MINI PROJECT REPORT

on

**Auto Insurance Fraud Detection using Machine Learning**

By

**Aditya Sadashiv (118A1003)**

**Dinesh Choudhary (118A1019)**

**Saravanan Iyer (118A1028)**

UNDER THE GUIDANCE OF

**Prof. Aarthi Boyanapalli**

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF

BACHELOR OF ENGINEERING

In

Computer Engineering



**DEPARTMENT OF COMPUTER ENGINEERING**

**SIES GRADUATE SCHOOL OF TECHNOLOGY**

**NERUL, NAVI MUMBAI – 400706**

ACADEMIC YEAR

2020 – 2021

**CERTIFICATE**

This is to certify that this is a bonafide record of Mini Project titled **“Auto Insurance Fraud Detection using Machine Learning”** carried out by the following students of third year in Computer Engineering.

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Name** | **Roll No.** |
| 1 | Aditya Sadashiv | 118A1003 |
| 2 | Dinesh Choudhary | 118A1019 |
| 3 | Saravanan Iyer | 118A1028 |

The report is submitted in partial fulfilment of the degree course of Bachelor of Engineering in Computer Engineering Department, of University of Mumbai during the academic year 2020-21.

Prof.Aarthi Boyanapalli Dr. Aparna Bannore Dr. Atul Kemkar

**Internal Guide         Head of Department                          Principal**

We have examined this report as per University requirements at SIES Graduate School of Technology, Nerul (E), Navi Mumbai on \_\_\_\_\_\_\_\_\_\_\_\_.

**Name of External Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature with Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Name of Internal Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature with Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACKNOWLEDGEMENT**

We would like to express our thanks to the people who have helped us the most throughout our project. We are grateful to our guide Prof. Aarthi Boyanapalli and coordinator Dr. Rajesh D. Kadu for nonstop support for the project.

A special thank of ours goes to each other who worked together as a team in completing the project, where we all exchanged our own interesting ideas, thoughts and made this possible to complete our project with all accurate information.

We also wish to thank our parents for their personal support and attention who inspired me to go my own way.

We would also like to extend our sincere gratitude to our Principal (Dr. Atul Kemkar) and our Head of the Department Dr. Aparna Bannore for their continuous support and encouragement.

We also would like to thank our other faculty members for providing us with all the required resources and references for the project.

**Project Group**

Aditya Sadashiv

Dinesh Choudhary

Saravanan Iyer

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| Sr. No. | Topic | Page No. |
| 1. | Abstract | i |
| 2. | List of Tables | ii |
| 3. | List of Figures | iii |
| 4. | - Intro Introduction | 1 |
| 5. | Literature Survey | 4 |
| 6. | Proposed System | 5 |
| 7. | System Design  Experimental/Modelling/ Designing/Programming/data collection etc. | 8 |
| 8. | Experimental Results and System Screenshots | 13 |
| 9. | Conclusion and Future Scope | 17 |
| 10. | References (in chronological order) | 18 |

**I. ABSTRACT**

Insurance fraud detection is a challenging problem, given the variety of fraud patterns and relatively small ratio of known frauds in typical samples. While building detection models, the savings from loss prevention needs to be balanced with cost of false alerts. Machine learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates. In this paper, multiple machine learning techniques for fraud detection are presented and their performance on various data sets examined. The purpose of this project is to work on Insurance Claims Dataset and to detect the Fraud Claims Insurance fraud has existed since the beginning of insurance as a commercial enterprise. Fraudulent claims account for a significant portion of all claims received by insurers, and cost billions of dollars annually. Types of insurance fraud are diverse, and occur in all areas of insurance. Insurance crimes also range in severity, from slightly exaggerating claims to deliberately causing accidents or damage. Fraudulent activities affect the lives of innocent people, both directly through accidental or intentional injury or damage, and indirectly as these crimes cause insurance premiums to be higher. Insurance fraud poses a significant problem, and governments and other organizations make efforts to deter such activities.. In recent days, Machine learning for spam classification is an important research issue. The proposed system classifies based on XGBoost and Linear Regression giving f-score of 84.31% and 62.75% respectively. A comparative analysis among the algorithms has also been presented in proposed system.

**II. LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Table Name** | **Page No.** |
| 1 | [Python Packages Used](#Packages) | 8 |
| 2 | [Performance Analysis](#Performance) | 15 |

**III. LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Figure Name** | **Page No.** |
| 1 | [Proposed System Block Diagram](#Block_Diagram) | 5 |
| 2 | Flowchart | 6 |
| 3 | Confusion Matrix | 9 |
| 4 | Performance of various algorithms | 16 |
| 5 | [Sample Output](#Samp_Restful_op) | 16 |

**CHAPTER 1**

**INTRODUCTION**

Insurance fraud refers to any claim with the intent to obtain an improper payment from an insurer. Motor and health insurance are the two prominent segments that have seen a spurt in fraud. Frauds can be classified from source and/or nature point of view. Sources can be policyholder, intermediary and/or internal with the latter two being more critical from internal control framework point of view. Frauds can be classified into nature wise, for example, application, inflation, identity, fabrication, staged/contrived/induced accidents etc.

Insurance frauds cover the range of improper activities which an individual may commit in order to achieve a favorable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation including the relevant actors and the cause of incident and ­finally the extent of damage caused.

Potential situations could include:

* Covering-up for a situation that wasn’t covered under insurance¬ (e.g. drunk driving, performing risky acts, illegal activities etc.)
* Misrepresenting the context of the incident: This could include¬ transferring the blame to incidents where the insured party is to blame, failure to take agreed upon safety measures.
* Inflating the impact of the incident: Increasing the estimate of loss¬ incurred either through addition of unrelated losses (faking losses) or attributing increased cost to the losses.

.

**1.1 Machine Learning Algorithms:**

Machine Learning Algorithms are classified into the three main types:

* + 1. **Supervised Machine Learning:** It is a type of Machine Learning in which Machines are trained using well “Labelled” training data. On the basis of this, the machine predicts the model.
    2. **Unsupervised Machine Learning:** Here models are not supervised using training dataset. Instead, the model itself finds hidden patterns and insights from the hidden data.
    3. **Reinforced Machine Learning:** Here output depends upon the state of the current input and the next input depends upon the output of the previous input.

The Machine Learning Algorithms which we have used in our project are XGBoost Classifier and Linear Regression which were used to calculate the Precision scores, Recall scores and the F-scores. Other algorithms include Random Forest algorithm, ADA, LDA ,KNN DT and Support Vector Machine. Machine Learning Algorithms are the engines of Machine Learning, i.e., it is the algorithms that turn a data set into a model.

Traditional approach consists of two ways:

1. Based on Certain rules: Here there would be a committee and that committee would create certain rules based on that rules, they would define weather the given insurance claim is fraud or genuine, if it is fraud then the it would be sent to investigation.
2. A checklist would be prepared based on the aggregation method: Here they would prepare a checklist which includes certain indicators along with the scores associated with that and what they would do is that , an aggregation of all the scores along with the value of the claim would be calculated and then this calculated value will be correlated with the boundary value, which would have been calculated earlier, if this calculated value is larger than the boundary value then the case would be sent to investigation else insurance would be provided.

These traditional approaches has certain limitation

1. Minimum number of parameters
2. Maximum human intervention So, to overcome these drawbacks we are developing a machine learning model which predicts whether the given claim is fraud or genuine.

**1.2 Objective:**

The main motive/objective of the project is to successfully detect and insurance fraud with the help of Machine Learning methods due to the upsurge in the volume of unwanted claims that has created an intense need for efficient and dependable robust system.The Machine Learning techniques have the capacity to learn and identify fraudulent claims by analysing loads of such messages throughout a vast collection of computers.

**1.3 Scope:**

* + Use of different imputers for missing values to see if any other gives superior performance such as iterative imputer
  + Application of sampling techniques to handle class imbalance such as SMOTE, random oversampling
  + Identify unique set of features that might always be leading to fraudulent claims.
  + Use of pipeline to streamline the modelling process
  + Adding features to dashboard and make it more user-friendly
  + There is a need to capture more data as not many useful business insights can be drawn from 1000 datapoints. More data, better results

**1.4 Outline:**

The remaining project is organized as follows; In chapter 2, we discuss the literature survey in detail; Chapter 3 covers block diagram of the process, methodology, function parameters, algorithms used for data analysis. Chapter 4 is the overview of System Design, packages used and steps taken for XGBoost. In chapter 5, we show screenshots of results after testing the model and also the analysis for different algorithms. It also shows output Finally, we concluded our work in chapter 6.

**CHAPTER 2**

**LITERATURE SURVEY**

1. A paper titled “Medicare Fraud detection using machine learning methods” [2]. This paper does an investigational survey by studying multiple supervised as well as unsupervised classification methods to discover the fraud cases. They have considered 3 groups or methods, 1. The Supervised learner which includes Random forest, Deep Neural Networks, Naïve Bayes Etc., 2. The Unsupervised learner includes KNN. Autoencoder etc., 3. The Hybrid learner includes NN Model. Their results show that Supervised learners are significantly better when compared to that of other type of learner.
2. A paper titled “Nearest Neighbours and Statistics Method based for detecting fraud in auto insurance” [3]. This study uses Nearest Neighbours method and statistics method to detect the occurrence of fraud and these methods are explained in detail in this paper. They made the comparison between SVM Method and the method that were used in this study, and found that when compared to SVM and interquartile range (statistic method), Nearest Neighbours method provides best performance.
3. A paper titled “Detecting insurance fraud by using Data mining Techniques” [4]. Makes use of 3 algorithms they are Bayesian Network, C4.5, Decision tree-based algorithm for predicting and analyzing fraud pattern from data. This model prediction will be aided by Bayesian naïve visualization, Decision tree visualization and rule-based classification. They looked at model performance matrices like recall, accuracy, and precision. This will be stronger when compared to that of class skew, by making it reliable performance matric in numerous prime insurance fraud detection functional areas.

**CHAPTER 3**

**PROPOSED SYSTEM**

When public claims for an insurance, branch in-charge receives the claim and performs some investigation like, whether the public has claimed for an insurance before, if so what type of claim was it whether it’s a fraudulent or genuine case, whether the premium has been paid properly then sends 8 parameters as input to our Insurance Fraud System. Our system performs some analysis and outputs as either fraud or genuine. If genuine then payment will be provided else case will be sent to investigation.

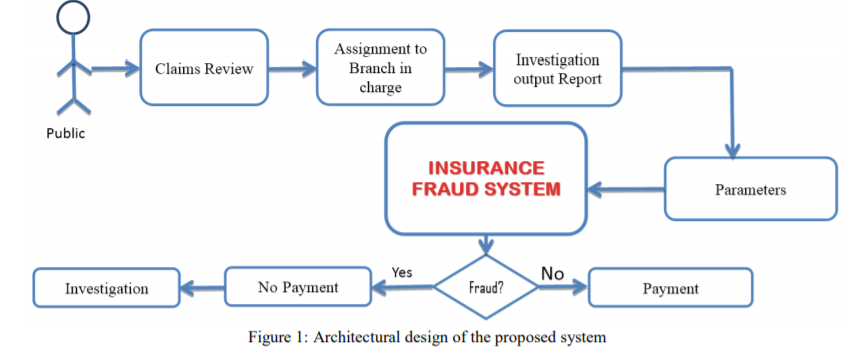


Figure 1: Proposed System Workflow

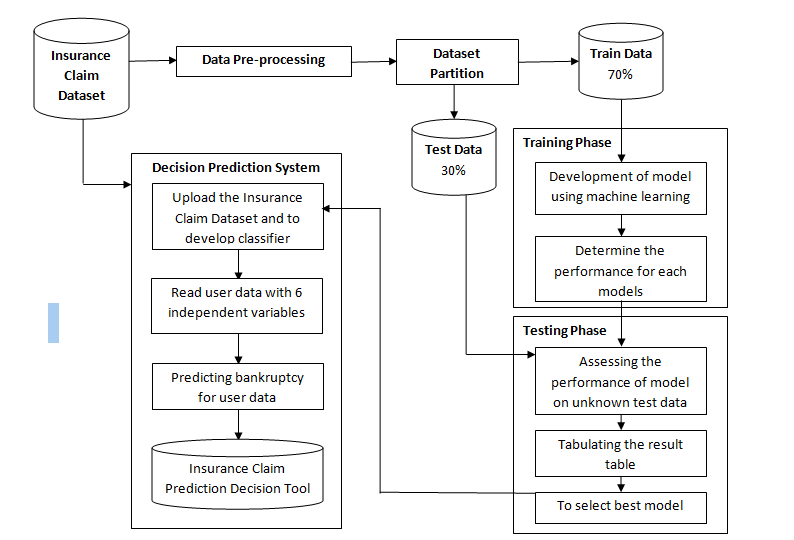
Advantages

* + The proposed method overcomes the low accuracy forecast problem.
  + Utilizing latest AI methods, the fraudulent transactions are recognized and the false alerts are reduced.
  + Fast and reliable solution is attained.
  + Less manual work needed for additional verification.
  + Higher accuracy of fraud detection.
  + Ability to identify new patterns and adapt to changes.
  + False declines or false positives happen when a system identifies a legitimate transaction as suspicious and wrongly cancels it.

**3.1 Methodology:**

When public claims for an insurance, branch in-charge receives the claim and performs some investigation like, whether the public has claimed for an insurance before, if so what type of claim was it whether it’s a fraudulent or genuine case, whether the premium has been paid properly then sends 8 parameters as input to our Insurance Fraud System.

Our system performs some analysis and outputs as either fraud or genuine. If genuine then payment will be provided else case will be sent to investigation.



* Once the dataset is obtained and cleaned, different models are tested on it.
* Based on the initial model performance, different features are engineered and re-tested
* Once all the features are engineered, the model is built and run using different β values and using different iteration procedures (feature selection process)
* In order to improve model performance, the parameters that affect the performance are tweaked and re-tested

Separate models generated for each fraud type which self-calibrate over time - using feedback, so that they adapt to new data and user behaviour changes

**3.2 Function Parameters:**

**data:**

(Required) This will process the dataset containing information.

**mode:**

Default: mode=2: It is used when data only contains content.

Otherwise, it is considered to contain sender information and mail history as well.

**classifier:**

(Default) classifier=’manual’: Only Linear Regression is used to classify.

classifier=’xgb’: Only XGBoost is considered to classify.

**Returns:**

Boolean: True if fraud is detected and False for otherwise.

**3.3 Algorithms Used:**

* + 1. **Random Forest Classifier:**

This algorithm is a part of supervised machine learning technique. It can be used for both: classification as well as regression problems.

It is a classifier that contains number of Decision trees on various subset of the given dataset.

* + 1. **XGBoost Classifier:**

It is one of the most popular and efficient implementation of gradient boosted trees algorithm.

XGBoost stands for Xtreme Gradient Boosting.It uses CPU cache to store calculated gradients to make necessary calculations fast.

* + 1. **Linear Regression**

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, we can represent a linear regression as:

y= a0+a1x+ ε

**3.3.4.** **Decision Tree**

* + - Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
    - In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
    - The decisions or the test are performed on the basis of features of the given dataset.

**3.3.5. SVM:**

* + - The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
    - SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane

**CHAPTER 4**

**SYSTEM DESIGN**

This project uses Python language for all algorithms

**4.1 Python Packages Used:**

|  |  |
| --- | --- |
| **P****ackage Name** | **Version Name** |
| Scipy [[6]](https://www.scipy.org/) | 1.5.2 |
| Numpy [[7]](https://numpy.org/) | 1.19.2 |
| Nltk [[8]](https://www.nltk.org/) | 3.5 |
| Pandas [[9]](https://pandas.pydata.org/) | 1.1.3 |
| Xgboost [[10]](https://xgboost.readthedocs.io/en/latest/) | 1.3.3 |
| Scikit\_learn [[12]](https://scikit-learn.org/stable/) | 0.24.1 |
| Matplotlib [[13]](https://matplotlib.org/) | 3.4.2 |

Table 1: Python Packages Used

**Source of Dataset** :Kaggle [[4]](https://spamassassin.apache.org/old/publiccorpus/)

In order to train our model, we had to perform various steps. These steps are divided into 4 phases:

1. **Data Cleaning and Pre-Processing**
2. **Exploratory Data Analysis**
3. **Feature Selection**
4. **Models and Approaches**

The above steps are explained in detail below.

1. **Data Cleaning and Pre-Processing**

Needed to change dtype of certain columns from numerical to categorical as they were categorical features by nature:

* 1. Number of Vehicles Involved
  2. Number of Bodily Injuries
  3. Number of Witnesses

Decided to drop 5 columns as had too many unique values, did not give any useful insights:

* 1. Policy Number
  2. Policy Bind Date
  3. Insured Zip
  4. Incident Date
  5. Incident Location

***Imputation of missing values***

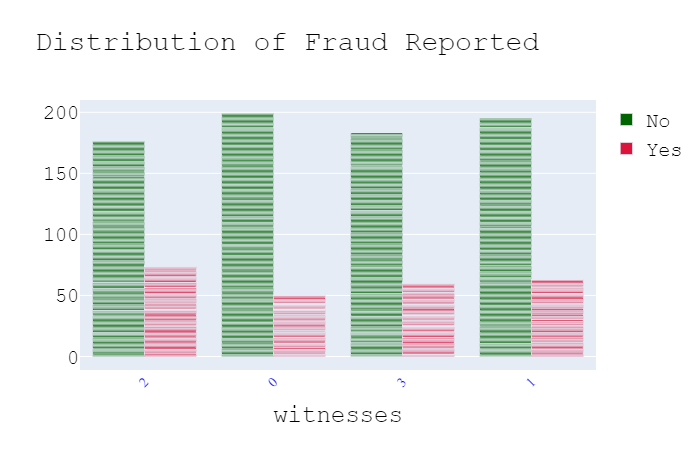
* Three columns had ‘?’ (missing values):
  + Collision Type
  + Property Damage
  + Police Report Available
* Imputed missing values using a KNNImuputer (K = 5)

***Label Encoding of Categorical Features***

* Categorical features were label encoded before inputting into the model.
* Did not one hot encode because of use of tree- based models.

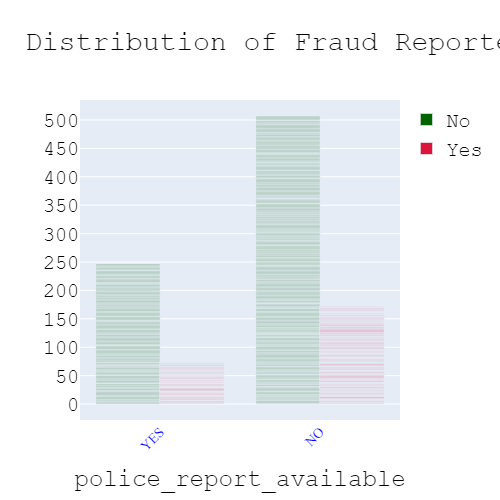
**B) Exploratory Data Analysis**

***Number Of Witnesses gives some information***

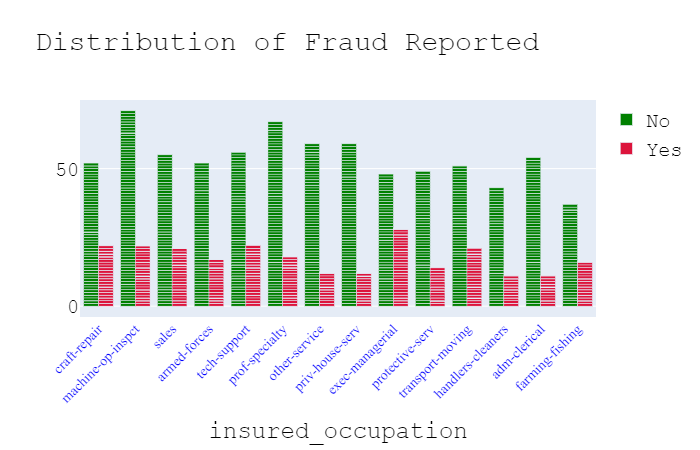
******

* Even the presence of one witness increase the chances of detecting a claim as fraudulent.
* Out of 247 fraudulent claims, 197 claims had atleast one witness present

***Police Report Is Important***

* **Availability of police report reduces the chances of fraudulent claims, although fewer cases actually have a police report available.
* Increase in number of datapoints would help to determine a more clear relationship between fraudulent claims and the availability of police reports.

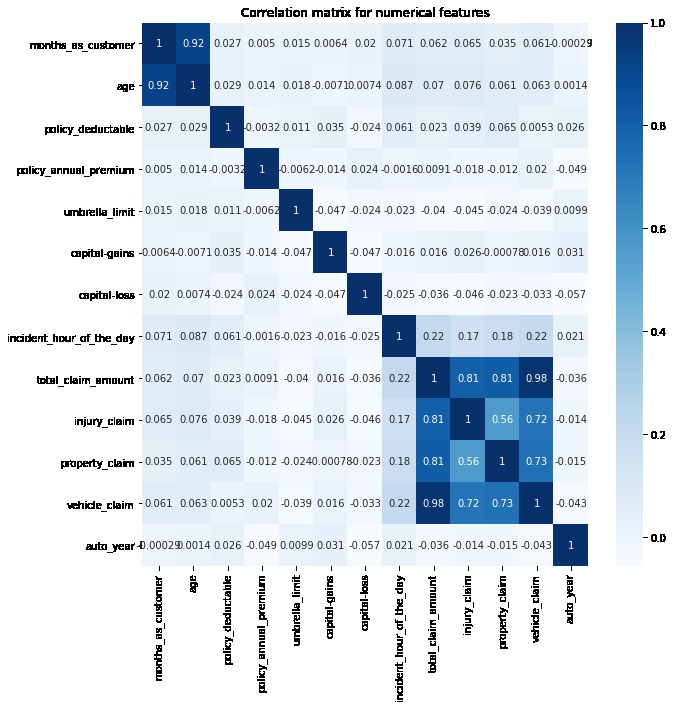
***Occupations Reveal an Interesting Fact***

**

* People who mentioned their occupation as Exec – Managers have the most likelihood of making a fraudulent claim
* 37% of claims made by this category was fraudulent.

**C) Feature Selection**

***Correlation Amongst Numerical Features***

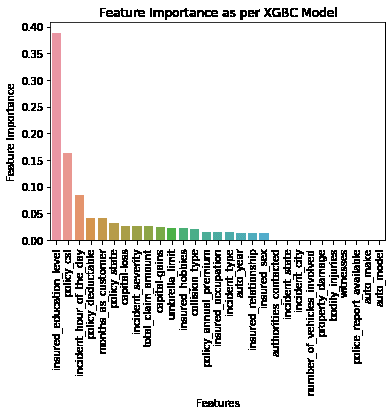
* + ******The following heatmap shows high correlation amongst the following columns:

Age and Months as customer

Total Claim Amount with its individual constituents. (Injury, Property, Vehicle)

* + Age and the individual Constituents were dropped from the dataset.

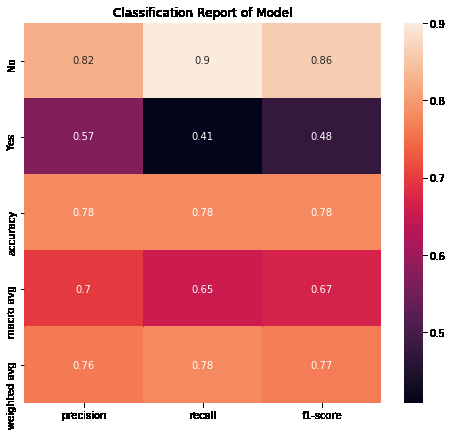
***Feature Importance as per XGBC***

**

* + As per Xgbc, insured education level and policy\_csl are the most important features.. This is followed by incident hour of the day.
  + However, choosing just three features does not allow the model to separate well between the classes
  + Decided to drop features with feature importance of less than 1%. This gives us 19 features.

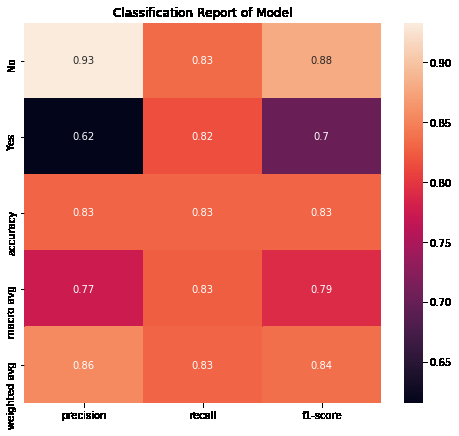
**D)Models and Approaches**

***Base model performance***

******

* The Base model performed fairly well on rightful claims, but not so great on the fraudulent claims.
* This shows the need to hypertune the model, by giving more weight to the underrepresented class. Could also be the case of overfitting on training dataset.

***Model Tuning***

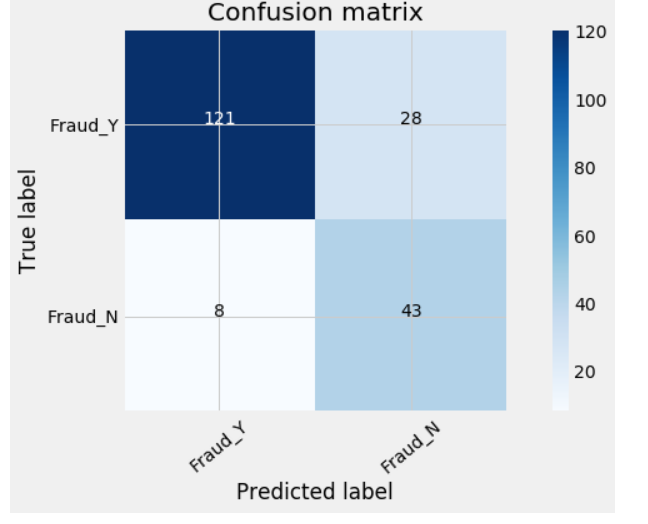
******

* Model performance has vastly improved because of hyperparameter tuning.
* Recall for both classes is above 80%.
* From a business point of view, recall of fraudulent class should be one, so still room for improvement.

**CHAPTER 5**

**EXPERIMENTAL RESULTS AND SYSTEM SCREENSHOTS**

After testing, we analyse the performance of our algorithm by finding the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). We plot those values along the decision boundary in order to visualise the performance of our model.



From the confusion matrix we see that,

* 121 transactions were classified as valid that were actually valid
* 8 transactions were classified as fraud that were actually valid (type 1 error)
* 28 transactions were classified as valid that were fraud (type 2 error)
* 43 transactions were classified as fraud that were actually fraud.

Err = {(28+8) / (121+8+28+43)}\*100 = 18%

So, the algorithm misclassified 18% fraudulent transactions.

**5.1 Performance Terminologies:**

**5.1.1:** TP: True Positives

**5.1.2:** TN: True Negatives

**5.1.3:** FP: False Positives

**5.1.4:** FN: False Negatives

**5.2 Performance Analysis and Formulae:**

**5.2.1 Accuracy:**

Accuracy is defined as the percentage of correct predictions for the test data.

**Formula**: Accuracy = [(TP + TN) / (Total Predictions)] \* 100

**5.2.2 Recall Score:**

Recall Score is the percentage of number of correct positive predictions made out of all positive predictions that could have been made. Recall Score provides an indication of missed positive predictions.

**Formula**: Recall Score = [(TP) / (TP + FN)] \* 100

**5.2.3 Precision Score:**

Precision Score is the percentage of number of correct positive predictions made out of all positive predictions made. Precision Score provides an indication of incorrect positive predictions.

**Formula**: Precision Score = [(TP) / (TP + FP)] \* 100

**5.2.4 F1 Score:**

F1 Score is the harmonic mean of precision and recall score. It is also called F-Score or F-Measure. In statistical analysis, of binary classification, the F-Score is a measure of a test’s accuracy.

**Formula**:

F1 Score = [(Precision Score \* Recall Score) / (Precision Score + Recall Score)] \* 100

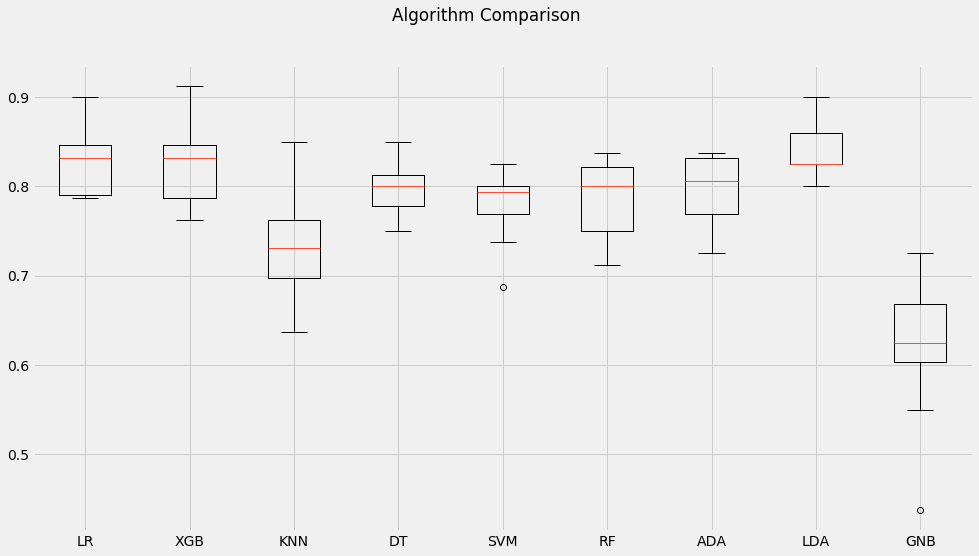


diagram 2: Performances of Various Algorithms

**Comparison:**

****

**Test Performance of model**

****

Hence we will keep the fitted XGB model as our final model

**CHAPTER 6**

**CONCLUSION AND FUTURE SCOPE**

The machine learning models that are discussed and applied on the datasets were able to identify most of the fraudulent cases with a low false positive rate i.e. with a reasonable precision. This enables loss control units to focus on new fraud scenarios and ensuring that the models are adapting to identify them. Certain datasets had severe challenges around data quality, resulting in relatively poor levels of prediction. Insurance fraud detection is a rough task, this industry has grappled with challenges of insurance claim fraud from the very beginning. Proposed system aims at developing a system that can help to recognize possible frauds with peak magnitude of accuracy. Proposed system predicts whether the claimed insurance is “FRAUD” or “GENUINE”. Thus helping the insurance companies to spot frauds with fewer amount of time and with good accuracy rate

The machine learning models that are discussed and applied on the datasets were able to identify most of the fraudulent cases with a low false positive rate i.e. with a reasonable precision. This enables loss control units to focus on new fraud scenarios and ensuring that the models are adapting to identify them. Certain datasets had severe challenges around data quality, resulting in relatively poor levels of prediction. System can be enhanced to predict the insurance fraud in bulk. System can be enhanced by adding feedback module, where users can posts feedbacks to clarify their doubts and also it can be upgraded by adding a report module, so whenever the output is Fraud, if branch in- charge clicks on the report button complete report must be sent to the nearby police station.

**REFERENCES**

* 1. https://ijarcce.com/wp-content/uploads/2020/08/IJARCCE.2020.9715.pdf
  2. <https://www.sciencedirect.com/science/article/pii/S2405844018353404#bib2>
  3. <https://spamassassin.apache.org/old/publiccorpus/>
  4. <https://www.python.org/doc/essays/blurb/>
  5. <https://www.scipy.org/>
  6. <https://numpy.org/>
  7. <https://www.nltk.org/>
  8. <https://pandas.pydata.org/>
  9. <https://xgboost.readthedocs.io/en/latest/>
  10. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
  11. <https://scikit-learn.org/stable/>
  12. <https://matplotlib.org/>