Research on Detecting Auto Insurance Fraud

Yuexin Quan*

Stamford American International School, Singapore, 357684, Singapore

*Corresponding Author Email: 24yuexinq@sais.edu.sg

Abstract. With the expansion of the insurance sector has come an increasing concern about insurance fraud. This issue has grown in prominence over time, necessitating the implementation of appropriate countermeasures. In response, this work conducts a thorough analysis of the available literature, with an emphasis on the incorporation of machine learning approaches for detecting fraudulent actions in the field of automobile insurance. By delving into the intricacies of the vehicle insurance fraud landscape, the research investigates the strategic deployment of expert systems and machine learning models to enhance fraud identification processes. Drawing from the wealth of high-quality auto insurance data spanning the last five years, this work explores the implementation of diverse machine learning algorithms, including Logistic Regression, Decision Tree, and Discriminative Analysis, in order to construct comprehensive predictive models. After the modeling was completed, comprehensive testing and analysis were conducted, and the results showed that the logistic regression model had the highest accuracy among these three models. Finally, the future research direction of automobile insurance fraud detection technology is discussed, and how to reasonably prevent automobile insurance fraud is proposed.

Keywords: Logistic regression, decision trees, discriminant analysis, auto insurance.

1. Introduction

In today's society, automobile insurance fraud has become one of the major challenges facing the insurance industry. Fraud not only causes financial losses to insurance companies but also jeopardizes the level playing field in the insurance market. The "disaster area" of insurance fraud has been auto insurance fraud, which is critical to the development of the insurance business. As a result, auto insurance fraud detection technology has become a hot topic among academics both at home and abroad. During times of economic uncertainty or recession, some people may try to obtain additional insurance coverage through fraudulent means because of increased financial stress. By analyzing fraud data, we can contribute to our understanding of patterns and trends in their fraudulent behavior. Also, by analyzing these data, key characteristics of fraudulent behavior can be identified so that fraud can be prevented. Groups characterized by fraudulent behaviors are subject to increased scrutiny, stricter claims review processes, and increased data sharing and collaboration. In the subsequent sections of the paper, we will present the fundamentals of each model and then evaluate their performance on fraud detection tasks through experiments and data analysis. We will consider metrics such as the accuracy, recall, and precision of the models, as well as their stability and interpretability. Through these comparisons, we hope to provide decision support to insurance companies regarding the selection of appropriate fraud detection models.

2. Literature Review

According to conservative estimates, at least 20% of China's claims for motor insurance are fraudulent [1]. Comparing vehicle insurance fraud to other forms of insurance fraud in China reveals how it differs from the others in several key ways. Notably, it often involves criminal concealment methods, employs a wide range of deceptive techniques, and is progressively manifesting as a modus operandi driven by organized criminal networks [2]. The landscape of insurance fraud is being reshaped by the evolving tactics employed in auto insurance fraud.

Moreover, the application of machine learning models to the prediction of auto insurance losses has gained traction due to several key factors. Machine learning models can accommodate raw data

with relatively minimal pre-processing requirements, a particularly pertinent advantage in the insurance context [3]. Additionally, the intricate interactions and non-linear relationships between various influencing factors can be effectively captured by machine learning models. Consequently, they have demonstrated superior predictive capabilities in estimating auto insurance losses. As the sophistication of these models grows, so does their potential impact on improving risk assessment and fraud detection strategies within the auto insurance domain. The detection of motor insurance fraud was thoroughly studied by Panigrahi et al., utilizing feature selection algorithms and machine learning methods [4]. Utilizing three separate feature selection algorithms, they concentrated on removing fundamental features from data on auto insurance fraud. The chosen characteristics were then subjected to a variety of machine learning algorithms in an effort to identify the most effective feature selection strategy for each model. Notably, the authors acknowledge that the study is hampered by the comparatively low incidence of auto insurance fraud in comparison to more widespread types, such as credit card fraud. They may need to take this into account in their study because it led to a major category imbalance. Using deep learning networks and hybrid modeling techniques, Yan, Yu, Xu, and their colleagues have achieved substantial advancements in the area of detecting auto insurance fraud [5]. Their work offers a fresh approach to this issue that might enhance the precision and effectiveness of fraud detection systems. The current state of the art in the field of fraud detection, where cutting-edge technologies like deep learning play a critical role, appears to be very much in line with this approach. It is interesting to note that Tuo Guozhu, Liu Xihua, and other researchers started out by exploring the theoretical foundations of auto insurance. Their investigation could have offered crucial contextual insights that advance knowledge of the complexities involved in identifying and preventing vehicle insurance fraud. This theoretical investigation might have served as the basis for later fieldwork implementation and empirical studies. Additionally, Guiping and his associates compiled a sizable body of international research on moral hazard in auto insurance before summarizing and drawing conclusions from their findings [6]. This research endeavor is probably going to help the body of information and best practices grow, which will result in a more thorough understanding of the moral and risk-related aspects of auto insurance fraud. Abroad, S. Viaene et al. [7], Hanafizadeh et al. [8], Kašćelan et al. [9], and Li et al. [10] have separately explored Bayesian modeling, clustering modeling, data mining, randomized Forest, and other technical techniques for auto insurance fraud detection.

