Project Title

Home Credit Default Risk (HCDR) Final Project

Team and Phase Leader Plan

Team Name: FP_Group_3

Phase 2 Leader: Katia Torres Sanchez

Team Members
Alicia Aaholm
Nicholas Chappell
Jaden Costa
Katia Torres Sanchez

Credit Assignment Plan

Phase	Task	Person Responsible	Estimated Person-Hrs
Phase 2	Jaden will succesfully download all datasets via Kaggle API	Jaden	1 hr
Phase 2	Alicia will perform visual and statistical EDA on all datato identify quality issues and important feature patterns	Alicia	1 hr
Phase 2	Nicholas will establish metrics that will be used to test the effectiveness of the model created by using latex and laymans terms	Nicholas	1 hr
Phase 2	The team (through collective efforts) will build the baseline pipelines by using block diagrams, code, and subsets of data	All	6 hrs
Phase 2	Team will provide a table of experimental results from the baseline pipelines	All	30 mins
Phase 2	Team will describe the baseline pipelines	All	30 mins
Phase 2	Katia will write a brief report that adheres to the guidelines in Canvas Phase 2 Assignmnet and its rubric	Katia	1 hr
Phase 2	Team will present a 2 minute status update via video	All	2 mins
Phase 2	Team will create 4-5 slides for the status update video that includes what has been accomplished in Phase 2, the present status, what is planned for Phase 3, and a analysis	All	30 mins

Project Abstract

The HCDR project is designed to tackle the problem of accurately predicting which applicants will default on their home credit loan. Phase 2 focuses on establishing a foundational understanding of the dataset through exploratory data analysis (EDA) and developing baseline machine-learning pipelines using the HCDR datasets. The primary goal of Phase 2 was to explore the structure and quality of the datasets to identify missing patterns and key correlations that will influence the model's performance. Evaluation metrics were defined in LaTeX and interpreted to clarify how each metric reflects model behavior. Then, baseline pipelines were implemented using subsets of the data. The pipelines included logistic regression, random forest, LGBM, XGBoost, and CatBoost. The results demonstrate that the LightGBM pipeline is the best model as of now, which can continue to be improved during Phase 3.

Introduction

For this phase, our team focused on performing Exploratory Data Analysis and building basic pipelines for our models. Our goal with performing these steps was to develop a better understanding of the patterns/trends that exist within the data so that we could establish basic pipelines that we believed would allow us to develop effective models.

