International Journal of Advanced Research and Review

www.ijarr.in

# INSURANCE CLAIM FRAUD DETECTION

**Purtee Kohli, Sonam, Yangchen Tshomo**

Jaypee Institute of Information Technology Noida

## ABSTRACT

In this paper we are exploring the significance of Fraud claim Detection in insurance project itself is significance and novelty project because of the data set used for detecting fraud is very huge and imbalanced or skewed. And the different types of techniques used to balance the data set like re sampling methods for better result, finding of best algorithms to classify the dataset and detect fraud is a very challenging and the some of the issues faced using these techniques and methods. The imbalance in original data set and most of the claims being non fraud. If we use the original data set as the base for our predictive models and analysis there is a high chances of getting lots of errors and algorithms to over fit since it will assume that most the claims as non fraud. So choosing best techniques, methods and algorithms to get better result is a novelty in this project medical insurance fraud detection.

## INTRODUCTION

Medical Insurance Fraud Detection project is a classification problem , With datasets available in kaggale so we took up the work of finding fraud claim holders or people liable for doing this fraud so the insurance companies can save their money  million of dollars or rupees
Back ground

"Medicare fraud detection using machine learning methods" by Richard A.Bauder and Taghi M.Khoshgoftaar of Florida Atlantic University College of Engineering and Computer Science, USA. In their paper, they looked at different methods of machine learning to detect fraud in the Medicare system. They used different sampling methods with four performance matrices and compared the results and reduced class imbalance through oversampling. This publication includes the calendar years 2012 to 2015, they specifically used all 2015 data over Doctors and other providers that describe the data on the applications for payment and use, as well as information about services and procedures for Medicare beneficiaries. There are ten model in supervised, unsupervised and hybrid is trained and tested. The results show a significant performance gap between the sampling methods. It also discovered that the 80-20 sampling technique is better for learning performance than sampling. The paper concludes that over-sampling leads to poor performance for learners.

“Health care fraud detection with community detection algorithms” by Song Chen and Aryya Gangopadhyay from Information Systems University of Maryland Baltimore explains various types of relationships, such as exclusive relationships, which are suspicious and could indicate potential fraud in the health sector. The relationship between patients and doctors. They developed two algorithms to recognize these small exclusive communities. They are aimed at some kind of fraud. These are communities of suspected vendors sharing or returning patient and

are generally small and have exclusive relationships within communities and no external relationships. The relationships between these communities are suspicious, but they cannot be fully believed that these communities carry out fraudulent activities. Writer says they have to take into account other factors such as incomplete data. These algorithms can be applied to a larger dataset and are highly scalable. They tested these algorithms with a series of synthesized datasets. These synthesized datasets were designed to resemble health insurance claims records and were used to test fraud detection algorithms. The results of the tests show that these algorithms are very efficient and make it possible to evaluate the community structures of 50000 suppliers in about a minute.

The research title "Predicting medical provider specialties to detect anomalous insurance claims" by Richard A.Bauder, Taghi M. Khoshgoftaar and Matthew Herland of Florida Atlantic University.In their work, they developed a machine learning model to detect abnormal behavior of physicians in their health insurance claims. This new study could help determine if and when physicians act outside the standard of their respective discipline, which could indicate abuse, fraud or lack of knowledge about billing procedures. They used a publicly available billing record published by the US Medicare system. Due to the size of the record, the record was sampled to capture all physicians practicing in one state. The model uses the multinomial algorithm of Naïve Bayes and is evaluated by calculation of precision, recall and Fscore with a cross validation of 5. The model is able to successfully predict several classes of physicians with an F-score greater than 0.9. These results show that it is possible to use machine learning in a new way to classify physicians in their respective disciplines only with the procedures they have billed. This article uses machine learning to effectively determine whether the use of procedural data accurately predicts the scope of a physician's job. This study examines the ability to create a machine learning model to assess vendors' fraudulent behavior based on their medical history.

