ISSN 2616-5775 Vol. 7, Issue 9: 25-31, DOI: 10.25236/AJCIS.2024.070904

A fraud detection method based on integrated machine learning

Yiye Zheng[1,a,#], Linze Li[2,b,#], Bin Gao[3,c,#], Hansheng Yang[4,d,#], Jiaqi Jiang[5,e,#]

[1]Artificial Intelligence Institute, Yango University, Fuzhou, Fujian, 351100, China
[2]School of Physics and Astronomy, Sun Yat-sen University, Zhuhai, Guangdong, 519000, China
[3]Wuxi Kuangyuan Billingual School, Wuxi, Jiangsu, 214000, China
[4]Teensen Genesis School, Nanchang, Jiangxi, 330103, China
[5]The High School Attached to JXNU, Nanchang, Jiangxi, 330038, China
[a]654494031@qq.com, [b]1301076233@qq.com, [c]2823585616@qq.com, [d]3190613912@qq.com,
[e]3492689503@qq.com
[#]Co-first author

Abstract: *While the insurance industry is growing quickly, it is also facing a growing fraud problem, which poses a threat to the operations and stability of insurance companies. This article proposes a machine learning-based approach aimed at identifying and preventing fraud in auto insurance. The method improves the accuracy through feature selection, sample equalization and model optimization. A new index is also proposed to measure the stability of the prediction algorithm. The experiment shows that compared with the existing methods, the method proposed in this paper promotes the accuracy of detecting fraud behavior, and the stability of the algorithm is improved. The proposal of this method can help to better protect the insurance companies from fraud and ensure the stable development of their business.*

Keywords: *Insurance fraud; Machine learning; Data analysis*

1. Introduction

The insurance industry is an indispensable part of the financial system and plays an irreplaceable role in the development of the society. Nevertheless, the recent surge in insurance fraud not only poses a threat to the financial security of insurance companies, but also may shake public trust in the insurance system, leading to an increase in insurance costs. Given the seriousness and universality of the insurance fraud problem, it is urgent for insurance companies to take effective measures to identify and prevent such fraud behavior. With the development of big data and artificial intelligence technology, insurance companies can use these advanced analysis tools to improve the accuracy of fraud identification, and reduce the cost and time of manual audit.

Machine learning technology plays an essential role in the field of fraud detection. By deeply analyzing the data and identifying patterns, these techniques significantly promote the accuracy and efficiency of detection. They are quickly able to adapt to the continuous evolution of fraud behavior, enabling real-time monitoring of fraudulent activities, providing a solid guarantee for businesses and consumers to reduce potential losses. Besides, machine learning has demonstrated a remarkable ability to handle large data sets, revealing subtle patterns of fraud. This not only improves the accuracy of the assessment, but also strengthens the effectiveness of preventive measures. As fraud becomes increasingly complex and changing, the continuous learning and adaptability of machine learning techniques has become particularly critical. They play an indispensable role in maintaining the stability of the insurance market and promoting the sustainable development of the economy.

The identification of insurance fraud is indeed a complex problem, and its complexity is mainly reflected in the following aspects. First of all, fraud is a minority in all insurance cases, which leads to an imbalance between positive and negative samples in the data set, this unbalanced state has a huge impact on the accuracy and effective operation of the model. Secondly, in terms of feature engineering, effective fraud detection requires the identification and utilization of the correct features and methods, which requires not only a lot of work, but also a deep understanding of the relevant fields. Finally, with the continuous evolution and update of fraud, anti-fraud technology must be updated to deal with new fraud.

For the above problems, this paper proposes a new method for fraud identification using machine learning techniques. Through data exploration and analysis, a series of improvements are put forward in data preprocessing and feature engineering, which is helpful to improve the prediction accuracy of the algorithm. This paper also proposes an ensemble learning algorithm, turning multiple unstable sets of algorithms into a stable algorithm. This paper also presents a new index to measure the stability of the prediction algorithm.

The organizational structure of the following part of this paper is as follows: The second part is the relevant work, which summarizes the existing relevant research results and methods, highlighting the innovations and differences of this research. The third part is the model method, which describes the proposed model and method in detail, including the theoretical basis, model structure, algorithm process, etc. The fourth part is experiment, which mainly introduces the experimental design, data set, experimental steps and evaluation indicators, and shows the experimental results and makes analysis and discussion. The last part is conclusion, which summarizes the main contributions and innovations of the research, reviews the main contents and conclusions of the paper, and points out possible research directions and future work.