Dataset

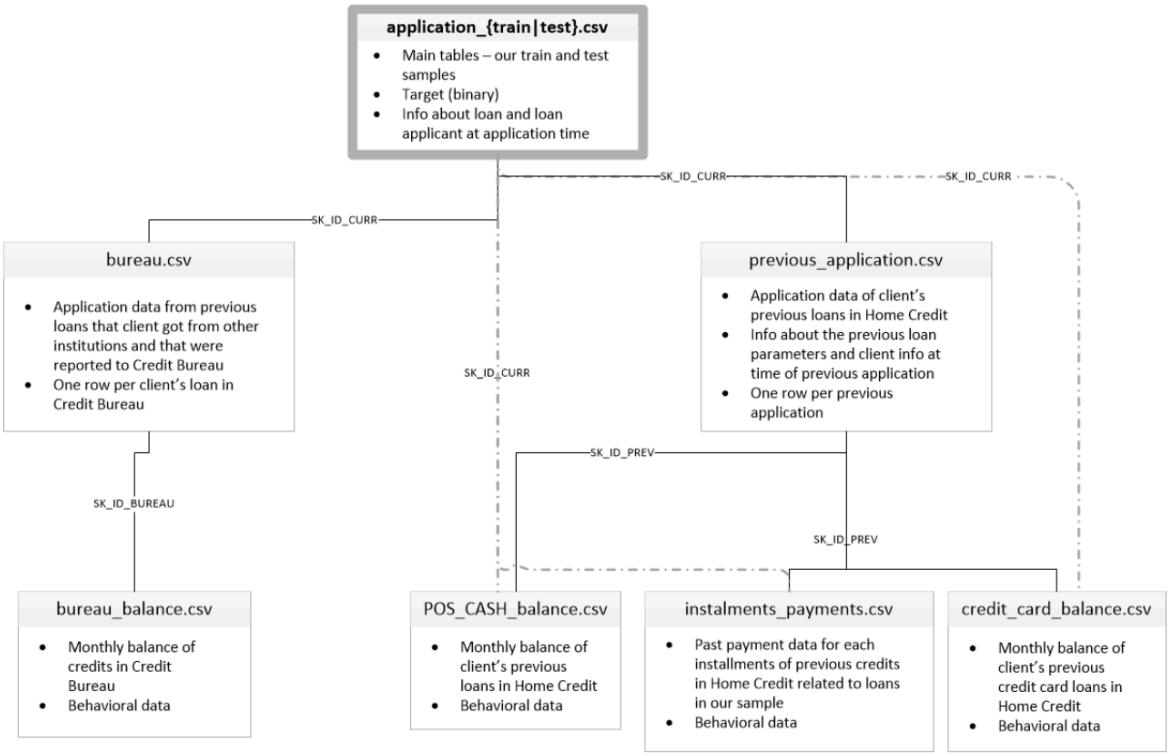
We plan to use the data sets that are provided by the HCDR Kaggle Competition file. The competition file contains multiple csv files, each of which is important for the functionality of the model being constructed. These csv files include:

- Application Data: Includes the majority of information regarding clients, such as the gender, income, family status, education, number of children, amount of income, and whether the client owns a car or not. The application data has split the main training data from the testing data.
- Bureau Data: This dataset contains data on clients' credit history.
- Bureau_Balance Data: This dataset breaks down the credit history by months.
- Previous_Application Data: Contains information on clients who have had previous applications for loans from Home Credit.
- POS_CASH_BALANCE Data: Breaks down point of sale and cash loan data by months.
- Credit_Card_Balance Data: Contains data for clients who have had a Home Credit credit card.
- Installments_Payment Data: Includes payment histories for clients who have had a previous Home Credit loan.

Tasks to be tackled include performing Exploratory Data Analysis on key features and datasets. EPA will provide foundational understanding of what features have correlation to increase success of predictions.

Workflow of Dataset

The following image is a workflow of all the datasets. The datasets connect with each other throguh the SK_ID_CURR, SK_ID_PREV, or SK_ID_BUREAU fields.



EDA

Data Dictionary of Raw Features

There are 221 raw features included in the Data Dictionary for the HCDR Kaggle Project. Below are some key raw features we are focusing on, pulled from the Data Dictionary CSV file. The identified features are not all-inclusive. The full list of raw features can be found in HomeCredit_columns_description.csv.

