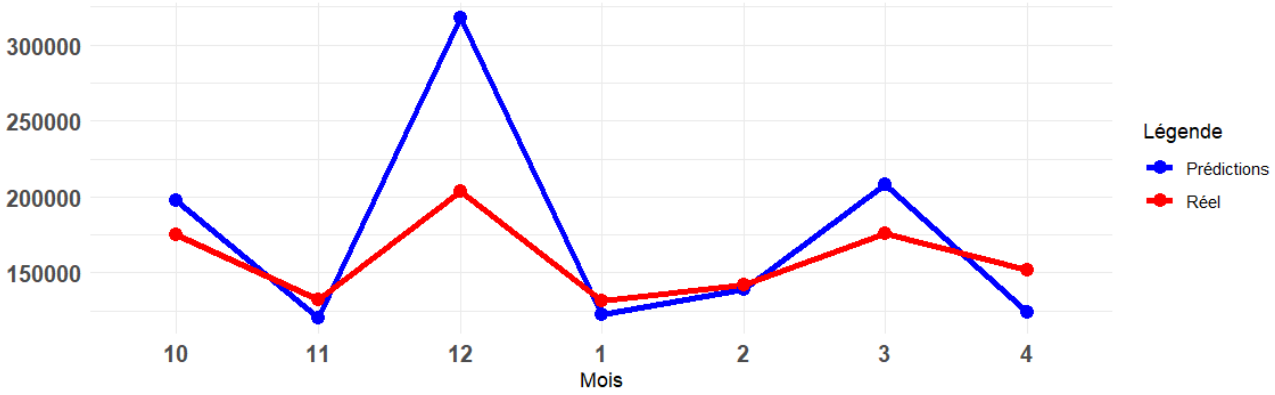

FIGURE 2.29 : Comparaison des données de charges de sinistres réelles et prédites (GLM) en fonction des mois de départ, à la date de vision juin 2024

Cette visualisation des données prédites en comparaison aux données réelles permet de comprendre que le modèle est pertinent. Néanmoins, ce graphique ne permet pas d'apporter une réponse complète à l'objectif de cette étude. Pour rappel, l'enjeu de ce mémoire est de réaliser une meilleure performance que la méthode actuellement utilisée. Comme le dit Jean de La Fontaine, "rien ne sert de courir, il faut partir à point", le lecteur est enfin amené à visualiser le graphique 2.30. Celui-ci compare les données réelles aux prédictions réalisées actuellement (courbe rouge) et aux prédictions réalisées par le GLM (courbe bleue).

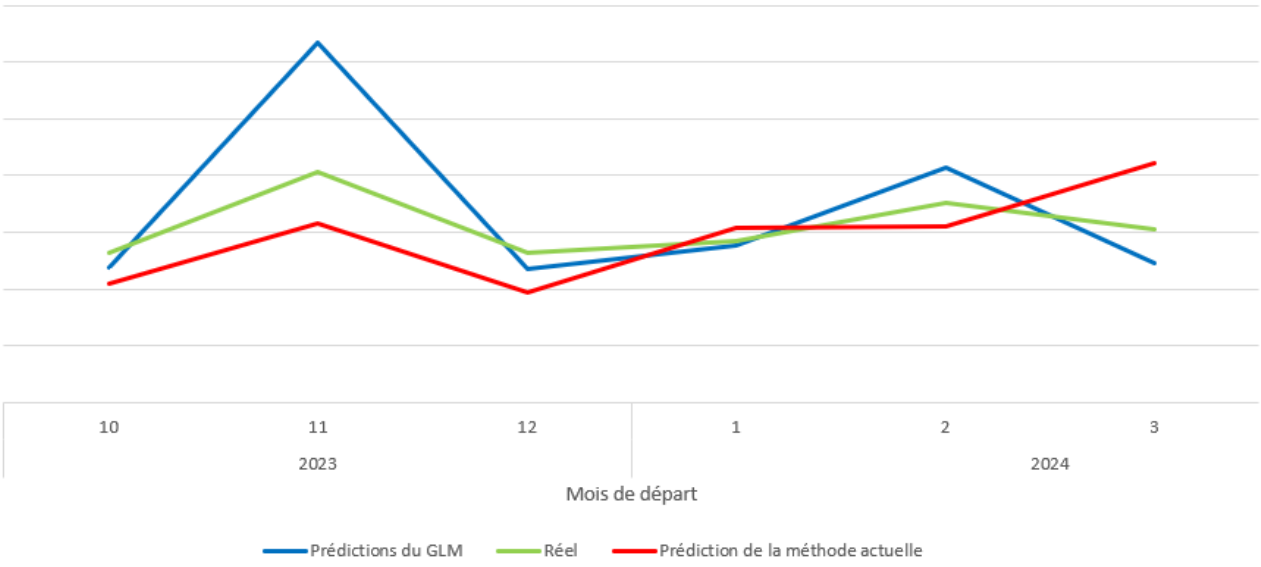

FIGURE 2.30 : Comparaison des prédictions de charge de sinistres en fonction des mois de départ, à la date de vision décembre 2023

Le graphique 2.30 permet de comparer les charges de sinistres prédites par la méthode actuelle et celle prédites par la méthode GLM. Il met en avant que le modèle GLM (courbe bleue) suit mieux les tendances imposées par les données réelles (courbe verte) que la méthode actuelle (courbe rouge). De plus, ses prédictions sont plus prudentes. Enfin, il s'adapte plus rapidement à un changement de répartition du portefeuille À titre d'exemple, l'analyse de sinistralité selon les pays de souscription montre que celle-ci est très variable (à l'instar de l'écart entre l'Allemagne et la Suisse). Dans le cas où,

le portefeuille se déséquilibre entre les pays, le modèle actuel réalisant seulement une moyenne s'adapte très lentement à cette modification. Alors que le modèle GLM prend en compte cette modification dès la souscription du contrat.

Cette figure met donc en exergue une bonne prédiction réalisée par le GLM. Il faut conserver en mémoire que le GLM ne peut prédire totalement la réalité des faits observés. Pour rappel, le modèle actuel réalise des prédictions grâce à des calculs de moyennes des variables intéressées sur les douze derniers mois glissants. L'analyse graphique permet de mettre en avant que les prédictions du GLM et du modèle actuel sont relativement proches. Néanmoins, le modèle GLM possède la qualité de pouvoir facilement s'adapter à un changement de comportement des voyageurs, lorsque la structure de ses données d'entrée se modifient. Il faut remarquer aussi que le GLM capte mieux les tendances que le modèle actuel.

Définition des courbes

L'objectif est de comparer les montants de sinistres réels avec les estimations réalisées. Deux types d'estimations sont réalisées :

- Méthode actuelle - courbe rouge : moyenne sur 12 mois glissant ($m-2$ à $m-13$ de la date de vision choisie), présentée dans le premier chapitre.
- Méthode GLM - courbe bleue : estimation par un modèle linéaire. L'objectif de cette estimation est de réaliser une meilleure performance prédictive que la méthode actuelle.

FIGURE 2.31 : Définition des courbes de prédiction

Conclusion

L'objectif de ce chapitre est de réaliser une prédiction de la charge de sinistres à partir des variables de tarification qui permettent alors de tarifier le contrat d'assurance annulation à l'aide d'une régression GLM avec une distribution Tweedie. Il s'agit de déterminer la prime pure pour un contrat donné. Les variables utilisées dans ce modèle sont des variables dites de souscription, variables disponibles lors de la souscription du contrat d'assurance par le voyageur. Le modèle permet de prédire la prime pure à partir du mois de départ, du pays d'achat de la police d'assurance, de la durée du voyage, de la *booking window* et de la tranche du coût du voyage.

Pour chaque contrat, la prime pure en pourcentage du coût du voyage est calculée en utilisant les coefficients estimés du modèle GLM. Le montant de la prime pure en euros est obtenu selon l'expression 3.19, multiplication entre l'IPTC prédit et le *trip cost* correspondant. Enfin, la mise en œuvre de la tarification se fait en ajustant la prime pure des composants de la prime commerciale. Elle est ainsi augmentée des frais généraux, du bénéfice de l'assureur, des commissions, etc. La relation entre la prime pure et la prime commerciale est rappelée dans l'équation 2.12.

D'un point de vue de la rentabilité qui s'apprécie en termes de *loss* ratio (LR), l'analyse des résultats de modélisation du modèle GLM de charges de sinistres permet de mettre en évidence des résultats cohérents performants puisqu'ils améliorent les prédictions de la méthode actuelle. En outre, ce modèle a été conçu pour s'intégrer à l'outil de gestion et remplacer la méthode actuelle de tarification. Il est donc façonné de manière à conserver les méthodologies déjà mises en place.

Chapitre 3

Modèle de suivi de rentabilité en assurance annulation

L'objectif de ce chapitre est de mettre en œuvre un modèle permettant de suivre la rentabilité du contrat établi avec ce partenaire commercial, en prenant en compte la sinistralité observée. Pour cela, il convient de prendre en compte les variables dites de souscription et des variables porteuses d'information sur la sinistralité. Ces dernières permettent de suivre le développement de la charge de sinistres. La variable cible devient alors l'IPTC ultime, c'est à dire, l'IPTC une fois que la charge de sinistres est entièrement développée. En considérant des souscriptions récentes, qui ont donc déjà eu lieu, certains contrats ont une sinistralité développée tandis que pour d'autres, elle n'est pas (entièrement) développée, comme le montre le graphique 3.1. Ce graphique montre au 31 décembre 2023 la charge de sinistres observée grâce à la courbe bleue, variable notée dans la modélisation IPTC observée (`IPTC_obs`). La courbe rouge indique cette charge à la vision du 31 juillet 2024, variable notée dans la modélisation IPTC ultime (`IPTC_ult`) . L'écart entre les deux reflète le développement de la sinistralité sur les mois de départ de janvier à juillet 2024.

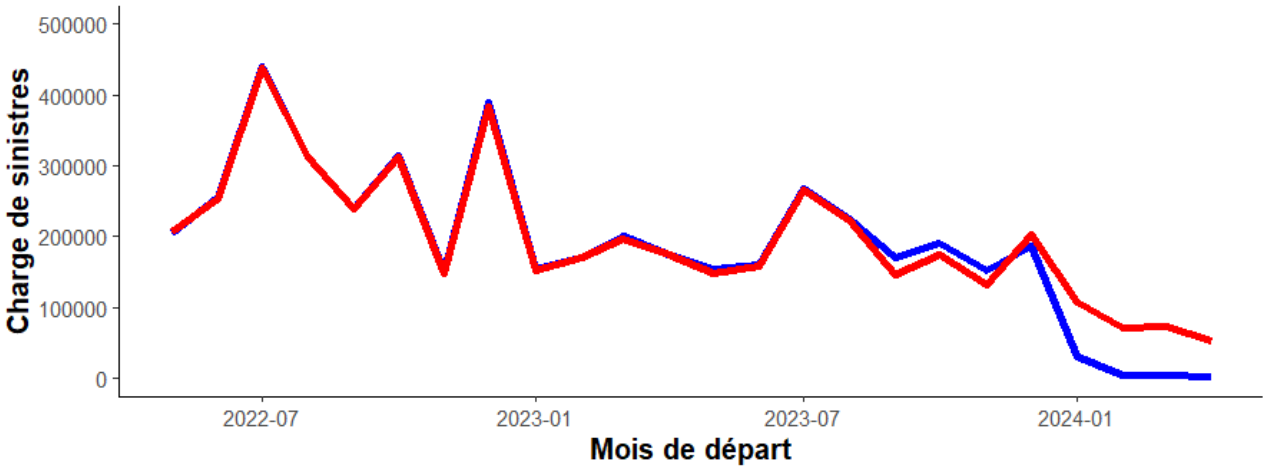

FIGURE 3.1 : Comparaison de la charge de sinistres observée au 31 décembre 2023 et développée au 31 juillet 2024

Dans ce contexte, il convient de différencier la sinistralité stabilisée de la sinistralité à estimer. La

décomposition de la sinistralité d'après ce critère se réalise grâce aux données de mois de départ. Cette variable permet de discriminer de la meilleure façon le développement de la charge de sinistres. Le raisonnement est expliqué de manière plus détaillée dans ce chapitre.

Cette estimation de la charge ultime est utilisée dans le calcul d'indicateurs de suivi de rentabilité. Elle permet alors de connaître la charge ultime de sinistres à comparer aux primes. L'enjeu est d'obtenir des valeurs clés du contrat d'assurance de manière efficace et rapide, afin de mettre à jour mensuellement l'outil de pilotage dédié à ce compte. Ces éléments permettent d'ajuster le tarif mis en œuvre et d'anticiper les négociations avec le partenaire commercial.

La rentabilité du compte s'apprécie avec le *loss* ratio, noté LR, et le *combined* ratio, noté COR. Ces indicateurs de rentabilité ont été définis précisément dans le premier chapitre, 1.3.1. Il est rappelé que le COR* est donné par l'expression $COR = \frac{S+F}{P}$ où S est la charge de sinistres nette , P sont les primes acquises nettes et F correspond aux frais généraux† correspondant aux coûts internes et aux commissions. Une étude de marché documentée par Adrien Suru (2020) montre que le COR sur le marché français de l'assurance Non-Vie est d'environ 100%. Ainsi, ce ratio se décompose en *Loss Ratio* : $LR = \frac{S}{P}$ et en *Expense Ratio*, le ratio de de frais : $ER = \frac{F}{P}$.

Il convient de rappeler que l'écart entre le LR et le COR provient du pourcentage, ER, qui inclus notamment en grande partie les commissions reversées au partenaire commercial. Lorsque l'assureur reverse un pourcentage de commissions au partenaire commercial, l'échange de flux financier se déroule au moment de la souscription du contrat par un voyageur et a lieu avant la date de départ en voyage, sans connaissance d'un quelconque sinistre. Néanmoins, la répartition des richesses créées par ce contrat d'assurance peut aussi se réaliser après cette date de départ grâce à un mécanisme de partage de bénéfice. Il s'agit d'un pourcentage appliqué au bénéfice réalisé sur le compte du partenaire commercial et reversé par l'assureur au partenaire commercial. Par conséquent, ces aspects sont à prendre en compte dans l'outil de pilotage de la rentabilité du compte.

Il est à noter que ce chapitre s'intéresse uniquement aux mesures de rentabilité du compte associé à ce partenaire commercial. Il n'a pas de visée globale ou réglementaire. Ainsi, les mesures de risque concernant le capital requis ou la marge de solvabilité ne sont pas étudiés dans ce mémoire.

3.1 Aspects théoriques de l'algorithme *Gradient Boosting Machine*

L'algorithme de *Gradient Boosting Machine* (GBM) est une technique d'apprentissage supervisé utilisée pour la régression et la classification. Dans ce contexte, il est appliqué à une tâche de régression où le modèle est chargé de prédire un pourcentage de charge de sinistres. Le GBM fonctionne en construisant à chaque étape un arbre de décision qui cherche à minimiser l'erreur de prédiction du modèle précédent grâce aux résidus. Il s'agit pour chaque modèle intermédiaire, appelé modèle faible, de prédire la variable cible (IPTC ultime) en ajustant les résidus du modèle précédent.

Cette méthode permet de capturer des relations non linéaires entre les variables explicatives. C'est un outil puissant pour modéliser des interactions complexes entre les données disponibles à la souscription et de sinistralité. Dans cette étude, il s'agit essentiellement de capter l'état de développement de la charge de sinistres. Grâce à ses mécanismes de régularisation intégrés, le GBM est également capable de limiter le surapprentissage, garantissant ainsi une bonne capacité de généralisation et une meilleure robustesse. Le surapprentissage se définit par une situation où le modèle apprend trop bien les spécificités du jeu de données d'entraînement, y compris le bruit. Par conséquent, il perd sa capacité de généralisation à un nouveau jeu de données.

*Le lecteur peut utiliser le lexique pour se référer aux définitions des sigles utilisés.

†Ces frais sont ventilés en plusieurs catégories de charges et prennent en compte notamment les commissions.

