**BAN 5753**

**ADVANCED BUSINESS ANALYSIS**

**MINI PROJECT 2**

GitHub link : <https://github.com/NaveenKrRai/BAN5753-MiniProject>

**TEAM - LAMDA**

A picture containing person, indoor, suit, posing

Description automatically generated A person in a suit

Description automatically generated with medium confidence  A person in a suit and tie

Description automatically generated with medium confidenceA person in a suit and tie

Description automatically generated with medium confidence

Srujana Talla

A

A

Gowtham Bhupathiraju

A

A

Kodjo Opoku Botchway

Naveen Kumar Rai

A

A

**Business Problem**

We are given the dataset of XYZ bank’s direct marketing campaign. We have to work on the following problems:

* 1. The objective of the classification is to identify clients who will subscribe (yes/no) for a term deposit. (Variable y: Target function).
  2. The Bank wants us to conduct Exploratory Data Analysis (EDA) to identify relationships, and trends in data. For example correlations, bivariate analysis of target versus input variables, facts, univariate patterns, missing data.
  3. Prescriptive recommendations
  4. K-means Clustering.
  5. Compare more than four different supervised algorithms

**Dataset Information:**

* Data is about an XYZ bank’s direct marketing campaign. Marketing campaigns were driven by telephone calls.
* Data Set In many cases, more than one contact for the same client was required., in order to access if the product (deposit) would be ('yes') or not ('no') subscribed
* The purpose of the classification is to forecast whether the customer will signup (yes/no) a term deposit (variable y).
* The dataset: XYZ\_Bank\_Deposit\_Data\_Classification.csv, 20 entries/columns, sorted by date between May 2008 and November 2010.

