CAPSTONE PROJECT 3
BANK MARKETING EFFECTIVENESS PREDICTION

TEAM MEMBERS:

K. YOGESH REDDY

D. PRUTHVI RAJ

INTRODUCTION:

Given the data of Direct marketing campaigns of Portuguese bank company and these marketing campaigns are primarly focused on phone calls. These phone calls are to check whether the client is subscribed to the short-term-deposit of the particular bank.

Problem Statement:

The data is related to direct marketing campaigns(phone calls) of a portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same clinet was required, inorder to access if the product (bank term deposit) would be (‘yes’) or not (‘no’) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data:

The Provided data has around 45000 obsevations client data portuguese banking coorporarions.

The dataset contains 45211 rows and 17 columns.

- Age(Numeric)
- Job : type of job (Categorical : Admin , Blue-Collar , Entrepreneur , Housemaid , Management , Retired , self-employed , Services , Student , Technician , Unemployed , Unknown)
- Martial : Marital Status(Categorical : Divorced , Single , Unknown)
- Education : (Categorical : Basic.4y , Basic.6y , Basic.9y , High school , Illiterate , Proffesional.Course , University.Degree , Unknown)
- Default : Has credit in default? (Categorical : No,Yes,Unknown)
- Housing : Has house loan? (Categorical : Yes,No,Unknown)
- Loan : Has personal loan? (Categorical : Yes,No,Unknown)

Related with last contact of current campaign:

- Contact : Communication type
- Month : last contact month of year
- Day_of_week : last contacted day of the week
- Duration : Last contact duration, in seconds. Important note: this attribute highly affects the output target. Yet,the duration is not known before a call is performed. Also, after the end of call y is obviously known.

Other attributes :

- Campaign : number of contacts performed during this campaign and for this client(numeric, includes last contact)
- Pdays : number of days that passedny after the client was last contacted from a previous campaign(999 means client was not previously contacted)
- Previous : number of contacts performed before this campaign and for this client
- Poutcome : Outcome of the previous marketing campaign

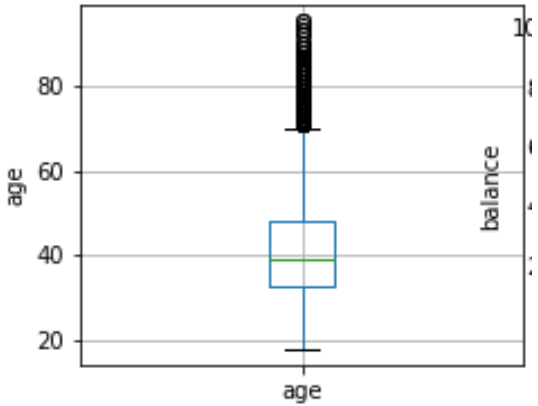
Output variable (desired target) :

- Y – Has the client subscribed a term deposit? (binary Yes,No)

EDA(EXPLORATORY DATA ANALYSIS):

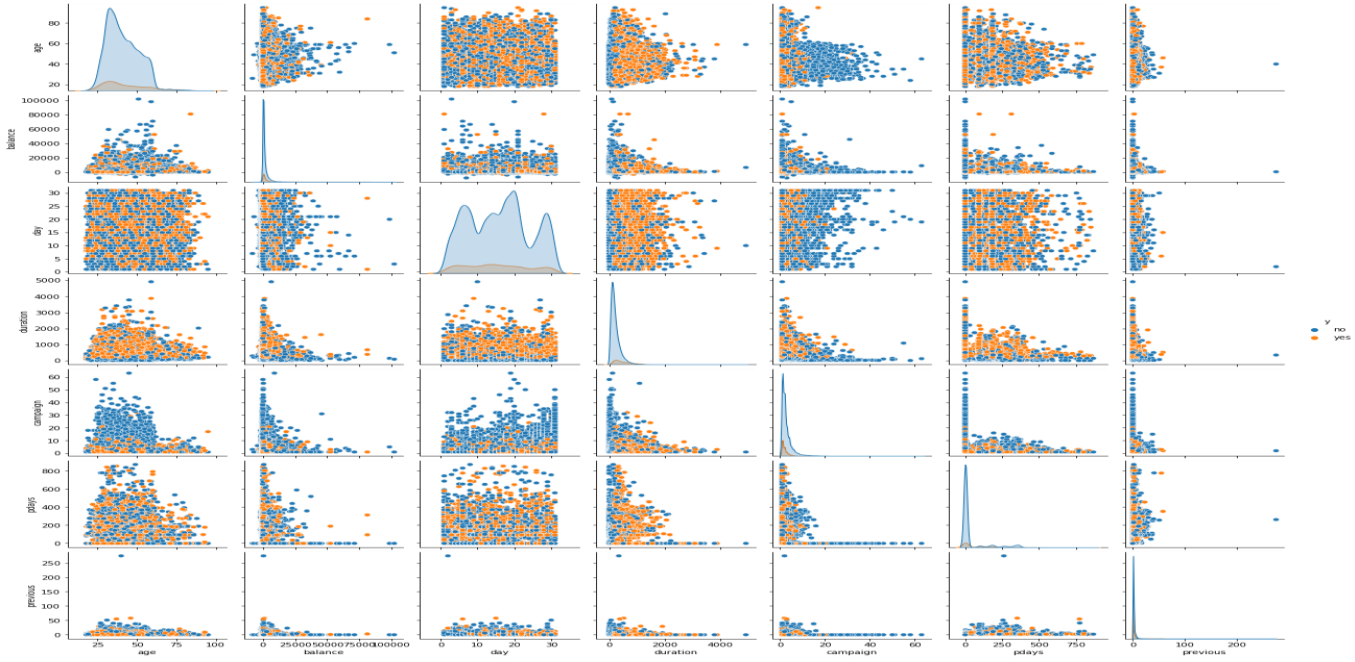
- The given data has 45211 rows and 17 columns.
- There were no null values in the data.
- There were no duplicate data in the dataset.
- There were 7 int datatype columns and 10 object datatype columns.
- Age column has outlier data, for this problem age can only be considered as outlier as other factors doesn’t depend on target feature.
- The given dataset is imbalanced, where target feature has imbalanced data.

