Practical guideline to efficiently detect insurance fraud in the era of machine learning

Denisa Banulescu-Radu * Meryem Yankol-Schalck †

February 18, 2023

Abstract

Detecting insurance fraud is a difficult process because the fraud phenomenon itself is complex and the techniques varied, the cases of fraud observed in the data sets are limited, and the human, financial, and time resources used for investigation are modest. The aim of this paper is to provide a clean and well structured study on modeling fraud on home insurance contracts, using real French data from 2013 to 2017. Several methods are developed to identify risk factors and unusual customer behaviors. Traditional econometric models as well as new machine learning algorithms with good predictive performance and high operational efficiency are tested, while maintaining method interpretability. Each methodology is evaluated on the basis of adequate performance measures and the issue of imbalanced databases is also addressed. Finally, specific methods are applied to interpret the results of the machine learning methods. Results show that, for instance, the frequency of reported losses has a very high importance as a predictor of fraud. The approach can be easily adopted as guideline by insurance companies and adapted to efficiently fight different types of fraud.[1]

JEL Codes: G22, G29, C10, C35, C38, C55.

Keywords: **Fraud detection, Household insurance, Machine learning, Imbalanced data, SHAP.**

*University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: denisa.banulescu-radu@univ-orleans.fr

†IPAG Business School, 4 Boulevard Carabacel, 06000 Nice, France. Email: meryem.schalck@ipag.fr

[1]Research conducted within the "Data Science and Insurance Fraud Detection" research initiative under the aegis of the Europlace Institut of Finance, a joint initiative by Laboratoire d'Economie d'Orléans and Thélem Assurances, and the "Fraud detection and Anti-Money Laundering" research project (REDFLAG APR-IA AE 2019-1850) funded by Centre-Val de Loire Region.

1 Introduction

Insurance fraud is an illegal act on the part of either the buyer or the seller of an insurance contract, and it is usually an attempt to exploit an insurance contract for financial gain. The majority of insurance fraud cases are exaggerated claims. It reflects external fraud such as automobile fraud, personal home fraud, healthcare or medical fraud and insurance crop fraud. *Automobile fraud* entails someone deceiving an insurance company about a claim involving their personal or commercial motor vehicle (see Nian et al., 2016), whereas *personal home fraud* covers possessions against fire, theft and other risks, such as accidental damage, and takes place when someone knowingly submits an inflated claim on their homeowners or renters policy for more than the actual value of the loss or damage (see Bentley, 2000; Von Altrock, 1996). As for *healthcare or medical fraud*, it is committed either by the insured individual or the provider of health services, and represents false or misleading information provided to a health insurance company in an attempt to have them pay unauthorized benefits to the policy holder, another party, or the entity providing services (see Kirlidog and Asuk, 2012). Lastly, *insurance crop fraud* is about making claims for crops that the farmers never planted or for crops they allege are destroyed but that are actually sold.[2]

Given its time-evolving character and its diversity, insurance fraud remains a hot topic with challenging issues for the detection and prevention processes. Nowadays, insurance fraud records impressive amounts and it still shows an up-trend due to the expansion of modern technologies. In 2019, the European Federation of insurance companies announced that fraud claims account for 10% of the total number of claims. The latest report of the French Agency for the fight against insurance fraud (ALFA) suggested that fraudulent claims cost France industry about 2.5 billion euros in 2014 in property damages. Insurers recovered only 219 million euros from this negative record. According to the Association of British Insurers (ABI), insurers detected 125,000 dishonest insurance claims valued at 1.3 billion pounds in 2016. Beside the financial cost, the insurance fraud Taskforce final report insisted also on the fact that "*the normalisation of fraudulent behaviour is socially corrosive and erodes trust*". In response to these stylized facts, there is an increasing interest in the way of managing information to help the insurer identify fraudulent claims.

Several studies have already attempted to develop decision support tools that allow investigators of insurance companies to be better prepared to fight fraud. For instance, practical models to sort out claims for insurance fraud investigation emerged in the 1990s with database organization and selection strategies (Major and Riedinger, 1992), fuzzy clustering (Derrig and Ostaszewski, 1995), simple regression scoring models (Weis-

[2]Definitions provided either by *www.nyccriminallawyer.com* or *http://www.helpstopfraud.org/*

berg and Derrig, 1998; Brockett et al., 1998) and probit and logit models (Artıs et al., 1999; Belhadji et al., 2000). Dionne et al. (2009) combines audit theory with scoring methods in an asymmetric information setting. Logistic regression methods were commonly used to describe data and to explain the financial fraud. However, logistic regression models with an important number of explanatory variables have less statistical power. Recently, traditional models have been replaced by data mining methods used for financial fraud to reduce the number of false positive and false negative decisions. Consequently, machine learning methods have emerged to solve fraud detection issues.[3] New methods such as decision tree, random forest, support vector machine have become popular research tools. Recently, Dhieb et al. (2020) used extreme gradient boosting method (XGBoost) to detect auto insurance fraud.

However, there are many shortcomings of the existing research. First, most studies do not take all available information into account. In particular, they do not use real data, which is the best way to identify risk factors and unusual behavior. Second, they do not fully deal with the issue of imbalanced data sets (i.e, the situation when one or more classes have very low proportions in the training data as compared to the other classes) for fraud detection modelling. Recent studies suggest the use of specialized techniques, data preparation tools, learning algorithms, and performance metrics for practical imbalanced classification (see Kuhn et al., 2013; He and Ma, 2013; Fernández et al., 2018; Brownlee, 2020). Third, another big challenge to overcome is the fact that machine learning techniques are veritable black boxes. The fraud investigator needs to understand the logic behind a fraud alert in order to take a decision, reason for which interpretability has become a vital concern in machine learning environment. This issue is nicely summarized by Ribeiro et al. (2016) as follows:

"*if the users do not trust a model or a prediction, they will not use it.*"

Several methods have been proposed to overcome this issue. For instance, Ribeiro et al. (2016) proposes the LIME approach, "*which explains the predictions of any classier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction*", while Lundberg and Lee (2017) introduced the SHAP (SHapley Additive exPlanations) technique. The latter is applied for interpreting the predictions, the method assigning for each of them an importance value to each feature. The interpretability issue is not treated often in the fraud detection setting. Fourth, the fraudulent files are generally detected in practice by the investigator. However, every day a large number of files must be analyzed by the anti-fraud team, which limits

[3]Two main types of machine learning models are mostly used for fraud detection: (i) unsupervised and (ii) supervised learning methods. (i) Unsupervised learning is the case where you only have input data or features and no corresponding output or target variables. These models generally aim to identify transactions or customers which are most dissimilar from a given norm by modelling the underlying structure or distribution in the data and learning more about it. They try hence to infer homogeneous classes based on some observable variables related to the different facets of fraud. (ii) Supervised techniques assumes the existence of an observable variable, which can be considered by its nature as a target of fraud (e.g., a binary variable taking value one if the claim is a fraud and zero otherwise). They are designed to detect ex-post fraud cases. In this case, both past fraudulent and non-fraudulent records are used to construct models which allow one to assign new observations into one of the two classes.

their capacity for action. Typically, practitioners are more interested in predicting the minority class than the majority class, as the minority class may carry a higher misclassification cost (Viaene et al., 2007). The solution is be to select automatically the most doubtful files. The techniques used must provide decision support by a fraud score associated to each incident, in order to prioritize the files that must be investigated first. Last but not least, data mining techniques have been applied most extensively so far to the detection of automobile insurance fraud and health insurance fraud, and there is an obvious lack of research on household insurance fraud. To our knowledge, there is no study on this subject treating together the issues mentioned above.

The aim of this study is to address these above-mentioned gaps by implementing recent models with good predictive performances and high operational efficiency. Consequently, this paper proposes a clean methodology to detect household insurance fraud based on a real database provided by a France insurance company. Based on studies related to other types of fraud, machine learning techniques seem to be relevant for this purpose. The contribution of this study is threefold: (i) first, we propose an empirical analysis of the household insurance fraud based on the use of *real data* instead of simulative data that allows to better identify risk factors and unusual behavior; (ii) second, a clear methodological framework of fraud detection for home insurance is applied based on the most recent machine learning techniques (corrected for the imbalanced issue) in order to provide a real-time fraud detection indicator; (iii) third, we assure the interpretability of the machine learning results by the means of SHAP technique. The ultimate goal of this study is thus to facilitate the task of the anti-fraud team by detecting suspicious claims, by providing relevant insights into indications of fraudulent behavior that necessitates further investigation. The most common fraudulent cases are clearly detected by the anti-fraud team and the adjusters based on their business expertise and knowledge. However, every day a significant number of alerts are also issued from automatic models and have to be analyzed by them, which limits their capacity to operate. If a fraud occurs, it is then extremely difficult to recover the payment. To avoid this, we propose a practical predictive approach to detect fraud at the first loss notification (FNOL), which allows a more vigilant management of the riskiest claims before payment occurs.

On the modeling side, we compare thirteen machine learning algorithms and select those that provide the best results for this configuration: XGBoost, Random Forest, Ridge Classifier, Linear Discriminant Analysis (LDA) and Multi-layer neural networks with 3 hidden layers (ML-3HL). Adequate metrics are calculated to compare their performances. The data sample consists of about 46,000 observations from a French insurance company over the period 2013-2017. Empirical results show that machine learning techniques, especially the XGBoost method in association with imbalanced datasets techniques, lead to a significant reduction of the false positive rate. In line with SHAP technique for model interpretability, results indicate that the loss frequency

at insurance policy level and the theft are the most important risk factor that are pushing up the probability of fraud, while the number of insurance policies and the water damage are associated with lower probabilities of fraud. These results are important as they contribute to the scarce literature on the household fraud detection, and the insurance industry can take advantage of it for the improvement of customer satisfaction, the retention of best customers and the reduction of external adjuster expenses. Moreover, our approach can be easily adopted as guideline by insurance companies and adapted to efficiently fight different types of insurance fraud.

The rest of the article is organized as follows: Section 2 presents the literature review related to fraud detection modelling in the insurance sector. Section 3 describes the methodology. Section 4 presents and discusses the results, and the final section provides some conclusive remarks.

2 Related work and motivation

Fraud detection is a research domain with a wide variety of different applications. Generally, in the academic literature on fraud detection, fraud is classified in three main categories of interest (Ngai et al., 2011; West and Bhattacharya, 2016) as summarized in Figure 1:

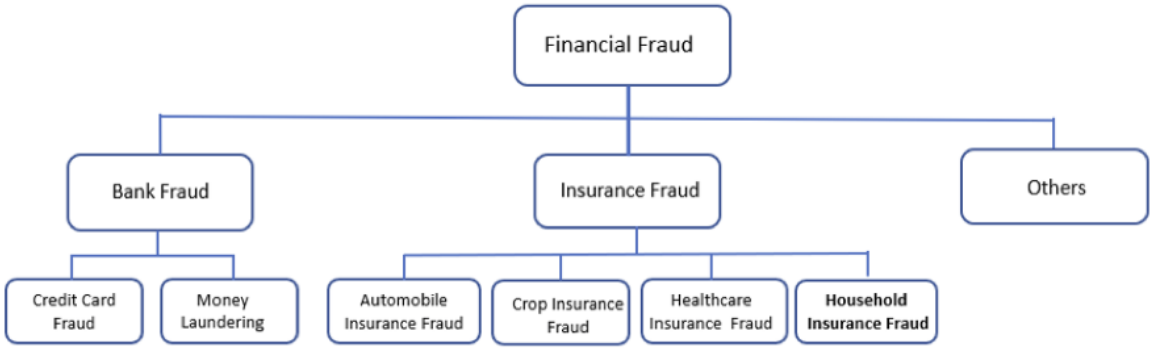

Figure 1: Types of Financial Fraud

Much of literature focuses on bank fraud. The most prominent bank fraud is money laundering. Money laundering is a companion crime, where individuals seek to legitimize assets derived from explicitly illegal activities. This is a crucial issue for financial institutions (Alexandre and Balsa, 2015), countries (Islam et al., 2017), and international communities (Levi, 2015). For instance, it is estimated that money laundering costs the UK 24 billion pounds a year.[4] Other studies investigate the second most prominent type of bank fraud that is the credit card fraud (see Rana and Baria, 2015; Van Vlasselaer et al., 2015; Bhattacharyya et al., 2011; Sánchez et al., 2009; Quah and Sriganesh, 2008, etc.). For instance, Van Vlasselaer et al. (2015) propose APATE, a specific approach to detect fraudulent credit card transactions conducted in online stores. In top of menaces

[4]https://www.nationalcrimeagency.gov.uk/what-we-do/crime-threats/money-laundering-and-illicit-finance

we can identify also the cybercrime, which has emerged lately as a topic of growing interest, especially for the online banking environment. Leukfeldt et al. (2017) examines this path of inquiry by providing a systematic analysis of 40 cases from The Netherlands, Germany, UK, and USA where criminal networks were involved in financial cybercrimes affecting the banking sector. The pandemic period has also been a factor favorable to the development of this type of fraud. Interpol reports that the fragile social and economic environments caused by COVID-19 encouraged cybercriminals to develop and boost their attacks to an alarming level.

