Project Title

Home Credit Default Risk (HCDR) Final Project

Team and Phase Leader Plan

Team Name: FP_Group_3

Phase 3 Leader: Jaden Costa

Team Members
Alicia Aaholm
Nicholas Chappell
Jaden Costa
Katia Torres Sanchez

Credit Assignment Plan

Phase	Task	Person Responsible	Estimated Person-Hrs
Phase 3	Jaden will conduct another round of feature engineering.	Jaden	2 hr
Phase 3	Nicholas will conduct any additional hyperparameter tuning.	Nicholas	2 hr
Phase 3	Alicia will decide on any additional feature selections to be added as well as ensemble methods.	Alicia	1 hr
Phase 3	The team will complete a project update, which will include a 2 minute video presentation, a slide deck, and a Jupyter notebook that will be submitted to the Canvas discussion portal	All	2 hrs
Phase 3	Katia will write up the report following the structure set by the rubric on Canvas for assignment Phase 3.	Katia	1 hr
Phase 3	Katia will upload the slide deck and Jupyter notebook to Canvas.	Katia	10 mins

Project Abstract

During Phase 3 of the HCDR Project, the team is focused on enhancing the models designed in Phase 2. The problems being tackled included models that are not performing at a optimal level, based on metrics such as F1 score and accuracy. The models produced in Phase 2 are not ideal for predicting which clients will default on a home credit loan. Thus, the goal of Phase 3 is to produce better predictions by performing additional feature engineering and hyperparameter tuning. In Phase 3, the team also performed feature selection based on a analysis of the importance of the features. An ensemble method was also used to enhance the model pipelines. The ensemble model has had the best performance. It is better than the logistic regression, tuned random forest, and random forest pipelines. The best score our team has produced is the 0.762 from the ensemble pipeline.

Introduction/Project Description

In Phase 3, the tasks to be tackled include feature engineering using recency, frequency, and monetary value fatures. There was also hyperparameter tuning and feature selction performed. The final task was to use ensemble methods to improve the model pipelines from Phase 2. The data used was from the bureau dataset and application_train dataset, which is further explained in the Data Lineage section of this report.

Data Lineage

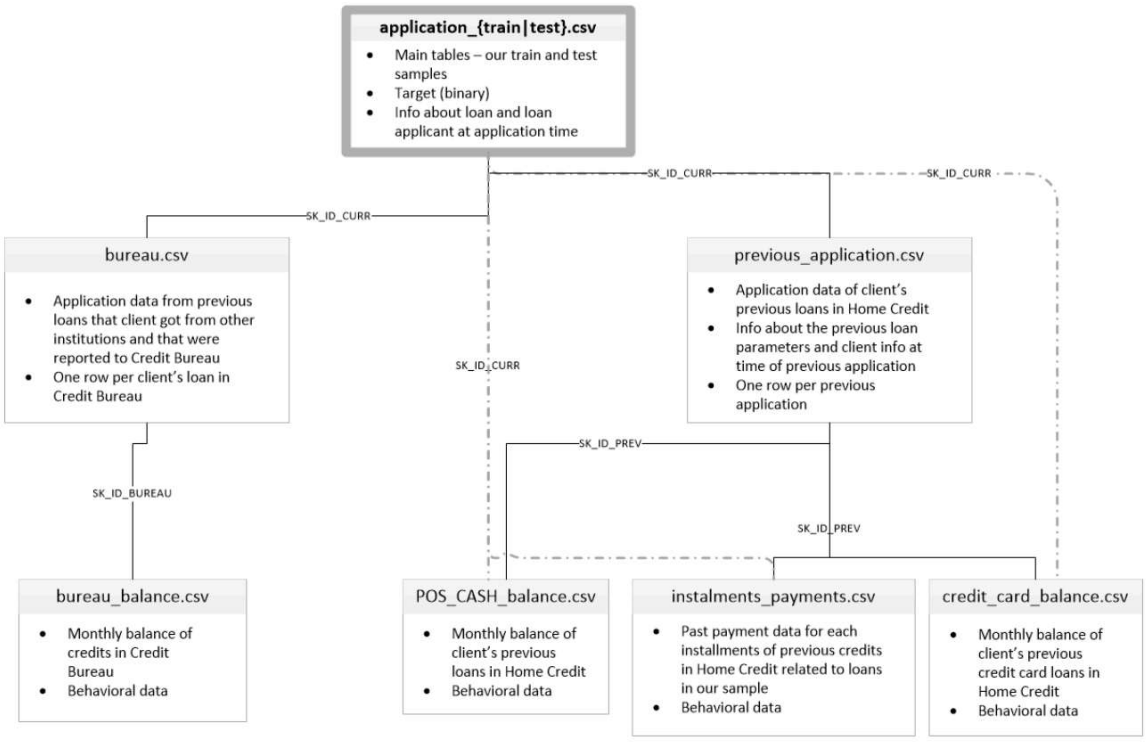
We continued to use the data sets that are provided by the HCDR Kaggle Competition file. During Phase 3, there was a focus on the application_train and bureau datasets. Below is a description of the datasets.

- Application Data: Includes the majority of information regarding clients, such as the gender, income, family status, education, number of children, amount of income, and whether the client owns a car or not. The application data has split the main training data from the testing data.
- Bureau Data: This dataset contains data on clients' credit history.

The datasets were merged in Phase 3 to perform the required tasks.