Ce modèle se distingue également par sa flexibilité rendue possible grâce au large éventail d'hyperparamètres ajustables. Pour rappel, comme l'explique Kassel (2023) dans son article, les hyperparamètres sont utilisés pour configurer les modèles, contrairement aux paramètres qui sont déterminés grâce à la modélisation. Un exemple connu et utilisé dans ce mémoire concerne les β d'une régression. Ainsi, le modèle GBM s'adapte aux spécificités des données de modélisation grâce une évaluation continue des erreurs à chaque étape de l'algorithme. Par ailleurs, cet algorithme permet d'estimer l'importance des variables, fournissant des informations supplémentaires sur les facteurs influençant les prédictions. Cependant, cette approche présente l'inconvénient d'être non explicite, demandant un grand nombre de calculs et un ajustement précis des hyperparamètres.

Le fonctionnement du GBM est explicité dans un schéma 15, en annexe de cette étude.

3.1.1 Fonctionnement du GBM

Le GBM fonctionne en trois étapes détaillées ci-après : initialisation, itérations, prédiction. Avant de procéder à la description de ces étapes, il convient de planter le décor. De manière théorique, un problème de *machine learning* (ML) se traite à l'aide du jeu de données $\mathcal{D}$, tel que :

$$\mathcal{D} = \left\{ (x_i, y_i) \mid \forall i \leq n,\ x_i \in \mathbb{R}^d,\ y_i \in \mathcal{Y} \right\}.$$

où n correspond au nombre d'observations et d, le nombre d'informations, autrement appelées variables explicatives. En outre, le vecteur $\mathbf{x}$ correspond au vecteur d'informations. Enfin, dans un problème de ML de régression, $\mathcal{Y}$ correspond à un ensemble de valeurs réelles prises par la variable cible. Ainsi, $y_i \in \mathcal{Y}$ est la valeur réelle associée à l'observation x_i du jeu de données $\mathcal{D}$.

Il est supposé qu'il existe une fonction f qui explique la variable cible à partir des informations disponibles dans la base de données. Ainsi, il est noté :

$$\exists f : \mathbb{R}^d \to \mathcal{Y},\ \forall i \leq n,\ y_i = f(x_i).$$

Soit la fonction de perte : $\mathcal{L} : \mathcal{Y} \times \mathcal{Y}$, une fonction qui permet de mesurer la performance d'apprentissage d'un modèle ML pour une tâche de régression ou de classification. La fonction de perte s'exprime différemment selon le type de tâche demandé. Un problème de *machine learning* se définit tel que

$$f^* = \arg \min_{f:\mathbb{R}^d \to \mathcal{Y}} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)). \tag{3.1}$$

Afin de limiter le surapprentissage du modèle GBM, il est possible de régulariser le problème en ajoutant une pénalisation. Le modèle pénalisé se définit alors comme :

$$f^* = \arg \min_{f:\mathbb{R}^d \to \mathcal{Y}} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \mathcal{P}_\lambda(f). \tag{3.2}$$

Pour un problème de régression, la fonction de perte considérée est la fonction de perte usuelle telle que $\mathcal{L}(y, f(z)) = \frac{1}{2}(y - f(z))^2$. A chaque itération, le modèle cherche à améliorer le modèle créé à l'itération précédente. Il apprend à partir des résidus créés dans les précédents modèles, ce qui constitue un avantage du GBM.

Pour rappel, la pénalisation est une technique de régularisation. La régularisation est un concept général utilisé dans le but de se prémunir contre le risque de surapprentissage afin de rendre le modèle plus robuste et généralisable. Il existe plusieurs méthodes de régularisation dont la pénalisation fait partie. Cette méthode consiste à ajouter un terme de pénalisation, comme le montre 3.2, à la fonction de coût que l'algorithme cherche à minimiser. Dans le GBM, d'autres méthodes de régularisation sont utilisées et décrites dans la partie 3.1.2.

Initialisation de l'algorithme

Le processus débute par l'initialisation d'un modèle simple. Arbitrairement, dans le cas d'une régression, le modèle prédit une constante souvent égale à la moyenne de la variable cible ou bien une valeur arbitraire égale à $0,5$. Ce modèle n'est évidemment pas performant et s'améliore ensuite grâce aux différentes itérations du modèle. Il s'écrit $f_0(\mathbf{x}) \equiv \text{Moyenne}(y)$.

Itérations de l'algorithme

En notant M, le nombre maximum d'itérations, à chaque répétition de l'algorithme, pour $m = 1, 2, \ldots, M$, il ajuste un modèle faible h_m pour minimiser la fonction de perte résiduelle. Les étapes 1 à 3, décrites ci-dessous, sont répétées pour un nombre M d'itérations. L'algorithme s'arrête lorsqu'un critère d'arrêt est atteint (par exemple, convergence de la fonction de perte ou un nombre maximum d'itérations atteint).

1. Calcul des résidus : Le résidu r_{im} est déterminé pour chaque observation i par l'expression

$$r_{im} = -\left[\frac{\partial l(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}. \tag{3.3}$$

   Dans le cas de la régression avec la fonction de perte quadratique $\mathcal{L}$, définie précédemment, le résidu de l'observation i devient simplement,

$$r_{im} = y_i - f_{m-1}(x_i), \tag{3.4}$$

   la différence entre la valeur réelle y_i et la valeur prédite par le modèle, $f_{m-1}(x_i)$.

2. Ajustement du modèle faible : dans cette étape, l'algorithme ajuste un modèle faible h_m en utilisant les résidus comme cibles

$$h_m = \arg\min_h \sum_{i=1}^{N} \mathcal{L}(r_{im}, h(x_i)). \tag{3.5}$$

   L'obtention de la fonction h_m est plus longuement explorée dans la partie 3.1.2. Cette étape fait intervenir un algorithme de descente de gradient, qui donne son nom au modèle, *gradient boosting*.

3. Mise à jour du modèle : Le modèle est ensuite mis à jour en ajoutant le modèle faible ajusté au modèle existant. Cette mise à jour est pondérée par un facteur de taux d'apprentissage η comme indiqué dans l'équation 3.6. Soit $\mathbf{x}$, le vecteur d'informations disponibles dans le jeu de données. Pour l'arbre m, la relation est donnée par

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \eta \cdot h_m(\mathbf{x}). \tag{3.6}$$

   Où $f_m(\mathbf{x})$ est le modèle final après l'ajout de l'arbre m.
   $f_{m-1}(\mathbf{x})$ est le modèle à l'itération précédente.
   η est le taux d'apprentissage.
   $h_m(\mathbf{x})$ est l'arbre à l'étape m.

   Le taux d'apprentissage η contrôle la contribution de chaque arbre au modèle final. Lorsqu'il est trop élevé, il peut conduire à un surapprentissage du modèle, aussi appelée *overfitting*. Tandis que lorsqu'il est trop faible, le modèle peut nécessiter un trop grand nombre d'itérations pour converger. La détermination de l'hyperparamètre est réalisée dans la partie 3.2.2 dédiée à cette analyse.

Prédiction du modèle

Pour un nombre d'itérations fixé, hyperparamètre du modèle, ces étapes sont répétées jusqu'à ce que le modèle atteigne un critère d'arrêt défini au préalable. Celui-ci peut être un nombre maximal d'arbres ou une erreur minimale sur un ensemble de validation. Pour une nouvelle observation $\mathbf{x}'$, la prédiction est donnée par l'équation

$$\hat{y} = f_M(\mathbf{x}'). \tag{3.7}$$

Le modèle final, après M itérations, est la somme pondérée des modèles faibles construits au cours des itérations. Il est donné par l'expression

$$f_M(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{m=1}^{M} \eta \cdot h_m(\mathbf{x}). \tag{3.8}$$

Le GBM inclut des techniques de régularisation pour éviter le surapprentissage, comme la limitation de la profondeur des arbres ou l'utilisation d'un sous-échantillonnage aléatoire des données, appelé (*bagging*). Cette technique n'est pas utilisée dans ce mémoire mais il convient de mentionner son existence.

3.1.2 Spécificités du `XGBoost`

Dans cette étude, la méthode `XGBoost` est utilisée en raison de ses performances élevées et de sa flexibilité. C'est une version optimisée, régulière et performante du GBM classique. À ce titre, `XGBoost` introduit les termes de régularisation L1 (Lasso) et L2 (Ridge)* pour contrôler la complexité des modèles, que le GBM classique ne prend pas explicitement en compte. Ces termes pénalisent les modèles avec trop de branches ou de feuilles, évitant ainsi le sur-apprentissage. En outre, `XGBoost` possède de nombreux paramètres permettant d'ajuster les arbres de manière fine. Ces hyperparamètres du modèle sont analysés lors de la mise en œuvre du modèle. Cette méthode est documentée par le cours de Théo Lopès-Quinta (2022) dispensé à l'Université Paris-Dauphine.

Définition de la fonction h_m

La fonction h_m est la fonction qui permet l'apprentissage du modèle. Elle définit un arbre de régression entraîné sur les résidus obtenus par le modèle faible précédent. Elle dépend, par conséquent, du modèle utilisé et est donc spécifique au modèle `XGBoost`. Elle se définit par l'expression qui cherche à minimiser la fonction de perte $\mathcal{L}$, avec une régularisation quadratique sur h_m, multiplié par λ, le paramètre de pénalité. La régularisation est destinée à limiter la complexité de la fonction h_m. Il faut noter que le terme γT permet de gérer l'élagage de l'arbre et est expliqué ultérieurement.

$$h_m = \arg \min_{h \in \mathcal{F}(\mathbb{R} \to \mathbb{R})} \sum_{i=1}^{n} \left(\mathcal{L}\left(r_i, h(x_i)\right) + \frac{1}{2}\lambda h(x_i)^2 \right) + \gamma T. \tag{3.9}$$

Pour résoudre ce problème de minimisation, l'algorithme utilise une approximation de la fonction h_m grâce aux approximations de Taylor. Le lecteur peut se référer à la page Wikipédia (2024c) du théorème de Taylor pour se documenter sur le sujet. Après réécriture de la fonction de pertes $\mathcal{L}$, l'expression devient

$$\mathcal{L}\left(r_i, h_m(x_i)\right) = \mathcal{L}\left(y_i - \hat{y_i}^{(m-1)}, h_m(x_i)\right) = \mathcal{L}\left(y_i, \hat{y_i}^{(m-1)} + h_m(x_i)\right).$$

*Le lecteur peut trouver dans l'ouvrage de Cornillon et al. (2019) les explications nécessaires à la compréhension de ces notions de pénalisation.

Par ailleurs, afin de simplifier les notations et améliorer la lecture, il faut considérer les expressions suivantes : Soit y_1 et y_2 tels que

$$\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)} + h_m(x_i)\right) = \mathcal{L}\left(y_1, y_2\right).$$

Après application du théorème de Taylor autour de la prédiction à l'itération précédente $\hat{y}_i^{(m-1)}$ (valeur connue), $h_m(x_i)$ s'exprime comme une petite variation autour de $\hat{y}_i^{(m-1)}$. En considérant,

- $\nabla\mathcal{L}$ est le gradient de la fonction de perte par rapport à la prédiction ;
- $\nabla^2\mathcal{L}$ est la dérivée seconde de la fonction de perte.

L'approximation suivante est obtenue :

$$\mathcal{L}(r_i, h_m(x_i)) \approx \mathcal{L}(y_i, \hat{y}_i^{(m-1)}) + \nabla_{y_2}\mathcal{L}(y_i, \hat{y}_i^{(m-1)})h_m(x_i) + \frac{1}{2}\nabla^2_{y_2}\mathcal{L}(y_i, \hat{y}_i^{(m-1)})h_m(x_i)^2. \tag{3.10}$$

La fonction h_m s'écrit alors comme

$$\begin{aligned} h_m \quad &= \arg\min_{h\in\mathcal{F}(\mathbb{R}\to\mathbb{R})} \\ &\sum_{i=1}^{n}\left(\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right) + \nabla_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right)h_m(x_i) + \frac{1}{2}\nabla^2_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right)h_m^2 + \frac{1}{2}\lambda h_m(x_i)^2\right)) \\ &+ \gamma T. \end{aligned} \tag{3.11}$$

La résolution du problème de minimisation permet donc de déterminer la valeur optimale pour une feuille donnée, qui est telle que

$$h_m^* = -\frac{\sum_{i=1}^{n}\nabla_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right)}{\sum_{i=1}^{n}\nabla^2_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right) + \lambda}. \tag{3.12}$$

Détermination du score

Une fois cette valeur déterminée, il s'agit de trouver la meilleure coupure permettant de faire grandir l'arbre. Il faut ainsi mettre en place un score pour déterminer si un nœud dans l'arbre de décision doit être divisé. D'après la forme optimisée h_m et décrite dans l'équation 3.11, il faut chercher à maximiser le gain de couper une racine en deux branches. Dans ce cas, il faut considérer un problème de maximisation et étudier l'expression de $-h_m$ puis en y intégrant la valeur optimale h_m^*, le score est tel que :

$$\text{Score} = \frac{1}{2}\frac{\left(\sum_{i=1}^{n}\nabla_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right)\right)^2}{\sum_{i=1}^{n}\nabla^2_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right) + \lambda} - \gamma. \tag{3.13}$$

avec λ et γ des paramètres de régularisation. La formule du score montre qu'il est calculé en comparant la réduction de l'erreur avant et après la division d'un nœud. Un noeud est divisé lorsque le score est positif : $Score > 0 \Leftrightarrow \frac{1}{2}\frac{\left(\sum_{i=1}^{n}\nabla_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right)\right)^2}{\sum_{i=1}^{n}\nabla^2_{y_2}\mathcal{L}\left(y_i, \hat{y}_i^{(m-1)}\right)+\lambda} > \gamma$. La valeur du paramètre γ décrit ci-dessous permet d'augmenter ou diminuer le nombre de noeuds créés encore appelés branches de l'arbre.

Régularisations de λ et γ

- Paramètre λ : régulateur de la magnitude des ajustements des valeurs associées aux feuilles. D'après l'expression du score de similarité dans l'équation 3.13, le terme λ divise la somme des dérivées secondes. Ainsi, lorsque λ augmente, il réduit l'importance de la dérivée seconde. Il pénalise les ajustements trop rapides. Ce paramètre correspond à une régularisation L2 (Ridge). Concrètement, λ pénalise les grandes valeurs d'ajustement attribuées aux feuilles de l'arbre de décision et évite les ajustements trop extrêmes pour certaines feuilles. En outre, il œuvre pour la stabilité du modèle en évitant que chaque coupure ne soit basée uniquement sur des petites variations des données d'entraînement, qui pourraient ne pas se généraliser aux données de test.

- Paramètre γ : contrôle de l'ajout de nouvelles branches à l'arbre de décision. D'après l'expression de la fonction h_m dans l'équation 3.9, le terme γ permet de contrôler l'élagage de l'arbre. Il impose une pénalité sur chaque nouvelle coupure d'une branche de l'arbre. Plus sa valeur est élevée, plus il est difficile pour l'algorithme d'ajouter de nouvelles branches. L'arbre reste donc moins profond. Durant la construction, l'algorithme `XGBoost` vérifie si le gain de la coupure est supérieur à γ. Dans le cas contraire, la coupure n'est pas effectuée. Après la construction de l'arbre, lorsqu'une une branche ne présente pas une amélioration significative en termes de réduction de la fonction de perte, elle est élaguée. Cet élagage permet de réduire la variance du modèle et d'augmenter sa capacité de généralisation.

3.1.3 Equilibre biais/variance

Comme le précise Maxence Jeunesse (2023), lors d'une conférence à l'Institut Louis Bachelier, le "*Gradient Boosting* est un algorithme itératif, dont chaque itération tend à réduire le biais mais à augmenter la variance."

