**Problem Definition**

**Goal of the project**

The goal is to find the Insurance fraud, which is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.



**Meaning of Automobile Fraud detection**

Insurance fraud occurs when an individual or group of individuals attempt to earn profit either through non-compliance or through finding ways and means to exploit loopholes in the terms and conditions of the insurance agreement. In the automobile sector the Insurance fraud occurs when collision occurs of the vehicle or there is vehicle theft. The type of fraud is not disclosed in this data set and could be false reports, inflating claims, staging accidents or submitting claim forms for damages or injuries that never occurred.

**Statistics of Fraud In India**

India’s insurance premium in 2018 for Life Insurance was US$73.74 billion and Non-Life Insurance was US$26.10 billion totalling US$99.84 billion

In FY2017-18 claims repudiated were 0.74, claims rejected were 0.43 of Life Insurance claims

According to a report, Insurance companies lose over US$6.25 billion to frauds which results in higher premiums for genuine consumers.

A media report stated that over 10% of claims in general insurance are fraudulent

According to the Insurance Regulatory and Development Authority (IRDA), every insurance company is required to set up a Fraud Monitoring Framework.



**Models aim**

The aim of the model built will predict the Automobile Insurance claimant is fraudulent or not.The model built to predict the claim will be aiming to obtain the highest Accuracy score. An accuracy score is fraction of predictions our model got right.

The model will also aim towards obtaining the highest ROC AUC score, precision score and recall.

**Methodology for execution of the modelling**

The data was cleaned and exploratory analysis was done

The data from the data set was split into the train and test data. The train data was then be checked for the best sample state. Using the sample state, the data was split again into the test and train data.

The data was tried to be balanced, but since its accuracy score was falling did not apply undersampling,oversampling or SMOTE.

This train data was modelled through various algorithms.The algorithms used here are Logistic regression, DecisionTreeClassifier, RandomForestClassifier and SVC. The model with least difference between accuracy score and Cross validation score was used for hyperparameter tuning and then the model was saved. The area under curve of the ROC (ROC AUC) will also be taken into consideration in model selection as a secondary criterion as it is important to distinguish between fraud and legit claims.

**Challenges in the project**

1. The dataset dependent variable is imbalanced
2. The sample size is small. tatistical models are more stable when data sets are larger. It also generalizes better as it takes a bigger proportion of the actual population.
3. the data only capture incident claims of 3 states
4. incidents between 2 months which may not be an accurate picture of the year

**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. The independent variables would be about the customer, through which the model would accurately predict if the claimant is fraud or not. The model is successful if it can predict the claim accurately.

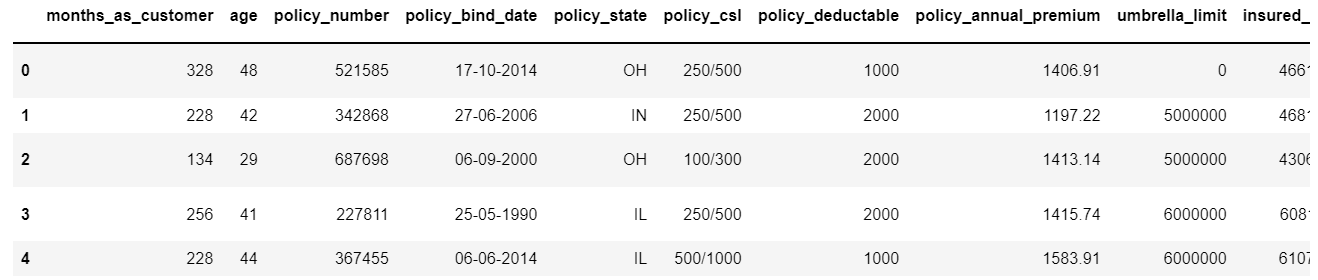
**Data Analysis**

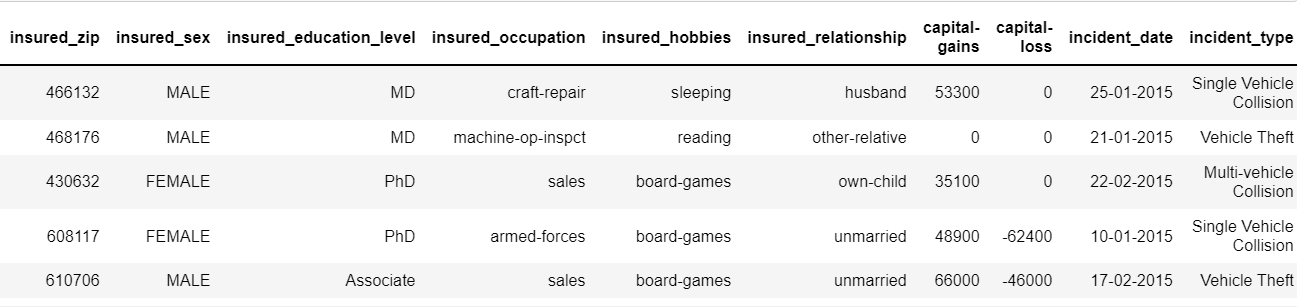
**Dataset Analysis**

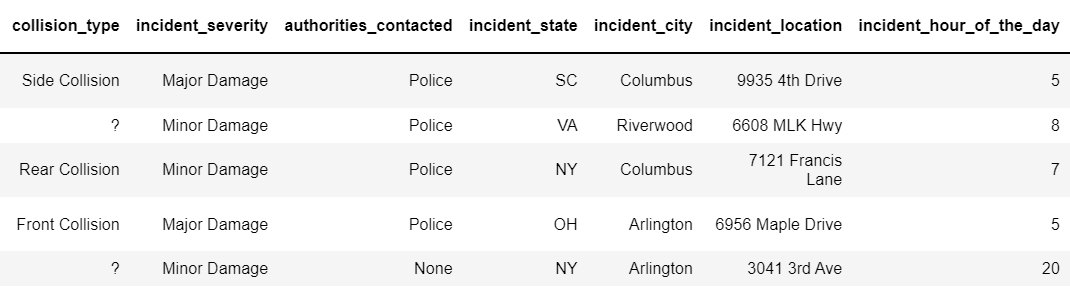
a.The dataset is obtained from:

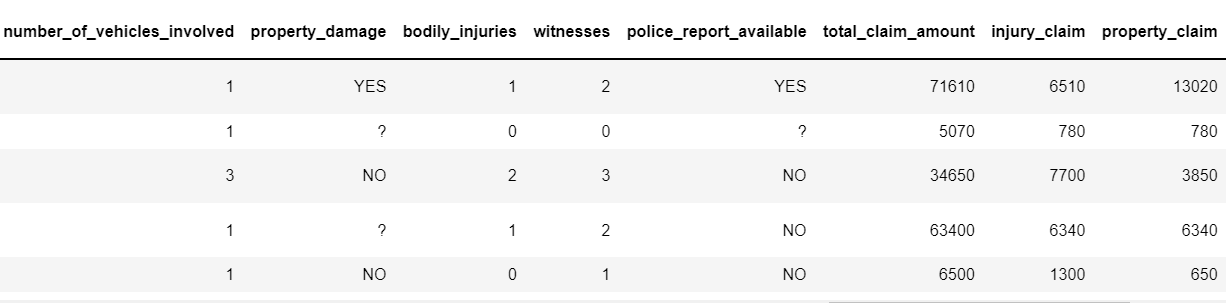
[**https://github.com/dsrscientist/DataScienceMLCapstoneProjects/blob/master/Automobile\_insurance\_fraud.csv**](https://github.com/dsrscientist/DataScienceMLCapstoneProjects/blob/master/Automobile_insurance_fraud.csv)

