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

Antoine HERANVAL[1,2], Olivier LOPEZ[2], Maud THOMAS[2]

April 21, 2020

Abstract

In this work, we propose a methodology to predict the total cost of a natural catastrophe shortly after its occurrence. Thanks to a large database provided through a partnership with Fédération Française d'Assurance, we manage to have access to a very large volume of claims (our database covers over 70% of the market). Using meteorological data, we measure the intensity of an event. Socioeconomic data provided by INSEE (French public statistical organization) allow to combine this information with a better knowledge of the exposure. In this work we propose the application of different machine learning methods to handle this big volume of data, from sparse Generalized Linear Models (Lasso and Elastic-Net penalties) to Random Forests.

Key words :

Natural Catastrophe, Generalized Linear Models,Lasso and Elastic-Net penalties, Extreme Gradient Boosting, Random Forests

[1]Mission Risques Naturels, 1 rue Jules Lefebvre 75009 Paris, France
[2]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, 75005 Paris, France
E-mails : antoine.heranval@mrn.asso.fr, maud.thomas@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr

1 Introduction

A study by the Fédération Française d'Assurance (FFA) [1] show that the cost of natural catastrophe will increase in the future year, due, mainly, to the general augmentation of wealth in France and climate change. The factor of climate change is very important for drought and its effect on individual houses. In this paper we propose a methodology to estimate the cost of the consequences of drought for the entire French market. We entirely focus on the effect of the shrinking and swelling of clay, that can cause damaged to houses [2]. This hazard is less known than other natural hazards in France, even though it is responsible for about 30 % of the total amounts of claims paid by the French regime CatNat (Catastrophe Naturel) [3]. We also see an intensification of the number of recent drought events, on the six most costful events, three occurs after 2010 [3].

The 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 of decree comes from an inter ministerial committee. 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. This criterion has already changed three times over the past 20 years. This process of decree and acknowledgment is one of the specificity of this system of compensation. A good description of the French regime, can be found in [4]. In our case we are mostly interested in the decision of the committee and the time between the occurrence of an events and its closure. The mean time between the occurrence and the decision is about 18 month [3], 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.

Thanks to a partnership with FFA, essentially with one of his dedicated technical body, the association of French insurance undertaking for natural risk knowledge and reduction (Mission Risques Naturels, MRN), we have access to a large volume of claims. This allows us to build a model based on the past, with an important depth. We use the meteorological data produced by Météo-France (French meteorological institute) in the project CLIMSEC, the SSWI (Standardized Soil Wet-ness Index), which is calculated as described in [5]. To characterize the sensitivity of the soil to the shrinking and swelling of clay, we use the cartography produced by the Bureau de recherches géologiques et minières (BRGM), a French Geological and mining research institute.

In this paper we use different machine learning methods, we started by using Generalized Linear Models with Lasso and Elastic-Net penalties [6], then we also applied Random Forest [7] from [8] and Extreme Gradient Boosting [9].

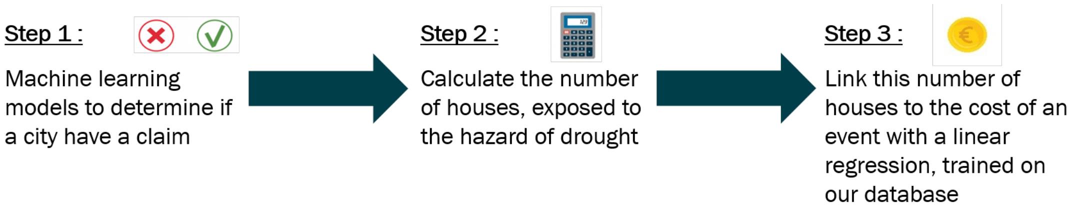
The paper is organized as follows, in section 2, we present the data, models and methods used. In section 3 we gathered the different results and the scoring associated.

2 Description of the variables and models

We first present the overall methodology in 2.1, the variables used in section 2.2, and then focus on the different models in section 2.3.

2.1 Overall methodology

The first step of our method is to predict the cities that will have a claim during the event of drought. For that we will use different models, described in section 2.2. Once we know the cities that are likely to be affected by drought, we calculated the number of houses of those cities, that are more sensible to shrinking and swelling of clay. To do that we use the cartography done by the BRGM. We then use a linear regression to link the number of houses, to the cost of the event. This linear model has been trained on our claim's database. We found a very good correlation on this two variables in our database, with a multiple $R^2 = 0.84$. This validates the use of this method for the cost prediction. Figure 1 summarizes this overall methodology.