L'équilibre entre la réduction du biais et l'augmentation de la variance dans un modèle de GBM est exprimé à travers l'erreur quadratique moyenne (*Mean Squared Error*, MSE), décomposée en trois principales composantes décrite ci-dessous.

$$\text{MSE} = \text{Biais}^2 + \text{Variance} + \text{Bruit}. \tag{3.14}$$

Le biais correspond à l'erreur due à la simplification excessive du modèle. Il se calcule comme la différence entre la prédiction moyenne du modèle et la valeur réelle comme l'indique l'équation 3.15. Chaque itération de l'algorithme crée un nouvel arbre de décision corrigeant les erreurs résiduelles laissées par les modèles précédents. Ainsi, avec chaque itération, le modèle devient plus complexe et le biais tend à diminuer. Le biais est défini par

$$Biais^2 = (\mathbb{E}[\hat{y} - y])^2, \tag{3.15}$$

où $\hat{y}$ est une prédiction du modèle et y est la vraie valeur associée.

La variance correspond à la variabilité des prédictions du modèle lorsqu'il est entraîné sur différents échantillons de données. Elle est calculée comme suit dans l'équation 3.16 et mesure la variation de la prédiction ($\hat{y}$) lorsque l'ensemble de données d'entraînement diffère. La variance est définie par

$$\text{Variance} = \mathbb{E}[(\hat{y} - \mathbb{E}[\hat{y}])^2]. \tag{3.16}$$

Le bruit est la dernière composante qui représente l'erreur intrinsèque dans les données qui ne peut être réduite, peu importe le modèle utilisé. Elle est due à des facteurs imprévisibles et aléatoires.

Dans un algorithme de GBM, chaque itération supplémentaire vise à réduire le biais en ajustant les prédictions. Néanmoins, la complexité du modèle augmente à chaque itération venant augmenter la variance. Il s'agit alors de déterminer un équilibre entre réduction du biais et augmentation de la variance. Il est possible de décomposer l'évolution de la métrique MSE entre deux itérations de *Gradient Boosting* par

$$\begin{aligned} \text{MSE} \quad &= (\text{Biais initial} - \text{Réduction de biais par Gradient Boosting})^2 \\ &+ (\text{Variance initiale} + \text{Augmentation de variance par Gradient Boosting}) \\ &+ \text{Bruit}. \end{aligned} \tag{3.17}$$

Ainsi, comme décrit ci-dessus, l'objectif du GBM est de déterminer un point d'équilibre où la réduction du biais compense l'augmentation de la variance sans entraîner un surajustement excessif. Lorsque ce point d'équilibre est dépassé, l'augmentation de la variance prime. Le modèle commence à surapprendre. Ainsi, une dégradation de la performance sur des données non vues se distingue.

Ainsi, le réglage des hyperparamètres de cet algorithme permet de résoudre cette problématique.

3.2 Mise en œuvre de la modélisation GBM

L'objectif de ce modèle est de prédire l'IPTC ultime à travers la variable cible `IPTC_ult`. Elle correspond à l'IPTC une fois les sinistres développés. Elle prend ainsi en compte la sinistralité passée. Cette approche se situe dans une logique de continuité avec l'approche mise en œuvre actuellement pour prédire la rentabilité du compte.

Par principe de continuité et de permanence des méthodes utilisées, il convient de comprendre qu'un certain nombre de paramètres et de variables est conservé entre les modèles de tarification et de rentabilité. Ainsi, le modèle de rentabilité permet de faire des projections sur la performance future de ce compte.

3.2.1 Présentation des variables de modélisation

Les variables utilisées dans le GLM construit dans le chapitre 2 sont conservées en tant que variables explicatives pour le modèle GBM.

Ce modèle a la particularité de prendre en compte la sinistralité passée comme variable explicative. À ce titre, la *cancellation window* et l'IPTC observée sont inclus dans le modèle. Dans ce cadre, il convient de considérer les mois de souscription pour travailler avec l'IPTC ultime. Pour chaque mois de souscription, il convient de déterminer la charge de sinistres à l'ultime.

Cancellation window La *cancellation window* est l'intervalle de temps entre la date d'annulation de la police d'assurance et la date de départ. Elle permet d'étudier le comportement des voyageurs qui utilisent leur assurance annulation.

La figure 3.2 ne prend en compte que les polices sinistrées et montre qu'une grande partie des voyageurs (50%) annulent avant leur départ et 20% entre 1 et 4 jours avant celui-ci. Les intervalles de CW choisis correspondent aux quantiles de la distribution des données. La présence de CW négatives est expliquée par le temps de latence entre la survenance du sinistre et la date de déclaration du sinistre. Il faut noter que la date de survenance a bien lieu avant le départ. Dans le cas contraire, EA ne rembourse pas le voyageur.

Les voyageurs peuvent annuler avant ou après leur départ, puisque la date retenue est celle de la survenance du sinistre. Néanmoins, l'assureur rembourse le montant de pénalité correspondant à la date d'annulation. Ainsi, le graphique met en lumière que 30% des voyageurs annulent dans la semaine après leur départ et 20% ultérieurement. Ce graphique mis en comparaison avec le graphique 1.8 du premier chapitre montre qu'il est important de prendre en compte la sinistralité développée dans le modèle de tarification.

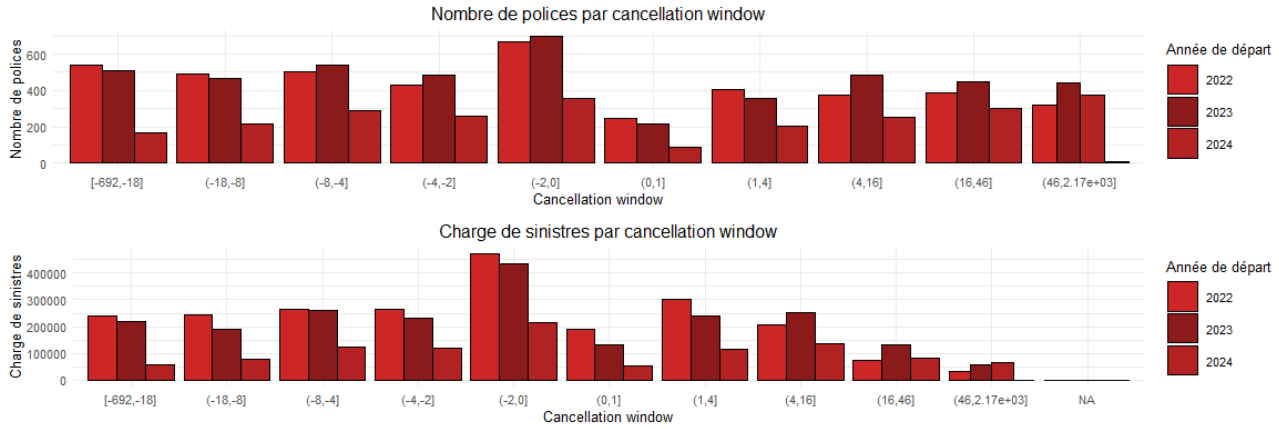

FIGURE 3.2 : Exposition selon la *cancellation window*

La mise en œuvre de modélisation GBM implique la construction de l'ensemble d'entraînement. Cet ensemble exige une jointure entre deux bases de données à date de vision différente, puisqu'il nécessite de récupérer l'IPTC ultime pour chaque catégorie de profil de risque (ligne de la base de données). La construction de l'ensemble d'entraînement met en exergue une limite dans la modélisation des variables de sinistralité, empêchant de prendre en compte la variable CW. En effet, certaines catégories de sinistres, à la date de vision la plus ancienne n'ont pas de sinistre, donc pas de valeur de CW. Néanmoins, à la date de vision la plus récente, base de données dans laquelle l'IPTC ultime est récupéré, la sinistralité de ces catégories s'est développée. Par conséquent, il n'est pas possible de réaliser la jointure de ces deux bases.

Une solution alternative est alors la prise en compte d'une variable de recul qui indique le développement de la sinistralité.

Variable de recul La variable de recul, notée dans le code `recul_m_dep`, permet de mesure l'intervalle de temps en mois entre la date vision et la date de départ. Sur le schéma expliqué, dans la figure 3.3, elle permet de connaître les sinistres dits développés à la date de vision du 31 juillet 2024 et ceux dont la sinistralité doit encore se développer. Ainsi, cette variable indique l'état de développement de la charge de sinistres. Lorsqu'elle est supérieure à 3, la charge de sinistres est considérée comme développée. La sinistralité est alors dite stabilisée. Lorsqu'elle est inférieure à 3, la charge de sinistres est en cours de développement. C'est cette charge de sinistres en cours de développement que le GBM cherche à prédire.

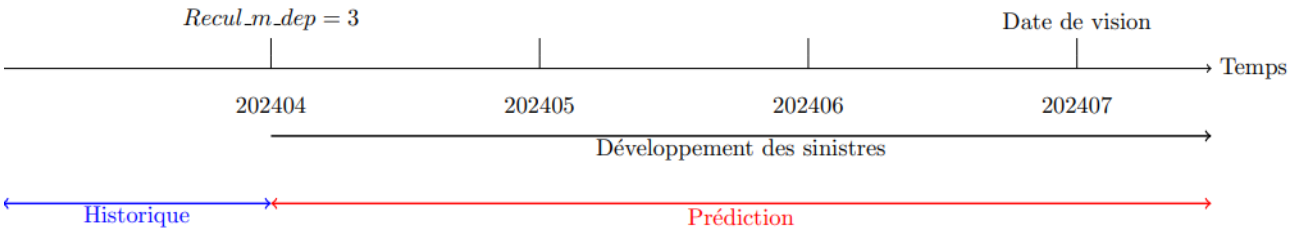

FIGURE 3.3 : Fonctionnement de la variable de recul

Dans ce graphique, la date de vision choisie pour illustrer ce propos est celle du 31 juillet 2024. Cette date est utilisée lors de la vérification des prédictions réalisées par le modèle dans la partie 3.3.2.

IPTC observé La sinistralité passée s'exprime aussi en termes d'IPTC observé. Cette variable correspond à l'IPTC observé à chaque date de vision, date d'extraction des données. Puisqu'il s'agit de prédire l'IPTC ultime, l'IPTC réel observé devient alors une variable explicative.

La figure 3.4 illustre les valeurs d'IPTC observées et calculées sur 12 mois de départ glissants, pour chaque date de vision comprise entre octobre 2023 et juin 2024. La valeur de l'IPTC à la date de vision juin 2024 est donc déterminée à partir des données de juin 2023 à mai 2024.

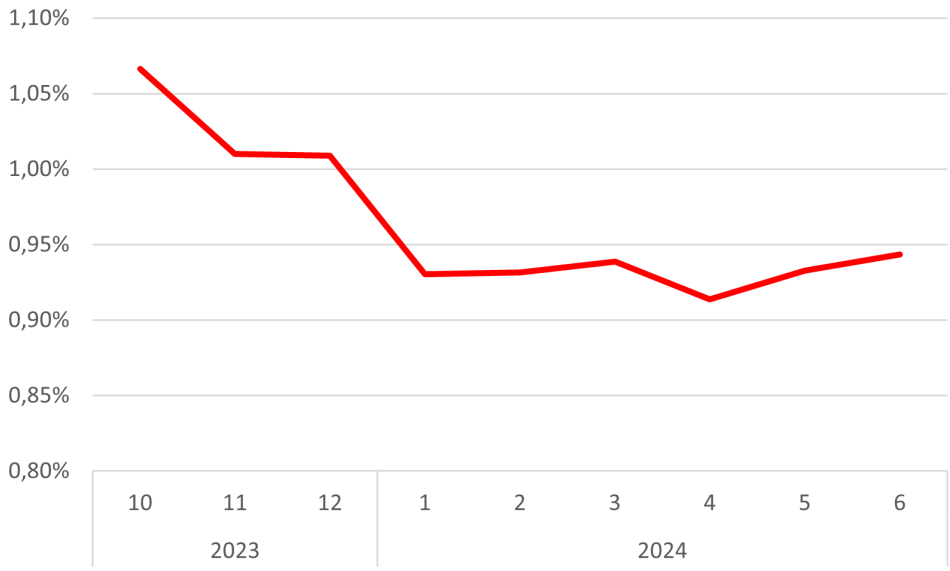

FIGURE 3.4 : Faible variation de l'IPTC observée pour chaque date de vision entre 202310 et 202406 sur 12 mois de départ glissants

L'IPTC reste stable dans le temps, il n'y a pas beaucoup de variations de la sinistralité. A partir de la date de vision de janvier 2024, les valeurs d'IPTC observées sont cantonnées entre $0,9\%$ et $1,1\%$. Les valeurs les plus importantes, correspondant aux dates de vision en fin d'année 2023, sont expliquées par la prise en compte dans le calcul de l'IPTC des données de 2022. Il est nécessaire de rappeler que l'épidémie du Covid-19 a entraîné une hausse des sinistres sur cette période. Il est pertinent d'utiliser les mois de départ pour calculer les 12 mois glissants car avec les mois de souscription, les valeurs d'IPTC sont plus faibles. Celles-ci témoignent d'un faible développement des sinistres lorsque les mois de souscription sont utilisés.

Le graphique 3.5 montre l'IPTC à la date de vision du 31 juillet 2024, en fonction des mois de

souscription. À cette date de vision, la majeure partie des départs a lieu durant les mois estivaux, en juillet et en août. Par conséquent, la sinistralité des contrats souscrits en 2024 (partie droite de la courbe) doit encore se développer. Les départs correspondant à ces souscriptions sont répartis d'une manière non visible sur le graphique. Néanmoins, il faut noter que lorsque la courbe se rapproche du point correspondant à la date de vision, la proportion de contrats souscrits non développés augmente, à cause des nouvelles souscriptions. Ainsi, ces nouveaux contrats ont une date de départ éloignée par rapport à la date de vision ayant pour conséquence une proportion plus faible* de sinistralité déjà déclarée. Pour les voyages souscrits en juin et juillet 2024, à la fin du mois de juin, plusieurs possibilités existent :

- Le voyage a eu lieu récemment : tous les sinistres sont survenus, certains sont déjà connus, mais d'autres ne sont pas encore été déclarés ;
- le voyage n'a pas encore eu lieu : une faible proportion des sinistres est survenue, dont une part encore plus faible est déjà connue.

Une tendance à la hausse d'IPTC se dégage pour les mois d'hiver, tandis que les mois d'été sont moins sinistrés comme le montrent les pics descendants pour les mois de souscription 7 et 8. Comme précédemment, les effets du Covid-19 sont visibles par des pics et des valeurs très élevées de l'IPTC observé pour les exercices 2021 et 2022. Cette étude prend en compte tous les pays de souscription.

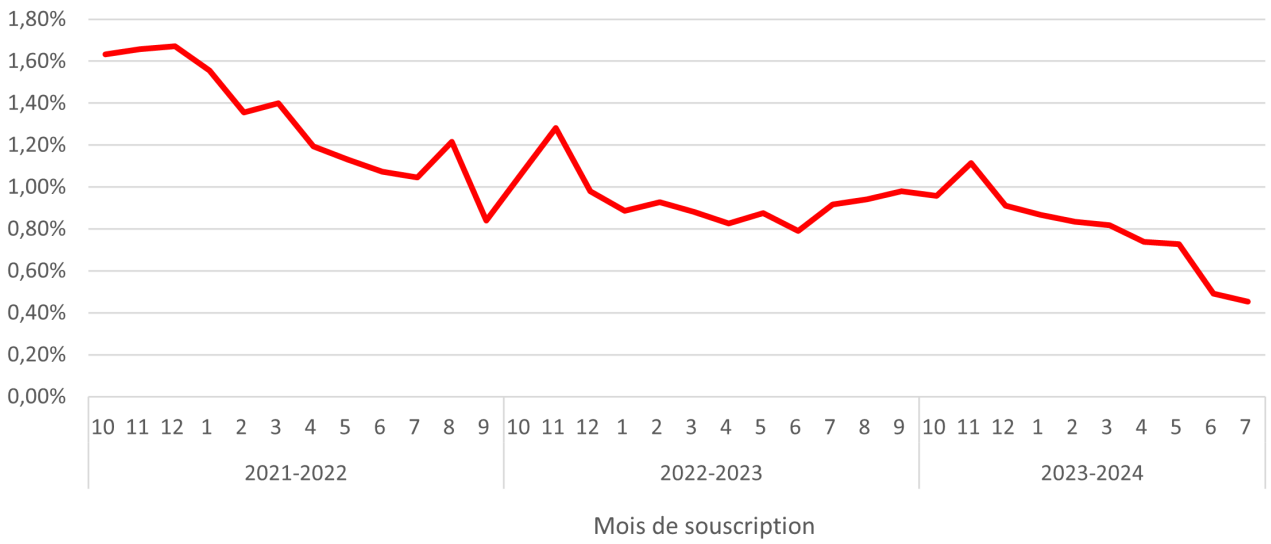

FIGURE 3.5 : IPTC observé à la date de vision 202407

IPTC ultime Il faut rappeler que l'IPTC ultime est la **variable cible** que le modèle cherche à prédire. Elle correspond au pourcentage de charges de sinistres lorsque les sinistres sont développés. Il s'agit finalement des données étudiées à une date de vision plus éloignée dans le temps. Il s'agit, par exemple, de prédire en juillet 2024 (date à laquelle est faite la prédiction), les valeurs de l'IPTC ultime pour tous les mois à venir jusqu'en décembre 2024.