"Investigating the effects of class imbalance in learning the claim authorization process in the brazilian health care market" written by Jackson Cunha Cassimiro from Brazil.He pointed out that fraud and abuse are two factors directly related to the high costs of healthcare, as they are expenditures that can be eliminated without affecting the quality of the services provided. In Brazil, health insurances are introducing a procedure for the approval of claims that can be used to detect fraud and abuse. This process consists of a preliminary analysis of the services requested by the suppliers to identify patterns of fraud and abuse. This analysis is usually done manually, which makes the execution costly and not measurable. Health insurance companies have invested in the use of data mining and machine learning to detect suspicious fraudulent behavior. However, the use of these techniques in the claim authorization process is hampered by the class imbalance problem because there are significantly more authorized service requests than unauthorized requests. This article presents the results of the study on the effects of class imbalances in the area of health insurance entitlement approval. An experiment measured the performance loss of several classifiers in different class distributions and the performance recovery achieved by the processing methods. The results show that the classification algorithms studied are affected differently by the class imbalance. They also show that the recovery performance is even lower when the class imbalance is high.

**PROPOSED WORK**

Classification of fraud and non fraud claims and the most essential thing here is not to classify fraud claims as non fraud claims and non fraud claims as fraud claims.

We have taken Exploratory Data Analysis the train beneficiary, train inpatient, train outpatient, train datasets contains 138556 rows and 25 columns, 40474 rows and 30 columns, 517737 rows and 27 columns and 5410 rows and 2 columns respectively.. The test beneficiary, test inpatient, test outpatient, test data contains 63968 rows and 25 columns, 9551 rows and 30 columns, 125841 rows and 27 columns and 1353 rows and 1 column respectively.

Since the CSV are divided and to perform data analysis, preprocessing and modeling of datasets, the repetition of same features(columns) in CSV and large number of categorical values , NA and IDs contain and imbalance dataset in this project is difficult to detect high accuracy score of fraud claims. So the merging of datasets, feature engineering (extraction), Binning of values, Dropping of NA and IDs Column , Encoding (label) on categorical values and SMOTE (Synthetic Minority Oversampling Technique) technique to balance the imbalance datasets has been performed as preprocessing of datasets.

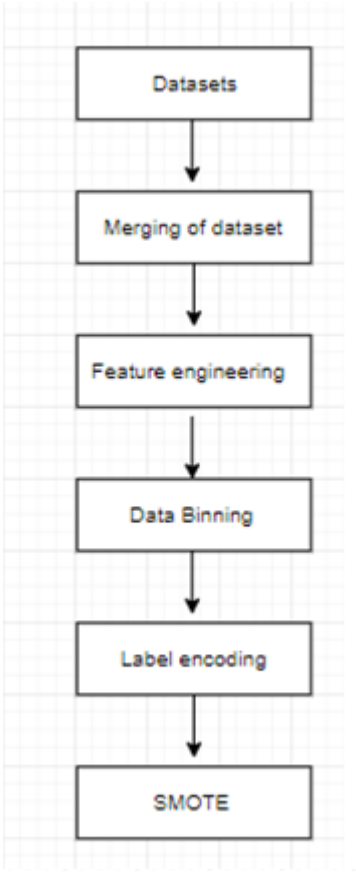


**Fig1 board steps form the data sets to finding frauds**

First Merging of dataset is done between inpatient data and outpatient data, second merging of dataset is done between the first merged data and beneficiary data and third merging of dataset is done between the second merged data and train dataset giving us the final train data set of size 558211 rows and 55 columns. From the final train dataset after merging feature engineering that

is the extraction of features (columns) has been done because when constructing a fraud detection model, it is very essential to extract the right features from the dataset to have high accuracy of fraud detected correctly. We have extracted the features that would help us in detection of fraud and we have deleted the columns after the extraction of features. Binning of data has been also performed on some of the features. Data Binning also known as bucketing, categorization or quantization to simplify and compress a column of data by reducing the number of possible values or levels represented in the data. Some of the advantages of data binning is that protection against minor data errors and outliers and for handling missing values.

After the binning of values we have dropped some of the attributes which contains NA and IDs giving us the final train datasets of 558211 rows out of which 212796 as potential fraud and 345415 as non fraud. A little Exploratory data analysis has been done in final train data in which we found that Claims belonging to column race 3 has high probability of being fraud, Insurance claims are mostly taken from the age group of 65-90, insurance fraud is seen in both gender female and male and if the amount to be reimbursed is greater than 60000 for a claim, it has higher probability of being fraud.