Step 1 :

Machine learning models to determine if a city have a claim

Step 2 :

Calculate the number of houses, exposed to the hazard of drought

Step 3 :

Link this number of houses to the cost of an event with a linear regression, trained on our database

Figure 1: Overall methodology

2.2 Variables

As mentioned before to build our machine learning models we used meteorological data produced by Météo-France, the Standardized Soil Wetness Index (SSWI), which is calculated on the mean of the Soil Wetness Index and on different depths. We have at our disposal four indexes, the mean calculated for one, three, six or twelve months. This give us four indexes for each month representing the drought of the soil. The standardization process is described in [5]. This normalization approach provides indices relative to a mean baseline, of the reference period taken for the computation (here 1982-2010). These indexes are given as a monthly observation. We also calculated four indexes on the events of drought, as described in [5]. The duration of an event is the number of months when the index is continuously negative, his severity is the absolute value of the minimum reached during the event. The magnitude is the absolute value of the sum of the indexes values during the event. The rarity is computed on the severity and correspond to the class of return period. Our events are calculated at the scale of the city for a whole year. In case of multiple events (in our case the maximum is four events) we give the information for all events.

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. This give us the surface and ratio of each zone (weak, medium, strong). We then computed the number of individual houses in each class regarding the data of INSEE (French public statistical organization). We use data from 2015. To take the evolution of the number of individual houses into account, we applied an augmentation or reduction of 1 % for each year. [10].

The last variables that we use, are indications on the decree of state of natural catastrophe. We used the information regarding the past acknowledgment. We also implemented a variable that specify the periods when the criteria were the same. In addition, we gave information on the decision, if the criteria was computed with our data. Indeed, the commission use a different meteorological data to say if a city is in state of natural catastrophe, so we cannot say if one city will be acknowledged. Nevertheless, we can say if a city meets the criteria regarding our SSWI. This gives an additional indication on the possibility of being acknowledged.

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, BD SILECC, (MRN), that represent about 70 % of the French market. We only gave an information on the occurrence, not on the number, this become a binary classification problem. We aggregate the claim for a year. The database is therefore a line by city by year, with the information on whether a claim has occurred during the year. This allows us to reduce the unbalanced nature of our data, in a month, city situation we have 1.02% of 1, whereas in the year, city situation the ratio is 5.58%.

2.3 Models

The situation we try to model is as mention above, a classification problem with two class, 0 or 1. We have a response variable $Y \in \mathbf{R}$, with $y_i = \{0, 1\}$ and $X \in \mathbf{R}^p$, our predictor. We have N observations.

First we used Generalized Linear Model with Elastic-Net regularization (GLMNET). In that case our observation are standardized. The problem to solve in the case of binomial models is [6] :

$$min_{\beta_0,\beta} - [\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot (\beta_0 + x_i^T\beta) - log(1 + exp(\beta_0 + x_i^T\beta)] + \lambda[\frac{(1-\alpha)}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1] \tag{1}$$

We have used five different methods of regularization :

- default GLMNET with $\lambda = 0$
- GLMNET with lasso penalties and $\lambda = \lambda_{min}$, $\alpha = 1$
- GLMNET with lasso penalties and $\lambda = \lambda_{1.se}$, $\alpha = 1$
- GLMNET with Elastic-Net penalties and $\lambda = \lambda_{min}$, $\alpha = 0.8$
- GLMNET with Elastic-Net penalties and $\lambda = \lambda_{1.se}$, $\alpha = 0.8$

The GLMNET are used with a cross validation to find the right λ_{min} and $\lambda_{1.se}$, as described in [6].

We also used Random Forest (RF) from the package [8] which respect to the original form developed by [7], and Extreme Gradient Boosting (XGBOOST) from in binary regression mode from [9]. In section 3 we will compare the results of the different model used.

