



Submitted by:- Ayush Pathak

**ACKNOWLEDGMENT**

I would like to express my gratitude towards Datatrained for giving me this opportunity to show case my talent and also for their constant support and guidance. Also, I would like to thank all data-trained support member’s for helping me all the time whenever needed. I express my deepest thanks to **Deepika Mam**, for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make my project easier. I choose this moment to acknowledge her contribution gratefully. I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Thanks & Regards

Ayush Pathak

**INTRODUCTION**

* Business Problem Framing

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, you are provided a dataset 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, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

* Conceptual Background of the Domain Problem

We are going to test multiple classifier algorithms with multiple evaluation metrics and select the best model based on proper metric, perform GridSearchCV for best parameter settings.

* Review of Literature

First of all data is loaded and then it can be observed that, this is a classification based problem. It is clearly visible that the dataset needs not that much of cleaning as there are no garbage & outliers values present in our dataset.Note, there’s missing value(null values present) in one column only, whole row of that column is empty.

* Motivation for the Problem Undertaken

As, after looking into the dataset, it is clearly visible that this pretty straigh forward but handling such datasets brings your perception, judgemental ideas and curiosity to the next level.And, after working on this assignment many challenges I have faced, that boosten up my confidence too and a lot more things I have to learn also.

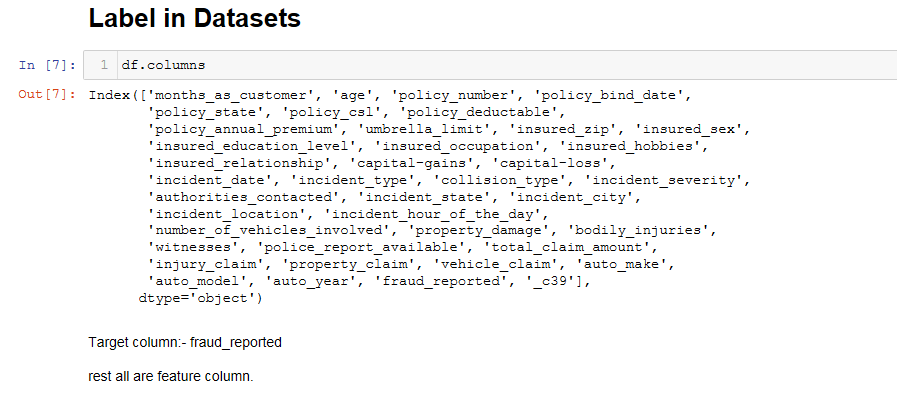
**Analytical Problem Framing**

The sample dataset is provided to us from our project database.

**Data Analysis:-**

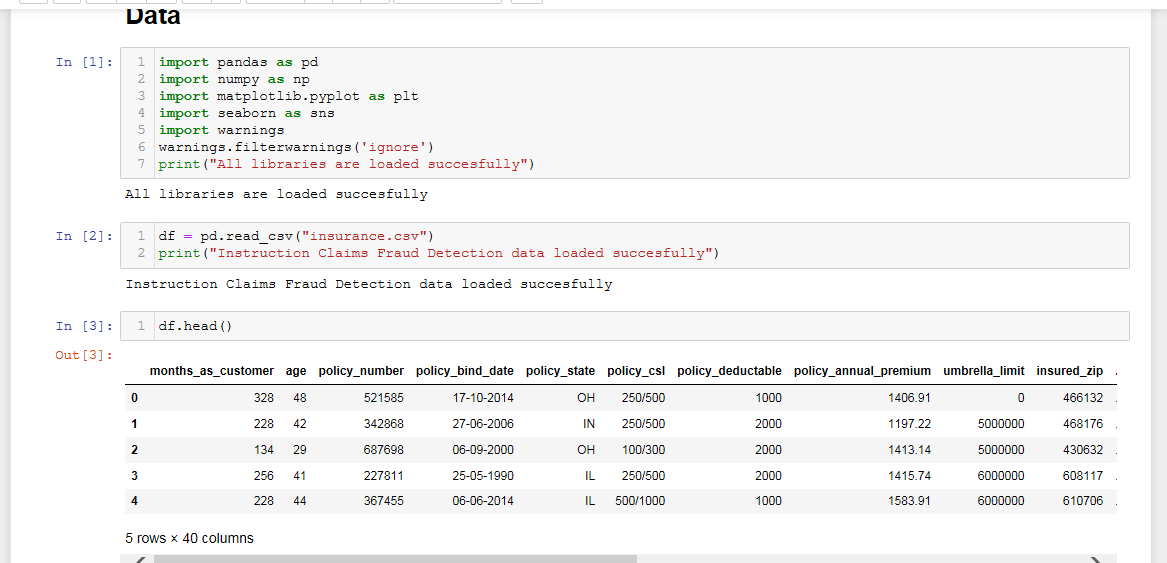
Source:   
<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile_insurance_fraud.csv>

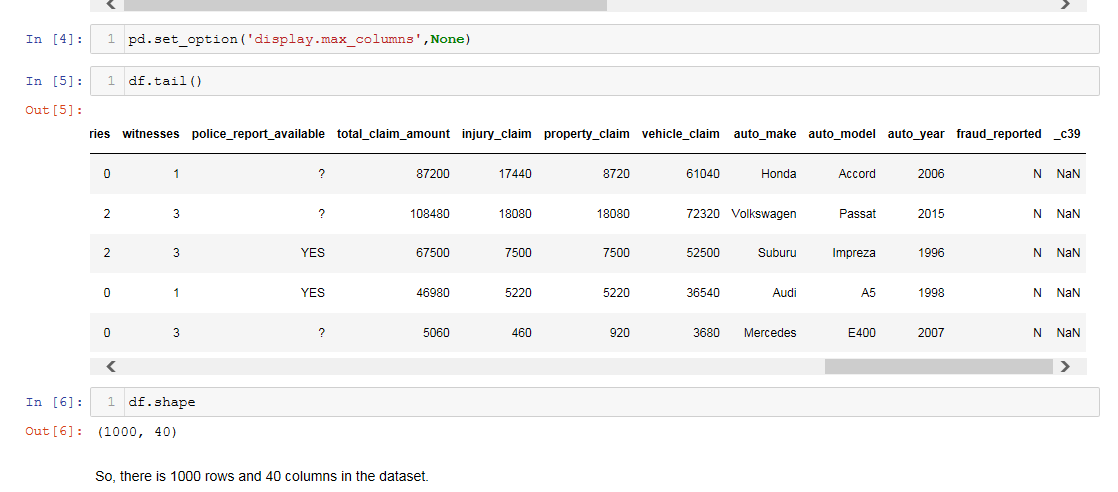
The data set have 38 independent variables and 1 target variable, i.e. Baseball dataset.We have 1000 rows and 39 columns in the dataset. The column names are as follows:-



