Home Credit Default Risk (HCDR)

Group 16
Aravind Sheru
Sai Charan Chintala
Seongbo Sim
Yun Joo An

Abstract

HomeCredit offers unsecured lending based on past credit history, repayment patterns and alternate data of the users using Machine Learning Modelling. Credit history is a measure explaining the credibility of a user generated using parameters like average/min/max balance maintained by the user, Bureau scores reported, salary etc and repayment patterns. As a part of this project, we use the datasets provided by kaggle to perform exploratory data analysis, build machine learning pipelines and evaluate the models across several evaluation metrics for a model to be deployed. In phase 1, we provide baseline logistic regression pipelines. We experimented with a baseline pipeline, 50000 non-defaulters balanced dataset, and 75000 non-defaulters balanced dataset. A baseline pipeline has high test accuracy, 91.9, but low AUC, 0.5. After balancing the dataset, 50000 non-defaulters dataset shows 71.4% of test accuracy and 0.62 of AUC. 75000 non-defaulters dataset shows 76% of test accuracy and 0.57 of AUC.

Data and Task Description

1. Data source

We are planning to use the existing datasets provided by Kaggle.
Source: https://www.kaggle.com/c/home-credit-default-risk/data

POS_CASH_balance.csv
This dataset gives information about previous credits information such as contract status, number of installments left to pay, DPD(days past due), etc… of the current application

bureau.csv
This dataset gives information about type of credit, debt, limit, overdue, maximum overdue, annuity, remaining days for previous credit, etc…

bureau_balance.csv
This dataset gives information about Status of Credit Bureau loan during the month, Month of balance relative to application date, Recoded ID of Credit Bureau credit

credit_card_balance.csv
This dataset gives information about financial transactions aggregated values such as amount received, drawings, number of transactions of previous credit, installments, etc…

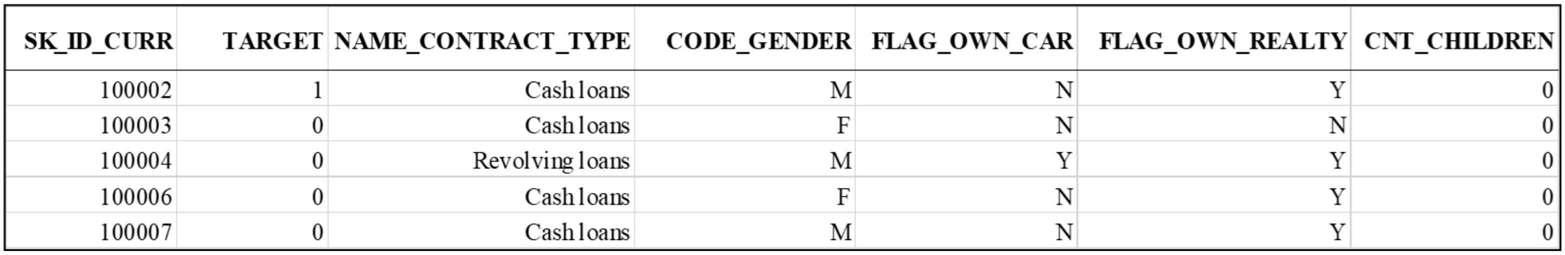
installments_payments.csv
This dataset gives information about payments, installments supposed to be paid and their details.

previous_application.csv
This dataset contains information about previous application details of an applicant.

2. **Train dataset in application. csv**

- Shape: (307511, 122)
- First five rows and seven columns look like:

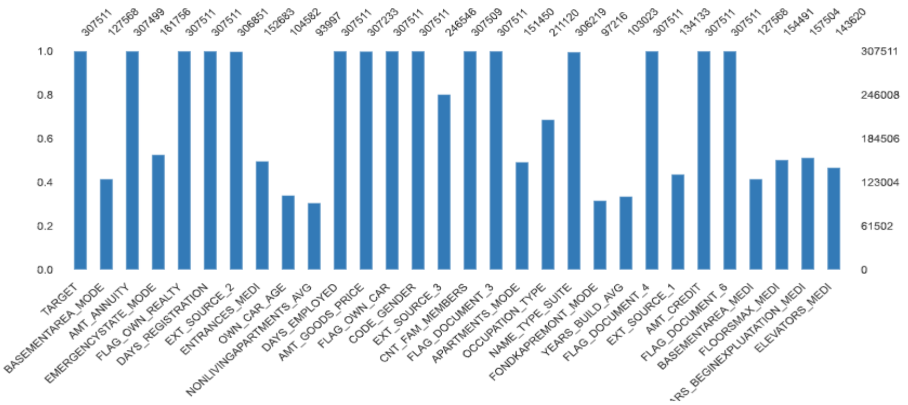
Table 1. Train dataset



SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
100002	1	Cash loans	M	N	Y	0
100003	0	Cash loans	F	N	N	0
100004	0	Revolving loans	M	Y	Y	0
100006	0	Cash loans	F	N	Y	0
100007	0	Cash loans	M	N	Y	0

- The null count and percentage of the features are as follows. We present ten features in table 2 as an example.

Table 2. Chart on nullity/ missing values by column.

