Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches

Mohamed Hanafy[1], Ruixing Ming[2]
School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China[1, 2]
Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Assuit University, Asyut, Egypt[1]

***Abstract*—Predicting the frequency of insurance claims has become a significant challenge due to the imbalanced datasets since the number of occurring claims is usually significantly lower than the number of non-occurring claims. As a result, classification models tend to have a limited ability to predict the occurrence of claims. So, in this paper, we'll use various data level approaches to try to solve the imbalanced data problem in the insurance industry. We developed 32 machine learning models for predicting insurance claims occurrence {(under-sampling, over-sampling, the combination of over-and under-sampling (hybrid), and SMOTE)× (three Decision tree models, three boosting models, and two bagging models) = 32}, and we compared the models' accuracies, sensitivities, and specificities to comprehend the prediction performance of the built models. The dataset contains 81628 claims, each of which is a car insurance claim. There were 5714 claims that occurred and 75914 claims that didn't occur. According to the findings, the AdaBoost classifier with oversampling and the hybrid method had the most accurate predictions, with a sensitivity of 92.94%, a specificity of 99.82%, and an accuracy of 99.4%. And with a sensitivity of 92.48%, a specificity of 99.63%, and an accuracy of 99.1%, respectively. This paper confirmed that when analyzing imbalanced data, the AdaBoost classifier, whether using oversampling or the hybrid process, could generate more accurate models than other boosting models, Decision tree models, and bagging models.**

Keywords—Machine learning; classification; insurance; imbalanced data problem; resampling methods

I. Introduction

The use of machine learning techniques and the transformation of the insurance market into a new level of digital applications are the insurance industry's current challenges. There are two types of insurance: life insurance and non-life insurance. Non-life insurance, specifically auto insurance, is the subject of this study.

A variety of variables influence automobile insurance pricing [1]. And these factors would affect the cost of a client's insurance policy. Credit history is an example of one of these factors; studies indicate that people with poor credit are more likely to file claims, commit fraud, or miss payments, putting the insurance company in a financial bind. Another factor to consider is the client's location; studies have shown that densely populated areas with heavy traffic have a higher rate of accidents, resulting in a higher number of claims. This would result in a significant rise in the customer's insurance premium. However, it is unjust for a good client to pay more simply because of where they live; this creates a problem for the consumer because if the insurance premium is raised, he will be unable to afford it, resulting in the insurance provider losing these clients. So, necessitating the creation of an appropriate method for evaluating the risk each client poses to insurers.

As a result, insurance rates should be adjusted based on a client's skill and other personal information, making car insurance more accessible to consumers. Where insurance companies should customize a custom premium for each customer because this will help the insurers to adjust to any situation and manage any loss. Since It would be unreasonable to expect a client with a good driving record to pay the same insurance premium as a client with a poor driving record; as a result, the model should classify which clients are unlikely to file claims, lower their insurance costs, and raise insurance costs for those who are likely to file claims.

Data imbalanced problem are more likely to arise in the case of insurance data since the number of occurring claims is usually significantly lower than the number of non-occurring claims. And one of the major problems with machine learning techniques is that they are affected by the data set's unequal binary class distribution. In other words, when the data is unbalanced, certain machine learning techniques will simply ignore the small class and allocate the majority of the unseen cases to the common class, resulting in high overall model accuracy. Nonetheless, the performance of the prediction models for the small class will be substantially reduced. To solve this problem, we will use resampling techniques, such as Over-sampling, under-sampling, the combination of over-and under-sampling (hybrid), and the synthetic minority over-sampling technique (SMOTE), to improve the classification efficiency for imbalanced data.

We used a large dataset given by a large automotive company based in Egypt. In this study, we apply data-level approaches that could reduce overfitting caused by data imbalance. And we built 32 machine learning models for predicting the occurrence of auto insurance claims ((under-sampling, over-sampling, hybrid of over-and under-sampling, and SMOTE) ×(three Decision tree models, three boosting models, and two bagging models) =32). And we compared the models' accuracy, sensitivities, and specificities to better understand the built models' prediction efficiency.

The following is the structure of this paper: Section II presents the previous studies. Section III explain the data collection, machine learning models, and data-level approaches. Section IV compared the results of the built thirty-two prediction models. Section V presents concludes. Section VI presents the future work.

