

# Bank Marketing ( Campaign ) Project Details

## Team member's details

Name: Batuhan YILMAZ  
Email: [batuhanyilmaz1999@hotmail.com](mailto:batuhanyilmaz1999@hotmail.com)  
Country: Turkey  
Company: A university in Turkey  
Specialization : Data Science

## Group Name

Datarpher

## Problem Description

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).

## Business Understanding

Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.

This will save resource and their time ( which is directly involved in the cost ( resource billing)).

## Project lifecycle

- **Business Understanding** ( week 1 )
- **Data understanding** ( week 1 )
- **Exploratory data Analysis** ( week 2 )
- **Data Preparation** ( week 3 )
- **Model Building ( Logistic Regression, ensemble, Boosting etc)** ( week 4 )
- **Model Selection** ( week 5 )
- **Performance reporting** ( week 6 )
- **Deploy the model** ( week 6 )
- **Converting ML metrics into Business metric and explaining result to business** ( week 7 )
- **Prepare presentation for non technical persons.** ( week 7 )

# Data Intake Report

**Name:** Bank Marketing (Campaign)

**Report date:** 19.07.2023

**Internship Batch:** LISUM22

**Version:**

**Data intake by:** Batuhan YILMAZ

**Data intake reviewer:**

**Data storage location:** <https://archive.ics.uci.edu/dataset/222/bank+marketing>

**Github repository link:** <https://github.com/Batuhan-Ylmz/Bank-Marketing-Campaign-Term-Deposit-Product-Purchase-Classification>

**Tabular data details:** bank-full

<b>Total number of observations</b>	45211
<b>Total number of files</b>	1
<b>Total number of features</b>	17
<b>Base format of the file</b>	csv
<b>Size of the data</b>	5.9+ MB

**Tabular data details:** bank-additional-full

<b>Total number of observations</b>	41188
<b>Total number of files</b>	1
<b>Total number of features</b>	21
<b>Base format of the file</b>	csv
<b>Size of the data</b>	6.6+ MB

Data files are same as each other. However, bank-additional-full.csv file includes more specific details regarding the customers. ( E.g; contact day with customer, consumer price index, consumer confidence index ...).

## Data and Business Understanding:

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).

Since there are some unnecessary kind of data that is likely to lower the quality of prediction model in the dataset with 21 features, the one dataset with 17 features ( 16 feature + 1 output feature) is selected as the data for model creation.

Each of 16 feature corresponds to a specific feature of that customer ( feature names can be modified for easier reading later on):

- 1- age ( “**age**” )
- 2- job ( “**job**” )
- 3- marital status ( “**marital**” )
- 4- education level ( “**education**” )
- 5- if customer has credit in default ( “**default**” )
- 6- yearly balance in euro ( “**balance**” )
- 7- if customer has housing loan ( “**housing**” )
- 8- if customer has personal loan ( “**loan**” )
- 9- how did the bank contacted the customer ( “**contact**” )
- 10- last contact day of the month ( “**day**” )
- 11- Last contact month of the year ( “**month**” )
- 12- Last contact duration ( “**duration**” ) → This feature highly affects the output. So, 2 models will be created with one including this feature and the other one not.
- 13- How many times customer was contacted for this campaign ( “**campaign**” )
- 14- How many days have passed since the customer was contacted for the previous campaign ( “**pdays**” )
- 15- Total number of contacts performed before this campaign ( “**previous**” )
- 16- Did the customer subscribed for the previous term deposit product ( “**poutcome**” )
- 17- Has the client subscribed a term deposit ( “**y**” )

## Balance of the dataset:

- The output of the data is **imbalanced**. While the ratio of people who have subscribed to the deposit product **is about 11% (“yes”)** , people who have not subscribed to the deposit product is about **%89 (“no”)**.
- Machine learning algorithms learn with the assumption that distribution of the provided labelled ( output ) data is symmetrical.
- In case of an imbalanced dataset, in order machine not to learn one class poorly compared to other one, **some handling methods will be applied during the EDA and Preprocessing stages.**

## Missing Values:

Dataset was splitted into 2 datasets as numerical and categorical features.

