Machine Learning with Applications 5 (2021) 100074

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata

Matheus Kempa Severino [a], Yaohao Peng [b,*]

[a] University of Brasilia, Campus Darcy Ribeiro, Brasilia, Distrito Federal, 70910–900, Brazil
[b] Brazilian Ministry of Economy, Brasilia, Distrito Federal, 70048–900, Brazil

ARTICLE INFO

Keywords:
Fraud detection
Insurance market
Risk management
Decision support systems
Supervised learning
Feature importance

ABSTRACT

This paper evaluated fraud prediction in property insurance claims using various machine learning models based on real-world data from a major Brazilian insurance company. The models were tested recursively and average predictive results were compared controlling for false positives and false negatives. The results showed that ensemble-based methods (random forest and gradient boosting) and deep neural networks yielded the best results, exhibiting superior average performance in comparison to the other classifiers, including the commonly used logistic regression. In addition, we compiled a general profile of confirmed fraudsters from the dataset and estimated the impact of each feature in the global classification performance and for prominent cases of false positive and false negative predictions using eXplainable Artificial Intelligence methods. The findings of this study can aid risk analysts and professionals in assessing the strengths and weaknesses of each model and to build empirically effective decision rules to evaluate future insurance policies.

1. Introduction

The insurance market is a highly profitable market that moves large sums of money over the years. In Brazil alone, about 10.8 billion USD was paid in insurance policies in 2017 (Brazilian National Confederation of Insurance Companies, 2017). Similarly, frauds can bring huge losses to the companies: In the same year of 2017, the total value of all occurred claims was around 10.0 billion USD, while the value of proven frauds totaled 221.2 million USD (Brazilian National Confederation of Insurance Companies, 2017).

Bearing in mind the economic relevance of this market and the challenge of fraud detection by professional analysts, the search for data mining and machine learning techniques had been showing its predicting potential in financial applications, as seen in works like Hsu et al. (2016), notably when involving complex problems and non-linear patterns (Huang et al., 2004; Soman et al., 2009). In a study that applied several machine learning models to predict the default rate of a government-funded housing finance program, de Castro Vieira et al. (2019) reported that the practical application of the proposed method would have significantly reduced the number of conceded non-performing loans and avoided approximately 3.0 billion USD from credit losses. Specifically for fraud detection, machine learning applications include Awoyemi et al. (2017), Chen et al. (2006), Hajek and Henriques (2017) and Raghavan and El Gayar (2019).

Therefore, this article aimed to verify whether the use of various machine learning models – ranging from regularized extensions of simple models to different structures of non-linear interactions and ensemble-based classifiers – can contribute to fraud identification for property insurance policies, comparing the performance of these models with the standard logistic regression. Furthermore, in this paper, we compiled an overall profile for confirmed fraudsters and analyzed the relative importance of the input variables according to each model using eXplainable Artificial Intelligence (XAI) methods, addressing as well the interpretation of the predictions for prominent false positive and false negative observations. These results were discussed in terms of their practical applicability to effectively aid risk management professionals in building data-driven decision rules based on the most prominent "signs" for potential frauds in future policies.

Moreover, this paper used real-world data at the individual level from a major Brazilian insurance company, containing information about the income level of the clients, as well as features like the time between contract start and claim and the number of past policy claims. In this sense, the use of real data represents a significant advantage over simulated data in terms of both model evaluation and practical applicability in decision-making. In addition, since scientific papers that analyze frauds in residential and business property insurance are relatively scarce in comparison to works about fraud detection in automobile insurance of credit card transactions, this paper further contributes to the current literature on machine learning applications.

* Corresponding author.
E-mail addresses: matheus.kempa@lamfo.unb.br (M.K. Severino), peng.yaohao@economia.gov.br (Y. Peng).
URL: https://www.lamfo.unb.br (M.K. Severino).

https://doi.org/10.1016/j.mlwa.2021.100074
Received 31 July 2020; Received in revised form 11 June 2021; Accepted 14 June 2021
Available online 22 June 2021

This article is structured as follows: Section 2 provides a review of the recent literature on insurance fraud prediction, with an emphasis on studies that applied machine learning techniques; Section 3 describes the explanatory variables, the data processing procedure and the steps of the empirical experiments, as well as metrics to evaluate the forecasts; Section 4 presents the results of the forecasts and their statistical significances, alongside a general profile of the fraud cases and model-agnostic methods of feature importance evaluation for global and local interpretation using eXplainable Artificial Intelligence, discussing the findings of this paper on practical risk analysis and fraud detection; finally, Section 5 discusses the limitations of this paper and points out suggestions for possible future developments.

2. Related literature

Machine learning methods are based on an inductive analytical paradigm, drawing conclusions based on the patterns observed from data without defining assumptions like probability distributions and functional form, such as linearity. As discussed in Peng and Nagata (2020), this flexibility demands additional caution to control the models' balance between generalization ability and complexity, since different models or even small variations on their hyperparameters can induce great impacts on the predicting performance while being applied to the same dataset.

One of the main applications of machine learning in business administration and finance is fraud detection, a topic of great relevance from the perspective of a decision-maker, given the fact that decision support systems that help risk analysts predict fraudsters have a direct impact on the economic performance of a company. In this sense, this topic has been investigated by a big number of researchers in recent years, as discussed in papers like Awoyemi et al. (2017),Ngai et al. (2011), Raghavan and El Gayar (2019) and Waghade and Karandikar (2018).

For instance, Triepels et al. (2018) proposed an automated system to detect frauds in shipping documents, which can be adulterated to overpass restrictions or to facilitate smuggling. The authors developed a model based on Bayesian networks to generate probabilistic discriminative models and predict the presence of goods on the shipments' cargo list, and then crossed with the documentation to determine whether a fraud is configured. The results showed that proposed automated systems considerably improved the detection of miscoding and smuggling compared to random audits, which are typically used by shipping companies to check these documentations, usually in a labor-intensive and non-scalable way.

Similarly, Dou et al. (2019) analyzed download frauds in Mobile App Markets, categorizing the frauds into three main classes based on their motivation: (1) boosting an App's front end downloads, (2) optimizing an App's search ranking, and (3) enhancing an App's user acquisition and retention rates. The authors applied the XGBoost model to predict frauds using different sets of features, reaching over 99% of accuracy in the most general case. In addition, the authors evaluated the predictions using performance metrics that consider the overall balance between false negatives and false positives, as well as generating a ranking of the features' importance for this predicting task. Both aspects were considered in this paper's empirical analysis as well.

Bearing in mind that the presence of fraud implies large profit losses for the insurance sector, Sheshasaayee and Thomas (2018) illustrates the main challenges of risk and fraud analysts to develop fraud identification mechanisms and decision rules, discussing the advantages of using Machine Learning methods to perform such tasks, especially concerning the most prominent features of fraudsters. In this line, Popat and Chaudhary (2018) listed recent researches about credit card fraud detection with machine learning models, discussing the strengths of this paradigm on mining patterns from high-dimensional data and assisting real-world decision-making.

Likewise, Dal Pozzolo et al. (2014) discussed the complexity involved in the development of a data-driven fraud detection algorithm, highlighting common issues like the non-stationary distribution of the data, highly imbalanced classes distributions, a continuous and massive flow of new transactions, and scarcity of available microdata due to confidentiality issues. Regarding these issues, the authors evaluated the predictive performance of three machine learning models (random forest, support vector machine, and neural network) using a real-world credit card dataset, as well as the overall impact of update periodicity, application of balancing techniques, and retainment of older observations in the training dataset. The results indicated that the random forest model performed consistently better than Support Vector Machine and neural networks for all training approaches; moreover, models that were updated with new data more often performed better, which indicates that the fraud distribution can quickly change over time. Concerning the issue of unbalanced classes, the application of balancing methods improved the performance over the "static" non-balanced dataset, in which the random forest yielded the worst performance. Finally, the procedure of discarding older observations exhibited a smaller marginal improvement in comparison to maintaining the dataset balanced.

Waghade and Karandikar (2018) used machine learning models to predict frauds in the healthcare sector, pointing out the costs of the manual identification process – which requires a long effort for reviewing auditors to evaluate medical insurance claims are fraudulent – and discussing the relevance of automated decision support systems for different types of fraud in this business branch. Similarly, Verma et al. (2017) applied outlier detection models to identify anomalies and potential frauds in healthcare systems.

Wang and Xu (2018) applied machine learning-based text mining algorithms to analyze the descriptions of car accidents in order to predict frauds for automobile insurance claims: the tested models were support vector machine (SVM), random forest, and deep neural network, and all three models managed to reach an F1 Score greater than 75%. Roy and George (2017), on the other hand, applied random forest and naive Bayes models to detect fraud in automobile claims, finding that the former performed better than the latter. Likewise, Yao et al. (2018) proposed a model to detect financial fraud combining feature selection and machine learning classification models. Starting from high-dimensional data, Principal Component Analysis and XGBoost were used to identify the most informative variables, after which several machine learning models were applied, amongst which the random forest had the best out-of-sample performance.

