Week 10 - Data Analyst: Cross-selling recommendation

Team Member Details: Individual project (no team)

Name: Asha K. C.

Email: aasha.kc@outlook.com

Country: India

Company: DataGlacier

Specialization: Data Analyst

Problem Description:

The XYZ bank is having difficulty cross-selling its products to existing customers. Customers are not buying additional products sold by their bank. Hence, as data analysts, we must provide the information required to enhance their cross-selling methods.

Github Repo link:

https://github.com/Asha-KC-07/Data-Glacier-Internship-2025---LISUM43/blob/main/Week%2010%20-%20Data%20Analyst_Cross-selling%20recomendation/EDA_v1.ipynb

EDA performed on the data:

1. Customer Demographics:
    a. Histogram of Age Distribution - To detect skewness, outliers, and age groups
    b. Bar Chart of Gender Counts - To see the male/female split
    c. Bar Chart of Top Provinces (nomprov) - To locate the highest customer concentrations
    d. Age by Gender Boxplot - To detect age differences between genders
    e. Heatmap of Age vs. Province (Top N) - To understand regional age variation.
2. Product Adoption Rates - Identify the most and least commonly used products by customers to understand baseline engagement and potential cross-sell opportunities. Columns used for this are ind_ahor_fin_ult1 → ind_recibo_ult1. All values are binary (0 = product not owned, 1 = owned).
3. Product Ownership by Segment - Identify how product ownership varies across **demographics and customer segments**, revealing which groups prefer which products. Variables used are sexo, age, segmento, all product columns (last 24 columns)
4. Customer Lifetime Value Proxies - Calculate proxy scores using a weighted combination of:
    a. Tenure (antiguedad)

    b. Income (renta)
    c. Product count (number of products held)
5. Churn Indicators - Use ult_fec_cli_1t to identify recent exits and compare patterns between churned and active customers in terms of:
    a. Age
    b. Segment
    c. Product ownership
    d. Tenure
6. Channel Effectiveness - Analyze acquisition channel effectiveness using:
    a. Customer count per channel
    b. CLV per channel
    c. Average product count by channel
7. Income Analysis - Explore income (renta) distribution and its relationship with:
    a. Age
    b. Segment
    c. Product ownership
8. Product Co-Ownership Patterns - Detect which products are often held together using correlation analysis.
9. Anomaly Detection - Identify outliers in:
    - Age (extreme values)
    - Income (extremely high/low)
    - Product mix (zero or full ownership)

Final Insights:

Demographics Analysis Insights

🔹 **1. Age Distribution Histogram**

- Most customers fall between **25 to 60 years**.
- There's a visible right skew due to older age values.
- Minor counts exist for extremely old ages (over 90).

🔹 **2. Gender Distribution**

- **Slightly more females** than males.
- Imbalance isn't significant but could affect segment-specific targeting.

🔹 **3. Top 10 Provinces by Customer Count**

- Certain provinces (like **Madrid**, **Barcelona**, etc.) dominate the customer base.
- Indicates geographic concentration — great for region-specific marketing strategies.

🔹 **4. Age by Gender Boxplot**

- Female customers tend to be **slightly older** on average.

- The age spread is similar for both genders but has a few more high-age outliers.

🔹 **5. Heatmap of Age vs. Province**

- Each province shows a **strong concentration around ages 40-55**.
- Useful to identify which regions have a younger vs older customer base.

Product Adoption Rate Insights

🔹 **1. Bar Chart: Product Adoption Rates**

- Most commonly held products:
  - `ind_recibo_ult1` (**12.1%**): Utility Bill Payments
  - `ind_nom_pens_ult1` (**5.5%**): Pension Deposits
  - `ind_nomina_ult1` (**5.1%**): Salary Deposits
  - `ind_tjcr_fin_ult1` (**3.8%**): Credit Cards
- Least commonly held products:
  - `ind_plan_fin_ult1`, `ind_pres_fin_ult1`, `ind_viv_fin_ult1`: All below **1%**

🔹 **2. Pie Chart: Total Product Holdings**

- A few product types dominate the customer portfolio space.
- Products like **salary/pension deposits** and **recurring payments** represent the bulk of ownership.

Key Takeaways:

- The dataset reflects a **basic banking usage pattern**, with limited product penetration.
- Many products have **<1%** ownership, indicating unexplored cross-selling opportunities.

Product Ownership by Segment Insights

🔹 **1. Gender-Based Ownership**

- **Females** slightly lead in products like:
  - `ind_nomina_ult1` (salary deposits)
  - `ind_recibo_ult1` (recurring bill payments)

- **Males** show marginally higher ownership in:
  - Investment-related products (`ind_valo_fin_ult1`, `ind_fond_fin_ult1`)

📌 Gender-based marketing strategies could be tuned around salary vs. investment product preferences.

🔹 **2. Age Group Trends**

- **Young adults (<25)** rarely hold any product.
- **Ages 35-54** dominate across almost every product — especially:
  - Salary deposits (`ind_nomina_ult1`)
  - Credit cards (`ind_tjcr_fin_ult1`)
- **Older groups (65+)** tend to show less ownership of credit or investment products.

📌 Age-based targeting: Focus younger on entry products, middle-aged on cross-sell, older on retention and service.