Insurance fraud and its prevention is also a major concern for companies because it impacts both the maintaining of their profitability and competitiveness levels, and the use of a fair price. Auto insurance is the most affected by the risk of fraud, including theft, fire and liability insurance which are very expensive for insurers. The number of automobile claims involving some kind of suspicious circumstance is high and has become a subject of major interest both for companies and academic research. This is the reason why there is a growing research literature on automobile insurance fraud, especially to develop decision support tools that allow investigators of insurance companies to be better equipped to fight this type of fraud. For instance, Belhadji and Dionne (1997); Belhadji et al. (2000) use Quebec data and develop a practical tool to aid insurance company adjusters in their decision-making. King and Zeng (2001) mention that popular statistical procedures, such as logistic regression, can sharply underestimate the probability of rare events. This is explained by the fact that fraud databases are highly imbalanced. However, despite its important impacts on model estimation and evaluation, this problem is rarely tackled in the academic papers on fraud detection.

Viaene et al. (2002) use different models based on a data set of personal injury protection claims from 1993 accidents collected by the Automobile Insurers Bureau of Massachusetts. He shows that relatively simple and efficient techniques such as linear logistic regression and linear kernel least squares support vector machine classification have excellent overall predictive capabilities, and (smoothed) naive Bayes also performs well, but decision tree operationalization results are rather disappointing. Artís et al. (2002) demonstrate the performance of binary choice models for fraud detection and implements models for misclassification in the response variable. A database from the Spanish insurance market that contains honest and fraudulent claims is used. They find that 5% of the fraudulent claims go undetected by the logistic regression model. Caudill et al. (2005) extend the study of Artís et al. (2002) by using logit model with missing information. A constrained version of this model is used to re-examine the Spanish insurance claim data. The results show how to identify misclassified claims. Going fruther, data mining techniques (i.e., finding patterns and anomalies in large amounts of data) have already proven useful in risk evaluation (Baesens et al., 2003a,b). However, fraud is an atypical example and requires built-in domain knowledge. Traditionally, insurance fraud detection relies heavily on auditing and

expert inspection. Since manually detecting fraud cases is costly and inefficient and fraud needs to be detected prior to the claim payment, data mining analytics is increasingly recognized as a key in fighting against fraud. Accordingly, Nian et al. (2016) show that the data mining and machine learning techniques have the potential to detect suspicious cases in a timely manner, and therefore potentially significantly reduce economic losses, both to the insurers and policy holders. More recently, Dhieb et al. (2020) use extreme gradient boosting method (XGBoost) to detect auto insurance fraud and show improved results compared to other state-of-the-art algorithms. However, they do not address the issue of interpretability of results and the imbalanced structure of the databases. Gomes et al. (2021) criticise the use of supervised learning techniques and use unsupervised deep learning approaches to identify the main driving factors of fraud.[5] The authors consider two insurance and one credit card transactions databases and proposes a new approach linking the variable importance methodology and two unsupervised models, namely, variational autoencoder (VAE) and autoencoder (AE). Li et al. (2021) introduce a class of nonparametric methods to study the misrepresentation issue in insurance applications. This approach can find its utility in identifying a very specific category of fraud, namely fraudulent beahviors (e.g., intentionally declaring untrue statements to alter insurance eligibility and/or lower the insurance premiums). As for Tumminello et al. (2022), they use social network analysis to detect communities of fraudsters by using microlevel data of subjects and vehicles involved in the same accidents. This category of methods might also be handy for fraud detection and it is also rigorously exposed by Baesens et al. (2015). Table 2 synthesizes some of previous studies having tackled insurance fraud and highlights the methods used.

In practice, after discussions with the fraud departments of different insurance companies, we identified two main groups of techniques used for the detection of fraud: detection based on *business rules* and detection based on *analytical methods*. The business rules are quite simple indicators, and it is quite easy to use by the anti-fraud team. However, they allow the detection of a limited number of fraudulent claims. Analytical technologies represent all the statistical learning techniques. These techniques are more sophisticated and detect better fraud cases. Analytical technologies used recently to detect fraud have proven to be very beneficial. Due to the complexity and sheer number of claims, the investigation team cannot verify all claims within the required timelines. The aim is to classify the cases by fraud potential and thus allow the anti-fraud team to focus on the most suspicious cases. Several business rules must then be coded in collaboration with the anti-fraud team (rules based on past fraud cases) to create a score and facilitate the investigation process.[6] Traditional scoring

[5] We consider that such a polemical discussion is beyond the scope of our paper, as both approaches have their advantages and limitations and the ideal would be to use them together.

[6] It was found that almost half of the fraudulent applications had no business rule outcome. In our case, the investigation team can exploit the information provided by SHAP values and, for each alert, investigate first the elements supposed to increase the probability of fraud and then proceed to the standard check.

Reference	Type of fraud	Methodology	Learning
Belhadji and Dionne (1997); Belhadji et al. (2000)	Automobile Insurance	Logistic regression	Supervised
Artís et al. (2002)	Automobile Insurance	Logistic regression	Supervised
Viaene et al. (2002)	Automobile Insurance	Logistic regression, decision tree, k-nearest neighbor, Bayesian learning multilayer perceptron neural network, least squares support vector machine, naive Bayes and tree augmented naive Bayes classification	Unsupervised & Supervised
Jin et al. (2005)	Crop insurance	Logistic model, probit model	Supervised
Caudill et al. (2005)	Automobile insurance	Multinomial logit model (MNL)	Supervised
Atwood et al. (2006)	Crop insurance	Yield-switching model	Supervised
Yang and Hwang (2006)	Healthcare Insurance	Association rules	Unsupervised
Viaene et al. (2007)	Automobile Insurance	Classification	Unsupervised
Xiaoyun and Danyue (2010)	Healthcare Insurance	Resolution based clustering	Unsupervised
Kirlidog and Asuk (2012)	Healthcare Insurance	Data mining and SVM	Supervised
Nian et al. (2016)	Automobile Insurance	Kernels; Spectral clustering; One-class svm	Unsupervised
Kowshalya and Nandhini (2018)	Automobile insurance	Random Forest, Naive Bayes	Supervised
Dhieb et al. (2020)	Automobile insurance	XGBoost	Supervised
Gomes et al. (2021)	Automobile insurance & Credit card	Deep Learning & Feature importance	Unsupervised
Li et al. (2021)	Behavioral fraud	Kernel quantile regression mixtures	Non-parametric
Tumminello et al. (2022)	Automobile fraud	Social network analysis	Unsupervised

Table 1: Supervised & unsupervised models for insurance fraud detection in chronological order

models, such as logistic regression, can also provide good statistical performance, but they have some practical limitations (e.g., among others, they require careful attention and processing of the database, which can be very time consuming, etc). Besides, there is a great demand for effective predictive methods which maximize the true positive detection rate and minimize the false positive rate. For these reasons, more and more machine learning techniques (less time consuming) are used and implemented, which explains also our choice.

Furthermore, our study can also be beneficial from an economic point of view in the competitive market. An illustrative example is the combination of the fraud score with the lifetime value of the customer in order to improve the retention of good customers. This approach can be thus used to improve marketing campaigns for household products and to influence market share ranking. The construction of a reliable score allows automation of fraud risk alerts and thus reduces the management delay. For instance, one of the factors that strongly influence the loyalty of policyholders after a claim is the time it takes for the claim to be handled and assisted. The best way to increase customer satisfaction is therefore to handle the claim as quickly and professionally as possible. Moreover, especially for small amounts of claims, a well-done score reduces expenses by avoiding

expertise. Overall, the adoption of such an approach generates economic gains through the cost optimization of the insurance company, a better customer retention (a 5% increase in customer retention produces more than a 25% increase in profit [7]), and the optimization of the marketing strategy. Customer retention is a very important economic issue for the insurance industry, as the cost of acquiring new customers in the insurance industry is constantly increasing.

3 Methodology

In this section, we introduce first the methodology to detect fraudulent claims and predict the client probability of fraud. Second, we discuss which are the adequate metrics to evaluate the performance of the models. Third, we present the challenges entailed by the use of an imbalanced data set and some solutions.

3.1 Technical tools

Consider $D = \{x_i, y_i\}_{i=1}^N$ the training set with input vectors $x_i \in \mathbf{R}^p$ and target labels $y_i \in \{0, 1\}$ with $y_i = 1$ when claim i is fraudulent, and $y_i = 0$ otherwise. Thirteen classification algorithms were used for this study to predict fraudulent claims: Logistic Lasso, Random Forest, XGBoost, AdaBoost Classifier, Extra Trees Classifier, Light Gradient Boosting Machine, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, KNeighbors Classifier, SVM Linear Kernel, Ridge classifier, Multi-layer with 3HL.[8] XGBoost, Random Forest, Ridge classifier, Linear Discriminant Analysis and Multi-layer with 3HL have proven empirically to be very effective for this classification problem.

Ridge Logistic regression. Logistic regression (Cox, 1958) is one of the most widely used statistical tools in many areas, with the binary response given by the probability of the response success:

$$\pi_i = \mathbf{P}(y_i = 1) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}, \tag{1}$$

where β is the column vector of the regression coefficients. The parameters estimates are obtained by maximizing

[7] https://media.bain.com/Images/BB_Prescription_cutting_costs.pdf

[8] For the choice of the classifiers, we opted for models that have proven their efficiency in similar contexts. For instance, Gunnarsson et al. (2021) have applied a set of algorithms for credit scoring. Their paper indicates that deep learning algorithms do not seem to be appropriate models for credit scoring based on this comparison and that XGBoost should be preferred to other methods. Multi-layer perceptron neural networks (MLP-NN) have been also popular in a number of fields such as computer vision, speech recognition and classification (Luo et al., 2017). However, these models generally have poorer predictive performance (Yankol-Schalck, 2022; Gunnarsson et al., 2021) and do not seem to be appropriate models for financial fraud problem.

the log-likelihood function:

$$l(\beta) = \sum_{i=1}^{N} [y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)] = \sum_{i=1}^{N} [y_i x_i \beta - log(1 + e^{x_i \beta})]. \quad (2)$$

Ridge Logistic regression Hoerl and Kennard (1970); Le Cessie and Van Houwelingen (1992), is obtained by maximizing the likelihood function with a penalized parameter ($\lambda \geq 0$)[9] applied to all the coefficients except the intercept, resulting in the following constrained maximization equation:

$$l_\lambda(\beta) = \sum_{i=1}^{N} [y_i x_i \beta - log(1 + e^{x_i \beta})] - \lambda \sum_{j=1}^{p} \beta_j^2. \quad (3)$$

When λ is sufficiently large, the ridge coefficient estimates will tend to approach zero.

Linear Discriminant Analysis. Linear Discriminant Analysis (LDA) is a commonly used technique for both data classification and dimensionality reduction. It is a generalization of Fisher's linear discriminant method proposed in 1936 and its objective is to find a linear combination of features that best separates two or more classes of events. Based on assumptions such as multivariate normality of independent variables, homoscedasticity, absence of multicolinearity and independence, it has been proven that LDA may still be reliable and performs quite well even if the these assumptions are not fully respected.

Random Forest. Introduced by Breiman (2001), the objective of the Random Forest method is to improve the decision tree method by combining the concepts of decision tree and bagging for the computation of a forest of decision trees. The main steps of the Random Forest algorithm are explained in Appendix A (Hastie et al., 2009). With a reduced number of parameters to be determined (B, the number of trees to combine and m, the number of predictors to sample at each splitting process when building trees), Random Forest is a good model candidate to fraud detection, with high adaptability to an important number of features and high accuracy.

Multi-layer neural networks. NNs are known as artificial neural networks (ANNs) and have been widely implemented in various fields. NNs rely on training data to learn and improve their accuracy over time. Multi-layer perceptron neural networks (MLP-NN) are trained on the back-propagation (BP) algorithm. This algorithm is performed using a learning procedure based on error correction. The network generates outputs by handling the input data received. Then the error value is calculated by comparing the target values and the network output. Following this, the weights and biases are adjusted to minimize the error and the training process is carried out until the network reaches a minimum error. Given its performance, the Multi-layer with 3 hidden layers model have retained our attention for this study.

[9]The estimator depends on the choice of a tuning parameter $\lambda \geq 0$ to be determined separately.

Extreme Gradient Boosting or XGBoost. XGBoost (Chen and Guestrin, 2016) is an optimized distributed gradient boosting framework with a highly efficient design, flexible and portable. It is an extension of gradient boosting machine (Friedman, 2001) with major improvements and highly used in machine learning and data mining challenges in many fields (for instance, in Kaggle competitions).[10] Designed for computational speed (e.g., it runs ten times faster than the existing popular solutions) and model performance, XGBoost is a scalable parallelized algorithm which optimizes the memory usage (e.g., it exploits out-of-core computation and enables the process of hundred millions of examples on a desktop) and handles *sparse data*.[11]

3.2 Imbalanced datasets

Class imbalance problems appear frequently in fraud detection as one class is more represented than another. It is also our case because only 0.8% of claims are fraudulent. It has been largely documented that the class imbalance issue heavily compromises both the process of learning[12] and the evaluation of its accuracy, which is jeopardized because the scarcity of data leads to poor estimates of the model's accuracy. To deal with this issue and see if model predictive performances are improved, we opt for the data resampling and use the following methods: SMOTE, ADASYN and ROSE.[13]

- *Synthetic Minority Oversampling TEchnique* (SMOTE) (Chawla et al., 2002) is one of the most renowned oversampling methods that generates new synthetic minority class examples by interpolating several minority class instances that lie together. It focuses hence on the "feature space", rather than on the "data space".

- *ADAptive SYNthetic sampling* (ADASYN) (He et al., 2008) is an improved version of SMOTE. This method is based on the idea of adaptively generating minority examples according to their distributions thus making it more realistic. In other words, instead of all the samples being linearly correlated to the parent they have a little more variance in them. It is designed to create synthetic instances in regions of

[10] Boosting is an algorithm that combines multiple weak learners into a stronger learner and corrects the errors made by the existing models. New learners are trained on the errors of the previous ones in order to increase the predictive power.

