Introduction

Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.

The goal of this notebook is to use the data to develop a strong model in order to predict which people the bank should market to for their marketing campaign to get people to sign up for a saving account.

What is a saving account?

A savings account is an interest-bearing deposit account held at a bank. Though these accounts typically pay only a modest interest rate, their safety and reliability make them a good option for depositing cash that you want available for short-term needs. For more detailed information what is the saving account click on this link: https://www.investopedia.com/terms/s/savingsaccount.asp
In general, datasets which contain marketing data can be used for 2 different business goals:

Prediction of the results of the marketing campaign for each customer and clarification of factors which affect the campaign results. This helps to find out the ways how to make marketing campaigns more efficient.

Finding out customer segments, using data for customers, who will subscribe to the saving deposit. This helps to identify the profile of a customer, who is more likely to acquire the product and develop more targeted marketing campaigns.

This dataset contains banking marketing campaign data and we can use it to optimize marketing campaigns to attract more customers to the deposit subscription.
Approach

In order to optimize the marketing campaigns with the help of the dataset, we will have to take the following steps:

Import data from dataset and perform initial high-level analysis: look at the number of rows, look at the missing values, look at dataset columns and their values respective to the campaign outcome.

Clean the data: remove irrelevant columns, deal with missing and incorrect values, and turn categorical columns into variables.

Use machine learning techniques to predict the marketing campaign outcome and to find out factors, which affect the success of the campaign.

Bank client data:

1. ***age:*** (numeric)

2. ***job -***type of job (categorical: 'admin.', 'bluecollar', 'entrepreneur', ‘ousemaid', 'management', 'retired', 'self-employed','services','student','technician','unemployed','unknown')

3. ***marital*** - marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4. ***education*** - (categorical: primary, secondary, tertiary and unknown)

5. ***default*** - has credit in default? (categorical: 'no','yes','unknown')

6. ***housing*** - has housing loan? (categorical: 'no','yes','unknown')

7. ***loan*** - has personal loan? (categorical: 'no','yes','unknown')

8. ***balance*** - Balance of the individual.

Related with the last contact of the current campaign:

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

10. ***month*** - last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

11. ***day*** - last contact day of the month

12. ***duration*** - last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.

Other attributes:

13. **campaign** - number of contacts performed during this campaign and for this client (numeric, includes last contact)

14. ***pdays*** - number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)

15. ***previous -*** number of contacts performed before this campaign and for this client (numeric)

16. ***poutcome*** - outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', success')

Output variable (desired target):

17. ***deposit*** - has the client subscribed a saving account? (binary: 'yes','no')

