PROJECT REPORT

SIX WEEK INDUSTRIAL TRAINING

**ON MACHNE LEARNING (ML)**

AT

C-DAC MOHALI



PROJECT - BANK DATA ANALYSIS USING MACHINE LEARNING

MENTORED BY - MR. PEEYUSH JAIN

MISS. HIMANI

PREPARED BY - BY PIYUSH BHATIA

(Batch 1)

Introduction:

Nowadays, Marketing itself is not a cheap task. It requires massive capital income and a lot of time.

Hence, it becomes essential for organizations to optimize marketing strategies and improve them

effectiveness. Understanding customer’s need leads to more effective marketing plans, returns and

customer satisfaction.

One example of this is the Banking Industry. In Banking Industry, it becomes important to analyse

the customer base and work accordingly for their satisfaction.

The Problem Statement of this Project is as follows:

On the basis of a direct Marketing Campaign Conducted by a Portuguese Bank,

***The Classification goal is to predict if an individual is going to subscribe in a Term***

***Deposit*.**

This problem becomes important to be discussed because marketing calls are not meant for all. For an individual, they have their own needs and here it becomes important for the bank to keep in mind the needs of their people before introducing them to new products.

To solve this problem, Machine Learning and Data Analysis would be used.

Basically, the data set would first be evaluated for the various variables it owns. A proper inside out Analysis of this data set will give us an idea about the useful and not so useful parts of the data set. The not so useful data would be cleaned and finally we will have a clean data which we would use for the ML Analysis and prediction.

The Basic aim of this project is to predict the class of people who would be interested in taking a term deposit.

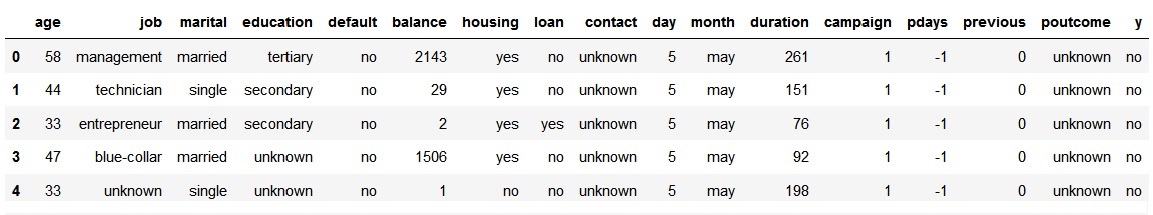
The final conclusion would be to give a range of features a person must have for his chances of taking term deposit are high.

The final result would be to predict either a yes or a no (Accurately) using ML algorithms on for an individual.

**PROBLEM DEFINATION, AlGORITHM & EXPERIMENTAL EVALUATION**

Task definition:

Our initial data set has 16 inputs features and 1 target feature.



The 16 inputs are age, job, martial …. Poutcome.

The 1 target is y.

We have to predict y for the 16 input values.

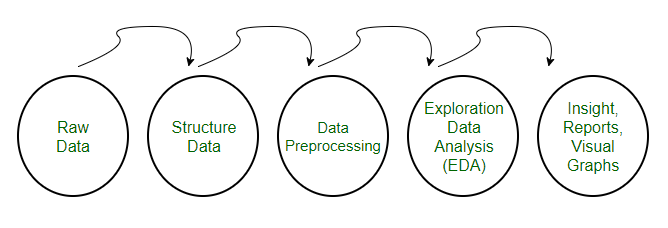
Algorithm Definition:

As previously discussed, it is a Classification problem and the Algorithm, we would be using for it would be Logistic Regression.

