GROUP-6
LOAN DEFAULT RISK PREDICTION

Shiva Keerthan Jalla, Aravindh Gowtham Bommisetty,
Krishna Mallik Nanduri, Koushik Billakanti

INTRODUCTION

- Due to weak or non-existent credit records, many people have difficulty obtaining loans.
- A crucial task for every lending organization in the banking sector is to evaluate whether an applicant has the capability to repay the loan or if the applicant is a potential loan defaulter.
- There are many underlying features that come to play if credit should be approved for the given client. The Data is provided by Home Credit, a service that provides lines of credit to the unbanked group of population.

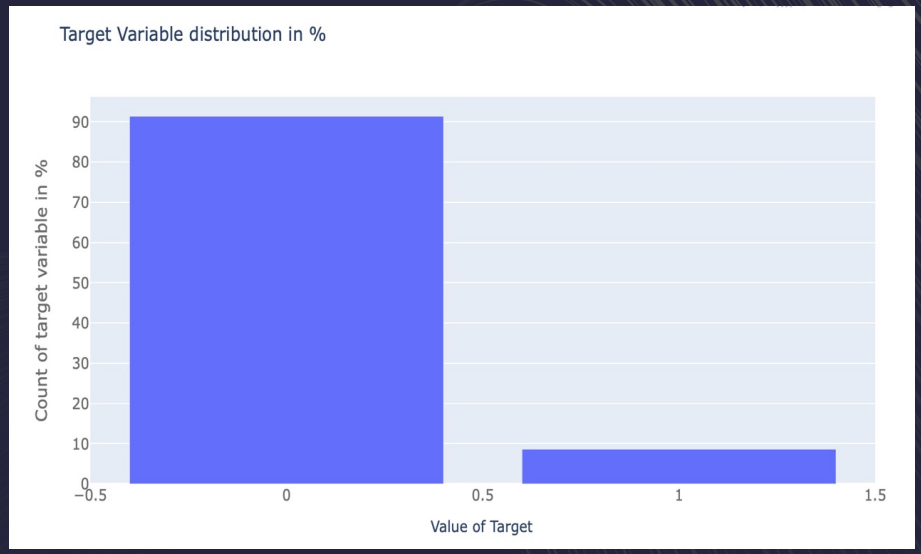
Our Goals

- To predict whether the applicant can repay the loan or not.
- Visualize the different patterns of the data.
- Reduce the rejections to the applicants who repays loan on time by predicting how probable the applicant can payback the loans on-time.
- Examine the data and build a model that can predict a person's ability to repay a loan and to ensure loans are given with terms that can be met. The success of the models allows for Home Credit to avoid losses and improves their potential for profits.

Info about our dataset

- The dataset we choose consists of 7 files.
- Source:https://www.kaggle.com/c/home-credit-default-risk/overview

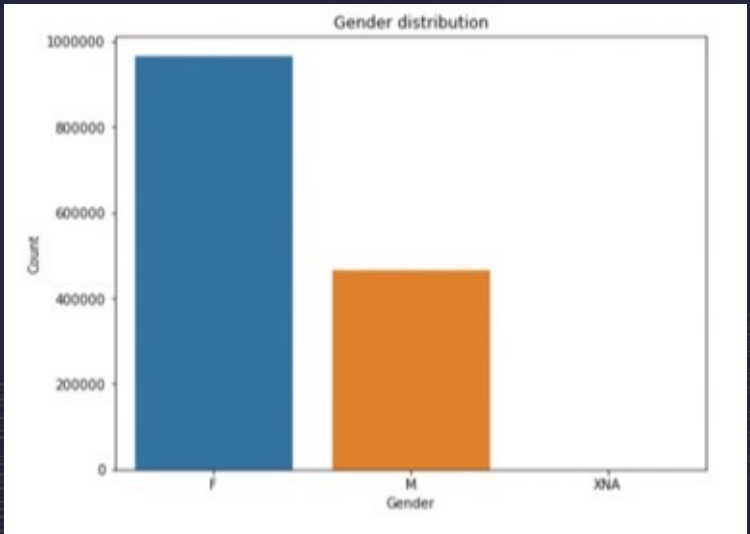
EXPLORATORY DATA ANALYSIS



Target Label distribution, where we have 0's around 90 percentage and 1's have 10 percentage of data.