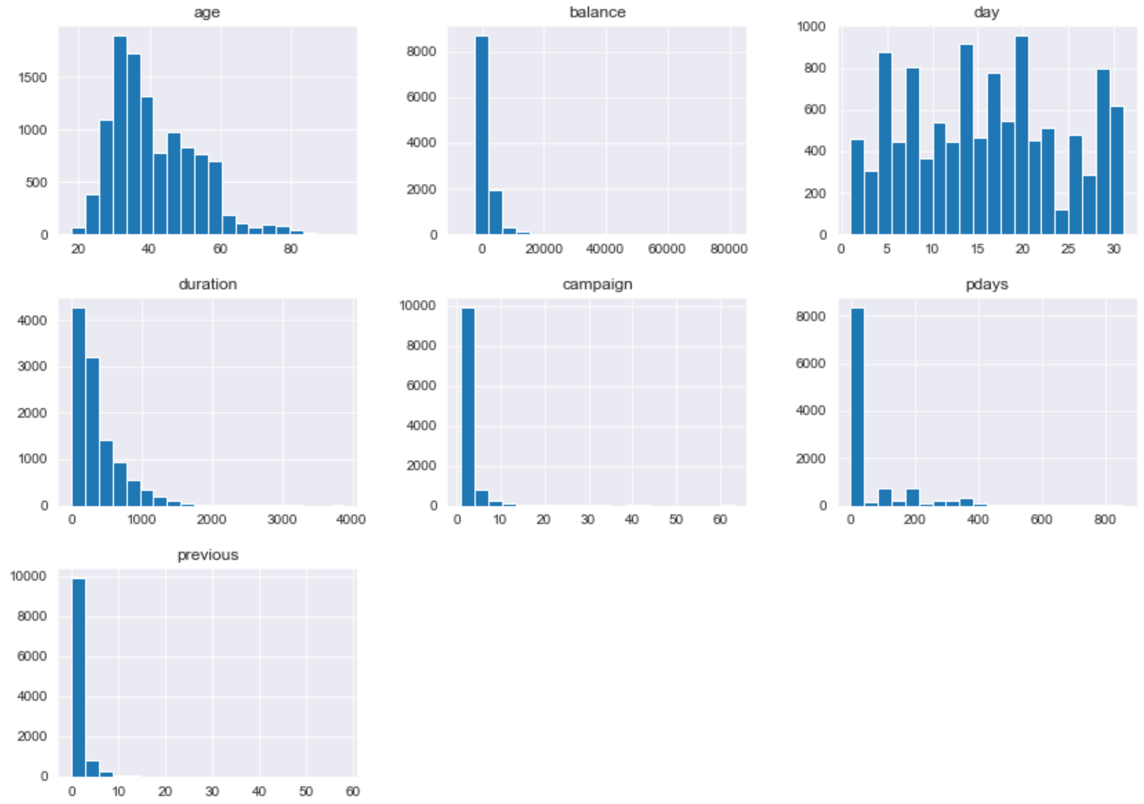
There are columns 11162 and 17 rows in a dataset and there was no missing data.
We found some interesting insights we can find from the data below:

1. 13.1% or 1462 of people have 0 or negative balance
2. 74.57% or 8324 has not been contacted for the first time

There are 7 numeric columns and 10 categorical columns.

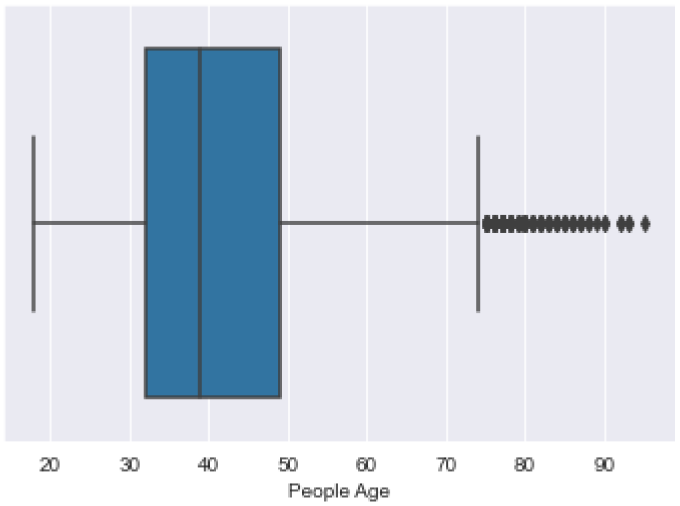
1. EDA

Some feature insights can be seen from the below histogram:



