Analysis and Discrimination of Insurance Fraud based on Data Mining

Tianqi Yang *, Yue Wu

authors: tqyoung@126.com*, wy18902292218@163.com

Beijing Normal University- Hongkong Baptist University United International College, ZhuHai, Guangdong, China

Abstract. Insurance is an important component of the financial system, playing an important role in social stability and ensuring people's livelihoods. With the vigorous development of the insurance industry this year, automobile insurance fraud has been a frequent occurrence. This article achieves effective discrimination of insurance fraud by constructing a model of insurance fraud. Firstly, preprocess the data from the insurance claims dataset and encode the text variables using average encoding; Then, this article conducts data correlation analysis on all attributes of the sample using Pearson correlation coefficient; Based on the results of the previous analysis, the K-means clustering method is used to achieve dimensionality reduction and enhancement of sample attributes; Finally, by training an SVM classifier with Gaussian kernel function, effective discrimination of insurance fraud is achieved. Through experimental verification, the model method is effective and can achieve accurate discrimination with an accuracy of 96%.

Keywords: Insurance Fraud, Data Mining, K-means, SVM

1 Introduction

Since the beginning of the 21st century, the living standards of many developing countries have significantly improved, and the purchase and possession of motor vehicles in these countries have also been increasing year by year. Insurance is an inseparable component of the financial industry in various countries, playing an indispensable role in safeguarding the lives of the country's people and maintaining social stability. However, it is also accompanied by the problem of insurance fraud, which has always been an urgent problem to be solved worldwide due to its high returns and low risks[1-3].

Firstly, it will affect normal claims processing, leading to a decrease in efficiency and an increase in the company's daily operating costs;

Secondly, insurance fraud claims lead to unnecessary financial flows, damage the financial interests of insurance companies, and increase the operational risks of the company;

Finally, it also seriously damages the rights and interests of other honest and trustworthy customers, causing them to pay for increased premiums, thereby disrupting the normal social order and market system.

Solving the problem of insurance fraud claims is an urgent task for the insurance industry. At present, many countries and the insurance industry have attached a certain degree of importance

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338047

to insurance fraud, and the rapid development of technology has also played a certain regulatory role in insurance fraud[3-6]. But the methods of insurance fraud have also become more complex and diverse, and even more covert. In order to address the occurrence of insurance fraud, detailed research has been conducted today, mainly divided into classic data analysis methods and neural network methods.

Classical data analysis methods mainly include: Verhulst first proposed the Logistic regression analysis model [7], which can be divided into multiple logic regression analysis and dual meta-logic regression analysis according to different values of the explained variables, and applied to the field of insurance fraud analysis. Artis and other scholars have updated and developed the Logit model to construct the AAG model [8], which has the advantage of handling missing claims sample data. Benedek utilizes the data mining tool SQL Server Analysis Services (SSAS) to identify key automotive insurance fraud indicators and compares the performance of decision tree and neural network data analysis methods [10]. A common indicator for measuring the correlation between two vectors is the Pearson correlation coefficient [11], which is more suitable for processing high-dimensional data. The Gini index is used as a method for selecting partition attributes in the CART (Classification And Regression Tree) decision tree[12], and can also be used as an indicator to measure the importance of features. Parallel methods that are independent of each other between machine learners, such as Bagging [14]. Bagging adopts a self-service sampling strategy, using simple voting for classification tasks and simple averaging for regression tasks. XGBoost [15] uses the second-order Taylor expansion of the loss function to more accurately fit the loss function. Provides greater space for future optimization. LightGBM can improve the training speed of traditional GBDT while achieving nearly the same accuracy [16].

