Insurance Fraud Detection with Machine Learning Models

horizontal line

# 

# Introduction

**Problem Definition:**

Insurance fraud is defined as a deception which is deliberately committed against/by a claimant/ an insurance company or agent with the motive of financial gain. At any step of a transaction or application the fraud may be committed. Inflating claims, distorting and/or misrepresenting facts on an insurance application, submitting claims for damages that may have never occurred as well as staging accidents are some of the most commonly occurring frauds.

Fraudulent Insurance claims have been a huge problem in the industry for well over a century. Over the due course of time it has been getting progressively difficult to identify fraud claims. According to the Federal Bureau of Investigation, over $40 Billion every year is the cost of insurance fraud transactions.

Machine Learning is a boon in helping the Auto Insurance industry with this problem.

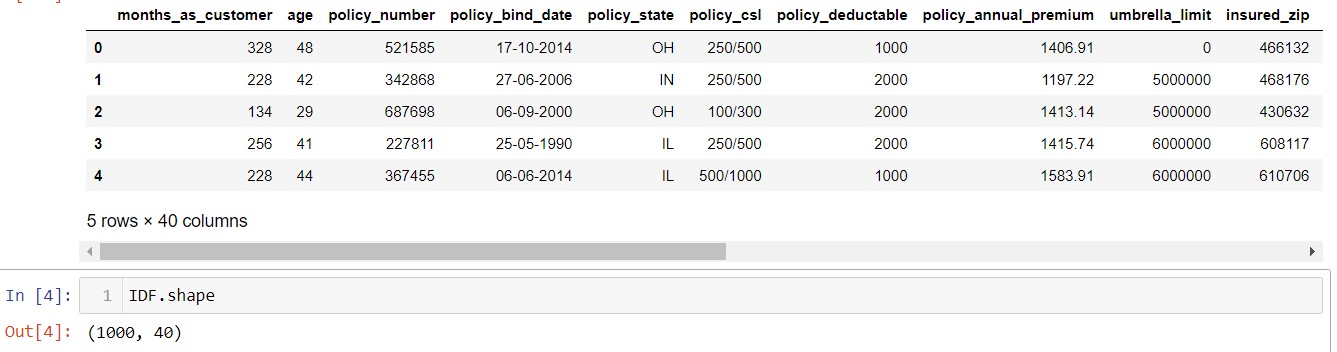
**Executive Summary:**

In this project, a dataset was provided with the details of the insurance policy along with the customer details, as well as details of the accident on the basis of which the claims have been made.

The Dataset was first cleaned, the various feature columns were analysed, then with feature engineering and based on strength of correlation and ANOVA f-score values, the feature columns were selected that would best predict the Target variable, to train and test machine learning models.

The auto insurance dataset was worked with to build a predictive model that best predicts if an insurance claim is fraudulent or not. Several models were trained and fitted with a part of the dataset and then tested with a different part of the dataset. The model that performed the best with the best confusion matrix performance, f1 score, ROC-AUC score and cross validation performance was then selected and tuned further with hyper parameter tuning techniques.

**About the Dataset:**

****

The given dataset consists of 40 columns and 1000 rows.

