2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) | 979-8-3503-2848-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICT4DA59526.2023.10302229

Detecting Fraud in Motor Insurance Claims Using XGBoost Algorithm with SMOTE

David Gichohi Maina
Department of Computer Science
Dedan Kimathi University of Technology
Nyeri, Kenya
david.gichohi@dkut.ac.ke

Juliet Chebet Moso
Department of Computer Science
Dedan Kimathi University of Technology
Nyeri, Kenya
juliet.moso@dkut.ac.ke

Patrick Kinyua Gikunda
Department of Computer Science
Dedan Kimathi University of Technology
Nyeri, Kenya
patrick.gikunda@dkut.ac.ke

***Abstract*—Fraudulent claims in motor insurance policies continue to be a big menace to insurance companies. Fraudsters are devising new tactics of fabricating claims to make them appear valid. This makes insurance companies register huge losses in billions of money every year. The insurance policyholders bear these losses through increased premiums thus having negative social and economic ramifications. Numerous approaches have been proposed and applied in detecting and preventing fraudulent claims. The traditional approaches have become complex, time-consuming, and with low success ratio. To improve on fraud detection, the existing historical data can be used to train prediction models. To optimize the performance, this data require feature engineering to ensure only relevant features are used and handling of class imbalance. In this paper, we propose a model that is built on XGBoost algorithm. In data preparation, we propose to handle class imbalance by oversampling, using SMOTE. We aim at comparing the effect of class imbalance and oversampling on the performance of our model. The results obtained reveals that XGBoost performs well with SMOTE compared to imbalanced training dataset and also compared to other algorithms. Once the model is deployed, insurance companies will be able to detect and identify perpetrators of fraud and take necessary action. This will reduce their loss adjustment expenses and thus increase their profits.**

***Index Terms*—fraudulent claims, class imbalance, XGBoost, SMOTE**

I. Introduction

The insurance industry has its basis on risk transfer, in which the insurer, i.e the insurance company, takes in the financial risks that may occur to the insured, i.e the owner of motor vehicle, in the future, in case of an accident. The insured pays premium as monetary compensation in exchange. Fraudulent claims in motor insurance policies continue to be a big menace to insurance companies affecting their economic operations.

Fraud is a deception act committed by a person or entity being aware that it may result in benefits that are adverse to the individual or others. By filing false documents or crafting an accident, the insured seeks financial gain from the insurer through a fraudulent claim [1]. Approximately 10% of all reported claims are fraudulent [2]; but only less than 3% is legally preceded [3]. Fraud can also be committed by agent, broker, staff, service providers, police officers, or drivers who are not policyholders. If an insurance company is able to identify perpetrators of fraud and take necessary action, the customer satisfaction rate will be increased and loss adjustment expenses reduced. The increase in customer satisfaction will reduce instances where customers register fraudulent claims. Reduced loss adjustment increases the profit made by insurance companies and, reduces the premium rating thus making insurance policies affordable. Complexity and dynamic nature of fraudsters make it difficult to eradicate fraud completely. Traditional approaches used by insurance companies include anti-fraud policy [4], whistle-blowing [5], staff rotation, and code of conduct [6]. These approaches are complex, time-consuming, expensive, and have low success ratio. They rely on domain knowledge, intuition, and expert's scrutiny.

These fraud detection techniques proposed, usually identify abnormalities in past motor insurance claim transactions. As the fraudsters evolve and change their tactics, these techniques become infeasible [7]. The historical data is usually complex and has class imbalance problem [8]. To leverage knowledge from this data, there is a need to have a proper mechanism for feature engineering and handle class imbalance in the data to distinguish fraudulent claims from valid claims and reduce the false positives.

In this paper, we propose the use of extreme gradient boosting (XGBoost) algorithm to classify claim as either fraudulent or not. XGBoost has the efficiency to resolve multiple computation problems in various fields. It requires less computational resources and its performance is good. Preprocessing of insurance data is performed to prepare data for model training. Also, to enhance performance of the model, class imbalance problem is addressed by oversampling where synthetic instances of the minority class are generated. A comparison with other classification algorithms such as Random Forest, Logistic Regression, LightGBM, indicates that XGBoost outperforms them.