The neural network method is a new type of information processing and computing system that belongs to artificial intelligence technology and has strong ability to process noisy data. Rumelhart et al. proposed the backpropagation algorithm, which has made the neural network method widely used [9]. SVM (Support Vector Machine) [13] is a binary classification model. Its basic idea is to find a hyperplane in the feature space that can correctly divide samples and has the largest spacing. The teams of Lu Bingjie and Li Weizhuo summarized the application of machine learning models in the field of insurance fraud detection. By using real insurance claims data from insurance companies to test and analyze different machine learning models, they predicted future insurance fraud [17]. Subudhi et al. [18] proposed to combine genetic algorithm and fuzzy C clustering method to generate a cluster with the optimal cluster center to obtain a balanced dataset, which was verified by different supervised learning methods DT, SVM, MLP and GMDH. Majhi et al. [19] used fuzzy mean clustering method for clustering and used an improved whale optimization algorithm to find the global optimal solution for a given dataset, thus proposing an insurance fraud detection system based on fuzzy clustering. Yan et al. [20] proposed a Kernel Ridge Regression (KRR) based on artificial bee colony algorithm for insurance fraud detection. Panigrahi et al. [21] used three feature selection algorithms to extract important representations from insurance fraud data and used machine learning algorithms for detection, in order to select the best feature selection method for different machine learning models.

The model proposed in this paper combines the advantages of classical data analysis methods and neural network methods. In terms of data processing, feature extraction and dimension reduction, classical data analysis methods are used. In the training process of classifiers, neural networks are used, hinge loss function constraints are used, and Gaussian kernel function is used to achieve accurate identification of insurance fraud. The above are the characteristics of the model in this paper.

The rest part of the paper is organized as follows. The Sec. II will introduce data preprocessing, including missing data processing in the dataset and encoding of text variables. The Sec. III introduces the process of building the model, including data normalization, K-means feature extraction, and SVM classification and discrimination. The Sec. IV will analyze the results of this study. Finally, a brief summary will be provided in Sec. V.

2 Data preprocessing

Data from https://tianchi.aliyun.com. This dataset provides car insurance data for customer claims, consisting of 700 samples, each with 39 variables. The last variable represents whether the customer has engaged in insurance fraud, while the remaining 38 variables are background information or accident information of the customer.

2.1 Data processing

To identify insurance fraud, it is necessary to select all variables related to insurance fraud and eliminate variables unrelated to insurance fraud. This article excludes two coding variables: policy id (insurance number) and insured zip (insured postal code), which have a significant weak correlation with insurance fraud.

Incident date (date of occurrence), policy bind date (insurance binding date), and auto year (automobile purchase date) are data in the form of dates and cannot be directly calculated. Considering practical significance, two time indicators are redefined: delta-time1 = time difference from insurance binding to accident occurrence=accident date - insurance binding date, delta-time2 = time difference from car purchase to accident occurrence=accident date - automobile purchase date, and delete the original 3 date indicators. The data types of delta-time1 and delta-time2 are real numbers.

2.2 Data missing processing

Of the 700 samples that have identified whether there is Insurance fraud, only some samples have missing Collision type values. By calculating whether the existence of Insurance fraud is independent of the Collision type at the significance level of 0.999, we can judge that the existence of Insurance fraud is strongly related to the Collision type. In the absence of Collision type, the judgment of Insurance fraud behavior is more inclined to no, which is mainly because people who commit Insurance fraud often design the process of fabricating accidents. Therefore, the missing Collision type will seriously affect the judgment of Insurance fraud. This article will uniformly encode the samples with missing Collision type to ensure the accuracy of the evaluation.

2.3 Coding of data categorical variable

In the data set of 700 samples, variables can be divided into numerical variables and categorical variable.

This article does not handle numerical variables.

Categorical variable refers to the definition of variables. They are meaningless for operations such as addition, subtraction, and averaging. Variables can be divided into multiple different data categories. Database categorical variable include policy Combined single limit, insured sex, insured education level, insured occurrence, insured hooks, insured relationship, Incident hour of the day, incident type, collaboration type, Incident severity, authorities contacted, incident state, incident city, property damage, policy report available, auto make, auto model.

For categorical variable, Mean Coding is used to prevent overfitting. If a feature is categorical and has a large number of possible values (high cardinality), then mean encoding is an efficient encoding method.

In posterior probability $\hat{P}(fraud = 1|variable_t = k)$, for the kth type of the tth categorical variable, the weight of the prior probability$\hat{P}(fraud = 1)$ is introduced to calculate the probability $\hat{P}$ used for coding：

$$\hat{P} = \lambda \times \hat{P}(fraud = 1) + (1 - \lambda) \times \hat{P}(fraud = 1|variable_t = k), \tag{1}$$

Where $\lambda(n) = \frac{1}{1+e^{(n-f)/f}}$, n is the number of times the feature category appears in the training set as input, and k and f are parameters, set k=2 and f=1 here.