METHODOLOGY & KERNEL: -

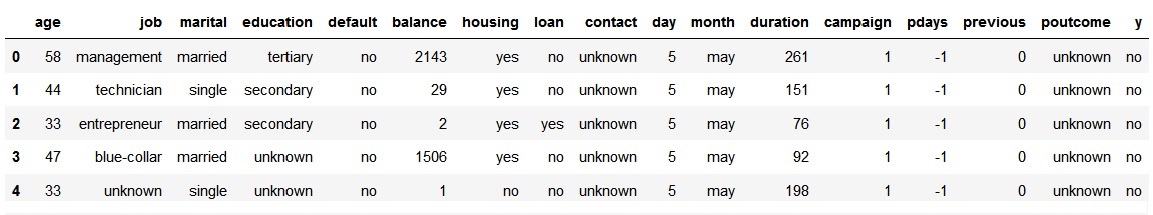
It is divided into 2 basic steps, Data Cleaning and Model Application

STEP 1: Data Cleaning



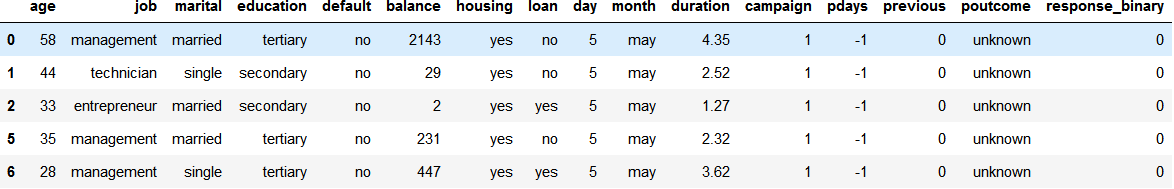
Raw data: -

Raw data is the initial dataset available with us.



Structure Data & Data Pre-processing: -

Our next step in Data cleaning is to organize the Data which involves, taking out unnecessary data from the dataset, identifying and cleaning outliers, dealing with missing data and filtering the data.



That is how our data set looks like after dropping the contact column, removing outliers, changing the y to response\_binary, changing the duration to minutes and doing the other necessary changes.

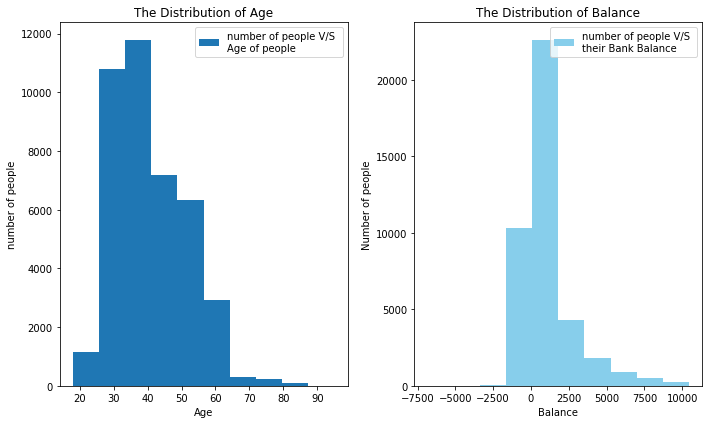
Exploration Data Analysis: -

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. test underlying assumptions;
5. determine optimal factor settings.

Next, to obtain a better understanding of the dataset, the distribution of key variables and the relationships among them were plotted.

1. Visualize the distribution of 'age' and 'balance'



**The distribution of age**: From this visualization we get to know that clients called by the bank have an extensive age range, from 18 to 95 years old. However, a majority of customers called is in the age of 30s and 40s (33 to 48 years old fall within the 25th to 75th percentiles).

**The distribution of balance**: From this visualization we get to know that After dropping outliers in balance, the range of balance is still massive, from a minimum of approximately -4500 euros to a maximum of approximately 10500 euros, giving a range of 6000 euros.

### Visualize the relationship between 'age' and 'balance

Based on this scatter plot, we cannot derive any clear relationship between client’s age and balance level.

though, over the age of 60, clients tend to have a significantly lower balance, mostly under 5,000 euros. This is due to the fact that most people retire after 60 and no longer have a reliable income source.

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\BBF71648.tmp

### Visualize the distribution of 'duration' & 'campaign'

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\39F8FB96.tmp

**The distribution of duration**: As observed from the box plot, the duration of contact has a median of 3 minutes, with an interquartile range of 1.73 minutes to 5.3 minutes.