Collectively, these efforts represent a multifaceted approach to advancing automobile insurance fraud detection. From feature selection and machine learning to deep learning network exploration and theoretical research, they collectively contribute to a comprehensive understanding of the field. This comprehensive approach lays the foundation for making informed decisions and developing effective fraud detection strategies in the automobile insurance domain.

3. EDA correlation

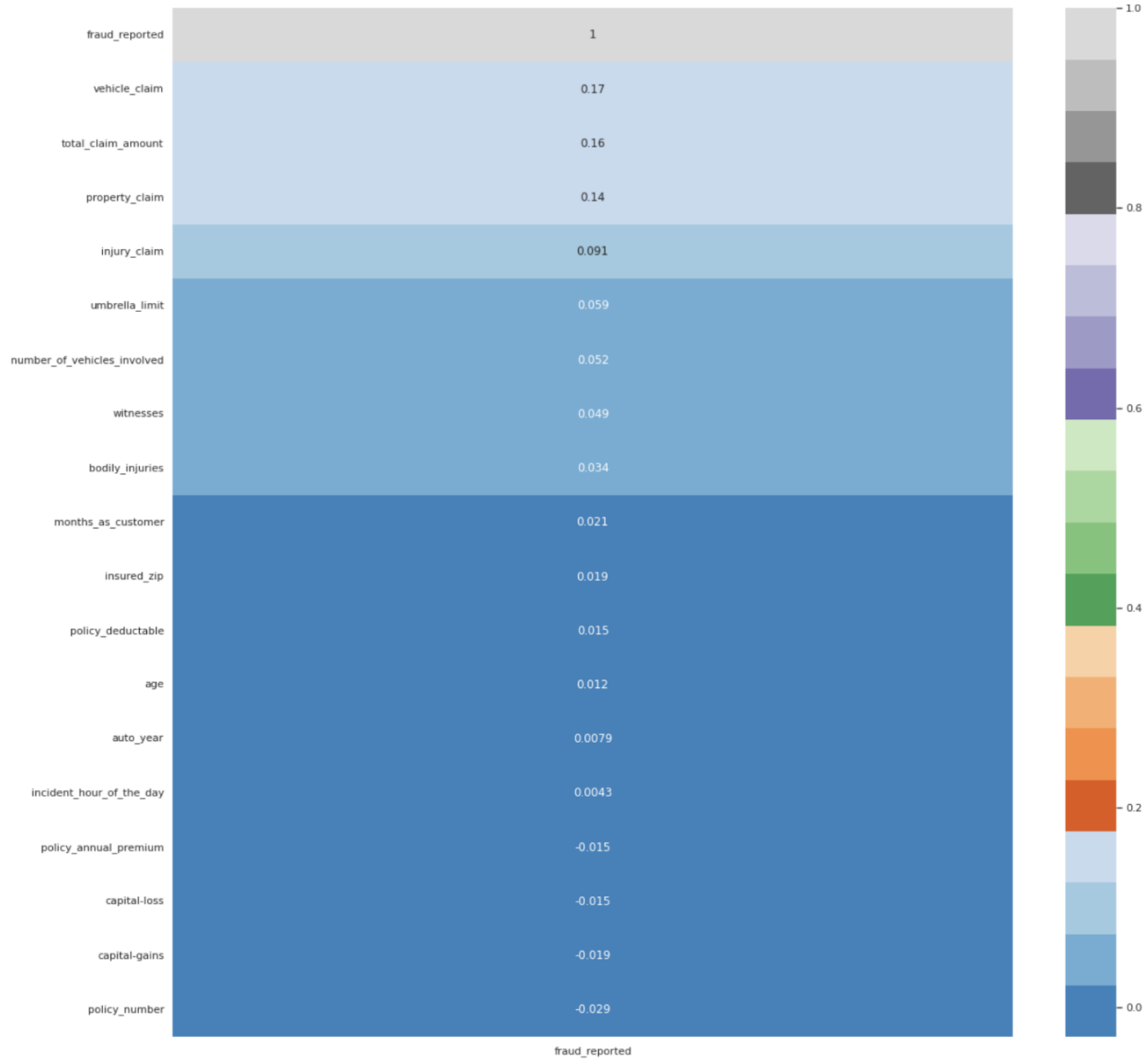


Fig. 1 Correlation Heatmap

The data on vehicle insurance fraud comes from the "Automobile Insurance Claims Fraud Detection" dataset on the Kaggle website. This dataset is an example dataset for fraud detection and contains a fictitious dataset of vehicle insurance claims. For the processing of the data first, the data to be submitted needs to be collected and organized. The data is usually in the form of a two-dimensional matrix, where the rows represent one dimension, i.e. the presence of fraud. The columns represent another dimension, i.e., information about the characteristics of the insurance policy. After the heat map is drawn, one can begin to interpret and analyze the results. By looking at the distribution of colors, patterns, relationships, and outliers in the data can be identified. This helps in extracting valuable insights from the data. With Figure 1, the correlation relationship between the insurance policy feature information and the presence of fraud can be seen.

4. Methods

4.1. Logistic regression

One popular statistical method that is very effective for binary classification issues is logistic regression. Because of this, it is a desirable option for fraud detection assignments. The basic idea

underlying logistic regression is to fit a logistic curve to the input features in order to describe the probability of a binary outcome. This curve applies the logistic function to convert the linear combination of input features, ensuring that the projected probabilities fall between 0 and 1. The coefficients for each input characteristic estimated by the logistic regression model show the influence of each feature on the likelihood of the desired outcome.