**The Independent Feature columns are:**

**months\_as\_customer:** Number of months for which the person has been a customer

**age:**  Age of Customer

**policy\_number**: Identification number of policy

**policy\_bind\_date**: Time period between effective date of coverage and policy issuance.

**policy\_state**: State where policy is active

**policy\_csl:**  Policy Combined single limit

**policy\_deductable:** Amount paid before the insurance company starts paying up.

**Policy\_annual\_premium**: The total amount of premium paid annually

**Umbrella\_limit:** Provides excess limits and gives additional excess coverage

**Insured\_zip:** Zip Code of the Insured address

**insured\_sex :** Gender

**Insured\_education\_level:** Education Background of Insured

**Insured\_occupation:** Occupation of Insured

**Insured\_hobbies:** Hobbies of the Insured

**Insured\_relationship:** Relationship of the Insured

**Capital-gains:** Capital Gains made from insurance

**Capital-loss:** Capital Loss incurred

**Incident\_date:** Date on which Incident Occured

**incident\_type:** Type of Incident

**Collision\_type:** Type of collision

**incident\_severity:** Severity of Incident

**Authorities\_contacted:** Whether authorities were contacted

**Incident\_state:** State where incident occurred

**incident\_city:** City where incident occurred

**incident\_location:** Location of incident

**Incident\_hour\_of\_the\_day:** Time of the day when incident occurred

**number\_of\_vehicles\_involved:** Number of vehicles involved in incident.

**property\_damage:** Whether there was property damage or not

**Bodily\_injuries:** Severity of bodily injuries

**witnesses:** Number of Witnesses

**Police\_report\_available:** Whether police reports are available

**Total\_claim\_amount:** Total amount of claim

**Injury\_claim:** Injury Claim amount

**Property\_claim:** Property Claim amount

**vehicle\_claim:** Vehicle Claim amount

**Auto\_make:** Make of Vehicle

**Auto\_model:** Model of Vehicle

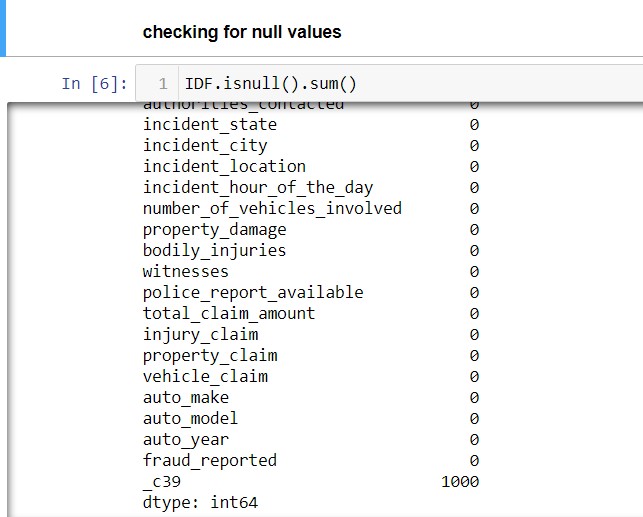
**Auto\_year:** Year of Vehicle Manufacture

**The Target Variable to predict is given in the column:**

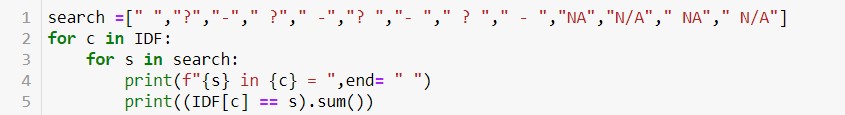
**Fraud\_reported:** Whether fraud was reported as Yes or No

**Data Cleaning:**

Upon inspecting all the columns in the dataframe, it is observed that column \_c39 has no usable data present as they were all NaN values. Other columns appear to have no NaN.



**Checking for blank spaces, random characters in each column**

****

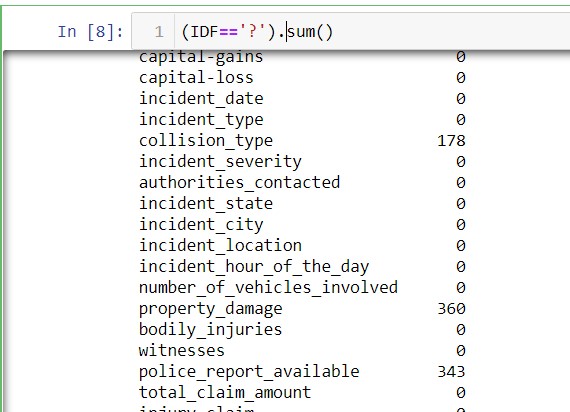
Upon inspecting all the columns, it is found that,

Column:

'police\_report\_available' contains 343 '?' character

'property\_damage'contains 360 '?' character

'collision\_type' contains 178 '?' character



**Therefore, Converting ' ?' to NaN values, which will later be imputed with values using various imputation techniques.**

****

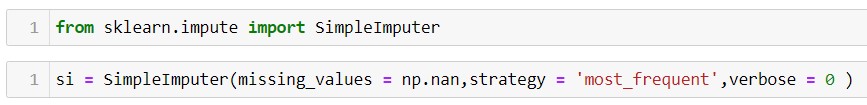
\_c39 has no usable data present. Other columns appear to have no null values. Therefore it will be dropped.

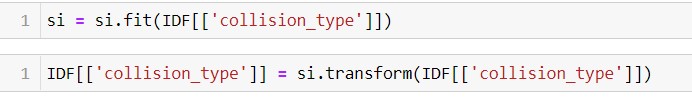
****

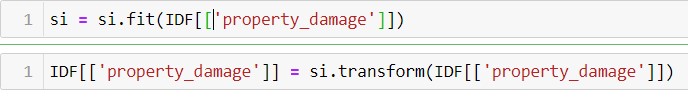
**Imputing Values to NaN values in Columns:‘collision\_type’,’property\_damage’,’police\_report\_available’**

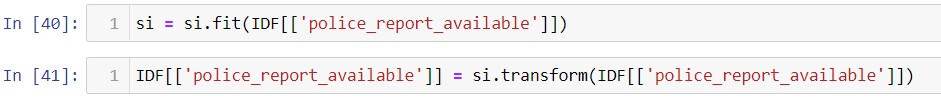
The most Frequently occurring value in each of the above columns was imputed to the NaN values of the respective columns of the Frequently occurring values.

