Week 8 Deliverables

Group Name: Repeat Sales
Group Member #1: Griffin Palfrey
Email: griffinpalfrey@gmail.com
Country: Canada
College: Columbia University
Specialization: Data Analyst
Group Member #2: Phanindra Kumar
Email: phanindrakumar1998@gmail.com
Country: United States of America
College/Company: University of Maryland, Baltimore County
Specialization: Data Analyst

Problem Description

XYZ Credit Union in Latin America are selling business products such as Credit Cards, Deposit Accounts, and Retirement Accounts, but they are not performing well in cross selling: persuading customers to acquire more than one of their banking products. This is resulting in reduced customer satisfaction, loyalty, and opportunities for revenue growth.

Impact: Reduced revenue from business products, missed opportunities for enhanced client engagement, and potential loss of market share to competitors who offer more comprehensive solutions.

Objectives: To enhance the effectiveness of cross-selling business products, thereby increasing client uptake and revenue.

Data Understanding

The datasets used for this analysis are sourced from canvas and are intended to explore a company's strategy for improving their cross sales. It includes data that breaks down each customer based on attributes relating to age, sex, income, history with the company, as well as many other traits.

There are two datasets titled 'train' and 'test'. The 'train' dataset comprises 13,647,310 rows and 48 columns, while the test dataset contains 929,616 rows and 24 columns. The columns for both were originally written in spanish, which we have translated and described below:

For 'train' dataset:

- 'date': the date of recorded input
- 'customer_code': the customer's code

- ‘employee_index’ : Employee index: A active, B ex employed, F filial, N not employee, P passive
- ‘country_of_residence’ : the customer’s country of residence
- ‘sex’ : the customer’s sex
- ‘age’: the customer’s age
- ‘holder_start_date’ : The date in which the customer became as the first holder of a contract in the bank
- ‘new_customer_index’ : New customer Index. 1 if the customer registered in the last 6 months.
- ‘seniority’ : Customer seniority (in months)
- ‘primary’: 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
- ‘last_date_as_primary’: Last date as primary customer (if he isn't at the end of the month)
- ‘customer_type_at_beginning_of_month’ : Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
- ‘customer_relation_type_at_beginning_of_month’ : Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
- ‘residence_index’ : Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
- ‘foreigner_index’ : Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
- ‘spouse_index’: 1 if the customer is spouse of an employee
- ‘channel’ : channel used by the customer to join
- ‘deceased_index’ : Deceased index. N/S
- ‘address_type’ : Address type. 1, primary address
- ‘province_code’: Province code (customer's address)
- ‘province_name’: Province name
- ‘activity_index’: Activity index (1, active customer; 0, inactive customer)
- ‘gross_income’: Gross income of the household
- ‘segmentation’: 01 - VIP, 02 - Individuals 03 - college graduated
- ‘saving_account’: Saving Account
- ‘guarantees’: Guarantees
- ‘current_accounts’: Current Accounts
- ‘derivada_accont’: Derivada Accounts
- ‘payroll_account’: Payroll Accounts
- ‘junior_account’: Junior Accounts
- ‘mas_particular_account’: Mas particular account

- ‘particular_account’: Particular Account
- ‘particular_plus_account’: Particular plus account
- ‘short_term_deposits’: short term deposits
- ‘medium_term_deposits’: medium term deposits
- ‘long_term_deposits’: long term deposits
- ‘e_account’: e-account
- ‘funds’: funds
- ‘mortgage’: mortgage
- ‘pensions’: pensions
- ‘loans’: loans
- ‘credit_card’: credit card
- ‘securities’: securities
- ‘home_account’: home account
- ‘payroll’: payroll
- ‘pensions’: pensions
- ‘direct_debit’: direct debit

For ‘test’ dataset:

- ‘date’: the date of recorded input
- ‘customer_code’: the customer’s code
- ‘employee_index’ : Employee index: A active, B ex employed, F filial, N not employee, P passive
- ‘country_of_residence’ : the customer’s country of residence
- ‘sex’ : the customer’s sex
- ‘age’: the customer’s age
- ‘holder_start_date’ : The date in which the customer became as the first holder of a contract in the bank
- ‘new_customer_index’ : New customer Index. 1 if the customer registered in the last 6 months.
- ‘seniority’ : Customer seniority (in months)
- ‘primary’: 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
- ‘last_date_as_primary’: Last date as primary customer (if he isn't at the end of the month)
- ‘customer_type_at_beginning_of_month’ : Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
- ‘customer_relation_type_at_beginning_of_month’ : Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)

- ‘residence_index’ : Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
- ‘foreigner_index’ : Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
- ‘spouse_index’: 1 if the customer is spouse of an employee
- ‘channel’ : channel used by the customer to join
- ‘deceased_index’ : Deceased index. N/S
- ‘address_type’ : Address type. 1, primary address
- ‘province_code’: Province code (customer's address)
- ‘province_name’: Province name
- ‘activity_index’: Activity index (1, active customer; 0, inactive customer)
- ‘gross_income’: Gross income of the household
- ‘segmentation’: 01 - VIP, 02 - Individuals 03 - college graduated

Data Problems

1. Mixed Data Types

There are several columns–namely ‘age’, ‘seniority’, ‘customer_type_at_beginning_of_month’, and ‘foreigner_index’– that contain mixed data types, which increases the difficulty of running and interpreting the information in each column.

2. Missing/Null values

Each dataset contains a significant number of null values. This is especially apparent in columns such as ‘last_date_as_primary’ or ‘spouse_index’, which individually contain over 13 million missing values, according to our count.

3. Outliers

Using both a z-score and IQR test, we have found 6,101,444 supposed ‘outliers’ in our data. This refers to any observation that lies an abnormal distance from the rest of the values.

4. Skewness

Certain columns report high skewness, both in the positive and negative direction. This means that the data is stacked more heavily on one side than the other, highlighting disparities in our demographics, and possibly alluding to more general concerns in access, inclusiveness, and equity. While there is little we can do to proactively manipulate these numbers, we must take them into account while conducting our exploratory data analysis and presenting our findings.

Approach

- For mixed data types, sort through the affected columns and make sure they contain strictly the data types that are expected. If they do not, either convert them to the same type or omit if they are unable to do so. We want to do this to make our data more accessible and readable, and to ensure that we can process each column without any difficulties or errors.
- For missing or null values, first check all the columns that are affected to ensure they have been properly calculated. If there is a mistake, try to convert the objects in each column into something that is readable. Otherwise, drop all missing values. While we never want to omit information, it merely makes the data sloppy when we have several missing values.
- For outliers, locate them using the z-score and IQR test. Collect and observe where most outliers are found, and remove them from the dataset if they do not fit in with the rest of the values. While it is difficult to closely inspect a dataset of this size, it is important that this process is selective. After all, it is entirely possible, given the size of this file, that the 6 million 'outlier' values are not worth omitting, but rather provide invaluable insight into some of the strengths and weaknesses of the company. Therefore, it is important to both locate and inspect these outliers, and program them into our EDA and understanding of the company.
- For skewness, inspect the columns in which the number is far from 0. For numbers greater than 0, understand that the data is right, or positively skewed, meaning that most of the data points are concentrated on the left side. For a column like 'age', a high positive skew would mean that the customer base is young, on average. For skewness numbers less than 0, be aware that the data is left, or negatively skewed. This means that most of the data points are concentrated on the right side of the distribution. For a column like 'seniority', this would mean that the bulk of the customer base has been there for plenty of months. To reiterate, we don't want to manipulate or mitigate skewness, but rather be cognizant of it while conducting our Exploratory Data Analysis.