**Insurance Claims-Fraud Detection**

In this article, I will be going through the process of building a machine learning models to predict if an insurance claim is fraudulent or not and it will take you through each and every steps in detail and helps you to understand the whole machine learning model building process.



**INTRODUCTION**

Auto insurance is a contract between you and the insurance company that protects you against financial loss in the event of an accident or theft. The auto insurance industry is complicated and involves millions of dollars changing hands every day. And whenever there is a large amount of money running through complex systems, there is opportunity for fraud**.**This fraud can be committed by professionals and company working in the industry. But it can also be committed against them. By reading this, first thing that we must know about what is insurance fraud.

What is insurance fraud?

Insurance fraud is an illegal act on the part of either the buyer or seller of an insurance contract. Insurance fraud involves any misuse of insurance policies or applications in order to illegally gain or benefit.

The insurance fraud can be broadly classified into 2 types.

* Soft Insurance Fraud
* Hard Insurance Fraud

**Soft Insurance Fraud:** Soft fraud is usually unplanned and arises when the opportunity presents itself. An example of this type of fraud would be getting into a car accident and claiming for your injuries that are worse than they really are, getting you a bigger settlement than you would get if you were telling the truth about your injuries.

**Hard Insurance Fraud:** Hard fraud is planning, scheming, and maybe even someone on the inside to help you to get money from an insurance company. An example of hard fraud would be getting into an accident on purpose so that you can claim the insurance money.

In this project we focuses on claim data of an Automobile insurance company. Because of the fraudulent claims the insurance companies are losing huge amounts of money, which indirectly affects the public. Therefore, it is important to know which claims are genuine and which are fraud.

In this article, we will walk through how to spot insurance fraud and the consequences of engaging in it by building machine learning models and getting prediction of which claims are likely to be fraudulent. This enables an insurer to detect more fraudulent claims.

1. **Problem Definition**

Insurance fraud 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.

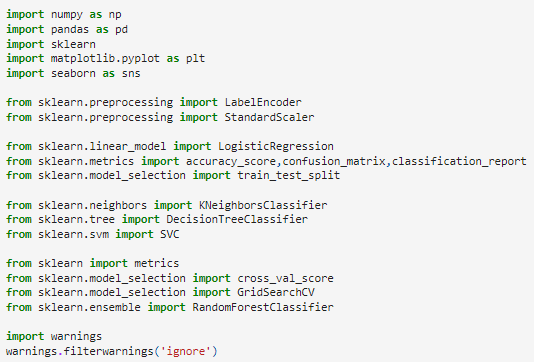
In this project, we are provided with a dataset which has the details of the insurance policy, customer details and also has the details of the accident on the basis of which claims have been made. we will be working with some auto insurance data to demonstrate how we can create a predictive model that predict if an insurance claim is fraudulent or not

The problem statement explains that the target variable is **Categorical**, so it is a **“Classification Problem”** where we need to predict whether an insurance claim is fraudulent or not.

1. **Data Analysis**

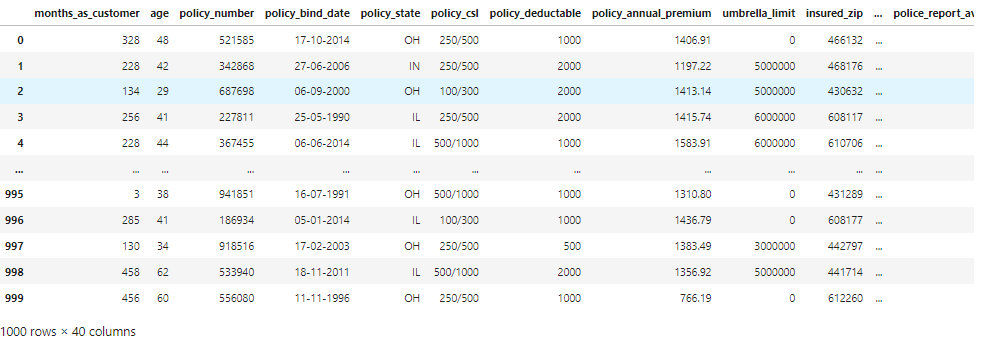
Data Analysis is the process of inspecting, cleaning, transforming and modelling data with the goal of discovering useful information.

**Importing necessary libraries and dataset**

****

****

The dataset that was given is CSV(comma-seperated values) format file, I imported that file using pandas.

****

The dataset contains 1000 rows and 40 columns. It is having both numerical and categorical data. We can observe the null values, some “?” signs and some of the columns are in date format (dd/mm/yy), so we need to perform lots of preprocessing, data cleaning to make our data perfect to build the Machine Learning models.

**Exploratory Data Analysis**

* Before entering into data preparation, first we need to check first 5 rows of the dataset, any sample row and last 5 rows of dataset using head, sample and tail method. This will help us to find any missing values, empty spaces or any special characters present in the dataset.
* After that we check our entire column names with columns function because we can’t go through entire column names in our dataset while importing and information of the dataset using info function.

**Checking Null Values**

****

With the help of isnull function and sum function we are able to see any null values present in the dataset. In our dataset column \_c39 is full of null values. So we can drop that column.

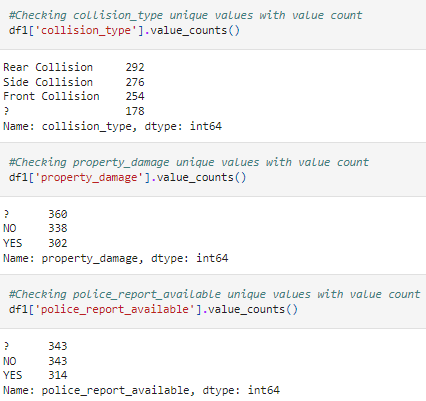


After dropping one column from the dataset now we are having 1000 rows and 39 columns.

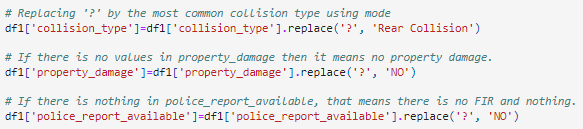
That we notice that our dataset have ‘?’ sign in our dataset, it does not shown while we checking null values. So we check it with isin([?]) function to view which columns having ? sign.



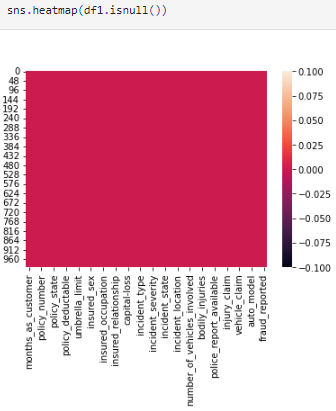
As a result of isin function we have output that 3 columns having ‘?’ sign collision\_type, property\_damage, police\_report\_available. Lets check value counts of these columns individually to replace ‘?’ sign.



As we see in the above code and result of value\_counts, we are able to see that most used collision\_type is Rear collision that have value 292, property\_damage have 338 for No count and police\_report\_available have 343 for No count, so we can replace it with mode values.



Now after replacing all the ‘?’ sign with mode values, again checking heatmap for checking Null values.



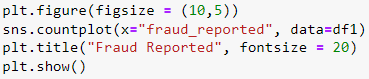
**Visualizations**

Data visualization is the graphical representation of the information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an easy way to see and understand trends, outliers, and patterns in data.

