**Insurance Claims- Fraud Detection**

1. Problem Statement: -

**Business case:**  
Insurance fraud is a huge problem in the industry. Fraudulent claims are difficult to identify. Machine learning is in a unique position to help an auto insurance industry to tackle this problem.

In this project, we will analyse a dataset that contains information about insurance policies as well as information about customers. It also has information about the accident on which the claims have been made.

In this Analysis, 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.

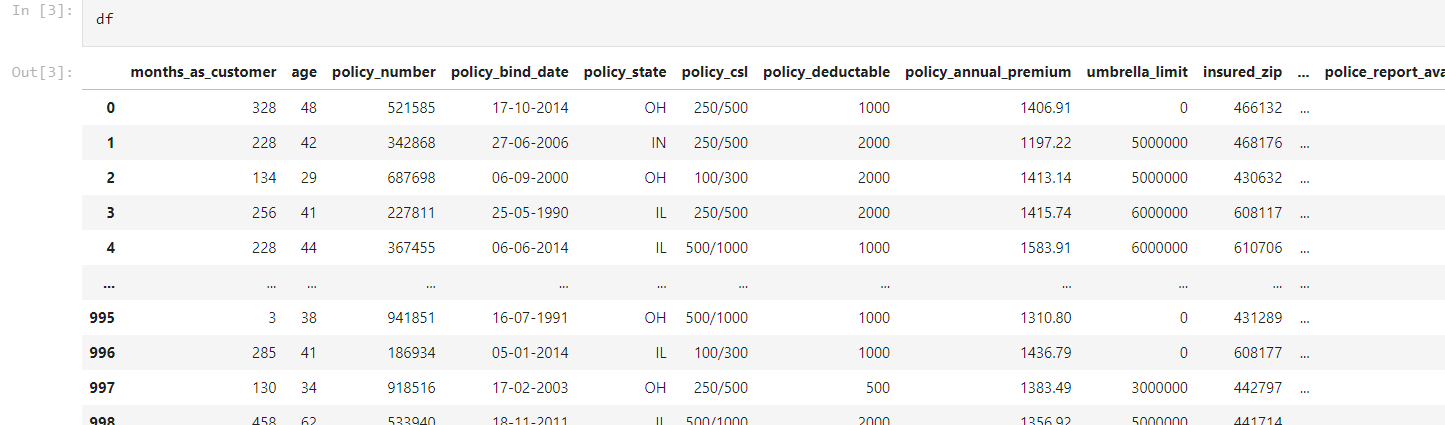
1. Data Analysis: -

Importing Libraries: - First, we will import few of the important libraries which will help us to analyse the dataset.

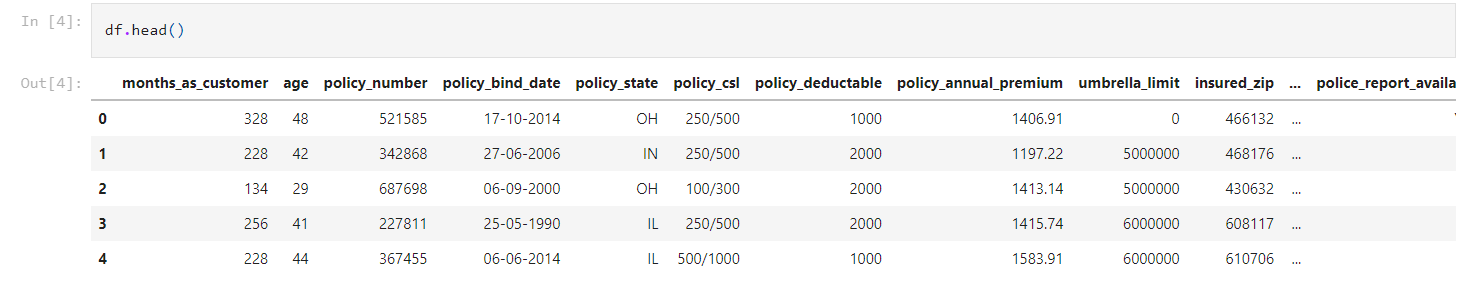
* import pandas as pd
* import numpy as np
* import seaborn as sns
* import matplotlib.pyplot as plt
* import warnings
* warnings.filterwarnings('ignore')

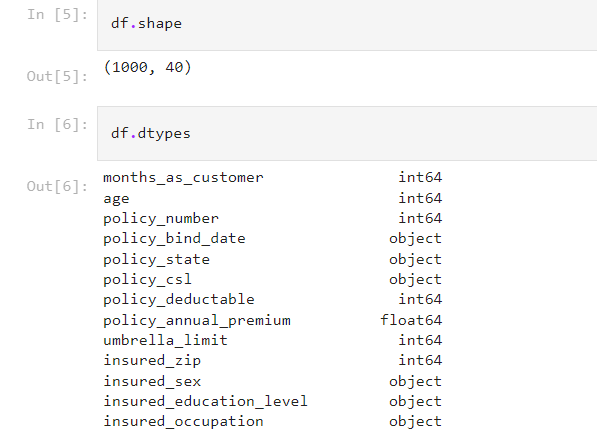
Getting the Data: - Here, with the help of pandas Library we will import the raw data.

* df=pd.read\_csv('https://raw.githubusercontent.com/dsrscientist/Data-Science-ML-Capstone-Projects/master/Automobile\_insurance\_fraud.csv')
* df

