Name : Enayat Kareem
Internship Program : Data science with Machine Learning and python Batch- June22 – Aug 22
Certificate Code : TCRIB3R149
Date of submission : 10/08/2022

Technical Coding Research Innovation, Navi Mumbai
Maharashtra, India-410206

Portuguese Banking Institution Campaign Prediction

A Case-Study Submitted for the requirement of
Technical Coding Research Innovation

For the Internship Project work done during
DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM

by
Enayat Kareem
TCRIB3R149

Rutuja Doiphode
CEO & CO-FOUNDER

Table of Contents

Portuguese Banking Institution Campaign Prediction

Enayat Kareem[#1]

[1]enayatkareem99@gmail.com

***Abstract*— Exploratory data analysis is crucial for understanding the underlying concepts present in the data and is the first step of every data-related project. In this report/paper, I used a dataset of a Portuguese Banking Institution Campaign, analyzed its features and explored its content through exploratory data analysis. Since the dataset contains 21 features and some highly correlated attributes, I implemented 11 Classification Machine Learning Algorithms and evaluated their accuracy. Via descriptive statistics and Graphical Plotting I showed the correlation between the attributes.**

***Keywords*— CSV, Data Cleaning, EDA, Classification, ROC**

I. INTRODUCTION

Exploratory data analysis (EDA) in statistics helps data analysts to understand the main characteristics of the data mostly through visual methods, which may further lead to the formulation of hypotheses, and the conduction of new experiments. EDA is one of the most important parts of data science. It consists of preprocessing, data visualization and insight extraction. In other words, EDA is the process of summarizing important characteristics of data in order to get a better understanding of the data. In most cases, the more data is preprocessed and visualized the more insights you can get and the better models you can train.

II. PROBLEM STATEMENT

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 to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

III. DATASET

Leveraging customer information is paramount for most businesses. In the case of a bank, attributes of customers like the ones mentioned below can be crucial in strategizing a marketing campaign when launching a new product.

The Dataset contains 21 attributes, can be categorized and sub-categorized into the following :

A. Bank - Client Data (8) :

The dataset contains attributes or features that can be classified into this bank-client data, and there are eight of them.

- age : (numeric)
  It is the age of a person associated in the bank-campaign.
- job : (categorical)
  type of job 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services','student', 'technician', 'unemployed', 'unknown')
- marital : (categorical)
  marital status 'divorced', 'married', 'single', 'unknown'. (note: 'divorced' means divorced or widowed)
- education : (categorical)
  'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate','professional.course', 'university.degree', 'unknown')
- default : (categorical)
  has credit in default?('no','yes','unknown')
- Balance : (categorical)
  Amount of money one has.
- housing : (categorial) has a housing loan? ('no', 'yes', 'Unknown')
- loan : (categorial) has a personal loan? ('no','yes','Unknown')

B. Features Related with the last contact of the current campaign (4) :

These attributes provide the knowledge of last contact of the current campaign.

- contact : (categorical)
  contact communication type ('cellular', 'telephone')

- month : (categorical)
  last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec')
- day : (categorical)
  last contact day of the week ('mon','tue','wed','thu','fri')
- duration : (numeric)
  last contact duration, in second. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not no before a call is performed. Also, after the end of the call y is obviously no. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

C. Other Attributes (4):

- campaign : (numeric)
  number of contacts performed during this campaign and for this client (includes last contact).
- pdays : (numeric)
  number of days dat passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
- previous : (numeric)
  number of contacts performed before this campaign and for this client.
- poutcome : (categorical)
  outcome of the previous marketing campaign ('failure','nonexistent','success').

D. Output variable (desired target):

- target - has the client subscribed a term deposit? (binary: 'yes','no')

