Customer Segmentation Project

**Report date**: 19-March-2022

**Internship Batch**: LISUM06

**Specialization**: Data Science

**GitHub link**: <https://github.com/saadbinmunir/Customer-Segmentation-Project>

# Team member details:

**Saad Bin Munir**

* Email: [saadmunir24@gmail.com](mailto:saadmunir24@gmail.com)
* Country: United Kingdom
* University: University of Central Lancashire

# Problem Description

Bank XYZ wants to offer Christmas offers to its customers. However, the bank does not want to offer the same offer to all its customers. Instead, they want to deploy the personalised offer to a particular group of customers. It is not effective to manually start understanding the category of the customer because they will not be able to uncover the hidden pattern in data. ABC analytics assigned this talk to their analytics team and instructed their team to come up with the approach and feature which group similar behavior customer in one category and others in different category.

# Data understanding

The dataset consists of details of Bank customers from 1995 to 2015. The observation in the dataset correspond to unique customer in the dataset. The attributes contain information of each customer such as gender, location, joining date, residence and the products utilised. An overview of the dataset can be seen below.

| **Column Name** | **Description** |
| --- | --- |
| fecha\_dato | The table is partitioned for this column |
| ncodpers | Customer code |
| ind\_empleado | Employee index: A active, B ex employed, F filial, N not employee, P pasive |
| pais\_residencia | Customer's Country residence |
| sexo | Customer's sex |
| age | Age |
| fecha\_alta | The date in which the customer became as the first holder of a contract in the bank |
| ind\_nuevo | New customer Index. 1 if the customer registered in the last 6 months. |
| antiguedad | Customer seniority (in months) |
| indrel | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| ult\_fec\_cli\_1t | Last date as primary customer (if he isn't at the end of the month) |
| indrel\_1mes | Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner ),P (Potential),3 (former primary), 4(former co-owner) |
| tiprel\_1mes | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential) |
| indresi | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| indext | Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country) |
| conyuemp | Spouse index. 1 if the customer is spouse of an employee |
| canal\_entrada | channel used by the customer to join |
| indfall | Deceased index. N/S |
| tipodom | Addres type. 1, primary address |
| cod\_prov | Province code (customer's address) |
| nomprov | Province name |
| ind\_actividad\_cliente | Activity index (1, active customer; 0, inactive customer) |
| renta | Gross income of the household |
| ind\_ahor\_fin\_ult1 | Saving Account |
| ind\_aval\_fin\_ult1 | Guarantees |
| ind\_cco\_fin\_ult1 | Current Accounts |
| ind\_cder\_fin\_ult1 | Derivada Account |
| ind\_cno\_fin\_ult1 | Payroll Account |
| ind\_ctju\_fin\_ult1 | Junior Account |
| ind\_ctma\_fin\_ult1 | Más particular Account |
| ind\_ctop\_fin\_ult1 | particular Account |
| ind\_ctpp\_fin\_ult1 | particular Plus Account |
| ind\_deco\_fin\_ult1 | Short-term deposits |
| ind\_deme\_fin\_ult1 | Medium-term deposits |
| ind\_dela\_fin\_ult1 | Long-term deposits |
| ind\_ecue\_fin\_ult1 | e-account |
| ind\_fond\_fin\_ult1 | Funds |
| ind\_hip\_fin\_ult1 | Mortgage |
| ind\_plan\_fin\_ult1 | Pensions |
| ind\_pres\_fin\_ult1 | Loans |
| ind\_reca\_fin\_ult1 | Taxes |
| ind\_tjcr\_fin\_ult1 | Credit Card |
| ind\_valo\_fin\_ult1 | Securities |
| ind\_viv\_fin\_ult1 | Home Account |
| ind\_nomina\_ult1 | Payroll |
| ind\_nom\_pens\_ult1 | Pensions |
| ind\_recibo\_ult1 | Direct Debit |

# Data type

The dataset provided contains 1000000 rows and 48 columns and is provided in csv format. It consists of categorical, numerical values as well as datatime format. The size of the dataset is 366 MB.

Most of the categorical values are binary however there are some columns with multiple categorical types. All the numerical columns contains integers and the ‘renta’ column consists of continuous values. Some of the values of the categorical columns contain float values. The complete details about the datatypes of each column as well as the overview of the dataset can been visualised in the figure 1 and figure 2 respectively.

Graphical user interface

Description automatically generated with medium confidenceTable

Description automatically generateds

Table

Description automatically generated  
Figure 1. Full detail of columns

Table

Description automatically generated  
Figure 2. Overview of the dataset

# Data Problems

The dataset is quite messy and contains a lot of missing and duplicated values. Moreover, the data is highly unbalanced which can affect the training of the model. The problem associated with the data can be seen in the figures below.

## Missing Data

A complete overview of missing data for all columns is shown in figure 3. It can be seen that there are some columns which have a huge amount of missing data.

Table

Description automatically generated  
Figure 3. Information about missing values

## Duplicate Data

It is interesting to see in figure 4 that there are a lot of duplicate values in the dataset. The data contains 1 million rows however only 626k of them are unique. This shows that 374k observations have been repeated in the dataset.

Graphical user interface, text

Description automatically generated  
Figure 4. Information about duplicate values

## Unbalanced data

Figure 5 shows that the data is highly imbalanced. Some of the categories only have negligible data whereas some of them Have high number of observations.

Chart, bar chart

Description automatically generated  
Figure 5. Information about distribution of data

## Outliers

The figure 6 below shows the distribution of gross income of a household we can see that there are so many outliers present in the data that can affect the machine learning model hence the removal of these outliers is necessary for better results.

Chart

Description automatically generated  
Figure 6. Information about outliers

# Approach Used

## Duplicated Observations

Since we have a lot of rows which have duplicate data hence, we will drop the duplicate values to make the data set unique. we will make sure that we drop the values which were taken earlier and retain the values which are taken later in the data set because they represent the latest data.

## Missing Values

To deal with the missing values we dropped the data which has more than 25% of missing values because it is not efficient to impute them. The values which have small amount of missing data shall be imputed using information of other features.

For columns which have small amount of missing data set we will perform imputation. The method of imputation depends on what variable we are performing the imputation for example find the imputation of household income we will make use of median imputation. Moreover, for categorical values we will make use of model imputation.

## Outlier

The outliers which are unrealistic can be removed as well as can be replaced by values. There are several techniques to replace the outliers for example in case of age we will use mean values to replace the outliers whereas in case of household income we will perform normal distribution which will help us to distribute the outliers.