Group Project Report

Group details:

Group Name	< DH-CrossSelling >
Member Name	<Dingyun Hu>
Email Address	<dh3517@nyu.edu>
School	<New York University>
Country	<USA>
Specialization	<Data Analyst: Cross selling recommendation >

Problem Description:
The project aims to increase cross-selling of banking products for XYZ Credit Union in Latin America. While the credit union excels in selling various banking products, the challenge lies in promoting additional offerings to existing customers. The primary objective is to analyze customer data and derive actionable insights that can enhance cross-selling strategies.

Data Understanding:
The dataset provided for analysis consists of customer information and binary indicators of whether a customer holds specific banking products. The dataset includes both numerical and categorical features, with a total of 48 columns and 13,647,309 rows. Several features have missing values, including 'ind_empleado, pais_residencia, sexo, fecha_alta, ind_nuevo, indrel, ult_fec_cli_1t, indrel_1mes, tiprel_1mes, indresi, indext, conyuemp, canal_entrada, cod_prov, nomprov, ind_actividad_cliente, renta, and segmento.

Type of Data for Analysis:
The data set is a 2.29GB csv file, includeing the following types of features:

- Numerical Features (32 in total): These features represent numerical values such as customer age, seniority, income, and various binary indicators of product ownership.
- Categorical Features (16 in total): These features include categorical variables like employee index, gender, customer segment, and more.

Data Problems:
Upon initial exploration, several data problems were identified:

- Missing Values (NA Values): Numerous columns have missing values, with varying degrees of missingness. Including: ind_empleado, pais_residencia, sexo, fecha_alta, ind_nuevo, indrel, ult_fec_cli_1t, indrel_1mes, tiprel_1mes, indresi, indext, conyuemp, canal_entrada, tipodom, cod_prov, nomprov, ind_actividad_cliente, renta, segmento
- Outliers: Outliers were detected in numerical features like age and income.

5. Approaches to Address Data Problems:

- Handling Missing Values: For categorical features, missing values were dealt based on certain data. Methods include imputing with the mode to ensure minimal data loss and replace "NA" with "Unknown". For numerical features, missing values were imputed with the mean to preserve data integrity.

- Outlier Detection and Treatment: Outliers in numerical features were identified using Z-scores and IQR methods. Extreme values beyond certain thresholds were treated as outliers and were either removed or transformed using appropriate techniques.

Github repo for week 8
“https://github.com/OliverHu726/DataGlacier-Intern/tree/main/week%208”