3 Methodology

Due to the increasing number of insurance fraud, among numerous research methods, this paper uses data mining methods to analyze insurance claim data and achieve accurate identification of insurance fraud. The specific steps are as follows:

1) Firstly, normalize the variables in the insurance sample library to eliminate the adverse effects caused by singular sample data;

2) Analyze data correlation to obtain the correlation between various variables in the sample library and insurance fraud. Through correlation verification, extract attributes with high sample correlation;

3) Conduct cluster analysis analysis on the attributes with high sample correlation, measure the similarity between different attributes, and classify the sample attributes into different clusters;

4) Using support vector machines, classify and train the clustering attributes of the sample library to achieve accurate discrimination of insurance fraud.

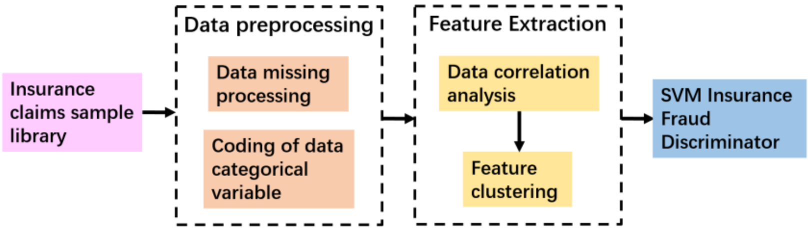
Insurance claims sample library
Data preprocessing
Data missing processing
Coding of data categorical variable
Feature Extraction
Data correlation analysis
Feature clustering
SVM Insurance Fraud Discriminator

Fig. 1. Design flowchart of insurance fraud discriminator

3.1 Data correlation analysis

Each sample in the insurance claims dataset has 39 attribute variables, some of which have a strong impact on the discrimination of insurance fraud, some have a weak impact on the discrimination, and some attribute variables have common correlations. These are all factors that affect insurance fraud. Therefore, how to identify the main judgment basis that affects insurance fraud among the numerous feature attributes and reduce the interference of weak impact terms is a major research issue in this article. This article proposes to use Pearson correlation coefficient to determine the correlation between 38 attribute variables in the insurance claims dataset and insurance fraud attributes.

The Pearson Correlation Coefficient (PCC) used in this paper is a correlation analysis method that can be used to measure the linear correlation between two variables. It can be used to describe the linear correlation between two variables, thus evaluating whether they have a linear relationship and whether the linear relationship between the two is positive or negative.

The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}, \tag{2}$$

Where x and y are two random variables, and $cov(x,y)$, which are the covariance of x and y, σ_x is the standard deviation of x, σ_y is the standard deviation of y.

For database samples, formula (2) can be rewritten as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3}$$

Where x and y are two random variables, and n is the number of samples, $x_{i,}y_i$ are the i-point observations corresponding to variables x, y, $\bar{x}$ is the average of x samples, $\bar{y}$ is the average of y samples.

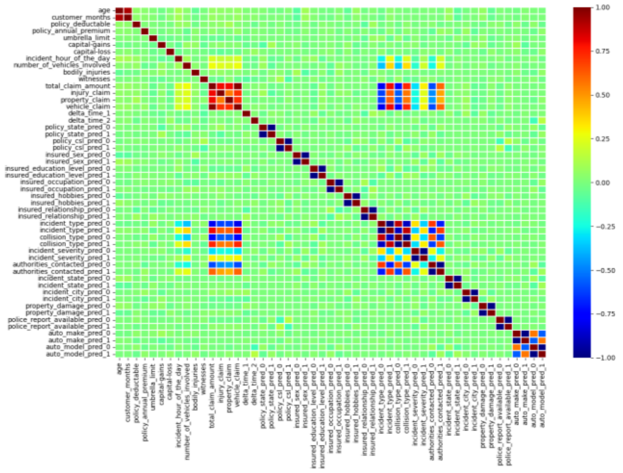

Fig. 2. Correlation matrix based on average encoding

This article conducted data correlation analysis using Pearson correlation coefficient, and the results are shown in Figure 2. Although there are many characteristic variables in the database, the variables with strong correlation with insurance fraud can be seen from the figure as total claim amount, property claim, injury claim, vehicle claim, incident type, incident severity, collision type, authorities contacted. Therefore, insurance fraud discrimination analysis will be conducted on these characteristic variables in the future.

