Insurance Claims Fraud Detection using Machine Learning

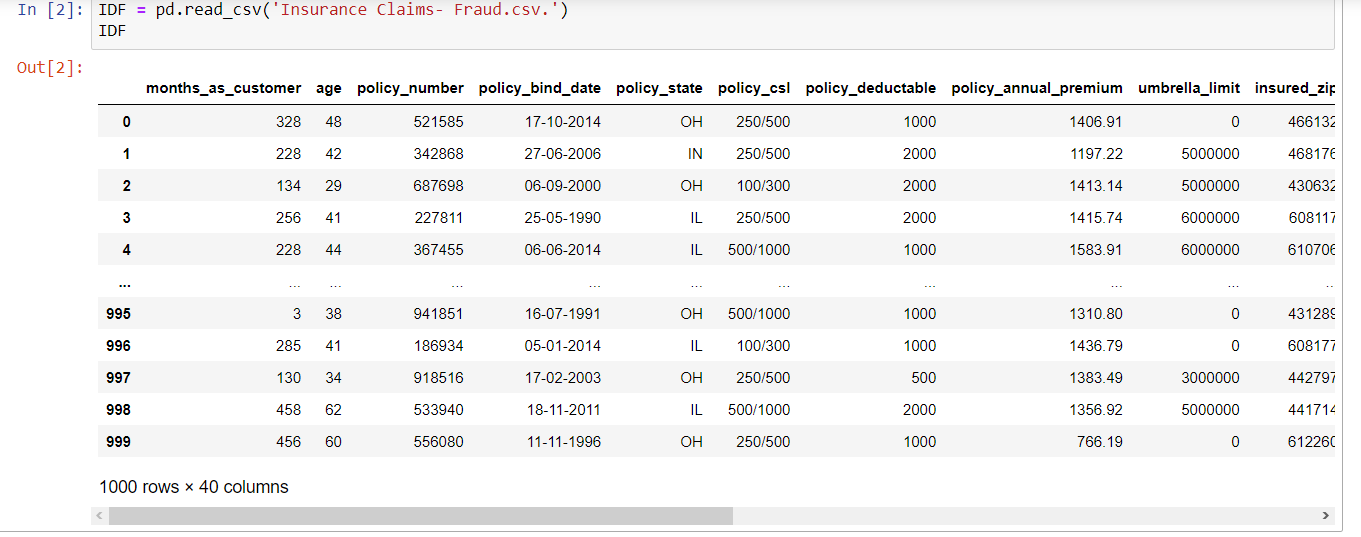
**Problem Definition:**

Business case: 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.

**Data Analysis:**

****

There are 1000 rows and 40 columns.

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

Target Variable :

Fraud reported: Whether fraud reported as Yes or No

**Next Step is Checking Null Values**:

\_c39 has no usable data present. Other columns appear to have no null values. So we will drop it.

# Now Checking for unique categories in the categorical columns with null values

By Using Simple Imputer Technique Imputing Values to NaN values in Columns:‘collision\_type’,’property\_damage’,’police\_report\_available’

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



**EDA (Exploratory data analysis):**

Exploratory Data Analysis (EDA) is an approach of analyzing

data sets to summarize their main characteristics, often with

visual methods, a statistical model can be used or not, but

primarily EDA is for seeing what the data can tell us beyond the

formal modelling or hypothesis testing task. we can say that EDA

is statisticians’ way of storytelling where you explore data, find

patterns and tell insights. EDA is a phenomenon under data

analysis used for gaining a better understanding of data aspects

like: - main features of data variables and relationships that hold

between them identifying which variables are important for our

problem.

# 

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.