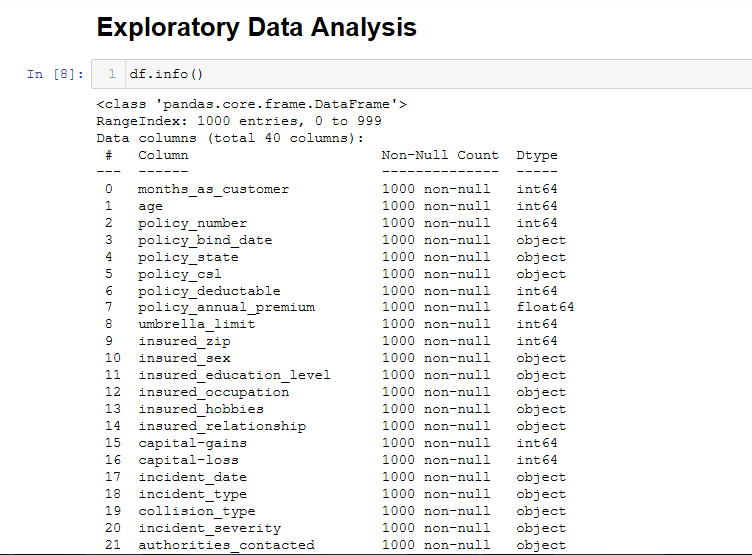
**Steps-**

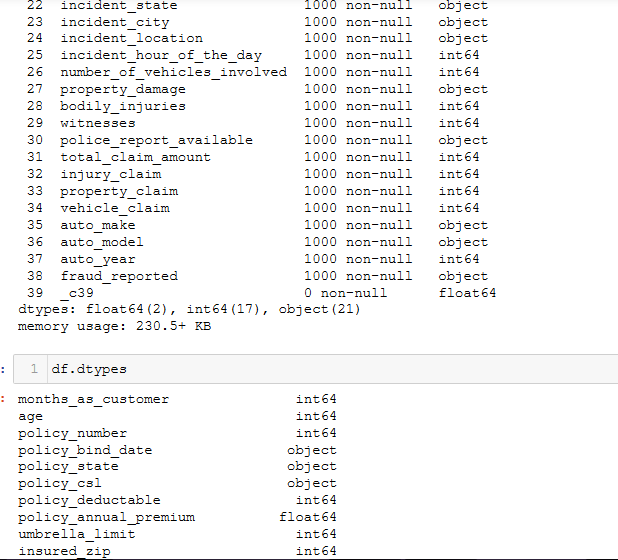
1) Firstly data is being loaded and previewing the first 5 rows or 5 tail rows.



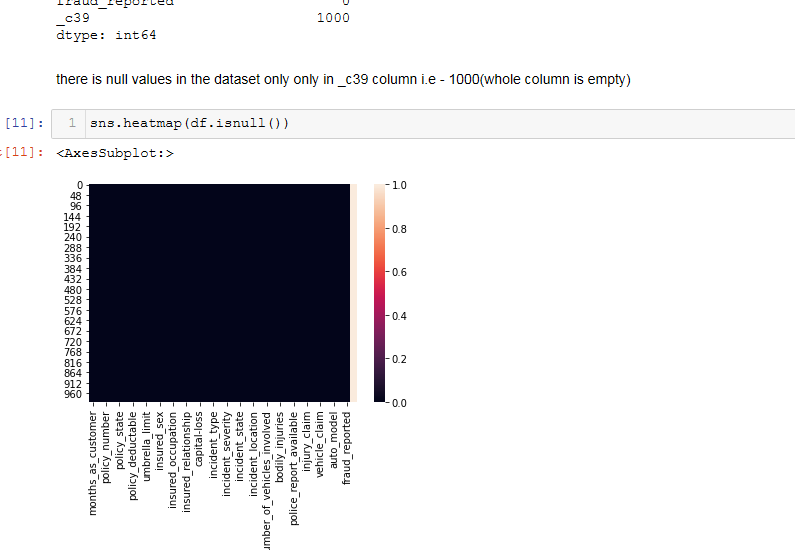


2) Analysing the columns present in the dataset and checking the basic information like data types(if null values are present or not).

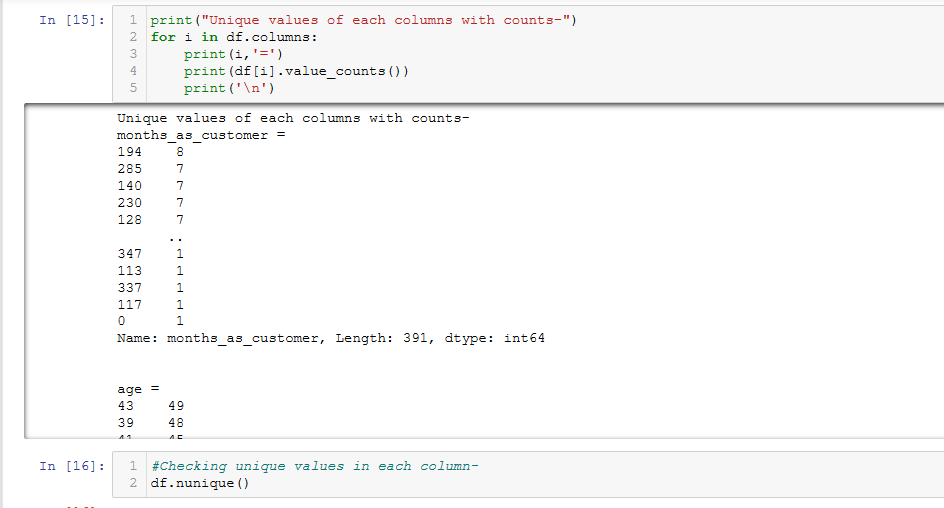




3)As, there is null values are present in one column i.e \_c39 whole column rows are null values(NaN). So, we’re going to drop this column from our dataset.