2. Related works

The explosive growth of technology and economy in today's society has led to an increase in the number of fraudulent activities, which exist in various aspects and industries of society. At this point, the research on fraud prediction technology becomes particularly important. With the efforts of scholars both domestically and internationally, this technology has made significant progress. Richard J. Bolton and David J. Hand [1] studied some possible methods for unsupervised credit card fraud detection through behavior outlier detection techniques. The article 'Unsupervised Profiting Methods for Fraud Detection' describes the early research stages to generate some frameworks for unsupervised fraud detection and presents some basic examples for illustration. By incorporating other information (not just the amount spent) into the anomaly detection process and determining the most useful and practical fraud detection methods.

Supervised learning is a more mainstream machine learning method for studying fraud prediction in modern times. M. Valavan and S Rita [2] compared rendering methods of different machine learning algorithms, such as decision trees, random forests, linear regression, and gradient boosting methods, for detecting and predicting fraud cases using loan fraud manifestations. Further model accuracy measurements were performed using confusion matrix and calculations of accuracy, precision, recall, and F-1 scores, as well as receiver operating characteristic (ROC) curves. Researchers have shown that the prediction accuracy of the random forest classifier is 80%, while the prediction accuracy of the logistic regression method is 70%. For this type of data, the random forest model seems to be a better choice. Compared with other technologies, one of them, called gradient boosting algorithm, has shown better results in integrated machine learning, and its accuracy has been shown to be over 90%. Among the three algorithms, the sine gradient boosting model has the highest efficiency. In fact, some scholars have compared the efficiency of different gradient boosting models. Yong Fang et al. [3] compared the Light Gradient Boosting Machine model with random forest and gradient boosting algorithms in their experiments. The results indicate that the Light Gradient Boosting Machine model has good performance. The credit card fraud detection experiment based on the Light Gradient Boosting Machine model achieved a total recall rate of 99% and fast feedback on a real dataset, demonstrating the efficiency of the new model in detecting credit card fraud. However, in most cases, the model used for fraud detection is XGBoost. Didrik Nielsen [4] provides evidence to explain why gradient boosting trees, especially XGBoost, appear so effective and versatile. The researchers creatively interpreted the boosting algorithm used by XGBoost as a Newton's method in the function space, and compared it with MART, demonstrating the superiority of XGBoost.

Meanwhile, Suneetha et al. [5] provided a detailed analogy between different supervised and unsupervised machine learning techniques used for detecting fraudulent activities. New schemes Cat Boost and Light Gradient Boosting Machine (LGBM) have been proposed to detect fraudulent behavior. Comparing the performance of these methods with autoencoder (AE), logistic regression, K-Means clustering, and neural network (NN) methods, it was found that Cat Boost and LGBM have high accuracy in fraud detection.

In addition, YI Dongyi et al. [6] proposed a medical insurance fraud detection model based on graph convolution and variational autoencoder (OCGVAE) to address the problems of insufficient fraudulent samples, high data annotation costs, and low accuracy in traditional Euclidean space models. This model effectively improves the accuracy of medical insurance fraud screening. Pooja Tiwari et al. [7] conducted a comprehensive review of various methods used to detect credit card fraud. These methods include Hidden Markov Models, Decision Trees, Logistic Regression, Support Vector Machines (SVM), Genetic Algorithms, Neural Networks, Random Forests, Bayesian Belief Networks. A comprehensive analysis was conducted on various technologies, and their advantages and disadvantages were listed. Almost all published fraud detection studies have been investigated and explored in the work of CLIFTON PHIA. It defines attackers, types and subtypes of fraud, technical properties of data, performance metrics, and methods and techniques. After identifying the limitations of fraud detection methods and techniques, it indicates that this field can benefit from other related areas. Specifically, unsupervised methods in counter-terrorism work, actual monitoring systems and text mining by law enforcement agencies, as well as semi supervised and game theory methods in intrusion and spam detection communities, can all contribute to future fraud detection research.

Today's scholars eliminate its limitations and achieve enhanced performance by creating a mixture of various techniques already used in fraud detection. J. Esmaily and R Moradinezhad [8] proposed a hybrid model of decision tree and neural network; R. Patidar and L Sharma[9] proposed a hybrid method of neural network and genetic algorithm; T. Kumar and S Panigrahi proposed a hybrid method of fuzzy clustering and neural networks; A. Agrawal et al. proposed a hybrid model consisting of Hidden Markov Models, Behavior Based Techniques, and Genetic Algorithms; Sam Maes proposed a hybrid Bayesian network and artificial neural network.

