Team Member Details

Team name: Team Muadh
Member count: 1

- Name: Muadh Faizan
- Email: muadhfaizan@gmail.com
- Country: United Arab Emirates
- Company: Global Management Consultants
- Specialisation: Data Science

Problem Description

The project chosen to be done is the Bank marketing project, which is regarding a term deposit product of a Portuguese bank. The bank wants to sell a term deposit product to its customers but want to know which of their customers will be more likely to buy it, based on the past interactions with the bank. A classification problem is at hand, where the dataset contains information of several customers who were already informed about the term deposit, such as their age, gender, and other information pertaining to their bank accounts and loans. Whether the customer had bought the product or not, is also given for each customer.

The aim of the project is to analyse the dataset and come up with a classification model which would be able to predict if a customer would buy the product or not.

Data Understanding

The dataset contains around 45,000 rows and 17 columns. The features include the customers' age, occupation, bank information, and other communication related information. Most of the features are categorical and binary.

There are no missing values, which is a good sign. Four features have 'unknown' as a value. The 'poutcome' value has nearly 36,000 unknown values, but that is okay since those customers were never contacted for a previous campaign, so there is actually no outcome. The value can be changed to 'Not applicable' for better understanding.

Outlier Treatment

There are many outliers in the numeric variables. Some are fine to keep, such as those in the Age or Campaign features. There are many outliers in the ‘pdays’ feature, but that is because nearly all customers were not contacted before so they all have a value of -1. This makes the customers that have been contacted and have many days that have past, far from the mean and displayed as outliers. These do not need to be treated. Some other features have outliers that are logical, but are still extremely further away from the mean which would be better off being treated.

For treating the outliers, I will only treat the outliers that are extremely far from the mean, and which would significantly cause skewness and maybe some issues in the model. I will clip them to a value that is more closer to the mean, which may still make them outliers but not as extreme as they were. This is so that the data still has good representation, and the integrity of the dataset is intact as much as possible.