Auto Insurance Fraud

Problem **Statement:**

The goal of this project is to build a model that can detect motor insurance fraud. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims. This type of problems is known as imbalanced class classification.

Frauds are unethical and are losses to the company. By building a model that can classify auto insurance fraud, I am able to cut losses for the insurance company. Less losses equates to more earning.

Relevance to businesses:

Imbalance class problems are common in many industries. Many a times, we are interested in a minority class against another much bigger class or classes. For instance, classification of other types of frauds, classification of defective goods, classification of at-risk teenagers, identifying high potential employees, identifying people of interest such as terrorist, just to name a few.

Criteria for success:

The model should be able to classify if a claim is a fraud or not on a data set that it has not seen, accurately. It is important to distinguish between fraud and legit claims. This is because investigations into frauds can be time consuming and expensive and may even negatively affect customer experience.

Executive Summary

The goal of this project is to build a model that can detect auto insurance fraud.

Several models were tested with different methods of handling imbalance datasets. The top models were also fitted and tested with different ensembles.

The final fitted model is a weighted XGBoost with the accuracy of score of 0.99. The model performed far better than the baseline. The model was able to correctly distinguish between fraud claims and legit claims with high accuracy.

Prior to modeling, the data was clean and exploratory data analysis was conducted. After which, the data was pre-processed for the modeling. After modeling, the models were evaluated.

About the Dataset

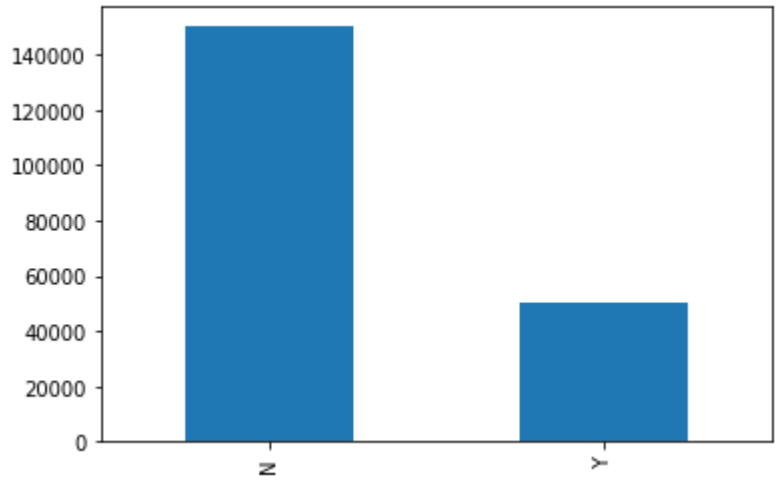
The inspiration for this project was to perform classification on imbalance class data sets, in particular fraud. Fraud data sets are very hard to come by and often unlabelled due to its sensitive nature.

The current data set was labelled with 200232 rows and 22 columns.

