*Journal of Mathematics and Modeling in Finance (JMMF)*
Vol. 5, No. 1, Winter & Spring 2025
Research paper

# Mitigating data imbalance for enhanced third-party insurance claim prediction using machine learning

**Maryam Esna-Ashari**[1], **Hamideh Badi**[2], **Majid Chahkandi**[3], **Hamid Saadatfar**[4]

[1] Insurance Research Center, Tehran, Iran
esnaashari@irc.ac.ir

[2] Department of statistics, University of Birjand, Birjand, Iran
hamideh.badi@gmail.com

[3] Department of statistics, University of Birjand, Birjand, Iran
mchahkandi@birjand.ac.ir

[4] Department of statistics, University of Birjand, Birjand, Iran
saadatfar@birjand.ac.ir

**Abstract:**
Accurate prediction of third-party insurance claims is critical for pricing policies and managing risk. However, the highly imbalanced nature of insurance data—where non-claim cases vastly outnumber claim case—poses significant challenges to standard predictive models. This study explores the use of machine learning algorithms to enhance claim prediction by directly addressing this imbalance. We use real data from the Insurance Research Center of Iran, incorporating variables such as driver characteristics, vehicle features, location, and claims history. Five models are evaluated: logistic regression, decision tree, bagging, random forest, and boosting. To handle the imbalance, we apply random undersampling, over-sampling, and SMOTE. Model performance is assessed using accuracy, sensitivity, specificity, precision, and F-score. Results indicate that when data imbalance is properly treated, ensemble method—particularly decision trees, bagging, and random fores—significantly outperform logistic regression and boosting, especially in detecting actual claim cases. The study underscores the importance of using appropriate resampling techniques and evaluation metrics in imbalanced settings. These findings can help insurers develop more reliable models for pricing and risk classification.

*Keywords:* Machine learning algorithms, Third-party insurance, Imbalanced data.
*Classification:* MSC2010 Classifications: 91B30, 68T09.

## 1 Introduction

In todays data-rich environment, machine learning has become a critical tool for analyzing complex datasets across various industries. Within the insurance sector, a primary challenge lies in accurately predicting claims made by policyholders. Precise claim forecasting is vital for setting equitable premiums and ensuring the financial health of insurance companies. The availability of extensive data in the insurance industry has driven the adoption of machine learning algorithms as an

[1]Corresponding author
Received: 10/03/2025 Accepted: 02/07/2025
**https://doi.org/10.22054/jmmf.2025.84807.1169**

efficient and accurate method for predicting third-party liability insurance claims.

This research focuses on automobile third-party insurance, a mandatory and significant segment of the insurance market. The ability to predict potential losses allows insurance companies to tailor policies to individual policyholders, making claim forecasting a cornerstone of premium estimation. The increasing frequency and severity of auto insurance claims necessitate innovative methods for assessing and predicting financial and physical damages, determining whether a claim will occur is essential. An effective model for determining premiums, which accounts for various risk factors, is crucial for preventing losses for insurance companies and ensuring customer satisfaction.

Machine learning (ML) is increasingly used in the insurance industry to analyze large datasets and improve decision-making. One of the critical applications of ML is predicting whether a third-party automobile insurance policy will result in a claim. This prediction is essential for fair premium pricing, effective risk assessment, and long-term financial sustainability of insurers. Third-party automobile insurance is mandatory in many countries and constitutes a major portion of insurers' portfolios. Predicting claim occurrence is complicated by the imbalanced nature of insurance data: most policyholders do not file claims, and the few that do represent the minority class. Traditional statistical models often struggle to handle such imbalance, resulting in low detection rates for actual claims. This research addresses the question: How can machine learning algorithms be effectively applied to predict third-party insurance claims under data imbalance conditions? To answer this, we analyze a real dataset from the Insurance Research Center of Iran, incorporating variables related to drivers, vehicles, and claim history. We apply a variety of ML algorithmsincluding logistic regression, decision trees, and ensemble methodsand compare their performance before and after using balancing techniques such as random undersampling, oversampling, and SMOTE. By focusing on the effect of data imbalance and the performance of various models, this study contributes to the development of more accurate, fair, and practical tools for insurers.