In our research, we primarily proposed enhancements in data preprocessing and feature engineering, significantly contributing to the enhancement of algorithm accuracy. Concurrently, we introduced an ensemble learning method to integrate multiple unstable algorithms into a stable one. Lastly, we proposed a new metric to assess the stability of algorithm predictions.

3. Our method

The method proposed in this paper consists of two steps: feature engineering and model construction, where feature engineering optimises the structure of the data to make it more conducive to the discovery of hidden information, and model construction mainly screens the models used and improves their predictive stability through integration. In order to characterise the stability of the model, this paper proposes a new metric of predictive stability. The details are discussed as follows.

3.1. Feature engineering

Feature Engineering (FE) is a crucial step in machine learning and data analytics that involves a series of processes and transformations of raw data to extract more useful features and convert these features into a format suitable for processing by machine learning models.

(1) Data preprocessing

First of all, the original data is cleaned, including the processing of missing values, outliers and duplicates, etc. For the missing values, according to the importance of the features and the distribution of the data, this paper adopts the method of filling in the plurality to deal with them, as well as picking a box-and-line diagram to deal with the anomalies and visualise them, and find the existence of variables with anomalies according to the box-and-line diagram Because of the special risk prediction, the anomalies are retained to avoid interfering with the forecasting Accuracy.

(2) Special data handling

Based on the case experience, it was considered feasible to dig deeper into the feature of time, so the time between accidents and weather-related months were put into the data as a new feature to be analysed, and some time features that were not relevant to the labels were removed, e.g., time of accidents, insurance policy number.

(3) Feature selection

Obviously irrelevant or redundant features, such as insurance numbers, insured postcodes, etc., were eliminated based on business experience and links between numerical variables such as Pearson Correlation Coefficients (PCCs) heat matrices.

(4) Feature Codes

Discrete values were coded as continuous values through MeanEncoder (MeanEncoder) to facilitate subsequent analyses. MeanEncoding is a coding method that maps category characteristics to mean values of the target variable. It takes advantage of the statistical properties of the target variable in terms of the values taken by different categories and assigns a corresponding coded value to each category. This coding method preserves information about the category features to some extent and usually provides a more compact representation than solitary heat coding. And using mean coding to code Object feature data, the following are the basic ideas and principles: mean coding, also called Target Encoding in some places, is a supervised coding method based on the statistics of target variables (Target Statistics). The method is based on the Bayesian idea, using the weighted average of the prior probability and the posterior probability as the coded value of the category feature value, which is suitable for classification and regression scenarios.

3.2. Modelling

(1) Model selection

In this paper, 26600 pieces of data were selected for integration and analysis, and set 70% as training set and 30% as validation set. After several algorithms such as Catboost, LightGBM, Random Forest etc., we finally selected two machine learning algorithms, GBDT and XGBoost for model construction, although these two algorithms perform well in classification problems and are suitable for insurance fraud detection scenarios, they are found to be unstable, so we will integrate and optimise these two models in the later research.

(2) Model training and tuning:

Divide the dataset into a training set and a test set, usually with a ratio of 70:30 or higher. The selected algorithms are trained using the training set data and an initial model is obtained. Subsequently, the model is evaluated using the test set data and the model performance is measured by metrics such as feature importance ranking. The model is tuned based on the evaluation results, including adjusting the algorithm parameters and selecting more appropriate features.

(3) Model selection and results:

In order to arrive at a more accurate prediction value, i.e., AUC (area under the curve), the principle is that given a random positive sample and a negative sample, a classifier is used to classify and predict the probability that the score of that positive sample is larger than the score of that negative sample. And after many experiments and evaluations, this paper finally selects ensemble learning as the optimal model. Ensemble learning performs well in both training speed and prediction accuracy, and is able to accurately identify most of the fraudulent behaviours, and the successful application of this model provides an effective technical means for insurance fraud detection.

4. Experiments

4.1. Introduction to data

This dataset comes from AliCloud Tianchi Competition, which contains 37 columns of feature data including age, education level, asset loss, time of enrolment, etc. of the participants, totalling about 38,000 pieces of data. We test and train these data according to a certain ratio to gradually optimise and improve the model's prediction correctness as well as higher stability.

4.2. Experimental design

Firstly, more in-depth digging on feature selection - for the original features to be analysed in detail, and added to the analysis of label correlation, and according to the following are the characteristics of the label and the reasons for the selection:

Time interval between date of insurance and date of accident:

Insurance fraud may be more common where an accident occurs quickly after the insurance is taken out. The time interval between the date of the policy and the date of the accident can help to identify this unusual pattern. Shorter time intervals may indicate a higher risk of fraud.