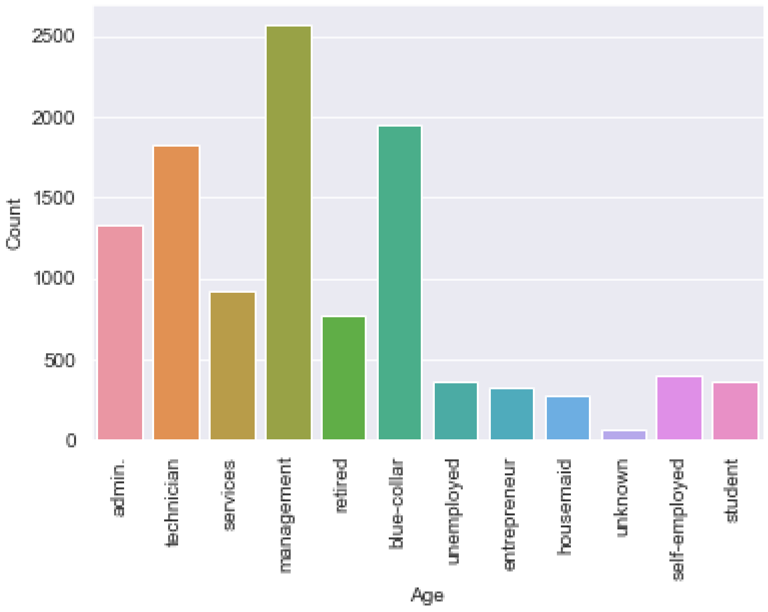
```
Then, we found that:
Ages above 74.5 are outliers
There are 171 outliers
```

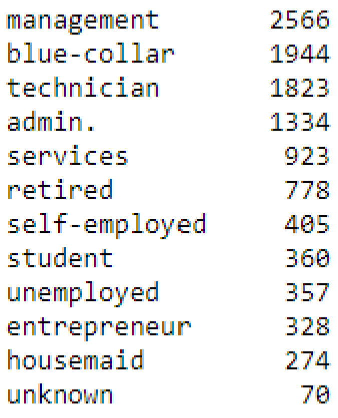
From the boxplot below you can see how the age data is distributed.



From the above age analyse we cannot conclude that the age of people has high effect whether they will deposit money or not. People of any age could open the account that's why we can fit our model with and without these outliers.

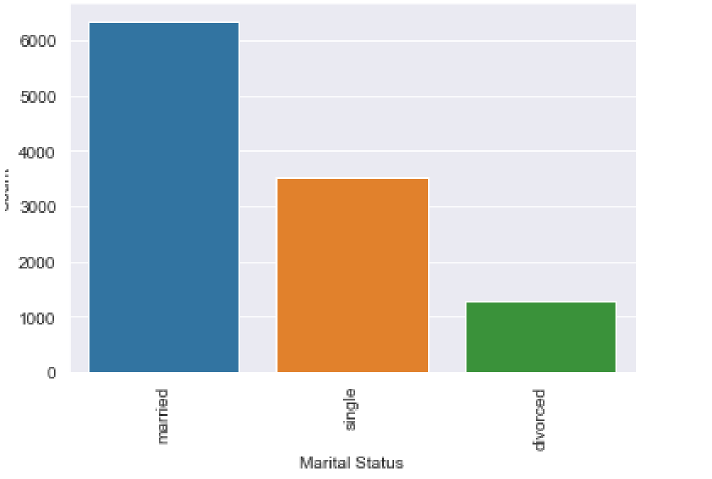
1.1 Jobs

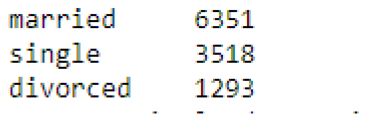




```
management       2566
blue-collar      1944
technician       1823
admin.           1334
services          923
retired           778
self-employed     405
student           360
unemployed        357
entrepreneur      328
housemaid         274
unknown            70
```