IV. EDA

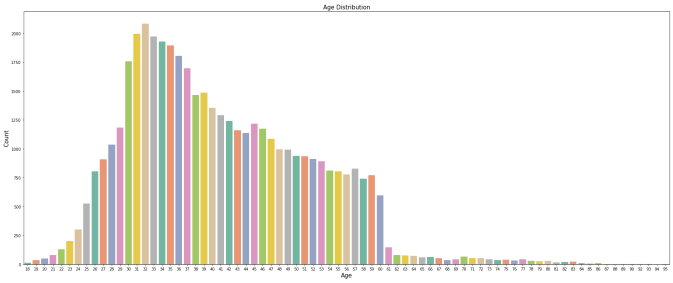
A. Read the dataset as a dataframe

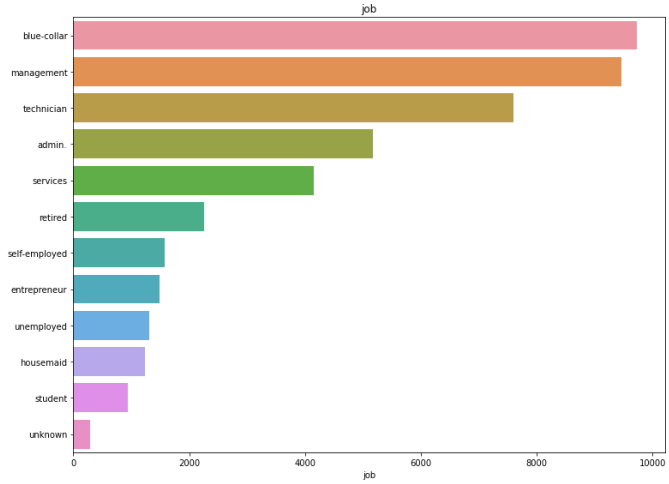
```
dset=pd.read_csv('/tcr_intern/bank-full.csv')
```

B. Dataset type, shape, info(), describe()

We have 17 columns, 7 of them are numerical data, the rest 10 are categorical in nature.

We start the EDA of the categorical data, EDA such as distribution graphs of our attributes, crosstabs and value counts.

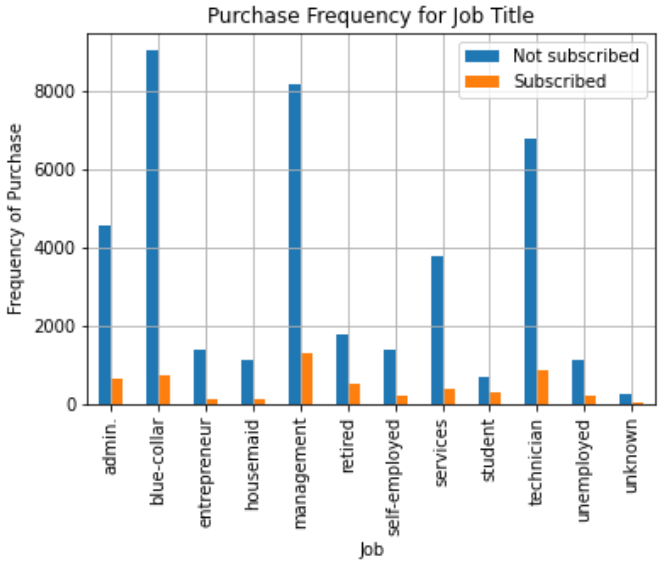




C. Input Categorical feature observation.

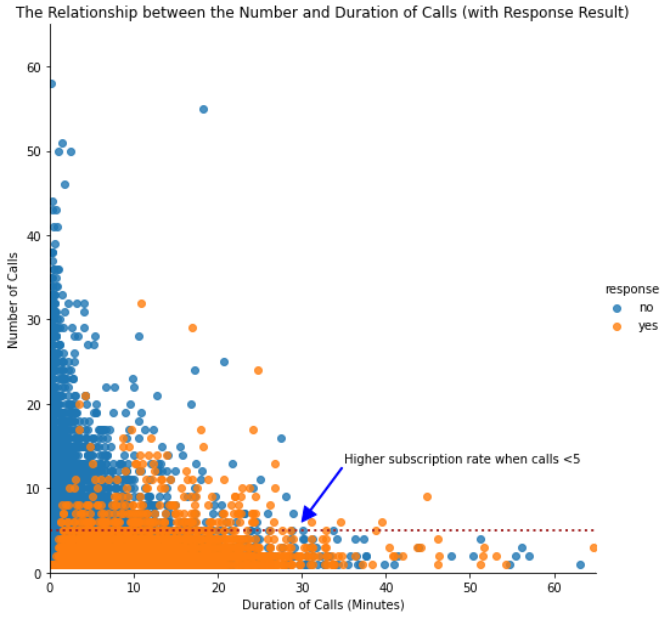
- Job - More Job types are Admin, Technician and blue-collar and it means bank targeting high salaried people.
- Marital - more people of type married - (#To Do - Check the target value distribution for high salaried married people)
- Education - more people count in university, degree people . And the illiteracy count is very low.
- default - most people have no credit default, which means they can be approached.
- housing - we must give more importance to people who have not taken any housing loan.
- loan - we must give more importance to people who have not taken any personal loan.