Outlier data:

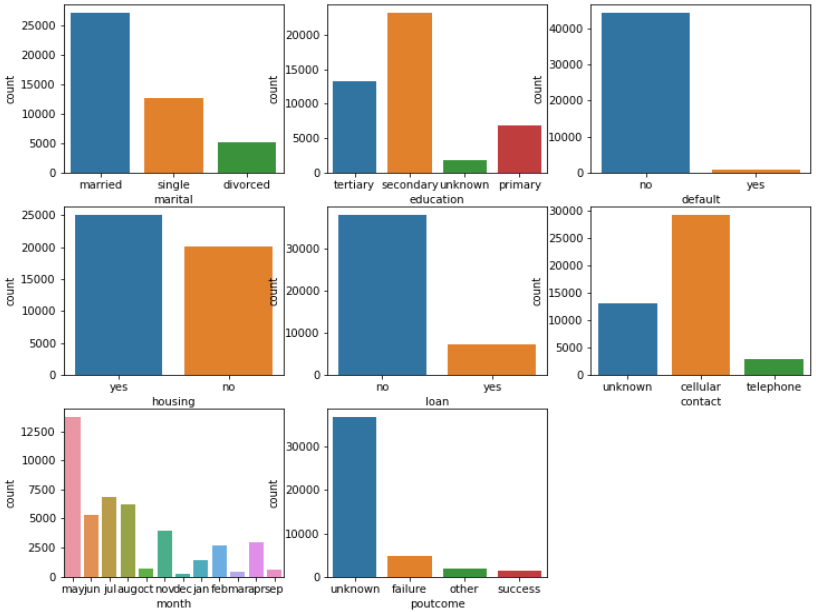


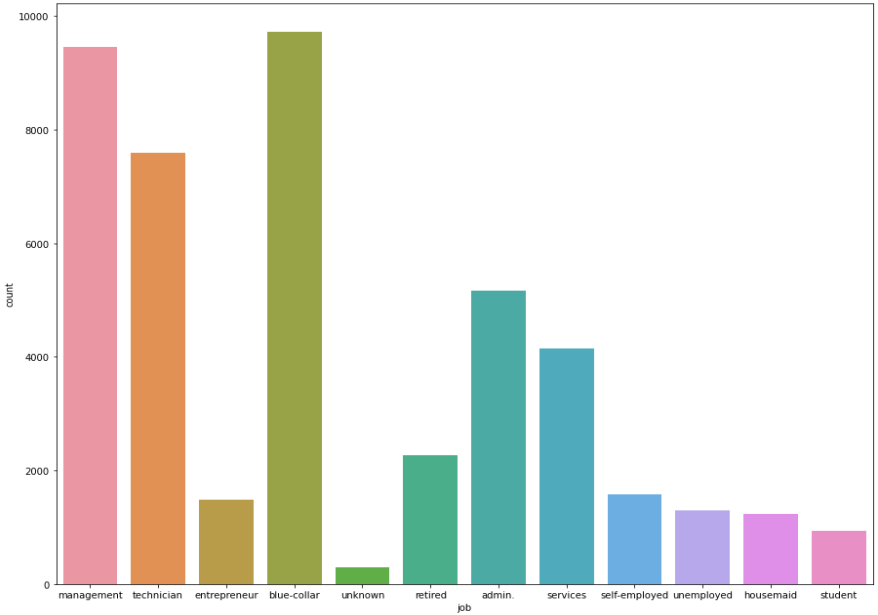
Here above 70 age are outliers, where customers above 70 age are not perfect for prediction, so we will remove these outlier data using iqr.

The data after outlier removal are 44724 rows and 17 columns.