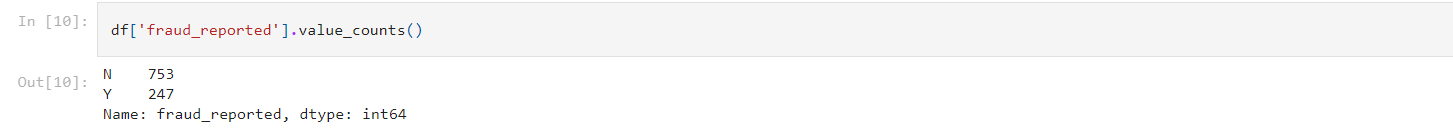


1. Analyze the Data and its Types:-

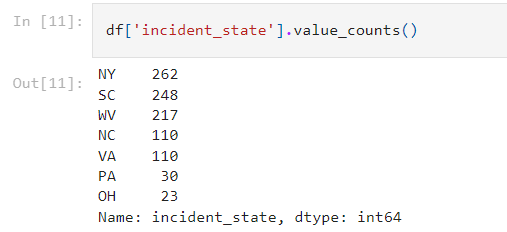




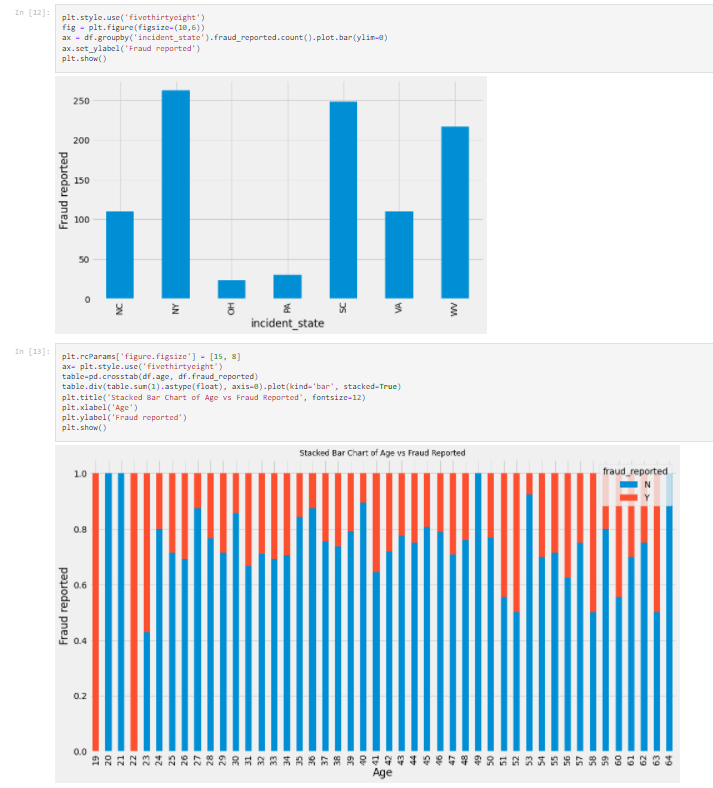


* 

\*\*Here we see that almost 25% fraud reported. Let’s try to look for an indicative variable. Let’s analyze location. This dataset only has information from the mid-Atlantic states from the USA.\*\*

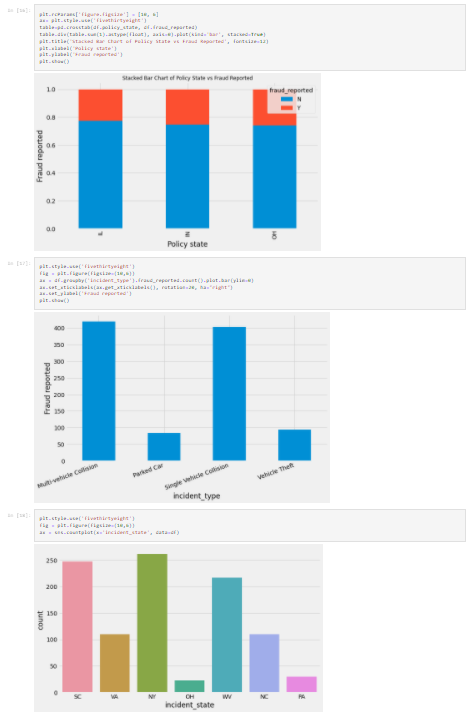


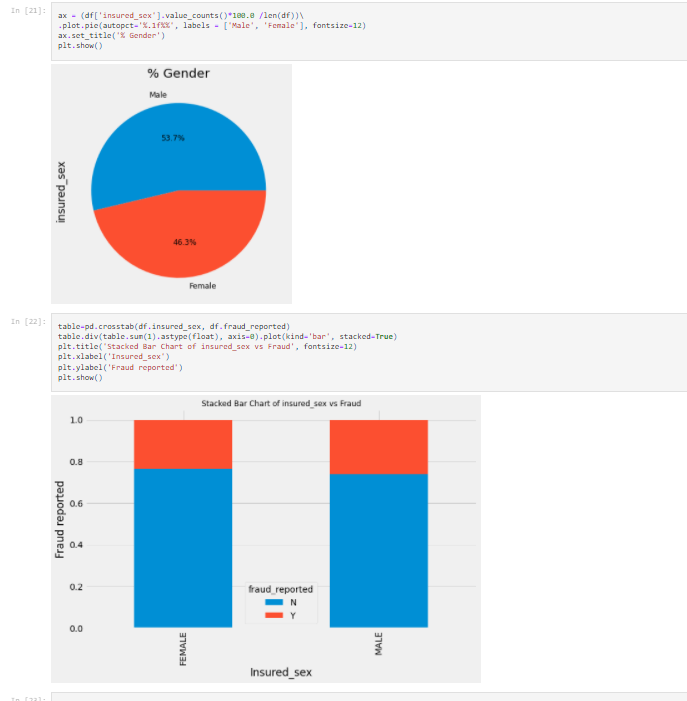
1. Data Visualization :-



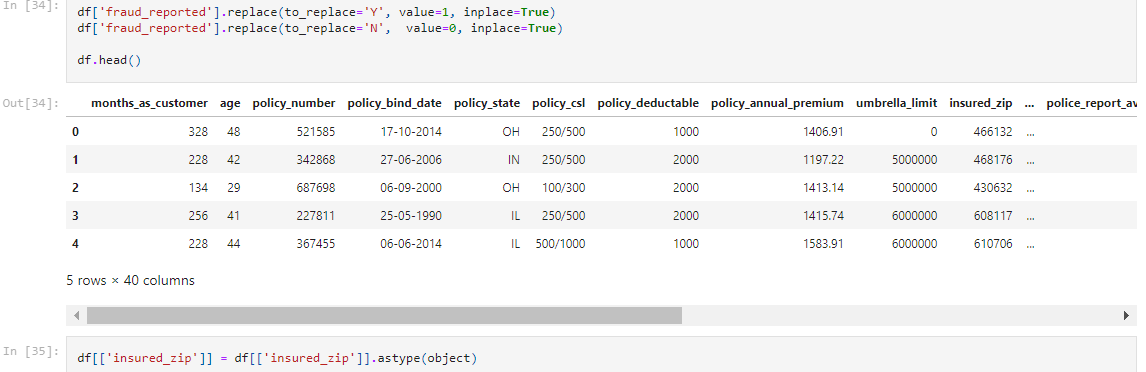
\*\*\* From above plot, it is obvious that, age is an important predictor for fraud reported. Age between 19-23 shows substantial number od fraud report.

