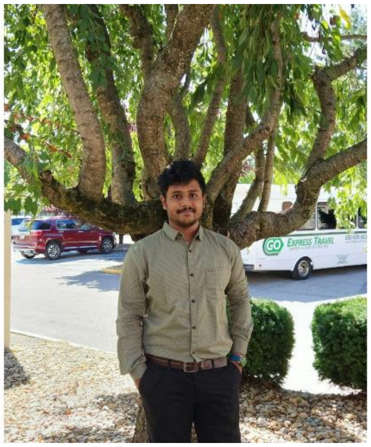
Home Credit Default Risk (HCDR)

Group 11
Anuj Mahajan
Siddhant Patil
Shashwati Diware
Shubham Jambhale

Team Members:

Shubham Jambhale
sjambhal@iu.edu



Siddhant Patil
sidpatil@iu.edu

Anuj Mahajan
anujmaha@iu.edu

Shashwati Diware
sdiware@iu.edu

Phase Leader Plan

Phase	Contributor	Contribution Description
Phase 1: Project Planning	Anuj Mahajan	Download Data, go through data, and load libraries. Create a pipeline diagram and describe the pipeline design. Describe Preprocessing,
Phase 1: Project Planning	Shashwati Diware	Project Abstract, ML Algorithm Names, and describe Metrics.
Phase 1: Project Planning	Shubham Jambhale(Phase Leader)	Understanding the problem statement, and writing table descriptions. Schedule meetings, coordinate tasks, plan phase
Phase 1: Project Planning	Siddhant Patil	Machine Learning Pipeline Steps and describes pipeline components.
Phase 2: Base Line Modelling and EDA	Anuj Mahajan (Phase Leader)	Creating Block Diagram EDA and one slide of the presentation. Schedule meetings, coordinate tasks, plan phase
Phase 2: Base Line Modelling and EDA	Shashwati Diware	Result Analysis EDA and one slide of the presentation.
Phase 2: Base Line Modelling and EDA	Shubham jambhale	Result Analysis and two slides of the presentation
Phase 2: Base Line Modelling and EDA	Siddhant Patil	Result Analysis and two slides of the presentation
Phase 3: Hyperparameter Tuning	Shashwati Diware (Phase Leader)	Testing Accuracy matrix and Schedule meetings, coordinating tasks, the planning phase
Phase 3: Hyperparameter Tuning	Siddhant Patil	Create and develop code for Hyperparameter tuning
Phase 3: Hyperparameter Tuning	Shubham Jambhale	Run and create analysis by testing the confusion / AUC matrix. Coordinate Tasks and one slide of the presentation
Phase 3: Hyperparameter Tuning	Anuj Mahajan	Run and analyze Lasso and ridge regression losses. Coordinate tasks and one slide of the presentation
Phase 4: Final Report Generation	Siddhant Patil (Phase Leader)	Plan Phase Schedule Meetings and Coordinate Tasks, analyze and go through the final results
Phase 4: Final Report Generation	Anuj Mahajan	Rearrange everything and go through the final documentation, list down the final recordings
Phase 4: Final Report Generation	Shashwati Diware	Prepare the final presentation
Phase 4: Final Report Generation	Shubham Jambhale	Check everything and submit the assignment before the deadline

Credit Assignment Plan

Phase 1:

Task	Task Description	Hours spent	Assigned to	Start	End
Understanding problem statement	Go through the problem statement to understand the requirements	6	Shubham	11/05/22	11/07/22
Data Exploration	Explore and analyze the data for a better understanding	6	Anuj	11/07/22	11/09/22
Project Proposal	Creating the project proposal and preparing a basic report with Abstract, ML models, and Gantt diagram	20	Group	11/09/22	11/14/22

Phase 2:

Task	Task Description	Hours Spent	Assigned to	Start	End
Creating Block Diagram	Creating the block diagram of the basic flow of execution.	5	Anuj	11/13/22	11/15/22
Creating Pipeline Diagram	Creating the pipeline diagram of the machine learning model from analyzing the data till the result analysis	5	Shashwati	11/13/22	11/15/22
Result Analysis	Analyzing the Result	10	Group	11/26/22	11/29/22
PowerPoint Presentation	Simultaneously prepare the PowerPoint presentation and add the analyzed data into it as per need	10	Group	11/20/22	11/29/22

Phase 3:

Task	Task Description	Hours spent	Assigned to	Start	End
Create and develop code for hyperparameter tuning	Design and develop python helper function for hyperparameter tuning	16	Siddhant	11/20/22	11/25/22
Result Analysis	Analysis of Obtained Result	2	Group	12/02/22	12/03/22
Testing Accuracy matrix	Analyzing accuracy using accuracy matrix	2	Shashwati	12/03/22	12/04/22
Testing f1 matrix	Analyzing accuracy using Confusion/AUC matrix score	2	Shubham	12/03/22	12/04/22
Lasso And Ridge Loss Functions	Analyzing the lasso and ridge loss function	2	Anuj	12/03/22	12/04/22

Phase 4:

Task	Task Description	Hours Spent	Assigned To	Start	End
Final Documentation	Rearrange everything and go through the final documentation, list down the final recordings	10	Anuj	12/03/22	12/08/22
Final Results	Analyze final results obtained after the final testing	6	Siddhant	12/05/22	12/08/22
Final Presentation	Prepare the final presentation	4	Shashwati	12/06/22	12/08/22
Assignment Submission	Check everything and submit the assignment before the deadline	1	Shubham	12/08/22	12/09/22

Abstract

Based on historical credit histories and repayment trends utilizing machine learning modeling, Home Credit offers unsecured lending. A user-generated credit score is calculated using criteria like the balance that the user has maintained. As part of this project, we are predicting the customer repayment status such as if the user is a defaulter or not using machine learning pipelines and models using the datasets provided by Kaggle. The data collection includes seven separate tables that aid in determining the user status, including bureau balance, credit card balance, home credit column detection, Installments payments, POS CASH balance, and previous applications. In phase 2, we provide feature engineering, EDA, and modeling pipelines. We experimented with categorizing baseline inputs and choosing features for Decision Trees, Random Forests, and Logistic Regression. The Random Forest baseline pipeline has the highest test accuracy, followed by Logistic Regression, then Decision Making tree.

Data and Task Description

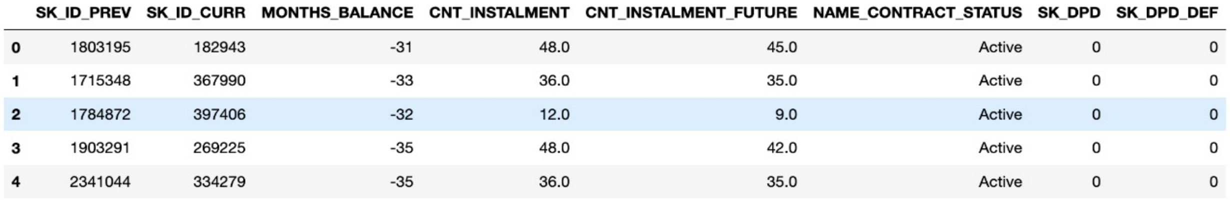
Data source

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 credit information such as contract status, the number of installments left to pay, DPD(days past due), etc. of the current application.

Table 1. POS_CASH_balance.csv

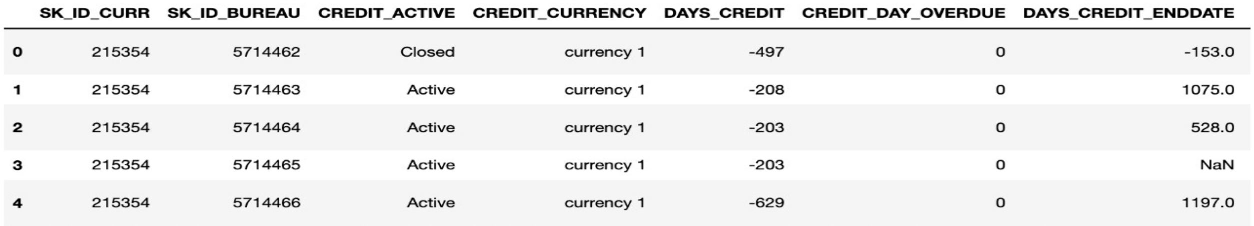


	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	CNT_INSTALMENT	CNT_INSTALMENT_FUTURE	NAME_CONTRACT_STATUS	SK_DPD	SK_DPD_DEF
0	1803195	182943	-31	48.0	45.0	Active	0	0
1	1715348	367990	-33	36.0	35.0	Active	0	0
2	1784872	397406	-32	12.0	9.0	Active	0	0
3	1903291	269225	-35	48.0	42.0	Active	0	0
4	2341044	334279	-35	36.0	35.0	Active	0	0

bureau.csv

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

Table 2. Bureau.csv

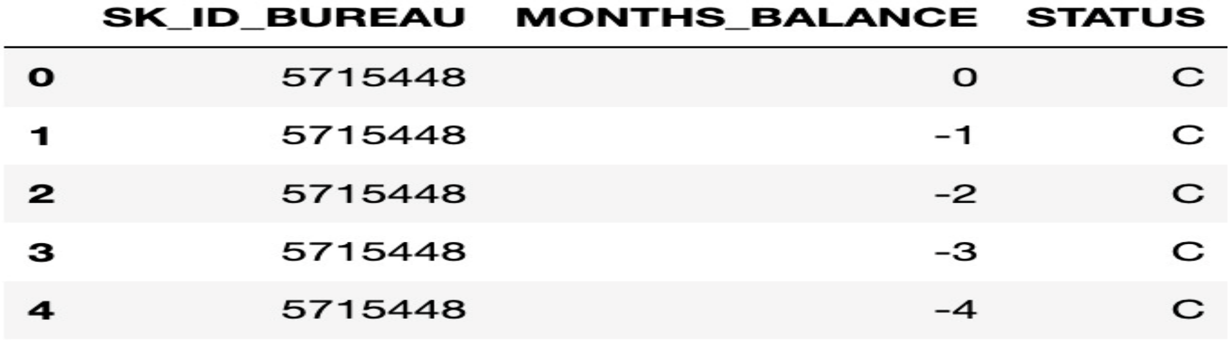


	SK_ID_CURR	SK_ID_BUREAU	CREDIT_ACTIVE	CREDIT_CURRENCY	DAYS_CREDIT	CREDIT_DAY_OVERDUE	DAYS_CREDIT_ENDDATE
0	215354	5714462	Closed	currency 1	-497	0	-153.0
1	215354	5714463	Active	currency 1	-208	0	1075.0
2	215354	5714464	Active	currency 1	-203	0	528.0
3	215354	5714465	Active	currency 1	-203	0	NaN
4	215354	5714466	Active	currency 1	-629	0	1197.0

bureau_balance.csv

This dataset gives information about the Status of the Credit Bureau loan during the month, the Month of balance relative to the application date, Recoded ID of the Credit Bureau credit. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.