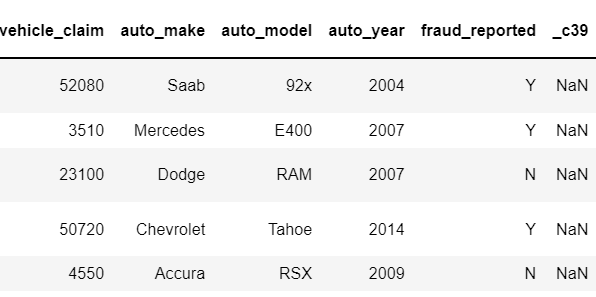
b.The dataset obtained is in CSV format



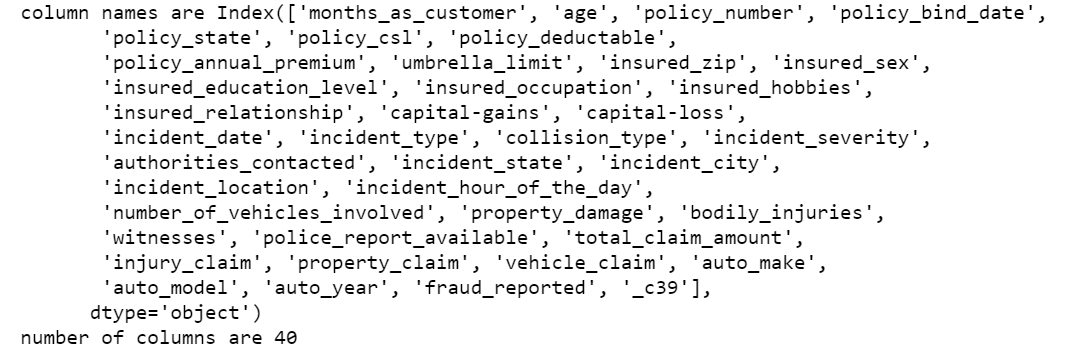








c.The column names/independent variables are :



d.Dataset has both categorical and numerical data. The label/dependent variable is categorical in nature.

e. Meaning of each of the column

|  |  |
| --- | --- |
| **'months\_as\_customer'** | Number of months the claimant was customer in the insurance agency |
| **'age’** | Age of the claimant |
| **'policy\_number'** | Number of the claimant applying for the policy |
| **'policy\_bind\_date'** | Date on which coverage got into place |
| **'policy\_state'** | State where the insurance policy centre is located |
| **'policy\_deductable'** | Deductible on the policy |
| **'policy\_annual\_premium'** | Annual premium of the policy |
| **'umbrella\_limit'** | Umbrella limit of the policy |
| **'insured\_zip'** | Zip code of the claimant |
| **'insured\_sex'** | gender of the claimant |
| **'insured\_education\_level'** | Education level of the claimant |
| **'insured\_occupation'** | occupation of the claimant |
| **'insured\_hobbies'** | Hobbies of the claimant |
| **'insured\_relationship'** | relationship of the claimant |
| **'capital-gains'** | Capital gain of the claimant of the claimant |
| **'capital-loss'** | Capital loss of the claimant of the claimant |
| **'incident\_date'** | Incident date of the accident/theft of the claimant |
| **'incident\_type'** | Incident type of the accident/theft of the claimant |
| **'collision\_type'** | collision type of the accident of the claimant |
| **'incident\_severity'** | Severity of the accident of the claimant |
| **'authorities\_contacted'** | Authoritites that were contacted when the incident occured |
| **'incident\_state'** | The state where the incident occured |
| **'incident\_city'** | The city where the incident occured |
| **'incident\_location'** | The location where the incident occured |
| **'incident\_hour\_of\_the\_day'** | The hour when the incident occured |
| **'number\_of\_vehicles\_involved'** | The number of vehicles involved in the vehicles |
| **'property\_damage'** | The damage to the property done during the incident |
| **'bodily\_injuries'** | The bodily injuries occurred during the incident |
| **'witnesses'** | The witnesses that were present during the incident |
| **'police\_report\_available'** | If police report was available of the incident |
| **'total\_claim\_amount'** | The total amount to be claimed by the claimant |
| **'injury\_claim'** | The injury claim to be claimed by the claimant |
| **'property\_claim'** | The property claim to be claimed by the claimant |
| **'vehicle\_claim'** | The vehicle claim to be claimed by the claimant |
| **'auto\_make'** | The automobile make of the claimant |
| **'auto\_model'** | The automobile model of the claimant |
| **'auto\_year'** | The year the automobile was taken |
| **'fraud\_reported'** | If there was fraud done by the claimant |
| **'policy\_csl'** | predetermined limit for the combined total of the Bodily Injury Liability coverage and Property Damage Liability coverage per occurrence or accident |

**Statistical analysis of dataset**

1.There is no column with single unique values

2.Null values is not present as all of them have 1000 datas

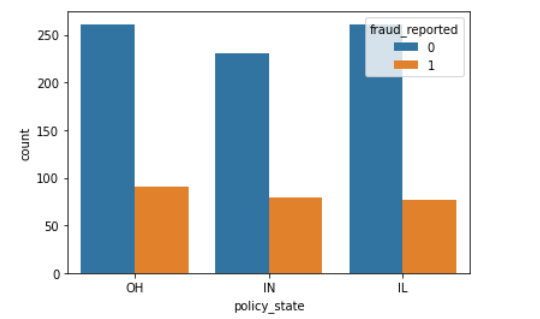
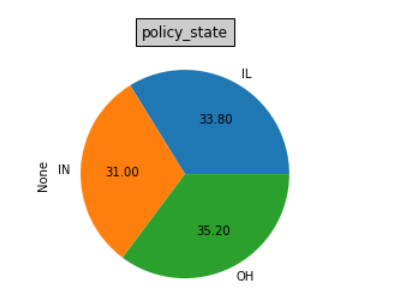
3.Outliers are seen in many columns as there is large difference between the mean and 50% data

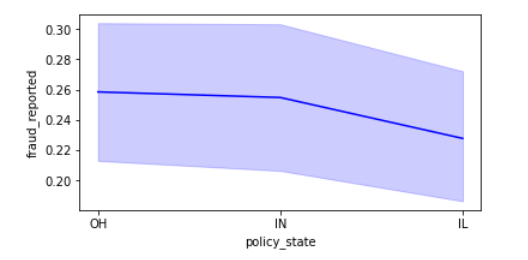
**Graphical analysis of dataset**

1. **policy state**

a.Highest frauds are seen in OH and least in IH

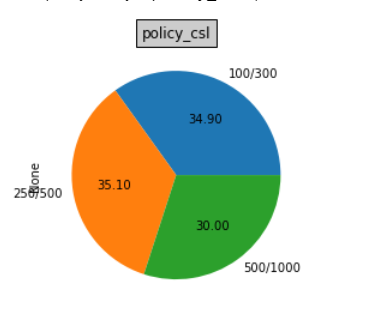
b.There almost equal customers from OH,IL and IN

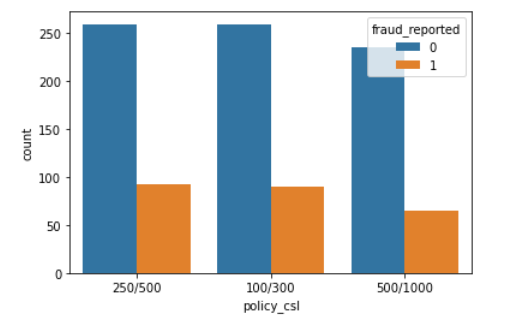
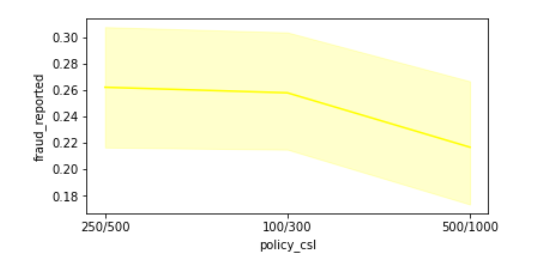