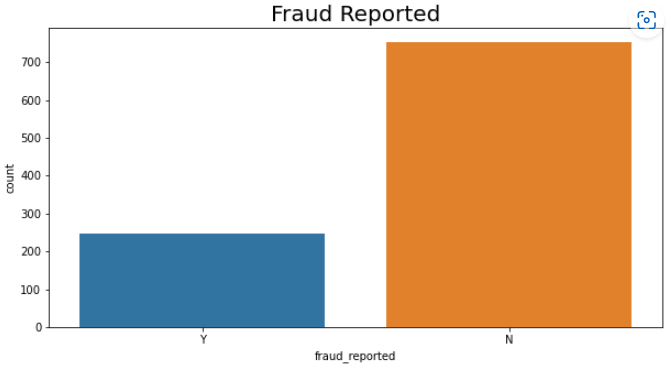
1. **Univariate Analysis**

Univariate analysis is the technique of comparing and analyzing the dependency of a single predictor and a response variable.

Code:

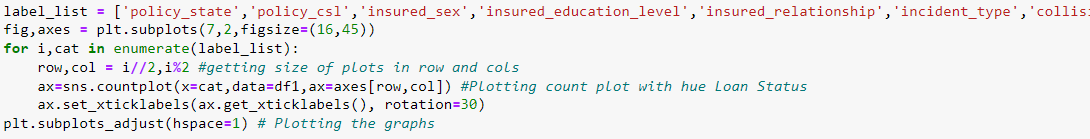


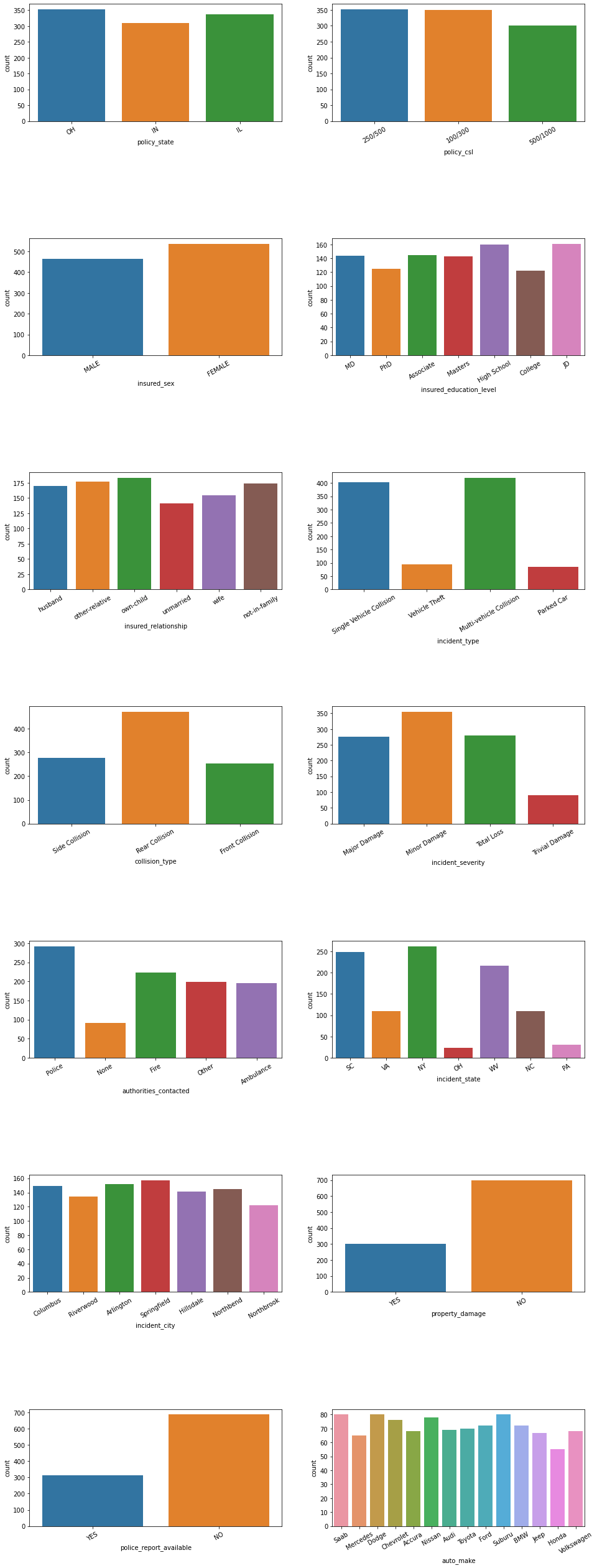
Output:



* In the above count plot, we can see that Y represents Yes means insurance claim is fraudulent and N represents No means insuramce claim is not fraudulent. No count is higher than yes count. Yes count is nearly 200-300, No count is above 700.
* Since we are dealing the classification problem, the target fraud\_reported it indicates the class imbalance issue. We need to balance the data before proceed with our models.

Then we can do more plots for some other counts using loop.

Code: 



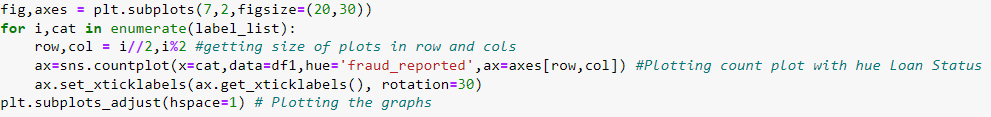
The above count plot shows that:

* Policy\_state is high in OH and IL where as low in IN.
* Policy\_csl is 250/500 and 100/300 is high where as 500/1000 is less.
* Female applied for Insurance more than Male.
* Mosty insured person have education level High School and JD.
* The person who have child, those peoples did insurances more.
* Singlie Vehicle Collision and Multi vehicle Collision had more incident where as Vehile Theft and Parked car incidents are less.
* Rear Collision is more than 400 where as Front collision is less.
* Minor Damages are happen more where as Trival Damages are less.
* People contacted more and in very less conditions people didn't call to anyone.
* More incident happen in NY where as in OH very less incident happen.
* Almost all cities have same number of incidents.
* Property damages are less.
* Police reports are available only in 30% cases.
* All cars are insured equally near 60-80.