Eshghi and Kargari (2019), on the other hand, stated that unsupervised methods like clustering and outlier detection techniques may not suffice for complex fraud detection tasks, and proposed a framework with Multi-Criteria Decision Analysis and intuitionistic fuzzy sets to incorporate the effect of behavioral uncertainties to model the propensity of a banking transaction to be a fraud. Similarly, Carcillo et al. (2019) stated in favor of integrating unsupervised and supervised learning techniques for credit card fraud detection, in order to better adapt to changes in customer behavior and fraudsters' ability to invent novel fraud patterns. The authors computed outlier scores for different levels of granularity based on clustering analysis, subsequently applying them on a real-world dataset and reporting an accuracy improvement on the detection performance.

Jurgovsky et al. (2018) presented a sequential learning approach to fraud detection in credit card transactions using LSTM recurrent neural networks, comparing it with a Random Forest classifier as a static benchmark. Using a real-world dataset and analyzing independently offline and e-commerce transactions, the authors found out that the frauds detected by the two learners were consistently different, which suggests the potential for the development of ensemble-based models that incorporate both approaches. In addition, the performance of both the static and the sequence learners benefited from manual feature aggregations, evidencing the importance of modeling aspects and feature engineering in fraud detection. Other recent studies on credit card fraud detection include Kim et al. (2016), which proposed a multi-class

algorithm to detect fraud intention in financial misstatements applying MetaCost (Domingos, 1999) to deal incorporate asymmetric misclassification costs to control for the classes' unbalance; and Varmedja et al. (2019), which applied Logistic Regression, Random Forest, Naive Bayes, and Neural Network as machine learning classifiers, combined with SMOTE (Synthetic Minority Oversampling Technique) to balance the training data.

For the prediction of corporate bankruptcy, Chen et al. (2020) combined two ensemble methods (namely bagging and boosting) with Support Vector Machines, using a scheme that assigns labels for unlabeled training data controlling for the bag-level relative proportion between the classes — this approach was shown to be efficient in terms of both data-labeling for large datasets and prediction performance improvement through the introduction of ensemble learning strategies; Nami and Shajari (2018), on the other hand, developed a method that involves two stages of detecting fraudulent payment card transactions, applying dynamic random forest and k-nearest neighbors as machine learning models: based on the primary data, additional transaction features are derived and a greater weight is assigned to most recent transactions, considering that the most recent behavior of credit card holders tend to have a larger impact on deciding whether a transaction is fraudulent or legitimate.

In an attempt to learn Complex Event Processing rules to extract relevant information from big-scale data streams, Bruns et al. (2019) proposed a model based on genetic algorithms, discussing as well heuristics about the choice of suitable parameters for the process. The empirical validation of this model was performed using real-world transportation data, which allowed the evaluation of the merits and weaknesses of the approach should it be applied in a real-world decision-making context. For a fraud detection application in finance, Eweoya et al. (2019), in turn, used real-world data from a financial institution and applied decision trees to predict frauds in bank loan administration and consequently diminish losses due to loan defaults.

Thus, an application using real-world data and machine learning methods is pertinent to investigate which machine learning models can better identify fraud patterns and accurately predict future felonies. Besides, given the great variety of possible frauds, with each category having its specificities and *modus operandi* (Gottschalk, 2010), in this paper we focused on frauds in which the consumer is the author of the fraud, delimited to policy claims over residential assets of individuals and firms, a relatively less explored segment in comparison to automobile or credit card insurance. Moreover, the data used in this paper were collected from a major insurance company, which provides additional insights to our conclusions regarding the relative importance of the database features, which can be very useful for real-world insurance policy evaluations.

As reported by the review papers of Ngai et al. (2011) and Sinayobye et al. (2018), the majority of the studies that apply machine learning to fraud detection problems focused on credit card and telecommunication frauds, while the applications for insurance frauds are mostly concentrated on healthcare and automobile applications, with few studies that tackle frauds for property insurances, especially for residential policies. In this sense, as mentioned at the end of the introduction, this paper contributes to the literature by testing the empirical strengths and weaknesses of various well-known machine learning models using real-world microdata for a relatively less explored insurance segment, potentially aiding market professionals and decision-makers on their respective model choices for similar tasks.

3. Empirical analysis

3.1. Data overview and preprocessing

We collected data from 2009 to 2018 of registered claims for residential and business insurance policies from one of Brazil's largest insurance companies. The labels for every policy – *i.e.:* whether the claim was a fraud – were assigned by human experts from the company's risk analysis sector. In this sense, we also included a few cases of detected frauds but had not yet been proven in court in the class "fraud" — we justify this decision based on a preemptive approach for risk management, as we find it important not to leave any real frauds (false negatives) out in a preliminary screening stage, which is the stage this study brings its main contributions. Our understanding is that refining a smaller set of most likely frauds for a posterior human inspection is an optimal strategy for fraud detection, and we evaluated the predictions using metrics that penalize false positives and false negatives bearing this in mind as well. The database used in this paper also contains information concerning the person involved in the claim, such as age, gender, wage level, timestamp of actions taken, past fraud occurrences, etc., adding extra value to the conclusions of this study. Fields that allow personal identification were accordingly suppressed.

Since most operations are not frauds, the dataset would be unbalanced if the time periods were the same for both frauds and non-frauds. In this sense, for this paper, we opted to collect all fraud claims registered between 2009 and 2018 and all non-fraud claims from 2015 onwards to keep the dataset roughly balanced, with a similar number of observations for the two labeled classes, totaling 851 observations. We opted for this treatment assuming that the overall fraudster profile did not change structurally in the observed years — based on the authors' practical experience in fraud analysis in the company which provided the microdata, this is a standard procedure for property fraud detection. Hence, whilst the literature has proposed many data mining methods to deal with imbalanced datasets and prediction of rare events, as presented by Haixiang et al. (2017), we did not apply oversampling techniques, such that we balanced the dataset by taking a longer period for confirmed frauds instead. Moreover, while balancing is not mandatory for any of the analyzed models, it helps to better interpret the predictions' accuracy, a metric that could yield high values due to non-informative classification in very unbalanced datasets: for instance, predicting only "non-frauds" for a dataset with 5% of frauds would lead to a 95% accuracy without generating any value in terms of fraud detection.

The description and motivation of each feature contained in our dataset are summarized below:

- **Product Type:** There were 3 classes for product type in our dataset: "Residential", and "Residential exclusive", for natural persons; and "Business", exclusive for legal persons;
- **Coverage type**: Coverage type is the protection granted by the insurance, each different type has its peculiarities regarding the operational procedure and analysis process for the claims. We summarized the coverage types into 6 classes: "electrical damage", "theft", "storm", "glass break", "fire/lightning/explosion", and "others";
- **Contract channel**: refers to the channel that the client used to contract his/her policy, with 3 possible classes: "physically" (at the counter), "online system" or "remote channel";
- **Automatic renewal**: Indicates whether the customer has opted to include a clause to automatically renew the policy after its expiration date. This field is important, as customers with intentions to commit frauds tend to not hire an insurance policy to renew it afterward;
- **Past renewal**: Indicates whether the policy is a new one or a past one renewed. For the company, a customer who renewed his/her policy indicates less risk of frauds, since the renewal approval depends on some procedural analyses;
- **Legal person**: Indicates whether the client is a natural or legal person;
- **Number of payment installments of the insurance value**;
- **Time of approval of the insurance policy**;

Table 1
Observed ranges for the numeric features, grouped by frauds and non-frauds.

Variable	Frauds		Non-Frauds	
	Min	Max	Min	Max
Days between contract end and claim	1	1095	2	1094
Days between contract start and claim	1	1924	4	1099
Time of approval	0	49	0	29
Insured amount	70700	2530000	30000	2000000
Insurance premium	101.36	5864.90	94.38	12848.10
Number of installments	1	30	1	36
Age	19.11	89.32	20.09	87.19
Previous claims	0	11	0	13

- **Differences of Days between contract term start/end and claim date**: These are considered to be relevant variables for many analysts — claims that take effect in less than 60 days after the policy starts are called "premature claims", and they often indicate high propensity of fraud;
- **Insured amount**;
- **Insurance premium**;
- **Age**;
- **Gender**;
- **Income range**;
- **Marital status**;
- **Number of previous claims of the customer in the company**.

Before proceeding to the empirical experiments, all variables were firstly converted to numerical format: for the binary variables the conversion was straightforward; for the ordinal variables "product type", "coverage type" and "contract channel", one-hot encoding was applied, with one class for each variable being dropped to avoid perfect multicollinearity — hence, the classes "product type: residential exclusive", "coverage type: others" and "contract channel: remote channel" were discarded in order to prevent the dummy variable trap.

