International Journal of
Mathematical Modelling & Computations
Vol. 12, No. 03, Summer 2022, 153- 162
DOI:10.30495/ijm2c.2022.1958403.1252

Risk Classification of Imbalanced Data for Car Insurance Companies: Machine Learning Approaches

F. Khamesian[a], M. Esna-Ashari[a], E. D. Ofosu-Hene[b] and F. Khanizadeh[a,*]

[a] *Insurance Research Center, Tehran, PO. Code 19395-4499, Iran,*

[b] *Department of Accounting and Finance, Faculty of Business and Law, De Montfort University, Leicester , PO. Code LE1 9BH, UK.*

Abstract. This paper presents a mechanism for insurance companies to assess the most effective features to classify the risk of their customers for third party liability (TPL) car insurance. Basically, the process of underwriting is carried out based on the expert experiences and the industry suffers from lack of a systematic method to categorize their policyholders with respect to the risk level. We analyzed 13,388 observations of an insurance claim dataset from body injury reports provided by an Iranian insurance company. The main challenge is the imbalanced dataset. Here we employ logistic regression and random forest with different resampling of the original data in order to increase the performance of models. Results indicate that the random forest with the hybrid resampling methods is the best classifier and furthermore, victim age, premium, car age and insured age are the most important factors for claims prediction.

Received: 09 May 2022, Revised: 01 August 2022, Accepted: 19 August 2022.

Keywords: Machine Learning; Supervised Learning; Imbalanced Data; Claim Risk; Classification.

AMS Subject Classification: 91B05, 68T09.

Index to information contained in this paper

1. Introduction

1.1 *Background*

As one of the most influential financial institutions in the economy and society, insurance companies need to have access to powerful risk analysis tools to be able

[*] Corresponding author. Email: khanizadeh@irc.ac.ir

http://ijm2c.ctb.iau.ir

to manage their risks properly. The output of risk analysis can be applied within risk management through various methods such as product pricing, marketing, and identification of customers risk level. In this regard, data mining includes very accurate and practical tools for risk assessment that can be used in various fields (see for example [17, 25]). These tools have also provided sound results in the field of insurance in evaluating and classifying the risk of insured (see [4, 13, 19, 27]).

In car insurance, data mining methods have been used to investigate the variables affecting the probability of an accident. Driving records, age, gender, marital status are the main factors commonly used to perform risk analysis ([9, 15, 20, 36, 38]). Four main features namely age, area of residence, vehicle type, and no-claim discount have been studied in [28]. In [5], the age is discussed in detail as an insurance rate class variable.

With the advent of telematics and usage-based insurance (UBI), some studies examine the impact of driver characteristics and behaviors to classify the risk of insured ([3, 40]). In [18] authors used logistic regression along with four machine learning algorithms: support vector machines, random forests, XGBoost, and artificial neural networks to model the risk probability taking driver behavior factors into consideration and employed a Poisson regression as claim frequency model.

1.2 *Related works*

One of the serious challenges associated with datasets of bodily third-party automobile insurance is the imbalance in the target variable. In [32] authors employed SVM models to overcome the issue of highly imbalanced classification. Cost-sensitive learning is another approach tackling imbalance dataset. The examples can be found in [12, 34, 42]. Some studies provide ensemble learning which try to reduce the result of variance on the test data [23, 31]. Fraud detection is a common imbalanced classification in which boosting [10, 26, 33] and bagging [30, 41] algorithms have been used to improve the model performance.

In some cases, algorithms are not solely able to overcome the challenge of imbalanced data. In this regard, first re-sampling methods are used to balance the distribution of the target variable and so that then one can use known classification algorithms. In over-sampling, the number of samples in the minority class is increased to achieve a relative balance between the two classes [11, 43]. On the other hand, under-sampling methods focus on the majority class and, based on different criteria, eliminate some samples in order to bring the distribution of the majority class to an appropriate level of balance with the minority class [6, 37, 39]].