[11] Moreover, XGBoost is implemented as an open source package, available in popular languages such as Python, R, Julia, C++, Scala. See https://github.com/dmlc/xgboost. For more technical details, see Appendix B

[12] Typically, many machine learning algorithms put more attention and perform better on the majority class and might ignore the rare events, which is exactly the class we care more about and which may carry a higher misclassification cost.

[13] Throughout the last years, many solutions have been proposed to deal with this problem, both for standard learning algorithms and for ensemble techniques. Three major categories stand out (Brownlee, 2020; Fernández et al., 2018): (i) data sampling: the training instances are altered in such a way as to produce a more balanced class distribution that allow classifiers to perform in a similar manner to standard classification; (ii) algorithmic modification: base learning methods are adapted to be more attuned to class imbalance issues; (iii) cost-sensitive learning: higher costs for the misclassification of the minority class instances with respect to the majority instances are considered either at the data level, at the algorithmic level, or at both levels jointly, in order to minimize higher cost errors.

Note that resampling techniques can be grouped into three families: (i) undersampling method: a subset of the original dataset is created by deleting instances (usually majority class instances); (ii) oversampling methods: a superset of the original dataset is created by replicating some instances or creating new instances from existing ones in order to gain importance; (iii) hybrids methods : combine both sampling approaches.This approach seems to be the dominant one in the community as it tackles imbalanced learning in a very straightforward manner.

the feature space where the density of minority examples is low, and fewer or none where the density is high.

- *Random OverSampling Examples* (ROSE) (Menardi and Torelli, 2014) is based on the generation of new artificial instances according to a smoothed bootstrap approach of re-sampling. Its theoretical basis are supported by the well-known properties of the kernel methods. ROSE is presented as a systematic and unified framework for dealing with imbalanced learning that jointly takes into account the effects of class imbalance in model training and model assessing.

Providing a complete review of the inherent literature on the class imbalance issues is beyond the scope of this paper (for excellent overviews, see, e.g. Brownlee (2020); Fernández et al. (2018); He and Ma (2013); He and Garcia (2009); Sun et al. (2009)).

3.3 Methods and metrics of comparison

Despite their intensive use in the classification environment, some metrics are not the most suitable for fraud detection technique because the dataset is highly imbalanced and most of standard metrics are insensitive to skewed distributions and their use in imbalanced domain can lead to suboptimal classification models and produce misleading conclusions (Branco et al., 2015). For instance, all fraudulent claims can be misclassified still with very high accuracy (i.e., *accuracy paradox*). Accuracy is no longer a proper measure in the imbalance scenario, since it does not distinguish between the number of correctly classified examples of different classes and the fact that the classification errors might imply different costs (i.e., misclassifying instances of the minority class is generally much costlier than misclassifying instances of the majority class). Thus, analysing the performance of learning algorithms in such a context becomes a difficult task. To alleviate this problem, more informative measures are required in order to assess the quality of the models.

To evaluate the performance of each classifier, we recommand the use of measures adapted for imbalanced dataset classification and for the specific context of fraud detection, such as: false discovery rate, false negative rate, F-measure, areas under the ROC and precision-recall curves (AUC-ROC and AUC-PR), cross-entropy and Brier Score (BS).[14] Following Brownlee (2020) and He and Ma (2013), they can be grouped into three categories: (a) threshold metrics (i.e., based on a threshold and a qualitative understanding of error, they quantify the classification prediction errors); (b) ranking metrics (i.e., used to evaluate classifiers based on how effective they are at separating classes); (c) probabilistic metrics (i.e., designed to quantify the uncertainty in

[14]For a more complete list, see Brownlee (2020); Baesens et al. (2015).

classifier's predictions). Table C.1 in Appendix C offers complete definitions and details on their computation. By using all these statistics, the analysis becomes hence more appropriate for the fraud detection framework, more robust and gives us a more complete view on the relative performances of the competing models.[15]

4 Empirical Application

4.1 Dataset and conduct of the analysis

The data correspond to a sample of house claims that occurred from 2013 to 2017 in a French insurance company. All historical fraud cases are well known and classified as proven fraud.[16] The output variable "Fraud" takes either value 1 if there is a detected fraud by experts, managers or others, or value 0 if the incident is not fraudulent. The input variables come both from internal and external databases. The variables that come from different internal databases have been processed and grouped together for the construction of the study base. The data set contains information on the claim (place, date, coverage, and so on), the insured (history of previous claims, etc.), and the house (residential or business use, construction materials, etc.). The information contained in the samples was obtained either from the claim statement or the police. External data originate from the French National Institute of Statistics (INSEE) and are related to sociodemographic and geographic data (including different levels of geographical granularity). Our database includes 46 302 observations of which only 348 fraudulent claims. We deal hence with a strong imbalanced data set, as fraudulent claims represent only 0.76% of observations. Figure 2 shows the data processes.

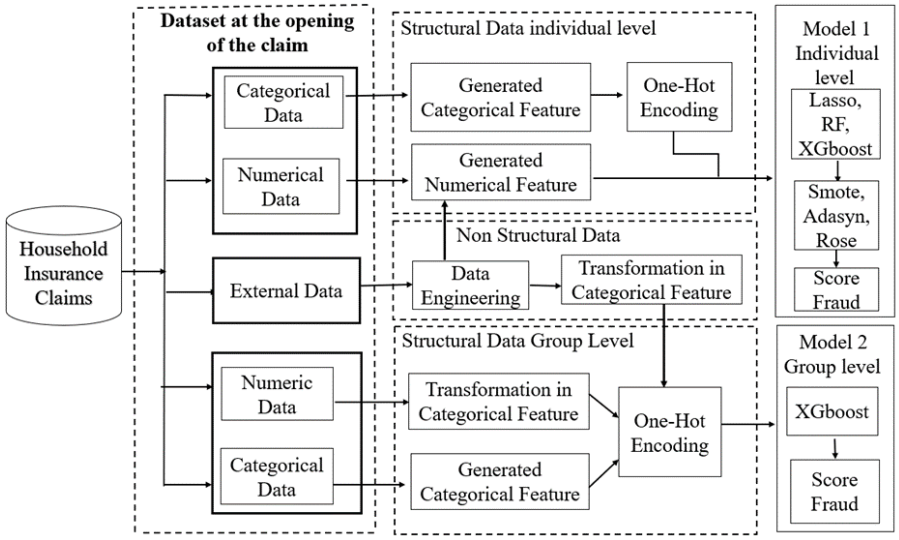

Figure 2: Analytical framework for household insurance fraud score

[15]Other performance measures based on unsymmetrical costs/rewards could be incorporated by means of a cost matrix (see Fernández et al., 2018).

[16]An amount greater than zero indicates that the fraud is certain and there is sufficient evidence.

The main steps of the empirical analysis are detailed in following:

1. *Data pre-processing.* Before the estimation process, data underwent some pre-processing steps such as: the anonymization of the database, the replacement of unknown values by the median for continuous variables.[17] For the missing values in the categorical variables we create a new category "NA" (Not Available), as the insurance company considers this information insightful as well.

2. *Feature engineering.* New features have been created based on the business user knowledge (e.g., age classes, the delay between the last amendment date of the insurance policy and the opening date of the claim, the delay between the occurrence date and the opening date of the claim, the opening of the claim shortly after the subscription of the insurance policy, the delay between the date of transformation in insurance policy and the effective date of the insurance policy, etc.). The objective was to introduce new important information while keeping their calculation as quick as possible.

3. *Feature/variable selection.* Feature selection strategies are implemented at two levels of the analysis: before the model is trained (e.g., the correlation filtering) and during the training of the model. The correlation filtering is intended to remove redundant features with statistically significant correlations. To this aim, depending of the type of the variables (continuous or categorical), the magnitude of the linear relationships between each couple of variables was calculated using specific statistics and their statistical significance was subsequently tested via appropriate statistical tests. For instance, Pearson coefficient of correlation and the associated statistical test were used to check the existence of statistically significant linear relationships between continuous variables, Cramer's V statistic and Chi-Square test were used for nominal categorical variables, and Spearman statistic and the Mantel-Haenszel Chi-Square statistical test in the case of ordinal categorical variables. The correlations results and the list of explanatory variables of interest after the correlation filters and the exclusion of redundant features are presented in Appendix D and Appendix E.[18] Globally, results show no strong association between variables. However, p-values are influenced by the sample size, as large datasets lead often to statistical significance (Cohen, 1988). Indeed, the error measures associated with the small sample make the correlation more "reliable" (see Nakagawa, 2004; Bauman et al., 2013).

   The model-based feature selection is operated during the model training and it is specific to each model

[17] There are several methods of handling missing values, such as by deleting rows, replacing with the mean or median, and predicting the missing values by complete features. We do not discuss each method in detail because it is outside the scope of the study.

[18] Figure D.1 shows the correlation matrix diagrams between numerical features. Figure D.3 indicate the correlation matrix for ordinal categorical variables and Figure D.2 presents the Cramer's V statistic for nominal categorical variables.

configuration used (e.g., Ridge technique forces some of the coefficients estimates to be exactly zero, Random Forests and XGBoost proceed to an automatic selection based on the total decrease in entropy or loss at all splits for a given feature as calculated by the gain measure (see Breiman et al., 1984)).[19]

4. *Model training and evaluation.* The initial step consists on the split of data into training (80% of the data) and test sets (20% of the data) by using stratified sampling techniques. *First*, all the competing models are first estimated on the features selected by the correlation filtering without any other transformation, and the models are evaluated based on the different performance measures previously presented. Logistic Lasso, Random Forest, XGBoost, AdaBoost Classifier, Extra Trees Classifier, Light Gradient Boosting Machine, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, KNeighbors Classifier, SVM Linear Kernel, Ridge Classifier, Multi-layer with 3 hidden layers (3HL) are applied. Each model is hence adjusted or learned only on the training base and assessed then on the test base. Thus, all performance measures of the model are calculated on the test sample, which is never used for estimation. The k-fold cross validation is applied for each method. As the total number of observations is not so consistent, five is chosen as k-value. The results of a five-fold cross-validation are summarized in Figure 2 with the mean and a measure of standard error for each model.

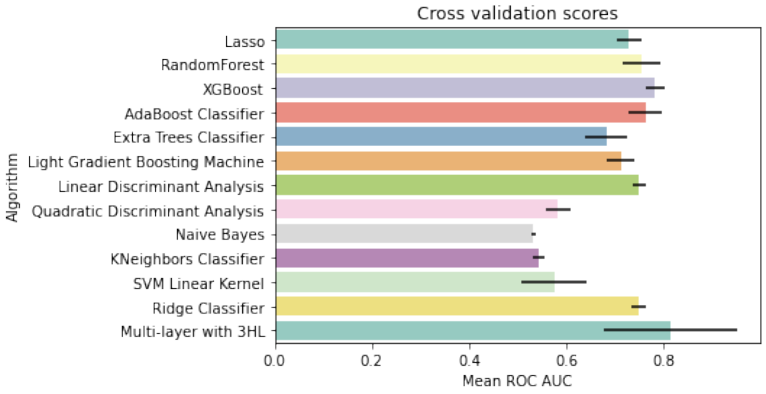

Figure 3: Five-fold cross validation scores for each method. The results are summarized by the mean and a measure of standard error for each model.

Ridge Classifier, XGboost, Random Forest, Linear Discriminant Analysis, Multi-layer with 3HL outperform for cross validation score, but the standard error measure is very high for Multi-layer with 3HL. A grid search is performed for finding the optimal hyperparameters. *Second*, given the imbalance in data we proceed to a resampling in order to made it more balanced by the means of the three oversampling techniques presented in Section 3.2. Different resampling ratios are tested (e.g., 20%/80%, 15%/85%,

[19]Generally, in the tree based ensemble models, the importance of each feature is indicated by a score, which allows to rank the variables and compare to each other. The more the attribute is used to make key decisions in the splitting process of decision trees, the higher its relative importance. The feature importance is calculated for each decision tree and then averaged across all the decision trees within the model. For more technical information on how to assess the feature importance, see Hastie et al. (2009).

10%/90%, 5%/95%). The competing models are then estimated on the transformed data.[20] *Third*, for the sake of interpretability, all continuous variables are discretized and the best performing model identified previously is also estimated on this transformed database. The interpretation of results is done using the SHAP values technique.

5. *Individual versus group-level analyses.* The entire analysis is done first on the original dataset with both categorical and numerical variables. In a second time, all numerical variables are converted into categorical variables and all the analysis is based on this new data set. This allows us to assign customers with specific characteristics to suspicious complaints, introduce more non-linearity into the model and facilitate interpretation.

4.2 Individual versus group level analysis

Individual-level analysis results. Given the results of the preliminary analysis, for the rest of the analysis we built only XGBoost, Random Forest, Ridge Classifier, Linear Discriminant Analysis and Multi-layer with 3HL models on the training set and evaluate their performances on the test set based on: F-measure, FDR, FNR, AUC-ROC, AUC-PR, Log loss and Brier score.[21] Best performing models are those with the lowest FDR, FNR, log-loss and Brier score, and biggest F-measure, AUC-ROC and AUC-PR. To avoid overfitting and improve the models performance, a five-fold cross validation procedure was applied for each model. Results are presented in Table 2 and we retain as best performing approach for this specific setting the one using XGBoost. As the differences in statistical performance between the different models selected may seem subtle and confusing to the reader, we considered several key performance indicators to label a model as the best: (i) its predictive performance, (ii) the interpretability of its results, (iii) its operational efficiency and its economic value. (i) First, XGBoost combines biggest AUC ROC and AUC PR and lowest log loss. The performance measures of the models with resampling technique are slightly improved, but XGBoost remains the best of them. Its performance (in terms of AUC and LogLoss) is improved slightly by applying SMOTE with the ratio 15%/85% and we retain thus this specification for the rest of the analysis. Besides, there is a great demand for effective predictive methods which maximize the true positive detection rate and minimize the false positive rate, as classification errors might imply different costs (i.e., misclassifying fraudulent claims is generally much costlier than misclassifying non-fraudulent claims). For instance, despite its reasonable predictive performance and ease

[20] Note that the resampling is only performed on the training dataset, and not on the holdout test dataset. The intent is to continue to evaluate the resulting model on data that is both real and representative of the target problem domain (He and Ma, 2013).