**The distribution of campaign**: About half of the clients have been contacted by the bank for the second time, while 25% was first introduced to the term deposit. Most clients have been reached by the bank for one to three times, which is reasonable. However, some clients have been contacted by as high as 58 times, which is not normal.

### Visualize the relationship between 'duration' & 'campaign': with response result

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\42B16414.tmp

What we can inference from this plot is that, “yes” clients and “no” clients are forming two relatively separate clusters. Compared to “no” clients”, “yes” clients were contacted by fewer times and had longer call duration.

More importantly, after five campaign calls, clients are more likely to reject the term deposit unless the duration is high.

This suggests that the bank should resist calling a client for more than five times, which can be disturbing and increase dissatisfaction.

1. Correlation Matrix

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\778BA742.tmpTo visualize about correlation of the features, a correlation matrix was plotted with all qualitative variables. Clearly, “response\_binary” has a strong correlation with “duration”, a moderate correlation with “previous contacts”, and mild correlations between “balance”, “month of contact” and “number of campaigns”. Their influences on campaign outcome will be investigated further in the machine learning part.

INSIGHTS AND VISUALIZATION

With a proper knowledge of the distribution of key variables, further analysis of each customer characteristic can be carried out to investigate its influence on the subscription rate.

### DATA VISUALISATION

### Visualize the subscription and contact rate by age

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\2A85D99.tmp**The inference from this graph is that people above the age of 60 have a very high chance of taking term deposit. Also, people under the age of 30 are a good target audience. This means target the youngest and the oldest instead of the middle-aged people.**

### Visualize the subscription rate by balance level

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\7B5A2C0F.tmp

The insight of this plot is that to **target clients with average or high balance as their probability of taking the term deposit is higher than the low or no balance class of people.**

### Visualize the subscription rate by job

### C:\Users\jack\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\D2BDDFF5.tmp

As noted from the horizontal bar chart, students and retired clients account for more than 50% of subscription, which is consistent with the previous finding of higher subscription rates among the younger and older and hence we confirm that is our target audience.

Our Step 1, I.e. the Data Cleaning part of the dataset is complete here.

Now, we move to the next part of the project, I.e. is the Model Application Part.

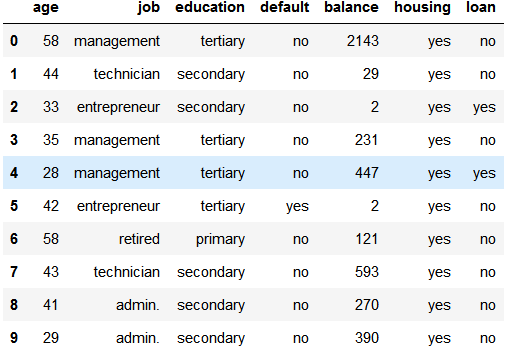
STEP 2: Model Implementation

As our main objective is to identity the most responsive customers before the marketing campaign, to achieve this objective, Classification algorithms will be employed. A classification model will be built to classify all clients into two groups: "yes" to term deposits and "no" to term deposits.