The rest of the paper is organized as follows. In section II, we provide a review of related work in fraud detection using machine learning algorithms. Section III presents the proposed methodology. In Section IV, experimental results and their discussion are carried out. Finally the paper is concluded and future work drawn in Section V.

979-8-3503-2848-6/23/$31.00 ©2023 IEEE

II. Related Work

In this section, we review some past work by researchers who have actively worked in building and developing fraud detection models to help mitigate this problem faced by the insurance industry.

Many insurance companies look for ways to predict and detect fraudulent claims to enable them take necessary action while reducing their loss adjustment expenses. This being one of the most interesting research areas in insurance industry and financial world at large, various methods have been tried in this domain from supervised learning, to ensemble learning, to deep learning [9]. Huge amount of labelled data is used to train a supervised learning model [10] and this model can be used for classification and regression problems.

A fraud detection study in auto insurance by [11] used Distance and Density Based (Nearest Neighbour) method and Interquartile range method. With 33 features of the original dataset, the accuracy were: SVM- 82%, Distance based- 94.4%, Density based- 35.2% and Interquartile range- 92.1%. Upon feature selection to 7 attributes, the accuracy changes to 82%, 99.9%, 82%, and 98% respectively. While the study attributed performance improvement to feature selection and also reduction of complexity of the model, it did not address the class imbalance in the dataset used.

A study by [12] observed that the number of independent states from the features with very high unique values could be reduced by feature selection. In data preprocessing, categorical values are converted to numerical and this improves the results of the classifier. XGBoost was observed to perform better than Decision Tree and KNN. Without resampling to balance the classes in the dataset, F1-score of classifiers used in the research were: XGBoost- 81%, Decision Tree- 71.86%, and KNN- 68%.

In a comparative study for machine learning methods, [13] compared 14 classifiers. Feature selection was based on correlation. SVM classifier had a good generalization in testing data, Bayesian Network had the highest True Negative Rate while Ensemble methods performed best in True Positive Rate.

An ensemble learning based approach was proposed by [14] using real-life data for impression fraud detection in mobile advertising. The dataset required feature extraction and generation since most features had different distributions.The study applied SMOTE to balance classes in the dataset. The proposed approach achieved an accuracy of 99.32%, with 96.29% precision and 84.75% recall. Despite good accuracy of the model, the recall was slightly lower.

A study for credit card fraud by [15] used Sliding-Window Method to extract some features that will assist in determining behavioural patterns. SMOTE dataset provides better results in the experiments compared to imbalanced dataset. The study also provided an alternative for handling class imbalance by use of one-class classifiers, using OCSVM. MCC (Matthews Correlation Coefficient) metric evaluated performance of the model. RF performed better than Logistic Regression, Decision Tree, and Local Outlier Factor.

An approach for credit card fraud detection proposed by [16], increases classification accuracy by performing feature engineering to create new attributes from existing features in the dataset. The data remained imbalanced in the study. XGBoost classifier performed well compared to other classifiers such as RF, LR and DT. This good performance of XGBoost, according to the research, was attributed to use of boosting method in ensemble learning technique. XGBoost achieved a higher accuracy, though it took more training and evaluation time, than Decision Tree and Naive Bayes, in a case study for fraud detection in automobile's body insurance [17].

Research by [18] observed that Decision Tree had a better efficiency than Naive Bayes and Support Vector Machine. It had an accuracy of 92.5% while Naive Bayes had 90.28%, and Support Vector Machine 30.28%. The dataset used had few instances and was highly imbalanced with only 360 damage file instances where 91 were fraudulent and 269 non-fraudulent. Class imbalance was not addressed as well as feature engineering which would improve the performance.

