Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2021-22

CUSTOMER BUYING HABITS IN BANKING DOMAIN

GROUP MEMBERS

UMANG KIRIT LODAYA 60009200032

JAY KANHAIYA BHANUSHALI 60009200047

Guided By

Dr. Kriti Srivastava

CHAPTER 1: INTRODUCTION

Aim: Design a classifier to get customer buying habits in banking domain

- **Problem description:**
  - Our problem statement was about customer buying habits in Banking. We used this dataset to classify whether a customer will purchase the *term deposit subscription* based on the existing data available with the bank and data collected from the campaigns.

CHAPTER 2: DATA DESCRIPTION

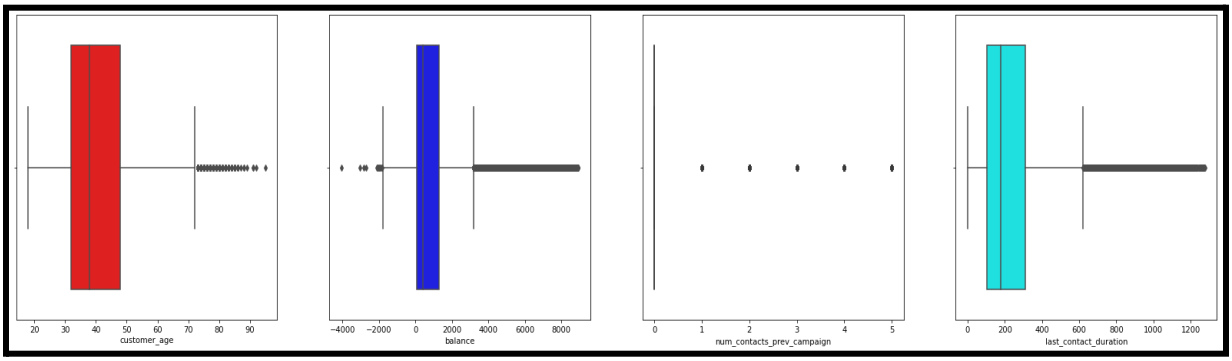
- The data is related to direct marketing campaigns of a Portuguese Banking Institution.
- The marketing campaigns were based on phone calls.
- Often, more than one contact with the same client was required, to assess if the product (Bank Term Deposit) would be subscribed (1) or not subscribed (0).

CHAPTER 3: DATA ANALYSIS

Pre-processing

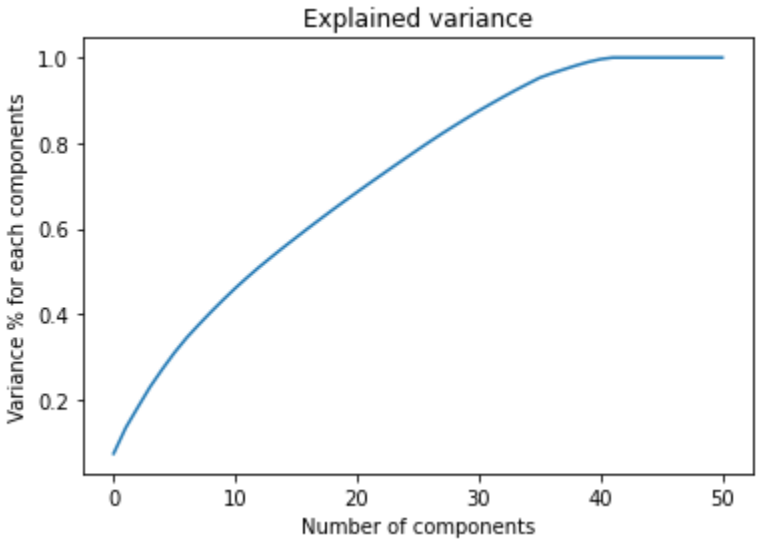
1. **Handled Duplicate Values**

2. **Preprocessing in Training Data**
   a. Filled Null Value With Median In Columns Dtype Float64
   b. Filled Null Values With Mode In Columns With Type Object
   c. Filled Null Values In Numerical Columns
   d. Filled Null Value With Mode In Columns With Dtype Object
   e. Handled Outliers