[21] For F-measure and FDR and FNR measures, we fix for each model a cutoff equal to the 99% percentile of the estimated probability series. As a result, if the probability of a claim estimated by the classification model is greater than this threshold value, then the claim is predicted as fraudulent, otherwise it is classified as non-fraudulent.

MODEL	F-measure	False discovery rate (FDR)	False negative rate (FNR)	AUC-ROC	AUC-PR	Log Loss	Brier score
Ridge							
Without resampling	0.086	0.925	0.9	0.757	0.043	0.325	0.077
SMOTE 20% / 80%	0.049	0.957	0.943	0.69	0.021	0.335	0.082
SMOTE 15% / 85%	0.061	0.946	0.929	0.689	0.019	0.332	0.081
SMOTE 10% / 90%	0.061	0.946	0.929	0.699	0.021	0.33	0.079
SMOTE 5% / 95%	0.074	0.935	0.914	0.718	0.024	0.327	0.078
ADASYN 20% / 80%	0.061	0.946	0.929	0.696	0.022	0.334	0.081
ADASYN 15% / 85%	0.061	0.946	0.929	0.694	0.022	0.332	0.08
ADASYN 10% / 90%	0.061	0.946	0.929	0.702	0.023	0.33	0.079
ADASYN 5% / 95%	0.074	0.935	0.914	0.716	0.027	0.327	0.078
ROSE 20% / 80%	0.086	0.925	0.9	0.76	0.041	0.407	0.114
ROSE 15%/ 85%	0.098	0.914	0.886	0.754	0.042	0.388	0.105
ROSE 10% / 90%	0.098	0.914	0.886	0.757	0.044	0.368	0.096
ROSE 5% / 95%	0.086	0.925	0.9	0.761	0.043	0.345	0.086
Linear Discriminant							
Without resampling	0.086	0.925	0.9	0.757	0.043	0.061	0.009
SMOTE 20% / 80%	0.049	0.957	0.943	0.69	0.021	0.15	0.008
SMOTE 15% / 85%	0.061	0.946	0.929	0.689	0.019	0.174	0.008
SMOTE 10% / 90%	0.061	0.946	0.929	0.698	0.021	0.209	0.008
SMOTE 5% / 95%	0.074	0.935	0.914	0.717	0.024	0.25	0.008
ADASYN 20% / 80%	0.061	0.946	0.929	0.696	0.022	0.153	0.008
ADASYN 15% / 85%	0.061	0.946	0.929	0.694	0.022	0.178	0.008
ADASYN 10% / 90%	0.061	0.946	0.929	0.702	0.022	0.211	0.008
ADASYN 5% / 95%	0.074	0.935	0.914	0.716	0.027	0.253	0.008
ROSE 20% / 80%	0.086	0.925	0.9	0.759	0.041	0.168	0.037
ROSE 15%/ 85%	0.098	0.914	0.886	0.754	0.042	0.139	0.029
ROSE 10% / 90%	0.098	0.914	0.886	0.757	0.044	0.108	0.02
ROSE 5% / 95%	0.086	0.925	0.9	0.761	0.043	0.08	0.014
Multi-layer with 3HL							
Without resampling	0.086	0.925	0.9	0.659	0.027	0.143	0.015
SMOTE 20% / 80%	0.123	0.892	0.857	0.703	0.044	0.074	0.012
SMOTE 15% / 85%	0.11	0.903	0.871	0.665	0.053	0.094	0.009
SMOTE 10% / 90%	0.074	0.935	0.914	0.663	0.04	0.142	0.013
SMOTE 5% / 95%	0.074	0.935	0.914	0.678	0.049	0.164	0.015
ADASYN 20% / 80%	0.11	0.903	0.871	0.709	0.058	0.068	0.012
ADASYN 15% / 85%	0.086	0.925	0.9	0.663	0.065	0.108	0.014
ADASYN 10% / 90%	0.086	0.925	0.9	0.665	0.041	0.145	0.013
ADASYN 5% / 95%	0.086	0.925	0.9	0.655	0.04	0.142	0.012
ROSE 20% / 80%	0.061	0.947	0.929	0.672	0.033	0.17	0.027
ROSE 15%/ 85%	0.074	0.935	0.914	0.656	0.062	0.211	0.023
ROSE 10% / 90%	0.074	0.935	0.914	0.657	0.033	0.206	0.018
ROSE 5% / 95%	0.074	0.935	0.914	0.649	0.05	0.205	0.016
RandomForest							
Without resampling	0.16	0.863	0.814	0.746	0.155	0.04	0.007
SMOTE 20% / 80%	0.16	0.86	0.814	0.762	0.105	0.082	0.011
SMOTE 15% / 85%	0.16	0.86	0.814	0.776	0.118	0.092	0.012
SMOTE 10% / 90%	0.16	0.86	0.814	0.751	0.149	0.064	0.009
SMOTE 5% / 95%	0.147	0.871	0.829	0.766	0.14	0.057	0.008
ADASYN 20% / 80%	0.16	0.86	0.814	0.747	0.124	0.084	0.011
ADASYN 15% / 85%	0.16	0.86	0.814	0.758	0.125	0.093	0.012
ADASYN 10% / 90%	0.172	0.849	0.8	0.754	0.16	0.057	0.008
ADASYN 5% / 95%	0.16	0.86	0.814	0.777	0.13	0.053	0.008
ROSE 20% / 80%	0.172	0.849	0.8	0.784	0.13	0.104	0.014
ROSE 15%/ 85%	0.147	0.871	0.829	0.764	0.135	0.09	0.012
ROSE 10% / 90%	0.172	0.849	0.8	0.771	0.153	0.073	0.01
ROSE 5% / 95%	0.16	0.86	0.814	0.795	0.142	0.056	0.008

MODEL	F-measure	False discovery rate (FDR)	False negative rate (FNR)	AUC-ROC	AUC-PR	Log Loss	Brier score
XGBoost							
Without resampling	**0.16**	**0.86**	**0.814**	**0.785**	**0.165**	**0.041**	**0.007**
SMOTE 20% / 80%	0.184	0.839	0.786	0.789	0.164	0.068	0.011
SMOTE 15% / 85%	0.196	0.828	0.771	**0.794**	0.175	**0.064**	**0.01**
SMOTE 10% / 90%	**0.209**	**0.817**	**0.767**	0.782	0.183	0.073	0.012
SMOTE 5% / 95%	0.196	0.828	0.771	0.781	**0.188**	0.064	0.01
ADASYN 20% / 80%	0.184	0.839	0.786	0.752	0.18	0.087	0.015
ADASYN 15% / 85%	0.196	0.828	0.771	0.755	0.172	0.082	0.013
ADASYN 10% / 90%	0.196	0.828	0.771	0.78	0.181	0.074	0.012
ADASYN 5% / 95%	0.184	0.839	0.786	0.788	0.185	0.064	0.01
ROSE 20% / 80%	0.184	0.839	0.786	0.765	0.184	0.273	0.074
ROSE 15%/ 85%	0.172	0.849	0.8	0.767	0.186	0.23	0.058
ROSE 10% / 90%	0.172	0.849	0.8	0.773	**0.202**	0.16	0.034
ROSE 5% / 95%	0.147	0.871	0.829	0.784	0.155	0.11	0.021

Table 2: Comparison metrics results with/without imbalanced dataset correction. The performance of the models (as measured by F1, FNR and FDR) is assessed by using as threshold the 99th percentile of the predicted probabilities on the test dataset. The best performing models are those with the lowest FDR, FNR, log-loss and Brier score, and biggest F-measure, AUC-ROC and AUC-PR.

of interpretation of results, Logistic regression should not be adopted systematically as the results provide a high false positive rate. In practice, an acceptable false positive rate for the anti-fraud team is less than 85%. (ii) Second, it is often argued that machine learning models are "black box" models difficult to interpret and that they tell nothing about statistical significance. For this reason, many studies defend the use of standard econometric methods. However, several methods have recently been proposed to deal with the interpretation of machine learning results (see Section 4.3 of the paper), and their functioning is quite easy to understand and implement. These solutions also make XGBoost results valuable and turn them into a useful decision support tool for the fraud investigation team. (iii) Third, the operational efficiency of a model is assessed in terms of the ease of deployment and execution, the processing time, the effort needed to collect, process and eventually integrate new data, evaluate and reestimate (if necessary) the model. For all these elements, XGBoost stands out also as the best choice.

To further improve the best specification identified previously, a two-step analysis is then performed: a first estimation of the XGBoost model to select the best features (which leads to 63 out of 125 features) and a second estimation on the pre-selected variables. Table 3 shows that the results are improved compared to the initial XGBoost. However, the SMOTE correction with ratio 15%/85% only improves slightly the results.

XGBOOST	F-measure	False discovery rate (FDR)	False negative rate (FNR)	AUC-ROC	AUC-PR	Log Loss	Brier score
Without resampling with pre-selected variables	0.170	0.849	0.800	0.801	0.171	0.045	0.008
SMOTE 15% / 85% with pre-selected variables	0.172	0.849	0.8	0.786	0.19	0.061	0.009

Table 3: XGBoost results without resampling and SMOTE 15% / 85% with pre-selected variables.

Group-level analysis results. All numeric variables are converted into categorical variables to identify

specific groups, introduce more non-linearity and facilitate interpretation. In this way, the group-level results can provide contrasting explanations between groups and allows us to identify the *precise categories* of the most important variables explaining fraud activity. This part of the analysis is preferred by the anti-fraud team, even though it implies a loss of information. The analysis is only performed here for the XGBoost and XGBoost with SMOTE 15%/85% correction. A pre-selection of the features is also done by XGBoost and 107 out of 209 features are selected. Table 4 shows the results at group level: criteria values show a less performant XGBoost model than previously, due to the loss of information induced by the transformation of the variables.

XGBOOST	**F-measure**	**False discovery rate (FDR)**	**False negative rate (FNR)**	**AUC-ROC**	**AUC-PR**	**Log Loss**	**Brier score**
Without resampling with pre-selected variables	**0.172**	**0.849**	**0.8**	**0.752**	**0.163**	**0.047**	**0.008**
SMOTE 15% / 85% with pre-selected variables	0.16	0.86	0.814	0.738	0.092	0.121	0.022

Table 4: XGBoost results without resampling and SMOTE 15% / 85% with pre-selected variables and transformed to categorical.

Features importance. Based on these specifications, we perform gain criterion of feature to identify which variables were the most relevant at predicting household fraud.[22] Figure 4 (panel A) shows the average training loss reduction gained when using a feature for splitting. We observe that the payment mode of the policy, the cause of the insurance claim declared either as theft, civil liability, water leaks or natural disaster, the loss frequency at policy level, the change in guarantee level, a basic insurance coverage, the number of total insurance policies and the construction year of the building are only a few variables having the highest importance as predictors of the probability of fraud. For instance, according to this approach clients who live in a flat in a building built in the 1950s are important elements to predict household fraud.

The extension of the analysis to the group-level stage goes further and can show if the predictive performances of the model are, for instance, more influenced by a specific category of customer age, or a high frequency of loss at the policy level (e.g., more than one claim per year).[23] Therefore, the average training loss reduction gained when using a feature for splitting is also performed within the group-level analysis (Figure 4, panel B). It shows not only the best relevant features for model prediction, but also the best category for each feature. Feature importance at group level is relatively consistent with feature importance at individual level, but it provides more information about each relevant feature. For instance, it is not only the number of insurance policies that is important for the predictive performance of the model, but the insurance contracts with one open policy.

[22]The gain is shown to be one of the most relevant attributes to interpret the relative importance of each feature.
[23]Appendix F provides the details of the transformed variables.

