IAA AAI

INSTITUT DES ACTUAIRES

SECTIONS VIRTUAL COLLOQUIUM | 2020

Application of machine learning methods for cost prediction of natural hazard in France

Antoine Heranval,
Mission Risques Naturels / Sorbonne Université

May 11th – May 15th 2020

About the speaker

Antoine Heranval

- PhD student at Mission Risques Naturels and Sorbonne Université (LPSM)
- Under the supervision of Olivier Lopez and Maud Thomas

Mission Risques Naturels

- Dedicated technical body of the Federation Française de l'Assurance (FFA)
- Association of French insurance undertaking for natural risk knowledge and reduction

Introduction

This work is part of a R&D project called CatClimData, the mains goals are to :

- Contribute to the FFA process of evaluation of natural events in France, soon after its occurrence.
- Improve the understanding of the damage due to natural catastrophe, at the scale of the house

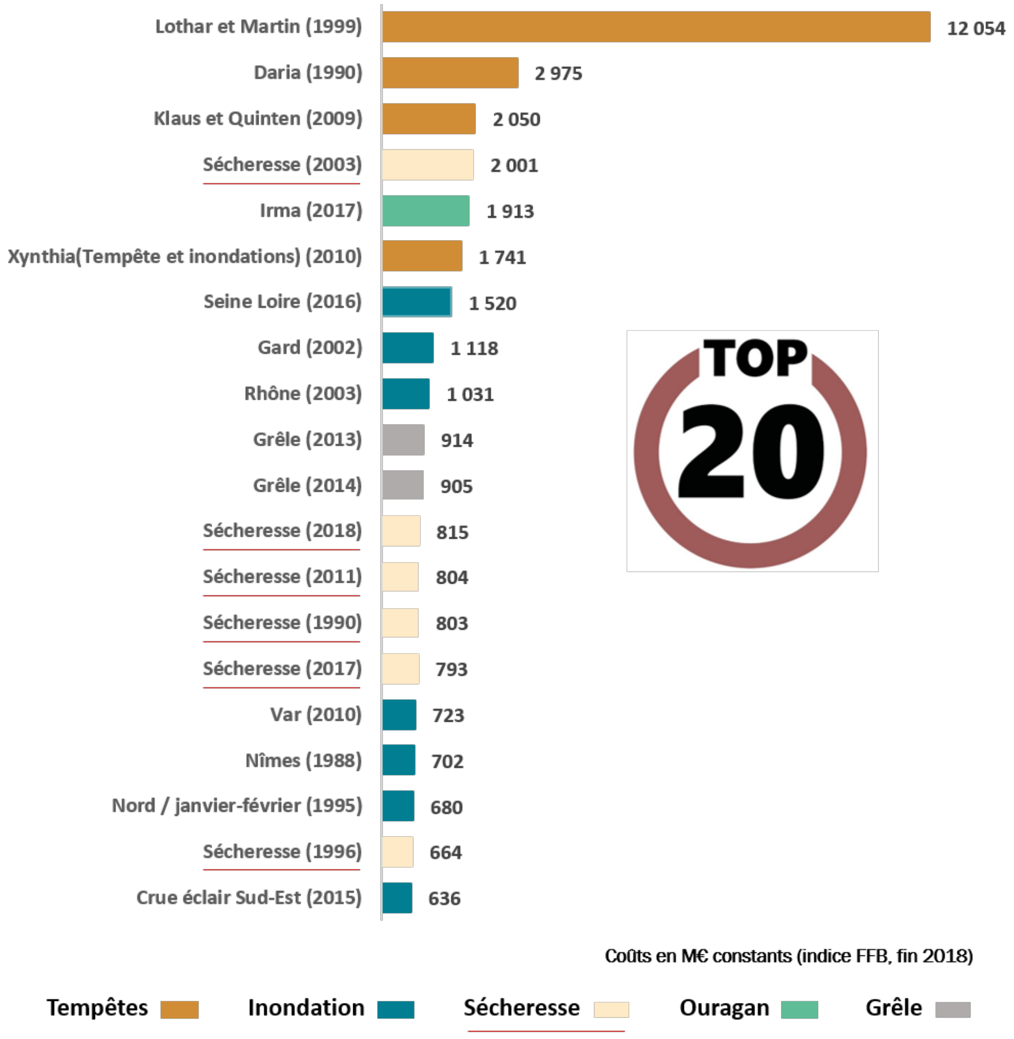
These two objectives are part of the missions of the MRN, contributing to the general interest of the insurance profession.

