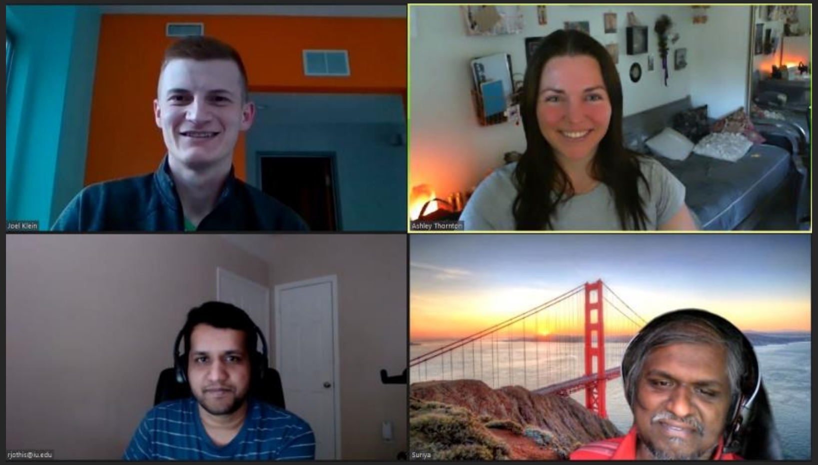


Home Credit Default Risk

Phase 2, Group 9

Rakesh Jothishankar, Joel Klein, Suriyadeepan Narayanasamy, & Ashley Thornton

Contents

Overview

1. Goal: Improve method of approving or declining loan applications.
2. Data: Home Credit data from Kaggle.
3. Methods: Logistic regression, XGBoost, & Light GBM.

Data Prep

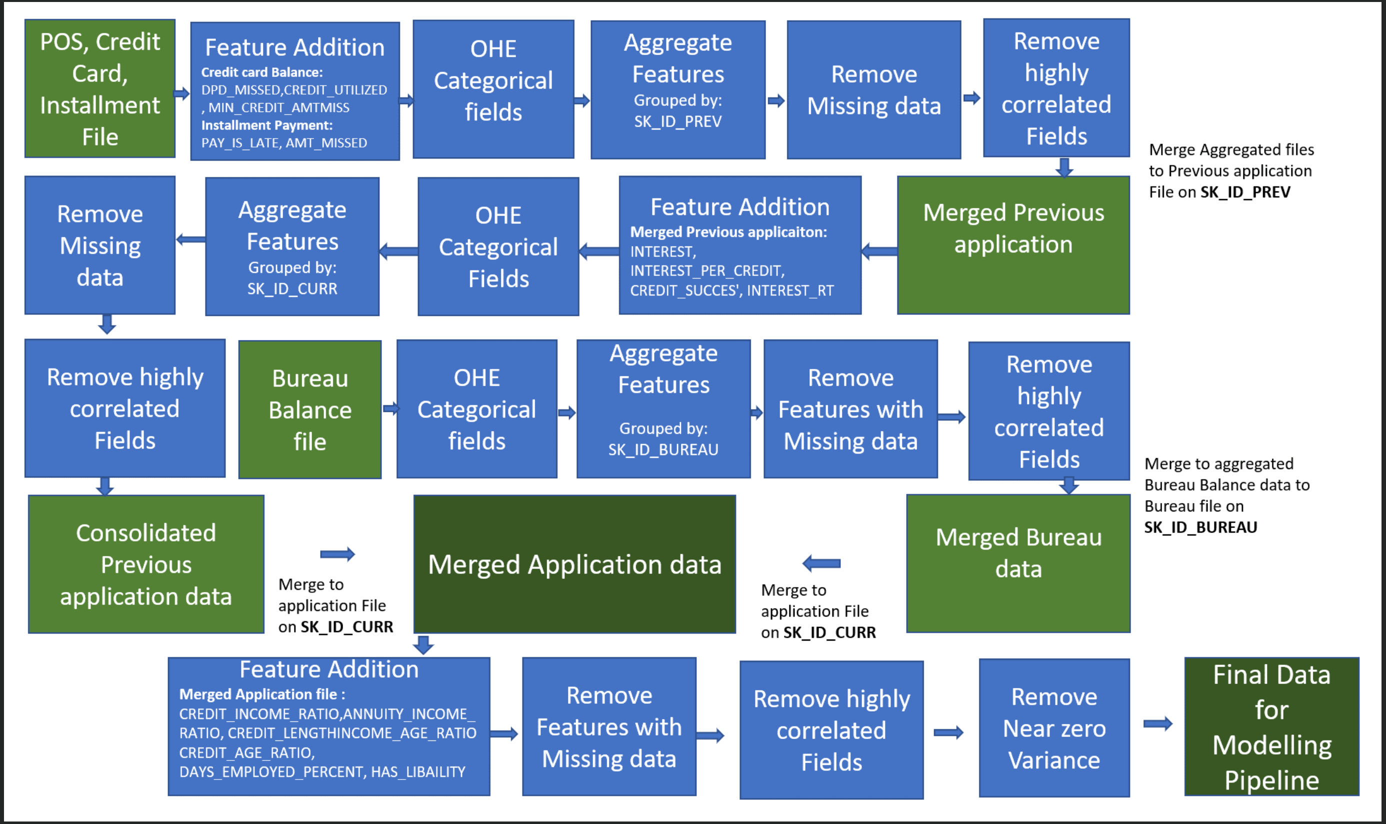
1. POS cash balance, installment payment, and credit card balance files get rolled up to the previous applications file joined by SK_ID_PREV.

