2019 4th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, UP, India. Nov 21-22, 2019

Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique

Sonakshi Harjai
Amity Institute of Information Technology
Amity University
Noida, India
sonakshi.harjai@student.amity.edu
harjaisonakshi@gmail.com

Sunil Kumar Khatri
Amity Institute of Information Technology
Amity University
Noida, India
skkhatri@amity.edu
sunilkkhatri@gmail.com

Gurinder Singh
Amity International Business School
Amity University
Noida, India
gsingh@amity.edu

***Abstract* –There has been a significant amount of growth in the number of fraudulent activities by the policy-holders over the last couple of years. Deliberately deceiving the insurance providers by omitting facts and hiding details while claiming for insurance has led to significant loss of money and customer value. To keeps these risks under control; a proper framework is required for judiciously monitoring insurance fraud. In this paper, we demonstrate a novel approach for building a machine- learning based auto-insurance fraud detector which will predict fraudulent insurance claims from the dataset of over 15,420 car-claim records. The proposed model is built using synthetic minority oversampling technique (SMOTE) which removes the class imbalance-ness of the dataset. We use random forests classification method to classify the claim records. The data used in our experiment is taken from a publically available auto insurance datasets. The outcomes of our approach were compared with other existing models based on various performance metrics.**

Keywords: Insurance fraud, Class Imbalance, Auto Insurance claims, Oversampling, Classification, Random Forests, Performance metrics

I. INTRODUCTION

The insurance industry has been facing numerous challenges due to fraud claims from the very beginning. Losses incurred due to frauds, impacts all the parties involved. Even one undetected fraud could lead to a huge loss [3]; resulting in increased premium-costs, process inefficiency and loss of trust. Though all insurance companies have their fraud- detection systems in place, still most of those processes are very inefficient and time consuming. Traditional mechanisms rely heavily on human intervention and hence are not adaptable to any changes or situation, if required. A long on-going investigation results in delay in pay-outs and has a negative impact on the customer. Uncaught fraudulent claims not only hinder the profitability of the firm but also encourage other policy holders to show similar behavior. Insurance fraud occurs when individuals attempt to profit by failing to fulfill the terms of the insurance agreement [1]. Frauds can be categorized under soft-fraud or hard-fraud [4]. If a policy holder intentionally plans an accident or invents a loss just to gain benefits from the insurance company then it is said to be a hard fraud. However, when an actual injury or theft occurs, and the insured exaggerates the claim to obtain more money from the company, then that is termed as a soft fraud.

The evolution of big-data and the growth of unstructured data has given rise to a lot of fraudsters exploiting the system. If the data is not analyzed thoroughly, there will be huge chances of occurrence of a fraud [3].

Data mining and analytics has changed the fraud detection scenario [2]. Data can be gathered from various sources and can be stored in a combined repository for further use. Implementing analytical solutions costs an initial investment to the insurance companies; thus they always resist implementing it. However, it has been observed that using machine learning and analytical capabilities have strengthened the insurance lifecycle in many forms.

It has been able to provide a lot of cost benefits to the companies by saving up a lot of money, by reducing the overall cost of fraud detection and improving the overall-ROI of fraud detection.

So, the insurers need to start leveraging their machine-learning capability in order to build more robust and risk-free systems. Hence, there is a crucial need to develop a system that can help the insurance industry to identify potential frauds with a high degree of accuracy, so that other claims can be cleared rapidly while the identified cases can be examined in detail.

The dataset used in this study is found to have a class imbalance problem; means that the number of instances of one class(positive) far exceeds the number of instances of other class(negative).The class having far less number of instances than the other becomes the minority class; other class being called as majority class[13]. Due to which, the minority class tends to be ignored during the classification process. To avoid minority data-instances to be treated as a noise and the classifier to be biased with the majority one, this data- imbalance needs to be fixed. A simple way to fix this imbalance is by balancing the data set, either by over-sampling instances of the minority class or under- sampling instances of the majority class.

