AML Final Project Report - Home Credit Default Risk

By Annika Hsi, Claire Chen, Jessica Marshall, Michael McGrath, Siyuan Kong

I. Exploratory Data Analysis (EDA)

I.i. Overview

We conducted an exploratory analysis of the main dataset (*application_train.csv*) and supplementary datasets (*bureau.csv, bureau_balance.csv, previous_application.csv, POS_CASH_balance.csv, installments_payments.csv, and credit_card_balance.csv*). The primary dataset uses *SK_ID_CURR* as a unique identifier, while supplementary datasets often include multiple records per client, complicating merging.

I.ii. Dataset Exploration

The main dataset *(application_train.csv*) contains 307,511 records with 121 features, with a highly imbalanced target variable (*TARGET*), where only 8.1% are positive cases. Supplementary datasets varied significantly in size, ranging from tens of thousands to millions of rows. Key features were aggregated by *SK_ID_CURR*, deriving metrics like credit counts, payment timeliness, and credit utilization. For example:

- **Bureau Data**: Summarized credit history with metrics like credit counts and average balances.
- **POS_CASH and Installments**: Aggregated payments and overdue counts.
- **Credit Card Transactions**: Captured metrics like late payments and credit utilization.
- **Previous Applications**: Derived features like total applications and approval rates.

I.iii. Feature Aggregation and Missing Values

Features were aggregated to make *SK_ID_CURR* the primary key, with numerical features summarized (e.g., mean, max) and categorical features one-hot encoded. Missing data strategies included:

- **Imputing** minimal missing values (<10%) using medians or modes.
- **Analyzing correlations** for moderate missing values (10%-50%).
- **Dropping features** with high missingness (>50%) to avoid noise.

I.iv. Insights from Visualizations

Visualizations provided valuable insights:

- **Density Plots**: Highlighted separations between positive and negative target classes for features like *EXT_SOURCE_3*.
- **Bar Charts**: Revealed trends in categorical variables, such as higher education levels correlating with lower default rates.
- **Correlation Matrices**: Identified multicollinearity, informing feature selection.

II. Implementation

II.i. Overview

In the previous section, we reviewed the main dataset (*application*) and supplementary datasets (*bureau, bureau_balance, previous_application, POS_CASH_balance, installments_payments, and credit_card_balance*). The main dataset has *SK_ID_CURR* as its primary key, while in supplementary datasets, it is not unique. This creates challenges in merging datasets effectively.

To address this, we considered two approaches:

1. **Merge and Aggregate:** Transform supplementary datasets by aggregating features over *SK_ID_CURR*, then perform a left join with *application_train*. However, this approach resulted in significant missing values, leading to sparsity, low variance, or incorrect distributions during imputation.
2. **Model-Specific Training:** Train separate models for each supplementary dataset, significantly reducing missing data. Predictions from these models are combined using a weighted average based on validation AUROC scores (see **Prediction & Evaluation**). We adopted this approach.

II.ii. Data Preprocessing

We performed two key steps during preprocessing (see *AML_Project_Data_Preprocessing.ipynb*):

1. **Split Training Dataset:** Split *application_train* into training and test subsets (stratified split with 20% as the test set). The resulting datasets (*train.csv* and *test.csv*) were used for model training and evaluation, respectively.
2. **Transform Supplementary Datasets:** Supplementary datasets were aggregated over *SK_ID_CURR* to make it their primary key. Depending on feature types, we applied:
    - **One-Hot Encoding (OHE):** Count occurrences of categorical values per `SK_ID_CURR`.
    - **Time Aggregation:** Extract first, last, and counts for specific time intervals (e.g., 6 months, 1 year, 5 years).
    - **Amount Aggregation:** Compute max, mean, min, and sum for numerical features.

Automated transformations were facilitated by recognizing feature types (*object* for categorical and *int64/float64* for numerical). Features with "DAYS" in their name were categorized as time-based.

For `bureau` and *bureau_balance*, we first aggregated *bureau_balance* over *SK_ID_BUREAU*, merged it into `bureau`, and then transformed *bureau* over *SK_ID_CURR*. Transformed datasets were saved as *.csv* files for use during training and prediction.

