



APPLIED MACHINE LEARNING FINAL PROJECT: HOME CREDIT DEFAULT RISK

Group 36:

Yashvanth Kumar Guntupalli
Sri Lekha Malraju
Revanth Sai Chowdary Rayala
Vinay Chandra Makkineni

CONTENTS:

- Team Profile
- Four P 's
- Final Project : Home Credit Default Risk.
 - Project Description
 - Exploratory Data Analysis (EDA).
 - Overview of Modelling Pipelines explored.
 - Results and Discussion.
- Conclusion and Next steps.

TEAM PROFILE:



Yashvanth Kumar
Guntupalli
(yakguntu@iu.edu)



Srilekha Malraju
(msrilekh@iu.edu)



Revanth Sai
Chowdary Rayala
(rerrayala@iu.edu)



Vinay Chandra Makineni
(vimakin@iu.edu)

FOUR P'S

Past:

- We are making the HCDR Project, which predict whether borrowers are defaulters or not based on various financial and non financial data.

Present:

- In Phase 1, We gathered the information, and did Exploratory Data Analysis. Additionally, we fabricated a model utilizing Logistic Regression and attempted to adjust the information by balancing the quantity of samples of non-defaulters.
- The pattern model gave a very high accuracy, however moderately low AUC. Adjusting information further developed AUC at the expense of accuracy.
- We have discovered that we lost information of samples. Which were prohibited in rebalancing, So we needed to find better method for further developing AUC while keeping the examples.

FOUR P'S CONT'D:

Planned:

- In Phase 2, We will try to introduce other candidate models including “ Decision tree's”, “Logistic Regression”, “Decision trees”, “Support Vector Machines (SVM)”, ”K- Nearest neighbors”, “Neural Networks” and “Random Forest”.
- Including to this, We are also planning to make our input data-set more balanced by Feature engineering and Feature importance analysis.
- For Candidate models, We will change the hyperparameters to further develop AUC and different measurements with higher accuracy.
- Likewise, We are wanting to change the subtleties of the models, by Additional feature engineering.
- At last, we will outfit our models to come by a superior outcomes.

Problems:

- The problem we are facing are, We might require some earlier information about the information, for example credit data.
- The information might assist us with assessing the course of feature engineering and feature importance analysis.

PROJECT DESCRIPTION:

Many people struggle to get loans due to insufficient or non-existent credit histories. Unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.


In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

There are multiple datasets that will be taken from Kaggle that are mentioned in the below section. We perform exploratory data analysis and feature engineering on the given datasets and train the data using multiple classification algorithms

PROJECT DESCRIPTION CONT'D:

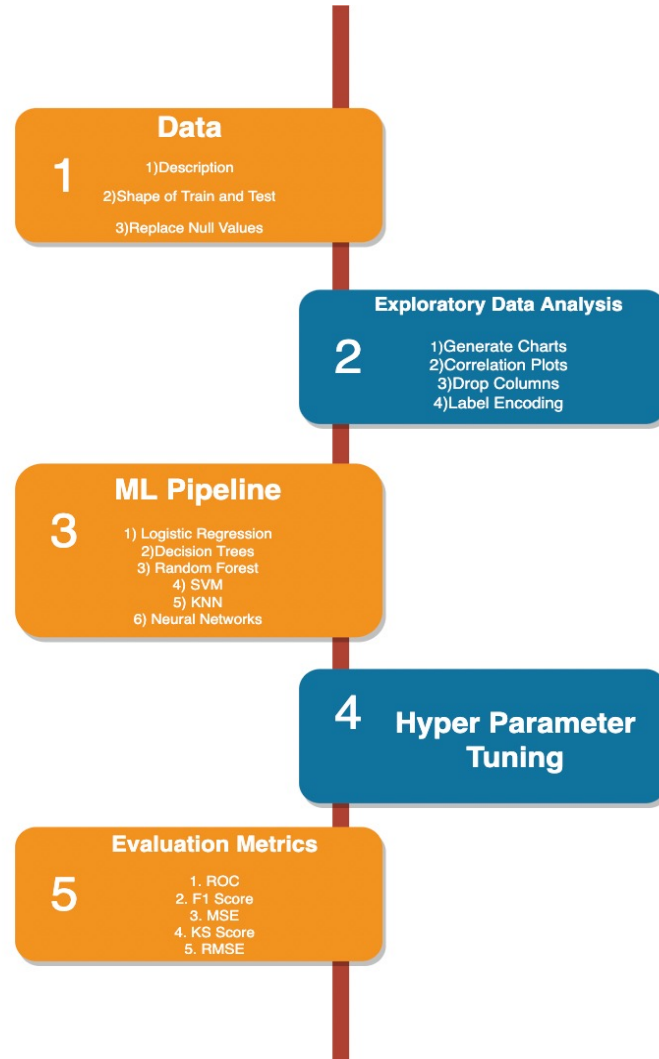
To find the best model, we train and evaluate several candidate models.