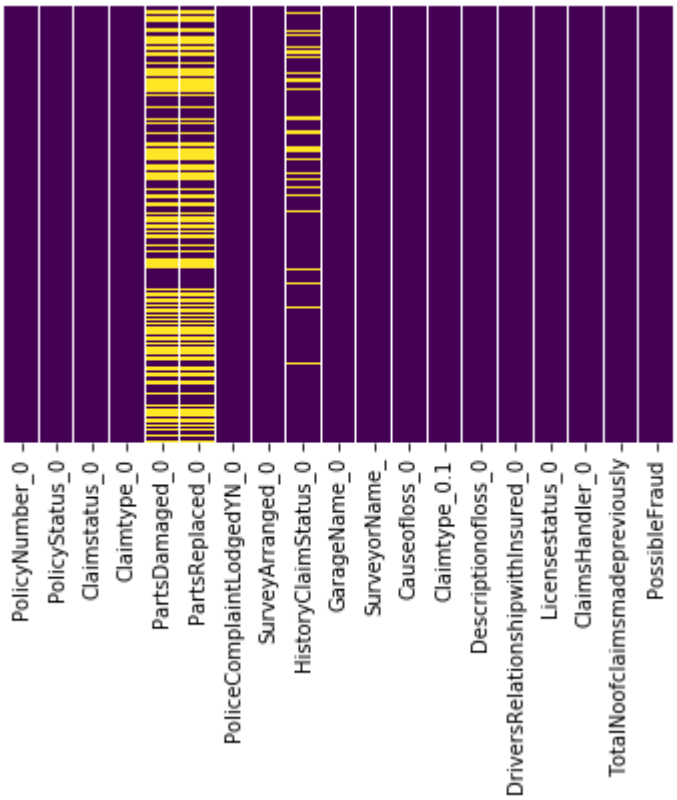
Exploratory Data Analysis

Dependent variable

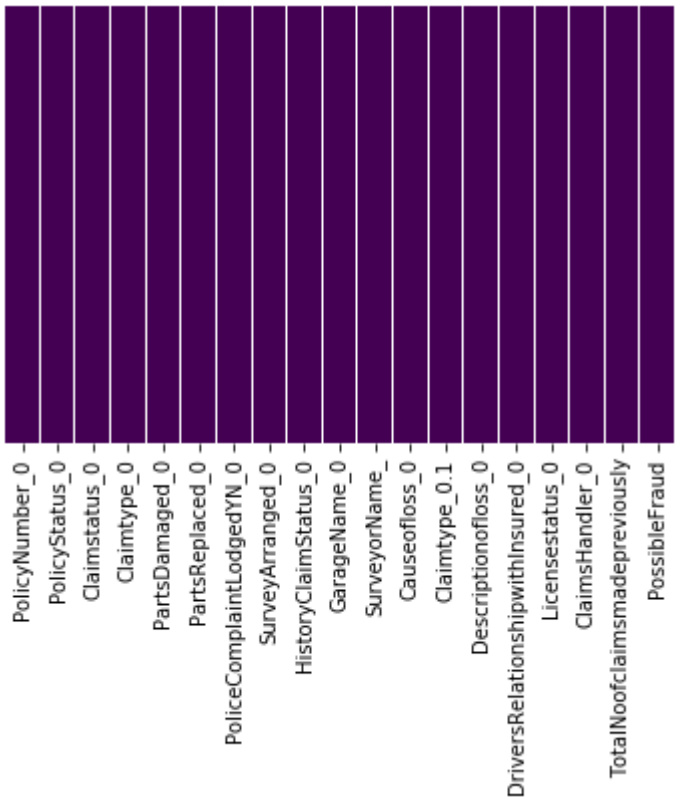
Exploratory data analysis was conducted started with the dependent variable, PossibleFraud. There were 50000 frauds and 150232 non-frauds. 24.7% of the data were frauds while 75.3% were non-fraudulent claims.



PartsDamaged_0 , PartsReplaced_0, HistoryClaimStatus_0, PossibleFraud, FIRno_0, and ClaimApprovedBy_0. Dropping FIRno_0 and ClaimApprovedBy_0 columns as there are maximum null values rest of the above mentioned columns are replaced the null values with mode.



From the below Heatmap we can see that there is no null values as all the null values are replaced by mode.