Le graphique 3.6 permet de comparer pour chaque mois de souscription, l'IPTC ultime (IPTC réel à la date de vision 202407) et l'IPTC observé à la fin de chaque mois de souscription.

*Cette proportion est plus faible que les autres points situés sur la courbe hors 2024, par exemple.

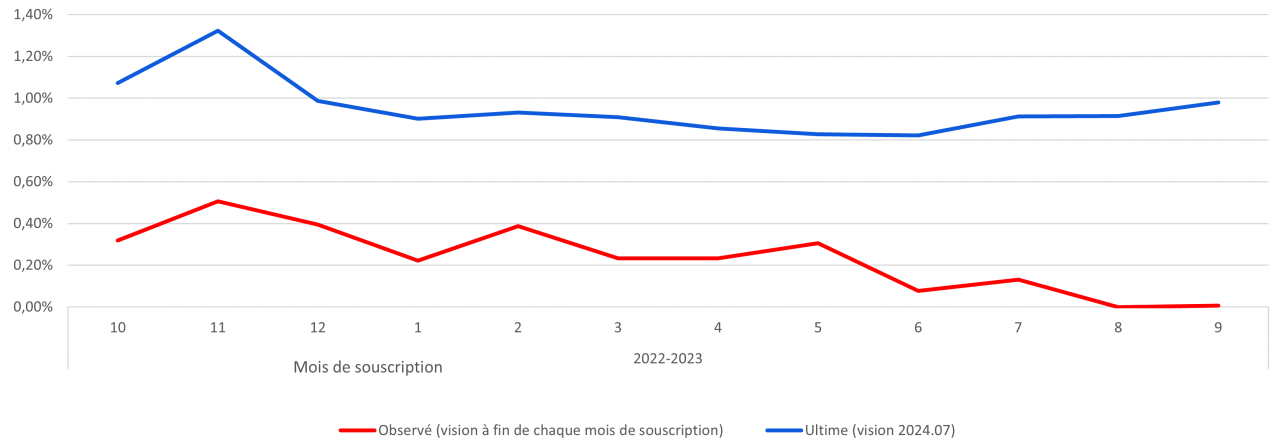

FIGURE 3.6 : Comparaison de l'IPTC par exercice et par mois de souscription ultime et observé

Ce graphique met en avant la différence entre l'IPTC observé à une date et l'IPTC ultime, une fois que les sinistres sont développés. La courbe bleue est donc située au-dessus de la courbe rouge puisque celle-ci représente l'IPTC avec des sinistres en cours de développement.

3.2.2 Présentation des hypothèses de modélisation

Afin de mettre en œuvre la modélisation GBM, le *package* `XGBoost` de `R` documenté par YUAN (2024) a été utilisé. Le *package* `caret`, expliqué par KUHN et al. (2023), ainsi que le *package* `Matrix`, expliqué par BATES et al. (2024), ont été nécessaires pour la réalisation de ce modèle.

Nettoyage des données

Cette modélisation utilise une base de données groupées par catégories selon les profils des voyageurs. Il ne s'agit plus d'une base de données individuelles (ligne à ligne). Les catégories correspondent aux croisements des modalités de chaque variable. Les retraitements effectués dans le premier chapitre, dans la partie 2.2.1, restent d'actualité pour un premier traitement de la base de données.

Le nettoyage des données et la construction d'une base de données adéquate à la réalisation d'un GBM nécessite de récupérer les IPTC ultimes d'une date de vision ultérieure dans le temps. Si le lecteur est perdu quant aux raisonnements sur les dates de vision, il peut se référer au schéma 2.23 permettant de comprendre les décalages entre les dates visualisation des données.

L'ensemble d'entraînement est construit sur 70% des données de la base. L'ensemble de test contient alors les 30% de données récentes afin de vérifier la performance du modèle avec de nouvelles données. Il convient de prêter attention à la taille de l'ensemble de test afin qu'il ne néglige pas une partie des données. La taille de l'ensemble est un paramètre important.

Enfin, il faut alors vérifier que le modèle prédit des données cohérentes avec les données réellement observées. Dans ce contexte, il est nécessaire d'extraire les données d'IPTC à une date de vision, par exemple, décembre 2023, puis de réaliser une jointure afin d'obtenir une colonne correspondant à l'IPTC ultime. Il faut rappeler que l'IPTC ultime est donc l'IPTC observé à une date de vision ultérieure dans le temps, qui pourrait être pour illustrer fin juin 2024.

Traitement des valeurs manquantes et des valeurs aberrantes Afin de pouvoir prédire de manière fiable la charge de sinistres associée à ce compte, les données doivent être nettoyées. À ce titre, aucune donnée manquante ne doit être présente dans la base de données disponible pour la modélisation.

Ainsi, la présence des variables de sinistralité comme la *cancellation window*, qui ne prend une valeur que lorsqu'il y a un sinistre, nécessite un retraitement. Les CW sont réparties en intervalles selon les quantiles de leur distribution. La valeur 0 est affectée à cette variable lorsqu'il n'y a pas de sinistre.

Les valeurs d'IPTC à prédire et IPTC observé trop élevées sont retirées de la base de données afin de réduire le bruit induit dans le modèle. Un traitement spécifique de ces sinistres, dits "graves" est réalisé en parallèle afin de quantifier la charge de sinistres associée à ceux-ci.

Encodage des données de modélisation L'encodage *one-hot* est l'une des méthodes les plus courantes pour traiter les variables catégorielles. Puisque l'algorithme GBM ne permet pas d'utiliser des données non numériques, les variables catégorielles sont alors converties en variables numériques avec *One-Hot Encoding*. Cette fonction de `R` permet de transformer les facteurs de la base de données en variables catégorielles numériques prenant les valeurs 0 ou 1 en fonction de la présence de la caractéristique. En d'autres termes, il s'agit d'une indicatrice. Cette méthode a l'inconvénient d'augmenter rapidement les dimensions de la base de données lorsque les variables catégorielles ont un grand nombre de modalités. Par conséquent, la division des groupes de BW, CW, durée du voyage et de coût du voyage selon les quantiles de leur distribution est une méthode pertinente.

Hyperparamètres du modèle

Le modèle GBM `XGBoost` nécessite la détermination d'hyperparamètres qu'il est possible de classifier en trois catégories.

- Paramétrage des arbres : permet de limiter la profondeur maximale d'un arbre, d'instaurer un nombre minimal d'observations dans une feuille ainsi qu'un nombre d'informations à considérer pour chaque coupure ;
- Paramétrage du *boosting* : permet la gestion du nombre d'arbres à construire dans la forêt, la réduction du poids des arbres successifs, de fixer le gain minimum qui autorise la construction d'un arbre, de fixer les régularisation ainsi que la fonction de perte ;
- Critère d'arrêt : permet de fixer la proportion des données d'entraînement à conserver pour tester l'*early-stopping*, le nombre minimal d'itérations sans améliorations avant d'arrêter l'apprentissage ainsi que la valeur minimale de modification de la fonction de perte qui déclenche l'arrêt prématuré.

Le résultat obtenu après l'entraînement du modèle avec la recherche de grille (*grid search*) est la meilleure combinaison d'hyperparamètres trouvée lors de la validation croisée sur les données d'entraînement.

Nombre d'itérations, `nrounds` $= 100$. Il correspond au nombre d'arbres réalisés par le modèle. Un nombre d'itérations trop élevé peut entraîner un surapprentissage, tandis qu'un nombre d'itérations trop faible peut sous-apprendre. La valeur de 100 est relativement modérée compte-tenu de la faible

valeur de η, le taux d'apprentissage. Le modèle a construit 100 arbres pour parvenir à une bonne performance sans surentraînement.

Taux d'apprentissage, `eta` $= 0,4$**.** Ce paramètre contrôle la contribution de chaque arbre au modèle final. Un taux d'apprentissage trop élevé peut entraîner une convergence prématurée avec un mauvais minimum local, tandis qu'un taux trop faible nécessite un très grand nombre d'arbres pour atteindre une bonne performance. La valeur de $\eta = 0,4$ est utilisée pour éviter le surapprentissage des données et permet d'obtenir un résultat cohérent en 44 itérations.

Profondeur des arbres, `max_depth` $= 10$. Elle doit être limitée pour éviter le surapprentissage. Une profondeur trop importante permet aux arbres de modéliser le bruit des données, ce qui peut conduire à une perte de généralisation. Une profondeur maximale égale à 10 indique que les arbres construits sont moyennement profonds rendant le modèle généralisable.

Réduction minimale de la fonction de perte, $\gamma = 0$ Ce paramètre est lié à l'élagage de l'arbre. Il régule la réduction minimale de la fonction de perte requise pour diviser un nœud. Plus gamma est élevé, moins il y a de divisions rendant le modèle plus simple, comme précisé dans la partie 3.1.2 dédiée à la régularisation d'un problème de ML avec un `XGBoost`. La valeur $\gamma = 0$ signifie que le modèle ne pénalise pas la création de nouveaux nœuds.

Fraction des colonnes, `colsample_bytree` $= 0,7$. La fraction des colonnes correspond aux caractéristiques à échantillonner lors de la construction de chaque arbre. Une valeur de ce paramètre égale à $0,7$ signifie que pour chaque arbre, seulement 70% des caractéristiques disponibles sont utilisées. Le modèle est alors moins exposé au surapprentissage. L'utilisation de cet hyperparamètre permet de faire varier la forme de la courbe bleue.

Poids minimal d'une feuille, `min_child_weight = 1`. Le poids minimal d'une feuille contrôle le nombre minimum d'observations nécessaire pour que la dernière branche d'un arbre soit divisée. Dans cette étude, la valeur déterminée est la valeur standard, égale à 1. Ainsi, chaque nœud doit avoir au moins une observation après chaque division. Cet hyperparamètre n'est donc pas contraignant dans notre cas.

Critère d'arrêt, `early_stopping_rounds` $= 10$ Le critère d'arrêt permet de mettre fin à l'algorithme lorsque la performance sur l'ensemble de validation ne s'améliore plus après un certain nombre d'itérations, ici égal à 10. La combinaison optimale trouvée montre une configuration prudente qui privilégie la généralisation du modèle plutôt que la complexité excessive.

Le tableau suivant récapitule les hyperparamètres utilisés dans cette modélisation *gradient boosting*.

Hyperparamètre	**Description**
`objective = "reg:squarederror"`	Fonction de perte utilisée pour la régression
`max_depth = 10`	Profondeur maximale des arbres
`eta = 0.3`	Taux d'apprentissage
`gamma = 0`	Paramètre de régularisation
`nthread = 2`	Nombre de *threads* utilisés pour l'entraînement
`eval_metric = "rmse"`	Indicateur de performance utilisé pour la régression
`subsample = 0.7`	Fraction des échantillons utilisés pour chaque arbre
`colsample_bytree = 0.7`	Fraction des colonnes utilisées par arbre, fait varier la forme de la courbe
`early_stopping_rounds = 10`	Critère d'arrêt
`nrounds = 100`	Nombre d'itération maximal

TABLE 3.1 : Paramètres pour le modèle de suivi de rentabilité par des variables de sinistralité et de souscription avec `XGBoost`

Après de nombreux tests sur les hyperparamètres du modèle, aucun ne semble réellement permettre d'obtenir un modèle optimal, comme le montre l'assemblage de graphiques 3.7. Il faut comprendre de ces graphiques que le modèle ne parvient pas à modéliser de manière correcte la période de temps dont une partie de la sinistralité est déjà développée. Cette partie correspond à l'intervalle de temps de octobre 2023 à janvier 2023. La charge de sinistres est simple à prédire puisqu'une proportion de celle-ci est déjà développée. Les courbes s'effondrent à partir de la date de vision. Pour rappel, la date de vision correspond à la date d'extraction des données, le 31 décembre 2024. A cette date, la charge de sinistres déjà développée pour les mois de départ à venir est presque nulle. Par conséquent, la charge de sinistres prédite est très faible.

La courbe, dans le carré situé au sud, semble être correcte. Néanmoins, en réalité, il s'agit d'une sortie résultant d'un tâtonnement de paramétrage. Les données de la courbe verte sont obtenues avec les données extraites fin juillet 2024, pour des données dites stables de octobre 2023 à avril 2024. La sinistralité se développe encore légèrement après la date du 31 juillet 2024.

Légende des graphiques 3.7 illustrant l'hyperparamétrage :

— Charge de sinistres réellement observée ;

— Charge de sinistres prédite.

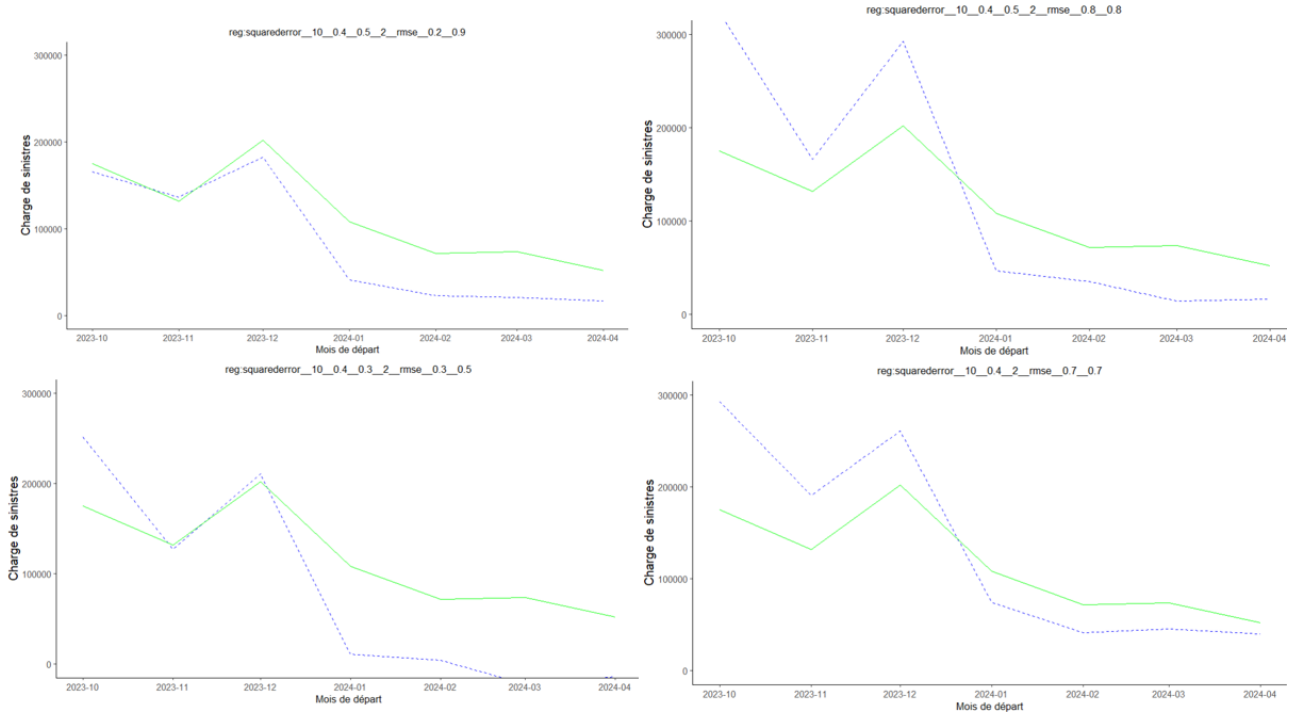

FIGURE 3.7 : Comparaison des résultats de modélisation GBM selon le paramétrage

L'hyperparamétrage retenu est celui permettant d'obtenir la courbe bleue sur le graphique 3.8. Avec ce paramétrage la courbe bleue modélise de la meilleure manière possible les sinistres dont la sinistralité est déjà en cours de développement. De plus, cette courbe bleue suit les tendances de la courbe verte.

