LOAN DEFAULT PREDICTION

Introduction

Predicting loan default is a critical task for financial institutions, helping mitigate risk and improve lending decisions. This project addresses the American Express "Loan Default Prediction" challenge on Kaggle. Participants must build models using anonymized customer data to predict default probability. The problem is complex due to the absence of semantic variable meanings and the need for robust preprocessing pipelines.

Goal

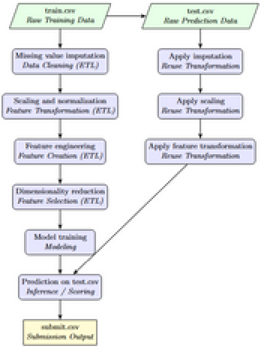
- **Research Question:** Can we develop a machine learning model that accurately predicts loan default risk from anonymized data?
- **Expected Outcome:** A competitive model that minimizes MAE (Mean Absolute Error) on the Kaggle test set.

Challenges

- Working with over 800 anonymized features
- Absence of semantic context for variables
- Maintaining consistency in preprocessing across train/test
- Evaluation metric sensitivity to outliers (MAE)

Proposed Solution

Our solution involves a modular machine learning pipeline including data cleaning, transformation, feature engineering, dimensionality reduction, and gradient boosting models. The figure below illustrates the architecture.



train.csv
Raw Training Data

test.csv
Raw Prediction Data

Missing value imputation
Data Cleaning (ETL)

Apply imputation
Reuse Transformation

Scaling and normalization
Feature Transformation (ETL)

Apply scaling
Reuse Transformation

Feature engineering
Feature Creation (ETL)

Apply feature transformation
Reuse Transformation

Dimensionality reduction
Feature Selection (ETL)

Model training
Modeling

Prediction on test.csv
Inference / Scoring

submit.csv
Submission Output

- **Key technical steps:**
- Imputation of missing values
- Scaling and normalization
- Feature engineering and selection
- Training with LightGBM and CatBoost
- Inference and generation of submit.csv

Conclusion

Our objective was to design a robust machine learning pipeline to predict loan default on anonymized data. The final LightGBM model achieved competitive MAE scores using advanced preprocessing and consistent transformations. While we met our goal, model interpretability remains limited due to feature anonymization.