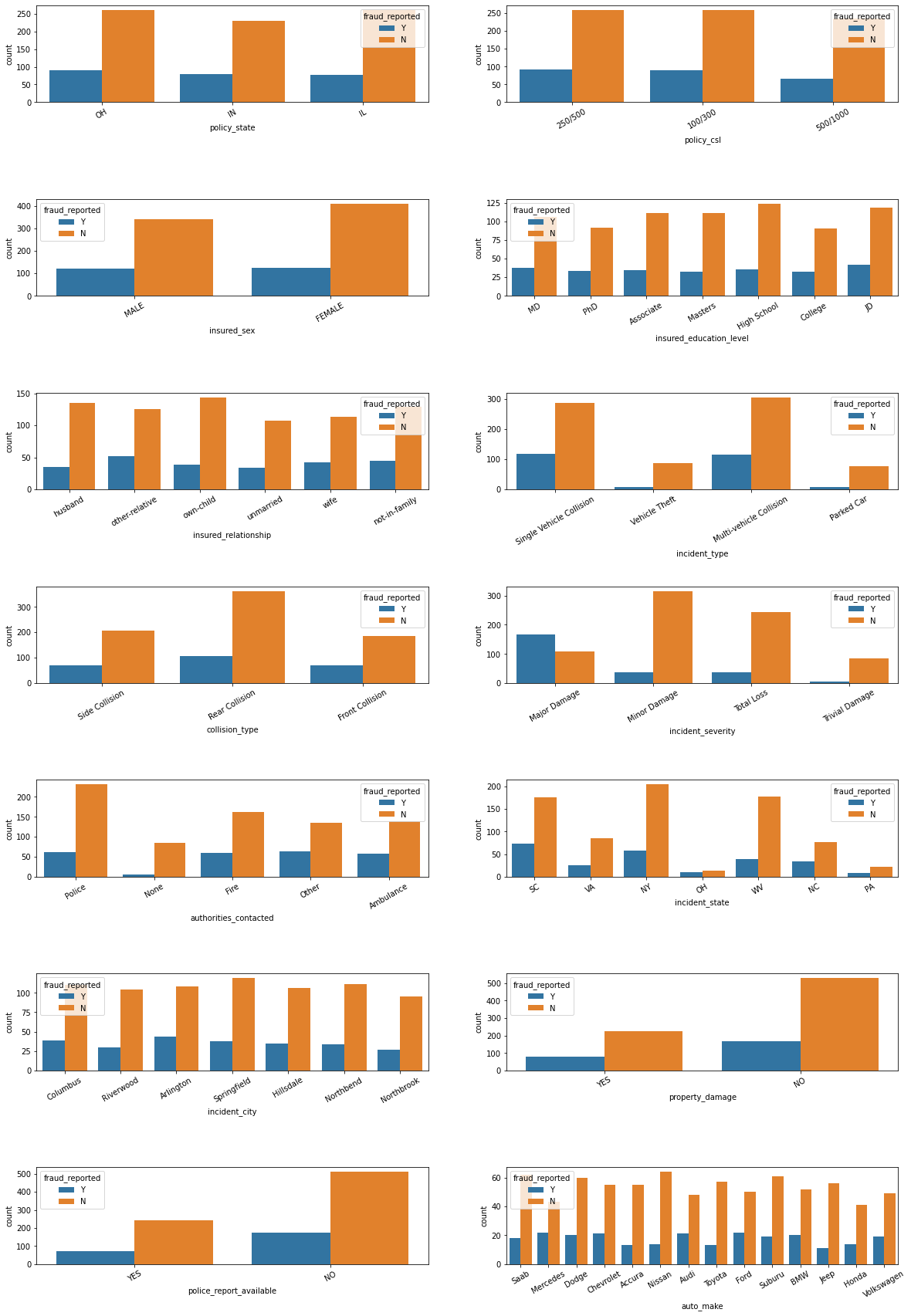
**b. Bivariate Analysis**

The relationship between each variable in the dataset and the target variable (or) using 2 variables and finding the relationship between them.

Code:



Output:

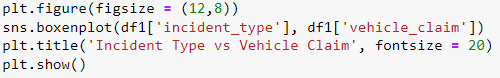


The above plot shows that:

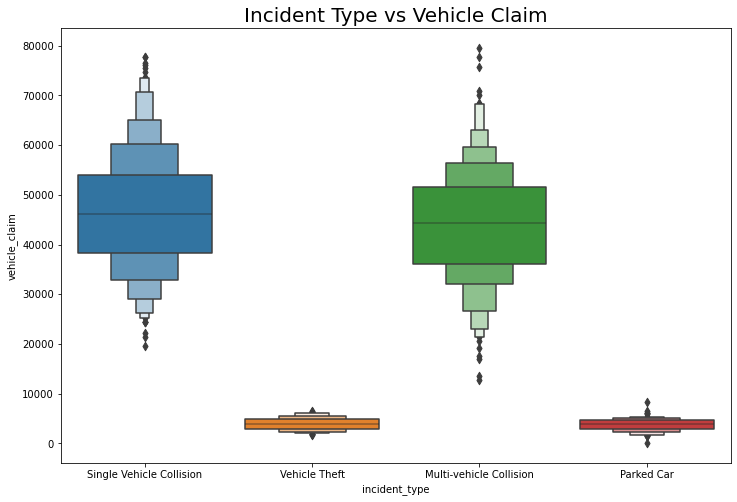
* In OH and IL Policy state Fraud didn't happen much when comapred with IN. In IN Fraud happen more than OH and IL.
* In Policy\_csl 250/500 and 100/300 Fraud reported more than 500/1000.
* More number of Fraud case reported against Female.
* Person who have education level JD did more frauds than others.
* People who have other relatives has reported highest fraude cases.
* In case of Singlie Vehicle Collision and Multi vehicle Collision maximum frauds happened.
* When Collision type is Rear, Fraud reported more.
* At the time of Major Damages, Fraud is reported too high.
* When People didn't call to anyone that time Fraud happen very less. If person contacted to anyone, Fraud happened.
* In SC state Fraue Reported more than other states.
* Almost all cities have same number of Fraud Reported.
* When Property is not damages that time Fraud happed more.
* When there is no Police report available that time Fraaud happen more.
* Mercedes and Ford have more number of Fraud Reported.

The next plot is to show boxen plot between incident\_type and vehicle claim.

Code:



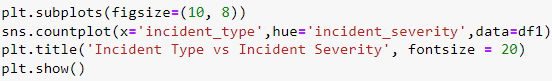
Output:



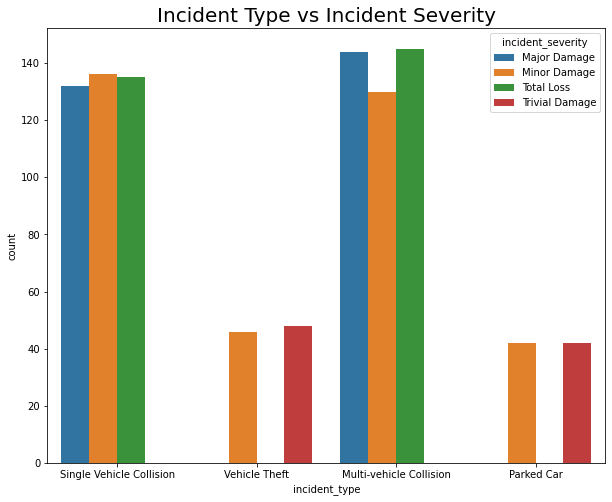
We can clearly see People do Vehicle claim when Single Vehicle Collision or Multi-vehicle Collision happen.

The next plot shows relationship between Incident type and Incident severity.

Code:

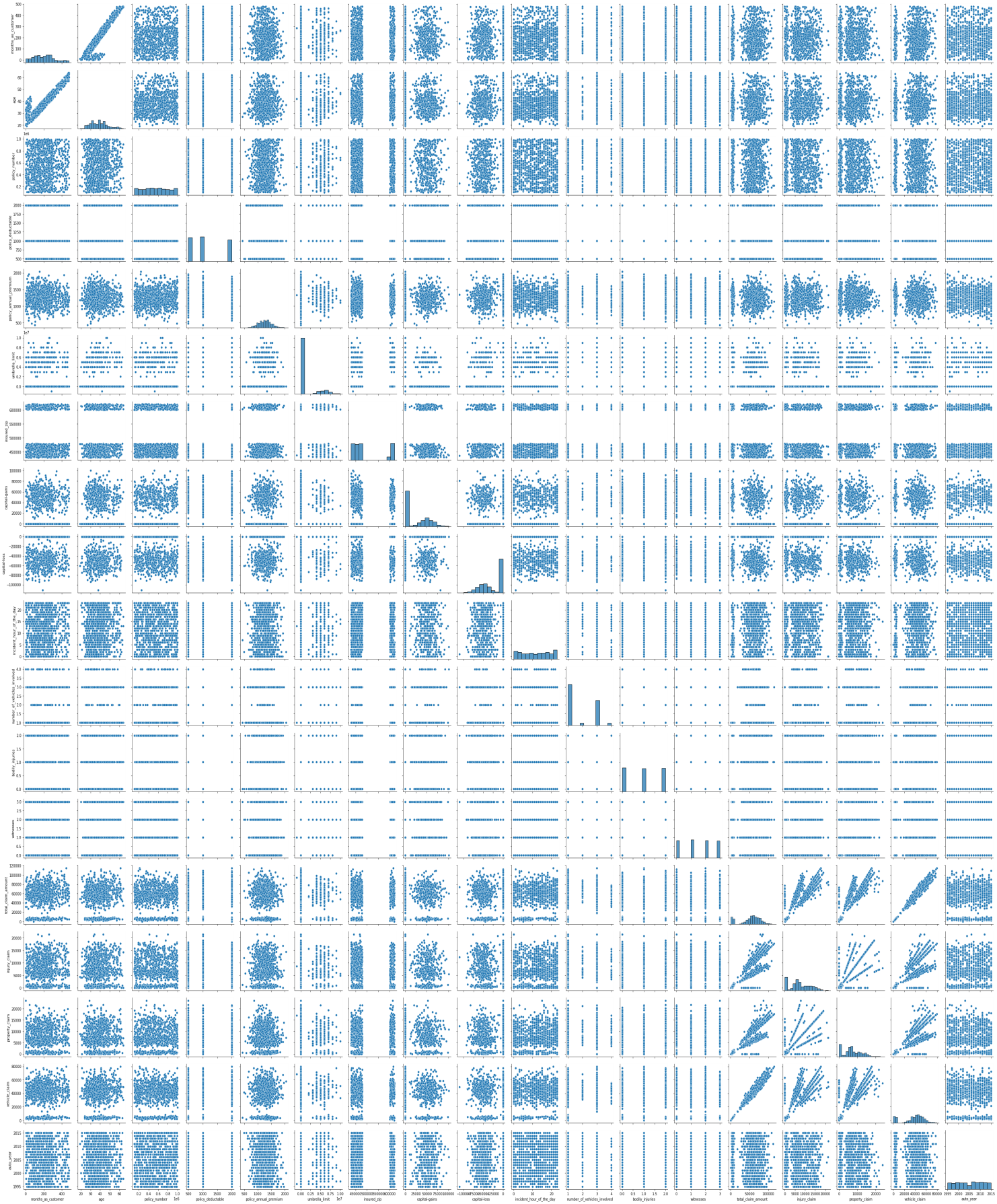


Output:

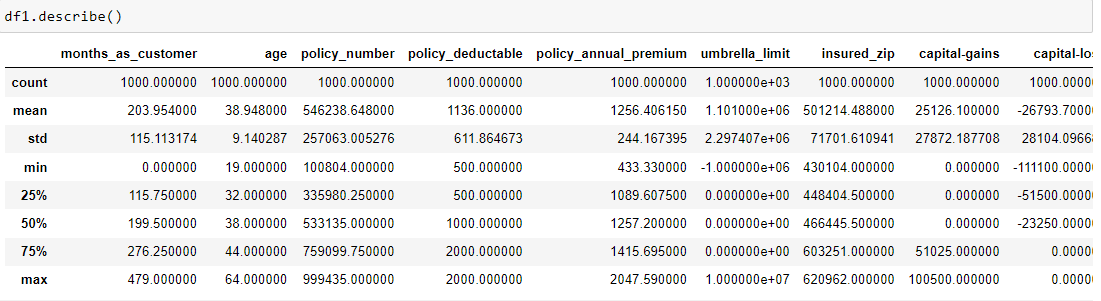


1. **Multivariate Analysis**

To understand interactions between different fields in the dataset.

****

**Summary Statistics**

****

The describe method gives the statistical information of the dataset. It gives the summary of numerical data.

From the above description we can observe the following things:

* Here the counts of all the columns are equal which means there are no missing values in the dataset.
* In some of the columns like months\_as\_customer, policy\_deductable, capital-gains, etc., we can observe the mean value is greater than the median (50%) which means the data in those columns are skewed to right.
* And in some of the columns like total\_claim\_amount, vehicle\_claim...etc we can observe the median is greater than the mean which means the data in the columns are skewed to left.
* And some of the columns have equal mean and median that means the data symmetric and is normally distributed and no skewness present.
* There is a huge difference in 75% and max it shows that huge outliers present in the columns.

Then we are checking unique values for the column because it has the value that separated with slash.



Then we are splitting policy\_csl column into two, csl\_per\_person and csl\_per\_accident by using string split with slash and expand=True functions.



Then we are converting Auto year column into Vehicle Age by subtracting 2020.



Then using drop we are going to drop some colomns from our dataset. Reasons behind droppin those columns:

* Policy number is unique for each policy.
* Insured Zip is pin code from where policy created.
* Policy bind date is policy date.
* Incident date is the date when incident happen.
* Incident location is where incident happen.
* Incident hour of the day is what what time incident happen.

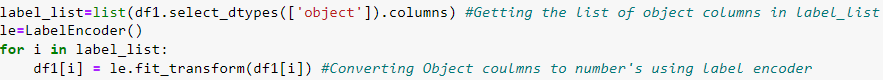
All of those columns are not gonna help us to predict something, So we’ll drop these columns. After dropping the above columns our dataset contains 1000 rows and 34 columns.

After that we are getting all integer columns in to int\_label list for making our encoding process of object variable easy.



**Encoding**

We are getting all the object variable into lable\_list and then using for loop we are doing the encoding process using LabelEncoder.



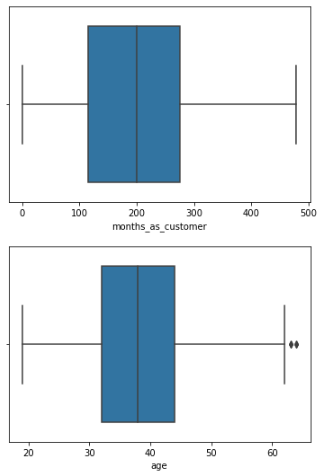
After encoding process all object variable is converted to integer datatype.

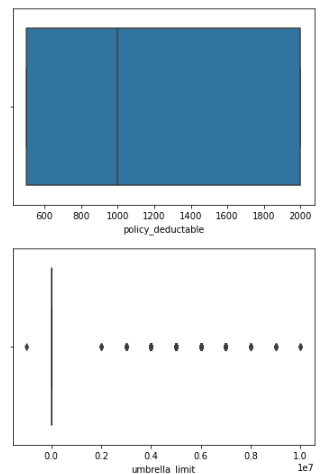
**Checking Outliers**

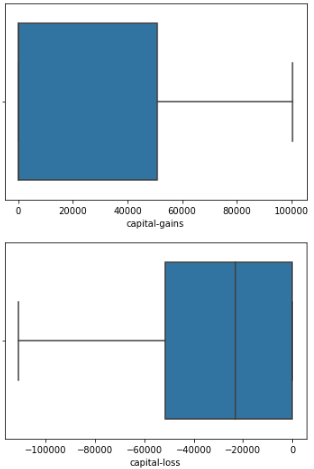
We are going to check outliers for integer datatypes using int\_label list with the help of for loop.