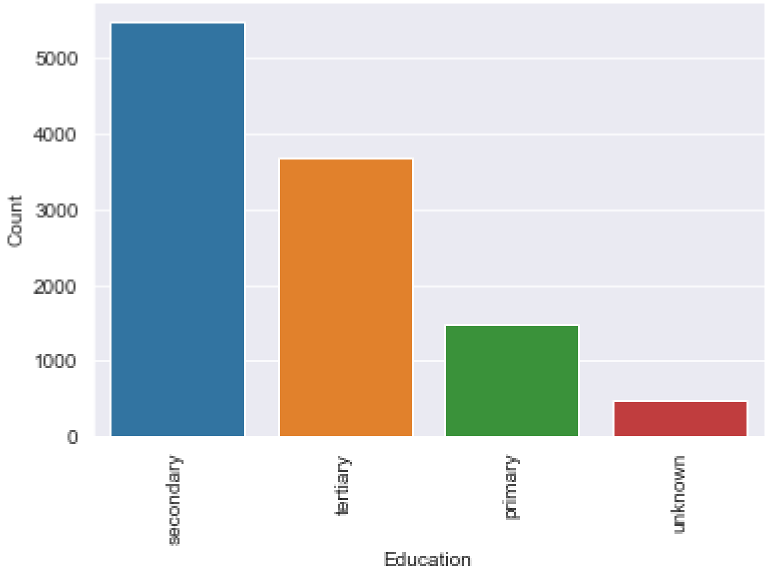
1.2 Marital Status

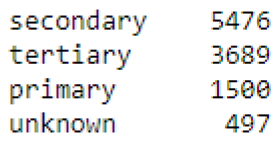




```
married     6351
single      3518
divorced    1293
```

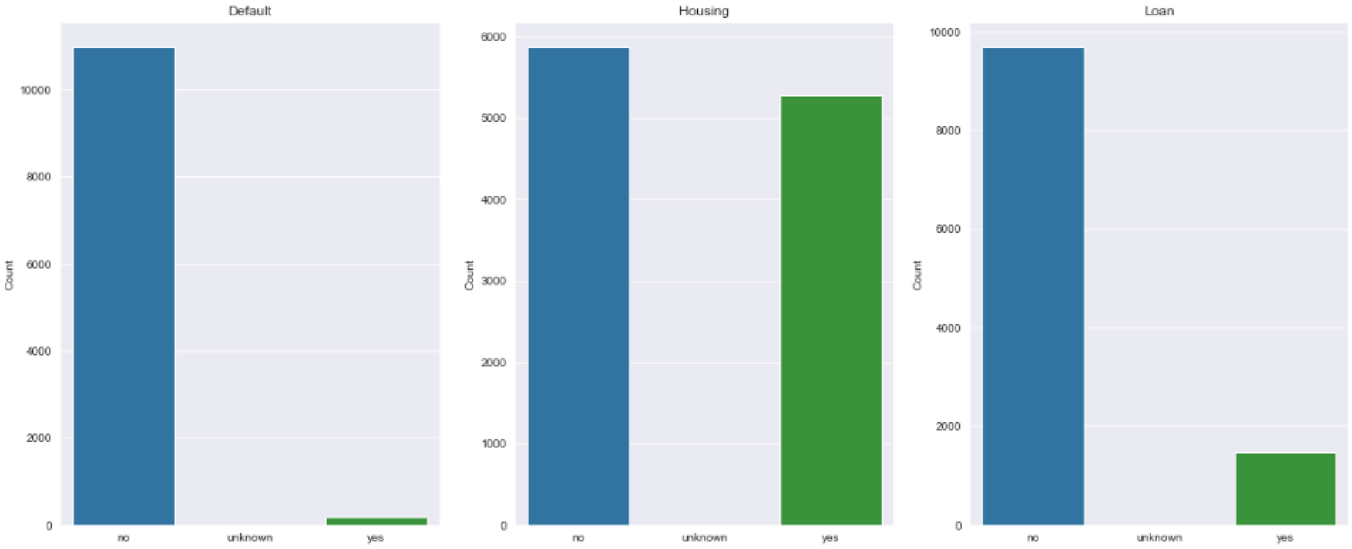
1.3 Education

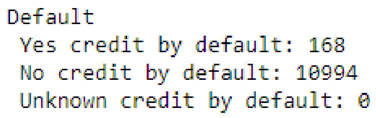




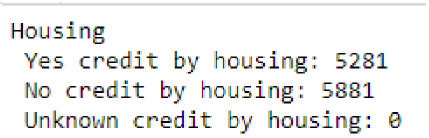
```
secondary    5476
tertiary     3689
primary      1500
unknown       497
```

Defining has a credit in default, and house and personal loans or not.

