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 CLEANSING AND TRANSFORMATION

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 renamed the columns to approximate English names to ease data interpretation.

2. The dataset is very large and exceeds my current computing capabilities. Instead, I have decided to take a simple random sample of the data of about 10% of the original dataset about 1.365 million rows of data. This is manageable with the computing power I have available. I used a random state equal to 42 to ensure reproducibility.

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. The columns that were dropped because of a high percentage of missing values are 'last_date_primary' and 'spouse_index' columns which had greater than 99% missing values. The number of rows dropped from the dataset were 285,013 which is approximately 20.88% of the original dataset.

4. Converted columns to appropriate data types. I cast some columns listed below into new data types to better match the data type that they contained.

   'customer_age': 'Int64',
   'customer_seniority': 'Int64',
   'date_partition': 'datetime64',
   'customer_join_date': 'datetime64',
   'province_code': 'Int64',
   'gross_income_household': 'float'

5. Replace irrelevant data. Replace all the values on the ‘customer_seniority’ column that were negative with a zero value. This is because you can’t have customers with less than zero seniority.

6. Feature engineering. Created a new feature called ‘total_products’ that shows the total number of products per customer.

Github Repo link :
https://github.com/Iandavidk/DataGlacier/tree/main/Week-9-Cross_Selling_Data_Cleaning