3.2 Feature clustering

The K-means algorithm is a clustering algorithm based on partitioning, with the optimization goal of keeping points of the same class as close as possible and points between classes as far as possible. What needs to be done is (1) given the number of clusters k, (2) selecting K initial points, which can be random values or random sample points, and (3) iterating to the termination condition.

In order to avoid the K-Means algorithm falling into local optima and achieve accurate description of k-values and cluster center points, this paper uses two evaluation indicators of k-means clustering: Calinski-Harabaz Index (CH) and Silhouette Coefficient, to jointly constrain the clustering process.

The essence of the Calinski Harabasz Index is the ratio of inter cluster distance to intra cluster distance, and the overall calculation process is similar to the variance calculation method, so it is also called the variance ratio criterion. Aggregating the dataset *x* with a capacity of *N* into *K* classes, the compactness within the class is measured by calculating the sum of the squares of the distances between each point within the class and the center of the class (intra class distance), and the separation of the dataset is measured by the sum of the squares of the distances between each center point of the class and the center point of the dataset (inter class distance). The higher the value, the better the clustering. The calculation formula for CH index is:

$$s = \frac{tr(B_k)(N-K)}{tr(W_k)(K-1)} \tag{4}$$

Where B_k is the covariance matrix between classes, W_k is the covariance matrix of data within class. The detailed formula is as follows:

$$B_k = \sum\nolimits_{q=1}^{k} n_q \left(c_q - c_e\right)\left(c_q - c_e\right)^T \tag{5}$$

$$W_k = \sum\nolimits_{q=1}^{k} \sum\nolimits_{x \in C_q} \left(x - c_q\right)\left(x - c_q\right)^T \tag{6}$$

Where c_q represents the center point of class *q*, c_e represents the center point of the dataset, n_q represents the number of data in class *q*, c_q represents a dataset of class *q*.

The Silhouette Coefficient is suitable for situations where the actual category information is unknown. For a single sample *i*, *a* (*i*) is the average distance between *i* and other samples in its category, and *b* (*i*) is the average distance between *i* and samples in different categories that are closest to it. The calculation method for the Silhouette Coefficient is as follows:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (7)$$

According to the top-2 method, attribute variables with high correlation in the insurance claims database are selected to perform k-means clustering and visualization by selecting the most relevant features of the target feature fraud.

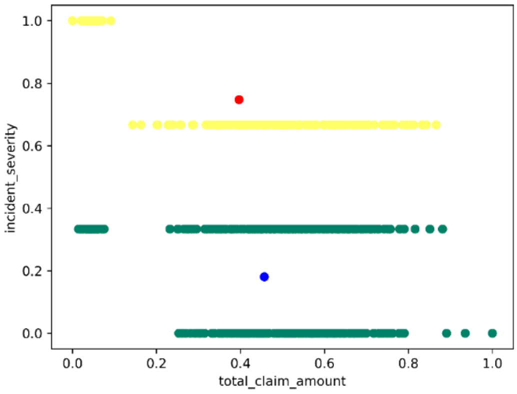

Fig. 3. K-means calculation of incident severity claim and total claim amount

Figure 3 shows the display of incident severity claim and total claim amount clustering. The evaluation indicators for these two sample attributes strongly related to insurance fraud after K-means clustering are shown in Table 1:

Table 1. K-means clustering parameter indicators

Calinski-Harabaz Index	748.2478769708888
Silhouette Coefficient	0.4960663561387431
Cluster Center	[[0.45657881 0.18077803], [0.39642576 0.747782]]