**2.policy csl**

Combined single limit (CSL):CSL is a single number that describes the predetermined limit for the combined total of the Bodily Injury Liability coverage and Property Damage Liability coverage per occurrence or accident

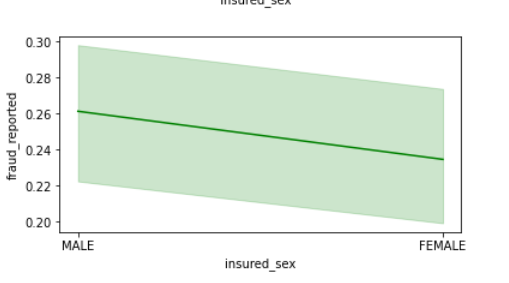
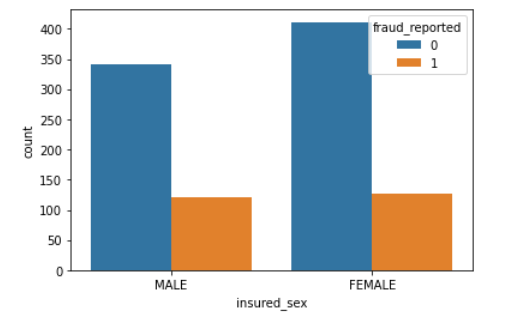
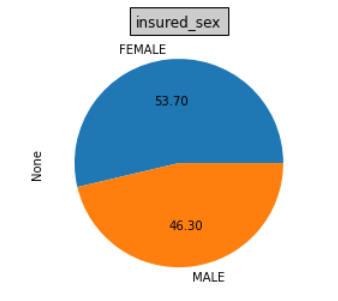


1.The CSL has options of 100/300, 500/1000 and 250/500, where the customers have chosen the 250/500 the most

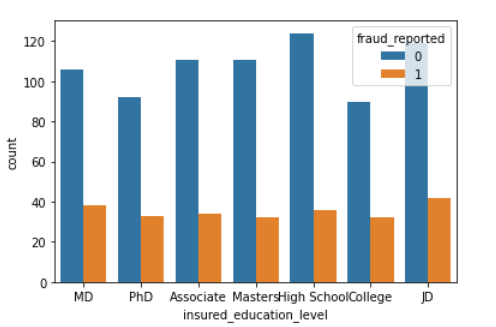
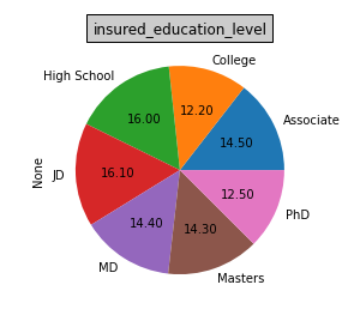
2.The 250/500 and 100/300 have shown higher frauds as compared to 100/300

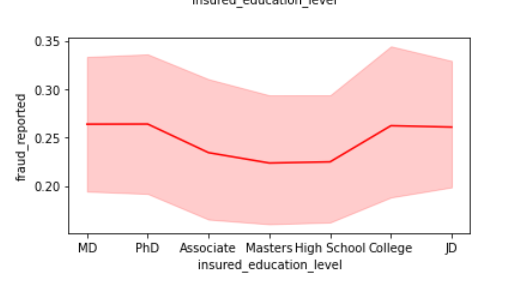
**3.Gender of the claimant**



1. There is larger number of females among the customers
2. Males have shown a higher number of frauds

**4.Education level of an insurer**

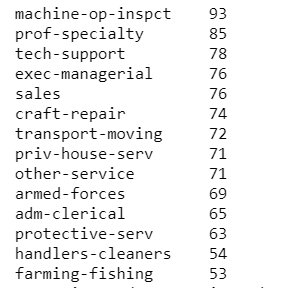
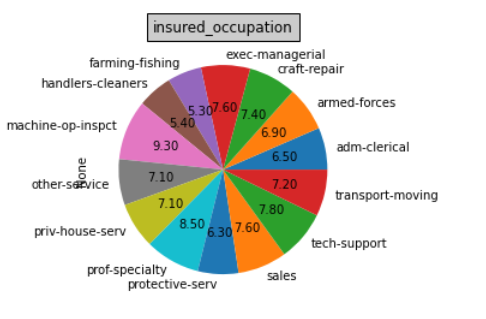




1.There is highest number of customers are high school graduates and None

2.Highest percentage of frauds are shown by JD, MD and High school

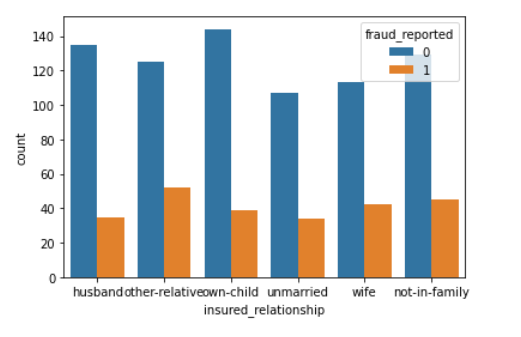
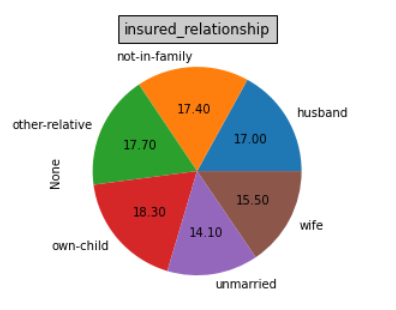
**5.Insurance claimant occupation**

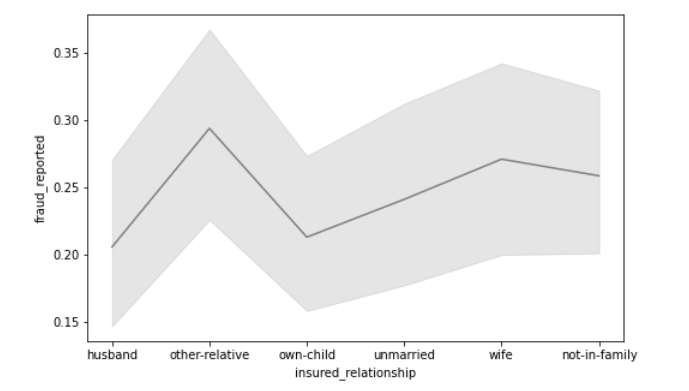
****



he largest number of customers are machine operating inspectors and the exec managerial have shown the highest fraud detection

**6. insurance claimants relationship**



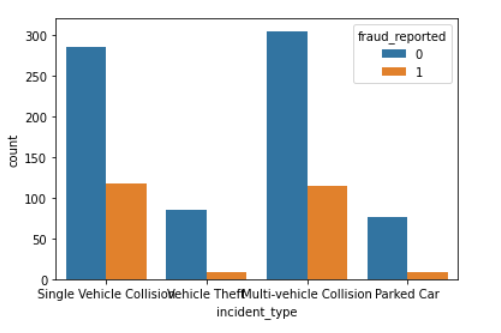
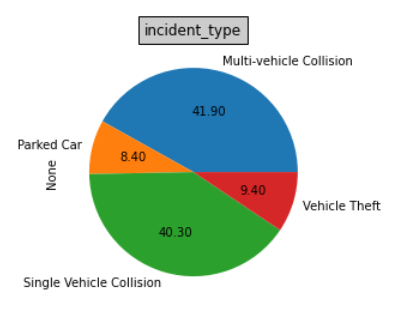


1.A code indicating the relationship of the patient to the identified insured is the insured relationship

2.The highest number of customers were having the insured relationship to be own child