II. Related Work

Over the last decade, many researchers have used machine learning algorithm to forecast the occurrence of auto insurance claims. And while machine learning models are efficient at predicting. But when the data is unbalanced, machine learning techniques will simply ignore the small class and allocate the majority of the cases to the common class, resulting in high overall model accuracy. Nonetheless, the performance of the prediction models for the small class will be substantially reduced. The following studies show a lack of using the resampling methods to solve the unbalanced data problem except for the study of [1] that only used the oversampling method.

In the study of machine learning approaches for auto insurance big data [1], they built eight classifiers to predict the occurrence of the claims using big insurance data, including XGBoost, J48, RF, C5.0, CART, K-NN, logistic regression, and naïve Bayes algorithms, and they handled the heavy imbalanced data using the over-sampler method. The RF model performed the best among the eight models. And [2] used two competing methods, XGBoost, and logistic regression, to predict the frequency of motor insurance claims. This study shows that the XGBoost model is slightly better than logistic regression. Furthermore, a model for predicting insurance claims was developed by [3]; they built four classifiers to predict the claims, including XGBoost, J48, ANN, and naïve Bayes algorithms. The XGBoost model performed the best among the four models. Another example of a similar and satisfactory solution to the same problem is the thesis “Research on Probability-based Learning Application on Car Insurance Data” by [4], which used only a Bayesian network to classify either a claim or no claim. And the [5] research also aims to look at data mining techniques for creating a predictive model for auto insurance claim prediction. And they compared three ML methods for predicting claims. Their findings showed that the best predictor was the neural networks.

In summary, despite the relevance of the imbalanced data problem in the insurance industry, there is a lack of comprehensive comparison among the prominent resampling approaches as a strategy to deal with it. The purpose of this study is to investigate the impact of the unbalanced data problem on the performance of machine learning models. This paper solves the unbalanced data problem with several resampling approaches and compares them while utilizing various machine learning classifiers to fill in the gaps in the literature.

In comparison to prior studies, the following are the novel innovations and vital procedures of this study:

- Applying and comparing several resampling algorithms, including the Random Over Sampler, Random Under Sampler, SMOTE, and the hybrid.
- Appling several machine learning models, such as three Decision tree models, three boosting models, and two bagging models, to compare the performance of resampling methods.
- Measuring the effectiveness of implemented machine learning models utilizing various performance measures such as Accuracy, Sensitivity, and Specificity.
- Demonstrating how resampling affects the performance of classifiers.

III. Methods

A. Data Collection

Our dataset is progressive record keeping, which usually updates over time to reflect the updated status of a particular customer, which means that provided data is the snapshot of some particular date, where all the records show the updated status of each customer. This dataset is updated on changes in circumstances of the customer, such as marital status, age, etc. The sample auto insurance claims dataset was collected between 2014 and 2018.

The data used in this study is real-life data obtained from an Egyptian car insurance firm; we end up with 81628 claims in the dataset, each of which is a car insurance claim. In total, there are 5714 claims that occurred, and 75914 non- occurred claims, suggesting that the data is heavily unbalanced. And as we mention, the performance of classification algorithms is greatly affected by imbalanced data. So, we apply four resampling techniques to solve the problem of data imbalance. Each claim comprises 17 features. Table I provides Attributes of the data.

B. Data Preprocessing

Numeric values are allocated to categorical variables. For example, instead of male or female as the gender of the insured, the "Male" component would be (1), and "female" would be (0). After this phase, we can apply our data to all machine learning models.

C. Machine-Learning for Auto Insurance Claims Occurrence:

1) Decision trees: A decision tree D, in more formal terms, is made up of two kinds of nodes:

- A leaf node that represents the response variable's given class/region.
- A decision node that defines a test on a single attribute (predictor variable) with one branch and subtree for each test outcome.

Using a recursive divide and conquer method, a decision tree can be used to classify an observation by beginning at the top decision node (called the root node) and going down through the other decision nodes until a leaf is encountered.

The following models are the most well-known methodology for constructing decision trees [6,7,8].

a) CART algorithm.

b) C5.0 algorithm.

c) C4.5 algorithm.