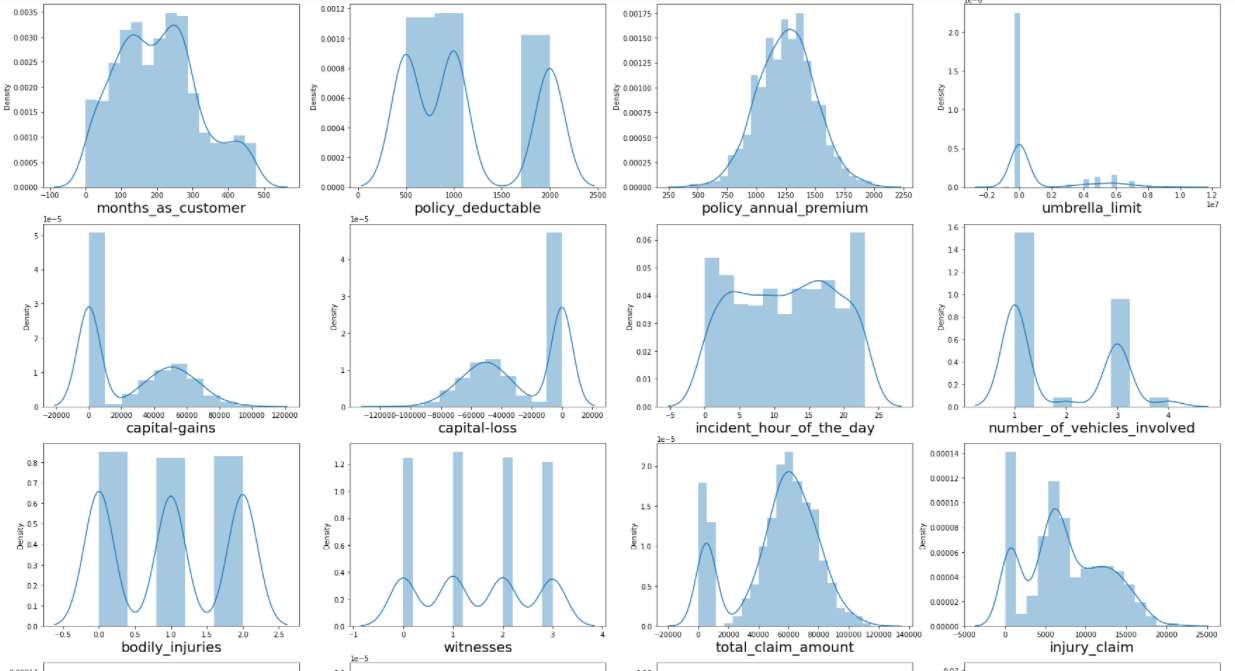
This is a Classification Problem since the Target variable / Label column ("fraud\_reported") has Catergorical type of Data

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

**Next Step is Analysing the Target Column:**

****

**Now Analysing Feature Columns with CATEGORICAL & CONTINUOUS Columns:**

****

Most continuous distributions are normally distributed and are multimodal.

# Now Interpreting Relationship between 'Income' vs Continuous/Discrete Data Columns

# 

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 the 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.

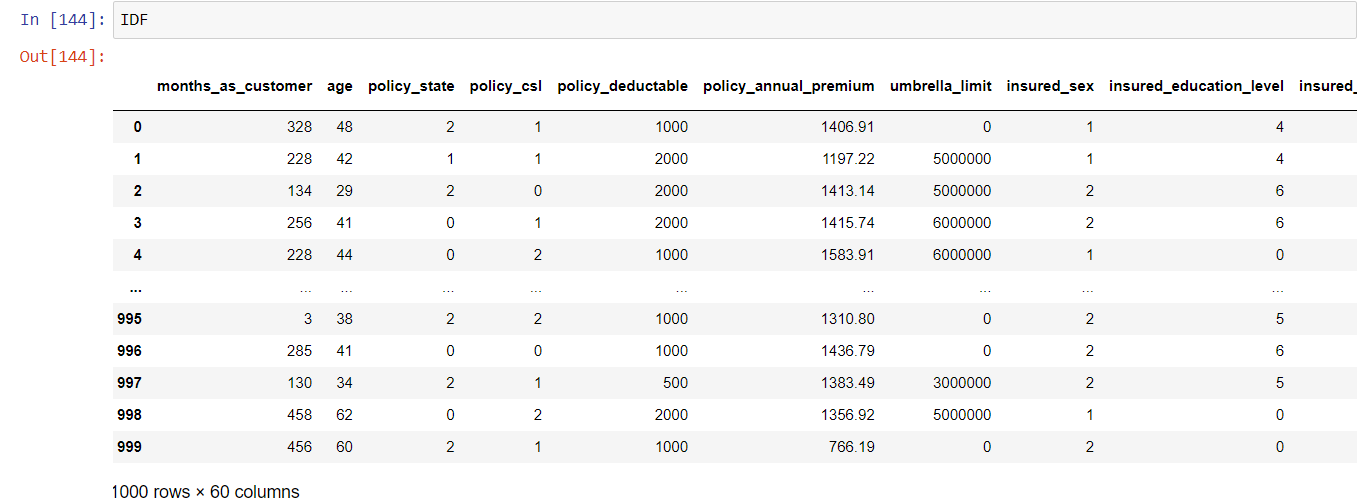
Higher the property claim amount, more the fraud claims are filed.