3.The highest frauds were seen where the insured relationship was other relative

**7.Incident type**

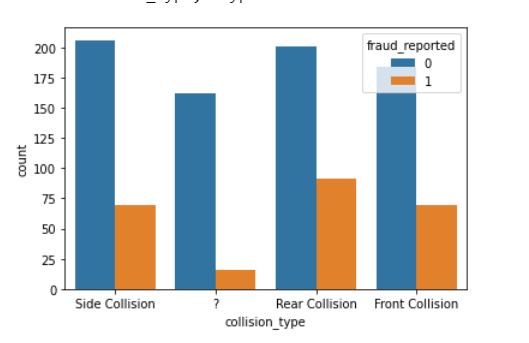
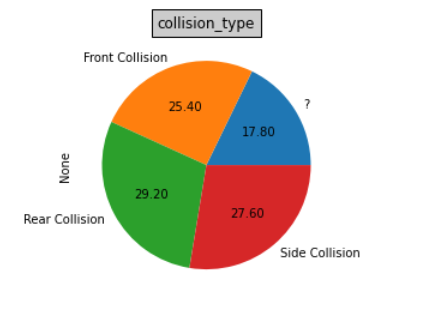


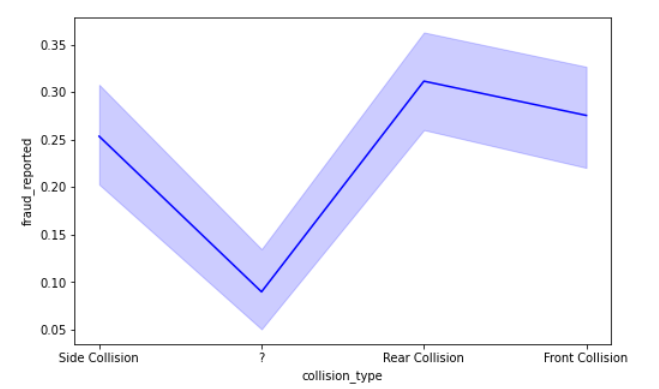


1.Highest number of customers faced Multi vehicle collision, followed by single vehicle collision

2.Single vehicle and multi vehicle collision have shown the highest Frauds reported

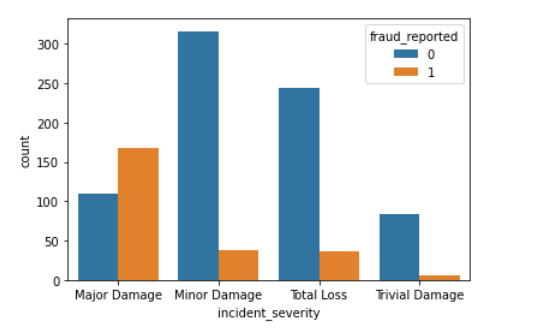
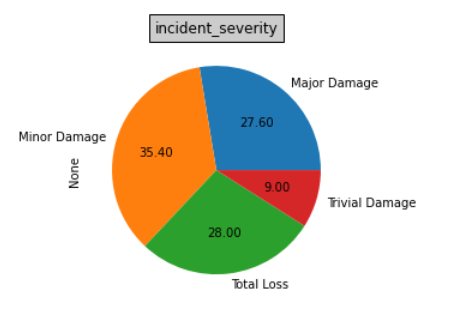
**8. collision type**

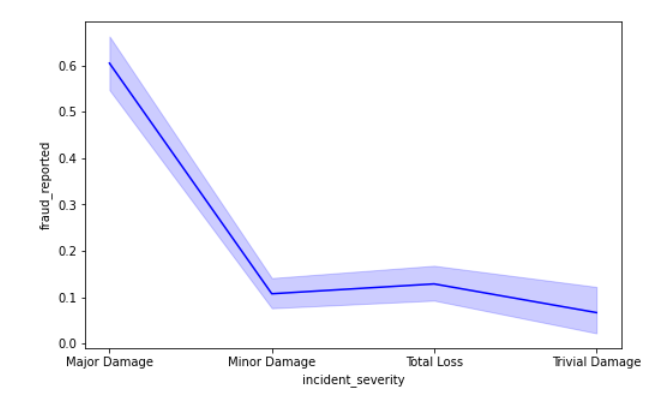




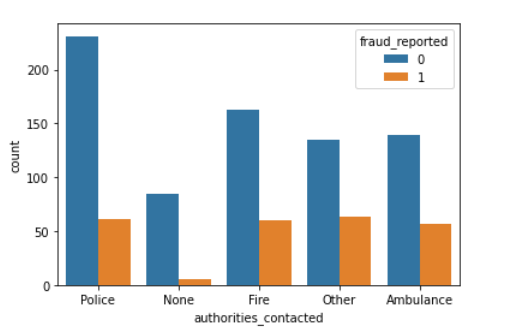
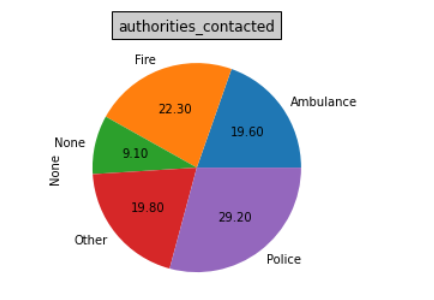
Rear and side collission has been the highest where the rear collission has shown the highest frauds

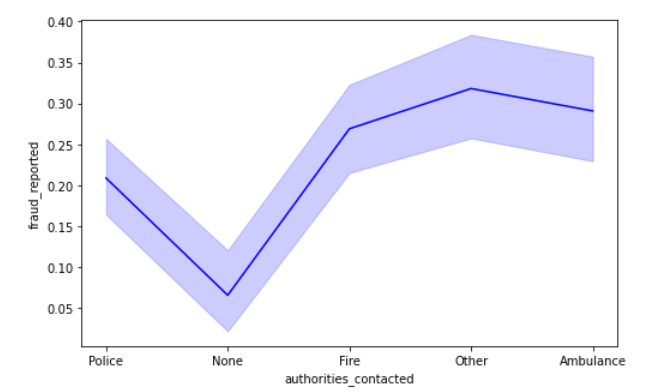
**9. incident severity**





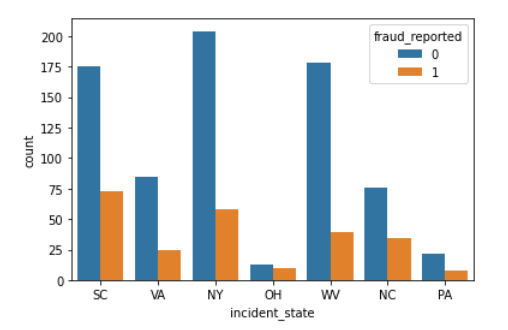
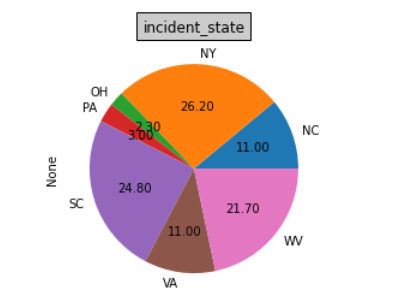
**10.Authorities contacted**

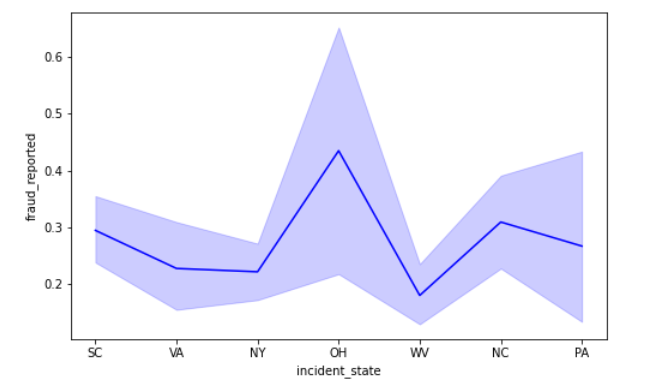




The highest frauds were reported by police, but the highest percentage of fraud among those reported were by OthersPolice reported cases shows lower frauds compared to those reported by police