II.iii. Training

Separate models were trained for each dataset. The training process involved:

1. **Input Specification:** Define model filename, correlation threshold (to reduce multicollinearity), number of trials, hyperparameter search space, and machine learning pipeline (e.g., XGBoost).
2. **Iterative Training:**
    - Save dataset-specific *SK_ID_CURR* values and preprocess missing data (median imputation for the main dataset and zero for supplementary datasets).
    - Filter features using a correlation threshold to reduce redundancy.
    - Perform hyperparameter tuning using *hyperopt* to maximize AUROC on the validation set.

Trained models and their metadata (preprocessors, hyperparameters, validation AUROC scores) were saved as *.pickle* files for later use.

II.iv. Prediction & Evaluation

Predictions were generated using a weighted average of probabilities from all models, with weights proportional to each model's validation AUROC. Steps included:

1. Split the test dataset into features and labels (`X` and `y`).
2. For each model:
    - Use the corresponding dataset (merged if supplementary) and apply the saved preprocessor.
    - Filter features based on the correlation threshold.
    - Generate predictions, weighted by validation AUROC, only for *SK_ID_CURR* values present in the dataset.
3. Compute the final weighted average of predictions across all models.

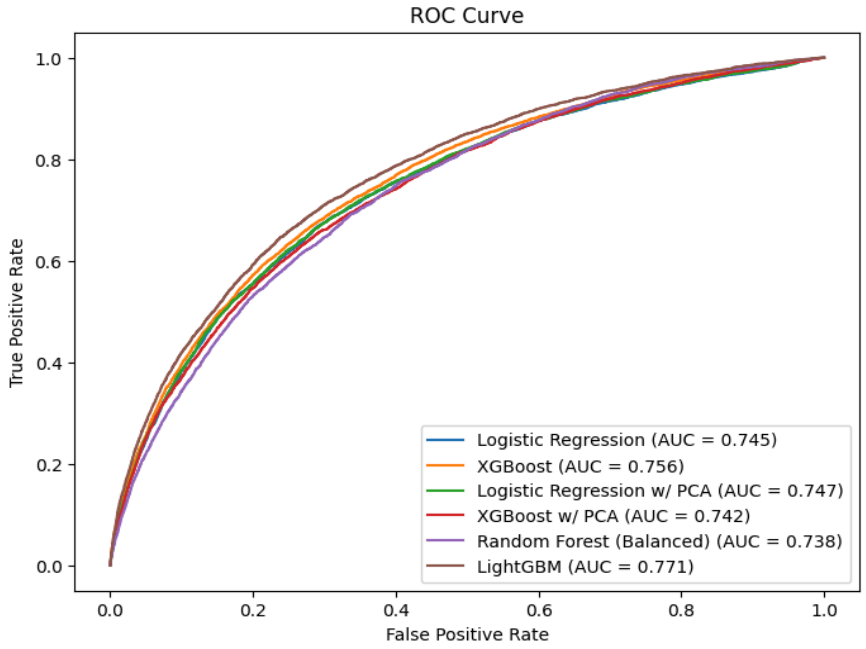
The performance was evaluated using AUROC on the test dataset. Experiments demonstrated that combining models outperformed using the main dataset alone, showcasing the effectiveness of this multi-dataset approach.

III. Results and Best Model

Given the large and highly imbalanced datasets on this binary classification problem, we trained and compared four types of models to identify the best model. To address highly correlated features and high dimensionality from preprocessing, we implemented following techniques:

1. Logistic Regression: Models were trained with and without PCA, using datasets filtered at varying thresholds to exclude highly correlated features.
2. XGBoost: Same approach as Logistic Regression
3. Random Forest - The *class_weight* parameter was set to *balanced* to handle imbalance
4. LightGBM - Handle sparse features and automatically manage class imbalance. The leave wise tree growth is used instead of the level wise tree growth, and feature importance ranking applied.

The best-performing models from each type were selected, and their AUROC curves are plotted below.



In conclusion, LightGBM achieved the highest AUROC score of 0.771 on the validation set. Therefore, LightGBM model is the most suitable for solving this credit risk default problem.