Following data preprocessing and feature selection, machine learning techniques, including decision trees, logistic regression, and ensemble methods, are applied to predict claims and assess the risk levels of policyholders ( [8], [26]). Insurance companies typically record policyholder information, and for claims, details about the at-fault drivers are also available, enabling the potential prediction of damage amounts. For other recent works on this topic we refer to [4, 14, 17, 20].

The performance of the various methods is evaluated and compared to identify the key factors that influence third-party auto insurance premiums, and a model for predicting customer claims is proposed. A significant challenge in this domain is the imbalanced nature of the data, where the number of policies with claims is much lower than those without claims. This imbalance can lead to poor performance of prediction models, as they tend to classify observations as the majority class. To

address this issue, the research employs several methods for handling imbalanced data ( [2], [1]). Unlike previous domestic research that often considers limited variables and models, this study uses a more diverse set of variables and advanced models based on prior research, utilizing data from the Insurance Research Center of Iran to predict claims. To account for the data imbalance between claimants and non-claimants, various criteria and methods, consistent with recent studies, are applied. Model calculations in this research are performed using Python.

## 2 Preliminaries

Third-party automobile insurance is a mandatory form of liability coverage that compensates third parties for bodily injury or property damage caused by the insured driver. The policyholder (first party) and the insurance company (second party) are contractually bound to provide coverage to affected individuals (third parties) ( [5]).
In this study, claim prediction is formulated as a binary classification task: the response variable indicates whether a claim occurred (1) or not (0). Given the high imbalance in class distribution, with far fewer claims than non-claims, the data must be carefully preprocessed and balanced before applying machine learning algorithms.
Machine learning methods used in this research follow a supervised learning framework, where models are trained using labeled data. The goal is to predict claim status based on input features related to driver characteristics, vehicle specifications, and policy history ( [15]).

## 3 Literature Review

This literature review examines both domestic and international research employing machine learning algorithms to predict risk in insurance policyholders. These studies explore a range of methodologies and variables, providing valuable insights into effective risk assessment and claim prediction.

- Domestic Research

[11] compared decision trees, neural networks, Bayesian networks, support vector machines, logistic regression, and discriminant analysis for predicting claim categories in comprehensive insurance, considering both policyholder and vehicle characteristics. The results indicated that neural networks and decision trees achieved the highest prediction accuracy, approximately 82%. [7] focused on identifying automobile insurance fraud using data mining. They evaluated six independent variables: insurance history, number of claims, incident reporting time, accident report status (injury or property damage), and claim amount. Employing decision trees, logistic regression, and Naïve Bayes, they found that the Naïve Bayes method was most

effective in detecting fraudulent claims. [24] utilized a neural network to predict potential policyholder loss and determine optimal rates. The model achieved an accuracy of 91% in estimating the loss category and 87% in predicting potential loss. [19] analyzed the impact of driver characteristics on claim occurrence, revealing that policyholders under 22 (passenger cars), 30 (trucks), and 25 (motorcycles) had significantly higher claim likelihoods, suggesting higher premiums for these age groups. Moreover, over 50% of personal injury claims involved uninsured drivers, indicating that vehicles with multiple drivers should also incur higher premiums. [6] used random oversampling to address data imbalance in auto body insurance and found that the Random Forest algorithm showed slightly better performance than XGBoost.