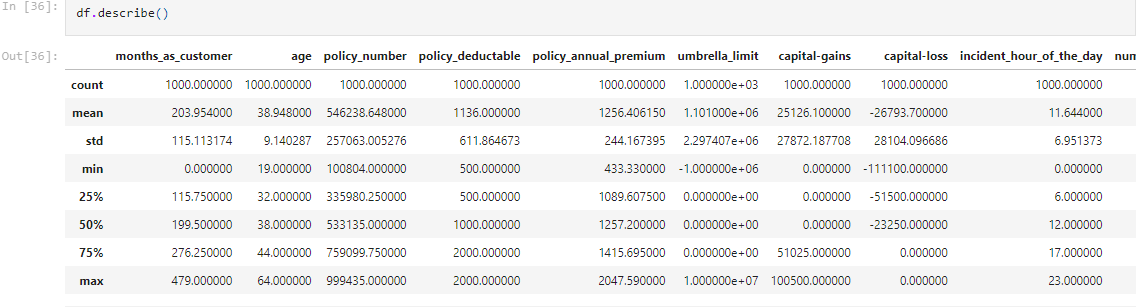






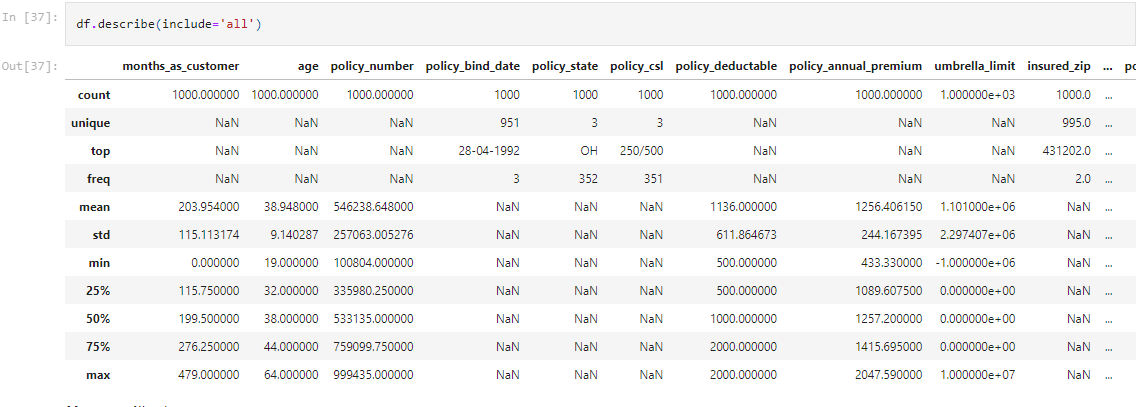
1. Data Processing**: -** Now we will clean the Data and prepare it for Machine Learning Model.

