Week 8 Deliverables: Data Understanding

Zakaria Arshad

Data Science Specialization

Submitted for Data Glacier

March 19, 2023

Table of Contents

1. Problem Description

ABC Bank is planning to launch a new term deposit product, but before they do, they want to develop a machine learning model that can predict whether a particular customer is likely to buy the product or not, based on the customer's past interactions with the bank or other financial institutions. The goal is to use this model to shortlist potential customers and focus their marketing efforts only on those customers who are more likely to buy the product. This would help the bank save time and resources that would otherwise be wasted on marketing to customers who are less likely to buy the product.

2. Data Understanding

2a. Type of Data in Dataset

We are provided a combination of both categorical and numeric data. The first seven variables provide data about the bank's client, such as age, job, marital status, and education. The next eight variables provide data about contact with the client, including last contact (month and day), duration, number of contacts performed, and outcome. Finally, the last five variables provide data on various broader economic variables, such as employment variation rate, consumer price index, and consumer confidence index.

From examining the data, we have 10 numeric variables and 11 categorical variables.

All variables and specific data types can be seen below. There are five integer data types, eleven object data types (representing categorical variables), and six float data types.


```
file.dtypes
```

```
age               int64
job              object
marital          object
education        object
default          object
housing          object
loan             object
contact          object
month            object
day_of_week      object
duration          int64
campaign          int64
pdays             int64
previous          int64
poutcome         object
emp.var.rate    float64
cons.price.idx  float64
cons.conf.idx   float64
euribor3m       float64
nr.employed     float64
y                object
dtype: object
```

2b. Identifying problems in dataset and methods of dealing with those problems

The dataset has no missing values, as seen below. Null values will not be an issue here.

```
file.isna().sum()
```

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

Possible outliers may exist in the numerical variables: determining whether these are part of natural variation of the data will be determined during the cleaning process of

the data. Methods to find outliers will include plotting data to visually detect outliers, or using z-scores and quartiles to detect outliers. These are simple yet effective methods of identification.

The most prevalent issue in this dataset is imbalanced data, specifically for if the client will subscribe to a term deposit. There is a much higher representation of no responses then yes responses, as seen below.

```
file['y'].value_counts()
```

```
no     36548
yes     4640
Name: y, dtype: int64
```

This is the variable we are trying to predict, or the target variable. Thus, it is necessary to apply some form of method to balance the data, as the model will be inaccurate otherwise.

Some possible techniques to deal with this imbalance are oversampling, undersampling, or Synthetic Minority Over-sampling Technique (SMOTE), which is a more specific type of oversampling..

- Oversampling and SMOTE is used to increase the amount of samples in the minority class. This would be the “yes” class.
- Undersampling reduces the amount of samples in the majority class. This would be the “no” class.

Oversampling and SMOTE work best with limited datasets, with very small minority classes. Undersampling works best with larger datasets where small amounts of data can be lost.

The minority class represents 11% of all data points. It is hard to predict which method will work best without examining the data further.

Overall, there are multiple data transformations that will be necessary to improve the quality of the data for model building.