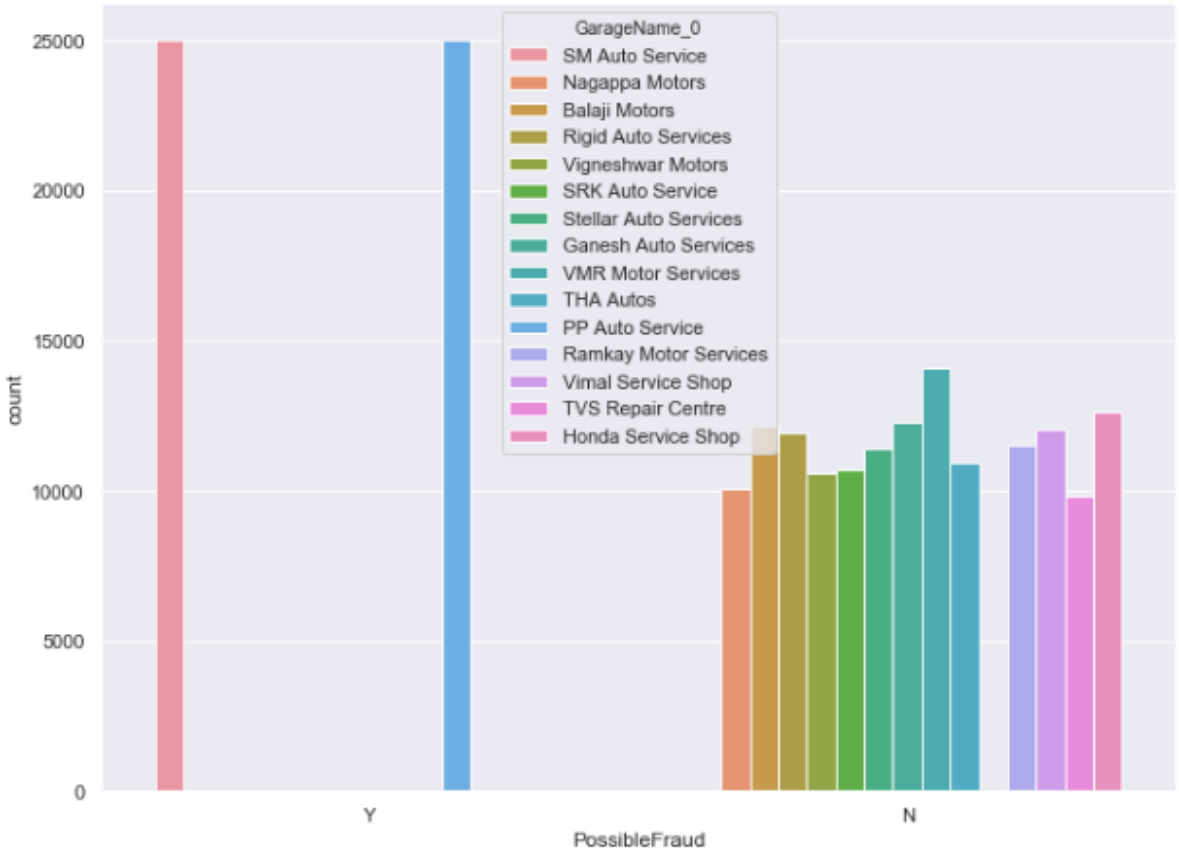


Visualizing variables against the PossibleFraud :

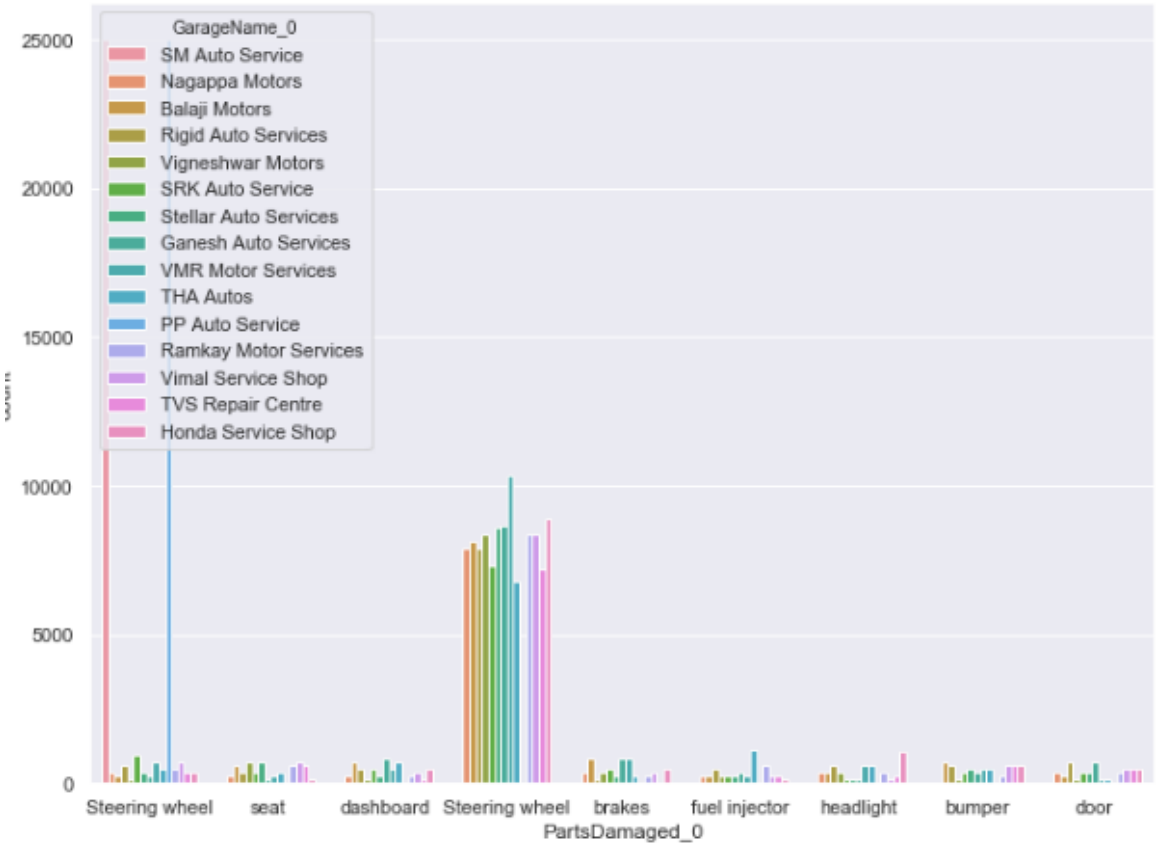
From the below visualisation we can say that total 50000 frauds are actually claimed in two auto services.