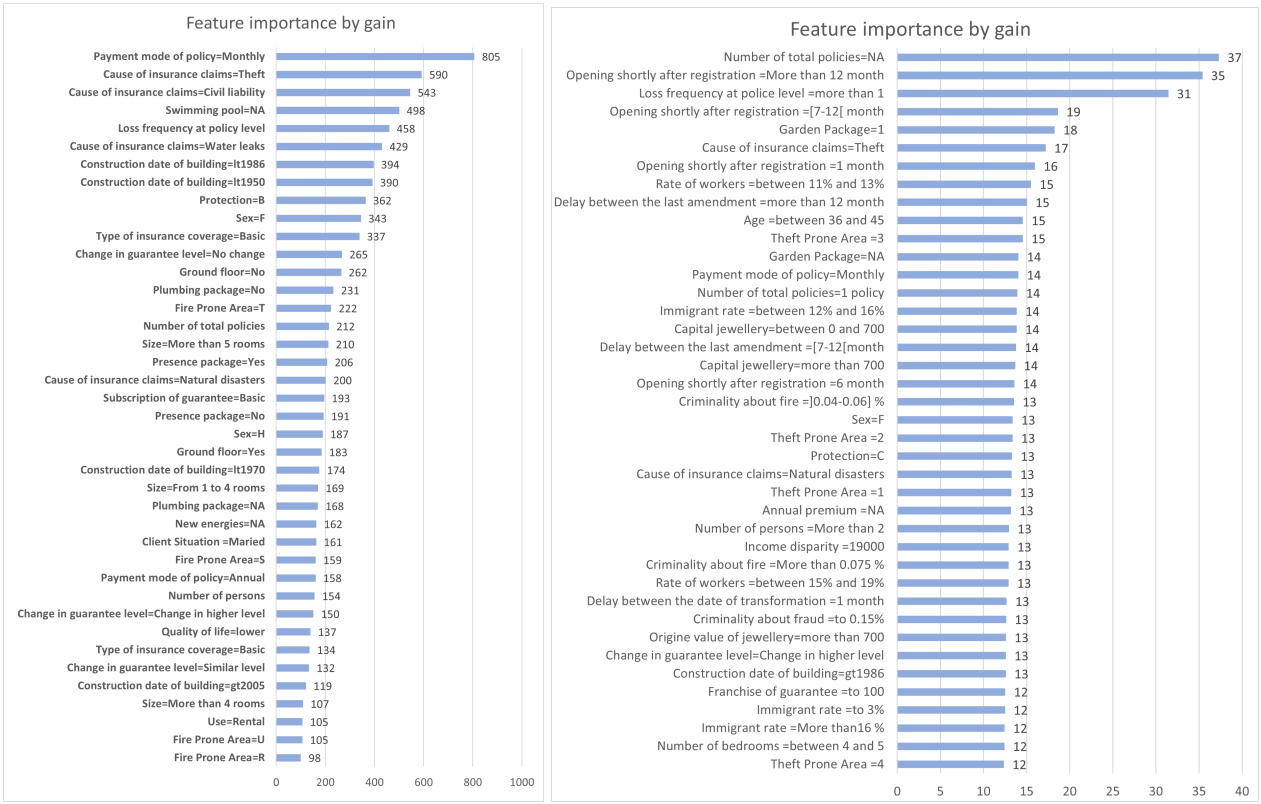

Figure 4: Feature importance by gain for the first 40 features according to XGBoost with SMOTE 15%/85% at individual level (panel A) and group level (panel B). The gain implies the relative contribution of each feature to the model. A higher value of this criterion compared to another feature indicates that the feature is more important for generating a prediction.

4.3 Interpretation via SHAP values

The feature importance plot seen previously is useful, but hold any information beyond the importances. Several other methods have been recently proposed to deal with the interpretation of machine learning results (e.g., the partial dependence plots, the individual conditional expectation plots, LIME, DeepLIFT, etc.). For this study, we use SHAP (SHapley Additive exPlanations) method introduced by Lundberg and Lee (2017); Lundberg et al. (2018), which comes as a generalization of the previous methods. For instance, SHAP feature importance is an alternative to permutation feature importance. There is a major difference between both importance measures: while permutation feature importance is based on the decrease in model performance when using a feature for splitting, SHAP is based on the importance of feature allocations to the predicted outcome (Molnar, 2020). Moreover, Lundberg and Lee (2017) propose TreeSHAP as a variant of SHAP for tree-based models, faster than KernelSHAP and optimized under Python.[24] As argued by its authors, this method proposes an unified framework for interpreting individual and global predictions (both in regression and classification settings), deals with the trade-off between the accuracy and the interpretability of a model's output, demonstrates better agreement with human intuition and better identification of influential features. The combination of a

[24]More recently, also based on Shapley values, (Sullivan et al., 2022) have developed a new methodology, called eXplainable PERformance (XPER), which measures the marginal contribution of a particular feature to the predictive or economic performance of a regression or classification model. XPER values decomposes thus a performance metric among features in a model.

solid theoretical justification and a fast-practical algorithm makes SHAP values a powerful tool for confidently interpreting tree models such as XGBoost. Global interpretability shows how much each feature contributes to the target variable, while local interpretability indicates the contribution of features to each observation.

Figure 5 is the SHAP summary plot of the first 30 features at individual (panel A) and group (panel B) level for XGBoost, ranked in descending order.[25] Each individual in the dataset is run through the model and a dot is created for each feature. It represents hence the SHAP value for a feature and an instance. The position on the y-axis is related to the feature and on the x-axis to the SHAP value. The color represents the value of the feature from low (blue) to high (red). Panel (A) shows that the first two most important risk factors that push the model to predict household fraud are the loss frequency at policy level and by the total number of policies. A high level of the loss frequency at policy level has a high and positive impact on the fraud probability. Additionally, panel (B) gives more details showing the exact categories of each feature that help the most to the fraud prediction. For instance, declaring more than one loss at policy level is strongly correlated to the fraud probability. Also, relevant features from INSEE variables, such as the divorce rate to 7%, an immigrant rate over 16%, and high levels of criminality fire by city postcode are associated with fraud probability.

Figure 6 is a simplified version of the previous plot and indicates the correlation between SHAP values and each feature across the data at individual and group levels (panels (A) and (B), respectively). The link with the target variable is highlighted trough different colours: the red (blue) colour means that the feature is positively (negatively) correlated to the household fraud probability. Thus, a strong level of loss frequency at policy level is associated with higher household fraud probabilities (panel (A)). The total number of policies and the water leaks are negatively correlated to the probability of fraud, while the occasion of theft, and the customer with at least one claim are positively correlated to the probability of fraud (panel (B)). Additionally, Figure 6 (A) indicates that globally the variable "Age" is not in the first 30 features correlated with the household fraud. However, Figure 6 (B) comes and completes the information, showing that within the age, the category between 36 and 45 has a positive impact on fraud probability, with an average SHAP value equal to 0.078. Similarly, a long time period (more than one year) between the opening of the claim and the subscription is associated with high a probability of fraud. In the same way, the insurance policies with one policy, that corresponds usually to a new policyholder, is also positively associated with fraud.[26]

[25]SHAP summary plots rank features by the sum of SHAP value magnitudes over all instances and uses SHAP values to show the distribution of the impact each feature has on the model output. Hence, they indicate how each feature affects the dependent variable. Standard feature importance bar charts give a notion of relative importance in the test dataset. The features are ordered according to their importance (the higher the mean SHAP value, the bigger the impact on the model output and more important the feature is in predicting the output).

[26]The term "Number of total policies" means the total number of policies that a client has actually underwritten for all different risks. Exceptionally, a missing value "NA" can be interpreted here as one policy.

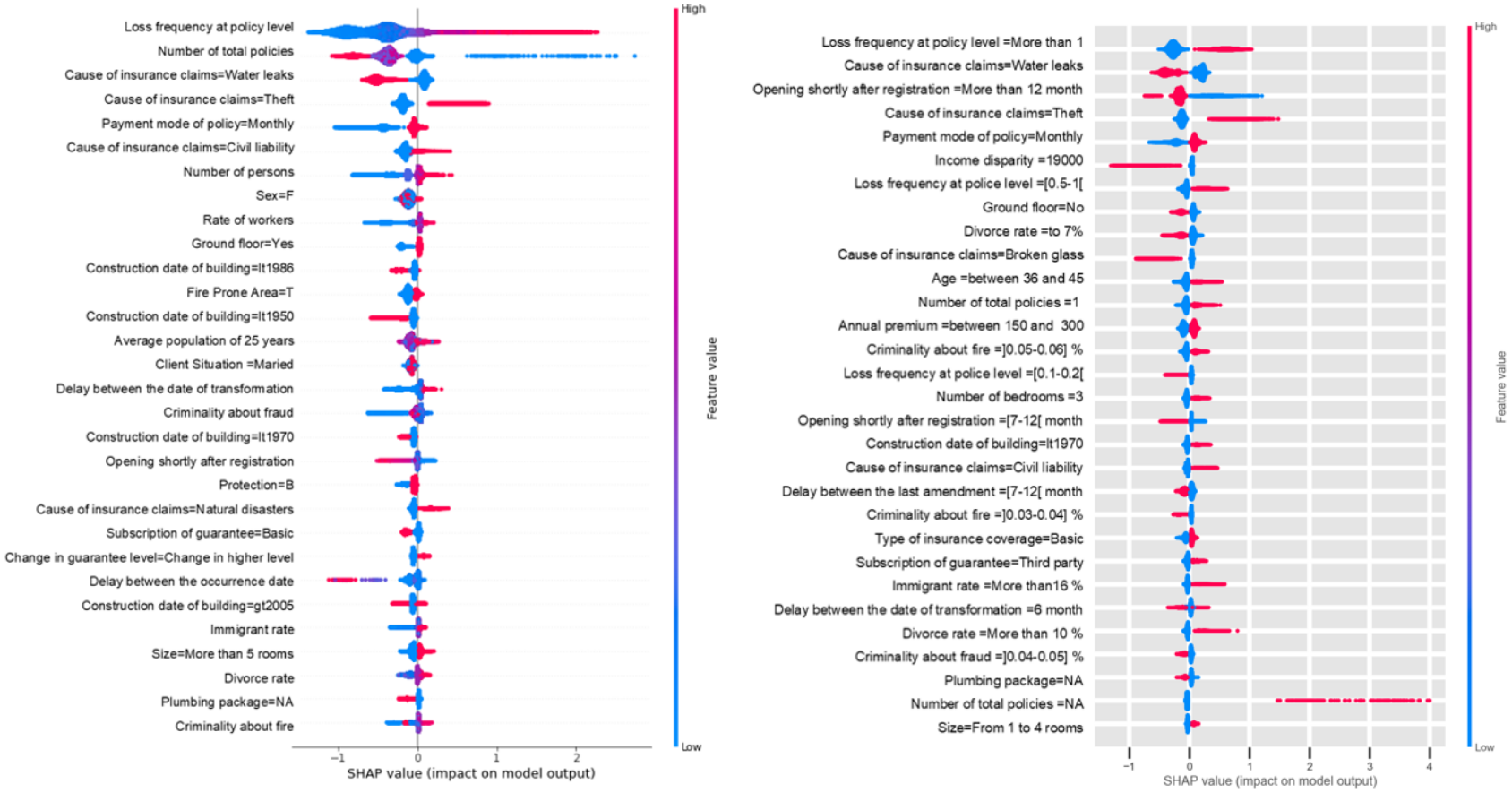

Figure 5: Global interpretability/ Individual SHAP values. Panel (A): Model for individual-level analysis of XGBoost with Smote 15%/85%; Panel (B): Model for group-level analysis of XGBoost with pre-selected variables. These graphs show the global importance of top 30 features in household fraud prediction based on SHAP values of each individual in the dataset. Features are ranked in descending order. The x-axe shows the signs of the effect on the prediction of fraud and the magnitude of its effect is indicated by different colors. The dot is blue if the effect is low while it is red if the effect is high.

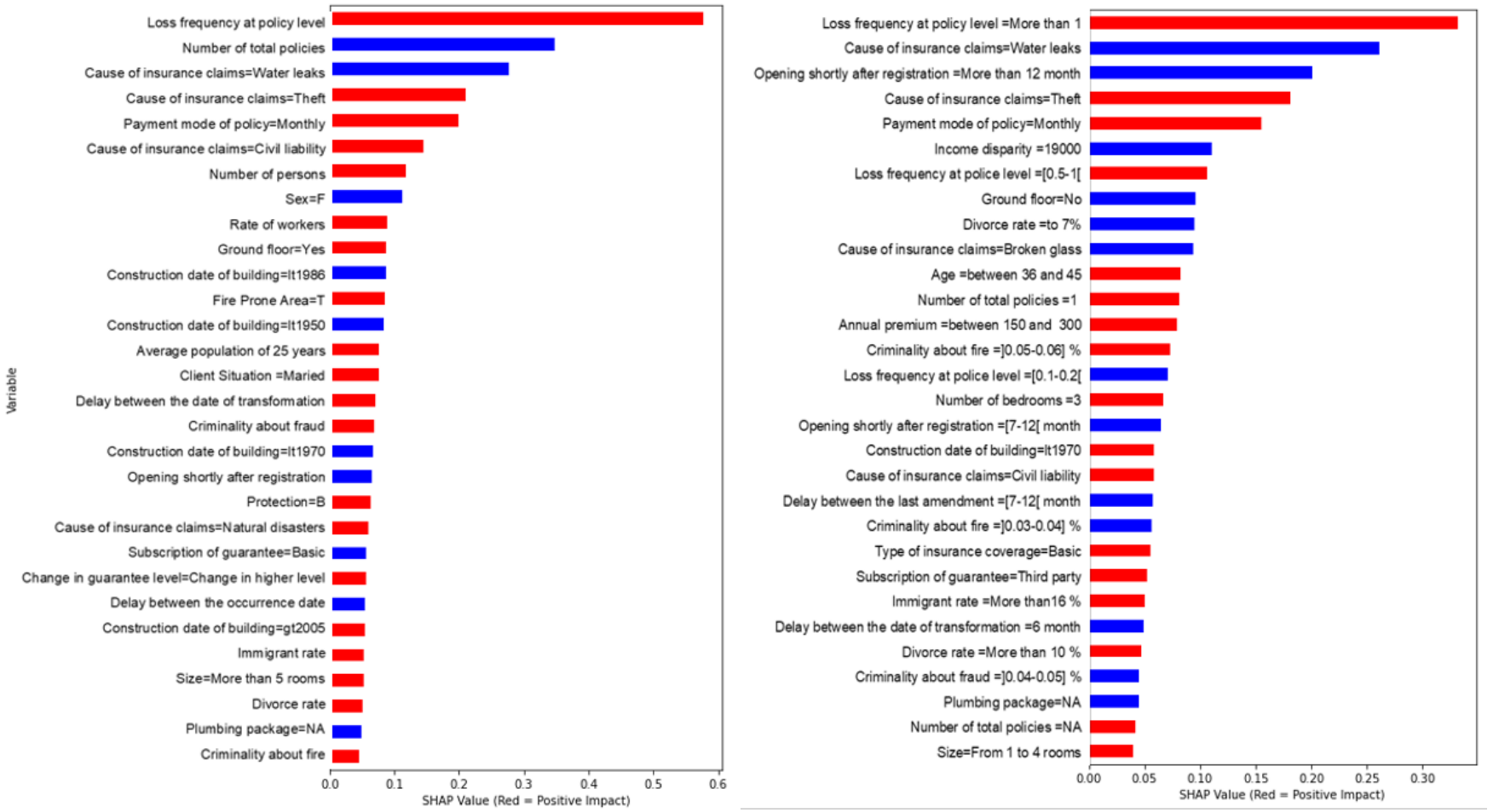

Figure 6: Global interpretability/Average SHAP values. Panel (A): Model for individual-level analysis of XGBoost with Smote 15%/85%; Panel (B): Model for group-level analysis of XGBoost with pre-selected variables. These graphs show the global importance of the top 30 features in household fraud prediction based on the average absolute SHAP values per feature across the data. Features are classified in descending order. The red color means that the feature is positively correlated with fraud prediction, while the blue color means that the feature is negatively correlated with fraud prediction.