Higher the vehicle claim amount, more the fraud claims are filed.

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

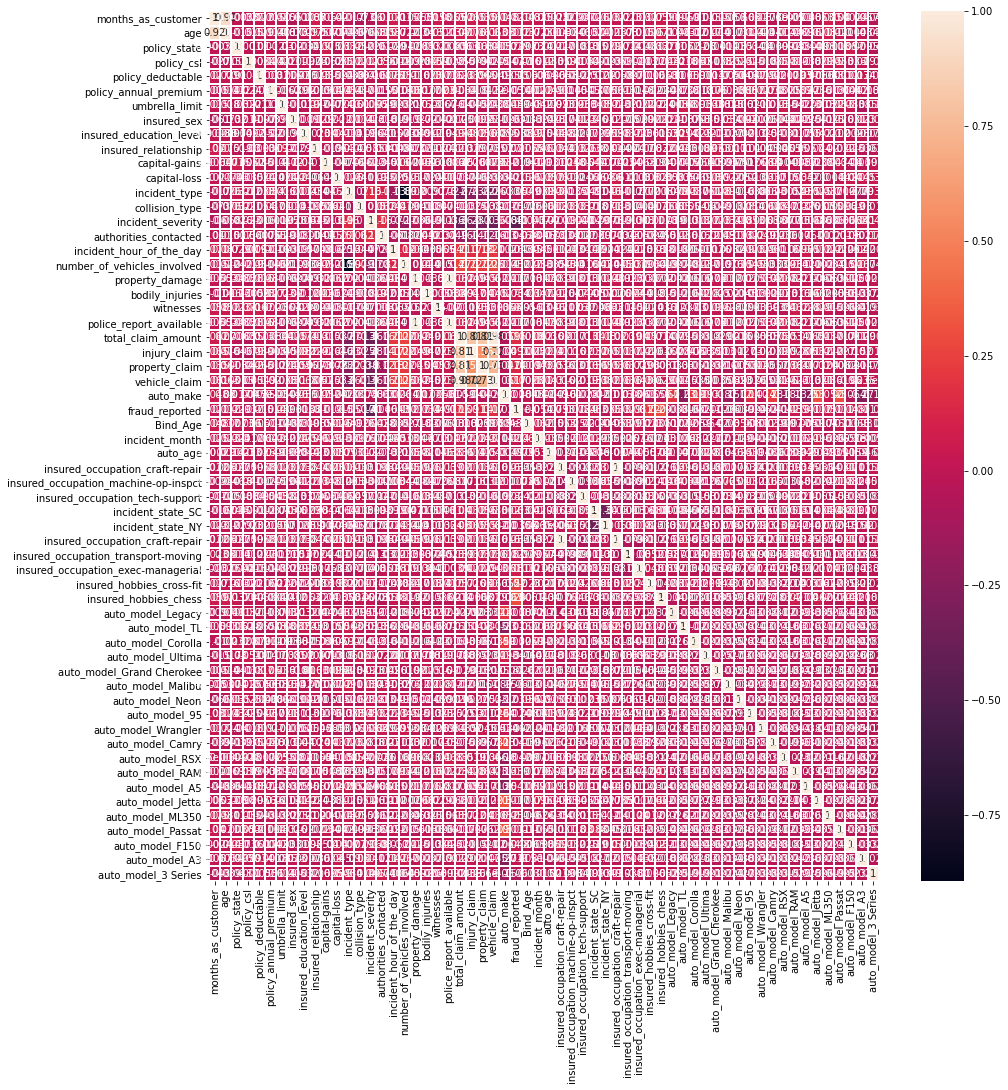
# Handling Categorical variables

# Most of the statistical models cannot take Objects / Strings as input they only takes numbers as inputs, with LabelEncoder () it is possible to categorize the string into Numbers as 1,2,3 and so on.

****

**Finding Correlation:**

Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables. corr() function is used to see the correlation among the dependent variable and independent variable you can see correlation in the following figure



**Pre-processing Pipeline**

Pipelines are the special way to simplify the code, Pipeline is generally used if we have to perform the code repeatedly usually when there is different train and Test data

Here Our EDA process is completed now moving towards next step.

**Building Machine Learning Models**

**Separating Features and Target column –**

It is necessary to separate the independent/Features column into a variable (x) and target column into a variable (y). here we have to separate all columns in x Data Frame (variable) and target variable in y Data Frame (variable).

**Splitting the Data for Training and Testing**

In ML the separated data is split into 4 parts for Training and Testing of features (x) and for Training and Testing of Target (y) like x\_train, x\_test, y\_train, y\_test. It is possible through a inbuilt library of sklearn’s train\_test\_model

**Training the Models**

To find the best model it is necessary to train 3-4 models, In the same way I have trained

RFC = Random Forest Classifier()

DTC = Decision Tree Classifier()

XGBC= XGB Classifier()

