Week 8 Deliverables

Group Name: Carpe-Diem group

Specialization: Data Science

Project Name: Bank Marketing (Campaign) -- Group Project

Team Members:

1. **Name**: Mohini Kalbandhe
   - **Email**:amohini099@gmail.com
   - **Country**: Canada
   - **Company**: Happy Orchard
   - **Specialization**: Data science
2. **Name**: Kashish Joshipura

- **Email**: kashishjoshipura@gmail.com
- **Country:** Canada
- **Collage**: The University of British Columbia (UBC)
- **Specialization**: Data Science

3. **Name**: Amir Shahcheraghian
   - **Email**: Amir.shahcheraghian@gmail.com
   - **Country:** Canada
   - **Collage**: University of Quebec , Canada
   - **Specialization**: Data Science, Energy Management analysis

Bank Marketing (Campaign)

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 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).

Problem description:

One of the most common marketing strategy In Banking sector is direct marketing campaigns through phone calls ,it is a form of advertising that allows organizations to communicate directly with customers to offer their services based on the client's existing bank profile .Here we will consider term deposit as a banking service .

Business Goal :

To build a list of target customers who are likely to subscribe a term deposit. The more targeted our campaigns, the more successful they are likely to be.

Project Objective:

By converting this problem into a machine learning classification problem we will build a model to predict whether a client will subscribe a term deposit or not so that the banks can arrange a better management of available resources by focusing on the potential customers "predicted" by the classifier .

Technique to be used: Classification

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.

Attribute Information:

Bank client data:

- **Age** (numeric)
- **Job** : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- **Marital** : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- **Education** (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- **Default**: has credit in default? (categorical: 'no', 'yes', 'unknown')
- **Housing**: has housing loan? (categorical: 'no', 'yes', 'unknown')
- **Loan**: has personal loan? (categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

- **Contact**: contact communication type (categorical: 'cellular','telephone')
- **Month**: last contact month of year (categorical: 'jan', 'feb', 'mar', …, 'nov', 'dec')
- **Dayofweek**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- **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. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- **Campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **Pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- **Previous**: number of contacts performed before this campaign and for this client (numeric)
- **Poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

- **Emp.var.rate**: employment variation rate - quarterly indicator, it defines as a measure of the extent to which available labor resources (people available to work) are being used. (numeric)
- **Cons.price.idx**: consumer price index - monthly indicator (numeric), it expresses the change in the current prices of the market basket in terms of the prices during the same month in the previous year.
- **Cons.conf.idx**: consumer confidence index - monthly indicator , CCI is a survey administered by The Conference Board, that measures how optimistic or pessimistic consumers are regarding their expected financial situation (numeric)
- **Euribor3m**: euribor 3 month rate - daily indicator (numeric), it is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months Nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

- **y** - has the client subscribed a term deposit? (binary: 'yes', 'no')

Data Understanding:

- Portuguese Bank wants to improve marketing campaign to recommend which customer is to target by analyzing their past marketing data.
- The Motivation is by devising such prediction algorithm the bank can be a better target for its customer and better channelize its customer.
- Bank of Portugal offered its customer fixed term products such as CD's. Data was collected about each client, type of contact and outcome.
- The classification goal is to predict if the client is subscribed or no for term deposit.
- We are going to use bank-additional-full.csv most for this project and it has Numerical Data Type in columns Age, Balance, Campaign, Day, Duration, Pdays, Previous. Categorical data type in column Job, Marital status, Education, Default, Housing ,Loan, Contact, Month, Poutcome. Binary data type used in Y.

Queries Related to Data Analysis:

Q1. What type of data you have got for analysis?

The data seems to be cleaned and it is heavily skewed however you can see most of the variables have outliers that needs to be cleaned with data cleaning process. There are Numerical, categorical as well as binary data types used in data.

Q2. What are the problems in the data (number of NA values, outliers , skewed etc)?

There are certain outliers in columns “age”, “duration”, “campaign” and most columns are skewed. The data has no missing value.

Q3. What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

Solution for NA values:

- We will replace missing values with mean median in numerical values.
- In categorical values we will replace missing value with the mode value and the most likely value of the missing one.

Solution for outliers:

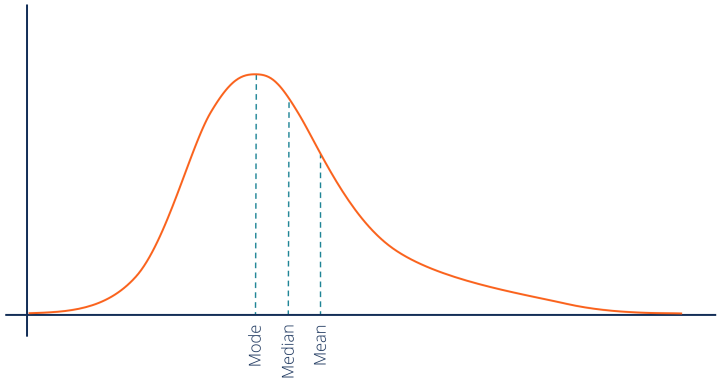
- As outlier differs significantly from other observations and change the meaning of data we need to remove all the outliers in the data from all the columns.

Solution for skewness:

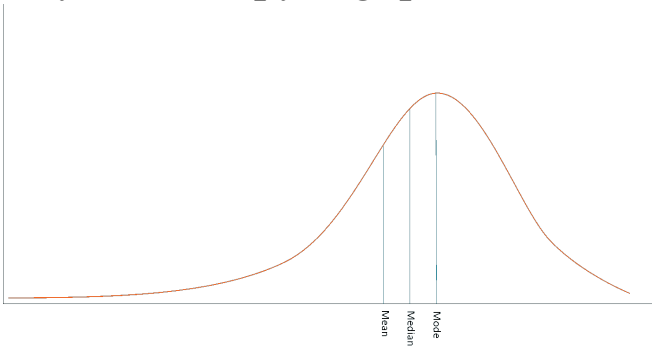
- Skewed column represents normal distribution, it can be right skewed or left skewed.
- Asymmetrical distribution have skewness of “0”.
- There are two types of skewness , positive skewness and negative skewness.

1. Positive skewness(or right skewed) distribution is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer. The positively skewed distribution is the direct opposite of the negatively

skewed distribution. The mean is greater than median and is greater than mode which means more clients are subscribed with no of term deposit by less than the average customers. For extreme positively skewed data we can use data transformation such as log transformation to reduce skew and overcome the problem.



2. Negative skewness (or left skewed) tail of left side is longer and fatter, the mean and median is less than mode, skewness differentiate in extreme value of one verses other tail.
   The high skewness may lead to misleading result in statistical test so the data goes through transformation process to make it close to normal distribution. You may use numpy.log1p on the column to remove it.



Github Repo link:

https://github.com/amohini099/Banco-de-portugal-marketing

https://github.com/amirshq/Bank-Marketing-Campaign.git