Bank Marketing (Campaign)

Assignment

Final Report

Prepared by:

Aly Medhat Moslhi

Submitted to:

Data Glacier Internship Program

Patch No.:

LISUM11

Sep. 24th, 2022

Table of Content

List of Figures

Team Members

Track: Data Science
Batch No.: *LISUM11*

Name: Aly Medhat Moslhi
Country: Egypt
Email: alymedhat10@yahoo.com
Company: AAST

Background

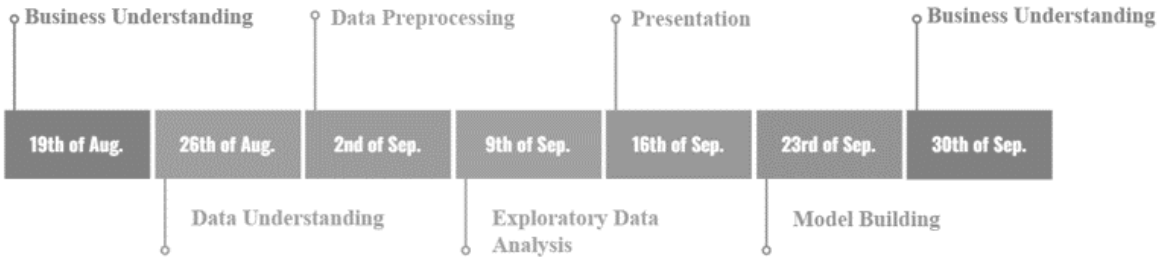
A. Problem Description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them in understanding whether a particular customer will buy their product or not (based on the customer's past interaction with the bank or other Financial institutions).
The bank wants to use the ML model to shortlist customers whose chance of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing, etc) can focus only on those customers whose chance of buying the product is more.

B. 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 with the same client was required, 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 (yes/no) to a term deposit (variable y)

C. Project **lifecycle**



Data Intake Report

Name: Bank Marketing (Campaign)
Report date:
Internship Batch: LISUM11: 30 June - 30 Sept 2022
Version: 1.0
Data intake by: Aly Medhat Moslhi
Data intake reviewer:
Data storage location:

Bank-additional-full data

Total number of observations	41188
Total number of files	1
Total number of features	20
Base format of the file	.csv
Size of the data	5.8 MB

Bank-additional data

Total number of observations	4119
Total number of files	1
Total number of features	20
Base format of the file	.csv
Size of the data	586 KB

Bank-additional-name data

Total number of files	1
Base format of the file	.txt
Size of the data	8 KB

Procedure

For solving the given problem by choosing the appropriate machine learning model, the following procedure was proposed. It includes data exploration, data cleaning, model deployment, and model evaluation. The five models that were used are:

1. Naïve Bayes.
2. Linear regression.
3. Ensemble model using max voting.
4. Ensemble models using Bagging.
5. Stacking classifier.

The accuracy of each model was calculated in more than a method to fully show the model capabilities and thus precisely choose the suitable model; the methods used are:

1. Accuracy.
2. Confusion matrix.
3. F1 Score.

In the following four sections, each of the steps in the procedure followed will be discussed.

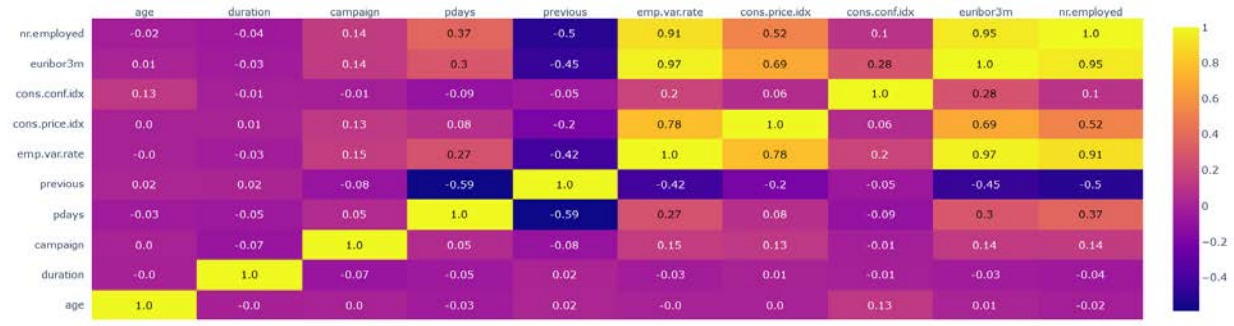
1- Analyzing and Cleaning the Data

A. Statistical analysis of the data results

1. No NaN values were found in the dataset
2. No duplicates were found
3. Statistical analysis was performed on the dataset (mean, median, standard deviation)
4. Outliers were found
5. Some columns will need to be encoded.

