Luddy School of Informatics, Computing, and Engineering

Phase 3- HOME CREDIT DEFAULT RISK- GROUP 06

Deepak Kasi Nathan
(dekasi@iu.edu)

Sai Sumanth Muvva
(saimuvva@iu.edu)

Viswa Suhaas Penugonda
(vpenugon@iu.edu)

Teja Naidu Chintha
(tnchinth@iu.edu)

INDIANA UNIVERSITY BLOOMINGTON

Presentation overview

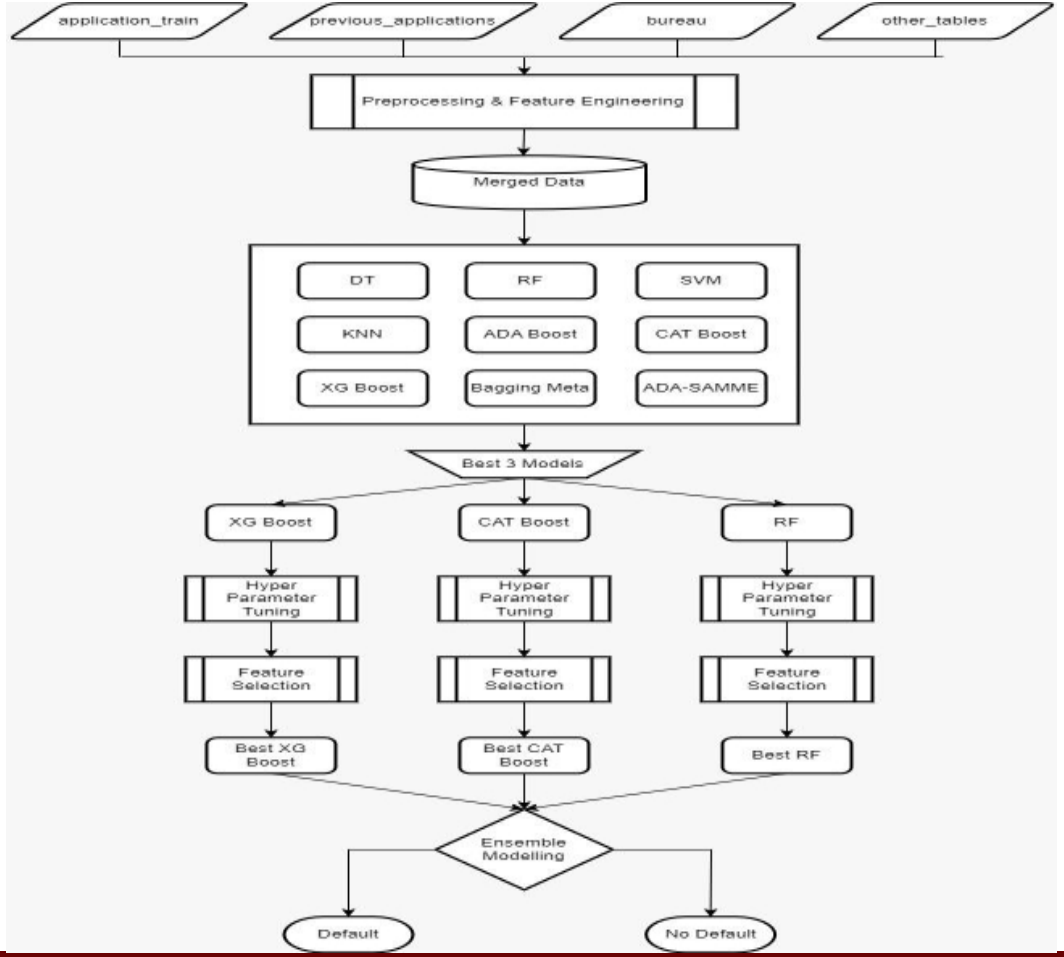
- Project Description
- Project Workflow
- Visual EDA
- Feature Engineering and Selection
- Hyperparameter Tuning
- Machine Learning Pipelines & Results
- Discussion and Future Work
- Conclusion

Project Description

In this project, we aim to predict the probability of default for Home Credit clients based on various features derived from historical data. Home Credit provides loans to clients but faces challenges in assessing the creditworthiness of clients with little or no credit history. Our primary objective is to use historical data from multiple sources and construct a robust machine learning model that can accurately predict the risk of default. To achieve this, we pre-processed and performed feature engineering, conducted EDA, and experimented with a range of machine learning algorithms such as logistic regression, random forests, KNN, decision trees, and ensemble methods like voting and stacking classifiers.