This research paper aims to develop a model to help the insurers take pro-active decisions and make them better equipped to combat fraud. In this paper, we propose a procedure for auto-fraud identification using Random-Forests Classification Technique, before which we remove the class imbalance-ness of our original dataset. This is done using synthetic minority oversampling technique (SMOTE).

978-1-7281-3651-6/19/$31.00 ©2019 IEEE

The paper is organized as following. Section II, gives a brief overview of the past research done in this area, along with the dataset and software used. Section 3 describes the proposed methodology. Section 4 includes the results and discussions. Section 5 contains the comparative analysis of our approach with other models. Section 6 holds conclusions after which list of references are given.

II. RELATED WORK

There have been various advancements in the area of fraud detection since the last decade. Some of the similar works include:

Subudhi et al. proposed a model for Auto Insurance Fraud Detection System (AIFDS) using adaptive oversampling technique (ADASYN) to study the effect of class- imbalanceness (skewness) on a dataset [5]. It has been proved that using machine-learning techniques can improve the accuracy of fraud detection in imbalanced samples.

Numerous ensemble methods have also been performed on insurance data-sets to detect fraudulent cases. In [11], Neural-Network Ensemble was used along with Random Rough Subspace method to perform Insurance Fraud Detection. Viaene et al. [6] applied a Boosting-Naïve-Bayesian Classifier for fraud diagnosis. This model collaborated the advantages of boosting and the power of a weight-scoring structure. Omar et al. gave a cost- effective approach for an AIFDS design to segregate fraudulent records from legitimate ones [7].

A survey was done of over 80 research papers and journals to review all the machine-learning approaches and advancements that were made over the last decade. The research showcased that Artificial Neural-Networks, SVM, Naïve-Bayes, Random Forests and k-NN were the most used classifiers for detecting automobile insurance fraud [8].In [9], Kowshalya et al. aims to detect suspicious claims and attempts to reduce the monetary losses occurred due to fraud policy claimers. The testing phase of this study, depicts that Random Forests outperforms J48 and Naïve Bayes classifier accuracies and proves to be a suitable choice for Fraud Detection. The finding of these studies was a motivation for the research.

A. *Software Used*

Weka Data Mining Software: Weka is free and open source software used as a data mining tool for machine learning and knowledge discovery [10]. Weka is used by various researchers and can be important to develop new machine learning schemes from different datasets. Datasets in Weka are loaded in .arff or .csv format. Being Open Source, we use this tool to experiment our proposed model under GNU General Public License.

B. *Dataset Description*

The dataset used that is used for the course of our research is “carclaims.txt” automobile insurance dataset that is provided by “Angoss Knowledge Seeker” software [12]. The data was recorded in year 1994-1996 in United States.

This dataset contains 15,420 auto-insurance claim records in total out of which 11,338 records were gathered from Jan’94 till Dec’1995 and remaining 4083 instances were recorded in the year 1996. Each record in the dataset has 33 attributes in total that were submitted while filing claims. Out of these 33 features, 32 are claim features that will help to predict the last 1 variable, called the class label.

FraudFound: - is our 'target variable' and it represents the presence and absence of a fraud claim.

- Fraudulent Claim is equivalent to "1" and
- Non-Fraudulent is equivalent to "0"

The complete statistical description of the dataset is given in Table I.

TABLE I: CAR INSURANCE DATASET [12]

DATA STATISTICS	
Number of Claim Records	15,420
Number of Attributes	33
Categorical (Nominal) Attributes	25
Numerical Attributes	6
# of Normal Claims (Non-fraudulent)	14,497 (94%)
# of Fraudulent Claims	923 (6%)
Number of Years of Data	3
Average Claims (per month)	430

The complete feature set (or class labels) of the dataset containing 33 attributes is described in Fig 1.

This dataset is an imbalanced class-distribution as it has approximately 94% of non-fraud claims and approximately 6% of fraud claims.