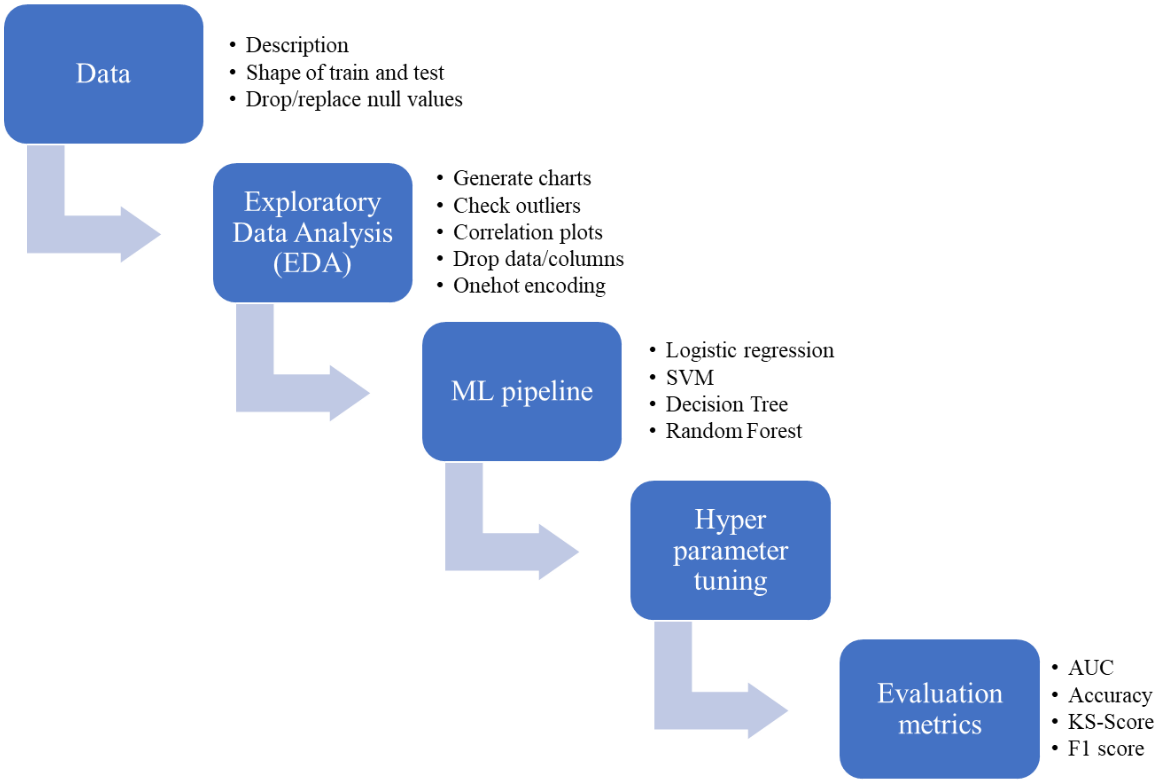


A simple visualization of nullity by column.

- We removed and replaced null values. We also removed infinitesimal large values.
- In particular, we dropped features with null values more than 50%.
- After dropping, the number of features is 121.

3. Diagram of workflow

Figure 1. Diagram of workflow



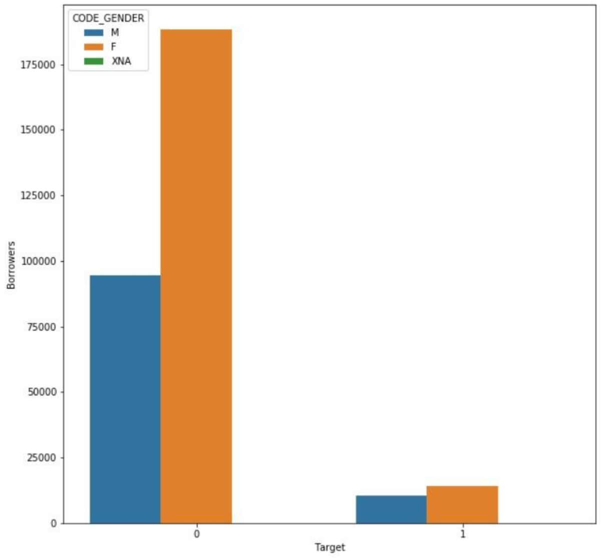
<u>Exploratory Data Analysis</u>

1. Descriptive statistics

- A data dictionary of the raw features
  - Pandas profiling in jupyter notebook
- We did descriptive analysis on the dataset such as data type of each feature, dataset size (rows and columns = 307511, 122), and summary statistics such as the number of observations, mean, standard deviation, maximum, minimum, and quartiles for all features and split of data is as follows Train: 70%, Test 20% , Validation 10%
- We generated charts on descriptive statistics of the target dataset.

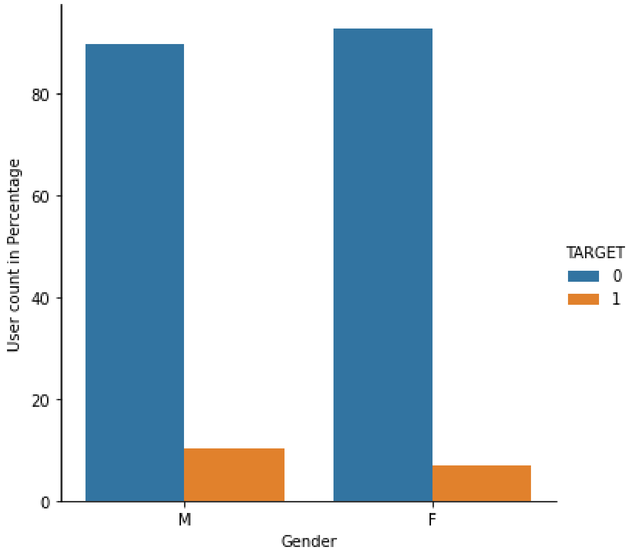
2. **Few Summary Statistics**

Figure 2. Target vs borrowers based on gender



- There are more women than men in both borrowers and targets. The difference between women and men was much larger in the borrower group.