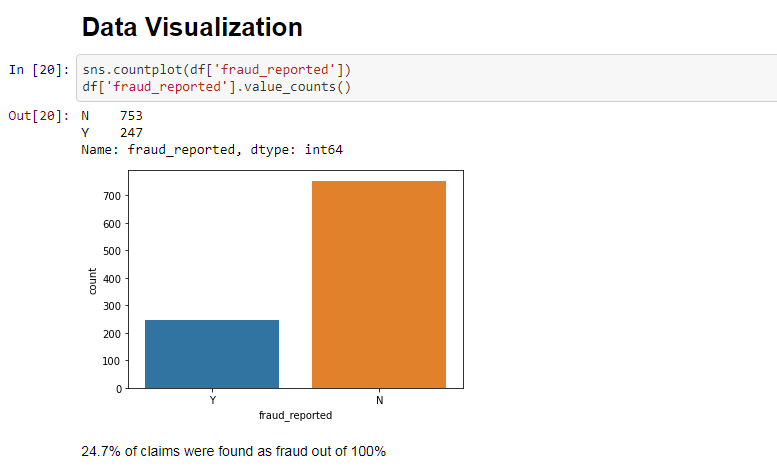
4) Checking uniqueness of each column with counts.



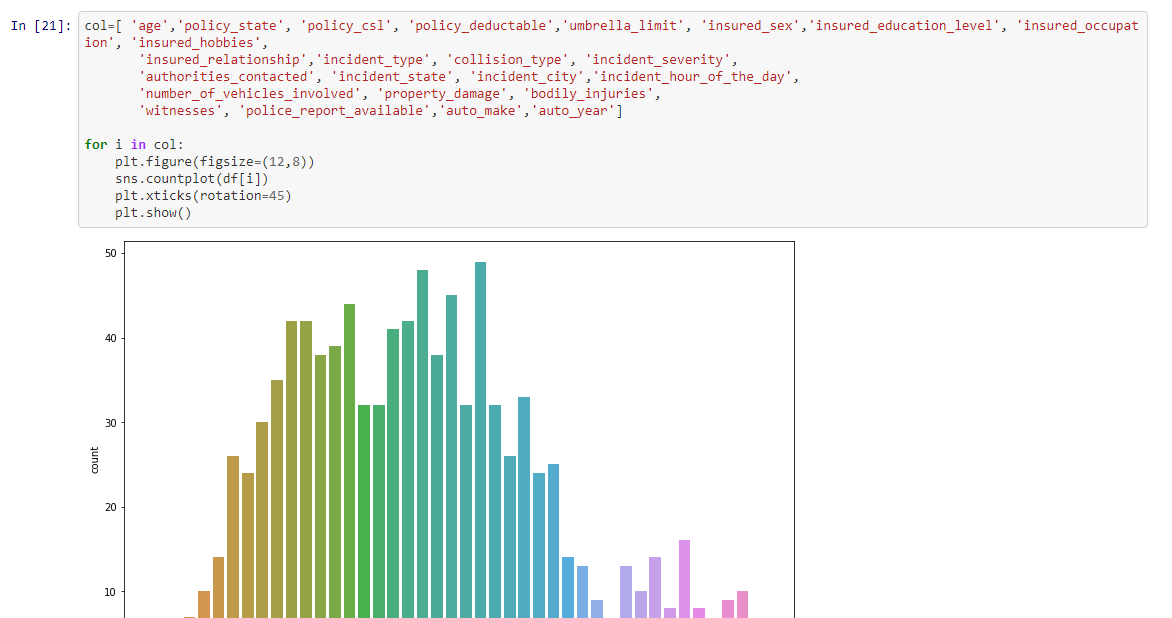
From the obervations we got, All the policy number, incident locations are having different data and also,policy\_bind\_date,policy\_annual\_premium & insured\_zip are having maximum no. of different data.Since, they are all unique for each row. we are dropping these columns. After this, dataset is with

1000 rows & 34 columns.

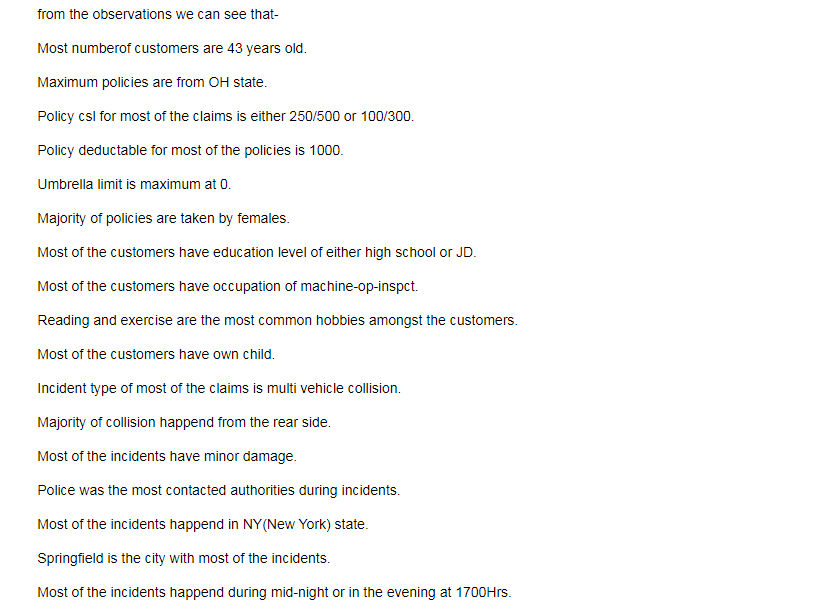
5)Visualizing the target variable(fraud\_reported), we see 24.7% were found as fraud in the dataset.



6) Data Visualizing for some columns, exploring univariate analysis in the dataset.