Introduction

In this work we propose a methodology to estimate the cost of the consequences of drought for the entire French market, annually.

Key figures of the drought in France :

- The aggregate cost of the drought is about 12 Bn€ (since 1982)
- Estimate mean cost of 16 300 € : highest of the non-life insurance
- 2 Bn€ for the exceptional drought of 2003
- 6 events of drought are in the top 20 most costing natural events
- 30 % of the total amounts of claims paid by the French regime CatNat (Catastrophe Naturelle)



Coûts en M€ constants (indice FFB, fin 2018)

MRN,2019

Introduction

Specificity of the regime of compensation CatNat

- One specificity of the French regime of compensation CatNat, is that before receiving the compensation, the city of the policyholder must be acknowledged by a decree as in state of natural catastrophe.

- This decision is based on a criterion that depends on both the exposition to shrinking and swelling of clay and the meteorological intensity of the drought in the city.

- The mean time between the occurrence of the event and the decision of the commission is about 18 months, which is a long time to wait for both the policyholder and the insurer.

The purpose of our method is to be able to anticipate the total cost an event, without waiting for the decree of the inter ministerial committee.

Description of the variables and models

Methodology :

The first step, is to predict the cities that will have a claim during the event of drought. We then, use a linear regression to link the number of houses, exposed to the hazard of drought, in these cities, to the cost of the event. We find a very good correlation with this two variables in our database.

Step 1 : ✗ ✓

Machine learning models to determine if a city has a claim

→

Step 2 :

Calculate the number of houses, exposed to the hazard of drought in these cities

→

Step 3 : €

Link this number of houses to the cost of an event with a linear regression, trained on our database (Multiple R^2= 0.84)

Description of the variables and models

Variables used for the machine learning models :

- Meteorological data produced by Météo-France to measure the intensity of an event **: the Standardized Soil Wetness Index (SSWI)**, which is calculated on the mean of the Soil Wetness Index. We calculated four indexes on the events of drought; The duration of an event, his severity, the magnitude and the rarity. Our events are calculated at the scale of the city for a whole year. All the data used are based on the work of the project ClimSec (see Vidal et al., 2010).

- To characterize the sensibility of shrinking and swelling of clay in the soil we use the three classes defined in the cartography done by the **BRGM.**

- We also used indications on the decree of state of natural catastrophe, information regarding the past acknowledgment. We also implemented a variable that specify the periods where the criteria were the same. In addition, we gave information on the decision, if the criteria was computed with our data.

- The variable that we want to predict is the occurrence of a claim in one city, for that we used the historical data on a database that represent about 70% of the French market : **BD SILECC, MRN.** The database used goes from 2003 to 2017.

Description of the variables and models

Models used :

The situation we try to model is, therefore a classification problem with two classes, 0 or 1. The two classes are unbalanced (5.5% of 1). We have used seven different methods that we will compare :

- Generalized Linear Model with Elastic-Net regularization from the package GLMNET (All the GLMNET model used are based on Friedman, 2010), with $\lambda = 0$
- GLMNET with lasso penalties and $\lambda = \lambda_{min}$
- GLMNET with lasso penalties and $\lambda = \lambda_{1se}$
- GLMNET with Elastic-Net penalties and $\lambda = \lambda_{min}$
- GLMNET with Elastic-Net penalties and $\lambda = \lambda_{1se}$
- Random Forest, classification mode from Breiman, 2001
- Extreme Gradient Boosting, from Chen, 2016

Results

Evaluation :