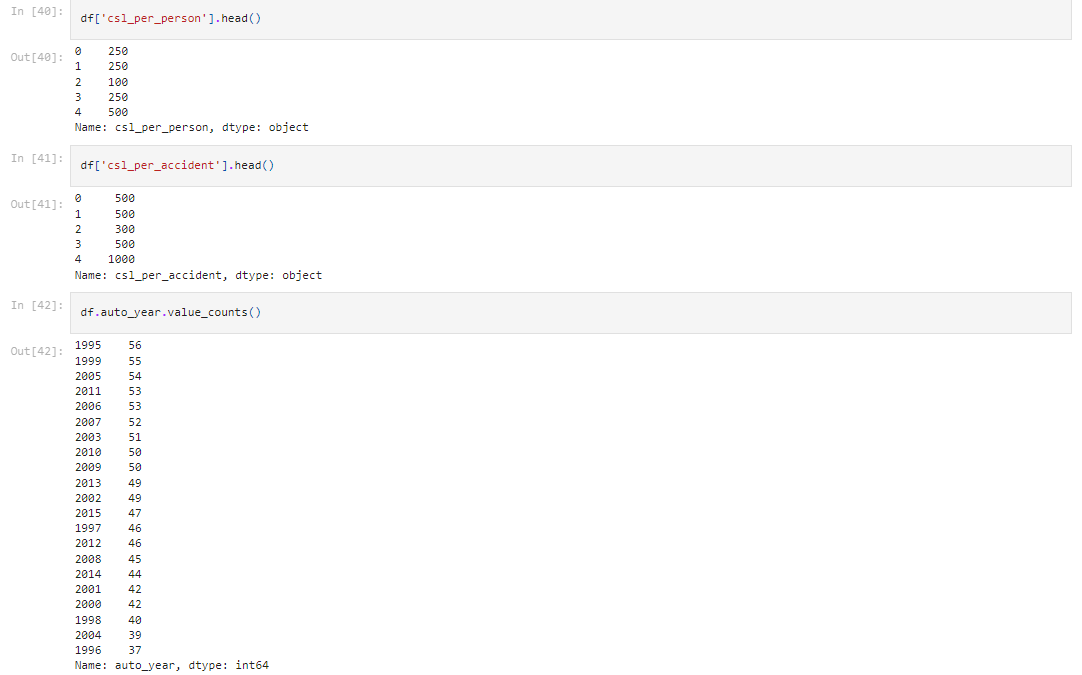




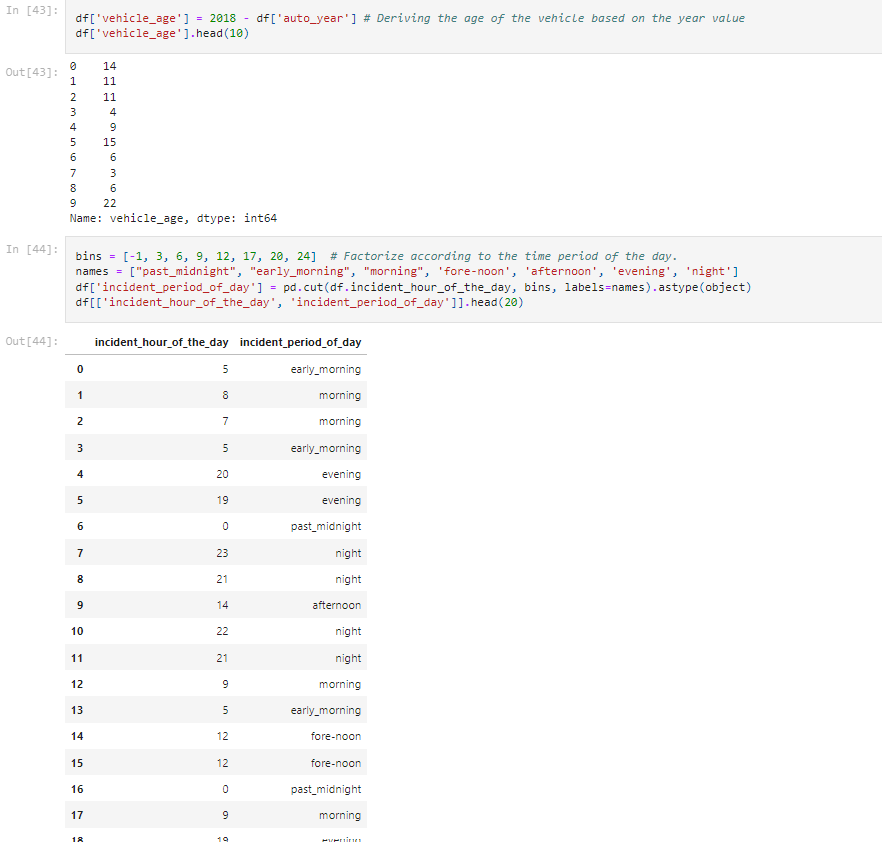
Note: - Some variables contain very high range. We will remove these columns for our purposes.

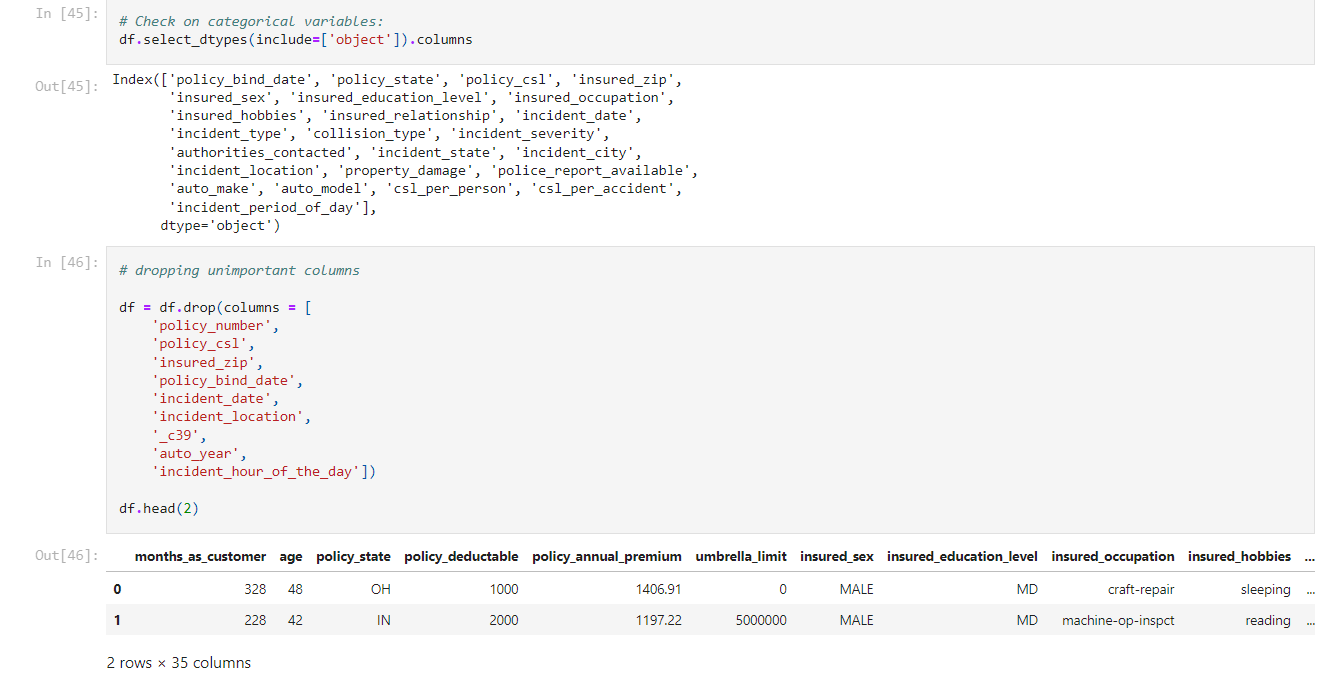
* Let use object data type to view the summary for all the column.

