Customer Segmentation

Team Name: DGI

Team Member Details

No	Name	Email	Country	College/ Company	Specialization
1	Rahma Mahjoub Abker Habeeb	Rahma.mahgoub@gmail.com	Kuwait	Computer Science	Data Science
2	Nonhlanhla L Luphade	nonnynyathi4611@gmail.com	Zimbabwe	University of Cape Town	Data Science
3	Ajaegbu Ebuka Emmanuel	ajaegbu35@gmail.com	Nigeria	Grand Treasury Ltd	Data Science
4	Robin Masawi	Masawirobin69@gmail.com	Zimbabwe	Econet	Data Science

Problem Statement:

The Bank XYZ wants to roll out personalized Christmas offers for certain customers instead of rolling out the same offers for all customers. As an alternative of trying to manually decide which customer is which category. The bank seeks an efficient approach that enables them to uncover hidden patterns in their customer data and categorize customers into a 5 unique groups.

Business Understanding:

- Customer segmentation is the process of categorizing the customers into various groups according to their characteristics or behaviors.
- This will help the companies effectively match their products to the exact customers groups.

Exploratory Data Analysis

Data Cleaning and Transformation:

The data set contained 1000000 observations and 48 columns.

Dealing with duplicates:

- About 37.3% of the data was duplicates, these were removed.
- The data was left with 626159 observations.

Dealing with missing values:

- Variables with more than 80% missing values were removed such as conyuemp and ult_fec_cli_1t. This was done because there is no use imputing 80% of the variables, this reduces the accuracy of our models.
- The customers with no age were also removed because customers with no age only had a customer id and no other data. Reduced the data to 619174 observations.
- For variables like ind_nom_pens_ult1 (73 missing values), ind_nomina_ult1 (73 missing variables), sexo (2 missing variable) and canal_entrada (51 missing variables), the missing values were replaced using the mode. Using the mode and KNNs were considered for imputing the missing categorical variables, however, the mode was the best the data set remained balanced.
- For variables like cod_prov (3682 missing values) and nomprov (missing values), the missing values were only for customers that stay outside Europe. These were replaced by the code 0 and 'NotEuropean' respectively.
- For numeric variables like renta (104731 missing values), the missing values were replaced by the mean gross income for all the customers.

Transforming features:

- The fecha_alta variable which is the date which the customer became the first holder of a contact at the bank was changed by subtracting the customer's date with the max date 2015-01-28 to get the number of years since first contract. This created a new column called fecha_alta_year.
- The 24 variables that start with the letters ind_ and end with the letter _ult1 which describe the different accounts or products the customers use was summed up to one variable which counts the number of products each customer makes use of in the bank. This created a new variable named number_of_accounts.

Dealing with outliers:

- Used the interquartile ranges to determine outliers in numeric variables. (lower quartile – 1.5*interquartile range < x<upper quartile + 1.5 * interquartile range).
- This reduced the dataset to 567592 observations and 46 columns.

Dealing with categorical variables:

The categorical variables can be separated to ordinal variables (some order) and nominal variables (no order). This was done because for some categorical variables order of importance matters. For ordinal variables, ordinal encoding was used. For nominal variables, one hot encoding was the best method for binomial variables because only 2 columns will be produced. For multi-categorical variables, label encoding, one hot encoding and frequency encoding were considered. However, one hot encoding was going to produce multiple columns increasing dimensionality. Label encoding introduces some order to categories that do not have order. Thus, the best method was the frequency encoding which makes use of the frequency of each category in the variable.

Ordinal variables were arranged in the order of importance.

1. Indrel – there is an order 1 is more important than 99.
2. Indrel_1mes – order of importance is 1(primary), 2 (co-owner), 3 (former primary)
3. Tiprel_1mes – order of importance is A (active), P (potential), I (inactive)

Nominal Variables:

1. Multicategory variables: Cod_prov, nomprov, canal_entrada and pais_residencia, these variables have more than 30 categories thus, one hot encoding would not be best. We made use of frequency encoding.
2. Binominal variables: The rest of the nominal variables have 2 categories; thus, we can use one hot encoding.

Variable Selection

- Useless variables dropped: Unnamed 0, tipodom and fecha_dato
- Variables dropped because of missing values: conyuemp and ult_fec_cli_1t
- Dependent variables: nomprov and cod_prov have a relationship and provide the same information thus were nomprov was prov was dropped.
- All variables beginning with 'ind_' and ending with 'ult1' were summed to one variable.

Thus, the final data set contained: 23 columns and 567592 observations.

Modelling and Model Selection

For the task we were given features only and no target variables. The aim of the task is to cluster the customers into 5 groups. Thus, clustering algorithms such as K Means clustering and Hierarchical clustering were considered for this task.

K Means:

Hierarchical Clustering:

Model Selection

The K Means was selected as the best model because it was able to handle a large data set while the hierarchical clustering threw a memory error.