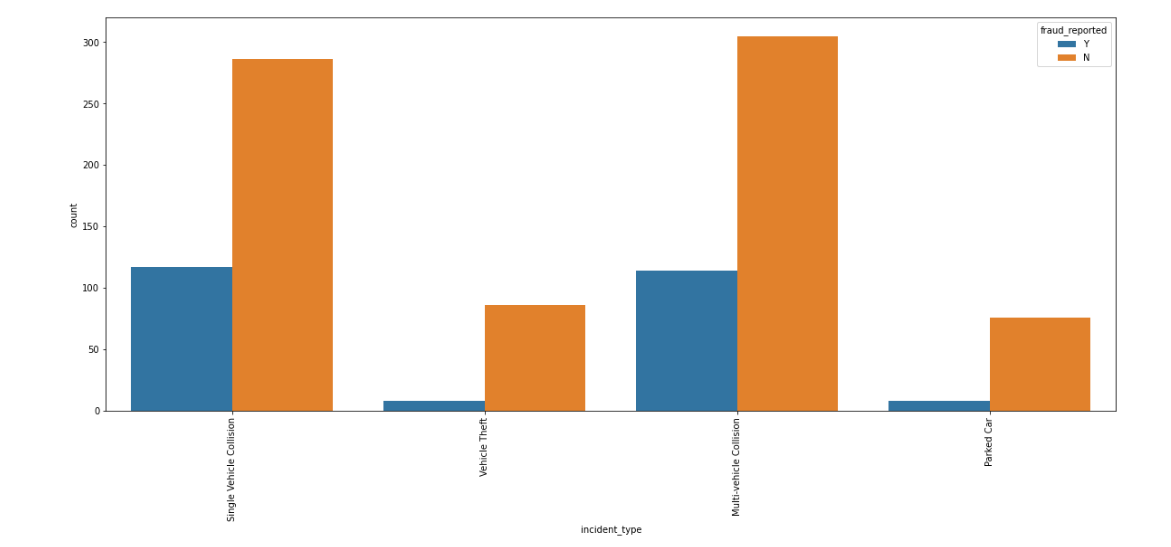


Similarly plotted for all the columns mentioned above in code.



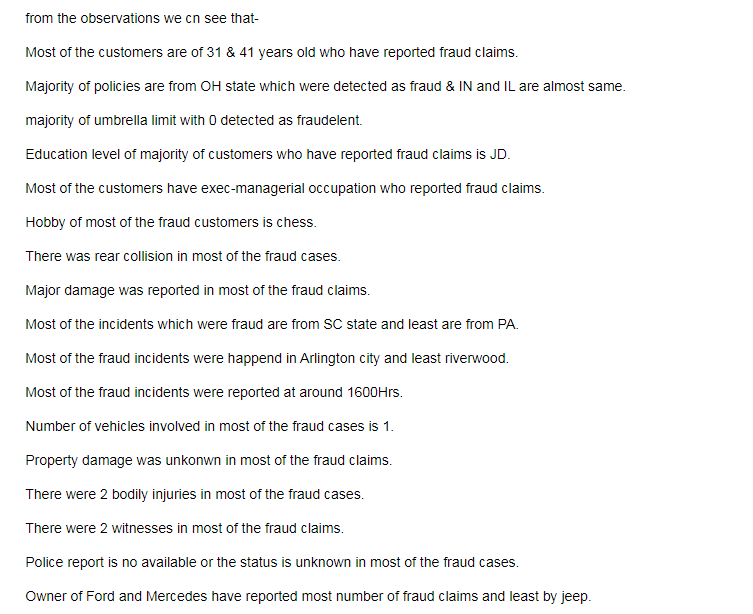
7) Now we are again going for data visualization using count plot.We are working here on bivariate analysis of dataset, with reference to target variable for each independent column.