Fig. 1. Attribute description of dataset [12]

III. PROPOSED METHODOLODY

We propose a model that aims to facilitate better decision- making of the insurers while making claim related decisions.

The proposed approach will work with any real-time data in spite of its class-distribution skewness as we transform the original dataset into a balanced dataset before performing any kind of classification algorithm on the variable.

In this paper, Synthetic Minority Oversampling Technique (SMOTE) is applied to re-balance the data followed by Random Forests classifier to classify the records into malicious or genuine claims.

The proposed methodology has been described in the following steps.

Step 1: Data Pre-processing

a) Data Cleaning

- After uploading the dataset, the data was checked for any missing values, redundant data, duplicates or any noise present.
- The original carclaims.txt dataset had no missing values into it.

b) Data Transformation

- The claims record sheet contained 4 columns for – Year, Month, Week of Month and Day of week respectively.
- We converted those values manually into the normal date-time format as YYYY-MM-DD to ease our further calculations.

c) Data Visualization

- The data was thoroughly analyzed by plotting various graphs and the features were grouped according to their categories to gain more insights of it.
- This was done to find out the dependencies between various features of the insurance dataset.
- We plotted graphs of the following categories to study their correlation:

  - ❖ Delay between AccidentDate vs DateOfClaim,
  - ❖ Age of PolicyHolders vs FraudFound &
  - ❖ Fault of the policy holder or third party v/s FraudFound

d) Data Resampling or Data Balancing (using SMOTE)

- Oversampling was done on classValue=1(i.e. the minority class) using SMOTE filter, keeping classValue=0 unchanged.

Step 2: Data Classification (using Random Forests)

- Applying Random Forests classification algorithm with a batch size of 100 and seedValue=1.
- Applied 10-folds cross validation methodology to implement the model.

Step 3: Training & Testing the Model

- Run the model on the train-test set which is in 80-20 ratio of the dataset.

Step 4: Model Validation

- Validate the results generated by the Confusion Matrix.

The architecture of the proposed model is illustrated in Fig 2.

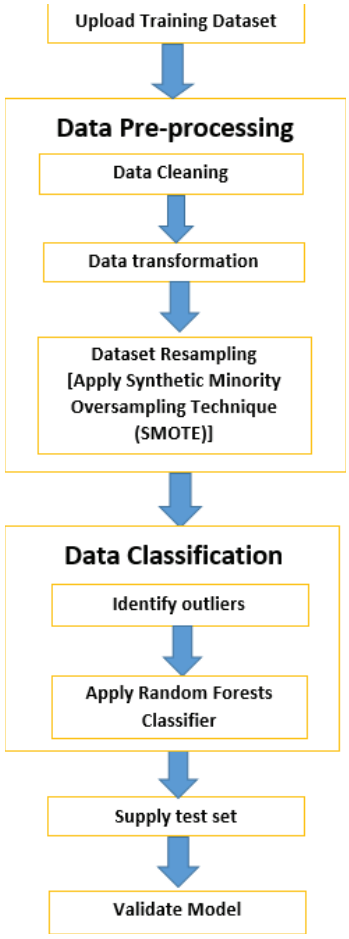

Fig. 2. Proposed Architecture for Auto-Insurance Fraud Detection System

A. Removing Class-Imbalance in Dataset using Synthetic Minority Oversampling Technique (SMOTE)

There are various data-balancing techniques that are being used to overcome the Class-Imbalance problem; broadly divided into- Over-sampling and Under- Sampling. (Fig 3)

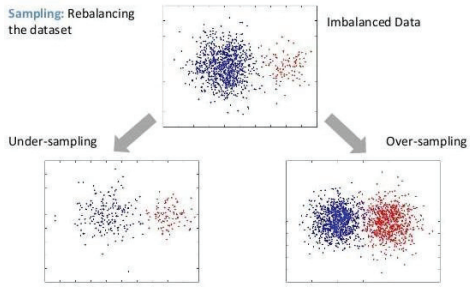

Fig. 3. Handling Imbalanced Class Distribution

Synthetic Minority Oversampling Technique (SMOTE) is used to resample the dataset in our proposed approach. SMOTE is a well-known technique of oversampling, it modifies an imbalanced dataset to generate a balanced one. It distributes the majority class and the minority class instances equally. SMOTE creates synthetic instances or similar samples of the minority-class and aims to reduce the biasness of the classifier's prediction towards the majority class [6].