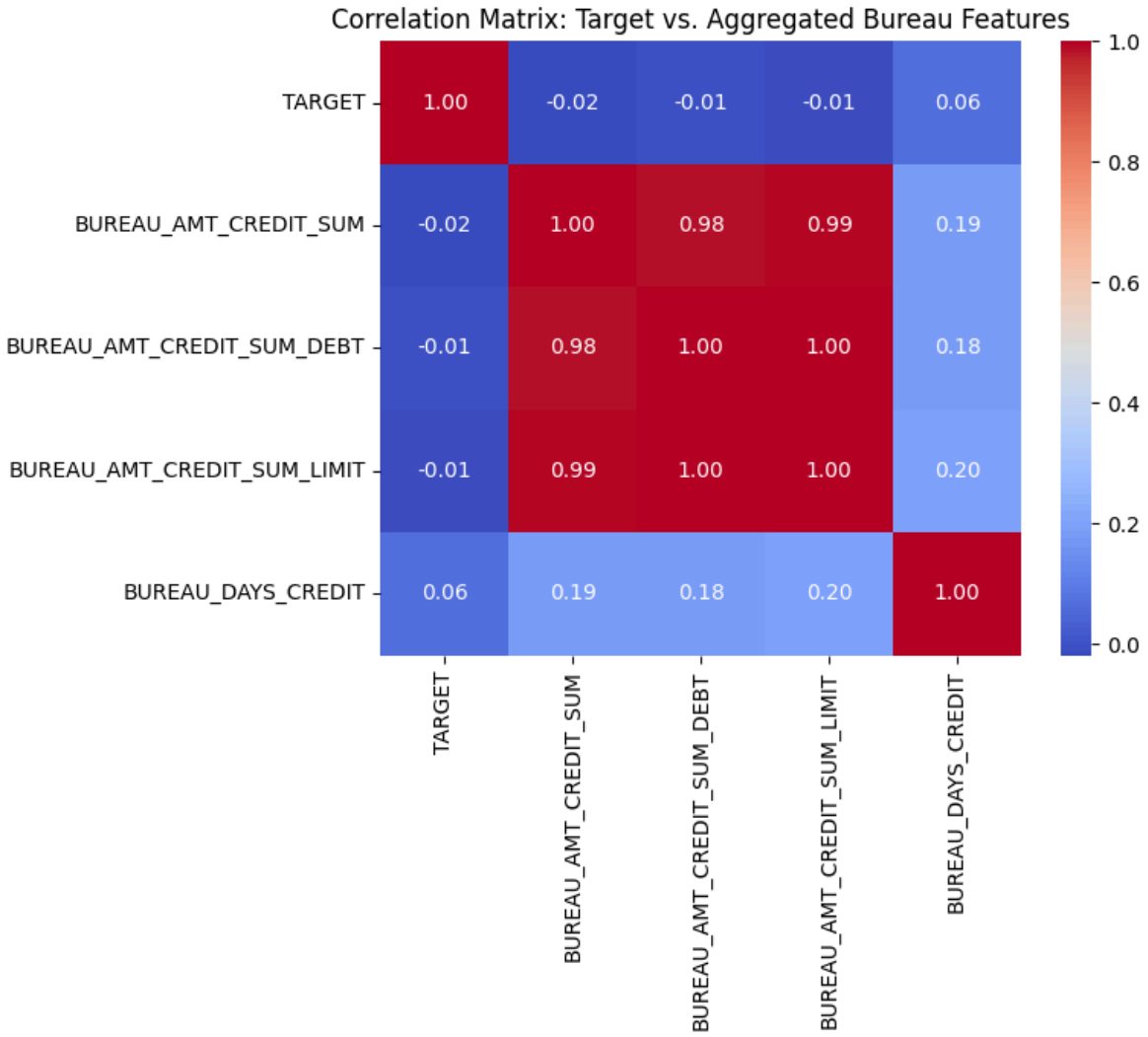
Raw Feature	Test Description	Data Type
SK_ID_CURR	ID of loan in the sample	Integer
TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)	Integer (1 for True, 0 for False)
NAME_CONTRACT_TYPE	Identifies if loan is cash or revolving	String/nvarchar
AMT_INCOME_TOTAL	Client's income	Float (rounded to 2 decimal places)
AMT_CREDIT	Loan's credit amount	Float(rounded to 1 decimal place)
AMT_ANNUITY	Loan annuity	Float(rounded to 1 decimal place)
NAME_INCOME_TYPE	Client's income type	String/nvarchar
NAME_HOUSING_TYPE	Client's housing situation	String/nvarchar
AMT_CREDIT_SUM_DEBT	Client's current debt on Credit Bureau credit	Float (rounded to 1 decimal place)
AMT_BALANCE	Balance during the month of previous credit	Float (rounded to 3 decimal places)
REGION_RATING_CLIENT	The rating of the region where client lives (1,2,3)	Integer