TABLE I. ATTRIBUTES OF THE DATA

No.	Description	Descriptive statistics
1	Representing the age of the insured.	Min. :18.00 Mean :44.84 Max. :81.00
2	Represent the Value of the car in Egyptian pounds.	Min. :75000 Mean :304003.5 Max. :6610100
3	Represent the car age in years.	Min. :0.000 Mean :10.298 Max. :28.000
4	Represent the car type.	Such as Mercedes, BMW, and Toyota, etc.
5	Represent the car use.	Commercial:30148 Private: 51480
6	Represent the insured education level.	High School :12244 Bachelors :32856 Masters :16325 PhD :7347 Other :12856
7	Represent the annual Income of the insured In Egyptian pounds.	Min. :0 Mean :61572 Max. :367030
8	Represent the number of dependants for the insured.	Min. :0.0000 Mean :3.000 Max. :5.0000
9	Represent the Marital status	Married:48977 not married:32651
10	Represent the insured's occupation.	The job of the insured
11	Represent the claim frequency.	Min. :0.0000 Mean :0.8007 Max. :5.0000
12	Represent if the insured license was revoked before.	No:71880 Yes:9748
13	Represent The insured gender.	MALE:38365 FAMLE:43263
14	Represent the Distance to work in km	Min. :5.00 Mean :33.42 Max. :142.00
15	Represent where the insured live urban vs. rural area.	Urban:66118 Rural:15510
16	Represent the number of Years on job for the insured(yoj).	Min. :0.00 Mean: 10.47 Max. :23.00
17	Claim occurred or not. (the target variable)	non occurred: 75914 occurred: 5714

2) Bagging trees: Bagging, or bootstrap aggregation, is an ensemble meta-algorithm. This algorithm increases the model's consistency and accuracy while also reducing overfitting. In classification, it weights the output to ensemble into a single output. Leo Breiman suggested bagging in 1996 [9] as a way to improve classification results.

The following are the two most common bagging algorithms:

a) Bagged CART.

b) Random forest.

3) Boosting: Boosting is an ensemble technique that, like training, creates several individual models sequentially. Each new model attempts to correct the errors of the previous group of models. Boosting, like bagging, can be used to improve any supervised machine learning algorithm. Boosting, on the other hand, is most effective when weak learners are used as sub-models. As a result, boosting has historically been used on shallow decision trees. By shallow, I mean a decision tree with a limited number of levels of depth or a single split. Boosting's aim is to bring together a group of weak learners to form a strong ensemble learner [10,11,12].

The following are the two most common bagging algorithms:

a) AdaBoost.

b) XGBoost.

c) Stochastic Gradient Boosting.

D. A Data-level Approach and Imbalanced Data

The imbalanced data problem exists in many datasets; as a result, classifiers models are biased against the minority class and are unable to predict it accurately [13]. In contrast, most machine learning models perform better when applied with balanced datasets [14,15,16,17].

Analysis of our database shows that they are extremely imbalanced, and the two forms of insurance claims are not balanced, with 93% (n=75914) of the auto insurance claims occurred, and those non-occurred were 7% (n=5714). As a consequence, the imbalanced data problem must be addressed. Many techniques have been developed to resolve the problem of unbalanced data. One of the most successful approaches for addressing unbalanced data is using a sampling-based approach, either Random Over Sampler [18], Random Under Sampler [19], and SMOTE [20].

We will use the ROSE and the ovun.sample function incorporates more conventional class inequality solutions, such as over-sampling the minority class, under-sampling the majority class, or a combination of over-and under-sampling. And also, we will use the DMwR package to apply SMOTE as a resampling method.

1) Over-sampling technique: This technique increases the weight of the minority class. It's important to note that the technique of over-sampling is typically used more than other methods.

a) Random over sampler: Random Over-Sampling is a technique based on bootstrap that supports the binary classification task in the presence of unbalanced classes by generating synthetic examples from a conditional density estimation of the two classes [21]. It handles both continuous and categorical data by randomly replicating samples from the minor class [22]. As a result of this process, the dataset grows in size. The argument is that no new samples are generated by a random over-sampler, and the variety of samples remains constant. Since the sample size grows, the oversampling technique takes longer to construct a model and can cause overfitting because it duplicates samples from a minor class. [23,24].

b) SMOTE: SMOTE is similar to random oversampling. However, it does not regenerate the same instance. It creates a new instance by appropriately combining existing instances, thus making it possible to avoid the disadvantage of overfitting to a certain degree. Moreover, SMOTE is an oversampling technique that produces new minority samples by combining two minorities and one of their K nearest neighbours [25]. This approach is a statistical technique for creating new instances to increase the number of minority samples in a dataset. This algorithm takes characteristic features for the target class and its closest neighbours, then produces new samples by combining the characteristics of a specific case with those of its neighbours.

