Customer Segmentation Project

Week 8

JULY 28, 2022

Contents

1. Group Information

Group Name**:** M.A.S

Specialization: Data Science

Submitted to: Data Glacier canvas platform

Internship Batch: LISUM10: 30

Group Members	**Three members**		
Name	**Email**	**Country**	**Collage/Company**
Moath Mohammed Bin Musallam	moathmusallam@gmail.com	Saudi Arabia	University of East Anglia
Andrew Kojo Mensah-Onumah	kojoakmo@gmail.com	Ghana	Data Glacier
Shaimaa Saleh Obad Al-khawlani	s.khawlany@gmail.com	Yemen	Data Glacier

2. Problem description

Most banks around the world have variant large customer base with different income levels, ages, characteristics, values and lifestyles.

XYZ bank wants to increase the production and the satisfactions of all customers categories by roll out Christmas offers to their customers.

But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also, they will not be able to uncover the hidden pattern in the data (pattern which group certain kind of customer in one category).

3. Data Understanding

The existing data, which was provided by the bank, is the bank's customers data. However, the data contains many columns that will help the analytics team analyze the data and build a customer segmentation approach for the bank.

Since the data does not contain a dependent variable or (Target), We believe that machine learning (clustering) techniques would be appropriate to use for this type of data.

Size: 1000000 records, 48 columns.

- **Columns Description:**

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

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

The dataset provided was CSV format. The dataset contains 1000000 rows and 48 columns. The datasets mostly contain numerical and categorical data types.

Since the data has no target value, the unsupervised learning (clustering) is the best algorithm to use for this kind of data.

Fewer categorical columns have higher cardinality, i.e, they have more than 10 categories. Most of the categorical columns are binary. Among the numerical features, only the `renta` variable is continuous. The rest are integers. It is important to note that some of the binary categorical columns are of float data type.

Below, we have attached snapshots of the datasets and its data types.

```
custSeg_ds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 47 columns):
 #   Column                                                  Non-Null Count    Dtype
---  ------                                                  --------------    -----
 0   data_date                                               1000000 non-null  object
 1   customer_code                                           1000000 non-null  int64
 2   employee_index                                          989218 non-null   object
 3   customer_country_residence                              989218 non-null   object
 4   customer_gender                                         989214 non-null   object
 5   age                                                     1000000 non-null  object
 6   bank_entry_date                                         989218 non-null   object
 7   new_customer_index                                      989218 non-null   float64
 8   customer_seniority                                      1000000 non-null  object
 9   first/primary_customer                                  989218 non-null   float64
 10  last_date_as_primary_customer                           1101 non-null     object
 11  customer_type_at_the_ beginning_of_the_month            989218 non-null   float64
 12  customer_relation_type_at_the_beginning_of_the_ month   989218 non-null   object
 13  residence_index                                         989218 non-null   object
 14  foreign_index                                           989218 non-null   object
 15  spouse_index                                            178 non-null      object
 16  type_of_channel                                         989139 non-null   object
 17  deceased_index_(N/S)                                    989218 non-null   object
 18  addres_type                                             989218 non-null   float64
 19  province_code                                           982266 non-null   float64
 20  province_name                                           982266 non-null   object
 21  activity_index                                          989218 non-null   float64
 22  gross_income_of_the_ household                          824817 non-null   float64
 23  saving_account                                          1000000 non-null  int64
 24  guarantees                                              1000000 non-null  int64
 25  current_account                                         1000000 non-null  int64
 26  derivative_account                                      1000000 non-null  int64
 27  payroll_account                                         1000000 non-null  int64
 28  junior_account                                          1000000 non-null  int64
 29  mas_particular_account                                  1000000 non-null  int64
 30  particular_account                                      1000000 non-null  int64
 31  particular_plus_account                                 1000000 non-null  int64
 32  short_term_deposits                                     1000000 non-null  int64
 33  medium_term_deposits                                    1000000 non-null  int64
 34  long_term_deposits                                      1000000 non-null  int64
 35  e-account                                               1000000 non-null  int64
 36  funds                                                   1000000 non-null  int64
 37  mortgage                                                1000000 non-null  int64
 38  pensions                                                1000000 non-null  int64
 39  loans                                                   1000000 non-null  int64
 40  taxes                                                   1000000 non-null  int64
 41  credit_card                                             1000000 non-null  int64
 42  securities                                              1000000 non-null  int64
 43  home_account                                            1000000 non-null  int64
 44  payroll                                                 994598 non-null   float64
 45  pensions                                                994598 non-null   float64
 46  direct_debit                                            1000000 non-null  int64
dtypes: float64(9), int64(23), object(15)
memory usage: 358.6+ MB
```

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

We started initial analysis and we can say that the dataset has some following problems:

1. Missing/null data:

```
draw_missing_data_table(custSeg_ds)
```

56]:

	Total	Percent
spouse_index	999822	0.999822
last_date_as_primary_customer	998899	0.998899
gross_income_of_the_ household	175183	0.175183
province_name	17734	0.017734
province_code	17734	0.017734
type_of_channel	10861	0.010861
customer_gender	10786	0.010786
customer_relation_type_at_the_beginning_of_the_ month	10782	0.010782
activity_index	10782	0.010782
addres_type	10782	0.010782
deceased_index_(N/S)	10782	0.010782
foreign_index	10782	0.010782
residence_index	10782	0.010782
customer_type_at_the_ beginning_of_the_month	10782	0.010782
first/primary_customer	10782	0.010782
new_customer_index	10782	0.010782
bank_entry_date	10782	0.010782
customer_country_residence	10782	0.010782
employee_index	10782	0.010782
payroll	5402	0.005402
pensions	5402	0.005402
loans	0	0.000000
e-account	0	0.000000
funds	0	0.000000
mortgage	0	0.000000
pensions	0	0.000000
data_date	0	0.000000
taxes	0	0.000000
credit_card	0	0.000000
securities	0	0.000000
home_account	0	0.000000
medium_term_deposits	0	0.000000
long_term_deposits	0	0.000000
saving_account	0	0.000000
short_term_deposits	0	0.000000
particular_plus_account	0	0.000000
particular_account	0	0.000000
mas_particular_account	0	0.000000
junior_account	0	0.000000
payroll_account	0	0.000000
derivative_account	0	0.000000
current_account	0	0.000000
guarantees	0	0.000000
customer_code	0	0.000000
customer_seniority	0	0.000000
age	0	0.000000
direct_debit	0	0.000000

```
custSeg_ds.isnull().sum().sum()
```

90]:

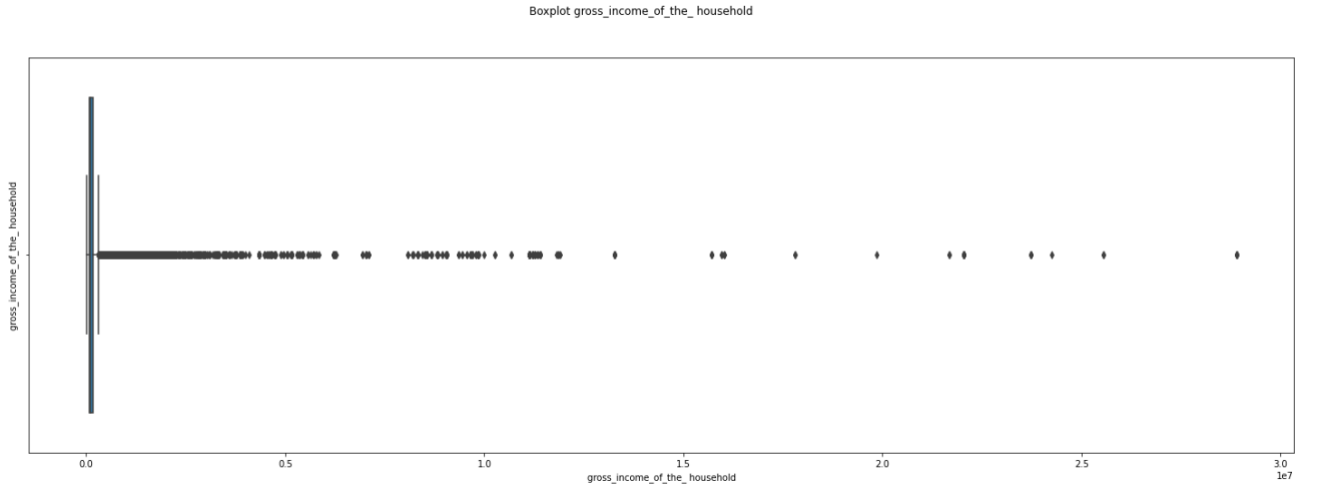
```
2371207
```

From the above analysis we found that the columns have 2371207 missing data, and the columns are listed above.

2. Outliers

```
fig,axes = plt.subplots(figsize=(25, 8), sharey=True)
fig.suptitle('Boxplot gross_income_of_the_ household')
sns.boxplot( x='gross_income_of_the_ household' , data=custSeg_ds ).set_ylabel("gross_income_of_the_ household")
```

```
: Text(0, 0.5, 'gross_income_of_the_ household')
```



As per the above there are significant outliers in “gross_income_of_the_ household” , Upper and Lower Limits of gross_income_of_the_ household is:
[-66219.10500000001, 301223.41500000004]

6. Approaches to apply to deal with identified problems

- Columns “spouse index” and “last_date_as_primary_customer” will be dropped because 99.8 percent of their values are missing.
- Imputing missing date values by either using interpolation, backfill or forward fill methods.
- Replacing some missing numerical values by mean, mode or median methods. Also, using KNN imputation technique and other approaches for numerical variables.
- Replacing missing categorical variables with mode and/or by using unsupervised techniques.