- International Research

[22] compared decision trees and neural networks for predicting whether an insured party would file a claim, demonstrating the superior performance of neural networks. [23] proposed a hybrid method for detecting insurance fraud in unbalanced datasets. [10] presented an automobile insurance claim prediction model using XGBoost, artificial neural networks, decision trees, and a Naïve Bayes, while also addressing missing data. Surprisingly, decision trees exhibited higher accuracy compared to other models in this study, which used a dataset of 30,240 observations. [25] explored data mining techniques for automobile insurance claim prediction, comparing three machine learning methods and identifying neural networks as the best predictor. [16] combined clustering methods, regression, and support vector machines for risk classification and automobile insurance claim amount prediction. Support vector regression was used to predict the expected claim amount, while kernel logistic regression was used to predict claim occurrence. [3] employed decision trees, logistic regression, and neural networks to identify suitable insured parties based on risk levels, aiming to select the best insured parties according to risk and the likelihood of insurance incidents. [9] reported significant results using decision tree methods for predicting automobile damage. [18] used machine learning to predict fraudulent claims and calculate premiums based on personal information. Random forests and Naïve Bayes classifiers were employed, with random forests demonstrating superior performance in fraud prediction. This study focused on fraudulent claims rather than general insurance loss prediction. [21] compared XGBoost and logistic regression for predicting motor insurance claim frequency, with XGBoost performing slightly better. Their database consisted of only 2,767 observations. They also developed a model for predicting insurance losses. [27] used decision trees and neural networks for risk modeling, prediction, and analysis of risk level patterns in automobile insurance. [12] employed logistic regression, K-nearest neighbors, random forests, decision trees, and Naïve Bayes to predict claim occurrence using large insurance datasets. Model performance was evaluated using the confusion matrix and metrics like accuracy, recall, F-score and

area under the curve. Random forests outperformed other methods. [13] utilized oversampling, undersampling, and combined resampling approaches to enhance the classification performance of imbalanced data. They concluded that oversampling provided a more accurate classifier based on sensitivity metrics.

# 4 Methodology

In this section, the necessary steps for implementing machine learning algorithms are described.

## 4.1 Research Phases

The execution phases of this research are as follows:

- Data collection from the Insurance Research Center's database, understanding and comprehending the data;
- Preprocessing and refining the data;
- Statistical analysis and dividing the data into balanced and random subsets in two categories: test data and training data;
- Pattern extraction using decision tree algorithms, logistic regression, and comparing the results obtained from these algorithms;
- Presenting the discovered model for classifying policyholders and identifying determining features;
- Evaluating the classification results and validating the model.

The data includes 21,294 issued insurance policies, and the selected final variables in line with the research objective, which is to predict claims, are presented in Table 1. To evaluate the generalization performance of the models, the dataset was split into training and testing subsets using an 80/20 ratio. Furthermore, to reduce variance and ensure stability in performance estimates, we applied 5-fold cross-validation during the training phase of each machine learning algorithm. This approach helps validate the robustness of the models across different random partitions and enhances the reliability of the reported metrics.

## 4.2 Data Imbalance and Methods for Dealing with

Data imbalance, a common challenge in real-world classification datasets, arises when there's a significant disparity in the number of samples across different classes. This uneven distribution of the response variable can hinder the performance of standard learning algorithms, which often assume or expect balanced class distributions.

Table 1: Variables Used in the Model for Claim Prediction

| Variable Name | Variable Description | Variable Type |
|---|---|---|
| Type_Car | Car Model | Categorical |
| Car_age | Car Age | Discrete Quantitative |
| CarGroupCode | Grouping Name in Third-Party Law | Categorical |
| Cylinder | Number of Cylinders | Discrete Quantitative |
| Capacity | Car Capacity | Discrete Quantitative |
| LfYrLosCnt | Number of Claim-Free Years | Discrete Quantitative |
| Financial_commitment | Financial Commitment Amount | Discrete Quantitative |
| FnYrLosCnt | Previous Policy Financial Losses | Binary Variable |
| LfYrLosCnt | Previous Policy Bodily Injury Losses | Binary Variable |
| Delay | Number of Days of Delay in Policy Renewal | Discrete Quantitative |
| gender_insured | Insured's Gender | Categorical |
| City | Insured's City | Categorical |
| claim | Claim | Binary Variable |