2) Random under sampler: Under-sampling is one of the simplest techniques to dealing with the problem of unbalanced data. It balances the majority and minority classes. The process of under-sampling includes arbitrarily deleting examples from the majority class in the training dataset, referred to as random under-sampling [26]. By reducing the amount of data, under-sampling can save time when building a model, but it comes at the cost of losing information [27,28].

3) Hybrid methods: There are several benefits and drawbacks of over-sampling and under-sampling. Combining these two strategies will add the strengths of these two methods to a new method [26].

E. Development of Prediction Models and Prediction Performance Evaluation

This study built thirty-two prediction models ((under-sampling, oversampling, the combination of over-and under-sampling (hybrid), and SMOTE)× (three Decision tree models, three boosting models, and two bagging models) =32). The accuracy, sensitivity, and specificity of each model are used to compare the prediction performance of the established models.

This study randomly divided the data into a training dataset and a test dataset at a ratio of 7:3. The hyperparameters are tuned using a 10-fold cross-validation that was performed only on the training dataset to get the best performance for each machine learning model. And the test dataset was used to evaluate the prediction performance. The Data-level Approaches must only be applied to the training set while the test set still unbalanced. We used R version 4.0.2 to conduct all analyses.

F. Evaluation Methods

Evaluation methods are essential in comparing and selecting the best model because they are assessing the efficiency of classifiers [1].

Table II shows the Evaluation methods used in this study. Where TP is the number of true positives, the number of false positives is FP, the number of true negatives is TN, and the number of false negatives is FN.

TABLE II. EVALUATION METHODS

Accuracy	Referred to the overall correctly prediction	$\frac{(TP + TN)}{(TP + FP + TN + FN)}$
Sensitivity	Referred to the correct rate of predicting the occur claims	$\frac{TP}{(TP + FN)}$
Specificity	Referred to the correct rate of predicting the non-occur claims.	$\frac{TN}{(FP + TN).}$

IV. RESULTS AND DISCUSSION

To show the difference between the ability of the machine learning classifiers to predict the insurance claims occurrence before and after handling the unbalanced data problem, we compared all applied models on the unbalanced data and also on the balanced data created by different resampling techniques. We measure the performance of models on the testing data using accuracy, sensitivity, and specificity.

A. Comparing the Performance of the Built Machine Learning Models

Tables III, IV, and V show the accuracy, sensitivity, and specificity respectively of the thirty-two prediction models, as well as Fig. 1.

Table III shows the Accuracy of each machine learning technique on unbalanced data and balanced datasets generated by four different resampling models. And we must consider that only if the data is balanced will Accuracy be a valuable metric, while when the data is unbalanced, the Accuracy would be meaningless. Because when the data is unbalanced, most machine learning techniques will simply ignore the small class and allocate most of the unseen cases to the majority class, resulting in high overall model accuracy and high specificity, while the sensitivity of the models will substantially be reduced. The AdaBoost classifier achieved 99.4 %accuracy by using the oversampling, which is the highest of all other classifiers. And the lowest accuracy outcome goes to the C5.0 with 74.84% by using the under-sampling.

Table IV refers to the Sensitivity of the machine learning models. Sensitivity relates to the ability to predict the occurrence of claims. We can note that the Sensitivity for all ML models with the unbalanced data is lower than the Sensitivity for balanced data created by different resampling methods because the occurred claims represent a small class with only 7% in our data. Therefore, before solving the unbalanced data problem, machine learning techniques will simply ignore the small class (occurred claims). Thus, resulting in very low Sensitivity in the case of the unbalanced data. While the Sensitivity is improved after applied the resampling methods. This refers to the effectiveness of using the resampling methods to handle the unbalanced data problem in the insurance industry. And the highest Sensitivity goes to the AdaBoost classifier with 92.94% using the oversampling, while the lowest one goes to the AdaBoost model with 0.46% using the unbalanced data.