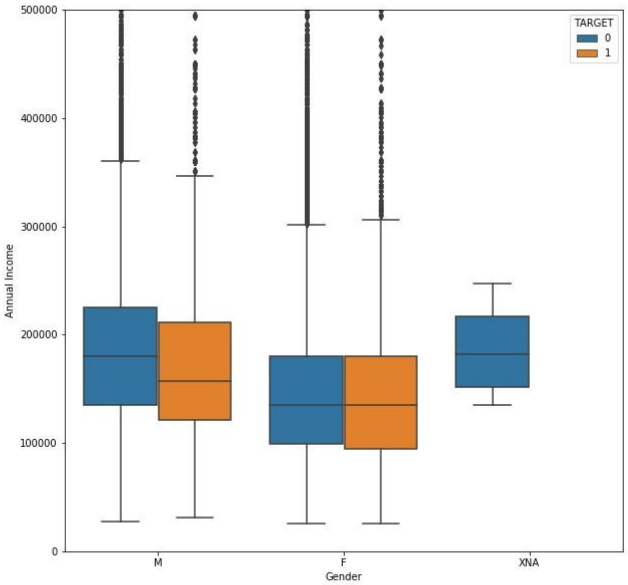
Figure 3. Target vs borrowers based on gender(in percentage)



-

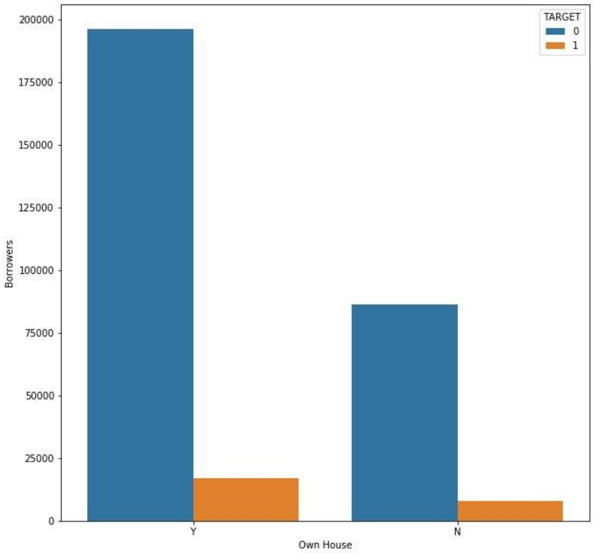
- Male most likely to default than Female based on percentage of defaulter_count(Second Graph)

Figure 4. Gender vs income based on target



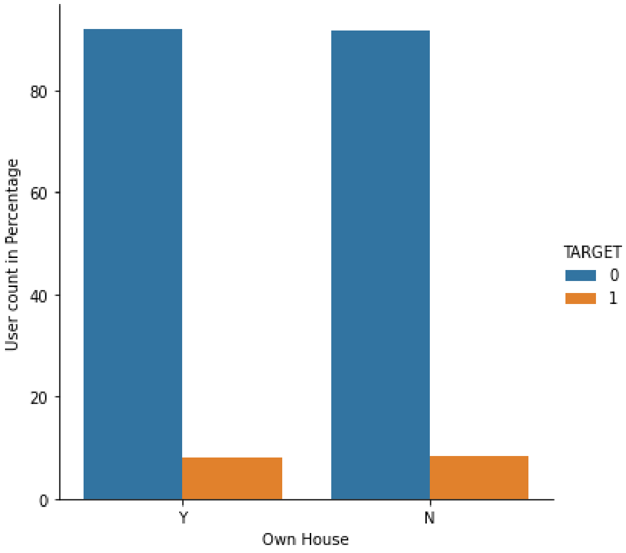
- Men had more income than women in both target and non-target. Men in non-target have higher income than men in target.

Figure 5. Own house count based on target



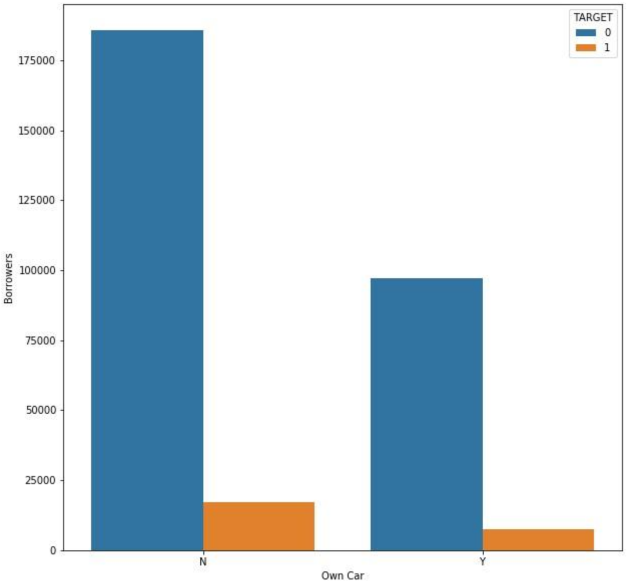
- There are more people who own houses than people who don’t own houses in non-target .
- There are more people who own houses than people who don’t own a house in the target.

Figure 6. Own house count based on target(in percentage)



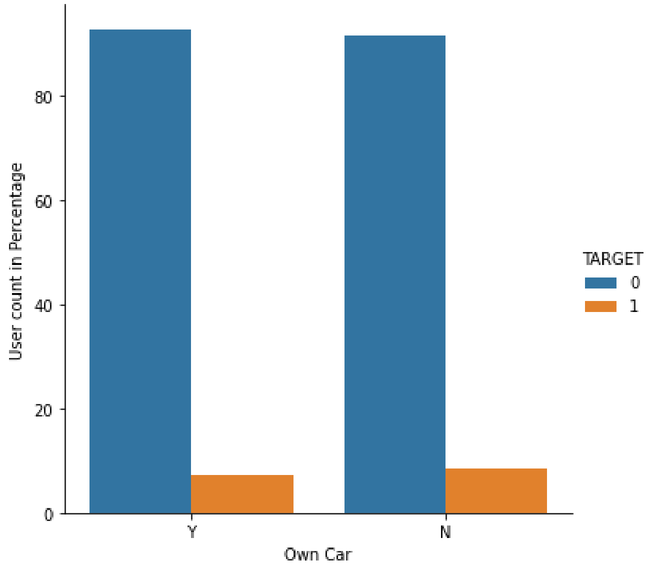
- Not a significant difference, but borrowers who own a house are more likely to pay

Figure 7. Own car count based on target



- More people don’t own cars than people who own cars in both target and non-target.