**Using Simple Imputer**

****

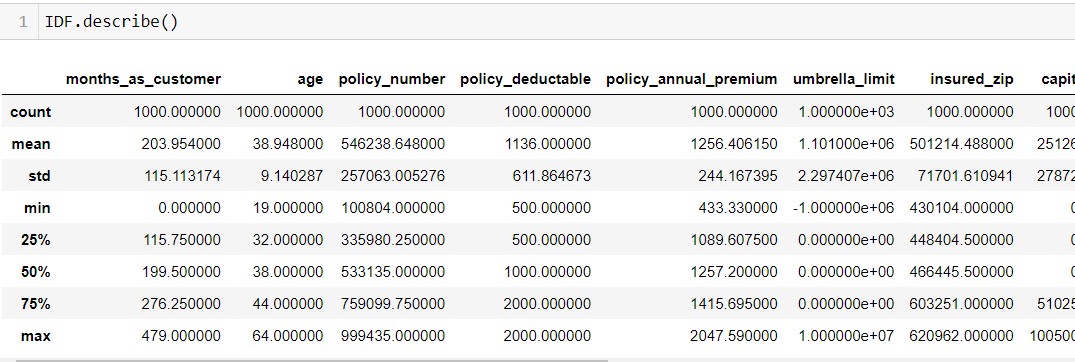
****

****

****

**After imputing the values it is observed that no more null values remain in the dataset.**

**Exploratory Data Analysis**

****

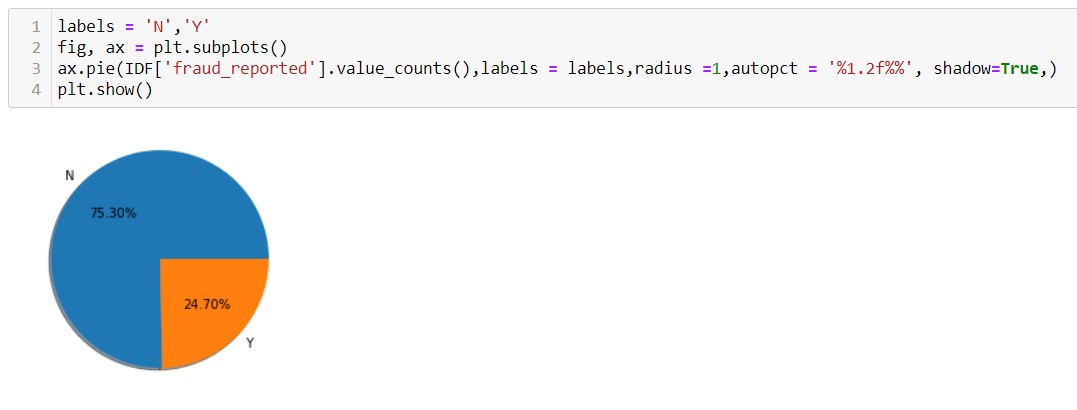
Difference in mean and 50% and considerable difference in 75% and max of columns months\_as\_customer,policy\_annual\_premium,capital-gains,total\_claim\_amount,injury\_claim and property\_claim suggests skewness in respective data distributions and presence of outliers.

### This is a Classification Problem since the Target variable / Label column ("fraud\_reported") has Categorical type of Data.

**Univariate Analysis**

**Analyzing the Target Class**

#### There are 2 unique categorical values in the Label column / target variable, viz. ‘Y’ and ‘N’.



**Class**

**'N' : Has 75.30% of total values**

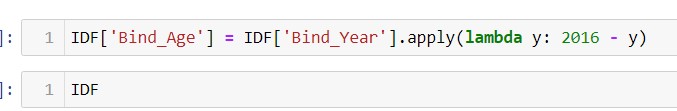
**'Y' : Has 24.70% of total values**

#### **Therefore, the Classes are imbalanced.**

**Feature Engineering**

#### **Extracting 'Bind Year' from policy\_bind\_date and storing it in 'Bind Year' column**

**Extracting 'Age' of policy bind from bind\_year to better help understand its relationship with Target Variable.**

****

**Extracting Age from auto\_year and storing it in 'Age ' column to better help understand its relationship with Target Variable.**

