**DATA MINING PROJECT**

**A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.**

The dataset ‘bank\_marketing\_part1\_Data’ contains **210 rows and 7 columns.**

spending 210 non-null float64

advance\_payments 210 non-null float64

probability\_of\_full\_payment 210 non-null float64

current\_balance 210 non-null float64

credit\_limit 210 non-null float64

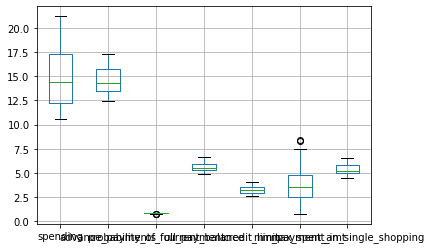
min\_payment\_amt 210 non-null float64

max\_spent\_in\_single\_shopping 210 non-null float64

The dataset has no duplicate values and null values.

Exploratory Data Analysis:

Visualising the five number summary with Boxplot.



The measure of dispersion IQR

spending 5.035000

advance\_payments 2.265000

probability\_of\_full\_payment 0.030875

current\_balance 0.717500

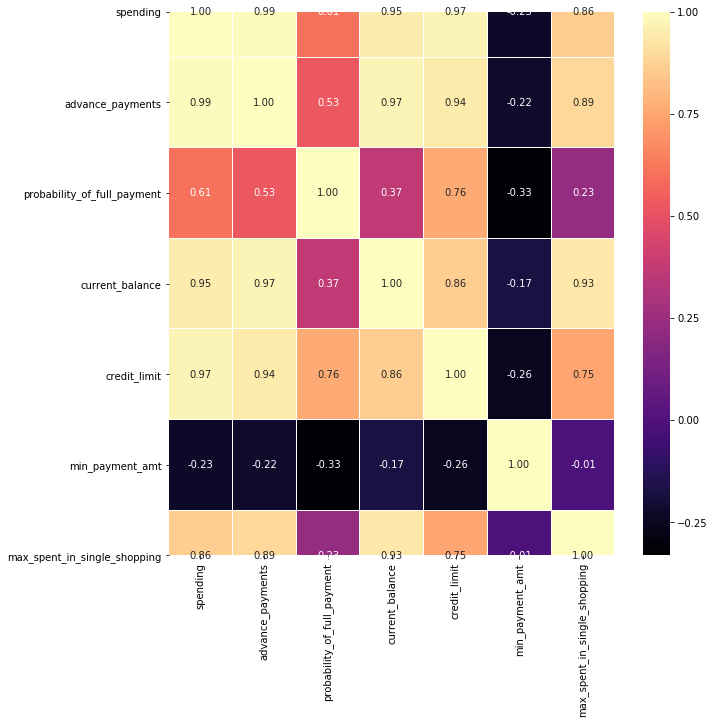
credit\_limit 0.617750

min\_payment\_amt 2.207250

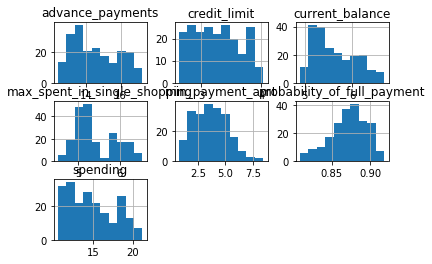
max\_spent\_in\_single\_shopping 0.832000

Visualizing the Heatmap for explaining the correlation between every attribute.

The heat map shows how each and every attribute showcase the correlation and have a positive high correlation of 0.99



To measure the skewness and visualize it by using Histogram



After data analyzed before moving on to clustering the dataset it requires ***Scaling.*** While using distance algo in clustering it requires to scale the dataset before applying the algorithm.

* Scaling removes redundant data and ensures the good quality custers.
* It helps to keep the data in a particular range.

In this case the values are in different forms as thousands and hundred, so to make it in a particular range scaling ins must in this scenario.

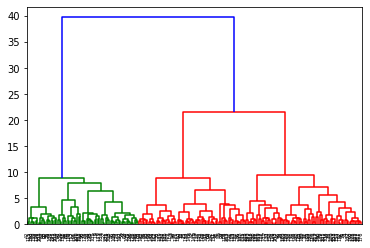
Hierarchical Clustering:

Here using **Ward method** for hierarchical clustering to group the data, it takes the measure of difference between clusters as the sum of squared differences within all clusters.

Looking at the dendrogram, the highest vertical distance that doesn’t intersect with any clusters is the Green one

Given that 5vertical lines cross the threshold ,the **optimal number of clusters is 2**.

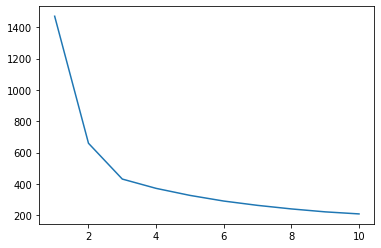
Dendrogram



K-means Clustering:

To understand how the dataset groups with each other using the K-means algorithm.

The **Elbow curve** will be plot based on K- means inertia which will help us to analyze the optimum number of clusters.



Here from the curve we can understand from the number of cluster 3 it slightly decreases, hence **the optimum number of clusters is three.**

The silhouette score determined how well each object lies within its cluster. The **silhouette score** for our cluster is **0.5863370.** Basically the score ranges from 1 to -1

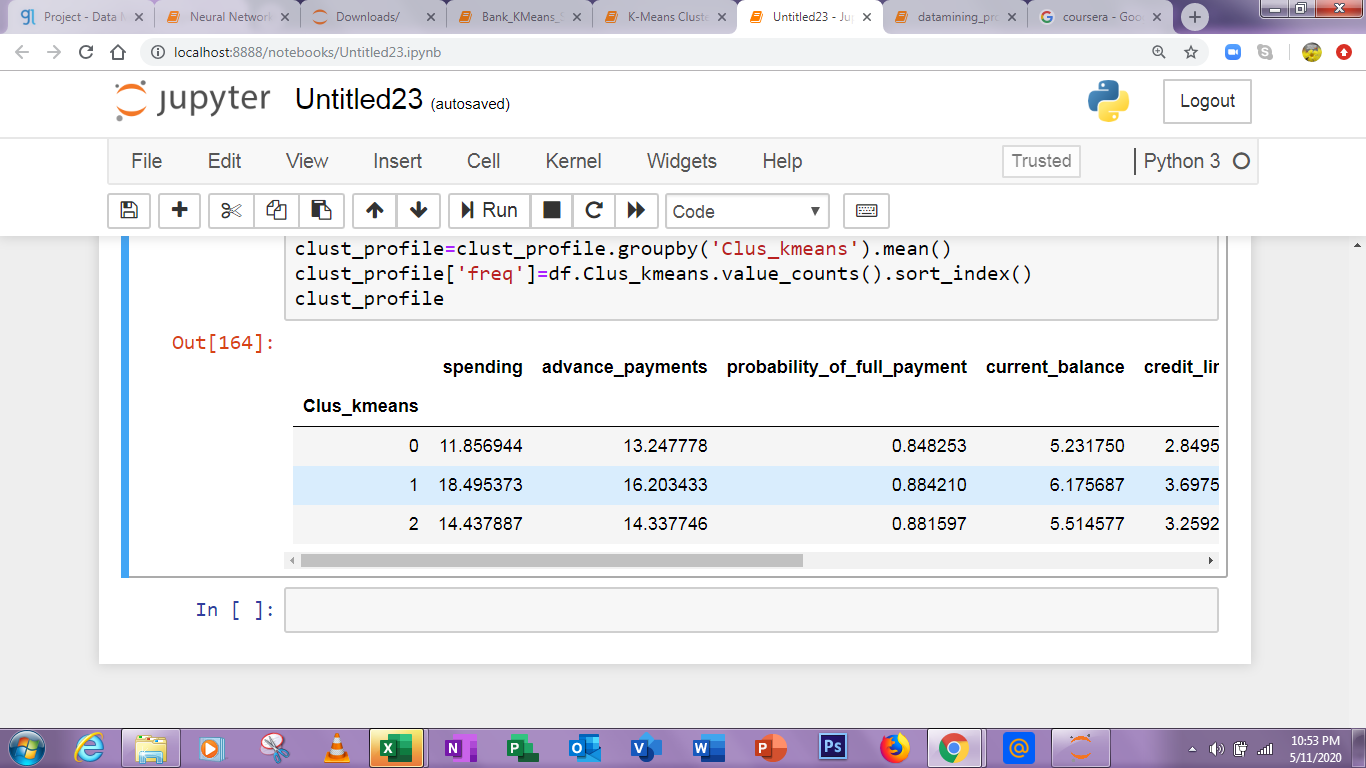
**Cluster Profiles:**

Based on profiling, it involves generation of description of clusters with reference to the input variables which are used for cluster analysis.