The instances are synthesized according to the k-nearest neighbor algorithm (Fig 4) , where any neighbor from k nearest neighbors are chosen at random and is inserted along the line segment(s) joining any or all of the k-minority class nearest- neighbors. [15] (Fig 5.) The number of new instances generated depends on the amount of oversampling required.

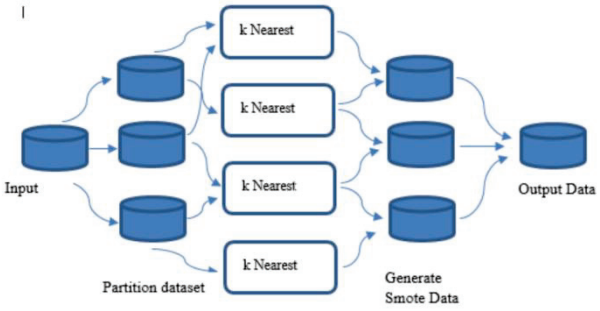

Fig. 4. SMOTE Approach

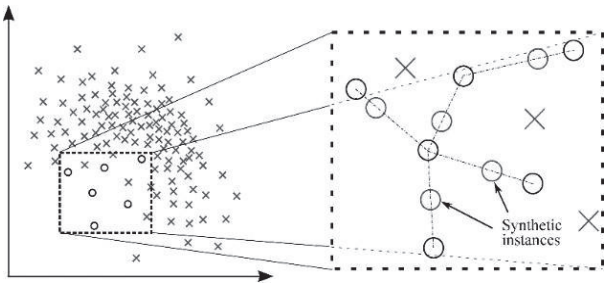

Fig. 5. Generating synthetic instances

O is the original data set
P is the set of positive instances (minority class instances)
For each instance *x* in *P*
    Find the *k*-nearest neighbors (minority class instances) to *x* in *P*
    Obtain *y* by randomizing one from *k* instances
    difference = *x* – *y*
    gap = random number between 0 and 1
    n = *x* + *difference* * *gap*
    Add *n* to *O*
End for

Fig. 6. SMOTE Re-sampling Algorithm

A. Random Decision Forests Classifier

Random Forests (RF), also termed as Random Decision Forests is a tree-based supervises machine learning algorithm that can be used for both classification and regression problems. RF creates Decision-Trees based on randomly selected data samples, gets prediction_value from each tree and picks out the best solution by the process of voting (Fig 7.). It is a really good indicator for obtaining important features from the dataset. Detecting fraudulent activities is one of the widely-used applications of RF technique.

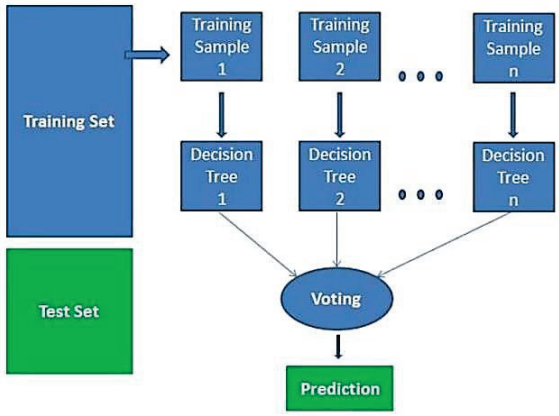

Fig. 7. Random Forests Classifier

IV. RESULTS AND DISCUSSION

A. Evaluation Criteria

Evaluating the success of a model is really crucial to determine how suitable it is for solving particular problem. Even slight improvements in performance can lead to large economic benefits.