****

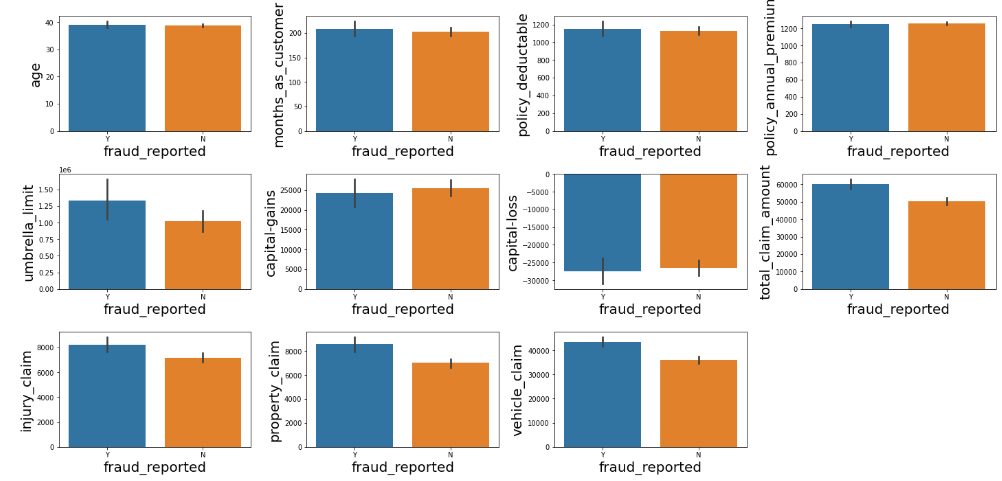
The columns 'policy\_bind\_date' and auto\_year are then dropped since it is no longer needed.

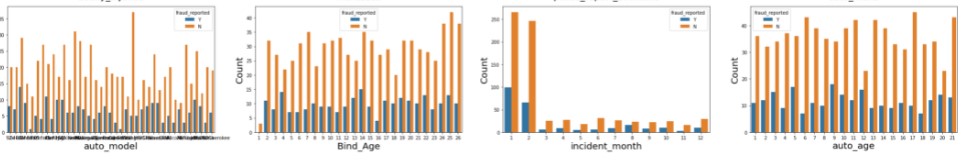
**Upon analyzing the rest of the Feature Columns, following observations are made:**

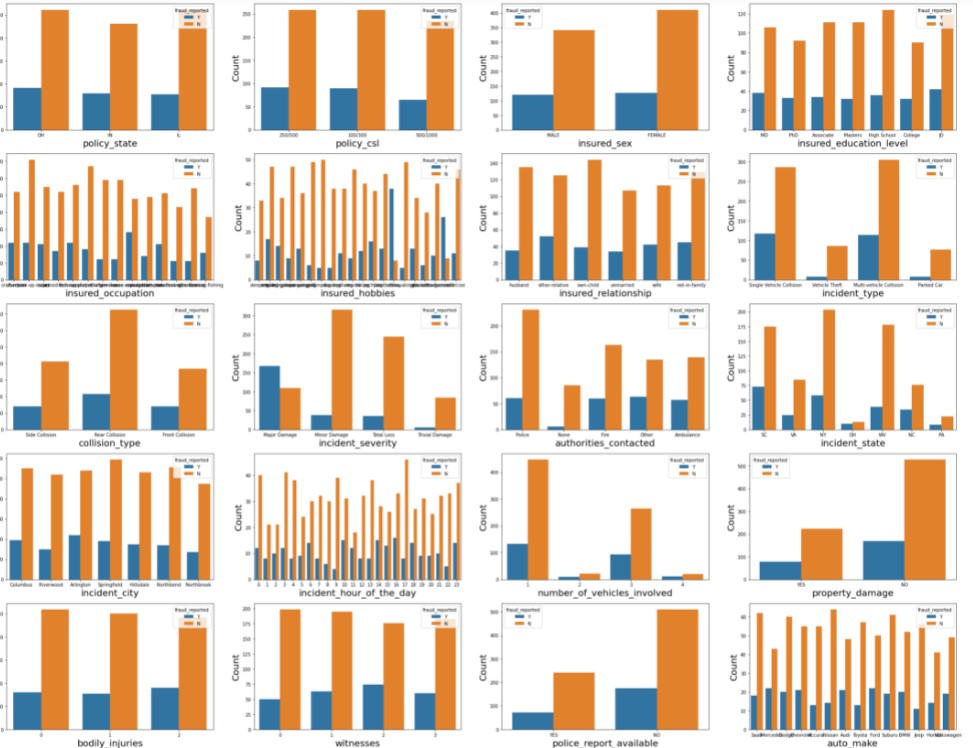
* Majority of the cases are Multi-vehicle Collision and Single Vehicle Collision and are Rear Collisions.
* Most incidents take place between January and February.
* Minor Damage is most common followed by Major Damage and Total loss.
* Most common authorities contacted were the Police followed by Fire force.
* Most of the incidents occured in states: SC,NY and WV
* Most incidents were reported from Columbus,Arlington, Springfield
* Majority reported no property damage.
* There are no police reports available for most cases.
* Most reports belong to models RAM,A3,Wrangler,Neon auto Models

