Journal of Computers Vol. 36 No. 5, October 2025, pp. 89-101
https://doi.org/10.63367/19911599202510360500 6

Analysis of Auto Insurance Fraud Prevention Model Based on CTGAN—Stacking Model

Yuqing Zheng*

School of Mathematics and Statistics, Zhengzhou Normal University,
Zhengzhou, Henan, China

zhengyq@zznu.edu.cn

Received 1 April 2025; Revised 25 April 2025; Accepted 26 April 2025

Abstract. With the increase in the number of motor vehicles in China, the scale of the auto insurance market has expanded, and the problem of auto insurance fraud has become increasingly serious, causing huge losses to insurance companies. Data mining technology is widely used in the field of auto insurance fraud prevention, but there is a lack of a detection model with strong robustness. This paper briefly analyzes the common types of auto insurance fraud and their causes, selects the public dataset "Vehicle Insurance Fraud Detection" provided by Angoss Knowledge Seeker, uses the CTGAN method to address the data imbalance problem, and constructs integrated features by combining multiple algorithms such as random forest, XGBoost, and analysis of variance. Furthermore, a Stacking model is built. This model selects random forest, SVM, and neural network as base learners and logistic regression as the meta-learner. Research shows that compared with a single model, the Stacking model constructed after processing data with CTGAN is more efficient and accurate when dealing with imbalanced data such as fraud data. At the same time, the research also points out the current research dilemmas such as data quality issues, constraints of privacy protection laws and regulations, and application barriers.

Keywords: vehicle insurance fraud, data mining, CTGAN, stacking model

1 Introduction

With the continuous growth of China's car ownership, the scale of the auto insurance market has been expanding, and the volume of auto insurance claim settlement business has also skyrocketed. According to the calculation of the International Association of Insurance Supervisors (IAIS), globally, about 20%-30% of insurance claims are suspected of fraud each year. Auto insurance fraud in China shows a high-incidence trend. According to relevant data, the leakage of auto insurance fraud in China accounts for up to 80% of insurance fraud, and it is conservatively estimated that the annual involved amount is as high as 20 billion yuan. The forms of fraud are diverse, covering more than 30 means such as staging accident scenes, secondary collisions, deliberate insurance fraud, false reporting of theft and robbery, driver substitution in cases of drunk driving or drug driving, and duplicate claims, which have a huge impact on the operation of insurance companies.

Data mining can automatically extract hidden patterns, associations, and laws from massive and complex data [1, 2]. Through in-depth analysis of multisource data such as historical claim settlement data, policyholder information, and vehicle driving records, it can accurately identify potential fraud behavior characteristics, providing scientific and efficient risk early warnings for insurance companies. This not only helps insurance companies detect fraud risks in a timely manner and take corresponding measures to reduce compensation losses but also optimizes the allocation of insurance resources, improves the overall risk management level of the industry, and promotes the healthy and sustainable development of the insurance industry.

Abroad, the application of data mining technology in the field of auto insurance fraud prevention started early and is relatively mature [3]. The United States, at the forefront of InsurTech, has seen many insurance companies collaborate with scientific research institutions to deeply explore the application of data mining algorithms in fraud detection. By integrating diverse data sources such as insurance industry data, police traffic records, and credit agency information, large and complex fraud detection models have been constructed.

In the European region, data mining research on auto insurance fraud prevention is also highly valued.

* Corresponding Author

German insurance enterprises focus on using data mining technology to optimize risk-pricing models. While preventing fraud, they achieve precise pricing of insurance products. Through detailed analysis of vehicle historical maintenance data and vehicle owners' driving behavior data, they not only identify fraud risks but also segment policyholders according to the risk level, providing differential insurance rates for customers at different risk levels and enhancing market competitiveness. The British Insurance Industry Association has taken the lead in establishing an industry-shared database. Insurance companies pool claim settlement data, fraud cases, and other information here and use data mining tools to conduct cross-company fraud analysis. This cooperation model breaks down the data barriers between enterprises. Through comprehensive mining of industry-wide data, it enables a more comprehensive understanding of the modus operandi of fraudsters and realizes the joint crackdown on fraud.

In Asia, insurance companies in Japan and South Korea actively draw on European and American experience and promote the application of data mining in line with their domestic market characteristics. Japanese insurance companies utilize their technological advantages in the automotive manufacturing field to incorporate vehicle operation data collected by in-vehicle sensors into the analysis, such as the frequency of sudden acceleration and braking, and the stability of driving routes, to infer the authenticity of accidents and prevent fraudulent claims. The South Korean insurance market, through the combination of government promotion and industry self-discipline, has established a unified auto insurance data platform, standardizing data formats and sharing mechanisms, providing a solid foundation for data mining. Insurance companies use this platform to carry out activities such as data mining competitions, stimulating innovation, and uncovering a series of efficient fraud detection models and algorithms.

In China, with the booming development of InsurTech, data mining research on auto insurance fraud prevention has also received increasing attention. In recent years, major insurance companies have increased their investment in this field, formed professional data teams, and explored fraud detection solutions suitable for the Chinese market. Some leading insurance companies have initially established a risk control system based on data mining, integrating data from multiple departments such as claim settlement, underwriting, and customer service within the company, and using algorithms such as decision trees and neural networks to construct fraud identification models. For example, Ping An Property Insurance's self-developed "AI Intelligent Damage Assessment Cloud Platform" relies on massive auto insurance claim settlement data and deep-learning algorithm models to achieve high-precision identification of common vehicle damage types, with a damage assessment accuracy rate of over 90%. By learning from historical fraud cases, the model can assign risk scores to new claim settlement applications and screen out high-risk cases for key investigations. Cross-field cooperation has gradually become a development trend in China's auto insurance fraud prevention. Insurance companies cooperate with public security departments, transportation departments, and fintech companies to achieve data sharing and coordinated operations.