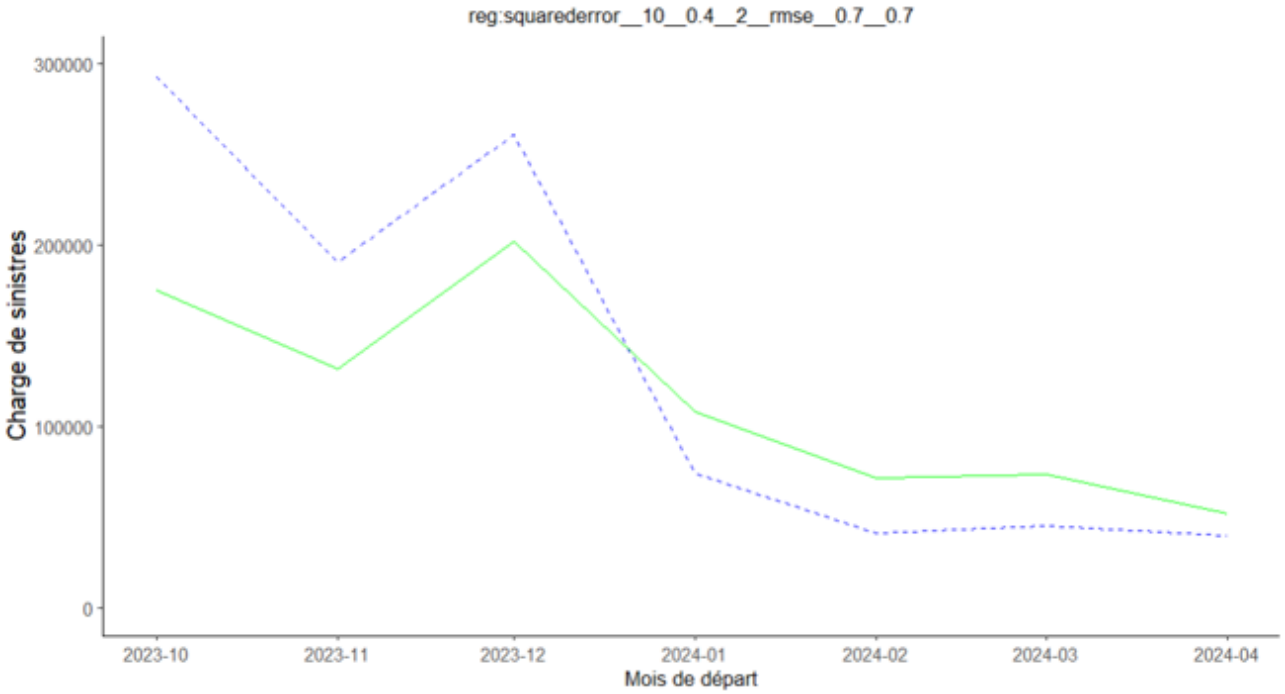

FIGURE 3.8 : Paramétrage optimal

3.3 Analyse des résultats de la modélisation

D'un point de vue opérationnel, il s'agit alors d'observer les résultats du modèle, c'est à dire les prédictions de la charge de sinistres.

3.3.1 Analyse des résultats de la modélisation GBM

Une première approche consiste à analyser les résultats des prédictions versus les valeurs réelles de la base de test afin de tester la qualité d'ajustement du modèle. Afin de comparer les résultats, il convient d'analyser la charge de sinistre prédite qui se calcule grâce à l'expression suivante, pour chaque catégorie i de profil de risque,

$$\text{Charge de sinistres}_i = IPTC_{\text{prédit}_i} \times Trip\ cost_i \tag{3.18}$$

La figure 3.9 permet de constater que le modèle ne prédit pas correctement les valeurs de l'IPTC comprises entre 0 et 1, valeurs généralement observées dans la réalité. Le modèle GBM ne semble pas pouvoir capturer les relations non linéaires et les interactions complexes entre les variables explicatives du jeu de données.

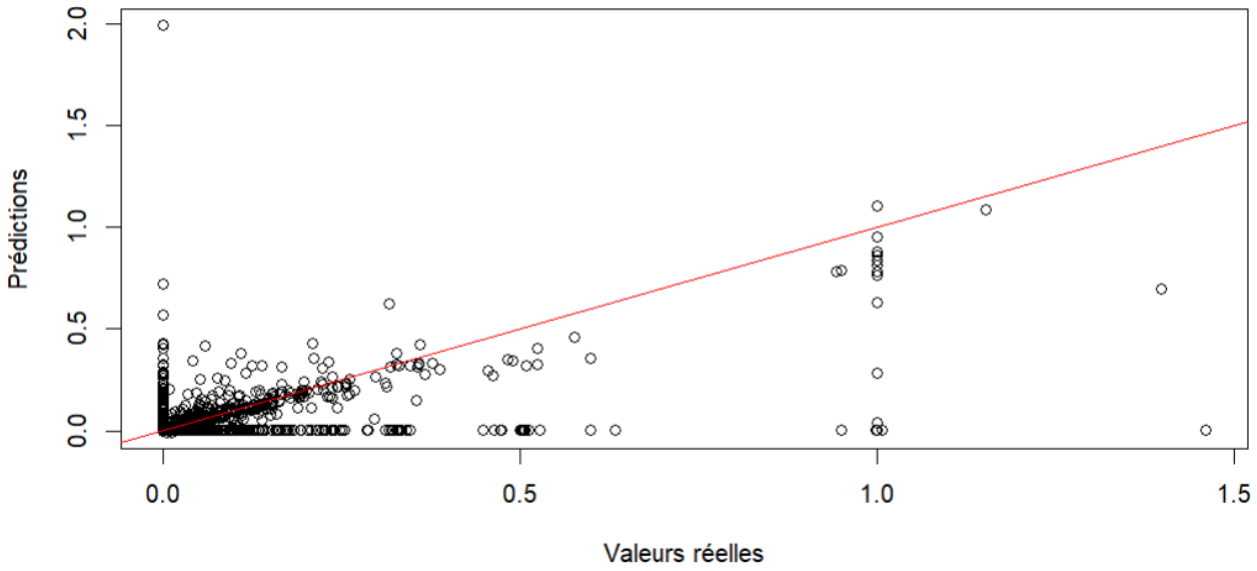

FIGURE 3.9 : Résultats prédiction de l'ensemble de test

Les quelques points situés loin de la ligne rouge indiquent la présence de quelques erreurs de prédiction. De nombreux points se trouvent sous la ligne rouge, indiquant que le modèle sous-estime les valeurs réelles. Les prédictions sont inférieures aux valeurs observées. Ce modèle n'a pas le bon goût d'être un modèle prudent, comme souhaité. En conclusion, cette méthode ne semble pas être pertinente pour un modèle de tarification en assurance. Une analyse plus profonde pourrait indiquer si cela est dû à des *outliers* ou à une capacité limitée du modèle à capturer des liens entre les données. Étant donné la capacité limitée du modèle à prédire correctement les points les plus fréquemment observés, cette analyse n'est pas mise en œuvre. Une première conclusion permet de mettre en avant que certains écarts s'expliquent par l'incapacité du modèle à séparer distinguer les sinistres développés des sinistres en cours de développement. La présence de groupements massifs de points noirs est expliquée par les pénalités* instaurées par le partenaire commercial de Europ Assistance.

Une brève analyse des métriques de performance du modèle, illustrées dans le tableau 3.2, montre que le modèle peut être performant car ces valeurs sont relativement faibles. Il faut noter que malgré ses faibles valeurs, une comparaison avec les résultats obtenus actuellement dans la partie 3.3.2 montre que ce modèle n'est pas pertinent.

Mesure	**Valeur**
Mean Absolute Error (MAE)	0,01
Mean Squared Error (MSE)	0,0033
Root Mean Squared Error (RMSE)	0,057

TABLE 3.2 : Mesure des erreurs de modélisation

Les expressions de ces métriques sont rappelées dans le chapitre précédent, figure 2.6, lors de l'utilisation de ces métriques. La différence entre le MAE et le RMSE indique qu'il existe quelques observations où les erreurs sont bien plus grandes que la moyenne, comme on le voit dans le graphique où certaines prédictions sont très éloignées de la ligne rouge.

*Cette notion est expliquée dans le premier chapitre, partie 1.1.4.

L'analyse de l'importance des variables grâce à une fonction de `XGBoost`, montre que l'IPTC observé est une variable majeure pour prédire l'IPTC ultime. Cette analyse montre que la charge de sinistres ultime est alors majoritairement prédite par la charge de sinistres observée à la date de visualisation des données. La variable `IPTC_dev` est l'interaction entre la variable de stabilité de la sinistralité et la variable de l'IPTC observé. Cette variable permet alors de prendre en compte le développement de la sinistralité.

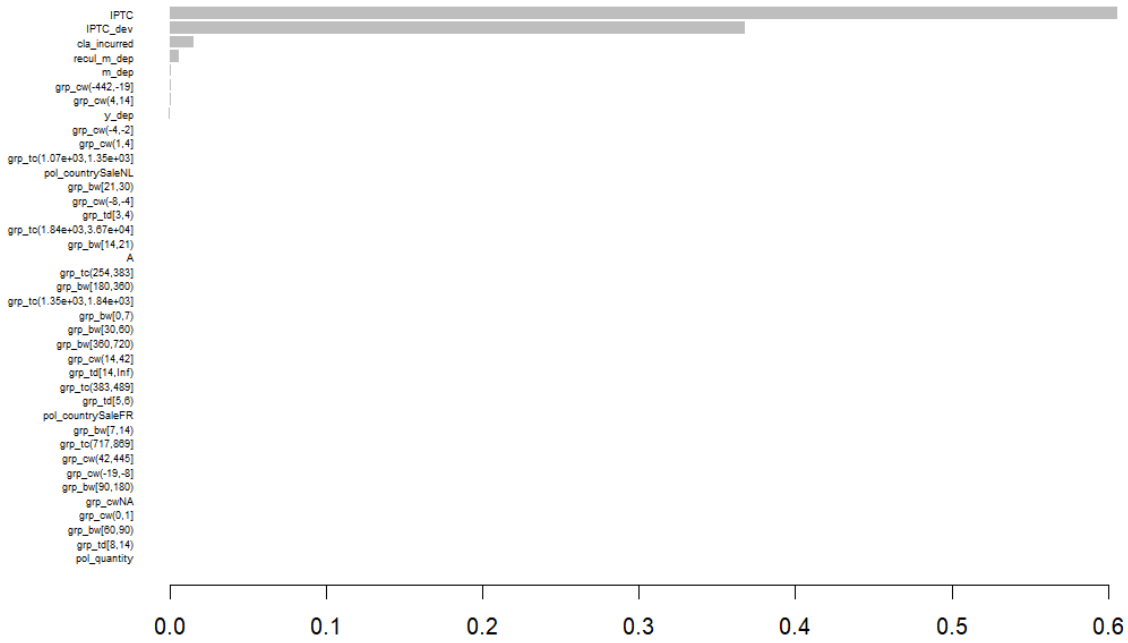

FIGURE 3.10 : Importance des variables dans le modèle `XGBoost`

La longueur de chaque barre indique l'importance relative de la variable dans le modèle. Plus la barre est longue, plus la variable est importante.

3.3.2 Utilisation des résultats du modèle GBM

Dans cette partie de l'étude, deux bases de données sont utilisées. Il s'agit alors de mettre en œuvre une méthodologie de *backtesting** avec les données prédites afin de comparer les prédictions à la réalité observée. Il convient de noter qu'une étude antérieure a permis de fixer une durée de développement de la charge de sinistres égale à trois mois après la date de départ. Trois mois après la date de départ, la charge de sinistres peut évoluer mais de manière non significative.

- Les prédictions sont réalisées sur des données à vision du 31 décembre 2023, comme le montre le graphique 3.11. L'ensemble d'entraînement est un ensemble où les données sont stables. Il correspond à l'intervalle de dates de départ antérieures à fin octobre 2023.

- Les résultats sont extraits à la date de vision de juillet 2024, noté 202407 et correspond à la date de vision ultime à laquelle les sinistres de la date de vision 202312 sont développés. Il s'agit de *backtester* les résultats en observant les données réellement observées au 31 juillet 2024. L'ensemble de test s'arrête en avril 2024 afin de tester le modèle sur des données dont la charge de sinistres est stabilisée.

*Le lecteur peut se référer à la partie 1.3.2 du premier chapitre, afin de comprendre le fonctionnement de cette méthodologie.

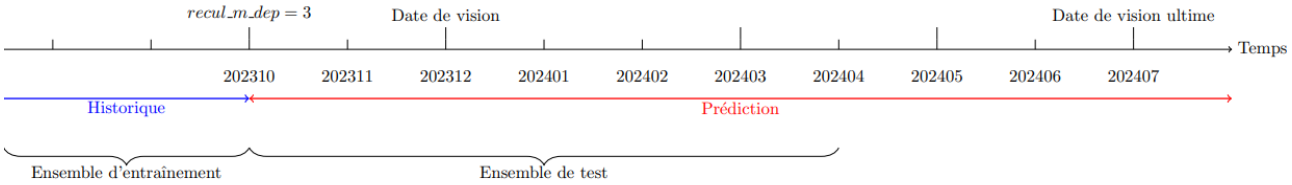

FIGURE 3.11 : Construction de l'ensemble d'entraînement et de test

Comparaison des performances avec le modèle existant

Quelques rappels de modélisation, afin de ne pas mettre la charrue avant les bœufs. Les prédictions sont réalisées à la date de vision 202312 pour des souscriptions réalisées avant cette date. Elles sont donc comparées uniquement à des charges de sinistres pour des souscriptions ayant eu lieu avant 202312. La charge ultime permettant une comparaison est obtenue à la date 202407.

Définition des courbes

L'objectif est de comparer les montants de sinistres réels avec les estimations réalisées. Deux types d'estimations sont réalisées :

- Méthode actuelle - courbe rouge : moyenne sur 12 mois glissants ($m-2$ à $m-13$ de la date de vision choisie), présentée dans le premier chapitre.
- Méthode GBM - courbe bleue : estimation par un *gradient boosting*. L'objectif de cette estimation est de réaliser une meilleure performance prédictive que la méthode actuelle.

Le graphique 3.12 compare les prédictions de charges de sinistres ultimes réalisées par la méthode actuelle (courbe rouge) et la méthode GBM (courbe bleue) aux données de charge de sinistres ultime (courbe verte). Les trois courbes de ce graphiques sont définies avec plus de précisions ci-après. La courbe rouge représente les prédictions de charges de sinistres réalisées par la méthode actuelle à la date de vision du 31 décembre 2023. Pour rappel, la charge de sinistres estimée par la méthode actuelle est calculée à partir d'une moyenne de la charge sur les douze derniers mois de départ glissants. La courbe bleue représente les prédictions réalisées par le GBM en décembre 2023. Enfin, la courbe verte représente la charge ultime stabilisée à la date de vision de 31 juillet 2024.