A research by [19] for detecting automobile insurance fraud used DT C4.5. The study achieved an accuracy of 93.6% and a specificity of 93.5% implying some false positives. Some valid cases were classified as fraudulent. Gradient boosted Decision Tree improved performance in fraud detection for medicare [20].

Pipelining and ensemble learning was used [21] in credit card fraud detection. The research applied comparative investigation with different classifiers including LR,NB, KNN, MLP, and RF. ADASYN method was used to handle class imbalance. The Ensemble Learning and Pipelining significantly performed better than other techniques with accuracy of 99.99% and 99.999% respectively. Random Forest was closer to them with an accuracy of 99.7%. In pipelining approach,a series of transformation started, followed by RF as the classifier; this improved the accuracy. In Ensemble Learning, bagging classification was applied with RF as the base classifier as well. To validate the good performance, the approach could be used in motor insurance claims fraud detection.

A study by [22] proposed an improved technique for detecting credit card fraud. The research used Support Vector Machine and Random Forest and a feature selection algorithm to identify anomalous transactions.The dataset is highly imbalanced where only 492 fraudulent transactions out of 284,807 transactions (0.17%). Classification of transactions as either legitimate or fraudulent was done by SVM. The research observed that SVM based on Random Forest Classifier has a good accuracy of 95%.

The models proposed to detect fraud in the related work register remarkable performance. However, there are some false positive errors where valid claims are classified as fraudulent. Complexity of data, class imbalance, and behavioral analysis for feature selection are some of the problems that affect the performance of the classifiers.

III. Methodology

In this section, we discuss our adopted methodology to formulate the problem, collect and prepare data, build the model and evaluate it for classifying a motor claim as either fraudulent or non-fraudulent. The proposed methodology is presented in Figure 1.

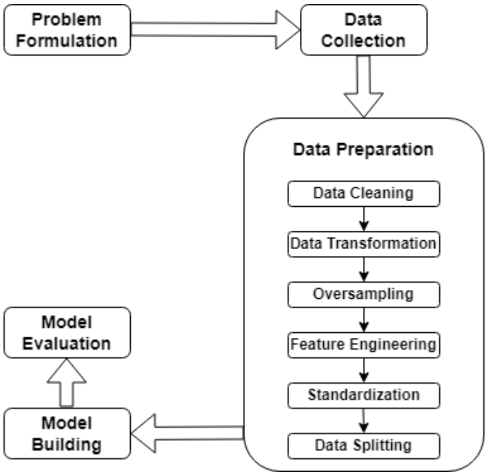


Fig. 1. Proposed Methodology

A. Problem Formulation

The fraud detection methods proposed typically reveal unusual activities in previous claim transactions. Due to dynamic tactics and evolution of fraudsters, these techniques become infeasible in detecting fraud. The historical data used is complex in determining the relevant features while instances of fraud claims are very few.

We formulate this problem as a detection function *f()* that maps X to Y, where $X = \{x_1, x_2, x_3, ..., x_M\}$ is a set of M input features for an insurance claim such as the date of loss, class policy, and type of loss, and $Y = \{0, 1\}$ is a binary output variable indicating whether the claim is fraudulent (1) or not (0). The function will be $Y = f(X, \epsilon)$, where ϵ is an error term representing the effect of accurately predicting the possibility of fraud. In most cases, not the complete set X determines the output Y. The problem is to determine a subset X_s of only relevant features, $X_s = \{x_1, x_2, x_3, ..., x_m\}$ where $m < M$.

The class imbalance problem occurs when a model predicts the majority class with high accuracy but poor performance on the minority class [23]. Such a model used a dataset with N instances and $N_{pos} << N_{neg}$ where N_{pos} is the number of instances in the positive(1) class and N_{neg} is the number of instances in the negative(0) class. The goal is to improve the function $Y = f(X_s, \epsilon)$ to maximize performance on both classes and reduce false positives by using a loss function $L(f(X_s), Y)$ that aims to reduce error $\epsilon = |f(X_s) - Y|$ where $f(X_s)$ is the predicted class and Y is the actual class.