Figure 8. Own car count based on target(in percentage)



-

- Borrowers owning a car are more likely to pay on time

Figure 9. Occupation type vs income based on Target

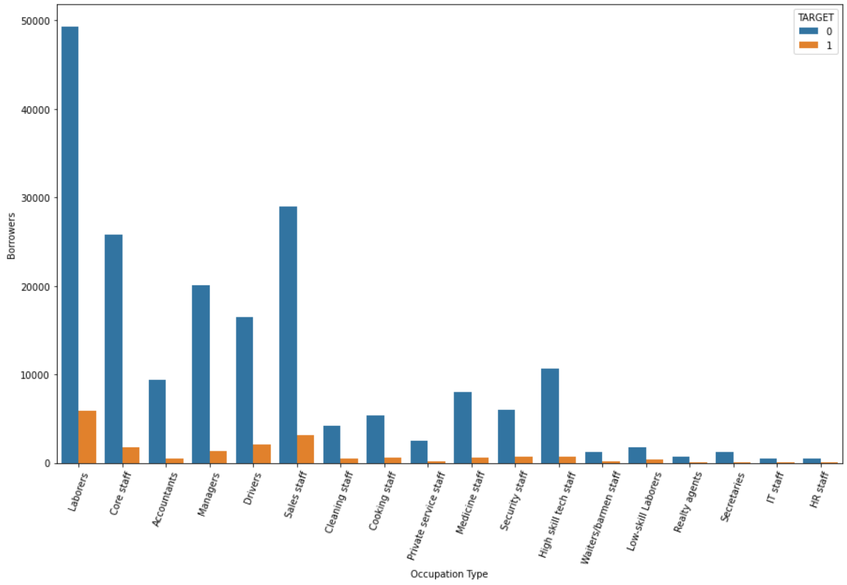
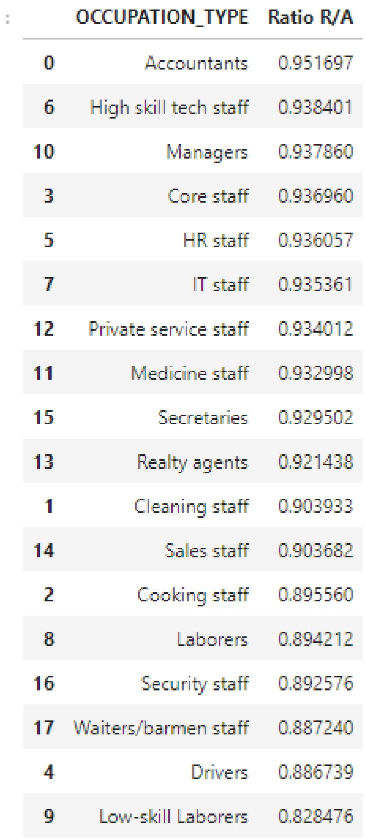


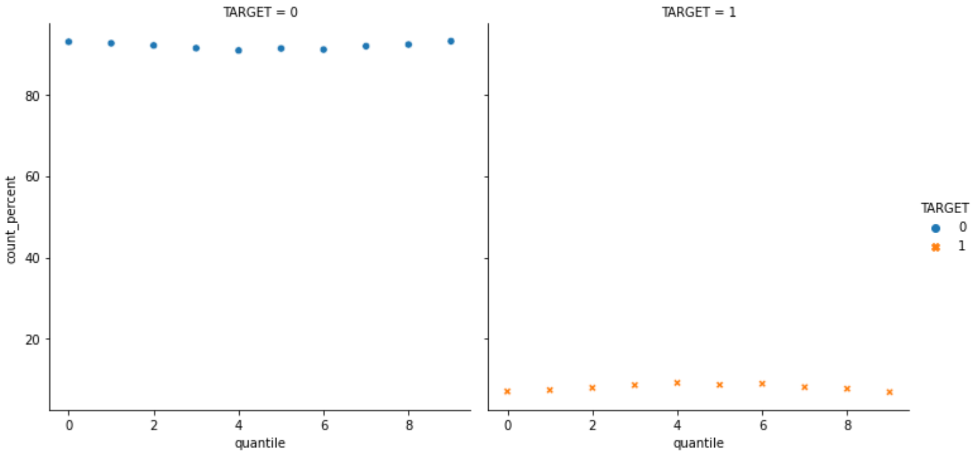
Figure 10. Repayers to Applicants Ratio



	OCCUPATION_TYPE	Ratio R/A
0	Accountants	0.951697
6	High skill tech staff	0.938401
10	Managers	0.937860
3	Core staff	0.936960
5	HR staff	0.936057
7	IT staff	0.935361
12	Private service staff	0.934012
11	Medicine staff	0.932998
15	Secretaries	0.929502
13	Realty agents	0.921438
1	Cleaning staff	0.903933
14	Sales staff	0.903682
2	Cooking staff	0.895560
8	Laborers	0.894212
16	Security staff	0.892576
17	Waiters/barmen staff	0.887240
4	Drivers	0.886739
9	Low-skill Laborers	0.828476