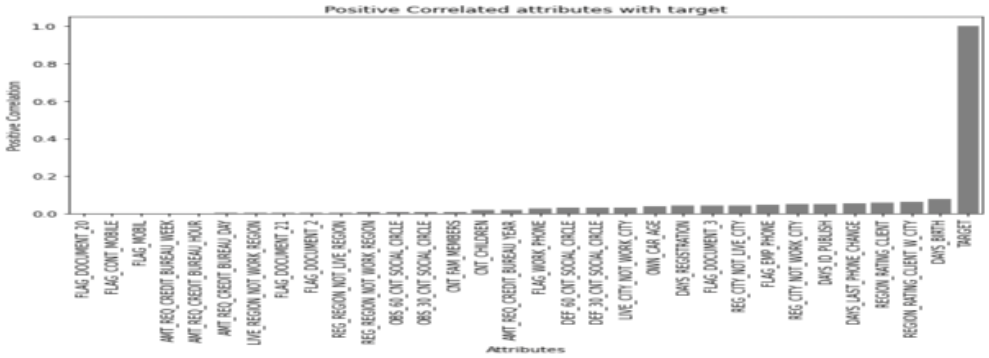
We fine-tuned these models using hyperparameter optimization and feature selection to select the best performing model. Our experiments involved comparing these models' performance and identifying the most effective pipeline. Ensemble learning methods, specifically voting and stacking classifiers, along with tuned random forests, demonstrated the highest test scores in metrics such as F1 score (0.6943) and AUC (0.7610). By implementing the best model, Home Credit will be able to make more informed lending decisions, minimize unpaid loans, and promote financial services for individuals with limited access to banking, ultimately fostering financial inclusion for underserved populations.

WORKFLOW


application_train
previous_applications
bureau
other_tables
Preprocessing & Feature Engineering
Merged Data
DT
RF
SVM
KNN
ADA Boost
CAT Boost
XG Boost
Bagging Meta
ADA-SAMME
Best 3 Models
XG Boost
CAT Boost
RF
Hyper Parameter Tuning
Hyper Parameter Tuning
Hyper Parameter Tuning
Feature Selection
Feature Selection
Feature Selection
Best XG Boost
Best CAT Boost
Best RF
Ensemble Modelling
Default
No Default

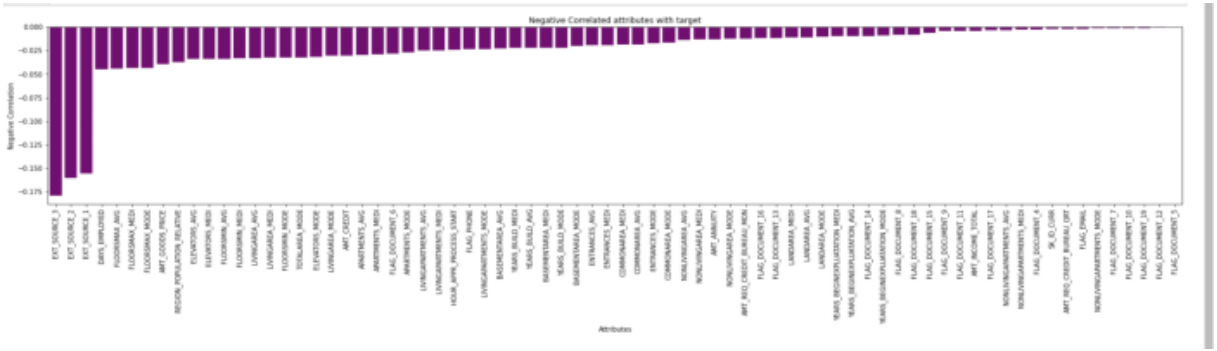
Visual Exploratory Data Analysis

Visual Exploratory data Analysis performed on Categorical values to understand their significance in data



The graph depicts the column features which are Positively correlated based on target

Visual Exploratory Data Analysis



The graph depicts the column features which are Negatively correlated based on target.

Feature Engineering & Selection

Feature Engineering is done on “previous_application”, “installments_payments”, and “Credit_card_balance tables.

Features selected in “previous_application” are "AMT_APPLICATION", "AMT_CREDIT", "AMT_ANNUITY", "approved_credit_ratio", "AMT_ANNUITY_credit_ratio", "Interest_ratio", "LTV_ratio", "SK_ID_PREV", "approved”.

