Data Glacier – Virtual Internship

Final Project

Name: Inesisa Primrose Mpofu

Email:
inesisap.mpofu@gmail.com

Country: South Africa

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:

Total number of observation	929 615
Total number of files	1
Total number of features	24
Format of file	csv
Size	107 706KB

   a. There are 9 set of features that are numeric and 15 which are data types.
   b. All the attributes of the data include:
   Index(['The table is partitioned for this column',
    'Employee index: A active, B ex employed, F filial, N not employee, P pasive',
    'Customer's Country residence', 'Customer's sex',
    'The date in which the customer became as the first holder of a contract in the bank',
   'Last date as primary customer (if he isn't at the end of the month)',
    'Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)',
    'Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)',
    '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 used by the customer to join', 'Deceased index. N/S',
    'Province name', 'Gross income of the household',
    'segmentation: 01 - VIP, 02 - Individuals 03 - college graduated']

   Numeric values include:
    Index(['Customer code', 'Age',
     'New customer Index. 1 if the customer registered in the last 6 months.',
     'Customer seniority (in months)',
     '1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)',
   'Addres type. 1, primary address',
     'Activity index

c. 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.

    iv. There are no duplicate values

(ii) Problems of the data:

    a. Number of records with most NA values: 27734

    b. Outliers:
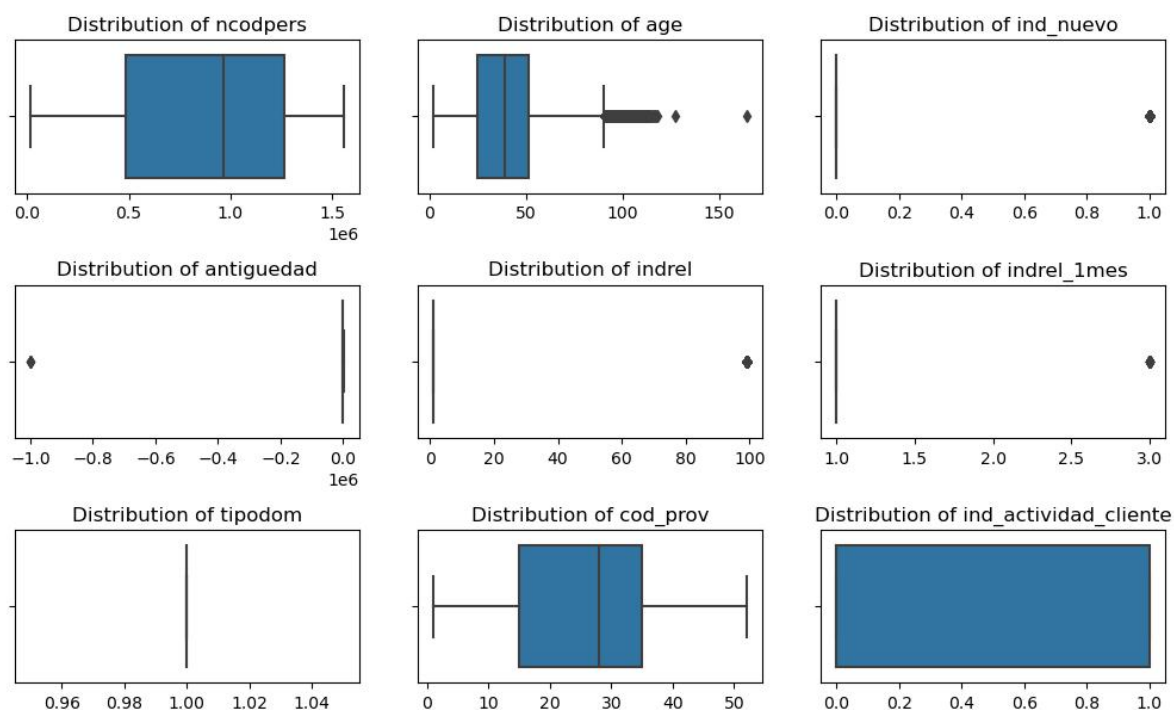
    c. Skew values:

       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.

Github Repo link: https://github.com/Ineprim12/Week-8-Cross-selling