It is also observed that if the reported causes of loss are theft, liability or natural disasters, or if there is a change in coverage to a higher level, more attention should be paid to the investigation. In practice, these observations can be easily transposed into investigation rules by the anti-fraud team, and most of them comply with the expert expectations. Typically, these findings provide alerts of the highest rated claims to the anti-fraud team the day after the declaration.

The SHAP value allows to visualize also individual predictions, as it analyses the importance of features for each observation. This is known as the *force plot* and identifies for each observation its prediction and the contributions of the predictors. On the right side, "blue" variables are the ones pushing the fraud probability towards lower values, while on the left side the "red" variables are increasing its value. Let us illustrate this with examples of randomly selected fraudulent and non fraudulent observations.

Figure 7 presents a fraudulent case as its SHAP value of 2.59 is higher than the base value of -0.634. For this case, the high level of protection of household is associated with a lower probability of fraud, while the loss frequency at the policy level, the theft and the total number of policies tend to increase the fraud probability.

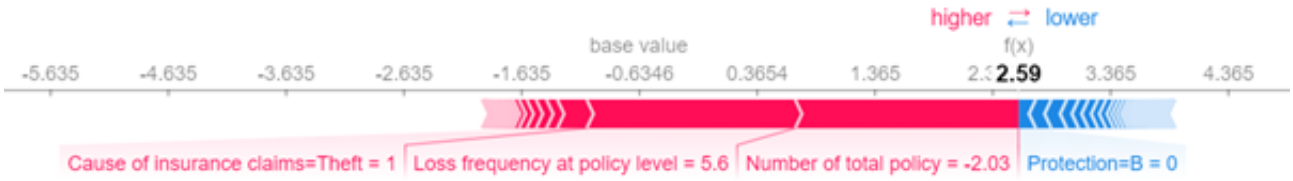

Figure 7: Local interpretability/individual SHAP values for a fraudulent case (A) : Model for individual-level analysis. The horizontal location shows whether the effect of that observation is associated with a higher or lower prediction. The red arrow shows that the feature pushes the prediction towards fraud, and the blue arrow shows that the feature pushes the prediction towards non fraud.

Figure 8 presents the analysis of the same case using the model at group level. The observation displays a SHAP value of -2.36, which is higher than the base value (SHAP value of -3.56) and implies a fraudulent case. The payment that is not done on a monthly basis is associated with a lower probability of fraud, while the cause of theft and the high immigrant rate recorded for this individual are associated with higher probability. This information complete the overview of Figure 7.

Figure 8: Local interpretability/individual SHAP values for a fraudulent case (B): Model for group-level analysis.

In the same way, Figure 9 presents a non-fraudulent case: this observation has the average predicted SHAP value of -3.18 which is lower than the base value. In this case, as the claim does not make the object of water

leaks, this variable is associated with a higher probability of fraud, while the objective of the claim related to a broken glass is associated with a lower probability of fraud. Figure 10 (B) presents the same analysis of this non-fraudulent case using the model at group level. The observation displays a SHAP value of -4.52 that is still lower than the base value. The building built in the 1970s is associated with a higher probability of fraud, while the broken glass and the loss frequency less than one are associated with lower probability of fraud.

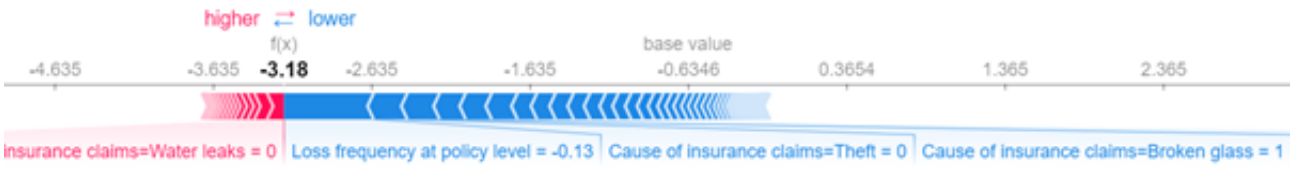

Figure 9: Local interpretability/individual SHAP values for a non-fraudulent case (A): Model for individual-level analysis.

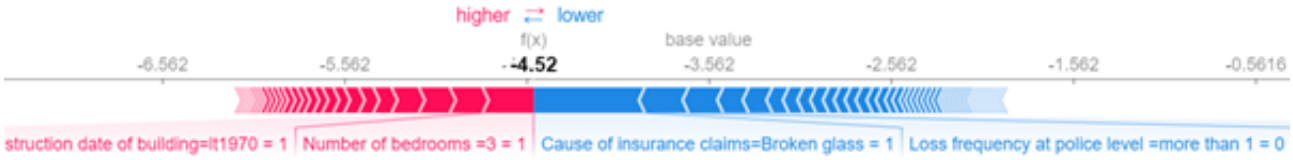

Figure 10: Local interpretability/individual SHAP values for a non-fraudulent case (B): Model for group-level analysis.

In summary, the XGBoost algorithm is the best model for predicting household fraud using a large dataset including company and external variables. Interpreting the results of the XGBoost algorithm using SHAP values allows the identification of conduct characteristics that are positively or negatively associated with the probability of fraud. Based on these results, the loss frequency at policy level, the total number of policies and the causes of insurance claims have great importance. Moreover, the use of external data highlights specific information by department or city postcode that is associated with the prediction of the probability of household fraud (e.g., an immigrant rate over 16%, a divorce rate over 10%).

5 Conclusion

In this paper, we use different supervised learning approaches to provide a study on fraud modelling on Personal Home policies. The objective is to propose a solution and facilitate the tasks of the anti-fraud team, which must analyse many files every day. However, the investigation team could not verify all claim declarations in the timelines required. The solution proposed orders claims by fraud potential according to a statistical approach and thus enables the team to concentrate on the most suspicious cases. The personal home fraud alerts are hence automated in order to improve the retention of good customers and reduce the expertise expenses and management delays of claims.

The first contribution of the paper consists in the use of a real-time database on home policy insurance. Secondly, we compare the performances of more than ten machine learning methods in a real-time processing environment in order to improve the performance of detection of fraudulent cases. Because fraud is a rare-event, the fraudulent cases in the database represent a very small proportion of fraudulent claims (0.76% in our case). To deal with this issue, we resampled our training data set and used the following methods: SMOTE, ADASYN and ROSE. Moreover, this study develops approaches with good predictive performances, while keeping the interpretability of the method. We used SHAP (SHapley Additive exPlanations) method to explain individual predictions. Our results show that on this specific case, XGBoost is the best performing method to detect home policy fraud insurance. The results are the same when resampling techniques are applied. Based on SHAP values, our results indicate that the loss frequency at policy level has the highest importance as predictor of fraud. It also important to closely monitor the total number of policies and the declared causes of the insurance claims.

This study has direct managerial implications because it allows the anti-fraud team to seek out high-risk fraud claims (according to the model) that the appraisers do not find necessarily suspicious, and to create alerts for the expert. An important part of providing improved risk scores is the ability to feed these scores to the direct complaints center in real time. One of the main benefits of including scores and rule results in the claim center will be the ability to have a risk summary in the report sent to the experts and the investigation team. Moreover, the construction of a reliable score makes possible the automation of fraud risk alerts and reduces the management delay. Especially for small amounts of claims, expenses can be reduced by avoiding expertise. Overall, the adoption of such an approach generates economic gains through the optimisation of the insurance company's expenses, better customer retention and more effective marketing strategy. This study could be improved by using spatial information, more precisely by taking into account the address of the policyholder. In addition, other over-sampling technique could be applied since optimal results of over-sampling technique depend on the chosen method.

A Appendix A

The main steps of the Random Forest method are the following (Hastie et al., 2009):

1. Given a data set with N observations and p inputs.

2. Fix m = a constant chosen on beforehand (corresponding generally to 1, 2 or $floor(log_2(p) + 1)$).

3. For i= 1 to B:

   (a) Draw a bootstrap sample i of size N from the training data.

   (b) Grow a full (without pruning) Random Forest tree k to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree:

      i. Select m variables at random from the p variables on which to base the splitting decision.

      ii. Split on the best of this subset.

      We denote by $\mathcal{D}_{k,j}$ the data at a given node or iteration j for the tree k. Let $\theta_{k,j} = (x_{k,j}, s_{x_{k,j}})$ be a candidate split, with $\{x_{k,j}\}_{j=1,\dots,p}$ a given predictive variable and $s_{x_{k,j}}$ a threshold value in the support of this variable. The algorithm splits the data $D_{k,j}$ into two subsets $\mathcal{D}_{k,j,1}(\theta_{k,j}) = (x_i, y_i)|x_{k,j} \leq s_{x_{k,j}}$ and $\mathcal{D}_{k,j,2}(\theta_{k,j}) = (x_i, y_i)|x_{k,j} > s_{x_{k,j}}$.

      The estimate $\hat{\theta}_{k,j}$ of $\theta_{k,j}$ is found such that:

$$\hat{\theta}_{k,j} = (\hat{x}_{k,j}, \hat{s}_{x_{k,j}}) = \underset{\theta_{k,j}}{argmax}\{\mathcal{H}(\mathcal{D}_{k,j}) - \mathcal{H}(\mathcal{D}_{k,j,1}(\theta_{k,j}), \mathcal{D}_{k,j,2}(\theta_{k,j}))\},$$

      with $\mathcal{H}(.)$ a measure of diversity (usually approximated by the Gini criterion).

   (c) Output the ensemble of trees $\{k_i\}_{i=1}^{B}$.

4. To make a prediction for a new observation x_n, let $\hat{C}_i(x_n)$ be the class prediction of the ith Random Forest tree. Then $\hat{C}_{rf}^{B}(x_n) = majority\ vote\ \{\hat{C}_i(x_n)\}_1^B$.

B Appendix B

For the description, let consider $\hat{y}_i$ the predicted value of the entry i, as defined by XGBoost:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}, \tag{4}$$

with f_k an independent tree in the space of regression trees $\mathcal{F}$, and $f_k(x_i)$ the predicted score given by the k-th tree of the i-th sample. To learn the set of functions f_k, XGBoost minimizes the following regularized objective function:

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{5}$$

with

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||\omega||^2. \tag{6}$$

$l(.)$ herein is a training loss function, which measures the difference between the predictive value $\hat{y}_i$ and the target y_i. The second term, $\Omega(.)$, is the penalization factor, which aims at avoiding the overfiting of the model. γ and λ are two regularization parameters, and T and ω are the numbers of leaves and the score of each leaf, respectively. The tree ensemble model is then trained in an additive manner by greedily adding for each instance i at each iteration t, the function of the regression tree, f_t, that mostly improves the previous prediction $\hat{y}_i^{(t-1)}$:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{7}$$

After using the second order Taylor expansion and removing the constant term, the objective function to minimize becomes:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{N} [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \Omega(f_t), \tag{8}$$

where g_i and h_i stand for the first and the second order gradient of the loss function $l(.)$. Consider $I_j = \{i|q(\mathbf{x}_i) = j\}$ the instance set of leaf j. For a fixed $g(\mathbf{x})$, the optimal weight $\omega_j^\star$ of leaf j and the corresponding optimal value of the objective function, $\tilde{\mathcal{L}}^{(t)}(q)$, are given by

$$\omega_j^\star = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \tag{9}$$

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T. \tag{10}$$

As stated by Chen and Guestrin (2016), $\tilde{\mathcal{L}}^{(t)}(q)$ can be used as a scoring function to measure the quality of the tree structure q (i.e., the smaller the score, the better the structure). As it is impossible to enumerate all the possible tree structures q, the authors propose the use of a greedy algorithm that starts from a single leaf and iteratively adds branches to the tree. The formula used to evaluate the candidate split is given by:

$$\mathcal{L} = \frac{1}{2}\Big[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda}\Big] - \gamma \tag{11}$$

where I_L and I_R are the instance sets of left and right nodes after the split and $I = I_L \cup I_R$. Note that when the regularization parameter of the model is set to zero, the objective falls back to the traditional gradient tree boosting (Friedman, 2001).

C Appendix C

The table below offers complete definitions and details on the computation of different performance metrics.