Features selected in “installments_payments” are "DAYS_INSTALMENT_DIFF", "AMT_PATMENT_PCT”.

Features selected in “AMT_DRAWINGS_PCT”, AMT_DRAWINGS_ATM_PCT, AMT_DRAWINGS_OTHER_PCT, AMT_DRAWINGS_OTHER_PCT, MT_PRINCIPAL_RECEIVABLE_PCT.

These features have been engineered and selected due to their high correlation with the target variable.

Hyper-parameter Tuning

Performed hyperparameter tuning for a Random Forest Classifier using GridSearchCV from sklearn.ensemble.

Defined a function ConductGridSearch that takes the training and test data as input, and then runs a grid search with different parameter combinations for the Random Forest Classifier.

Computed the training accuracy of the best estimator, and then used the score method of the best estimator to compute the test accuracy.

Machine Learning Pipelines & Results

Model Name	Basis	Test Accuracy	Test AUC	F1 Score
Baseline LR	Undersampled data	0.7777	0.7477	0.3636
Baseline LR	Undersampled data - Selected Features	0.6904	0.7535	0.69
KNN	KNN with undersampled data-2 124 features	0.6184	0.655	0.6226
Decision Tree	Undersampled data 124 features	0.6591	0.7129	0.673
Random Forest	Undersampled data (124 features)	0.6661	0.7273	0.664
Extra Trees	Undersampled data (124 features)	0.6482	0.7023	0.6536
Bagging Meta Estimator	Undersampled data (124 features)	0.6446	0.6983	0.6164
ADABoost SAMME	Undersampled data (124 features)	0.695	0.7594	0.6948
CATBoost	Undersampled data (124 features)	0.6926	0.7591	0.6906
CatBOOST -Feature & HyperParameter Tuning	CatBOOST Tuned with x>0 116 features	0.6801	0.7409	0.6816
Random Forest -Feature & HyperParameter Tuning	Random Forest Tuned with x>0 116 features	0.6801	0.7409	0.6816
Random Forest -Feature& HyperParameter Tuning	Random Forest Tuned with x>0.1 19 features	0.6631	0.7218	0.6648
Random Forest -Feature & HyperParameter Tuning	Random Forest Tuned with x>0.005 58 features	0.6808	0.743	0.6822
Ensemble Learner - Voting Classsifier	Tuned and selected XgBoost, catboost, random forest (124 features)	0.6948	0.761	0.6943
Ensemble Learner - Stacking Classifier	Tuned and selected XgBoost, catboost, random forest (124 features)	0.6967	0.7629	0.6958

Discussion & Future Work

Bagging meta estimator (Model 8) shows overfitting with high training accuracy (0.9844) and F1 score (0.9842), but poor validation (accuracy: 0.6477, F1 score: 0.6210) and test performance (accuracy: 0.6446, F1 score: 0.6164).

KNN (Model 3) and SVM (Model 4) underfit the data, displaying lower accuracy and F1 scores on both training and validation sets (e.g., KNN: training accuracy 0.6950, F1 score 0.6992; validation accuracy 0.6155, F1 score 0.6205).

Ensemble learners (voting classifier - Model 22, stacking classifier - Model 23) and tuned random forests (Models 18, 19, and 20) outperform other models, achieving higher accuracy, AUC, and F1 scores.

Further optimization and tuning of the promising ensemble learners and tuned random forest models could lead to improved performance in terms of accuracy, AUC, and F1 scores.

Conclusion

In this project, we aimed to predict the probability of default for Home Credit clients based on historical data. Our objective was to construct a robust machine learning model that could accurately predict the risk of default and help Home Credit make informed lending decisions. We performed pre-processing, feature engineering, EDA, and experimented with a range of machine learning algorithms. Our experiments help compare these models' performance and identify the most effective pipeline which helps Home Credit make more informed lending decisions, minimize unpaid loans, and promote financial services for individuals.

In ensemble learning methods (Model 22), which used a Voting Classifier to combine the predictions of (XGBoost, CatBoost, and Random Forest) achieved the highest test F1 score (0.6943), which outperformed all other models. This model also has the highest test AUC (0.7610), which suggests that it was better at distinguishing between positive and negative classes. And the results of this phase suggest that ensemble learning methods may be effective for this classification problem. In future phases, we will perform further tuning and optimization involving neural networks which could potentially lead to even better results.

Thank you

INDIANA UNIVERSITY BLOOMINGTON