In the academic field, many scholars have conducted extensive research on motor vehicle insurance fraud detection. Viaene et al. evaluated multiple algorithms using the 1993 Massachusetts Personal Injury Protection Claim dataset, covering algorithms such as Logistic regression, C4.5 decision tree, and KNN. The research results show that the C4.5 decision tree performed poorly in this research context [4]. Phua et al. used backpropagation (BP), Naive Bayes (NB), and decision tree models for prediction and integrated these basic classifiers using the bagging technique. The results show that the model after bagging is slightly better than the single decision tree model [5]. Pérez et al. deeply analyzed the performance of classification trees in the fraud detection of auto insurance companies, compared the merged tree with the C4.5 tree, and conducted a more comprehensive analysis of errors [6].

Bhowmik used the Naive Bayes and decision tree models in auto insurance fraud detection and analyzed potential moral risks with the help of visualization tools [7]. Du et al. introduced a WOE (Weight of Evidence) encoding method within the vertical federated learning framework. Using this method, they constructed a classification task model for insurance anti-fraud and employed a series of evaluation metrics to demonstrate the effectiveness of the proposed method compared with other unsupervised coding methods. Meanwhile, it was also verified that federated learning can integrate data from multiple parties, thus improving the classification ability of anti-fraud models [8]. Kašćelan et al. applied clustering, decision tree, and regression methods to the auto insurance risk assessment dataset and found that compared with traditional standard methods, data mining technology can predict results more accurately, significantly improving the accuracy and reliability of risk assessment [9].

Subudhi and Panigrahi proposed a hybrid method for auto insurance fraud detection, which integrates the fuzzy C-means clustering technology based on the genetic algorithm and multiple supervised learning models such as DT, SVM, MLP, and GMDH [10]. This method consists of two stages: the first is the training stage,

where under-sampling is performed using FCM clustering optimized by the genetic algorithm; the second is the fraud detection stage, where four different supervised learning methods are used to verify suspicious samples. The research results fully confirm the effectiveness of this hybrid method in auto insurance fraud detection. Majhi combined the improved MWOA and FCM methods to propose an optimized clustering strategy for the auto insurance fraud identification system. This system first performs outlier removal and pruning on the sample data and then inputs the processed data into advanced classifiers such as CatBoost and XGBoost [11]. Experiments show that the system AIFDS composed of MWOA-based fuzzy clustering and CatBoost is significantly better than other methods in performance.

Yan et al. proposed an Apriori algorithm based on simulated annealing genetic fuzzy C-means (SAGFCM Apriori). The empirical research results on multiple datasets show that SAGFCM Apriori can obtain better fuzzy data in the data preprocessing process, and the proposed algorithm can reduce the mining time of association rules and improve mining efficiency [12]. Dhieb et al. proposed a secure and automated insurance system framework SISBAR, which uses a new-type insurance fraud detection system based on a permissioned blockchain and machine-learning algorithms [13]. In this framework, XGBoost is used to execute insurance services, and its performance is compared with other advanced algorithms. The experimental results show that on the auto insurance dataset, XGBoost obtains higher performance gains compared with other existing learning algorithms and demonstrates high efficiency in predicting customers' future behaviors and claim amounts.

At the application level, different countries have distinct approaches. The US combines InsurTech with data integration for fraud detection models. Germany optimizes risk pricing in Europe, while the UK creates industry-shared anti-fraud databases. In Asia, Japan uses vehicle-mounted sensor data, and South Korea depends on a government-industry platform. In China, the insurance industry invests in risk control systems and cross-sectoral cooperation is increasing. In the academic realm, scholars are dedicated to exploring more efficacious fraud detection solutions through diverse algorithms and methodologies, aiming to enhance accuracy and introduce innovative frameworks. Nevertheless, research on car insurance fraud detection models with robust algorithm-level stability remains relatively scarce. Thus, researchers should shift their focus to the data domain, conduct comprehensive analyses of diverse data features, and formulate targeted strategies to surmount the challenges posed by these features. This approach is crucial for further optimizing the model's performance.

The "CTGAN stacking model" is chosen for auto insurance fraud research for several reasons. Auto insurance data often has an imbalance issue, with few fraud samples. CTGAN can generate synthetic data, alleviating this imbalance and helping the model learn more features. The stacking model combines multiple base models and a metamodel, capturing complex data relationships to enhance fraud pattern recognition and generalization. Given the multi-source and complex nature of auto insurance data, the CTGAN stacking model's strong data processing capabilities can integrate and analyze data, providing more accurate fraud detection solutions and reducing industry losses.

2 Analysis of the Types and Causes of Auto Insurance Fraud

2.1 Common Types of Auto Insurance Fraud

False Reporting Fraud. False reporting fraud is a relatively common type of auto insurance fraud. It refers to the situation where the insured reports a claim to the insurance company when no real insurance accident has occurred or deliberately exaggerates the extent of the accident loss to obtain a higher compensation amount. This kind of fraud not only damages the interests of insurance companies but also disrupts the normal order of the insurance market. The main forms include individuals attempting to defraud insurance money by fabricating accidents to cover the costs of vehicle maintenance, and vehicle owners reporting to the insurance company that minor malfunctions or natural wear-and-tear of the vehicle are caused by traffic accidents to avoid bearing the repair costs themselves.

False reporting fraud has the following notable characteristics: First, the reporting time is rather random, usually not in line with the normal accident reporting rules. There may be a situation where a large number of reports are made when the insurance is about to expire, or the vehicle frequently has accidents shortly after being renewed. Second, the accident description is vague, and the details are unconvincing. When describing the accident process, the vehicle owner often speaks vaguely and cannot accurately provide key information such as the exact time, location, and collision angle of the accident. Third, there is a contradiction between the on-site evi-

dence and the reported content. When the surveyor conducts an on-site inspection, it is easy to find that the actual traces and physical evidence do not match the accident situation described by the vehicle owner.

Deliberate Accident-Making Fraud. Deliberate accident-making fraud is a rather malicious means of auto insurance fraud. Fraudsters usually deliberately plan and artificially create traffic accidents such as vehicle collisions and scratches with the aim of obtaining high-value insurance compensation, and they may even resort to dangerous and illegal methods to achieve their goal of defrauding insurance.