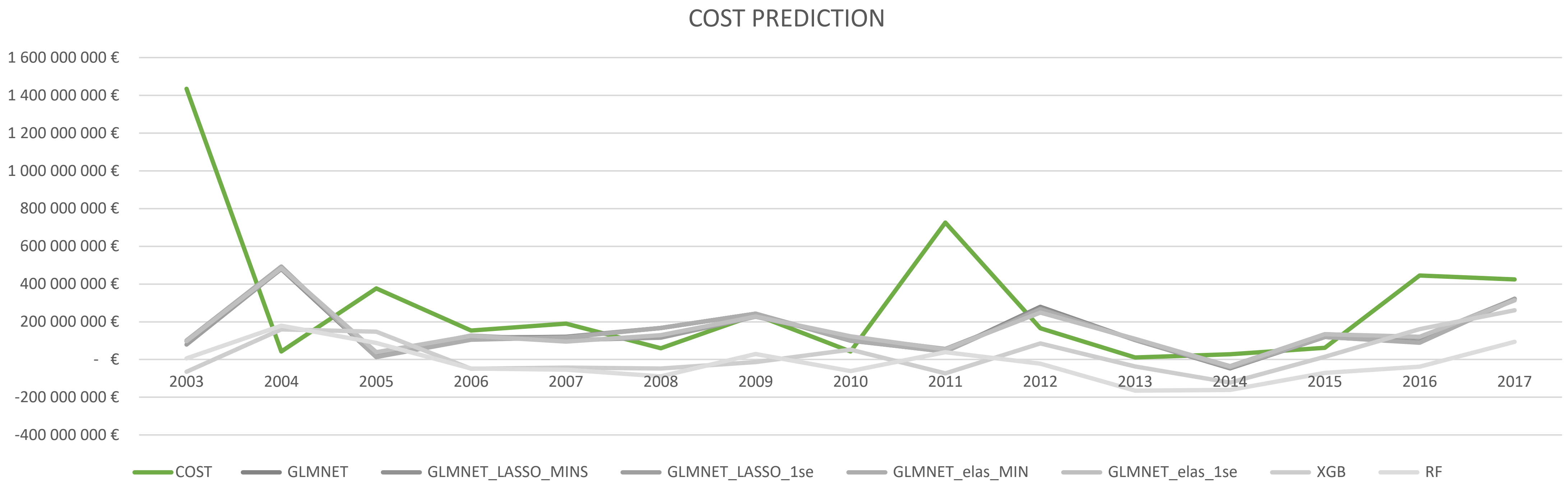
To evaluate our model we separated our data in train and test set. We also evaluated the generalization error by leaving one year out of our model ant then testing it, that way we can see how the model is doing on a whole year, which will be the use case.

To measure the performance we use :

