

# Applied Machine Learning Final Project: Home Credit Default Risk

Indiana University

December 2022

## **Group 11**

Anuj Mahajan

Shubham Jambhale

Shashwati Diware

Siddhant Patil

# Team Members:

Shubham Jambhale  
[sjambhal@iu.edu](mailto:sjambhal@iu.edu)



Siddhant Patil  
[sidpatil@iu.edu](mailto:sidpatil@iu.edu)



Anuj Mahajan  
[anujmaha@iu.edu](mailto:anujmaha@iu.edu)



Shashwati Diware  
[sdiware@iu.edu](mailto:sdiware@iu.edu)



# Contents

- Four P's
- Project Description
- Summary EDA
- Overview of modeling Pipelines
- Results and discussion (Accuracy, AUC, Kaggle)
- Conclusion and Next steps

# Four P's

- **Past:**

- We began the HCDR Project, which uses a variety of financial and nonfinancial variables to determine whether borrowers will fail or not.
- In the initial stage, we performed fundamental EDA and data collection.
- We decide on the baseline models that will be applied in Phase 2 ( Decision Tree, Random Forest, and Logistic Regression)

- **Present:**

- First, we chose more pertinent features and imputed missing values from them.
- We used these datasets to develop baseline models (Logistic Regression, Decision Tree, and Random Forest).
- AUC and the Confusion Matrix were used to assess the accuracy.
- We calculated loss using log loss.

# Four P's

- **Planned:**

- We will tune the hyperparameters of the models, and we will try to find the best parameters using Grid Search.
- In the steps above, we used the best parameters for each model, to evaluate the models.
- We shall attempt to carry out extensive feature engineering to enhance the results.

- **Problems:**

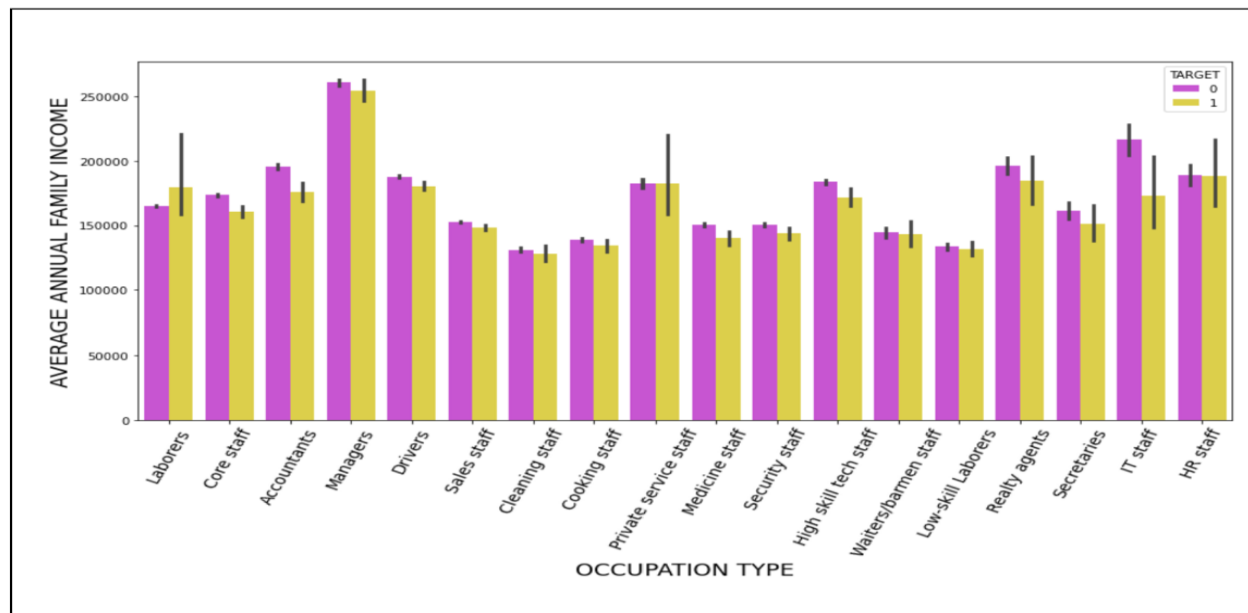
- With the given feature selection, we could not improve the test accuracy or AUC of the baseline model beyond a certain level.
- Even if the decision tree model's AUC is on the upper side, log loss is still significant.
- We may need some prior knowledge about the data, for example, credit data

