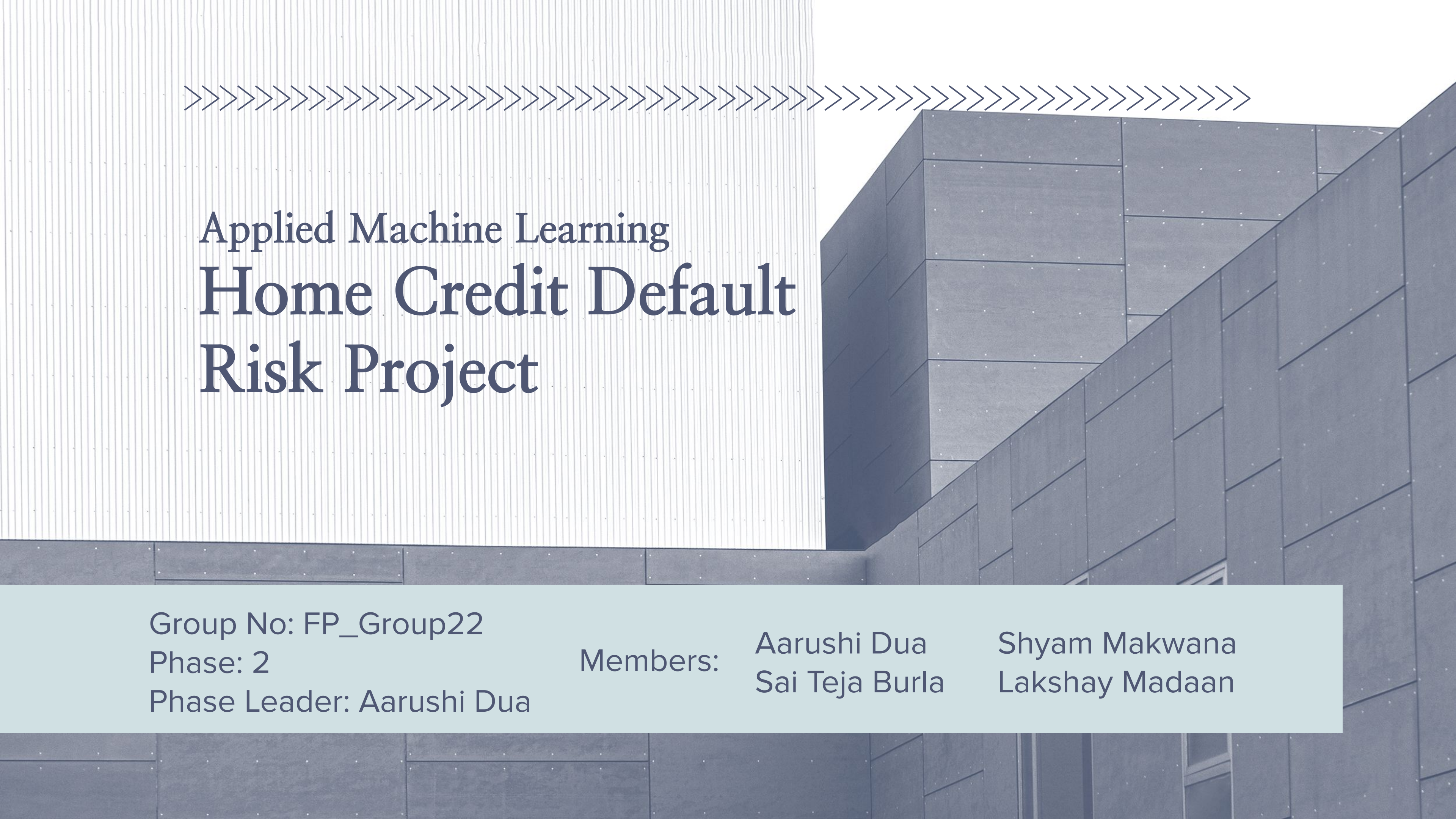


Applied Machine Learning

Home Credit Default Risk Project

Group No: FP_Group22
Phase: 2
Phase Leader: Aarushi Dua

Members:
Aarushi Dua
Sai Teja Burla
Shyam Makwana
Lakshay Madaan

01 Team Names and Photo

02 Project Description

03 EDA

04 Modelling Pipeline

05 Results

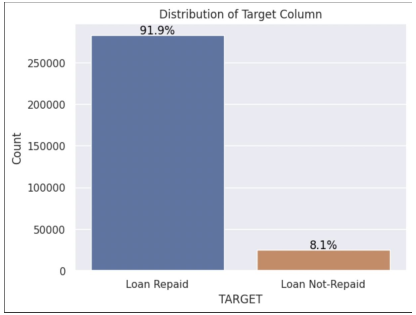
06 Conclusion and Next Steps

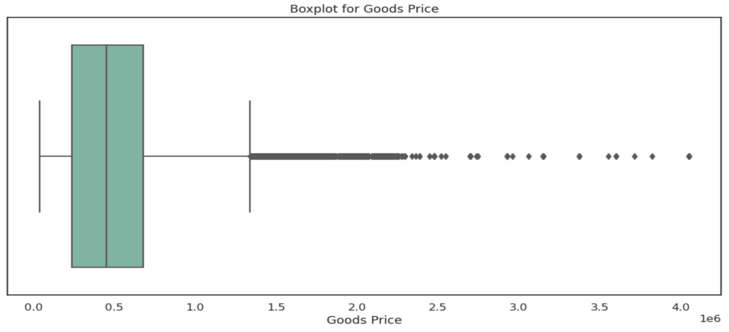
Team Names and Photo

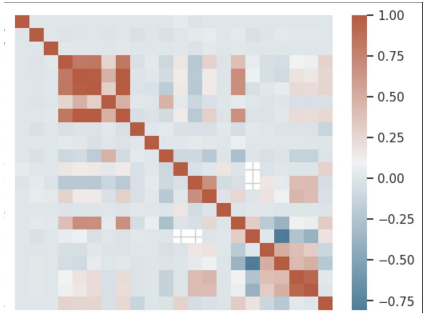
Project Description

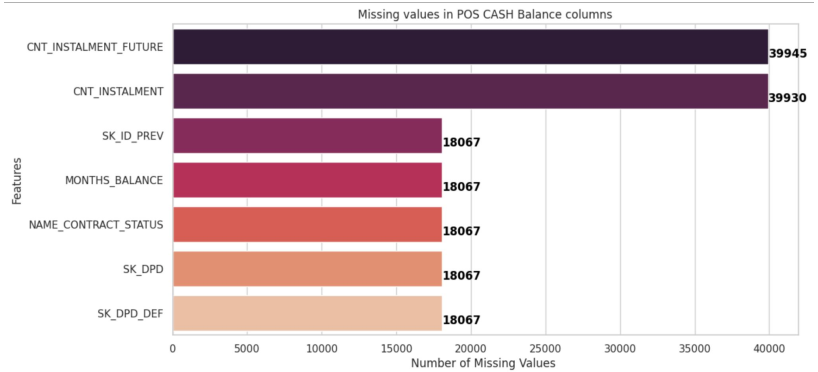
- The objective of the HCDR project is to assess a borrower's capacity to repay a loan, based on the features in the dataset
- To accomplish this objective, we will examine multiple factors of the applicant’s profile in addition to their credit history
- During this phase, we conducted exploratory data analysis on all the 7 datasets
- After EDA, we experimented with table features for model building, and then implemented machine learning pipelines utilizing multiple classifiers
- We also took note of the accuracy on train and test, the AUC score for train, test and validation sets and also plotted an ROC curve for comparing the test AUC scores of different algorithms