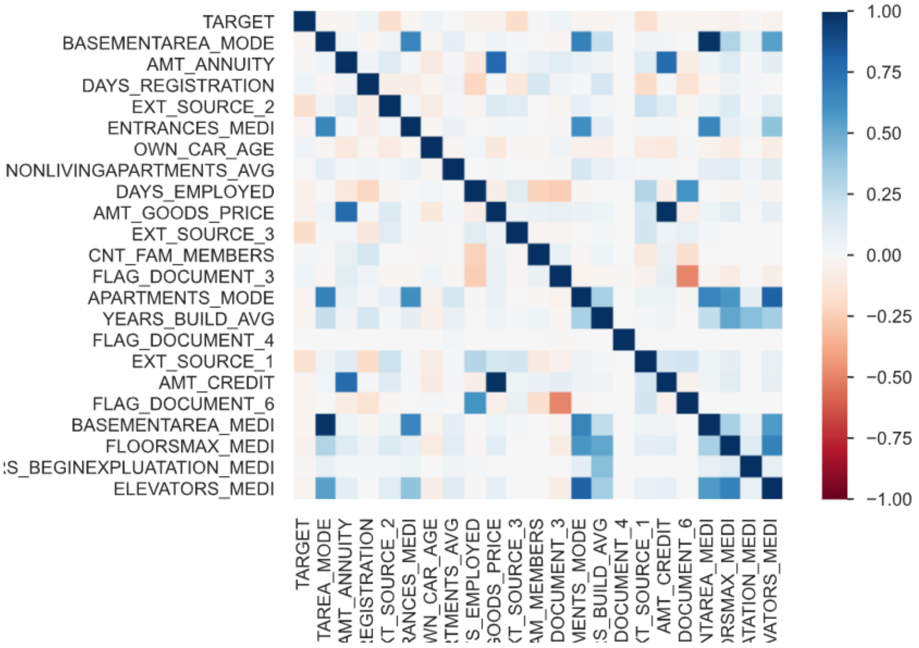
- Above figure gives the ratio of repayment based on occupation type.

Figure 11. Quantiles vs Income to Credit ratio

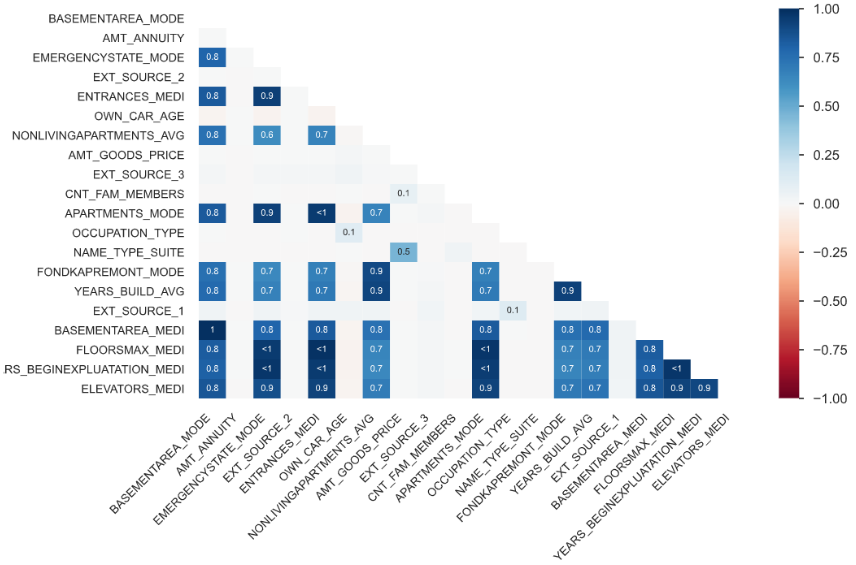


- Defaulters percentage is less when IC_ratio is either Low or High

Correlation Analysis and Visual EDA:

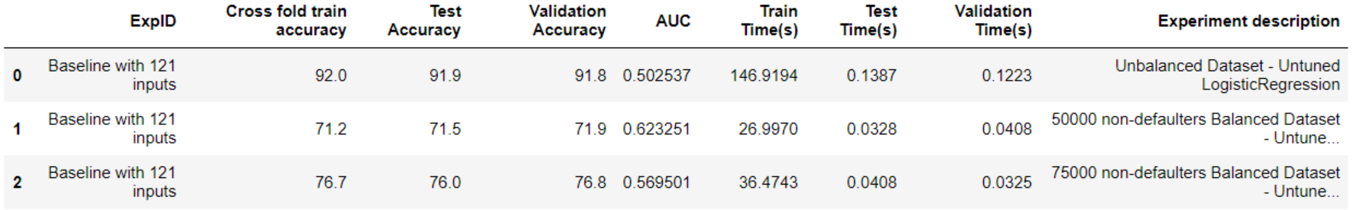


Heat Map - Null values



Detail visual EDA has been provided in the html/ipynb notebook(EDA)

Modeling Pipelines



	ExpID	Cross fold train accuracy	Test Accuracy	Validation Accuracy	AUC	Train Time(s)	Test Time(s)	Validation Time(s)	Experiment description
0	Baseline with 121 inputs	92.0	91.9	91.8	0.502537	146.9194	0.1387	0.1223	Unbalanced Dataset - Untuned LogisticRegression
1	Baseline with 121 inputs	71.2	71.5	71.9	0.623251	26.9970	0.0328	0.0408	50000 non-defaulters Balanced Dataset - Untune...
2	Baseline with 121 inputs	76.7	76.0	76.8	0.569501	36.4743	0.0408	0.0325	75000 non-defaulters Balanced Dataset - Untune...