After being converting to numerical format, all variables were centered to zero-mean and scaled to unit-variance. In order to provide a quick descriptive assessment for the data, we displayed in Table 1 the ranges for the raw observed values (before centering and scaling) of the explanatory variables that were numerical before conversion, grouped by both frauds and non-frauds:

3.2. Methods and experiment procedure

In Ngai et al. (2011)'s survey paper, the logistic regression was the most common model for insurance fraud researches using data mining techniques; in recent papers like Varmedja et al. (2019) and Yao et al. (2018), this model is still widely used due to its easy implementation and interpretation. However, as discussed in Hsu et al. (2016), a large number of recent studies have shown that non-parametric models based on machines have shown better empirical performance in several classification problems (such as stock market forecasting, portfolio allocation, and asset pricing) in comparison with classic econometric techniques. In this sense, we tested the empirical classification performance of nine models, namely: (1) Standard logistic regression; (2) Logistic regression with elastic-net regularization; (3) Naive Bayes; (4) K-Nearest Neighbors (KNN); (5) Support Vector Machine with Polynomial Kernel; (6) Support Vector Machine with Gaussian Kernel; (7) Deep Neural Network; (8) Random Forest; and (9) Gradient Boosting Machine. Based on the related works discussed in Section 2, the main references that applied each aforementioned model in fraud detection tasks is summarized in Table 2 below:

For the empirical experiments, we chose to randomly take 200 observations as the in-sample training set, while the remaining 651 cases were allocated as out-of-sample data. We opted to use the smaller set as training data to better simulate a real-world decision-making scenario in which fraud data are scarce. For each training set, we applied 10-fold cross-validation to search for the best combination of hyperparameters for each model and applied the decision function with the optimal hyperparameters on the remaining 651 observations. The neural network model followed a feedforward architecture trained using a stochastic gradient descent algorithm proposed by Niu et al. (2011), with dropout layers after each of the dense layers before the output and the Rectified Linear Unit (ReLU) as activation functions. The hyperparameters that were tuned for each model during the validation step and their respective grid-search ranges are summarized in Table 3:

Furthermore, in order to strengthen the robustness of our results, we repeated the training-validation-test procedure 1000 times, randomly selecting different observations for the training and test sets on each evaluation round. The average out-of-sample prediction results and their standard deviations are summarized in Table 4. Finally, we applied model-agnostic eXplainable Artificial Intelligence methods to estimate the relative importance of each feature using a permutation-based approach and to perform a local analysis on prominent observations frequently classified as false positive and false negative for the best-performing models using Shapley Additive Explanation, discussing the results in light of the usual practices of human analysts and the applicability of our experiments in real-life risk analysis and fraud detection.

3.3. Evaluation of the results

In classification problems, the most commonly used evaluation metric is accuracy (Henrique et al., 2019). However, this metric disregards potential imbalances in relation to false positives or false negatives — for example, on an unbalanced basis with 95 non-frauds and 5 frauds, a classifier would obtain the apparently "high" accuracy of 95% for simply classifying all observations as non-fraud, which in practical facts would not imply any real action. In this sense, the literature has proposed many other evaluation metrics for machine learning models in terms of error and performance fitness, as summarized in Naser and Alavi (2020). Thus, in this paper, we evaluated the quality of the forecasts using five other metrics besides accuracy, as listed below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = \frac{2}{(1/Recall)+(1/Precision)}$$

$$\text{Kappa} = \frac{Accuracy - p_e}{1-p_e}$$

$$\text{MCC} = \frac{(TP\cdot TN)-(FP\cdot FN)}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$$

where TP are the true positives (frauds predicted as frauds), FP are false positives (non-frauds predicted as frauds), TN are the true negatives (non-frauds predicted as non-frauds), FN are false negatives (frauds predicted as non-frauds), and $p_e = \left(\frac{[(TP+FP)\cdot(TP+FN)]+[(TN+FP)\cdot(TN+FN)]}{(TP+FN+TN+FP)^2}\right)$ is the probability of random agreement between the predictions and the actual classes.

The precision metric penalizes false positives (Type I error), allowing to evaluate the percentage adjustment only for those predicted as fraud; similarly, the recall penalizes false (Type II error), thus providing the relative frequency of a fraud to be identified in the universe of claims that were in fact fraud. Finally, the F1 Score is the harmonic mean between precision and recall, therefore being a conservative measure, in the sense of presenting a high value only when both accuracy and recall are high — that is, the F1 Score reflects the quality of classifications penalizing both type I and type II errors. In the context of fraud detection algorithms, this issue is further underlined in the work of Kim et al. (2019), which compared deep learning model and a hybrid ensemble model that aggregates decision tree, logistic

Table 2
Main references of applications of machine learning techniques in fraud detection.

Model	References
Logistic Regression	Caudill et al. (2005), Viaene et al. (2002), Yeh and Lien (2009) Awoyemi et al. (2017), Varmedja et al. (2019), Yao et al. (2018)
Naive Bayes	Awoyemi et al. (2017), Viaene et al. (2002), Yeh and Lien (2009) Roy and George (2017), Varmedja et al. (2019)
KNN	Awoyemi et al. (2017), Viaene et al. (2002), Yeh and Lien (2009) Nami and Shajari (2018), Raghavan and El Gayar (2019)
SVM	Chen et al. (2006), Dal Pozzolo et al. (2014), Viaene et al. (2002) Raghavan and El Gayar (2019), Wang and Xu (2018), Yao et al. (2018) Chen et al. (2020)
Neural Networks	Dal Pozzolo et al. (2014), Viaene et al. (2002), Yeh and Lien (2009) Jurgovsky et al. (2018), Wang and Xu (2018), Yao et al. (2018) Raghavan and El Gayar (2019), Varmedja et al. (2019)
Random Forest	Dal Pozzolo et al. (2014), Nami and Shajari (2018), Roy and George (2017) Jurgovsky et al. (2018), Wang and Xu (2018), Yao et al. (2018) Raghavan and El Gayar (2019), Varmedja et al. (2019)
GBM	Dou et al. (2019), Gupta et al. (2019), Majhi (2019) Dhieb et al. (2019), Taha and Malebary (2020)

Table 3
Grid-search intervals of the hyperparameters.

Model	Hyperparameter	Interval
Logistic Regression	No hyperparameters	
Penalized Logistic Regression	Elastic-net regularization weight	$\{0.1, 0.2, \ldots, 0.8, 0.9\}$
Naive Bayes	Laplace correction factor	$\{0, 0.1, \ldots, 0.9, 1\}$
KNN	Number of neighbors	$\{1, 3, \ldots, 13, 15\}$
Polynomial Kernel SVM	Polynomial degree	$\{2, 3, 4\}$
	Misclassification cost	$\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$
	Tolerance band for the ε-insensitive loss function	$\{0, 0.05, \ldots, 0.95, 1\}$
	Bias term for the Kernel function	$\{0, 0.1, \ldots, 1.9, 2\}$
Gaussian Kernel SVM	Misclassification cost	$\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$
	Inverse bandwidth for the Kernel function	$\{0, 0.05, \ldots, 1.95, 2\}$
Deep Neural Network	Number of hidden layers	$\{3, 5, 7\}$
	Learning rate	$\{0.1, 0.2, \ldots, 0.8, 0.9\}$
	Input layer dropout ratio	$\{0, 0.1, 0.2, 0.3\}$
	Hidden layer dropout ratio	$\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$
Random Forest	Number of trees	$\{300, 400, \ldots, 900, 1000\}$
	Number of sampled features at each split	$\{2, 3, \ldots, 9, 10\}$
GBM	Learning rate	$\{0.1, 0.2, \ldots, 0.8, 0.9\}$
	Maximum depth of each tree	$\{3, 4, 5, 6, 7, 8, 9\}$
	Minimum loss reduction for splits	$\{0.1, 0.2, 0.3, 0.4, 0.5\}$

Table 4
Mean and standard deviation of performance metrics for 1000 rounds of out-of-sample forecasts.

Model	Accuracy	Precision	Recall	F1 Score	Kappa	MCC
Logistic Regression	80.67% (1.60%)	80.56% (2.94%)	78.99% (3.79%)	79.67% (1.81%)	61.26% (3.21%)	61.41% (3.18%)
Penalized Logistic Regression	81.40% (1.72%)	81.27% (3.36%)	79.95% (3.99%)	80.48% (1.88%)	61.34% (3.19%)	62.93% (3.36%)
Naive Bayes	71.16% (5.66%)	73.18% (8.64%)	73.02% (5.53%)	72.39% (3.72%)	47.66% (10.28%)	49.51% (8.78%)
KNN	75.74% (2.51%)	77.74% (3.76%)	69.77% (4.67%)	73.39% (2.88%)	51.25% (5.01%)	51.66% (4.98%)
Polynomial Kernel SVM	81.34% (0.75%)	79.22% (1.15%)	82.98% (1.06%)	80.93% (0.84%)	62.92% (1.48%)	63.00% (1.48%)
Gaussian Kernel SVM	79.56% (1.71%)	79.07% (3.08%)	78.41% (4.17%)	78.53% (1.98%)	58.81% (3.36%)	59.00% (3.31%)
Deep Neural Network	81.88% (1.58%)	78.41% (3.11%)	86.28% (3.06%)	82.06% (1.32%)	63.84% (3.08%)	64.32% (2.80%)
Random Forest	84.56% (1.43%)	84.72% (2.60%)	82.77% (3.72%)	83.61% (1.65%)	69.05% (2.97%)	69.24% (2.88%)
GBM	83.21% (1.61%)	83.55% (2.97%)	81.73% (3.96%)	82.44% (1.82%)	66.20% (3.08%)	66.39% (3.02%)