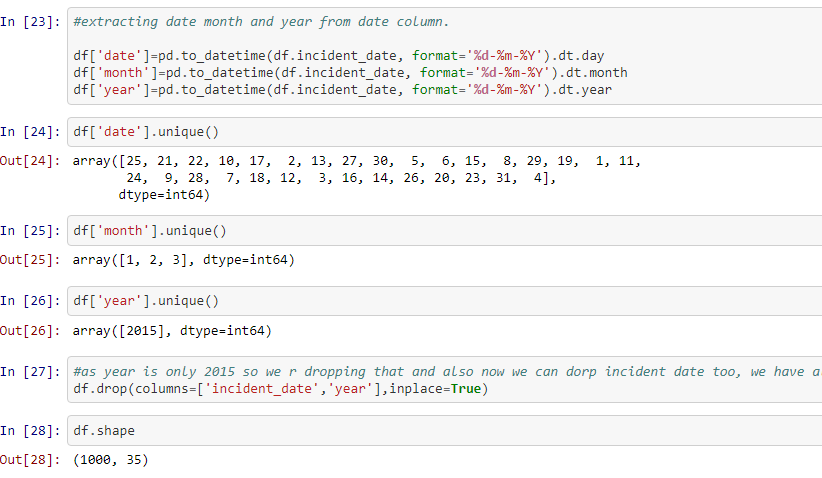


Similarly plotted for all the other columns

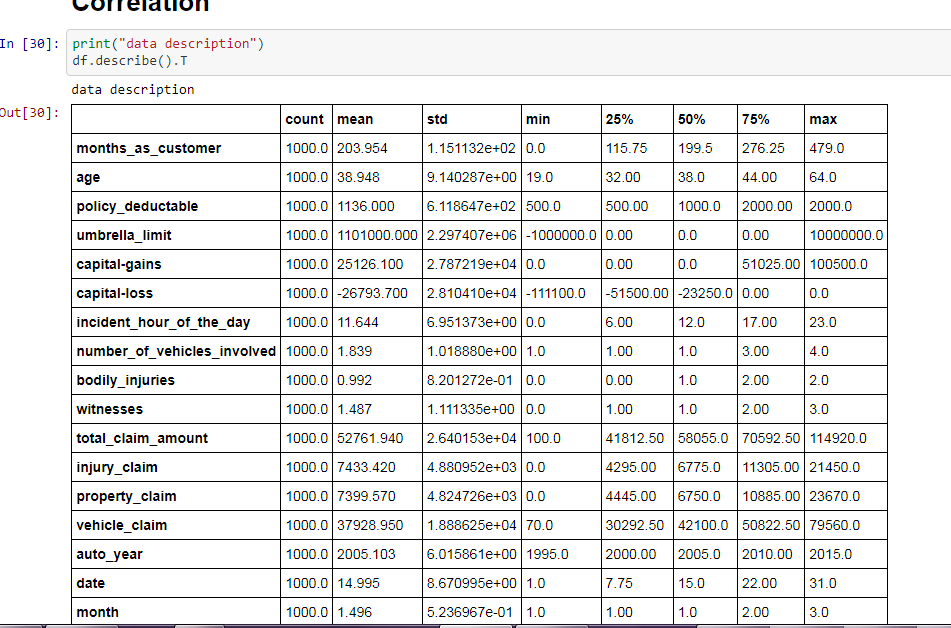
Observations we got after this-

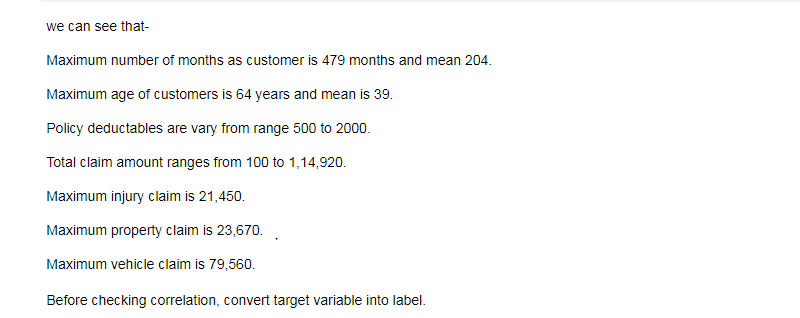


8) Before correlation, extracting incident\_date into date, month & year separately.Then dropping off incident\_date and year(same for all), from the dataset.

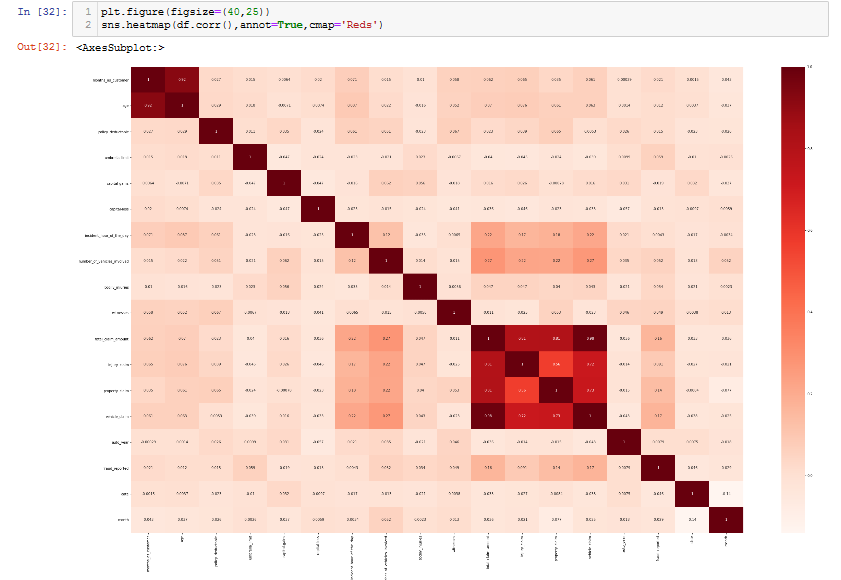


9)Checking the statistical description of the dataset. This report’s helps us to decide which transformation will improve the model’s accuracy,outlier and range of particular column.

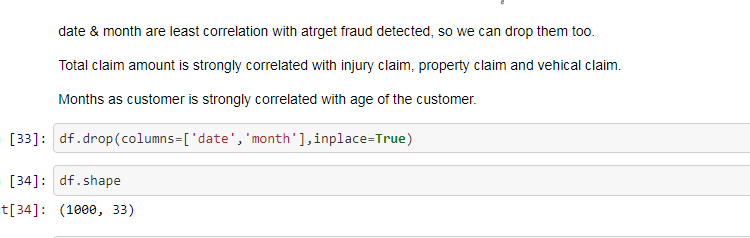




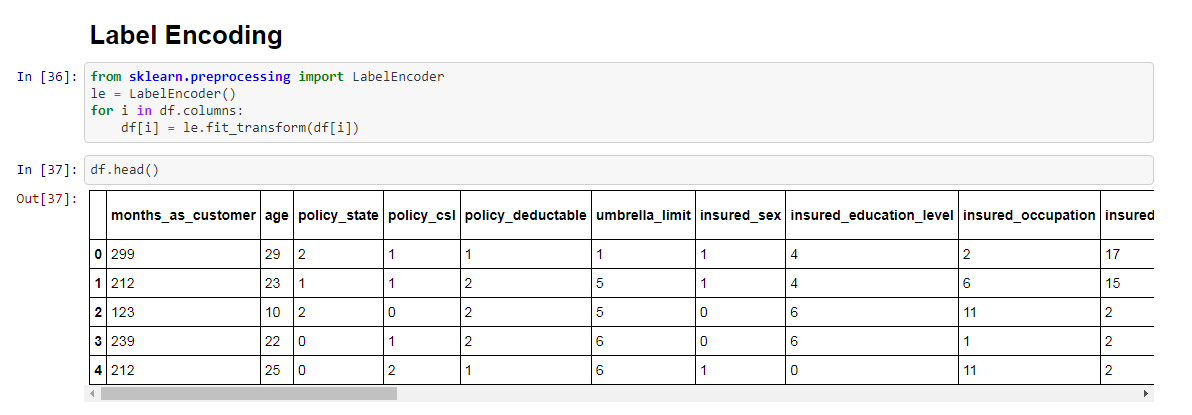
10) Now we are checking the correlation by using heatmap from which we can figure out some columns that we need to drop on the basis of correlation factor.