**Our Approach:**

**Exploratory Data Analysis**

* Read the data

**Text

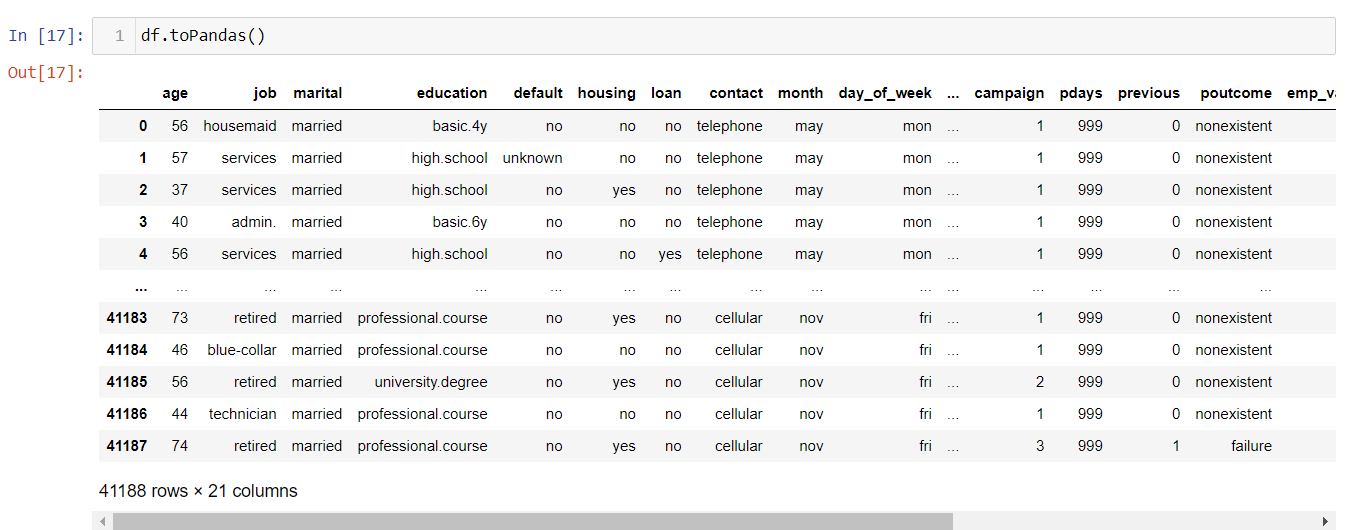
Description automatically generated**

* Rename the columns

**Text

Description automatically generated**

* Displays the Data frame



**Summary Statistics:**

Graphical user interface, application, table

Description automatically generated

**Schema:**

**Text

Description automatically generated**

**Missing Data:**

* We did not find any missing values in the data

Text

Description automatically generated

**Numeric Features:**

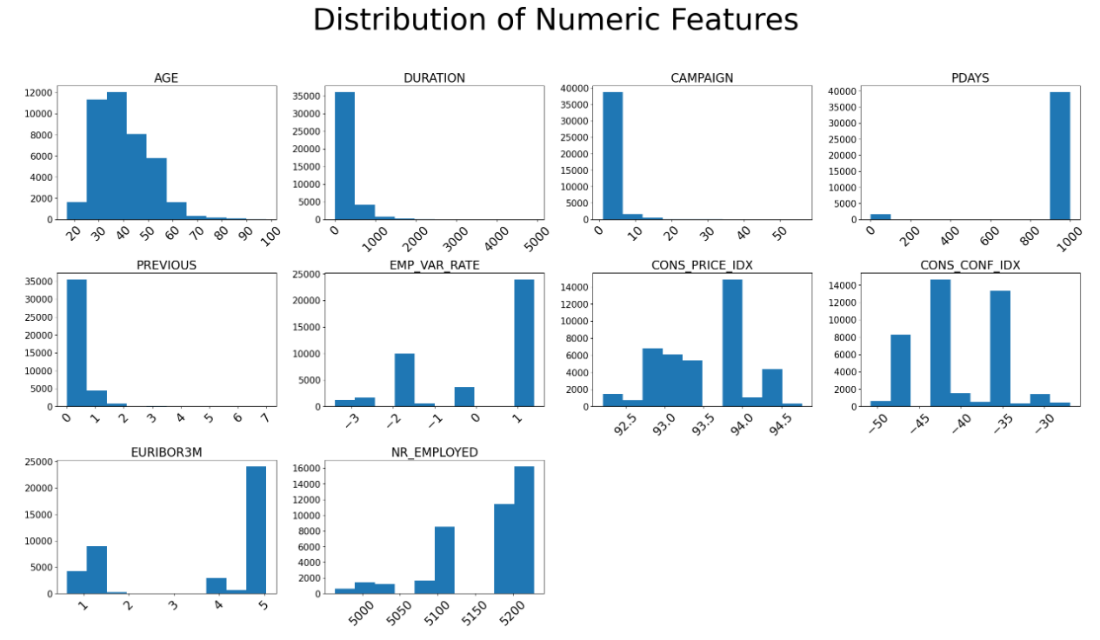
* There are a total of 10 numeric variables in the data set

Graphical user interface, application

Description automatically generated

* Univariate Patterns of the Numeric variables

From the patterns we can observe that the is not normal distribution of these variables. Most of them are skewed to the left and we can also see few BiModal distributions



* Correlations between Numeric Features

**A picture containing text

Description automatically generated**

* The variables emp\_var\_rate and euribor3m have the highest correlation.

**Categorical Features:**

* We have a total of 11 categorical variables of which y is the target variable
* Below snippet explains the distinct values in each of the categorical variable

Graphical user interface, application

Description automatically generated

* Univariate Patterns of the categorical variables

**Scatter chart

Description automatically generated**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Bivariate Analysis of Target vs Input Variables:**

**Timeline

Description automatically generated**

**Timeline

Description automatically generated**

**Modelling**

* Split the data into 80% train and 20% test data.