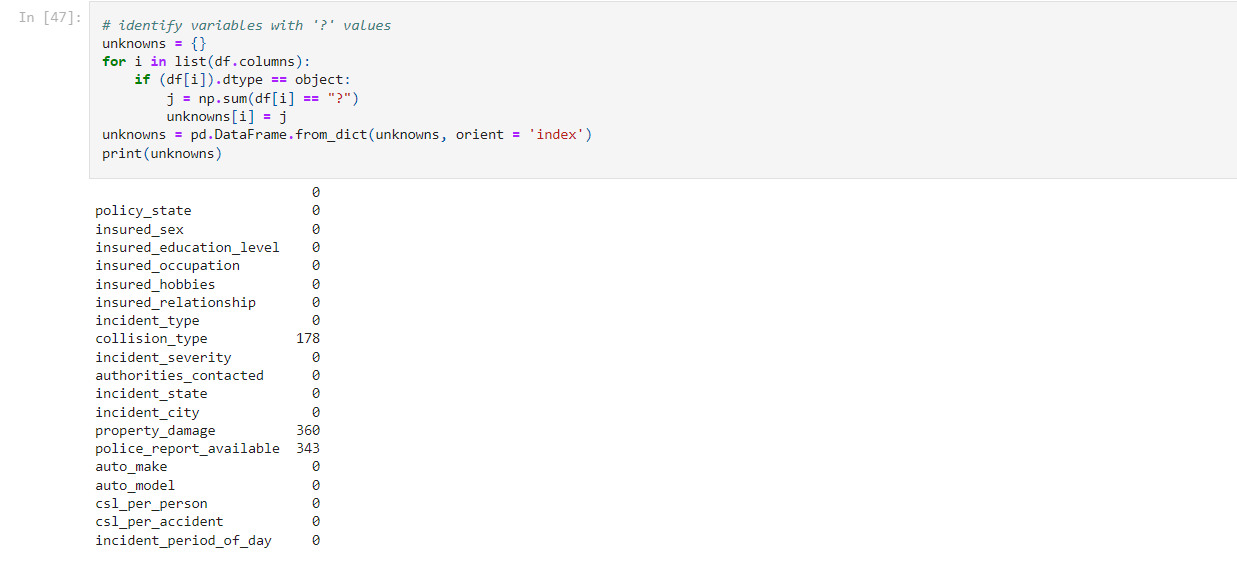




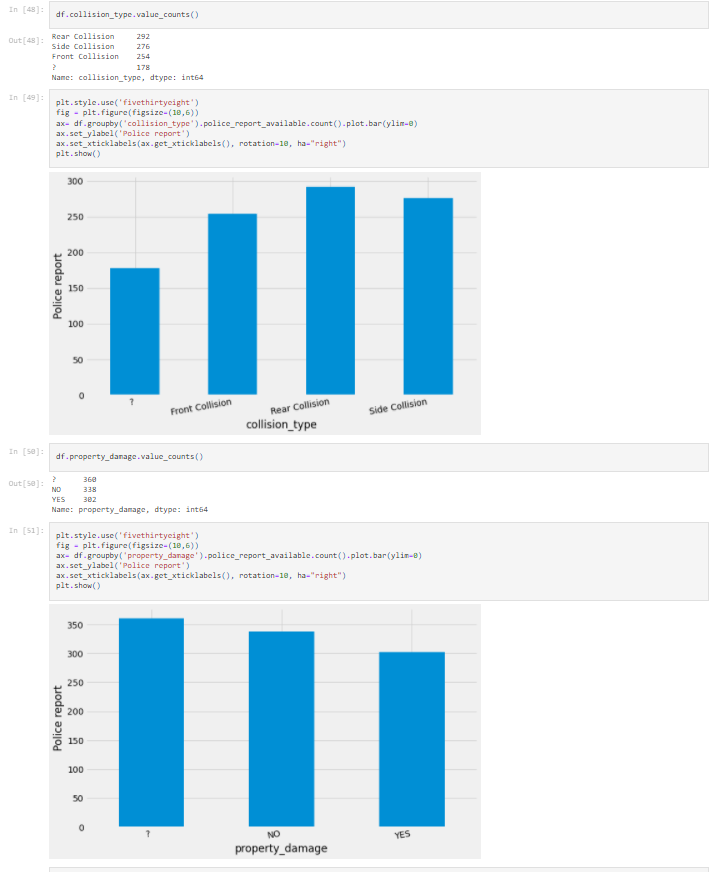
Note: -Here, Auto year has 21 levels and the number of records for each level is quite significant, given that the data size is not so large. We will apply some feature engineering by using these variables, given that the year of manufacture of the automobile shows the age of the vehicle and may contain valuable information about the insurance premium or fraud.