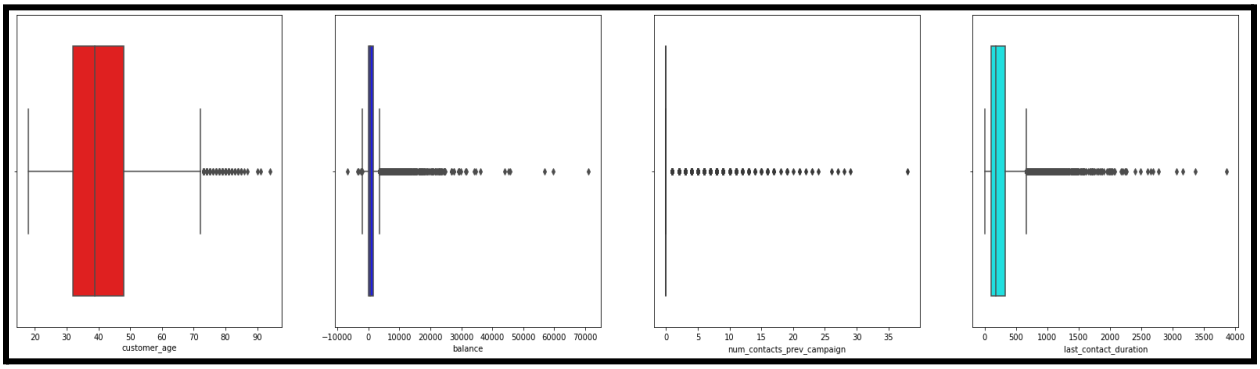
Removed 10 percentile outliers from dataset

f. Label Encoded with One-Hot Encoding
g. Performed PCA and got 40 number of components



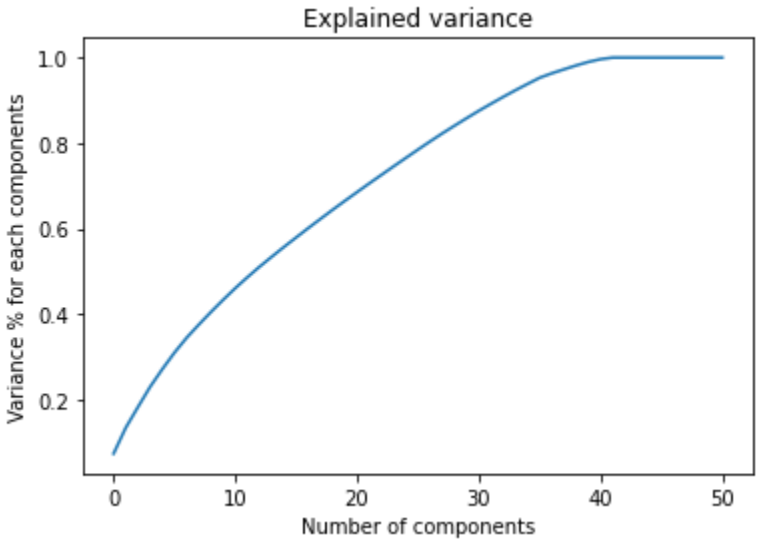
3. Preprocessing in Testing Data

a. Filled Null Value With Median In Columns Dtype Float64
b. Filled Null Values With Mode In Columns With Type Object
c. Filled Null Values In Numerical Columns
d. Filled Null Value With Mode In Columns With Dtype Object
e. Handled Outliers



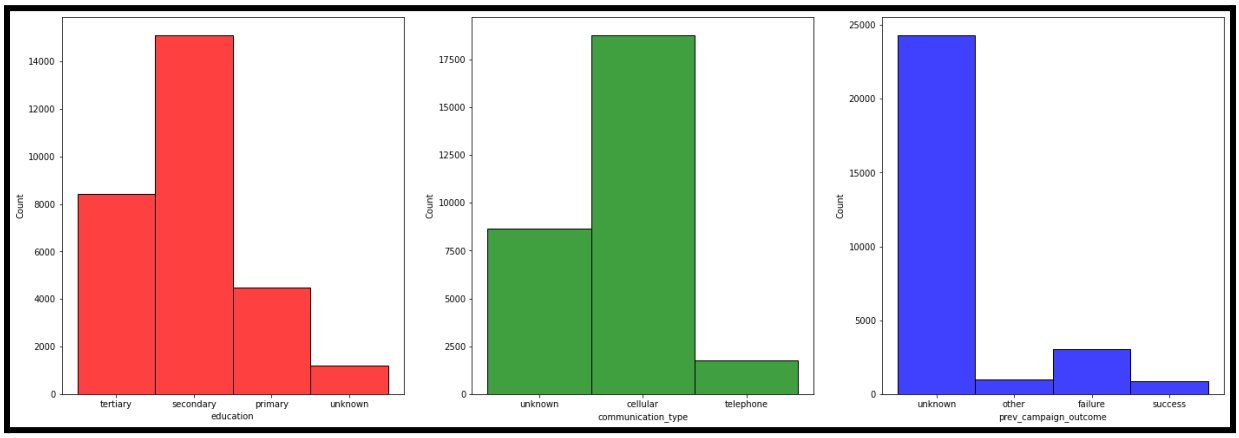
Removed 10 percentile outliers from dataset

f. Label Encoded with One-Hot Encoding
g. Performed PCA and got 40 number of components