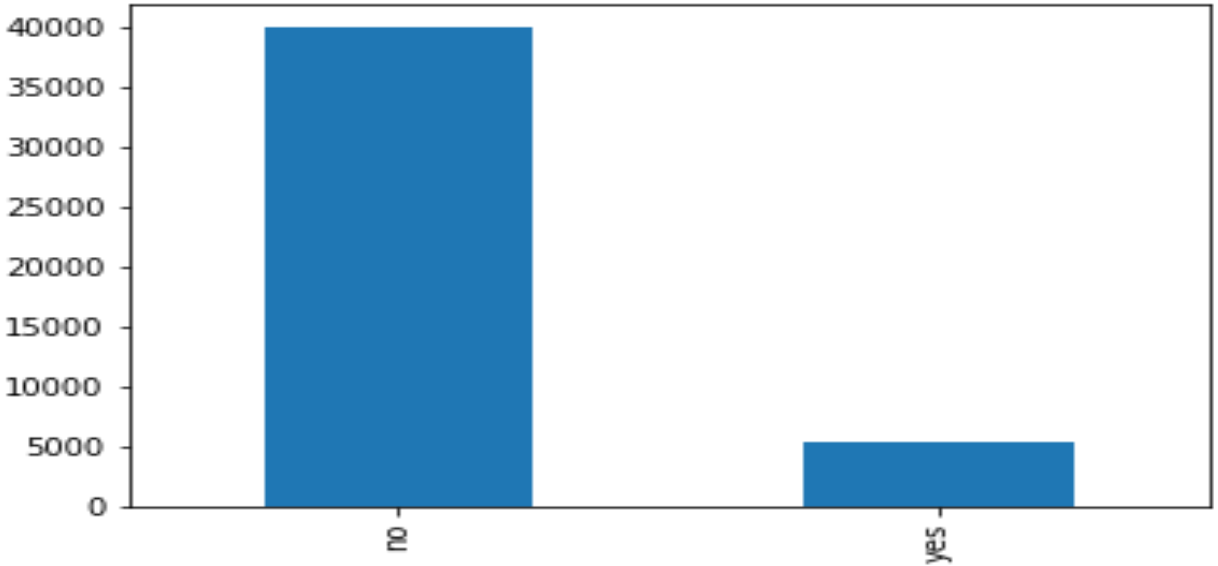


Here we have performed pair plot on data where this plot gives dependency of features with eachother.

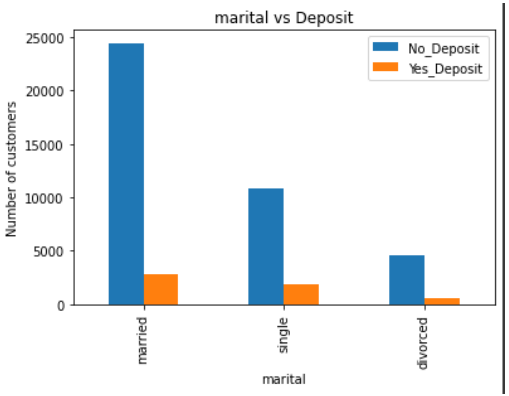


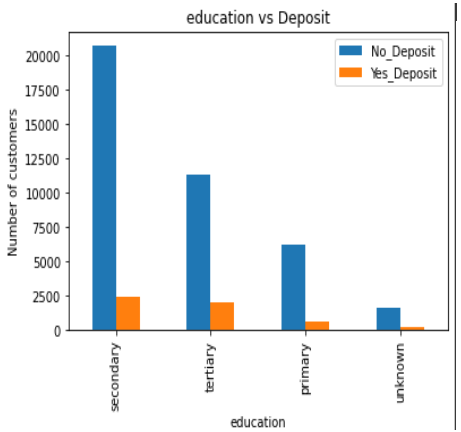


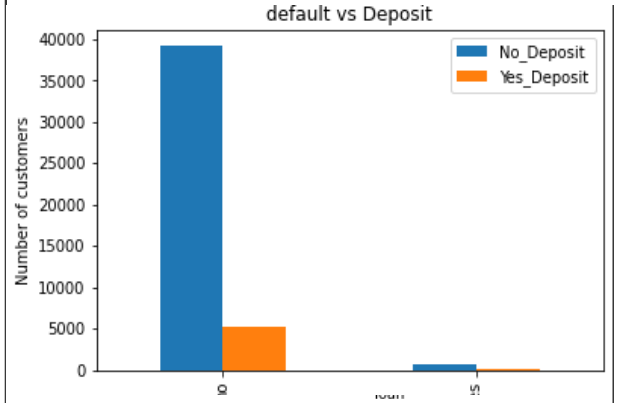
The above count plots gives total categories and their count in each categorical features.

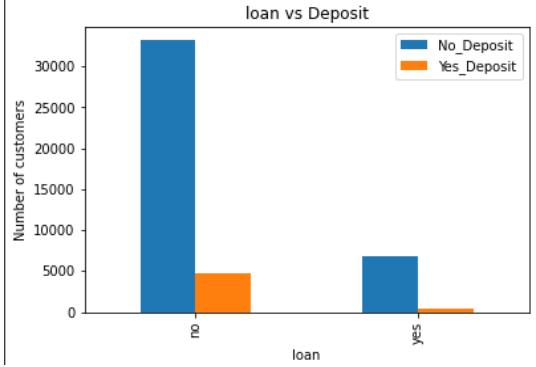


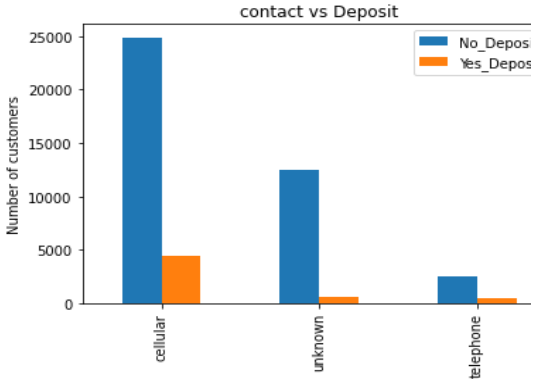
The above target feature count plot shows the total number of yes's and no's. We can clearly say that the data is imbalanced we have to resample further and implement Ml models and should compare with Ml models created on imbalanced data.