Type	Measure	Formula
Threshold metrics	**Sensitivity** or **true positive rate (TPR)** or **Recall** or **Hit rate** describes how good the model is at predicting the positive class when the actual outcome is positive, or how many of the fraudsters are correctly labelled by the model as a fraudster.	$Sensitivity = Recall = Hit\ rate = \frac{TP}{TP+FN}$
	Precision indicates how many of the predicted fraudsters are actually fraudsters.	$Precision = \frac{TP}{TP+FP}$
	F-measure (F1) focuses on the analysis of positive class by combining the precision and the recall measures.	$F-measure = 2 \times (Precison \times Recall)/(Precision + Recall)$
	False discovery rate (FDR) is the fraction of negative cases incorrectly predicted to be in the positive class out of all predicted positive cases. FDR represents the probability of a fraudster with a high score to actually have a low score.	$FDR = \frac{FP}{TP+FP}$
	False negative rate (FNR) is the fraction of positive cases incorrectly predicted to be in the negative class out of all actual positive cases. FNR represents the probability of a fraudster with a good predicted score to actually have a bad score.	$FNR = \frac{FN}{TP+FN}$
Ranking metrics	**AUC-ROC** (Bradley, 1997; Hanley and McNeil, 1982) represents the area under the ROC (receiver operating characteristic) curve, which plots the sensitivity versus 1-specificity. Compared to the previous measures, it is not dependent of the cut-off evaluating hence the overall discriminatory performance of a model or classier. AUC-ROC is very close to 1 for a better model.	
	AUC-PR represents the area under the precision-recall curve.	
Probabilistic metrics	**Log loss** or **cross-entropy loss** is a measure of dissimilarity between the predicted probability of fraud for individual i (p_i) and the actual label ($y_i \in \{1\ if\ fraud,\ 0\ otherwise\}$). The smaller the better.	$CE = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i) + (1-y_i) log(1-p_i))$
	Brier score (BS) is another measure of dissimilarity between the predicted probability of fraud and the actual label (similar to the Mean Square Error in the regression analysis) and it is bounded between 0 and 1. BS is a loss function, so lower values indicate better discrimination.	$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$

Table C.1: Metrics of comparison

D Appendix D

		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
V1	Age	1																							
V2	Criminality about fraud	-0.04 0.00	1																						
V3	Criminality about fire	-0.10 0.00	0.03 0.00	1																					
V4	Income disparity	-0.03 0.00	0.26 0.00	-0.08 0.00	1																				
V5	Number of persons	-0.23 0.00	-0.09 0.00	0.06 0.00	-0.07 0.00	1																			
V6	Origine value of jewellery	0.06 0.00	0.01 0.00	-0.02 0.00	0.03 0.00	0.14 0.00	1																		
V7	Capital jewellery	0.18 0.00	0.02 0.00	-0.03 0.00	0.04 0.00	0.13 0.00	0.62 0.00	1																	
V8	Origine of valuables	0.08 0.00	0.03 0.00	-0.02 0.00	0.02 0.00	0.00 0.79	0.16 0.00	0.18 0.00	1																
V9	Capital valuables	0.07 0.00	0.02 0.00	-0.01 0.01	0.02 0.00	0.01 0.23	0.19 0.00	0.20 0.00	0.64 0.00	1															
V10	Annual premium	0.20 0.00	0.04 0.00	-0.03 0.00	0.00 0.89	0.30 0.00	0.30 0.00	0.38 0.00	0.14 0.00	0.13 0.00	1														
V11	Loss frequency at police level	-0.06 0.00	0.00 0.41	0.01 0.06	0.00 0.41	-0.01 0.01	-0.01 0.15	-0.03 0.00	-0.01 0.08	0.00 0.37	-0.07 0.00	1													
V12	Divorce rate	0.06 0.00	0.02 0.00	-0.21 0.00	0.12 0.00	-0.13 0.00	-0.02 0.00	-0.03 0.00	0.00 0.52	0.00 0.83	-0.09 0.00	0.01 0.02	1												
V13	Immigrant rate	-0.04 0.00	0.34 0.00	0.00 0.35	0.40 0.00	-0.06 0.00	0.01 0.19	0.00 0.29	0.00 0.91	0.00 0.67	-0.04 0.00	0.00 0.35	0.21 0.00	1											
V14	Rate of marry	0.07 0.00	-0.13 0.00	-0.05 0.00	-0.29 0.00	0.15 0.00	0.08 0.00	0.09 0.00	0.02 0.00	0.02 0.00	0.21 0.00	-0.02 0.00	-0.47 0.00	-0.45 0.00	1										
V15	Rate of workers	-0.04 0.00	-0.28 0.00	0.17 0.00	-0.64 0.00	0.10 0.00	-0.04 0.00	-0.06 0.00	-0.02 0.00	-0.02 0.00	-0.05 0.00	0.00 0.43	-0.24 0.00	-0.21 0.00	0.16 0.00	1									
V16	Average population of 25 years	-0.13 0.00	0.19 0.00	0.34 0.00	0.16 0.00	-0.02 0.00	-0.03 0.00	-0.04 0.00	-0.02 0.00	-0.02 0.00	-0.09 0.00	0.01 0.08	-0.04 0.00	0.43 0.00	-0.56 0.00	-0.01 0.25	1								
V17	Number of total policy	0.10 0.00	-0.14 0.00	-0.04 0.00	-0.07 0.00	0.25 0.00	0.07 0.00	0.08 0.00	0.01 0.15	-0.01 0.01	0.09 0.00	-0.02 0.00	-0.06 0.00	-0.11 0.00	0.17 0.00	0.05 0.00	-0.13 0.00	1							
V18	Delay between the last amendment	0.06 0.00	0.03 0.00	-0.01 0.01	0.01 0.06	-0.02 0.00	-0.01 0.01	0.01 0.00	0.00 0.93	0.00 0.73	-0.01 0.04	-0.12 0.00	0.00 0.29	0.01 0.00	-0.01 0.22	-0.01 0.03	0.00 0.32	-0.01 0.04	1						
V19	Delay between the occurrence date	0.00 0.57	0.04 0.00	-0.01 0.00	0.01 0.02	0.00 0.80	-0.02 0.00	-0.02 0.00	0.00 0.36	0.00 0.43	-0.03 0.00	-0.02 0.00	0.02 0.00	0.02 0.00	-0.02 0.00	-0.01 0.00	0.01 0.04	-0.02 0.00	0.50 0.00	1					
V20	Delay between the date of transformation	0.17 0.00	-0.07 0.00	0.01 0.02	-0.05 0.00	0.06 0.00	0.03 0.00	0.05 0.00	0.00 0.92	-0.02 0.00	0.03 0.00	-0.01 0.01	-0.02 0.00	-0.04 0.00	0.06 0.00	0.05 0.00	-0.03 0.00	0.17 0.00	0.01 0.27	-0.01 0.05	1				
V21	Opening shortly after registration	0.30 0.00	0.00 0.37	-0.07 0.00	0.00 0.64	0.04 0.00	0.00 0.68	0.16 0.00	0.01 0.01	-0.01 0.01	0.38 0.00	-0.16 0.00	-0.03 0.00	-0.01 0.18	0.07 0.00	-0.04 0.00	-0.03 0.00	0.06 0.00	0.10 0.00	0.03 0.00	-0.02 0.00	1			
V22	Franchise of guarantee	0.01 0.02	-0.01 0.09	0.01 0.01	0.01 0.09	0.01 0.05	0.08 0.00	-0.03 0.00	0.02 0.00	0.02 0.00	0.01 0.20	-0.04 0.00	0.00 0.79	-0.01 0.12	-0.01 0.24	0.00 0.40	0.00 0.51	-0.03 0.00	0.05 0.00	0.01 0.03	0.05 0.00	-0.06 0.00	1		
V23	Number of guarantees	-0.01 0.22	0.00 0.33	0.02 0.00	0.01 0.00	0.00 0.31	0.04 0.00	0.04 0.00	0.00 0.30	0.01 0.17	0.01 0.14	0.00 0.34	-0.01 0.00	0.02 0.00	0.02 0.00	-0.02 0.00	0.01 0.00	0.00 0.46	-0.02 0.00	-0.05 0.00	0.00 0.65	-0.01 0.14	0.00 0.71	1	
V24	Number of bedrooms	0.15 0.00	-0.27 0.00	0.04 0.00	-0.16 0.00	0.47 0.00	0.24 0.00	0.30 0.00	0.09 0.00	0.09 0.00	0.60 0.00	-0.03 0.00	-0.20 0.00	-0.26 0.00	0.33 0.00	0.14 0.00	-0.17 0.00	0.24 0.00	-0.01 0.07	-0.03 0.00	0.10 0.00	0.16 0.00	0.02 0.00	0.02 0.00	1

Figure D.1: Pearson correlation matrix. P-values are indicated on the second line of each observation. Source: Authors' calculations (size: 46 302 observations)

		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22
V1	Payment mode of policy	1																					
V2	Subscription of guarantee	0.12	1																				
V3	Change in guarantee level	0.08	0.53	1																			
V4	Type of insurance coverage	0.06	0.66	0.44	1																		
V5	Construction date of building	0.09	0.1	0.15	0.07	1																	
V6	Cause of insurance claims	0.04	0.1	0.03	0.12	0.05	1																
V7	Type of house	0.04	0.35	0.12	0.3	0.14	0.41	1															
V8	Nature of the insurance claim	0.14	0.02	0	0.02	0.01	0.21	0.03	1														
V9	New energies	0.08	0.38	0.28	0.23	0.2	0.08	0.12	0.01	1													
V10	Swimming pool	0.08	0.37	0.28	0.24	0.2	0.12	0.17	0.01	0.61	1												
V11	Security package	0.08	0.48	0.24	0.28	0.15	0.05	0.13	0.02	0.6	0.7	1											
V12	Garden Package	0.07	0.27	0.21	0.24	0.13	0.16	0.86	0.03	0.6	0.52	0.58	1										
V13	Plumbing package	0.1	0.26	0.18	0.18	0.11	0.12	0.4	0.01	0.45	0.46	0.38	0.39	1									
V14	Plan	0.02	0.1	0.07	0.05	0.04	0.01	0.01	0.06	0.08	0.08	0.08	0.09	0.06	1								
V15	Presence package	0.06	0.15	0.07	0.13	0.11	0.1	0.12	0.01	0.33	0.44	0.39	0.26	0.14	0.03	1							
V16	Quality subscriber	0.18	0.4	0.11	0.36	0.06	0.27	0.47	0.01	0.14	0.18	0.12	0.46	0.52	0.06	0.05	1						
V17	Ground floor	0.02	0.21	0.07	0.18	0.21	0.24	0.63	0.04	0.07	0.1	0.08	0.5	0.23	0.02	0.11	0.4	1					
V18	Sex	0.06	0.07	0.03	0.04	0.02	0.04	0.08	0.02	0.03	0.03	0.03	0.06	0.06	0.05	0.06	0.1	0.05	1				
V19	Client Situation	0.06	0.15	0.05	0.14	0.72	0.08	0.33	0.03	0.05	0.07	0.08	0.15	0.12	0.02	0.07	0.28	0.2	0.2	1			
V20	Size	0.04	0.26	0.09	0.21	0.1	0.24	0.65	0.03	0.1	0.15	0.1	0.5	0.24	0.02	0.14	0.5	0.58	0.07	0.22	1		
V21	Use	0.07	0.65	0.1	0.51	0.06	0.08	0.04	0.02	0.28	0.28	0.33	0.28	0.18	0.02	0.1	0.16	0.02	0.03	0.02	0.07	1	
V22	Professional use	0.02	0.05	0.03	0.03	0.03	0.03	0.06	0.01	0.02	0.03	0.33	0.05	0.04	0.03	0.32	0.03	0.04	0.02	0.06	0.06	0.03	1

Figure D.2: CramerV statistic. It varies between 0 and 1 without any negative values. A value close to 0 means that there is no association. Source: Authors' calculations (size: 46 302 observations)

		V1	V2	V3	V4
V1	Theft Prone Area	1			
V2	Quality of life	0.04	1		
		0.00			
V3	Protection level of household	0.19	0.18	1	
		0.00	0.00		
V4	Fire Prone Area	0.47	-0.21	0.05	1
		0.00	0.00	0.00	

Figure D.3: Spearman correlation matrix. Source: Authors' calculations (size: 46 302 observations)

E Appendix E

The features are presented in the table below for individual level.