3 Results

3.1 Evaluation

To evaluate our model we separated our data in train set and test set. We also evaluated the generalization error by leaving one year out of our model and then testing it, that way we can see how the model is doing on a whole year, which will be the use case. The first indication that we used is to calculate the F1score, introduced by [12], on the tests sets. It is defined as

$$F1score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

For the generalization on the past year we can compare the predicted cost with the historical cost that we have in our database. Since our data are very unbalanced, in addition to the Receiver Operating Characteristics we also used the Precision and Recall Curves. We used the methodology described in [13], with the associated package.

3.2 Results

In this section we will describe the main results of our test. First we compare F1score of all the models on the test set.

Table 1: Fscore of the Test set, with all the different test on the variables

FSCORE	GLMNET.TEST	GLMNET_LASSO_MIN_TEST	GLMNET_LASSO_1se_TEST	GLMNET_elas_MIN_TEST	GLMNET_elas_1se_TEST	.XGB.TEST	.RF.TEST
Test	0,38	0,38	0,36	0,38	0,36	0,52	0,50

These results show that the XGBOOST and RF seems to work better on our data. In figure 2, we computed the cost predicted, using the method described in section 2.1. We see that the models does not seem to work well on the years 2003 and 2011, but those two years are exceptional in term of cost.

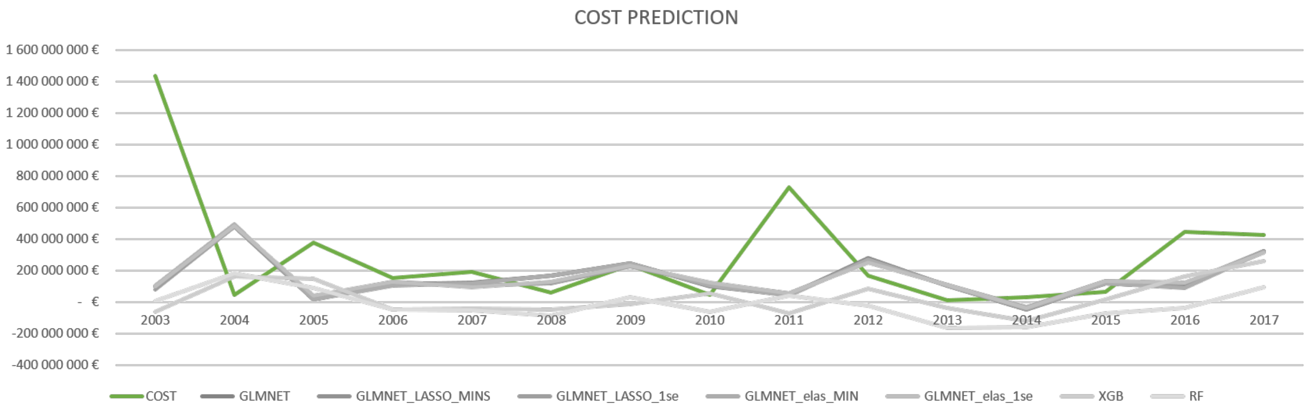


Figure 2: Cost prediction

In figure 3, we can also look at the evolution of the F1score of generalization over the years.

We see that the F1score is highly fluctuating. It goes from 0.09 to 0.45. The year with the lowest F1score are also the year with the biggest residuals prediction. The penalization does not seem to have much impact on both the F1score and the predicted cost. For the rest of the studies we concentrate only on the default GLMNET with $\lambda = 0$.

In figure 4, 5, 6, we plotted the ROC and precision and recall curves for the models. We can also calculate the area under the curve for each model, as shown in table 2.