TABLE III. THE ACCURACY OF THE DEVELOPED MODELS

Models	Unbalanced	OVER	UNDER	HYBRID	SMOTE
Decision trees models					
C5.0	0.9393	0.9426	0.7484	0.9822	0.8441
C4.5	0.9432	0.96	0.794	0.9544	0.8117
CART	0.9441	0.7791	0.7586	0.8076	0.8316
Bagging models					
Bagged CART	0.939	0.928	0.7487	0.9044	0.8266
Random forest	0.94	0.978	0.7807	0.9806	0.8533
Boosting models					
AdaBoost	0.9385	0.994	0.7516	0.9919	0.8689
XGBoost	0.9386	0.8786	0.7951	0.8715	0.8641
Stochastic Gradient Boosting	0.9396	0.7865	0.7648	0.796	0.8461

TABLE IV. THE SENSITIVITY OF THE DEVELOPED MODELS

Models	unbalanced	OVER	UNDER	HYBRID	SMOTE
Decision trees models					
C5.0	0.1422	0.8415	0.7768	0.8878	0.8098
C4.5	0.1268	0.713	0.6333	0.8633	0.8064
CART	0.1463	0.7098	0.7024	0.6707	0.8
Bagging models					
Bagged CART	0.161	0.1904	0.7309	0.3873	0.6499
Random forest	0.0456	0.8223	0.7244	0.8428	0.8064
Boosting models					
AdaBoost	0.0046	0.9294	0.7175	0.9248	0.8109
XGBoost	0.1139	0.7744	0.7153	0.795	0.8337
Stochastic Gradient Boosting	0.1162	0.7358	0.7585	0.7289	0.7722

Table V refer to the Specificity of the machine learning models. Specificity refers to the ability to predict non-occurred claims. We can note that the Specificity for all models with unbalanced data is highest than the Specificity for balanced data created by different resampling methods because the non-occurred claims represent the majority class with 93% in our data. Therefore, before solving the unbalanced data problem, machine learning techniques will allocate the most unseen cases to the majority class (non-occurred claims). This is resulting in very high overall model Specificity in the case of the unbalanced data. But our objective is to detect MINOR class more accurately than MAJOR class; therefore, we are interested in Sensitivity more than Specificity. SO, we need to resample the dataset to force algorithms to identify both classes with equal importance. And the highest Specificity in the dataset belongs to AdaBoost classifiers with 99.93% using the unbalanced data, and the lowest one goes to the C5.0 model with 74.65 % using the under-sampling.

Last but not least, from Tables III, IV, V, and Fig. 1, we can conclude that using the resampling methods is very effective for handle the unbalanced data problem in the insurance industry, because the best results are achieved after applied the data-level approaches.

And the best models are AdaBoost with the over and hybrid methods, then the C5.0 model with the hybrid method, and then the random forest model with the hybrid method. Where AdaBoost with the over and hybrid methods achieved a sensitivity of 92.94%, a specificity of 99.82%, and an accuracy of 99.4%. And a sensitivity of 92.48%, a specificity of 99.63%, and an accuracy of 99.19%, respectively. And the C5.0 model with the hybrid method has a sensitivity of 88.78%, a specificity of 98.79%, and an accuracy of 98.22%. Then there's the random forest model with the hybrid method, which has a sensitivity of 84.28%, a specificity of 98.96%, and an accuracy of 98.06%.

TABLE V. THE SPECIFICITY OF THE DEVELOPED MODELS

	unbalanced	OVER	UNDER	HYBRID	SMOTE
Decision trees models					
C5.0	0.9935	0.9487	0.7465	0.9879	0.8477
C4.5	0.9924	0.9761	0.8045	0.9604	0.812
CART	0.9922	0.7832	0.7619	0.8159	0.8335
Bagging models					
Bagged CART	0.9859	0.9781	0.7499	0.9396	0.8386
Random forest	0.9982	0.9881	0.7843	0.9896	0.8564
Boosting models					
AdaBoost	0.9993	0.9982	0.7538	0.9963	0.8726
XGBoost	0.9923	0.8854	0.8003	0.8913	0.8654
Stochastic Gradient Boosting	0.9932	0.7898	0.7652	0.8003	0.8509