A classic example is car insurance claims, where the majority of policyholders typically do not file claims, leading to fewer claim instances (minority class) compared to non-claim instances (majority class). Identifying the minority class accurately is crucial in these scenarios. Standard learning algorithms often struggle with imbalanced data. Their primary goal is to increase accuracy and reduce errors, but they are trained under the assumption of equally distributed categories. This can result in poor prediction of the minority class because the algorithm is not adequately trained on these less frequent samples. The incorrect classification of a minority sample can have serious consequences in various applications. In claim prediction, the number of insurance policies with claims is typically much smaller than those without claims. An example dataset of 21,294 insurance policies has 1,575 claim-incurred samples (coded as 1) and 19,719 non-claim-incurred samples (coded as 0), clearly demonstrating the imbalance. This imbalance is shown in Figure 1.

Several methods exist to address the imbalanced data problem:

**Random Undersampling**: This involves randomly removing samples from the majority class in the training dataset to create a balance. It's a simple technique best suited for very large datasets where reducing the number of training samples improves execution time and storage. However, a disadvantage is the potential removal of valuable information.

**Random Oversampling**: This involves randomly duplicating samples from the minority class and adding them to the training dataset, increasing its size. While it doesn't lead to information loss, oversampling simply adds duplicate observations, potentially leading to overfitting. Although training accuracy may be high, accuracy on unseen test data may be worse.

**Synthetic Minority Oversampling Technique (SMOTE)**: SMOTE is a widely

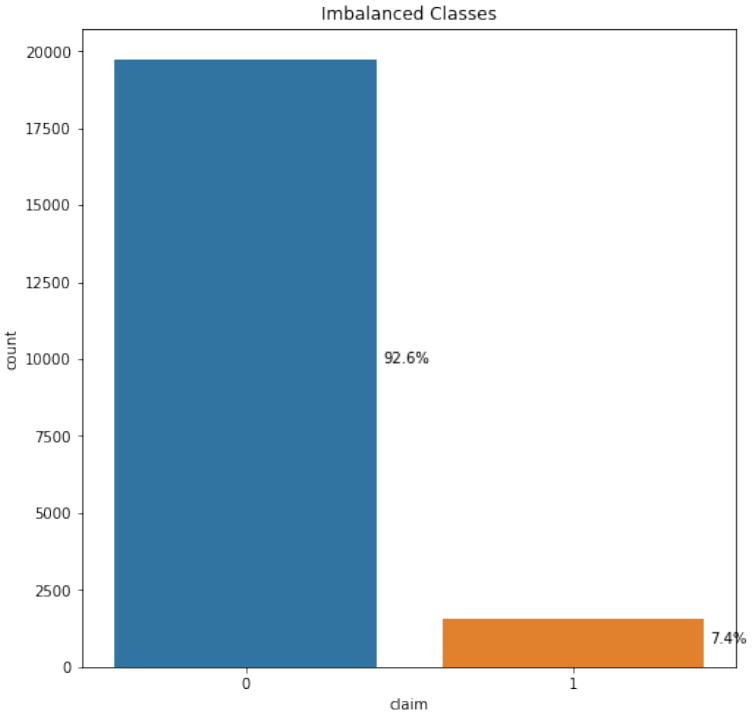


Figure 1: Imbalance in the target variable

used oversampling technique that generates new synthetic data points for the minority class. It selects a sample from the minority class, calculates the Euclidean distances to its k nearest neighbors, randomly selects one of these neighbors, and then creates a synthetic data point using a specific formula. This process helps to diversify the minority class samples.

## 4.3 Confusion matrix

The performance of the algorithm is computed by a confusion matrix shown in Figure 2.

In the matrix, TP and TN represent the number of correctly classified positive and negative instances, respectively, while FP and FN represent incorrectly classified positive and negative instances, respectively. In this paper, the positive class consists of individuals who have filed a claim, and the negative class consists of individuals who have not filed a claim.