**Bivariate Analysis**

### Interpreting Relationship between Dependent Variable and Independent Variables



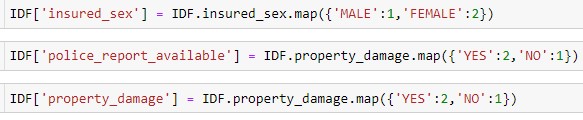


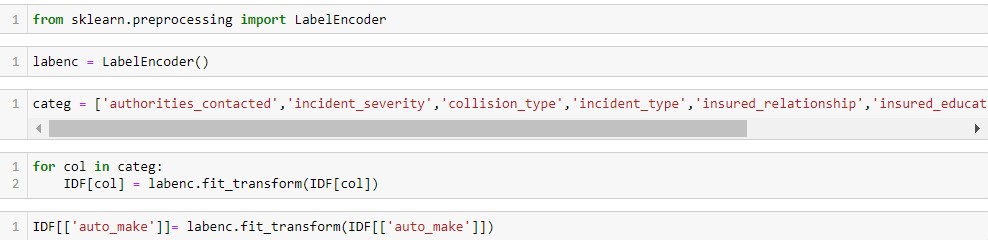


**Following observations can be made from above graphs:**

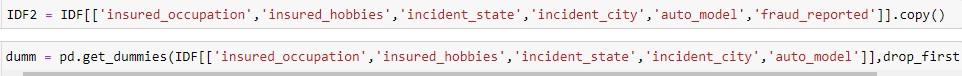
* 'age','months\_as\_customer','policy\_deductable','policy\_annual\_premium','capital-gains','capital-loss', don't seem to contribute to fraud probability.
* Higher the umbrella limit, more fraud claims are filed.
* Higher the total claim amount, more the fraud claims are filed.
* Higher the injury claim amount, more the fraud claims are filed.
* Higher the property claim amount, more the fraud claims are filed.
* Higher the vehicle claim amount, more the fraud claims are filed.
* policy state,policy csl,insured sex,authorities contacted,bodily injuries,incident city, witnesses don't seem to contribute to fraud probability.
* Education levels of JD and Highschool and MD contribute most to the fraud claims filed.
* Relationships - other relative and not in family contribute most to the fraud claims filed.
* Single vehicle collision and multi vehicle collision contribute most to the fraud claims filed.
* Incidents in states SC and NY contribute most to the fraud claims filed.
* fraud claims are more for 1 and 3 vehicles involved in accident
* fraud claims are more for rear collision in accident
* fraud claims are most for Major damage reported
* fraud claims are most for hours 10,14,16,18(office rush hours) and 23 of the day
* fraud claims are more when no property damage is reported
* fraud claims are more when no police report is available
* fraud claims are more during months 1(january) and 2(february).
* fraud claims are policy bind ages 2,4,13 and 14
* fraud claims are most for car age 3,5,9,12,20,21.
* Ram,A5,Jetta,ML350,Passat,F150,A3 have the highest fraud insurance claims, while 3 series,RSX,Camry have the lowest.
* Wrangler,Passat,95,Neon,Malibu,Grand Cherokee,auto\_model\_Ultima,Corolla,TL,Legacy have the highest legitimate claims.
* Mercedes,Dodge,Chevrolet,Audi,Ford,Volkswagen have the highest fraud insurance claims.
* Most fraud reports were filed by exec managerial,Transport moving and Craft repair.
* Most fraud claimants have chess and cross fit as hobbies

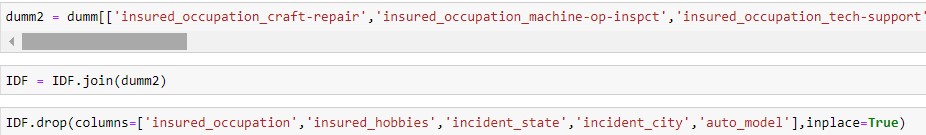
**Before Proceeding with finding the correlations of the columns, The data of the categorical columns needs to be encoded.**

****

****

Encoding the rest of the columns using get\_dummies and retaining only the most important features in the dataframe.

