DATA ANALYST: CROSS SELLING RECOMMENDATION FINAL PROJECT

TEAM MEMBER’S DETAILS

Group Name: Individual
Name: Ian Kihara Wangui
Email: eandavid6@gmail.com
Company: DataGlacier
Specialization: Data Analyst

PROBLEM DESCRIPTION

XYZ credit union in Latin America is performing very well in selling the Banking products (eg: Credit card, deposit account, retirement account, safe deposit box etc) but their existing customer is not buying more than 1 product which means bank is not performing good in cross selling (Bank is not able to sell their other offerings to existing customer). XYZ Credit Union decided to approach ABC analytics to solve their problem. Can you tell us how this can be solved?
My role as a data analyst is to inspect the data and suggest what action bank can take to increase cross selling (without using ML)

DATA UNDERSTANDING

DATA SOURCES

We’ve been provided with data in a zip archive that had been uploaded to google drive. The zip archive contains two csv files; `Train.csv` and `Test.csv’. We are going to use Train.csv for purposes of data analysis because it contains all the data features. It has 13.65 million rows of data and 48 columns. Test.csv on the other hand has 0.93 million rows of data and 24 columns.

TYPE OF DATA

#	**Column Name**	**Description**
0	fecha_dato	The table is partitioned for this column

#	Column Name	Description
1	ncodpers	Customer code
2	ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
3	pais_residencia	Customer's Country residence
4	sexo	Customer's sex
5	age	Age
6	fecha_alta	The date in which the customer became as the first holder of a contract in the b
7	ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
8	antiguedad	Customer seniority (in months)
9	indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end o
10	ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)
11	indrel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co co-owner)
12	tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P
13	indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the
14	indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different th
15	conyuemp	Spouse index. 1 if the customer is spouse of an employee
16	canal_entrada	channel used by the customer to join
17	indfall	Deceased index. N/S
18	tipodom	Addres type. 1, primary address
19	cod_prov	Province code (customer's address)
20	nomprov	Province name
21	ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
22	renta	Gross income of the household
23	segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated

#	Column Name	Description
24	ind_ahor_fin_ult1	Saving Account
25	ind_aval_fin_ult1	Guarantees
26	ind_cco_fin_ult1	Current Accounts
27	ind_cder_fin_ult1	Derivada Account
28	ind_cno_fin_ult1	Payroll Account
29	ind_ctju_fin_ult1	Junior Account
30	ind_ctma_fin_ult1	Más particular Account
31	ind_ctop_fin_ult1	particular Account
32	ind_ctpp_fin_ult1	particular Plus Account
33	ind_deco_fin_ult1	Short-term deposits
34	ind_deme_fin_ult1	Medium-term deposits
35	ind_dela_fin_ult1	Long-term deposits
36	ind_ecue_fin_ult1	e-account
37	ind_fond_fin_ult1	Funds
38	ind_hip_fin_ult1	Mortgage
39	ind_plan_fin_ult1	Pensions
40	ind_pres_fin_ult1	Loans
41	ind_reca_fin_ult1	Taxes
42	ind_tjcr_fin_ult1	Credit Card
43	ind_valo_fin_ult1	Securities
44	ind_viv_fin_ult1	Home Account
45	ind_nomina_ult1	Payroll
46	ind_nom_pens_ult1	Pensions
47	ind_recibo_ult1	Direct Debit

- The data in columns 0-24 are mostly categorical variables, and some date variables. They are either discrete or nominal variables, and dimension data that describe customers' demographics, account information, and transaction records.
- Column 5 "age" is categorical, but it can be converted into continuous or ordinal data.
- Columns 25-47 are mostly binary variables indicating whether a customer has a particular financial product or not. They are discrete, and measures data.

PROBLEMS IN THE DATA AND PROPOSED SOLUTIONS

1. The first is that the column names are in Spanish. This makes it hard to interpret the data and derive insights unless you are competent in Spanish. I will rename the columns to approximate English names to ease data interpretation.

2. The dataset is very large and exceeds my current computing capabilities. The ‘Train.csv’ dataset is 2.29GB and has over 13.65 million records and 48 columns. This considerably slows down the computer and makes it hard to manipulate the data. Instead, I have decided to take a simple random sample of the data of about 10% of the original dataset. This is manageable with the computing power I have available.

3. The data also has presence of missing values. Missing values are a problem because they can stop certain python functions from running and therefore inhibit data analysis. I am going first to drop columns with a high percentage of missing values (>95%) and subsequently drop the rows with missing values too from the remaining columns.

4. Outlier detection. The dataset also has outliers especially on the certain columns. It is important to exercise judgement and consult domain experts because this can skew the results of your data analysis. Therefore, I will drop outliers only on a case-by-case basis. The columns with the highest percentage of outliers are `"gross_income_household"` (7.68%), `"customer_age"` (2.22%), and `"customer_seniority"` (1.98%).

5. Skewness. Due to the nature of the data, after conducting a thorough analysis, we did identify that the dataset contains skewed data. The skewness of the data will affect how we interpret the data. Such as is a data distribution right-skewed or left-skewed or symmetrical. In the context of this specific dataset, the histograms with the highest degree of skewness are `"gross_income_household"` (right skewed) and `"customer_age"` (right skewed).

Github Repo link :
https://github.com/Iandavidk/DataGlacier/tree/main/Week-8-Cross_Selling_Data_Understanding