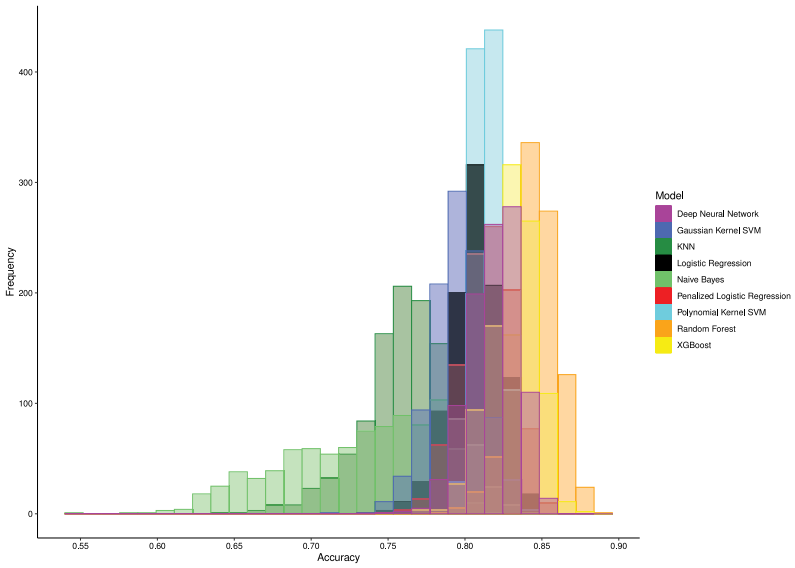

Fig. 1. Out-of-sample accuracy histogram for the analyzed models.

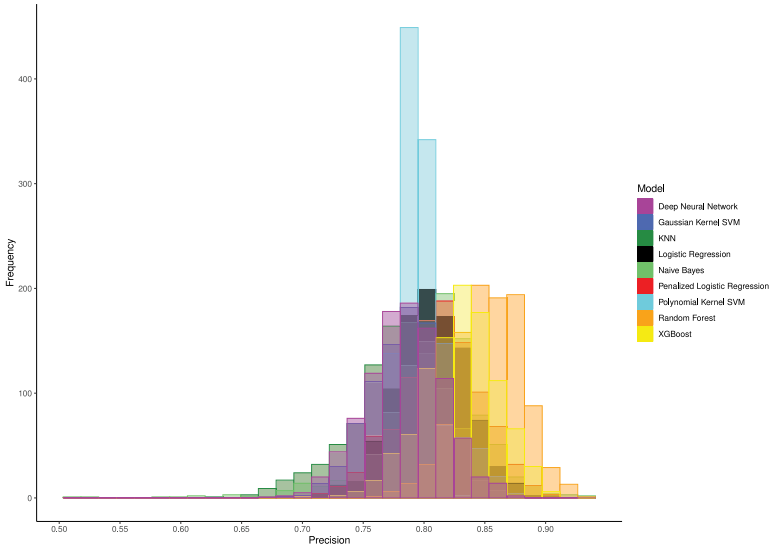

Fig. 2. Out-of-sample precision histogram for the analyzed models.

regression, and neural network for credit card fraud detection; in their research, the authors also emphasized the importance of bearing in mind the different costs associated with false alarms (false positives) and missed frauds (false negatives).

To further evaluate the reliability of the results and controlling by the possibility of the models having predicted outcomes correctly by chance, we also calculated the Cohen's Kappa coefficient (Cohen, 1960) of the models after each round of training-validation-test, as well as the Matthews correlation coefficient — MCC (Matthews, 1975), also known as "Pearson's phi coefficient", which is a scaled version of the test statistic for Pearson's chi-squared test on a 2×2 contingency table. As pointed out in Chicco and Jurman (2020), MCC tends to be more informative than metrics like accuracy and F1 Score because it takes into account the balance ratios of the four confusion matrix categories; although our dataset was a balanced one, we calculated the MCCs for each model to further verify their empirical performance. The average values for each evaluation metric and their standard deviations are displayed in Table 4 and plotted in Figs. 1to 6.

4. Results and discussion

4.1. Performance metrics and statistical significance

At a descriptive level, we first summarized a macro-profile of the 409 cases of fraud:

1. 60.14% of the fraudsters were male;
2. 48.16% of the frauds were premature claims;
3. 52.81% of the fraudsters were non-married;
4. 79.95% of the total coverage amount was for electrical damage or theft claims;
5. The average age of fraudsters was 41 years;
6. 72.61% of the frauds were new insurance policies;
7. Fire/lightning/explosion coverage had the highest average payment amount for detected but unproven frauds.

The results of the predictions provided by the machine learning algorithms are summarized in Table 4 below:

The results indicate that the standard logistic regression showed a middling overall performance in comparison to the other models,

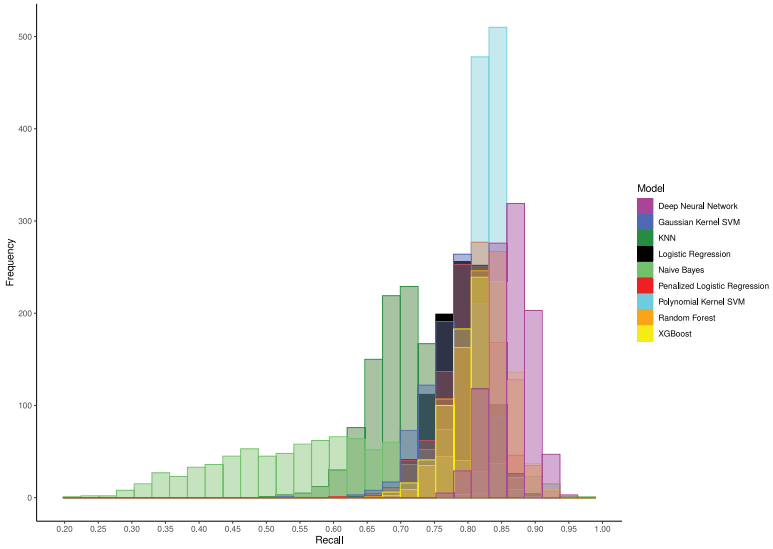

Fig. 3. Out-of-sample recall histogram for the analyzed models.

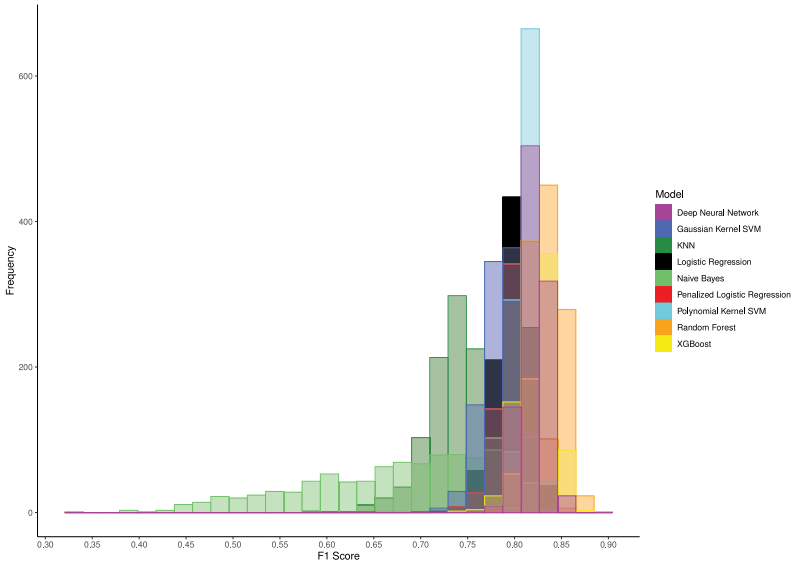

Fig. 4. Out-of-sample F1 Score histogram for the analyzed models.

being outperformed by the penalized logistic regression and the two ensemble-based machine learning models (random forest and gradient boosting) for all six evaluation metrics. On the other hand, naive Bayes and KNN showed the worse results, with their low recall values suggesting a high presence of false negatives, consequently hindering the F1 Score as well — from a risk manager's perspective, this implies in a large proportion of frauds that went unnoticed, which in turn means that the company would be bearing losses in favor of fraudsters. Concerning the Gaussian Kernel SVM, which is able to theoretically generalize nonlinear interactions with arbitrarily high dimensionality, its out-of-sample was actually worse than the logistic regression, which is a sign of overfitting of this model, similar to the reported in Peng and Nagata (2020).

A visual synthesis of Table 4 is given by Figs. 1to 6, which give away the histograms of all tested models for each performance metric over the 1000 rounds of training-validation-test. In overall terms, it can be observed that the distributions for all 6 metrics of KNN (in dark green) and naive Bayes (in light green) were shifted to the left in comparison to the others, while the values for Deep Neural Networks (in magenta), GBM (in yellow) and random forest (in orange) were mostly concentrated on larger values for all metrics, with the standard logistic regression (in black) staying in the middle-ground. It can also be noted that the polynomial Kernel SVM (in cyan) had the lowest variance, and naive Bayes had the greatest variance, causing a heavy tail to the left in its histogram.