In the context of auto insurance fraud detection, logistic regression offers interpretability and simplicity. By quantifying the influence of each feature on the likelihood of fraud occurrence, insurers can gain insights into the factors contributing to potential fraudulent activities. The relatively low complexity of logistic regression also makes it computationally efficient, suitable for real-time applications in the insurance industry.

4.2. Decision trees

Non-linear models called decision trees divide the feature space into regions based on the values of the input characteristics. Each partition is governed by a set of rules that divide the data into subsets iteratively until leaf nodes representing class labels are produced. Because they can efficiently handle non-linear relationships and capture complicated interactions among features, decision trees are promising for fraud detection.

In the auto insurance context, Decision Trees can help identify intricate patterns indicative of fraudulent behavior. By selecting the most discriminative features at each branching point, Decision Trees can excel at detecting fraud cases that might not adhere to linear relationships. Furthermore, the hierarchical nature of Decision Trees allows for a transparent representation of decision-making processes, enhancing the model's interpretability.

4.3. Discriminant analysis

Discriminant Analysis encompasses techniques designed to find a discriminant function that maximizes the separation between classes in the feature space. For example, Linear Discriminant Analysis (LDA) constructs a linear combination of features that optimally separates different classes. Quadratic Discriminant Analysis (QDA) allows for quadratic decision boundaries, accommodating more complex data distributions.

Choosing Discriminant Analysis for auto insurance fraud detection stems from its ability to handle the class imbalance issue. Given that auto insurance fraud is less prevalent than typical fraud cases, the pronounced category imbalance poses a challenge for accurate detection. Discriminant Analysis aims to find decision boundaries that maximize the separation between classes, which can be advantageous in scenarios where fraud cases are underrepresented.