2. Bureau balance file gets rolled up to the bureau file joined by SK_ID_BUREAU.

3. Joined previous application data and joined bureau data get rolled up to the applications file on SK_ID_CURR.

4. Features with a large amount of missing data or highly correlated to other features were removed.

Data Prep



POS, Credit Card, Installment File
Feature Addition
Credit card Balance:
DPD_MISSED,CREDIT_UTILIZED , MIN_CREDIT_AMTMISS
Installment Payment:
PAY_IS_LATE, AMT_MISSED
OHE Categorical fields
Aggregate Features
Grouped by: SK_ID_PREV
Remove Missing data
Remove highly correlated Fields
Merge Aggregated files to Previous application File on SK_ID_PREV
Merged Previous application
Feature Addition
Merged Previous applicaiton:
INTEREST, INTEREST_PER_CREDIT, CREDIT_SUCCES', INTEREST_RT
OHE Categorical Fields
Aggregate Features
Grouped by: SK_ID_CURR
Remove Missing data
Remove highly correlated Fields
Bureau Balance file
OHE Categorical fields
Aggregate Features
Grouped by: SK_ID_BUREAU
Remove Features with Missing data
Remove highly correlated Fields
Merge to aggregated Bureau Balance data to Bureau file on SK_ID_BUREAU
Consolidated Previous application data
Merge to application File on SK_ID_CURR
Merged Application data
Merge to application File on SK_ID_CURR
Merged Bureau data
Feature Addition
Merged Application file :
CREDIT_INCOME_RATIO,ANNUITY_INCOME_RATIO, CREDIT_LENGTHINCOME_AGE_RATIO CREDIT_AGE_RATIO, DAYS_EMPLOYED_PERCENT, HAS_LIBAILITY
Remove Features with Missing data
Remove highly correlated Fields
Remove Near zero Variance
Final Data for Modelling Pipeline