Month of accident date:

Different seasons may affect the frequency of accidents. For example, winter may have a higher incidence of accidental road accidents due to poor weather conditions. By including month features, the model can capture these seasonal patterns.

(3)The educational qualifications of the policyholder:

The level of educational qualifications may be associated with a pattern of policyholder behaviour. Policyholders with certain levels of education may be more inclined to commit insurance fraud. For example, those with lower levels of education may be more likely to engage in fraudulent behaviour, while those with higher levels of education may be more cautious. The use of educational qualifications as a feature may help the model to better distinguish between fraudulent and non-fraudulent behaviour.

Secondly, in terms of model selection using integrated learning method (Ensemble learning) will GBDT and XGBoost according to the AUC value and the stability of the high and low values of the model weight adjustment, after data analysis decided to 0.5: 0.5 weights of the two models for integration, and then the different features selected for the arrangement of the data, and do not join any features of the AUC value as the basis, to determine whether adding different features is to have a higher AUC value.

4.3. Experimental results

The performance comparison of the experimental schemes is shown below:

According to Figures 1 and 2, although the AUC obtained by the integrated learning method is slightly smaller than the XGBoost algorithm and higher than the GBDT algorithm but its variance is smaller than both algorithms, which implies that the integrated learning algorithm can obtain higher stability at the expense of a certain model prediction accuracy.

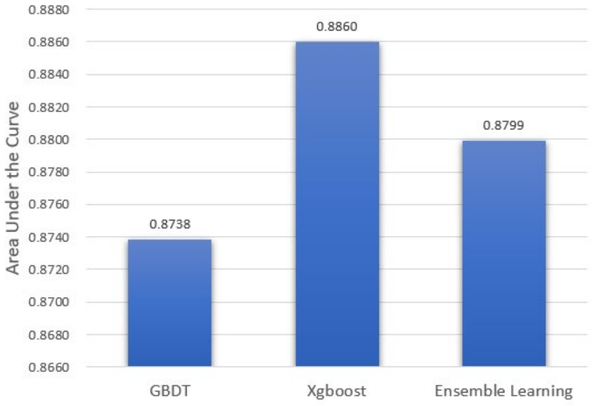


Figure 1: AUC of different algorithms

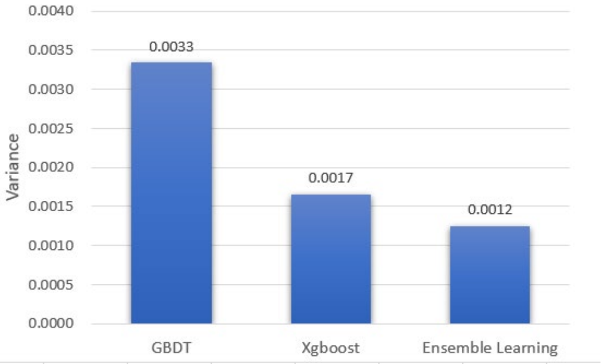


Figure 2: Magnitude of variance of different algorithms

According to Figures 3 and 4, adding both new features, time and education, to the integrated algorithm significantly reduces the variance of the algorithm and improves the stability; at the same time, the area of the AUC is greatly improved. This proves the success of feature selection and the superiority of the integrated algorithm