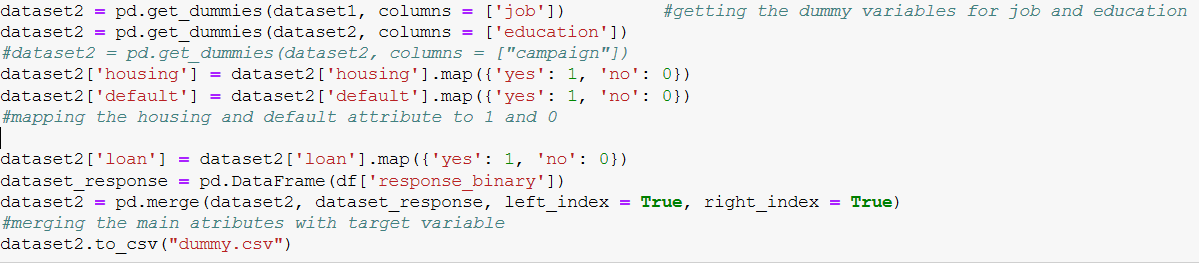
Loading Clean data

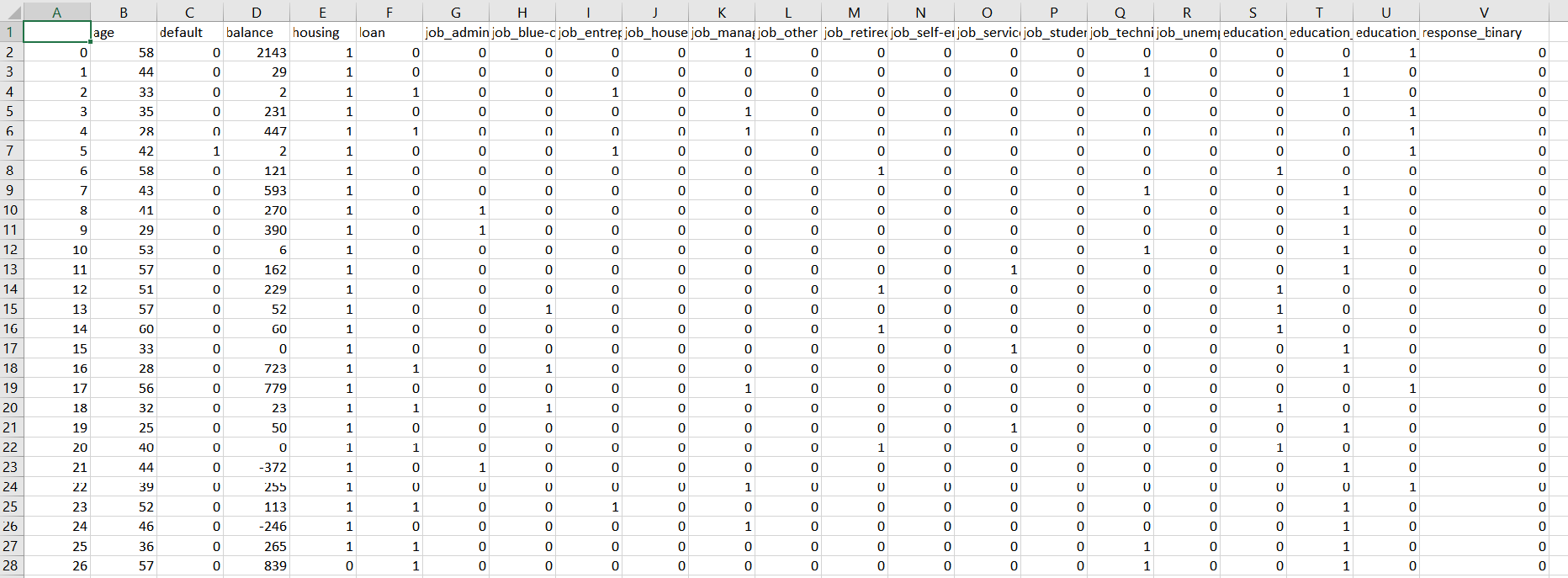
Our first task in the model application is to select the features on which we are going to apply the ML Algorithms.

These are the 7 features selected which would be used for the training of our model.