True-Negative: The individual has not filed a claim, and the model correctly predicts that the individual has not filed a claim;

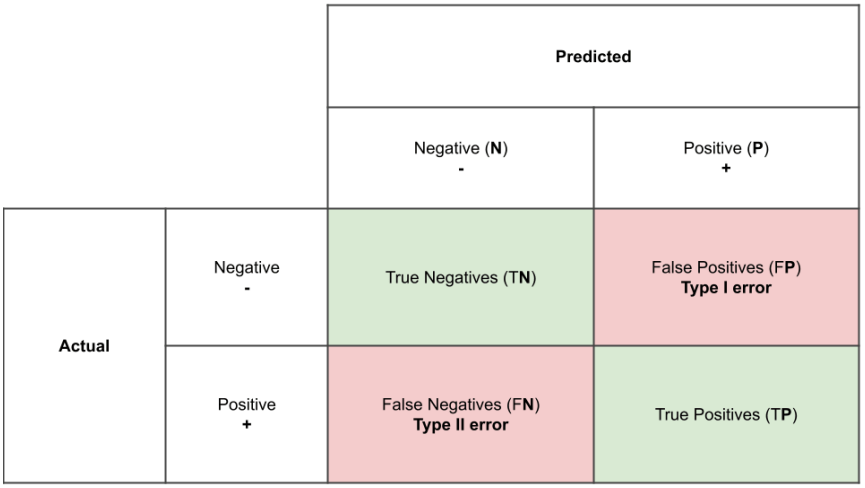


Figure 2: Confusion Matrix

True-Positive: The individual has filed a claim, and the model correctly predicts that the individual has filed a claim;

False-Positive: The individual has not filed a claim, and the model incorrectly predicts that the individual has filed a claim;

False-Negative: The individual has filed a claim, and the model incorrectly predicts that the individual has not filed a claim.
The performance of the algorithm is measured using accuracy, sensitivity (also known as recall), specificity, precision, and the F-score, which is the harmonic mean of precision and sensitivity. The associated formulas are listed below. The greater value, the greater performance:

$$\begin{aligned} accuracy &= (TP+TN)/(TP+FP+TN+FN) \\ sensitivity &= TP/(TP+FN) \\ specificity &= TN/(FP+TN) \\ precision &= TP/(TP+FP) \\ F-score &= (2.precision.sensitivity)/(precision+sensitivity). \end{aligned}$$

# 5 Findings

Classification is a fundamental problem in the field of machine learning, and various methods are used to address it. In these types of problems, the data is labeled, and the goal is to predict the labels. In this research, the data labels indicate whether policyholders have experienced a claim or not.

This section focuses on performing classification of insurance policies based on the presence or absence of a claim. The methods and models introduced in the previous section are implemented, and then the results are compared.

The algorithms used here to build the predictive model are:

- Logistic Regression
- Decision Tree
- Bagging
- Random Forest
- Boosting

The results after implementing the algorithms are shown in Table 2 for the imbalanced data.

| | Confusion Matrix | accuracy | specificity | sensitivity | precision | F_score |
|---|---|---|---|---|---|---|
| Logistic Regression | $\begin{bmatrix} 3932 & 0 \\ 327 & 0 \end{bmatrix}$ | 0.92 | 1 | 0 | 0 | 0 |
| Decision Tree | $\begin{bmatrix} 3602 & 330 \\ 316 & 11 \end{bmatrix}$ | 0.85 | 0.92 | 0.03 | 0.03 | 0.03 |
| Bagging | $\begin{bmatrix} 3772 & 160 \\ 324 & 3 \end{bmatrix}$ | 0.89 | 0.95 | 0.01 | 0.01 | 0.01 |
| Random Forest | $\begin{bmatrix} 3777 & 155 \\ 327 & 0 \end{bmatrix}$ | 0.89 | 0.9608 | 0 | 0 | 0 |
| Boosting | $\begin{bmatrix} 3931 & 1 \\ 326 & 1 \end{bmatrix}$ | 0.92 | 0.99 | 0 | 0.05 | 0.01 |

Table 2: Classification Results with Imbalanced Data

One of the main challenges in insurance loss prediction is the imbalanced nature of the data classes. As can be seen in Table 2, the accuracy metric is high for all models, but despite this high accuracy, the model lacks any operational value. In

other words, in this case, the model only learns based on outputs related to non-loss data, and loss data is ignored. Therefore, selecting an appropriate performance metric is an important aspect of working with imbalanced data.