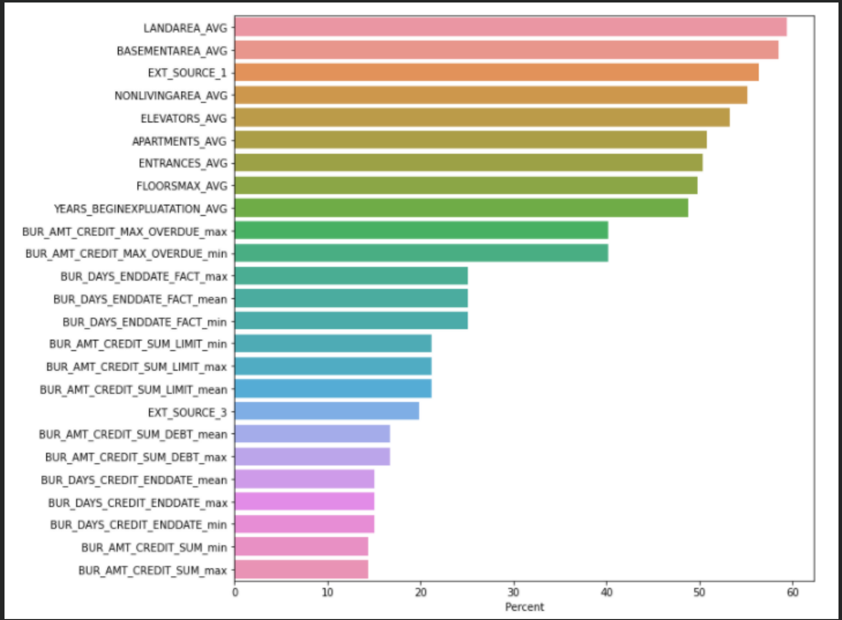
Feature List

Feature Types

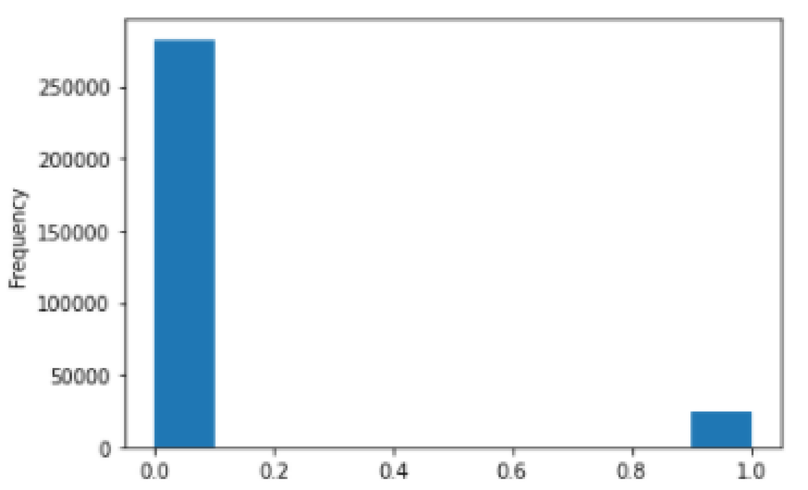
Loan (6)	Surrounding DPD (4)	Previous Application (36)
Date (5)	Document forms (20)	Previous Monthly POS/Cash Balance (6)
Contact info (6)	Credit bureau inquiries (6)	Previous Loan Installment Payments (6)
Family (3)	Demographics (3)	Previous Monthly Credit Card Loan Balance (21)
Region (9)	Occupation (2)	Bureau Previous Credits (15)
External (3)	Process Time (2)	Bureau Previous Credits Monthly Balance (2)
Housing (48)	Other Assets (3)	