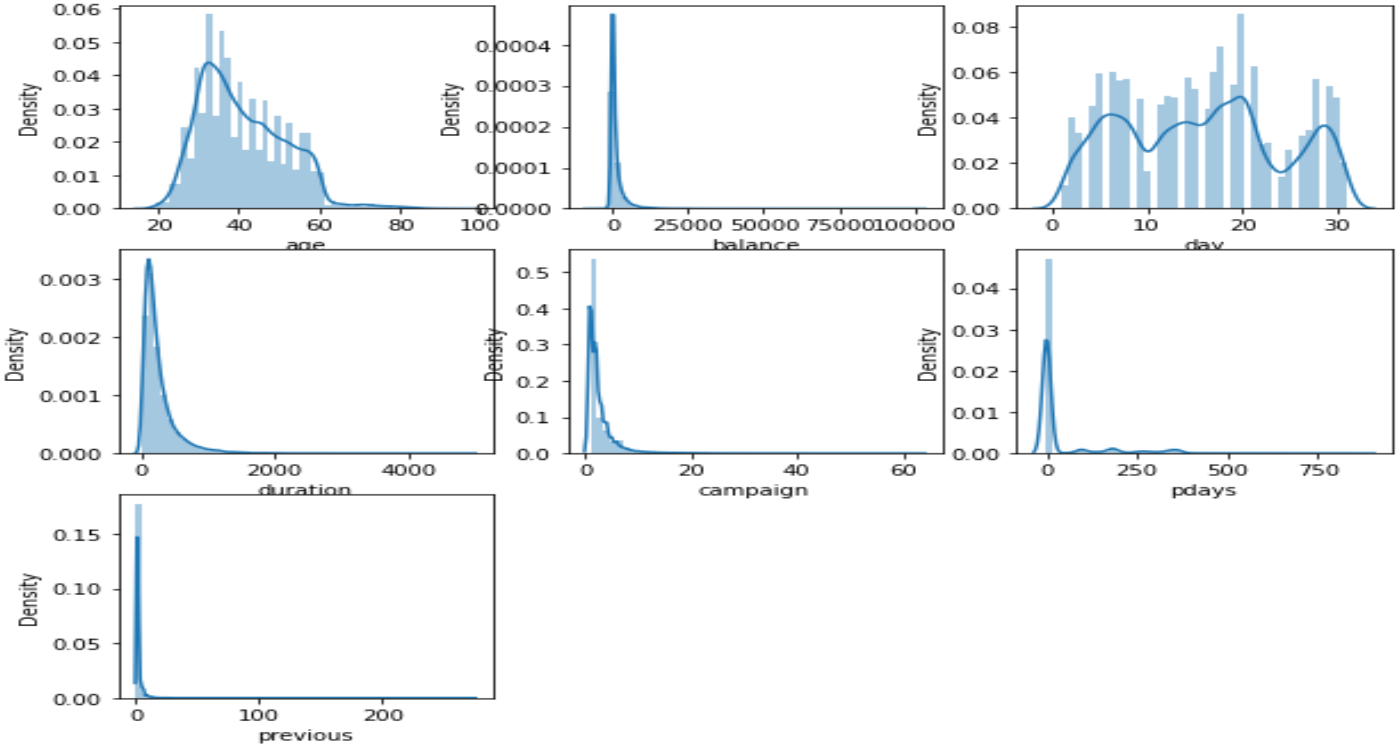




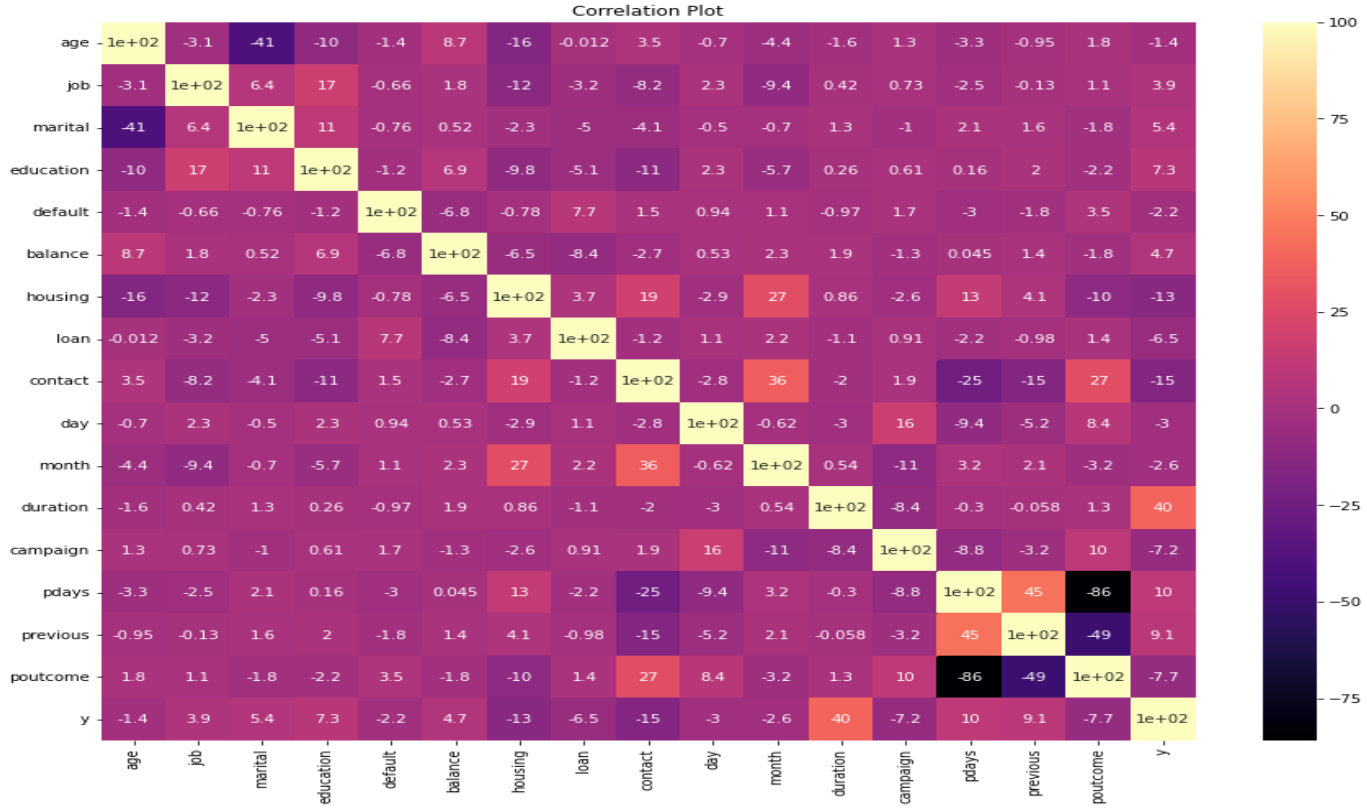




The above bivariate plots gives information how each feature is dependent on deposit.

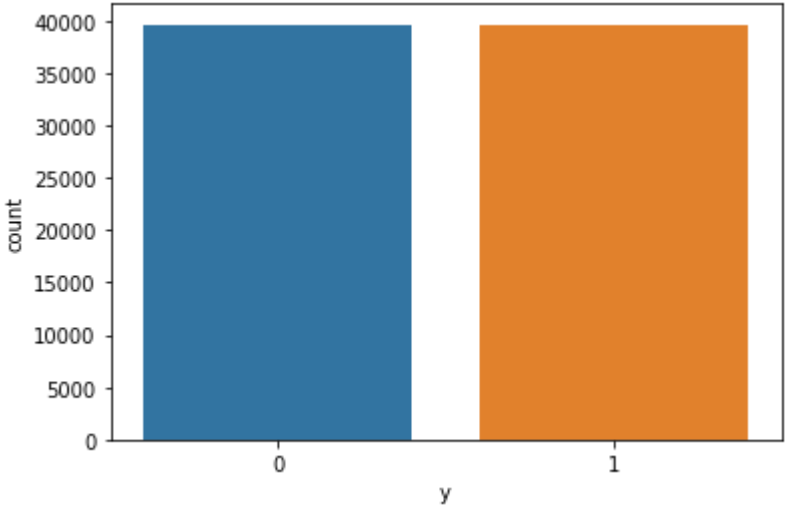


The above plot shows how each feature is distributed, out of all this features balance, pdays and previous are normally distributed.