Dans ce graphique, deux périodes se distinguent, à partir de la date de vision du 31 décembre 2023. Une première période entre les mois de départ d'octobre à décembre 2023 où la charge de sinistres semble prédite correctement par le GBM. La courbe bleue suit les tendances imposées par la courbe verte. Néanmoins, sur la période entre janvier 2024 et avril 2024, la charge de sinistres observée en décembre 2023 est insignifiante (presque nulle) puisque la plupart des annulations ont lieu une semaine avant le départ en voyage. Par conséquent, la charge de sinistres estimée par le GBM (courbe bleue) s'effondre brutalement. Cette chute de la charge de sinistres s'explique, comme mis en avant dans la figure 3.10, parce que le modèle GBM utilise en grande partie la charge de sinistres observée pour prédire l'ultime. Ainsi, pour ces mois de départ où peu de sinistres sont déclarés, la charge de sinistres ultime prédite est presque nulle.

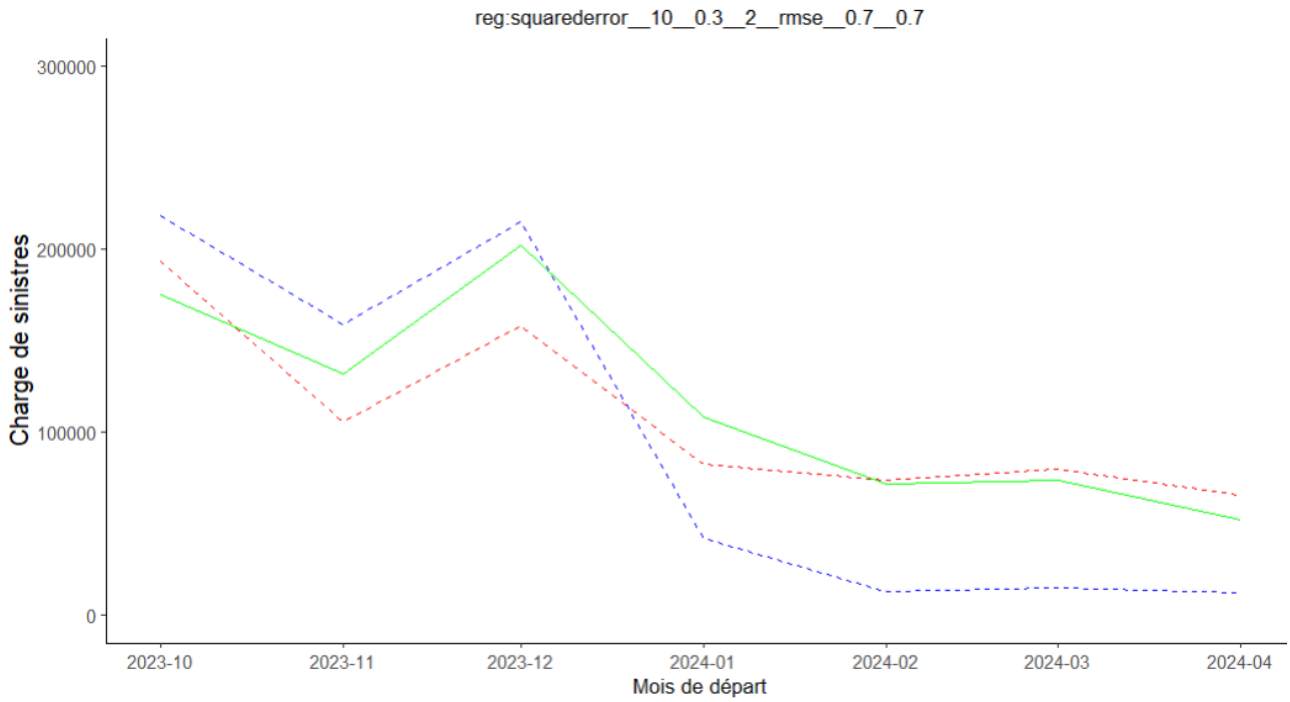

FIGURE 3.12 : Comparaison des résultats de prédiction de l'IPTC ultime avec des variables de sinistralité

Ce graphique permet de conclure que l'insertion des variables de sinistralité observée pour prédire la sinistralité ultime pour chaque mois de départ avec un modèle de *gradient boosting* n'est pas pertinent. Cette modélisation n'est pas viable pour un assureur qui souhaite estimer la charge de sinistres pour les mois de départ à venir. Ainsi, il semble que la prise en compte des variables de sinistralité pour prédire la charge de sinistres n'est pas pertinente et ne permet pas d'améliorer les résultats obtenus par la méthode actuelle. Néanmoins, la méthode de modélisation offerte par le *gradient boosting* est une modélisation intéressante et prometteuse qui semble pouvoir pallier les inconvénients de la méthode actuelle. En effet, ce modèle contient les données de souscription, ce qui lui permet de s'adapter rapidement aux modifications de composition du portefeuille. Par conséquent, afin de vérifier cette hypothèse, un modèle de *gradient boosting* contenant uniquement les variables de souscription est réalisé.

Utilisation du modèle de suivi de rentabilité

Il convient de rappeler que l'objectif de cette étude est de concevoir un modèle performant pour prédire la charge de sinistres et affiner le suivi de la rentabilité du compte de ce partenaire commercial. Afin d'étudier la rentabilité du compte, l'outil de pilotage de Europ Assistance permet, chaque mois, après réception des données de sinistres, de mettre à jour la sinistralité observée pour ce compte. L'outil actuel permet de suivre la rentabilité du compte du partenaire commercial grâce au graphique 1.16 présenté dans le premier chapitre. Il permet la comparaison entre le *loss* ratio (LR) projeté et réel, ainsi qu'avec le COR cible et le COR réel.

L'objectif du modèle GBM est d'affiner les prédictions afin d'offrir un meilleur suivi de la rentabilité du compte. À ce titre, il s'agit d'intégrer cette méthode à l'outil de suivi de performance actuellement mis en place. Le suivi de cette rentabilité permet d'évaluer la pertinence du tarif mis en place ou encore de quantifier les impacts endogènes et de saisonnalité sur la charge de sinistres en temps réel.

La capacité à prédire la rentabilité permet d'ajuster la tarification ou encore d'agir sur le partage des bénéfices créés par un contrat.

La prédiction de la rentabilité du compte est utilisée dans l'établissement des provisions techniques, provisions pour sinistres à payer ou provisions pour primes non acquises. Une mauvaise estimation de la rentabilité peut entraîner des erreurs de provisionnement. Celles-ci exposent l'assureur à des risques financiers ou à un surplus de provision. Dans une optique commerciale, la prédiction de la rentabilité du compte est une aide à la prise de décisions stratégiques telles que l'élargissement d'un partenariat, le lancement de nouveaux produits ou encore le ciblage de nouveaux segments de marché. Il s'agit aussi d'allouer de manière optimale le capital humain et financier qui sont à disposition de l'assureur.

Dans ce chapitre, il s'agit de prédire la charge de sinistres ultime à partir de variables de souscription et de sinistralité. Néanmoins, après étude du modèle, la combinaison de ces deux types de variables ne semble pas être une méthode optimale pour évaluer la rentabilité du compte. En conséquence, un modèle GBM à partir des variables de souscription est alors mis en place pour prédire la charge de sinistres et montrer la pertinence du modèle de *gradient boosting*.

Pour rappel, dans un modèle dit de souscription, la variable cible est l'IPTC observé à une date, ici le 31 juillet 2024, à partir des données de souscription disponibles, à la date du 31 décembre 2023. Le montant de la charge de sinistres se calcule d'après l'expression 3.18. Le montant de prime pure pour chaque contrat est

$$\text{Prime Pure}_i = IPTC_{\text{prédit}_i} \times \textit{Trip cost}_i. \tag{3.19}$$

Ainsi, à partir des variables de souscription utilisées dans le GLM, examiné sous toutes les coutures dans le chapitre 2, un modèle GBM a été construit. Ce modèle permet d'obtenir des résultats cohérents qui améliorent la prédiction réalisée par le modèle actuel. Il devient alors un modèle complémentaire à la régression linéaire généralisée mise en place dans le chapitre 2. Les résultats de ce modèle sont étudiés brièvement. À ce titre, le graphique ci-dessous montre que le modèle GBM réalisé à partir des variables de souscription capte les tendances imposées par la courbe verte. Il prédit la charge de sinistres de manière raisonnée et prudente. Les résultats performants de ce modèle permettent de mettre en avant l'intérêt du modèle de *gradient boosting*. Comme expliqué précédemment, ce modèle est conçu à partir des données de souscription, lui permettant de s'adapter rapidement aux modifications de composition du portefeuille, telles qu'une modification des habitudes de réservations ou une différence de répartition des souscriptions selon le pays d'achat de la police. Cette modélisation peut alors être utilisée par Europ Assistance et intégrée dans son processus de suivi de rentabilité.

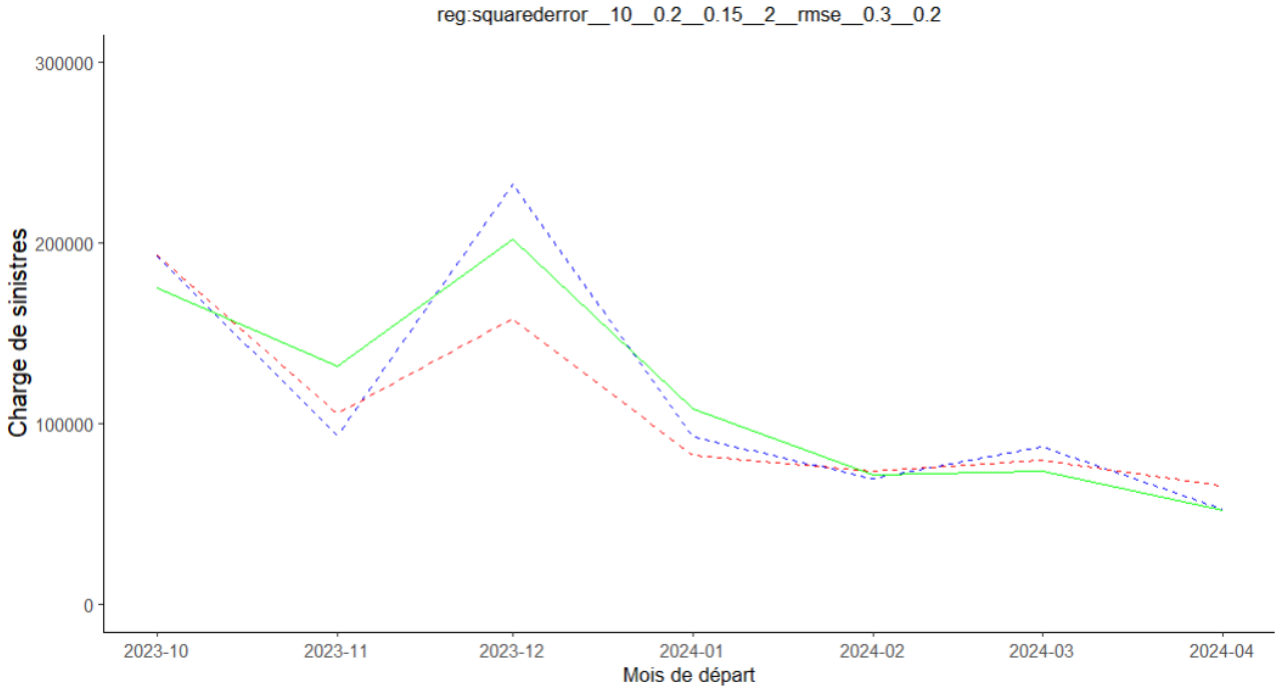

FIGURE 3.13 : Comparaison des résultats de prédiction de l'IPTC ultime avec des variables de souscription

Les hyperparamètres, dont les définitions sont explorées en détails dans la partie 3.2.2, sont présentés dans le graphique ci-après. Néanmoins, le nombre d'hyperparamètres, souvent considéré comme un avantage de cette méthode de *gradient boosting* est ici un inconvénient. Ce surnombre complexifie l'explication et la réplicabilité du modèle aux autres comptes de partenaires commerciaux.

Hyperparamètre	**Description**
`objective = "reg:squarederror"`	Fonction de perte utilisée pour la régression
`max_depth = 10`	Profondeur maximale des arbres
`eta = 0.3`	Taux d'apprentissage
`gamma = 0.3`	Paramètre de régularisation
`nthread = 2`	Nombre de *threads* utilisés pour l'entraînement
`eval_metric = "rmse"`	Indicateur de performance utilisé pour la régression
`subsample = 0.2`	Fraction des échantillons utilisés pour chaque arbre
`colsample_bytree = 0.2`	Fraction des colonnes utilisées par arbre, fait varier la forme de la courbe
`early_stopping_rounds = 10`	Critère d'arrêt
`nrounds = 100`	Nombre d'itération maximal

TABLE 3.3 : Paramètres pour le modèle de suivi de rentabilité par des variables de souscription avec `XGBoost`

Les variables de souscription utilisées et leur contribution à l'explication de l'IPTC sont indiquées dans le graphique 3.14.

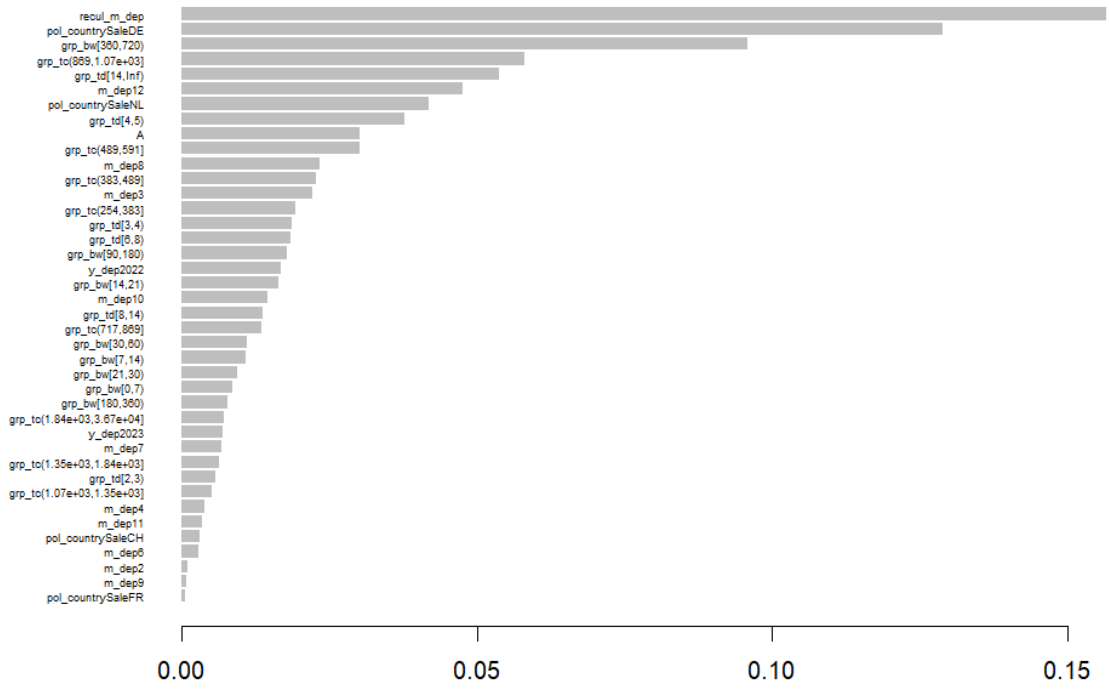

FIGURE 3.14 : Comparaison des résultats de prédiction avec des variables de souscription

Le pays de souscription et le recul du mois de départ par rapport à la date de vision sont des variables importantes qui permettent de prédire l'IPTC observé. Les intervalles de BW et de durée de voyage, ainsi que les tranches tarifaires de *trip cost* permettent de discriminer les profils des voyageurs et d'établir un pourcentage de charge de sinistres pour chaque contrat. Cette prédiction représente alors la prime pure du contrat. Ce modèle "ligne à ligne", bien que pertinent pour les résultats demande un temps de calcul beaucoup plus important que le GLM. En outre, l'aspect communément appelé "boîte noire" des GBM, difficulté à interpréter son fonctionnement interne et sa prise de décisions, est un enjeu, malgré les différentes méthodes disponibles pour éclaircir les processus. Ces limites sont des éléments non négligeables dans la prise de décision de l'intégration de ce modèle dans le processus de suivi de rentabilité de Europ Assistance.