As the contribution of this paper, a combination of machine learning models and sampling methods has been systematically used to solve this challenge. To the best of our knowledge this is the first paper to tackle the imbalance dataset in third-party liability car insurance. Studies based on Iranian data (see for example: [21, 29]) follow the similar strategy as ones carried out using overseas data (see for example [1, 18, 44]). These papers consider combination of individual, car and behavioral features. This can be a reasonable approach in cases where suitable data is available. Thus the main contributions of this paper are: (1) thorough study focusing on the most challenging issue namely imbalanced dataset; (2) comprehensive algorithms suitable for typical datasets collected in Iran insurance companies and provide a systematic applied approach to assess high and low risk customers.

The rest of the paper is structured as follows. In Section 2, we cover the essentials of the data set and methods to overcome imbalanced issue. In Section 3, we provide the results of the model and finally in Section 4 we discuss the results and present the guideline to assess the risk level.

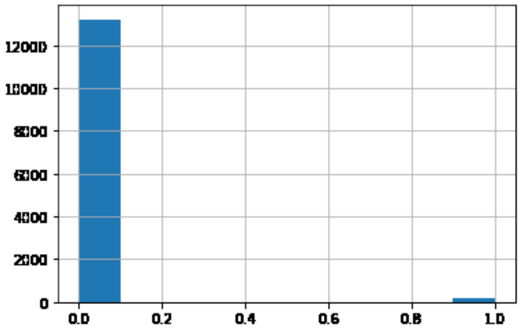

Figure 1. Distribution of class imbalance in the dataset.

2. Data and methods

Throughout this section we introduce dataset, and the methods and strategies in getting the results are provided. The algorithms are coded in Python programming language.

2.1 *The dataset*

For this study, an Iranian insurance company provided the dataset concerning bodily injury liability motor insurance. The dataset contains four features and a target, each with the following descriptions:

- Insured Age – age of the driver involved in an accident
- Victim Age – age of the person injured in an accident
- Premium – premium paid for one year of car body insurance
- Car Age – age of the car that caused the injury
- Target – binary variable indicating the occurrence/non-occurrence of the claim

An important point to note about the dataset is the imbalance of the target variable. This occurs when the dataset has many more instances of certain classes than of the others. In our case, the dataset consists of a negative class (0) of 98.7% and positive class (1) of 1.3%.

As most machine learning algorithms assume balanced distributions, building decision boundaries to classify the observations accurately is a challenge. Basically, samples from the minority class are most often misclassified. This makes an issue since we are mainly interested in the minority class.

In the next section we provide solutions to overcome the imbalanced issue and to build the most efficient machine learning algorithm.

2.2 *Methods*

One of the main strategies to approach imbalanced dataset is the data level method. In this approach one modifies the distribution of the data in a way to obtain either more observations in the minority class (over-sampling) or less data point in the majority class (under-sampling). In fact, these methods yield similar ratio for each class. In under-sampling, our focus is on the majority class and we reduce the number of samples within this class. There are basically two main categories of under-sampling namely; fixed and cleaning. Fixed methods reduce majority class to the equal number of observations in the minority class: On the other hand, in the cleaning case, we delete datapoints from majority class according to some

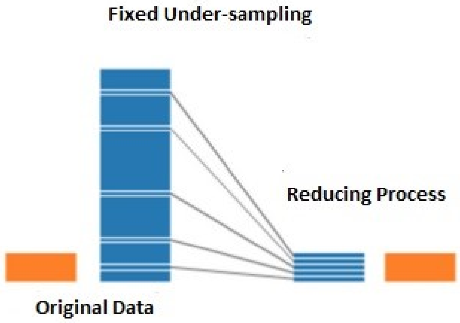

Figure 2. Fixed under-sampling outcome.