The above correlated plot shows how each feature is correlated with other features.

SMOTE(Synthetic Minority Oversampling Technique):

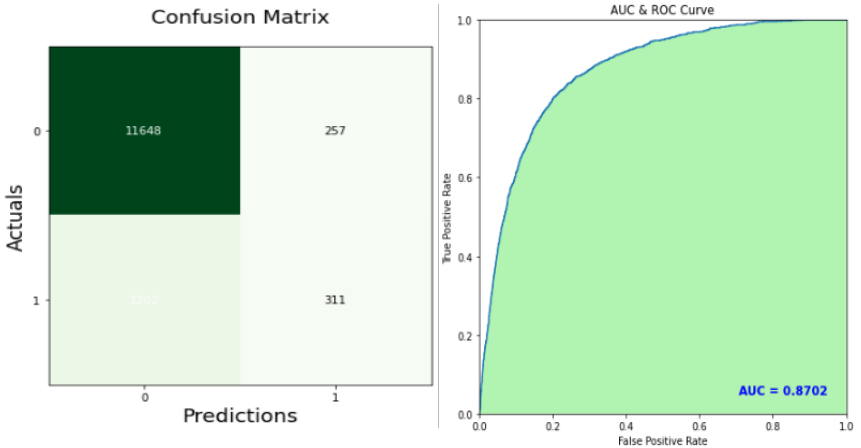


Before resampling there were 39653 0’s and 5071 1’s. After performing SMOTE we can see that 1’s are now increased to 39653 which is same of total 0’s. Here SMOTE randomly takes rows with 1’s and creates similar rows.