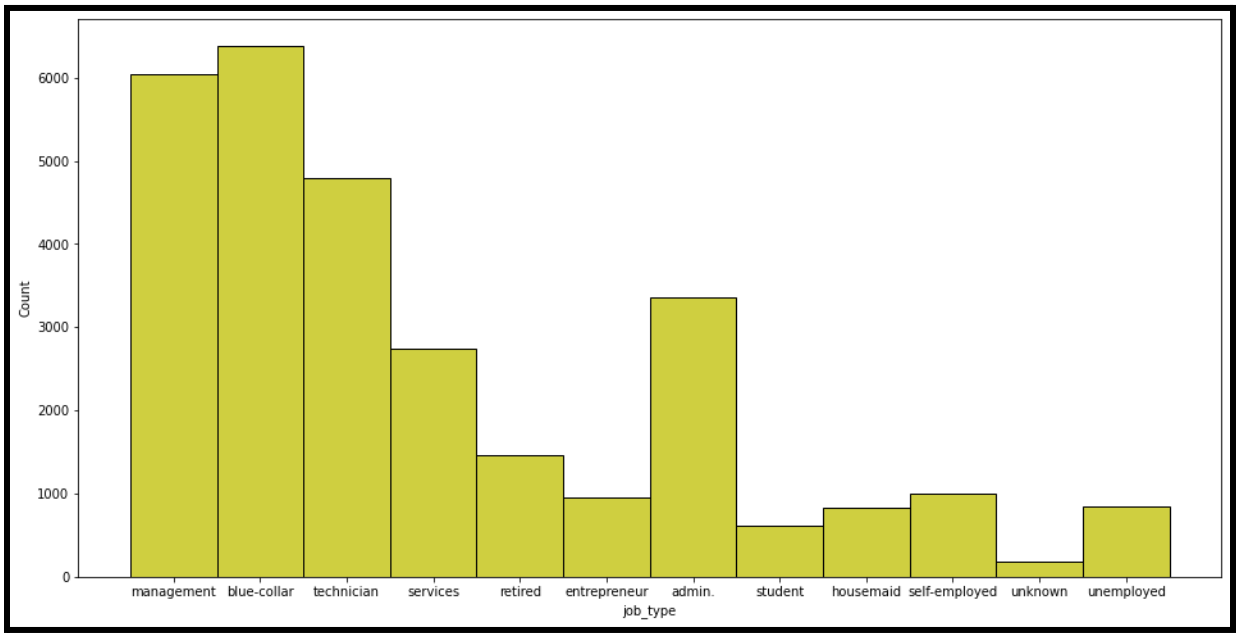


Exploratory data analysis

1. Plotted Histogram of features

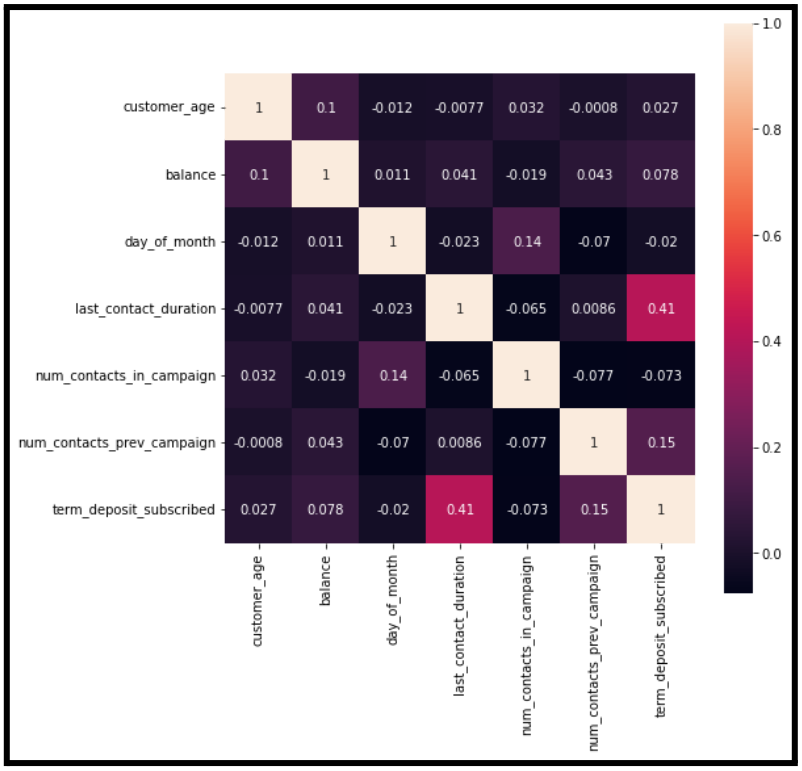


Histogram of Education, Communication_Type, and Prev_Campaign_Outcome



Histogram of Job_Type

2. Plotted Heatmap, to get dependencies of features



CHAPTER 4: MODEL MAKING

Models used:

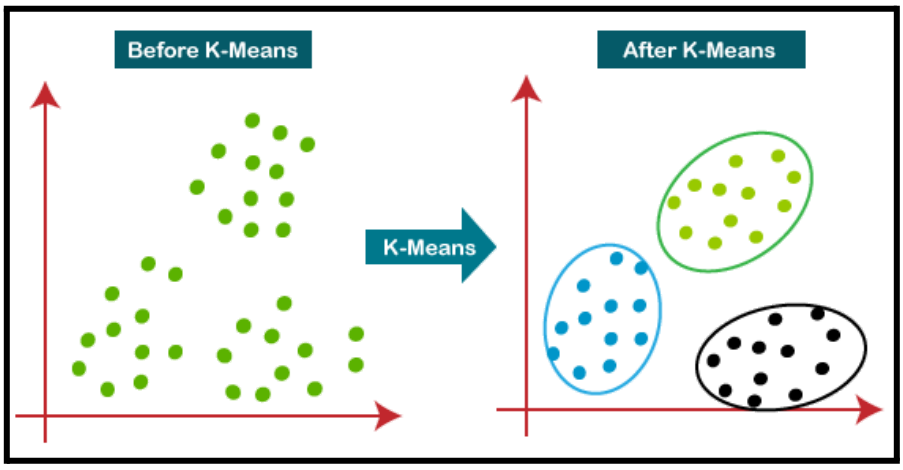
1. K Means
2. Random Forest
3. Random forest with K-fold cross validation

Reason to select machine learning model:

1. Testing dataset was not labeled, hence we used **K-Means** to cluster the testing dataset and add labels to it.
2. We used the **random forest classifier** because our data was **imbalanced**. It also contains such **outliers** that cannot be removed, **Random forest handles outliers by essentially binning them**. It is also indifferent to non-linear features. **It has methods for balancing error in class population unbalanced data sets.**

