Problem Description

XYZ Credit Union in Latin America is performing well in selling banking products such as credit cards, deposit accounts, retirement accounts, and safe deposit boxes. However, the bank struggles with cross-selling; most customers own only one product. XYZ Credit Union has approached ABC Analytics to devise a solution to enhance cross-selling among its existing customers.

Data Understanding

Type of Data Received for Analysis:

The dataset provided by XYZ Credit Union contains the following information:

1. **Customer Information:**
    - **ncodpers**: Customer code
    - **sexo**: Customer's sex
    - **age**: Age
    - **pais_residencia**: Customer's country of residence
    - **renta**: Gross income of the household
    - **segmento**: Customer segmentation (VIP, Individuals, College Graduated)
2. **Employment Information:**
    - **ind_empleado**: Employee index (A: active, B: ex-employed, F: filial, N: not employee, P: passive)
    - **conyuemp**: Spouse index (1: spouse of an employee)
3. **Customer Relationship Data:**
    - **fecha_dato**: Data partitioning column
    - **fecha_alta**: The date the customer became the first holder of a contract in the bank
    - **ind_nuevo**: New customer index (1: customer registered in the last 6 months)
    - **antiguedad**: Customer seniority (in months)
    - **indrel**: Primary relationship indicator (1: first/primary, 99: primary customer during the month but not at the end)
    - **ult_fec_cli_1t**: Last date as primary customer (if not at the end of the month)
    - **indrel_1mes**: Customer type at the beginning of the month (1: primary customer, 2: co-owner, P: potential, 3: former primary, 4: former co-owner)

- **tiprel_1mes**: Customer relation type at the beginning of the month (A: active, I: inactive, P: former customer, R: potential)
- **indresi**: Residence index (S: yes, N: no if the residence country is the same as the bank country)
- **indext**: Foreigner index (S: yes, N: no if the customer's birth country is different than the bank country)
- **canal_entrada**: Channel used by the customer to join
- **indfall**: Deceased index (N: no, S: yes)
- **tipodom**: Address type (1: primary address)
- **cod_prov**: Province code (customer's address)
- **nomprov**: Province name
- **ind_actividad_cliente**: Activity index (1: active customer, 0: inactive customer)

4. **Product Usage Information:**
   - Information on ownership of various products (e.g., saving accounts, current accounts, credit cards, loans, etc.)

Data Quality Issues

Missing Values (NA values):

The dataset contains several columns with missing values. Below is the summary of missing values in terms of count and percentage:

Column	Missing Values	Missing Values (%)
age	0	0.0%
antiguedad	0	0.0%
canal_entrada	1881	23.65%
cod_prov	2744	34.51%
conyuemp	795092	99.99%
ind_actividad_cliente	1	0.01%
ind_nuevo	0	0.0%
indext	1	0.01%
indfall	1	0.01%
indrel	0	0.0%
indrel_1mes	24	0.30%
indresi	1	0.01%
ncodpers	0	0.0%

Column	Missing Values	Missing Values (%)
nomprov	2744	34.51%
renta	1	0.01%
segmento	2021	25.41%
sexo	5	0.06%
tipodom	1	0.01%
tiprel_1mes	24	0.30%
ult_fec_cli_1t	793640	99.80%

Outliers:

Outliers are detected using both the Interquartile Range (IQR) method and the Z-score method. The counts of outliers in numerical columns are as follows:

Column	Outliers (IQR method)	Outliers (Z-score method)
age	10118	5619
antiguedad	3	3
canal_entrada	NaN	NaN
cod_prov	0	0
conyuemp	NaN	NaN
ind_actividad_cliente	0	0
ind_nuevo	25297	25269
indext	NaN	NaN
indfall	NaN	NaN
indrel	1556	1555
indrel_1mes	26	26
indresi	NaN	NaN
ncodpers	0	0
nomprov	NaN	NaN
renta	NaN	NaN
segmento	NaN	NaN
sexo	NaN	NaN
tipodom	0	0
tiprel_1mes	NaN	NaN
ult_fec_cli_1t	NaN	NaN

Skewed Data:

Some columns have highly skewed data, which can affect the model's performance. The skewness values for numerical columns are:

Column	Skewness
age	NaN
antiguedad	-513.821226
canal_entrada	NaN
cod_prov	NaN
conyuemp	NaN
ind_actividad_cliente	NaN
ind_nuevo	5.335482
indext	NaN
indfall	NaN
indrel	22.540082
indrel_1mes	174.873251
indresi	NaN
ncodpers	NaN
nomprov	NaN
renta	NaN
segmento	NaN
sexo	NaN
tipodom	NaN
tiprel_1mes	NaN
ult_fec_cli_1t	NaN

Summary:

Missing Values: Significant missing values are present in canal_entrada, cod_prov, conyuemp, nomprov, segmento, and ult_fec_cli_1t.

Outliers: Notable outliers are present in age, ind_nuevo, and indrel.

Skewed Data: High skewness is observed in antiguedad, ind_nuevo, indrel, and indrel_1mes.

Approaches to overcome these data issues include imputation for missing values, outlier treatment using capping or transformation, and normalization or transformation for skewed data.

Approaches to Overcome Data Issues

To handle the data quality issues such as missing values, outliers, and skewed data, we will implement the following strategies:

1. Missing Values (NA values)

Approaches:

- **Imputation:**
  - **Mean/Median Imputation:** For numerical columns, impute missing values with the mean or median. For example, the **renta** column can be imputed with the median since income data might be skewed.
  - **Mode Imputation:** For categorical columns, impute missing values with the mode. For instance, columns like **sexo** and **canal_entrada** can be imputed with the most frequent value.
- **Special Handling:**
  - **Columns with High NA Percentage:** Columns like **conyuemp** and **ult_fec_cli_1t**, which have over 99% missing values, may be dropped if deemed irrelevant or imputed with a special category indicating 'unknown'.

Why?

- Imputation helps to retain all data and avoid dropping rows, which is crucial for maintaining a large dataset for training models. It ensures that the models have a complete dataset without losing potentially valuable information.

2. Outliers

Approaches:

- **IQR Method (Capping):**
  - For columns with outliers detected by the IQR method, cap the outliers to a fixed percentile. For instance, cap the **age** values beyond the 1.5 * IQR range to the upper and lower bounds defined by the IQR.
- **Z-score Method (Transformation):**
  - For columns with outliers detected by the Z-score method, apply transformations such as log or square root to reduce the impact of extreme values. For example, log transformation can be applied to the **renta** column if it contains extreme outliers.

Why?

- Handling outliers prevents models from being overly influenced by extreme values, leading to better model performance and more stable predictions.

3. Skewed Data

Approaches:

- **Log Transformation:**
    - Apply log transformation to skewed numerical columns like **antiguedad** and **renta** to reduce skewness and make the distribution more normal.
- **Square Root Transformation:**
    - Apply square root transformation to columns with moderate skewness to stabilize variance and normalize the data distribution.

Why?

- Transformations help to normalize skewed data, making the data distribution more symmetric and suitable for modeling, especially for algorithms that assume normally distributed data.