Data Glacier Virtual Internship – Final Project (Week 8 Report)

Cross Selling Recommendation for XYZ Credit Union

1. Team Member Details

Group Name: Data Vision

Name	**Email**	**Country**	**College/Company**	**Specialization**
Naga Pavithra Jajala	pavithrajajala8naga@gmail.com	United Kingdom	Birmingham City University	Data Analyst
Raj Pawar	rajpawar32646@gmail.com	United Kingdom	University of Liverpool	Data Analyst

2. Problem Description

XYZ Credit Union in Latin America has shown good performance in selling individual banking products like credit cards, deposit accounts, and retirement accounts. However, data shows that most customers own only one product, indicating weak cross-selling performance.

The business aims to improve customer engagement and profitability by increasing the number of products held by each customer. Our goal as Data Analysts is to explore the dataset and recommend actionable strategies (without using machine learning) to improve cross-selling opportunities.

3. Data Understanding

The dataset provided for this project is named Train.csv and contains extensive information on customer demographics, product usage history, and relationship data.

- **Total Records: 13,647,309**
- **Total Features: 48 columns**
- **File Size: ~229 MB**
- **Format: CSV**

Due to memory limitations in Google Colab, initial data understanding was conducted on a 1,000-row sample using nrows=1000.

4. What Type of Data We Have

The dataset contains a mix of data types and categories:

- **Numerical Features:**
  - **renta (estimated income)**
  - **age**
  - **antiguedad (seniority)**
  - **cod_prov (province code)**
- **Categorical Features:**
  - **sexo, pais_residencia, segmento, ind_empleado, canal_entrada, etc.**

- **Date/Time Features:**
  - **fecha_dato (record month)**
  - **fecha_alta (account creation date)**
  - **ult_fec_cli_1t (last contact date)**
- **Target Variables – Product Ownership Flags:**
  - **24 binary columns such as ind_cco_fin_ult1, ind_hip_fin_ult1, ind_nomina_ult1, ..., ind_recibo_ult1 indicate whether the customer holds that specific banking product.**

5. Data Quality Issues Identified

Based on the 1000-row sample:

Column	**Missing Values (out of 1000)**	**Action Suggested**
conyuemp	1000	Drop (100% null)
ult_fec_cli_1t	999	Drop or ignore
renta	165	Impute (mean/median/segment)
Other columns	1 each	Simple imputation

Other Observations:

- age, antiguedad, and similar numeric fields are stored as text (object type) – need conversion to numeric.
- renta and age show signs of outliers and skewed distribution.
- Product flag columns are clean binary (0 or 1) and ready for EDA.

6. Approaches to Handle Identified Issues

Issue	**Approach**
Missing values (conyuemp)	Drop the column (100% null)
Missing values (renta)	Impute using mean/median or segment-based imputation
Minor nulls in other fields	Use mode (for categorical) or mean (for numeric)
Object-type numerics	Convert columns like `age`, `antiguedad` to integers
Outliers in numeric columns	Identify using IQR and handle via capping or flooring
Skewed distributions	Apply **log transformation** for visual clarity and comparison
Large dataset size	Use chunk processing via `nrows=1000` to avoid memory errors in Colab

7. GitHub Repository Link

https://github.com/nagapavithrampl/Data-Glacier---Final-Project

https://github.com/nagapavithrampl/Data-Glacier---Final-Project/tree/main/Week-8