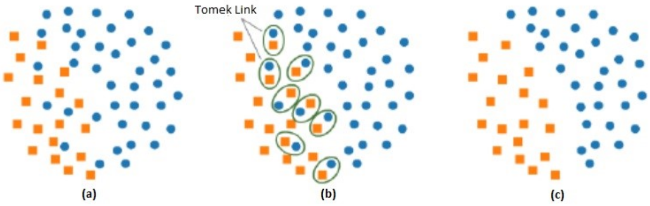

Figure 3. (a) original data. (b) detecting Tomek link. (c) data after under-sampling.

constraint. Basically, we define a balancing ratio to reach to the desired balance level:

$$R(class) = \frac{n(minority)}{n(majority)}.$$

Tomek link are pairs of opposite-class samples in close proximity. We remove the points of the majority class from the Tomek link. This approach results in a better decision boundary for a classifier. This under-sampling process belongs to the category of cleaning types as it removes observations that sit in the boundary of the classifier. These are the points that are very similar to each other but from different classes. In this method, we believe that the observations on the boundary are noises or not very important to focus on.

In over-sampling techniques, the aim is to increase the number of observations in the minority class to obtain some extent of balance between two classes. Similar to the under-sampling methods we can have two different versions namely; fixed and cleaning. In this section we introduce the Synthetic Minority Over-sampling Technique (SMOTE) as an approach to the construction of classifiers from imbalanced datasets. For more information regarding over-sampling methods one can read ([2, 14, 24]).

SMOTE creates new samples from the minority class through interpolation in which new data points are added within the range of known observations. In this way new samples are different from the original ones and therefore it prevents duplication.

2.3 *ROC curve*

The ROC Curve, known as the ”Receiver Operating Characteristics” is a graphical diagram that demonstrates the ability of a binary classification model to distinguish between classes. The curve is a performance measurement tool which can be used

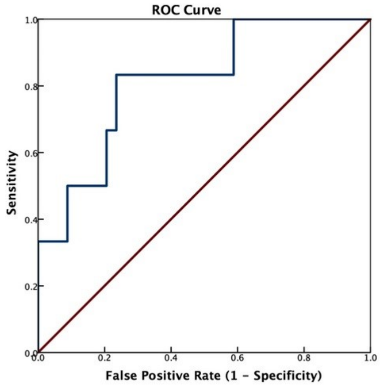

Figure 4. Plot of ROC curve.

to examine concepts such as cutoff, sensitivity and specificity. The Rock Curve (ROC Curve), as shown below, is used to express the True Positive Rate (TPR) versus the False Positive Rate (FNR). The vertical and horizontal axis of the curve denote sensitivity and its (1-specificity) respectively. A bisector line can also be seen in Figure 4.

As shown in Figure 4, the ROC Curve is divided into three parts as follows:

(1) Above the bisector line: There are points in this area that have a higher sensitivity or true positive rate (TPR) than false positive rate (FPR). The higher the ROC Curve above the bisector line, the better model performance and more reliable results will be achieved.
(2) On the bisector line: In this area, the numerical values of the true positive rate (TPR) and the false positive rate (FPR) are equal. This means the model has not learned anything and is predicting the classes by chance.
(3) Below the bisector line: There are points in this area whose sensitivity or true positive rate (TPR) is lower than false positive rate (FPR). When the rock curve (ROC Curve) is below the bisector line, the model has a very poor performance and the outcomes cannot be the basis of any analysis.

Therefore to evaluate the model performance we have the notion of AUC (area under the curve). The numerical value of AUC is clearly a real number between zero and one and indicates the power of detection or accuracy of the model. If the number is close to one, it means the point are generally at the top of the bisector line, and the model possesses good detection ability or accuracy.

3. Results

In this paper, the classification model (logistic regression) and ensemble model (random forest) are used to investigate the effect of variables on the probability of car accident injuries in third party liability insurance. Due to the imbalance of the target variable, before applying the algorithms, over-sampling and under-sampling techniques and their combination have been used on the dataset.