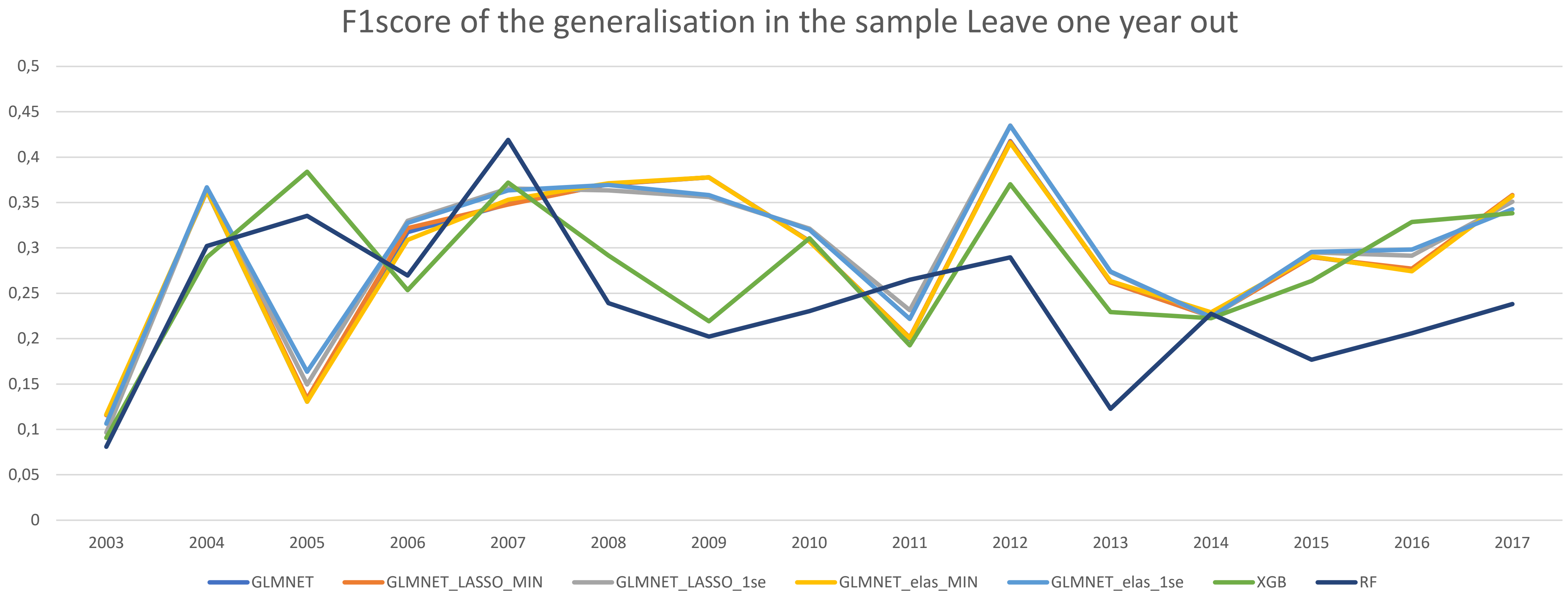
- F1 score, with $F1 = 2 \times \frac{precison \times recall}{precison + recall}$, and $precison = \frac{TP}{TP+FP}$ $recall = \frac{TP}{TP+FN}$
- Plot and AUC of the ROC (Receiver Operating Characteristics)
- Plot and AUC of the precision recall curve
- Cost on the leaving one year out samples

General results

F1.SCORE	GLMNET	GLMNET_LASSO_MIN	GLMNET_LASSO_1se	GLMNET_elas_MIN	GLMNET_elas_1se	XGB	RF
Test	0,38	0,38	0,36	0,38	0,36	0,52	0,50