Now, we apply the three methods discussed in the previous section for balancing data and the results are in Tables 3, 4 and 5, respectively.

| | Confusion Matrix | accuracy | specificity | sensitivity | precision | F_score |
|---|---|---|---|---|---|---|
| Logistic Regression | $\begin{bmatrix} 3821 & 144 \\ 279 & 48 \end{bmatrix}$ | 0.96 | 0.96 | 0.15 | 0.25 | 0.18 |
| Decision Tree | $\begin{bmatrix} 3821 & 111 \\ 113 & 214 \end{bmatrix}$ | 0.95 | 0.94 | 0.65 | 0.66 | 0.66 |
| Bagging | $\begin{bmatrix} 3795 & 137 \\ 93 & 234 \end{bmatrix}$ | 0.95 | 0.96 | 0.72 | 0.63 | 0.67 |
| Random Forest | $\begin{bmatrix} 3809 & 123 \\ 97 & 230 \end{bmatrix}$ | 0.95 | 0.94 | 0.70 | 0.65 | 0.67 |
| Boosting | $\begin{bmatrix} 3828 & 104 \\ 294 & 33 \end{bmatrix}$ | 0.91 | 0.97 | 0.10 | 0.24 | 0.14 |

Table 3: Classification results with balanced data using the undersampling method

The results show that using models designed to address data imbalance, decision tree, bagging, and random forest models perform better than logistic regression and boosting methods. The results also indicate that eliminating data imbalance improves metrics such as sensitivity, precision and F-score.

# 6 Conclusions

This study demonstrated that properly addressing data imbalance significantly improves the performance of machine learning models in predicting third-party automobile insurance claims. Using a real-world dataset, we evaluated logistic regression, decision trees, bagging, random forest, and boosting models under different data balancing techniques, including random undersampling, oversampling, and SMOTE. The results revealed that ensemble methodsespecially bagging and

| | Confusion Matrix | accuracy | specificity | sensitivity | precision | F_score |
|---|---|---|---|---|---|---|
| Logistic Regression | $\begin{bmatrix} 3757 & 175 \\ 283 & 44 \end{bmatrix}$ | 0.89 | 0.13 | 0.13 | 0.20 | 0.16 |
| Decision Tree | $\begin{bmatrix} 3821 & 111 \\ 113 & 214 \end{bmatrix}$ | 0.95 | 0.94 | 0.65 | 0.66 | 0.66 |
| Bagging | $\begin{bmatrix} 3795 & 137 \\ 93 & 234 \end{bmatrix}$ | 0.95 | 0.96 | 0.72 | 0.63 | 0.67 |
| Random Forest | $\begin{bmatrix} 3809 & 123 \\ 98 & 229 \end{bmatrix}$ | 0.95 | 0.96 | 0.70 | 0.65 | 0.67 |
| Boosting | $\begin{bmatrix} 3828 & 104 \\ 294 & 33 \end{bmatrix}$ | 0.91 | 0.96 | 0.10 | 0.24 | 0.14 |

Table 4: Classification results with balanced data using the oversampling method

random forestconsistently outperformed logistic regression and boosting models in terms of sensitivity and F-score when the class imbalance was handled appropriately. This indicates that such models are more capable of identifying true claimants in highly skewed insurance datasets. From a practical standpoint, these findings provide actionable insights for insurance companies. Incorporating balanced machine learning models into claim prediction systems can enhance underwriting accuracy, reduce loss ratios, and allow for more personalized premium pricing. The use of features such as vehicle type, claim-free years, and delay in policy renewal were shown to contribute significantly to the predictive power of the models. Furthermore, the study emphasizes the importance of evaluating classification performance with appropriate metrics beyond accuracyparticularly sensitivity and F-scorewhen working with imbalanced data. For future research, we recommend exploring advanced techniques such as deep learning architectures, cost-sensitive learning, and hybrid ensemble approaches. Moreover, the integration of telematics data or unstructured data (e.g., claim narratives) may further enhance model performance and provide a richer risk assessment framework.

