Assignment 1 Report

Simrat Kaur Randhawa
20031112
srandhaw@stevens.edu
https://www.kaggle.com/simratkr

1. Introduction

This report details the data preprocessing, feature engineering, and the introduction of a quadratic variable within a financial dataset.

2. Data Loading and Initial Exploration

The dataset is loaded using Pandas from a CSV file (lc_loan.csv). Initial steps include:

- Identifying categorical and numerical columns.
- Checking for missing values.
- Conducting statistical analysis and visualization to understand distributions and correlations.
- Analyzing categorical variable distributions and identifying features for encoding.

3. Data Preprocessing

3.1 Handling Categorical Variables

Categorical variables are processed using different encoding techniques:

- **Label Encoding** for ordinal variables such as grade, sub_grade, emp_length, and verification_status.
- **One-Hot Encoding** for nominal variables like home_ownership, purpose, addr_state, and loan_status.

3.2 Handling Missing Values

Missing values are analyzed and handled appropriately based on their nature and potential impact on modeling.

3.3 Feature Scaling

Standardization techniques such as StandardScaler are applied to numerical features to improve model performance.

4. Feature Engineering

To enhance predictive power, additional features are created:

- **Financial Ratios**:
    - debt_to_income: Ratio of loan amount to annual income.
    - payment_to_income: Ratio of installment amount to monthly income.
    - loan_to_value: Ratio of loan amount to funded amount.
- **Interest Rate Features**:
    - interest_income: Product of loan amount and interest rate.
    - effective_rate: Estimated effective annualized return.

5. Introduction of a Quadratic Variable

To capture potential non-linear relationships, a quadratic feature is introduced:

- debt_to_income_squared: The square of the debt_to_income ratio, acknowledging that its impact may not be strictly linear.

This transformation allows the model to capture more complex relationships between debt levels and financial outcomes.

6. Data Visualization

Key visualizations include:

- Histograms to inspect the distribution of numeric features.
- Correlation heatmaps to identify multicollinearity and feature relationships.
- Categorical value counts to analyze the distribution of different categories.

7. Conclusion

This preprocessing and feature engineering pipeline enhances the dataset for modeling by:

- Ensuring numerical features are properly scaled.
- Encoding categorical variables effectively.
- Creating meaningful financial ratios and transformations.
- Introducing a quadratic feature to capture non-linearity.

These steps improve the dataset's quality and model interpretability, leading to better predictive performance in downstream analysis.

Generative AI help:

I have used generative AI for help in debugging in a few places I was stuck and in writing up this report. Other than this a few places for help with syntax in python.