New EDA

Features with Most Missing Data



Target Frequency Distribution



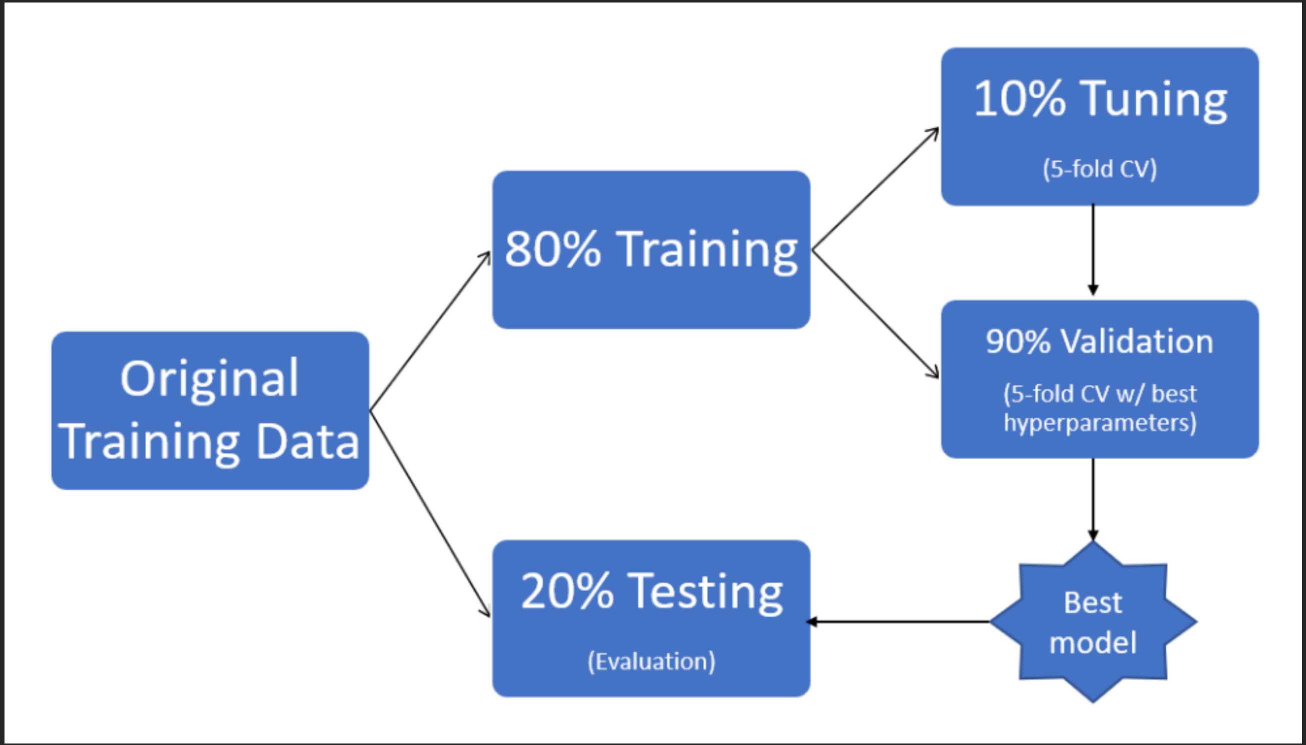
Data Handling Pipeline

1. Imputed missing numeric values with median.
2. Standardized numeric features.
3. Imputed missing categorical values with “Unknown.”
4. OHE categorical features.
5. Featured engineered new features.
6. Removed near zero variance features.
7. Removed features with zero importance from previous model (for some test runs).

New Engineered Features:

- Late payment
- Amount missed
- Credit utilized
- Min credit amount missed
- Interest
- Interest per credit
- Credit success
- Interest rate
- Credit to income ratio
- Annuity to income ratio
- Credit length
- Income to age ratio
- Credit to age ratio
- Percent of days employed
- Liability

Sampling Method



Modeling Pipeline

Models:

- Logistic Regression
- XGBoost
- Light GBM
- Random Forest

Preprocessing:

- All features
- PCA (95% of variance)
- Feature Selection
- Synthetic Minority Over-sampling TEchnique (SMOTE)