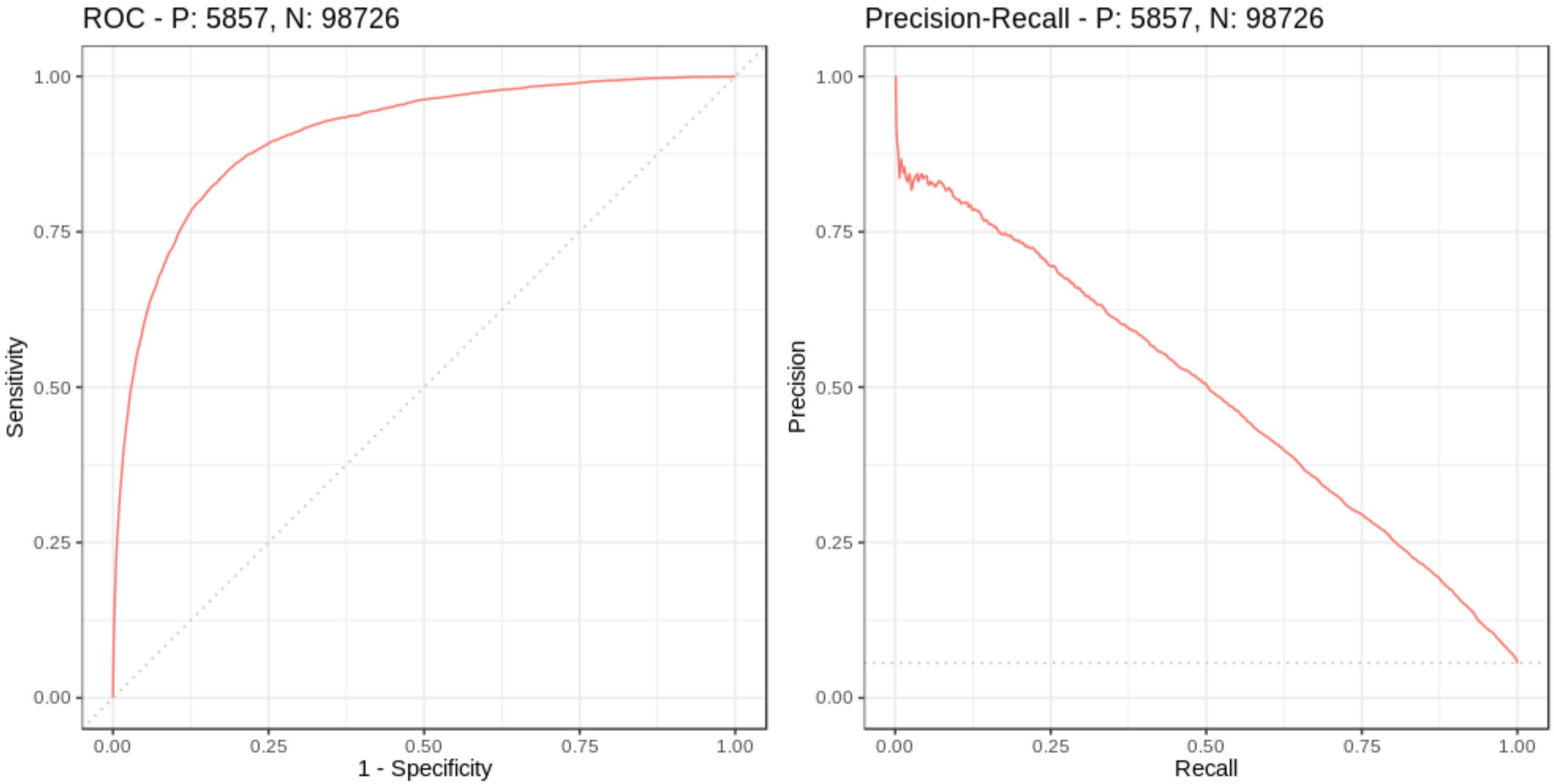
Results



- We can see that the F1score is changing over the year, the year with lowest F1score are also the year with the biggest residual prediction
- The penalization doesn’t seem to have much impact

