TEAM 7 SANTANDER PRODUCT RECOMMENDATION SYSTEM

ZEPPELIN: EXPLORATORY ANNAYSIS

CSYE 7200 Big Data Engineering using Scala

Northeastern University | Spring 2018

Exploratory Analysis

1. Total number of rows in raw data: **13647309**

2. Renamed all Spanish columns names to English names and then took null count for each column. Below table describes information about the same:

Column Name [English]	Column Name [Original]	Null Values	Description
Partitioned_Date	fecha_dato	0	The table is partitioned for this column
Cust_Code	Ncodpers	0	Customer code
Emp_Index	ind_empleado	27734	Employee index: A active B ex employed F Filial N not employee P passive
Cust_Residence	pais_residencia	27734	Customer's Country residence
Cust_Gender	Sexo	27804	Customer's sex
Cust_Age	Age	0	Age
First_Holder_Date	fecha_alta	27734	The date in which the customer became as the first holder of a contract in the bank
New_Cust_Index	ind_nuevo	0	New customer Index. 1 if the customer registered in the last 6 months.
Cust_Seniority	Antiguedad	0	Customer seniority [in months]
Cust_Index_Primary	Indrel	0	1 [First/Primary] 99 [Primary customer during the month but not at the end of the month]
Last_Date_Primary_Cust	ult_fec_cli_1t	13622516	Last date as primary customer [if he isn't at the end of the month]
Cust_Type	indrel_1mes	149781	Customer type at the beginning of the month 1 [First/Primary customer] 2 [co-owner] P [Potential]3 [former primary] 4[former co-owner]
Cust_Relation_Type	tiprel_1mes	149781	Customer relation type at the beginning of the month A [active] I

			[inactive] P [former customer] R [Potential]
Residence_Index	Indresi	27734	Residence index [S [Yes] or N [No] if the residence country is the same than the bank country]
Foreigner_Index	Indext	27734	Foreigner index [S [Yes] or N [No] if the customer's birth country is different than the bank country]
Spouse_Index	Conyuemp	13645501	Spouse index. 1 if the customer is spouse of an employee
Channel	canal_entrada	186126	channel used by the customer to join
Deceased_Index	Indfall	27734	Deceased index. N/S
Addres_Type	Tipodom	0	Address type. 1 primary address
Province_Code	cod_prov	0	Province code [customer's address]
Province_Name	Nomprov	93591	Province name
Activity_Index	ind_actividad_cliente	0	Activity index [1 active customer; 0 inactive customer]
Gross_Income	Renta	2794375	Gross income of the household
Cust_Identification	Segmento	189368	segmentation: 01 - VIP 02 - Individuals 03 - college graduated
Saving_Acc	ind_ahor_fin_ult1	0	Saving Account
Guarantees	ind_aval_fin_ult1	0	Guarantees
Current_Acc	ind_cco_fin_ult1	0	Current Accounts
Derivada_Acc	ind_cder_fin_ult1	0	Derivada Account
Payroll_Acc	ind_cno_fin_ult1	0	Payroll Account
Junior_Acc	ind_ctju_fin_ult1	0	Junior Account
Mas_Acc	ind_ctma_fin_ult1	0	Más particular Account
Particular_Acc	ind_ctop_fin_ult1	0	particular Account

Particular_Plus_Acc	ind_ctpp_fin_ult1	0	particular Plus Account
Short_Term_Deposit	ind_deco_fin_ult1	0	Short-term deposits
Medium_Term_Deposits	ind_deme_fin_ult1	0	Medium-term deposits
Long_Term_Deposits	ind_dela_fin_ult1	0	Long-term deposits
e_Acc	ind_ecue_fin_ult1	0	e-account
Funds	ind_fond_fin_ult1	0	Funds
Mortgage	ind_hip_fin_ult1	0	Mortgage
Pensions	ind_plan_fin_ult1	0	Pensions
Loans	ind_pres_fin_ult1	0	Loans
Taxes	ind_reca_fin_ult1	0	Taxes
Credit_Card	ind_tjcr_fin_ult1	0	Credit Card
Securities	ind_valo_fin_ult1	0	Securities
Home_Acc	ind_viv_fin_ult1	0	Home Account
Payroll	ind_nomina_ult1	0	Payroll
Nom_Pensions	ind_nom_pens_ult1	0	Pensions
Direct_Debit	ind_recibo_ult1	0	Direct Debit

3. Datatype for each column variable:

Column name [English]	Column name [Original]	DataType
Partitioned_Date	fecha_dato	timestamp [nullable = true]
Cust_Code	Ncodpers	double [nullable = true]
Emp_Index	ind_empleado	string [nullable = true]
Cust_Residence	pais_residencia	string [nullable = true]
Cust_Gender	Sexo	string [nullable = true]
Cust_Age	Age	string [nullable = true]

First_Holder_Date	fecha_alta	timestamp [nullable = true]
New_Cust_Index	ind_nuevo	string [nullable = true]
Cust_Seniority	Antiguedad	string [nullable = true]
Cust_Index_Primary	Indrel	string [nullable = true]
Last_Date_Primary_Cust	ult_fec_cli_1t	timestamp [nullable = true]
Cust_Type	indrel_1mes	string [nullable = true]
Cust_Relation_Type	tiprel_1mes	string [nullable = true]
Residence_Index	Indresi	string [nullable = true]
Foreigner_Index	Indext	string [nullable = true]
Spouse_Index	Conyuemp	string [nullable = true]
Channel	canal_entrada	string [nullable = true]
Deceased_Index	Indfall	string [nullable = true]
Addres_Type	Tipodom	string [nullable = true]
Province_Code	cod_prov	string [nullable = true]
Province_Name	Nomprov	string [nullable = true]
Activity_Index	ind_actividad_cliente	string [nullable = true]
Gross_Income	Renta	double [nullable = true]
Cust_Identification	Segment	string [nullable = true]
Saving_Acc	ind_ahor_fin_ult1	integer [nullable = true]
Guarantees	ind_aval_fin_ult1	integer [nullable = true]
Current_Acc	ind_cco_fin_ult1	integer [nullable = true]
Derivada_Acc	ind_cder_fin_ult1	integer [nullable = true]

Payroll_Acc	ind_cno_fin_ult1	integer [nullable = true]
Junior_Acc	ind_ctju_fin_ult1	integer [nullable = true]
Mas_Acc	ind_ctma_fin_ult1	integer [nullable = true]
Particular_Acc	ind_ctop_fin_ult1	integer [nullable = true]
Particular_Plus_Acc	ind_ctpp_fin_ult1	integer [nullable = true]
Short_Term_Deposit	ind_deco_fin_ult1	integer [nullable = true]
Medium_Term_Deposits	ind_deme_fin_ult1	integer [nullable = true]
Long_Term_Deposits	ind_dela_fin_ult1	integer [nullable = true]
e_Acc	ind_ecue_fin_ult1	integer [nullable = true]
Funds	ind_fond_fin_ult1	integer [nullable = true]
Mortgage	ind_hip_fin_ult1	integer [nullable = true]
Pensions	ind_plan_fin_ult1	integer [nullable = true]
Loans	ind_pres_fin_ult1	integer [nullable = true]
Taxes	ind_reca_fin_ult1	integer [nullable = true]
Credit_Card	ind_tjcr_fin_ult1	integer [nullable = true]
Securities	ind_valo_fin_ult1	integer [nullable = true]
Home_Acc	ind_viv_fin_ult1	integer [nullable = true]
Payroll	ind_nomina_ult1	string [nullable = true]
Nom_Pensions	ind_nom_pens_ult1	string [nullable = true]
Direct_Debit	ind_recibo_ult1	integer [nullable = true]

4. Detail Exploratory Analysis:

- Number of Female Customers [46%] are less than Number of Male customers [54%]
- Last Date being Primary Customer the max value is 2016-05-30

- Individual Customers [58%] are more than VIP [6%] and Students [36%] of total number of customers
- Customer type has values of data type String Double and Integer
- Average gross income of VIP customers is more than Individual and Student customers
- Average gross income of female is lesser than male
- Male with Individual customer types are maximum
- Count of Guarantees is maximum for male individual type customers with average gross income approx. 139593
- Number of savings account maximum for male individual type of customers with average gross income 139593
- Number of derivada account maximum for male individual type of customers with average gross income 139593
- Further analysis shows that all the products are consumed maximum by male individual type of customer
- There are more university students as customer than other types [Individual and VIP]

Approach to follow towards data cleaning activities:

Total 15 columns have missing [null] values, also for few columns need to check for datatypes. The approach we will follow is as below:

1. **Age [Cust_Age]:**
   Null value count = 0
   Step 1 - Remove outliers firsts.
       - Divide age in to sections replace with mean values. This will help us fix the age distribution.
          mean [age >=18 and <=30] - Replace rows with age < 18
          mean [age >= 30 and age <= 100] - Replace rows with age > 100
    Step 2 - Fix data type
       - round[age] to convert it into integer from float

2. **fecha_alta [First_Holder_Date]**
    Null value count - 27734
    Step – replace null values with median [middle value] of fetcha_alta column

3. **nomprov [Province_Name]**
    Null value count - 93591
    Step - replace NULL values with Unknown

4. **sexo [Cust_Gender]**
    Null value count - 27804
    Step - we will replace the NULL values with ratio

5. **indfall [Deceased_Index]**
 Null value count - 27734
 Step - we will replace NULL values with ratio

6. **indresi [Residence_Index]**
 Null value count - 27734
 Step - we will replace NULL values with ratio

7. **ind_empleado [Emp_Index]**
 Null value count - 27734
 Step - we will replace NULL values with ratio

8. **indrel_1mes [Cust_Type]**
 Null value count - 149781
 Step - we will replace NULL values with ratio

9. **tiprel_1mes [Cust_Relation_type]**
 Null value count - 149781
 Step - we will replace NULL values with ratio

10. **indext [Foreigner_Index]**
 Null value count - 27734
 Step - we will replace NULL values with ratio

11. **canal_entrada [Channel]**
 Null value count - 186126
 Step - we will replace NULL values with ratio

12. **renta [Gross_Income]**
 Null value count - 2794375
 Step - we will replace NULL values with average income of the customers in same province

13. **segmento [Cust_Identification]**
 Null value count - 189368
 Step - we will replace the NULL values with Unknown

14. **Columns to remove**
    - **tipodom** [Address_Type]: not useful
    - **cod_prov** [Province_Code]: we already have this information in nomprov [Province_Name], so can be removed
    - **conyuemp** [Spouse_Index]: approx. 99% values are null, so can be removed
    - **ult_fec_cli_lt** [Last_Date_Primary_Cust]: approx. 99% values are null, so can be removed