Dataset Sizes

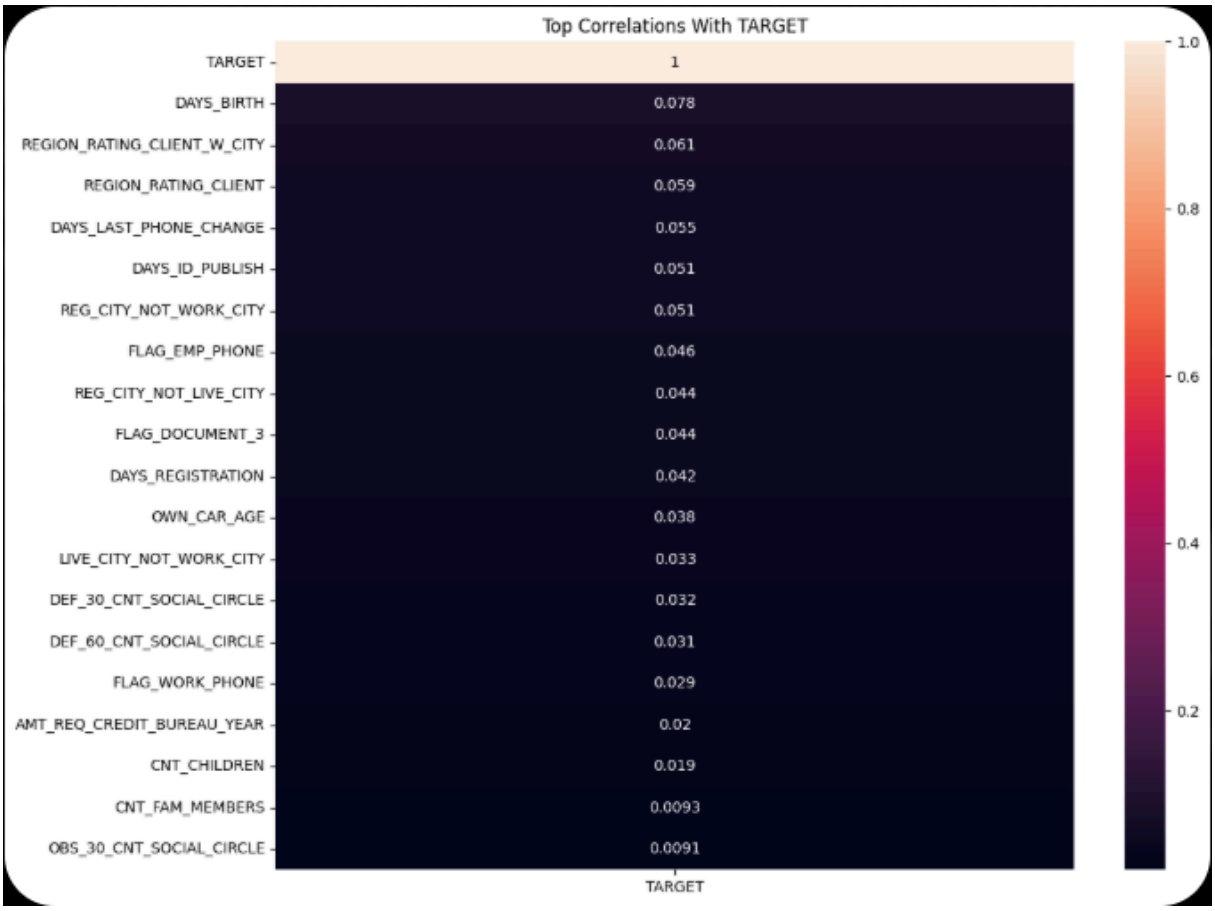
Dataset	Row Count	Column Count	Train/Test
application_train	307,511	122	Train
application_test	48,744	121	Test
bureau	1,716,428	17	N/A, used for feature engineering
bureau_balance	27,299,925	3	N/A, used for feature engineering
credit_card_balance	3,840,312	23	N/A, used for feature engineering
installments_payments	13,605,401	8	N/A, used for feature engineering
previous_application	1,670,214	37	N/A, used for feature engineering
POS_CASH_balance	10,001,358	8	N/A, used for feature engineering