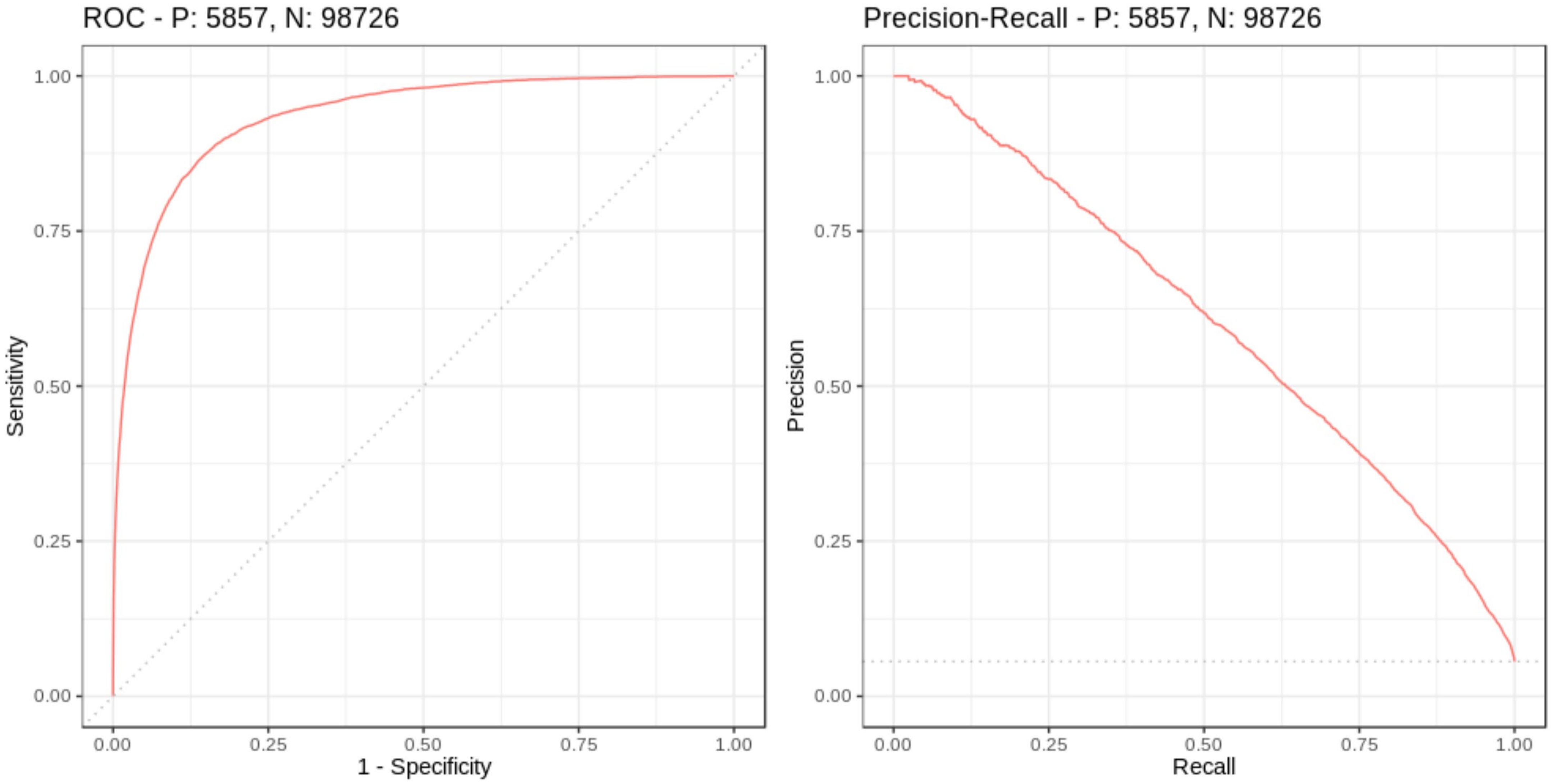
Results

GLMNET :
AUC ROC : 0.90 / AUC PRC : 0.49