5. Results

5.1. Logistic regression

Table 1. The results of the Logistic Regression model

	precision	recall	f1-score	support
N	0.86	0.88	0.87	226
Y	0.59	0.55	0.57	74
Accuracy			0.80	300
Macro avg	0.73	0.72	0.72	300
Weighted avg	0.79	0.80	0.79	300

Based on the findings of the Logistic Regression model, the model performed as follows in this automated insurance fraud detection test. The model's accuracy is 0.796, indicating that it has a relatively good success rate in correctly categorizing the samples. It is crucial to note, however, that

the accuracy rate is not the sole statistic of interest, because in the situation of category imbalance, the accuracy rate may be pushed up by a high percentage of true negative samples.

The model has a precision rate of 0.594, which means that about 59.4% of all samples predicted by the model to be positive (fraud) are in fact actual cases of fraud. The recall rate is 0.554, indicating that the model correctly identified 55.4% of the genuine fraud cases. The balance of precision and memory is frequently a key problem, since increasing precision may result in a loss in recall and vice versa.The F1 Score of 0.573, which combines precision and recall, is a key indicator for assessing the model's overall performance.

The classification of the model is displayed in the confusion matrix. The program properly identified 198 samples in the non-fraud category (N), but wrongly identified 28 samples as fraudulent. The model accurately identified 41 samples in the fraud category (Y), but misidentified 33 samples in the non-fraud group. This information aids in comprehending how the model performs across the various categories.

5.2. Decision trees

Table 2. The results of the decision tree model

	precision	recall	f1-score	support
N	0.88	0.84	0.86	226
Y	0.57	0.64	0.60	74
Accuracy			0.79	300
Macro avg	0.72	0.74	0.73	300
Weighted avg	0.80	0.79	0.79	300

The results based on the decision tree model show that the model performs as follows in the automated insurance fraud detection task. The accuracy of the model on the training data is 84.82%, while the accuracy on the test data is 79.00%. This means that the model is still able to achieve relatively good predictive performance when dealing with unseen data.

The confusion matrix shows the classification of the model. In the non-fraud category (N), the model correctly classified 190 samples, but incorrectly classified 36 samples as fraudulent. In the fraud category (Y), the model correctly classified 47 samples, but incorrectly classified 27 samples as non-fraudulent. These data provide information about the performance of the model on the different categories as well as misclassification.

The performance of the decision tree model is more balanced in terms of precision and recall. For the non-fraud category, the precision rate is as high as 0.88, indicating that the model accurately captures the majority of non-fraud cases. For the fraud category, the precision rate is 0.57, implying that about 57% of the cases predicted as fraud by the model are indeed true fraud cases. For recall, the recall rate is 0.84 for the non-fraud category and 0.64 for the fraud category, indicating that the model is able to capture most of the true fraud cases.

5.3. Discriminant analysis

Table 3. The results of the discriminant analysis model

	precision	recall	f1-score	support
N	0.84	0.85	0.84	220
Y	0.57	0.55	0.56	80
Accuracy			0.77	300
Macro avg	0.70	0.70	0.70	300
Weighted avg	0.77	0.77	0.77	300

The results based on the discriminant analysis model are shown below. In the automated insurance fraud detection task, the training accuracy of the discriminant analysis model was 83.43%, while the

testing accuracy was 77.00%. This shows that the model is able to achieve some predictive performance on unseen data.

According to the discriminant analysis model's classification report for the test data, the model has a precision of 0.84, a recall of 0.85, and an F1-score of 0.84 for the non-fraud category (N). Precision, recall, and F1-score are all 0.57 for the fraud category (Y), while recall is 0.55 and F1-score is 0.56. These measures integrate the model's performance across various categories.

Combining the results of the discriminant analysis model, although it performs better in the non-fraud category, its performance in the fraud category is relatively low. The model may need to further adjust the parameters, optimize the feature selection, or consider other methods to improve the detection of fraud cases. Meanwhile, the interpretability of the model is also a factor of concern, and it may be necessary to weigh the performance and interpretability of the model in practical applications.