Conclusion

Pour conclure, le modèle GBM avec l'insertion des variables de sinistralité n'est pas un outil efficace pour prédire l'IPTC ultime. La prise en compte de la charge de sinistres observée ne permet pas d'obtenir de meilleurs résultats de prédiction pour le suivi de la rentabilité. Contrairement à ce qui était espéré, le modèle ne détecte pas de lui-même la stabilisation des sinistres via la variable de recul indiquant l'état du développement de la sinistralité. Une solution alternative est alors de construire un modèle GBM dont les entrées ne contiennent que des informations disponibles lors de la souscription. Cette méthodologie est pertinente et permet de compléter l'étude menée dans le chapitre

2. Néanmoins, compte-tenu des différents inconvénients d'un modèle de *gradient boosting*, malgré des résultats probants, cette méthode avec des variables de tarification n'est pas retenue par Europ Assistance pour affiner son suivi de rentabilité. Cette approche permet alors à Europ Assistance de posséder une ébauche de modèle de suivi de rentabilité adaptée à la structure de ses données, si le besoin de suivi de rentabilité évolue.

Conclusion

Ce présent mémoire s'est attelé à l'élaboration de modèles prédictifs visant à améliorer la tarification et le suivi de la rentabilité d'un partenaire commercial de Europ Assistance. L'objectif de cette étude est de répondre à la question, "comment prédire la sinistralité des garanties annulation pour un compte spécifique ?", en s'appuyant sur des modèles statistiques. Afin de répondre à cette problématique, des modèles de régression et de *gradient boosting* sont étudiés pour estimer la sinistralité en cours de développement des contrats vendus par le biais du partenaire commercial de Europ Assistance.

Dans un premier chapitre, cette étude se consacre à l'approfondissement des notions liées à l'assurance voyage et plus précisément à la garantie annulation. Les notions importantes de vocabulaire et de calculs de rentabilité y sont expliquées. Dans cette étude, deux types de variables sont distingués : les variables de souscription (disponibles lors de la souscription d'une police d'assurance) et les variables de sinistralité (porteuses d'information sur la sinistralité passée). Suite à cet exposé, la méthode de calcul actuelle est présentée afin de comprendre l'apport d'un nouveau modèle.

Le deuxième chapitre est consacré à la construction d'un modèle linéaire généralisé (GLM) afin d'analyser les risques associés à la souscription d'une assurance. Ce modèle est conçu, uniquement, à partir des variables disponibles à la souscription d'un contrat d'assurance. Il prend en compte plusieurs variables comme le mois de départ, la durée du voyage, l'intervalle entre la réservation et le départ, appelée *booking window* ainsi que le coût du voyage. Une variable importante est le mois de départ. Elle permet de discriminer les contrats dont la sinistralité est stabilisée versus ceux dont la sinistralité est en cours de développement. Cette distinction permet d'utiliser des modèles de régression afin d'estimer cette sinistralité. Une analyse statistiques des données permet d'abord de comprendre comment ces facteurs influencent la probabilité de survenance d'un sinistre. Les conclusions de cette analyse ont permis de construire un modèle GLM pertinent. Ce modèle, par son approche paramétrique, produit des résultats satisfaisants en termes de performance prédictive, pour les contrats dont la sinistralité n'est pas encore développée. Ainsi, il rend possible la mise en place d'un modèle de tarification adéquat qui s'incorpore dans les outils déjà utilisés pour le suivi de ce compte. Ce modèle établit la prime pure payée par l'assuré en fonction de ces caractéristiques de souscription et permet de suivre ainsi la rentabilité du compte étudié.

Enfin, le troisième chapitre est consacré à l'étude de la prise en compte par les modèles prédictifs des variables de sinistralité observée. Pour ce faire, un modèle plus sophistiqué est utilisé, le *gradient boosting*. L'objectif de ce modèle est de prendre en compte des variables porteuses d'informations sur la sinistralité passée. Il s'agit alors de détecter le niveau de stabilisation des sinistres et d'adapter les prédictions en fonction de cette observation et des montants déjà payés. Il apparaît cependant, que le comportement du modèle n'est pas celui espéré puisqu'il ne permet pas de distinguer l'état de développement du sinistre. Ainsi, la qualité des prédictions ne s'améliore pas avec l'introduction de ces informations. Les données de sinistralité semblent plutôt apporter de la confusion en demandant au modèle de prédire des valeurs très différentes de la variable cible. En outre, le caractère communément appelé "boîte noire" de ce modèle de *machine learning* est un frein à l'utilisation de ce modèle par

Europ Assistance. Cette partie de l'étude permet néanmoins d'explorer cette nouvelle approche de modélisation pour Europ Assistance. Elle constitue alors une première ébauche destinée à être approfondie en fonction des besoins de l'entreprise.

Puisqu'à ce stade, le modèle GBM mis en œuvre dans cette étude ne permet pas de détecter la stabilisation de la charge de sinistre via une variable discriminante (distance entre le mois de départ et la date d'extraction des sinistres), une solution alternative est alors choisie pour intégrer le processus actuel de suivi de rentabilité. En combinant la modélisation GLM avec la modélisation actuelle, qui sépare manuellement les sinistres selon l'état de développement de leur charge, le modèle obtenu est pertinent. Dans cette modélisation, Europ Assistance a choisi de retenir l'utilisation du GLM pour prédire la sinistralité des contrats dont la sinistralité n'est pas développée. Tandis que pour les contrats ayant une sinistralité développée, il s'agit d'appliquer la méthode actuelle.

Pour conclure, le modèle GLM, en se combinant à la méthode actuelle permet d'améliorer significativement les estimations réalisées mensuellement, en se basant sur les informations de souscription. L'ajout des informations de sinistralité, présentées dans le chapitre 3, ne permet pas au modèle de détecter la stabilisation de la sinistralité et n'apporte pas d'amélioration des résultats de prédiction. Néanmoins, cette étude propose une solution concrète pour le suivi de la tarification et de la rentabilité de ce contrat d'assurance annulation en respectant les contraintes imposées par la modélisation actuellement utilisée. Elle peut alors être utilisée pour suivre la rentabilité des autres partenaires commerciaux de Europ Assistance.

Bibliographie

Adobe (2023). What is B2B2C Ecommerce? URL : https://business.adobe.com/fr/blog/basics/what-is-b2b2c-ecommerce.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19.6, p. 716-723.

Akerlof, G. (1970). The Market for Lemons: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* 84.3, p. 488-500.

Akur8 (2024). Risk Modelling and Pricing Solution. URL : https://fr.akur8.com/pricing/risk.

Allianz Partners (2023). Customer Lab. URL : https://www.allianz-partners.com/fr_FR/services/customer-lab.html#customer-lab.

Anderson, B. (2024a). Comment effectuer un test de Breusch Pagan dans R ? URL : https://statorials.org/breusch-pagan-test-r/.

Anderson, B. (2024b). Erreurs robustes de White. URL : https://statorials.org/blanc-test-in-r/.

Anderson, B. (2024c). Régression et Multicolinéarité. URL : https://statorials.org/regression-multicolinearite/.

Arrow, K. J. (1950). The Economics of Information. *The American Economic Review* 40.2, p. 94-111.

AXA Partners (2023). Les challenges à venir pour l'assurance voyage. Rapp. tech. AXA Partners. URL : https://www.axapartners.com/fr/blog/vue-expert-prochaines-tendances-assurance-voyage.

Bates, D., Maechler, M., Jagan, M. et Davis, T. A. (2024). Matrix: Sparse and Dense Matrix Classes and Methods. Version 1.7-0. R Project. URL : https://Matrix.R-forge.R-project.org.

Cornillon, P.-A., Hengartner, N., Matzner-Løber, E. et Rouvière, L. (2019). Régression avec R. EDP Sciences. Presses Universitaires de France.

Delong, L., Lindholm, M. et Wüthrich, M. V. (juin 2021). Making Tweedie's compound Poisson model more accessible. *European Actuarial Journal* 11.1. Creative Commons Attribution 4.0 International, p. 207-226.

Dunn, P. K. (17 août 2022a). Package 'tweedie'. Version 2.3.5. Available at CRAN. URL : https://cran.r-project.org/web/packages/tweedie/index.html.

Dunn, P. K. (2022b). Package Tweedie - Evaluation of Tweedie Exponential Family Models. Version 2.3.5. URL : https://cran.r-project.org/web/packages/tweedie/tweedie.pdf.

Europ Assistance (2018). Baromètre des vacances 2018. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr.

Europ Assistance (2019a). Baromètre des vacances 2019. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr.

Europ Assistance (2019b). Notre histoire. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr.

Europ Assistance (2020). Baromètre des vacances 2020. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr.

Europ Assistance (2021). Baromètre des vacances 2021. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr.

Europ Assistance (2022). Baromètre des vacances 2022. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr.

Europ Assistance (2023). Baromètre vacances 2023. Rapp. tech. Europ Assistance. URL : https://www.europ-assistance.fr/fr/partenaires/media-room/publications/barometre-des-vacances-2023-ipsos-europ-assistance.

Europ Assistance SA (2019). Rapport sur la solvabilité et la situation financière. Rapp. tech. Europ Assistance SA. URL : https://www.europ-assistance.com/publications/#:~:text=Discover%20a%20comprehensive%20repository%20of%20essential%20Company%20documents,.

Europ Assistance SA (2020). Rapport sur la solvabilité et la situation financière. Rapp. tech. Europ Assistance SA. URL : https://www.europ-assistance.com/publications/#:~:text=Discover%20a%20comprehensive%20repository%20of%20essential%20Company%20documents,.

Europ Assistance SA (2021). Rapport sur la solvabilité et la situation financière. Rapp. tech. Europ Assistance SA. URL : https://www.europ-assistance.com/publications/#:~:text=Discover%20a%20comprehensive%20repository%20of%20essential%20Company%20documents,.

Europ Assistance SA (2022). Rapport sur la solvabilité et la situation financière. Rapp. tech. Europ Assistance SA. URL : https://www.europ-assistance.com/publications/#:~:text=Discover%20a%20comprehensive%20repository%20of%20essential%20Company%20documents,.

Europ Assistance SA (2023). Rapport sur la solvabilité et la situation financière. Rapp. tech. Europ Assistance SA. URL : https://www.europ-assistance.com/wp-content/uploads/2024/04/S2_SFCR_S-FR-9695008E6296RUK5A776-SIG1_SFCR_S-2023-12-31.pdf.

Forbes France (2024). Phenomen révolutionne l'assurance annulation - Forbes France. Rapp. tech. Forbes France. URL : https://www.forbes.fr/business/phenomen-revolutionne-lassurance-annulation/.

France Assureurs (2024). Séjour à l'étranger, assurances et assistance. Rapp. tech. France Assureurs. URL : https://www.franceassureurs.fr/lassurance-protege-finance-et-emploie/lassurance-protege/lassurance-en-pratique-pour-les-particuliers/sejour-etranger-assurances-et-assistance/.

France TV Info (août 2023). Assistance voyageurs : "Un porteur de carte bancaire basique peut se croire couvert, il ne l'est pas", avertit le président d'Europ Assistance France. URL : https://www.francetvinfo.fr/replay-radio/l-interview-eco/assistance-voyageurs-un-porteur-de-carte-bancaire-basique-peut-se-croire-couvert-il-ne-l-est-pas-avertit-le-president-d-europ-assistance-france_6542264.html.

Frédéric Planchet, Antoine Miseray (2023). Tarification IARD - Introduction aux techniques avancées. ISFA.

Gnonlonfoun, H. (2023). Construction d'une loi d'acquisition des primes en assurance voyage. Mémoire d'actuariat. URL : https://www.institutdesactuaires.com/se-documenter/memoires/memoires-d-actuariat-4651?id=9b2be2ee7bd36b6fe19035c0e74a10be.

Grasland, C. (2000). Initiation aux Méthodes Statistiques en Sciences Sociales. Université Paris VII / UFR GHSS.

Institut des Actuaires (2017). Cours Non-Vie, Séance 2 - Construction des primes : Les principes et les méthodes. Rapp. tech. Institut des Actuaires. URL : https://www.institutdesactuaires.com/global/gene/link.php?news_link=2017115414_2017-cea-cours-non-vie-seance-2.pdf&fg=1.

Institut des Actuaires (2024). 10 raisons de devenir actuaire. https://www.institutdesactuaires.com/devenir-actuaire/10-raisons-de-devenir-actuaire-29.

Ipsos (2024). Près de 7 Français sur 10 partiront en vacances cet été malgré un contexte économique et international tendu. Rapp. tech. Ipsos. URL : https://www.ipsos.com.

Jeunesse, M. (2023). Early Stopping et Gradient Boosting : Extraits de la Littérature. Conférence RE2A : Gradient Boosting et Applications, Initiative de Recherche Risques Émergents ou Atypiques en Assurance, mai 2023, https://www.institutdesactuaires.com/evenements/1071/details.

Kassel, R. (2023). Hyperparamètres : Qu'est-ce que c'est ? À quoi ça sert ? URL : https://datascientest.com/hyperparametres-tout-savoir#:~:text=Un%20hyperparam%C3%A8tre%20est%20un%20param%C3%A8tre%20qui%20est%20utilis%C3%A9,par%20l%E2%80%99utilisateur%20avant%20de%20commencer%20l%E2%80%99entra%C3%AEnement%20du%20mod%C3%A8le..

Kuhn, M. et al. (2023). caret: Classification and Regression Training. Version 6.0-94. URL : https://cran.r-project.org/web/packages/caret/index.html.

L'Argus de l'Assurance (août 2023). We Connect, le nouveau plan stratégique d'Europ Assistance. URL : https://www.argusdelassurance.com/acteurs/assisteurs/we-connect-le-nouveau-plan-strategique-d-europ-assistance.99639.

Leconte, P. (1978). Les Bronzés. https://youtu.be/hxgrUAHgIzs. Extrait du film.

Lopès-Quinta, T. (2022). Python-M2-ISF - Cours de Python pour le master ISF à Paris-Dauphine. https://github.com/theo-lq/Python-M2-ISF.

Marceau, E. (2013). Modélisation et évaluation quantitative des risques en actuariat : Modèles sur une période. Collection Statistique et probabilité appliquées. Springer.

Marri, F. (2016). Chapitre 3 : Les méthodes de provisionnement. https://www.actuarialab.net/wp-content/uploads/partie4.pdf.

Next Move Strategy Consulting (2023). Travel Insurance Market to Reach USD 58.93 Billion by 2030, Growing at a CAGR of 16.7% from 2023 to 2030. URL : https://www.nextmsc.com/news/travel-insurance-market.

OpenAI (2024). ChatGPT. https://www.openai.com. Utilisé à des fins d'aide à la programmation, de reformulation et d'enrichissement de l'étude.

Perrin, G. (14 juin 2024). Le courtage sans frontières, L'ASSURTECH DE LA SEMAINE : Yupwego veut changer les règles de l'assurance voyage. *L'Argus de l'assurance* 7862. SPÉCIAL RENDEZ-VOUS DE LYON, p. 61.

Québec Centre for Biodiversity Science (2023). Chapitre 4 Régression linéaire avec R. URL : https://r.qcbs.ca/workshop04/book-fr/r%C3%A9gression-lin%C3%A9aire-avec-r.html.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL : https://www.R-project.org/.

RE, C. (2021). Qu'est-ce que le ratio combiné ? https://blog.ccr-re.com/fr/qu-est-ce-que-le-ratio-combine.

Suru, A. (2020). Assurance IARD - Les dessous d'un secteur qui vous protège. Economica. Economica.

Travel Weekly (2024). Cover-More pulls insurance product in Australia and New Zealand. Rapp. tech. Travel Weekly. URL : https://www.travelweekly.com.

Travelers Insurance (2024). Travelers History. Rapp. tech. Travelers Insurance. URL : https://www.travelers.com.