B. Multivariate analysis of the data

The heatmap shown in figure 1 shows the correlation between the numeric variables. It shows that there is a very high correlation between employment variation rate, the number of employees, Euro Interbank Offered 3-month rate, and the consumer confidence index.



C. Categorical vs Categorical analysis

It is difficult to analyze most of the features manually since most of the features are categorical. Therefore, it is better to compare them using feature engineering after data cleaning.

D. Numeric vs Categorical analysis:

There are 9 numeric columns and 11 categorical columns. Showing the correlation between them gives an insight into the data. For instance, comparing the job and the age of a client and how it can affect the decision.

1. To deal with outliers two methods were proposed

a. The IRQ method was proposed, and the results can be seen in fig. 1

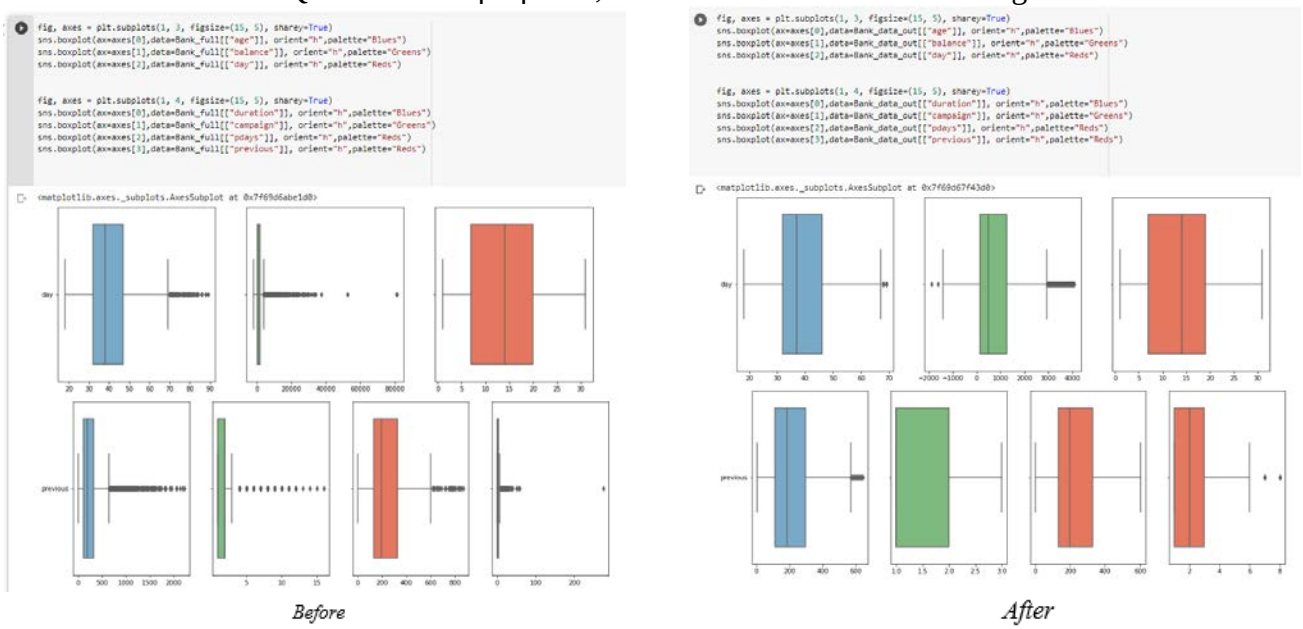


Figure 1- Boxplot of the categorical columns before and after removing outliers

b. The Slandered deviation method shown in fig 2

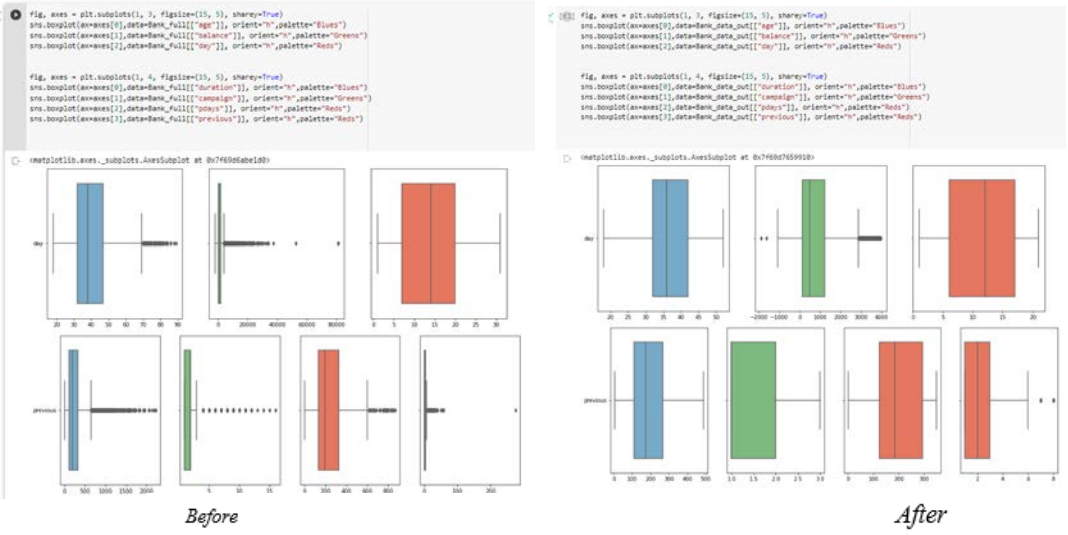


Figure 2- Boxplot of the categorical columns before and after removing outlier

2 - Exploratory data analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. The EDA was used to first understand the data and give recommendations for the classification. The results of the step were as follows:

A. The relation between the bank balance and the client's age

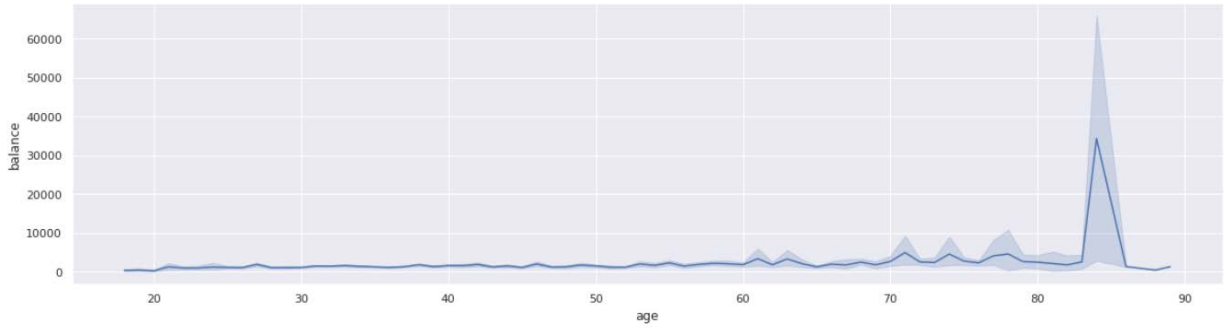


Figure 3 - The relation between the bank balance and the client's age

The graph shows the relation between the account balance and the age of the client. This may not help in the final prediction but gives an insight into the data.