As seen in Fig. 2, random forest and GBM had the best out-of-sample values for precision, indicating that those models had a small amount of false positive predictions; conversely, the model with the best performance for the recall was the neural network, as illustrated by Fig. 3, indicating that this model performed better with regard to avoiding false negatives. When jointly evaluating the two types of error using F1 Score, Cohen's Kappa and MCC, GBM, and random forest stood out as the best models for our experiments; nonetheless, since a false negative (failing to predict an actual fraud) often lead to larger financial losses than false positives, our experiments point out that the deep neural network approach is also recommended as a prominent model to support risk analysts and decision-makers.

Moreover, in order to statistically evaluate which of the tested models had the best predicting performance across the experiments, we applied Hansen et al. (2011)'s Model Confidence Set procedure (MCS).

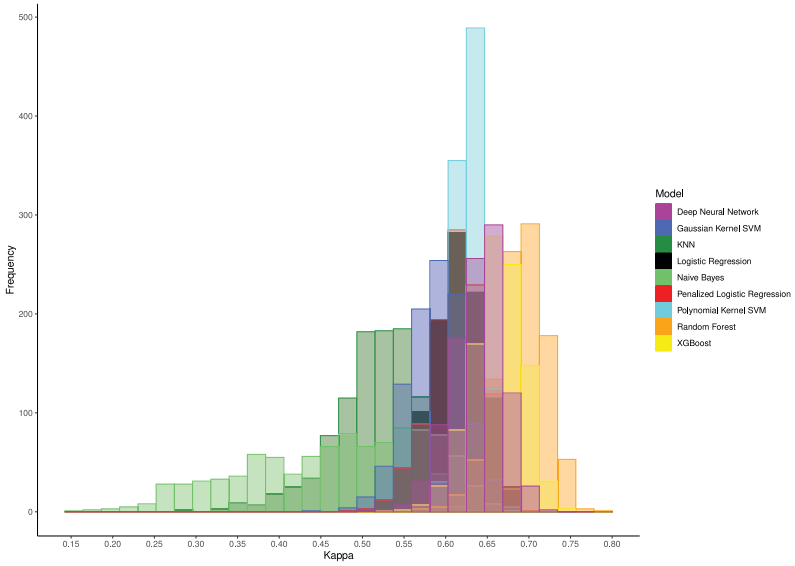

Fig. 5. Out-of-sample Kappa histogram for the analyzed models.

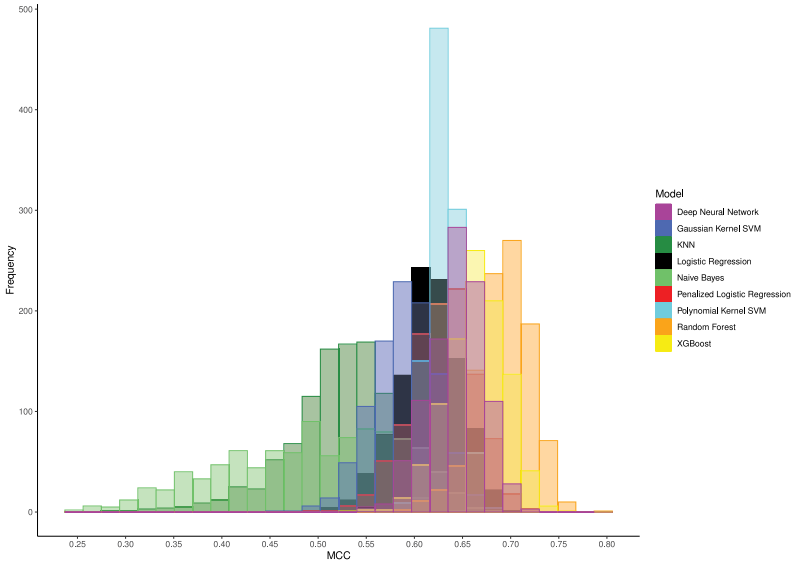

Fig. 6. Out-of-sample MCC histogram for the analyzed models.

Starting from the set of all tested models, MCS provides, at a given significance level α, a subset of "superior models" that contain the best model with probability greater than $1 - p$ by recursively testing the null hypothesis of equal predictive ability for all remaining models using a block bootstrap approach. The elimination rule evaluates a loss function and recursively removes the model with the worst relative performance in comparison to the average across all other models until all remaining models are statistically equal in terms of predictive power or until only one model remains. We applied MCS for all evaluation metrics described in subSection 3.3, defining the loss function as 1 minus the respective metric (given that all of them are naturally bounded between 0 and 1) and using the usual confidence level of 95% (*i.e.*: $\alpha = 0.05$).

As shown by the results displayed in Table 5, for all evaluation metrics, MCS identified only one model showing superior performance over all others, with the random forest standing out as the superior model for all metrics except recall, for which the deep neural network model statistically outperformed the other models. Once again, the results argue in favor of the random forest model for all metrics that balance both type I and type II errors (F1 Score, Kappa, and MCC), but suggest that the neural network model may perform better in avoiding false negatives. This finding can aid professional risk analysts interested in constructing fraud prediction systems based on machine learning models, depending on the relative cost of a false negative over a false positive.

4.2. Global interpretation: permutation-based variable importance

Besides the evaluation of the performance metrics from Table 4 and the statistical significances from Table 5, we proceed to yield a ranking of the features' relative importance, with a model-agnostic, permutation-based approach, proposed by Fisher et al. (2019) as an extension of Breiman (2001)'s feature importance measurement for the random forest model. As Fisher et al. (2019) pointed out, besides providing an estimate for the model class reliance, this more general approach is able to be applied for general models instead of tree-based or ensemble models since permutations on inputs are performed to the overall model instead of the individual ensemble members. Intuitively, if a specific feature is important to model the target variable, the predictive performance of the learner is expected to undergo sharper

Table 5
Set of superior models for each performance metric according to Hansen et al. (2011)'s Model Confidence Set procedure at the 95% confidence level.

Model	Evaluation metric					
	Accuracy	Precision	Recall	F1 Score	Kappa	MCC
Logistic Regression						
Penalized Logistic Regression						
Naive Bayes						
KNN						
Polynomial Kernel SVM						
Gaussian Kernel SVM						
Deep Neural Network			✓			
Random Forest	✓	✓		✓	✓	✓
GBM						

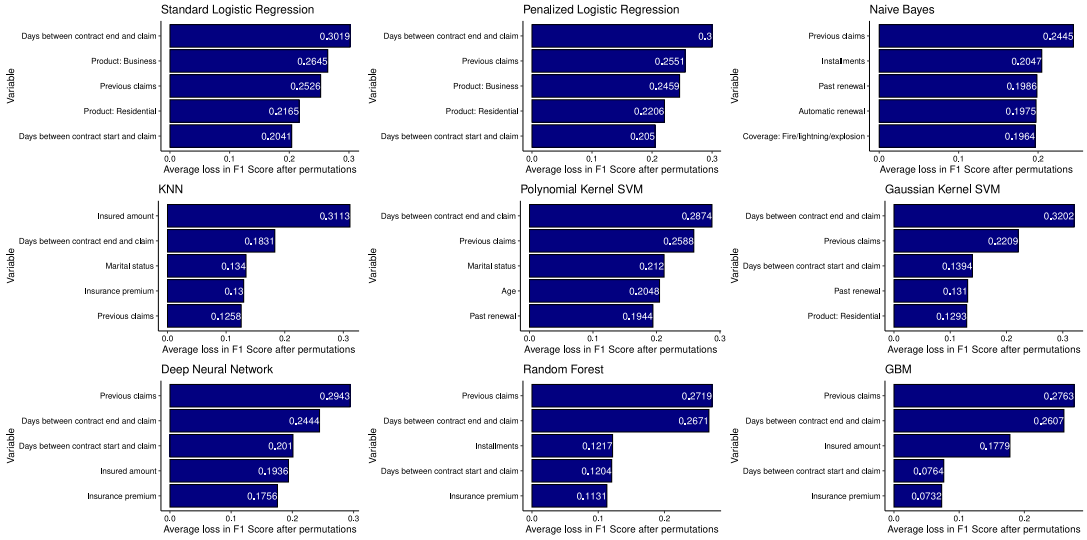

Fig. 7. Most important features for each model, given by the average loss in F1 Score after permutations.

drops after that feature is permuted; on the other hand, on a less relevant feature, the performance loss is expected to be smaller. In this sense, the difference between the observed values for the loss function before and after the permutations on a feature can be regarded as a proxy for its overall importance.

