**Insurance Claims- Fraud Detection**

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, the first thing that comes into your mind is, what is insurance fraud. Let’s understand this.

What is insurance fraud?

“Improper community committed by an individual in order to obtain favourable outcomes from the insurance company”.

The insurance fraud can be broadly classified into 2 types.

1. Soft Insurance Fraud
2. Hard Insurance Fraud

**Soft insurance fraud**: A common example for this is, if the accident has taken place, but the amount of damage what has happened to the vehicle is very less. In such cases, the individual claims to the insurance company that huge amount of damage have occurred to the vehicle with the goal of charging the insurance company a higher bill.

**Hard insurance fraud**: In this type of fraud, an individual intentionally plans and invest the loss so that he can claims for the insurance from the company. A common example for this type of fraud is staging a car wreck with the goal of benefitting from the resulting claim.

This project focuses on claim data of an Automobile insurance company. Because of 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’ll 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.

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, a dataset was provided which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

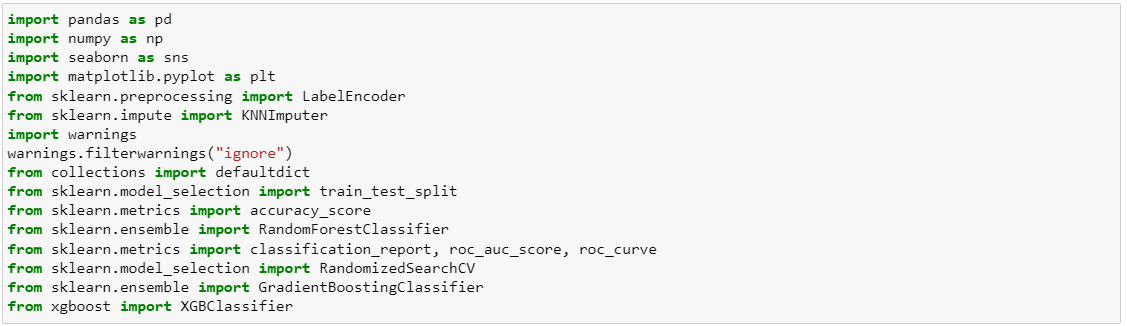
In this example, we will be working with some auto insurance data to demonstrate how we can create a predictive model that predicts if an insurance claim is fraudulent or not.

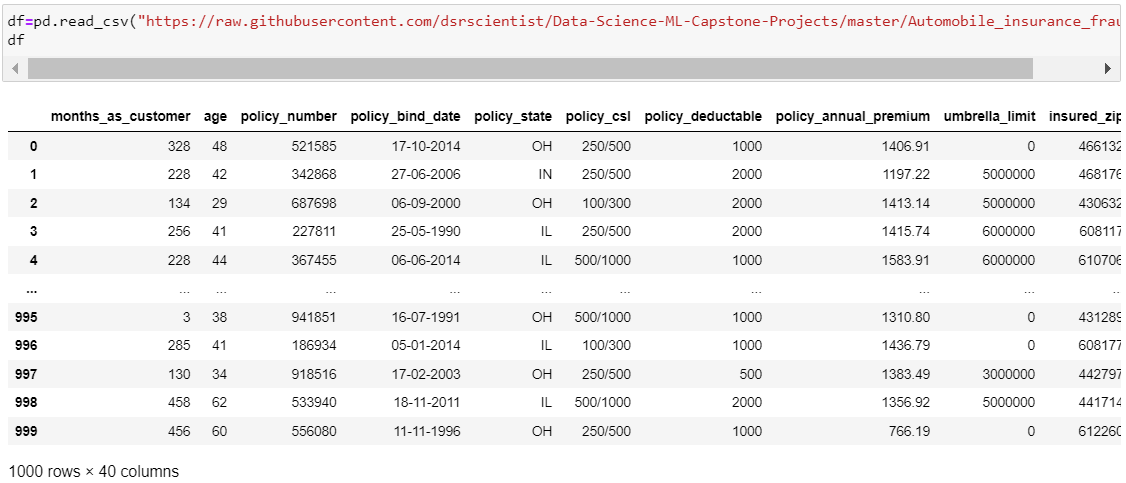
The problem statement explains that the target variable contains the categories, so it is a **“Classification Problem”** where we need to predict whether an insurance claim is fraudulent or not.

Data Analysis

The process of cleaning, transforming and extracting data to discover the useful information for business decision making is called data analysis.

Importing necessary libraries and 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:**

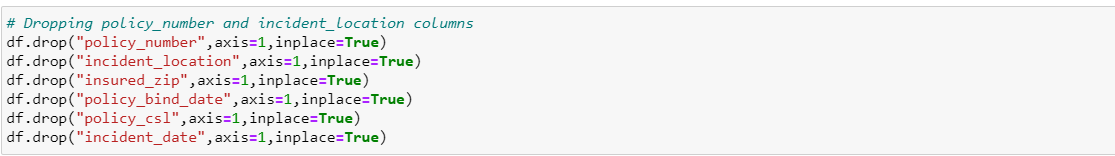
**Fraud\_reported:** Whether fraud was reported as Yes or No

For data pre-processing, we need to check some statistical information about the dataset like checking shape, datatypes, nunique, value counts, info() etc.

While running df.info(), I found c\_39 column having one unique count as NAN throughout the dataset and it is of no use, I dropped that column.

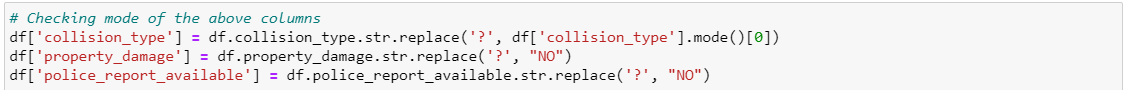


We have dropped few more columns which holds more unique values and not required for our analysis as well.



From the value counts we can observe that umbrella\_limit, capital-gains and capital-loss contains zero values. Since the umbrella\_limit columns more zero values, let's drop the same. Also, we can observe some columns have "?" values, we need to fill them.