Correlation Analysis with Visualizations

The following correlation matrix demonstrates the correlation analysis between the Target and Aggregated Bureau Features. There was nearly no relationship between the target variable, loan default, and BUREAU_AMOUNT_CREDIT_SUM, BUREAU_AMOUNT_CREDIT_SUM_DEBT, and BUREAU_AMT_CREDIT_SUM_LIMIT. There was very weak correlation between the loan defaul target variable and BUREAU_DAYS_CREDIT variable. Hoever, there were strong correlations between the variables foe Aggregated Bureau Features.



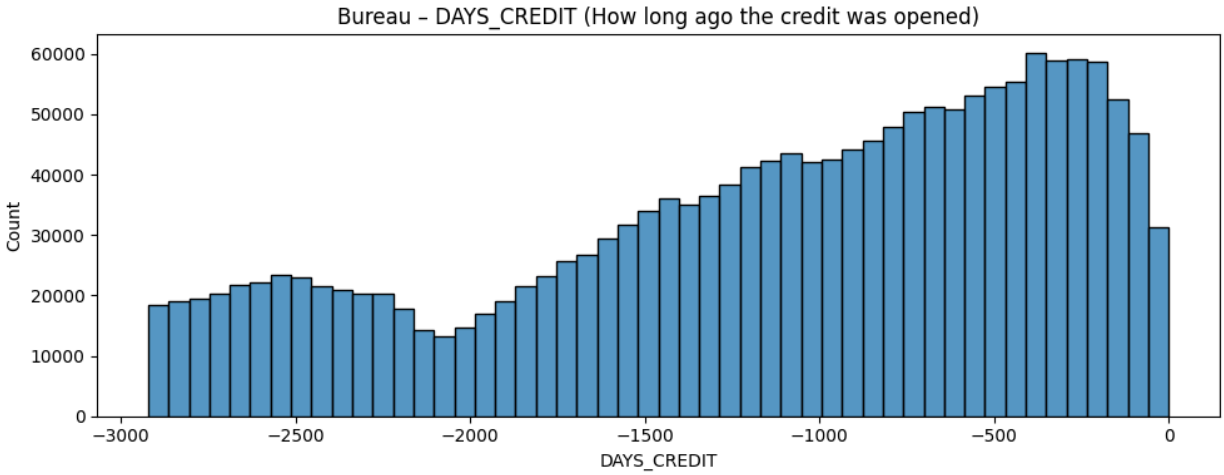
The Target variable showed weak correlation with the variables in the image. The DAYS_BIRTH variable had the highest relationship with Target, but it was still very weak.