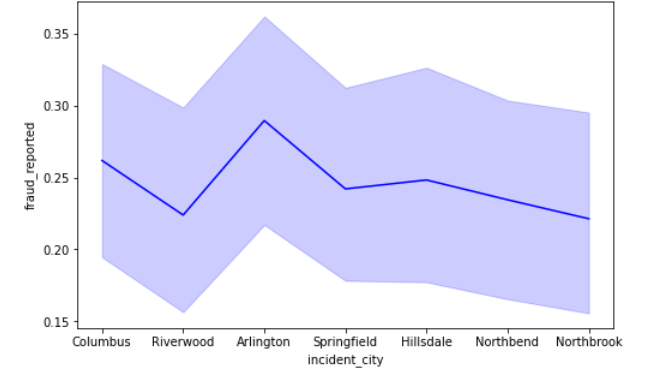
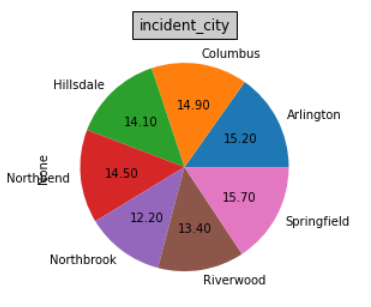
**11.Incident state**

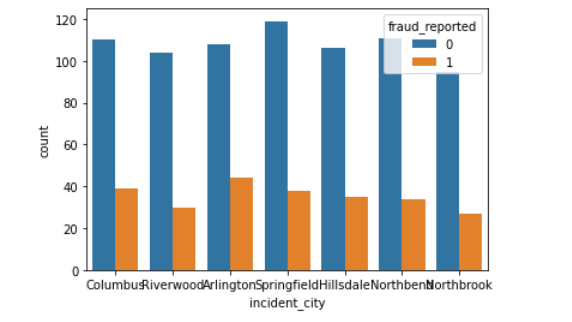




* Highest incident state was SC
* the highest fraud detected were from SC
* The highest percentage of those detected were from OH

**12.Incident city**

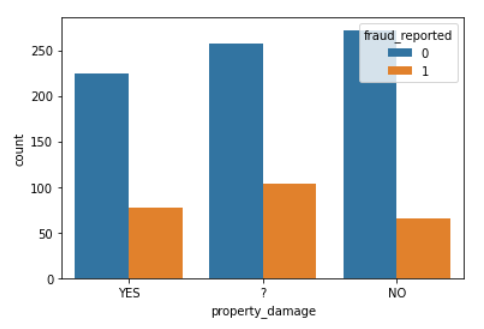
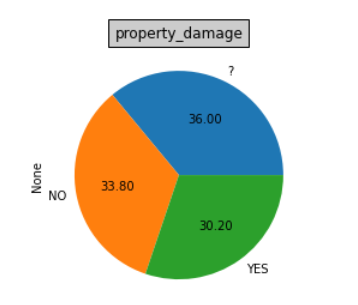


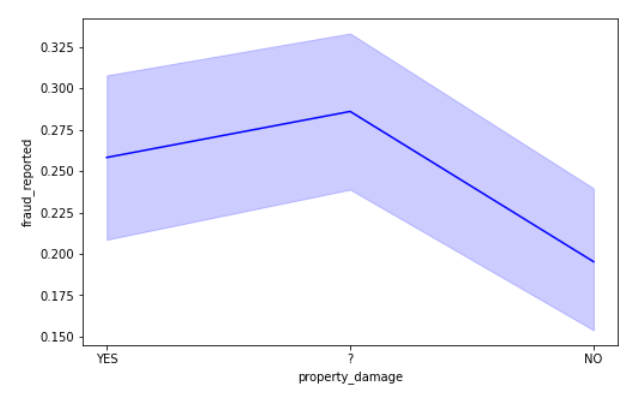


1. Largest number of customers were from Arlington and springfield

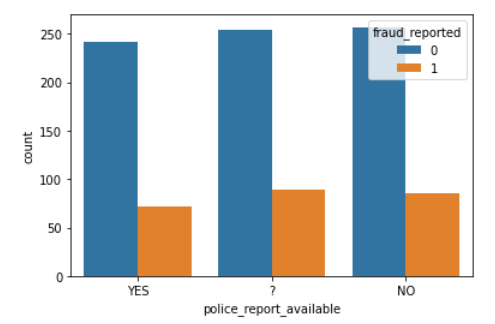
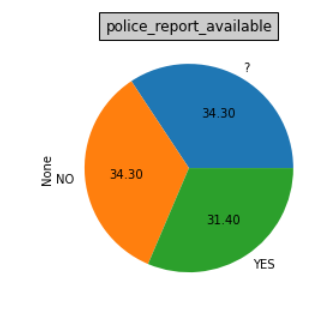
2.Largest percentage of frauds reported were from Arlington

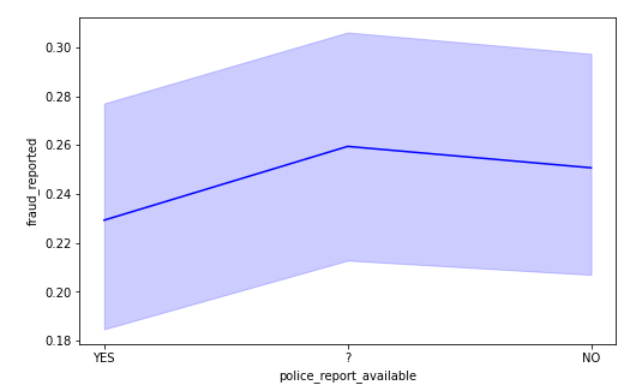
**13.Property Damage**



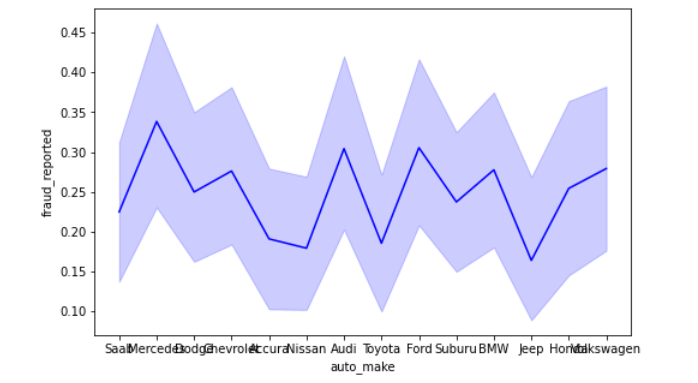
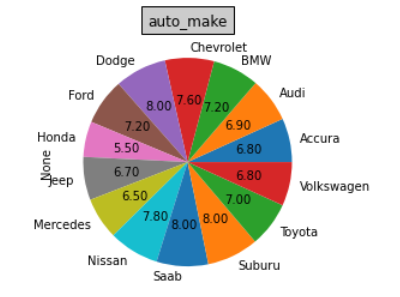