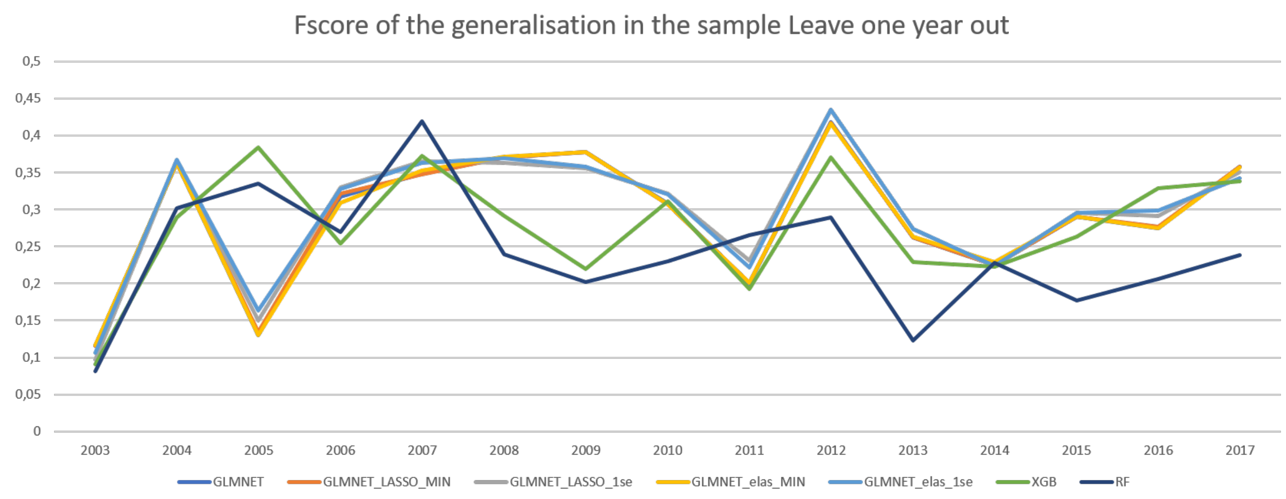


Figure 3: F1score of generalization

Table 2: AUC of the different models

MODEL	AUC ROC	AUC PRC
GLMNET	0.90	0.49
XGBOOST	0.93	0.60
XGBOOST	0.93	0.58

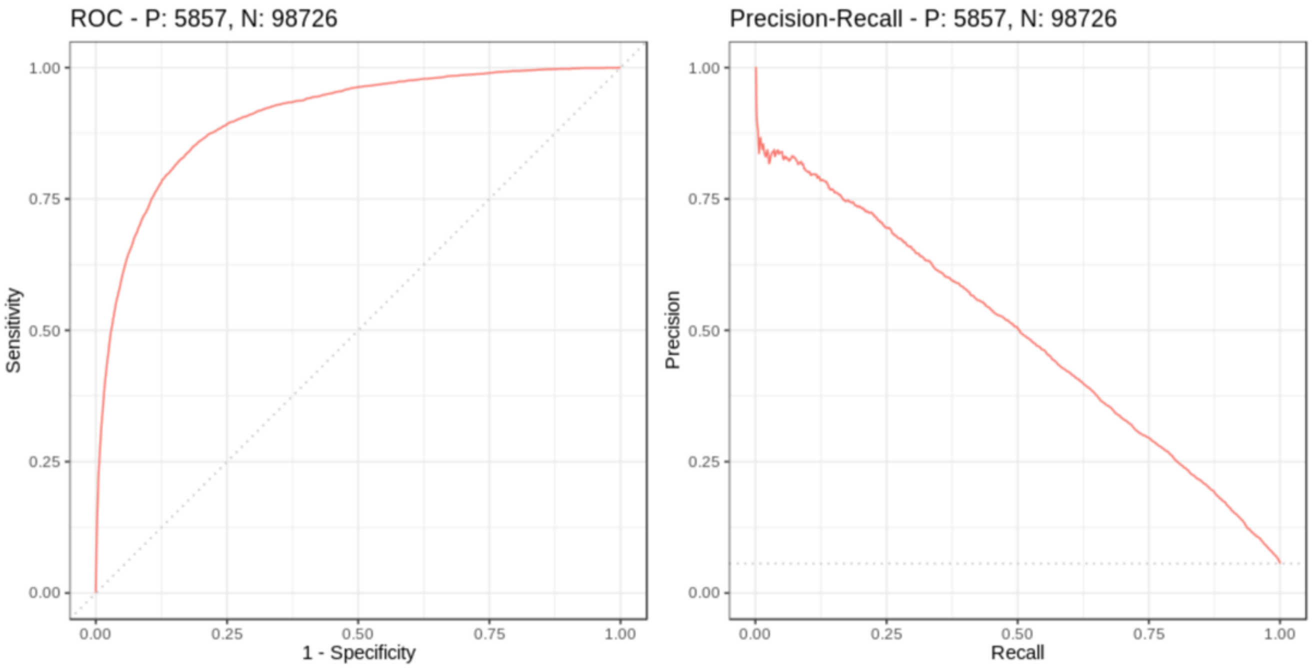


Figure 4: GLMNET

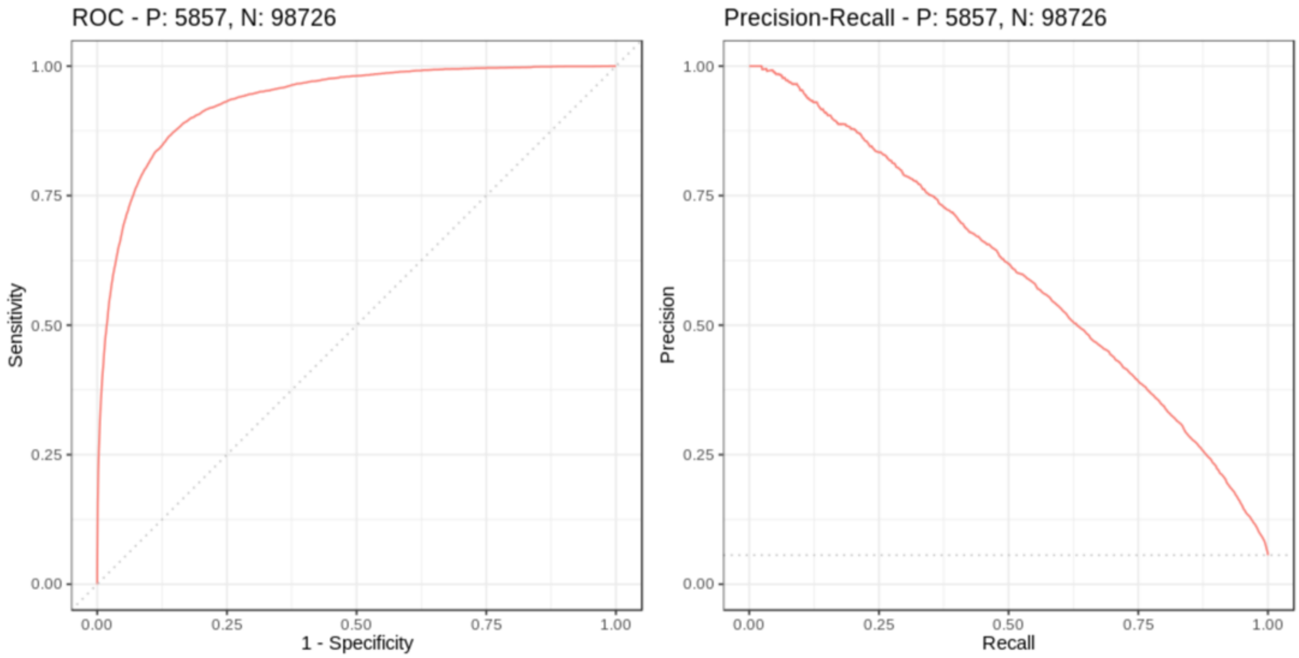


Figure 5: XGBOOST

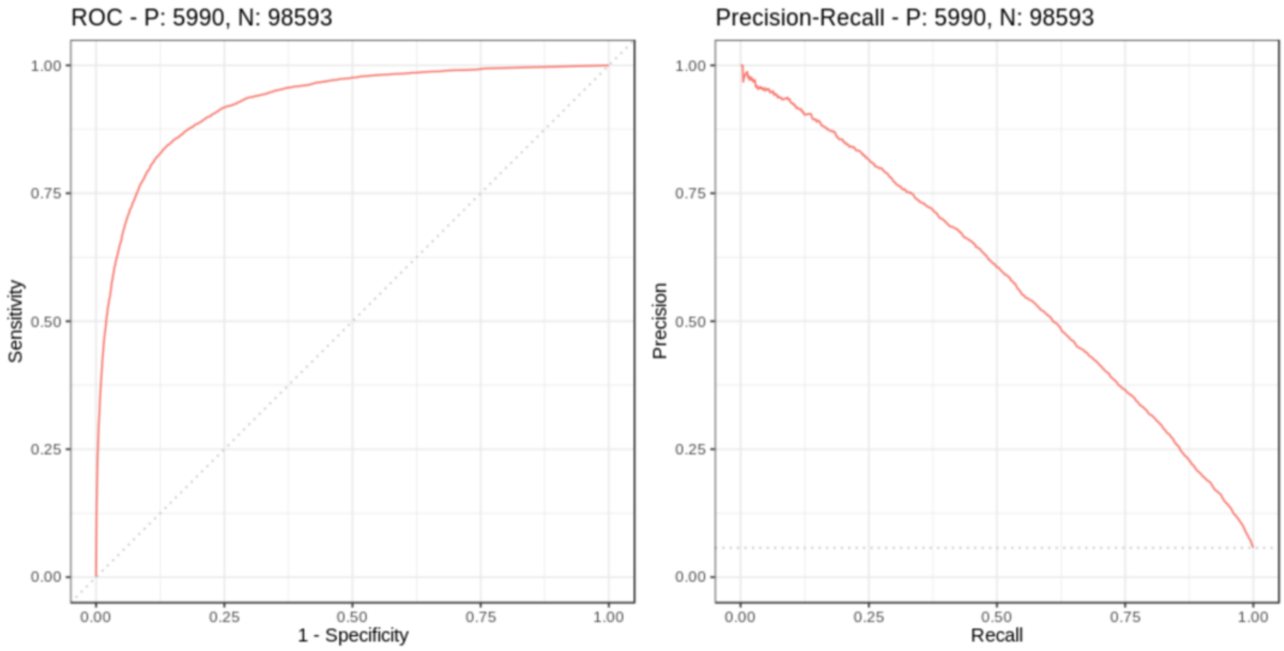


Figure 6: RF

Table 3: Impact of different threshold for 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

Table 4: Impact of different threshold for 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

As described in [13] we see that the use of the precision recall curves is more appropriate to our unbalanced data. Indeed, based on the AUC ROC we have a very good discrimination, always over 0.9 but the AUC PRC nuanced those results with values from 0.49 to 0.60.