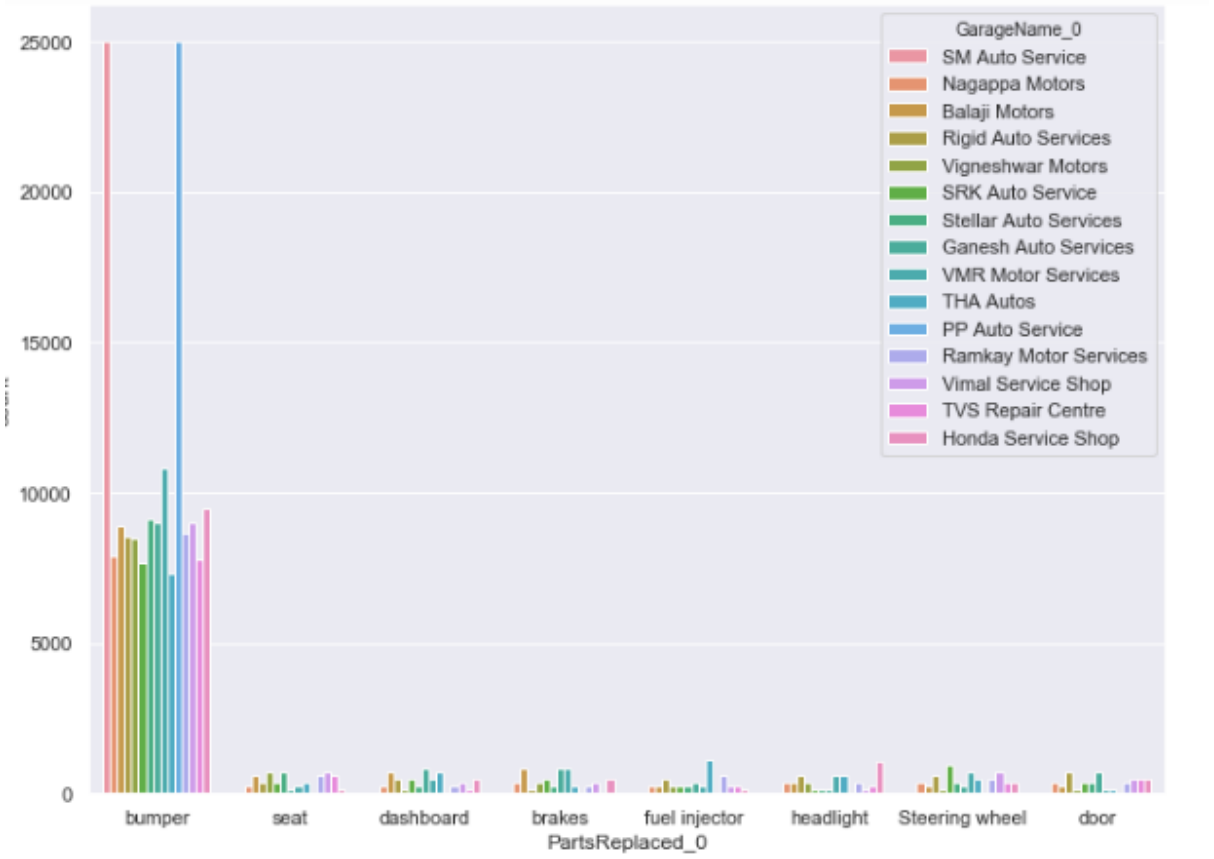
1. PP Auto service

2. SM Auto service



After visualizing parts damaged and parts replaced with the Gurage names its clearly say that both the auto services they do some paper work in that they false claim as steering wheel of the car is damaged and they replace with bumper which is a fraud statement.


GarageName_0
SM Auto Service
Nagappa Motors
Balaji Motors
Rigid Auto Services
Vigneshwar Motors
SRK Auto Service
Stellar Auto Services
Ganesh Auto Services
VMR Motor Services
THA Autos
PP Auto Service
Ramkay Motor Services
Vimal Service Shop
TVS Repair Centre
Honda Service Shop
25000
20000
15000
10000
5000
0
Steering wheel
seat
dashboard
Steering wheel
brakes
fuel injector
headlight
bumper
door
PartsDamaged_0