The following graph depicts the outcome of random forest on the dataset along with over-sampling, under-sampling and under-over-sampling techniques on the dataset:

In Figure 5, the first column from the left shows the original data without applying any resampling technique. The resampling methods are divided into over-sampling and under-sampling techniques which are abbreviated in Figure 4 and for

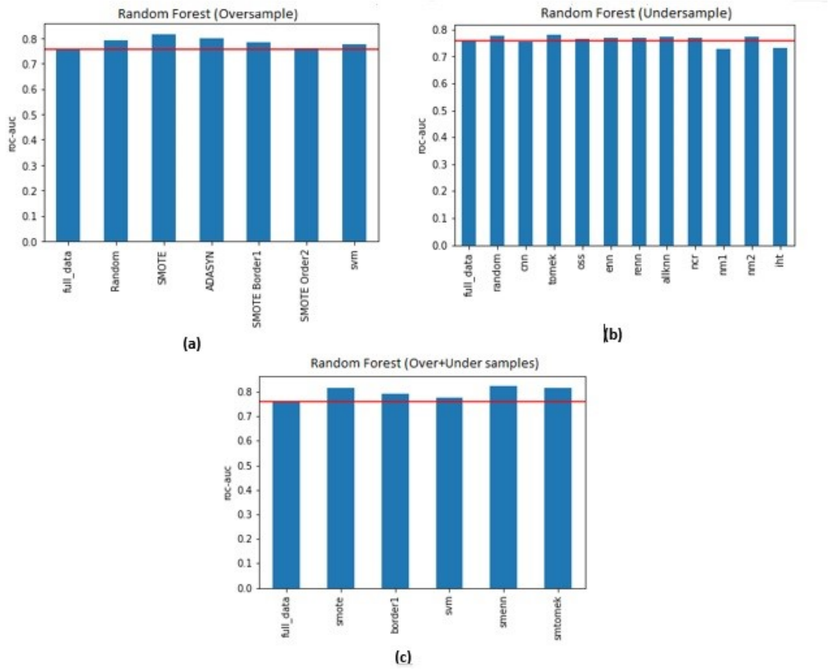

Figure 5. (a) random forest with over-sampling methods. (b) random forest with under-sampling methods. (c) random forest with combination of under and over-sampling methods.

over-sampling we have; Random= random oversampling, SMOTE= synthetic minority over-sampling technique, ADASYN= adaptive synthetic, SMOTE Border= Borderline SMOTE, Svm= SVM SMOTE. The abbreviations for under-sampling methods stand for; Random= random under-sampling, cnn=condensed nearest neighbors, tomek= Tomek links, oss=one sided selection, enn=edited nearest neighbors, renn=repeated edited nearest neighbors, allknn=all k-nearest neighbors, ncr=neighborhood cleaning rule, nm=near miss, iht=instance hardness threshold.

The vertical axes demonstrates the area under the curve (AUC) as a summary of the ROC curve for different resampling methods which measures the ability of a classifier to distinguish between classes. The red horizontal line indicates the performance of the random forest model on the original data and other outputs can be evaluated based on this benchmark. Furthermore, for each over-sampling method, hyperparameters are tuned through k-fold cross validation (k=10) and grid search approach. In cross validation the training data is randomly split into 10 folds in which one fold is used iteratively as a validation set to evaluate the performance of the model.

Figure 6, shows the best result for random forest model in the case that we apply SMOTE (over-sampling) and ENN (under-sampling) as a hybrid resampling method on the data. Moreover, hyperparameters are set as number of estimators (=300), maximum features (=sqrt) and maximum depth (=6).