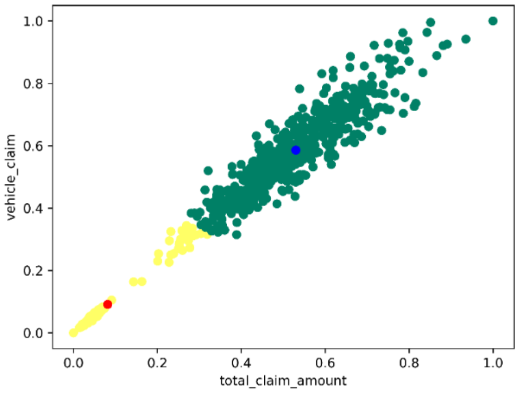

Fig. 4. K-means calculation of vehicle claim and total claim amount

Figure 4 shows the display of vehicle claim and total claim amount clustering. The evaluation indicators for these two sample attributes strongly related to insurance fraud after K-means clustering are shown in Table 2:

Table 2. K-means clustering parameter indicators

Calinski-Harabaz Index	1800.1131007504207
Silhouette Coefficient	0.6894023562976334
Cluster Center	[[0.53005815 0.58590357], [0.08168621 0.09154894]]

3.3 Classification and discrimination based on SVM

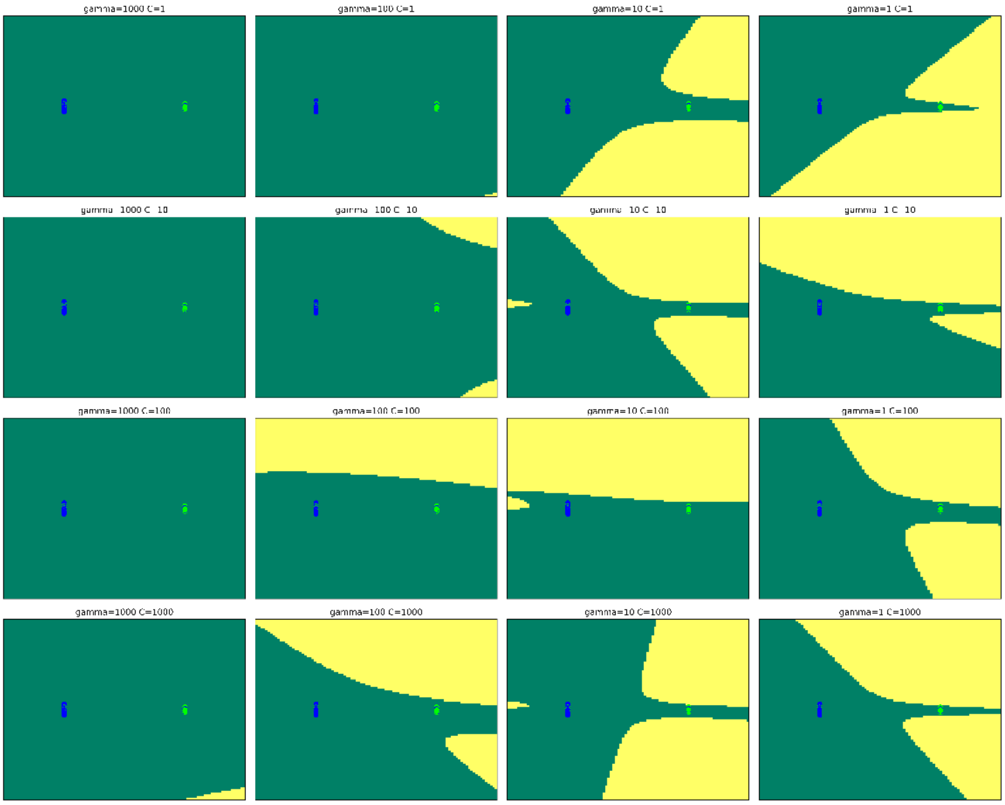

Fig. 5. SVM Gaussian Kernel Function Parameter Experiment

Support vector machines (SVM) is a binary classification model. Its basic model is the Linear classifier defined in the feature space with the largest interval, which makes it different from perceptron; SVM also includes kernel techniques, which makes it a virtually nonlinear classifier. The learning strategy of SVM is to maximize the interval, which can be formalized as a problem of solving convex Quadratic programming. This article uses an SVM classification model to identify insurance fraud in the insurance claims database. The kernel function of SVM adopts Radial Basis Function (RBF), which can map a sample to a higher dimensional space. At the same time, both large and small samples have good performance, and the kernel function has fewer parameters and high implementation accuracy.