Based on those graphics we then tried different threshold values to see how it can improve our results. For that we tested each threshold we a step of 0.1 and look the impact on the F1score for the test and the prediction of the cost on the generalization year. We computed the RMSE and MAE of the predicted cost to see and select the best threshold.

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. In GLMNET and RF we tried to have the best RMSE and MAE rather than the best F1score.

Those new thresholds give us new prediction for the historical cost and as we can see in figure 7. The new prediction are more close to the reality. With the models selected we can do prediction for the year 2018 and 2019, we find predictions of the same order of magnitude as the prediction done with another method.

Table 5: Impact of different threshold for 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

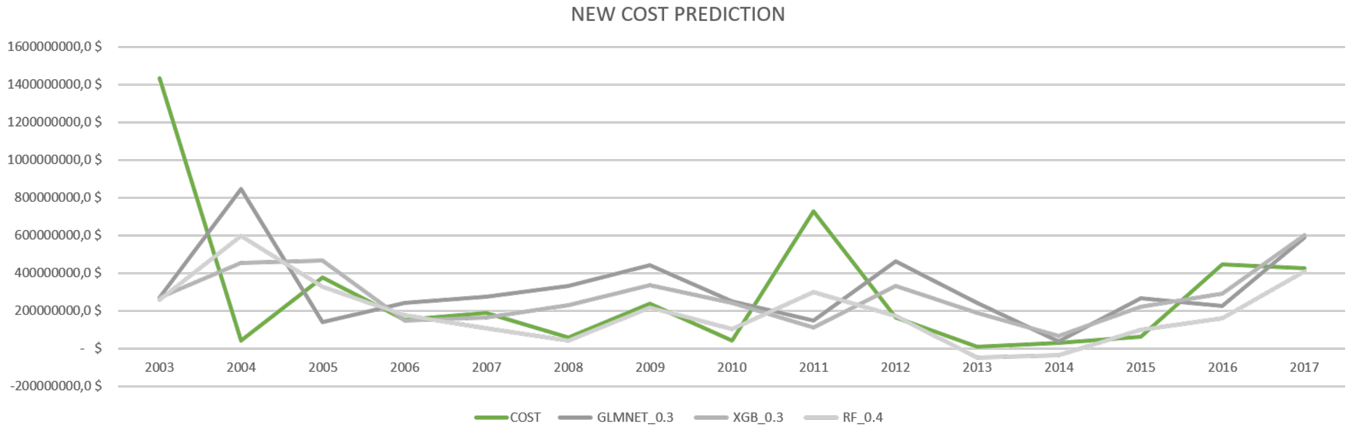


Figure 7: New cost prediction based on the modified threshold

4 Conclusion and discussions

In this work we developed a method to estimate the cost of the consequences of drought for the entire French market, fitting a GLMNET, XGBOOST and RF model with different threshold. This gave us cities that could have a claim in it. Based on those cities we calculated the number of houses sensible to the shrinking and swelling of clay and then computed the total cost using a linear regression.

