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

   Given the high class imbalance in the dataset (bad loans are $< 5\%$), we evaluated model performance based on precision-recall metrics rather than accuracy or ROC AUC alone. In particular, we focused on the F1 score and PR AUC for the bad loan class to ensure that the model is both effective at catching risky applications (high recall) and not overly conservative (reasonable precision).

   ROC AUC was included for completeness, but due to class imbalance, we relied more heavily on PR AUC and class-wise classification reports.