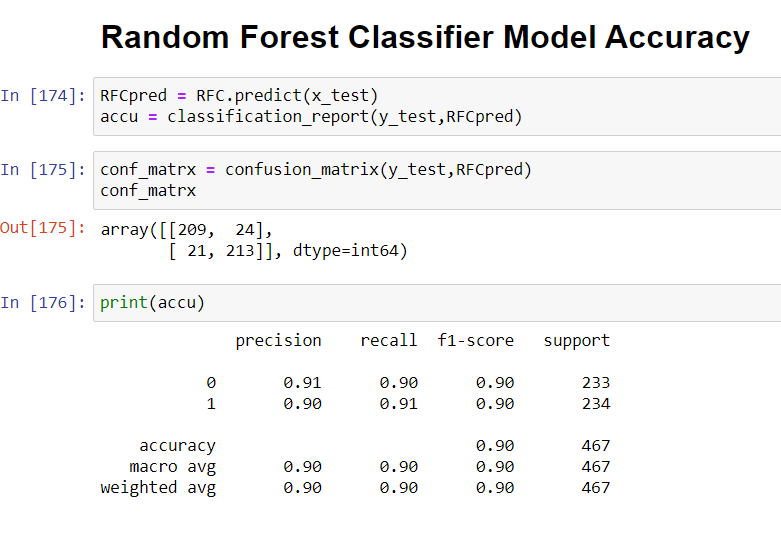
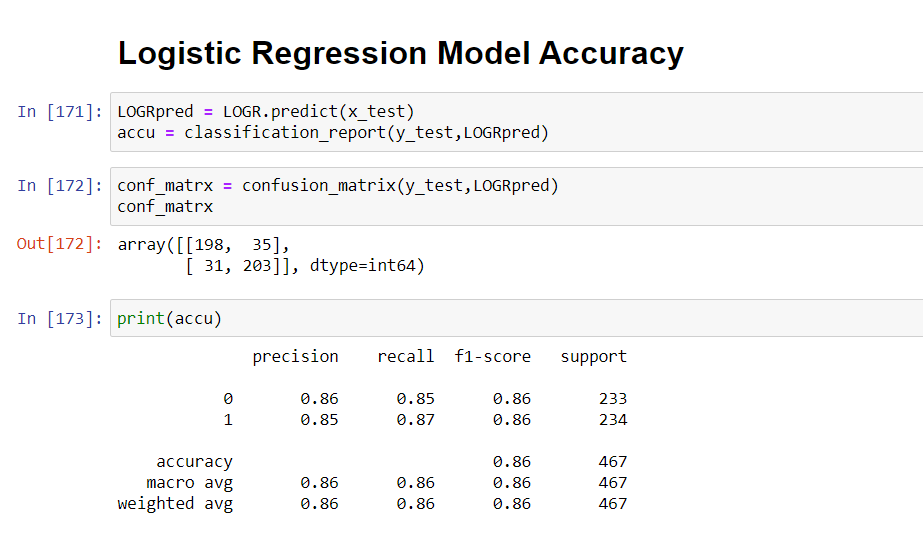
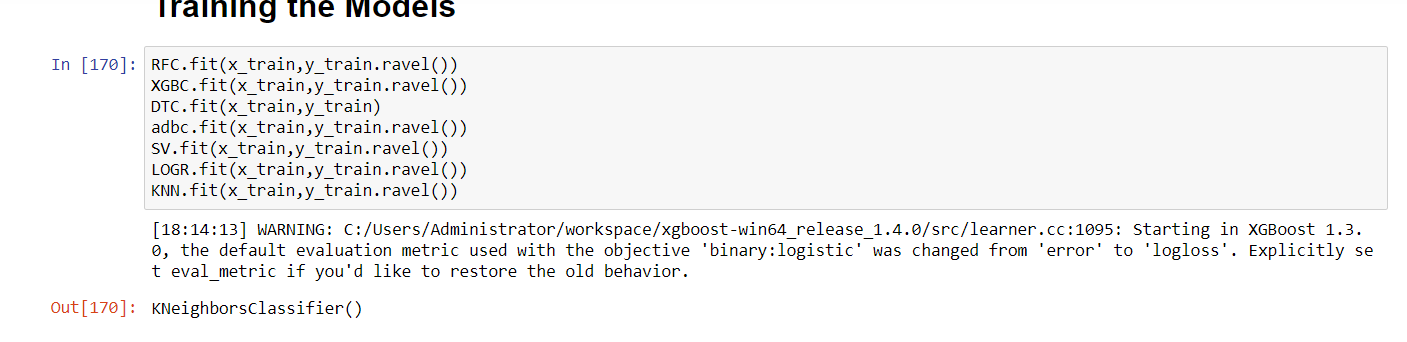
adbc = AdaBoost Classifier()