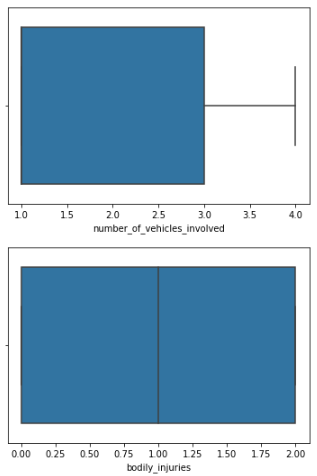


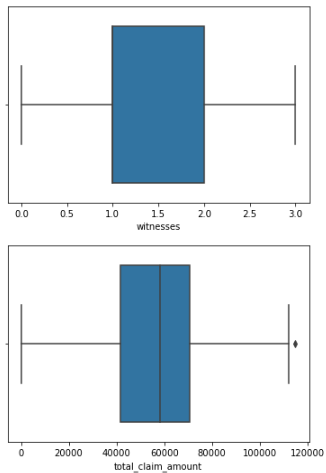
The below box plots are the view of outliers.

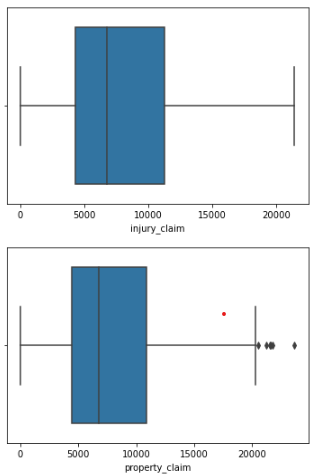


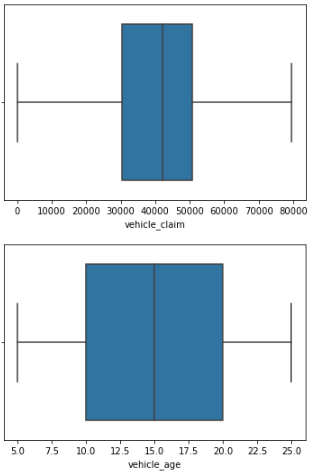




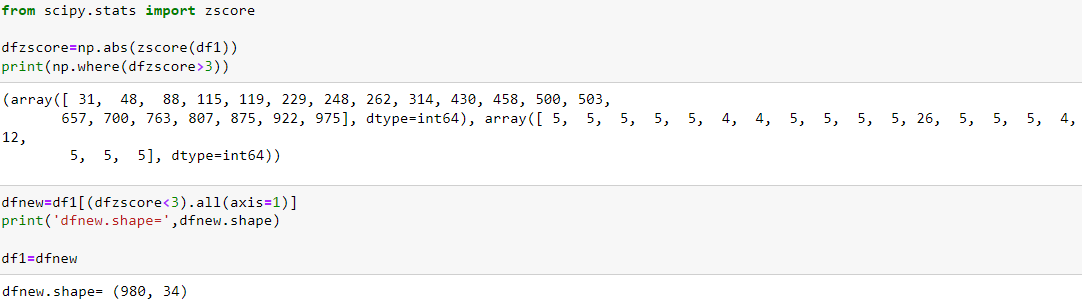




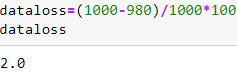




Umberalla limit have many outliers and Property\_claim, total\_claim\_amount,age few outliers. To remove Outliers I use Z-score method.



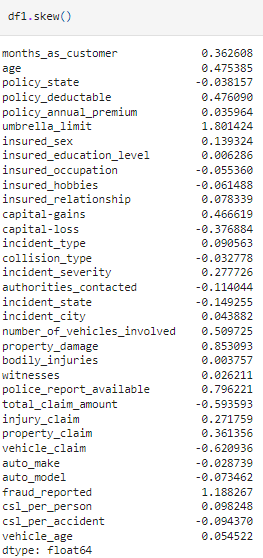
After removing Outliers using Z-score we are having 980 rows. I think percentage loss of data is less when we see the result of removing outliers.



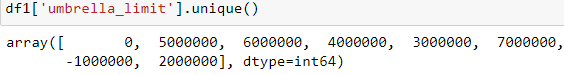
Dataloss is 2%. So that we can proceed further.

**Skewness**

We can check skewness of dataset using skew function.



Now we check unique values of umbrella\_limit.

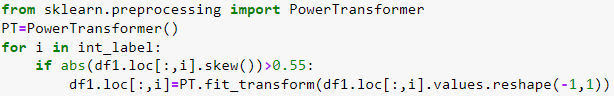


As we noticed from the aove code umbrella\_limit have only few values and feels like categorical. So we will not change its skewness.

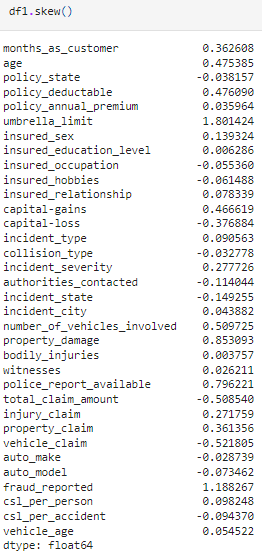
We remove it from integer datatype i.e., from init\_label.



Now we are removing skewness of all object type variables using power transform method.



Now skewness is reduced from our dataset.



After the skewness next process is our correlation.

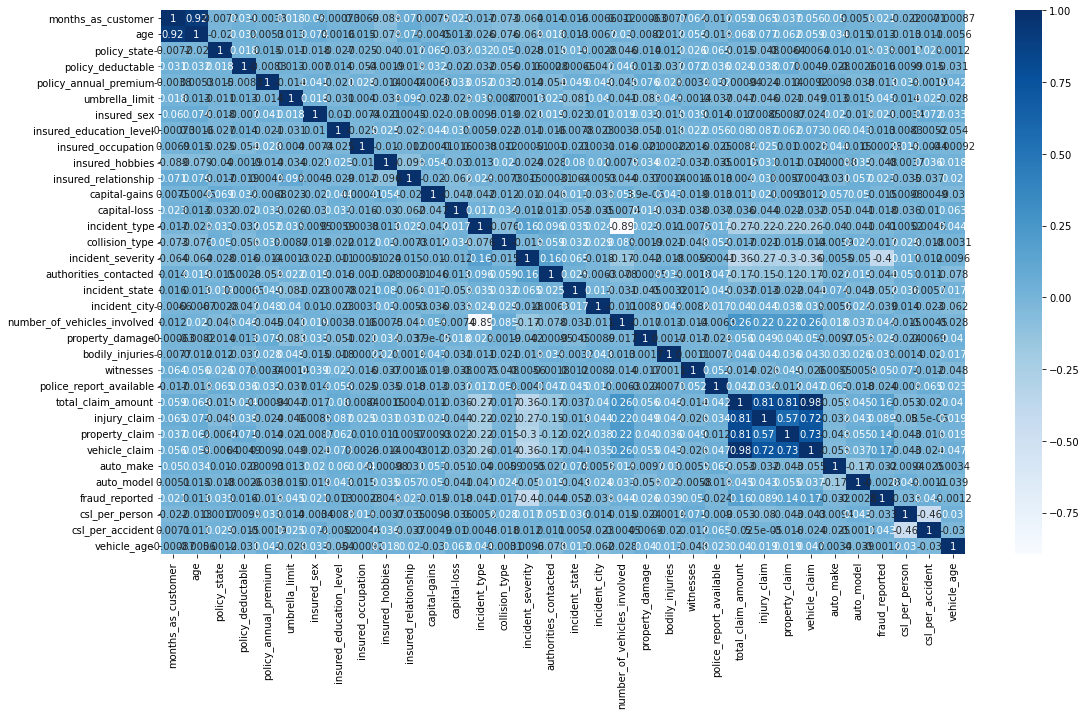
**Correlation**

We are going to check correlation of our datatset using corr() function. The below code is to show heatmap for correlation.