For example, a criminal gang in Chengdu, led by Hou, specifically purchased second-hand luxury cars as "props" and insured huge vehicle damage insurance with multiple property insurance companies. After insuring, they deliberately created insurance accidents that caused heavy damage or total loss of the vehicles. By carefully planning the collision scenes, they chose blind spots of surveillance or remote roads where they were less likely to be detected, and deliberately made the vehicles collide violently with other obstacles or pre-arranged vehicles to create the illusion of serious vehicle damage and then claimed compensation from the insurance companies. This gang illegally profited by earning rebates on repair costs given by the insurance companies or the price difference between the claim settlement amount and the vehicle purchase price, with the involved amount exceeding 4.6 million yuan.

In cases of drunk-driving accidents, some vehicle owners, to avoid the legal consequences of drunk driving and defraud insurance compensation, choose to find someone to take the blame. After an accident, they quickly collude with friends or relatives who have not been drinking. The latter pretend to be the driver and report the accident to the insurance company and fabricate false accident processes when the traffic police and insurance company surveyors arrive at the scene. This behavior not only seriously violates the insurance contract but also interferes with the fairness of justice and brings great difficulties to the investigation and handling of the case.

Deliberate accident-making fraud often shows some characteristics: First, there are usually many artificially arranged traces at the accident scene. For example, the collision location and angle do not conform to the mechanical principles of normal traffic accidents, the degree of vehicle damage does not match the accident description, and the brake marks and debris left at the scene are too regular or have obvious signs of forgery. Second, the behavior of the involved people is abnormal. The person taking the blame often gives inconsistent answers and looks nervous when questioned by the traffic police and the insurance company due to their lack of knowledge of the real accident situation, while the real perpetrator deliberately avoids key issues. Third, such fraud cases often involve gang-related crimes. The members have a clear division of labor, with some responsible for creating accidents, some for contacting insurance companies, and some for forging evidence materials. The criminal chain is relatively complex and highly concealed.

Duplicate Claim Fraud. Duplicate claim fraud means that the insured makes repeated claim applications to multiple insurance companies for the same insurance accident or divides the losses caused by one accident into multiple stages and makes repeated claims to the same insurance company, attempting to obtain compensation exceeding the actual losses.

The difficulties in preventing duplicate claim fraud are as follows: First, the information-sharing mechanism within the insurance industry is not perfect. The data of each insurance company is relatively independent, and there is a lack of a real-time and comprehensive information-interaction platform, which gives fraudsters an opportunity to hide the fact of duplicate claims among different companies. Second, during the claim settlement review process, insurance companies find it difficult to accurately identify whether there are staged claims in some complex cases, especially when vehicle owners provide seemingly reasonable explanations and supplementary materials, which are likely to mislead the review. Third, some fraudsters take advantage of the time lag and review loopholes in the insurance company's claim settlement process to skillfully arrange the claim time, increasing the difficulty of identification and prevention.

2.2 Exploration of the Causes of Auto Insurance Fraud

Social Credit Environment Factors. The imperfect social credit system is an important breeding ground for auto insurance fraud. In the current social environment, some individuals have a weak sense of integrity and lack respect for the spirit of contract. They regard auto insurance as a tool to obtain improper benefits. Some vehicle owners, to avoid premium payments or seek high-value compensation, do not hesitate to cross moral and legal boundaries, fabricating false accidents, exaggerating the extent of losses, and committing other forms of fraud.

Moreover, the punishment cost for fraud is relatively low when it is discovered, and the legal sanctions are insufficient. This makes some lawbreakers take chances, believing that even if their fraud is revealed, the cost they pay is within an acceptable range. Once this bad social atmosphere is formed, it is likely to trigger imitation behavior, further exacerbating the spread of auto insurance fraud.

Internal Loopholes in the Insurance Industry. Many internal loopholes in the insurance industry provide opportunities for auto insurance fraud. On the one hand, the complexity and ambiguity of insurance clauses make it easy for policyholders to have misunderstandings. Some unethical practitioners take advantage of this to mislead customers and even deliberately conceal key information, laying the groundwork for potential fraud. Some insurance clauses have obscure descriptions of claim settlement conditions and exemption scopes. Policyholders do not fully understand them when insuring and only find that they do not meet their expectations after an accident, thus inducing the motivation for fraud and attempting to obtain compensation through improper means.

On the other hand, the ineffective supervision of the claim settlement process is a key factor in the frequent occurrence of fraud. Insurance companies are not strict enough in reviewing claim settlement materials in the claim settlement link. Some staff members have insufficient professional qualities and cannot accurately identify fraud signs such as false evidence and fabricated accident scenes. Some insurance companies pay too much attention to business expansion and market share, while neglecting risk control. The "lenient entry and strict exit" business model allows fraud cases to mix into the normal claim settlement process.

In addition, the phenomenon of collusion between internal insurance industry employees and external lawbreakers also occurs from time to time. Internal employees such as loss adjusters and claim settlement staff take advantage of their positions to collude with auto repair shops and vehicle owners, defrauding insurance money by inflating loss assessment amounts and forging claim settlement materials. This kind of internal-external collusion seriously disrupts the normal order of the insurance market and exacerbates the complexity and severity of auto insurance fraud.

Information Asymmetry. In the realm of auto insurance, information asymmetry is conspicuously prominent, and it stands as one of the fundamental reasons underlying the persistent prevalence of fraud, despite numerous attempts to curtail it. A natural information disparity exists between insurance companies and customers. Customers possess in-depth knowledge of aspects such as their vehicle conditions, driving habits, and accident history. Conversely, when it comes to underwriting and claim-handling processes, insurance companies predominantly rely on the limited materials furnished by policyholders and the restricted investigation methods at their disposal to gather information. As a result, it is arduous for them to discern the actual risk profiles of vehicles comprehensively and precisely.

Certain vehicle owners capitalize on this information advantage. During the insurance-purchasing phase, they withhold crucial information like the vehicle's prior accident records and modification details to secure lower premiums. Post-accident, they inflate the extent of losses and fabricate accident causes in a bid to defraud higher compensation amounts. Moreover, the data-sharing mechanism within the insurance industry remains imperfect. The claim-settlement data, customer credit data, and other relevant information of different insurance companies are relatively insular, failing to form an effective information-sharing network. This situation presents opportunities for fraudsters to engage in repeated insurance purchases and claim-filing across multiple insurance companies. Owing to the lack of shared information, insurance companies encounter difficulties in promptly detecting fraudulent activities, leading to the continuous accumulation of fraud risks within the industry.