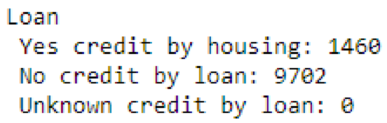




```
Default
 Yes credit by default: 168
 No credit by default: 10994
 Unknown credit by default: 0
```



```
Housing
 Yes credit by housing: 5281
 No credit by housing: 5881
 Unknown credit by housing: 0
```



```
Loan
 Yes credit by housing: 1460
 No credit by loan: 9702
 Unknown credit by loan: 0
```

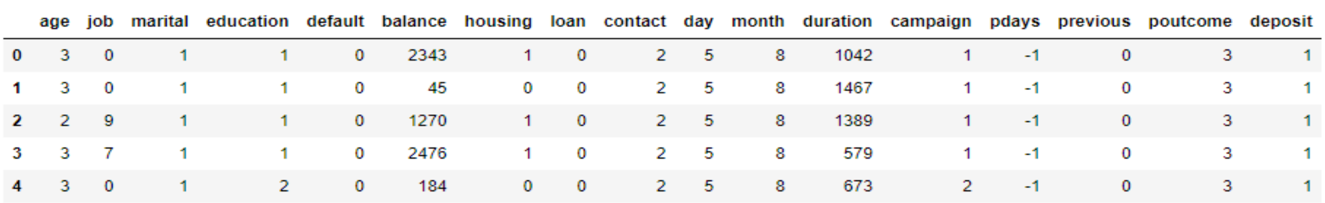
Dividing the ‘age’ column into four groups:

Ages below 32 = 1
Ages between 32 and 47 years old = 2
Ages between 47 and 70 years old = 3
Ages above 70 years old = 4

There are other visualizations and insights I found in this data. The detailed EDA of Panda’s Profile Report can be seen in the model.

2. ***Classification Mode***

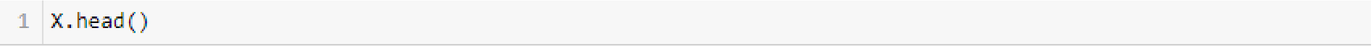
We will use Label encoder and convert categorical columns.



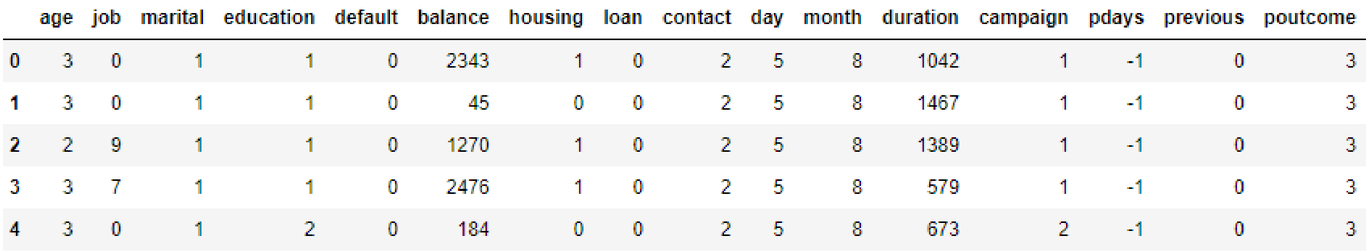
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	3	0	1	1	0	2343	1	0	2	5	8	1042	1	-1	0	3	1
1	3	0	1	1	0	45	0	0	2	5	8	1467	1	-1	0	3	1
2	2	9	1	1	0	1270	1	0	2	5	8	1389	1	-1	0	3	1
3	3	7	1	1	0	2476	1	0	2	5	8	579	1	-1	0	3	1
4	3	0	1	2	0	184	0	0	2	5	8	673	2	-1	0	3	1

Then, we will split the data into train and test sets. The train set has 80% of data and test is 20%. We do this for training our dataset in 'train' set and testing the results in 'test' set. Our target variable is y, which equal to 'deposit'.