B. The relation between the bank balance and the client's age

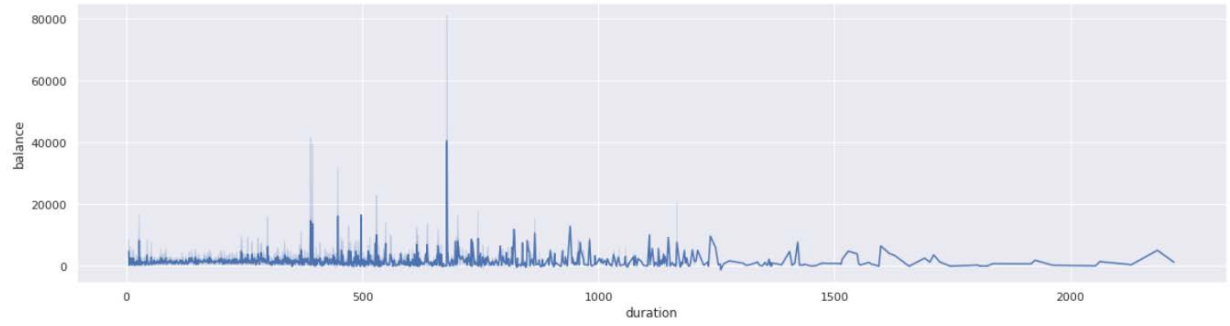


Figure 4 - The relation between the bank balance and the client's age

The graph shows the relation between the account balance and the duration. This may not help in the final prediction but gives an insight into the data.

C. The relation between the total number of applications and the client's age

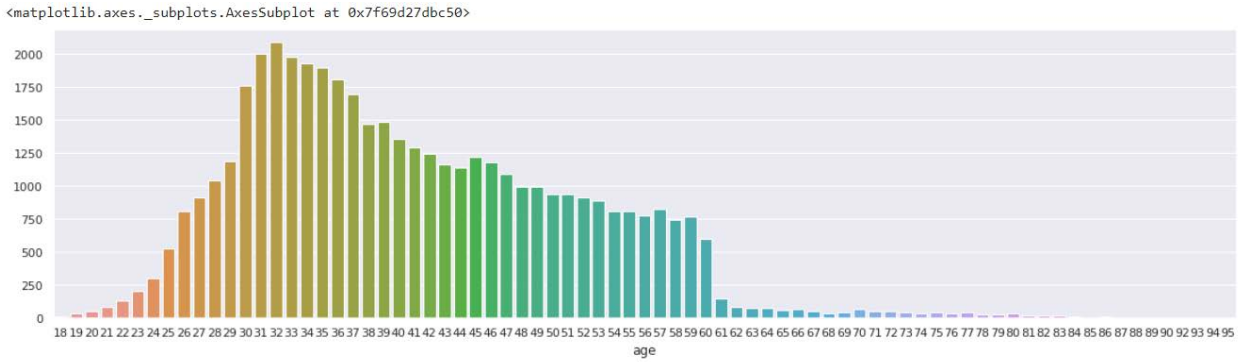


Figure 5 - The relation between the total number of applications and the client's age

The graph shows the relation between the total number of applications the bank received and the age of the applicant. It can be seen that most applicants' age varies from 25-60 years old.

D. The relation between the accepted applications and the client's age

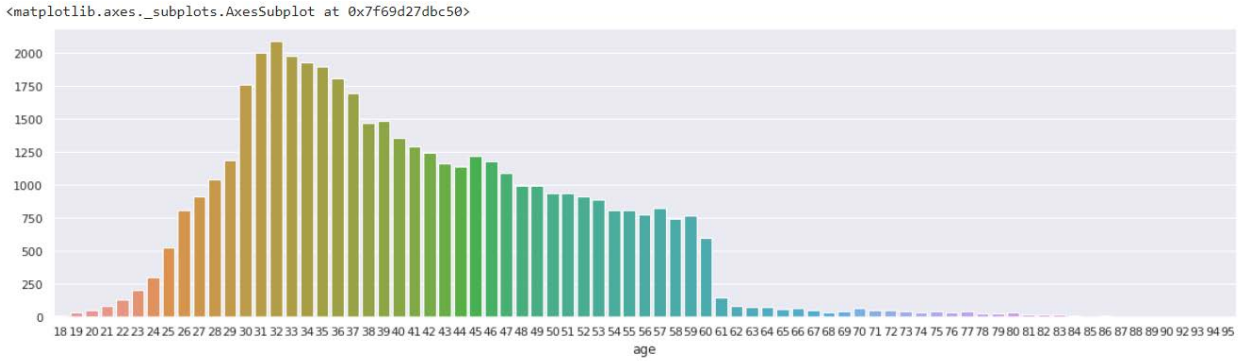


Figure 6 - The relation between the accepted applications and the client's age

The graph shows the relation between the number of accepted applications and the age of the applicant. The graph is similar to the previous graph meaning that there is no dependency on accepting the application and the applicant's age.