**RESULTS**

| **Model** | **Train** | **Test** |
|---|---|---|
| Logistic Regression | 0.866 | 0.897 |
| Support Vector Machine ( SVM ) | 0.875 | 0.888 |
| Random Forest | 0.975 | 0.883 |
| Naïve Bayes | 0.724 | 0.913 |
| Decision Tree | 0.928 | 0.843 |
| XGB | 0.958 | 0.891 |

| **Model** | **Fraud** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|---|
| Logistic Regression | No | 0.9751 | 0.9110 | 0.9429 | 989 |
| | Yes | 0.4430 | 0.7527 | 0.5578 | 93 |
| Support Vector Machine (SVM) | No | 0.9769 | 0.8989 | 0.9363 | 989 |
| | Yes | 0.4186 | 0.7742 | 0.5434 | 93 |
| Random Forest | No | 0.9716 | 0.8989 | 0.9338 | 989 |
| | Yes | 0.4012 | 0.7204 | 0.5154 | 93 |
| Naïve Bayes | No | 0.9552 | 0.9494 | 0.9523 | 989 |
| | Yes | 0.4949 | 0.5269 | 0.5104 | 93 |
| Decision Tree | No | 0.9756 | 0.8504 | 0.9087 | 989 |
| | Yes | 0.3273 | 0.7742 | 0.4601 | 93 |
| XGB | No | 0.9708 | 0.9090 | 0.9389 | 989 |
| | Yes | 0.4231 | 0.7097 | 0.5301 | 93 |