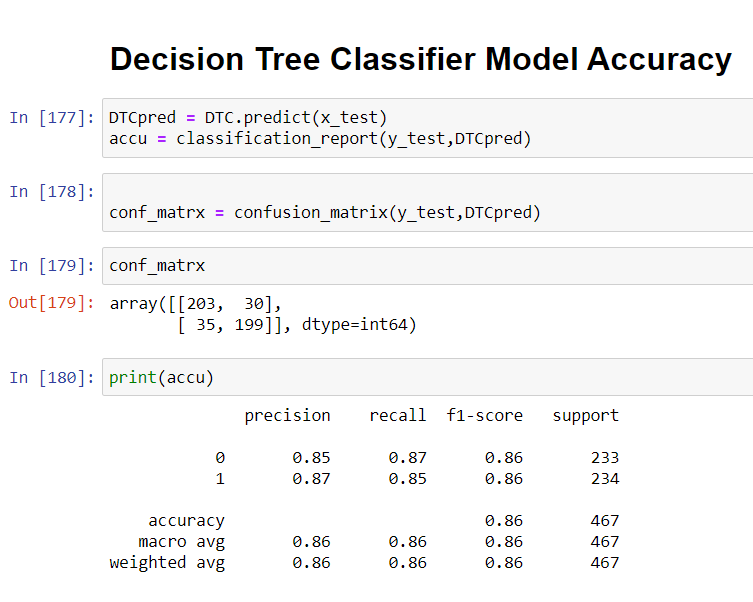
LOGR= Logistic Regression

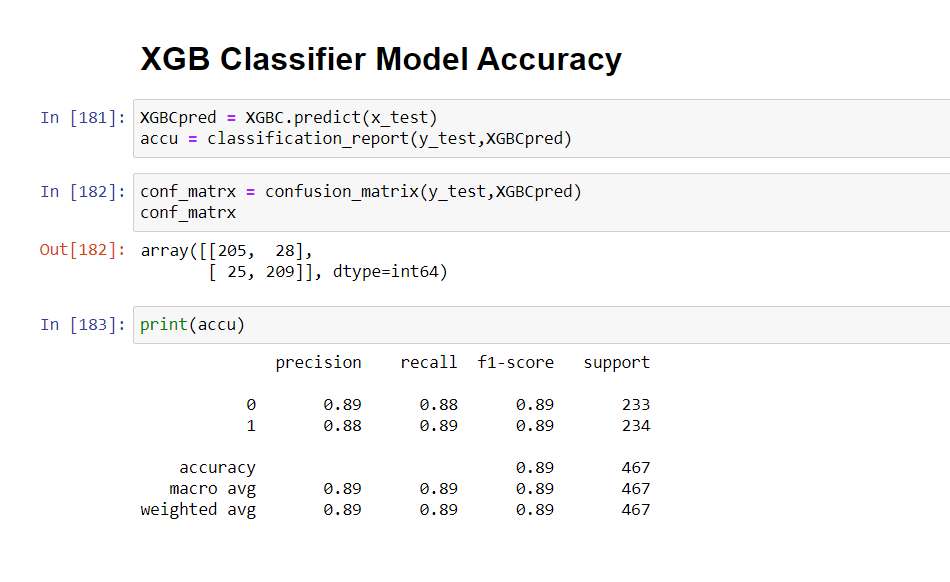
SV = SVC()