Code:



Output:



This heatmap shows the correlation matrix by visualizing the data. we can observe the relation between one feature to other. This heat map contains both positive and negative correlation. There is very less correlation between the target and the label.

1. **EDA Concluding Remark**

* I have checked the null values in the dataset and there was no missing values found.
* I have dropped some of the irrelevant columns to overcome with the multicollinearity problem.
* I have extracted some new features from the existing features to get better results without any hindrance. And dropped the old columns, if I keep them as it is they will act as duplicates and that leads to multicollinearity problem.
* Replaced the corrupted entries “?” in the columns with their respective mode values.
* To get the better insights about the features, I have used box plots, scatter plots, count plots and distribution plots.

1. **Pre-Processing Pipeline**

First, I have to separate the label and features to process my dataset.





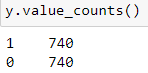
I have separated independent and dependent features and stored them in x and y respectively.

* Now, I have to scale the data containing independent variables (x) in order to overcome with the data biasness.
* Since I have removed the skewness and outliers and my data is also normal so I can use Standard Scaler method to scale the data. If it is not the case then we could use Min Max Scaler.
* Before that I use SMOTE method to overcome the imbalance of the dataset.

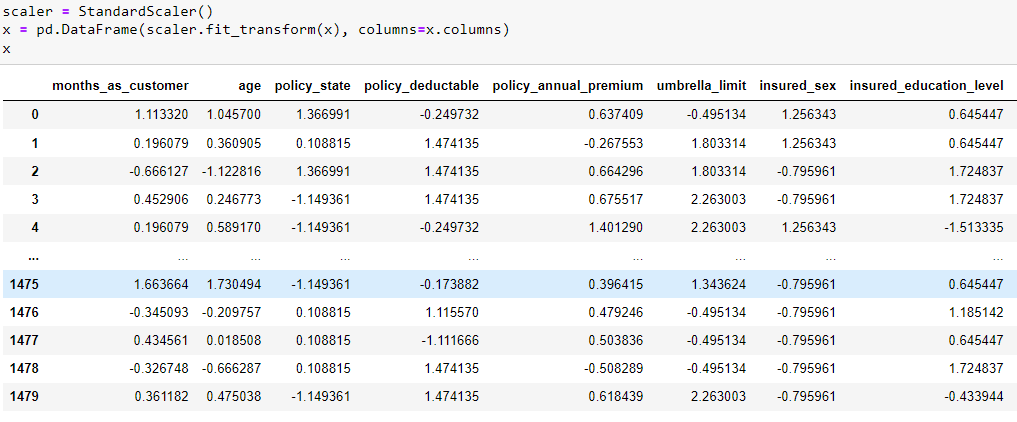
**SMOTE**



The result of SMOTE is given below.



**Scaling**

****

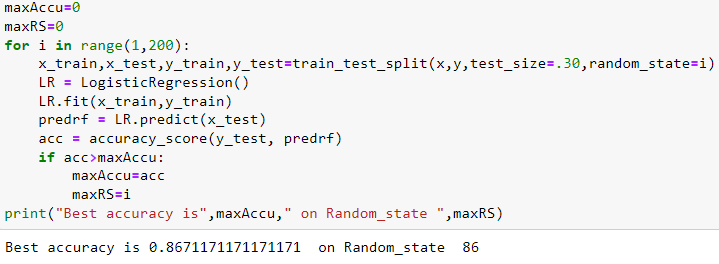
I have scaled the data using standard scaler method to overcome with the issue of data biasness.

Since I have done all the pre processing and data cleaning, now my data is ready to build the model. Let’s get the predictions by creating some classification algorithms as it is a classification problem.

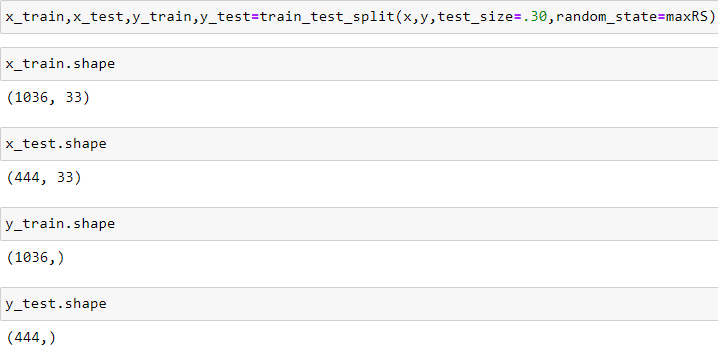
1. **Building Machine Learning Models**

Before building the models, first we need to find the best random state and accuracy using any one of the classification models.

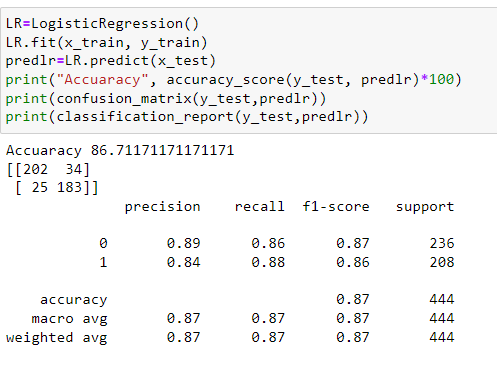
**Finding the best random state**



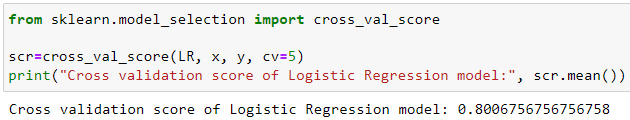
* I have got best random state as 86 and best accuracy as 86.7% using Random Forest Classifier.
* Now let’s create new train and test and fit them into the models to find our ideal model.