Table 3. bureau_balance.csv

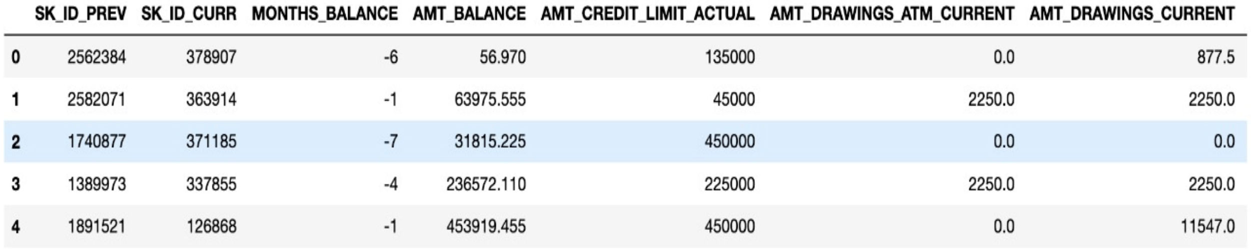


	SK_ID_BUREAU	MONTHS_BALANCE	STATUS
0	5715448	0	C
1	5715448	-1	C
2	5715448	-2	C
3	5715448	-3	C
4	5715448	-4	C

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. Each row is one month of a credit card balance, and a single credit card can have many rows.

Table 4. credit_card_balance.csv

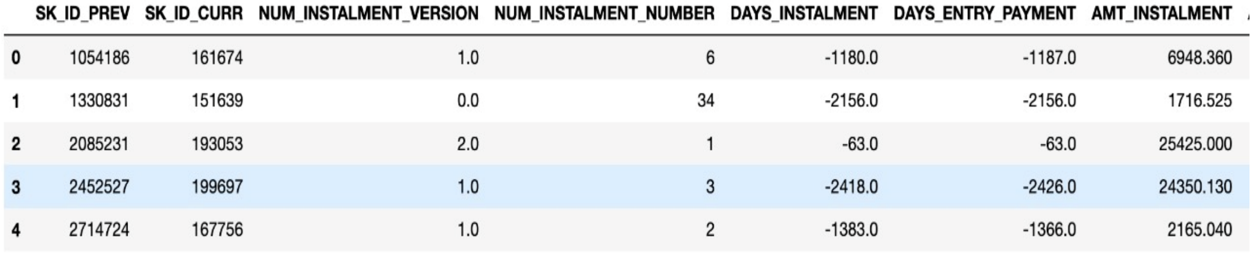


	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE	AMT_CREDIT_LIMIT_ACTUAL	AMT_DRAWINGS_ATM_CURRENT	AMT_DRAWINGS_CURRENT
0	2562384	378907	-6	56.970	135000	0.0	877.5
1	2582071	363914	-1	63975.555	45000	2250.0	2250.0
2	1740877	371185	-7	31815.225	450000	0.0	0.0
3	1389973	337855	-4	236572.110	225000	2250.0	2250.0
4	1891521	126868	-1	453919.455	450000	0.0	11547.0

installments_payments.csv

This dataset gives information about payments, installments supposed to be paid, and their details. There is one row for every made payment and one row for every missed payment.

Table 5. Installments_payments.csv

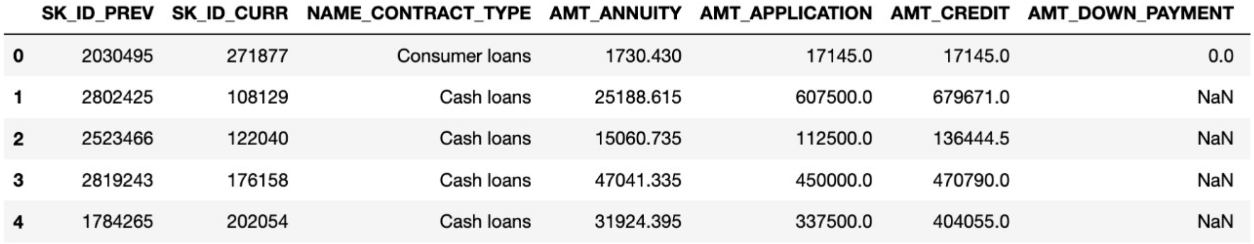


	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	DAYS_INSTALMENT	DAYS_ENTRY_PAYMENT	AMT_INSTALMENT
0	1054186	161674	1.0	6	-1180.0	-1187.0	6948.360
1	1330831	151639	0.0	34	-2156.0	-2156.0	1716.525
2	2085231	193053	2.0	1	-63.0	-63.0	25425.000
3	2452527	199697	1.0	3	-2418.0	-2426.0	24350.130
4	2714724	167756	1.0	2	-1383.0	-1366.0	2165.040

previous_application.csv

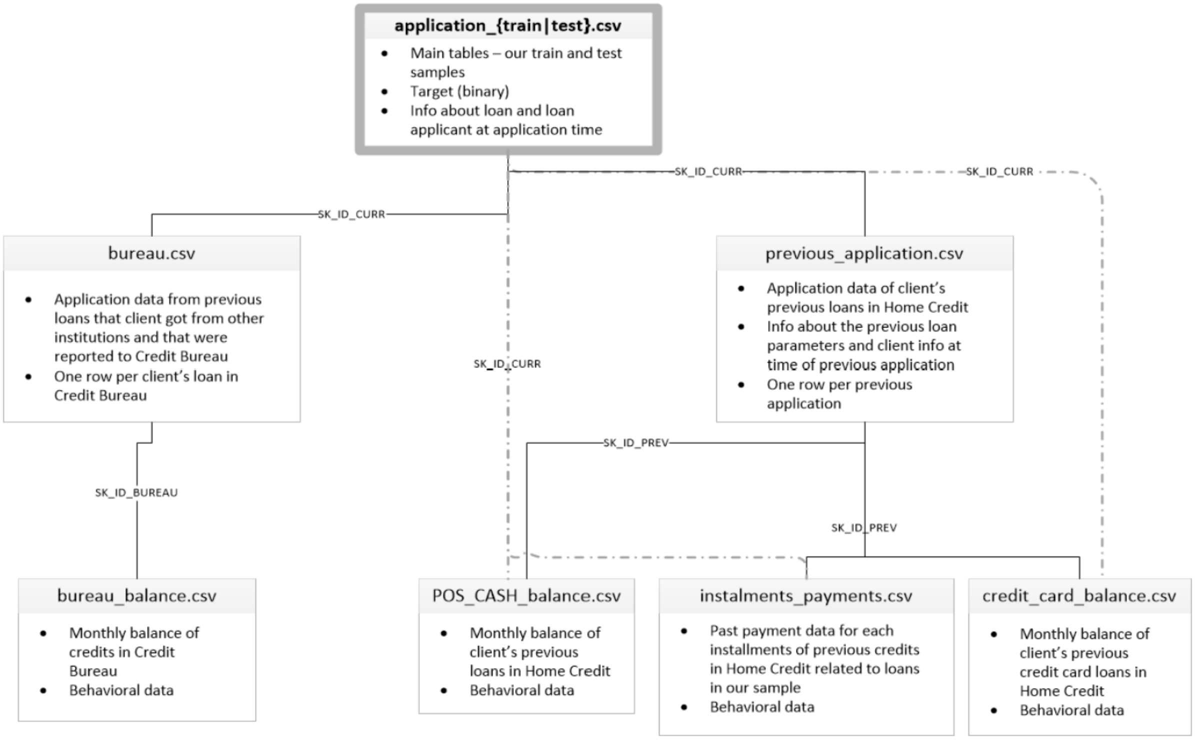
This dataset contains information about previous application details of an application. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK_ID_PREV.

Table 6. previous_application.csv



	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN

Figure 1: Data Description Diagram

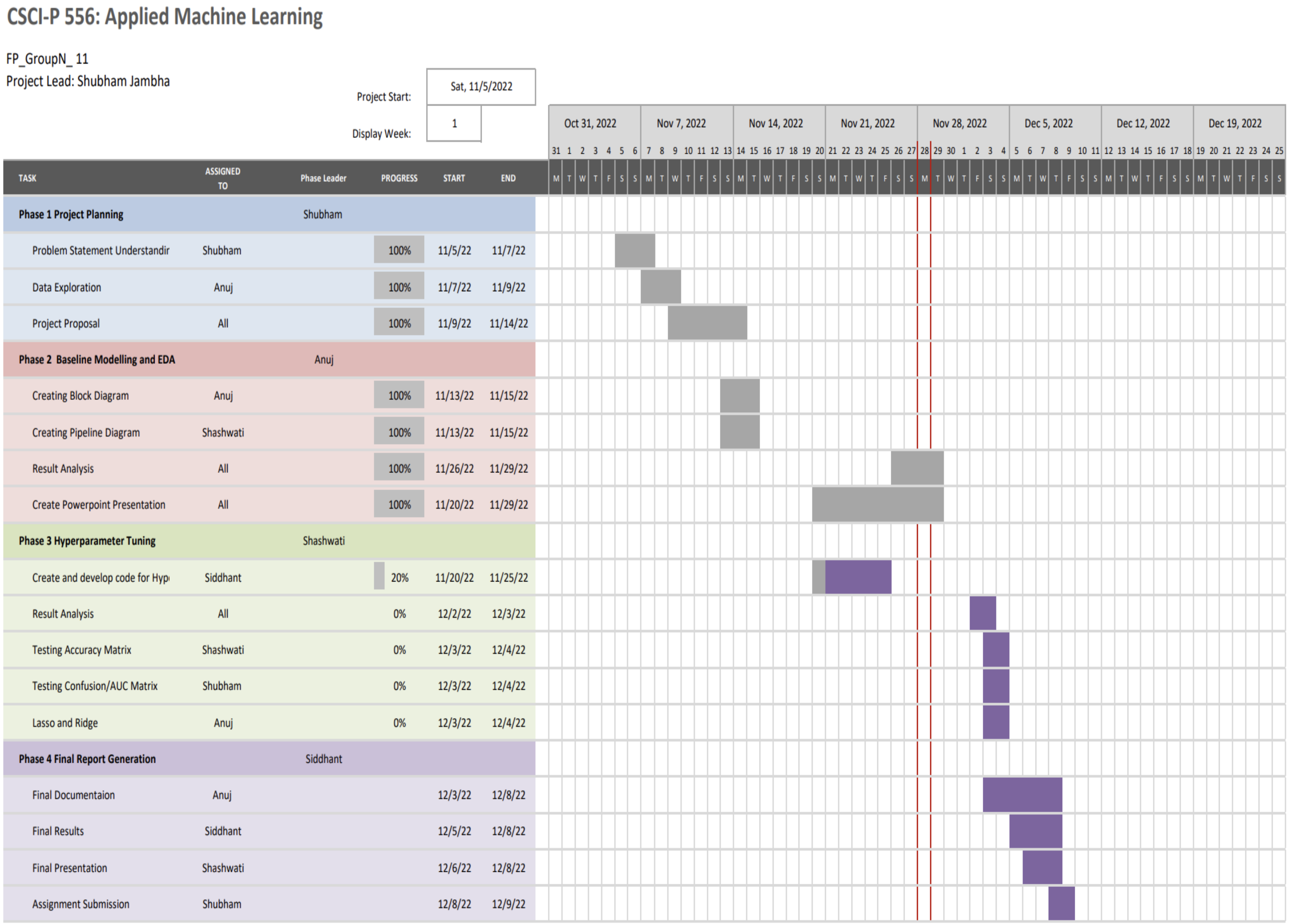

application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Task To Be Tackled:

- We need to predict the customer repayment status such as if the user is a defaulter or not using different baseline machine learning pipelines and models using the datasets provided by Kaggle.