1. Machine learning algorithm and metrics

Outcome of the risk analysis project is to predict whether the customer who has reached out to HomeCredit for a loan is a defaulter or not. Thereforth this is a classification task where the outcome of the event is binary 0 or 1. Although in industry it's defined as 0,1,2 - non defaulter, temporary defaulter and permanent defaulter respectively we are going with only 0 and 1 defaulter and non-defaulter. Our team built the following below pipelines for machine learning modeling:

- Logistic Regression

Futuristic(Phase 2 & 3)

- Decision Making Trees
- Random Forest
- SVMs
- Neural networks(RNN,LSTMs)

Deep learning neural networks could also be used for improving accuracy of the prediction model, but we would not be able to provide the features responsible for predicting a customer as a defaulter or not. This would further lead to compliance issues where we need to provide the exact features responsible for rejection of the loan for the probable non-defaulters.

Metrics to be used:

- Accuracy
- F1 Score
- AUC

KS-Score in particular could be more important in our case as we could have different models with same accuracy and F1 score but Kolomogorov Smirnov score helps us in deciding the best model across those, where the results with lower confidence level (probability prediction) could be moved forward to a different model for improving performance.

2. Machine Learning Pipeline Steps

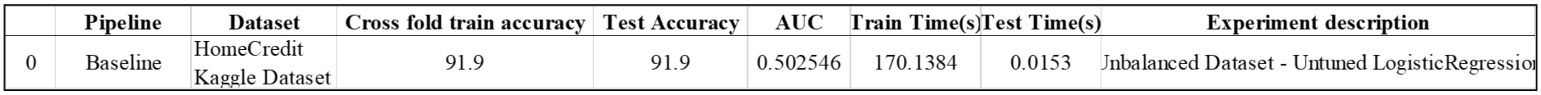
We used the typical machine learning pipeline introduced in the class.

- **Data Preprocessing**
    a. Gather raw data from Kaggle.
    b. Do exploratory data analysis.
    c. Prepare the proper input dataset, i.e. feature engineering.
- **Model Selection**
    d. Train and evaluate different candidate models including “Logistic Regression”, “Decision Making Trees”, “Random Forest”, and “SVMs”.
    e. Choose the best model based on the evaluation.
    f. Use different evaluation metrics such as “Accuracy”, “F1 Score”, “AUC”, and “KS-Score”.
- **Prediction Generation**
    g. Prepare the new data, and extract the features as before.
    h. Once the winning model is selected, use it to make predictions on the new data.

3. Baseline Logistic Regression Pipeline

- First, as we learned in class, we split train and test data. We split 20% test data with random seed set to 42 for correct results
- Next, we built a logistic regression baseline pipeline. We build a numerical pipeline based on numerical attributes and standard scaler. We impute the missing values using median. We do a logistic regression with this numeric pipeline.
- Lastly, we compute cross validation splits by using 30 splits and a test size of 0.3. We compute test accuracy and AUC using these cross validation.

Table 3. Logistic Regression Baseline Results



	Pipeline	Dataset	Cross fold train accuracy	Test Accuracy	AUC	Train Time(s)	Test Time(s)	Experiment description
0	Baseline	HomeCredit Kaggle Dataset	91.9	91.9	0.502546	170.1384	0.0153	Jnbalanced Dataset - Untuned LogisticRegressio

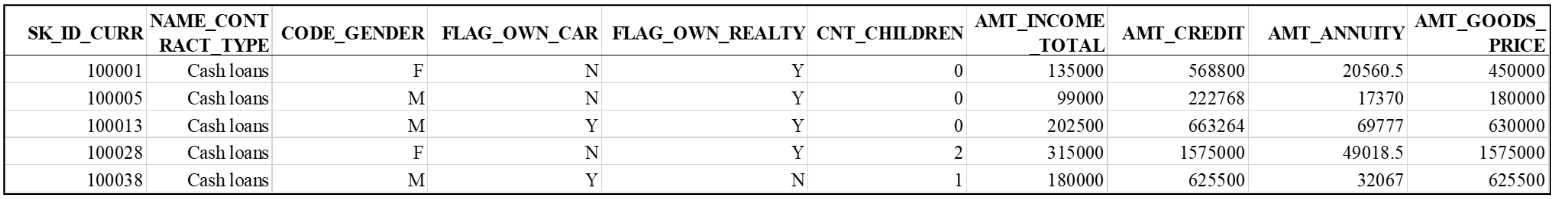
- Table 3 reports test accuracy and AUC of the logistic regression baseline results. The AUC is 0.502546. The test accuracy is 91.9.

4. Improving AUC

(a) _50000 non-defaulters Balanced Dataset

- To improve the AUC, we check across the test dataset and balance the dataset.
- Table 4 presents the first five rows and ten columns of the test dataset.

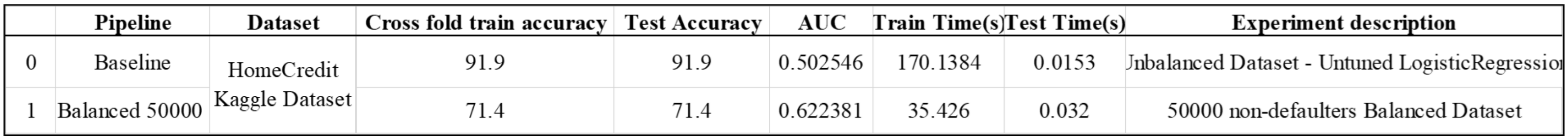
Table 4. Test dataset



SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
100001	Cash loans	F	N	Y	0	135000	568800	20560.5	450000
100005	Cash loans	M	N	Y	0	99000	222768	17370	180000
100013	Cash loans	M	Y	Y	0	202500	663264	69777	630000
100028	Cash loans	F	N	Y	2	315000	1575000	49018.5	1575000
100038	Cash loans	M	Y	N	1	180000	625500	32067	625500

- In the target dataset, the value count of 0 is 282686 and the value count of 1 is 24825.
- We balance the dataset of appending the values of 1 and 0 in the target dataset.
- e build 50000 non-defaulters Balanced Dataset.
- We split train and test data with 20% test data and random seed set to 42.
- Train X dataset is a shape of 59860 by 121 matrix, and train Y dataset is a shape of 59860 row matrix. Refer to cell #13 for the head of the balanced dataset.
- Table 5 presents the results for the baseline and balanced datasets. In particular, the newly added second row provides the results for the balanced dataset.