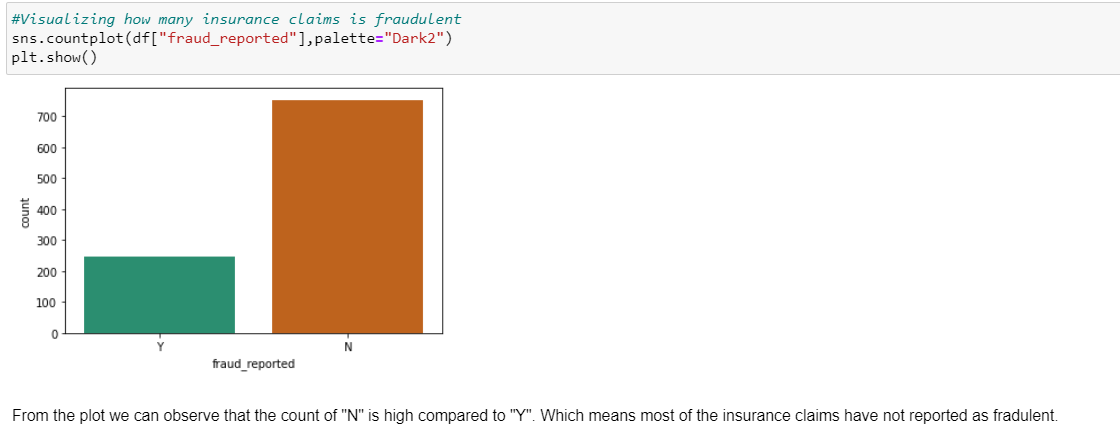




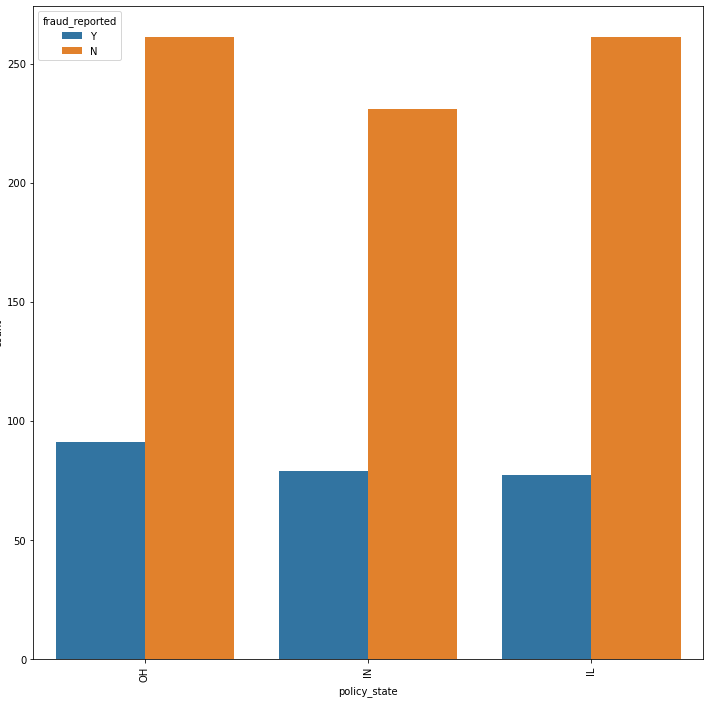
Exploratory Data Analysis

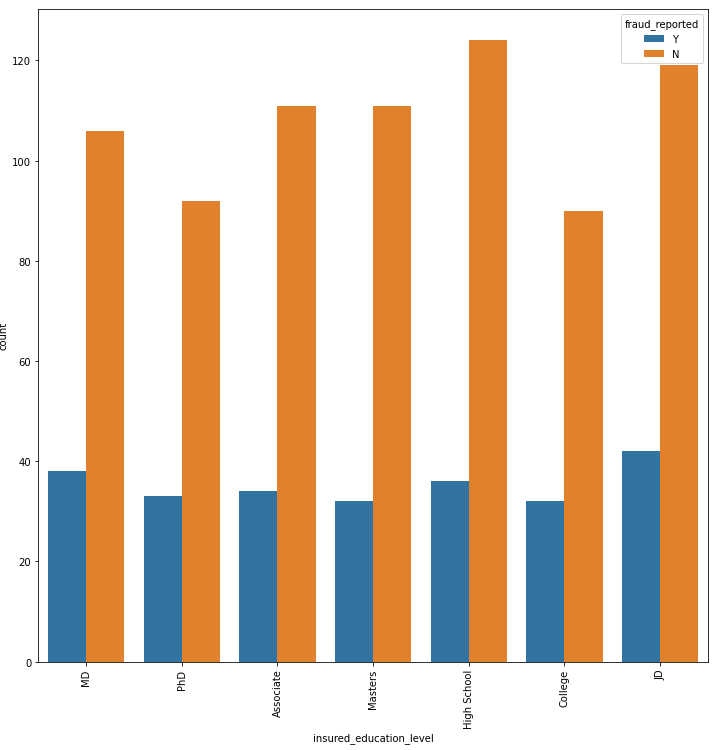
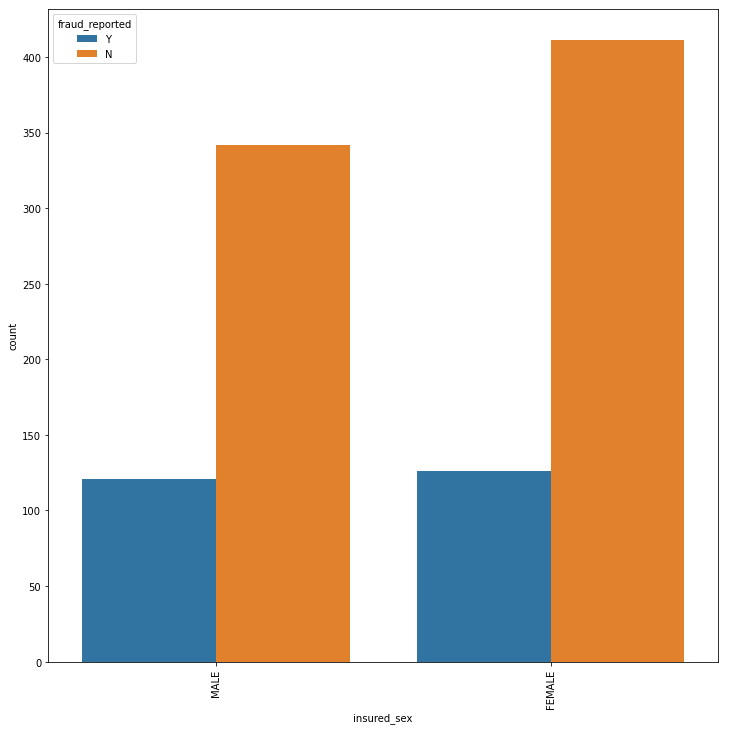
**Analyzing the Target Class**