- We will introduce candidate models like “Decision tree’s”, “Logistic Regression”, “Decision trees”, “Support Vector Machines (SVM)”, “K-Nearest neighbors”, “Neural Networks” and “Random Forest” to decide the best model.



We decide which model is the best among all. The output of the model will be either 0 or 1 where 0 corresponds to saying that the customer will repay the loan and 1 meaning that there's some risk to the customer repaying to the lender.

PROJECT WORKFLOW :



EXPLORATORY DATA ANALYSIS:

We are doing Exploratory data analysis to check the below attributes of the data.

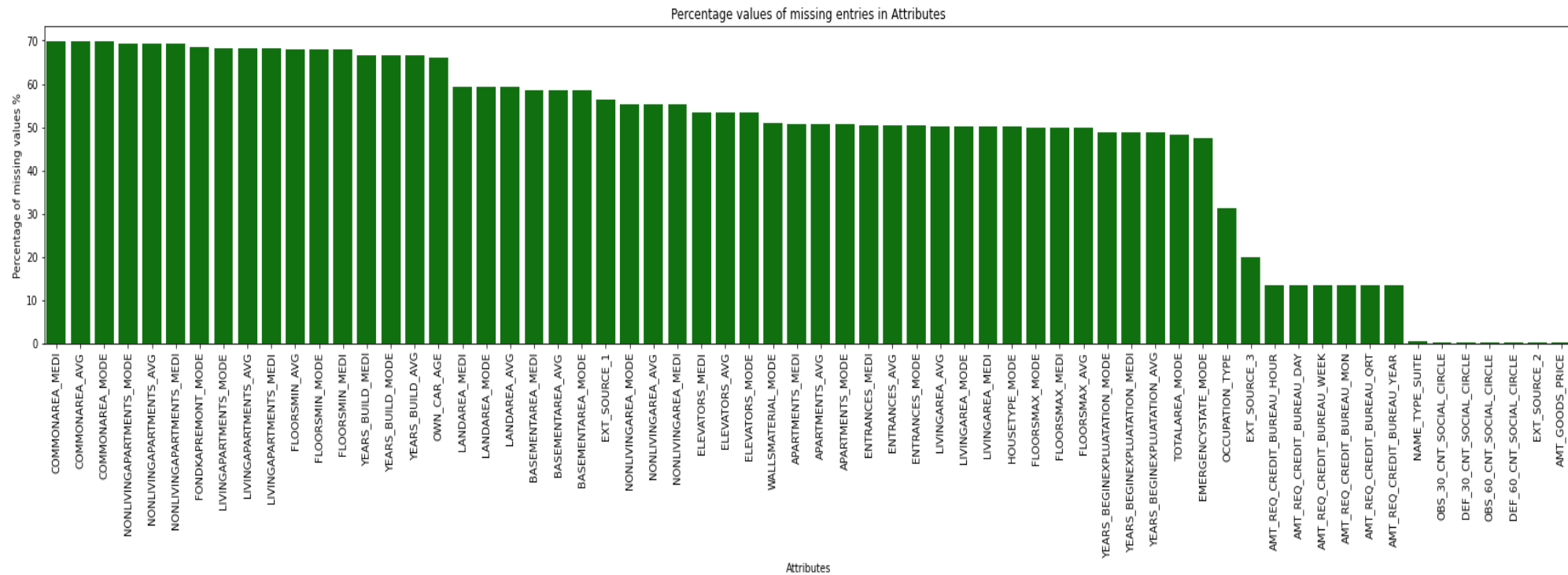
- “Test description”, “Data size”, “Summary statistics”, “Correlation analysis”, “Checking missing values”,etc..