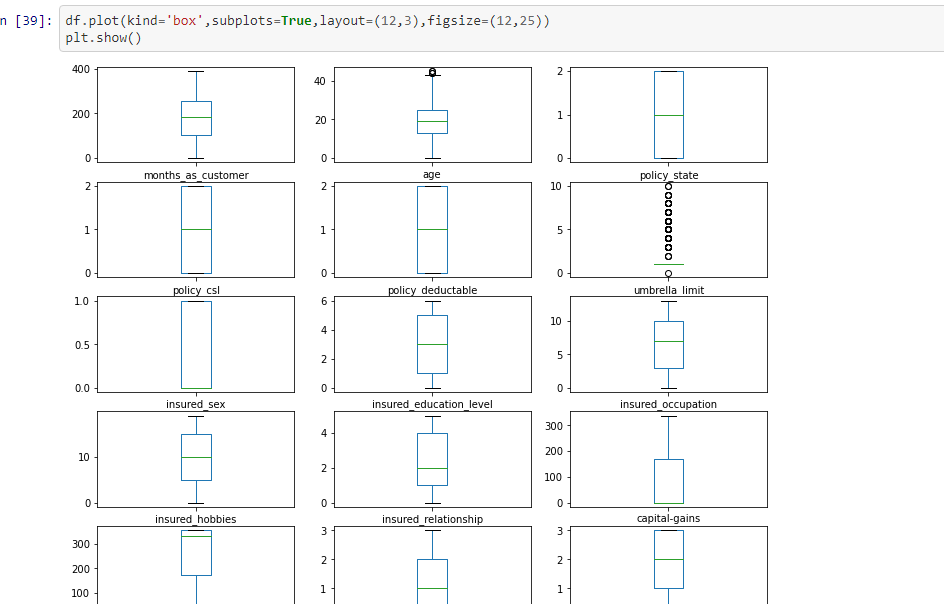
Observation we got after this-

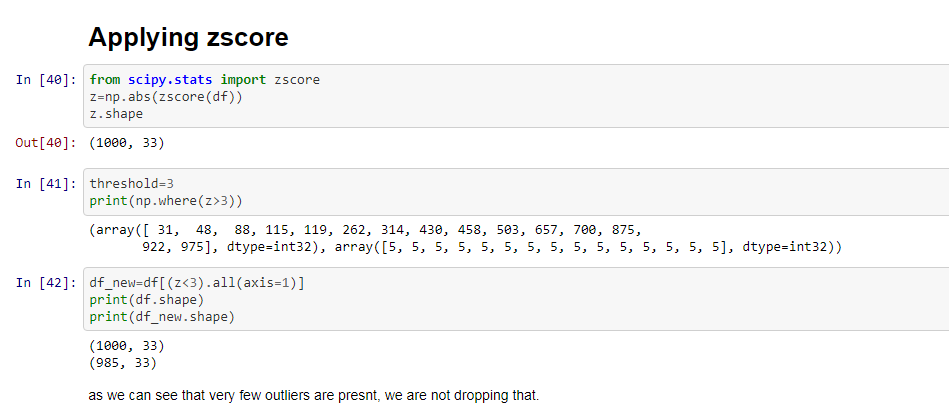


11) As, still some categorical columns are present in the dataset.Therefore Label encoding we are enforcing here.



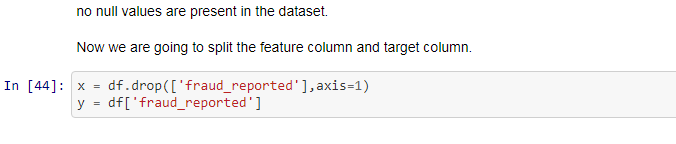
12)We’ll go for data cleaning, firstly going to check outliers through plots and then by applying z-score(factor3) on the dataset.



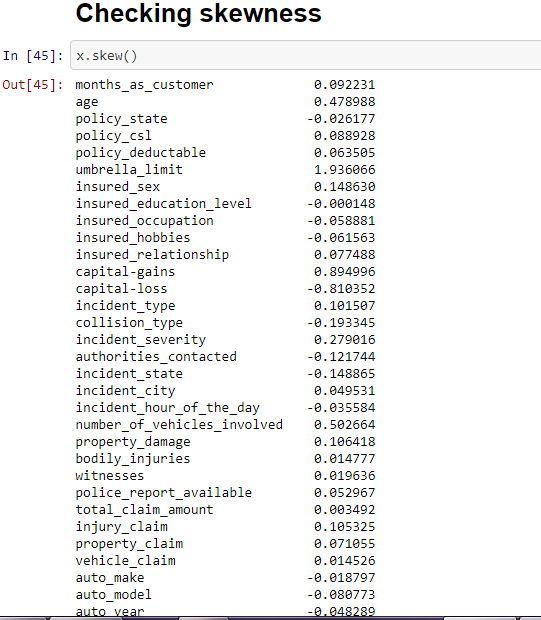


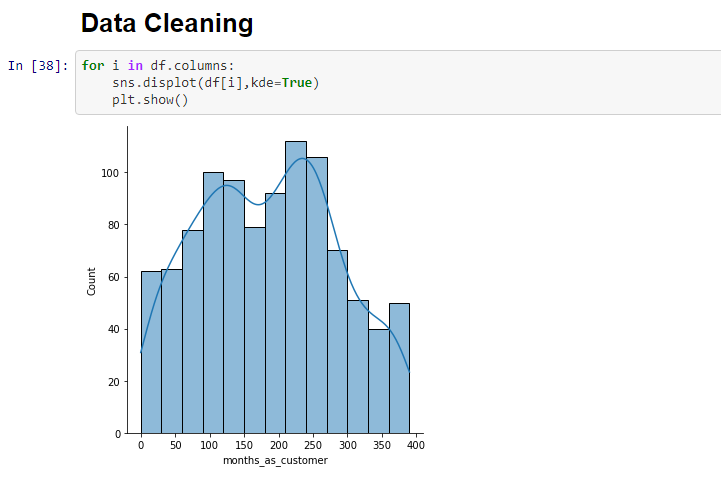
No outliers as such present in the dataset.Very few that can be neglected.

13) Let’s go for dividing the feature and target, separating the input and output variable.