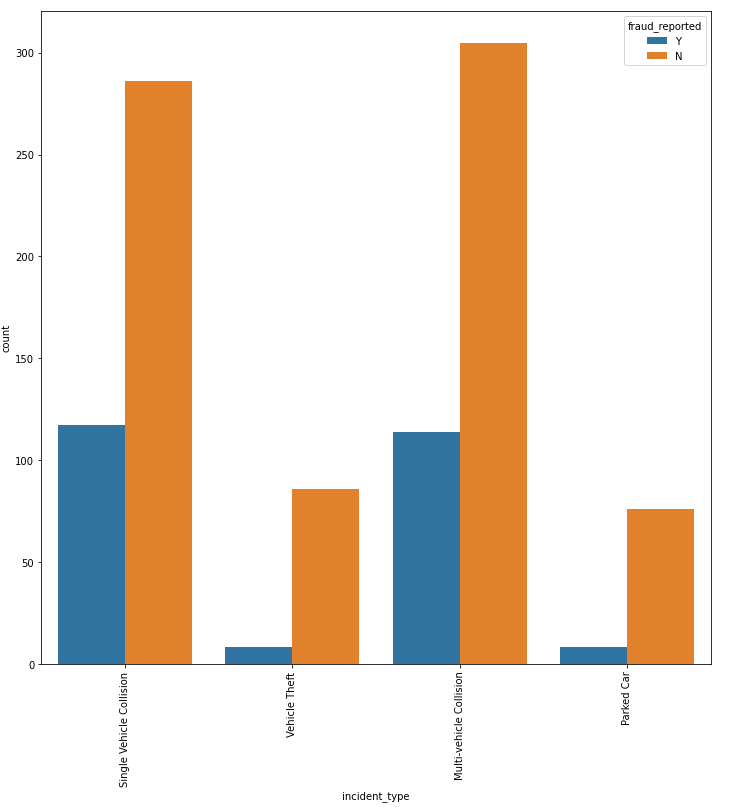
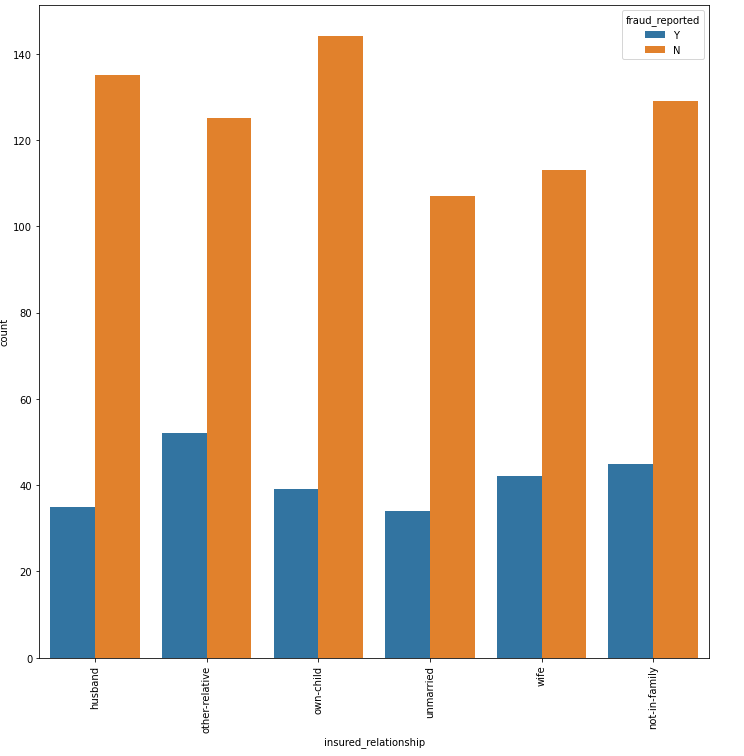
#### There are 2 unique categorical values in the Label column / target variable, viz. ‘Y’ and ‘N’.

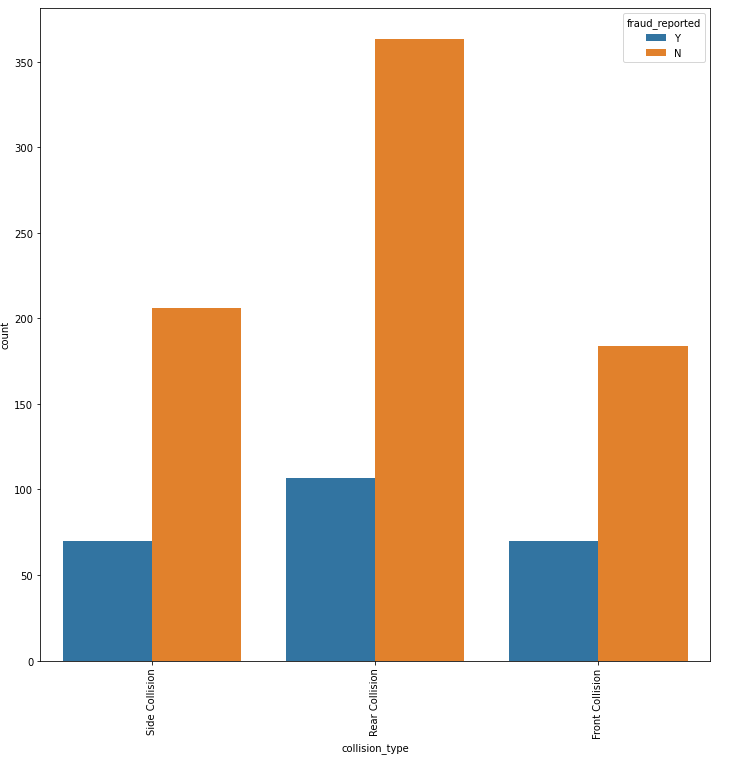


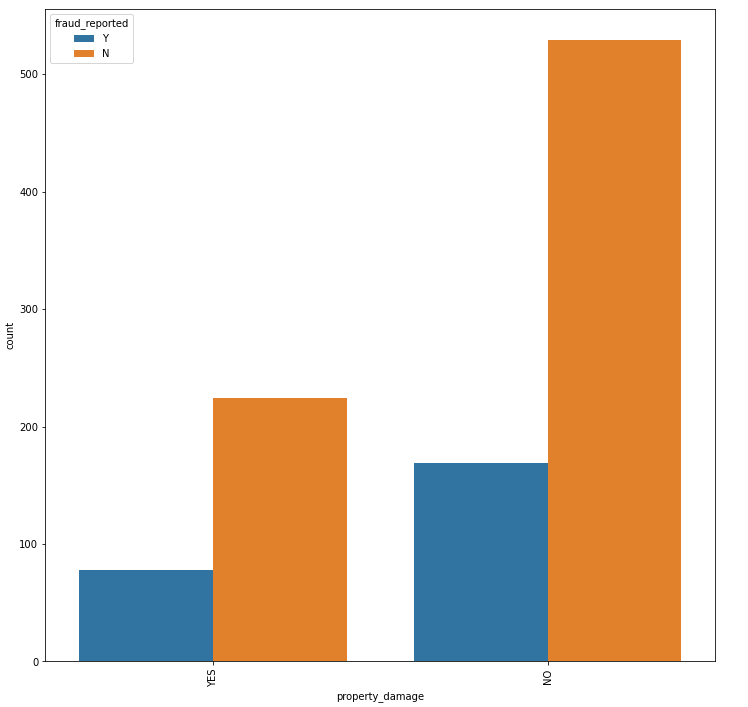
Checking categorical columns with target:

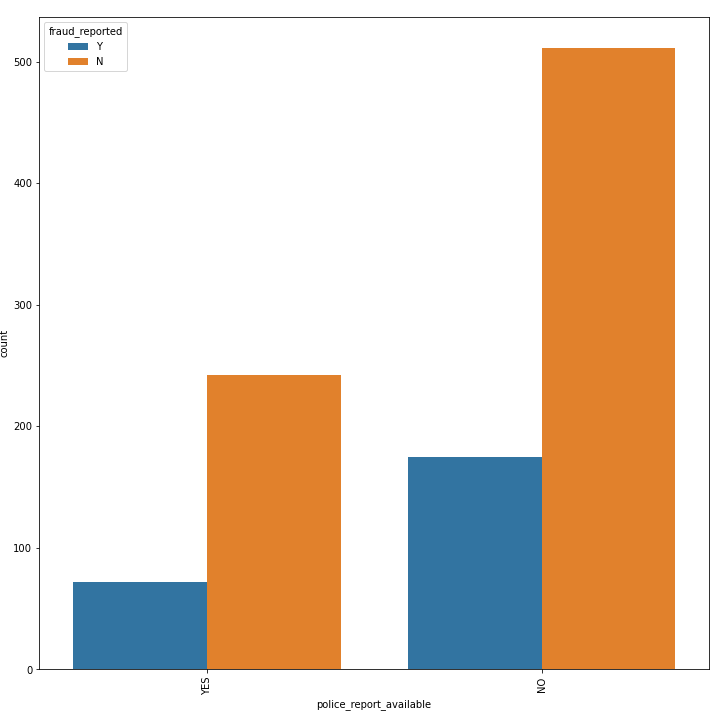




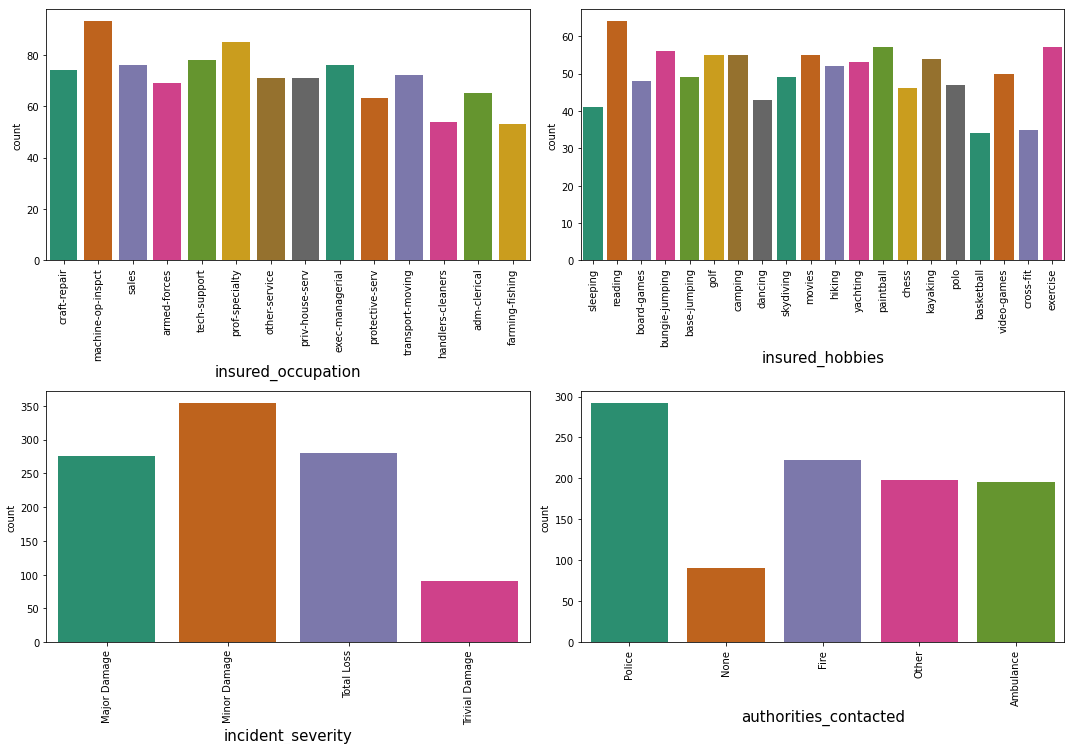


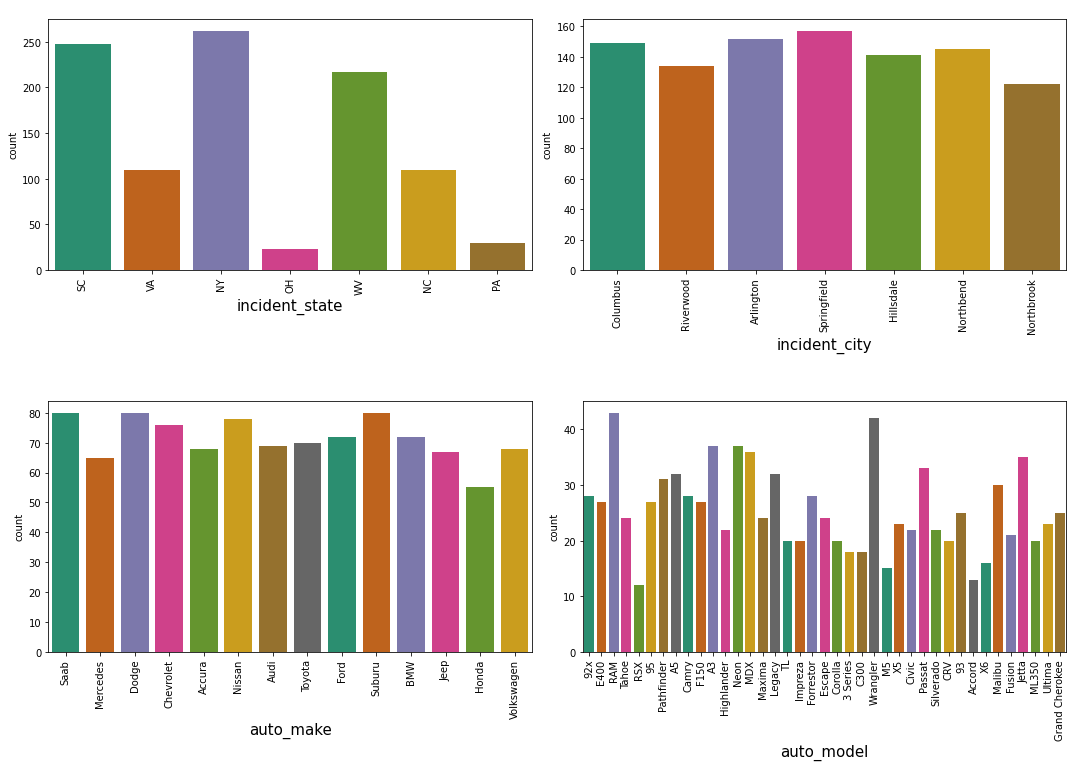






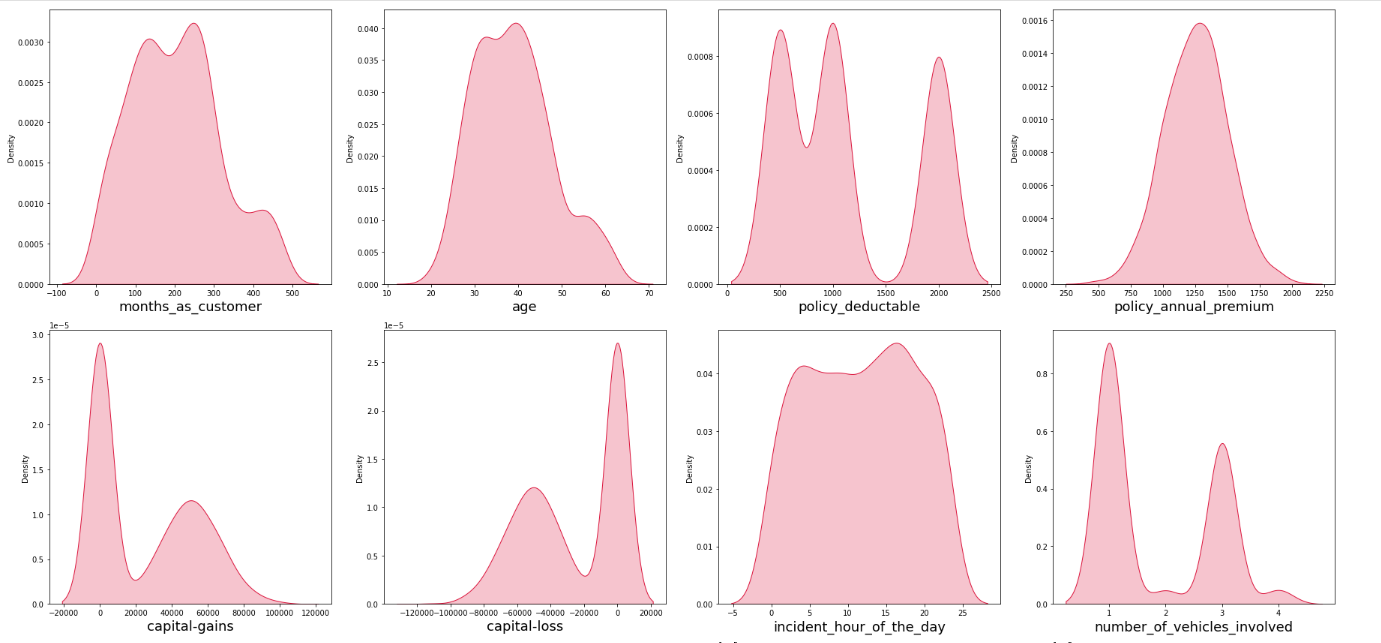
* The types of the policies claimed by the customers are almost same - still the policy state type OH has bit high counts and the type IL has bit less count.
* Both male and female have insurance but the count for Female is little higher than Male counts.
* The count is pretty much same for all the education level but still the people who have completed their college,masters and PhD have less count compared to others.
* While looking at the incident type, Multi-vehicle collision and Single Vehicle Collision are pretty much similar. But the count is very less for Parked car and Vehicle Theft.
* The collision type, the count is high for Rear collision and the other two types have almost equal counts.