Next we run logistic regression along with different resampling methods and the following graph demonstrate the outcome:

In Figure 7, the first column from the left shows the original data without applying any resampling technique. The red horizontal line indicates the performance of the random forest model on the original data and other outputs can be evaluated based on this benchmark. Furthermore, for each over-sampling method, hyperparameters are tunned through k-fold cross validation (k=10) and the grid search approach. Figure 8, shows the best result for the logistic regression model in the

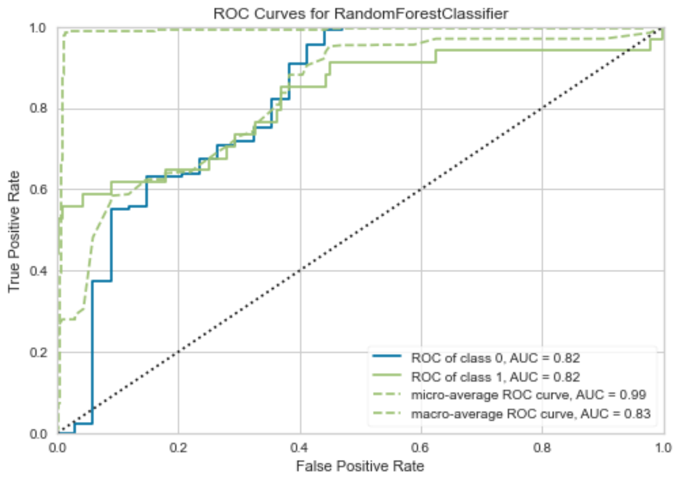

Figure 6. ROC AUC for the random forest model with SMOTENN method.

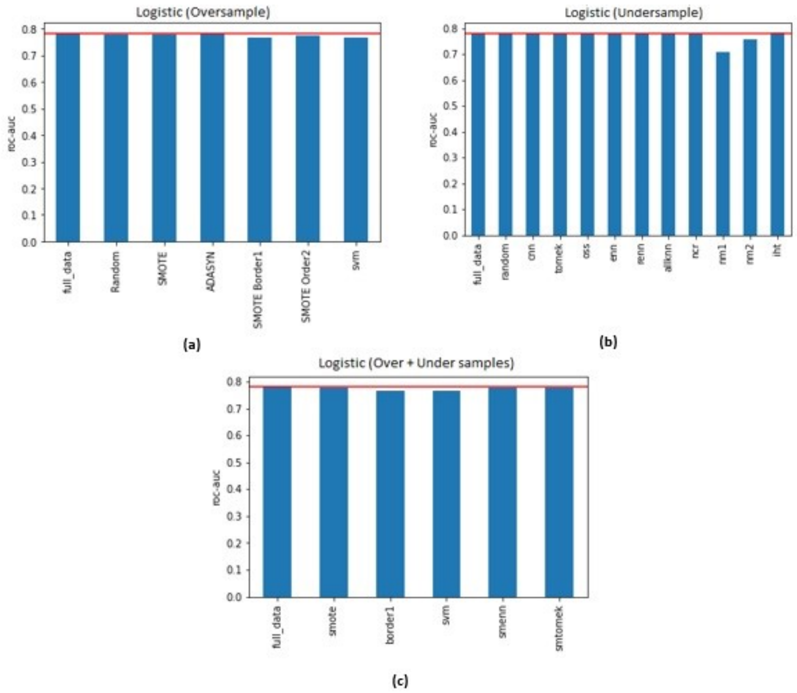

Figure 7. (a) logistic regression with over-sampling methods. (b) logistic regression with under-sampling methods. (c) logistic regression with combination of under and over-sampling methods.

case that we apply ADASYN as the over-sampling method on the data. Moreover, hyperparameters are set as the C parameter (=1000), and the penalty (=L2).

For a better understanding of the model results, the following diagram compares the performance of logistic regression and random forest with respect to the ROC AUC metric: Given that the random forest model has shown the best performance, it has been used to prioritize the effective variables, which is presented in the figure below:

Figure 10 shows the impact of various features on the occurrence of bodily injured accidents in third party liability insurance. The next section provides a detailed interpretation of how the features affect the prediction.