Therefore, for each of the 1000 rounds of training-validation-test and for all models, after computing the out-of-sample F1 Score we performed random permutations on the values of each feature and estimated the respective F1 Score using the optimal hyperparameters tuned after the cross-validation step, storing the difference between the metrics for every feature. Then, by taking the difference between the F1 Scores before and after the permutations, we computed the performance loss after the permutations for each feature. Finally, after computing the F1 Score differences between all features across the 1000 rounds, we took their average value and sorted them by descending importance. The five most relevant features for each model and their respective estimated impact on the F1 Score are displayed in Fig. 7:

As seen in Fig. 7, the number of previous claims and the variables that indicate premature claims (days between contract start/end and claim) were among the most important ones according to the permutation-based approach; "previous claims", specifically, was ranked as one of the 5 most relevant features for all 9 tested models. This result is aligned with the average expectation of a professional risk analyst, as discussed in the list of variables displayed in Section 3: based on our practical professional experience in fraud analysis, there are many necessary steps to analyze the information described by the client and evaluate the veracity of the claim: in this process, the variables which are usually considered the most relevant ones by analysts are the customer's history of previous claims and whether the claim is premature (claims in the first 60 days after the policy start date). By analyzing the customer's claims history, it is possible to see all the events described in previous claims, which are useful to understand not only the customer's behavior in previous claims but also most likely patterns of frauds, in the potential case of reincidence.

With regard to premature claims, on the other hand, this variable is commonly used by analysts to evaluate whether the policy was contracted right from the start with the expressive intention of reporting a claim, especially for new clients. The customer's history and the starting/ending dates of the policy are considered to be "primary variables", and are usually analyzed before the data related to the specific claim, such as the customer's description of the event, the reported damages, and the reports from the regulators, which are responsible for carrying out the necessary inspections to verify the claim. However, the analysts usually emphasize more on the time period between the start of the contract and the claim, instead of the end date of the contract, which is more regarded as a "check-up" variable of the term of the contract to verify whether the claim date is valid for the purposes of claim rejection or receipt of financial restitution, instead of an indicator of fraud. The fact that this variable having a high impact on the predictions may suggest that fraudsters could plan the "timing" of the fraudulent claim based on the contract duration, potentially anticipating the well-established "suspicion level" of a premature claim by making a delayed claim instead, which tend to arouse less suspicion on professional risk analysts, since the fraudster would be paying for the product for a longer period.

Other variables that had a high overall relevance across the models include the insurance premium and the insurance amount, which are also commonly used variables by risk analysts in fraud detection. Variables like income range, marital status, and age were ranked in intermediate positions, reaching top-5 on importance only for KNN

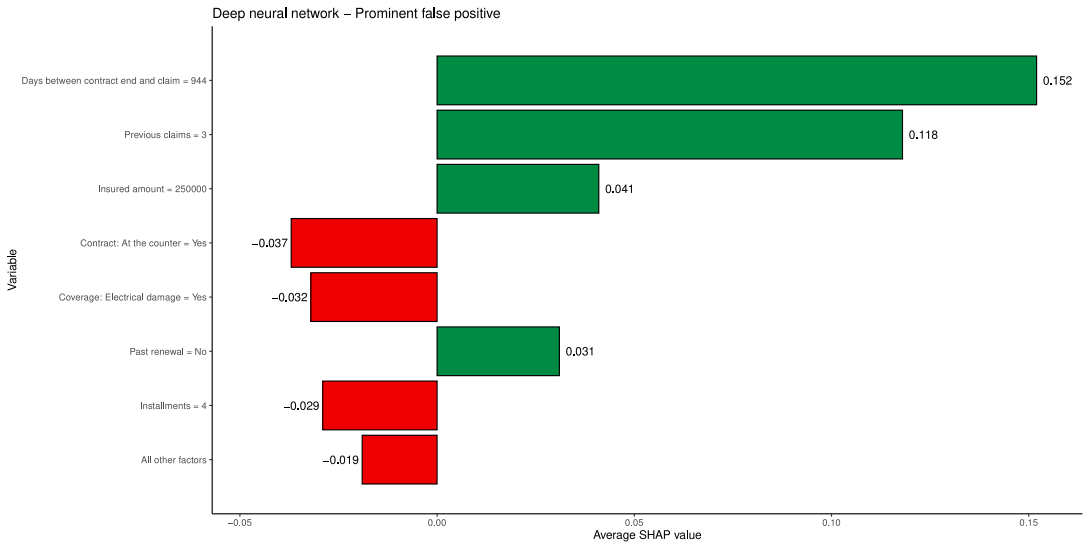

Fig. 8. Average SHAP values of deep neural network for the prominent false positive observation.

and Polynomial Kernel SVM, suggesting the existence of more complex patterns that determine the propensity of fraud. The indicators for the product type exhibited high importance for the logistic regression and its penalized version while having a much more timid relevance according to the majority of the other models, including the ones with the best predictive performance according to Table 4 (deep neural network, random forest GBM).

The importance of the variable "number of installments" varied across the nine tested models, having a high impact on the F1 Score after permutations for one of the best-performing models (random forest) and one of the worst-performing models (Naive Bayes) at the same time. A possible explanation is the fact that this variable is often evaluated jointly with other features, such as income range and insurance premium, suggesting that there may be cross interaction between those variables that culminate in relevant patterns for fraud identification. The variable "automatic renewal" was considered as a variable with low importance, while, in contrast, "past renewal" was assigned fairly high importance values for some models — this may have captured the effect of the document analysis involved in the renewal process, which is expected to make the fraudster more exposed to the detection of inconsistencies or abnormal behaviors. Indicators for legal person and contract channel were other variables that were in general considered as less relevant features across the models.

The feature importance rankings presented in this subsection provided by the machine learning models enable a better understanding of the relative strengths and weakness of each machine learning model, and can be used to generate decision rules applicable to real-world insurance policy evaluation, being applicable in practice as a screening stage to assist human analysts' posterior evaluation, with potential gains of speed and efficiency. Since we fitted the models with data from real fraud claims from an insurance company, the proposed methods are not only empirically effective but also have high applicability on corporate or governmental decision-making, integrating a good out-of-sample fraud prediction performance without losing the practical interpretability of those algorithms.

4.3. Local interpretation: Shapley additive explanation

As a complementary analysis to exemplify the potential use of eXplainable Artificial Intelligence framework to interpret predictions for individual observations, we performed an additional exercise using Shapley Additive Explanation values (henceforth SHAP values), a model-agnostic method introduced by Lundberg and Lee (2017) that aims at explaining individual machine learning model predictions, inspired by Shapley (1953)'s work on cooperative game theory. As discussed in Lundberg and Lee (2017), SHAP unifies a wide class of additive feature attribution techniques used for machine learning model explanations, such as LIME (Ribeiro et al., 2016), which approximates linear interpretable models near a given prediction; and Shapley sampling values (Štrumbelj & Kononenko, 2014), which provide estimates for feature importance in linear models under the presence of multicollinearity, by approximating the effect of removing each feature from the learner as a weighted average of differences between the predictions of a model trained with and without the respective feature. In this sense, while being computationally expensive, SHAP values assign importance values for each feature for a particular prediction, thus allowing to decompose the impact of each variable in the predicted outcome compared to the average prediction for the sampled observations.

In this sense, in addition to the global variable importance analysis presented in the previous subsection, we performed a local analysis on the non-fraud observation classified as a fraud the most times across all models and the fraud observation classified as a non-fraud the most times across all models — we shall call those observations as "prominent false positive" and "prominent false negative", respectively. We calculated the SHAP values associated with those two observations for the models that had the best overall performances – deep neural network, random forest, and GBM – using 1000 training rounds, each of them with a sample of 200 randomly selected training examples and 50 variable orderings, using the implementation of Biecek (2018). The average SHAP values for the 1000 rounds are displayed in Figs. 8to 13. The observed values for each variable were reported in their actual values (instead of centered and scaled) for better understanding.

In general terms, it was observed that the average SHAP values of two variables stood out for all three models: for the false positive observation, the variable "days between contract end and claim", was the one that contributed the most for predicting the prominent observation as a fraud; on the other hand, the variable "number of previous claims" was responsible for the strongest contribution for the prediction of the false negative case as a non-fraud. These results are aligned with the estimated variable importance displayed in Fig. 7 and, based on our practical experience in fraud detection, also aligned with the intuition of human analysts, since clients that issue frequent claims in property insurance usually raise the warnings for probable fraudulent behavior. Moreover, as reported at the beginning of Section 3, a significant proportion of the frauds are premature claims, which is measured by the interval between the contract start or end and the claim, variables that analysts tend to look at with emphasis.

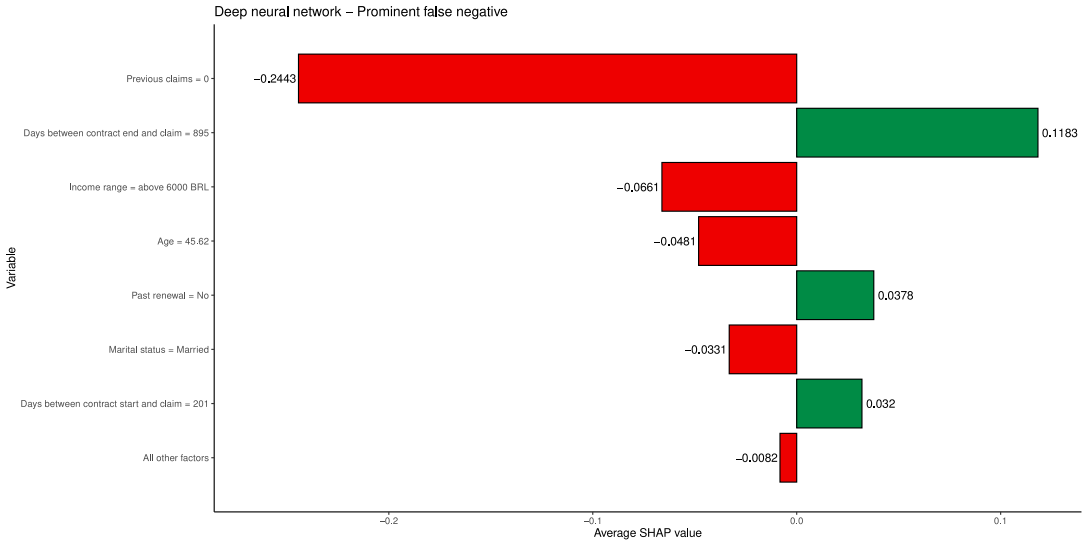

Fig. 9. Average SHAP values of deep neural network for the prominent false negative observation.