* More than half of the people did not face any property damage.
* More than half of the people were produced the police reports.

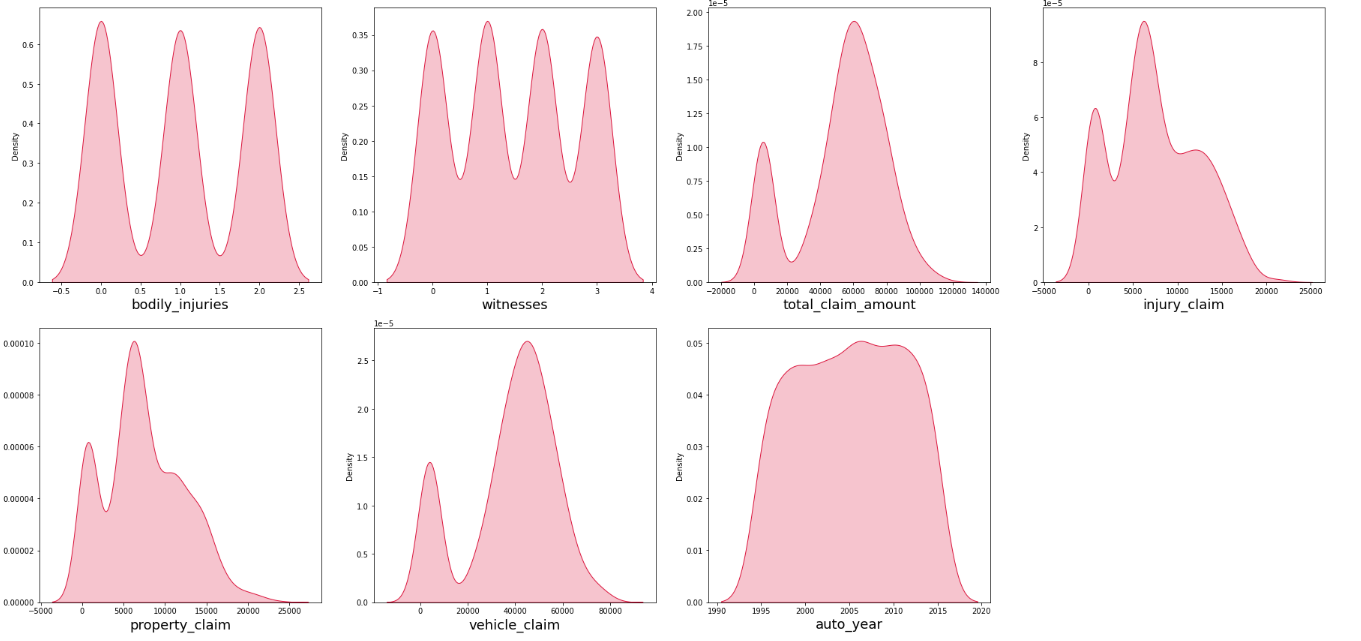




* In insured occupation, most of the data is covered by machine operation inspector followed by professional speciality.
* In insured hobbies, we can see reading holds the highest data followed by exercise.
* The incident severity count is high for Minor damages and trivial damage data has very less count compared to others.
* Most of the authorities contact police and Fire having the second highest count.
* With respect to the incident state, New York, South Carolina and West Virginia states have highest counts.
* In incident city, almost all categories have similar counts.
* with respect to automobile companies, the categories Saab, Suburu, Dodge, Nissan and Volkswagen have high count.
* With respect to vehicle models RAM and Wrangler have high count.