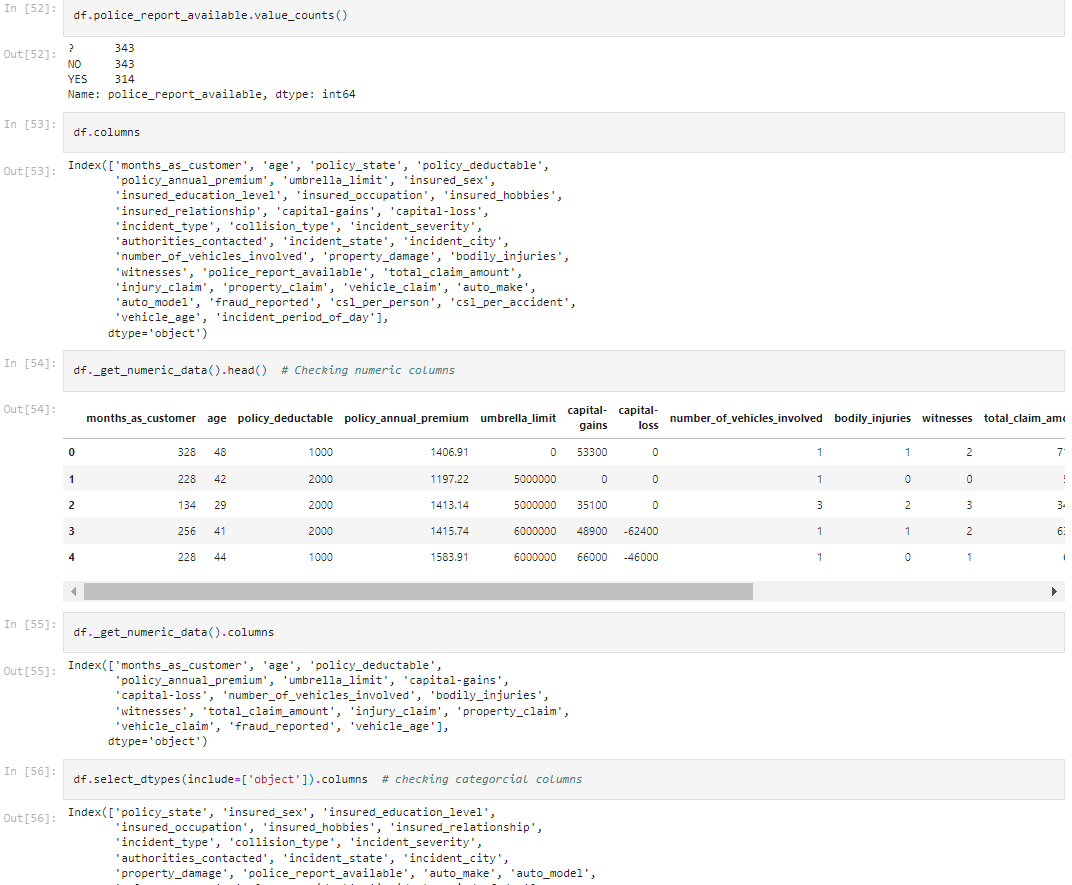






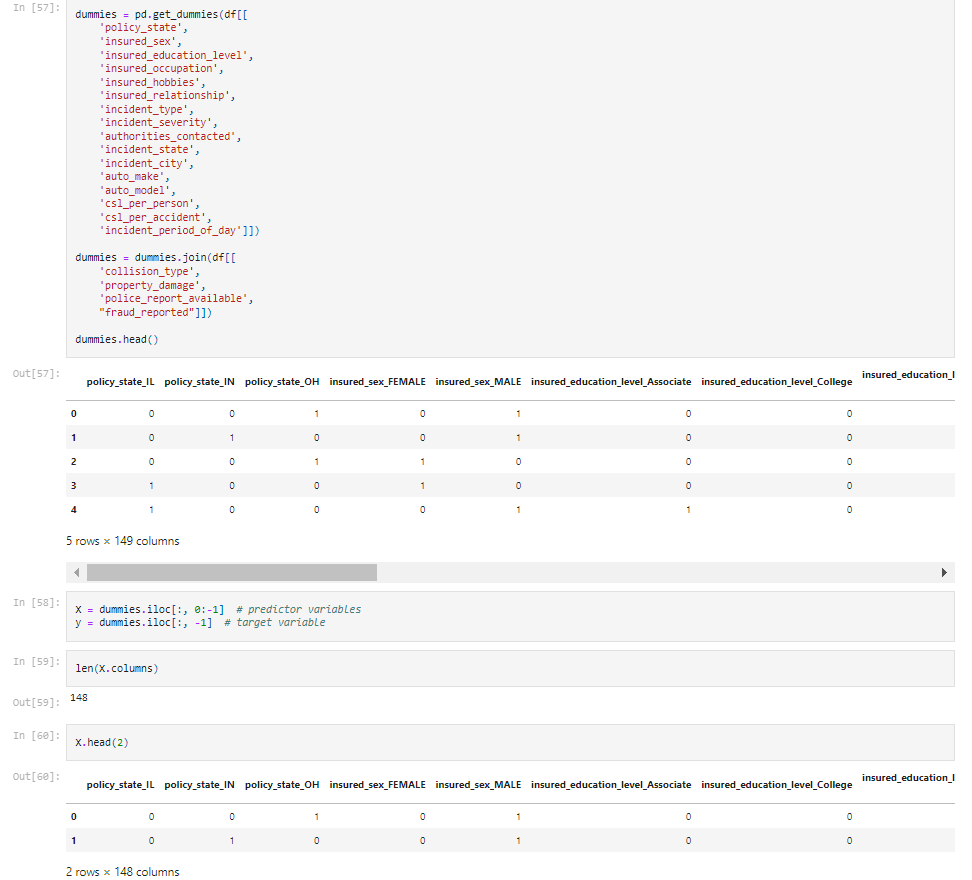
Note: - collision\_type, property\_damage, police\_report\_available contain many missing values. So, first isolate these variables, inspect these individually for spread of category values.

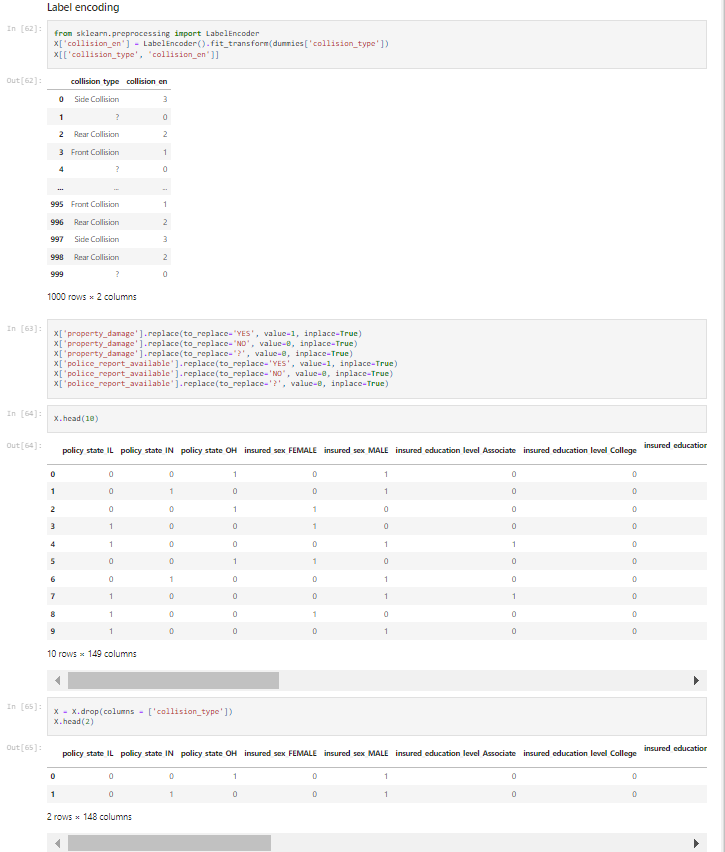




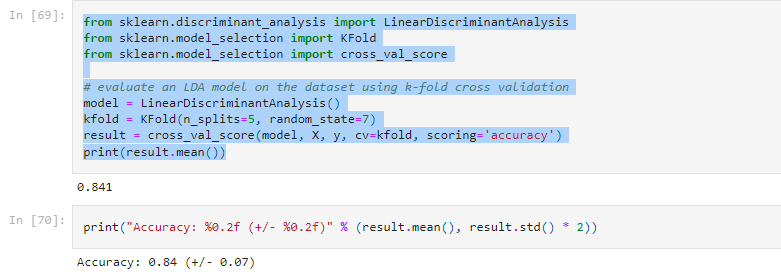
* convert all categorical variables except out target variables

'collision\_type', 'property\_damage', 'police\_report\_available', 'fraud\_reported'



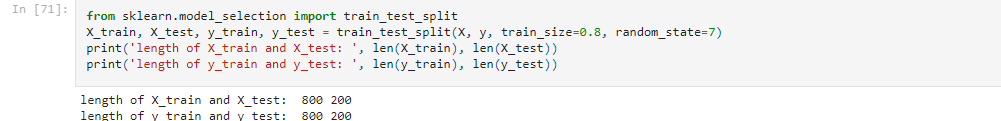


#### We now have a dataset that we could use to evaluate an algorithm sensitive to missing values like LDA.

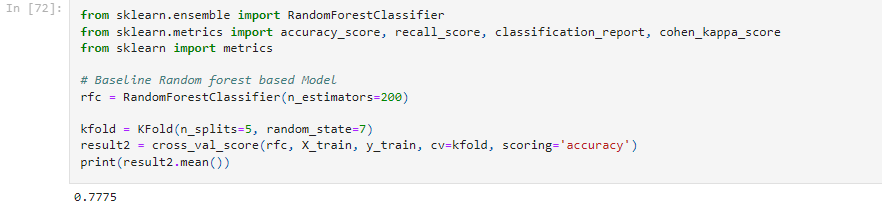


Note: Without Standardizing the Data it scored 84% Cross- Validation Score. Above it reflects 95% confidence interval of the score estimate and also the mean score. Data seems good, we can move forward for other classification methods.