E. The relation between the accepted applications and if they had a load

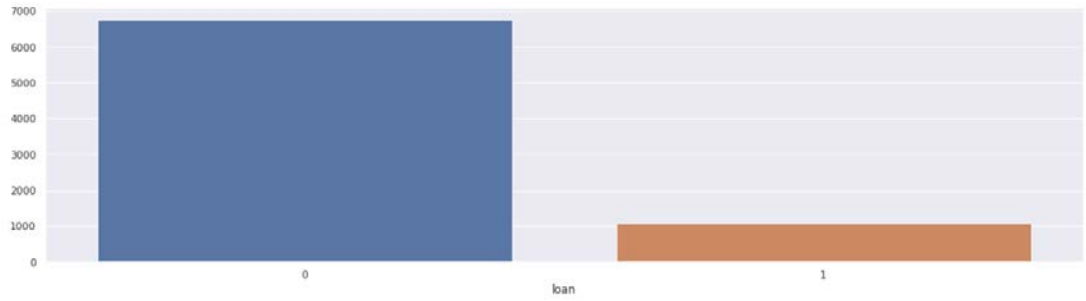


Figure 7 - The relation between the accepted applications and if they had a load

It can be seen that most of the accepted applications didn't have a previous loan. Meaning that it can be taken into consideration in decision making

F. The relation between the accepted applications and if they have a house

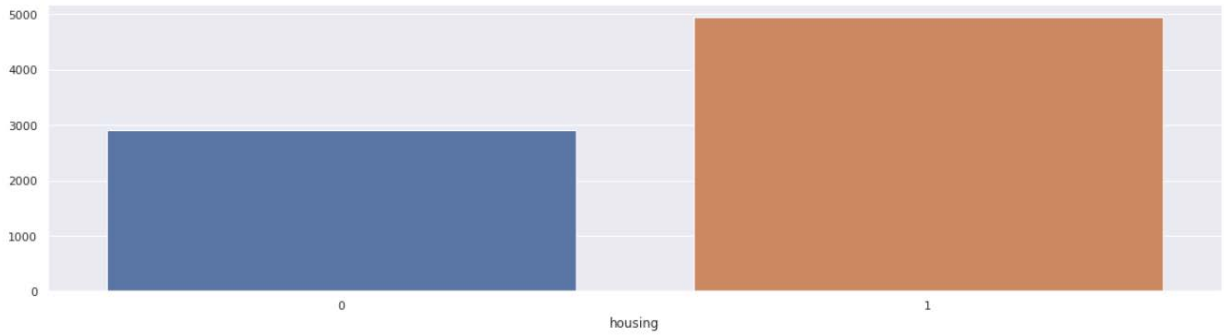


Figure 8 - The relation between the accepted applications and if they have a house

It can be seen that most of the accepted applications have the house. The relation is not that high still it should be taken into consideration in decision making

G. The relation between the accepted applications and the applicant's job

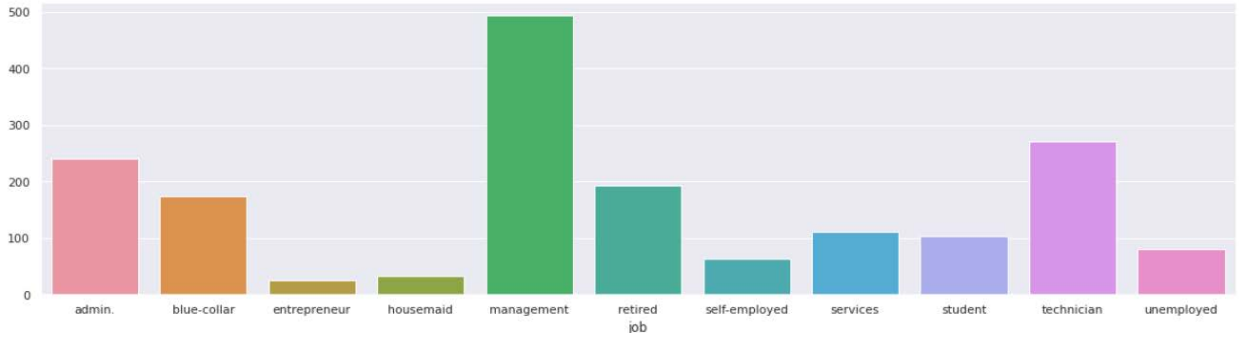


Figure 9 - The relation between the accepted applications and the applicant's job

It can be seen some jobs have a higher probability of getting accepted. This means it should be taken into consideration.