3 Modeling Process Based on GAN and Stacking Model

3.1 Data Introduction

In consideration of the imperative to safeguard customer privacy, domestic insurance companies exercise stringent management over user data, rendering it challenging to obtain. Consequently, this paper utilizes the publicly available dataset titled "Vehicle Insurance Fraud Detection," sourced from Angoss Knowledge Seeker, as the experimental data (Hargreaves et al. [14], Abakarim et al. [15], Ablad et al. [16]). This dataset has been extensively employed by numerous scholars both domestically and internationally in the research domain of auto insurance fraud identification, and it exhibits a certain degree of accuracy and reliability.

The dataset encompasses motor-vehicle insurance claim-settlement records from diverse regions spanning the years 1994 to 1996. Comprising a total of 15,420 data entries, it includes 14,497 legitimate claim-settlement records and 923 instances of fraud-related data. The fraud-related data accounts for 6.36% of the dataset, indicating an imbalanced data distribution. The data encompasses 32-dimensional feature variables as shown in Table 1, primarily covering accident-related information, vehicle specific details, and claimant-associated data. The target variable is denoted as "Fraud Found," which has two possible states: "Yes" and "No." "Yes" indicates the presence of fraud within the sample, while "No" indicates the absence of fraud.

Table 1. Data characteristics of the dataset

Feature name	Type
Make	Categorical
Vehicle Category	Categorical
Vehicle Price	Categorical
Age Of Vehicle	Categorical
Number Of Cars	Categorical
Fault	Categorical
Witness Present	Categorical
Accident Area	Categorical
Address Change-Claim	Categorical
Police Report Filed	Categorical
Sex	Categorical
Age	Numerical
Marital Status	Categorical
Driver Rating	Categorical
Past Number Of Claims	Categorical
Age Of Policy Holder	Categorical
Policy Type	Categorical
Deductible	Numerical
Policy Number	Numerical
Rep Number	Numerical
Agent Type	Categorical
Number Of Suppliments	Categorical
BasePolicy	Categorical
Year	Numerical
Month	Numerical
WeekOfMonth	Numerical
DayOfWeek	Numerical
Month Claimed	Numerical
Week Of Month Claimed	Numerical
Day Of Week Claimed	Numerical
Days: Policy-Accident	Categorical
Days: Policy-Claim	Categorical

3.2 Data Preprocessing

Elimination of Irrelevant Features. The process of eliminating irrelevant features involves removing those feature variables from the original dataset that contribute minimally to the outcomes during the modeling and prediction procedures. This operation serves multiple purposes: it enhances the computational efficiency of the

model, mitigates the risk of overfitting, and simplifies the interpretability of the model. In the dataset utilized in this study, the feature variables slated for removal fall into two categories. The first category consists of truly irrelevant features, such as the policy number (Policy Number). The second category encompasses redundant and complex data, mainly involving time-series feature information, specifically 'Month', 'Year', 'Week of Month', 'Day of Week', 'Month Claimed', 'Week of Month Claimed', and 'Day of Week Claimed'. Given the chaotic nature of this time-related feature information and the fact that the dataset already contains the time-interval features between policy-related accidents and claims, these two groups of time-series feature information can be safely removed.

Data Transformation and Standardization. Data transformation generally refers to the process of converting or manipulating data in a specific manner to optimize its suitability for the modeling task at hand. The dataset employed in this paper contains a substantial number of categorical variables. However, machine-learning algorithms often encounter difficulties in directly processing such categorical data. Therefore, it is essential to encode the original data and transform it into a numerical format that can be readily accepted by classifiers. This ensures that the model can accurately interpret and effectively utilize these features, thereby enhancing the model's predictive performance. In this paper, the one-hot encoding technique is employed for data transformation.

The dataset used in this research also exhibits the issue of inconsistent data scales. In data-mining operations, data with larger scales tend to dominate the analysis. Hence, data standardization is a necessary step. This paper adopts the Z-score standardization method, which transforms the data into a standard normal distribution with a mean of 0 and a standard deviation of 1 through linear transformation.

3.3 Modeling Analysis of Imbalanced Data

Following data preprocessing, the random-forest algorithm is applied to analyze the data. To mitigate the issues of overfitting and data imbalance, the data is processed using a pruning strategy and a multi-stage approach. We employ multiple evaluation metrics to comprehensively assess the quality and performance of the current model.

Accuracy represents the proportion of correctly predicted samples to the total number of samples, calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, where TP (True Positive) refers to the number of samples that are actually positive and predicted as positive; TN (True Negative) denotes the number of samples that are actually negative and predicted as negative; FP (False Positive) indicates the number of samples that are actually negative but predicted as positive; and FN (False Negative) signifies the number of samples that are actually positive but predicted as negative. However, in imbalanced datasets, accuracy can be misleading, necessitating the consideration of other metrics for a complete model performance evaluation.

Precision, calculated as $\frac{TP}{TP+FP}$, is the ratio of true positives among the samples predicted as positive, measuring the accuracy of positive predictions. Recall, expressed as $\frac{TP}{TP+FN}$, represents the proportion of true positives among the actual positive samples, reflecting the model's ability to correctly identify positive instances. The F1-score, given by $2\times\frac{Precision\times Recall}{Precision+Recall}$, serves as the harmonic mean of precision and recall. It balances these two metrics, providing a more comprehensive assessment of the model's performance.

The specific results are presented as illustrated in Table 2.