- Both dataset were investigated and it was seen that there were **no missing values in the entire dataset**.
- However, some categorical columns have “**unknown**” values. Since sometimes customers with unknown features are likely to subscribe as well, they were not removed from the dataset.

## Outliers:

- Numerical dataset were summarized using the 5 number summary ( **min, 25%, median, 75%, max**) in order to see the distribution of each feature.
- Impact of current feature on the output was visualized to see how much the outliers influences it.
- The correlation between the numerical features and the output was investigated with a **heat-map**.
- **Observations have shown that:**
  - o *“pdays” and “previous” columns have quite less impact upon the output.*
  - o *Their variance are pretty less and consist of single values.*
  - o *People with the age range 20-60 are more likely to subscribe the deposit product.*
  - o *Duration feature has a strong impact on the output. For a healthy prediction model, 2 model will be created where first the one includes it and the second one does not.*
  - o *There is a positive skewness mostly in remaining columns.*

## Approach to the outliers

- First, “**pdays**” and “**previous**” columns are dropped.
- Then, 4 different methods were picked in order to handle outlier data in the numeric\_data dataset.
- As first method **Interquartile Range method ( IQR )** was used.
  - o Since there is always a certain range for the remaining numerical features, coming up with an approach that eliminates the outliers that fall beyond this acceptable range can be useful.
- As second method “**winsorization**” method which is an approach based on the percentile for specific ranges was chosen.
  - o A specific percentile is selected ( eg. %90, then data points smaller than %5 and greater than %95 of whole data are considered as outlier and replaced with the nearest data points).
- As third method “**MinMax Scaling**” which is an approach similar to Z-score was chosen.
  - o Rather than using the mean for outlier detection in calculation, it uses median to set the limits. Then, scales the data to a specific range. ( Mostly [0,1]).
- As fourth method “**Z-score (Standard Scaling)** ” which is an approach based on re-scaling the data points that fall beyond the calculated z-score.
  - o Z-score is calculated by subtracting the current data point from the mean value and dividing it by std value for each data points

- 4 different numerical datasets were generated with the name of related outlier handling methods:
  - o **Numeric\_data\_without\_outliers\_IQR,**
  - o **Numeric\_data\_without\_outliers\_winsorization,**
  - o **Numeric\_data\_without\_outliers\_MinMaxScaling,**
  - o **Numeric\_data\_without\_outliers\_StandardScaling,**
- All these numeric datasets will be separately combined with the categorical data dataset that was created in the beginning while one version of the combined dataset having the “duration” column and the other version does not.

## Cleaning and Transformation of Categorical Data

- Even though there are no missing values in categorical columns, high amount of “unknown” category is exist in one of the features of categorical\_data. However, this “unknown” type category makes up more than the %80 of whole data whose modification might cause huge misleads for the learning process of model. So, it is treated as a separate and valid category like the others.

## Approach for Encoding the Categorical Columns

- For handling the encodings of categorical data, the most common encoding approaches will be used:
  - o **One-hot encoding**
  - o **Label Encoding**
- Since one-hot encoding creates new columns as many as the number of categories in that feature, using one-hot encoding for more than a threshold value that of the specific dataset–e.g, depending on the feature numbers in dataset this can be around 10- it may lead to a high-dimensional feature space which
  - o **Can be computationally more expensive,**
  - o **Lead to the situation known as “curse of dimensionality”.**
    - **Such as:**

Education_primary	education_high_school	education_college
0	1	0
1	0	0
.....	.....	.....
- In case there are too many different categories for the features, Label Encoding is used to specify an integer for the corresponding category.
  - **Such as:**

Education	month
0	3
7	8
2	6
.....	.....
- In the categorical dataset, only the “month” and “job” columns contains more than 10 categories.

- Hence, only these 2 categorical features ("**month**", "**job**") were encoded using "**LabelEncoder**" in order to prevent high-dimensionality.
- Rest of the categorical features encoded using **OneHotEncoder**.

Currently there are **4 different numeric\_features** without outliers processed with **4 different outlier handling algorithms** and categorical\_features where features with more than 10 are encoded with LabelEncoder and less than 10 are encoded with OneHotEncoder.

During the EDA step, data imbalance problem will be handled and all the dataset will be prepared to be trained by diverse kind of models.