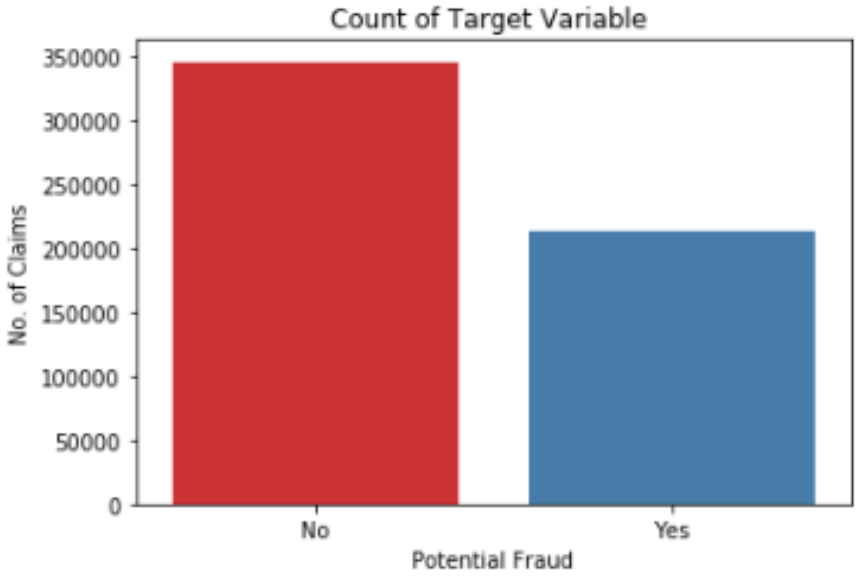


**Fig 4. Fraud claims**

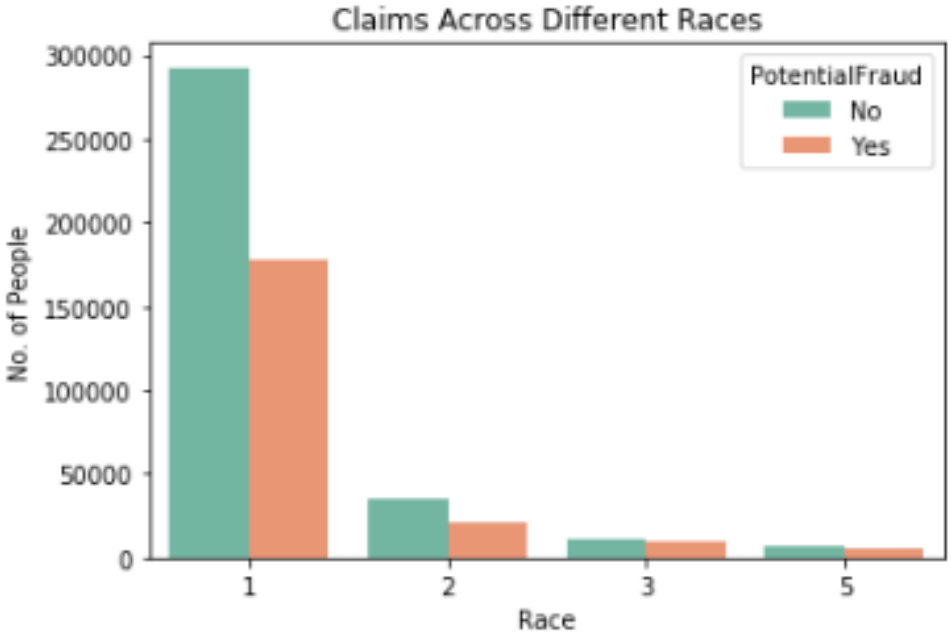


**Fig 5. Fraud claims across race**

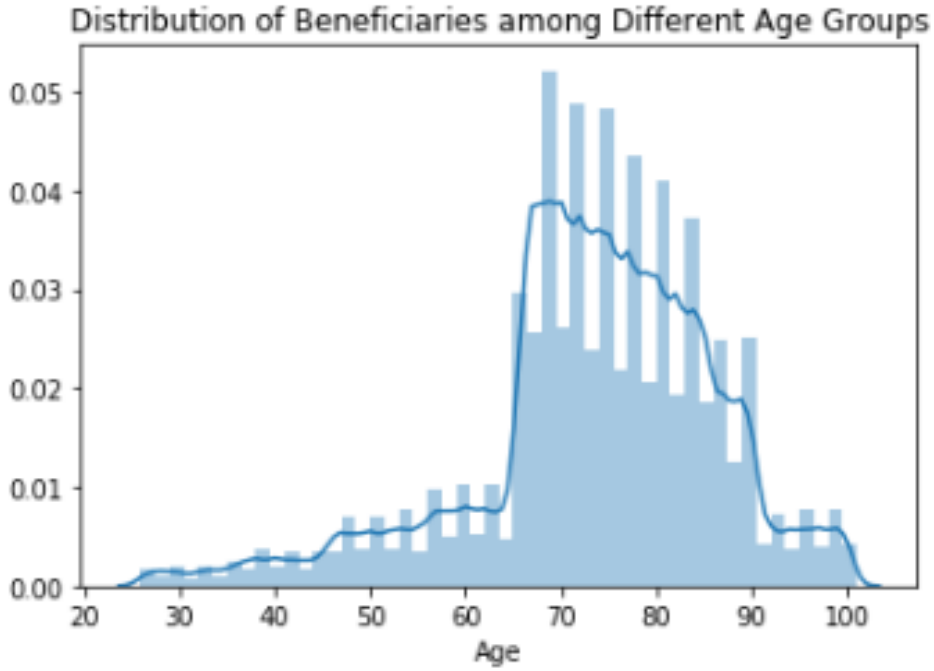


**Fig 6. Distribution of Beneficiaries among Different Age group**

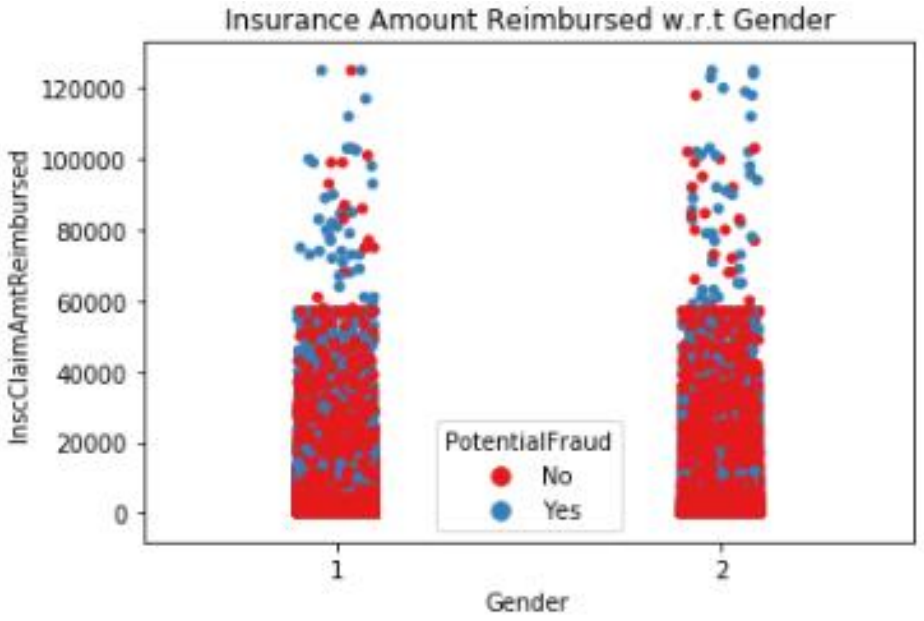


**Fig 7. Insurance amount with respect to gender**

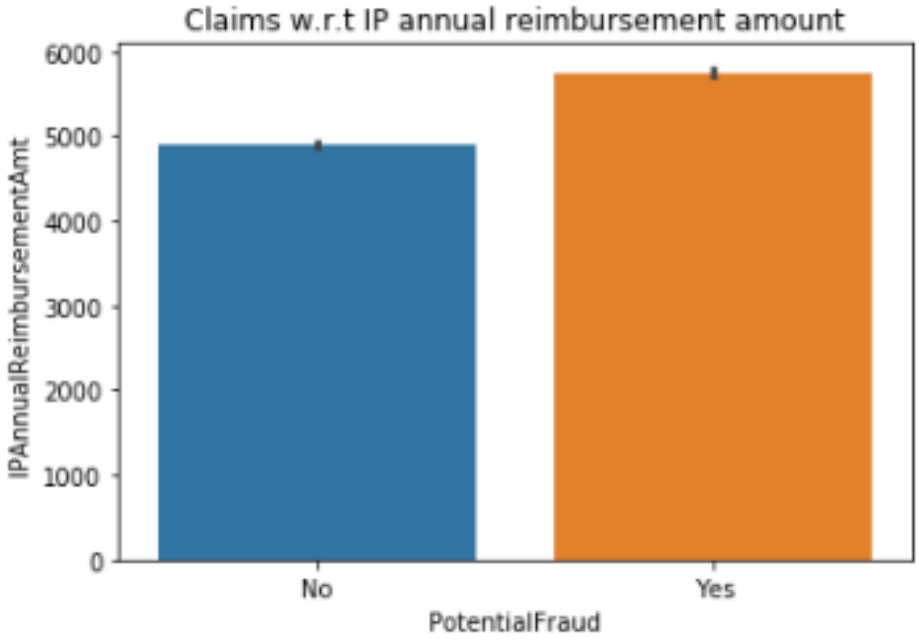


**Fig 8. Claims with respect to annual reimbursement amount**

## CONCLUSION

To improve detection of fraud in insurance claims in an automatic and effective way, building an accurate and efficient fraud detection system is one of the key tasks for insurance institutions. In this project six classification methods were used to build fraud detecting models. The work demonstrates the advantages of applying the data mining techniques including Logistic Regression and Support Vector Machine.

The result shows from the experiment that Random Forest has high accuracy score in training set and naïve bayes in testing set and naïve bayes has least accuracy score in training set and random forest in testing set. We have used random forest model in training set and naïve bayes model in testing set to classify the fraud claims since both model objective is to classify fraud claims