Results

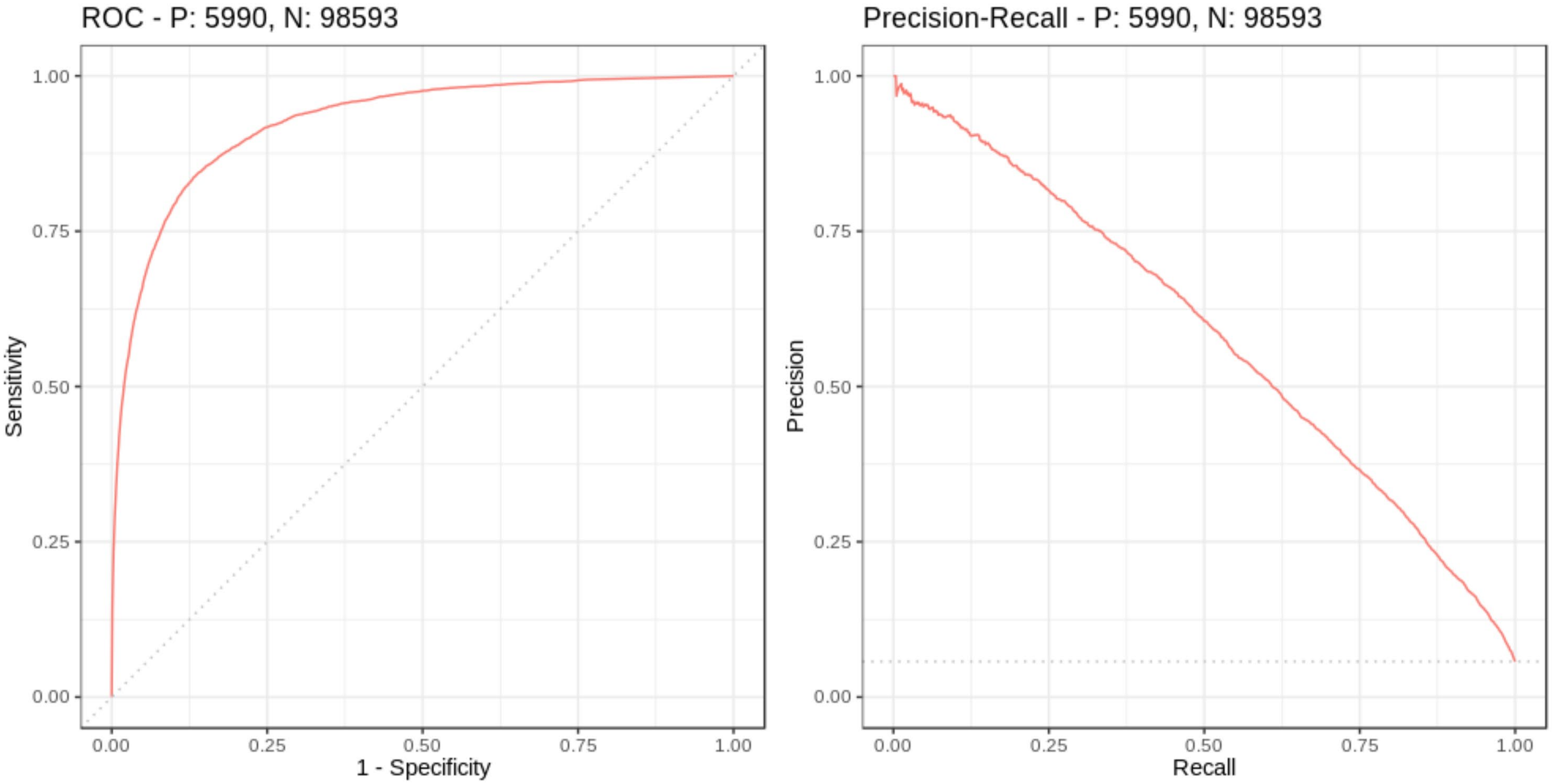
XGBOOST :
AUC ROC : 0.93 / AUC PRC : 0.60