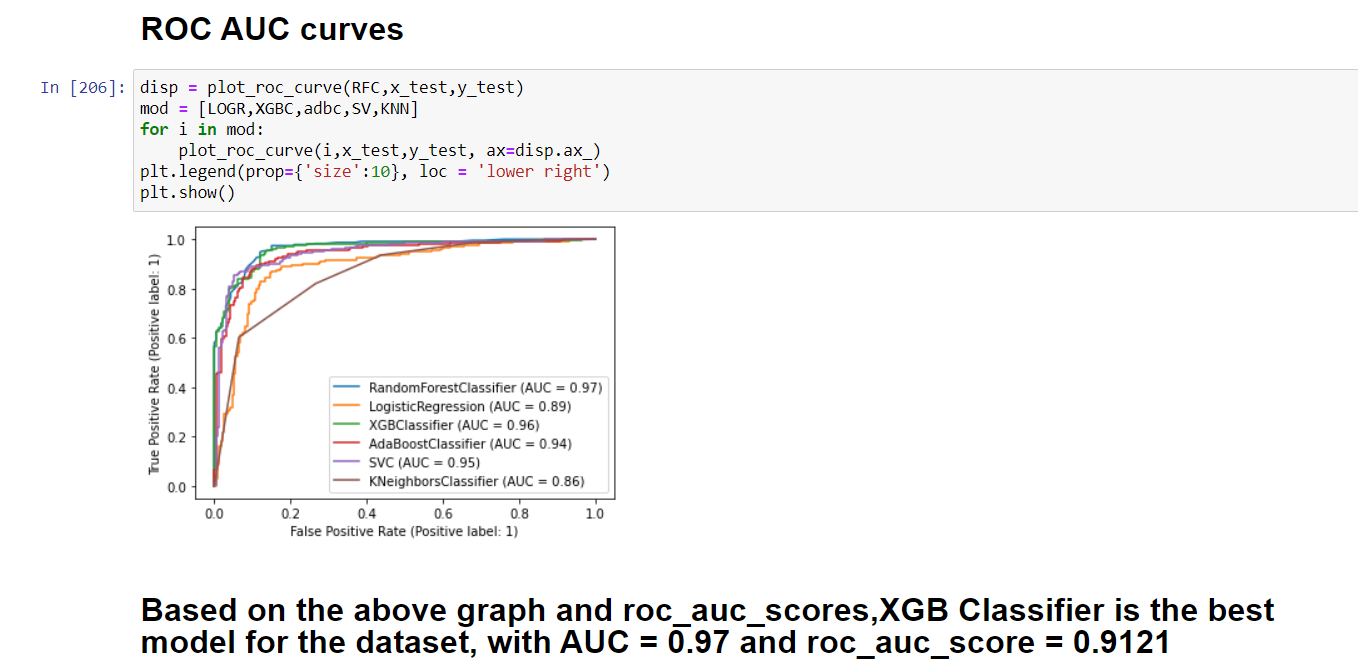
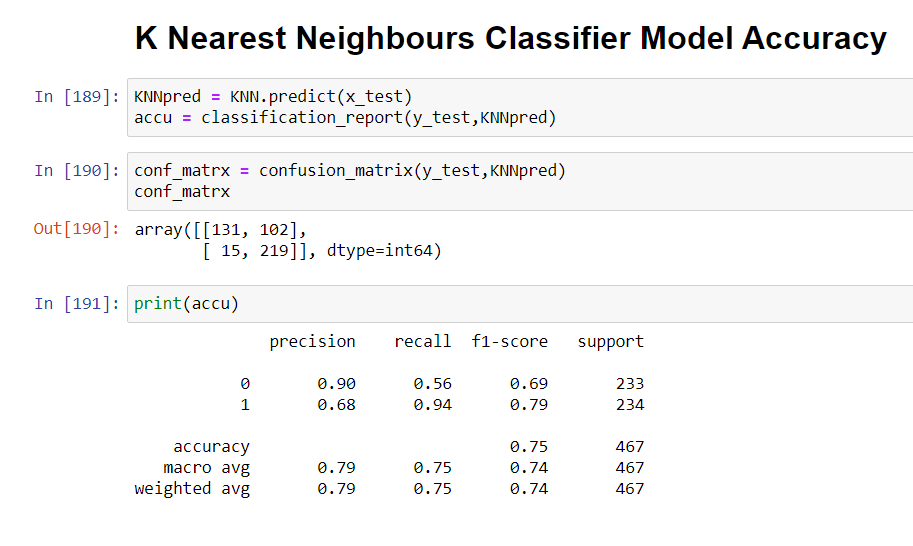