The Gaussian kernel function formula is as follows:

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \tag{8}$$

When different values are selected for the parameters of the Gaussian kernel function, it will have different effects on the classification results. This article analyzes the significant features of the insurance claims database by testing different parameters. The test results are shown in Figure 5. It can be seen that the classification effect is best when the Gaussian kernel function parameters gamma=1 and C=1000.

4 Results & discussion

This article divides 700 samples from the insurance claims database, with a training sample set consisting of 560 samples and a validation sample set consisting of 140 samples. The sample division adopts a random selection method.

The discriminative model of insurance fraud, after extracting the attributes of samples with strong correlation between the database and insurance fraud, trained by SVM model, gets the ROC curve on the validation dataset as shown in Figure 6.

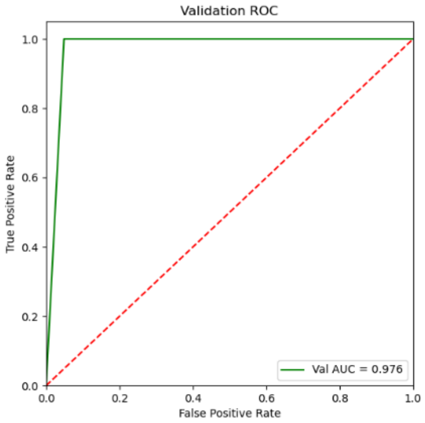

Fig. 6. SVM insurance fraud discriminative model ROC curve

The classification report of SVM discriminative model for insurance fraud are shown in Table 3.

The SVM classifier for Insurance fraud discrimination is tested through the test sample set, and the AUC value of the classifier reaches 0.976. According to the classification standard of AUC value, it has reached an excellent level. The results show that the SVM classifier has good discrimination ability in the prediction of Insurance fraud behavior. At the same time, the accuracy rate of the Insurance fraud discriminator is 0.9642, and the recall rate is 0.95, which is a good level. The results show that the Discriminative model of Insurance fraud can well predict insurance claims with fraud.

Table 3. Insurance Fraud SVM discriminative model Classification Report

	precision	recall	f1-score	support
0	1.00	0.95	0.98	103
1	0.88	1.00	0.94	37
accuracy			0.96	140
macro avg	0.94	0.98	0.96	140
weighted avg	0.97	0.96	0.96	140

This article analyzes the correlation between insurance fraud and concludes that these attribute variables, total claim amount, property claim, entry claim, vehicle claim, incident type, incident severity, conflict type, and authorities contacted, play a decisive role in the judgment of insurance fraud. Therefore, in addition to utilizing the insurance fraud judgment classification model proposed in this article, insurance companies need to pay special attention to appeal type data. At the same time, it is necessary to ensure the accuracy of the data, which is crucial for the judgment of insurance fraud.

5 Conclusion

In summary, this article investigates the problem of modeling insurance fraud discrimination based on insurance claims databases. Specifically, insurance fraud should be related to various attribute variables of the sample. However, through correlation analysis, it appears that only a small number of attribute variables have a strong correlation with insurance fraud, while the majority of other attribute variables are discriminant interference terms. At the same time, the focus of research is on utilizing the K-means clustering algorithm to achieve feature extraction of strongly correlated attribute variables. Finally, an SVM classifier is used to effectively identify insurance fraud. In the future, the continuous improvement of insurance fraud models will help avoid the occurrence of insurance fraud, ensure the positive flow of financial funds, and enhance social stability. Overall, after experimental verification and result analysis, the construction of this model will provide effective guidance for avoiding insurance fraud.

References

[1] Q. Zhu. Feature Selection Based on the Discriminative Significance for Sparse Binary-Valued and Imbalanced Dataset[J].International Journal of Pattern Recognition and Artificial Intelligence, 2023, 37(03). DOI:10.1142/S0218001423500088.

[2] A. Jeffrey, V. S. Caroline..Fraud detection in motor insurance: privacy and data protection concerns under EU Law.International Data Privacy Law, 2022(3):3. DOI:10.1093/idpl/ipac009.

