Loan Default Prediction Project | Preliminary Data Summary (Milestone 2)

Executive Summary Report

Prepared by: PhD Aleksandar Osmanli

OVERVIEW

I'm currently developing a data analytics project aimed at decreasing overall loan default among borrowers by predicting which client attributes most contribute their loan default. For the purposes of this project, loan default means borrowers failed to pay their loans to the financial institution. **This report offers a preliminary data summary, information on the project status and key insights of Milestone 2, which impact the future development of the overall project.**

PROJECT STATUS

Milestone 2 - Compile Summary Information

🎯 **Target Goal:** Inspect user data to learn important relationships between variables.

🎯 **Methods:**

- Built a dataframe
  - Each row represents a single observation, and each column represents a single variable
- Collected preliminary statistics
- Analyzed user behavior

🎯 **Impact:** important relationships were determined between variables that will guide further analysis of user data.

NEXT STEPS

➔ **It is recommended to analyze the impact of some variables on the client's loan default**. It's obvious that variables 'Age', 'InterestRate', 'Income', 'MonthsEmployed' and 'CreditScore' have negative impact on the loan default, while 'InterestRate', 'LoanAmount', 'NumCreditLines' and 'DTIRatio' have positive impact on the loan default.

➔ **The immediate next step is to conduct thorough EDA and feature engineering** to reveal the correlation between features and the loan default.

KEY INSIGHTS

- I was provided with two datasets:

1. Train dataset contains 18 variables and 70% of the overall sample, and will reveal whether the client was loan defaulted or not (the "ground truth").
2. Test dataset contains 17 variables with the exact same information about the remaining sample of 30%, but does not disclose the "ground truth" for each loan. I should predict the outcome.

- The types of variables include 8 objects (7 categorical variables), 2 floats, and 8 integers; one categorical variable contains space and single quotes between its possible values, so I concatenated and simplify them to avoid future problems with analysis.
- There were no missing values in both train and test sets.