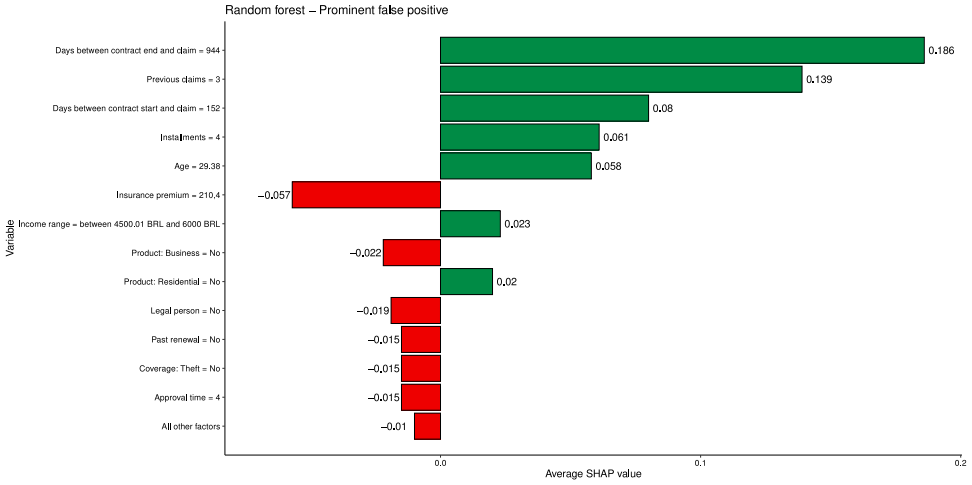

Fig. 10. Average SHAP values of random forest for the prominent false positive observation.

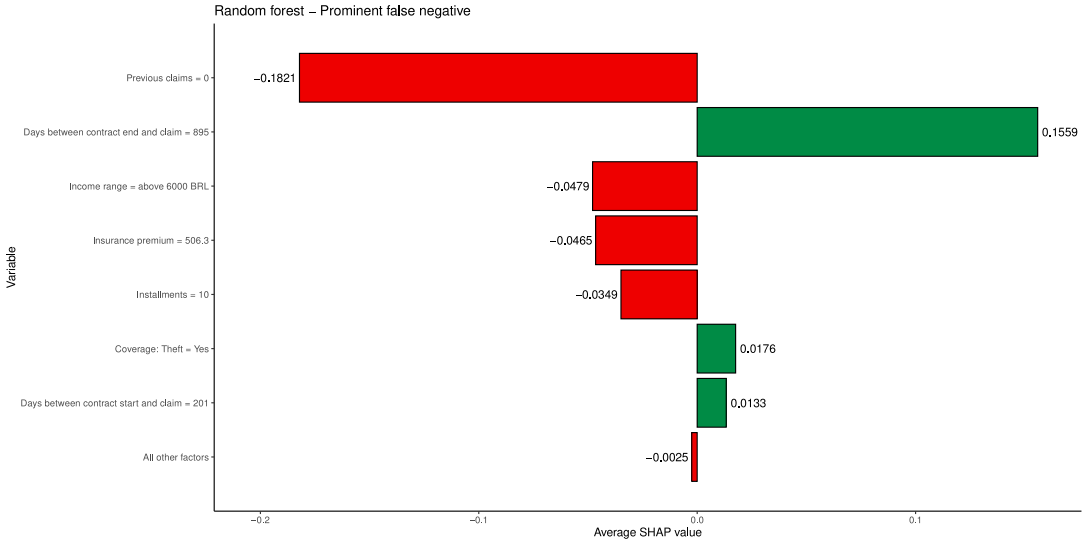

Fig. 11. Average SHAP values of random forest for the prominent false negative observation.

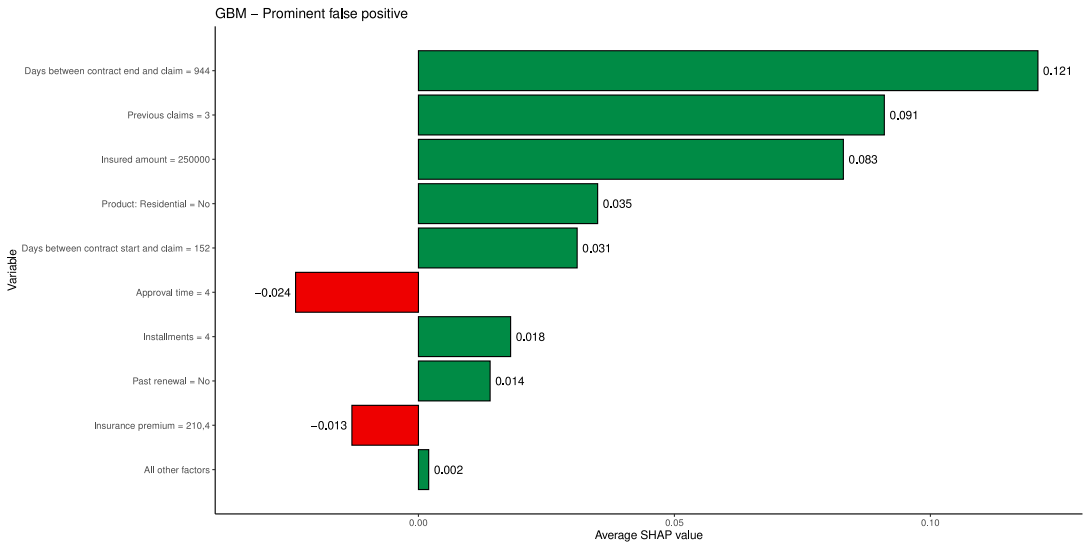

Fig. 12. Average SHAP values of GBM for the prominent false positive observation.

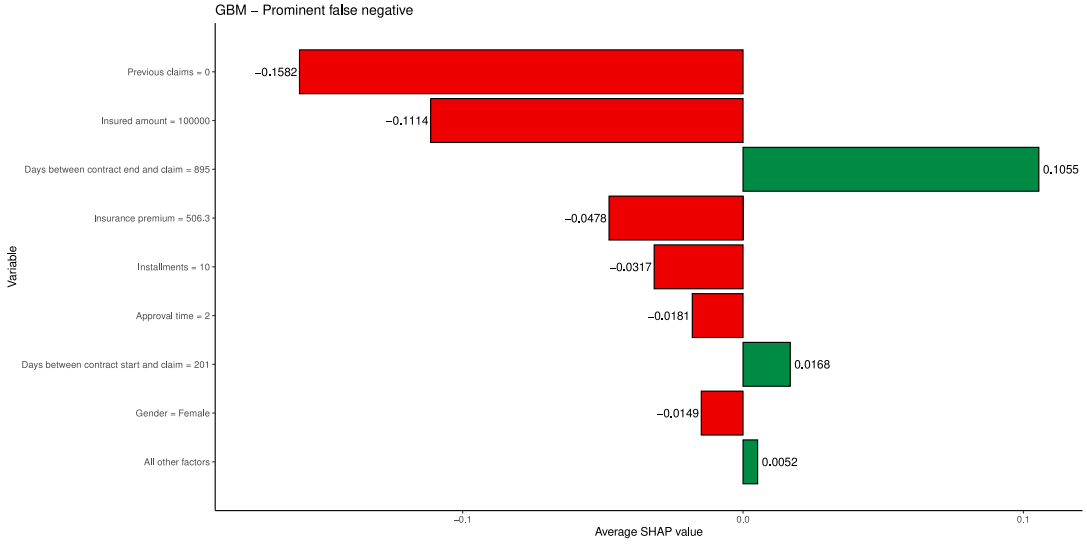

Fig. 13. Average SHAP values of GBM for the prominent false negative observation.

For the neural network model (Figs. 8 and 9) the variable "insured amount" played a relevant role "pushing up" the predicted fraud probability for the false positive observation, while the covered event (electrical damage) and the contract channel were pointed as important for reducing the predicted fraud probability, probably because the physical contract channel is the most common modality and the policies for electrical damage having in general low values for the insurance premium. For the prominent false negative observation, apart from the number of previous claims, the variables "income range", "age" and "marital status", which are variables usually jointly analyzed by human professionals, had important contributions for the erroneous classification of this particular policy, although none of them reached the top-5 most important features in the permutation-based analysis of the previous subsection. Bearing in mind the fact that deep neural networks had the best overall performance for false negatives (as seen by the recall values in Fig. 3), it can be inferred that those variables still play an important role in detecting actual frauds, alongside the insured amount, which also had a strong average SHAP value towards "non-fraud" prediction. No variables apart from the ones that indicate premature claims showed significant positive SHAP contributions.