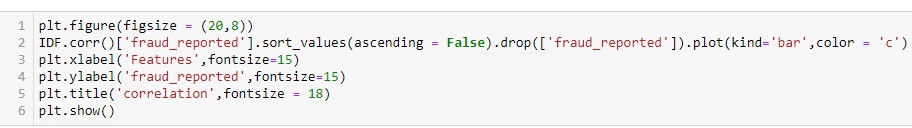


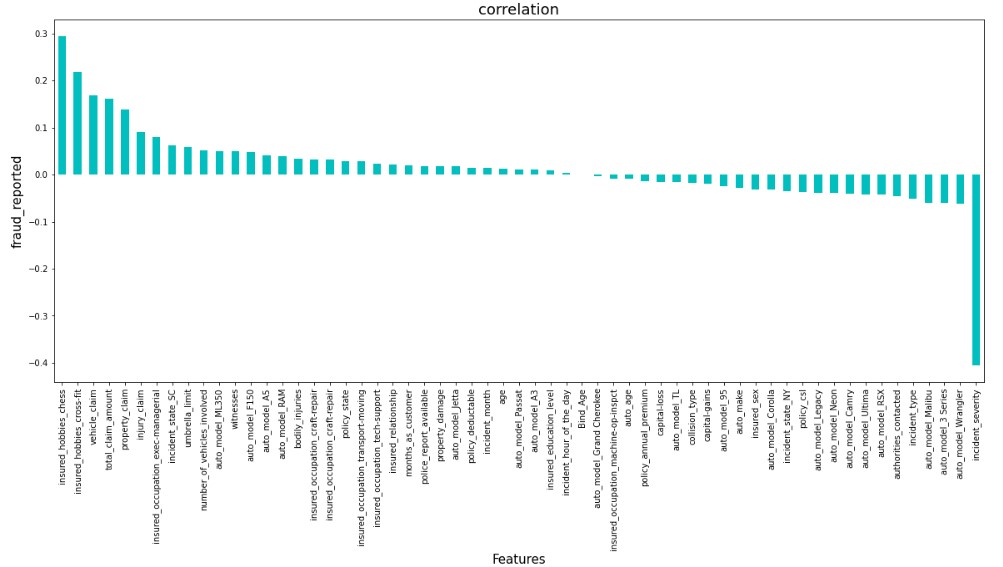


**Encoding the Label Column**

****

**Finding the correlations**

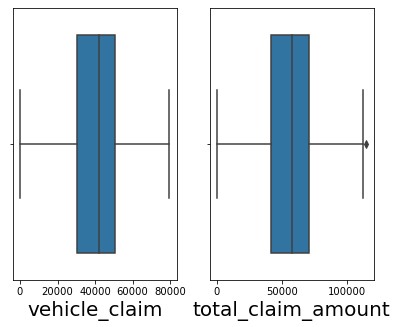
****

****

Incident\_severity has the highest negative correlation with fraud\_reported while, insured\_hobbies\_chess,insured\_hobbies\_cross-fit,vehicle\_claim,total\_claim\_amount,property\_claim have the highest positive correlation with fraud\_reported.

**Data Pre-Processing**

**Checking for Outliers in columns with continuous distribution**



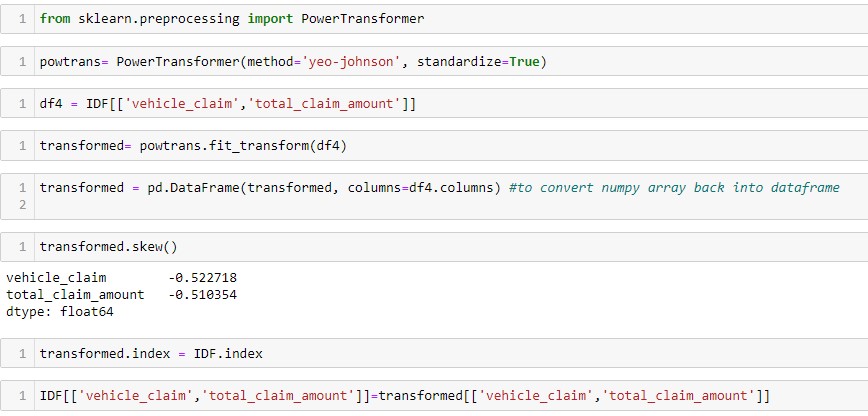
Upon observing the above plots, it is concluded that there no considerable outliers in the columns.

#### **Checking for Skewness in Data**



It is observed that vehicle\_claim,total\_claim\_amount are skewed.