Below, how the dataset will look like after the conversion:

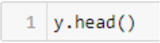


```
X.head()
```

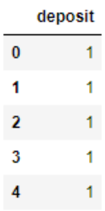


	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	3	0	1	1	0	2343	1	0	2	5	8	1042	1	-1	0	3
1	3	0	1	1	0	45	0	0	2	5	8	1467	1	-1	0	3
2	2	9	1	1	0	1270	1	0	2	5	8	1389	1	-1	0	3
3	3	7	1	1	0	2476	1	0	2	5	8	579	1	-1	0	3
4	3	0	1	2	0	184	0	0	2	5	8	673	2	-1	0	3

Our target is y = 'deposit' column



```
y.head()
```



	deposit
0	1
1	1
2	1
3	1
4	1

There we will scale our X (dependant variables) with 'Standard Scaler'.

Before we start modelling we need avoiding overfitting. Below, brief description of overfitting:

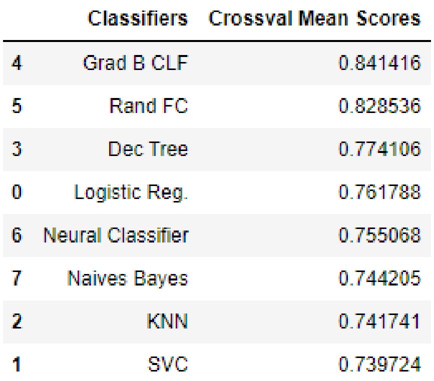
This is an error in the modelling algorithm that takes into consideration random noise in the fitting process rather than the pattern itself. You can see that this occurs when the model gets an awesome score in the training set, but when we use the test set (Unknown data for the model) we get an awful score. This is likely to happen because of overfitting of the data (taking into consideration random noise in our pattern). What we want our model to do is to take the overall pattern of the data in order to correctly classify whether a potential client will subscribe to a saving account or not. It is most likely that overfitting could give us nearly perfect scores (100% and 99%) accuracy scores.

How can we avoid Overfitting?

The best alternative to avoid overfitting is to use cross validation. Taking the training test and splitting it. For instance, if we split it. We will do the testing process three times. This algorithm will iterate through all the training and test sets and the main purpose of this is to grab the overall pattern of the data.