Gantt Chart

Figure 2. Gantt Chart



CSCI-P 556: Applied Machine Learning

FP_GroupN_ 11
Project Lead: Shubham Jambha

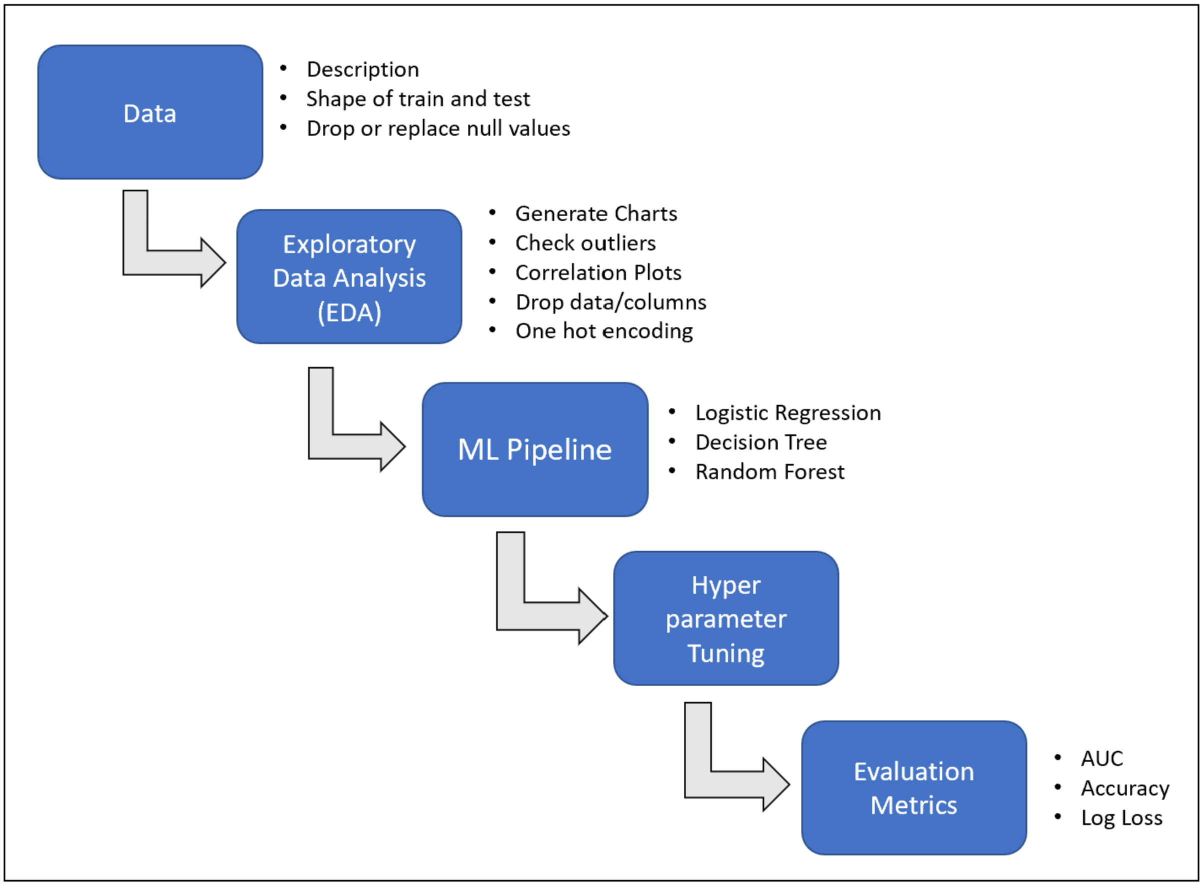
Project Start: Sat, 11/5/2022

Display Week: 1

TASK	ASSIGNED TO	Phase Leader	PROGRESS	START	END
Phase 1 Project Planning		Shubham			
Problem Statement Understandir	Shubham		100%	11/5/22	11/7/22
Data Exploration	Anuj		100%	11/7/22	11/9/22
Project Proposal	All		100%	11/9/22	11/14/22
Phase 2 Baseline Modelling and EDA		Anuj			
Creating Block Diagram	Anuj		100%	11/13/22	11/15/22
Creating Pipeline Diagram	Shashwati		100%	11/13/22	11/15/22
Result Analysis	All		100%	11/26/22	11/29/22
Create Powerpoint Presentation	All		100%	11/20/22	11/29/22
Phase 3 Hyperparameter Tuning		Shashwati			
Create and develop code for Hyp	Siddhant		20%	11/20/22	11/25/22
Result Analysis	All		0%	12/2/22	12/3/22
Testing Accuracy Matrix	Shashwati		0%	12/3/22	12/4/22
Testing Confusion/AUC Matrix	Shubham		0%	12/3/22	12/4/22
Lasso and Ridge	Anuj		0%	12/3/22	12/4/22
Phase 4 Final Report Generation		Siddhant			
Final Documentaion	Anuj			12/3/22	12/8/22
Final Results	Siddhant			12/5/22	12/8/22
Final Presentation	Shashwati			12/6/22	12/8/22
Assignment Submission	Shubham			12/8/22	12/9/22

Machine Learning Workflow:

Figure 3: Workflow



Machine Learning Algorithm and Metrics

The outcome of this project is to predict, whether the customer will repay the loan or not. That's why this is a classification task where the outcome is 0 or 1. To classify this problem we will be building the following machine-learning models:

1. **Logistics Regression**:
   - In our case, the number of features is relatively small i.e. <1000, and no. of examples is large. Hence logistic regression can be a good fit here for the classification.

2. **Decision Tree**:
   - Decision trees are better for categorical data and our target data is also categorical in nature that's why decision trees are a good fit.

3. **Random Forest**:
   - Random Forest works well with a mixture of numerical and categorical features.
   - As we have a good amount of mixture of both types of features random forest can be a good fit.

Loss Function:

Log Loss:

- How closely the forecast probability matches the associated real or true value is indicated by log-loss (0 or 1 in case of binary classification). The higher the log-loss number, the more the predicted probability deviates from the actual value.

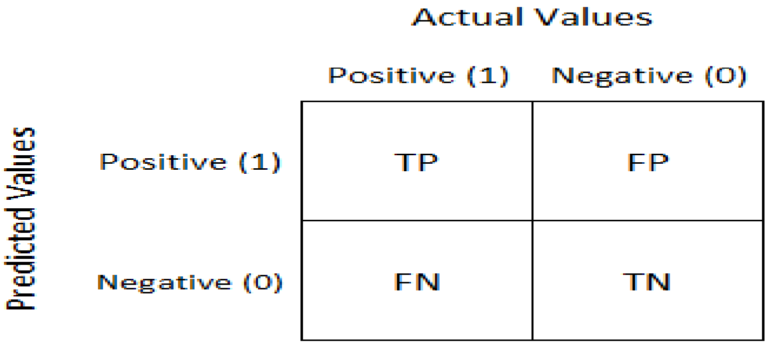
$$\mathrm{logloss}() = \frac{-1}{m} \sum (y \log(p) + (1 - y) \log(1 - p))$$

Metrics:

1. **Confusion Metrics:**

- A confusion matrix, also called an error matrix, is used in the field of machine learning and more specifically in the challenge of classification. Confusion matrices show counts between expected and observed values. The result "TN" stands for True Negative and displays the number of negatively classed cases that were correctly identified. Similar to this, "TP" stands for True Positive and denotes the quantity of correctly identified positive cases. The term "FP" denotes the number of real negative cases that were mistakenly categorized as positive, while "FN" denotes the number of real positive examples that were mistakenly classed as negative. Accuracy is one of the most often used metrics in classification.

Figure 4: Confusion Matrix



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

2. **AUC:**

- AUC stands for "Area under the ROC Curve." It measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). It is a widely used accuracy method for binary classification problems

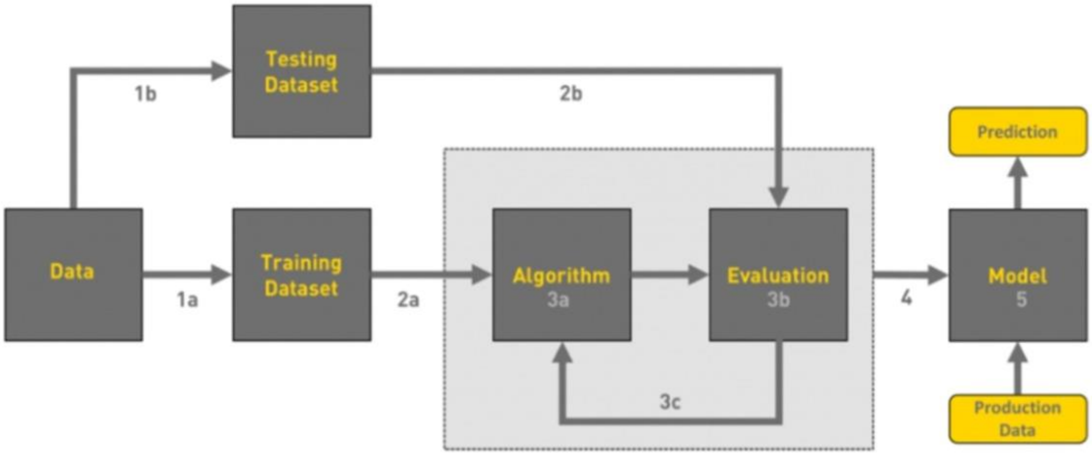
3. **Accuracy:**

- The accuracy score is used to gauge the model's effectiveness by calculating the ratio of total true positives to total true negatives across all made predictions. Accuracy is generally used to calculate binary classification models.

$$\mathrm{Accuracy}() = \frac{True_{P}ositives + True_{N}egatives}{True_{P}ositives + True_{N}egatives + False_{P}ositives + False_{N}egatives}$$

Block Diagram:

Figure 5: Block Diagram of Project



Overview of the Workflow of ML

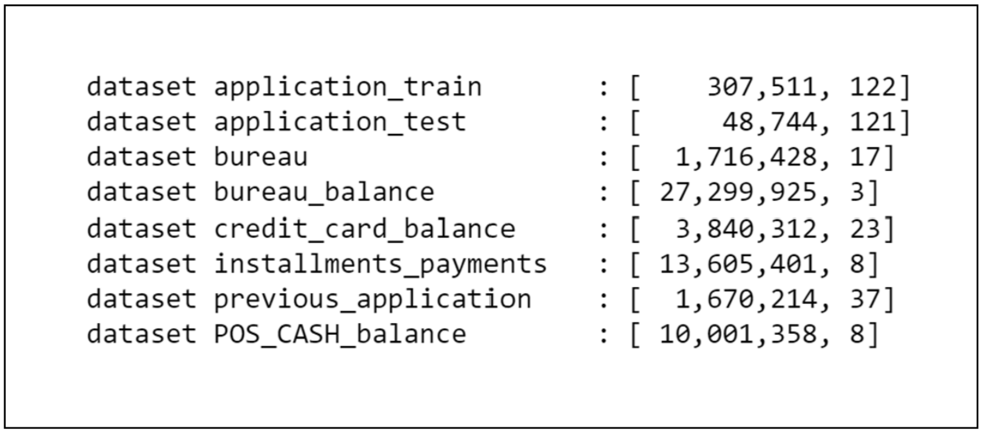
Referenced from: https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94

Exploratory Data Analysis (EDA):

1. Data Set Dictionary and size:

- The table below describes the dataset size like the number of rows and the number of columns.