Table 5. 50000 non-defaulters Balanced Dataset



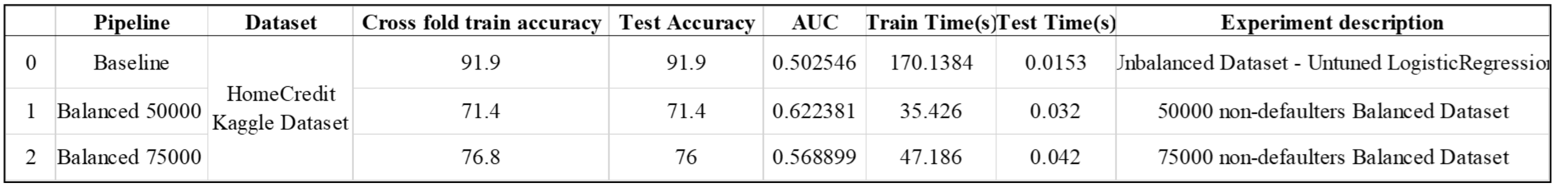
	Pipeline	Dataset	Cross fold train accuracy	Test Accuracy	AUC	Train Time(s)	Test Time(s)	Experiment description
0	Baseline	HomeCredit Kaggle Dataset	91.9	91.9	0.502546	170.1384	0.0153	Jnbalanced Dataset - Untuned LogisticRegressio
1	Balanced 50000		71.4	71.4	0.622381	35.426	0.032	50000 non-defaulters Balanced Dataset

- The AUC for 50000 non-defaulters balanced dataset is 0.62. The test accuracy is 71.4.

(b) 75000 non-defaulters Balanced Dataset

- In table 6, we use an alternative approach to balance the dataset. We append the values of 1 and 0 in the target dataset and this time build 75000 non-defaulters Balanced Dataset.
- In the 75000 non-defaulters balanced dataset, the train X dataset is a shape of 79860 by 121 matrix, and train Y dataset is a shape of 79860 row matrix. Refer to cell #20 for the head of the balanced dataset.

Table 6. 75000 non-defaulters Balanced Dataset



	Pipeline	Dataset	Cross fold train accuracy	Test Accuracy	AUC	Train Time(s)	Test Time(s)	Experiment description
0	Baseline	HomeCredit Kaggle Dataset	91.9	91.9	0.502546	170.1384	0.0153	Jnbalanced Dataset - Untuned LogisticRegressio
1	Balanced 50000		71.4	71.4	0.622381	35.426	0.032	50000 non-defaulters Balanced Dataset
2	Balanced 75000		76.8	76	0.568899	47.186	0.042	75000 non-defaulters Balanced Dataset

- The AUC for 75000 non-defaulters balanced dataset is 0.57. The test accuracy is 76.

Additional Results and Discussions

Tables 3,5, and 6 report the test accuracy and AUC of the logistic regression baseline results, rebalancing dataset of 50000 non-defaulters, and rebalancing dataset of 75000 non-defaulters, respectively. For logistic regression baseline pipeline, the AUC is 0.502546 and the test accuracy is 91.9. After rebalancing, for the 50000 non-defaulters balanced dataset, the AUC is 0.62 and the test accuracy is 71.4. So, by re-balancing, we can earn higher AUC but lose test accuracy. By increasing the sample size of non-defaulters, we gain explanatory power of the test set.

The baseline model prior to rebalancing shows the highest test accuracy but the lowest AUC. Thus, our attempt to improve the AUC was to rebalance the dataset in different sample sizes. Because AUC represents the quality of the model's predictions, our goal was to build a model that has similar test accuracy with the baseline model but the higher AUC. Since this is a baseline model without feature engineering and hyper-parameter tuning, we have room to improve with other candidate models.

Our baseline model shows a great test accuracy but relatively low AUC. Re-balancing is helpful in improving AUC at the cost of accuracy. We believe there would be a way to improve AUC along with improving test accuracy with other candidate models such as hyper-parameter tuning and additional feature engineering.

Conclusion

The object of the HCDR project is to predict the repayment ability of the financially under-served population. This project is important because well-established predictions are necessary to both the loaner and borrower. In real-time Homecredit is able to display loan offers to its customers with the maximum amount and APR using their ML pipelines where fetching data from the data providers via APIs, performing EDA and fitting it to the model to generate scores occurs in microseconds of time. Hence Risk analysis becomes very critical in this regard where NPA(Non-Performing Asset) expected is less than 5% in order to run a profitable business.

Credit history is a measure explaining the credibility of a user generated using parameters like average/min/max balance maintained by the user, Bureau scores reported, salary etc and repayment patterns could be analysed using the timely defaults/repayments made by the user in the past. Alternate data includes other parameters like geographic data, social media data, calling/SMS data etc. As part of this project we would be using the datasets provided by kaggle to perform exploratory data analysis, build machine learning pipelines and evaluate the models across several evaluation metrics for a model to be deployed.