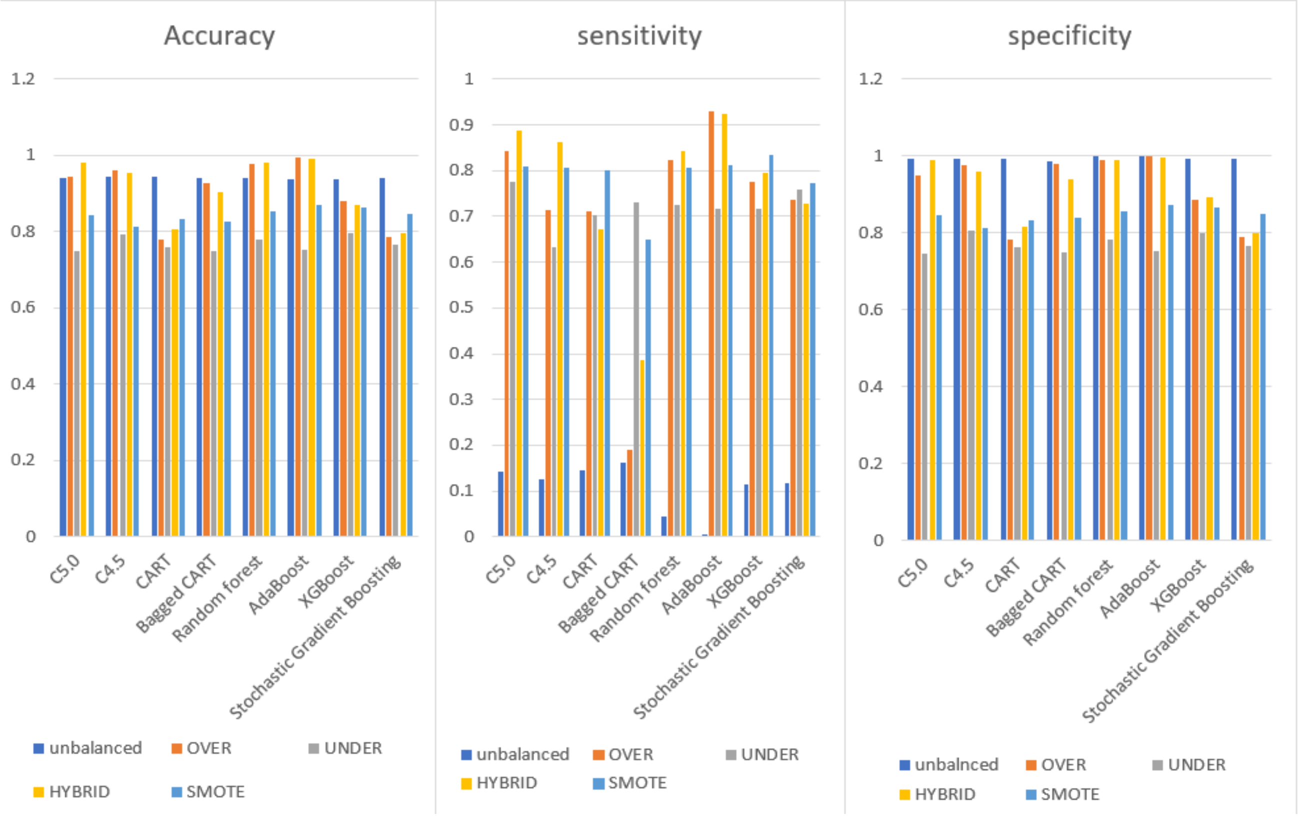

Fig. 1. Comparison between the Developed Models based on the Accuracy, Sensitivity and Specificity.

B. *Variables Importance for Auto Insurance Claims Classification in the AdaBoost with the Oversampling*

The importance of the variables of the final model (AdaBoost with the oversampling) is presented in Fig. 2.

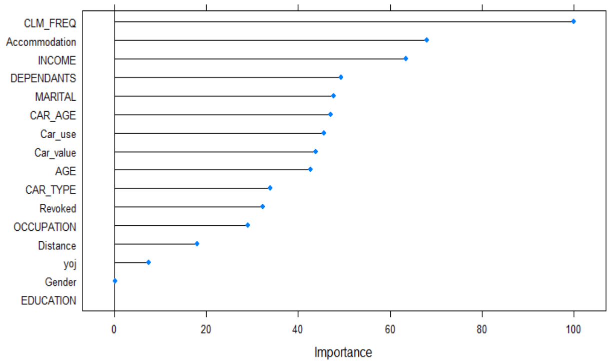

Fig. 2. The Importance of Variables in the AdaBoost Classifier with Oversampling-based Claims Occurrence Prediction Model.