Figure 6: Data Set Dictionary

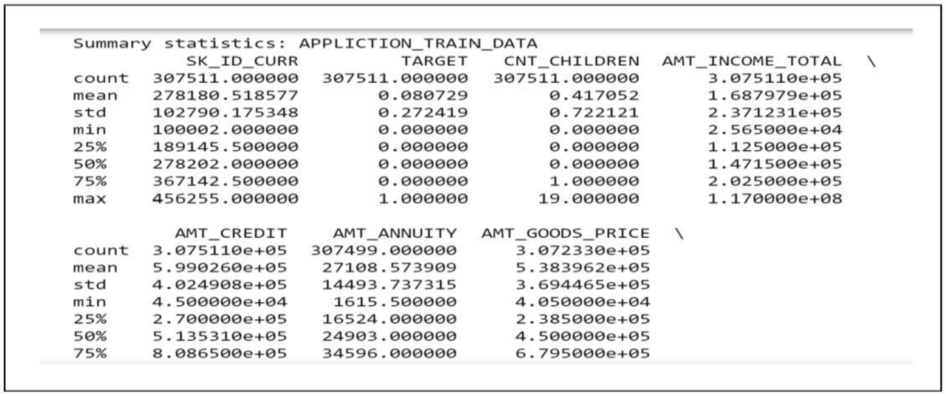


```
dataset application_train       : [    307,511, 122]
dataset application_test        : [     48,744, 121]
dataset bureau                  : [  1,716,428, 17]
dataset bureau_balance          : [ 27,299,925, 3]
dataset credit_card_balance     : [  3,840,312, 23]
dataset installments_payments   : [ 13,605,401, 8]
dataset previous_application    : [  1,670,214, 37]
dataset POS_CASH_balance        : [ 10,001,358, 8]
```

2. Summary Statistic:

- The Table below describes the contents of the table.

Figure 7: Summary Statistic

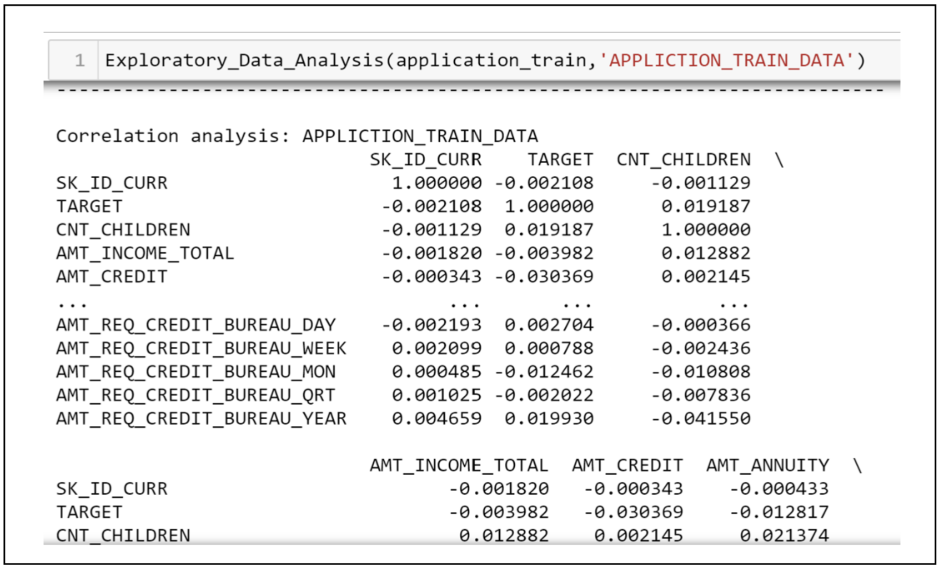


```
Summary statistics: APPLICTION_TRAIN_DATA
           SK_ID_CURR         TARGET   CNT_CHILDREN  AMT_INCOME_TOTAL  \
count   307511.000000  307511.000000  307511.000000      3.075110e+05
mean    278180.518577       0.080729       0.417052      1.687979e+05
std     102790.175348       0.272419       0.722121      2.371231e+05
min     100002.000000       0.000000       0.000000      2.565000e+04
25%     189145.500000       0.000000       0.000000      1.125000e+05
50%     278202.000000       0.000000       0.000000      1.471500e+05
75%     367142.500000       0.000000       1.000000      2.025000e+05
max     456255.000000       1.000000      19.000000      1.170000e+08

          AMT_CREDIT    AMT_ANNUITY  AMT_GOODS_PRICE  \
count   3.075110e+05  307499.000000     3.072330e+05
mean    5.990260e+05   27108.573909     5.383962e+05
std     4.024908e+05   14493.737315     3.694465e+05
min     4.500000e+04    1615.500000     4.050000e+04
25%     2.700000e+05   16524.000000     2.385000e+05
50%     5.135310e+05   24903.000000     4.500000e+05
75%     8.086500e+05   34596.000000     6.795000e+05
```

3. Correlation Analysis:

Figure 8 : Correlation Analysis



```
Exploratory_Data_Analysis(application_train,'APPLICTION_TRAIN_DATA')
```

```
--------------------------------------------------------------------------

Correlation analysis: APPLICTION_TRAIN_DATA
                            SK_ID_CURR    TARGET  CNT_CHILDREN  \
SK_ID_CURR                    1.000000 -0.002108     -0.001129
TARGET                       -0.002108  1.000000      0.019187
CNT_CHILDREN                 -0.001129  0.019187      1.000000
AMT_INCOME_TOTAL             -0.001820 -0.003982      0.012882
AMT_CREDIT                   -0.000343 -0.030369      0.002145
...                                ...       ...           ...
AMT_REQ_CREDIT_BUREAU_DAY    -0.002193  0.002704     -0.000366
AMT_REQ_CREDIT_BUREAU_WEEK    0.002099  0.000788     -0.002436
AMT_REQ_CREDIT_BUREAU_MON     0.000485 -0.012462     -0.010808
AMT_REQ_CREDIT_BUREAU_QRT     0.001025 -0.002022     -0.007836
AMT_REQ_CREDIT_BUREAU_YEAR    0.004659  0.019930     -0.041550

                            AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
SK_ID_CURR                         -0.001820   -0.000343    -0.000433
TARGET                             -0.003982   -0.030369    -0.012817
CNT_CHILDREN                        0.012882    0.002145     0.021374
```

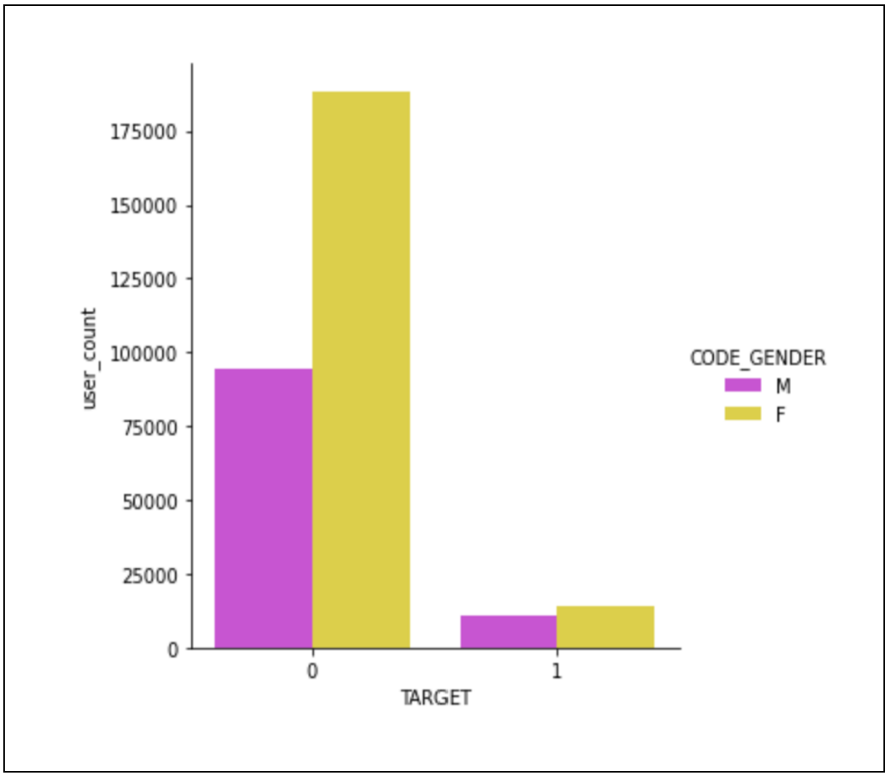
Visual Exploratory Data Analysis (EDA)

- **Descriptive Analysis**

- We performed descriptive analysis on the dataset, identifying the data type for each feature, its size (rows and columns = 307511, 122), and summary statistics for all features, including the number of observations, mean, standard deviation, maximum, minimum, and quartiles.
- We generated charts on descriptive statistics of the target dataset.

