Week 10

Name	Email	Country	University	Specialization
Ilyas Nayle	ilyasnayle5@gmail.com	Turkey	TED University	DATA Science

Problem Statement:

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Data Set Information:

The data is related with 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 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) a term deposit (variable y).

Number of Columns	17
Number of rows	45212
Number of columns with missing values	Null
Now of rows with missing values	Null
Total number of categories	10
Total number of integers	7
Output	No: Client did not subscribe to a term Deposit Yes: Client did subscribe to a term Deposit

Since the data is about the information gathered from a phone call made by the bank during previous market research. The response variable is categorical, with 'No' representing that such client has not registered to a term deposit and 'Yes' representing that such client has subscribed to a term deposit.

The column 'Duration' is significant since it is considered to explain the variability in the response with certainty. There are 17 features in all, 10 of which are categorical and the rest are numerical.

Problem:

The Problem here is that the Data sets are outliers in AGE, DURATION. Where duration is right skewed as most of the data is close to 0. some of the data type features needs to be changed as it is in numerical. The data sets seems imbalanced because there are more “NO” than “YES”

Since the Duration can differ from customer to customer this outliers won’t be altered while age isn’t having any significant outliers. But, the data needs to perfected by “NUMBER OF DAYS PASSED BY after the client last contacted during previous research” .

Final Recommendation:

The following are the final recommendation that is finalized after implying the EDA on the Bank data set file;

Note: The word “ Those” refers to clients.

I. Those Retired are more likely to buy policy than others
II. Those who are between age 20-50 are more likely those who bought the policy and also there are those above 70 also have opted the policy.
III. Those Married are more likely to buy between the age of 60 – 80.
IV. Those without “no” as their option in default column are more likely to buy the policy.
V. Those who are in management section are more likely to get the policy
VI. Those in the Secondary are more than others in term of subscription.
VII. Those who will most likely opt for policy are between the month of may – October as well as September. Moreover the contacts before the campaign brings more clients.
VIII. Those with the cell-phones have higher rate of subscription than those with none and telephone.

IX. There is an average of 426 seconds for those who opt the policy and for those who did not where the call lasted for 146 seconds which is fewer than those who opt for the policy. Although they spend more time on the communication when they opt for the policy.

X. Those with “no” in their default and more likely between the age of 20-60 years old.