**14.POLICE REPORT**

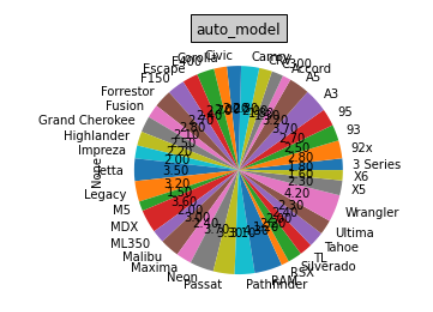


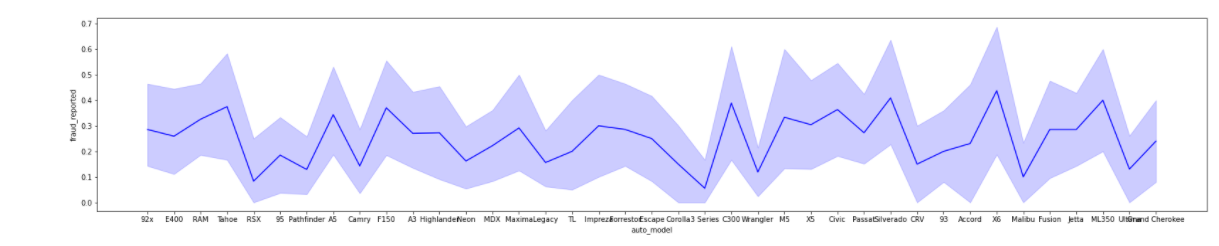


**15.AUTO MAKE**

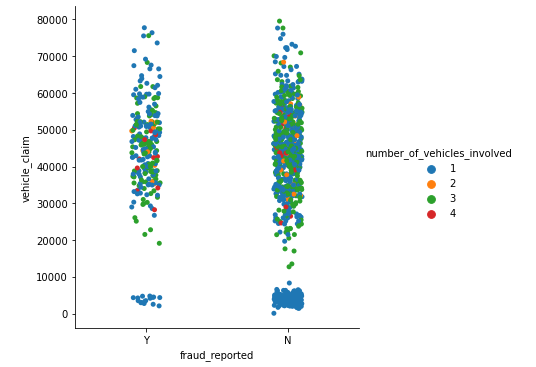


**16.AUTO MODEL**

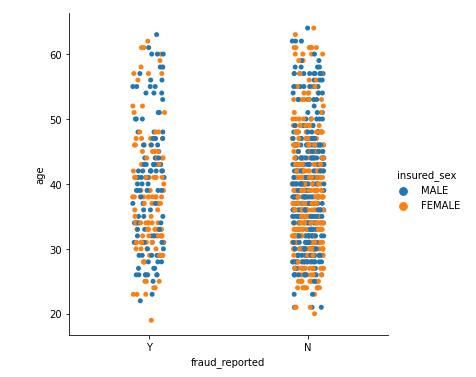




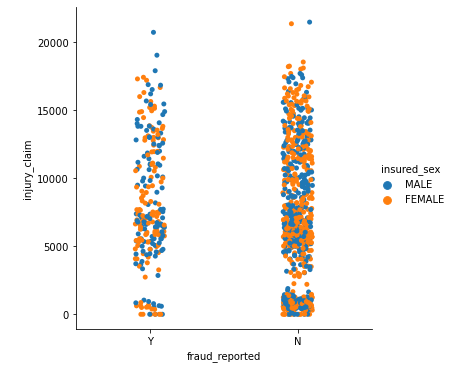
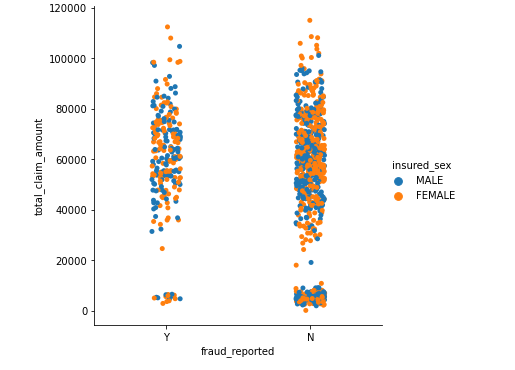
**17. various claims**

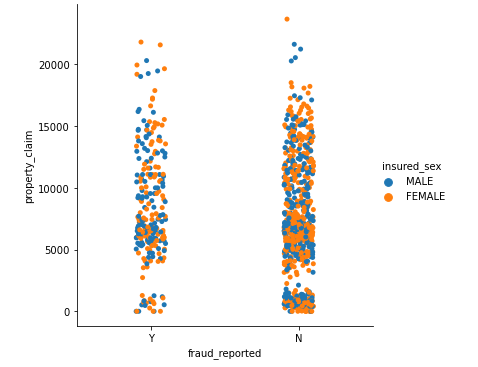


1 and 3 vehicles, **between rs3000 and rs6000** of vehicle claims have been detected for fraud



Both male and female, **between ages 30 to 48** have been highest to detect fraud

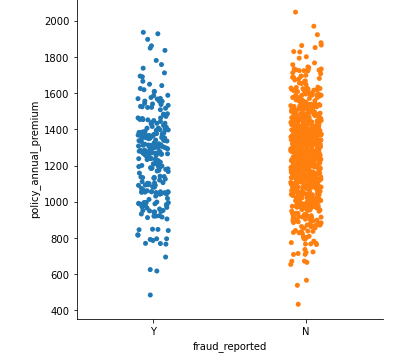




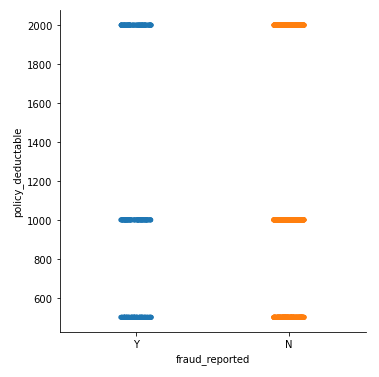
total claim has shown more between 4000 and 8000 is highest in frauds

injury claim has shown highest frauds between 4500 and 6000

property claim has shown highest frauds between 4500 and 6000



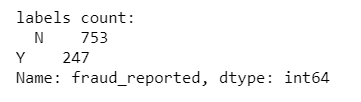
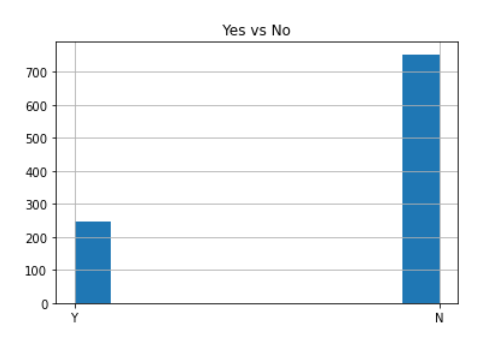
Policy annual premium is showing highest frauds between 1000 and 14000



The policy deductible doesn’t impact the frauds much.

**EDA Concluding Remark**

1.Dependent/label:



There were 247 frauds and 753 non-frauds. 24.7% of the data were frauds while 75.3% were non-fraudulent claims.

The label is imbalanced.

2.There is **no null values** present

3.There is columns which have large number of unique values, which need to be dropped

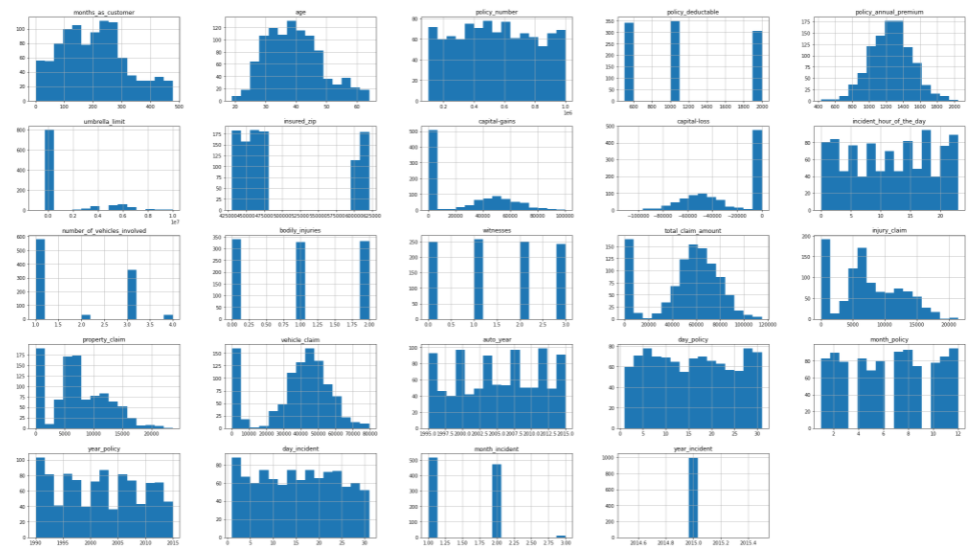
a. **policy\_bind\_date has 951 unique values**