Results

RF :

AUC ROC : 0.93 / AUC PRC : 0.58



Results

We then tried different threshold values to see how it can improve our results

GLMNET			
Threshold	FSCORE	RMSE	MAE
0.1	0,44	4 504 354 000	3 692 635 000
0.2	0,50	4 459 676 000	3 816 729 000
0.3	**0,48**	**3 179 992 000**	**2 369 805 000**
0.4	0,43	3 687 355 000	2 995 208 000
0.5	0,38	4 425 543 000	3 474 320 000
0.6	0,32	3 996 661 000	3 425 562 000
0.7	0,25	3 872 756 000	2 892 608 000
0.8	0,19	4 899 530 000	4 361 896 000
0.9	0,11	5 531 263 000	4 452 161 000

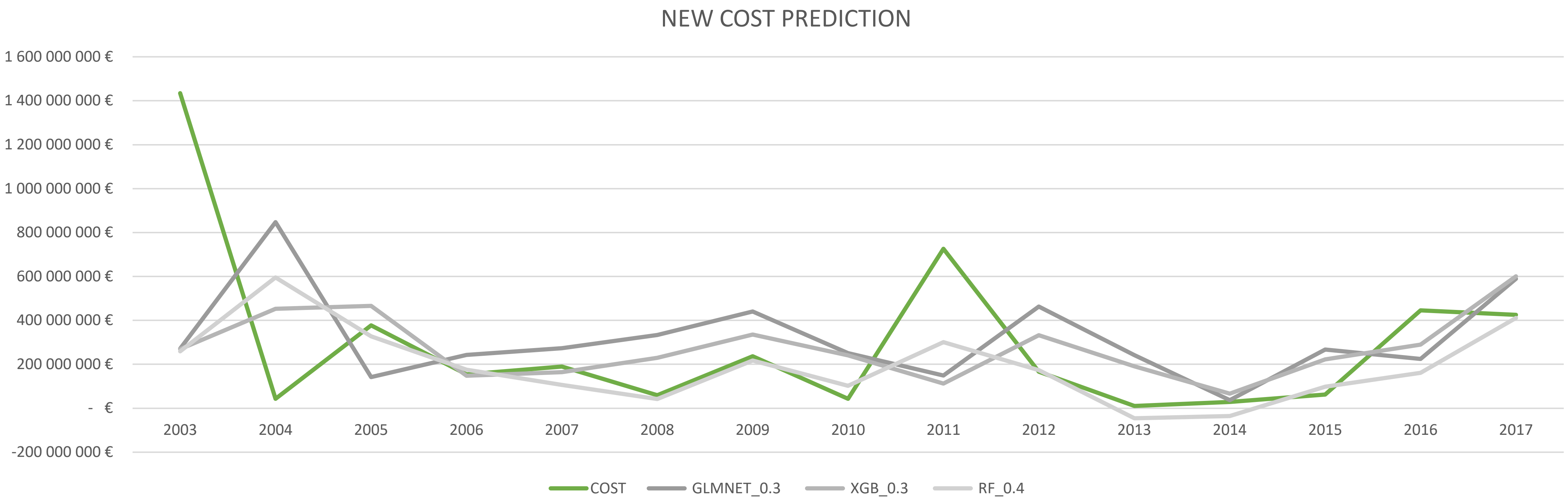
XGBOOST			
Threshold	FSCORE	RMSE	MAE
0.1	0,49	6 763 935 000	5 927 135 000
0.2	0,55	4 295 736 000	3 809 327 000
0.3	**0,56**	**3 214 239 000**	**2 601 030 000**
0.4	0,54	5 784 566 000	4 847 457 000
0.5	0,52	5 049 103 000	4 041 230 000
0.6	0,46	5 333 997 000	4 582 741 000
0.7	0,40	2 438 740 000	2 080 158 000
0.8	0,31	2 929 533 000	2 517 322 000
0.9	0,18	2 302 804 000	2 139 509 000

RF			
Threshold	FSCORE	RMSE	MAE
0.1	0,45	1 547 409 000	1 462 452 000
0.2	0,53	6 128 543 000	5 164 632 000
0.3	0,55	4 380 765 000	4 010 604 000
0.4	**0,54**	**3 675 284 000**	**2 694 463 000**
0.5	0,49	4 529 252 000	3 421 479 000
0.6	0,42	2 748 887 000	2 018 742 000
0.7	0,33	2 557 110 000	2 454 839 000
0.8	0,23	2 673 923 000	2 503 413 000
0.9	0,11	1 276 931 000	925 126 500

- On this basis we determine a new classification threshold
- We choose the value of 0.3 for GLMNET and XGBOOST, and 0.4 for RF

Results

New prediction summary



With the models selected we can do prediction for the year 2018 and 2019, we find predictions of the same order of magnitude as prediction done with another method.

-> We will be able to verify our prediction in one or two years

Conclusion / Discussion

- In this work we developed a method to estimate the cost of the consequences of drought for the entire French market, fitting a GLMNET, a XGBOOST and a RF model with different threshold.

- We obtained encouraging results for such a complex phenomenon. The database used, the process of state of natural catastrophe and the nature of this hazard make the modeling very complex and uncertain.

- We also faced difficulties to evaluate our model.

- In future work we will try to improve the cost prediction based on the cities that have a claim in it.

Thank you for your attention

Contact details :

Antoine Heranval
Mission Risques Naturels
1 rue Jules Lefebvre
75009 Paris
France

Antoine.heranval@mrn.asso.fr

https://www.actuarialcolloquium2020.com/

Disclaimer:

The views or opinions expressed in this presentation are those of the authors and do not necessarily reflect official policies or positions of the Institut des Actuaires (IA), the International Actuarial Association (IAA) and its Sections.

While every effort has been made to ensure the accuracy and completeness of the material, the IA, IAA and authors give no warranty in that regard and reject any responsibility or liability for any loss or damage incurred through the use of, or reliance upon, the information contained therein. Reproduction and translations are permitted with mention of the source.

Permission is granted to make brief excerpts of the presentation for a published review. Permission is also granted to make limited numbers of copies of items in this presentation for personal, internal, classroom or other instructional use, on condition that the foregoing copyright notice is used so as to give reasonable notice of the author, the IA and the IAA's copyrights. This consent for free limited copying without prior consent of the author, IA or the IAA does not extend to making copies for general distribution, for advertising or promotional purposes, for inclusion in new collective works or for resale.