Data Glacier – Virtual Internship

Final Project

Name: Isha Panjwani

Email: ishapanjwani5@gmail.com

Country: Canada

Company: Data Glacier

Specialization: Data Analytics

Deliverable: Week 8

Group: Individual

Problem Statement: In this project, our client is a Latin American credit union company XYZ. They are having issues in cross-selling banking products such as credit cards, savings accounts, retirement accounts, and safe deposit boxes. It can take a significant amount of research and business knowledge to increase cross-selling. To succeed in the cross-selling area of the business, Data Analyst at ABC analytics is searching for the best technique to be recommended.

Business statement: The goal of ABC analytics company is to perform Exploratory data analysis on the data provided by the client and gain some meaningful insights. As a data analyst intern, my job was to perform EDA on the credit union's dataset and create visualizations to analyse the data and to provide recommendations to the company to increase effective cross-selling of banking products.

Data Understanding:

(i) Attributes of data: The attributes gathered for each observation are

- **a.** The data gathered for observation is in the form of a csv file.
- **b.** It includes more than 13M observations collected in 18 months for nearly 937k customers of the credit union.
- **c.** All the attributes of the data include
    - **i.** Customer Demographics
    - **ii.** Customer Join-Leave Date
    - **iii.** Customer Type
    - **iv.** Relations with the employees
    - **v.** Number and kinds of accounts they hold

(ii) Date Volume: Raw data acquired from the XYZ Credit Union is with around 13M records and 48 fields.

- **a.** The volume of the data has been reduced to nearly 1 million records by just keeping the latest observation of each customer to remove or reduce duplicacy/redundancy.
- **b.** Data Attributes:
    - **i.** Continuous variables - very few variables such as customer seniority joining and leaving date, age, etc are continuous variables.
    - **ii.** Categorical variables – other variables are categorical variables like types of customers, some continuous variables are transformed into categorical variables such as age is converted into age groups for better understanding of number and kinds of accounts sold.
    - **iii.** Missing values: Records with majority of fields being missing or with null values are being dropped for getting accurate data for analysis and better decision making.

(iii) Type of data for analysis: We are working with ordinal data

(iv) Problems of the data:

- **a.** Number of records with most NA values: 27734
- **b.** Outliers: There are 17 outlier values in employee index field
- **c.** Mean, median, mode: The mean(), median(), mode() and other statistical values are being calculated in the python notebook with the query: df.describe()
- **d.** Skew values:
    - **i.** Negative skew values: Some negative skew values included fields like customer code, customer address, and current accounts.

ii. Positive skew values: Some positive skew values included fields included fields like Saving account, guarantees, derivative account.

(v) Approach for cleaning data: Missing values:

a. In Customer Code field has been removed as customer code is always a unique code assigned to each individual customer.

b. In Employee Index field outliers are dropped

c. In Gender field, random gender is assigned in place of nulls.

d. Customer leave date – Null values are replaced by last date of collecting data.

e. Gross Income: Missing values in gross income field has been replaced by average value of the gross income of all the fields.

Github Repo link: https://github.com/isha1912/Cross-Selling-Recommendation-EDA