Data Glacier

Week 9: Deliverables

Group Name: Solo Analyst

Name: Munirah Alfehaid

Email: munirah9hamad@gmail.com

Country: Saudi Arabia

Specialization: Data Analyst

Problem Description

XYZ Credit Union in Latin America has been successful in selling individual banking products like credit cards, deposit accounts, and retirement accounts. However, they are facing a challenge in cross-selling, as existing customers are typically not purchasing more than one product. The goal of this project is to analyze the current situation and suggest strategies to increase cross-selling opportunities among existing customers without relying on machine learning solutions.

Data Cleansing and Transformation

To address the issue of missing values in our dataset, the following approaches were employed:

Categorical Data Imputation:

> For the categorical column sexo, which contains missing values, the most frequent value (mode) was used to fill in the missing entries. This is a common practice for categorical variables where the most frequent category is assumed to be the best replacement for missing values.

Numeric Data Imputation:

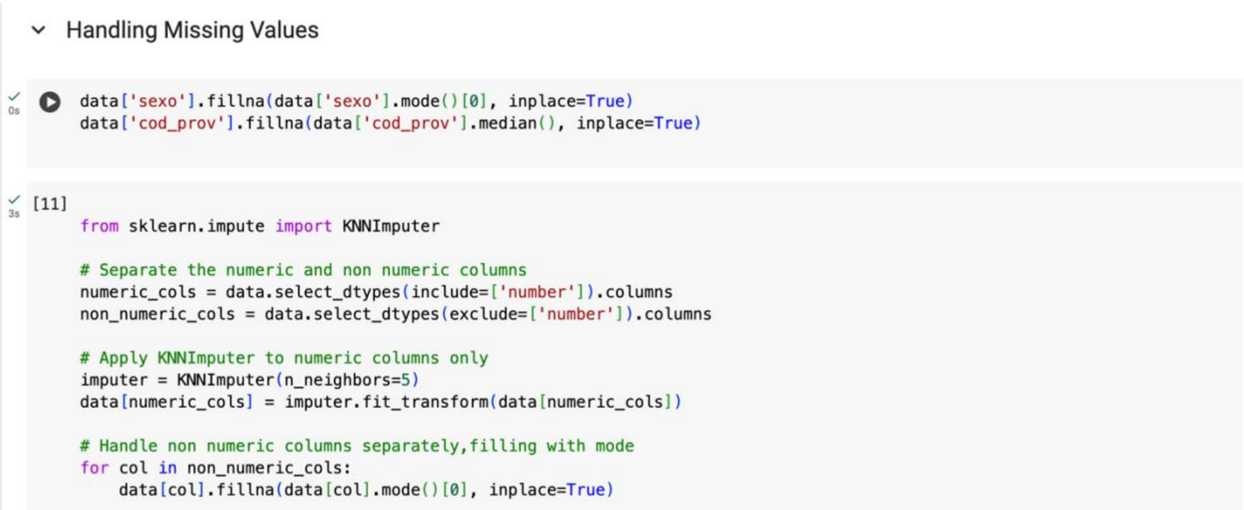
> For the numeric column `cod_prov`, which had missing values, the median of the column was used to fill in the missing entries. The median is preferred over the mean because it is less affected by outliers, providing a more robust central tendency measure.

KNN Imputation for Numeric Columns:

> To handle missing values in other numeric columns, the K-Nearest Neighbors (KNN) imputation method was applied. This method imputes missing values by finding the nearest neighbors (in terms of other features) and averaging their values. This technique is useful for capturing the underlying structure in the data.

Handling Non-Numeric Columns Separately:

> For non-numeric columns that were not suitable for KNN imputation, the mode was used to fill in the missing values. This ensured that categorical data were treated appropriately, maintaining the integrity of the dataset.



Handling Missing Values

```
data['sexo'].fillna(data['sexo'].mode()[0], inplace=True)
data['cod_prov'].fillna(data['cod_prov'].median(), inplace=True)
```

```
from sklearn.impute import KNNImputer

# Separate the numeric and non numeric columns
numeric_cols = data.select_dtypes(include=['number']).columns
non_numeric_cols = data.select_dtypes(exclude=['number']).columns

# Apply KNNImputer to numeric columns only
imputer = KNNImputer(n_neighbors=5)
data[numeric_cols] = imputer.fit_transform(data[numeric_cols])

# Handle non numeric columns separately,filling with mode
for col in non_numeric_cols:
    data[col].fillna(data[col].mode()[0], inplace=True)
```