3. ***Modelling.***

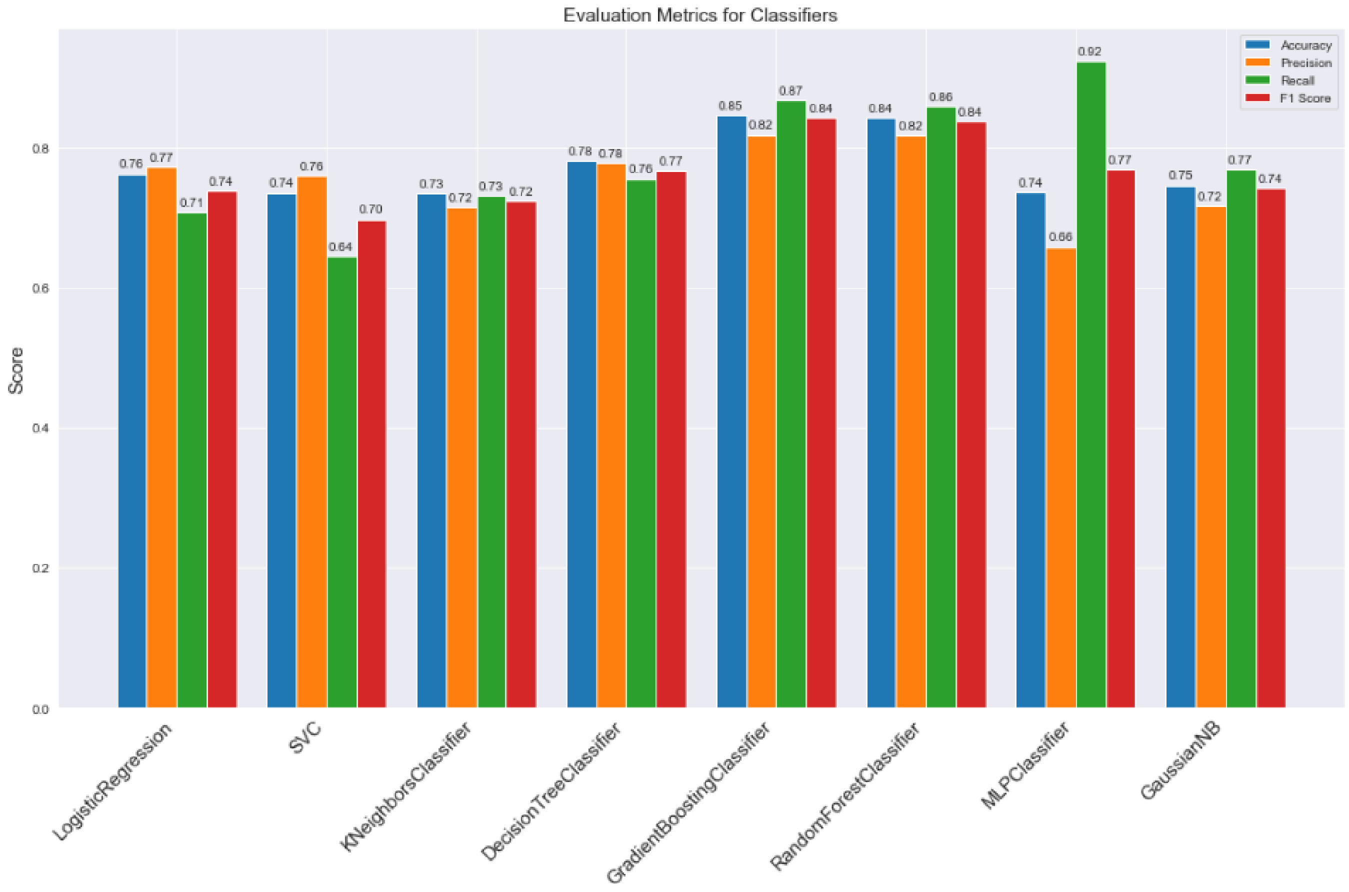
After building several different models and validating them with cross validating scores, we have the models with the scoring results sorted by descending order. The best performing model (by accuracy score) below:



	Classifiers	Crossval Mean Scores
4	Grad B CLF	0.841416
5	Rand FC	0.828536
3	Dec Tree	0.774106
0	Logistic Reg.	0.761788
6	Neural Classifier	0.755068
7	Naives Bayes	0.744205
2	KNN	0.741741
1	SVC	0.739724