Our model will be compared on the basis of its relative performance as well as absolute performance with respect to various performance metrics like-Accuracy, Sensitivity, and Specificity. A list of some commonly used performance metrics is given in Table II.

TABLE II: PERFORMANCE METRICS

Predicted		Actual	
		Normal	Fraud
	Normal	True Negatives (TN)	False Negatives (FN)
	Fraud	False Positives (FP)	True Positives (TP)

While constructing a detection-model, the savings from loss prevention needs to be balanced with the cost of false alerts. Whether an alert will be generated or not that depends on the TN, TP, FP and FN's [14]. Refer to Table III.

TABLE III: PREDICTION AND ALERT GENERATION [14]

Prediction	Fraud	Legal
Alert	Hits	False Alarms
No Alert	Misses	Normal

B. Experimental Results

After oversampling was done on the minority class using SMOTE filter (Fig 8.), the number of instances increased to 16,343 where minority class instances increased to 1846 keeping the number of positive class instances equal to 14,497.

Total number of instances:-

- In imbalanced dataset =15420 (Fig.9)
- In balanced dataset = 16343 (Fig. 10)

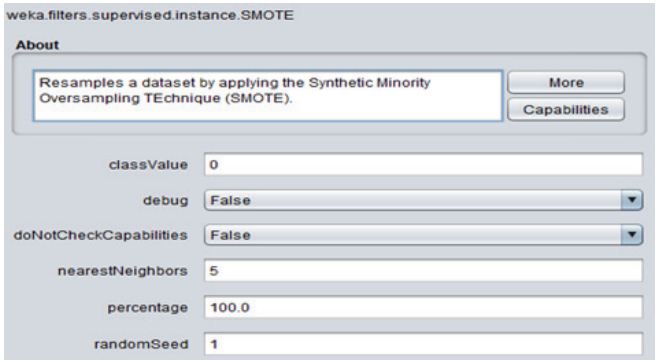

Fig. 8 .SMOTE Filter in Weka

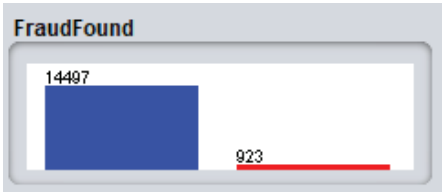

Fig. 9. Class distribution of data before re-sampling with SMOTE in Weka

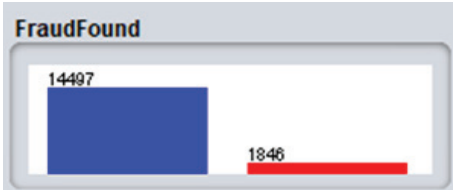

Fig. 10. Class distribution of data after re-sampling with SMOTE in Weka

The exploratory data analysis (EDA) of our dataset gave the following outcomes. Refer to Fig 11, 12, 13 and 14.

- 82% of the cases which turned out to be fraud involved vehicles with 6 to 8 years of age. That proves that old vehicles tend to be involved more in frauds.
- 99.6% fraudulent claims have no witness while in case of non-fraudulent claims 83% of them have witnesses.

