IS 590 MD Final Project Proposal
Loan Default Prediction

Karan Chhabra, Erick Li

The dataset is from the Kaggle Competition--“Loan Default Prediction” (https://www.kaggle.com/c/credit-default-prediction-ai-big-data/overview). The aim of the competition is to predict whether a given person will default the loan or not by using various predictors. The training set contains 7,500 rows and 18 columns, where 17 columns serve as predictors and one as response variable. Additionally, the testing set contains 2,500 rows without the response variable. It will be used as an holdout cross validation set to check the prediction accuracy.

The descriptions of the indicators and response are summarized in Table 1.

Table 1. Description of the Columns

Columns	Description	Types
Id	Primary key. Non-predictive	Numerical
Home.Ownership	Types of accommodation	Categorical
Annual Income	Annual income of the person	Numerical
Years.in.current.job	Years the person is working in his current job	Ordinal
Tax.Liens	Number of times liens imposed on the property	Numerical
Number.of.Open.Accounts	Number of open accounts the person currently have	Numerical
Years.of.Credit.History	Length of the person’s credit history	Numerical
Maximum.Open.Credit	Maximum credit limit	Numerical
Number.of.Credit.Problems	Number of credit problems	Numerical
Months.since.last.delinquent	Number of months since the last delinquent activity	Numerical
Bankruptcies	Number of times the person has declared bankruptcy	Numerical
Purpose	The purpose of the loan	Categorical
Term	Term length, either short or long	Categorical
Current.Loan.Amount	Current loan the person has	Numerical
Current.Credit.Balance	Current available credit	Numerical
Monthly.Debt	Monthly debt	Numerical
Credit.Score	Credit score	Numerical
Credit.Default	Whether the person defaults. For prediction	Boolean

Research Questions

Q1: Given the information from the predictors, can banks predict if the person is going to default before granting loans?

This will help banks to determine whether an investment on the customer is good or bad. Banks can use the model to improve their decision-making and reduce risks.

Q2: Which factors are important in influencing whether a person will default or not?

To determine whether a person will default or not, banks are heavily dependent on background verification process. Background verification process incurs cost for banks which is ultimate paid by customers in the form of application fee. If a bank charges too high in terms of application fee, it will demotivate the customers to apply for the loan. Therefore, banks try to minimize the application fee for customers, which in turn requires banks to reduce its background verification costs. It can be done effectively if the bank knows which factors are key influencers in determining whether the customer will default or not.

Data Analysis

For the data analysis following steps will be followed:

- Exploratory Data Analysis (EDA)
  To get a better understanding of the data, EDA will be conducted
- Data Cleaning and Transformation
  Missing value and outlier will be tackled and log transformation will be used for highly skewed values on some variables.
- Feature Engineering
  Based on the EDA new feature or predictors will be created.
- Collinearity Check
  Collinearity will be checked using correlation and Variance Inflation Factor(VIF).
- Training-test Split
  The data will be split into training set and validation set. And the kaggle test data will be used as a holdout set.
- Prediction
  Various prediction models will be applied on the dataset starting with the simpler models which have higher interpretability.
  Variable selection will be done on the basis of EDA and P-Value.
  Scaling and hyperparameter tuning will be consider while fitting different models.
- Model Evaluation Metric
  For model selection and threshold tuning the ROC-AUC curve will be used.
  And based on the F1 score the model will be evaluated.
- Model Fitting
  After the best model has been selected, the whole training dataset will be used to fit the model and the model performance will be checked on the holdout validation set (Kaggle test set).