Project Report: Data Wrangling for Santander Value Prediction

Introduction

This report presents the initial stages of the "Santander Value Prediction" project, which aims to predict transaction values and provide valuable business insights. The project involves data preprocessing, exploratory data analysis (EDA), and initial feature engineering.

Data Collection and Exploration

The project began by importing essential Python libraries (Pandas and NumPy) and loading the provided datasets (train.csv and test.csv). These datasets contain anonymized customer transactional data. The dimensions of the datasets were examined, revealing that the training dataset has 4459 rows and 4993 columns, while the testing dataset has 49342 rows and 4992 columns. The 'train.csv' dataset includes a 'target' column representing transaction values.
The summary statistics of the testing dataset were calculated and printed, providing an overview of the data's distribution and variability. The 'info()' method was used to inspect the data types and identify potential missing values. Both datasets were observed to have columns with numeric data types (float64 and int64) and one object type column ('ID' in both cases).

Data Cleaning

No missing values were found in either dataset, which is a positive sign for the data quality. However, the 'train.csv' dataset was observed to contain columns with constant values across all rows. A list of these constant columns was generated, and a total of 256 columns were identified. These columns might not provide meaningful information for modeling and analysis, so they were dropped from the 'train.csv' dataset, resulting in a new dataset with 4737 columns.
Additionally, duplicate rows were checked in the filtered training dataset, and no duplicates were found.

Project Proposal Reflection

The project proposal highlighted the problem of accurately predicting transaction values to gain customer-centric insights for various industries such as commercial banking, e-commerce, and retail. It emphasized the importance of accurate predictions for tailoring services and optimizing marketing efforts. The proposal also outlined the stakeholders' interests, client needs, context, criteria for success, and constraints.

Next Steps

The data preprocessing and initial exploration phases have been completed successfully. The next steps in the project involve feature engineering, model selection, model training, evaluation, implementation, and reporting. The project aims to create a predictive model that accurately estimates transaction values and provides actionable insights for client businesses.

Conclusion

The initial stages of the "Santander Value Prediction" project have set a strong foundation for further analysis and modeling. The data has been preprocessed, constant value columns removed, and the project's context and objectives revisited. The project is now poised to move forward with feature engineering, model development, and ultimately, providing valuable insights to the client stakeholders.

This project report was prepared by Sang Thai, a student, and Bernard Chan, a Data Scientist & Machine Learning Specialist.