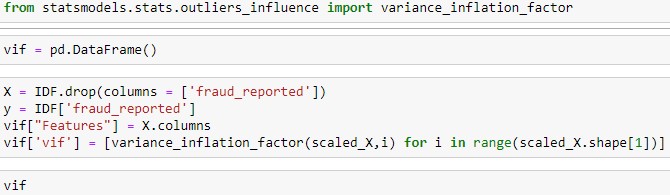
#### **Reducing skewness using PowerTransformer**

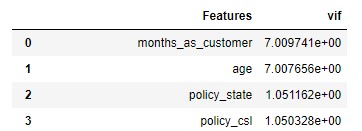


Skewness has been greatly reduced.

Next step is to select the best features which would build the most accurate Machine Learning Models to predict the target variable.

### Checking for Multicollinearity using Variance Inflation Factor

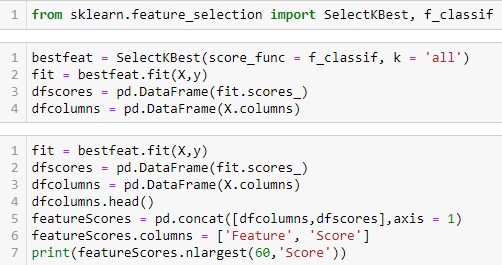
Variance inflation factor measures how much the variance of an independent variable is influenced / inflated, by its interaction/correlation with other independent variables. 

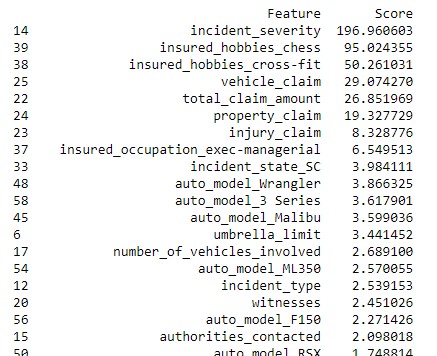


It is found that months\_as\_customer, age are highly multicollinear

### Selecting Kbest Features

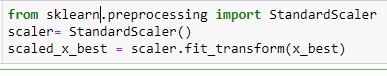
Based on the respective ANOVA f-score values, the feature columns are selected that would best predict the Target variable, to train and test machine learning models.





Upon analyzing the scores of each column, it is decided that the columns with the lowest scores will be dropped.

**Feature scaling**

**Scaling the values in the feature columns using StandardScaler inorder to normalize the range of data.**

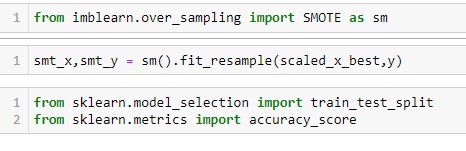
### 

### Balancing out classes in the Label column using SMOTE technique.

The classes in the target column are heavily imbalanced. Inorder to ensure that the precision and recall accuracies of the models for both of the classes are balanced, the classes of the target column need to be balanced.

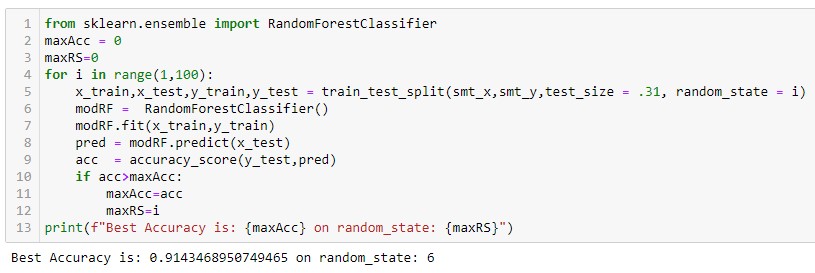
SMOTE technique will be implemented to balance them out.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.



**Classification Model Building**

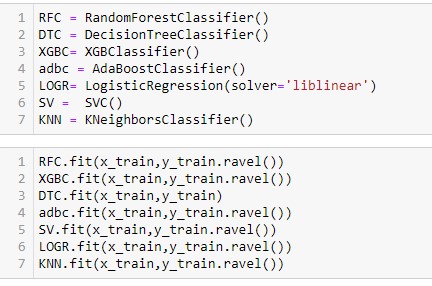
**Finding the Best Random State**



**Creating Train-Test split based on random state obtained above:**

****

### Training the Models

****

**Analyzing Model Accuracies**

### Logistic Regression Model Accuracy

The trained Logistic Regression Model shows

F1 score of 0.83

Roc\_auc score of 0.8329

Cross validation score of 0.8320

Sensitivity(Recall for ‘ Fraud’ cases) is 0.86 and Specificity (recall of non-fraud cases) is 0.81

