# 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.
   - Apply Log transformation for skewed variables
   - Remove one of the multi-collinear variables.
2. Feature Engineering
   - Evaluate feature importance using permutation importance.
3. Model Selection and Training
   - Train multiple classifiers:
     (a) Logistic Regression (baseline model).
     (b) Random Forest Classifier.
     (c) XGBoost Classifier.
   - Ensemble model and stacking models.

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

   The dataset is highly imbalanced, with approved rate (majority class) far outnumbering rejected rate (minority class), thus ROC-AUC curve can be misleading as it considers true negatives, which dominates the dataset.

   Precision-Recall Curve: focuses only on positive classes predicitions, making it more informative.