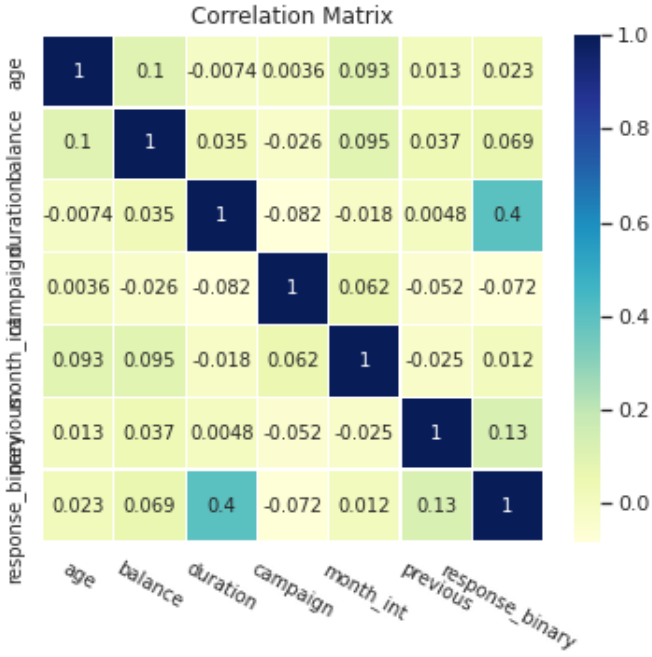
- month - Seems May is a busy season in Portugal.
- Day_of_week - Seems every day is busy but not on weekends.
- p_outcome -outcome of the previous marketing campaign- Success is small rate.



We can see that Students are more likely to subscribe to the term deposit.

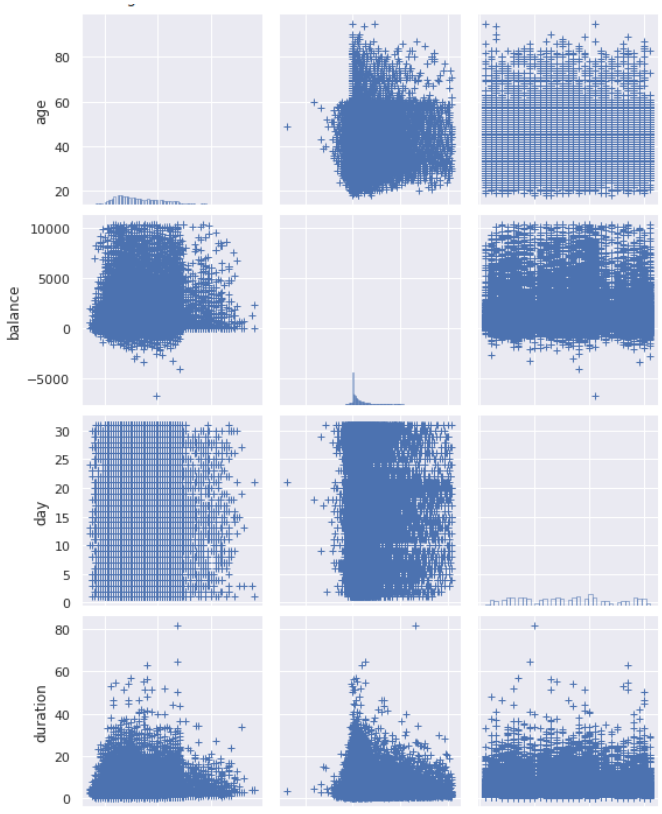
D. After cleaning the dataset a little i checked the relationship between the number of calls and duration of calls.