**Text

Description automatically generated**

* Preprocessing like – String Indexing, Standard Scaler and Vector Assembler

**A picture containing text

Description automatically generated**

* Train and test data set with the processed features

**Text

Description automatically generatedGraphical user interface, text, application

Description automatically generated**

**Predictive Model 1- Logistic Regression**

* The first classification model is Logistic Regression. Below are the predictions from the model

**A picture containing table

Description automatically generated**

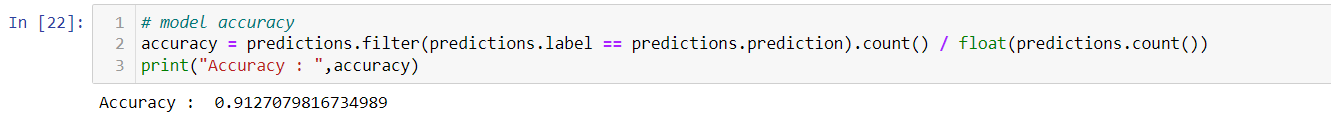
**Confusion Matrix:**

* The True Positives is 400 and True Negatives is 7170 = 7570
* The False Positives is 182 and False Negative is 542 = 724

**Chart, waterfall chart

Description automatically generated**

**Model Accuracy: 91.27%**

****

**ROC Curve:**

**Graphical user interface, text, application

Description automatically generated**

**Test AUC: 93.85%**

**Text

Description automatically generated**

**Predictive Model 2 – Decision Tree**

* The second classification model is Decision Tree - classifier. Below are the predictions from the model

**A picture containing text

Description automatically generated**

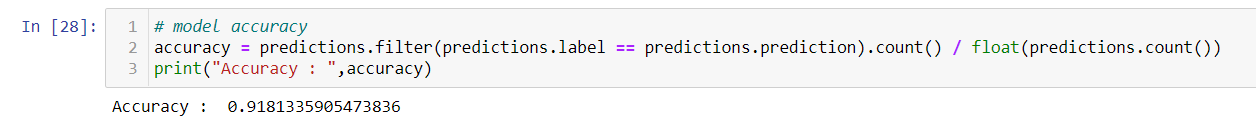
**Confusion Matrix:**

* The True Positives is 517 and True Negatives is 7098 = 7615
* The False Positives is 254 and False Negative is 425 = 679

**Chart, waterfall chart

Description automatically generated**

**Model Accuracy: 91.81%**

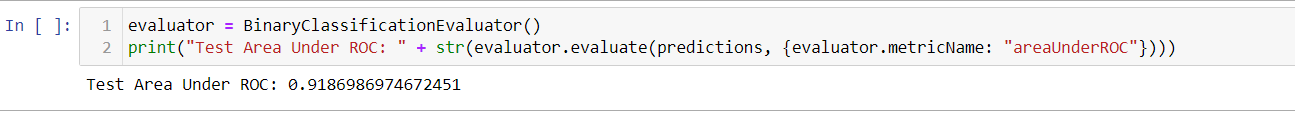
****

**ROC Curve:**

**Chart

Description automatically generated**

**Test AUC: 91.87%**

****

**Predictive Model 3 – Random Forest**

* The Third classification model is Random Forest - classifier. Below are the predictions from the model

**A picture containing text

Description automatically generated**

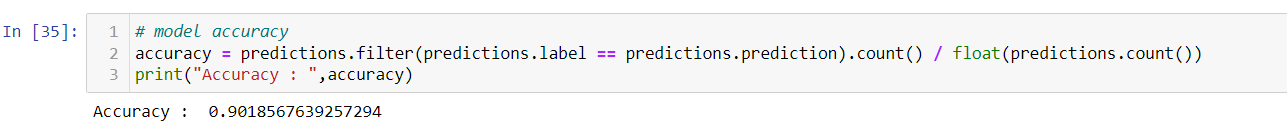
**Confusion Matrix:**

* The True Positives is 161 and True Negatives is 7319 = 7480
* The False Positives is 33 and False Negative is 781 = 814