Table 2. Prediction results of the random forest model (imbalanced data)

Evaluation index	Score
Accuracy	0.340636
Precision	0.146272
Recall	0.792746
F1-score	0.246973

The accuracy of this random-forest model is relatively low. A value of merely 0.34 implies that the proportion of correctly predicted samples by the model, overall, is not substantial. In the context of auto-insurance fraud

prevention, this may suggest that the model is prone to making numerous misjudgments when differentiating between fraudulent and non-fraudulent cases. The precision is equally low, indicating that among the cases predicted as fraudulent by the model, the proportion of genuinely fraudulent cases is rather small. In other words, the model may misclassify a significant number of normal cases as fraudulent, which, in practical applications, will impose additional investigation costs on insurance companies and cause inconvenience to customers. The recall rate, on the other hand, is relatively high, signifying that the model can identify most actual fraud cases. However, when considered in conjunction with the low precision, it appears that this may be achieved at the expense of accuracy, with many normal cases being misidentified as fraud cases. The F1-score, which combines precision and recall, is low due to the significant disparity between these two metrics. This indicates that the overall performance of the model requires improvement, and a more optimal balance needs to be struck between precision and recall. Overall, the predictive ability of this model is subpar, and its quality is unsatisfactory.

The factors contributing to the poor performance of the model can be attributed to two aspects: the inherent imbalance of the data and the selection of a single-model approach. Insurance fraud is typically a binary-classification problem, i.e., determining the presence or absence of fraud. The data used in this paper is imbalanced, with fraud-related data constituting only 6.36% of the total, resulting in a relatively small number of fraud cases. Data imbalance is thus a critical issue. Given the data imbalance, it is challenging to effectively mine data features using a single-model strategy, and the construction of combined features may be necessary.

3.4 Constructing the Stacking Model

Feature-Set Construction Process. Based on the foregoing analysis, it is imperative to address the data-imbalance problem initially. Conventionally, oversampling methods such as SMOTE and generative-adversarial-network-based methods are employed for data balancing (Chawla et al. [17]). Compared with traditional oversampling techniques, generative adversarial networks (GANs) have the advantage of generating more complex and real-distribution-like samples, which can enhance the model's ability to recognize minority classes (Goodfellow et al. [18]). Therefore, in this paper, GAN is utilized to process the data. A GAN is composed of two sub-networks: a generator and a discriminator. The generator is tasked with generating new data samples, while the discriminator assesses the similarity between the generated samples and the real data. In this study, the CTGAN is trained to generate fraud-case samples, aiming to balance the dataset.

A single method is often insufficient for effectively mining data features. Thus, this paper explores the use of multiple methods for feature selection to generate integrated features, with the goal of identifying the features that have the most significant impact on auto-insurance fraud detection. The random-forest algorithm, XGBoost, recursive feature elimination (RFE), analysis of variance, and correlation analysis are employed in combination to generate integrated features. Each of these four algorithms has its unique strengths. The random-forest algorithm evaluates the importance of features by calculating the average importance score of each feature across all decision trees, enabling it to handle high-dimensional data and non-linear relationships effectively (Breiman [19]). XGBoost, an efficient gradient-boosting framework, calculates feature importance based on the influence of features on the target variable during the training process, allowing it to capture complex feature-target variable relationships (Bentéjac et al. [20]). The RFE algorithm starts with all available features and iteratively removes the features that have the least impact on the model's performance until a preset number of features is reached, thereby identifying the feature subset that maximizes the model's prediction effectiveness (Guyon et al. [21]). The analysis of variance determines whether the values of a feature vary significantly by calculating its variance, helping to identify features that contain more information. Finally, correlation analysis is conducted to identify linear relationships between features, and highly correlated redundant features are removed. The flowchart of this process is shown in Fig. 1, and the importance results of specific classification features are shown in Fig. 2. This process constructs integrated features and reduces the complexity of the model.

Stacking Model Construction. The Stacking model integrates multiple individual models to comprehensively uncover data features and patterns, aiming to enhance predictive performance. It consists of base learners and a meta learner. When selecting learners, two key aspects are considered.

For base learners, models with diversity and high predictive accuracy are preferred. The random forest can handle high-dimensional and non-linear data well. SVM excels in small-sample and non-linear classification by finding the optimal hyperplane. Neural networks have strong learning and generalization capabilities. Thus, these three are chosen as base learners in this study.