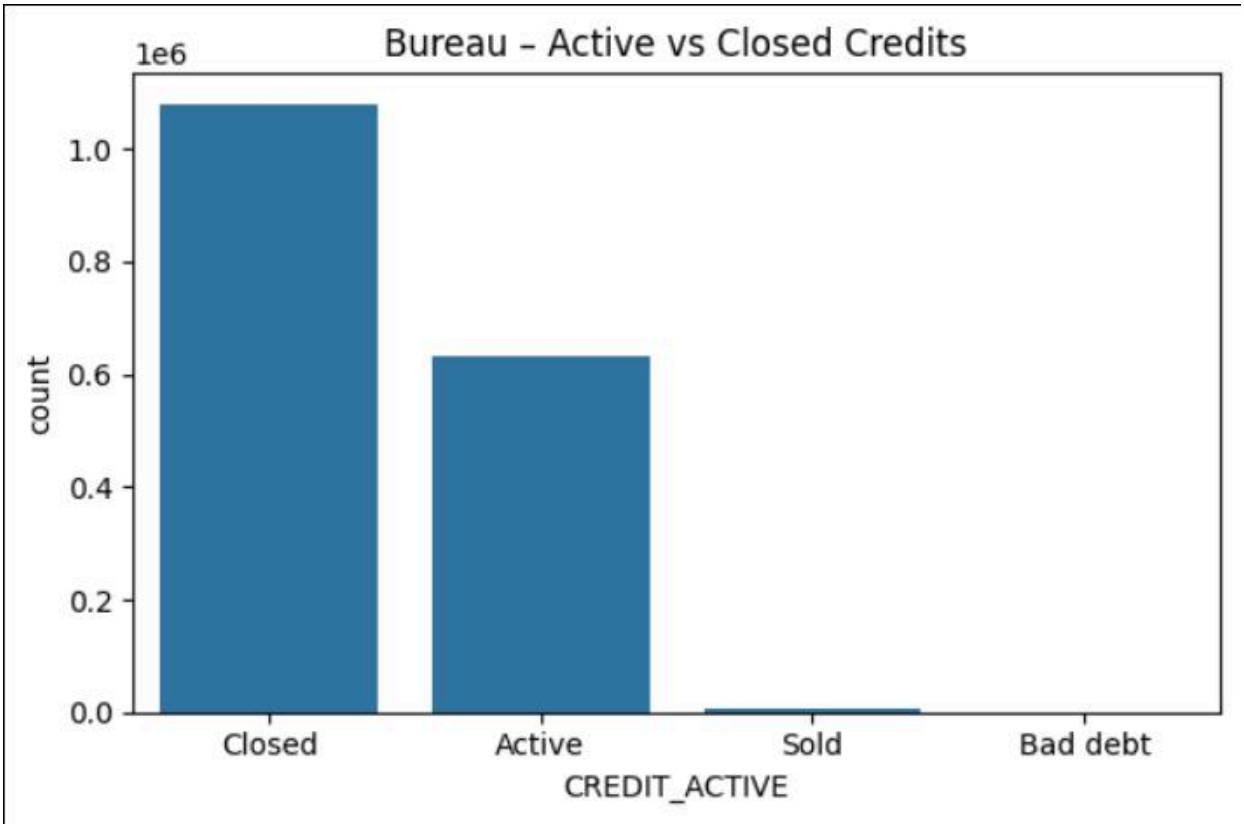
To have better accuracy on predictions created by the pipeline, the team will aim to use variables with higher correlation to the Target variable.

Pair-based Visualization of the Input and Output Features

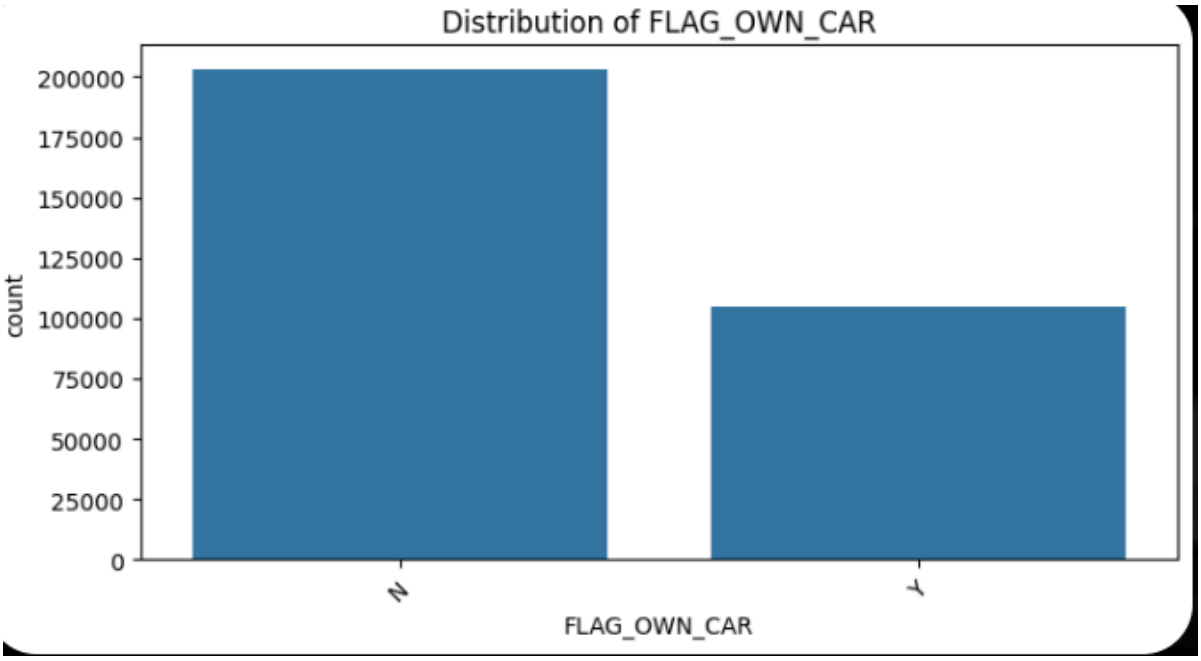
The following visual shows the amount of clients that each DAYS_CREDIT value had.



