

Week 8

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

Problem Statement:

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 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	4521
Number of columns with missing values	Null
Number 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.

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 be perfected by "NUMBER OF DAYS PASSED BY after the client last contacted during previous research" .