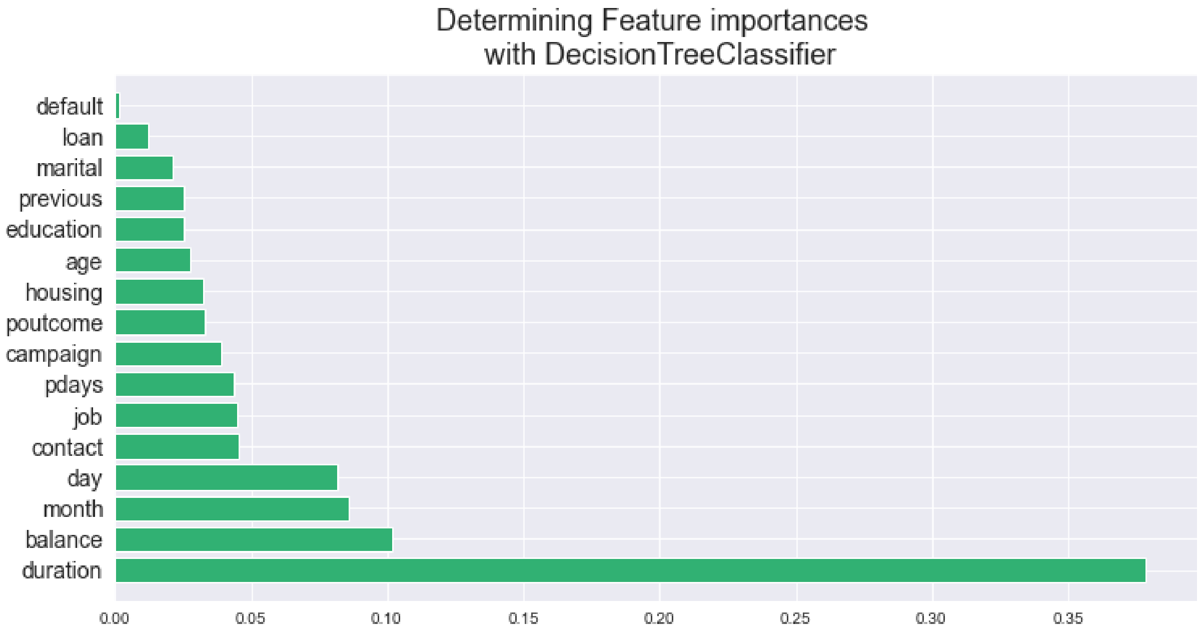
We want to see each classifier’s name in the chart to see the best performing model in terms of the accuracy precision, recall, f1 scores.
We will predict our ‘y’ variable through the models we built and see the results.

The detailed result of the best performing model can be seen the following chart:



The last step of our analysis is determine feature importance of the bank data. The meaning of this is that we are going to see which feature of our data is crucial and can have a bigger impact in deciding which feature of our data we need to target in order to have a successful campaign.

4. ***Feature importance***



We chose Gradient Boosting Classifier model as it has the best result in all aspect. The above chart shows that ‘duration’, followed by ‘balance’ have the biggest impact.

Conclusion

1) Months of Marketing Activity: We saw that the month of highest level of marketing activity was the month of May. However, this was the month that potential clients tended to reject saving deposits offers. For the next marketing campaign, it will be wise for the bank to focus the marketing campaign during the months of March, September, October and December. (December and March should be under consideration because there were the months with the lowest marketing activity, there might be a reason why.)

2) Seasonality: Potential clients opted to subscribe to the account during the seasons of fall and winter. The next marketing campaign should focus its activity throughout these seasons.

3) Campaign Calls: A policy should be implemented that states that no more than 3 calls should be applied to the same potential client in order to save time and effort in getting new potential clients. Remember, the more we call the same potential client, the likely he or she will decline to open a saving account.

4) Age Category: The next marketing campaign of the bank should target potential clients in their 20s or younger and 60s or older. It will be great if for the next campaign the bank addressed these two categories and therefore, increase the likelihood of more saving accounts to open.

5) Develop a Questionnaire during the Calls: Since duration of the call is the feature that most positively correlates with whether a potential client will open the account or not, by providing an interesting questionnaire for potential clients during the calls the conversation length might increase. Of course, this does not assure us that the potential client will open it. Nevertheless, we don't lose anything by implementing a strategy that will increase the level of engagement of the potential client leading to an increase probability of opening, and therefore an increase in effectiveness for the next marketing campaign the bank will execute.

6) Target individuals with a higher duration: Target the target group that is above average in duration, there is a highly likelihood that this target group would open for opening the new account. This would allow that the success rate of the next marketing campaign would be highly successful.

Here are the few recommendations for the bank than can help improve the deposit rate

- Listen to the leads and extract more information to deliver the best deposit plan, which can increase the duration of calls and that can lead to a deposit.

- Approaching the leads during the start of new bank period (May-July) will be a good choice as many have shown positive results from data history.