## REFERENCES

[1] Taghi M.Khoshgoftaar, Richard A.Bauder"*Medicare Fraud Detection Using Machine Learning Methods*" 16 th IEEE International Conference on Machine learning and application. 18-20 th December 2017.

[2] Richard A.Bauder, Taghi M.Khoshgoftaar "*A Survey of Medicare Data Processing and Integration for Fraud Detection.*" IEEE international Conference on Information Reuse and Integration.INSPEC Acessionm number (1799050) .July 2018.

[3] Charles.Francis, Noah.Pepper and Hormer.Strong "*Using Support Vector Machines to DetectMedical Fraud and Abuse*" Annual International Conference of the IEEE EMBS Boston,Massachusetts USA, August 30 - September 3, 2011

[4] A. Gangopadhyay, S. Chen, and Y. Yesha, "*Detecting healthcare fraud through patient sharing schemes,*" in Information Systems, Technology and Management. Springer, 2012.

[5] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "*Predicting medical provider specialties to detect anomalous insurance claims,*" in Tools with Artificial

Intelligence (ICTAI), 2016 IEEE 28th International Conference on. IEEE, 2016

[6] V. Chandola, S. R. Sukumar, and J. C. Schryver, "*Knowledge discovery from massive healthcare claims data*," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[7] Saba kareem, Dr. Rohiza Binti Ahmad, Dr. Aliza Binit Sarla"Framework for the *Identification of Fraudulent Health Insurance Claims using Association Rule Mining*" 2017 IEEE Conference on Big Data and Analytics (ICBDA)