The Bureau - Active vs Closed Credits chart showed that most clients had more closed credits than active credit.

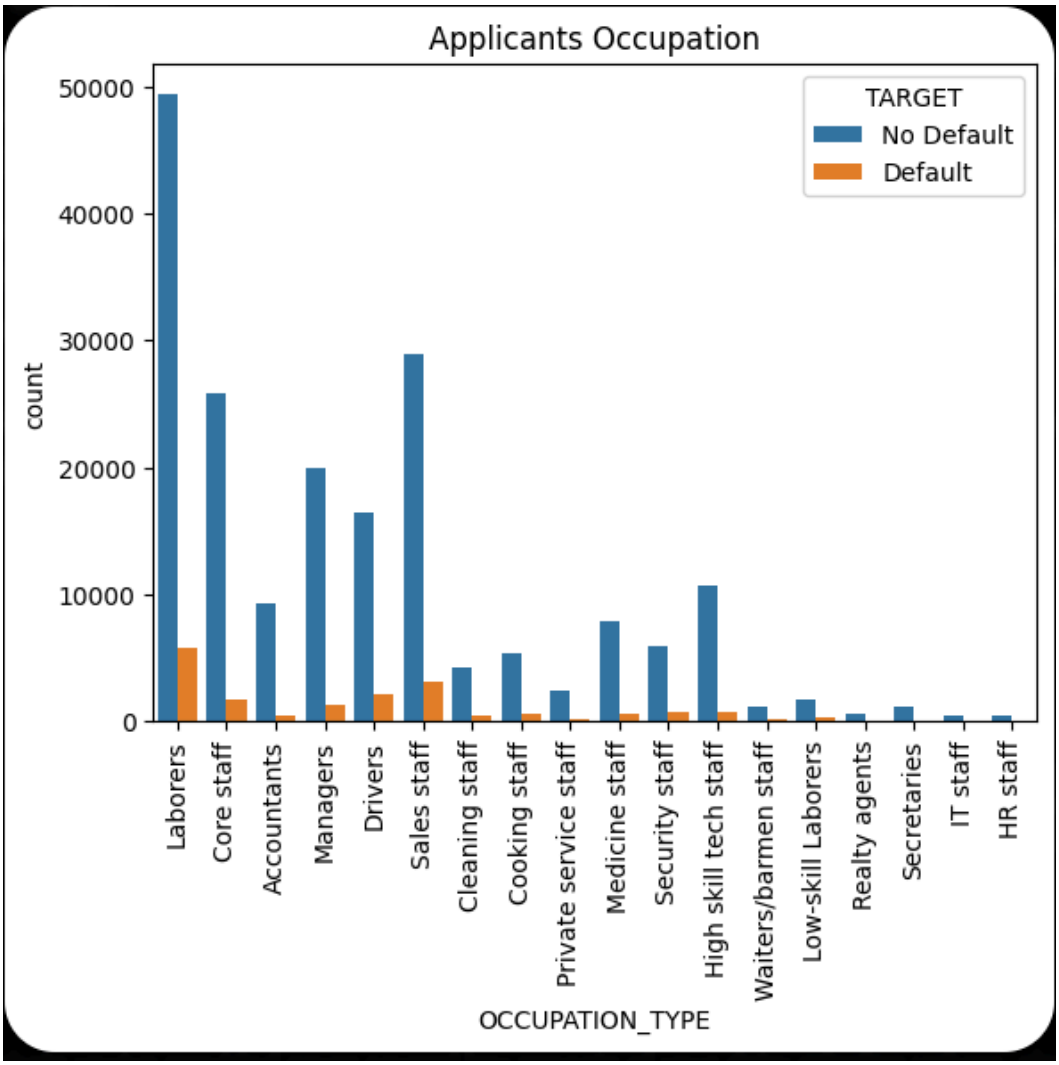


The following visual shows the distribution of clients that own a car. Nearly 200,000 do not own cars and nearly 120,000 own cars.



Visualization of each of the input and target features

The following visual considers the applicants document and the target feature. The occupation with the highest default was Laborers. Whereas, Realty agents, secretaries, IT staff, and HR staff had no defaults. However, Laborers also had the highest no defaults amongst all occupations.



Experiments

After our EDA, our experiments included training and testing our models on the relevant data. Models were graded on various metrics to provide a substantial evaluation, and to provide insights moving into the next phases of engineering and testing. These metrics include:

Metric Name	Description
Confusion Matrix	Derives false positives/negatives in evaluation
F1 Score	Provides a single balanced score for models evaluated on imbalanced data
ROC Score	Measures the model's ability to distinguish between the classes

Results and Discussion

The best pipeline produced during Phase 2 is LightGBM. The model had the highest ROC_AUC, F1 score, and recall. Although precision was not the highest, it had the best overall performance. The following data frame shows the ROC_AUC, F1 score, Precision, and Recall for each pipeline tested.

There was not much variance on the ROC_AUC metric of each model, with all scoring above .70. The Random Forest model had the lowest F1 score, with all other models resulting in similar F1 scores. The precision was low for all models, however the Random Forest outperformed all other models. The Random Forest model continued to perform low in the recall metric, whereas all other models had similar recalls.