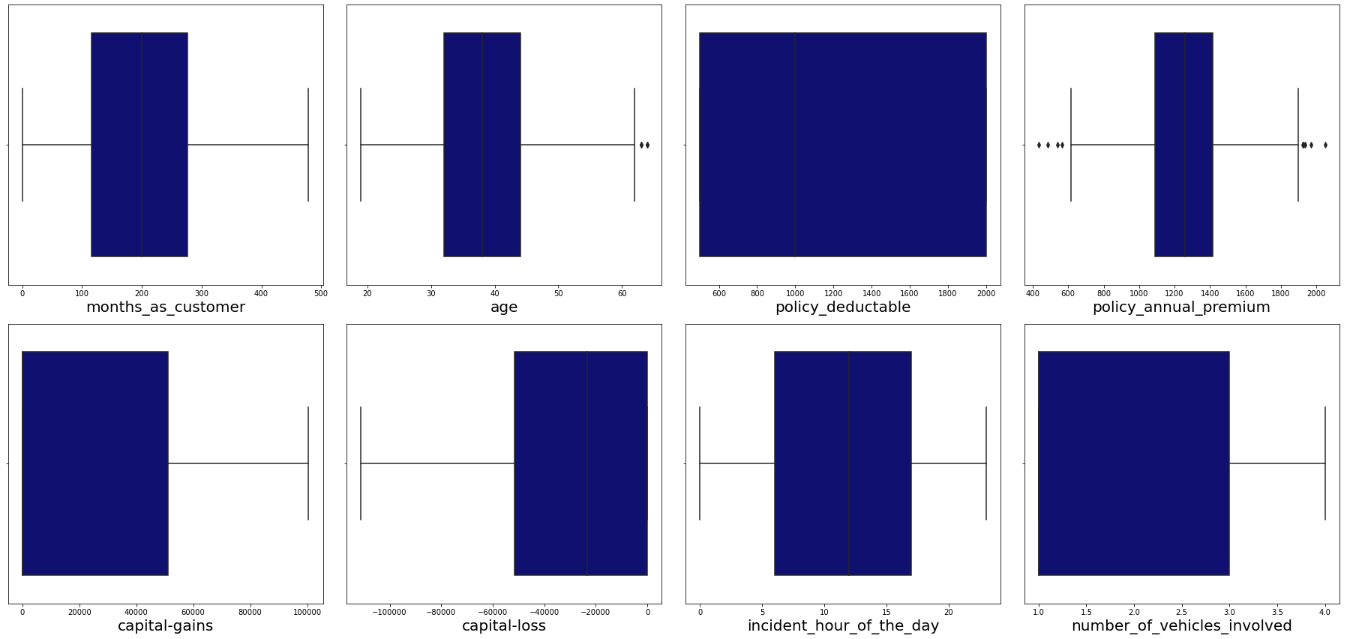
Checking the numerical features:

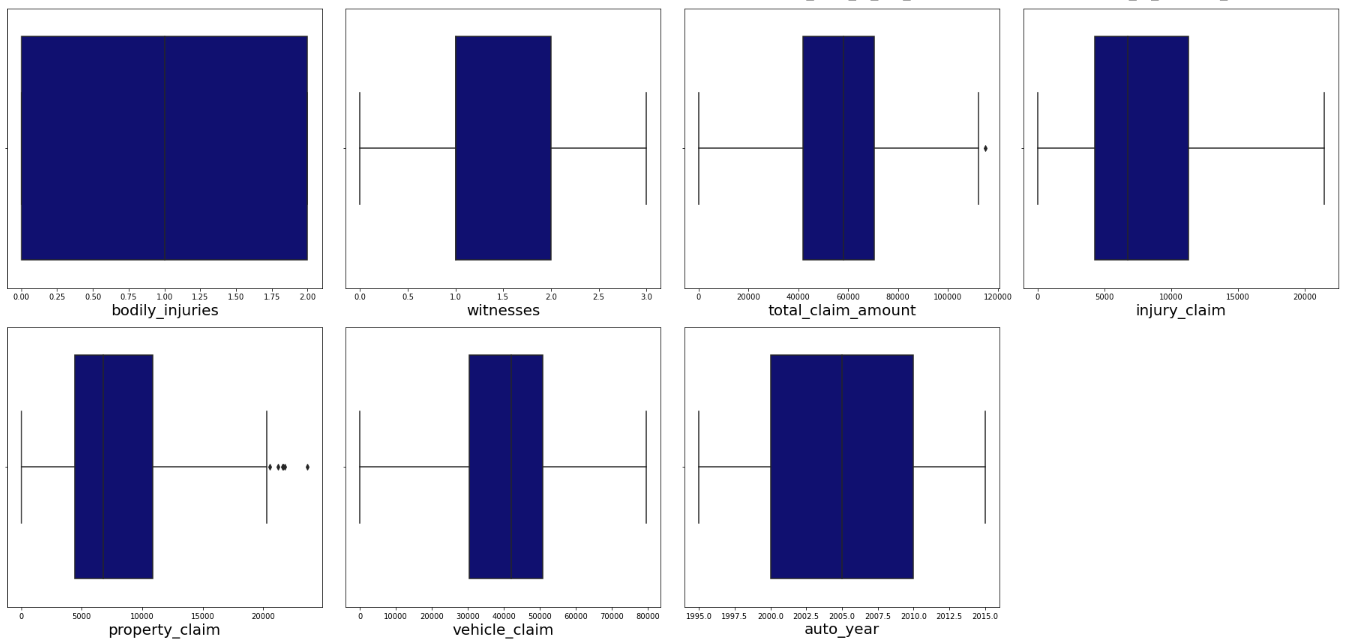




We can see the data in capital gains is skewed right and in capital loss it is skewed left.

# Next Step Involves Checking of Outliers, Skewness and then reducing it by Power Transformer Method

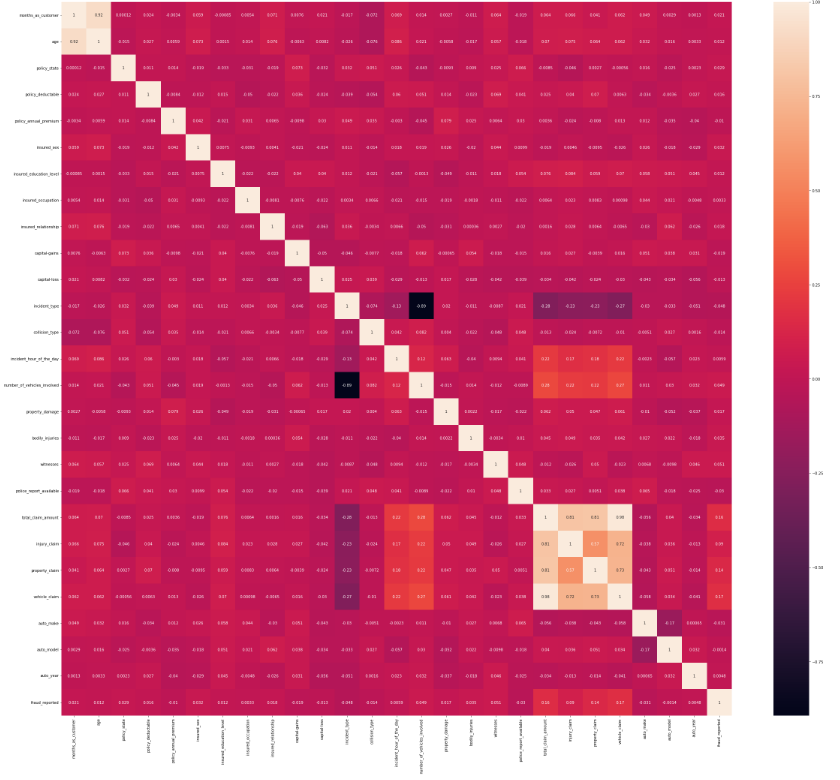




We can see outliers in age,policy\_annual\_premium,total\_claim\_amount,property\_claim. After removing the same we have a loss of 0.4% which is acceptable.