**b. incident\_location has 1000 unique values**

**2.Removing the outliers using zscore**

The zscore is applied for standard deviation of 3. After the zscore was applied the data became more normalised

Dataset before applying zscore:

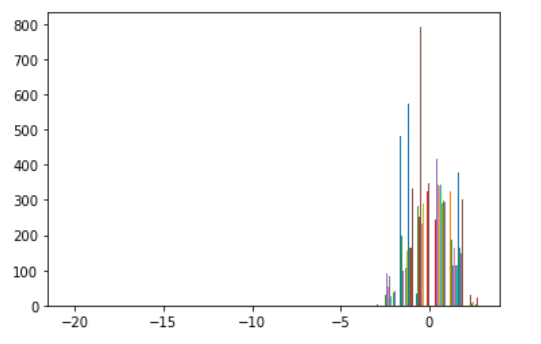


Dataset after applying zscore:



4.Used power transform to **remove skewness**

Histogram after using power transform

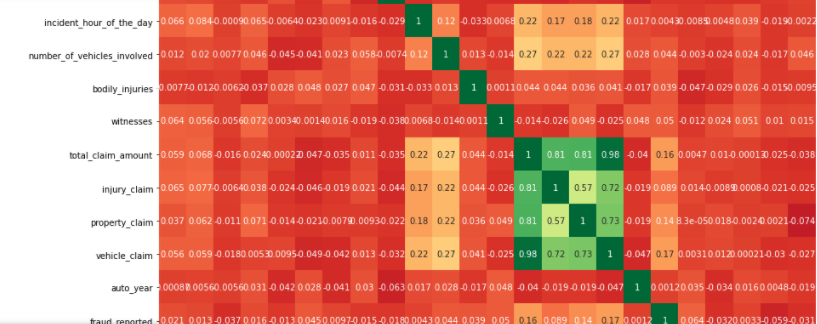
****

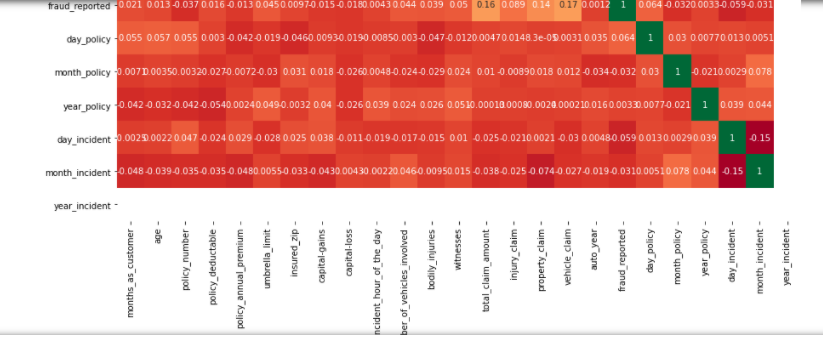
5.Data after EDA, was less skewed and more normalised

**Preprocessing Pipeline**

**feature engineering**







**Heatmap was plotted for variables**:

1.total claim amount and vehicle claim has high multicollinearity

2.month as customer and age has high multicollineairity

Apart from that, there don’t seem to be much correlations in the data.

There don’t seem to be multicollinearity problem except maybe that all the claims are all correlated, and somehow total claims have accounted for them.

However, the other claims provide some granularity that will not otherwise be captured by total claims. Thus, these variables were kept.

**Feature engineering**

1.The DV, fraud\_reported was label coded 1 for fraud and 0 for non-fraud.

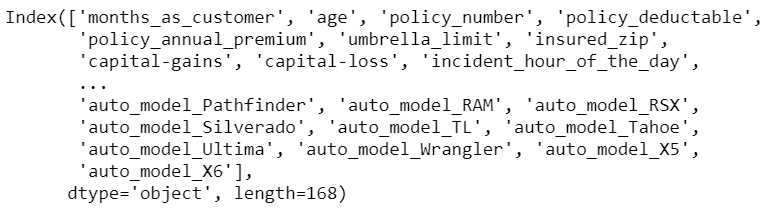
2.The collision type, property damage, police report available has ? in the data

3.the date will be broken down into more usable format

4.dropping the columns which have large number of unique values, policy\_state, incident\_location

5. \_c39 has full null values which will be dropped

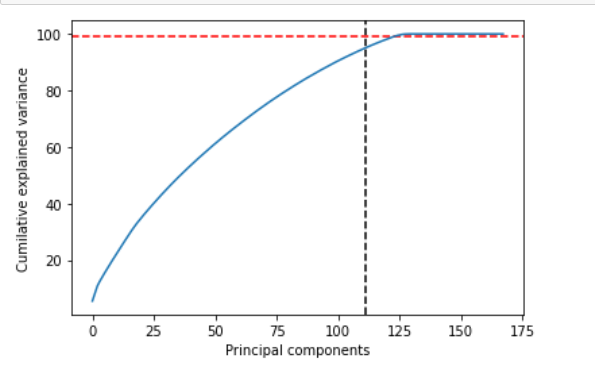
6.**one hot encoding** was applied on the independent variables, after which 168 columns were obtained with the some names as:



**Principal component analysis (PCA)**

1.The independent variables was standardised using scalar standard

1. Principal component analysis(PCA) was applied on the standardised data
2. Number of components explaining 95% variance is 111



1. The final data set of independent variables is made out of 111 columns

**Baseline Score**

As our dataset is imbalance, accuracy is not a good measure of success. A high **accuracy** can be achieved by a poor model that only selects the majority class, hence, not detecting and measuring the accuracy of classifying the class of interest. In fact, predicting only the majority class will give an accuracy of 75%, specificity of 100% but a sensitivity of 0%

Thus, **ROC AUC score** will also be used to measure how well we distinguish between Fraud and legit claims.

**Building machine learning model**

**Modeling**

**1.Split the data into test and train**

a.Both the dependent variable and independent variable was split into test and train data(75% train and 25% test)

b. Best sample state was found to be 39 and the data was split as per it.

2. Four different classifiers were used in this project:

- logistic regression

- K-nearest neighbours

- Random forest

- DecisionTreeClassifier

i.Accuracy score,f1 score,precision,classification matrix ,f1 score and recall was obtained for each of it:

|  |  |
| --- | --- |
| **a.Logistic regression** |  |
| **b.Decision tree classifier** |  |
| **c.RandomForestClassifier** |  |
| **d. SVC** |  |

**ii.Cross validation score of each of the algorithm:**

|  |  |
| --- | --- |
| Logistic Regression | 0.8091 |
| Decision Tree Model | 0.6653 |
| Random forest Model | 0.7571 |
| SVC Model | 0.7744 |

**iii.ROC AUC Score**

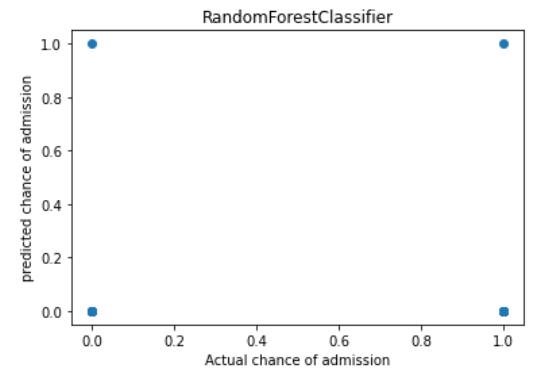
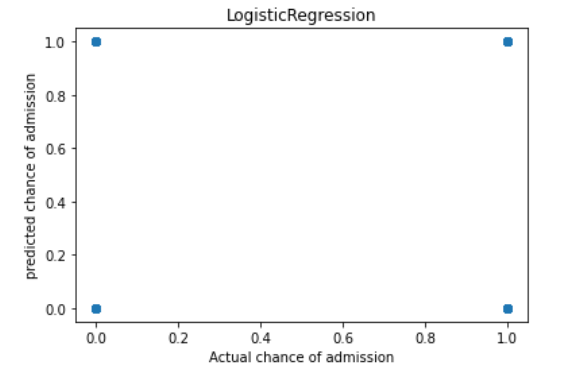
The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’.