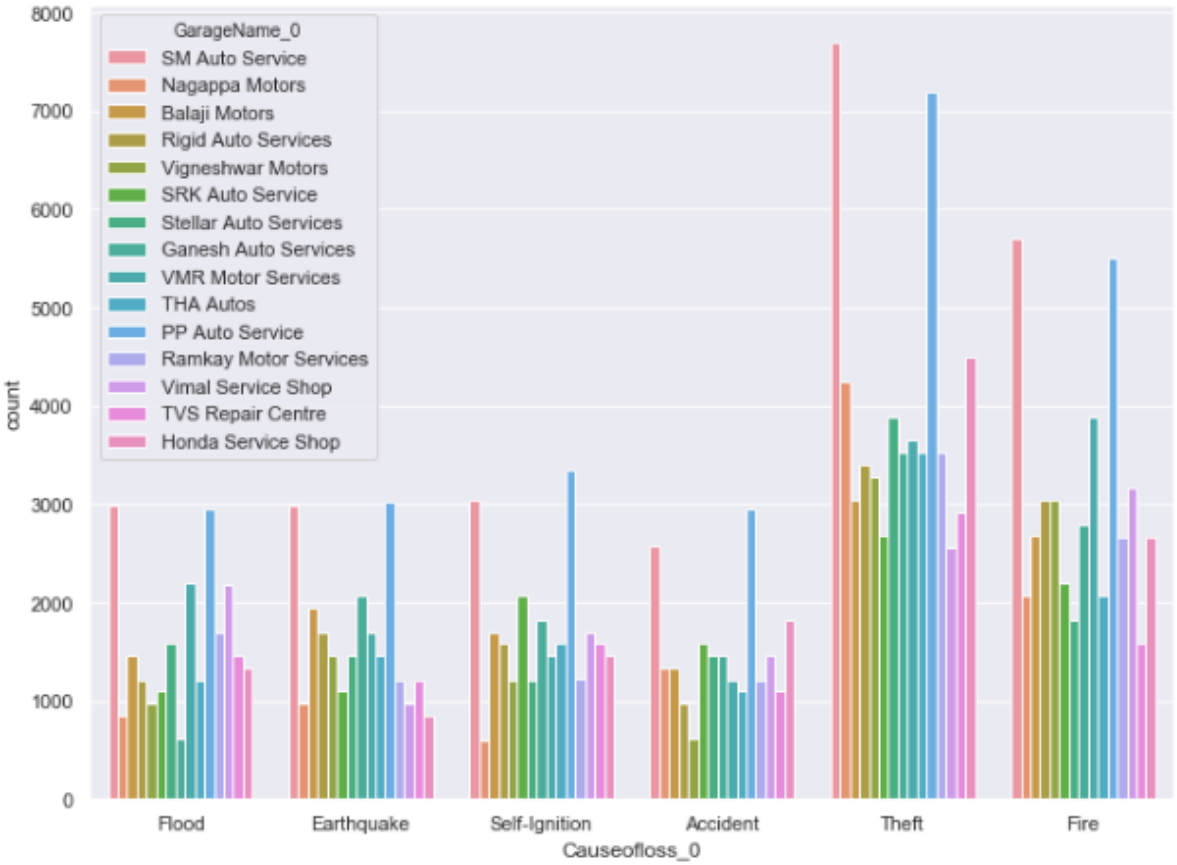

GarageName_0
SM Auto Service
Nagappa Motors
Balaji Motors
Rigid Auto Services
Vigneshwar Motors
SRK Auto Service
Stellar Auto Services
Ganesh Auto Services
VMR Motor Services
THA Autos
PP Auto Service
Ramkay Motor Services
Vimal Service Shop
TVS Repair Centre
Honda Service Shop
25000
20000
15000
10000
5000
0
bumper
seat
dashboard
brakes
fuel injector
headlight
Steering wheel
door
PartsReplaced_0

And then comes to the Surveyors. Let me tell the role of Surveyors. Surveyors are professionals who assess