Cluster 0 72 nodes

Cluster 1 67 nodes

Cluster 2 71 nodes



Promotional Strategies:

Cluster 0 : Has small spending, Small advance payments,small current balance, credit limit but highest Minimum amount payment.

Cluster 1 : High in current balance and highest in spending, high minimum amount payment and advance payment.

Cluster 2 : Medium in spending, advance payment, current balance, Credit limit, but minimum in minimum amount payment.

Recommendations:

Cluster 0 Has used its credit card less as far as the result of spending, promotional offers need to give for the customers in cluster 0.

Cluster 1 Customers have widely used their credit card and the credit limit is more and it shows that the usage is more frequent and also the Payment activities are perfectly fine with this group. Hence more promotional offers can be given.

In cluster 0 the minimum Payment amount is higher compared to others, hence there may be producing more EMI options can be useful for the group of people to widely use it.

Above all, the more number of people using credit cards is less as far as clustering, hence adding more promotional offers will grab the customer to use it.

**An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.**

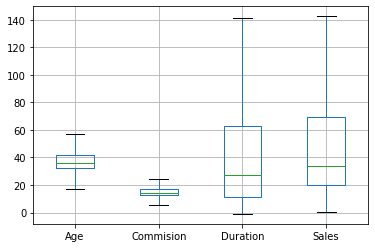
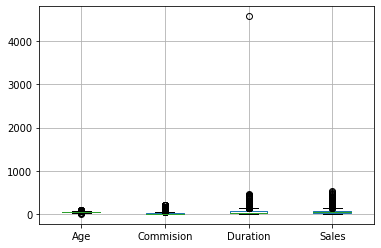
Descriptive Statistics:

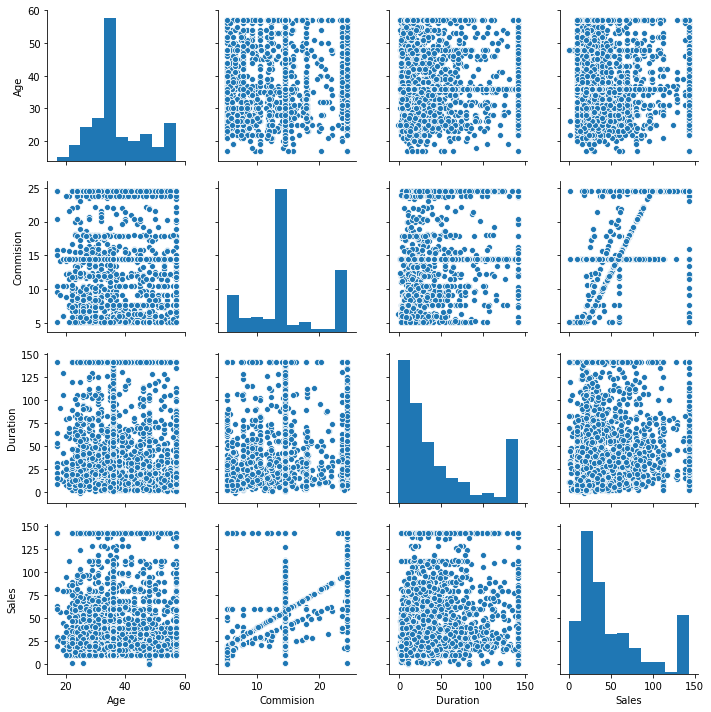
The dataset ‘insurance\_part2\_data’ has **3000 rows and 10 columns.**

**Summary :**

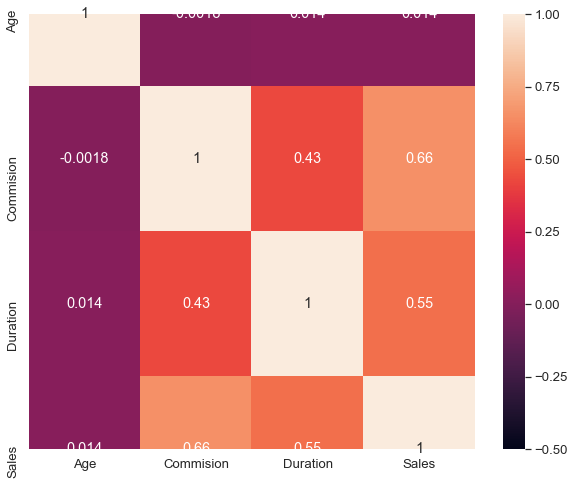
Check for duplicate rows the dataset has 139 rows duplicated, in this case we don't need to remove the duplicates.

Check for Null values and treat outliers with the IQR value.

