Loan Default Prediction Report

Dataset Description and Preprocessing Steps

The Lending Club Loan Dataset contains data on peer-to-peer loans issued by the Lending Club. It includes various borrower features such as credit history, loan amount, interest rate, employment status, and payment records. Preprocessing steps involved:

- Handling missing values through imputation techniques.

- Encoding categorical variables.

- Addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

- Feature scaling using StandardScaler for models like SVM.

Models Implemented with Rationale for Their Selection

1. LightGBM: A gradient boosting framework that is efficient and provides high accuracy on structured data.

2. SVM (Support Vector Machine): Chosen for its ability to handle high-dimensional feature spaces and provide clear decision boundaries.

Both models are known for their robustness and are suitable for binary classification problems such as loan default prediction.

Key Insights and Visualizations

Key findings from the data include:

- Higher interest rates correlate with increased default probability.

- Shorter loan terms tend to have fewer defaults.

- Features such as annual income, credit score, and loan purpose are strong predictors.

Visualizations like bar charts and correlation heatmaps were used to support these insights.

Challenges Faced and Solutions

Loan Default Prediction Report

1. Imbalanced classes: The number of defaulted loans was significantly lower than non-defaulted ones. Solution: Applied SMOTE.

2. Missing data: Several columns had missing entries. Solution: Used appropriate imputation techniques based on data types.

3. Model overfitting: Initial models overfit on training data. Solution: Used cross-validation and regularization techniques.

These challenges were overcome with methodical preprocessing and model tuning strategies.