Among all the Exploratory data analysis we did, Some interesting EDA's are:

- Target VS Borrowers based on their children

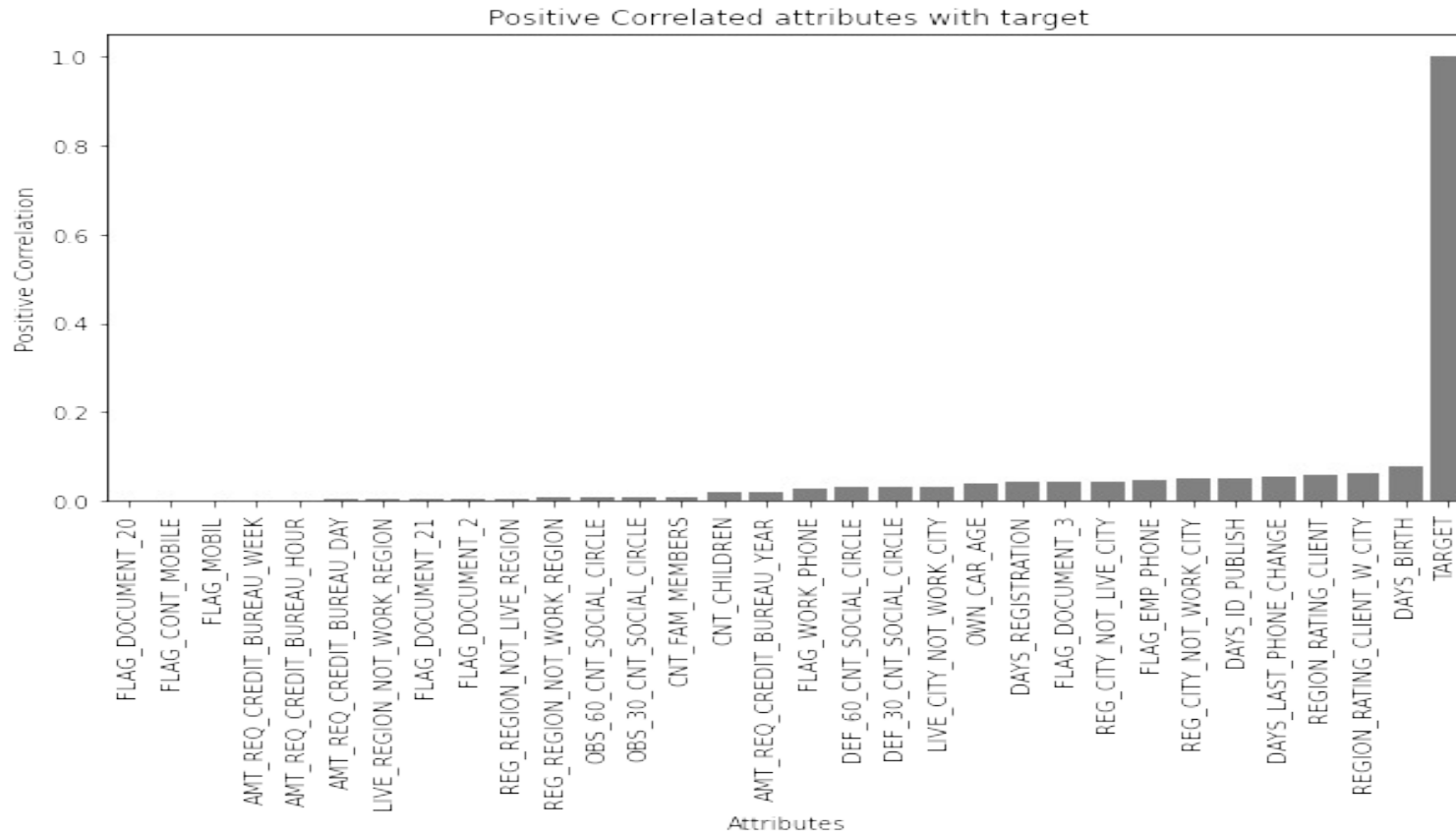
EXPLORATORY DATA ANALYSIS CONT'D:

- Plot based on Missing Values :
- The above figure 1 gives us the Null Count of each feature of the training data set.
- So for the Null Values, We replace them with the Mean.