Check the pairwise distribution of Continuous variables

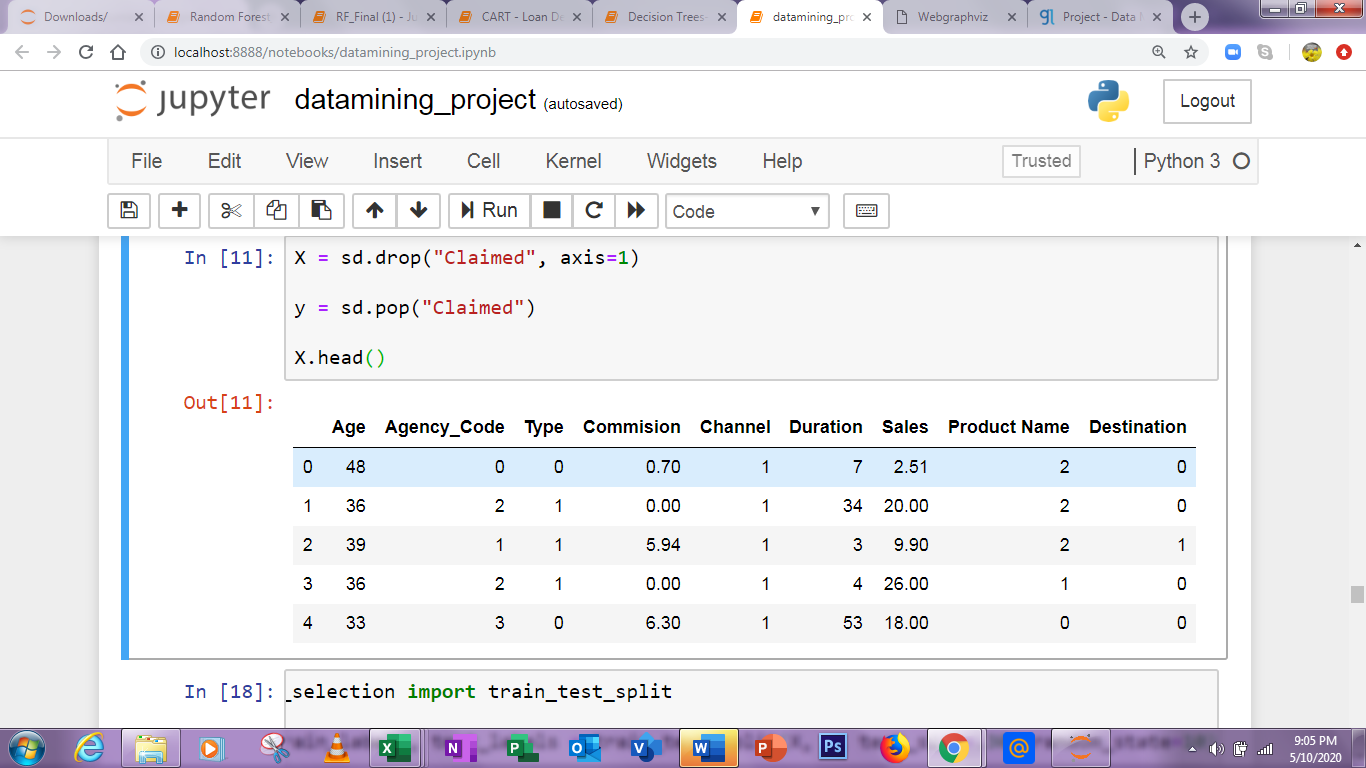


Finding the correlation between the continuous variables



In this case we have columns like **Commision, Sales, Duration** which has Zeros in it, Imputate them with their mean for further proceedings.Thus checking them with null value condition checks will help us to model a best one.

After doing basic descriptive statistics, convert the attributes to categorical so that it will be easy to perform a model.



**Data Split:**

While predicting something, split the dataset into training set and test set.

* Here **Claimed is a prediction** variable for the dataset to analyze the status of the customer.
* Splitting the claimed column to a new variable to test the data on its prediction.
* Using the Train Test Split algorithm, separating it into two parts.

**X\_train (2100, 9)**

**X\_test (900, 9)**

**train\_labels (2100,)**

**test\_labels (900,)**

**CART:**

Trying Classification and Regression tree for analyzing the insurance claim and recommending to management.

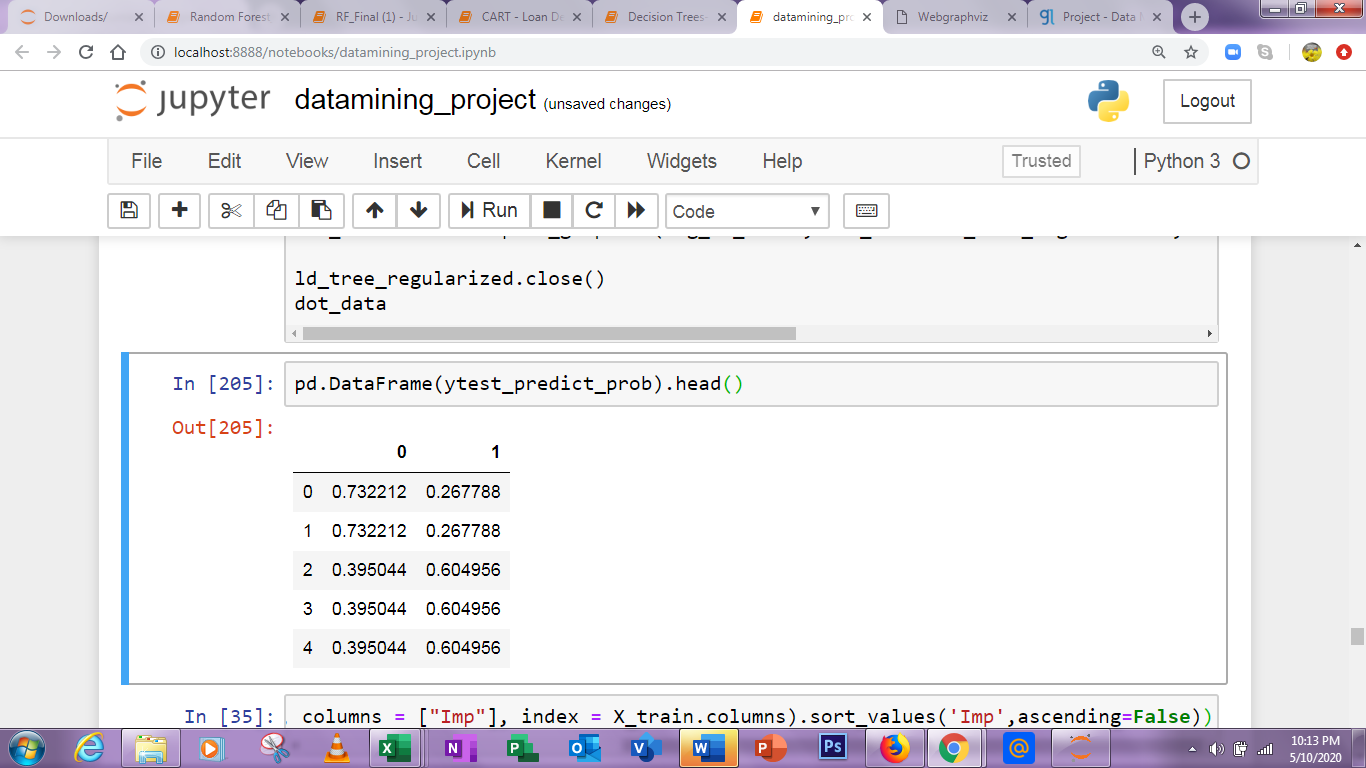
Observations are equally distributed between the train and test dataset.

**Training labels Testing labels**

**0 0.676667 0 0.727778**

**1 0.323333 1 0.272222**

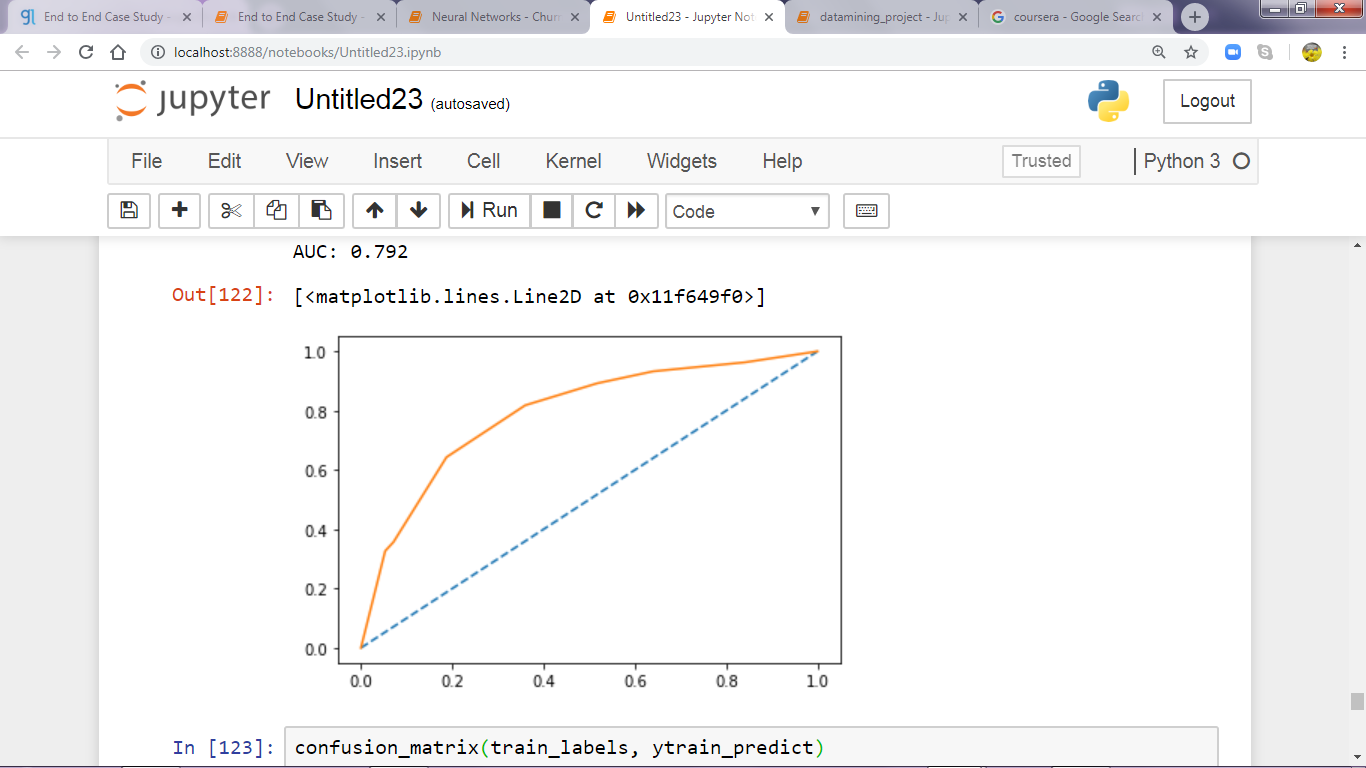
Finding the predicted probability



AUC of Train set



AUC of Test set



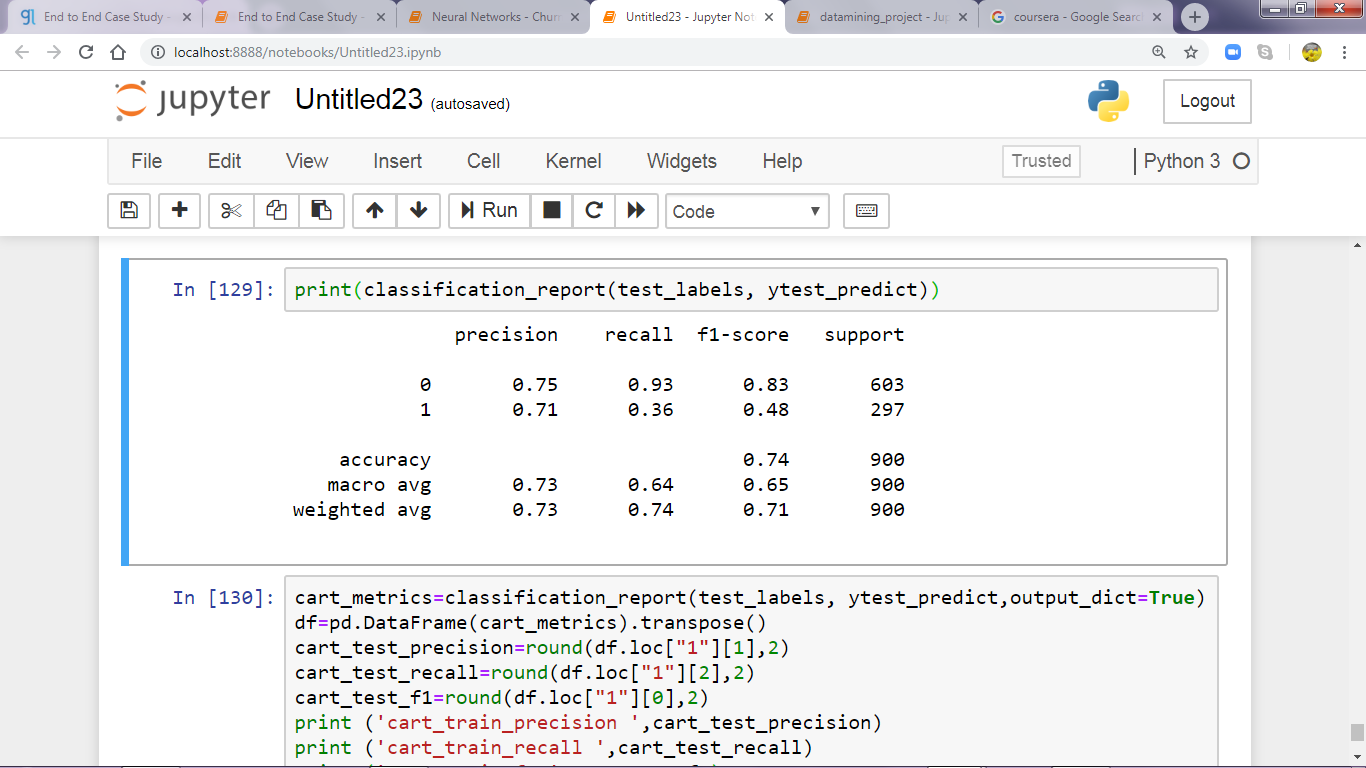
**Confusion matrix - Train Data Accuracy = 0.7452380952380953**

**Classification Report: Train data**

****

**Confusion matrix : Test data Accuracy = 0.7855555555555556**

**Classification Report : Test Data**



**Train Data: TEST Data**:

AUC: 80.6% AUC: 79.2%

Accuracy: 79.6% Accuracy: 74%

Sensitivity: 41% Sensitivity: 36%

Precision: 52% Precision: 48%

f1-Score: 73% f1-Score: 71%

**RF -Random forest**

Observations are equally distributed between the train and test dataset.

Training labels Testing labels

0 0.676667 0 0.727778

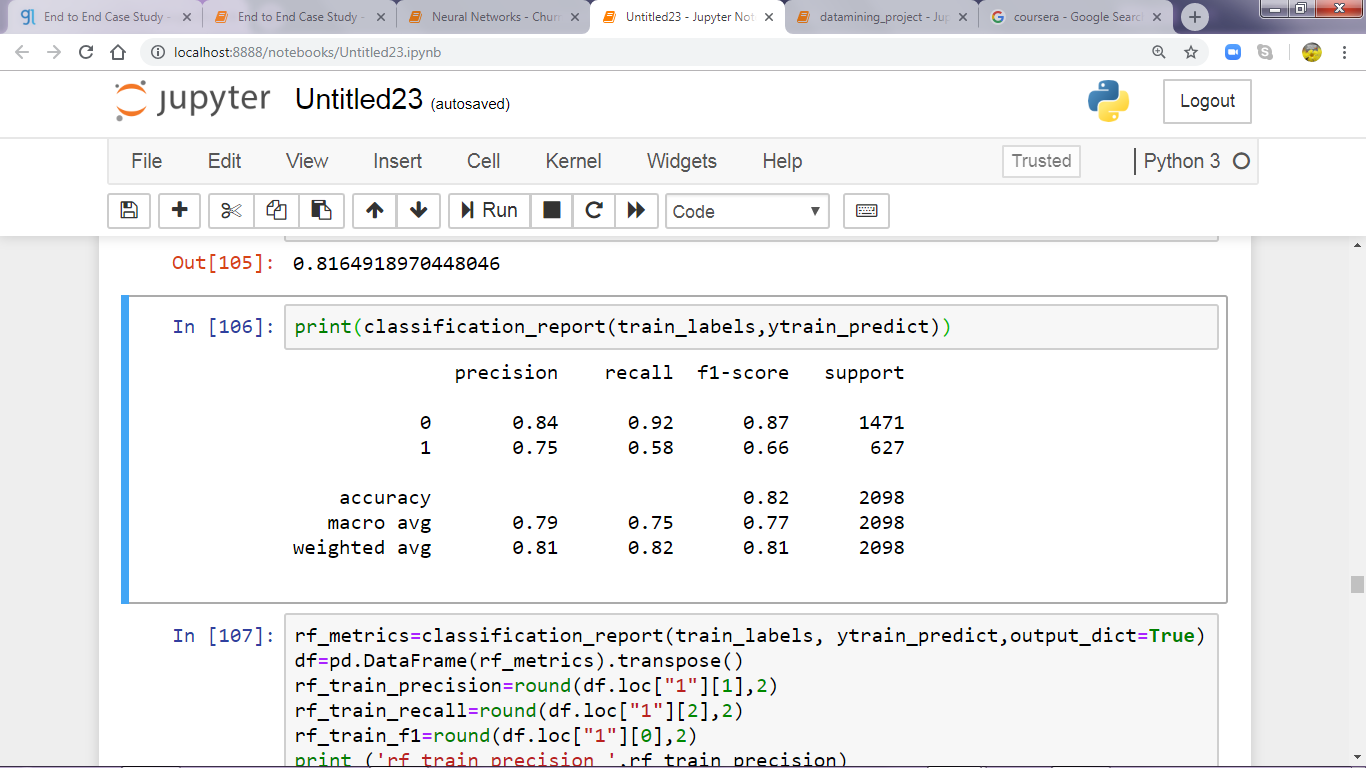
1 0.323333 1 0.272222

**Confusion Matrix - Train Data Accuracy**

[[1347, 124],

[ 261, 366]]

**Classification Report- Train Data**

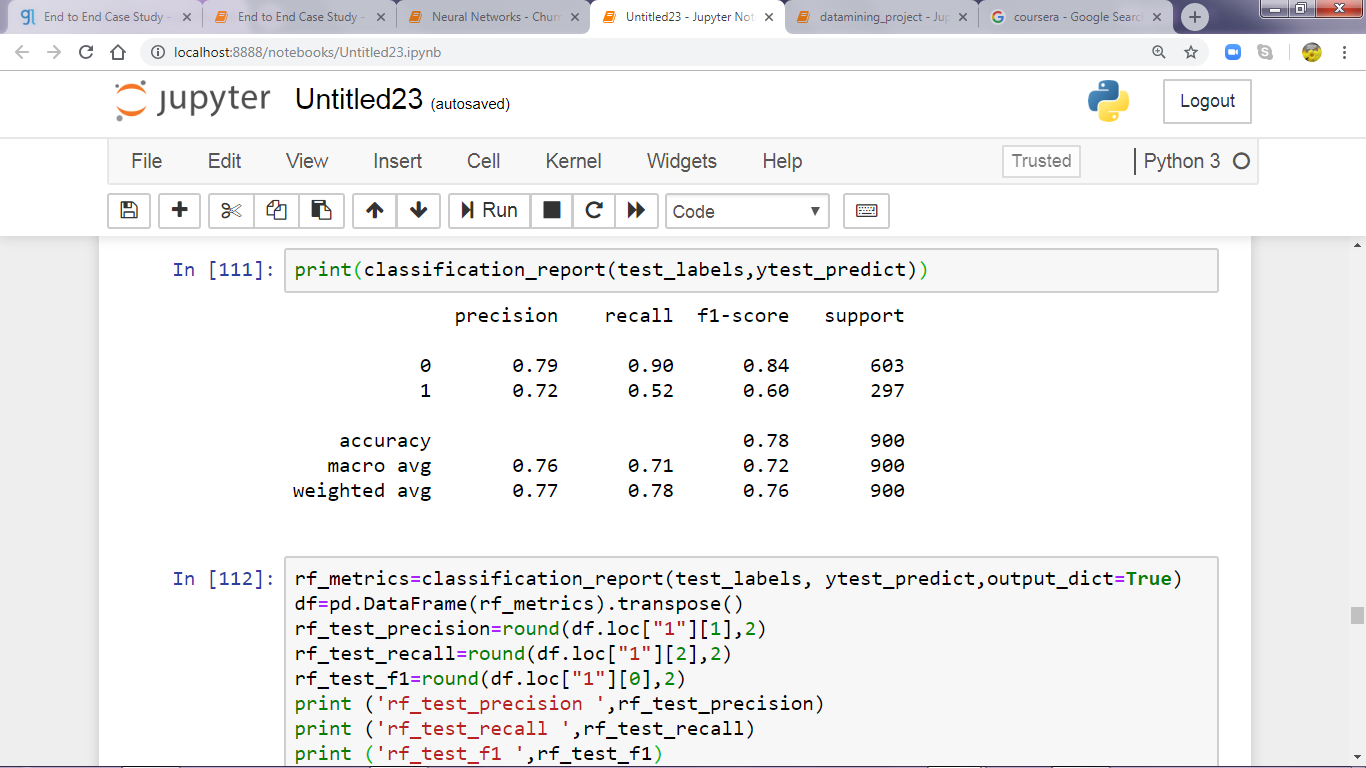
****

**Confusion Matrix : Test data**

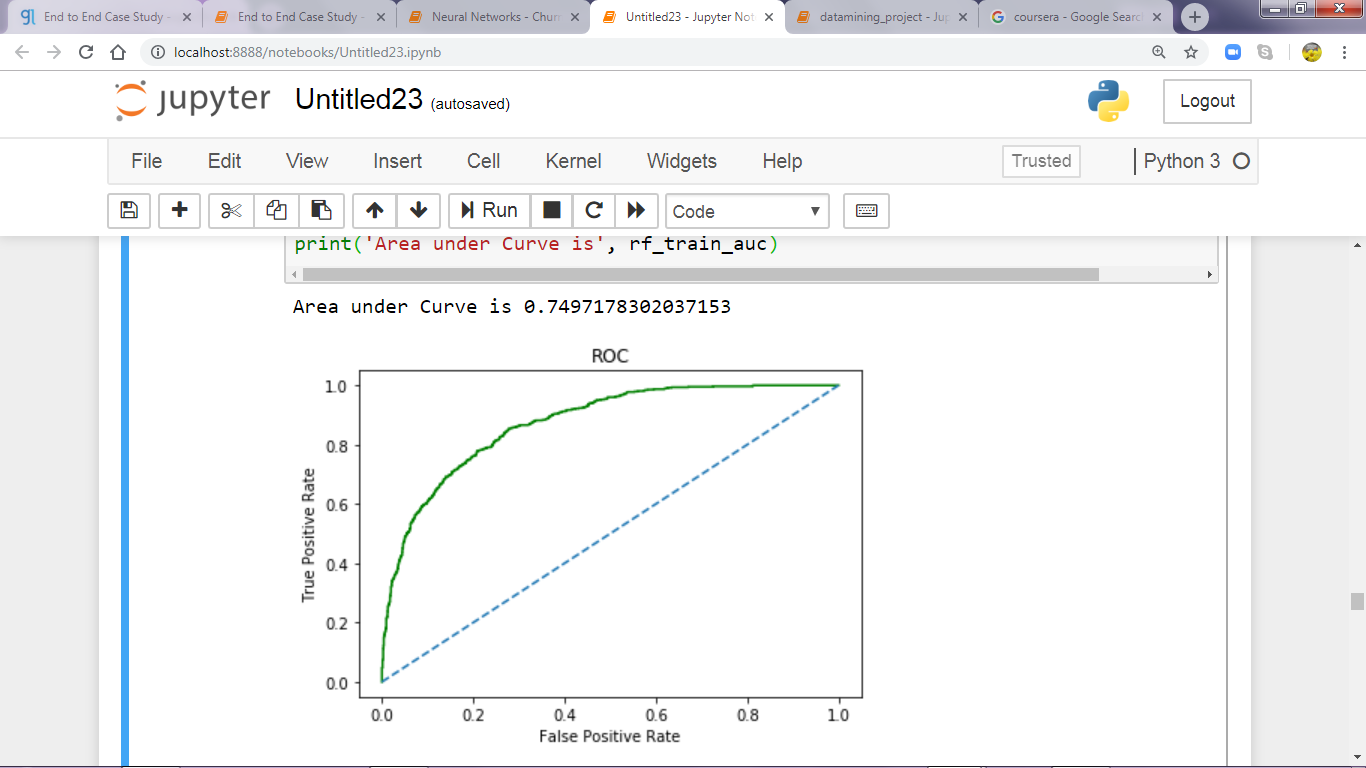
**[[544, 59],**

**[143, 154]]**

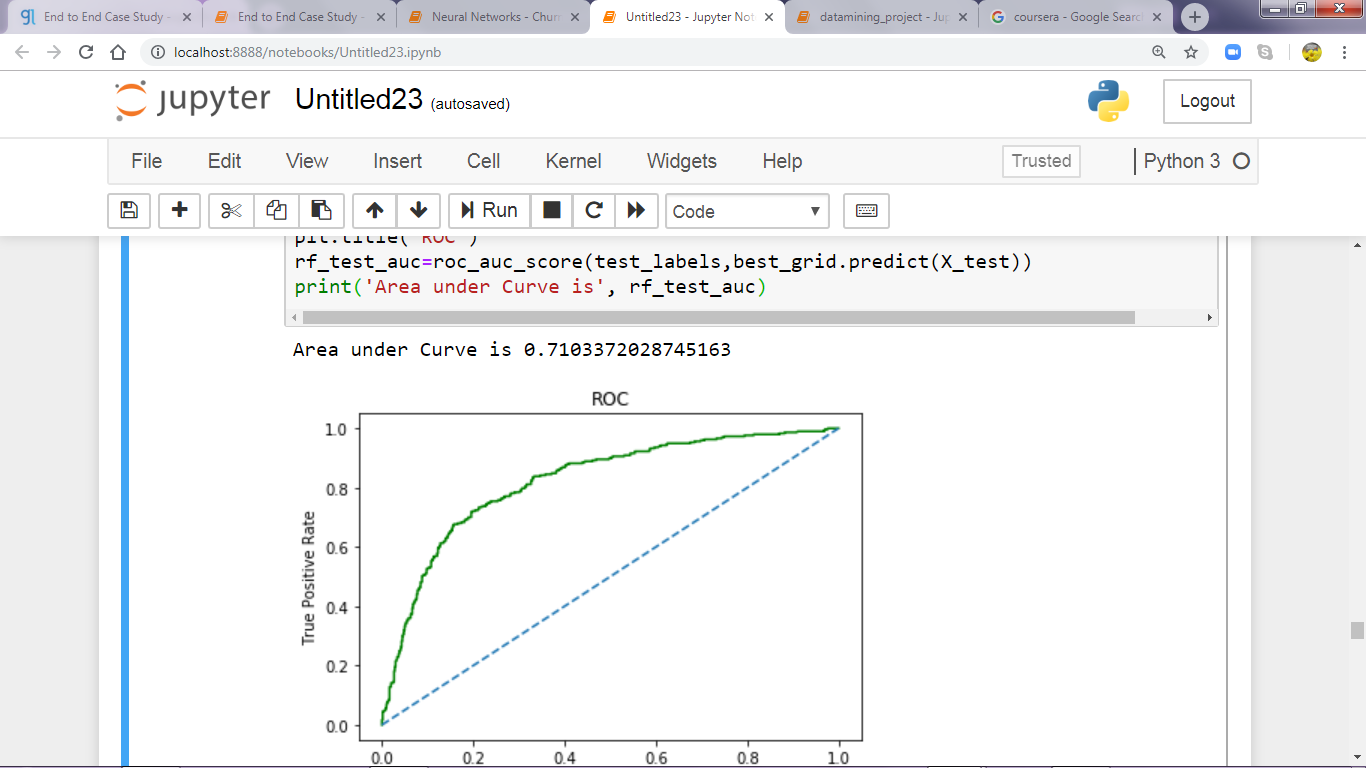
**Classification Report : Test data**

****

**AUC for Train Data**

****

**AUC for Test data**

****

**Train Data: TEST Data**:

AUC: 74% AUC: 71%

Accuracy: 81% Accuracy: 77.5%

Sensitivity: 58% Sensitivity: 52%

Precision: 66% Precision: 60%

f1-Score: 75% f1-Score: 72%

Importance of Variable

Agency\_Code 0.334551

Sales 0.202768

Training and tests are almost similar, with its high measures as a GOOD model.

***Agency Code*** is again the important variable in predicting the industry claim

**Artificial Neural Network**

Applying the predicted weight over the model will result Neural network.

Training labels Testing labels

0 0.676667 0 0.727778

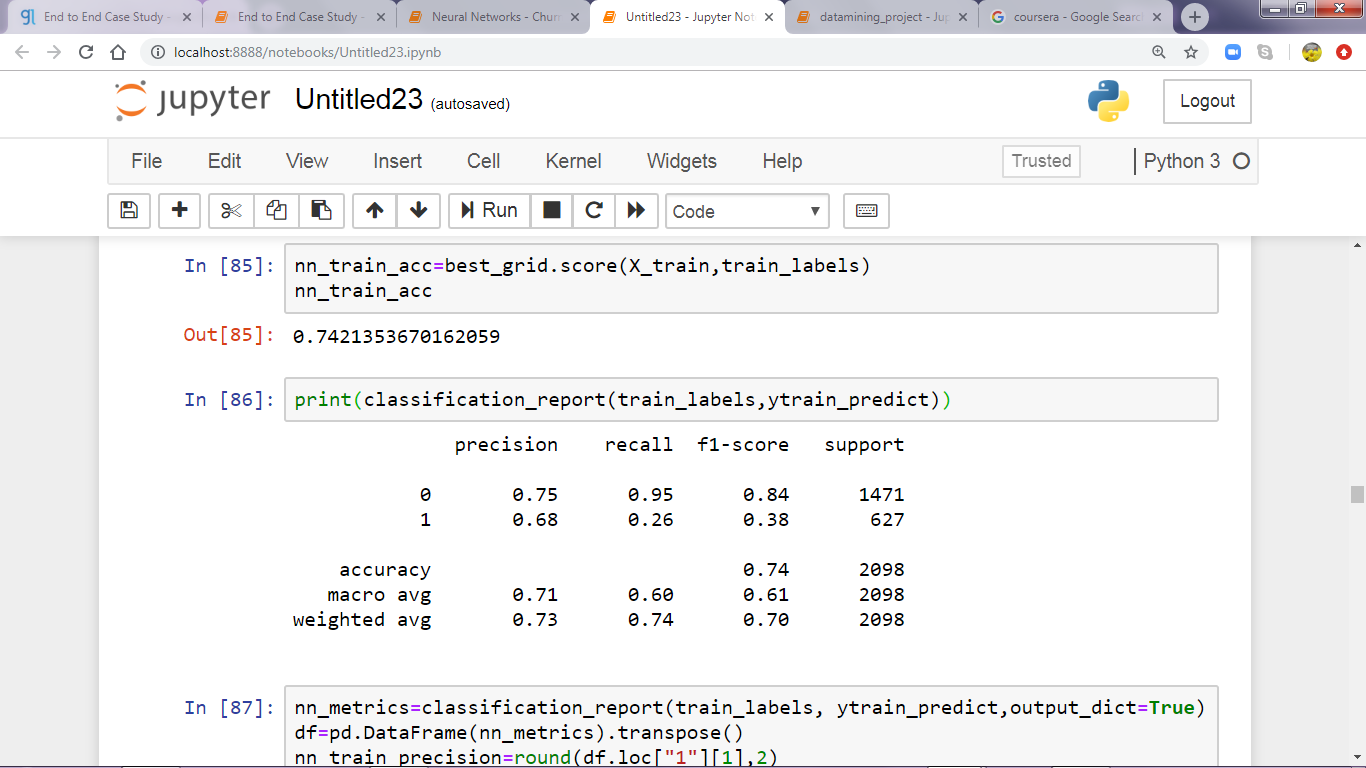
1 0.323333 1 0.272222

**Confusion matrix - Train data accuracy**

**[[1392, 79],**

**[ 462, 165]]**

**Classification report -Train data**

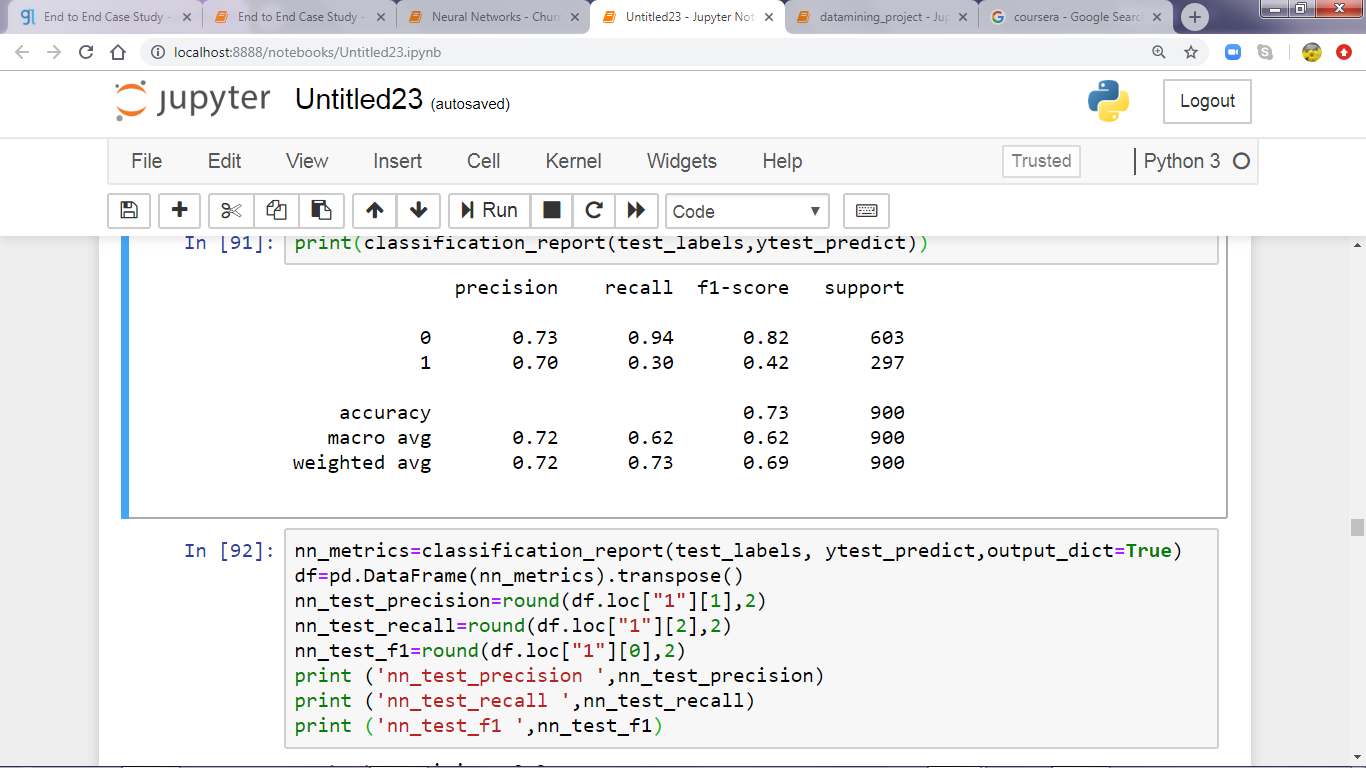
****

**Confusion matrix - Test data**

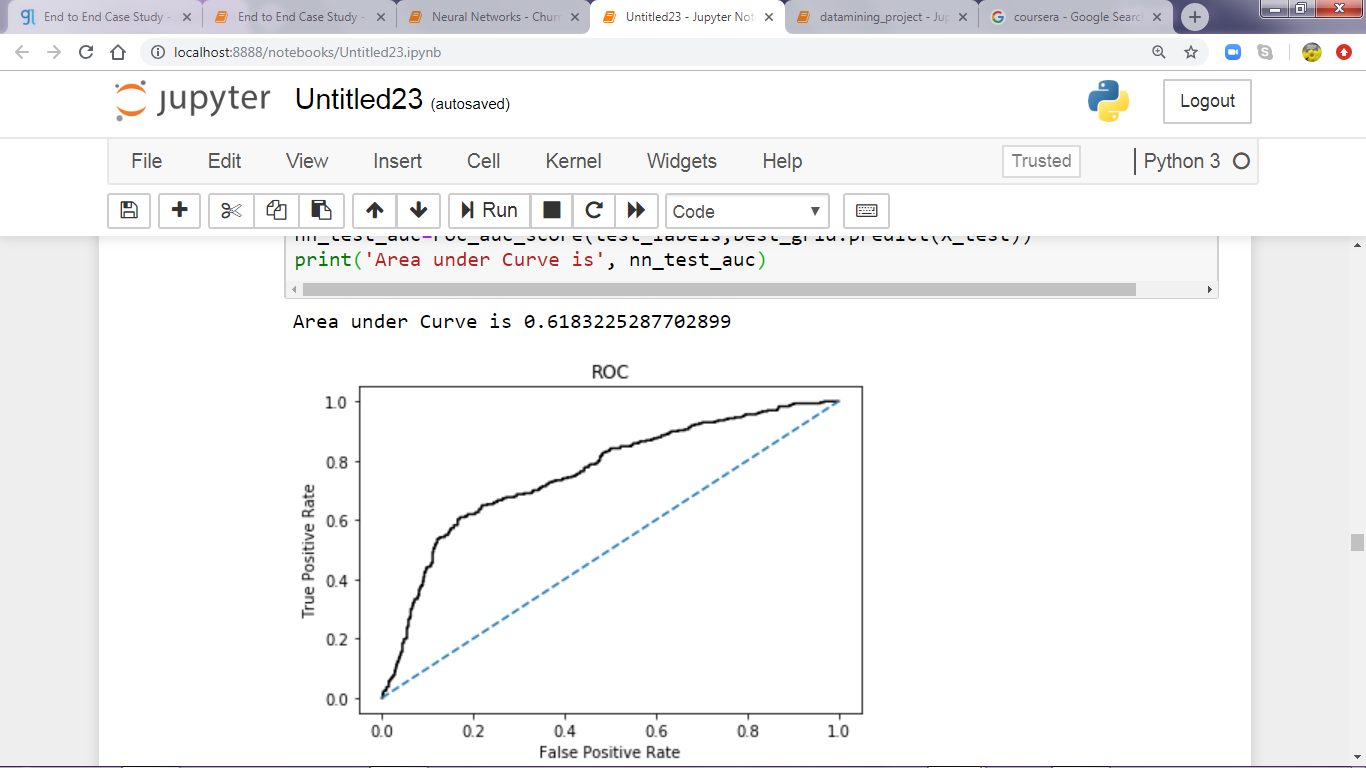
**[[565, 38],**

**[208, 89]]**

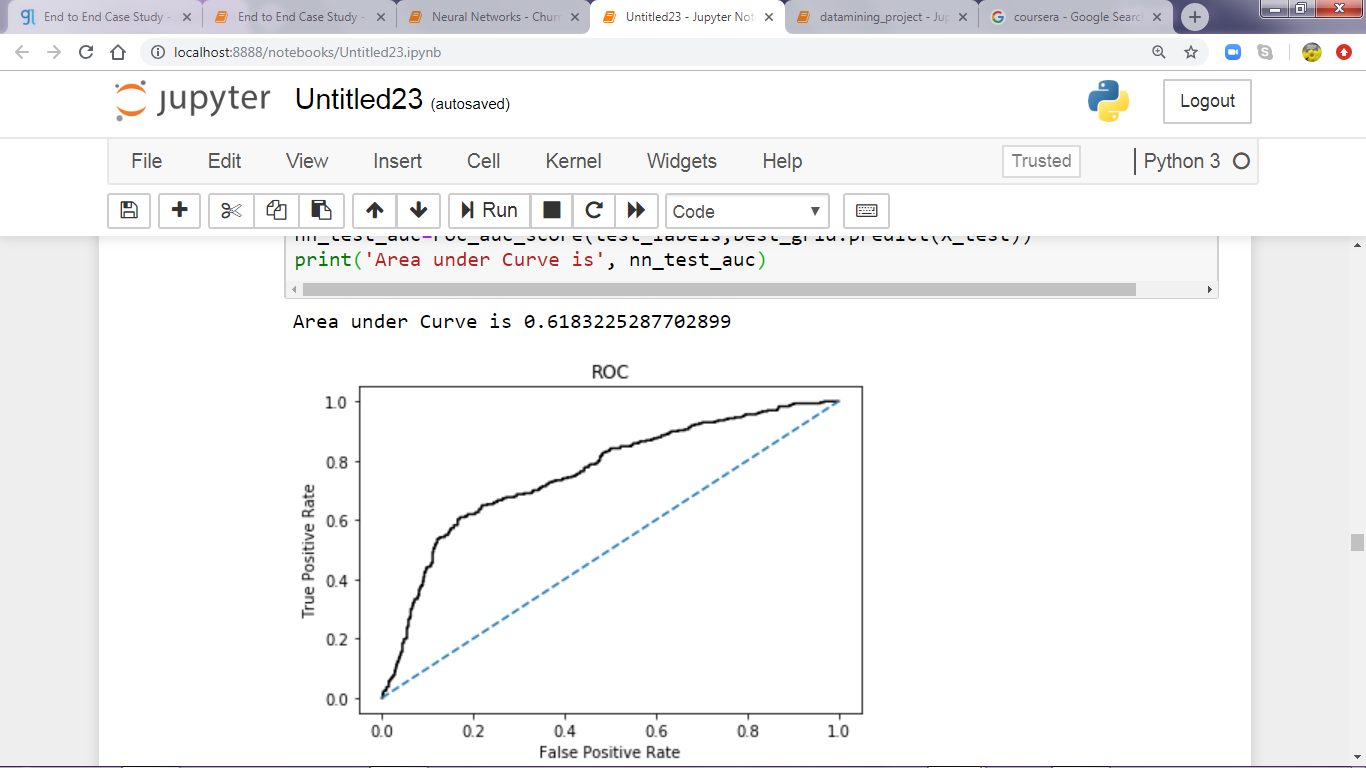
**Classification Report : Test data**

****

**AUC for train data**

****

**AUC for test data**

****

Train data Test data

AUC :60% AUC : 61%

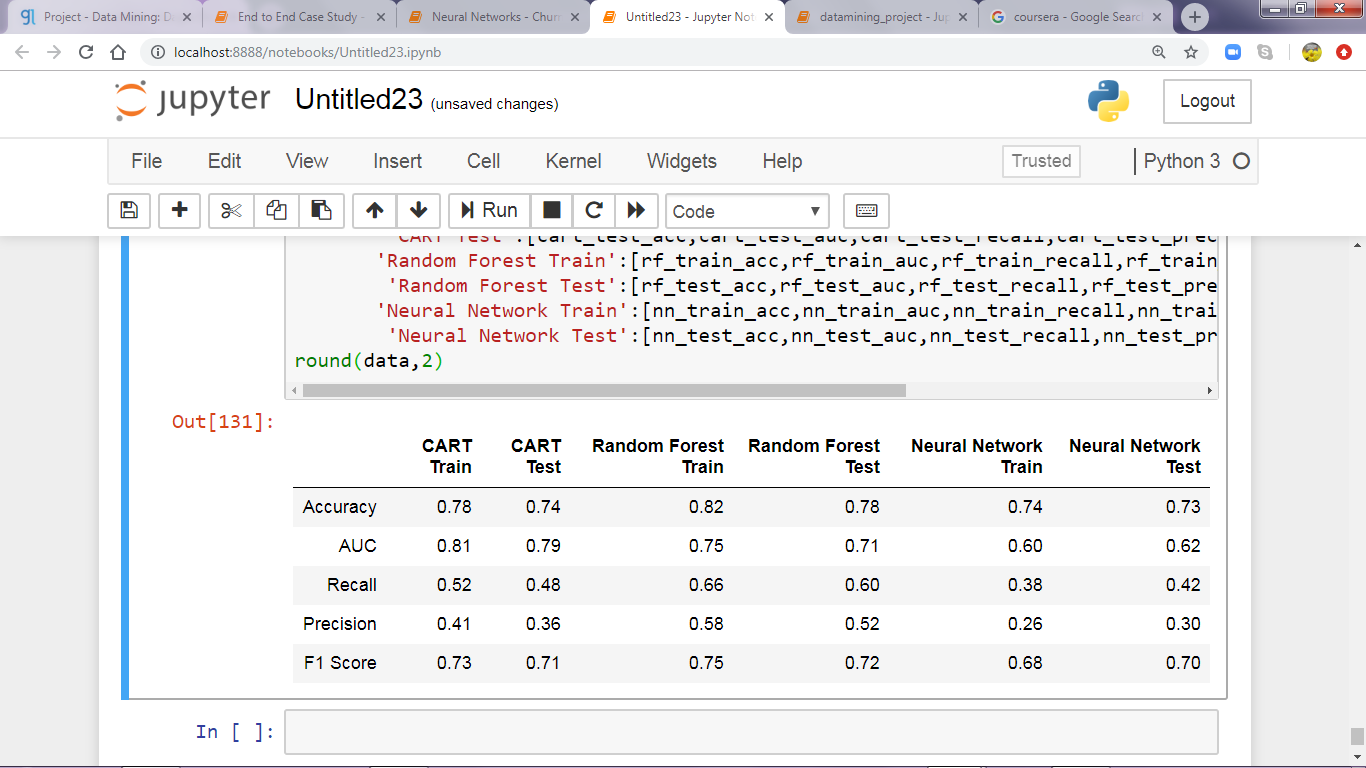
Accuracy : 74% Accuracy : 72%

Sensitivity: 26% Sensitivity :30%

Precision : 38% Precision :42%

f1-score:68% f1-score : 70%