| | Confusion Matrix | accuracy | specificity | sensitivity | precision | F_score |
|---|---|---|---|---|---|---|
| Logistic Regression | $\begin{bmatrix} 3788 & 144 \\ 279 & 48 \end{bmatrix}$ | 0.90 | 0.96 | 0.15 | 0.25 | 0.18 |
| Decision Tree | $\begin{bmatrix} 1464 & 2468 \\ 97 & 230 \end{bmatrix}$ | 0.40 | 0.86 | 0.70 | 0.09 | 0.15 |
| Bagging | $\begin{bmatrix} 3799 & 133 \\ 93 & 234 \end{bmatrix}$ | 0.95 | 0.96 | 0.72 | 0.64 | 0.67 |
| Random Forest | $\begin{bmatrix} 3810 & 122 \\ 97 & 230 \end{bmatrix}$ | 0.95 | 0.96 | 0.70 | 0.65 | 0.70 |
| Boosting | $\begin{bmatrix} 3828 & 104 \\ 294 & 33 \end{bmatrix}$ | 0.91 | 0.97 | 0.10 | 0.24 | 0.14 |

Table 5: Classification results with balanced data using the SMOT

# Acknowledgement

This article is an excerpt from a student thesis Prediction of Third-Party Insurance Claim with Machine Learning Algorithms for the degree of Master of Science at University of Birjand, which has been conducted with the cooperation and support of the Insurance Research Center. We also express our sincere thanks to the anonymous reviewers and the Editor for their incisive comments on an earlier version of this manuscript which led to this improved version.

## Bibliography

[1] E.M., ALDAHASI, R.K.,ALSHEIKH, F.A., KHAN, G., JEON, *Optimizing fraud detection in financial transactions with machine learning and imbalance mitigation*, Expert Systems, 42 (2025), e13682.

[2] A., ABDALLAH, M.A., MAAROF, A., ZAINAL, *Fraud detection system: A survey. Journal of Network and Computer Applications*, 68 (2016), pp. 90-113.

[3] P., BAECKE, L., BOCCA , *The value of vehicle telematics data in insurance risk selection processes*, Decision Support Systems, 98 (2017), pp. 69-79.

[4] K., DING, B., LEV, X., PENG, T., SUN, M.A., VASARHELYI, *Machine learning improves accounting estimates: Evidence from insurance payments. Review of accounting studies*, 25 (2020), pp. 1098-1134.

[5] G., DIONNE (ED.), *Handbook of Insurance*, 2nd ed. Springer, 2013.

[6] M., Esna-Ashari, *Using a new data mining method for automobile insurance fraud detection: a case study by a real data from an Iranian insurance company*, International Journal of Mathematical Modeling Computations, 14 (2024), pp. 15-20.

[7] M., Firuzi, M., Shakouri, L., Kazemi, S., Zahedi, *A data mining approach to auto insurance fraud*, Iranian Journal of Insurance Research (Sanaat-e-Bimeh). 26 (2011), pp. 103-128. Available from: https://sid.ir/paper/100794/en (in Persian).

[8] E.W., Frees , *Regression modeling with actuarial and financial applications*, Cambridge University Press, 2014.

[9] N.K., Frempong, N., Nicholas, M.A., Boateng, *Decision tree as a predictive modeling tool for auto insurance claims*, International Journal of Statistics and Applications, 7 (2017), pp. 117-120.

[10] I., Goodfellow, Y., Bengio, A., Courville, *Machine learning basics*, Deep Learning, 1 (2016), pp. 98-164.