B. Data Collection

This involves obtaining data that is relevant for the study, assessing its quality and having basic understanding of the data for model training. Two datasets were considered for this research. These datasets have several features and are labelled with a target variable of whether a claim was fraudulent or not. Based on the target variable, the datasets are highly imbalanced. These makes them relevant to the study.

1) Dataset 1: This is an online dataset, obtained from Kaggle [24]. The dataset was published by Oracle and had been collected by Angoss Knowledge Seeker software from January 1994 to December 1996 and stored in CSV format. It contains 32 predictor variables and a target variable, FraudFound with values 1 when claim is fraudulent, and 0 when the claim is valid. The features include: Insured details (age, marital status, gender, etc), Vehicle details (make, age, price, etc), Accident details (day, area, police report filed, witness present, no. of vehicles involved, etc), and Policy details (type, number, year, agency, etc).

2) Dataset 2: This is a dataset, obtained from an insurance company in Kenya. The dataset was anonymized by the company and it included all claims reported in the year 2022. The dataset has 15 predictor variables including claim number, policy class and subclass (private, commercial, PSV, etc), policy number, loss date, report date, loss description, and LOP amount. FraudFound is labelled either 'Claim paid' or 'Claim repudiated due to fraud'.

The Table I below summarises the two datasets.

TABLE I
Summary of Datasets used

No.	No. of Claims	No. of Features	0 (No Fraud)	1 (Fraud)
1	15420	32	14496 (94%)	923 (6%)
2	10856	15	10286 (95%)	570 (5%)

C. Data Preparation

To ensure that the model developed from this research produces accurate and insightful results, this step is crucial. It involves data cleaning, transformation, handling null values, removal of duplicate and irrelevant data, leaving only the bits that improves the data quality for efficient and effective classification.

1) Data Cleaning: In this stage, duplicate records and missing values are checked. For both datasets, there are no missing values or duplicate records.

2) Data Transformation: This involves placing data in formats interpretable by the machine learning classifiers. Textual values (strings) were converted into integer values. Using One Hot Encoding method, categorical data is formatted into integral data. The transformed data is then integrated with the columns that initially had numerical values.

3) Handling Class Imbalance: The first dataset has 14,496 (94%) Non Fraud instances dominating Fraud instances which are 923 (6%) the second dataset have 95% and 5% respectively. With the presence of such unbalanced distribution, the classifier will tend to be biased towards the majority class. The model will learn more about majority class and fail to learn about minority class.

Synthetic Minority Over-sampling Technique (SMOTE) developed by [25] was applied. It is one of the most widely used and effective oversampling technique [26]. SMOTE rebalances data by creating synthetic instances of minority class by interpolating between nearest features without adding new information to the data.
Since our interest was on fraudulent claims (minority class), we handled the class imbalance by oversampling. The observations from minority class are replicated to balance the ratio between minority and majority sample. There is no loss of important information from the samples unlike undersampling technique.

4) Feature Engineering: When data has high dimensionality and heterogeneity, there is need to perform feature selection [27]. The aim is to extract features from dataset that are only relevant and contains rich details of fraudulent and non-fraudulent claims. In this research, feature selection was done by using filter method, where we used correlation to check how features are relating to the output (target variable). Using Chi-Squared Test we measured each feature's score in relation to target variable. The method ranked the features with respect to their scores. Those features with low scores were dropped [28].

5) Standardization: In this phase, the features are rescaled to standard normal distribution with a mean of 0 and a standard deviation of 1 [29]. This is important since the values in features have different units and scales. It assists in avoiding higher weightage to features with higher magnitude. Standardization is achieved by computing z-score as below:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where z is the z-score, μ is mean and σ is standard deviation.

6) Data Splitting: By convention, the data was split into two parts: 80% training set and 20% test set for both datasets.