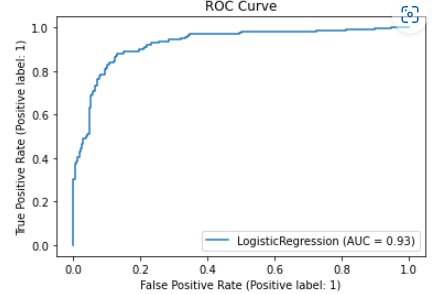
**Logistic Regression**

****

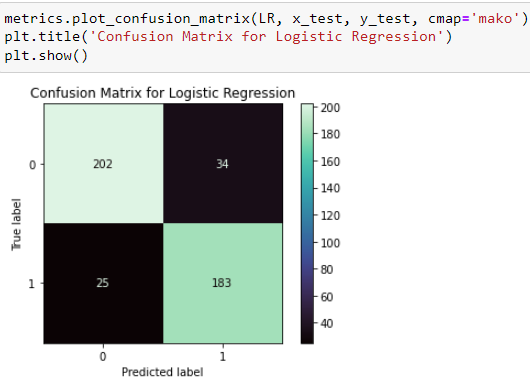
**Cross Validation Score**

****

**ROC Curve**

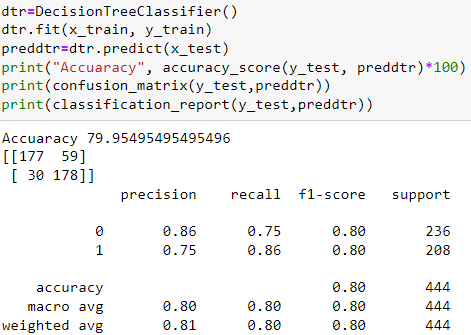
****

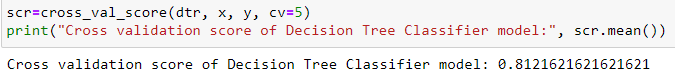
**Confusion Matrix**

****

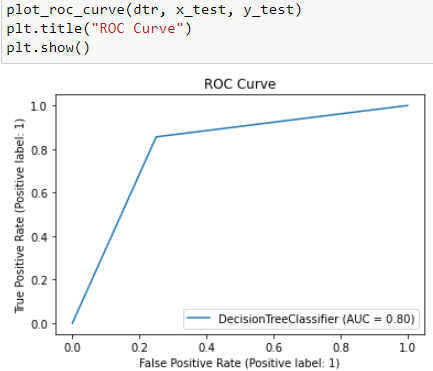
Logistic Regression giving accuracy of 86 % and cross validation score is 80 %.

**Decision Tree Classifier**

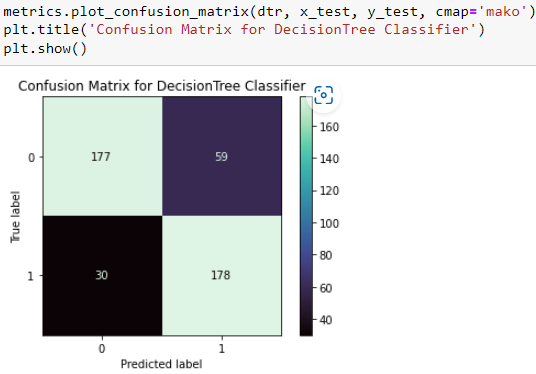
****

**Cross Validation Score **

**ROC Curve**

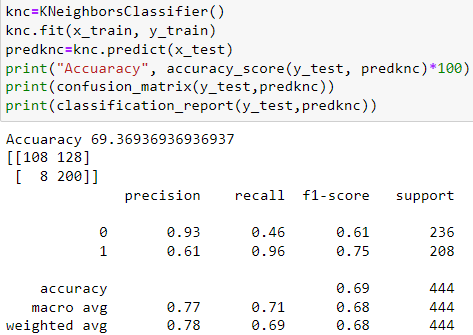
****

**Confusion Matrix**

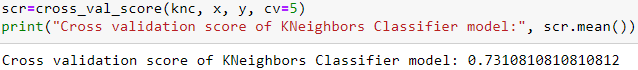
****

Decision tree classifier gives the accuracy as 79 % and cross validation score is 81 %.

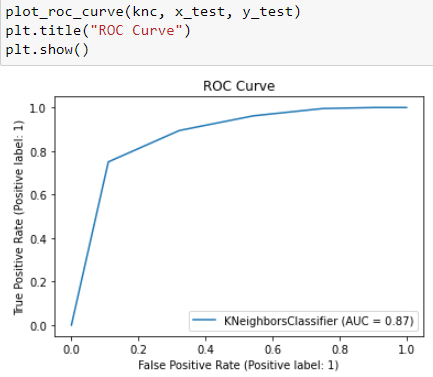
**KNearest Neighbors Classifier**



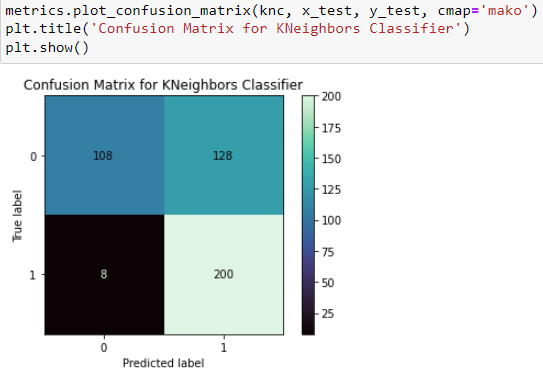
**Cross validation Score**

****

**ROC Curve**

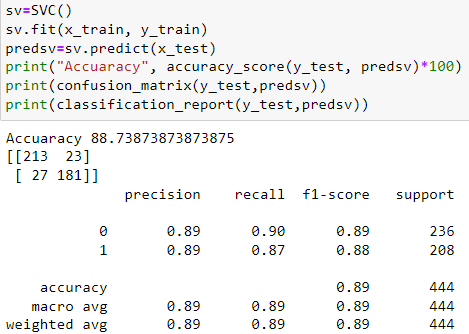
****

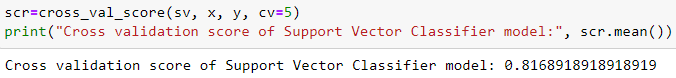
**Confusion Matrix**

****

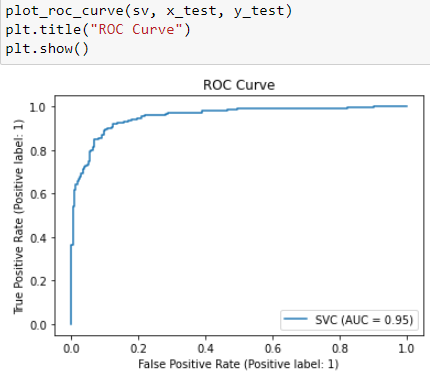
KNearest Neighbors Classifier gives the accuracy value 69 %, and cross validation score as 73 %.

**Support Vector Classifier**

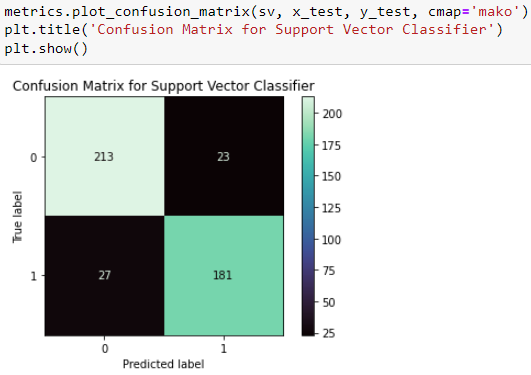
****

**Cross Validation Score **

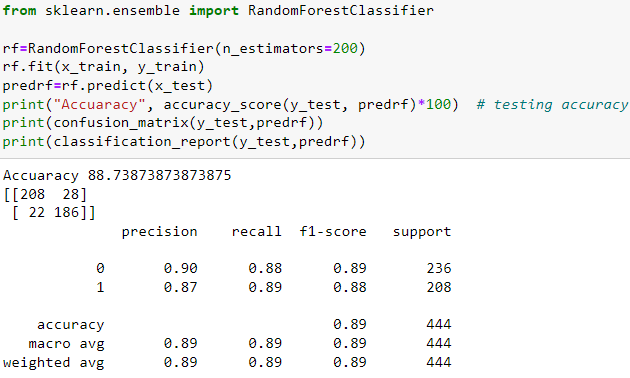
**ROC Curve**

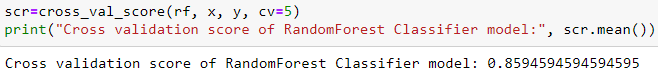
****

**Confusion Matrix**

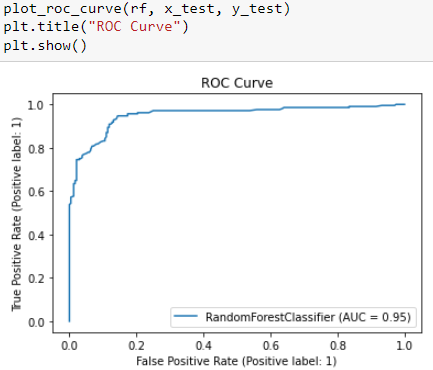
****

Support Vector Classifier gives the accuracy score as 88 % and cross validation score as 81 %.