EXPLORATORY DATA ANALYSIS CONT'D:

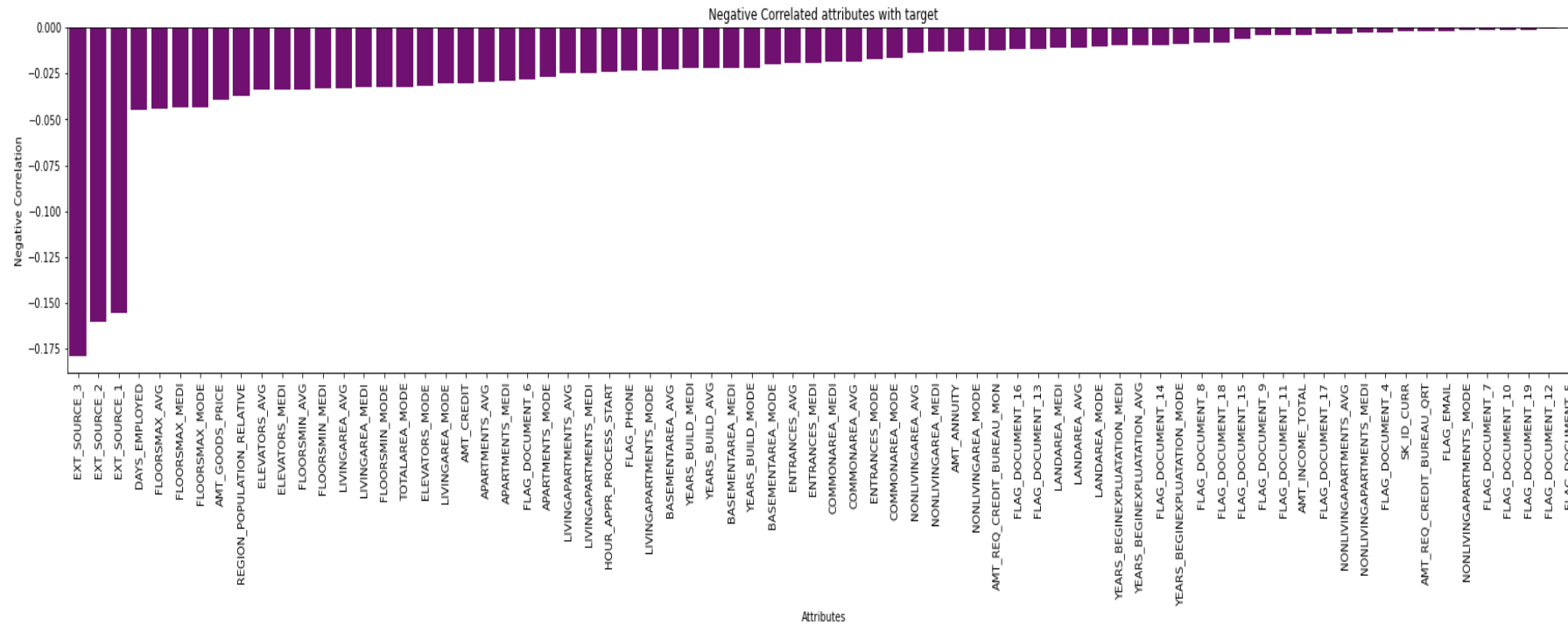
➤ Positively Correlated features based on Target.



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

EXPLORATORY DATA ANALYSIS CONT'D:

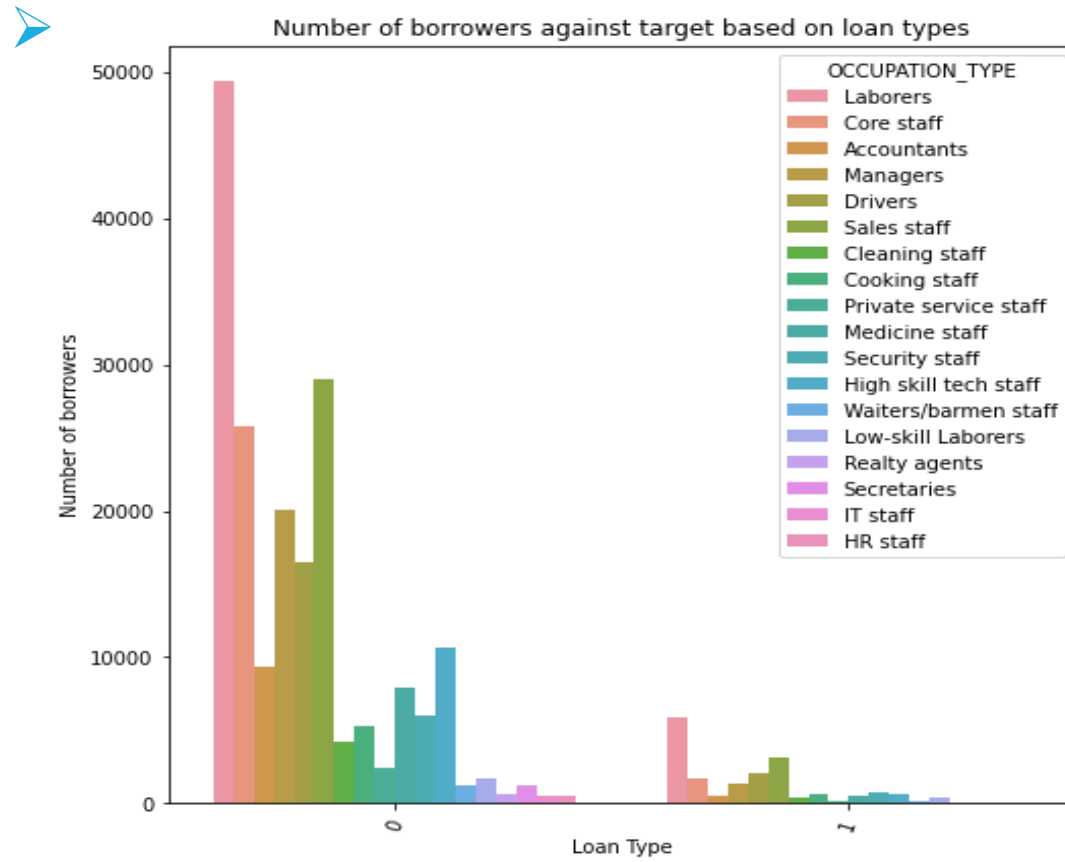
➤ Negatively Correlated features based on Target.