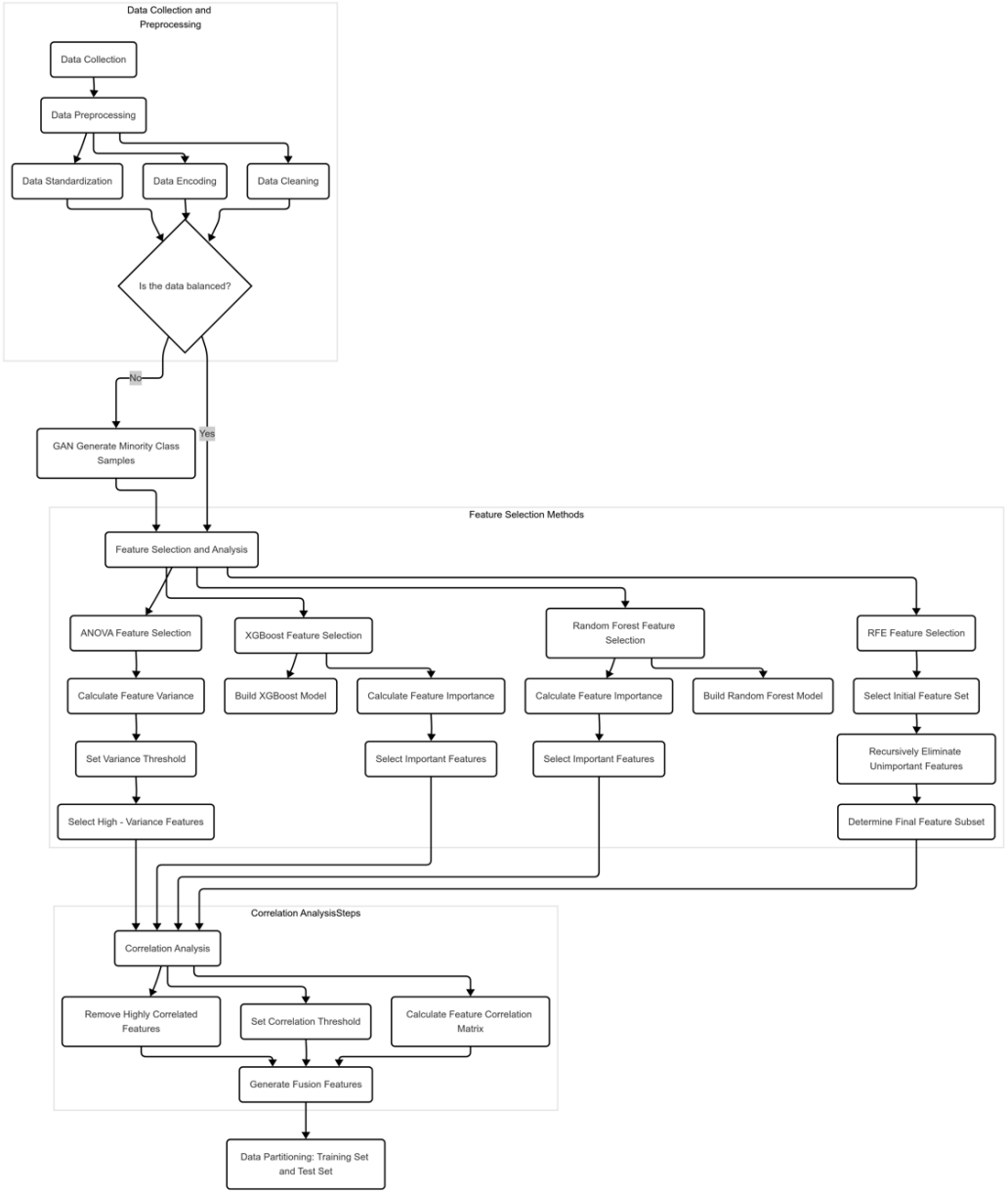

Fig. 1. Flowchart of feature set construction

Regarding the meta learner, a simple and stable model is favored. The logistic regression model, with its simplicity and interpretability, can effectively integrate base learners' outputs. So, it is selected as the meta-learner.

The algorithm process, as shown in Fig. 3, includes:

- Dataset Partition: Split the balanced dataset into a training set (Train1) and a test set (Test1) at an 8:2 ratio.
- Base Learners' Training and Prediction: Train the random forest, SVM, and neural network using the training set, and then use them to predict on the training set.
- Meta-learner Training and Prediction: Use the base learners' prediction results as the input for the logistic regression (meta-learner) and train it with the training set.
- Model Evaluation: Evaluate the Stacking model's performance using accuracy, precision, recall, and F1-score.