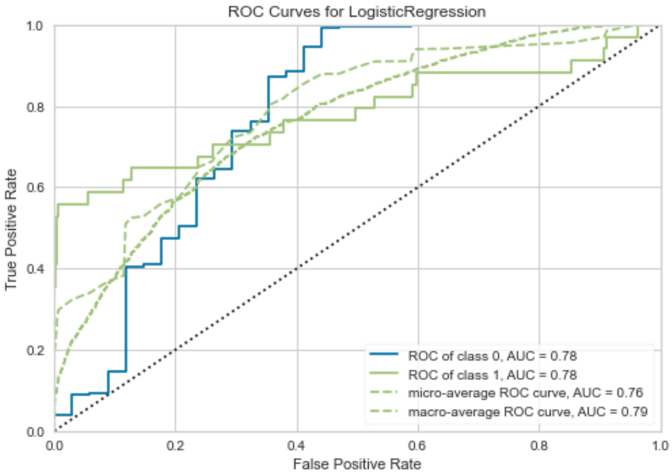

Figure 8. ROC AUC for the logistic regression model with SMOTENN method.

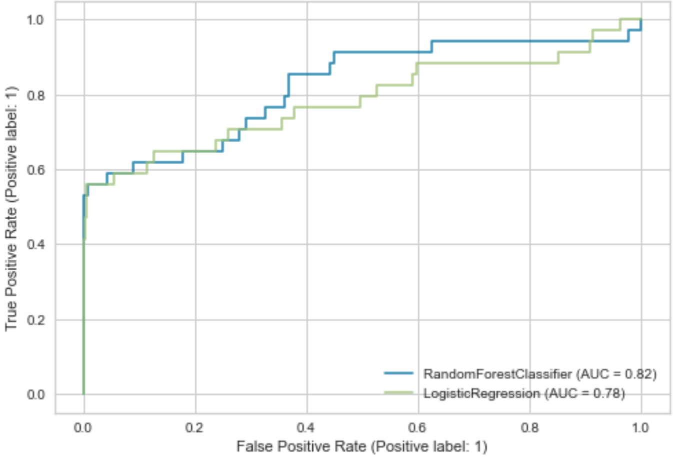

Figure 9. Comparison of logistic regression and random forest.

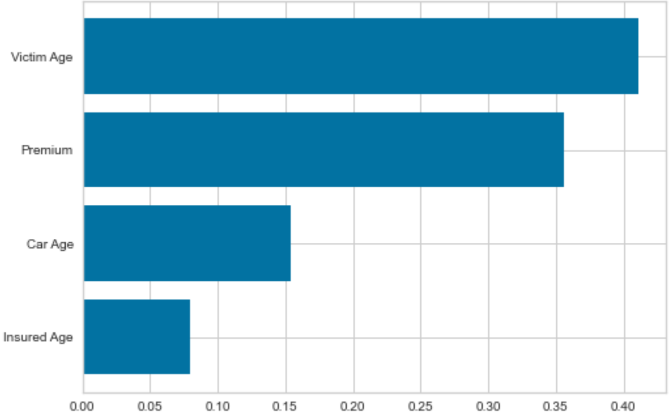

Figure 10. Importance of features on classification task.

4. Discussion

As can be deduced from the results, victim age is the main factor explaining the number of injured in traffic accidents. Car body premium is the second most important factor affecting the number of injured. In fact this variable indicates the value of the car, or in other words, the quality of the car. As we would expect, cars with more modern control and braking systems (i.e. higher priced cars) play a crucial role in preventing injury.

Another significant variable in this study is the car age, which is logically similar to the car value and is somewhat correlated with the change in car value. The last important variable explaining the extent of the injury is the age of the driver involved in an accident. Since most injuries are related to low-cost and high-consumption cars such as taxi drivers, we expect the age distribution, especially for the tails, look similar. As a result, it is expected that insured age is not a very important and determining factor in predicting the occurrence of an accident.

According to the results, it can be said that older ages of the victims are the main cause of injuries and another important factor is the quality of the culprit vehicles. The third and fourth priorities are related to the car age and the age of the insured, respectively.

In this regard, due to the high importance of vehicle quality as a controllable factor in the rate of injuries, it is expected that by changing the quality of insured vehicles, one can observe a significant reduction in the rate of injuries in the insurance portfolio.