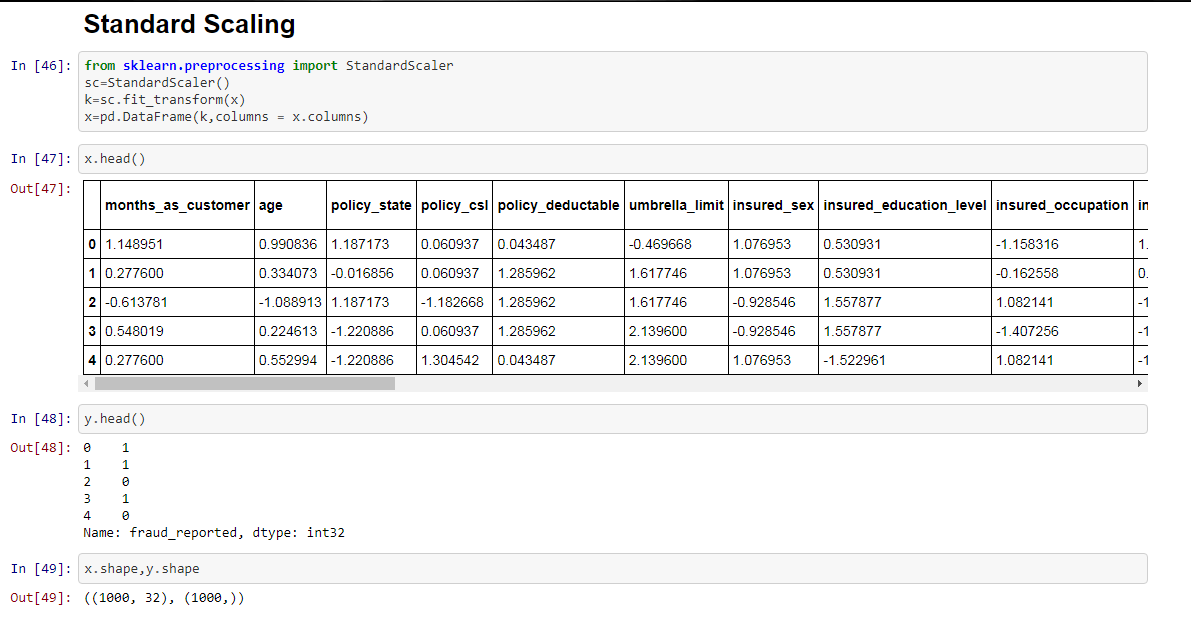
14)While checking skewness, we find that skewness is acceptable & also plotted distplot for columns during data cleaning





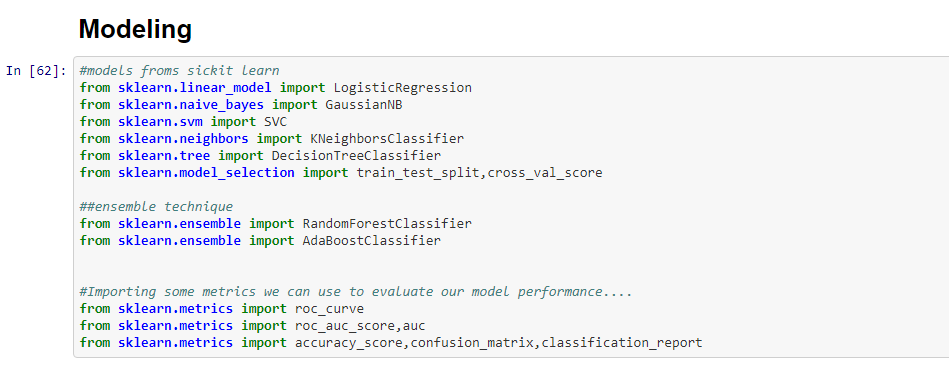
Similarly plotted for each columns-

15)Standardizing the input dataset(Standard scaler makes the mean of the distribution 0).

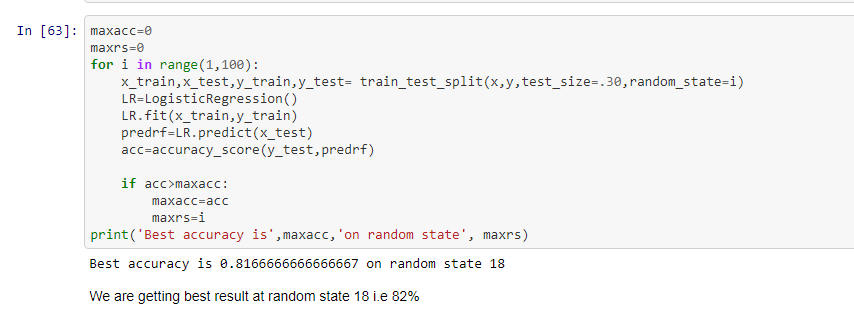


**Model/s Development and Evaluation**

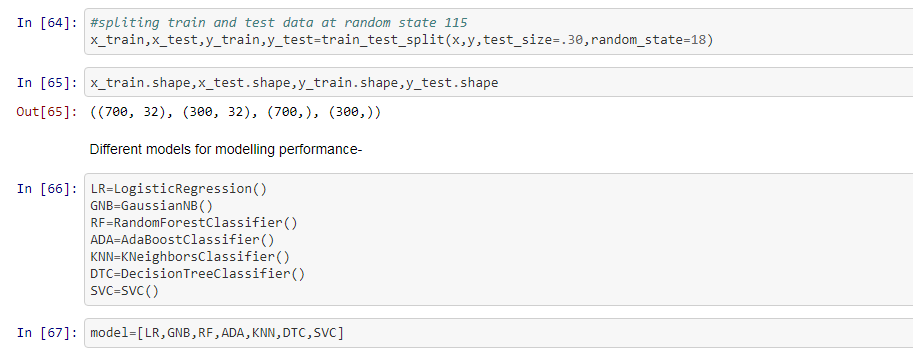
1) Loading all the libraries required for modeling as well as evaluation metrics.



2) Find the best random state with best accuracy score for modelling.