[3] R. Y. Gupta, S. S. Mudigonda, P. K. Baruah. TGANs with Machine Learning Models in Automobile Insurance Fraud Detection and Comparative Study with Other Data Imbalance Techniques.

International Journal of Recent Technology and Engineering, 2021, 9(5):236-244. DOI:10.35940/ijrte.E5277.019521.

[4] A. Macedo, C. Cardoso, J. Neto, et al. Car insurance fraud: The role of vehicle repair workshops. International Journal of Law Crime and Justice, 2021, 65. DOI:10.1016/j.ijlcj.2021.100456.

[5] Skarsdóttir María, W. Ahmed, K. Antonio,et al. Social Network Analytics for Supervised Fraud Detection in Insurance. Risk Analysis, 2021(4). DOI:10.1111/risa.13693.

[6] R. Y. Gupta, S. S. Mudigonda, P. K. Baruah,.et al. Markov model with machine learning integration for fraud detection in health insurance. 2021. DOI:10.48550/arXiv.2102.10978.

[7] M. Vogels, R. Zoeckler, D. M. Stasiw, et al. P. F. Verhulst's "notice sur la loi que la populations suit dans son accroissement" from correspondence mathematique et physique. Ghent, vol. X, 1838. Journal of Biological Physics, 1975, 3(4):183-192. DOI:10.1007/BF02309004.

[8] M. Artis, M. Ayuso, M. Guillen. Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims. The Journal of Risk and Insurance, 2002, 69(3): 325-340. DOI:10.1111/1539-6975.00022

[9] D. E. Rumelhart, G. E. Hinton, R. J. Williams. PDP: Computational models of cognition and perception. Journal of Political Economy, 1986, 83-97.

[10] B. Botond, Ede Laszlo. Identifying Key Fraud Indicators in the Automobile Insurance Industry Using SQL Server Analysis Services. Studia Universitatis Babes-Bolyai Oeconomica, 2019, 64(2): 35-42. DOI: 10.2478/subboec-2019-0009

[11] E. S. Pearson, B. A. S. Snow. Tests for rank correlation coefficients. Biometrika, 1962(1-2):1-2. DOI:10.1093/biomet/49.1-2.185.

[12] L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, et al. Classification and Regression Tree. Monterey, California, U.S.A.: Wadsworth International Group, 1984. DOI: 10.2307/2530946

[13] C. Cortes, V. Vapnik. Support vector networks. Machine Learning, 1995, 20(3): 273-297. Doi:10.1007/BF00994018

[14] L. Breiman. Bagging Predictors. Machine Learning, 1996, 24(2): 123-140.

[15] T. Chen, C. Guestrin. XGBoost: A Scalable Tree Boosting System. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785- 794. DOI：10.1145/2939672.2939785

[16] G. Ke, Q. Meng, T. Finley, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30, 2017: 3146-3154.

[17] B. J. Lu, W. Z. Li, Z. N. Na, et al. Research progress of machine learning models in vehicle insurance fraud detection. Computer Engineering and Applications,2022,58(05):34-4. DOI：10.3778/j.issn.1002-8331.2109-0312

[18] S. SUBUDHI，S. PANIGRAHI. Use of optimized fuzzy C-means clustering and supervised classifiers for auto-mobile insurance fraud detection. Journal of King Saud University- Computer and Information Sciences， 2020，32（5）：568-575. DOI：10.1016/j.jksuci.2017.09.010

[19] S. K. Majhi. Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. Evolutionary Intelligence，2021, 14（1）：35-46. DOI：10.1007/s12065-019-00260-3

[20] Y. C. A，Y. LI，W. LIU，et al. An artificial bee colonybased kernel ridge regression for automobile insurance fraud identification. Neurocomputing，2020，393：115-125. DOI：10.1016/j.neucom.2017.12.072

[21] A. JB , B. Atdaf, D.Amc, A. Ms, E. Rsa. Auto loan fraud detection using dominance-based rough set approach versus machine learning methods - ScienceDirect. Expert Systems with Applications, 163. DOI：10.1016/j.eswa.2020.113740