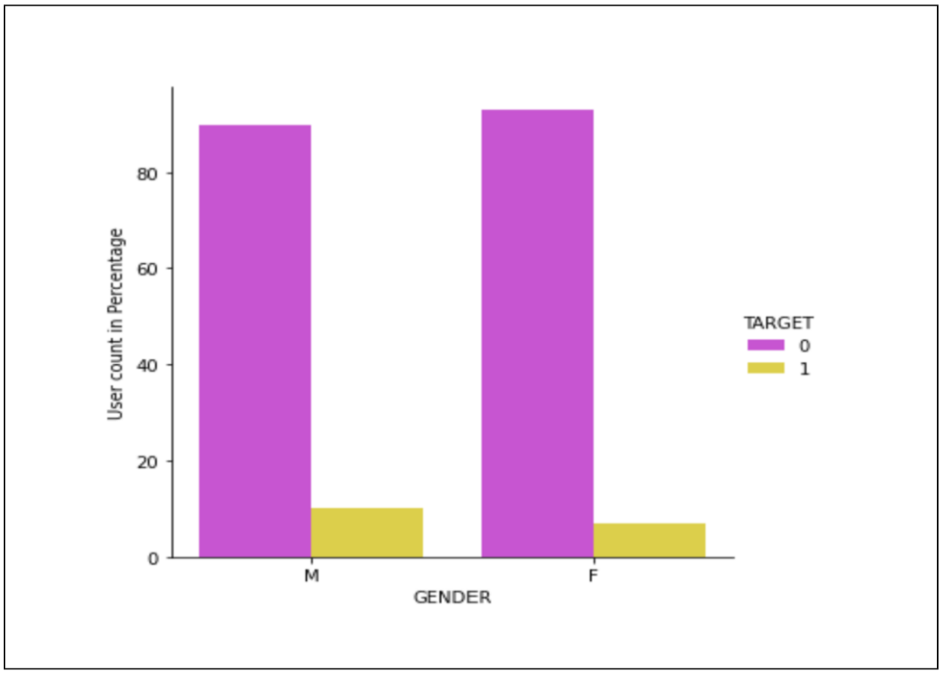
- **Summary Statistics**

Figure 9: Target VS Borrowers based on Gender



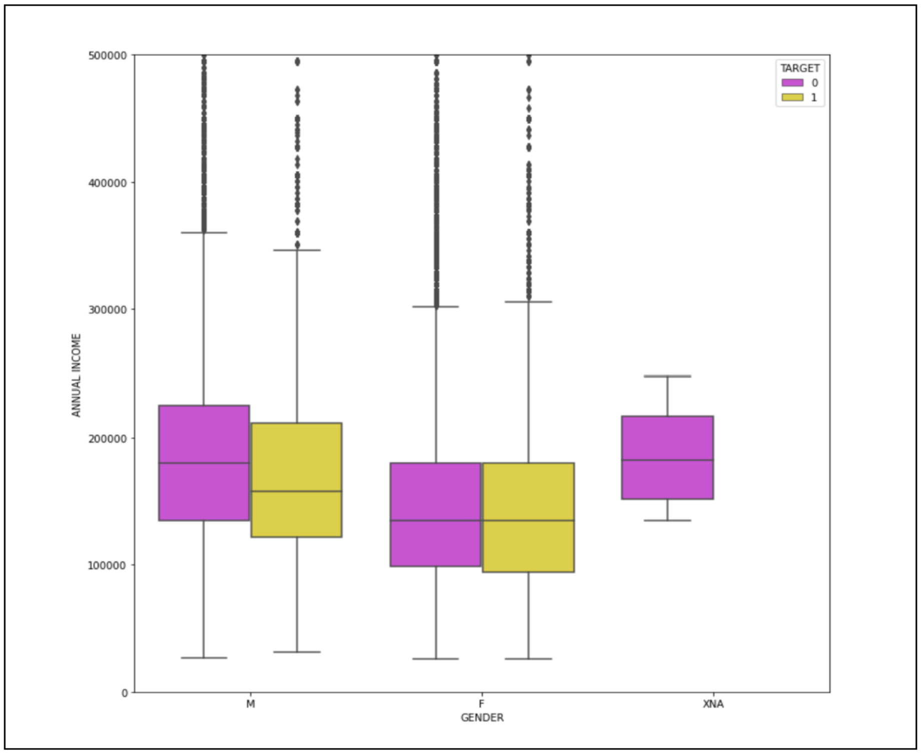
- ➢ Both borrowers and targets are more female than male.
- ➢ In the borrowing group, there was a significantly bigger disparity between men and women.

Figure 10: Target vs Borrower based on gender



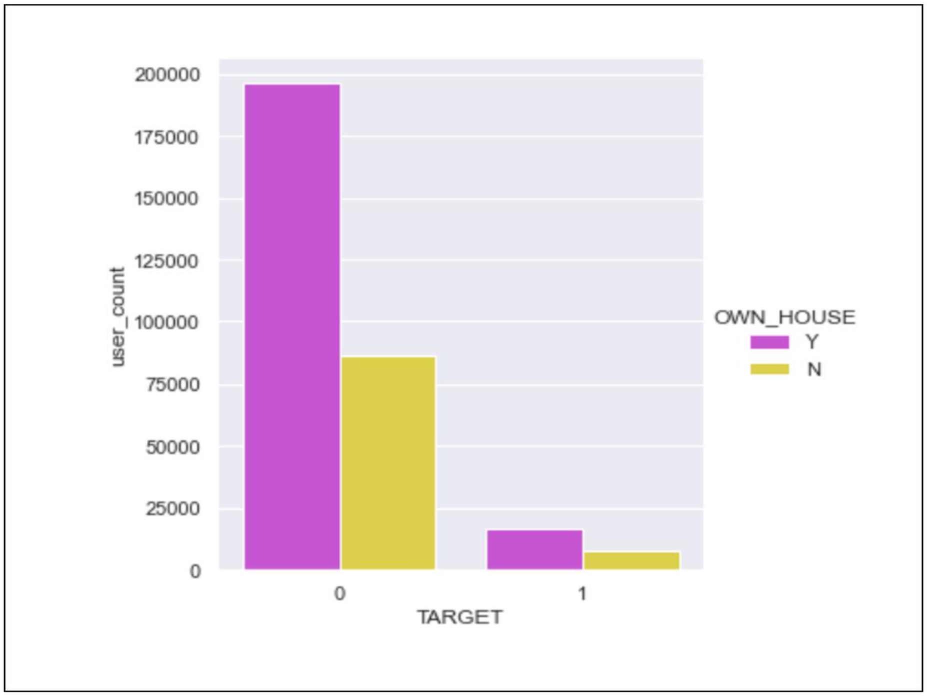
- ➢ Based on the proportion of defaulters count (Second Graph), men are more likely than women to default.

Figure 11: Gender vs Income based on target



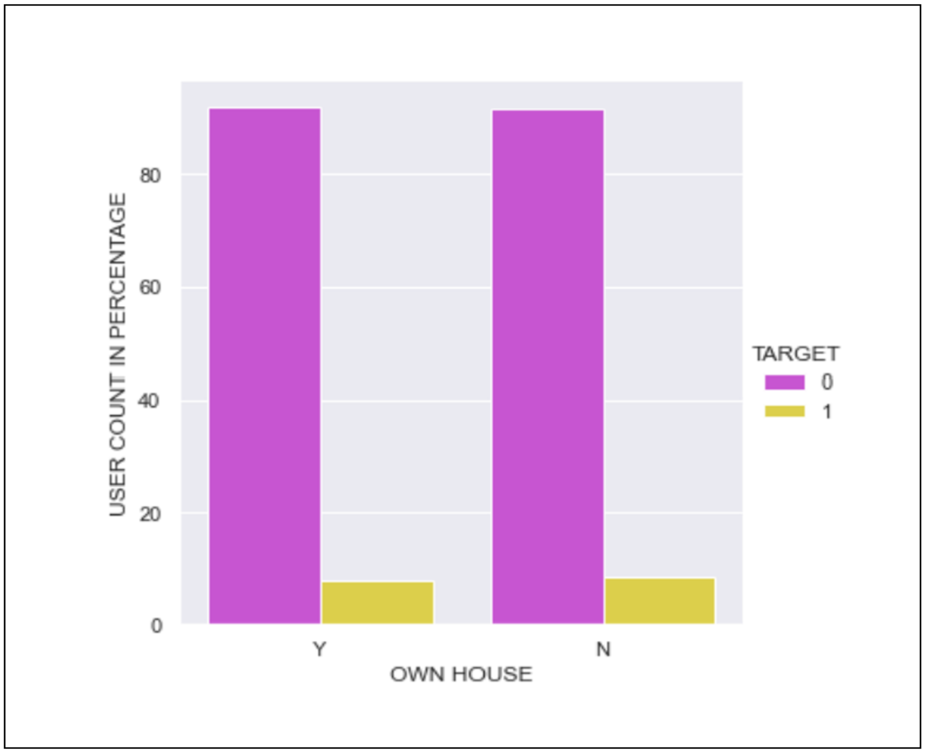
- ➢ In both target and non-target, men earned more than women. Men in non-target groups earn more money than men in target groups.

Figure 12: Own House Count based on Target



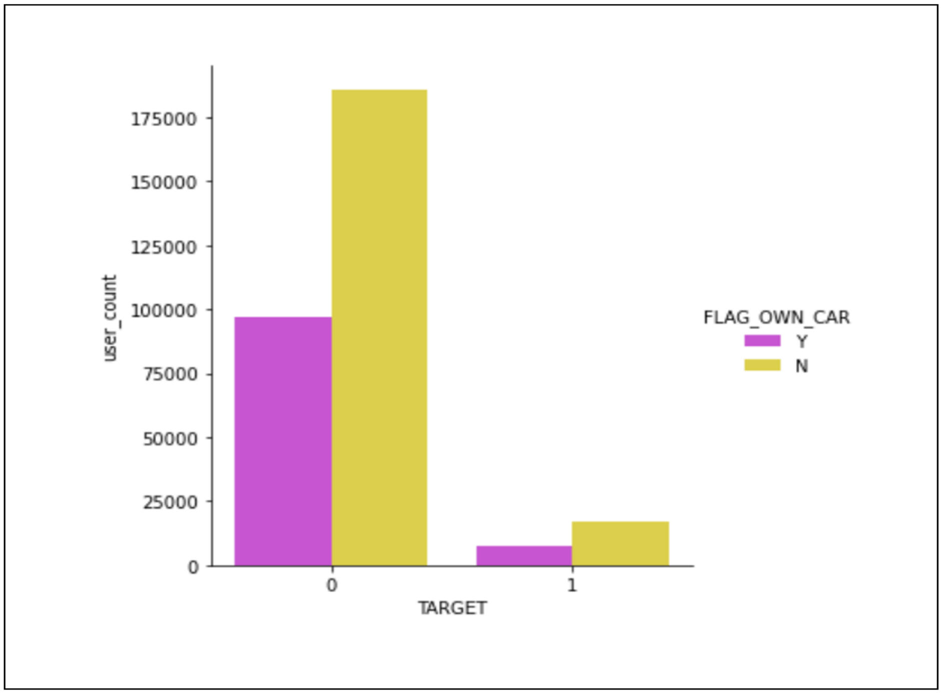
- The number of homeowners in non-target is higher than the number of renters.
- The target has a higher percentage of homeowners than renters.