5.4. Compare three models

For the automated insurance fraud detection task, we used three different machine learning models, Logistic Regression, Decision Trees and Discriminant Analysis. When comparing and analyzing these three sets of data together, we can observe differences and similarities in the following key areas.

The three models show slight differences in terms of accuracy, with the Logistic Regression model having a slightly higher accuracy (0.796), the Decision Tree model having the next highest accuracy (0.79), and the Discriminant Analysis model having an accuracy of 0.77. It is important to note that accuracy is not a single performance metric, and therefore we need to evaluate these models more comprehensively.

In terms of the trade-off between precision and recall, logistic regression and decision trees perform better and can strike a relative balance between precision and recall. However, the discriminant analysis model performs relatively poorly in terms of precision and recall and needs to be further optimized. The F1-Score, as a composite metric, suggests that there is little difference in the overall performance of the three models.

Interpretation of the confusion matrix shows that the logistic regression model performs poorly in misclassifying non-fraudulent samples as fraudulent, the decision tree model is slightly less effective in this regard, and the discriminant analysis model performs poorly in misclassifying fraudulent samples as non-fraudulent. The balance between precision and recall is critical in fraud detection tasks.

The logistic regression and decision tree models have higher accuracy on the test data, 0.80 and 0.79 respectively, while the discriminant analysis model is slightly lower at 0.77. This may imply that the logistic regression model has better generalization ability and is able to maintain high predictive performance on unseen data.

In addition, the interpretability and complexity of the models are also considered. Logistic regression and decision trees are more interpretable compared to discriminant analysis models, while decision tree models are relatively complex and may be prone to overfitting.

6. Discussion

When a new insurance claim is made, relevant data is entered into the model. These may include vehicle information, accident details, etc. The model will make predictions based on the characteristics of the input data and output whether there is a suspicion of fraud. Regarding the interpretation and processing of the results, the output of the model can be the probability of fraud suspicion or classification labels. Based on the results, the insurance company can take appropriate actions, such as further investigation, verification of information, etc. It should be noted that the machine learning model is not absolutely accurate, and there may be misjudgments. This is just a brief reference, and factors such as data privacy and legal compliance need to be considered in

practical applications. At the same time, the performance of the model also needs to be regularly monitored and updated to maintain accuracy and reliability.

Reference

[1] Yu, W., Feng, G. F., & Zhang, W. J. (2017). Motor vehicle insurance fraud detection system and group identification research. Insurance Research, 2017(2), 63-73.

[2] Joint Project Group on Auto Insurance Anti-fraud. (2021). Research on auto insurance fraud and anti-fraud issues and regulatory suggestions. Insurance Research, 2021(6), 3-10.

[3] Zhang, B., Xiao, Y., & Zeng, Y. (2019). A Comparative Study of Risk Factor Importance Measurement in Auto Insurance Pricing Based on Integrated Learning Methods and Generalized Linear Regression Model. Insurance Research, 2019(10), 73-83.

[4] Panigrahi, S., & Palkar, B. (2018). Comparative analysis of classification algorithms for auto-insurance fraud detection based on feature selection algorithms. International Journal of Computer Science and Engineering, 6(9), 72-77.

[5] Yan, C., Li, Y. Q., & Sun, H. T. (2017). Auto insurance fraud identification based on an ant colony algorithm optimizing a stochastic forest model. Insurance Research, 2017(6), 114-127.

[6] Guo, G. Z., & Duan, J. X. (1999). Game analysis of insurance fraud. Journal of the Capital University of Economics and Business, 1999(3), 51-54.

[7] Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. Expert Systems with Applications, 29(3), 653-666.

[8] Hanafizadeh, P., & Rastkhiz Paydar, N. (2013). A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry. International Journal of Strategic Decision Sciences (IJSDS), 4(1).

[9] Kašćelan, L., Kašćelan, V., & Novović-Burić, M. (2014). A Data Mining Approach for Risk Assessment in Car Insurance: Evidence from Montenegro. International Journal of Business Intelligence Research (IJBIR), 5(3), 11-28.

[10] Li, Y., Yan, C., Liu, W., et al. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Applied Soft Computing, 70, 1000-1009.