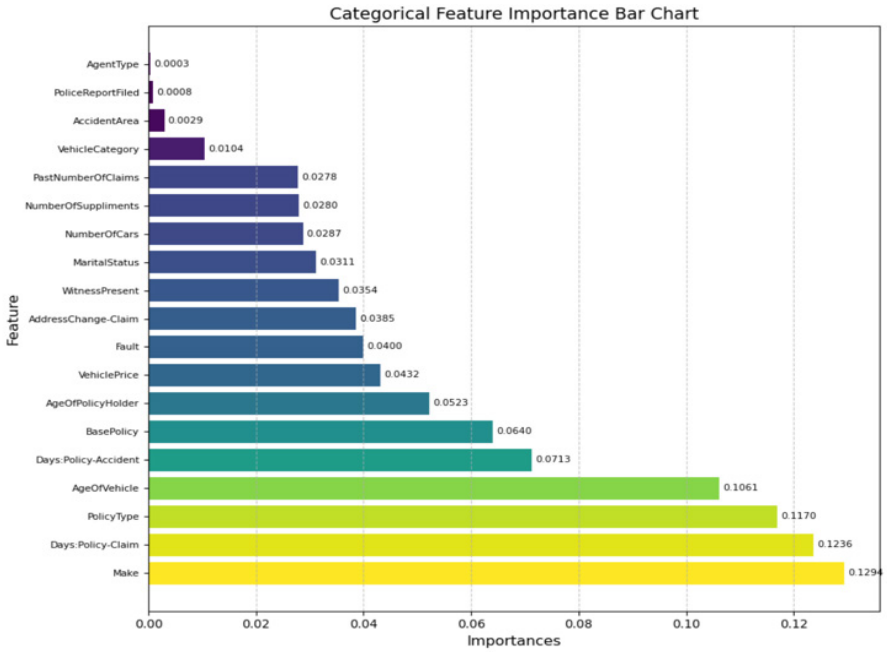

Fig. 2. Categorical feature importance bar chat

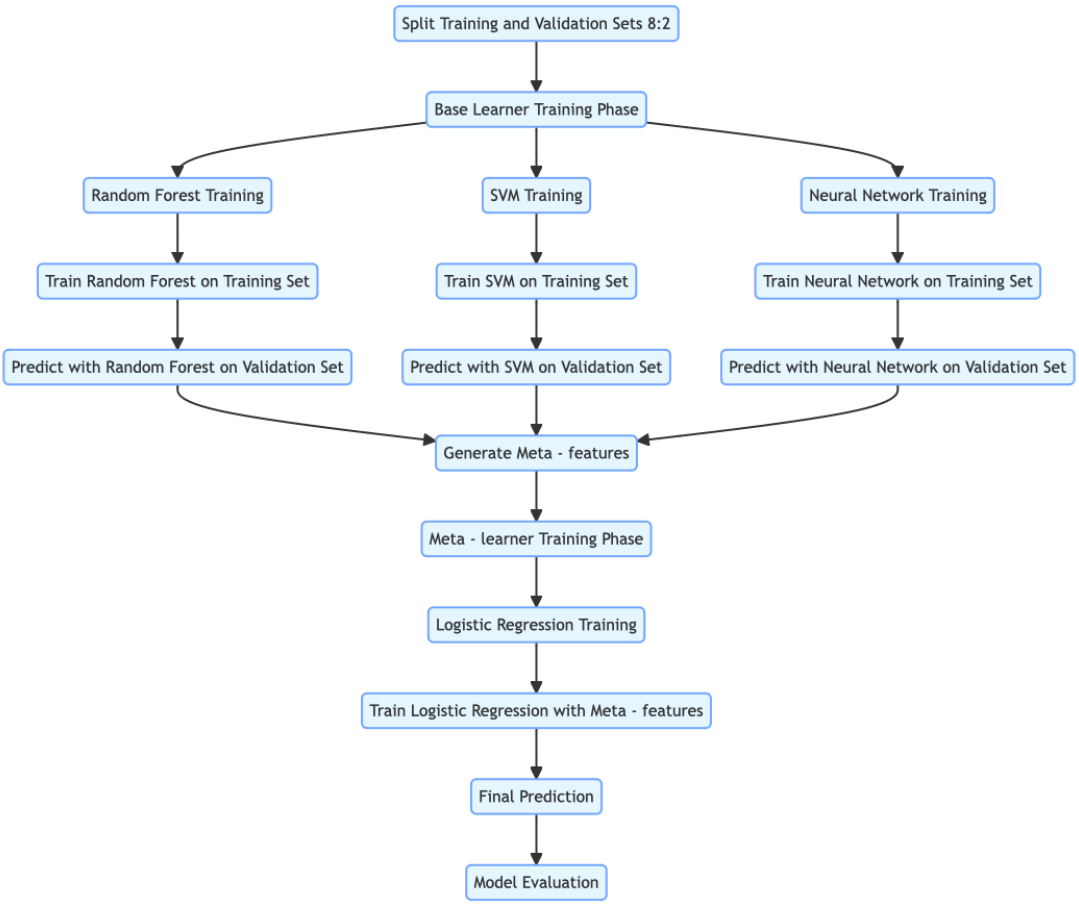

Fig. 3. Flowchart of the stacking model

Evaluation metrics are criteria used to measure the performance of machine learning models on specific tasks and are crucial for assessing a model's predictive ability, generalization ability, and overall effectiveness. In traditional classification tasks, accuracy is usually widely used as a benchmark for measuring model performance. However, when dealing with imbalanced class data, relying solely on accuracy to evaluate a model's performance may be inaccurate and incomplete. This paper uses accuracy, precision, recall, and F1-score as the criteria for model evaluation. The specific evaluation scores of the model are shown in Table 3.

Table 3. Evaluation of the stacking model

Evaluation metrics	Stacking model
Accuracy	0.9484393861010519
Precision	0.9479848432655873
Recall	0.9489655172413793
F1 Score	0.9484749267620196

4 Conclusion and Challenges

This research undertakes an in-depth exploration of the types and causes of auto insurance fraud. It clearly defines common fraud types, including false reporting, intentional accident fabrication, and duplicate claims, and uncovers the underlying factors such as deficiencies in social integrity, loopholes within the industry, and information asymmetry. Leveraging empirical research methods, this study comprehensively presents the application process of the Stacking model in auto insurance fraud prevention. This process encompasses multiple crucial stages, ranging from the selection of data sources, data cleaning and transformation, and the construction of feature engineering, to model selection, training, and the establishment of evaluation indices.

By meticulously comparing evaluation metrics such as accuracy, precision, recall, and the F1-score, it has been determined that when dealing with imbalanced data related to fraud, constructing a Stacking model subsequent to data processing using the CTGAN approach proves to be a more efficient and accurate strategy compared to relying solely on a single model. This finding underscores the superiority of the proposed method in enhancing the effectiveness of auto insurance fraud detection.

Challenges for Future Research. The quality of data generated by CTGAN is a crucial aspect. Although CTGAN can generate synthetic data, insurance data usually consists of complex tabular data, such as a mixture of numerical and categorical features, and may also contain time-series information. Ensuring that the generated data is both realistic and privacy-preserving while maintaining consistency in statistical properties can be challenging. Additionally, it needs to be verified whether the generated data can truly enhance the performance of detection models, especially in cases of extreme data imbalance.

Then, there is the complexity of the stacking model itself. Integrating multiple base models and metamodels may lead to increased computational costs, especially when dealing with large-scale insurance data. The interpretability of the model is also an issue. Since stacking models are generally complex, insurance companies may require transparent decision-making processes to meet regulatory requirements. Further research may be needed to explore how to improve the interpretability of the model while maintaining high accuracy.

Data Quality Issues. In the process of case-based analysis, data quality problems emerge as significant obstacles to data feature extraction. Inaccurate data has the potential to misdirect the model, causing it to learn incorrect fraud patterns. For example, erroneous claim amounts can lead the model to misinterpret the probability of fraud. Incomplete data, lacking essential information, poses challenges to the precise identification of fraud features. For instance, the absence of policyholders' occupation information may prevent the detection of occupation-specific fraud risks. Inconsistent data, on the other hand, not only increases the complexity and cost of data preprocessing but also reduces the efficiency of model training. In severe cases, it can even result in incorrect judgments during model application, thereby undermining the accuracy and timeliness of auto insurance fraud detection. These data quality issues stem from a combination of human-induced errors and management-related shortcomings, and they exert multiple adverse effects on fraud detection efforts.

Constraints Imposed by Privacy-Protection Laws and Regulations. In an era of heightened awareness regarding data privacy protection, a plethora of stringent laws and regulations have been promulgated both domestically and internationally. China's "Personal Information Protection Law" explicitly mandates that insurance companies must obtain explicit consent from customers when collecting, using, or storing auto insurance-related customer data. They are also required to adhere to the principle of minimum necessity, ensuring that data collection is restricted to what is directly relevant to business operations. Additionally, the "Cybersecurity Law" obligates insurance companies to fortify their data security protection measures to safeguard against data leakage and tampering.

On the international front, the European Union's "General Data Protection Regulation" (GDPR) serves as a prominent example. It confers numerous rights upon data subjects, including the right to be informed, the right of access, and the right to be forgotten. Moreover, it imposes strict restrictions on cross-border data transfer. Even multinational insurance companies must comply scrupulously with these regulations when handling the auto insurance data of customers within the EU.