Next, we have used label encoder for all the categorical features to convert it to numerical data.

# Correlation:



We can see from the above correlation heat map that correlation is high between month\_as\_customer and age as they both represent no. of months. We can also see there is a high correlation for total\_claim\_amount, injury\_claim, property\_claim, and vehicle\_claim as total\_claim is the sum of injury\_claim, property\_claim and vehicle\_claim. Therefore dropping them will not affect the dataset.

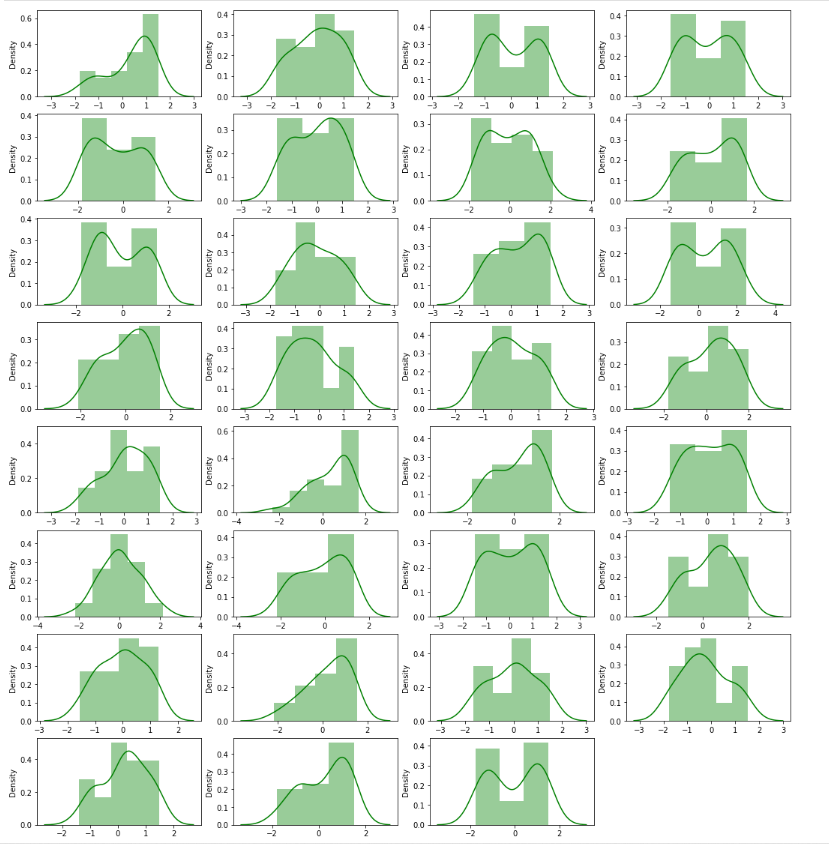
Pre-Processing Pipeline

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

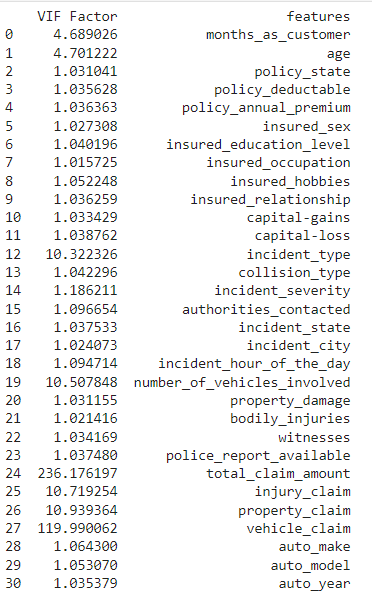


Skewness:

Next, we removed the skewness from the data by using power transformation (yeo-johnson method) method and normalised the data by standard scaler technique.



Post that we checked the VIF (Variance Inflation Factor) value for each feature for checking multicolinearity.

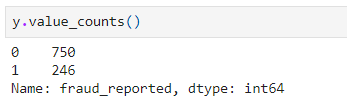


We can see vehicle\_claim,total\_claim\_amount,injury\_claim,property\_claim,

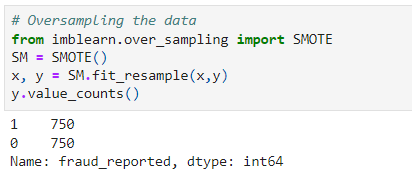
number\_of\_vehicles\_involved,incident\_type has VIF values more than 10. First, we removed total\_claim\_amount and checked the vif again. We can still see high value in number\_of\_vehicles\_involved,incident\_type has VIF values more than 10. Now we dropped number\_of\_vehicles\_involved from the data and checked vif values again.

Now all values are less than 10, so there is no multicollinearity.

Next, we would check if the data is balanced.



We applied SMOTE technique to balance the data.



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.

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.



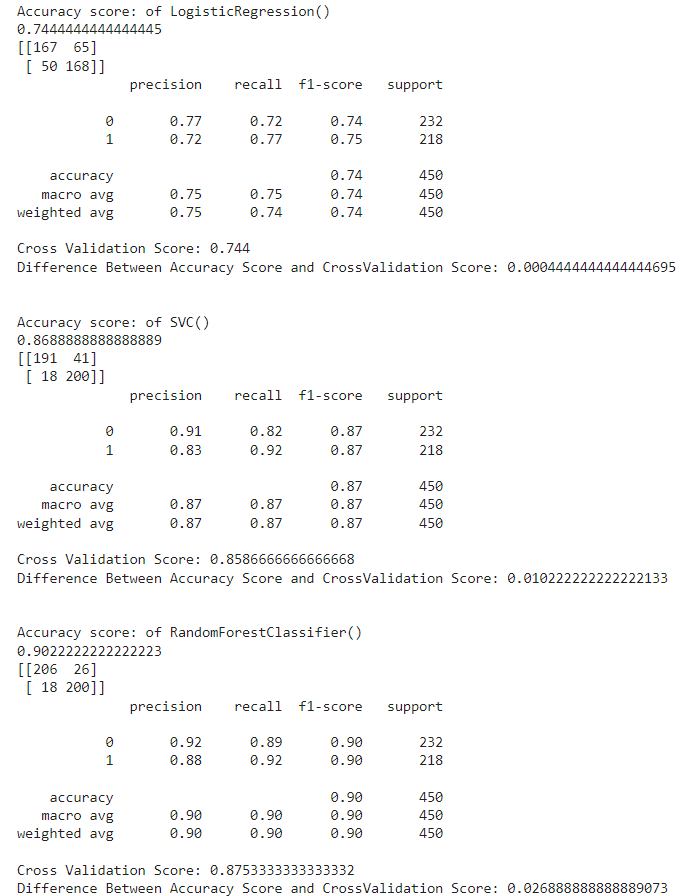
I have got best random state as 73 and best accuracy as 91.5% using Random Forest Classifier.