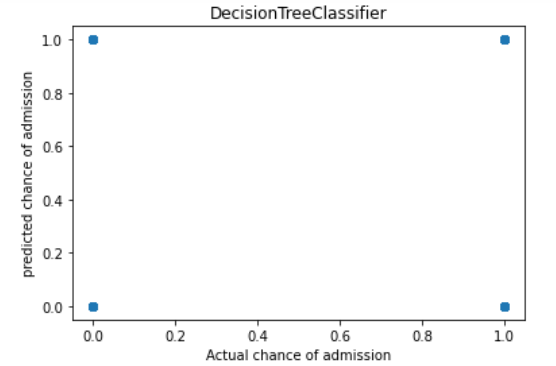
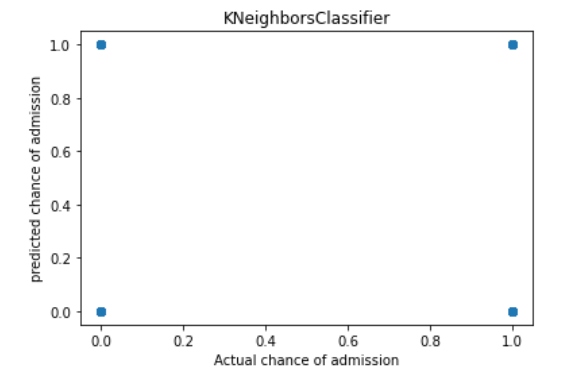
|  |  |
| --- | --- |
| Logistic Regression | 0.7521205974552831 |
| Decision Tree Model | 0.5812281025262769 |
| Random forest Model | 0.554121335054398 |
| SVC Model | 0.5124008851189379 |

**iv.ROC AUC curve**

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



**3.Chose Logistic regression for hyper parameter tuning**

Reason for choosing Logistic regression model:

1.Least difference is present between the accuracy score and cv score

2.The ROC-AUC score is higher for logistic regression

**4. Hyperparameter tuning on Logistic regression**

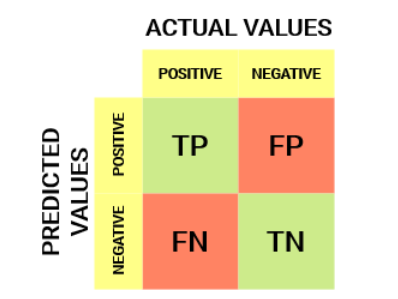
* Hyperparameter tuning and selection was done for Logistic Regression Model using GridSearchCV.
* After a 6-fold GridSearchCV, the model with its selected hyperparameters were fitted on the training set
* Model was created using the best parameters and saved as a pickle file

**5.Model scores obtained are:**

|  |  |
| --- | --- |
| Cross validation score | 81.02 |
| Accuracy score | 86.12 |
| ROC AUC score | 75.44 |

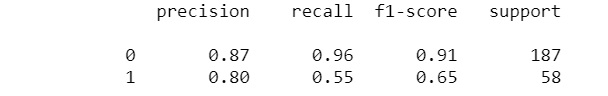
**6.Summary of Classification scores**

Classification table :

The summary of the classification report is presented below.

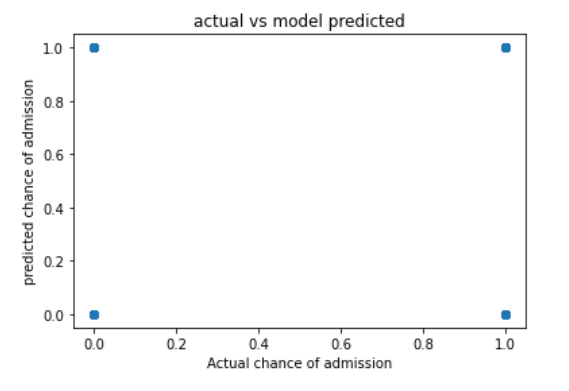




|  |  |  |  |
| --- | --- | --- | --- |
|  | **Formula** | **Definition** | **Value** |
| **Sensitivity (recall of fraud cases)** | derived from:  True positive/(True positive + False negative) | Sensitivity summarizes our true positive rate, which is how many we got correct out of all the positive cases. | 55% |
| **Specificity (recall of non-fraud cases)** | True negative/(True negative + False positive) | Specificity summarizes our true negative rate, which is how many we got correct out of all the negative cases. | 96% |
| **Precision of fraud cases** | True positive/(True positive + False positive) | Precision of fraud cases summarize the accuracy of fraud cases detected. That is, out of all that I predicted as fraud, how many are correct. | 80% |
| **Precision of non-fraud cases** | True negative/(True negative + False negative) | Precision of non-fraud cases summarize the accuracy of non-fraud cases detected. That is, out of all that I predicted as non-fraud, how many are correct. | 85% |
| **F1 scores** | (2 x recall x precision)\(recall + precision) | As we are interested in fraud cases, only the F1 scores on fraud cases are reported. | 65% |

**5.ROC AUC Curve of final model**

The ROC curve below summarizes how well our model is at balancing between the true positive rate(sensitivity) and the false positive rate(1-specificity). Ideally, we want to have a 100% true positive rate of predicting fraud and a 100% true negative rate of predicting non-frauds (or a 0% false positive which is 100% — 100% true negative rate). This means we have perfect prediction for both classes. However, in imbalance class problems, this is extremely hard to achieve in the real world. On top of that, there is a trade of between the true positive rate and the true negative rate and conversely the false positive rate.



**CONCLUDING REMARKS**

This project has built a model that can detect auto insurance fraud. In doing so, the model can reduces loses for insurance companies. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

Five different classifiers were used in this project: - logistic regression, K-nearest neighbours, Random forest and DecisionTreeClassifier. The best model for the dataset chosen is Logistic regression model. The Model can predict with an accuracy of 86.12%, if the claimant is fraud or not.

The parameters that influence if the claimant is fraud or not are vehicle claim,total claim amount and property claim among numerical values.

Among categorical variables it shows that highest frauds are seen in 0H

1. with 250/500 and 100/300 CSL
2. among males
3. among high school graduates
4. among exec managerial
5. who have hobbies chess and cross fit
6. where the insured relationship was other relative
7. among Single vehicle and multi vehicle collision
8. having rear collision
9. customers with major damages
10. other authorities were contacted and not police
11. by customers with Saab
12. by X6 auto model

The advantage of the prediction model is that, it makes it easier for the Automobile fraud company to not go through police to check if fraud is present and also not give a bad experience to the customers by questioning them unnecessarily. Hence it increases the customer satisfaction for the company and also doesn’t need to spend the time and resources on Police cases.

However the study having its own limitation is small sample size, restriction to few location and also in small area has its inability to easily adapt to every area and every time period.

**Github link:**

https://github.com/Akashpandey268/DataTrained-Evaluation-Projects/blob/main/Week%203%20Projects/Evaluation%20Project%20-%209/Automobile%20Insurance%20fraud-Evaluation%20Project%209.ipynb

**Reference:**

<https://www.policybazaar.com/motor-insurance/car-insurance/articles/fraud-car-claims-and-auto-insurance-fraud-in-india/>

<https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc>