DATA GLACIER VIRTUAL INTERNSHIP

CROSS SELLING RECOMMENDATION-GROUP PROJECT

WEEK 8: DELIVERABLES

GROUP NAME: HEGY

Team members:

Name: B. Harika
Email: harikabreddy444@gmail.com
Country: India
College/ Company: Data Glacier
Specialization: Data Analyst

Name: Yusuf Yuhan
Email: yusufyuhan98.yy@gmail.com
Country: Srilanka
College/ Company: The Open University of Srilanka
Specialization: Data Analyst

Name: Ebaghae Imhanlahimi
Email: imhanlahimiw@gmail.com
Country: America
College/ Company: Data Glacier
Specialization: Data Analyst

Name: Gladys Kalas
Email: gladys@kalas.me
Country: USA
College/ Company: Data Glacier
Specialization: Data Analyst

Contents

Problem description:

XYZ Credit Union is a financial institution based in Latin America that offers a variety of banking products to its customers, including credit cards, deposit accounts, retirement accounts, and safe deposit boxes. While the credit union has been successful in selling these products individually, it has not been as successful in cross-selling its products to existing customers.

The lack of success in cross-selling suggests that there may be several barriers preventing XYZ Credit Union from selling additional products to its existing customers. To address this problem, XYZ Credit Union has decided to work with ABC Analytics, a data analytics consulting firm, to identify the barriers to cross-selling and develop strategies to overcome them.

ABC Analytics will work with the credit union to analyze Customer data and information to identify patterns and trends, and develop targeted marketing strategies that are designed to increase their possibilities and revenues in the credit union's quest to cross sell banking products to the customers.

Data understanding:

The data available for analysis was obtained from the data bank of XYZ credit union. It contains information about XYZ bank customers and the financial product holdings that XYZ offers to its customers.

The data for cross-selling recommendation is a large csv file, with file size on disk of 2.13 GB.

Upon primary understanding the features appear in Spanish literacy. It comprises of 48 features and 13647309 observations (The feature names are changed to English for better understanding). The dataset contains both numerical and categorical variables.

- Data contains demographic characteristics of the customers:

  - Age
  - Address
  - Income
  - Etc.
  - Gender
  - Nationality
  - Life status

Financial products offered by the bank:

- ind_ahor_fin_ult1/ Saving Account
- ind_cco_fin_ult1/Current Accounts
- ind_ctju_fin_ult1 / Junior Account
- ind_deme_fin_ult1/Medium-term deposits
- ind_ecue_fin_ult1 / e-account
- ind_hip_fin_ult1 / Mortgage
- ind_pres_fin_ult1 / Loans
- ind_tjcr_fin_ult1 / Credit Card
- ind_viv_fin_ult1 / Home Account
- ind_nom_pens_ult1 / Pensions
- ind_aval_fin_ult1 / Guarantees
- ind_cno_fin_ult1/Payroll Account
- ind_deco_fin_ult1/Short-term deposits
- ind_dela_fin_ult1 / Long-term deposits
- ind_fond_fin_ult1 / Funds
- ind_plan_fin_ult1 / Pensions
- ind_reca_fin_ult1 / Taxes
- ind_valo_fin_ult1 / Securities
- ind_nomina_ult1 / Payroll

Type of data:

File name	Train.csv
No. of observations	13647309
No. features	48
File Size	2.13 GB
File type	CSV
No. of files	1

Column_Names	**Data Types**
fecha_dato/ Date	Object
ncodpers/ Customer_code	Int64
ind_empleado/ Employee_index	Object
pais_residencia/ Country	Object
Sexo/ Gender	Object
age/ Age	Object
fecha_alta/ Customer_join_date	Object
ind_nuevo/ Customer_index	Float64
antiguedad/ Customer_senoirity	Object
indrel/ primary_customer	Float64
ult_fec_cli_1t/ Customer_leave_date	Object
indrel_1mes/ Customer_type	Object
tiprel_1mes/ Customer_relation	Object
indresi/ Residence_index	Object
indext/ Foreign_index	Object
conyuemp/ Spouse_index	Object
canal_entrada/ Channel	Object
indfall/ Deceased_index	Object
tipodom/ Primary_address	Float64
cod_prov/ Customer_address	Float64
nomprov/ province_name	Object
ind_actividad_cliente/ Activity_index	Float64
renta/ Gross_income	Float64
segmento/ Segmentation	Object
ind_ahor_fin_ult1/ Saving_account	Int64
ind_aval_fin_ult1/ Guarantees	Int64
ind_cco_fin_ult1/ Current_account	Int64
ind_cder_fin_ult1/ Derivative_account	Int64
ind_cno_fin_ult1/ Payroll_account	Int64
ind_ctju_fin_ult1/ Junior_account	Int64
ind_ctma_fin_ult1/ More_private_account	Int64
ind_ctop_fin_ult1/ Private_account	Int64
ind_ctpp_fin_ult1/ Private_plus_account	Int64
ind_deco_fin_ult1/ Short_term_deposits	Int64
ind_deme_fin_ult1/ Medium_term_deposits	Int64
ind_dela_fin_ult1/ Long_term_deposits	Int64
ind_ecue_fin_ult1/ E_account	Int64
ind_fond_fin_ult1/ Funds	Int64
ind_hip_fin_ult1/ Mortgage	Int64
ind_plan_fin_ult1/ Pensions	Int64
ind_pres_fin_ult1/ Loans	Int64
ind_reca_fin_ult1/ Taxes	Int64
ind_tjcr_fin_ult1/ Credit_card	Int64
ind_dela_fin_ult1/ Securities	Int64
ind_dela_fin_ult1/ Home_account	Int64
ind_dela_fin_ult1/ Payroll	Float64
ind_dela_fin_ult1/ Pensions_2	Int64
ind_dela_fin_ult1/ Direct_debit	Int64