the loss or damage and serve as a link between the insurer and the insured. Once the surveyor has compiled their report, they will present this to the insurer

From PP auto service Peorsan is the Surveyors who do false claim the report and present it into the insurer.

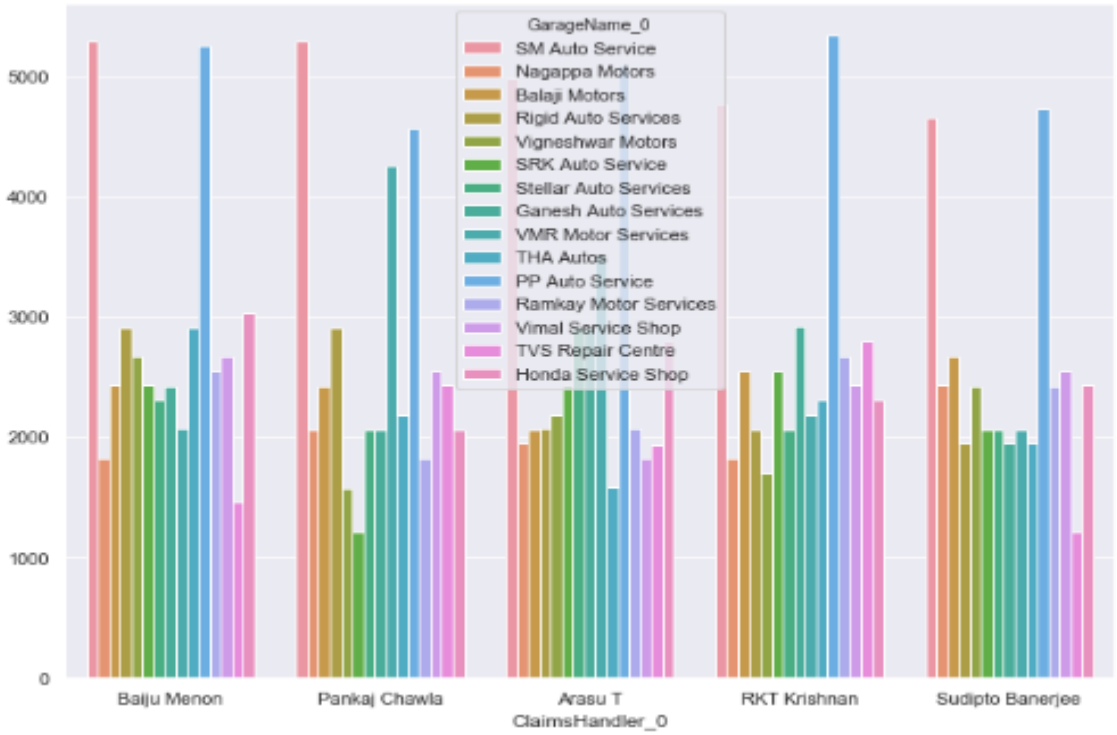
From SM auto service Nobert is the Surveyors who do false claim the report and present it into the insurer.

And the loss they described was Hit and run, theft, accident, run over pedestrian, collision .All this are actually false claims.