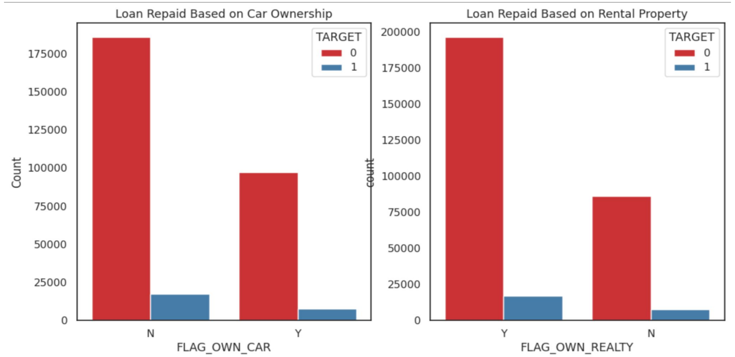
Exploratory Data Analysis







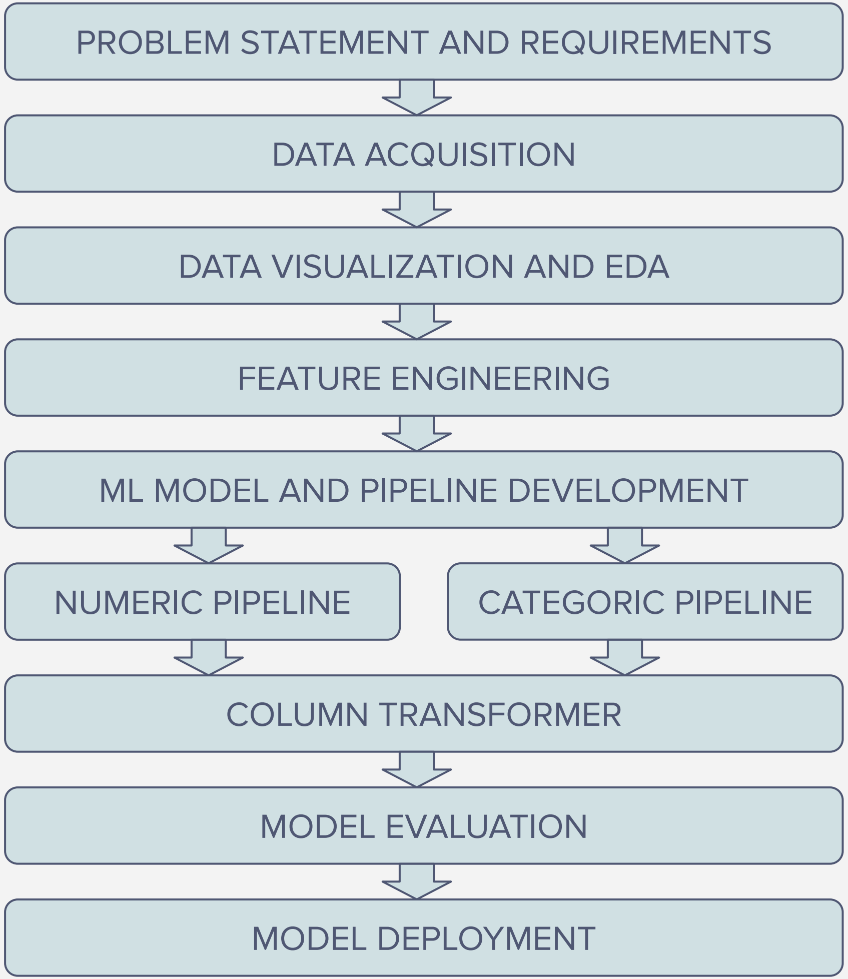




Modelling Pipeline

- Final set of features were selected after performing EDA and visual EDA.
- Separate imputer pipelines were developed for numerical and categorical data to handle missing values and numeric transformation in each data type.
- For our baseline pipeline, we did feature union for pre-processing the numerical and categorical features.
- The various Machine Learning algorithms that we experimented with are as follows:
  - Gaussian Naive Bayes
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier

Pipeline Overview



PROBLEM STATEMENT AND REQUIREMENTS

DATA ACQUISITION

DATA VISUALIZATION AND EDA

FEATURE ENGINEERING

ML MODEL AND PIPELINE DEVELOPMENT

NUMERIC PIPELINE

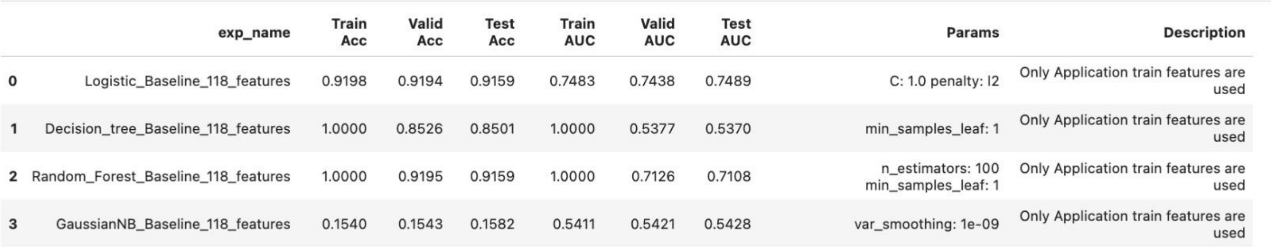
CATEGORIC PIPELINE

COLUMN TRANSFORMER

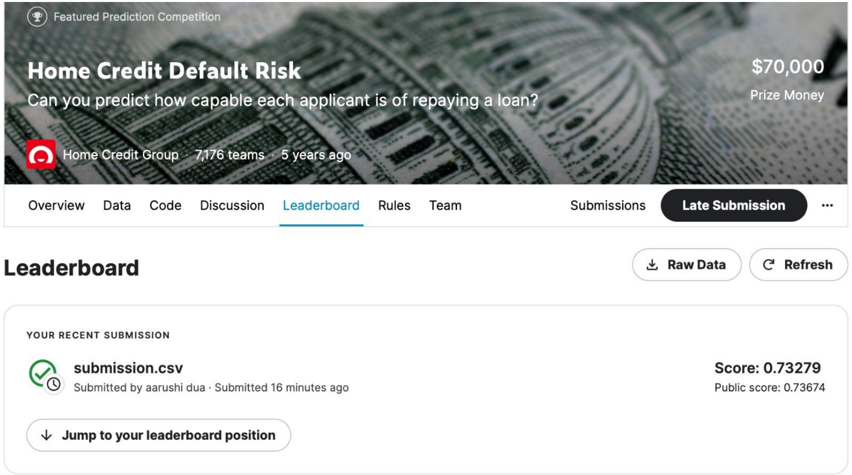
MODEL EVALUATION

MODEL DEPLOYMENT

Results



	exp_name	Train Acc	Valid Acc	Test Acc	Train AUC	Valid AUC	Test AUC	Params	Description
0	Logistic_Baseline_118_features	0.9198	0.9194	0.9159	0.7483	0.7438	0.7489	C: 1.0 penalty: l2	Only Application train features are used
1	Decision_tree_Baseline_118_features	1.0000	0.8526	0.8501	1.0000	0.5377	0.5370	min_samples_leaf: 1	Only Application train features are used
2	Random_Forest_Baseline_118_features	1.0000	0.9195	0.9159	1.0000	0.7126	0.7108	n_estimators: 100 min_samples_leaf: 1	Only Application train features are used
3	GaussianNB_Baseline_118_features	0.1540	0.1543	0.1582	0.5411	0.5421	0.5428	var_smoothing: 1e-09	Only Application train features are used



Featured Prediction Competition

Home Credit Default Risk

Can you predict how capable each applicant is of repaying a loan?

$70,000 Prize Money

Home Credit Group · 7,176 teams · 5 years ago

Leaderboard

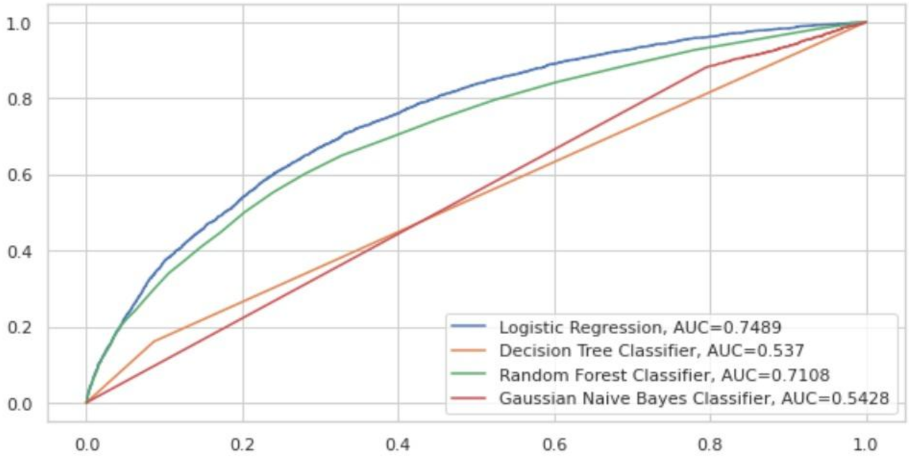
YOUR RECENT SUBMISSION

submission.csv

Submitted by aarushi dua · Submitted 16 minutes ago

Score: 0.73279

Public score: 0.73674



Conclusion and Next Steps

- EDA was used to understand and identify significant features which were then used to implement the modelling pipelines
- We created the ML pipelines by experimenting with various classifiers
- We observed that there might be underfitting in Gaussian Naive Bayes and overfitting in Decision Tree and Random Forest implementations.
- In the next phase, we intend to implement advance feature engineering techniques and perform hyperparameter tuning for our existing models.
- We aim to gain a more comprehensive understanding of the data, which will help us to determine the most suitable model and further enhance the evaluation metrics.

Thank You!!!