Acknowledgements

We express our sincere thanks to the anonymous reviewers for their incisive comments on an earlier version of this manuscript which led to this improved version.

References

[1] P. Baecke and L. Bocca, The value of vehicle telematics data in insurance risk selection processes, Decision Support Systems, **98** (2017) 69–79.

[2] R. Barandela, R. M.Valdovinos, J. S. Snchez and F. J. Ferri, The imbalanced training sample problem: Under or over sampling?, In Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR), Springer, Berlin, Heidelberg, (2004) 806–814.

[3] Y. Bian, C. Yang, J. L. Zhao and L. Liang, Good drivers pay less: A study of usage-based vehicle insurance models, Transportation research part A: policy and practice, **107** (2018) 20–34.

[4] N. Boodhun and M. Jayabalan, Risk prediction in life insurance industry using supervised learning algorithms, Complex & Intelligent Systems, **4 (2)** (2018) 145–154.

[5] R. L. Brown, D. Charters, S. Gunz and N. Haddow, Age as an Insurance Rate Class Variable, University of Waterloo, (2004) 103–114.

[6] L. Cao and H. Shen, Imbalanced data classification using improved clustering algorithm and under-sampling method, In 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), IEEE, (2019) 358–363.

[7] N. V. Chawla, Data mining for imbalanced datasets: An overview, Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA, (2009) 875–886.

[8] D. Devi, S. K. Biswas and B. Purkayastha, A review on solution to class imbalance problem: Under-sampling approaches, In 2020 International Conference on Computational Performance Evaluation (ComPE), IEEE, (2020) 626–631.

[9] G. Dionne and C. Vanasse, Automobile insurance ratemaking in the presence of asymmetrical information, Journal of Applied Econometrics, **7 (2)** (1992) 149–165.

[10] K. Divakar and K. Chitharanjan, Performance evaluation of credit card fraud transactions using boosting algorithms, Int. J. Electron. Commun. Comput. Eng. IJECCE, **10 (6)** (2019) 262–270.

[11] G. Douzas, F. Bacao and F. Last, Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE, Information Sciences, **465** (2018) 1–20.

[12] A. Fernndez, S.Garca, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, Cost-sensitive learning, In Learning from Imbalanced Data Sets, Springer, Cham, (2018) 63–78.

[13] Y. L. Grize, W. Fischer and C. Ltzelschwab, Machine learning applications in nonlife insurance, Applied Stochastic Models in Business and Industry, **36 (4)** (2020) 523–537.

[14] H. Han, W. Y. Wang and B. H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, In International Conference on Intelligent Computing, Springer, Berlin, Heidelberg, (2005) 878–887.

[15] S. E. Harrington and H. I. Doerpinghaus, The economics and politics of automobile insurance rate classification, Journal of Risk and Insurance, **60 (1)** (1993) 59–84.

[16] P. Hart, The condensed nearest neighbor rule (corresp.), IEEE Transactions on Information Theory, **14 (3)** (1968) 515–516.

[17] J. Hegde and B. Rokseth, Applications of machine learning methods for engineering risk assessmentA review, Safety Science, **122** (2020) 104492.

[18] Y. Huang and S. Meng, Automobile insurance classification ratemaking based on telematics driving data, Decision Support Systems, **127** (2019) 113156.

[19] R. Jain, J. A. Alzubi, N. Jain and P. Joshi, Assessing risk in life insurance using ensemble learning, Journal of Intelligent & Fuzzy Systems, **37 (2)** (2019) 2969–2980.

[20] M. Kelly and N. Nielson, Age as a variable in insurance pricing and risk classification, The Geneva Papers on Risk and Insurance-Issues and Practice, **31 (2)** (2006) 212–232.

[21] S. B. Khakbaz, N. Hajiheydari and M. Pourestarabadi, Car insurance risk assessment with data mining for an Iranian leading insurance company, International Journal of Business and Economics Research, **3 (3)** (2014) 128–134.