Algorithm:

1. **K-means:**
   a. Determines the best value for K center points or centroids by an iterative process.
   b. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

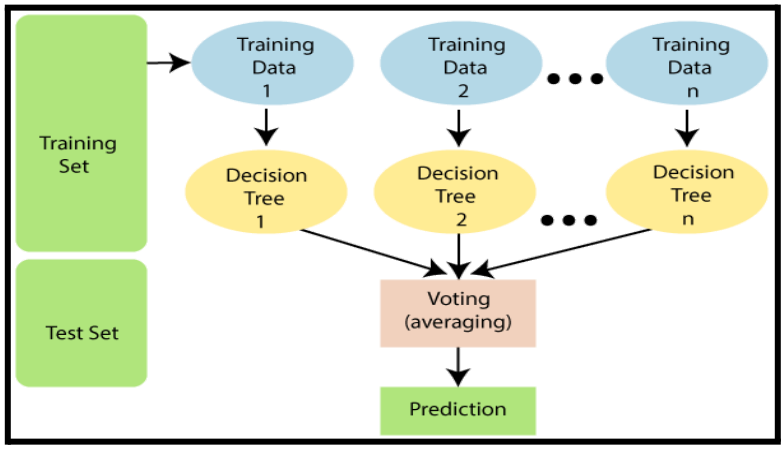


Graphical illustration of K Means clustering

2. Random Forest:

The working process can be explained in the below steps and diagram:

1. Select random K data points from the training set.
2. Build the decision trees associated with the selected data points (Subsets).
3. Choose the number N for decision trees that you want to build.
4. Repeat Step 1 & 2.



Random forest algorithm

3. Result Analysis:

MODEL	RANDOM FOREST	RANDOM FOREST WITH K-FOLD CROSS VALIDATION
ACCURACY	100%	90.2%

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

This model can be performed by using other classification methods e.g. SVM, NN etc.

This model can be used to classify other labels in other domains too.