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

EXPLORATORY DATA ANALYSIS CONT'D:

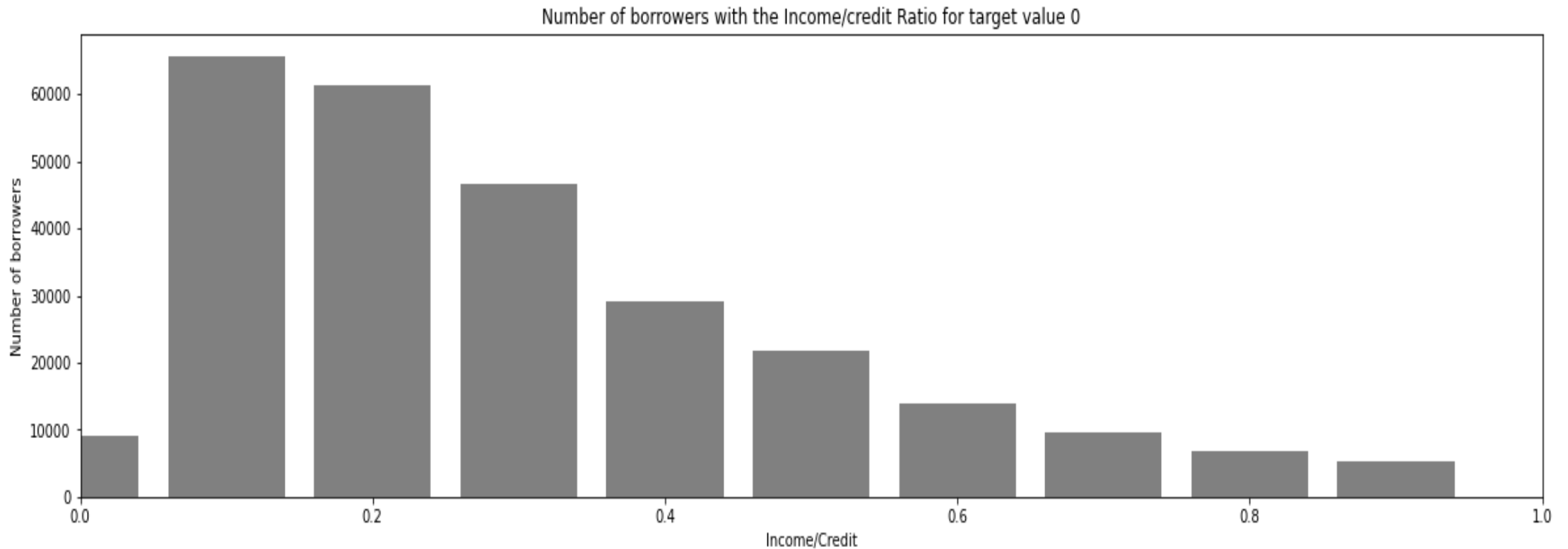
➤ Target VS Borrowers based on Gender.



➤ The above graph depicts the multiple kinds of occupation types and what type of loan (Cash loan/revolving loan) they have taken.