Figure 13: Own house count based on Target (in percentage)



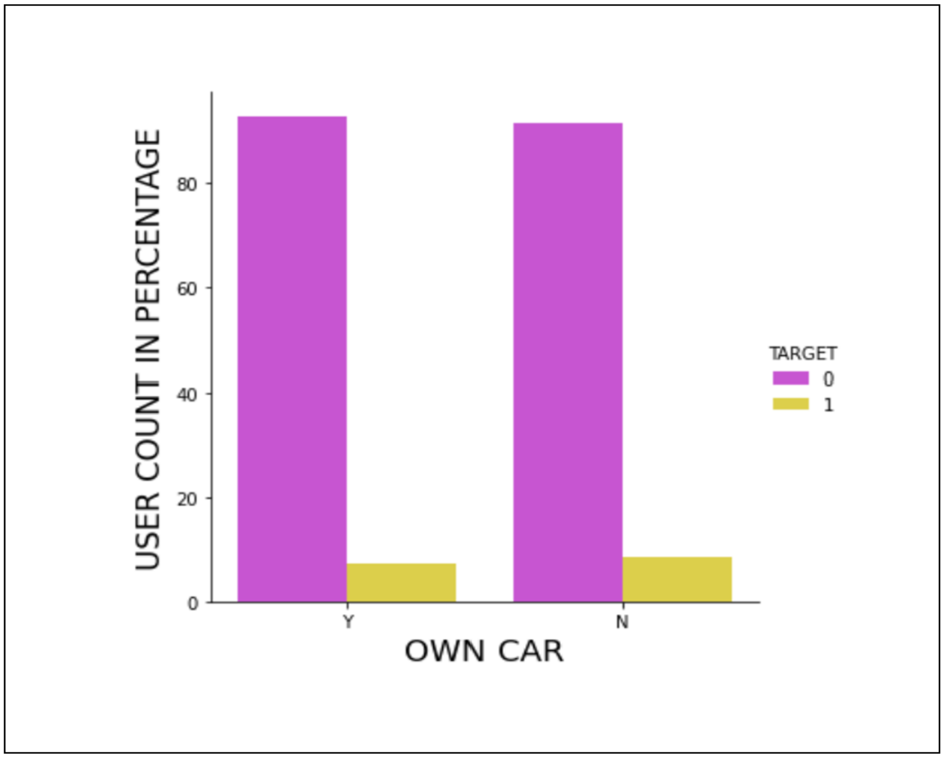
- Borrowers who own a home are more likely to pay back loans, while the difference is not great.

Figure 14: Own Car count based on Target



- ➢ Both in the target and non-target, there are more persons without automobiles than those who do.

Figure 15: Own Car count based on target (in percentage)



- ➢ Borrowers who own cars are more likely to make timely payments.

Figure 16: Occupation Type Count based on Target

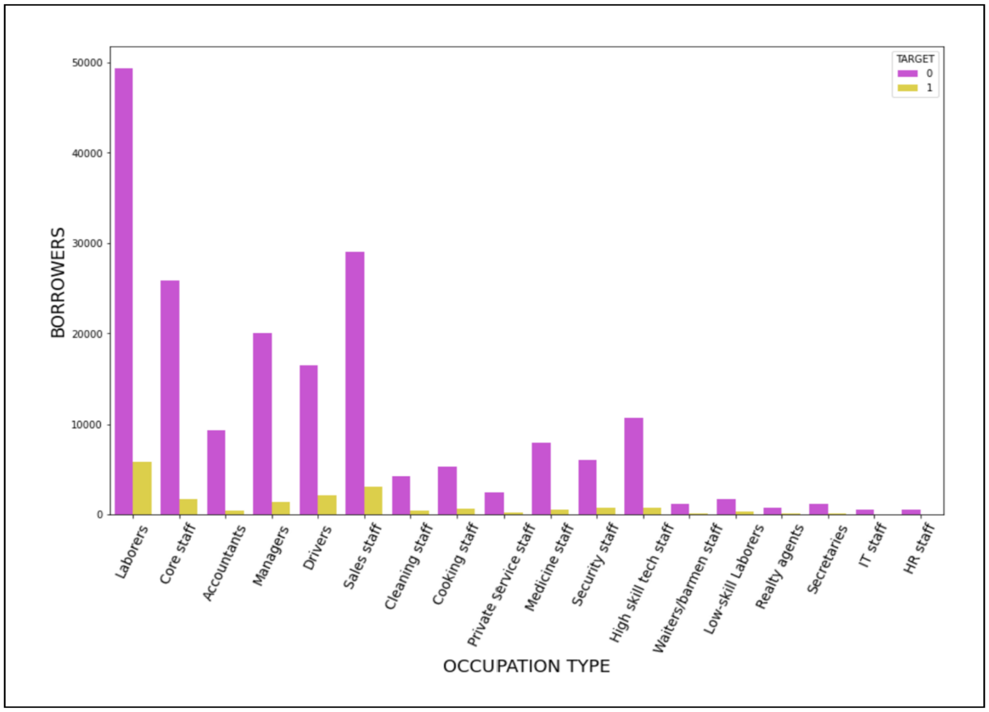


Figure 17: Occupation Type vs Income based on Target

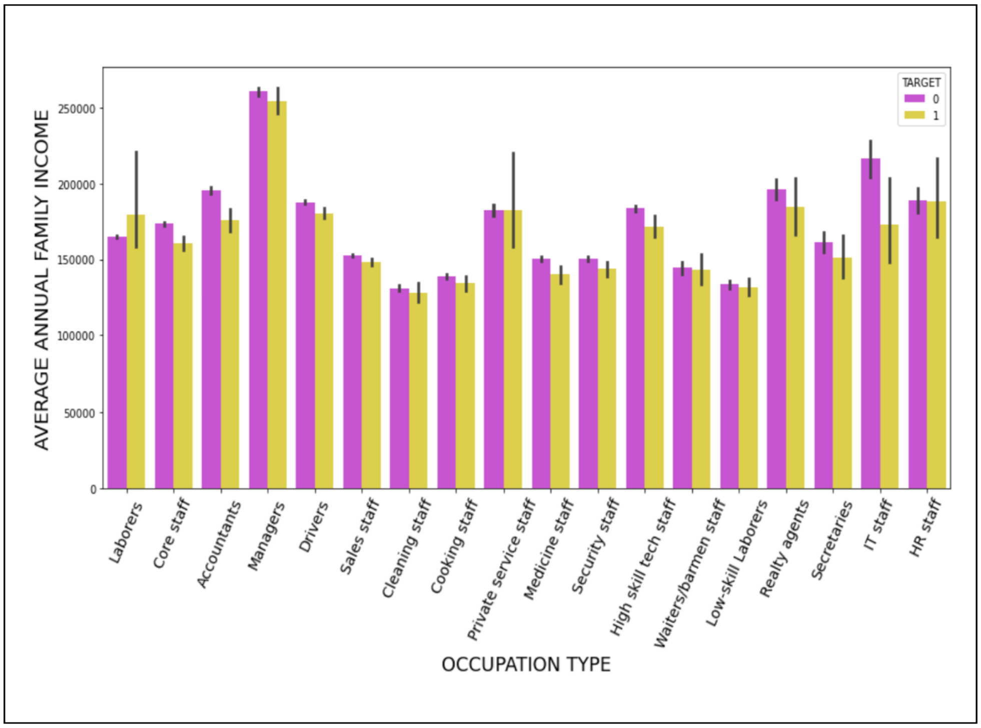
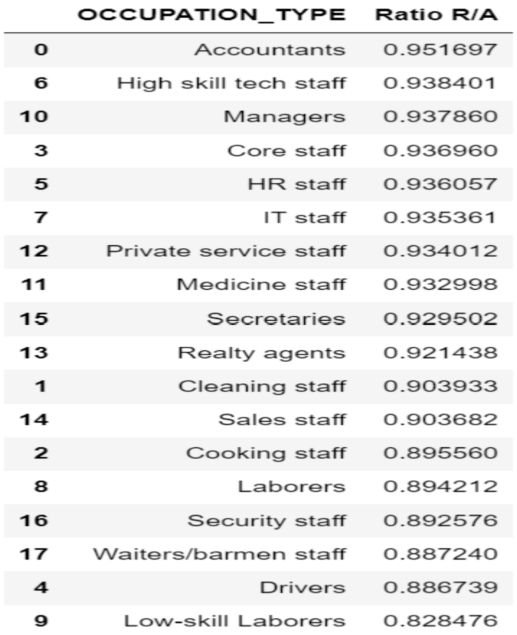


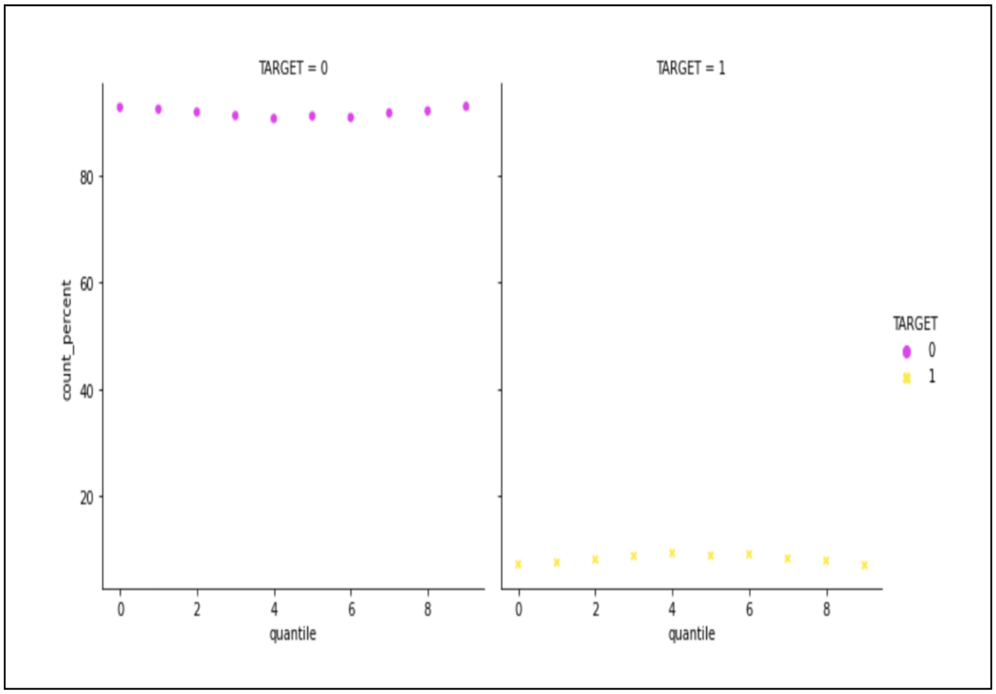
Figure 18: Re Payers to Application Ratio



	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

- The above figure describes the ratio of repayment based on occupation type.