Precision for ‘ Fraud’ cases is 0.82 and Precision for non-fraud cases is 0.85

### Random Forest Classifier Model Accuracy

The trained Random Forest Classifier Model shows

F1 score of 0.91

Roc\_auc score of 0.9057

Cross validation score of 0.8765

Sensitivity(Recall for ‘ Fraud’ cases) is 0.93 and Specificity (recall of non-fraud cases) is 0.88

Precision for ‘ Fraud’ cases is 0.89 and Precision for non-fraud cases is 0.93

**XGB Classifier Model Accuracy**

The trained XGB Classifier Model shows

F1 score of 0.91

Roc\_auc score of 0.9121

Cross validation score of 0.8745

Sensitivity(Recall for ‘ Fraud’ cases) is 0.94 and Specificity (recall of non-fraud cases) is 0.88

Precision for ‘ Fraud’ cases is 0.89 and Precision for non-fraud cases is 0.94

### AdaBoost Classifier Model Accuracy

The trained AdaBoost Classifier Model shows

F1 score of 0.89

Roc\_auc score of 0.8886

Cross validation score of 0.8500

Sensitivity(Recall for ‘ Fraud’ cases) is 0.87 and Specificity (Recall of non-fraud cases) is 0.91

Precision for ‘ Fraud’ cases is 0.91 and Precision for non-fraud cases is 0.87

### SV Classifier Model Accuracy

The trained SV Classifier Model shows

F1 score of 0.90

Roc\_auc score of 0.8971

Cross validation score of 0.8951

Sensitivity(Recall for ‘ Fraud’ cases) is 0.93 and Specificity (Recall of non-fraud cases) is 0.86

Precision for ‘ Fraud’ cases is 0.87 and Precision for non-fraud cases is 0.93

### K Nearest Neighbours Classifier Model Accuracy

The trained K Nearest Neighbours Classifier Model shows

F1 score of 0.73

Roc\_auc score of 0.7276

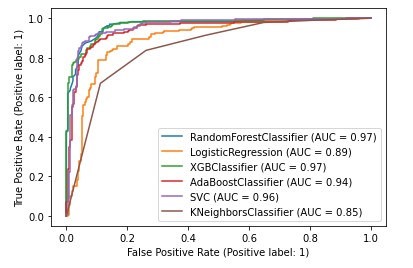
Cross validation score of 0.7377

Sensitivity(Recall for ‘ Fraud’ cases) is 0.91 and Specificity (Recall of non-fraud cases) is 0.54

Precision for ‘ Fraud’ cases is 0.67 and Precision for non-fraud cases is 0.86

### 

### ROC AUC curves



Random Forest Classifier and XGB Classifier have the largest area under their respective curves both scored 0.97 on AUC

Since , Sensitivity summarizes the true positive rate, ie. how many we got correct out of all the positive cases and Specificity summarizes our true negative rate, which is how many we got correct out of all the negative cases. The model that performs the best in those criteria will be chosen.

Both Random Forest Classifier and XGB Classifier have performed the best on fraud and non fraud detections based on the fact that their Sensitivity, Specificity and Precision scores are the highest amongst all the model performances.

### Therefore, Based on the above graph and roc\_auc\_scores,XGB Classifier is the best model for the dataset, with AUC = 0.97 and roc\_auc\_score = 0.9121[¶](http://localhost:8888/notebooks/InsuranceFraud_proj.ipynb#Based-on-the-above-graph-and-roc_auc_scores,XGB-Classifier-is-the-best-model-for-the-dataset,-with-AUC-=-0.97-and-roc_auc_score-=-0.9121)

### Hyper Parameter Tuning

GridSearchCV was used for Hyper Parameter Tuning of the XGB Classifier model.

Based on the input parameter values and after fitting the train datasets,

The XGB Model was further tuned based on the parameter values yielded from GridsearchCV.

The Tuned XGB Model displayed accuracy of 91.43%

F1 score of 0.91

Sensitivity(Recall for ‘ Fraud’ cases) is 0.91 and Specificity (recall of non-fraud cases) is 0.91

Precision for ‘ Fraud’ cases is 0.91 and Precision for non-fraud cases is 0.91

### Concluding Remarks

In conclusion, XGB Classifier Model is able to correctly distinguish between Fraud claims and legitimate claims with high accuracy.

The dataset had very limited data which is problematic as models show greater stability when the dataset is of a good size. However, a large set of feature columns enabled selecting a smaller feature size that provides the best accuracy for the model and obtaining results in optimal time.