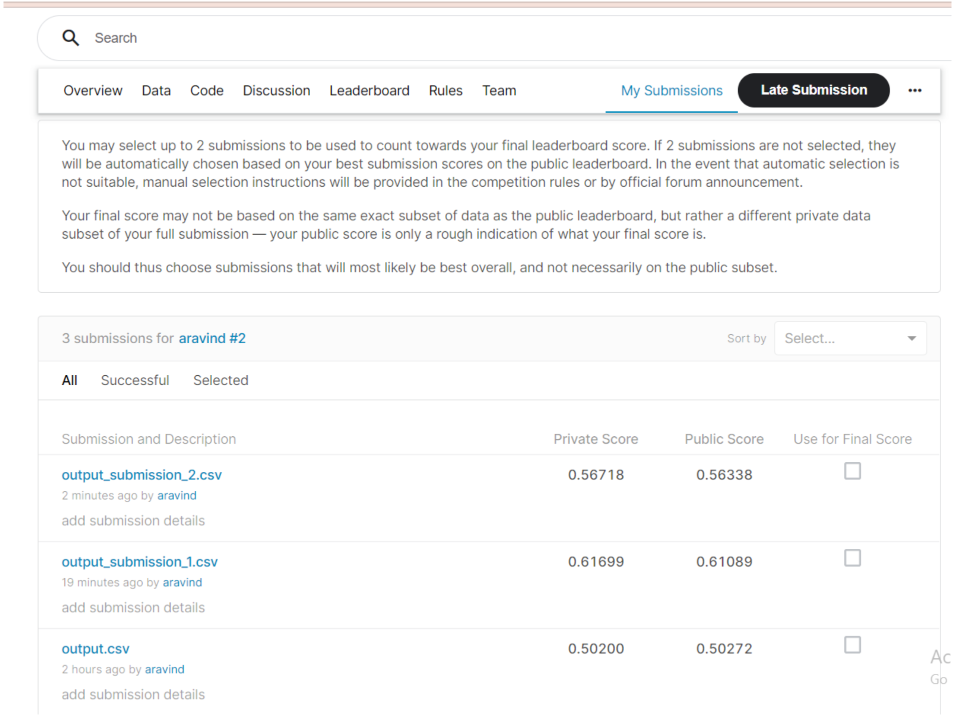
Our EDA analysis shows characteristics of the target dataset. We measure the gender difference, income difference, and whether they own a house or a car, and the occupations of the data in target vs non-target. Our results indicate that there are more women than men in borrowers and targets. Men have higher income than women. More people own houses than people who do not own houses. More people own cars than people who do not own cars.

In phase 1, we provide baseline logistic regression pipelines. We experimented with a baseline pipeline, 50000 non-defaulters balanced dataset, and 75000 non-defaulters balanced dataset. A baseline pipeline has high test accuracy, 91.9, but low AUC, 0.5. After rebalancing the dataset, 50000 non-defaulters dataset shows 71.4% of test accuracy and 0.62 of AUC. After increasing the sample size, the 75000 non-defaulters dataset shows 76% of test accuracy and 0.57 of AUC. Thus, our results indicate that by re-balancing, we earn higher AUC but lose test accuracy. These regressions provide a useful implication that there is a tradeoff between the AUC and test accuracy, thus researchers should carefully consider which one to improve.

Since these are baseline models without feature engineering and hyper-parameter tuning, we believe that we should consider other candidate models in the future. That is, in phase 2, we would start considering additional feature engineering, hyper-parameter tuning, feature selection and importance, and ensemble methods. These additional methods will hopefully improve both AUC and test accuracy of the models.

Kaggle submissions

- Our output dataset on the target has 45673 values of 0, and 3071 values of 1.
- We completed Kaggle submissions as follows:



You may select up to 2 submissions to be used to count towards your final leaderboard score. If 2 submissions are not selected, they will be automatically chosen based on your best submission scores on the public leaderboard. In the event that automatic selection is not suitable, manual selection instructions will be provided in the competition rules or by official forum announcement.

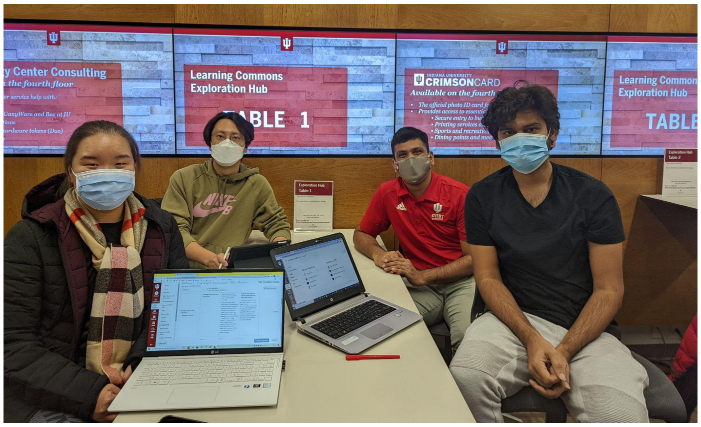
Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

3 submissions for aravind #2

Submission and Description	Private Score	Public Score	Use for Final Score
output_submission_2.csv 2 minutes ago by aravind add submission details	0.56718	0.56338	□
output_submission_1.csv 19 minutes ago by aravind add submission details	0.61699	0.61089	□
output.csv 2 hours ago by aravind add submission details	0.50200	0.50272	□

Team profile



Herman Wells library, Nov 11th Thursday, 2021.

Individual Profile

Aravind Reddy Sheru
Email: asheru@iu.edu

Sai Charan Chintala
Email: sachin@iu.edu

Seongbo Sim
Email: simseo@iu.edu

Yun Joo An
Email: yunjooan@iu.edu