Figure 19: Quantiles vs Income Credit Ratio



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

Feature Extraction

- **Step 1:**
  - We eliminated columns with more than 50% Null Values. The table below displays the number of NA values and their percentages for the remaining columns.
- **Step 2:**
  - Following columns give us the number of enquiries done.
  - AMT_REQ_CREDIT_BUREAU_HOUR
  - AMT_REQ_CREDIT_BUREAU_DAY
  - AMT_REQ_CREDIT_BUREAU_WEEK
  - AMT_REQ_CREDIT_BUREAU_MON
  - AMT_REQ_CREDIT_BUREAU_QRT
  - AMT_REQ_CREDIT_BUREAU_YEAR

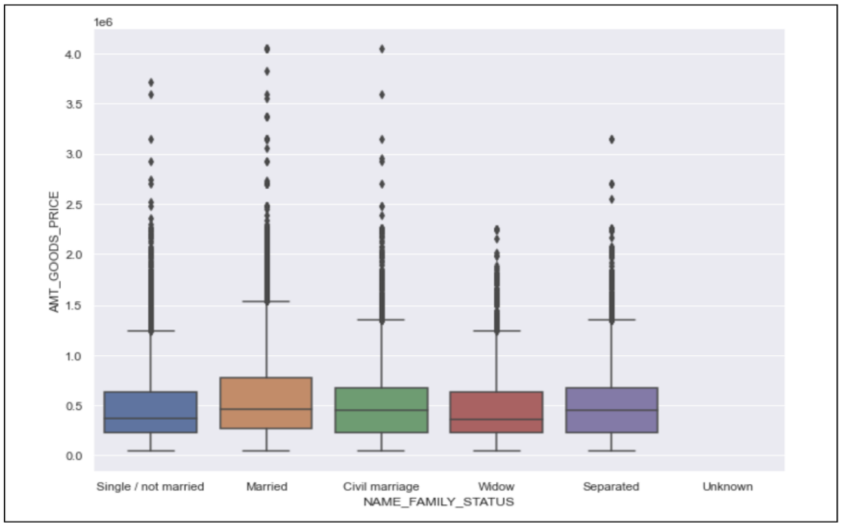
In all these columns for some entries values are not mentioned, so we can assume that there are no enquiries for those inputs. We can replace null values in such inputs by 0

- **Step 3:**
  - Following columns give us the number of immediate connections who have a loan in Home Credit.
  - OBS_30_CNT_SOCIAL_CIRCLE
  - DEF_30_CNT_SOCIAL_CIRCLE
  - OBS_60_CNT_SOCIAL_CIRCLE
  - DEF_60_CNT_SOCIAL_CIRCLE

In all these columns for some entries, values are not mentioned, so we can assume that there are no immediate connections for those inputs. We can replace null values in such inputs with 0.

- **Step 4:**
  - AMT_GOODS_PRICE values are depending on NAME_FAMILY_STATUS categories. So we replaced null/N/A values with medians with respect to NAME_FAMILY_STATUS. We can see in the below figure that AMT_GOODS_PRICE depends on NAME_FAMILY_STATUS.

Figure 20: Name_Family Status



- **Step 5:**
    - Used median to fill in the empty or missing values for CNT FAM MEMBERS

- **Step 6:**
    - In order to replace the EXT SOURCE 2 null or N/A values, we discovered the some of the highly associated variables. We found out that there is high correlation between EXT SOURCE 2 and REGION RATING CLIENT. Due to the categorical nature of REGION RATING CLIENT, we fill null or NA values with the median depending on categories.

- **Step 7:**
    - Same as step 8, we need to replace EXT SOURCE 3 null or N/A values. For this, we discovered some of the highly associated variables. We found out that there is a high correlation between EXT SOURCE 3 and DAYS BIRTH. Given that DAYS BIRTH is numerical, we used Linear Regression to fill in the null or NA values.

Including new features in the training data.

We tested and trained on a few chosen columns mentioned below,

- Salary-to-Credit Ratio
- Total External Source

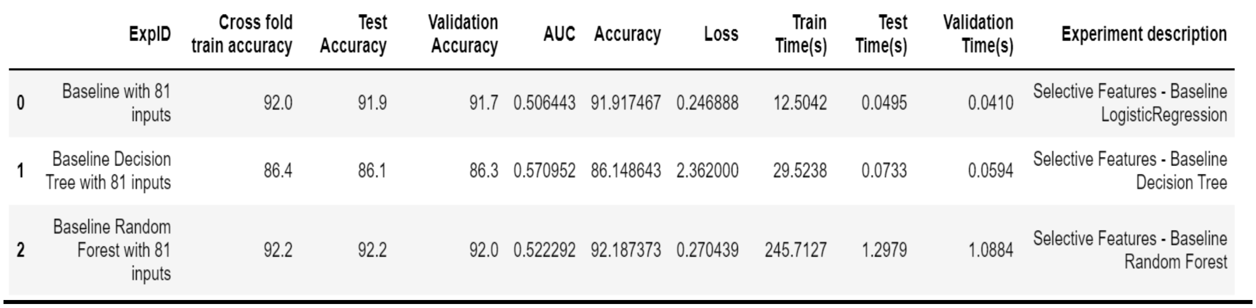
We also tried a number of other features, such as segregated bins based on AMT ANNUITY and AMT SALARY based on percentile, however, they were ineffective for us in terms of improving accuracy or AUC. We also tried performing modeling over Baseline using one-hot encoding rather than label encoding and compared the results, however, we found no improvement in AUC or accuracy.

Why we choose the technique and strategy:

- We removed some of the features which has most number of null values, as they will contribute a very little in the predictions.
- For other features, It is best to handle filling in null values for categorical and continuous variables individually. It won't be possible to fill all category variables with the most or least frequent values and all continuous variables with the median value.
- A reliable metric to assess a person's reliability and repayment capacity would be ratios between income, credit requested, and credit to be paid per year. So, we thought about putting above.

Experiments:

Table 7: Experiment Logs

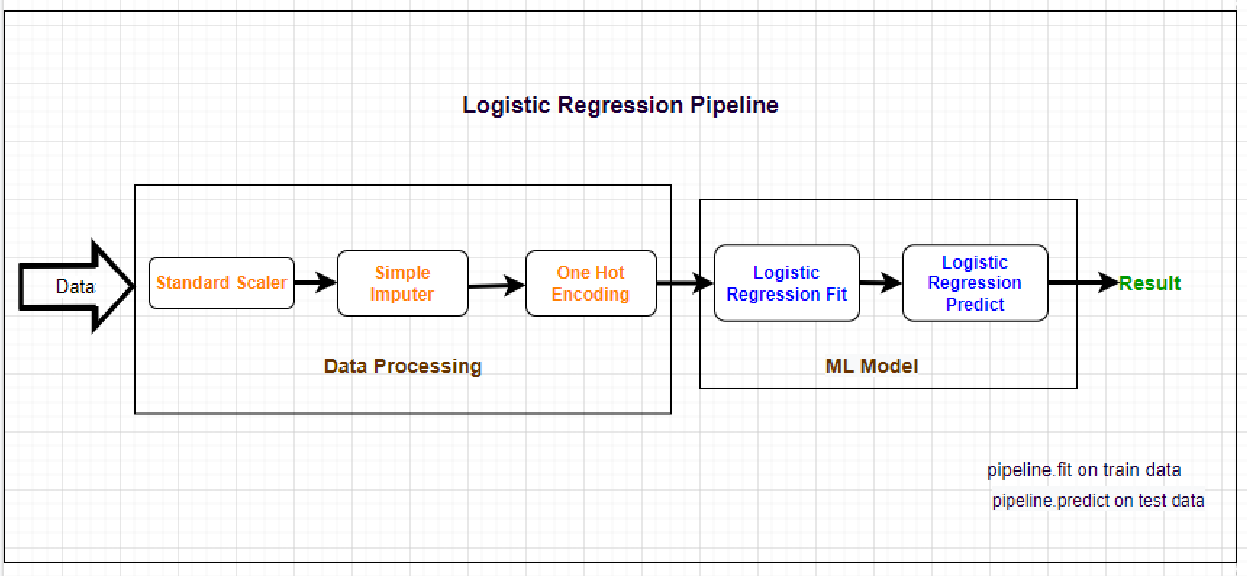


	ExpID	Cross fold train accuracy	Test Accuracy	Validation Accuracy	AUC	Accuracy	Loss	Train Time(s)	Test Time(s)	Validation Time(s)	Experiment description
0	Baseline with 81 inputs	92.0	91.9	91.7	0.506443	91.917467	0.246888	12.5042	0.0495	0.0410	Selective Features - Baseline LogisticRegression
1	Baseline 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

1. Logistic Regression:

- Train and test data were separated. With a random seed set to 42, we divided the 20% test data for accurate findings.
- Next, a logistic regression baseline process was constructed. Based on numerical properties and a common scaler, we construct a numerical pipeline. We use the median to impute the missing data. With this numerical pipeline, logistic regression is performed.
- Finally, using 5 splits and a test size of 0.3, we generate cross-validation splits. We use this cross-validation to compute test accuracy and AUC.
- In Logistic Regression we can see that the Testing Accuracy is quite high and a descent AUC, so logistic regression can be a good model for this dataset
- The Log Loss (0.25) for logistic regression is on the lower side which means our model is predicting accurate results.

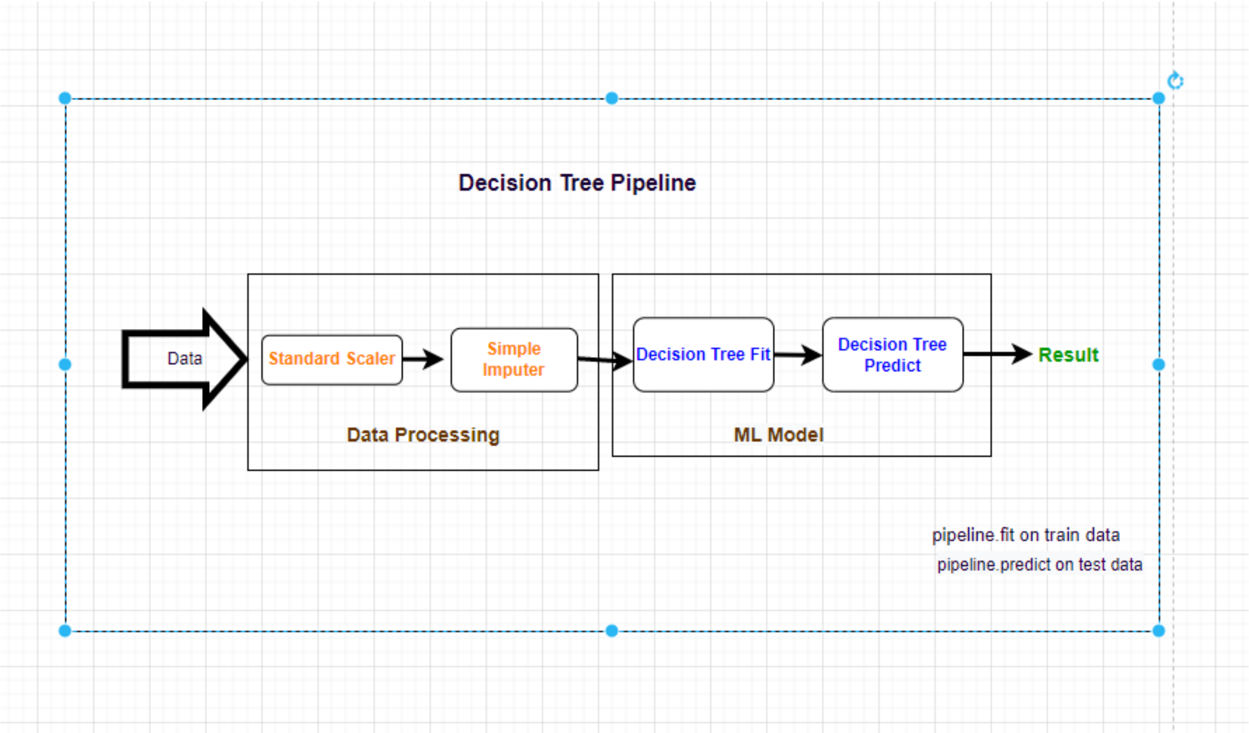
Figure 21: Logistic Pipeline


Logistic Regression Pipeline
Data
Standard Scaler
Simple Imputer
One Hot Encoding
Data Processing
Logistic Regression Fit
Logistic Regression Predict
ML Model
Result
pipeline.fit on train data
pipeline.predict on test data