Selected Variables	Details	Type	Source
Age of client in 2019 (2019 - year of birth)		Numerical	Company
Annual premium	Integer with Value-Added Tax (bounded from 0 to 1000)	Numerical	Company
Average population of 25 years old by zip code town		Numerical	INSEE
Capital jewellery of the house (Indexed Value)	Euros	Numerical	Company
Capital valuables of the house (Indexed Value)	Euros	Numerical	Company
Cause of insurance claims	NA / Broken glass / Water leaks / Natural disasters / Fire / Civil liability / Theft	Categorical	Company
Change in guarantee level	No change / Change in higher level / Similar level / Change in lower level / NA	Categorical	Company

Client situation at the first policy signature	Maried / Single / Divorced / NA	Categorical	Company
Construction date of building	lt1950 /]1950 ,1970] /]1970 ,1986] /]1986 ,2000] /]2000, 2005] / More than 2005	Categorical	Company
Criminality about fire by department		Numerical	INSEE
Criminality about fraud by department		Numerical	INSEE
Delay to turn propects into customers	Days	Numerical	Company
Delay between the last amendment date of policy and the opening date of claim	Days	Numerical	Company
Delay between the occurrence date of the incident and the opening date of claim	Days	Numerical	Company
Divorce rate by zip code town		Numerical	INSEE
Fire-Prone Area	R to V (defined by company to indicate the fire risk regions)	Categorical	Company
Franchise of guarantee	Euros	Numerical	Company
Garden Package Level	Yes / No / NA	Categorical	Company
Ground floor	Yes / No	Categorical	Company
Immigrant rate in 2007 by zip code town		Numerical	INSEE
Income disparity (1st and 3rd quantile) by zip code town	Euro	Numerical	INSEE
Loss frequency at policy level		Numerical	Company
Nature of the insurance claim	Damage to the equipment / Personal injury	Categorical	Company
New energies (Home Package)	Yes / No / NA	Categorical	Company
Number of bedrooms	1 to 10	Numerical	Company
Number of guarantees	1 to 3	Numerical	Company
Number of people by household	1 to 10	Numerical	Company
Number of total policies	1 to 13	Numerical	Company
Opening of the claim shortly after the policy registration	Days	Numerical	Company
Initial value of capital	Euros (the valuables reimbursed at their initial purchase value in case of an incident)	Numerical	Company
Initial value of jewelery	Euros (the jeweleries reimbursed at their initial purchase value in case of an incident)	Numerical	Company
Payment mode of policy	Annual / Quarterly / Semi-annual / Monthly	Categorical	Company
Plan	Insurable risk / Submit to agreement / Non insurable risk	Categorical	Company
Plumbing package level	Yes / No / NA	Categorical	Company
Presence of package for household policy	Yes / No	Categorical	Company
Professional use of household	Yes / No / Childminder (1%-98%)	Categorical	Company
Protection level of household	A / B / C (for instance. the presence of an alarm system. etc.)	Categorical	Company

Quality of life	High / Low / Medium (defined by company based on revenues)	Categorical	Company
Quality of subscriber	Owner / Tenant	Categorical	Company
Rate of workers in 2007 by zip code town		Numerical	INSEE
Security package	Yes / No / NA	Categorical	Company
Gender	Man / Woman	Categorical	Company
Size of household	From 1 to 2 rooms / From 3 to 4 rooms / More than 4 rooms	Categorical	Company
Subscription of guarantee	Premium / Basic	Categorical	Company
Swimming pool package	Yes / No / NA	Categorical	Company
Theft-Prone Area	1 to 6 (defined by company to indicate the theft risk regions)	Numerical	Company
Type of house	House / Apartment	Categorical	Company
Type of insurance coverage	Basic / Premium / Student / NA	Categorical	Company
Use	Principal residence / Rental / Second home	Categorical	Company

Table E.1: Explanatory variables

F Appendix F

The features used for the group-level analysis are presented in the table below.

Selected Variables	Details	Type	Source
Age of client in 2019 (2019 - year of birth)	[18, 30[/ [30, 35[/ [36, 45[/ [46, 50[/ More than 50	Categorical	Company
Annual premium	Integer with Value-Added Tax (to 150 /]150, 300] /]300, 400] / More than 400)	Categorical	Company
Average population of 25 years old by zip code town	to 10 /]10, 13] /]13, 15] /]15, 18] / More than 18 (in %)	Categorical	INSEE
Capital jewelery of the house (Indexed Value)	to 700 / More than 700 (Euros)	Categorical	Company
Capital valuables of the house (Indexed Value)	to 12000 / More than 12000 (Euros)	Categorical	Company
Cause of insurance claims	NA / Broken glass / Water leaks / Natural disasters / Fire / Civil liability / Theft	Categorical	Company
Change in guarantee level	No change / Change in higher level / Similar level / Change in lower level / NA	Categorical	Company
Client situation at the first policy signature	Married / Single / Divorced / NA	Categorical	Company
Construction date of building	Before 1950 /]1950 ,1970] /]1970 ,1986] /]1986 ,2000] /]2000, 2005] / After 2005	Categorical	Company
Criminality about fire by department	to 0.02 /]0.02, 0.03] /]0.03, 0.04] /]0.04, 0.06] /]0.06, 0.075] / More than 0.075 (in %)	Numerical	INSEE
Criminality about fraud by department	to 0.15 /]0.15, 0.2] /]0.2, 0.3] /]0.3, 0.5] /]0.5, 0.7] / More than 0.7 (in %)	Categorical	INSEE
Delay to turn prospects into customers	to 1 /]1, 2] /]2, 3] /]3, 6] /]6, 12] / More than 12 (in month)	Categorical	Company

Delay between the last amendment date of policy and the opening date of claim	to 1 /]1, 2] /]2, 3] /]3, 6] /]6, 12] / More than 12 (in month)	Categorical	Company
Delay between the occurrence date of the incident and the opening date of claim	to 1 /]1, 2] /]2, 3] /]3, 6] /]6, 12] / More than 12 (in month)	Categorical	Company
Divorce rate by zip code town	to 7 /]7, 10] / More than 10 (in %)	Categorical	INSEE
Fire-Prone Area	R to V (defined by company to indicate the fire risk regions)	Categorical	Company
Franchise of guarantee	to 100 / More than 100 (Euros)	Categorical	Company
Garden Package Level	Yes / No / NA	Categorical	Company
Ground floor	Yes / No	Categorical	Company
Immigrant rate in 2007 by zip code town	to 3 /]3, 5] /]5, 8] /]8, 12] /]12, 16] / More than 16 (in %)	Categorical	INSEE
Income disparity (1st and 3rd quantile) by zip code town	to 11500 /]11500, 12500] /]12500, 13000] /]13000, 13500] /]13500, 14500] /]14500, 15000] /]15000, 16000] /]16000, 17500] /]17500, 19000] /]19000, 20000] / More than 20000 (Euro)	Categorical	INSEE
Loss frequency at policy level	to 0.1 /]0.1, 0.2] /]0.2, 0.5] /]8, 12] /]0.5, 1] / More than 1	Categorical	Company
Nature of the insurance claim	Damage to the equipment / Personal injury	Categorical	Company
New energies (Home Package)	Yes / No / NA	Categorical	Company
Number of bedrooms	1 / 2 / 3 / from 4 to 5 / More than 5	Categorical	Company
Number of guarantees	1 / More than 2	Categorical	Company
Number of people by household	to 2 / More than 2	Categorical	Company
Number of total policies	1 / 2 / More than 2	Categorical	Company
Opening of the claim shortly after the policy registration	to 1 /]1, 2] /]2, 3] /]3, 6] /]6, 12] / More than 12 (in months)	Categorical	Company
Initial value of capital	to 12000 / More than 12000 (Euros) (the valuables reimbursed at their initial purchase value in case of an incident)	Categorical	Company
Initial value of jewelery	to 700 / More than 700 (Euros) (the jeweleries reimbursed at their initial purchase value in case of an incident)	Categorical	Company
Payment mode of policy	Annual / Quarterly / Semi-annual / Monthly	Categorical	Company
Plan	Insurable risk / Submit to agreement / Non insurable risk	Categorical	Company
Plumbing package level	Yes / No / NA	Categorical	Company
Presence of package for household policy	Yes / No	Categorical	Company
Professional use of household	Yes / No / Childminder (1%-98%)	Categorical	Company
Protection level of household	A / B / C (for instance. the presence of an alarm system. etc.)	Categorical	Company
Quality of life	High / Low / Medium (defined by company based on revenues)	Categorical	Company
Quality of subscriber	Owner / Tenant	Categorical	Company
Rate of workers in 2007 by zip code town	to 9 /]9, 11] /]11, 13] /]13, 15] /]15, 19] / More than 19 (in %)	Categorical	INSEE

Security package	Yes / No / NA	Categorical	Company
Gender	Man / Woman	Categorical	Company
Size of household	From 1 to 2 rooms / From 3 to 4 rooms / More than 4 rooms	Categorical	Company
Subscription of guarantee	Premium / Basic	Categorical	Company
Swimming pool package	Yes / No / NA	Categorical	Company
Theft-Prone Area	1 to 6 (defined by company to indicate the theft risk regions)	Categorical	Company
Type of house	House / Apartment	Categorical	Company
Type of insurance coverage	Basic / Premium / Student / NA	Categorical	Company
Use	Principal residence / Rental / Second home	Categorical	Company

Table F.1: Explanatory variables

References

Alexandre, C. and Balsa, J. (2015). Client profiling for an anti-money laundering system. *arXiv preprint arXiv:1510.00878*.

Artıs, M., Ayuso, M., and Guillen, M. (1999). Modelling different types of automobile insurance fraud behaviour in the spanish market. *Insurance: Mathematics and Economics*, 24(1-2):67–81.

Artís, M., Ayuso, M., and Guillén, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, 69(3):325–340.

Atwood, J. A., Robison-Cox, J. F., and Shaik, S. (2006). Estimating the prevalence and cost of yield-switching fraud in the federal crop insurance program. *American journal of agricultural economics*, 88(2):365–381.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003a). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, 49(3):312–329.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003b). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635.

Baesens, B., Van Vlasselaer, V., and Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Bauman, S., Toomey, R. B., and Walker, J. L. (2013). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of adolescence*, 36(2):341–350.

Belhadji, E. and Dionne, G. (1997). Development of an expert system for the automatic fraud detection of automobile insurance fraud. *Canada, Montreal: Ecole des Hutes Etudes Commerciales*.

Belhadji, E. B., Dionne, G., and Tarkhani, F. (2000). A model for the detection of insurance fraud. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 25(4):517–538.

Bentley, P. J. (2000). " evolutionary, my dear watson" investigating committee-based evolution of fuzzy rules for the detection of suspicious insurance claims. In *Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation*, pages 702–709.

Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Branco, P., Torgo, L., and Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brockett, P. L., Xia, X., and Derrig, R. A. (1998). Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance*, pages 245–274.

Brownlee, J. (2020). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery.

Caudill, S. B., Ayuso, M., and Guillén, M. (2005). Fraud detection using a multinomial logit model with missing information. *Journal of Risk and Insurance*, 72(4):539–550.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences, 2nd edn.(hillsdale, nj: L. erlbaum associates).

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Derrig, R. A. and Ostaszewski, K. M. (1995). Fuzzy techniques of pattern recognition in risk and claim classification. *Journal of Risk and Insurance*, pages 447–482.

Dhieb, N., Ghazzai, H., Besbes, H., and Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access*, 8:58546–58558.

Dionne, G., Giuliano, F., and Picard, P. (2009). Optimal auditing with scoring: Theory and application to insurance fraud. *Management Science*, 55(1):58–70.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gomes, C., Jin, Z., and Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, 88(3):591–624.

Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., and Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1):292–305.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

He, H. and Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Islam, M. S., Eva, S. A., and Hossain, M. Z. (2017). Predicate offences of money laundering and anti money laundering practices in bangladesh among south asian countries. *Studies in Business and Economics*, 12(3):63–75.

Jin, Y., Rejesus*, R. M., and Little, B. B. (2005). Binary choice models for rare events data: a crop insurance fraud application. *Applied Economics*, 37(7):841–848.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Kirlidog, M. and Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, 62:989–994.

Kowshalya, G. and Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 1338–1343. IEEE.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):191–201.

Leukfeldt, E. R., Kleemans, E. R., and Stol, W. P. (2017). Cybercriminal networks, social ties and online forums: social ties versus digital ties within phishing and malware networks. *The British Journal of Criminology*, 57(3):704–722.

Levi, M. (2015). Money for crime and money from crime: Financing crime and laundering crime proceeds. *European Journal on Criminal Policy and Research*, 21(2):275–297.

Li, H., Song, Q., and Su, J. (2021). Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. *Journal of Risk and Insurance*, 88(3):625–663.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Luo, C., Wu, D., and Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65:465–470.

Major, J. A. and Riedinger, D. R. (1992). Efd: A hybrid knowledge/statistical-based system for the detection of fraud. *International Journal of Intelligent Systems*, 7(7):687–703.

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Nakagawa, S. (2004). A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology*, 15(6):1044–1045.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, 50(3):559–569.

Nian, K., Zhang, H., Tayal, A., Coleman, T., and Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, 2(1):58–75.

Quah, J. T. and Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert systems with applications*, 35(4):1721–1732.

Rana, P. J. and Baria, J. (2015). A survey on fraud detection techniques in ecommerce. *International Journal of Computer Applications*, 113(14).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Sánchez, D., Vila, M., Cerda, L., and Serrano, J.-M. (2009). Association rules applied to credit card fraud detection. *Expert systems with applications*, 36(2):3630–3640.

Sullivan, H., Christophe, H., Christophe, P., and Sébastien, S. (2022). Explainable performance. *arXiv preprint arXiv:2212.05866*.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719.

Tumminello, M., Consiglio, A., Vassallo, P., Cesari, R., and Farabullini, F. (2022). Insurance fraud detection: A statistically validated network approach. *Journal of Risk and Insurance*.

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., and Baesens, B. (2015). Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, 75:38–48.

Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., and Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, 176(1):565–583.

Viaene, S., Derrig, R. A., Baesens, B., and Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, 69(3):373–421.

Von Altrock, C. (1996). *Fuzzy logic and neurofuzzy applications in business and finance*. Prentice-Hall, Inc.

Weisberg, H. I. and Derrig, R. A. (1998). Quantitative methods for detecting fraudulent automobile bodily injury claims. *Risques*, 35(July–September):75–99.

West, J. and Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. *Computers & security*, 57:47–66.

Xiaoyun, W. and Danyue, L. (2010). Hybrid outlier mining algorithm based evaluation of client moral risk in insurance company. In *2010 2nd IEEE International Conference on Information Management and Engineering*, pages 585–589. IEEE.

Yang, W.-S. and Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31(1):56–68.

Yankol-Schalck, M. (2022). The value of cross-data set analysis for automobile insurance fraud detection. *Research in International Business and Finance*, 63:101769.