For random forest (Figs. 10 and 11) the number of installments played an important role for both the false positive and false negative observations with strong average SHAP values, as well as age and income range. This is an interesting result because the number of installments usually does not play a decisive role for risk classification in property insurance; in the global interpretation subsection, this variable was also assigned a large importance for the random forest model. Another noteworthy result is the sign of the average SHAP value for the variable "past renewal", which is typically expected to reduce the fraud propensity when the policy is a renewed one, since it involves a screening process, while also having a lower rate of premature claims; however, premeditated fraudsters may also anticipate this decision rule to hide their intentions from the analysts. Given the overall good performance of the random forest model, future researches and professional analysts are recommended to further study the implications of these particular features on fraud detection tasks.

Finally, for GBM (Figs. 12 and 13) the relatively high insured amount was associated with a higher probability of fraud for the false positive case, as well as the fact of that policy being a residential one, which usually has a smaller proportion of frauds in comparison to the other categories. The number of installments also had a positive average SHAP for the false positive case, just like in random forest, and the strongest contribution for a non-fraud prediction was the small

approval time, which usually occurs in policies for a common event with low insurance premium — indeed, the insurance premium value also had negative average SHAP. For the prominent false negative observation, insurance amount and insurance premium had large contributions for predicting it as a non-fraud, while all variables apart from the number of days between contract start/end and claim (*i.e.*: the variables associated with the detection of premature claims) had small contributions in indicating a higher probability of fraud.

5. Conclusion and remarks

This article evaluated machine learning-based predictive models to detect frauds in property insurance policy claims, comparing the predictive results of nine predictive models using data from a major Brazilian insurance company. The results indicated that the random forest model achieved significantly better performance than the standard logistic regression and other machine learning methods, as evidenced by the metrics of accuracy, precision, F1 Score, Cohen's Kappa, and MCC, while the deep neural network model outperformed the other models for the recall metric. Moreover, based on the documented fraud cases, we listed a macro profile of the fraudsters and ranked the relative importance of the explanatory variables according to a permutation-based approach, highlighting the features that contributed the most to the models' overall predictive power and for the prediction of prominent false positive and false negative observations.

The findings of this paper can contribute to the literature of machine learning applications to fraud detection for residential and business insurance policies, a segment with relatively fewer works that follow this paradigm. In special, the fact that we tested the models using real-world data strengthens the relevance of the results over exercises that use simulated data, and further evidences the feasibility of converting the proposed models to operational tools for decision-making support in risk management, potentially assisting in the creation of data-driven and interpretable decision rules or being integrated into the evaluation process itself. Analogously, human analysts can benefit from this kind of product and also refine the algorithms by feeding more human-validated data into past datasets, which can be valuable to rectify mistakes made by the machine learning models.

The models proposed in this paper can also be adapted to a probabilistic approach, yielding not only if an insurance policy is more likely to be a fraud or a non-fraud, but also the probability of that specific policy to be a fraud. This probability can then be used as an input to evaluate the expected return (or loss) of a given insurance policy or to mathematically estimate the insurance premium adjusted to the fraud risk, bearing in mind operational, legal, and ethical constraints.

As future developments, we believe that a spatial analysis can be performed, adding to the current model variables such as the distance of the claimer's home to town center/working place, the overall income level and demographic variables of his/her neighborhood, among other potentially useful features that can potentially further enhance the model's quality. Other popular machine learning algorithms such as deep belief networks and restricted Boltzmann machines can also be applied in similar experiments on fraud detection. Finally, additional improvements can also be made augmenting the number of replications performed for each model, testing for a larger volume of data using methods for imbalanced classification, such as SMOTE (Chawla et al., 2002) and other methods described in Haixiang et al. (2017), as well as performing additional tuning of the hyperparameters' values for each model.

6. Disclaimer

Disclaimer 1: The views expressed in this work are of entire responsibility of the authors and do not necessarily reflect those of their respective affiliated institutions nor those of its members.

Disclaimer 2: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Matheus Kempa Severino: Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Yaohao Peng:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1–9). IEEE.

Biecek, P. (2018). DALEX: explainers for complex predictive models in R. *Journal of Machine Learning Research*, *19*(1), 3245–3249.

Brazilian National Confederation of Insurance Companies (2017). Dados básicos. http://cnseg.org.br/cnseg/estatisticas/mercado/dados-basicos/.

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

Bruns, R., Dunkel, J., & Offel, N. (2019). Learning of complex event processing rules with genetic programming. *Expert Systems with Applications*, *129*, 186–199.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*.

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, R.-C., Chen, T.-S., & Lin, C.-C. (2006). A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition and Artificial Intelligence*, *20*(02), 227–239.

Chen, Z., Chen, W., & Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications*, *146*, Article 113155.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928.

de Castro Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. *Applied Soft Computing*, *83*, Article 105640.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety (ICVES)* (pp. 1–5). IEEE.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 155–164).

Dou, Y., Li, W., Liu, Z., Dong, Z., Luo, J., & Philip, S. Y. (2019). Uncovering download fraud activities in mobile app markets. In *2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 671–678). IEEE.

Eshghi, A., & Kargari, M. (2019). Introducing a new method for the fusion of fraud evidence in banking transactions with regards to uncertainty. *Expert Systems with Applications*, *121*, 382–392.

Eweoya, I., Adebiyi, A., Azeta, A., & Azeta, A. E. (2019). Fraud prediction in bank loan administration using decision tree. *Journal of Physics: Conference Series*, *1299*(1), Article 012037.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Gottschalk, P. (2010). Categories of financial crime. *Journal of Financial Crime*, *17*(4), 441–458.

Gupta, R. Y., Mudigonda, S. S., Kandala, P. K., & Baruah, P. K. (2019). Implementation of a predictive model for fraud detection in motor insurance using gradient boosting method and validation with actuarial models. In *2019 IEEE international conference on clean energy and energy efficient electronics circuit for sustainable development (INCCES)* (pp. 1–6). IEEE.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*.

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., & Johnson, J. E. (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, *61*, 215–234.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, *37*(4), 543–558.

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, *100*, 234–245.

Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, *62*, 32–43.

Kim, E., Lee, J., Shin, H., Yang, H., Cho, S., Nam, S.-k., Song, Y., Yoon, J.-a., & Kim, J.-i. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications*, *128*, 214–224.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Majhi, S. K. (2019). Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. *Evolutionary Intelligence*, 1–12.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems with Applications*, *110*, 381–392.

Naser, M., & Alavi, A. (2020). Insights into performance fitness and error metrics for machine learning. ArXiv Preprint arXiv:2006.00887.

Ngai, E., Hu, Y., Wong, Y., Chen, Y., & 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.

Niu, F., Recht, B., Re, C., & Wright, S. J. (2011). HOGWILD! a lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of the 24th international conference on neural information processing systems* (pp. 693–701).

Peng, Y., & Nagata, M. H. (2020). An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos, Solitons & Fractals*, Article 110055.

Popat, R. R., & Chaudhary, J. (2018). A survey on credit card fraud detection using machine learning. In *2018 2nd international conference on trends in electronics and informatics (ICOEI)* (pp. 1120–1125). IEEE.

Raghavan, P., & El Gayar, N. (2019). Fraud detection using machine learning and deep learning. In *2019 international conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 334–339). IEEE.

Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144).

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. In *2017 international conference on circuit, power and computing technologies (ICCPCT)* (pp. 1–6). IEEE.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Sheshasaayee, A., & Thomas, S. S. (2018). Usage of r programming in data analytics with implications on insurance fraud detection. In *International conference on intelligent data communication technologies and internet of things* (pp. 416–421). Springer.

Sinayobye, J. O., Kiwanuka, F., & Kyanda, S. K. (2018). A state-of-the-art review of machine learning techniques for fraud detection research. In *2018 IEEE/ACM symposium on software engineering in africa (SEiA)* (pp. 11–19). IEEE.

Soman, K., Loganathan, R., & Ajay, V. (2009). *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd..

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665.

Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, *8*, 25579–25587.

Triepels, R., Daniels, H., & Feelders, A. (2018). Data-driven fraud detection in international shipping. *Expert Systems with Applications*, *99*, 193–202.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit card fraud detection-machine learning methods. In *2019 18th international symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1–5). IEEE.

Verma, A., Taneja, A., & Arora, A. (2017). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. In *2017 tenth international conference on contemporary computing (IC3)* (pp. 1–7). IEEE.

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

Waghade, S. S., & Karandikar, A. M. (2018). A comprehensive study of healthcare fraud detection based on machine learning. *International Journal of Applied Engineering Research*, *13*(6), 4175–4178.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95.

Yao, J., Zhang, J., & Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. In *2018 international conference on artificial intelligence and big data (ICAIBD)* (pp. 57–61). IEEE.

Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480.