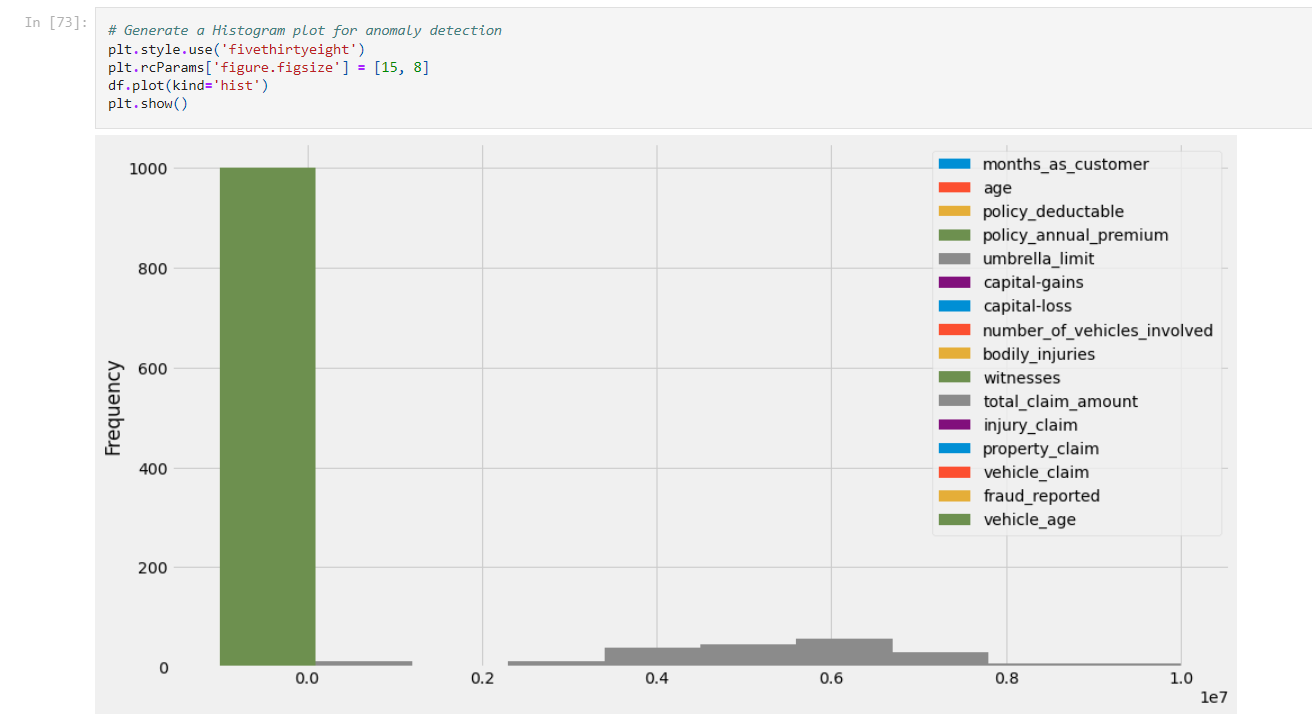
1. Creating a Training Set for the Data Set: -



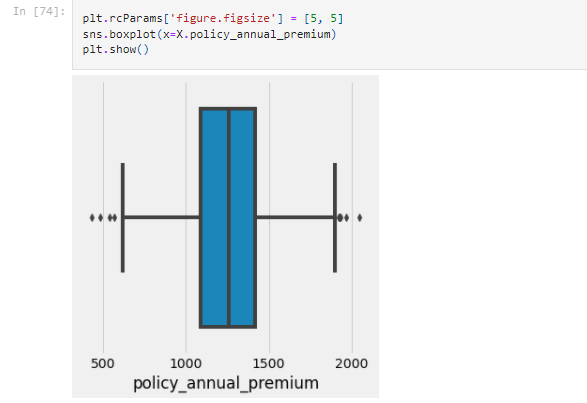
1. Random Forest Classification: -



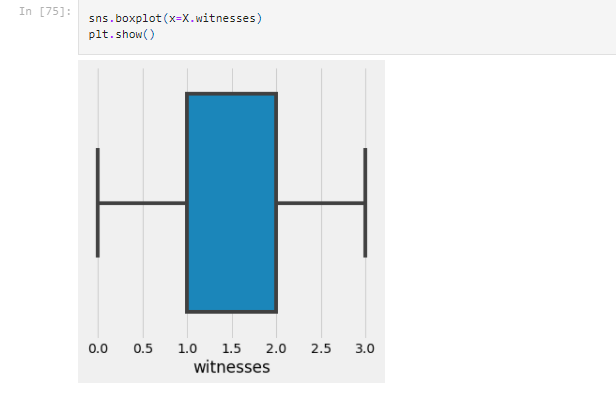
Note: - Here, we see that, Random Forest baseline model unable to provide greater accuracy. We will check on their classifier to compare. Before doing so, let’s check if any anomalies/outliers are present in data.



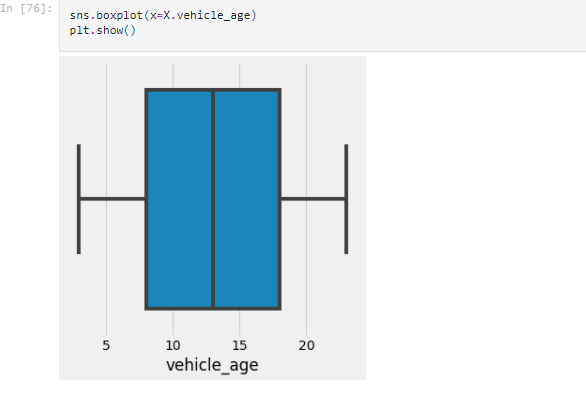
Note: - The green bar standing tall and away from all signifies anomalies in either of policy\_annual\_premium, witnesses or vehicle age. Let's draw box-and-whisker plot on each to check the presence of outliers.



Note: - Outliers are visible from the above plot from both Q1 and Q3 quartiles above the whiskers.