# Project Description:

- The object of the Home Credit Default Risk (HCDR) project is to predict the repayment abilities of the financially under-served population.
  - The well-established prediction is necessary for both Home Credit and borrowers.
  - Lend money to whom can pay back and give them a chance to build credit.
- We use a variety of data, such as past credit information, credit type, past credit days left, payments, past application information, etc.
  - We use categorical and numerical data to improve the accuracy of our predictions.
  - We provide appropriate datasets for machine learning models using EDA.
- We trained and assessed a number of potential models before selecting the best one.
  - Our potential models include Random Forest, Decision Making Trees, and Logistic Regression.
  - Accuracy, AUC, Confusion Matrix, and Log Loss are just a few of the measures we employ to evaluate candidate models specifically.
- We observed that the Random Forest baseline pipeline has the highest test accuracy which is then followed by the Logistic Regression and Decision-making Tree.

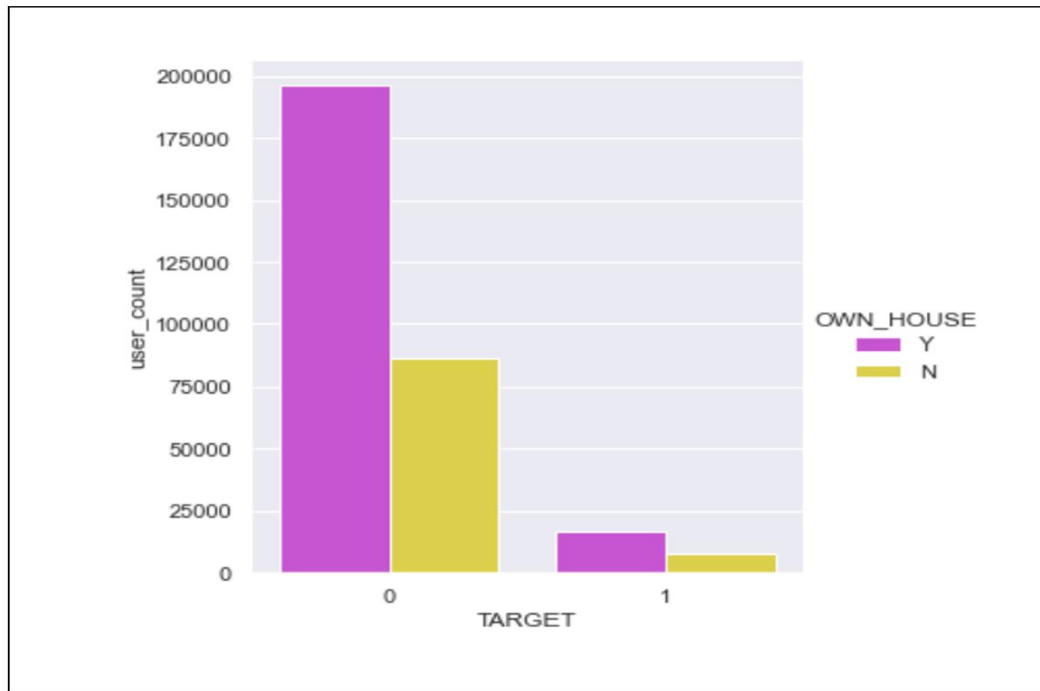
# Exploratory Data Analysis (EDA):

- We perform EDA and examine the following attributes of the data
  - Test Description, Dataset size, Summary Statistics, Correlation analysis, Checking missing values, etc.
- Some Interesting EDA findings:
  - Occupation Type Vs Annual Family Income