**Chart, waterfall chart

Description automatically generated**

**Model Accuracy: 90.185%**

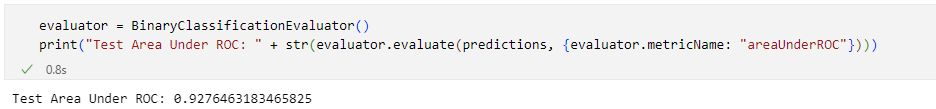
****

**ROC Curve:**

**Chart

Description automatically generated**

**Test AUC: 92.76%**

****

**Predictive Model 4 – Gradient Boosting**

* The Fourth classification model is Gradient Boosting - classifier. Below are the predictions from the model

**Graphical user interface, text, application

Description automatically generated**

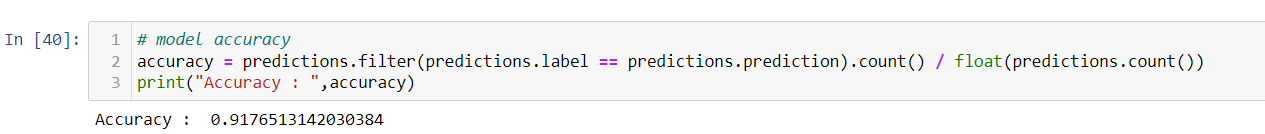
**Confusion Matrix:**

* The True Positives is 501 and True Negatives is 7110 = 7611
* The False Positives is 242 and False Negative is 441 = 683

**Chart, waterfall chart

Description automatically generated**

**Model Accuracy: 91.76%**

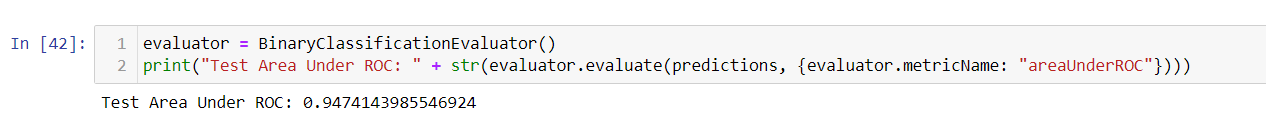
****

**ROC Curve:**

**A picture containing graphical user interface

Description automatically generated**

**Test AUC: 94.74%**

****

**Champion Model**

|  |  |  |
| --- | --- | --- |
| **Model Name** |  | **Test AUC Value** |
| Logistic Regression |  | 0.9385 |
| Decision Tree |  | 0.9186 |
| Random Forest |  | 0.9276 |
| Gradient Booster |  | 0.9474 |

* On comparing the four models **Gradient Boosting** has the best Model accuracy of – 91.76%, True Positives and Negatives of 7611 and Test AUC of 94.74%.
* Hence, we choose this model as our best model for classification of identify clients who will

subscribe (yes/no) for a term deposit.

* Here are the top 10 Important Features of our Gradient Boosting Model

**Table

Description automatically generated**

**K Means Clustering**

* We clustered the given customers into 5 clusters based on the features.