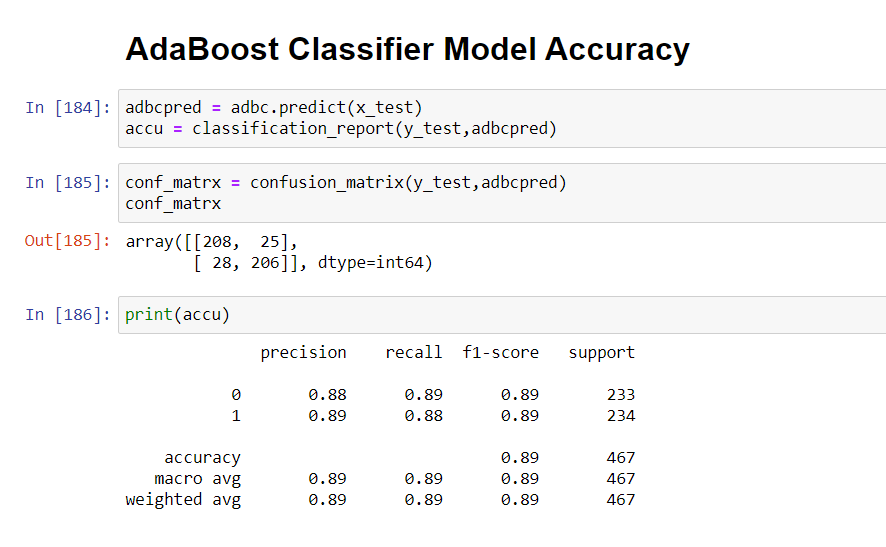
KNN = KNeighbors Classifier()