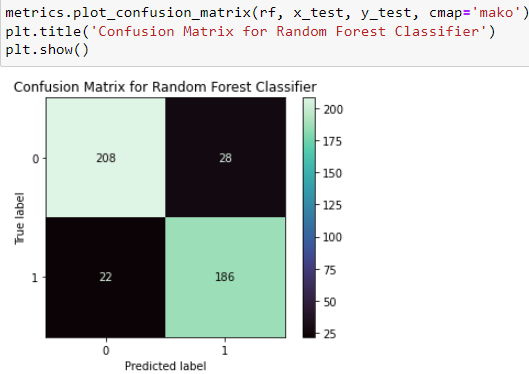
**Random Forest Classifier **

**Cross Validation Score **

**ROC Curve**

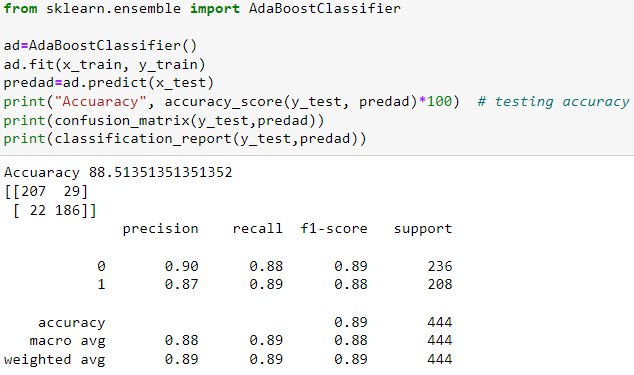
****

**Confusion Matrix**

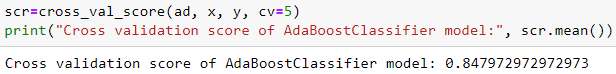
****

Random Forest Classifier gives accuracy score as 88 % and cross validation score is 85%.

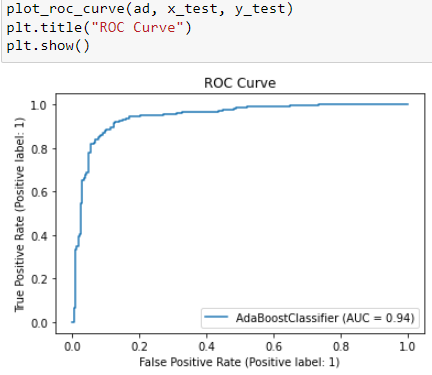
**AdaBoost Classifier**



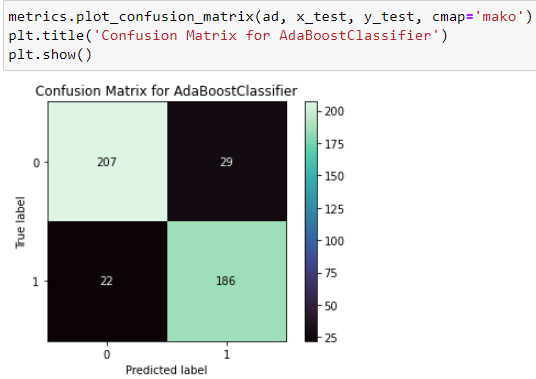
**Cross Validation Score**

****

**ROC Curve**

****

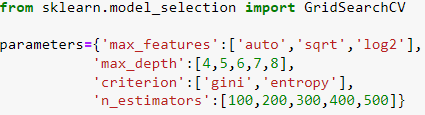
**Confusion Matrix**



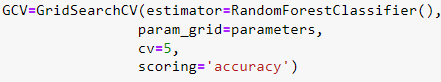
AdaBoost Classifier gives the accuracy of 88 % and cross validation score as 84 %.

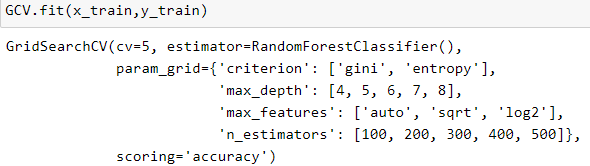
* From the difference between the accuracy score and cross validation score, we can conclude that Random Forest Classifier as our best fitting model for prediction.
* Based on these results I am performing hyper parameter tuning to improve my accuracy.

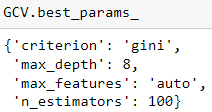
**Hyper Parameter Tuning**



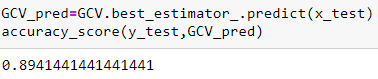
By using above parameters, I am tuning my best model and after tuning I have to choose the best parameters from the above list. Let’s get the best parameters.





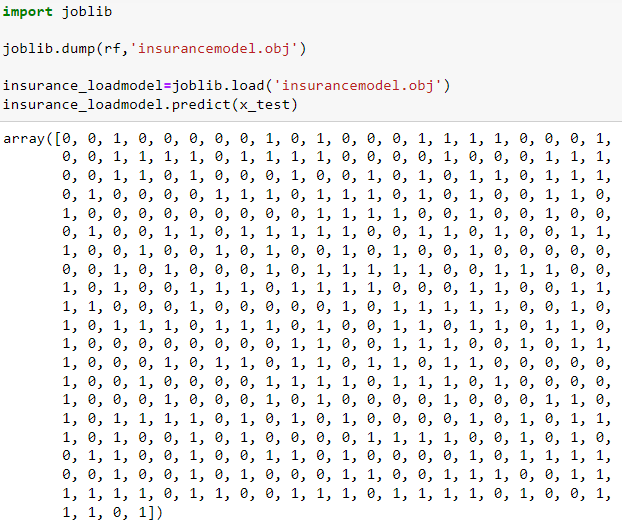


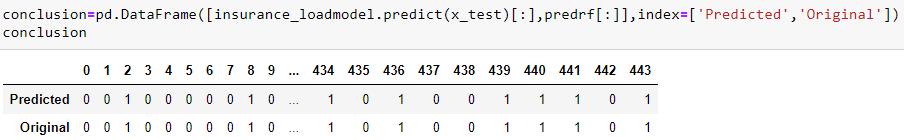
I ran GridSearchCV and find the best parameters. Now I will use these parameters to get the accuracy of the best model.



After tuning the model, the accuracy score increased a little.

Since I have done with the model building, it’s time to save the model. Here I have saved my model and predicting the results as shown below.





The actual and predicted values are almost same, that means our model worked well.

1. **Conclusion Remark**

In this project we have gone through the feature engineering which is the most crucial thing and removed the outliers and skewness. Also handled the categorical columns by encoding the data, scaled the data and at last, we built different classification models to predict whether the insurance claim is fraudulent or not and performed the hyper tuning to improve the model by using different parameters.

With the help of above techniques, our model is able to predict the fraudulent report with the accuracy of 89.41 %. Also, we have seen the actual and predicted values are almost same that means our model worked correctly. Building machine learning models for such problems can help the insurance companies to choose the correct insurer. So, Machine learning techniques are very useful to solve these kind of problem.