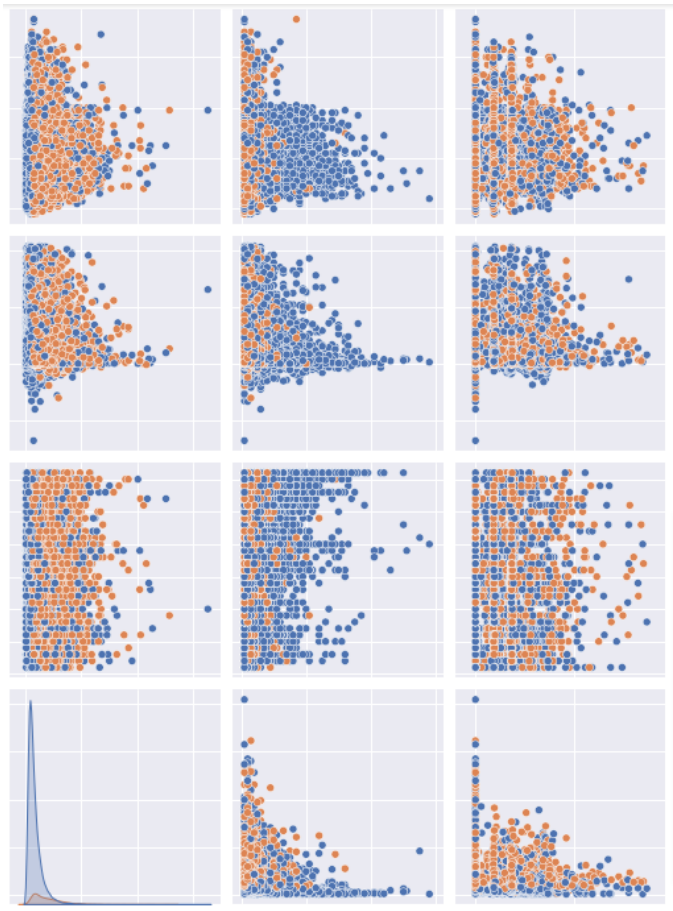




Correlation by heat mapping the relationship between the features. Dark color represents a positive correlation, Light color/ white is towards the negative correlation.

E. Pair Plotting





Observation : From the above pair plot we can infer the association among the attributes and target column as follows:

- 'Age' column is slightly right skewed. Most of the client's age is between 25 to 65 years.
- 'Job' Here the number of clients who have 'blue-collar','management' and 'technician' are more in our dataset and it is skewed towards the right.
- 'marital' from above we can see that married clients are more likely to subscribe for the term deposit.
- 'Education' clients with higher education are more and also there are many clients whose education level is unknown.
- 'Balance' there is a long tail towards the right side in the balance column which would be the outliers.
- 'Housing' clients who don't have a housing loan are more likely to subscribe for the term deposit.
- 'Duration' the skewed in the duration indicates that most of the calls are relatively short aslo large number of outliers presence in the dataset.
- 'Campaign' most of the clients have been contacted by the bank between 1 to 5 times. There are some clients who have been contacted by the bank more than 20 times and also we can see that their probability towards the term deposit is very low or almost 0. Higher subscriber rate when calls < 5.

F. Feature Selection

Prepare the data to train a model – check if data types are appropriate, get rid of the missing values etc.

- I have already converted the data types to appropriate data types in EDA.
- Also I did not find any missing value in the EDA of each column.
- Now we are good to go with model building after generating the dummy variables.

Generating Dummy Variables :-
There are three columns 'job', 'marital', 'poutcome' for which I will be converting to a dummy variable.

Splitting our dataset into a training set and test set.

- 20% train test split for model fitting

```
X_train,X_test,Y_train,Y_test= train_test_split(X,
                                Y, test_size=0.2,random_state=42)
print(X_train.shape,X_test.shape)
```

```
(36168, 21) (9043, 21)
```

V. TRAINING & PREDICTION

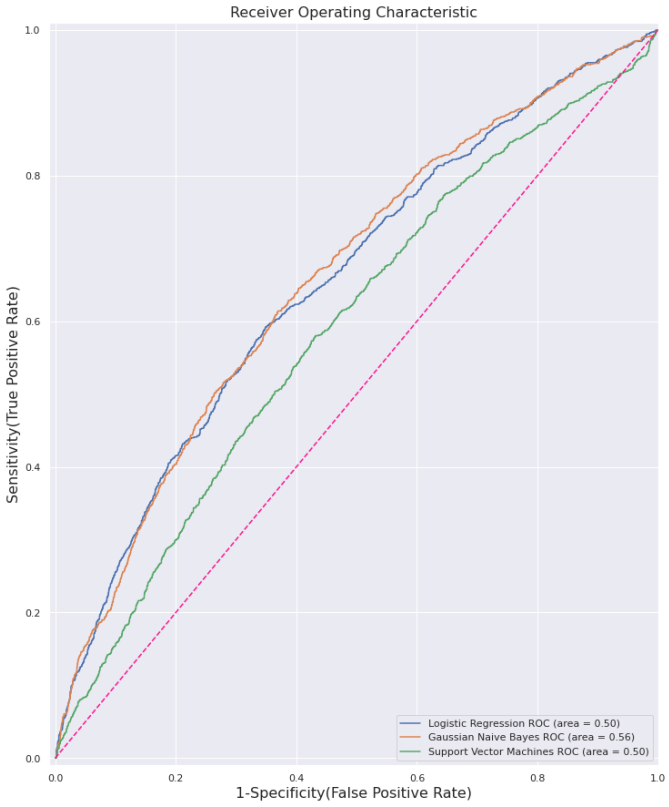
I will be testing the following models with my training data :