Wikipédia (2023). Cumulant statistique. URL : https://fr.wikipedia.org/wiki/Cumulant_(statistiques).

Wikipédia (mars 2024a). Cramér's V. URL : https://en.wikipedia.org/wiki/CramÃľr%27s_V.

Wikipédia (2024b). Prêt à la grosse aventure. URL : https://fr.wikipedia.org/wiki/Pr%C3%AAt_%C3%A0_la_grosse_aventure.

Wikipédia (mars 2024c). Théorème de Taylor. URL : https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_Taylor.

Wikipédia (2024d). Tweedie distribution. URL : https://en.wikipedia.org/wiki/Tweedie_distribution.

Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Chapman et Hall/CRC.

Yuan, J. (2024). Package 'xgboost'. Version 1.7.8.1. URL : https://github.com/dmlc/xgboost.

Zuer, M. (2024). Data Split - Machine Learning. URL : https://mzuer.github.io/machine_learning/data_split.

Annexe

Démonstration : appartenance à la famille exponentielle

Cette partie a pour objectif de démontrer que les distributions de Tweedie appartiennent à la famille exponentielle. Les notions liées à ce type de distribution sont expliquées dans le chapitre 2.

Définition d'une distribution de la famille exponentielle

Une variable aléatoire possède une densité de probabilité appartenant à la famille exponentielle si elle peut s'écrire sous la forme : $f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$ où :

- θ est le paramètre canonique ;
- ϕ est le paramètre de dispersion ;
- $b(.)$ est trois fois dérivable et b' inversible ;
- $c(y, \phi)$ est une fonction qui ne dépend que de y et de ϕ.

Si Y admet une densité appartenant à la famille exponentielle , alors :

- $\mathbb{E}_\theta[Y] = b'(\theta)$,
- $\mathrm{Var}_\theta[Y] = b''(\theta)\phi$.

Il faut noter par ailleurs que cette décomposition permet de déterminer la fonction de lien adéquate.

Pour démontrer que les distributions de Tweedie appartiennent à cette famille, il n'est pas possible d'utiliser la méthode usuelle à partir de la densité de probabilité. Il est nécessaire de partir de la relation variance moyenne spécifique à ces distributions. Ensuite, il faut utiliser le lien avec la fonction génératrice de cumulants, afin d'obtenir la forme générale d'une densité de distribution appartenant à la famille exponentielle.

Forme générale de la distribution de Tweedie

Les distributions de Tweedie sont définies par une relation de variance-moyenne de la forme suivante : $\mathrm{Var}(Y \mid \mu) = \phi\mu^p$ où ϕ est le paramètre de dispersion et p est le paramètre de puissance de la distribution de Tweedie. Cette relation indique que la variance d'une variable aléatoire Tweedie est proportionnelle à une puissance de sa moyenne.

Afin de déterminer la forme de sa distribution, il est nécessaire de passer par la fonction génératrice de cumulants.

Calcul de la fonction de log-vraisemblance pour une distribution de Tweedie

La fonction de log-vraisemblance des distributions de Tweedie peut être dérivée de la relation variance-moyenne. Si Y suit une distribution de Tweedie avec moyenne μ, la log-vraisemblance d'une observation y s'écrit comme suit :

$$\log L(y \mid \mu, \phi) = \frac{1}{\phi}\left(\frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right) + c(y, \phi),$$

où ϕ est le paramètre de dispersion et $c(y, \phi)$ est une fonction qui ne dépend que de y et ϕ. Cette log-vraisemblance est dérivée à partir de la structure de variance-moyenne et peut être mise en relation avec la structure de la famille exponentielle.

La log-vraisemblance est construite à partir de la relation variance-moyenne et des propriétés des distributions de Tweedie. Pour les distributions de Tweedie, la fonction de variance est proportionnelle à μ^p, ce qui nous permet de définir une densité pour Y conditionnellement à μ. En intégrant cette relation avec des fonctions logarithmiques et cumulatives, nous obtenons la forme ci-dessus. Il est important de noter que la forme de la log-vraisemblance varie en fonction du paramètre de puissance p, mais elle conserve la structure générale qui permet de l'écrire comme une distribution de la famille exponentielle.

Mise sous la forme de la famille exponentielle

Pour montrer que la distribution de Tweedie appartient à la famille exponentielle, nous devons exprimer la densité sous la forme,

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right).$$

Par identification, il y a :

- Le paramètre canonique θ est lié à la moyenne μ et dépend du paramètre de puissance p. On a $\theta = \mu^{1-p}$.
- La fonction $b(\theta)$ est liée à la variance. Elle s'exprime comme $b(\theta) = \frac{\mu^{2-p}}{2-p}$.
- La fonction $c(y, \phi)$ dépend uniquement de y et de ϕ. Elle ne modifie pas la structure exponentielle de la fonction de log-vraisemblance.

Ainsi, la distribution de Tweedie peut être exprimée dans la forme de la famille exponentielle :

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\mu^{1-p} - \frac{\mu^{2-p}}{2-p}}{\phi} + c(y, \phi)\right)$$

Par conséquent, la distribution de Tweedie appartient à la famille exponentielle.

Coefficients du GLM

Les coefficients du modèle final pour le GLM sont répertoriés dans le tableau suivant.

TABLE 4 : Résultats du modèle GLM

Variable	***Estimate***	***Std. Error***	***t-value***	$Pr(>\|t\|)$
Intercept	-1.878	525.780×10^{-3}	-3.572	354.000×10^{-6}
m_dep2	-72.190×10^{-3}	95.520×10^{-3}	-756.000×10^{-3}	450.000×10^{-3}
m_dep3	87.780×10^{-3}	90.770×10^{-3}	967.000×10^{-3}	334.000×10^{-3}
m_dep4	-411.340×10^{-3}	103.740×10^{-3}	-3.965	73.500×10^{-6}
m_dep5	-339.890×10^{-3}	101.930×10^{-3}	-3.334	855.000×10^{-6}
m_dep6	-335.170×10^{-3}	103.280×10^{-3}	-3.245	1.170×10^{-3}
m_dep7	-344.140×10^{-3}	107.890×10^{-3}	-3.190	1.420×10^{-3}
m_dep8	-349.630×10^{-3}	112.420×10^{-3}	-3.110	1.870×10^{-3}
m_dep9	-40.450×10^{-3}	99.890×10^{-3}	-405.000×10^{-3}	686.000×10^{-3}
m_dep10	-93.210×10^{-3}	101.460×10^{-3}	-919.000×10^{-3}	358.000×10^{-3}
m_dep11	142.130×10^{-3}	94.050×10^{-3}	1.511	131.000×10^{-3}
m_dep12	642.180×10^{-3}	83.830×10^{-3}	7.661	18.500×10^{-18}
m_sub2	167.470×10^{-3}	78.410×10^{-3}	2.136	32.700×10^{-3}
m_sub3	88.320×10^{-3}	84.620×10^{-3}	1.044	297.000×10^{-3}
m_sub4	109.230×10^{-3}	95.600×10^{-3}	1.143	253.000×10^{-3}
m_sub5	92.980×10^{-3}	99.190×10^{-3}	937.000×10^{-3}	349.000×10^{-3}
m_sub6	47.830×10^{-3}	98.820×10^{-3}	484.000×10^{-3}	628.000×10^{-3}
m_sub7	-50.890×10^{-3}	102.390×10^{-3}	-497.000×10^{-3}	619.000×10^{-3}
m_sub8	17.250×10^{-3}	93.860×10^{-3}	184.000×10^{-3}	854.000×10^{-3}
m_sub9	69.150×10^{-3}	85.330×10^{-3}	810.000×10^{-3}	418.000×10^{-3}
m_sub10	-35.600×10^{-3}	84.280×10^{-3}	-422.000×10^{-3}	673.000×10^{-3}
m_sub11	109.250×10^{-3}	80.090×10^{-3}	1.364	173.000×10^{-3}
m_sub12	20.910×10^{-3}	85.820×10^{-3}	244.000×10^{-3}	808.000×10^{-3}
grp_tc(254,383]	717.180×10^{-3}	83.710×10^{-3}	8.568	200.000×10^{-18}
grp_tc(383,489]	651.160×10^{-3}	86.250×10^{-3}	7.550	43.700×10^{-15}
grp_tc(489,591]	693.490×10^{-3}	86.730×10^{-3}	7.996	1.290×10^{-15}
grp_tc(591,717]	554.710×10^{-3}	90.160×10^{-3}	6.152	765.000×10^{-12}
grp_tc(717,869]	543.420×10^{-3}	90.920×10^{-3}	5.977	2.280×10^{-9}
grp_tc(869,1.07e+3]	479.900×10^{-3}	94.250×10^{-3}	5.092	355.000×10^{-9}
grp_tc(1.07e+3,1.35e+3]	397.870×10^{-3}	97.830×10^{-3}	4.067	47.600×10^{-6}
grp_tc(1.35e+3,1.84e+3]	290.030×10^{-3}	103.970×10^{-3}	2.790	5.280×10^{-3}
grp_tc(1.84e+3,3.67e+4]	212.170×10^{-3}	110.130×10^{-3}	1.927	54.000×10^{-3}
grp_td[2,3)	-3.985	652.950×10^{-3}	-6.103	1.040×10^{-9}
grp_td[3,4)	-3.854	438.140×10^{-3}	-8.796	200.000×10^{-18}
grp_td[4,5)	-4.085	437.650×10^{-3}	-9.334	200.000×10^{-18}
grp_bw[7,14)	370.950×10^{-3}	129.230×10^{-3}	2.871	4.100×10^{-3}
grp_bw[14,21)	631.330×10^{-3}	124.900×10^{-3}	5.055	431.000×10^{-9}
grp_bw[21,30)	705.860×10^{-3}	119.510×10^{-3}	5.906	3.500×10^{-9}
grp_bw[30,60)	591.290×10^{-3}	109.400×10^{-3}	5.405	64.900×10^{-9}
grp_bw[60,90)	576.610×10^{-3}	111.930×10^{-3}	5.151	259.000×10^{-9}
grp_bw[90,180)	426.100×10^{-3}	109.560×10^{-3}	3.889	101.000×10^{-6}
grp_bw[180,360)	276.870×10^{-3}	115.770×10^{-3}	2.392	16.800×10^{-3}
grp_bw[360,720)	465.490×10^{-3}	200.860×10^{-3}	2.317	20.500×10^{-3}
grp_bw[720,6e+3)	-9.311	88.283	-105.000×10^{-3}	916.000×10^{-3}
NL1	237.650×10^{-3}	294.900×10^{-3}	806.000×10^{-3}	420.000×10^{-3}
BE1	560.810×10^{-3}	295.170×10^{-3}	1.900	57.400×10^{-3}
DE1	876.030×10^{-3}	292.580×10^{-3}	2.994	2.750×10^{-3}
FR1	439.340×10^{-3}	294.320×10^{-3}	1.493	136.000×10^{-3}

Fonctionnement du GBM

Ce schéma permet de visualiser comment fonctionne un algorithme de *gradient boosting*. Dans cette figure, les rectangles bleus représentent les entrées et les sorties du modèle, le rectangle rouge correspond au calcul des résidus tandis que les rectangles blancs correspondent aux étapes.

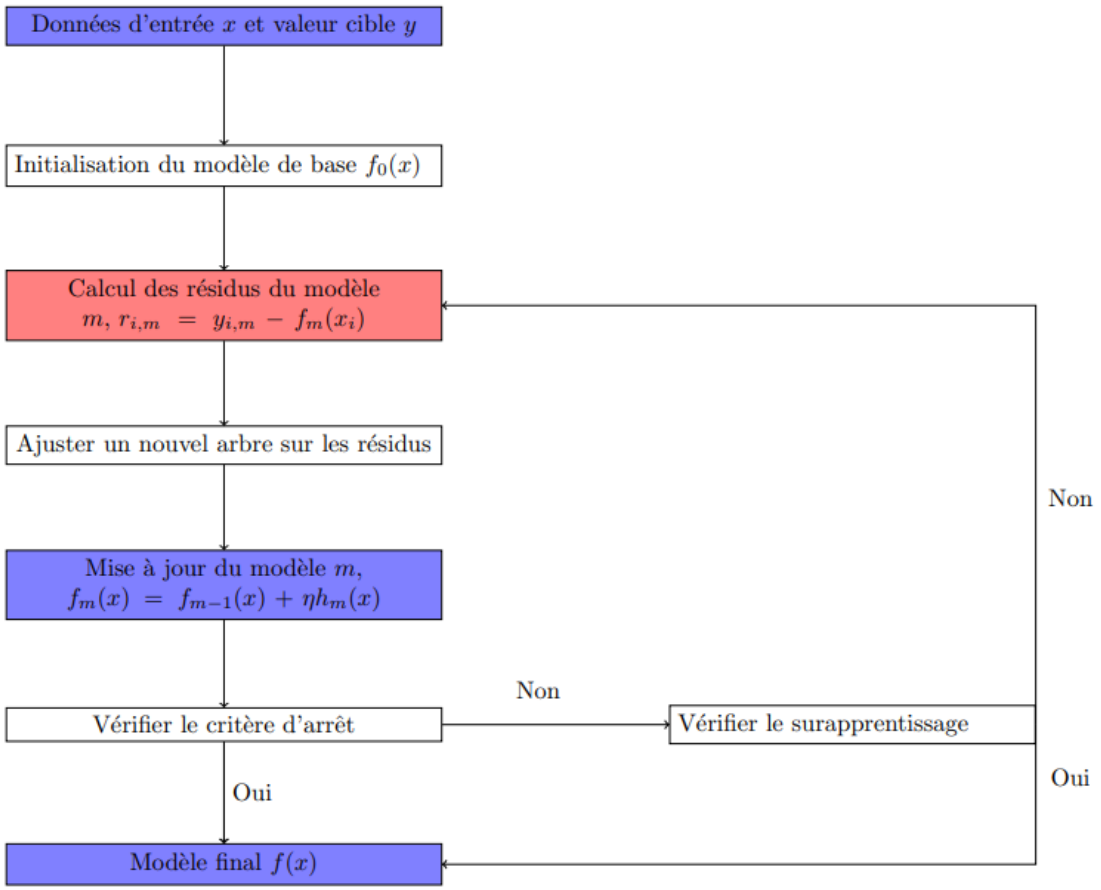

FIGURE 15 : Explication du fonctionnement du GBM

Utilisation de l'intelligence artificielle (IA)

L'objectif de cette annexe est de comprendre comment l'IA générative a été intégrée dans le cadre de la conception de ce mémoire. L'utilisation de cet outil ne se substitue pas à la réflexion et à la production personnelle nécessaire pour répondre à la problématique de cette étude et proposer des solutions adéquates.

Dans le cadre de la rédaction de ce mémoire, l'outil *ChatGPT* a été utilisé pour diverses tâches, notamment pour générer rapidement ou optimiser la structure du code, donner des idées de reformulations plus claires de certains passages du corps du texte ou encore faciliter la mise en forme. Cet outil, lorsqu'il est correctement utilisé, permet également d'explorer de nouvelles pistes de réflexion,

non prévues initialement.

L'emploi de cette IA générative s'inscrit dans une logique d'utilisation optimale des ressources mises à disposition. En tant que future actuaire, il est important de savoir maîtriser les nouvelles innovations techniques, compétence mise en avant par l'Institut des Actuaires (2024). L'apprentissage de l'usage de ce type d'outil permet de gagner en efficacité et d'enrichir la réflexion. Une bonne maîtrise de l'outil repose sur une compréhension solide du sujet et des notions mathématiques utilisées dans les modélisations mises en place.

Dans le cadre de l'exercice d'un métier de modélisation et de traitement de données, la capacité à utiliser de nouvelles méthodes et à intégrer les avancées technologiques dans sa pratique est indispensable. Par conséquent, l'utilisation de *ChatGPT* ne se limite pas à un simple recours à un outil de génération de texte ou de code. Il met en lumière la volonté d'adopter des méthodes de travail modernes. Cette utilisation est complémentaire d'un travail de réflexion et de compréhension du sujet et de ses enjeux.