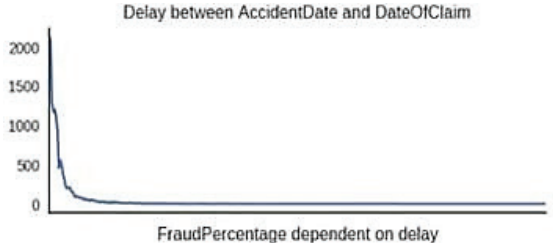

Fig. 11. Delay between AccidentDate & DateOfClaim

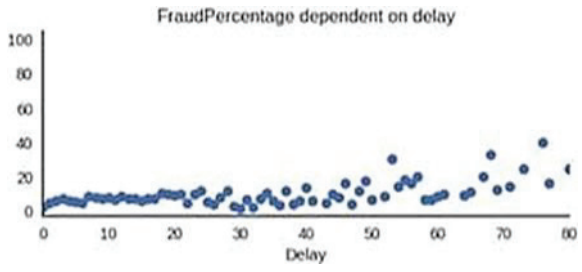

Fig. 12. Fraud percentage and its dependency on Delay

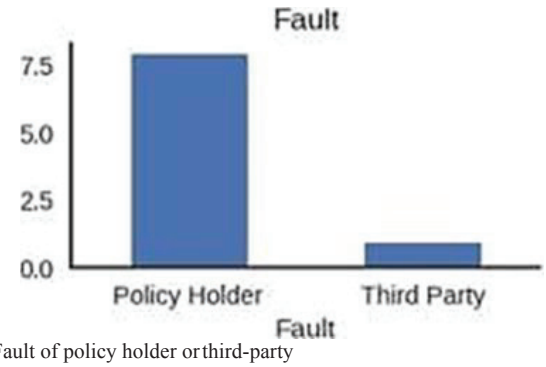

Fig. 13. Fault of policy holder or third-party

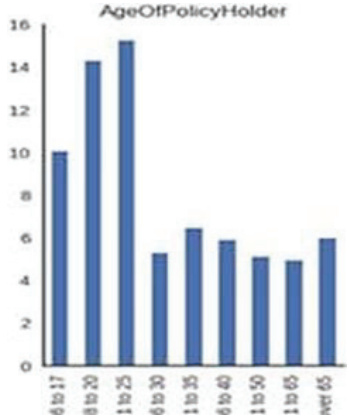

Fig. 14. Age of Policy Holder v/s Percentage of Fraud

We applied the classification technique on the balanced dataset containing 16343 instances. It was found that 15,318 instances were correctly classified which is 93.72% of the whole dataset. The confusion matrix is represented in Fig 15.

```
=== Confusion Matrix ===

    a     b   <-- classified as
14485    12 |     a = No
 1013   833 |     b = Yes
```

Fig. 15. Confusion Matrix for the proposed model

The detailed accuracy of the proposed model is given in Table [IV] with the corresponding weighted averages. The proposed approach gave us 99.9% accuracy & recall value, which proved to be better than various existing fraud detection systems.

TABLE IV: PERFORMANCE METRICS OF PROPOSED APPROACH

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
No	0.999	0.549	0.935	0.999	0.966	0.933
Yes	0.451	0.001	0.986	0.451	0.619	0.933
Weighted Avg.	0.937	0.487	0.940	0.937	0.927	0.933

V. COMPARATIVE ANALYSIS

We carry out a detailed relative and comparative analysis between our proposed model and the other existing approaches – the SMOTE-SVM model, Decision tree and MLP classifiers. Refer to Table V for the comparative analysis of results.

It is observed that the proposed method provides us better efficiency than the other models [5] in terms of sensitivity as

well as other factors. Our model out-performed other models as in [5].

TABLE V: COMPARISON OF VARIOUS MODELS WITH THE PROPOSED APPROACH [5]

Performance Metrics (in %)	Support Vector Machine (SVM)	Decision Tree	Multi-layer Perceptron (MLP)	Proposed Mo
Accuracy	58.41	57.39	74.98	94.33
Sensitivity (or Recall value)	90.53	86.94	47.83	**99.9**
Specificity	36.86	38.14	18.75	45.1

VI. CONCLUSION

Evaluation of a Fraud Detection Model is based on numerous factors. We found that, algorithm selection and performance metrics performs a key role in Model Evaluation and affect the expected results of the proposed methods. Data balancing has been an advantageous method to improve the predictive accuracy of the classifier.

Using SMOTE re-sampling method to balance the class distribution and then applying respective classification filter has given us exceptional results. Our model outperformed three different supervised classifiers, namely, Support Vector Machine (SVM), Decision-Tree and Multi-layer Perceptron (MLP) and gave 99.9 % sensitivity value. It took 1.43 seconds to build the model.