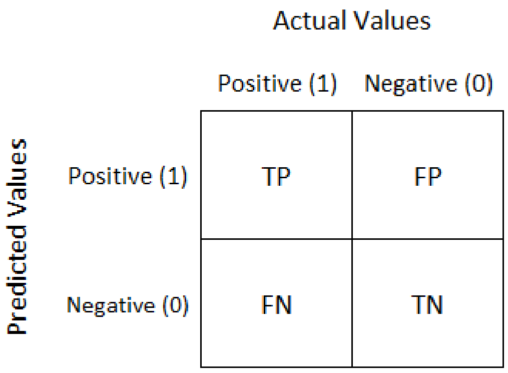
D. Model Building

In this phase, we used an experimental approach to develop the models. The classifiers used are LightGBM, Random Forest, Logistic Regression, KNN, and XGBoost. Each classifier is trained separately by the two datasets before balancing. The building is done in Jupyter Notebook using Python libraries such as sklearn.linear_model, sklearn.ensemble, xgboost, and lightgbm. After training the classifiers, their testing accuracy are evaluated using sklearn.metrics. XGBoost classifier is observed to have a slightly better accuracy than the other classifiers and further experiment is conducted with the classifier, incorporating SMOTE for class balancing.

E. Model Evaluation

In this phase, analysis is done to determine the appropriateness of the models in detecting fraudulent motor insurance claims. Various metrics were used for assessing the models. The metrics include: Confusion Matrix, AUC-ROC, Precision, Recall, F1-Score, and Accuracy.

1) Confusion Matrix: This is a common metric in predictive analysis due to its understandability [30] and also is used in computing other metrics. It is composed of statistics: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) which are calculated using the combination of actual and predicted values.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 2. Confusion Matrix

True Positive (TP) is a case where the actual value was positive (e.g., fraud) and the predicted value is also positive.
False Positive (FP) is a case where the actual value was negative (e.g., non_fraud) but the predicted value is positive.
True Negative (TN) is a case where the actual value was negative (e.g., non_fraud) and the predicted value is also negative.
False Negative (FN) is a case where the actual value was positive (e.g., fraud) but the predicted value is negative.

2) AUC-ROC: Area Under Curve Receiver Operating Characteristic curve tells how good a model performs when used at different probability thresholds. By default, the threshold is usually 0.5. It is a plot between True Positive Rate, TPR (also Sensitivity), and False Positive Rate, FPR (computed as 1-Specificity).
$Sensitivity = \frac{TP}{TP+FN}$
$Specificity = \frac{TN}{TN+FP}$

3) Precision: Precision is the percentage of correctly classified claims in relation to the total number of classified claims.
$Precision = \frac{TP}{TP+FP}$

4) Recall: Recall is the percentage of correctly classified claims out of all classified claims. $Recall = \frac{TP}{TP+FN}$

5) F1 Score: Also known as F-Measure is the harmonic mean of precision and recall. It ranges from 0 to 1 with 0 being worst and 1 considered best.
$F1 = \frac{2*(Precision*Recall)}{Precision+Recall}$

6) Accuracy: This is the total number of correct predictions made, divided by the total number of all predictions.
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

IV. Results and Discussion

In this section, we evaluate the performance of our models to detect fraud in motor insurance claims.

In the first experiment, the preprocessed data for both datasets was used to train the five classifiers. The experiment was done with unbalanced datasets. Table II show the testing accuracy of the classifiers.

TABLE II
RESULTS: TESTING ACCURACY

Model	Dataset 1	Dataset 2
LR	94.01	94.43
RF	94.44	94.24
KNN	94.30	94.43
LightGBM	94.77	91.77
XGBoost	95.17	94.86

From the results, there is no specific technique that would perform extremely better than other techniques in both datasets. Also, in all techniques, there is no dataset that is giving better results than the others. However, XGBoost is observed to have slightly higher performance among the classifiers. In view of this observation, XGBoost classifier was used for further experiments that incorporated SMOTE. One experiment was done before oversampling, and the other after oversampling.
The results obtained before oversampling are as below:

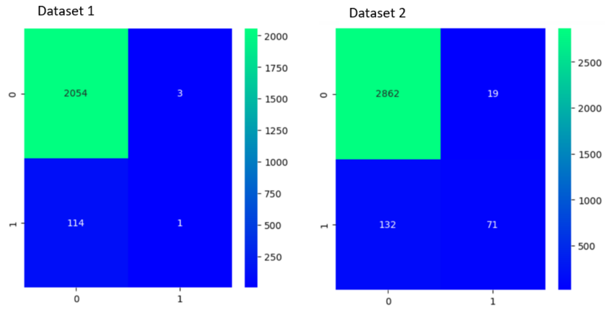
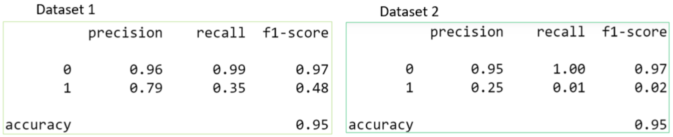


Fig. 3. Confusion Matrix before Oversampling



```
Dataset 1
              precision    recall  f1-score

           0       0.96      0.99      0.97
           1       0.79      0.35      0.48

accuracy                               0.95
```

```
Dataset 2
              precision    recall  f1-score

           0       0.95      1.00      0.97
           1       0.25      0.01      0.02

accuracy                               0.95
```

Fig. 4. Classification Report before Oversampling

The overall accuracy for both datasets is at 95%. However, from the confusion matrix in Figure 3, the True Negatives are many while the True Positives are few. In both results, in Figure 4, the model is performing well in negative class 0 with precision, recall and F1-score being over 95%. The models performs poorly in positive class 1 (fraudulent claims) with as low as 1% in recall for Dataset 2. The model is able to correctly identify non fraudulent claims and unable to identify fraudulent claims due to imbalance in the dataset. AUC-ROC curve for Dataset 2 is at 50% as seen in Figure 5.

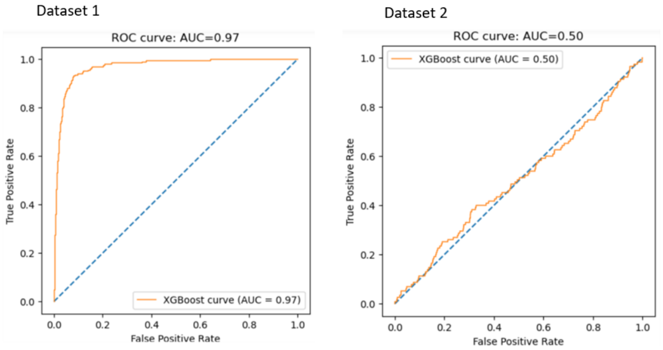


Fig. 5. AUC-ROC before Oversampling

SMOTE was applied on both datasets to balance the classes and the experiment repeated with the resampled data. The results obtained after oversampling are as below:

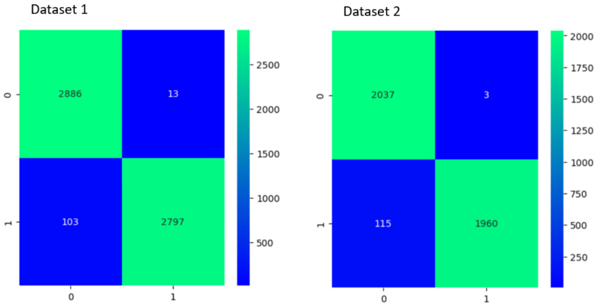
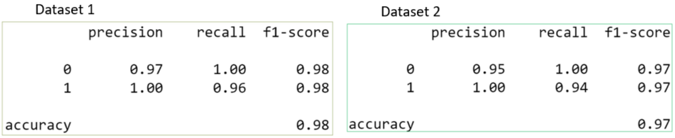


Fig. 6. Confusion Matrix After Oversampling



```
Dataset 1
              precision    recall  f1-score

           0       0.97      1.00      0.98
           1       1.00      0.96      0.98

accuracy                               0.98
```

```
Dataset 2
              precision    recall  f1-score

           0       0.95      1.00      0.97
           1       1.00      0.94      0.97

accuracy                               0.97
```