[11] N., Hajiheidari, S., Khaleie, A., Farahi, *The insured risk classification in auto collision insurance using data mining algorithms: evidence from an Iranian insurance company*, Iranian Journal of Insurance Research (Sanaat-e-Bimeh). 26 (2012), pp. 107-129. Available from: https://sid.ir/paper/100920/en (in Persian).

[12] M., Hanafy, R., Ming, *Machine learning approaches for auto insurance big data*, Risks, 9 (2021), pp. 42.

[13] M., Hanafy, R., Ming, *Improving imbalanced data classification in auto insurance by the data level approaches*, Journal of Advanced Computer Science and Applications, (2021), pp. 493-499.

[14] J.T., Hancock, T.M., Khoshgoftaar, J.M., Johnson, *Evaluating classifier performance with highly imbalanced big data*, Journal of Big Data, 10 (2023), pp. 1-31.

[15] G., James, D., Witten, T., Hastie, R., Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed. Springer, 2021.

[16] V., Kaelan, L., Kaelan, M., Novovi Buri, *A nonparametric data mining approach for risk prediction in car insurance*, Economic Research-Ekonomska Istraivanja. 29 (2016), pp. 545-558.

[17] F., Khamesian, M., Esna-Ashari, E., Dei Ofosu-Hene, F., Khanizadeh, *Risk classification of imbalanced data for car insurance companies: Machine learning approaches*, International Journal of Mathematical Modelling & Computations, 12 (2022), pp. 153-162.

[18] G., Kowshalya, M., Nandhini, *Predicting fraudulent claims in automobile insurance*, In: Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), (2018), pp. 1338-1343.

[19] M., Manteqipour, V., Ghorbani, M., Aalaei, *Classifying age of policyholders according to the claim rates in Iran*, Journal of Applied Economics Studies in Iran, 39 (2021), pp. 141-175.

[20] R., Ming, O., Mohamad, N., Innab, M., Hanafy, (2024). *Bagging Vs. Boosting in Ensemble Machine Learning? An Integrated Application to Fraud Risk Analysis in the Insurance Sector*, Applied Artificial Intelligence, 38 (2024), 2355024.

[21] J., Pesantez-Narvaez, M., Guillen, M., Alcañiz, *Predicting motor insurance claims using telematics dataXGBoost versus logistic regression*, Risks, 7 (2019), 70.

[22] K.A., Smith, R.J., Willis, M., Brooks, *An analysis of customer retention and insurance claim patterns using data mining: a case study*, Journal of the Operational Research Society, 53 (2002), pp. 532-541.

[23] G.G., Sundarkumar, V., Ravi, *A novel hybrid under-sampling method for mining unbalanced datasets in banking and insurance*, Engineering Applications of Artificial Intelligence, 37 (2015), pp. 368-377.

[24] M., Torkestani, A., Dehpanah, M.T., Taghavifard, S., Shafiee, *Providing a framework for reforming premium rates of vehicle collision coverage using neural networks model: a case study of Asia Insurance Company*, Journal of Information Technology Management, 8 (2017), pp. 711-732. Available from: https://sid.ir/paper/140340/en (in Persian).

[25] K.P.M.LP., Weerasinghe, M.C., Wijegunasekara, *A comparative study of data mining algorithms in the prediction of auto insurance claims*, European International Journal of Science and Technology, 5 (2016), pp. 47-54.

[26] M.V., Wüthrich, M., Merz, *Statistical foundations of actuarial learning and its applications*, Springer Nature, 2023.

[27] S., WUYU, P., CERNA, *Risk assessment predictive modelling in insurance industry using data mining*, Software Engineering, 6 (2019), 121.

*How to Cite:* Maryam Esna-Ashari[1], Hamideh Badi[2], Majid Chahkandi[3], Hamid Saadatfar[4], *Mitigating data imbalance for enhanced third-party insurance claim prediction using machine learning*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 5, No. 1, Pages:175–187, (2025).