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. Indeed, our database is based on the past claims, reported by different insurers and can contain error, particularly on the localization. The second difficulty is the process of decree of state of natural catastrophe. To be able to obtain a compensation, and therefore to appear in our database, the city must be acknowledged, there may be claims in cities not acknowledged. Our models can find such claims, but we are not able to tell if is true or not. Also, there have been in the past 20 year, three changes in the criterion to be acknowledged, those changes have induced changes in the meteorological characteristic of the cities affected by drought. Therefore, in our train database we can have different characteristics that will have different effect depending on the criterion. The last effect is the complicated nature of this hazard, as a matter a fact, for an individual house to be damage by drought you need more that just a dry soil. There are many structural factors that can have an effect, in particular the type of foundations of the house. The interaction between the structure of the house and the composition of the ground play an important role to determine if the house will be damage by the drought. We took into account the nature of the soil with the BRGM's indicators but it is very difficult to take into account the structure of the house due to the lack of data on the different type of foundations, especially at a local scale.

We also faced difficulties to evaluate our model. As mention above there is an uncertainty on the results due the acknowledgment but more generally it is difficult to find the right marker to judge a model, especially with such data. We used multiple indexes to evaluate our model but there is always uncertainty. Furthermore, the prediction that we make can only be verified in one or two year, and even more if we want to have all the claim.

In future work we will try to improve the cost prediction based on the cities that have a claim in it. We will try to examine more closely the link between the cities with a claim and the total cost of an event. A possible analysis is to fit a specific model (GLMNET with a Gamma distribution for example) to those city to see the total or mean cost at the scale of the city, or maybe at a wider scale. Another area of improvement is to concentrate on the extreme events like 2003 and 2011, to understand why we failed to do precise prediction. We could then try a specific method, involving extreme value theory for instance, to improve our results.

References

[1] Fédération Française de l'assurance, *Etude : Changement climatique et assurance à l'horizon 2040*, (2015)

[2] Agence Qualité Construction , *Avant de construire – Prendre en compte les risques du terrain*, (2014)

[3] Mission Risques Naturels, *Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti*, (2018)

[4] Nussbaum, R. *Involving public private partnerships as buildingblocks for integrated natural catastrophes country risk management—Sharing on the French national experiences of economic instruments integrated with information and knowledge management tools.* IDRiM Journal5(2): 82–100., (2015)

[5] Vidal, J.-P., Martin, E., Franchistéguy, L., Habets, F., Soubeyroux, J.-M., Blanchard, M. and Baillon, M., *Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcou hydrometeorological suite*, Hydrology and Earth System Sciences, 14(3), 459-478. DOI : 10.5194/hess-14-459-2010, (2010)

[6] Friedman, J., Hastie, T., Tibshirani, R., *Regularization Paths for Generalized Linear Models via Coordinate Descent*, J Stat Softw, 33(1):1-22, ISSN 1548-7660., (2010)

[7] Breiman, L.,*Random forests*, Machine Learning Vol. 45, p. 5-32 (2001)

[8] Wright , M.-N.,Ziegler,A., *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.*, Journal of Statistical Software, 77(1), 1–17. doi: 10.18637/jss.v077.i01, (2017)

[9] Chen ,T.,Guestrin,C., *XGBoost: A Scalable Tree Boosting System*, In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785-794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503- 232-2. doi: 10.1145/2939672.2939785. event-place: San Francisco, California, USA., (2016)

[10] Arnold, C.,*Le parc de logements en France au 1er janvier 2018*, (2018)

[11] Fédération Française de l'assurance,*L'assurance des événements naturels en 2018*, (2019)

[12] Chinchor, N.,*MUC-4 evaluation metrics*,in Proc of the Fourth Message Understanding Conference, pp. 22–29, (1992)

[13] Takaya Saito and Marc Rehmsmeier *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PLoS One. 10(3):e011843, (2015)