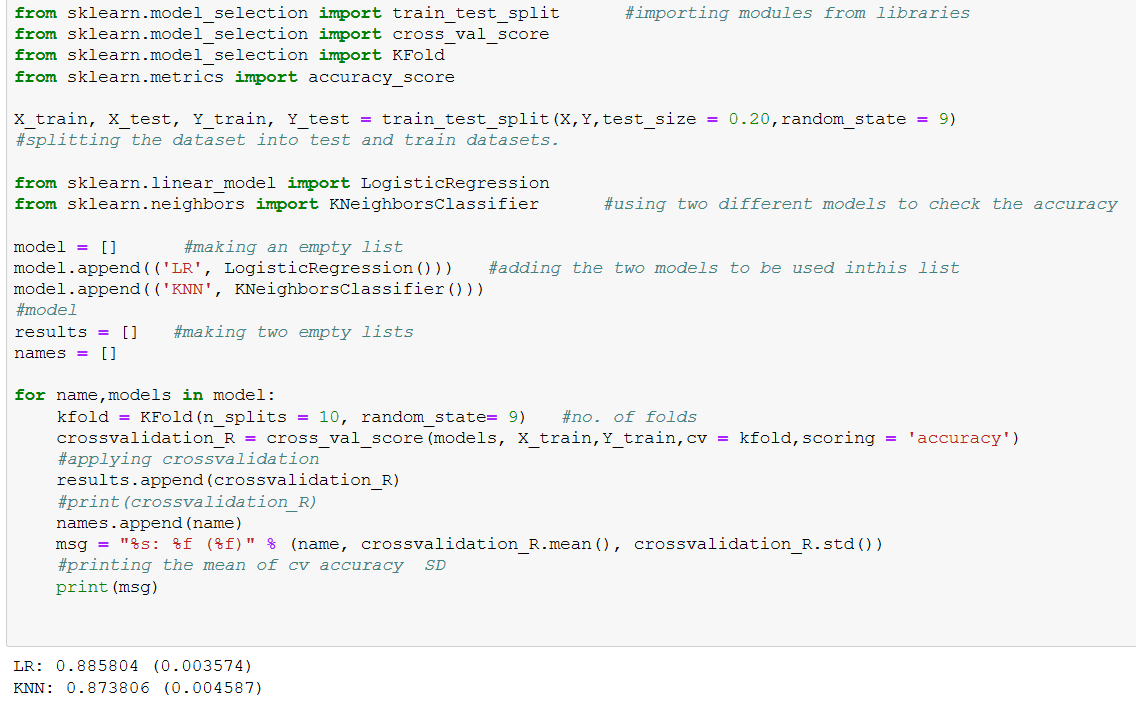
One thing to be noted here is that, for all the classification problem, the value of features must be numerical rather the qualitative. To encounter this problem, the concept of One Hot Encoding & dummy variables is used.



Dataset after the application of One Hot Encoding and addition of Target variable:

Next, what we did was split the dataset into two parts, Test dataset and train dataset using the test\_train\_spilt function from the sklearn library.

Next, two different classification algorithms (Logistic Regression, K-Neighbors Classifier are run on the dataset and the best-performing one will be used to build the classification model.

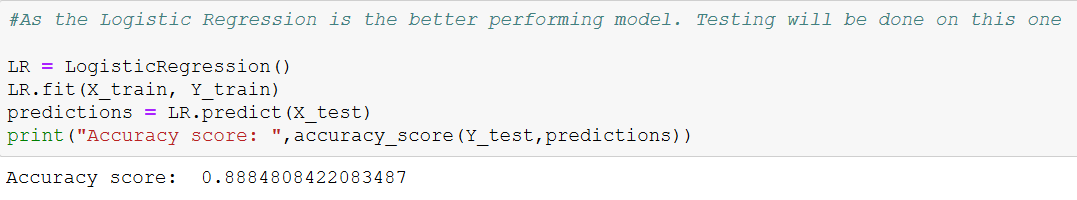


We also used the K-fold model for splitting the training dataset into “n” different sets for the application of Cross-validation.

From the model’s accuracy, what we can conclude is that Logistic regression is the best performing model**.**

Between the two, logistic regression achieved an accuracy of about 88%, suggesting a high level of strength of this model to classify the customer response given all the defined customer features.

Next, we need to implement and evaluate our LR model.



After implementing my LR on the test dataset,

Accuracy score is the percentage of correct predictions out of all predictions made. The LR algorithm achieves an accuracy of 88.84%, suggesting high level of strength of this model to classify the customer response given all the defined customer features.