This model can be improved by applying other data-balancing techniques or other classifiers that are not affected by the class imbalance-ness; in future work.

ACKNOWLEDGMENT

Authors of the paper express a hearty sense of gratitude to Dr. Ashok K. Chauhan, Founder President of Amity University for promoting research in Amity University which is a huge opportunity for us to reach great heights.

REFERENCES

[1] James E. Whitaker, “Insurance Fraud Handbook”, Association of Certified Fraud Examiners, Inc., 2018, Available at: https://www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/Insurance-Fraud-Handbook.pdf

[2] Carol Anne Hargreaves and Vidyut Singhania , “Analytics for Insurance Fraud Detection an Empirical Study” : Vol. 1,No.3,2015,pp.227-232 https://www.academia.edu/33431675/Analytics_for_Insurance_Fraud_Detection_An_Empirical_Study.pdf

[3] Ruchi Verma and Sathyan Ramakrishna Mani ,”Using Analytics for Insurance Fraud Detection”,2013,Pg 2-5,7,8, Available at : https://www.the-digital-insurer.com/wp-content/uploads/2013/12/53-insurance-fraud-detection.pdf

[4] Insurance Fraud; Wikipedia, the free encyclopedia; Available at: https://en.wikipedia.org/wiki/Insurance_fraud

[5] Sharmila Subudhi and Suvasini Panigrahi , “Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud” , 2nd International Conference on Data Science and Business Analytics ,2018

[6] S. Viaene, R.A. Derrig, and G. Dedene, “A case study of applying boosting naive Bayes to claim fraud diagnosis,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 5, pp. 612-620, 2004.

[7] L. A. Rodrigues and N. Omar, “Auto claim fraud detection using multi classifier system,” Journal of Computer Science & Information Technology, vol. 14, 2014.

[8] Sinayobye Janvier Omar Makerere, Kiwanuka, Fred Makerere, Kaawaase Kyanda Swaib Makerere, “A State-of- the-Art Review of Machine Learning Techniques for Fraud Detection Research” ,ACM/IEEE Symposium on Software Engineering in Africa, 2018

[9] G.Kowshalya and Dr.M.Nandhini, “Predicting Fraudulent Claims in Automobile Insurance” ,2018

[10] Weka 3.8.1: Data mining software tool [Online Available]: http://www.cs.waikato.ac.nz/ml/weka/

[11] Wei Xu, Shengnan Wang, Dailing Zhang and Bo Yang, “Random Rough Subspace based Neural Network Ensemble for Insurance Fraud Detection” 2011 Fourth International Joint Conference on Computational Sciences and Optimization

[12] Shengnan Wang, Dailing Zhang and Bo Yang, “Auto insurance fraud detection using unsupervised spectral ranking for anomaly”, Volume 2, Issue 1, School of Information Renmin University of China Beijing , March 2016, Pages 58-75 [Available at]: https://doi.org/10.1016/j.jfds.2016.03.001

[13] “Prediction Model Framework for Imbalanced Datasets”, DATA ANALYTICS 2014: The Third International Conference on Data Analytics https://www.thinkmind.org/download.php?articleid=data_analytics_2014_2_20_60051

[14] Clifton Phua, Damminda Alahakoon, and Vincent Lee ,”Minority Report in Fraud Detection: Classification of Skewed Data”, [Available at]: https://sci2s.ugr.es/keel/pdf/specific/articulo/phua.pdf

[15] Raisul Islam Rashu, Naheena Haq, Rashedur M Rahman. "Data mining approaches to predict final grade by overcoming class imbalance problem", 2014 17th International Conference on Computer and Information Technology (ICCIT), 2014

[16] Felix Last, “Oversampling for Imbalanced Learning Based on k-means and SMOTE” , NOVA Information Management School, [Available at]: https://run.unl.pt/bitstream/10362/31042/1/TAA0010.pdf