V. Conclusion

This study specifically established models for improving the classification efficiency of imbalanced data by using oversampling, under-sampling, the combination of over-and under-sampling, and SMOTE as resampling approaches. ((under-sampling, oversampling, a combination of over-and under-sampling (hybrid), and SMOTE) × (three Decision tree models, three boosting models, and two bagging models) =32) for predicting auto insurance claims occurrence. According to the findings of this analysis, the AdaBoost model with over and hybrid could generate more accurate models than other boosting models, Decision tree models, and bagging models, then the C5.0 model with the hybrid method, and then the random forest model with the hybrid method.

VI. Future Work

Further research is required to compare the accuracy using various datasets from various fields to prove the prediction efficiency of an AdaBoost classifier with resampling methods to solve the imbalanced data problem. And Future work may be done in the following directions: Using hybrid machine learning classifiers to improve comparison and performance. And also, use different feature selection approaches to enhance model results and gain a deeper understanding of the important features.

References

[1] Hanafy, Mohamed, and Ruixing Ming. 2021. Machine learning approaches for auto insurance big data. *Risks* 9: 42.

[2] Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks* 7: 70.

[3] ABDELHADI, SHADY, KHALED ELBAHNASY, and MOHAMED ABDELSALAM. 2020. A PROPOSED MACHINE MODEL TO PREDICT AUTO INSURANCE CLAIMS USING LEARNING TECHNIQUES. *Journal of Theoretical and Applied Information Technology* 98.

[4] Jing, Longhao, Wenjing Zhao, Karthik Sharma, and Runhua Feng. Year. Research on Probability-based Learning Application on Car Insurance

Data. Paper presented at the 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017).

[5] Weerasinghe, KPMLP, and MC Wijegunasekara. 2016. A comparative study of data mining algorithms in the prediction of auto insurance claims. *European International Journal of Science and Technology* 5: 47-54.

[6] Breiman, L. (1984). Classification and Regression Trees. Routledge.

[7] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[8] Quinlan, J. R. et al. (1996). Bagging, boosting, and c4. 5. In AAAI/IAAI, Vol. 1, pages 725–730.

[9] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.

[10] Kégl, Balázs. 2013. The return of AdaBoost. MH: multi-class Hamming trees. *arXiv preprint arXiv:1312.6086*.

[11] Chen, Tianqi, and Carlos Guestrin. Year. Xgboost: A scalable tree boosting system. Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

[12] Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*: 1189-232.

[13] Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30: 25-36.

[14] Hussain, Lal, Kashif Javed Lone, Imtiaz Ahmed Awan, Adeel Ahmed Abbasi, and Jawad-ur-Rehman Pirzada. 2020. Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. Waves in Random and Complex Media: 1-24.

[15] H. Byeon, Development of a physical impairment prediction model for Korean elderly people using synthetic minority over-sampling technique and XGBoost. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 36-41, 2021.

[16] R. Mohammadi, R. Javidan, M. Keshtgari, and N. Rikhtegar, SMOTE: an intelligent SDN-based multi-objective traffic engineering technique for telesurgery. IETE Journal of Research, vol. 1-11, 2021.

[17] H. Byeon, Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 74-79, 2021.

[18] L. Abdi, and S. Hashemi, To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, 238-251, 2015.

[19] S. J. Yen, and Y. S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, vol. 36, no. 3, pp. 5718-5727, 2009.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[21] Menardi, Giovanna, and Nicola Torelli. 2014. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery* 28: 92-122.

[22] M. Wang, X. Yao, and Y. Chen, An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients. IEEE Access, vol. 9, pp. 25394-25404, 2021.

[23] H. Byeon, Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 74-79, 2021.

[24] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. AlaizMoretón, and I. García-Rodríguez, Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine, vol. 202, pp. 105968, 2021.

[25] Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. Year. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Paper presented at the Pacific-Asia conference on knowledge discovery and data mining.

[26] Lunardon, Nicola, Giovanna Menardi, and Nicola Torelli. 2014. ROSE: A Package for Binary Imbalanced Learning. *R journal* 6.

[27] H. Byeon, Development of a physical impairment prediction model for Korean elderly people using synthetic minority over-sampling technique and XGBoost. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 36-41, 2021.

[28] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, SMOTEWENN: Solving class imbalance and small sample problems by oversampling and distance scaling. Applied Intelligence, vol. 51, no. 3, pp. 1394-1409, 2021.