Under the purview of these regulatory frameworks, insurance companies encounter substantial challenges when leveraging data-mining techniques for auto insurance fraud prevention. Firstly, data acquisition has become more restrictive. Certain potentially sensitive data sources, which were previously accessible through third-party channels, are now difficult to obtain in a compliant manner. For example, vehicle owners' travel records and consumption preferences data sourced from social media platforms, although valuable for fraud risk assessment, cannot be incorporated into analysis due to privacy concerns. Secondly, data sharing has been impeded. Insurance companies could potentially enhance their collective risk-control capabilities by sharing fraud cases and high-risk customer information. However, privacy regulations have complicated the data-sharing process. To meet compliance requirements, comprehensive data de-identification and encryption measures are necessary. Failure to do so may expose companies to legal risks, thereby hampering the efficiency of joint efforts in auto insurance fraud prevention.

To mitigate these challenges, insurance companies can adopt encryption technologies to safeguard sensitive information, such as policyholders' ID numbers and bank card details, during data storage and transmission, ensuring data security throughout its lifecycle. Anonymization techniques can be employed prior to data-mining analysis to remove personally identifiable information, transforming the data into an anonymous dataset. This approach not only protects customer privacy but also preserves the statistical characteristics of the data for use in fraud-detection model training. By doing so, insurance companies can fully exploit the value of data within the framework of compliance, thereby achieving a balance between auto insurance fraud prevention and privacy protection.

Application Barriers. In practical applications, business personnel typically rely on the risk scores predicted by models to identify auto insurance claim cases with a high likelihood of fraud. However, they often lack insight into the specific features utilized by the model and the underlying logical reasoning behind the predictions. This lack of transparency creates a knowledge gap, which in turn generates hesitancy among business personnel when conducting subsequent investigations and making decisions based on model outcomes. As a result, the trust in these models is eroded, and this has a detrimental impact on the efficient advancement of auto insurance fraud-prevention initiatives.

Acknowledgement

The work described in this paper was supported by the Key Project of Henan Provincial Education Science Planning in 2025 (No. 2025JKZD36).

References

[1] M. Singh, G. Jindal, A. Oberoi, R. Dhangar, Improving Crime Detection Through Geo-MDA: A Hybrid Linear Regression Approach in Data Mining. International Journal of Performability Engineering 20(8)(2024) 469–477. https://doi.org/10.23940/ijpe.24.08.p1.469477

[2] M. Agarwal, B.B. Agarwal, Methodical Implementation of Data Mining Classifiers and ANN for Prediction of Accomplishment of Student Education, International Journal of Performability Engineering 19(9)(2023) 587–597. https://doi.org/ 10.23940/ijpe.23.09.p4.587597

[3] Z. Wang, L. Li, B. Wang, R. Luo, L. Ye, The Application of the Teaching Model of "Promoting Learning Through Competition, Promoting Teaching Through Competition" in the Course of Data Analysis and Mining, in: Proc. 2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion, 2024. https://doi.org/10.1109/QRS-C63300.2024.00146

[4] S. Viaene, R.A. Derrig, B. Baesens, G. Dedene, A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection, Journal of Risk and Insurance 69(3)(2022) 373–421. https://doi.org/10.1111/1539-6975.00023

[5] C. Phua, D. Alahakoon, V. Lee, Minority report in fraud detection: classification of skewed data, ACM SIGKDD Explorations Newsletter 6(1)(2004) 50–59. https://doi.org/10.1145/1007730.1007738

[6] J.M. Pérez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, J. Martín, Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance, in: Proc. Pattern Recognition and Data Mining (ICAPR 2005), 2005. https://doi.org/10.1007/11551188_41

[7] R. Bhowmik, Detecting Auto Insurance Fraud by Data Mining Techniques, Journal of Emerging Trends in Computing and Information Sciences 2(4)(2011) 156–162. http://www.cisjournal.org/Download_April_pdf_1.aspx

[8] W. Du, H. Wang, J. Shen, G. Meng, H. Li, W. Zhou, Insurance Anti-fraud based on FL-WOE Encoding for Vertical Federated Learning, in: Proc. 2024 IEEE International Conference on Big Data (BigData), 2024. https://doi.org/10.1109/BigData62323.2024.10825724

[9] L. Kašćelan, V. Kašćelan, M. Novović-Burić, A data mining approach for risk assessment in car insurance: evidence from Montenegro, in: Big Data: Concepts, Methodologies, Tools, and Applications, IGI Global Scientific Publishing, 2016. https://doi.org/10.4018/978-1-4666-9840-6.ch094

[10] S. Subudhi, S. Panigrahi, Use of optimized fuzzy c-means clustering and supervised classifiers for automobile insurance fraud detection, Journal of King Saud University-Computer and Information Sciences 32(5)(2020) 568–575. https://doi.org/10.1016/j.jksuci.2017.09.010

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

[12] C. Yan, J. Liu, W. Liu, X. Liu, Research on automobile insurance fraud identification based on fuzzy association rules. Journal of Intelligent & Fuzzy Systems 41(6)(2021) 5821–5834. https://doi.org/10.3233/JIFS-201301

[13] N. Dhieb, H. Ghazzai, H. Besbes, Y. Massoud, A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement, IEEE Access 8(2020) 58546–58558. https://doi.org/10.1109/ACCESS.2020.2983300

[14] C.A. Hargreaves, V. Singhania, Analytics for insurance fraud detection: an empirical study, American Journal of Mobile Systems, Applications and Services 1(3)(2015) 227–232.

[15] Y. Abakarim, M. Lahby, A. Attioui, A Bagged Ensemble Convolutional Neural Networks Approach to Recognize Insurance Claim Frauds, Applied System Innovation 6(1)(2023) 20. https://doi.org/10.3390/asi6010020

[16] M. Ablad, B. Frikh, B. Ouhbi, Uncertainty quantification in deep learning context: Application to insurance, in: Proc. 2020 6th IEEE congress on information science and technology (CiSt), 2020. https://doi.org/10.1109/CiSt49399.2021.9357201

[17] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research (16)(1)(2002) 321–357.

[18] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proc. Advances in neural information processing systems, 2014. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf

[19] L. Breiman, Random Forests, Machine Learning 45(1)(2001) 5–32. https://doi.org/10.1023/A:1010933404324

[20] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artificial Intelligence Review 54(3)(2021) 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[21] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 46(1-3)(2002) 389–422. https://doi.org/10.1023/A:1012487302797