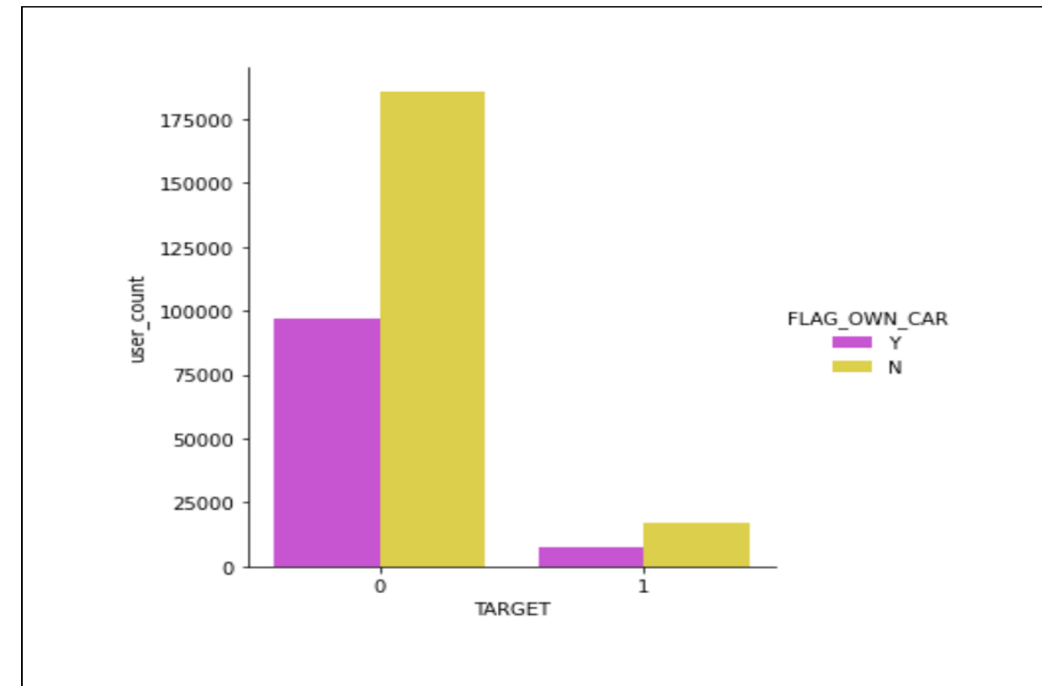


# Exploratory Data Analysis (EDA):

Defaulters and Non-defaulter in terms of Home Ownership



Defaulters and Non-defaulter in terms of Car Ownership





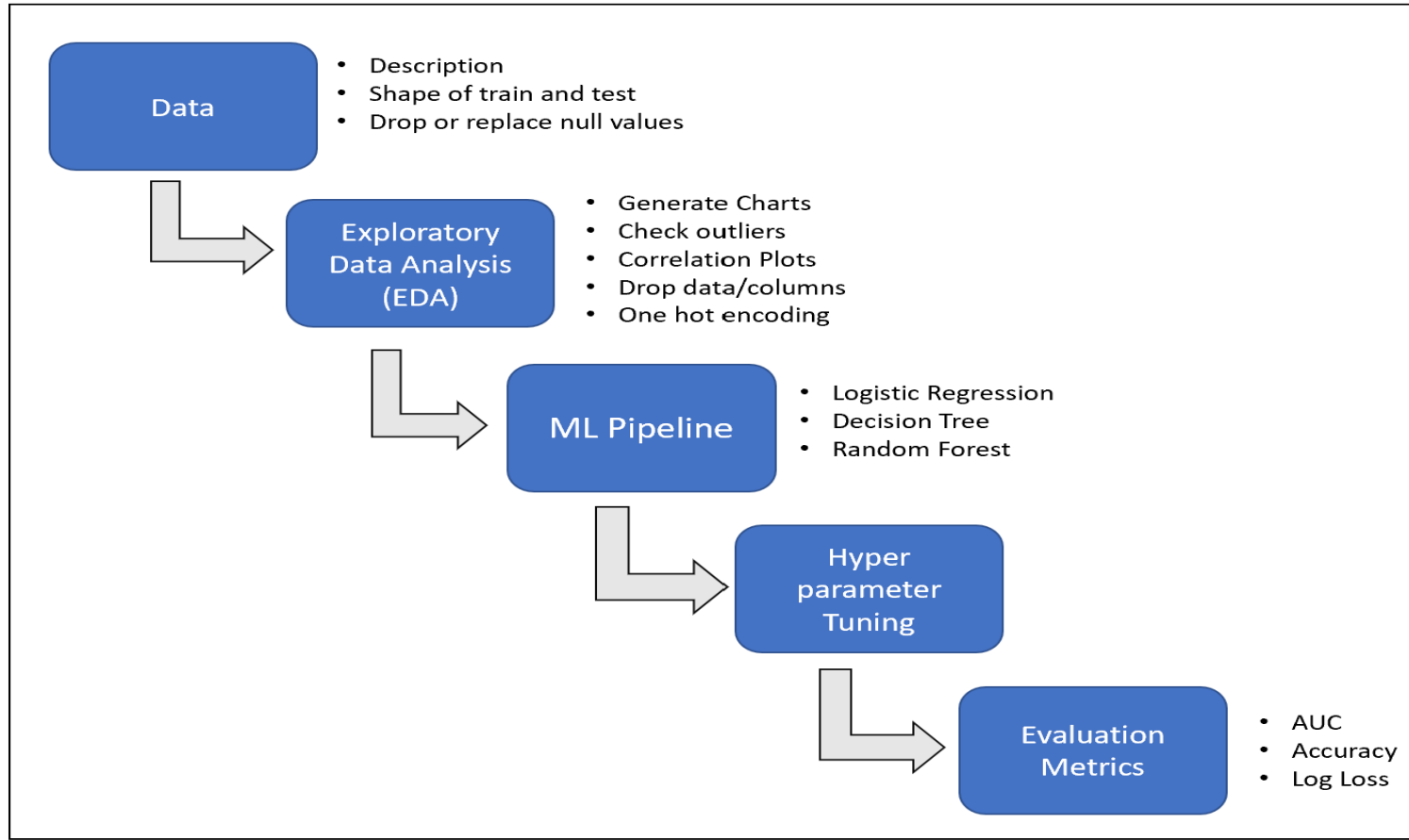
# Feature Engineering:

- Feature selection and imputation
  - We discarded features with missing values of more than 50% and dropped irrelevant features.
  - We filled in the missing values of selected features.
- Adding more relevant data
  - We decided to add relevant features based on our prior knowledge.
  - e.g. Salary-to-Credit Ratio, Total External Source

# Modelling Pipeline:

- The goal is to predict whether the borrower is a defaulter or not. Phase 2 involves creating baseline models (Logistic Regression, Decision Tree, Random Forest)
  - Separate the dataset into training and testing data
  - Prepare the input dataset by converting missing values and scaling the data.
  - Perform a baseline model.
  - Analyze the baseline model using measures for accuracy, AUC, and log loss.
  - Perform steps 1–4 for all the baseline models.

# Modelling Pipeline Flow:



# Result and Discussion:

- Logistic Regression performs well on the given dataset.
  - AUC and Accuracy (91.9) for Logistic Regression are on the higher side.
- Random Forest outperforms baseline Logistic Regression and Decision Tree.
  - It provides the highest accuracy of 92.2.
- Decision Tree seems to have low test accuracy of 86.1.
  - It might be due to the short depth of the decision tree.
  - In turn AUC has increased significantly.

	ExpID	Cross fold train accuracy	Test Accuracy	Validation Accuracy	AUC	Accuracy	Loss	Train Time(s)	Test Time(s)	Validation Time(s)	Experiment description
0	Baseline with 81 inputs	92.0	91.9	91.7	0.506443	91.917467	0.246888	12.5042	0.0495	0.0410	Selective Features - Baseline LogisticRegression
1	Baseline Decision Tree with 81 inputs	86.4	86.1	86.3	0.570952	86.148643	2.362000	29.5238	0.0733	0.0594	Selective Features - Baseline Decision Tree
2	Baseline Random Forest with 81 inputs	92.2	92.2	92.0	0.522292	92.187373	0.270439	245.7127	1.2979	1.0884	Selective Features - Baseline Random Forest

# Result and Discussion:

OverviewDataCodeDiscussionLeaderboardRulesTeam

SubmissionsLate Submission...




## Submissions

You selected 0 of 2 submissions to be evaluated for your final leaderboard score. Since you selected less than 2 submission, Kaggle auto-selected up to 2 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.

☒ Submissions evaluated for final score

AllSuccessfulSelectedErrors

Recent ▾

Submission and Description	Private Score ⓘ	Public Score ⓘ	Selected
<div> <b>output4.csv</b> Complete (after deadline) · now</div>	<b>0.50048</b>	<b>0.50122</b>	<input type="checkbox"/>
<div> <b>output3.csv</b> Complete (after deadline) · 1s ago</div>	<b>0.53589</b>	<b>0.52885</b>	<input type="checkbox"/>
<div> <b>output2.csv</b> Complete (after deadline) · 3h ago</div>	<b>0.50195</b>	<b>0.50271</b>	<input type="checkbox"/>

# Conclusion and Next steps:

- In Phase 2 we conclude that the decision tree model cannot outperform the other baseline model with the lowest test accuracy.
- Even though the test accuracy is low, for Decision Tree AUC is highest.
- Out of all the baseline models, the random forest model performs the best followed by logistic regression.
- In the future, we intend to put all models into practice in phase 3 by performing below steps:
  - Additional Feature Engineering
  - Hyper-parameter Tuning