H. The relation between the accepted applications and the applicant's education

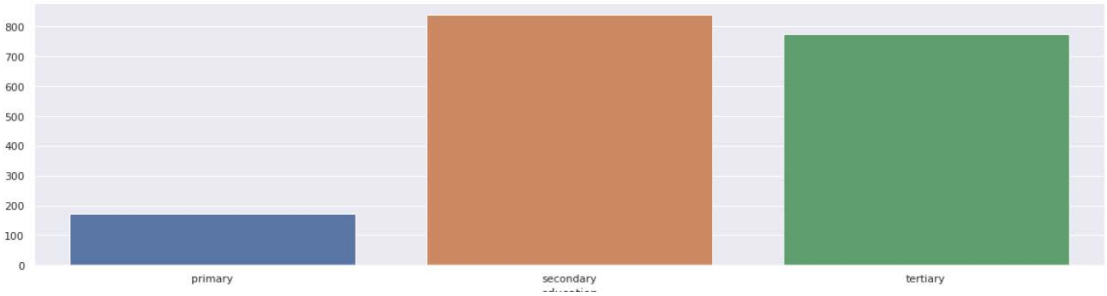


Figure 10 - The relation between the accepted applications and the applicant's education

It can be seen education influences the probability of getting accepted. This means it should be taken into consideration.

I. The relation between the accepted applications and the application day

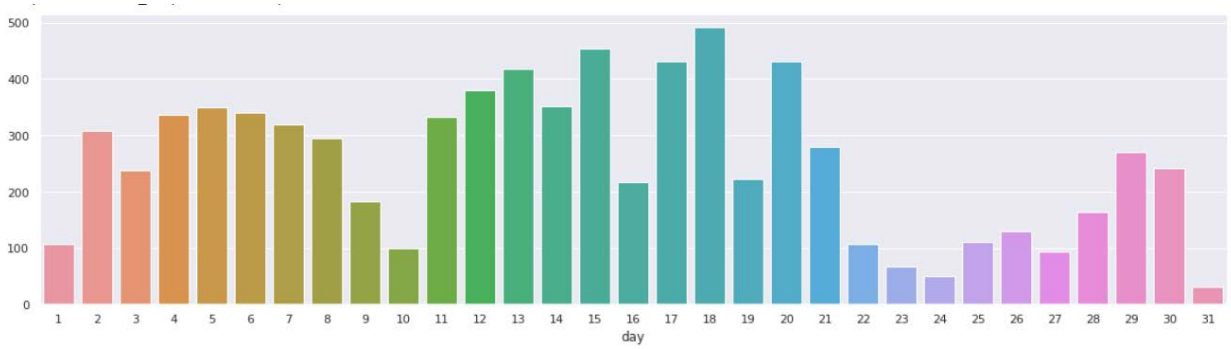
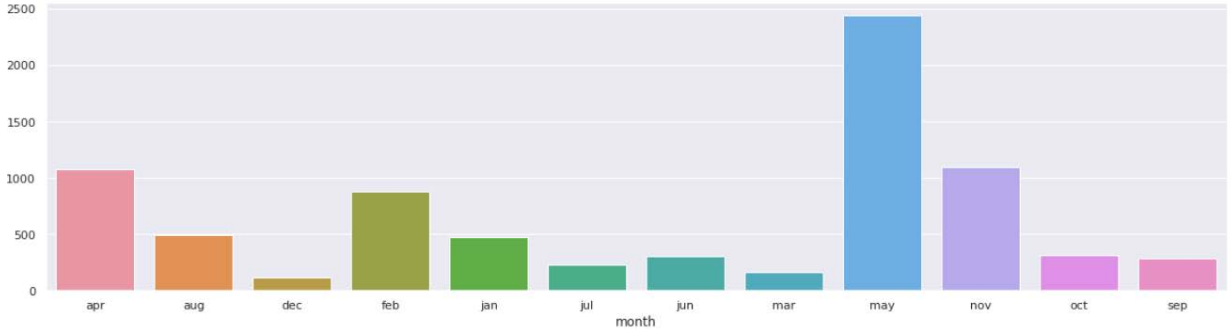


Figure 11 - The relation between the accepted applications and the application day

It can be seen that the day of application does not influence the probability of getting accepted. This means it should not be taken into consideration.