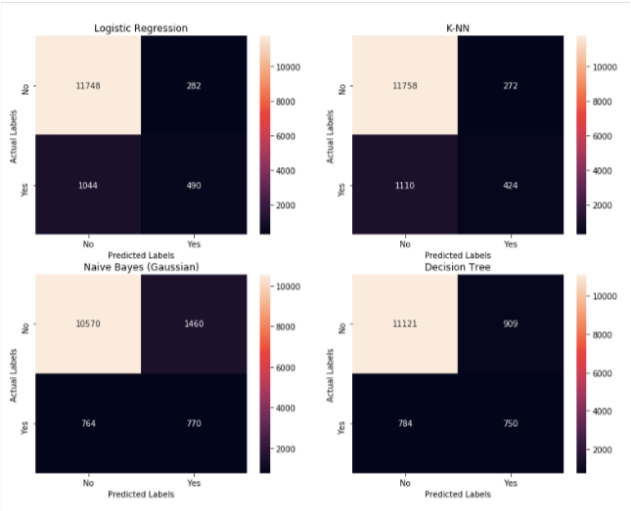
- Gaussian Naive Bayes
- Logistic Regression
- Support Vector Machines
- Perceptron
- Decision Tree Classifier
- Random Forest Classifier
- KNN or k-Nearest Neighbors
- Stochastic Gradient Descent
- Gradient Boosting Classifier

For each model, we set the model, fit it with 80% of our training data, predict for 20% of the training data and check the accuracy.

	Model	Score
2	Support Vector Machines	87.94
3	Bagging (SVM)	87.94
4	Linear Support Vector Machines	87.94
8	Bagging(Random Forest)	87.92
0	Logistic Regression	87.90
11	Gradient Boosting Classifier	87.90
10	Stochastic Gradient Descent	87.74
9	KNN	87.23
7	Random Forest	86.80
6	Decision Tree	82.82
1	Gaussian Naive Bayes	82.01
5	Perceptron	81.39

	Model	Accuracy	Accuracy %	Precision	Recall	F1 Score	ROC
0	Logistic Regression	0.879022	87.90	0.000000	0.000000	0.000000	0.499811
0	Gaussian Naive Bayes	0.820082	82.01	0.236739	0.220898	0.228544	0.561593
0	Support Vector Machines	0.879354	87.94	0.000000	0.000000	0.935805	0.500000
0	Bagging (SVM)	0.879354	87.94	0.000000	0.000000	0.000000	0.500000
0	Linear Support Vector Machines	0.369789	87.94	0.140000	0.821265	0.239220	0.564556
0	Perceptron	0.813889	81.39	0.067251	0.042163	0.051831	0.480966
0	Decision Tree	0.828154	82.82	0.300946	0.320807	0.310559	0.609284
0	Random Forest	0.867964	86.80	0.422556	0.257562	0.320046	0.604636
0	Bagging(Random Forest)	0.879244	87.92	0.000000	0.000000	0.000000	0.499937
0	KNN	0.872277	87.23	0.396774	0.112741	0.175589	0.544612
0	Stochastic Gradient Descent	0.877364	87.74	0.050000	0.000917	0.001800	0.499264





VI. CONCLUSION

- Support Vector Machines and Logistic Regression gives us best accuracy of 87.9% approx.
- The area in ROC curve for Support Vector Machine is around 0.50 and the ROC curve is highest for random forest.
- Hence among the above 11 algorithms applied on the underlying dataset, Support Vector Machine and Logistic Regression would be the best choice to predict the clients who will subscribe for the term deposit.

VII. REFERENCES

- Samvelyan, A., Shaptala, R., & Kyselov, G. (2020). *Exploratory data analysis of Kyiv city petitions. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC).* doi:10.1109/saic51296.2020.9239185
- https://www.kaggle.com/code/pritech/predicting-the-term-deposit-subscription
- https://research.com/research/how-to-write-a-research-paper-abstract