Visualization of the Modeling Pipelines

```
Starting model evaluation on X_valid...
-> Logistic Regression evaluated.
-> Random Forest evaluated.
-> XGBoost evaluated.
-> LightGBM evaluated.

========================================================
           MODEL PERFORMANCE COMPARISON
========================================================
```

	Model	ROC_AUC	F1_Score (T=0.5)	Precision (T=0.5)	Recall (T=0.5)
3	LightGBM	0.747238	0.259692	0.160880	0.673123
2	XGBoost	0.737612	0.258813	0.163254	0.624159
0	Logistic Regression	0.735750	0.250522	0.154518	0.661555
1	Random Forest	0.717900	0.011161	0.456522	0.005650

```
========================================================
```

Conclusion

The HCDR Project is important because it allows Home Credit to accurately predict which clients will repay the loan. It looks at several variables to increase accuracy. We hypothesize that machine learning pipelines with feature engineering can accurately predict the repayment of a loan. In this phase, we performed EDA, created basic pipelines, and evaluated results based on key metrics. The results are significant because they establish realistic baselines for model performance and highlight which variables and preprocessing steps are most likely to influence prediction accuracy. These findings help us understand where the current approach succeeds, where it struggles, and what areas will benefit most from deeper feature engineering and tuning. In Phase 3, we plan to build on the pipelines by continuing feature engineering and hyperparameter tuning, including additional feature selection, as well as using ensemble methods. Overall, Phase 2 has provided the foundation for further improvement in predictions.