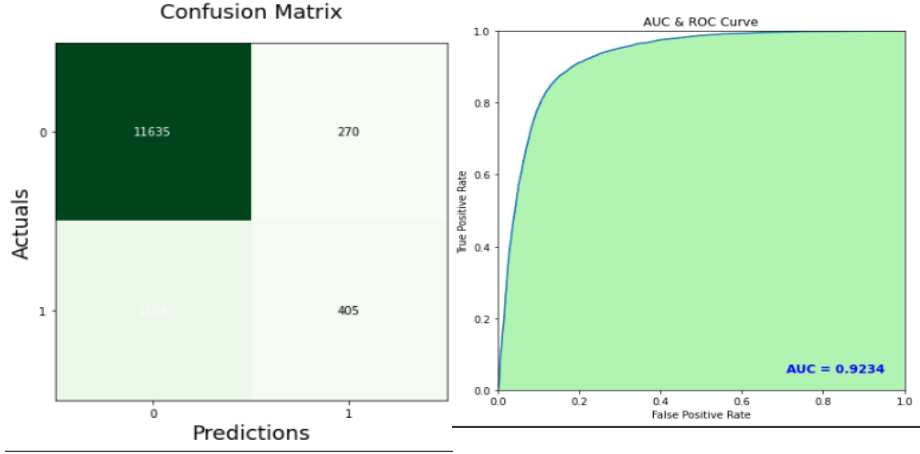
Model Building and Training Modles:

Logistic Regresson:

Here we have build models on imbalanced and data and resampled data.



Imbalanced Data

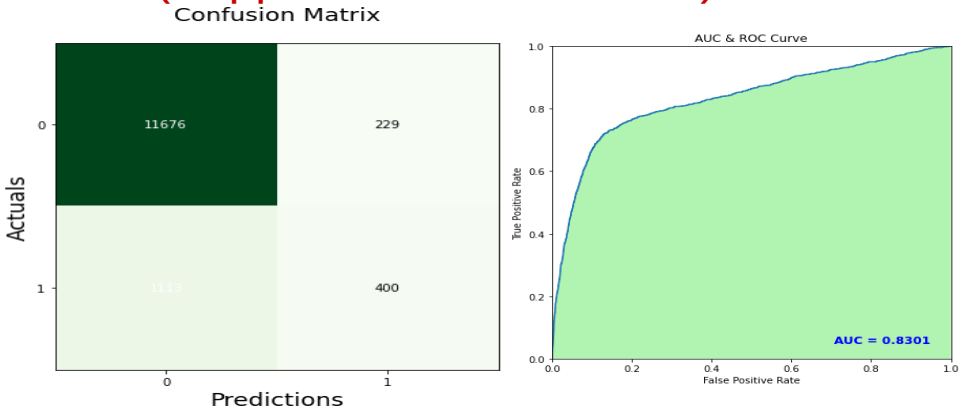


Resampled Data

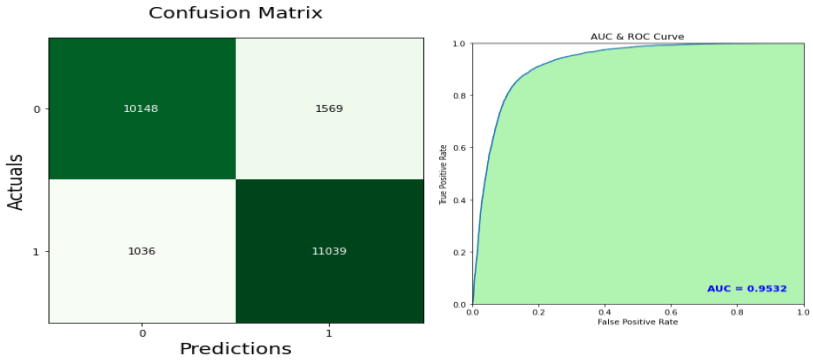
The accuracy of imbalanced data is 89.1%

The accuracy of resampled data is 86%.