[22] M. Kubat and S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, In Icml, **97 (1)** (1997) 197.

[23] R. Malhotra and J. Jain, Handling imbalanced data using ensemble learning in software defect prediction, In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, (2020) 300–304.

[24] H. M. Nguyen, E. W. Cooper and K. Kamei, Borderline over-sampling for imbalanced data classification, International Journal of Knowledge Engineering and Soft Data Paradigms, **3 (1)** (2011) 4–21.

[25] N. Paltrinieri, L. Comfort and G. Reniers, Learning about risk: Machine learning for risk assessment, Safety Science, **118** (2019) 475–486.

[26] C. V. Priscilla and D. P. Prabha, Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection, In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, (2020) 1309–1315.

[27] S. Rawat, A. Rawat, D. Kumar and A. S. Sabitha, Application of machine learning and data visualization techniques for decision support in the insurance sector, International Journal of Information Management Data Insights, **1 (2)** (2021) 100012.

[28] D. Samson and H. Thomas, Linear models as aids in insurance decision making: the estimation of automobile insurance claims, Journal of Business Research, **15 (3)** (1987) 247–256.

[29] Z. Shams Esfandabadi and M. M. Seyyed Esfahani, Identifying and classifying the factors affecting risk in automobile hull insurance in Iran using fuzzy Delphi method and factor analysis, Journal of Industrial Engineering and Management Studies, **5 (2)** (2018) 84–96.

[30] V. Sobanadevi and G. Ravi, Handling data imbalance using a heterogeneous bagging-based stacked ensemble (HBSE) for credit card fraud detection, In Intelligence in Big Data TechnologiesBeyond the Hype, Springer, Singapore, (2021) 517–525.

[31] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu and Y. Zhou, A novel ensemble method for classifying imbalanced data, Pattern Recognition, **48 (5)** (2015) 1623–1637.

[32] Y. Tang, Y. Q. Zhang, N. V. Chawla and S. Krasser, SVMs modeling for highly imbalanced classification, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **39 (1)** (2008) 281–288.

[33] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi and M. Asadpour, Boosting methods for multi-class imbalanced data classification: an experimental review, Journal of Big Data, **7 (1)** (2020) 1–47.

[34] N. Thai-Nghe, Z. Gantner and L. Schmidt-Thieme, Cost-sensitive learning methods for imbalanced data, In The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, (2010) 1–8.

[35] I. Tomek, Two modifications of CNN, IEEE Trans. Systems, Man and Cybernetics, **6** (1976) 769–772.

[36] P. Tryfos, On classification in automobile insurance, The Journal of Risk and Insurance, **47 (2)** (1980) 331–337.

[37] C. F. Tsai, W. C. Lin, Y. H. Hu and G. T. Yao, Under-sampling class imbalanced datasets by combining clustering analysis and instance selection, Information Sciences, **477** (2019) 47–54.

[38] W. A. Wiegers, The use of age, sex, and marital status as rating variables in automobile insurance, The University of Toronto Law journal, **39 (2)** (1989) 149–210.

[39] S. J. Yen and Y.S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications, **36 (3)** (2009) 5718–5727.

[40] J. L. Yin and B. H. Chen, An advanced driver risk measurement system for usage-based insurance on big driving data, IEEE Transactions on Intelligent Vehicles, **3 (4)** (2018) 585–594.

[41] M. Zareapoor and P. Shamsolmoali, Application of credit card fraud detection: Based on bagging ensemble classifier, Procedia Computer Science, **48 (2015)** (2015) 679–685.

[42] S. Zhang, Cost-sensitive KNN classification, Neurocomputing,**391** (2020) 234–242.

[43] Z. Zheng, Y. Cai and Y. Li, Oversampling method for imbalanced classification, Computing and Informatics, **34 (5)** (2015) 1017–1037.

[44] K. Zhuang, S. Wu and X. Gao, Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms, Tehniki vjesnik, **25 (6)** (2018) 1783–1791.