to check which model is giving the best max score.

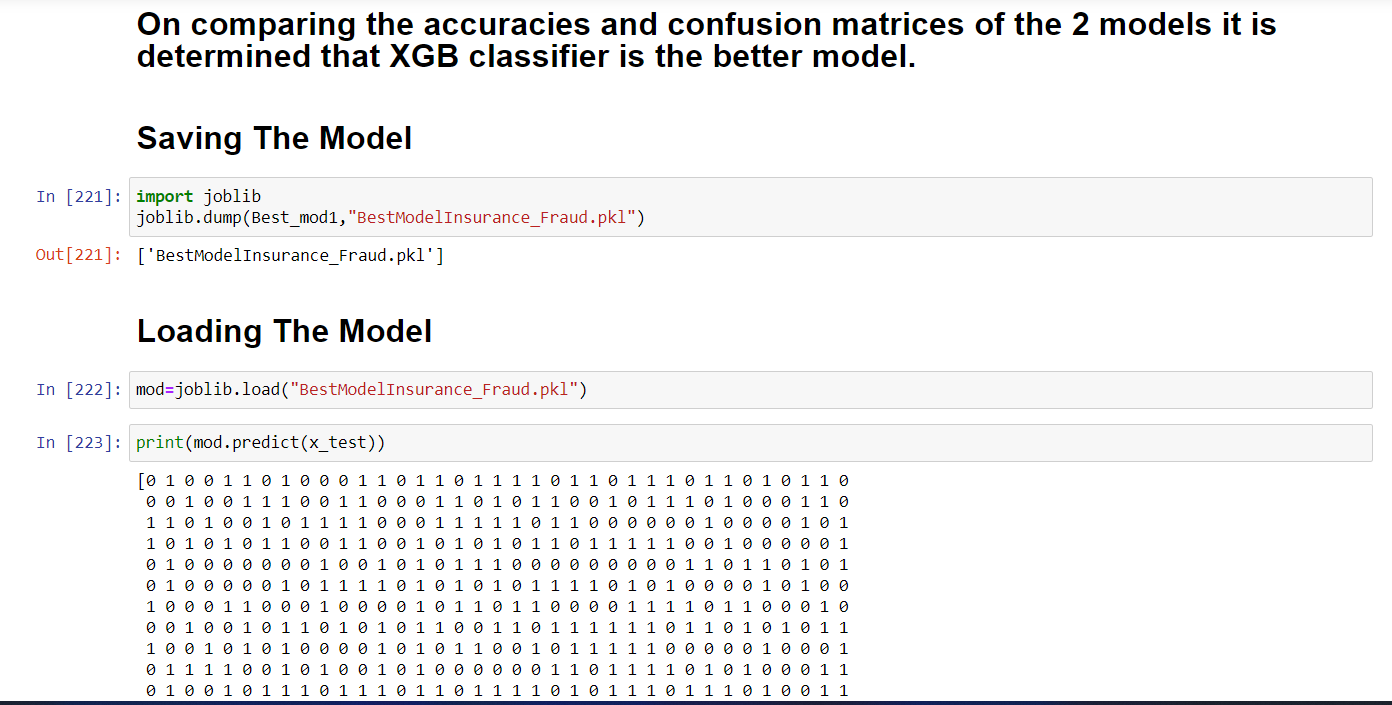








**Saving and Loading the Model:**



**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.