Workflow of Dataset

The following image is a workflow of all the datasets. However, the team focused on application_train.csv and bureau.csv datasets to create the recency, frequency, and monetary features. The recency feature captured the most recent loan, or max of the DAYS_CREDIT variable. The frequency feature captured the number of past loans. The monetary feature captured total past credit. Any missing values for these features had a impute missing value performed.



There are 221 raw features included in the Data Dictionary for the HCDR Kaggle Project. Below are some key raw features we are focusing on, pulled from the Data Dictionary CSV file. The

identified features are not all-inclusive. The full list of raw features can be found in HomeCredit_columns_description.csv.

Raw Feature	Test Description	Data Type
SK_ID_CURR	ID of loan in the sample	Integer
TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)	Integer (1 for True, 0 for False)
NAME_CONTRACT_TYPE	Identifies if loan is cash or revolving	String/nvarchar
AMT_INCOME_TOTAL	Client's income	Float (rounded to 2 decimal places)
AMT_CREDIT	Loan's credit amount	Float(rounded to 1 decimal place)
AMT_ANNUITY	Loan annuity	Float(rounded to 1 decimal place)
NAME_INCOME_TYPE	Client's income type	String/nvarchar
NAME_HOUSING_TYPE	Client's housing situation	String/nvarchar
AMT_CREDIT_SUM_DEBT	Client's current debt on Credit Bureau credit	Float (rounded to 1 decimal place)
AMT_BALANCE	Balance during the month of previous credit	Float (rounded to 3 decimal places)
REGION_RATING_CLIENT	The rating of the region where client lives (1,2,3)	Integer

Feature Engineering

Feature engineering was performed to improve the predictions the models were producing. It is important to perform feature engineering to provide the training data with additional features that can be used to optimize the model. Appropriate feature engineering can provide a positive impact to the model. However, the features selected must be correlated to the loan default target variable to have a significant impact on the model. In Phase 3, the team added features for recency, frequency, and monetary, as stated in the data lineage section. This approach was chosen becasue RFM metrics tend to be vital indicators of how a customer may behave. As mentioned in the Canvas assignment, each feature can disclose a customer's lifetime value, retention, or engagement. All of which are important to consider when reviewing loan applications.

Hyperparameter Tuning

After completing feature engineering and selecting the Random Forest classifer as the base model, hyperparameter tuning was performed to identify the optimal parameter configuration. The goal of hyperparameter tuning was to imporve the model's generalization performance by systematically exploring combinations of key Random Forest parameters, such as the number of trees, the tree depth, and minimum sample requirements for splitting the nodes.

RandomizedSearchCV was chosen for tuning. This method randomly samples a defined number of hyperparameter combinations from the search space. It provides a balance between computational efficiency and model performance, which is important since the parameter space is larage. The RandomizedSearchCV was configured to evaluate 10 random combinations using 3 fold cross validation. The ROC-AUC is used as the scoring metric.

The tuning process explored variations in n_estimators, max_depth, min_samples_split, and min_samples_leaf. The best performing configuration was selected based on the highest average ROC-AUC score.

Modeling Pipelines

Our modeling pipeline combined feature engineering, model tuning, and ensemble learning to improve loan repayment prediction. We began by creating Recency, Frequency, and Monetary features from the bureau dataset, then merged them into the application data to expand the behavioral signal available to the model. After preprocessing and imputing missing values, we tuned a Random Forest model using RandomizedSearchCV to explore variations in tree count, depth, and split thresholds. This process allowed us to evaluate multiple configurations through cross-validation and select a more stable model. We then introduced a soft-voting ensemble that combined Logistic Regression with the tuned Random Forest to capture both linear patterns and more complex nonlinear relationships. Finally, we reviewed feature importance scores to understand which inputs had the strongest influence on predictions and to validate the usefulness of the engineered RFM features.

Results and Discussion

The best model during Phase 3 was the ensemble model. The hyperparameter used was soft voting. It produced a 0.762 ROC AUC score, which is also higher than the ROC_AUC scores produced during Phase 2. The two lowest performing models used the default hyperparameters. Although the tuned random forest model had the best single model, the ensemble model had a better overall ROC_AUC score. The RFM features lifted the AUC_ROC by over 0.015, showcasing that it did impact the model in a positive manner. Frequency was the top feature.

Visualization of the Model Pipelines Experiment Log

Experiment ID	Model	Hyperparameters	ROC AUC	Notes
1	Logistic Regression	Default	0.720	Baseline
2	Random Forest	Default	0.745	Moderate Overfit
3	Tuned Random Forest	n=200, depth=10	0.755	Best Single Model
4	Ensemble	Soft Voting	**0.762**	Best Overall

Gap Analysis

Although these results represent a meaningful improvement over the baseline Logistic Regression model, they still fall short of the leaderboard benchmark of 0.795. This gap suggests that the pipeline may benefit from more expressive models or richer feature sets, especially around

installment and credit-use behavior. Some likely improvements can gradient boosting and installment featu

Conclusion

This project focused on building an effective machine learning pipeline to predict loan repayment outcomes using the HCDR dataset. Our goal was to test whether custom feature engineering and systematic model tuning could improve predictive accuracy. The results support this hypothesis: RFM features added valuable behavioral insight, and the tuned models consistently outperformed simpler baselines. These findings highlight the importance of feature quality and model selection when working with credit-risk data. While our current best model performs well, we see opportunities for further improvement through gradient boosting methods, deeper hyperparameter searches, and expanded behavioral features. In future work, we will refine these approaches and prepare the model for a more production-ready deployment.