Written Proposal of Modeling Approach

March 21, 2025

1. Data Preprocessing
   - Handle missing values using SimpleImputer.
   - Cap the outliers at 99th percentile to reduce their impact.
   - Scale numerical features using RobustScaler to mitigate the effect of outliers.
   - Encode categorical variables using OnHotEncoder.
   - Remove one of the multi-collinear variables.
2. Feature Engineering
   - Evaluate feature importance.
3. Model Selection and Training
   - Train multiple classifiers:
     (a) Logistic Regression (baseline model).
     (b) Random Forest Classifier.
     (c) XGBoost Classifier.
   - Stacking model.

   Note that we performed GridSearchCV for hyperparameter tuning and used StratifiedKFold cross validation to ensure robust performance.
4. Model Evaluation

   Due to the highly imbalanced nature of our dataset where bad loans account for less than 5% we prioritized precision-recall (PR) metrics over traditional accuracy and ROC AUC. While the ROC AUC of 0.81 suggests good general discrimination, it is less informative under class imbalance.

   Instead, we focused on the PR AUC and the recall for the bad loan class. In credit risk modeling, maximizing recall for bad loans is crucial to avoid approving high-risk applicants. You can tolerate rejecting a few good loans, but you can't afford to approve bad ones.