**Table

Description automatically generated**

* Maximum number of clusters are in the group 0 followed by group 2

**Table

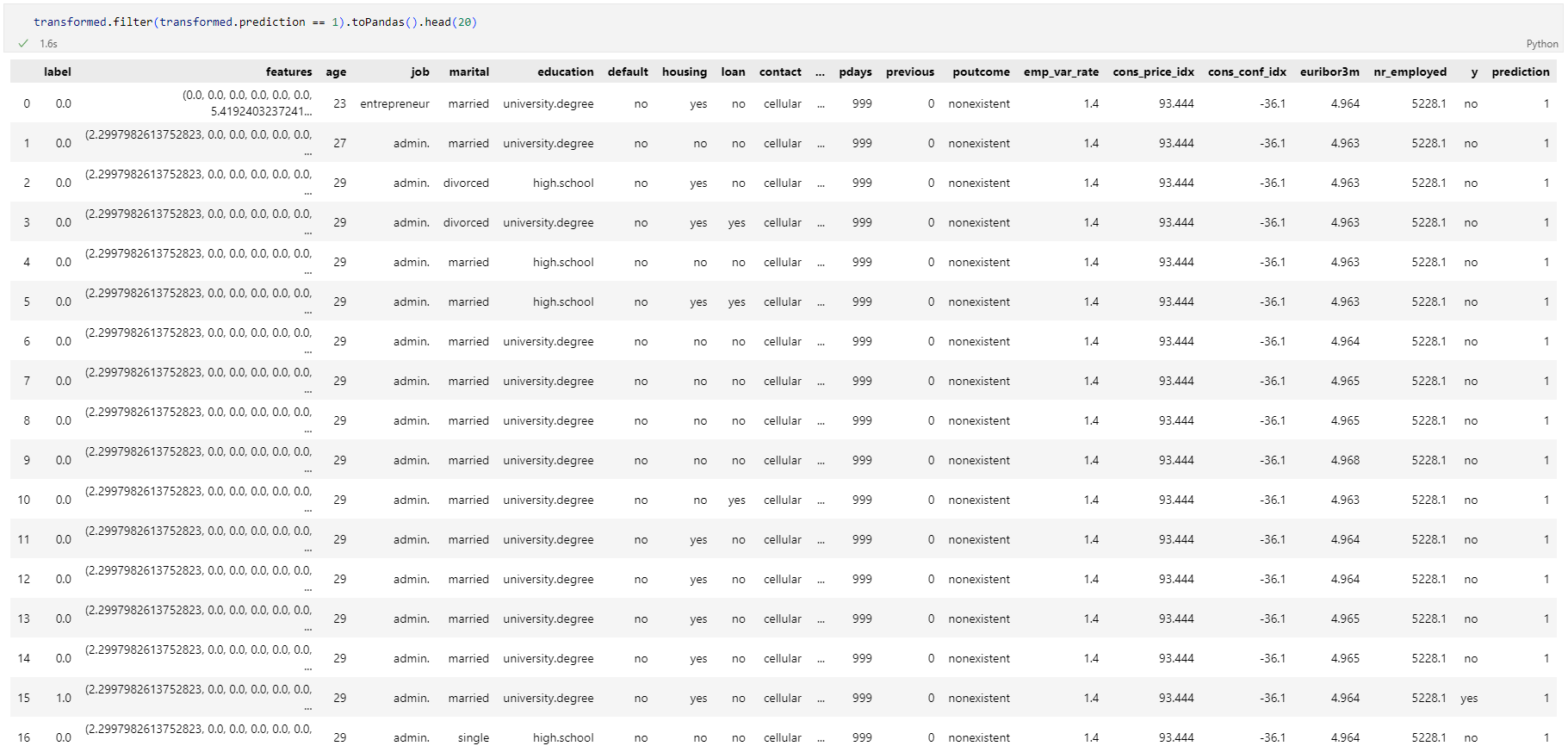
Description automatically generated with medium confidence**

* In cluster 0 most customers are single and have basic or high school education. They usually are admins and blue collared employees.

**Table

Description automatically generated**

* In cluster 1 most customers are married or divorced and have a university degree education. They usually are admins.

****

* In cluster 2 most customers are single and have a basic or high school education. They usually are mostly students.

**Graphical user interface, application, table, Excel

Description automatically generated**

* In cluster 3 most customers are single or divorced and have a high school education. They usually are mostly students and technicians.

**Table

Description automatically generated**

* In cluster 4 most customers are single and have a university or professional education. They usually are working people who use the telephone.

**Table

Description automatically generated**

**Our Recommendations**

* Since most of the customers for the bank lie in the age group 30-50, educating this age group will be more beneficial to make them sign up for term deposit.
* Contact duration less than 2000 seconds will be beneficial in conversion of clients.
* Contacting less than 20 times during the campaign for a particular customer would end the bank with a high number of clients subscribing for term deposit.
* Clients who don’t have credit are the target segment.
* Most of the clients are people with a job – admins, Blue-collared, service.
* May is the month of most conversion – focus should be laid around that duration.
* From the Prediction Model we also see that duration, Number of employees, age, cons\_conf\_idx, campaign are the top features for prediction. Working on them would yield better results.