2. Decision Tree:

- Train and test data were separated. With a random seed set to 42, we divided the 20% test data for accurate findings.
- Next, a decision tree baseline process was constructed. We use the median to impute the missing data.
- Finally, using 5 splits and a test size of 0.3, we generate cross-validation splits. We use this cross-validation to compute test accuracy and AUC.
- For the decision tree, as compared to logistic regression the test accuracy is on the lower side but AUC has increased significantly.
- This loss of test accuracy may be due to the short dept of the decision tree, compared to the number of variables we have used.
- The Log Loss for the Decision Tree is relatively high as compared to Logistic regression and Random Forest

Figure 22: Decision Tree Pipeline

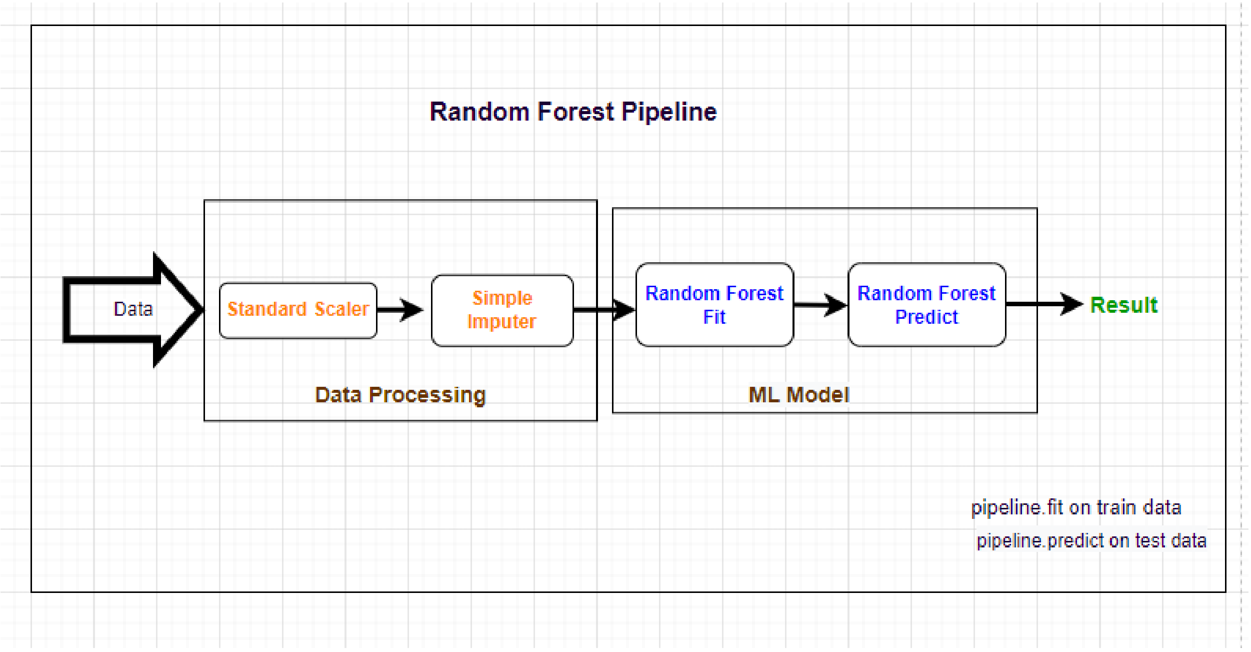


3. Random Forest:

- Train and test data were separated. With a random seed set to 42, we divided the 20% test data for accurate findings.
- Next, a random forest baseline process was constructed. We use the median to impute the missing data.
- Finally, using 5 splits and a test size of 0.3, we generate cross-validation splits. We use this cross-validation to compute test accuracy and AUC.
- After running the experiments on the baseline algorithm, the random forest provides us with the best test accuracy (92.2).

- As compared to the decision tree, AUC is decent. So random forest is also the best fit for the provided dataset
- The Log Loss (0.27) for Random Forest is on the lower side which means our model is predicting accurate results.

Figure 23: Random Forest Pipeline



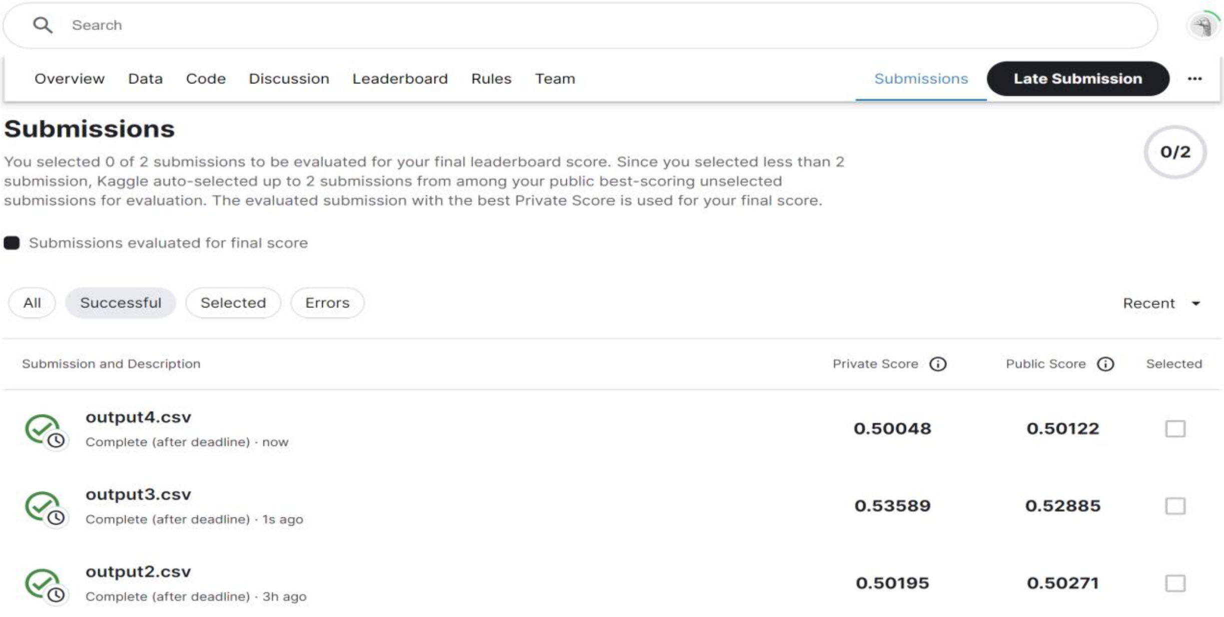
Result and Discussion:

Table 7 describes the accuracy, AUC, and loss of baseline machine learning model logistic regression, Decision Tree, and random forest. For the baseline logistic regression model, we can see that the train (92.0) and test (91.9) accuracy is on the higher side, which means logistic regression is performing well on the provided dataset. The log loss for logistic regression is on the lower side which is 0.25 which means most of our predictions are correct. For logistic regression the AUC of 0.5 is significant and the accuracy of 91.94 is on the higher side. So, the algorithm is performing well for given inputs.

Both Random Forest and logistic regression have approximately the same train and test accuracy and log loss. But baseline Random Forest remains the best-fit algorithm as it beats the logistic regression by a very small margin in all the criteria. We observed a slight increase of 0.02 in AUC, 0.3 in test accuracy, and 0.3 in overall accuracy. The log loss for random forest (0.27) is on the lower side and hence it beats the baseline logistic regression model.

The decision tree has comparatively low train and test accuracy. This loss of test accuracy may be due to the short dept of the decision tree, compared to the number of variables we have used even though we see the drop in test accuracy, in return, we can see a little increase of 0.07 in AUC as compared to baseline random forest and logistic regression. So, the decision tree can also be a good fit for a given dataset, but we might need to finetune our datasets a little extra.

Figure 24: Kaggle 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

Submission and Description	Private Score	Public Score	Selected
output4.csv Complete (after deadline) · now	0.50048	0.50122	□
output3.csv Complete (after deadline) · 1s ago	0.53589	0.52885	□
output2.csv Complete (after deadline) · 3h ago	0.50195	0.50271	□

Conclusion

The HCDR project's goal is to forecast the population's capacity for payback among those who are underserved financially. Because both the lender and the borrower want reliable estimates, this project is crucial. Real-time Home credit's ML pipelines, which acquire data from the data sources via APIs, run EDA, and fit it to the model to generate scores, which allows them to present loan offers to their consumers with the greatest amount and APR.

Hence if NPA expected to be less than 5% in order to maintain a profitable firm, risk analysis becomes extremely important. Credit history is an indicator of a user's trustworthiness that is created using parameters such as the average, minimum, and maximum balances that the user maintains, Bureau scores that are reported, salary, etc. Repayment patterns can be analysed using the timely defaults and repayments that the user has made in the past. Other criteria such as location information, social media data, calling/SMS data, etc. are included in alternative data. As part of this project, we would create machine learning pipelines, do exploratory data analysis on the datasets provided by Kaggle, and evaluate the models using a variety of evaluation measures before deploying one.

Phase 2 involved the estimation of several models. Data imputation and feature selection were done. We started by selecting features and imputed values. The values of certain features that were missing were filled in. Then, based on our past understanding, we chose to include pertinent features. We trained and assessed several models, including Random Forest, Decision Tree Model, and Logistic Regression, to discover the best one.

We have concluded from phase 2 that the decision tree model is unable to defeat the baseline model. The random forest model performs the best out of all the models. In phase 3 we plan to implement all models through hyper-tuning of their parameters.

Bibliography:

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition

- by Aurélien Géron

Lab-End_to_end_Machine_Learning_Project

- by James Shahnan