**Performance Metrics:**

****

**Inference:**

* Based on the evidence andreasoning ***Random Forest*** performs best as compared to others.
* Greater transparency and statistical understanding has offered the Forest to do better with usability and prediction.
* Random Forest simplicity is used wisely for many classification and regression scenarios.
* Here the Random Forest avoids the Overfitting model.
* Thus Random Forest has found the most Important Feature from the training dataset.
* Here ‘AGENCY\_CODE’ is the most important feature for Predicting the Claimed status.

**Business Insights and Recommendations:**

* In ML algorithm it must to select the right approach and perspective, and also the data which we are using must be proper and cleaned one for building a model.
* The data must have a predictor variable, at the end there will be some allowance of error after building a model.
* As Business insights these algorithms improves the Data Quality check.
* Adding more data and Algorithm tuning will lead to Better model.
* The more the trees in it, will never allow the model to overfit.
* Random Forest can also handle Large Dataset with Higher Dimensionality.

R**ecommendations:**

* Based on these predictions Random Forest can be widely used in
  + Banking
  + E-Commerce
  + Stock Market
  + Medicines
* For High Accuracy it can be used.
* As Random Forest is a stochastic guessing method used widely in credit risk analysis.
* In this Insurance claim scenario it check with the data set and results that with the help of Agency code we can predict that insurance claimed or not and report to the management.