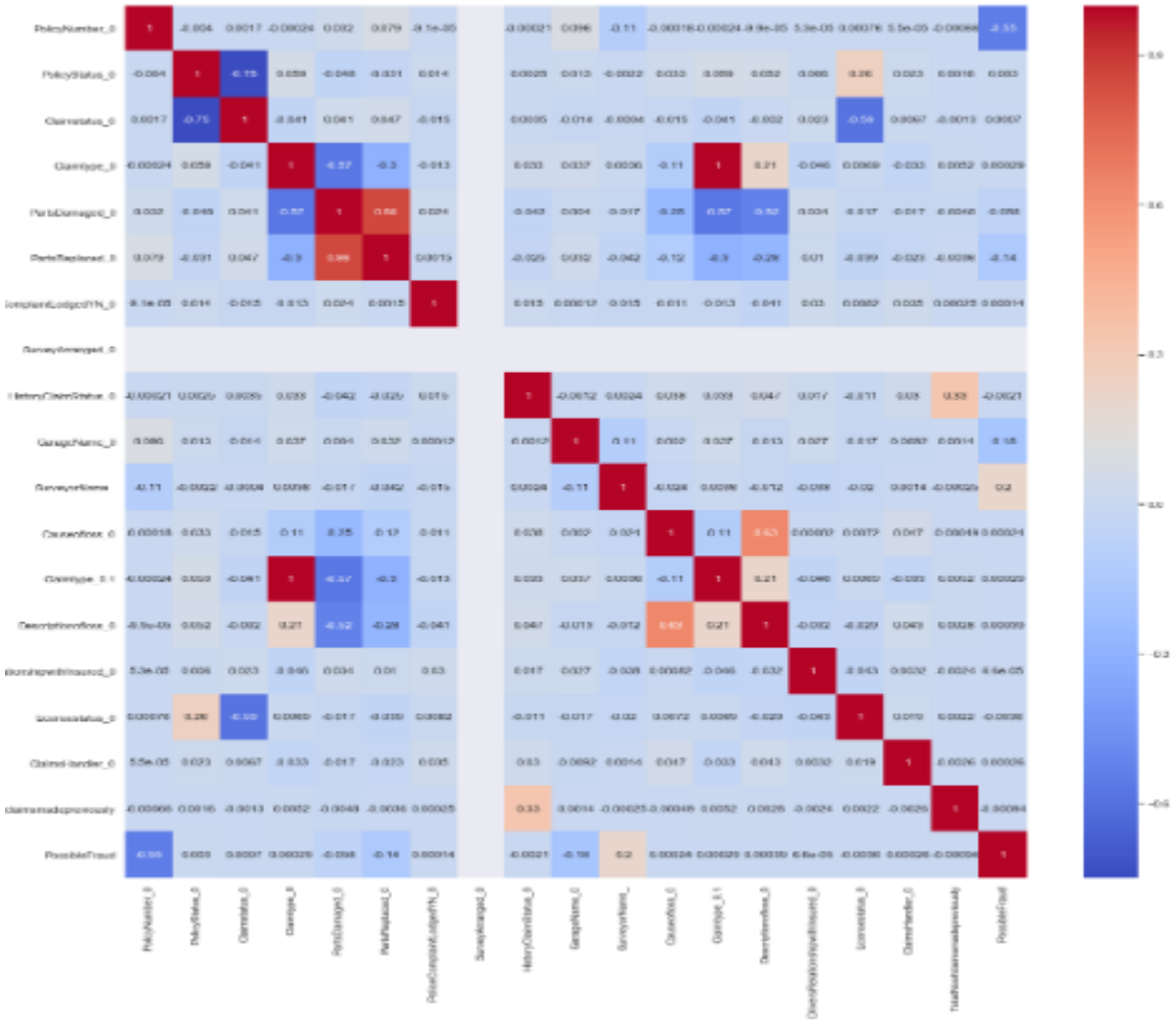
Then it comes to the Claims handlers. Claims handlers are responsible for processing and investigating insurance claims relating to customers' policies.

Krishnan, Baiju Menon, Arasu , Banerjee and Pankaj Chawla are the claim Handlers who do false clai ms.



All the columns are object data type,with the help of label encoding all are converted to integer dat a type.

Visualizing correlation with the attributes:



Model Building :

Five different classifiers were used in this project:

- Logistic regression

- Gradient Classifier

- Random forest

- XGBoost

- Isolation Forest and Local Outlier Factor

Algorithm steps:

Step 1: Read the dataset.

Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.

Step 4: Feature selection are applied for the proposed models.

Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

Step6: Then retrieve the best algorithm based on efficiency for the given dataset.

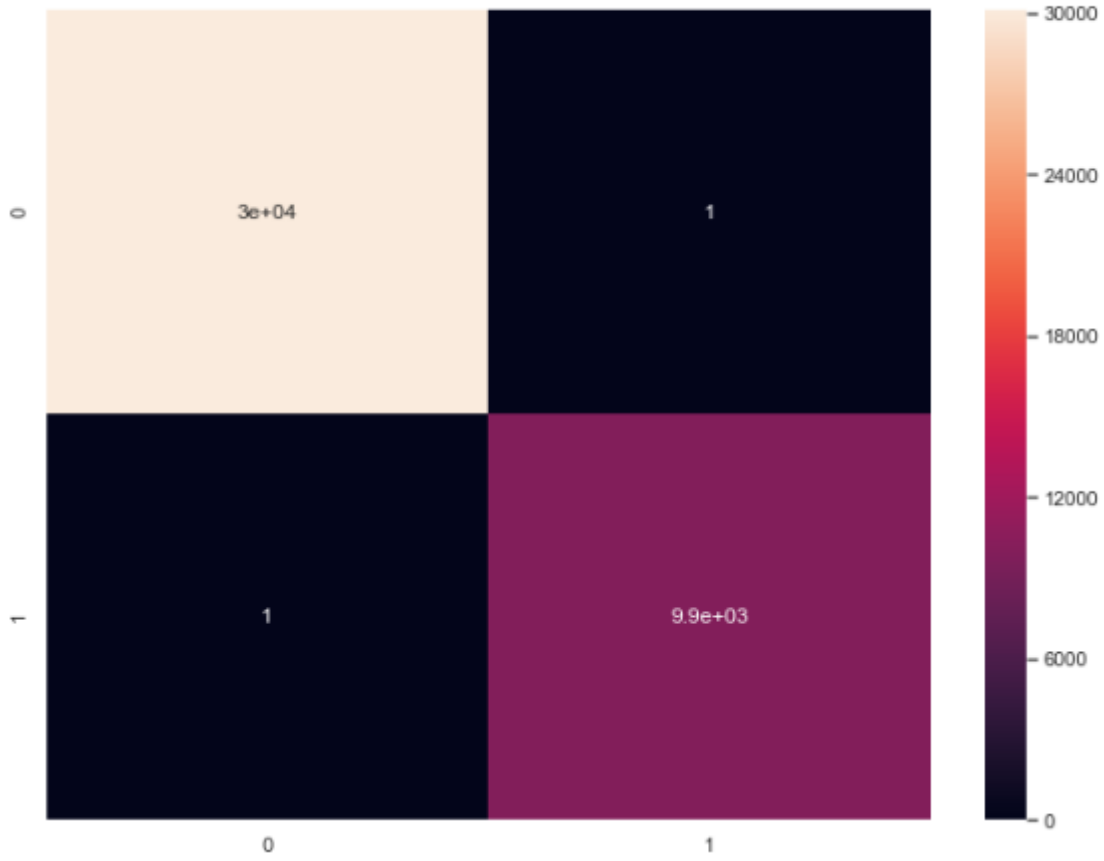
Confusion matrix:

Now the confusion matrix which shows 30141+9904 = 40045 correct predictions.

True Positives: 30141

True Negatives: 9904

```
array([[30141,     1],
       [    1,  9904]], dtype=int64)
```



In this project, Machine learning technique like Logistic regression, Gradient boosting, and XG Boost were used to detect the fraud in Motor Insurance system. Sensitivity, Specificity, accuracy and error rate are used to evaluate the performance for the proposed system. The accuracy for logistic regression, Decision tree and random forest classifier are 0.82, 0.99, and 0.99 respectively. By comparing all the three method, found that random forest classifier is better than the Gradient boosting, and XG Boost

The End