EXPLORATORY DATA ANALYSIS CONT'D

➤ Target VS Borrowers based on Income/Credit ratio.



MODELLING PIPELINES:

Baseline Linear 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 mean. We do a logistic regression with this numeric pipeline.

We compute test accuracy and AUC using this model.

The detailed setup is documented in the report and omitted here.

RESULTS:

➤ Results of Logistic regression :

- In .pynb Section 4 reports the test accuracy and AUC Score of logistic regression baseline. The AUC Score is 0.5039 and a testing accuracy of 91.9%

```
train size X : (246008, 121)
train size y : (246008,)
test size X : (61503, 121)
test size y : (61503,)
0.91910963692828 : is the accuracy score
The training score is: 0.9190310884198888
The testing score is: 0.91910963692828
The Area Under the Curve: 0.5039186543640561
The Confusion Matrix is: [[56483    71]
 [ 4904    45]]
```


RESULTS OF LOGISTIC REGRESSION ON BASELINE MODEL -1 :



Since the dataset is unbalanced, we balance it by building a Non-Defaulters Dataset. We do this to improve the Area Under the Curve score.



We split the dataset into testing and training dataset where testing = 20% of the full dataset.



We take 25000 Non-Defaulters dataset initially and train the model with this dataset.



The results are as in Section 4.1, The Area Under the Score increased to 0.68 but the accuracy of the model fell considerably down to 68%.

RESULTS OF LOGISTIC REGRESSION ON BASELINE MODEL -1 CONT'D :

➤ The following are the results of the model :

```
(49825, 122)
(39860, 121) (39860,)
train size X : (39860, 121)
train size y : (39860,)
test size X : (9965, 121)
test size y : (9965,)
0.6873055694932263 : is the accuracy score
The training score is: 0.683843452082288
The testing score is: 0.6873055694932263
The Area Under the Curve: 0.6872928184362408
The Confusion Matrix is: [[3454 1550]
(49825, 122)
(39860, 121) (39860,)
train size X : (39860, 121)
train size y : (39860,)
test size X : (9965, 121)
test size y : (9965,)
0.6873055694932263 : is the accuracy score
```

RESULTS OF LOGISTIC REGRESSION ON BASELINE MODEL -2:

➤ In the second model, we take a dataset with 80000 Non-Defaulters and train the model with this dataset. The results are as in Section 4.2, The Area Under the Score came out to be 0.57 and the accuracy of the model is still poor with an accuracy of 77.5%.

```
(104825, 122)
(83860, 121) (83860,)
train size X : (83860, 121)
train size y : (83860,)
test size X : (20965, 121)
test size y : (20965,)
0.7754352492248986 : is the accuracy score
The training score is: 0.7814094920104937
The testing score is: 0.7754352492248986
The Area Under the Curve: 0.5784691600353616
The Confusion Matrix is: [[15245  739]
 [ 3969 1012]]
```

ADDITIONAL RESULTS AND DISCUSSIONS:

- In Section 8.2.1, 8.2.2, and 8.2.3 of the pynb it reports the accuracy and Area Under the Curve score of the model. In 8.2.1, we split the numerical and categorical attributes into different pipelines and perform Feature Union on those pipelines and train it using Linear Regression by splitting the training and testing data in 80% and 20% ratio respectively.
- After training we test the model using Area Under the Curve metric and testing the model. The AUC score came out to be 0.5039 and the model score is 91.9%. We got a decent accuracy score. But the AUC score is pretty, this is due to the reason that the dataset is imbalanced.
- Since the dataset is imbalanced, we try to balance the dataset by down sampling the dataset and taking only 25000 Non-Defaulters initially. We then train this model using Logistic Regression again and apply the same metrics.
- The AUC score came out to be 0.68 and the model score is 0.68. We got a good AUC score but at the expense of a bad accuracy which is not acceptable.
- Now, we try to take a higher sample with 80000 Non-Defaulters dataset. We follow the same process on this dataset as well. The AUC score came out to be 0.57 and the model score is 0.77. The AUC score decreased and it improves the accuracy as well. But the scores aren't satisfying.

CONCLUSIONS:

- The object of the HCDR project is to foresee the repayment capacity of the financially under-served populace. This project is important because well-established predictions are necessary to both the loaner and borrower.
- Home Credit will need accuracy of the highest level to make sure the data is predicted correctly as it involves people's lives.
- Our aim for the next phases is to increase the the AUC score and accuracy of the model to more than 95% and all of this should be done within milliseconds of time.

THANK YOU.