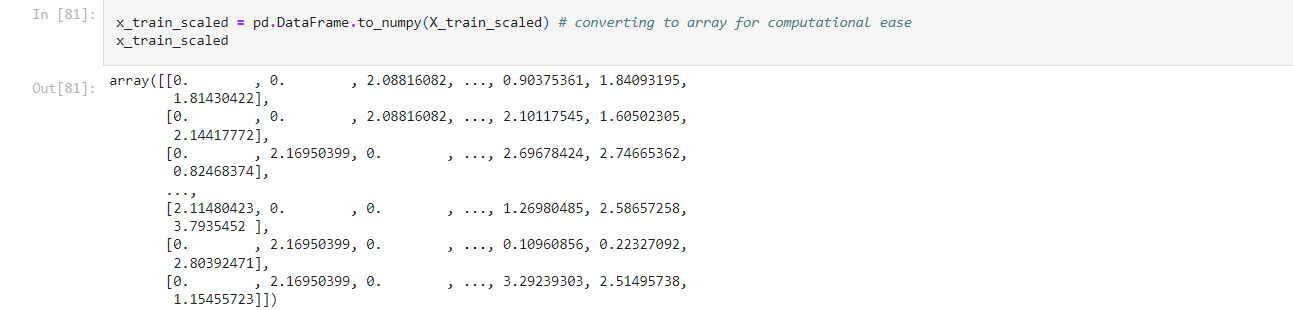


Note: - Missing median line represents data distribution is highly imbalanced.



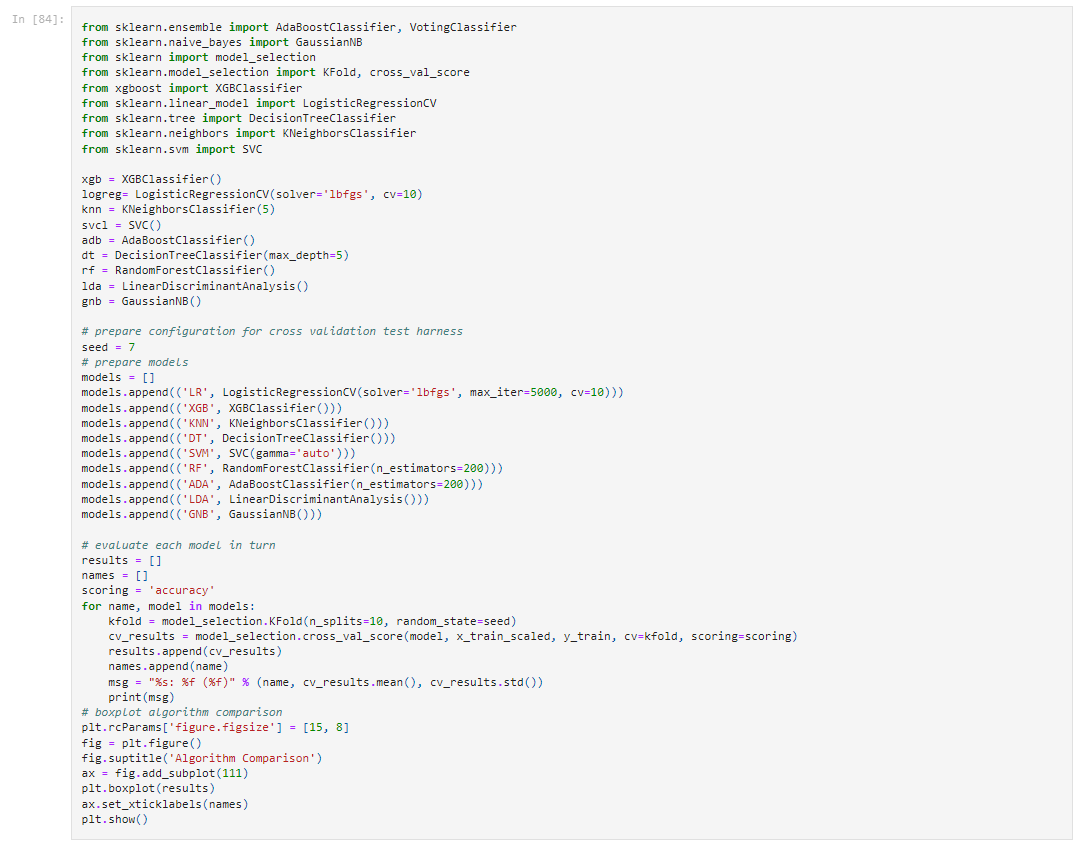
1. Standardizing the data and recheck the data distribution

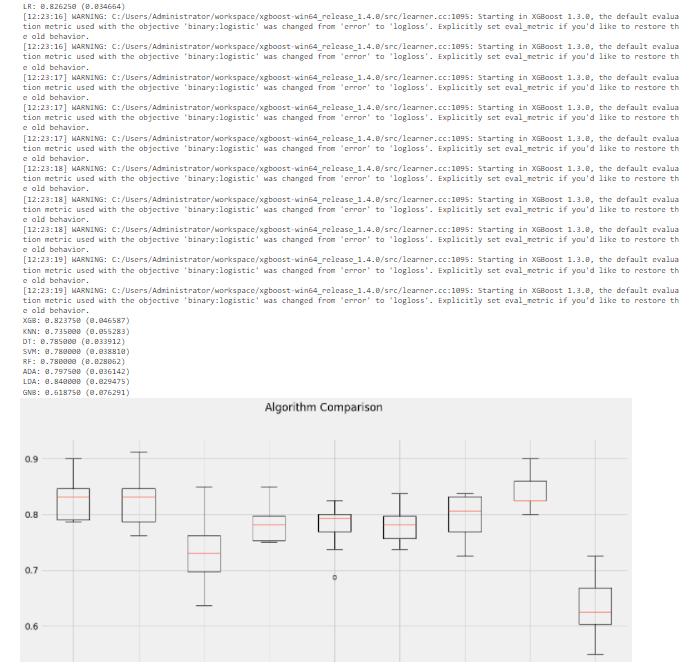




Note: - Here, data is distributed and the anomalies are gone after standardization.

* The 10-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way.



­

**Conclusion: -**

Above a list of each algorithm, the mean accuracy and the standard deviation accuracy and a box & whisker plot showing the spread of the accuracy scores across each cross-validation fold for each algorithm. It is clear that the linear Discriminant Analysis (82%) is leading the list. Logistics regression and XGB are almost close (82.62% and 82.87% respectively). We could see some noise / outlier in data in case of XGB. The LR box-plot is skewed one side with longer tail.