Fig. 7. Classification Report After Oversampling

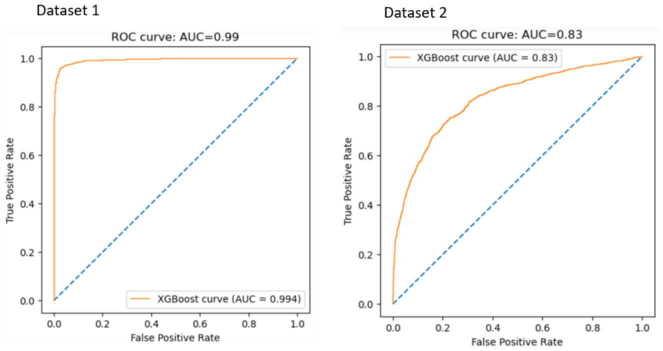


Fig. 8. AUC-ROC After Oversampling

From the confusion matrix in figure 6, the True Positive instances have increased. The model is now able to detect positive claims after oversampling. The number of false negatives and false positives have reduced slightly.The overall accuracy increases from 95% in Dataset 1 and Dataset 2 to 98% and 97% respectively. Performance of class 1 also improved to as high as recall of 94% in Dataset 2. AUC-ROC curve for Dataset 1 increased from 97% to 99% while that of Dataset 2 increased from 50% to 83%. The oversampling did not affect performance of class 0.

V. Conclusion and Future Work

In our study, we investigated fraud detection in motor insurance using XGBoost algorithm with SMOTE. The performance of the proposed model is compared to state-of-the-art solutions, and XGBoost without SMOTE. The algorithms are evaluated for different metrics. The results show that XGBoost classifier, combined with SMOTE to handle class imbalance, has better performance in detecting fraudulent claims. For future work, we will experiment with different insurance datasets and other insurance-related prediction problems to validate XGBoost's performance and ensure there is no false positive or false negative. By using it as a base classifier, a series of transformations can be performed on the training process to improve model's performance. When the model is deployed in the insurance industry, they will be able to identify perpetrators of fraud and take necessary action. The customer satisfaction rate will be increased and loss adjustment expenses reduced. The increase in customer satisfaction will reduce instances where customers register fraudulent claims. Reduced loss adjustment increases the profit made by insurance companies and, reduces the premium rating thus making insurance policies affordable.

References

[1] N. Remli, F. Salleh, and J. Arifin, "Motor insurance fraudulent claims: An overview reconnaissance," *International Journal of Business, Economics and Law*, vol. 25, 2021.

[2] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection," in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*. IEEE, 2019, pp. 1–4.

[3] G. Kowshalya and M. Nandhini, "Predicting fraudulent claims in automobile insurance," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018, pp. 1338–1343.

[4] M. A. Rashid, A. Al-Mamun, H. Roudaki, and Q. R. Yasser, "An overview of corporate fraud and its prevention approach," *Australasian Accounting Business & Finance Journal*, vol. 16, no. 1, pp. 101–118, 2022.

[5] U. Rani, O. L. Pramudyastuti, and A. P. Nugraheni, "Disclosing the practice of whistleblowing system in indonesiaâ€™ s public listed companies," *INOVASI*, vol. 18, pp. 79–87, 2022.

[6] M. H. AYBOGA and F. Ganji, "Detecting fraud in insurance companies and solutions to fight it using coverage data in the covid 19 pandemic," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 15, pp. 392–407, 2021.

[7] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, p. 100402, 2021.

[8] M. Guillen, J. P. Nielsen, and A. M. Pérez-Marín, "Near-miss telematics in motor insurance," *Journal of Risk and Insurance*, vol. 88, no. 3, pp. 569–589, 2021.

[9] A. I. Alrais, "Fraudulent insurance claims detection using machine learning," 2022.

[10] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements using Machine Learning and Data Mining: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 72 504–72 525, 2021.