🔹 **3. Segment-Based Trends**

- `segmento` indicates customer types like "01 - VIP", "02 - Individuals", "03 - College students":
  - **VIPs** show highest ownership in multiple financial products, especially credit and investment tools.
  - **College students** exhibit minimal product ownership — mostly basic accounts.

📌 A clear case for segmented offerings: upscale for VIPs, simplified for students.

Customer Lifetime Value Proxy Insights

🔹 **1. Distribution of CLV Proxy**

- Most customers fall within **mid** CLV proxy scores.
- Right-skewed: a smaller segment shows high-value potential (top 10–15%).

🔹 **2. Boxplot by Segment**

- **VIPs** (`01 - VIP`) have the highest median and spread of CLV.
- **College students** (`03 - UNIVERSITARIO`) score lowest in CLV — expected due to low tenure, product count, and income.
- **Mass Market Individuals** (`02 - PARTICULARES`) span the full spectrum, indicating a diverse customer base.

🔹 **3. Correlation Heatmap**

- **Product count** (`norm_products`) is most strongly correlated with overall CLV score.
- **Income and tenure** also contribute but less dominantly.

📌 Suggests focusing product penetration for increasing lifetime value, especially among mid-tier customers.

Churn Indicator Insights

🔹 1. Churned vs Retained Customers

- A **very small fraction** of the customers are marked as "churned".
- Indicates the dataset mostly includes **active customers**.

🔹 2. Age, Tenure, and CLV Comparisons

- **Churned customers** tend to:
  - Be **older on average**
  - Have **shorter tenure** (surprising — might reflect new users abandoning)
  - Have **lower CLV scores** overall
- Retained users dominate in higher tenure and product count distributions.

🔹 3. Product Ownership Drop-Off

- Churned users have **lower ownership rates across nearly all products**.
- Largest relative drop in products like:
  - Credit Cards (`ind_tjcr_fin_ult1`)
  - Salary/Pension Accounts

📌 Insight: Churn is tightly linked with low engagement. Preemptive outreach to low-product-count users could reduce attrition.

Channel Effectiveness Insights

🔹 1. Top Acquisition Channels by Customer Count

- A few channels dominate customer acquisition:
  - `KHE`, `KAT`, and `KFC` are among the most used.
- These likely represent physical or digital acquisition pathways.

🔹 2. Average CLV Proxy by Channel

- High customer volume **does not always mean high value**:
    - Channels like `KAT` and `KFA` show **higher CLV**, despite smaller customer bases.
- Mass channels may acquire many users, but **niche or referral-based ones attract higher value clients**.

🔹 **3. Average Product Count by Channel**

- Channels with higher CLV also typically yield **more product engagement**.
- `KFA` and `KAT` again appear as effective **quality acquisition routes**.

📌 Strategic Focus:

- Maintain mass channels (e.g., `KHE`) for volume.
- Invest in high-value channels (`KFA`, `KAT`) for profitability.

Income Analysis Insights

🔹 **1. Income Distribution (Log Scale)**

- Highly **right-skewed**: majority of incomes lie below ~60,000.
- A few customers report **extremely high income** (>100,000), suggesting income outliers.

🔹 **2. Income by Customer Segment**

- **VIPs** (`01 - VIP`) predictably show the **highest income range**.
- **College students** (`03 - UNIVERSITARIO`) have the lowest and most compact income distribution.
- Segments are well-separated, validating the segmentation strategy by income.

🔹 **3. Income vs. Product Count**

- Positive trend: **higher product count tends to correlate with higher income**.
- However, some **high-income customers own few products**, indicating potential for upselling.

📌 Strategy:

- Target under-engaged high-income users.
- Customize offerings to segment-specific income brackets.

Product Co-Ownership Insights

🔹 **Key Product Correlations**

- **High Positive Correlation Pairs**:
  - `ind_nomina_ult1` (salary) ↔ `ind_recibo_ult1` (bill payments): 0.66
    ➤ Customers receiving salaries tend to set up recurring payments.
  - `ind_nom_pens_ult1` (pension) ↔ `ind_recibo_ult1`: 0.45
    ➤ Similar trend with pension-based income.
- **Investment-related Products** like `ind_fond_fin_ult1`, `ind_valo_fin_ult1`, `ind_deco_fin_ult1` are often held together:
  - Moderate correlations (~0.3–0.5), indicating bundled behaviors.
- **Minimal or Near-Zero Correlations**:
  - Savings accounts (`ind_ahor_fin_ult1`) and insurance (`ind_plan_fin_ult1`) don't strongly align with others.

📌 Use-case:

- Suggesting new products based on current holdings becomes data-driven (e.g., customers with payroll should be targeted for bill setup or investment services).

Anomaly Detection Insights

🔹 **1. Age Outliers**

- Some customers are recorded as **under 18** and **over 100**.
- Likely data entry or formatting errors — recommend flagging or excluding these from sensitive analysis.

🔹 **2. Income Outliers**

- Very high-income values (>99th percentile) sharply diverge from the median.
- Income is **extremely skewed**, requiring **log scaling** for meaningful analysis.
- These cases might be legitimate high-value clients or input anomalies.

🔹 **3. Product Count Extremes**

- Many customers have **zero products**, indicating passive or new accounts.
- A very small number of customers hold **all products**, possibly internal test users or high-value clients.

📌 Recommendation:

- Clean or filter age/income outliers for modeling or reporting.
- Investigate zero-product holders for onboarding improvements.
- Review full-product holders for potential upsell benchmarks or audits.