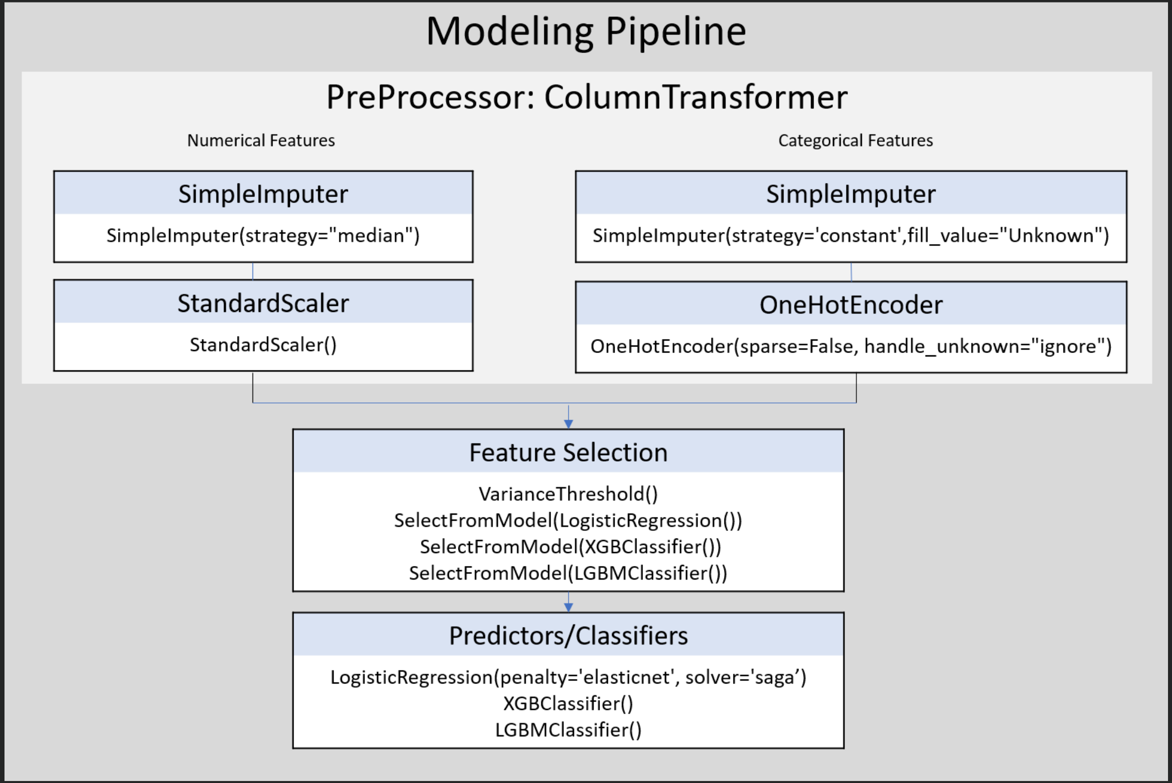
Hyperparameter Tuning:

```
# set logistic parameter grid
logistic_params= {'logistic__C': (100, 10, 1, 0.1, 0.01),
                  'logistic__l1_ratio': (0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1.),
                 }

# set xgboost parameter grid
xgb_params = {'xgb__n_estimators': [300, 500, 700],
              'xgb__learning_rate': [0.01, 0.1],
              'xgb__max_depth': range(3, 10),
              'xgb__colsample_bytree': [i/10.0 for i in range(1, 3)]
             }

# set lightgbm parameter grid
lgbm_params = {'lgbm__boosting_type': ['goss','dart'],
               'lgbm__n_estimators': [3000, 5000, 7000],
               'lgbm__learning_rate': [0.005, 0.001, 0.05, 0.01],
               'lgbm__max_depth': [2, 6, 10],
               'lgbm__colsample_bytree': [0.1, 0.3, 0.5]
              }
```

Modeling Pipeline

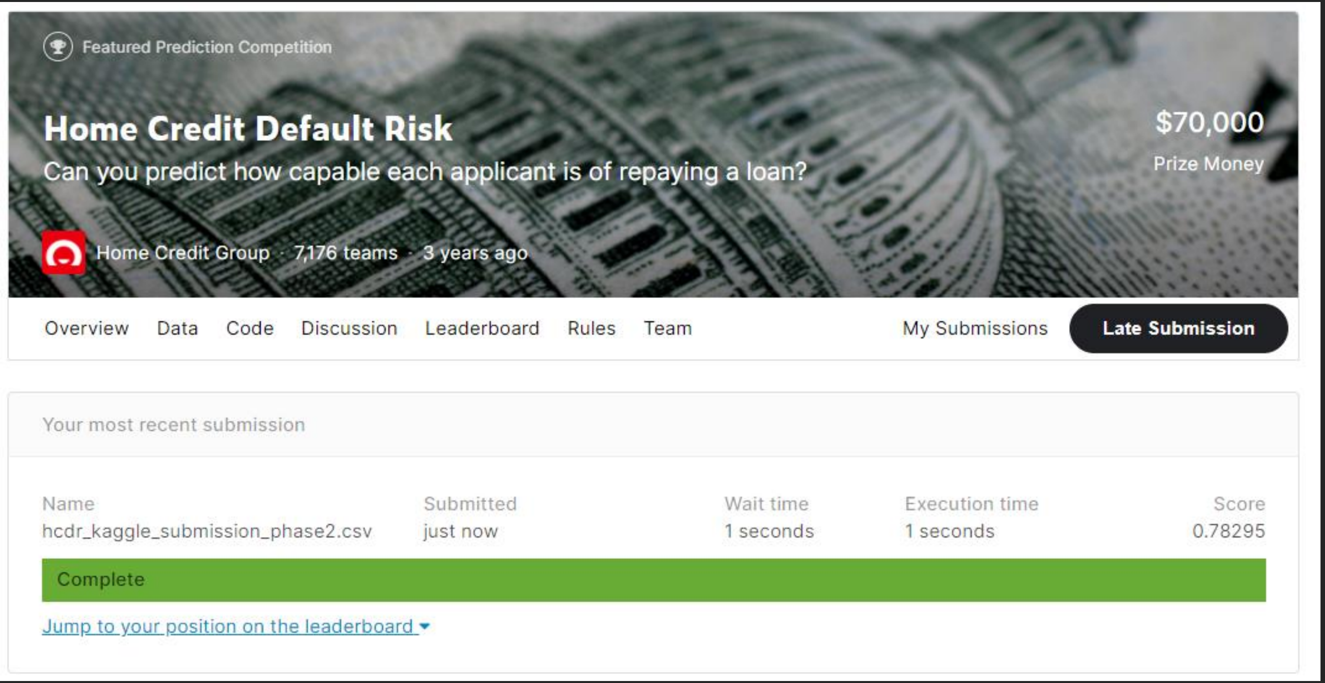


Results

Model	Experiment	Train Accuracy	Test Accuracy	Train Area under ROC	Test Area under ROC	Best Parameters
XGBoost	App, agg prev app & bal, agg bureau & bal data w/ feature selection	91.99	92.02	78.27	78.67	{'xgb__colsample_bytree': 0.1, 'xgb__learning_rate': 0.1, 'xgb__max_depth': 3, 'xgb__n_estimators': 300}
LightGBM	App, agg prev app & bal, agg bureau & bal data w/ feature selection	91.99	92.01	78.02	78.48	{'lgbm__boosting_type': 'goss', 'lgbm__colsample_bytree': 0.1, 'lgbm__learning_rate': 0.005, 'lgbm__max_depth': 2, 'lgbm__n_estimators': 7000}
XGBoost	App, agg prev app & bal, agg bureau & bal data w/ NZV features	91.98	92.03	78.27	78.7	{'xgb__colsample_bytree': 0.1, 'xgb__learning_rate': 0.1, 'xgb__max_depth': 3, 'xgb__n_estimators': 300}
LightGBM	App, agg prev app & bal, agg bureau & bal data w/ NZV features	91.98	92.01	78.03	78.48	{'lgbm__boosting_type': 'goss', 'lgbm__colsample_bytree': 0.1, 'lgbm__learning_rate': 0.005, 'lgbm__max_depth': 2, 'lgbm__n_estimators': 7000}
Logistic Regression	App, agg prev app & bal, agg bureau & bal data w/ NZV features	91.93	91.92	76.56	76.81	{'logistic__C': 0.01, 'logistic__l1_ratio': 0.2}
XGBoost	App, agg prev app & bal, agg bureau & bal data w/ PCA	91.93	91.95	75.26	75.36	{'xgb__colsample_bytree': 0.2, 'xgb__learning_rate': 0.1, 'xgb__max_depth': 3, 'xgb__n_estimators': 300}
Logistic Regression (Baseline)	All application data features	91.91	91.93	74.17	74.48	{'logistic__C': 0.1, 'logistic__l1_ratio': 0.6}

Kaggle Submission

Place: 3,834
out of 7,176



Issues

1. Size of data
2. Time constraint
3. Resource constraints

Conclusion & Next Steps

- Past
- Defined project, performed EDA, and ran baseline model.
- Present
- Improved model through hyperparameter tuning and additional algorithms.
- Future
- Neural Network with PyTorch
- Perceptron and SVM