**The result of accuracy score can possibly yield misleading result if the data set is unbalanced, because the number of observations in different classes largely vary.** A confusion matrix gives a detailed breakdown of prediction result and error types.

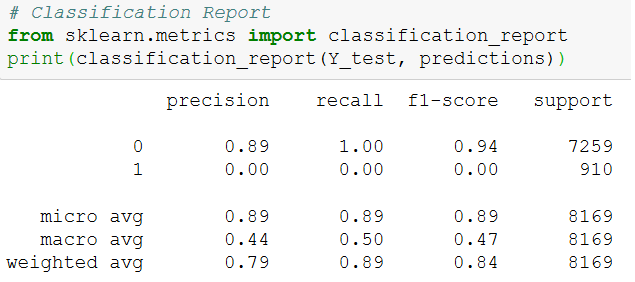
Confusion Matrix: -

**A problem revealed by this confusion matrix is that the dataset is highly unbalanced, with nearly all client actually decline to subscribe.**



This infers that the accuracy score is biased, and further evaluation should be carried out to determine the accuracy of logistic regression model.

Further Evaluation: -



Classification report shows the precision, recall, F1-score and support.

* Precision of 0 represents that for all cases when prediction was as no, the percentage of clients that actually said no is 89%.
* Recall is the ability of a classifier to find all positive instances. Recall of 0 tells us that for all people who actually said no, the model predicts 100% correctly that they would decline the offer.

CONCLUSION

**The main objective of this project was to increase the effectiveness of the bank's telemarketing campaign, which was met successfully through data analysis, visualization and analytical model building. A target customer profile was established with the help of classification model which was built to predict customers' response to the term deposit campaign.**

**According to the Data Analysis, a target customer profile which can be established.**

**The most responsive customers hence are:**

* Feature 1: age < 30 or age > 60
* Feature 2: students or retired people
* Feature 3: a balance of more than 5000 euros

By applying logistic algorithm, classification model was successfully built. With this model, the bank will be able to predict a customer's response to its telemarketing campaign before calling this customer. In this way, the bank can allocate more marketing efforts to the clients who are classified as highly likely to accept term deposits, and call less to those who are unlikely to make term deposits.

In addition, predicting duration before calling and adjusting marketing plan benefit both the bank and its clients. On the one hand, it will increase the efficiency of the bank’s telemarketing campaign, saving time and efforts. On the other hand, it prevents some clients from receiving undesirable advertisements, raising customer satisfaction. With the aid of logistic regression model, the bank can enter a virtuous cycle of effective marketing, more investments and happier customers.

RECOMMENDATIONS AND FUTURE SCOPE:

1. **More appropriate timing**

When implementing a marketing strategy, external factors, such as the time of calling, should also be carefully considered. The previous analysis did not deal with the month of calling. A further study of seasonal activity and time of calling should be done to improve the effectiveness of the model.

1. **Smarter marketing design**

By targeting the right customers, the bank will have more and more positive responses, and the classification algorithms would ultimately eliminate the imbalance in the original dataset. Hence, more accurate information will be presented to the bank for improving the subscriptions. Meanwhile, to increase the likelihood of subscription, the bank should re-evaluate the content and design of its current campaign, making it more appealing to its target customers.

1. **Better services provision**

With a more granular understanding of its customer base, the bank has the ability to provide better banking services. For example, marital status and occupation reveal a customer's life stage while loan status indicates his/her overall risk profile. With this information, the bank can estimate when a customer might need to make an investment. In this way, the bank can better satisfy its customer demand by providing banking services for the right customer at the right time.

BIBLOGRAPHY:

Search Engine: -

1. [www.google.co.in](http://www.google.co.in)

Python Implementation/Machine Learning

1. <https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/>
2. <http://algolytics.com/tutorial-how-to-determine-the-quality-and-correctness-of-classification-models-part-1-introduction/>
3. <https://www.geeksforgeeks.org/ml-dummy-variable-trap-in-regression-models/>
4. <https://towardsdatascience.com/linear-regression-python-implementation-ae0d95348ac4>

**Dataset From: -**

[http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)