[11] T. Badriyah, L. Rahmaniah, and I. Syarif, "Nearest neighbour and statistics method based for detecting fraud in auto insurance," in *2018 International Conference on Applied Engineering (ICAE)*. IEEE, 2018, pp. 1–5.

[12] A. Urunkar, A. Khot, R. Bhat, and N. Mudegol, "Fraud detection and analysis for insurance claim using machine learning," in *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, vol. 1. IEEE, 2022, pp. 406–411.

[13] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods," *Knowledge-Based Systems*, vol. 128, pp. 139–152, 2017.

[14] C. M. R. Haider, A. Iqbal, A. H. Rahman, and M. S. Rahman, "An ensemble learning based approach for impression fraud detection in mobile advertising," *Journal of Network and Computer Applications*, vol. 112, pp. 126–141, 2018.

[15] V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms," *Procedia computer science*, vol. 165, pp. 631–641, 2019.

[16] D. X. Cho, D. N. Phong, and N. Duy Phuong, "A new approach for detecting credit card fraud transaction," *International Journal of Nonlinear Analysis and Applications*, 2023.

[17] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "Extreme gradient boosting machine learning algorithm for safe auto insurance operations," in *2019 IEEE international conference on vehicular electronics and safety (ICVES)*. IEEE, 2019, pp. 1–5.

[18] L. Goleiji and M. J. Tarokh, "Fraud detection in the insurance using decision tree, naive bayesian and support vector machine data mining algorithms (case study-automobile's body insurance)," 2016.

[19] I. M. N. Prasasti, A. Dhini, and E. Laoh, "Automobile insurance fraud detection using supervised classifiers," in *2020 International Workshop on Big Data and Information Security (IWBIS)*. IEEE, 2020, pp. 47–52.

[20] J. T. Hancock and T. M. Khoshgoftaar, "Gradient boosted decision tree algorithms for medicare fraud detection," *SN Computer Science*, vol. 2, no. 4, p. 268, 2021.

[21] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, "Credit card fraud detection using pipeling and ensemble learning," *Procedia Computer Science*, vol. 173, pp. 104–112, 2020.

[22] N. Rtayli and N. Enneya, "Selection features and support vector machine for credit card risk identification," *Procedia Manufacturing*, vol. 46, pp. 941–948, 2020.

[23] E. Alogogianni and M. Virvou, "Handling class imbalance and class overlap in machine learning applications for undeclared work prediction," *Electronics*, vol. 12, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/4/913

[24] S. Bansal, "Vehicle insurance claim fraud detection — kaggle," 2021. [Online]. Available: https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection

[25] 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.

[26] H. Ahmad, B. Kasasbeh, B. Aldabaybah, and E. Rawashdeh, "Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (sbs)," *International Journal of Information Technology*, vol. 15, no. 1, pp. 325–333, 2023.

[27] Y.-Y. Hsin, T.-S. Dai, Y.-W. Ti, M.-C. Huang, T.-H. Chiang, and L.-C. Liu, "Feature engineering and resampling strategies for fund transfer fraud with limited transaction data and a time-inhomogeneous modi operandi," *IEEE Access*, vol. 10, pp. 86 101–86 116, 2022.

[28] T. Parlar and G. Cinarer, "Feature selection methods for intrusion detection using machine learning methods," *Selcuk University Journal of Engineering Sciences*, vol. 21, no. 2, pp. 63–68, 2022.

[29] J. Ma, Z. Bo, Z. Zhao, J. Yang, Y. Yang, H. Li, Y. Yang, J. Wang, Q. Su, J. Wang *et al.*, "Machine learning to predict the response to lenvatinib combined with transarterial chemoembolization for unresectable hepatocellular carcinoma," *Cancers*, vol. 15, no. 3, p. 625, 2023.

[30] E. Bayhan, A. G. Yavuz, M. A. Güvensan, and M. E. Karslıgıl, "The effect of feature selection on credit card fraud detection success," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021, pp. 1–4.