SVC(Support Vector Classifier):



Imbalanced Data

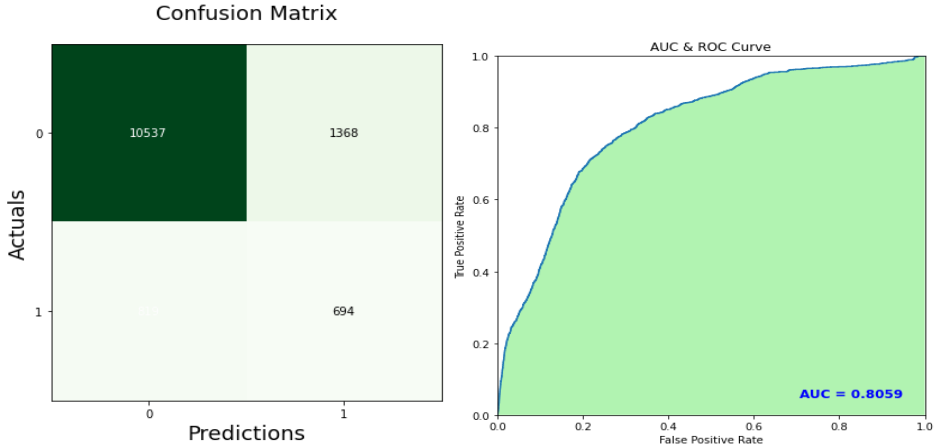


Resampled Data

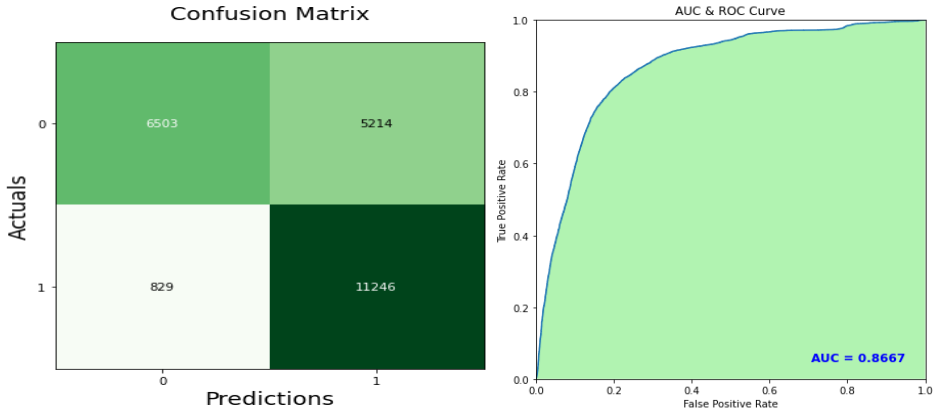
The accuracy of the model with imbalanced data is 89.9%.

The accuracy of the model with resampled data is 89%.

Gaussian Naive Bayes:



Imbalanced Data

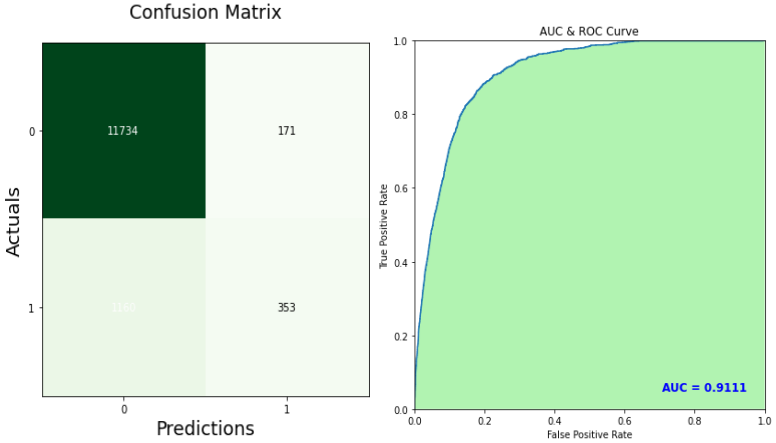


Resampled Data

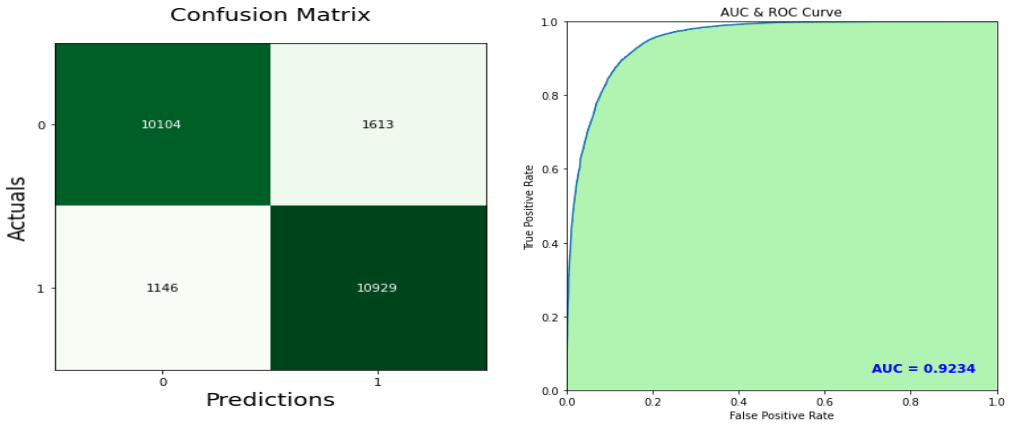
The accuracy with imbalanced data is 84%.

The accuracy with resampled data is 74%.