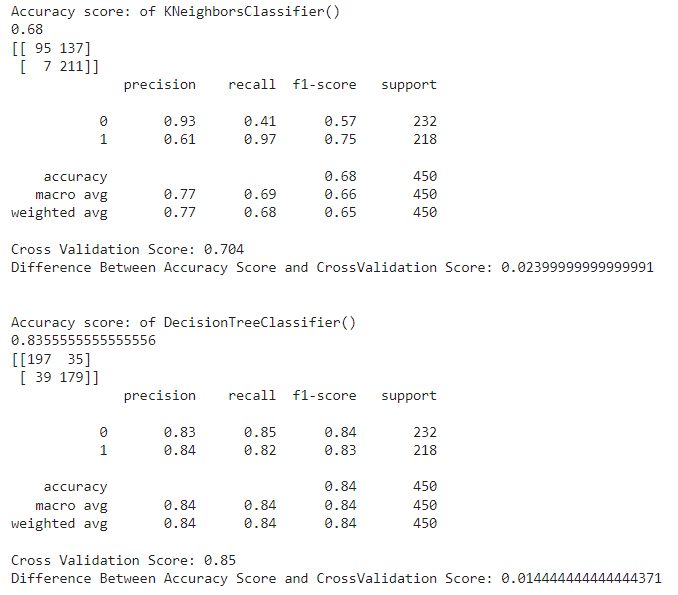
Now let’s create new train and test and fit them into the models to find our ideal model.



I have used evaluation metrics like classification report, confusion matrix, roc score and accuracy score. Also used cross validation score to get the difference from the model accuracy.

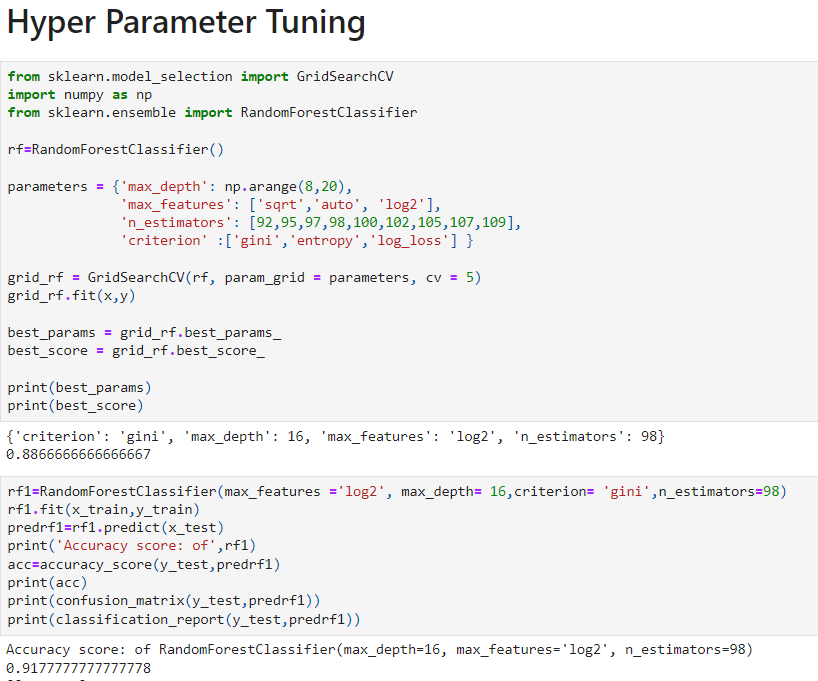




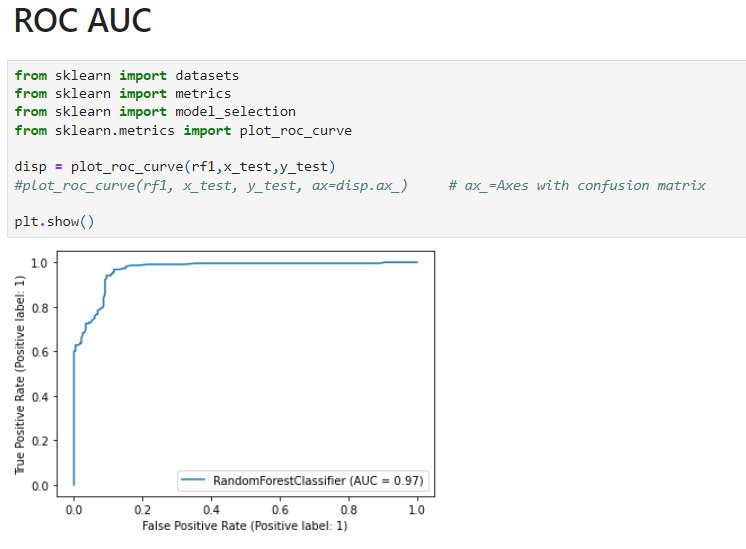


We observe that random forest classifier is performing better than other models.

By using below parameters, we have tuned the best model and after tuning we have to choose the best parameters from the list. Let’s get the best parameters.



Applying the best parameter values we get 91.77% accuracy score. Now we will plot ROC curve and compare the AUC for the best model.

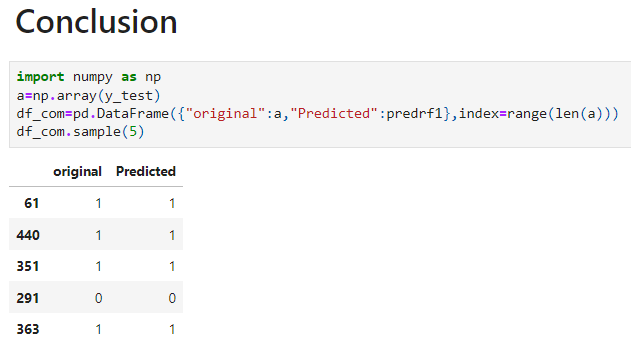


We see the score is 0.97 which very good, so we can save the model.

Saving the model –



Now I can predict whether the insurance claim is fraudulent or not.



The actual and predicted values are almost same, that means our model has good performance.

Concluding Remarks

From the above results of the data modelling and prediction we can see that the Random Forest Classifier Model is performing well as the accuracy score, cross val score and ROC score are good, also the maximum of the area under the curve fall under true positive rate. Therefore, we can save the model as .obj/.pkl file so that it can be used to predict the result of different data sets.

In this kind of problems Pre-processing and data-cleaning is the most important thing. We need to handle both the categorical and numerical data properly and also need to check by building different ML model on the same dataset. We need to check accuracy and cross val score of each model and chose the one which has the best of the same.