J. The relation between the accepted applications and the application month



It can be seen that the month of application does not influence the probability of getting accepted. This means it should not be taken into consideration.

Final recommendations

1. Outliers are not to be removed.
2. There is a direct relationship between the loan and the acceptance of the application (An important feature for classification).
3. Some jobs have higher priority for acceptance (an important feature for classification).
4. Some features do not affect the output as day and month (Not to consider).

3 – Models training

This section will focus on the machine learning models used in the procedure followed. There were five models tested, and each of them will be explained and the results of each model will be briefly discussed.

1. Naïve Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. The main advantages of using this algorithm:

- It is very fast at classifying unknown records.
- It is simple and easy to implement.
- It doesn't require as much training data.
- It handles both continuous and discrete data.
- It is highly scalable with several predictors and data points.

2. Linear regression:

Linear regression is perhaps one of the most well-known and well-understood algorithms in statistics and machine learning. Its advantages are:

- It is easier to implement, and interpret, and very efficient to train.
- It can interpret model coefficients as indicators of feature importance.
- It can interpret model coefficients as indicators of feature importance.

3. Ensemble model using max voting.

It is mainly used for classification problems. The method consists of building multiple models independently and getting their output called 'vote'. The class with maximum votes is returned as output. Here three classification models (logistic regression, xgboost, and random forest) are combined. The advantages of using this method are:

- Reduce the spread in the average skill of a predictive model.
- Improve the average prediction performance over any contributing member in the ensemble.
- Has better performance than the singles model.

4. Ensemble model using Bagging.

It is also known as a bootstrapping method. Base models are run on bags to get a fair distribution of the whole dataset. A bag is a subset of the dataset along with a replacement to make the size

of the bag the same as the whole dataset. The final output is formed after combining the output of all base models.

- Allows many weak learners to combine efforts to outdo a single strong learner.
- Higher predictive accuracy, compared to the individual models.

5. Stacking classifier (RF, KNN - DT, RF):

It is an ensemble method that combines multiple models (classification or regression) via meta-model (meta-classifier or meta-regression). The base models are trained on the complete dataset, then the meta-model is trained on features returned (as output) from the base models. The base models in stacking are typically different. The meta-model helps to find the features from base models to achieve the best accuracy. Here the classifiers that will be used are (RF, KNN - DT, and RF). The advantages of using this technique are:

- Stacking can yield improvements in model performance.
- Stacking reduces variance and creates a more robust model by combining the predictions of multiple models.

Models Evaluation

This section will focus on the results of each of the trained models. Different evaluation matrices were used in evaluating the models, the accuracy, the confusion matrix, and the F1 score.

1. Naïve Bayes

Accuracy	Confusion Matrix	F1 score
0.893853	7539 139 778 183	0.28526

2. Linear regression

Accuracy	Confusion Matrix	F1 score
0.87486	7190 488 593 368	0.405063

3. Ensemble model using max voting

Accuracy	Confusion Matrix	F1 score
0.9007	7490 188 669 292	0.405274

4. Ensemble models using Bagging.

Accuracy	Confusion Matrix	F1 score
0.9018	7494 209 639 297	0.4119

5. Stacking classifier.

Accuracy	Confusion Matrix	F1 score
0.8996	7446 232 692 269	0.404123

Discussion

The models had relevantly close results. The ensemble algorithms as it was expected outperformed the other models. The highest results were obtained from the ensemble model using bagging (the fourth model) then comes after it the ensemble models using max voting (third model) with very close results. Then in order comes the stacking classifier, followed by the Naïve Bayes followed by the linear regression model.

GitHub Repository link

https://github.com/alymedhat10/Bank-Marketing-Campaign-.git