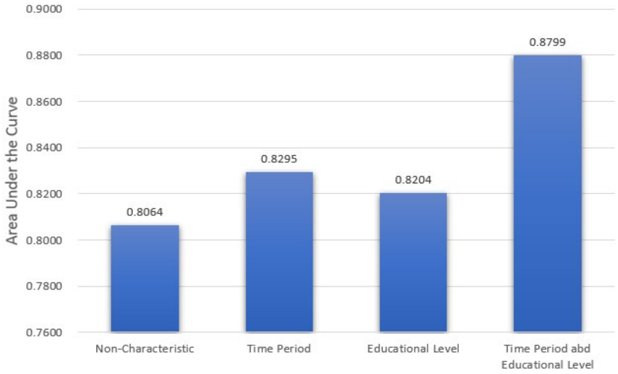


Figure 3: Relationship between different features and AUC

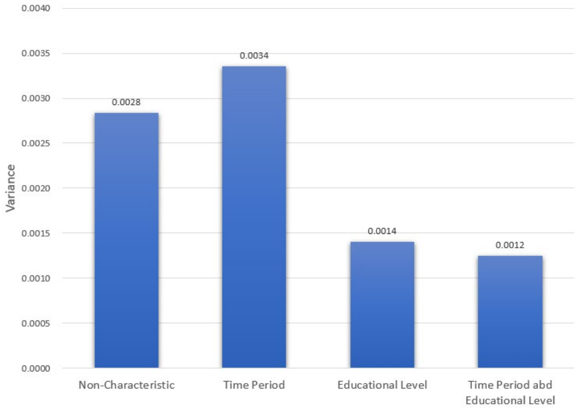
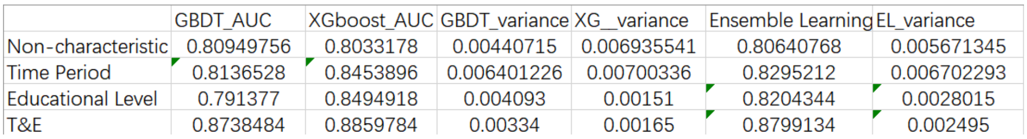


Figure 4: The magnitude of the variance of different features



	GBDT_AUC	XGboost_AUC	GBDT_variance	XG__variance	Ensemble Learning	EL_variance
Non-characteristic	0.80949756	0.8033178	0.00440715	0.006935541	0.80640768	0.005671345
Time Period	0.8136528	0.8453896	0.006401226	0.00700336	0.8295212	0.006702293
Educational Level	0.791377	0.8494918	0.004093	0.00151	0.8204344	0.0028015
T&E	0.8738484	0.8859784	0.00334	0.00165	0.8799134	0.002495

Figure 5: Test analysis

As shown in Figure 5, we add data with different characteristics to three models, sample the data and run multiple times to obtain the final analysis value, and calculate the variance value of each model according to the different AUC values of each model. According to the numerical table, it can be seen that the AUC value measured by adding the T&E (Time and Education level) feature is higher than the other values, and the variance value is controlled in a very small range, which proves its stability. In terms of model selection, under the premise of considering both performance and stability, the ensemble learning model is the best choice.

5. Conclusions

Aiming at the increasingly serious fraud problem in the insurance industry, this study presents a fraud detection method based on machine learning technology. This method significantly improves the identification accuracy of fraud behavior by deeply analyzing the data, identifying the patterns, and solving the imbalance problem in the data set. At the same time, the model adopts a variety of models, which can timely and effectively adapt to the continuous evolution of fraud means. The innovation of this study is to propose an ensemble learning method that turn multiple unstable sets of unstable algorithms into a stable algorithm. A new index of prediction stability is also presented to measure the stability of the prediction algorithm.

Nevertheless, this approach still faces challenges in feature engineering and real-time updating, and requires further optimization to improve its accuracy in complex environments. Future work will focus on further optimizing the process of model diversity and improving the model response to emerging fraud patterns. Through these efforts, we expect to provide the insurance industry with more efficient and accurate fraud detection tools to maintain the stability and fairness of the market.

References

[1] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. Credit scoring and credit control VII, 235-255.

[2] Valavan, M., & Rita, S. (2023). Predictive-Analysis-based Machine Learning Model for Fraud Detection with Boosting Classifiers. Computer Systems Science & Engineering, 45(1).

[3] Fang, Y., Zhang, Y., & Huang, C. (2019). Credit Card Fraud Detection Based on Machine Learning. Computers, Materials & Continua, 61(1).

[4] Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win" every" machine learning competition? (Master's thesis, NTNU).

[5] Ramani, K., Suneetha, I., Pushpalatha, N., & Harish, P. (2022). Gradient boosting techniques for credit card fraud detection. Journal of Algebraic Statistics, 13(3), 553-558.

[6] Yi, D., Deng, G., Dong, C., Zhu, M., Lyu, Z., & Zhu, S. (2020). Medical insurance fraud detection algorithm based on graph convolutional neural network. Journal of Computer Applications, 40(5), 1272.

[7] Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J., & Singh, A. K. (2021). Credit card fraud detection using machine learning: a study. arXiv preprint arXiv:2108.10005.

[8] Esmaily, Jamal, Reza Moradinezhad, and Jamal Ghasemi. "Intrusion detection system based on multi-layer perceptron neural networks and decision tree." 2015 7th Conference on Information and Knowledge Technology (IKT). IEEE, 2015.

[9] Behera, T. K., & Panigrahi, S. (2015). Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In 2015 second international conference on advances in computing and communication engineering, IEEE, pp. 494-499.