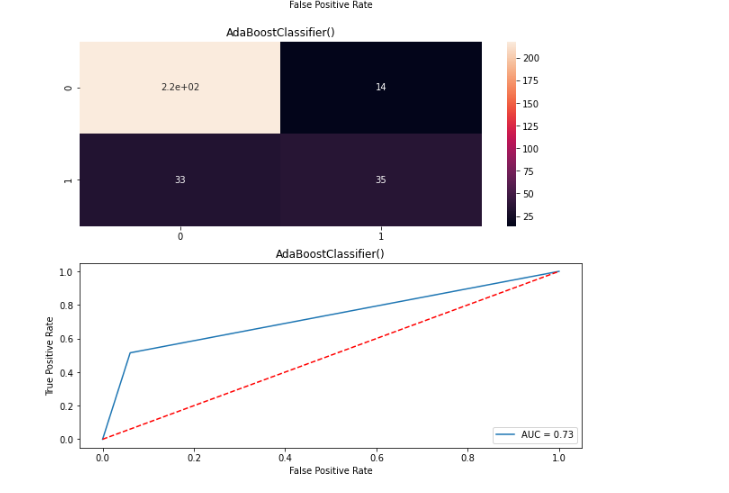


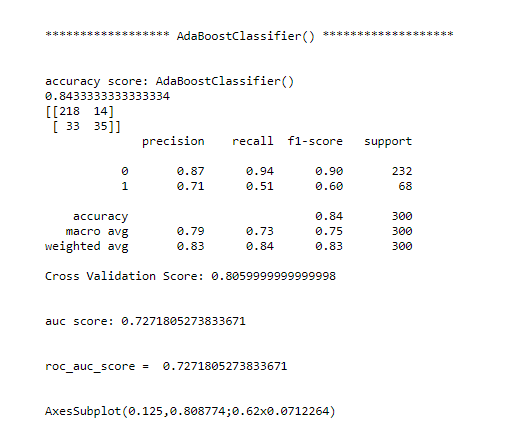
3) Basic models were chosen to check the best performing model and splitting into training & testing data’s.



4) Already training & testing data splitted, models were trained and accuracy score, confusion matrix, classification report , roc auc curve and auc score were evaluated.

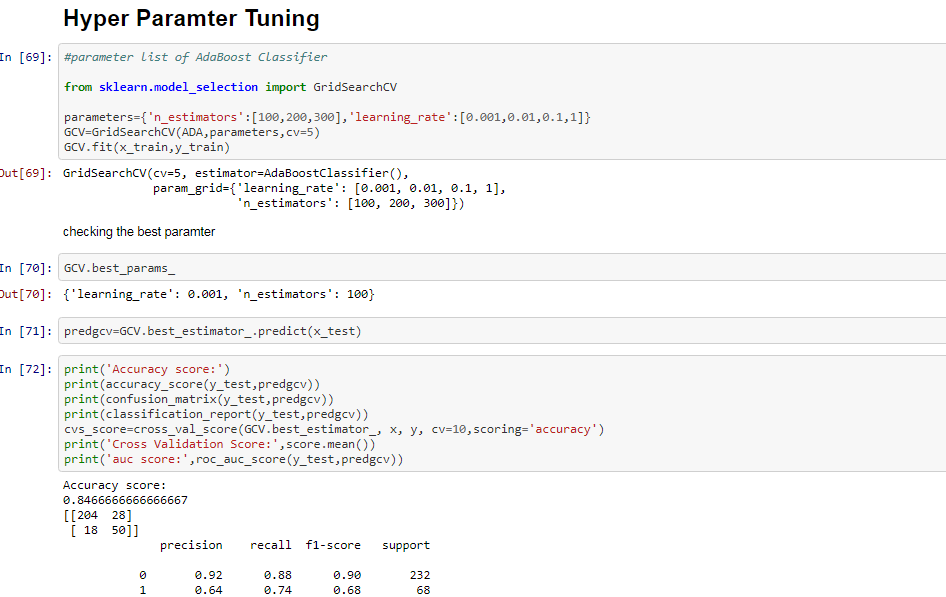


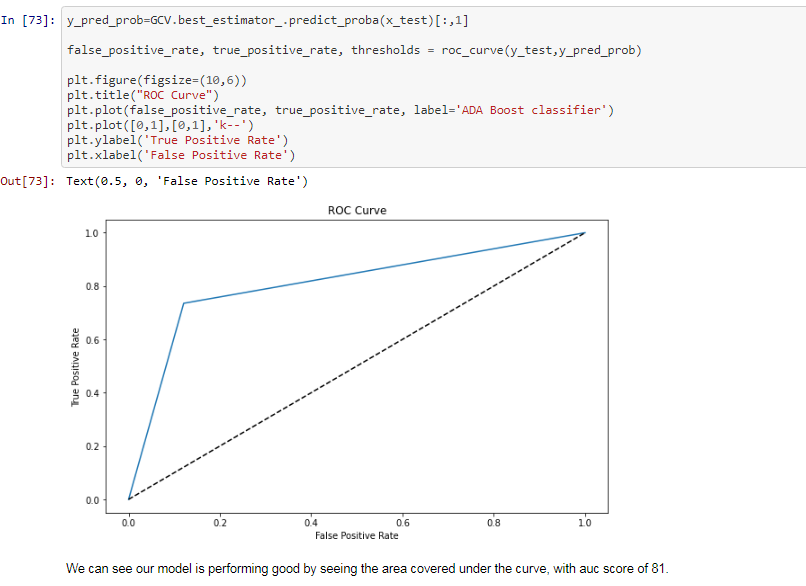




# 5) After performing modelling, AdaBoost Classifier is performing best for our data set with accuracy of 84%, cross validation score of 81%, auc score of 73%.

6)Therefore, best model I have chosen for hyper tuning i.e AdaBoost Classifer is having highest accuracy score, cross validation score & Auc score. So, I am choosing this for hyper tuning.

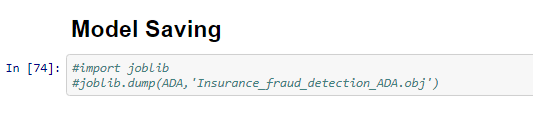
  
As we can see most like,results are with accuracy and cvs score, but good increase in auc score with 81.



**CONCLUSION**

From the auc\_score observations, accuracy\_score & cross\_val\_score this model is performing very well with the data’s. Note- auc score increasing very well after hyper tuning the ada boosting classifer model.

**Model Saving-**



* Limitations of this work and Scope for Future Work

As, I don’t have high end processor laptop, unable to do label sampling because of memory reference value error. This is the major limitation I faced as my chipsets are not compatible. Apart from that, more learning, alayzing & exploration is needed in Data science field that I am continuing. Apart from that, more learning , alayzing & exploration is needed in Data science field that I am continuing. Because, **The Journey to Data Science is a Marathon, not a Sprint.**

For more information, please visit:- <https://github.com/ayushpathak0912/Datatrained-Evaluation--Blog_Writing->

For more information, please visit github .ipynb file- <https://github.com/ayushpathak0912/DT-Evaluation-3rd-week/blob/main/Insurance%20Fraud%20Project.ipynb>