Problems and solutions for the Data:

S/N	Problem	Proposed Solution	Reason
1	Column names recorded in Spanish	Rename column names in English interpretation	For ease in understanding and analyzing the data.
2	Column data types interpreted wrongly	Convert column data types	To ensure accuracy in analysis and improve memory efficiency
3	Some Column information recorded in Spanish	Replace records with English interpretation	For ease in understanding and analyzing the data.
4	Employee_index / ind_empleado has 27734 null values	Values to be deleted.	It contains categorical data and requires further information from the company.
5	Country / pais_residencia has 27734 null values	Values to be deleted	It contains demographic data and requires more information from the company.
6	Gender / sexo has 27804 null values	Values to be deleted	It contains demographic data and requires accurate information from the company.
7	Customer_join_date / fecha_alta has 27734 null values	Value to be imputed.	Values can be imputed based on existing records.
8	Customer_index / ind_nuevo has 27734 null values	Value to be imputed	Values can be imputed based on existing records.
9	Primary_customer / indrel has 27734 null values	Value to be imputed	Values can be imputed based on existing records.
10	Customer_leave_date / ult_fec_cli_1t has 13622516 null values	Value to be imputed	Values can be imputed based on existing records.
11	Customer_type / indrel_1mes has 149781 null values	To be deleted	It contains categorical data and requires accurate information from the company.
12	Customer_relation / tiprel_1mes has 14781 null values	Values to be deleted	It contains categorical data and requires accurate information from the company.

13	Residence_index / indresi has 27734 null values	Values to be deleted	It contains demographic data and requires accurate information from the company.
14	Foreigner_index / indext has 27734 null values	Values to be deleted	It contains demographic data and requires accurate information from the company.
15	Spouse_index / conyuemp has 13645501 null values	Values to be imputed	Values can be imputed based on existing records.
16	Channel / canal_entrada has 186126 null values	Values to be deleted	It contains categorical data and requires accurate information from the company.
17	Deceased_index / indfall has 27734 null values	Values to be deleted	It contains demographic data and requires accurate information from the company.
18	Primary_address / tipodom has 27735 null values	Value to be imputed	It contains demographic data and requires accurate information from the company.
19	Customer_address / Cod_prov has 93591 null values	Values to be deleted	It contains demographic data and requires accurate information from the company.
20	Province_name / nomprov has 93591 null values	Values to be deleted	It contains demographic data and requires accurate information from the company.
21	Activity_index / ind_actividad_cliente has 27734 null values	Values to be imputed	Values can be imputed based on existing records.
22	Gross_income / renta has 2794375 null values	Values to be imputed	Values can be imputed based on existing records.
23	Segmentation / segmento has 189368 null values	Values to be imputed	Values can be imputed based on existing records.
24	Payroll / ind_nomina_ult1 has 16063 null values	Values to be imputed	Values can be imputed based on existing records.
25	Pensions_2 / ind_nom_pens_ult1 has 16063 null values	Values to be imputed	Values can be imputed based on existing records.

Data Cleaning:

- Several missing values have been dropped from the variables.
- Column names are translated and renamed accordingly.
- The mean, mode, median, and zeroes are used to impute null values.
- Columns like gender, residence index, spouse index, customer relations, employee index, etc.; variables are assigned to their respective categories.
- Outliers are detected using different methods and treated accordingly.

GitHuB Repo Link:
https://github.com/HarikaReddyB/Cross_selling_recommendation---Group_Project/tree/main/Week%208