Random Forest Classifier:



Imbalanced Data

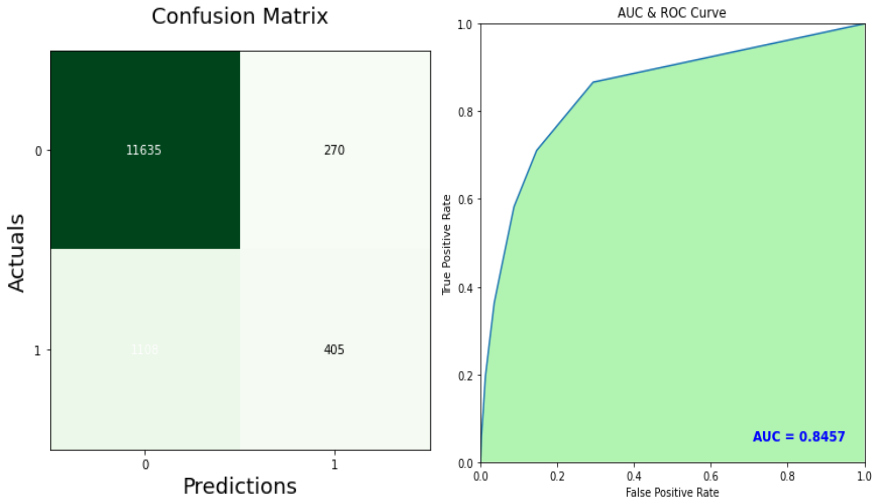


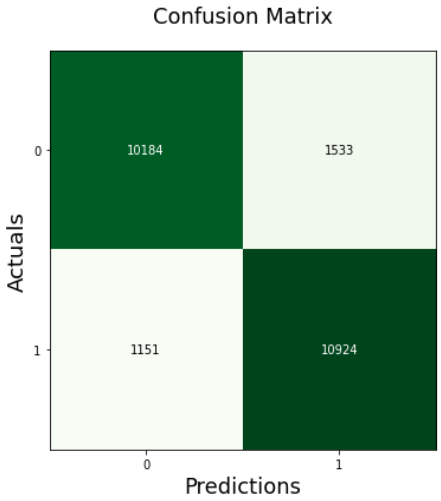
Resampled Data

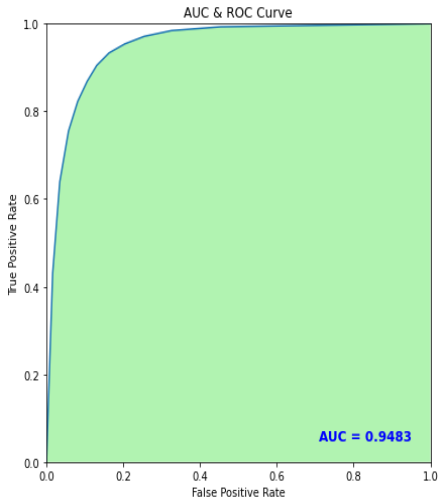
The accuracy of the model with imbalanced data is 90%,

The accuracy of the model with resampled data is 88%.

KNN:







Imbalanced Data

Resampled Data

The accuracy of the model with imbalanced data is 89.7%.

The accuracy of the model with resampled data is 88%.

Boosting Techniques:

AdaBoost Classifier:

AdaBoost classifier has given accuracy of 90% for imbalanced data.

AdaBoost classifier has given accuracy of 62.5% for resampled data.

Gradient Boost Classifier:

Gradient Boost classifier has given accuracy of 90.6% for imbalanced data and for resampled data accuracy is 91.5%

XGBoost Classifier:

XGBoost Classifier has given accuracy of 90.5% for imbalanced data and for resampled data accuracy is 93%.

Conclusion:

- The Data related to marketing campaign of a Portuguese bank company has been successfully visualized and analysed. After drawing all the possible insights the data has been worked on pre-processing to undergo all the possible machine learning classification models after carefully fine tuning it.
- Metrics of the classification models with normal sampling of dependant variable :

➢ Logistic Regression – Accuracy : 89%
➢ SVM Classification – Accuracy : 89.9%
➢ Gaussian Naive Bayes – Acuuracy : 83.9%
➢ Random Forest Classifier – Accuracy : 90%
➢ KNN Classifier -- Accuracy : 89.7%

Boosting Methods :

➢ ADA Boost Classifier – Accuracy : 90%
➢ Gradient Boost Classifier – Accuracy : 90.6%
➢ XG Boosting Classifier – Accuracy : 90.5%

- Metrics of the classification models after over-sampling the dependant variable :

➢ Logistic Regression – Accuracy : 86.2%
➢ SVM Classification – Accuracy : 89%
➢ Gaussian Naive Bayes – Acuuracy : 74%
➢ Random Forest Classifier – Accuracy : 88.4%
➢ KNN Classifier -- Accuracy : 88.7%

- Boosting Methods :

➢ ADA Boost Classifier – Accuracy : 62.5%
➢ Gradient Boost Classifier – Accuracy : 91.5%
➢ XG Boosting Classifier – Accuracy : 93%

After closely observing all the performed classification models and we can clearly see the over-sampled XG Boosting classifier gave the best results and the predicted values were on par on the test dataset.

THANK YOU