0 - Paid back the loan

1 - Loan Defaulter

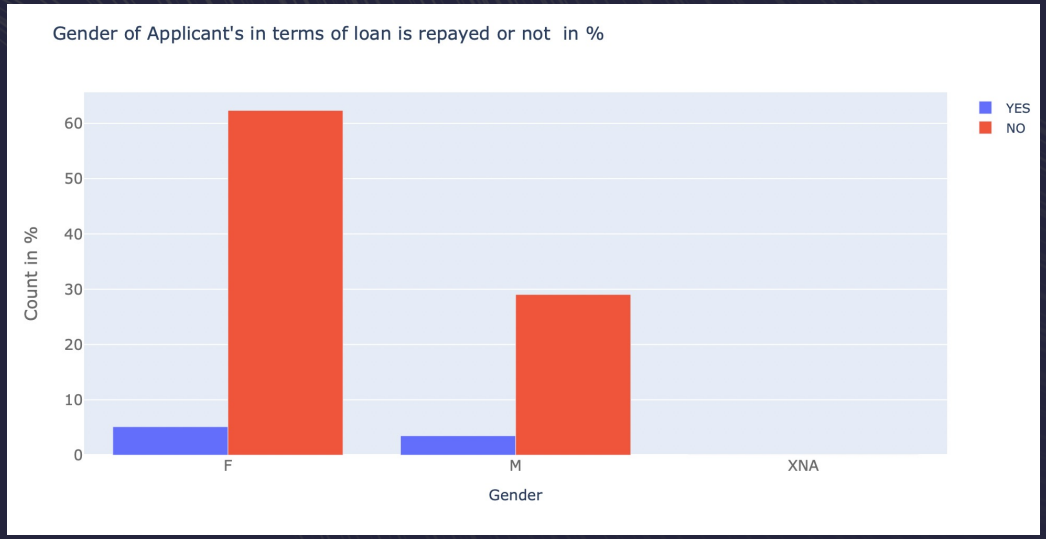


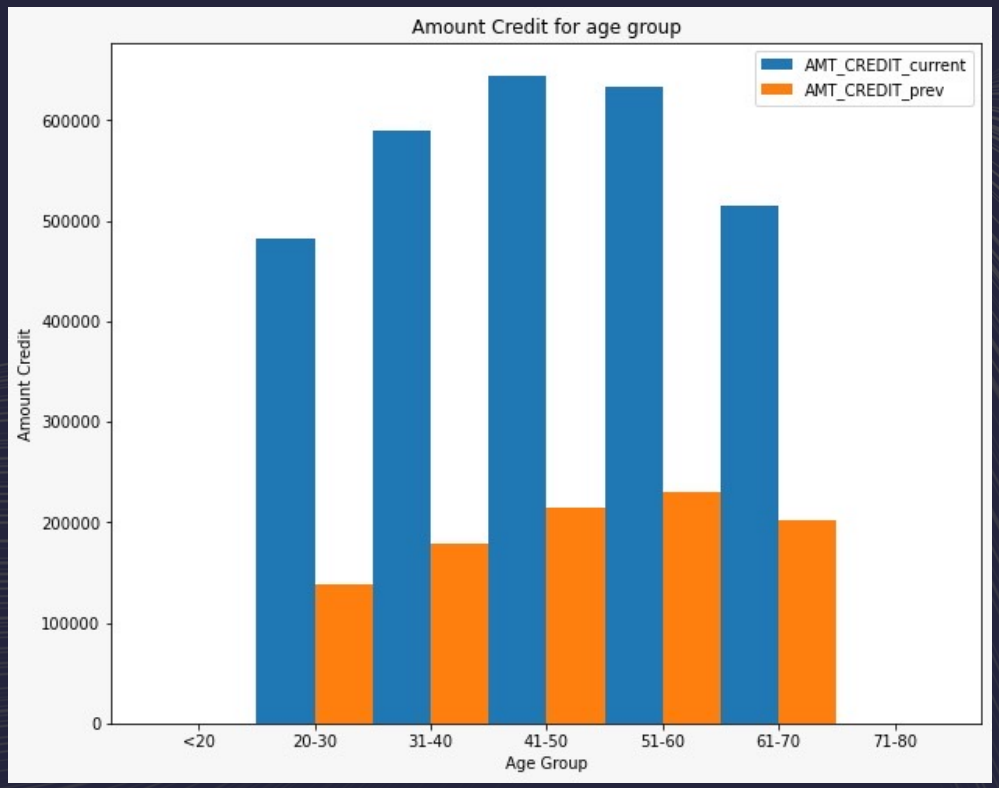
Which gender group usually obtain loans from vendors?

Female > Male > XNA(Other)

Borrowers who repay the loan grouped by gender.

Female – Less than 10% pay back,
Male – Less than 5% pay back.

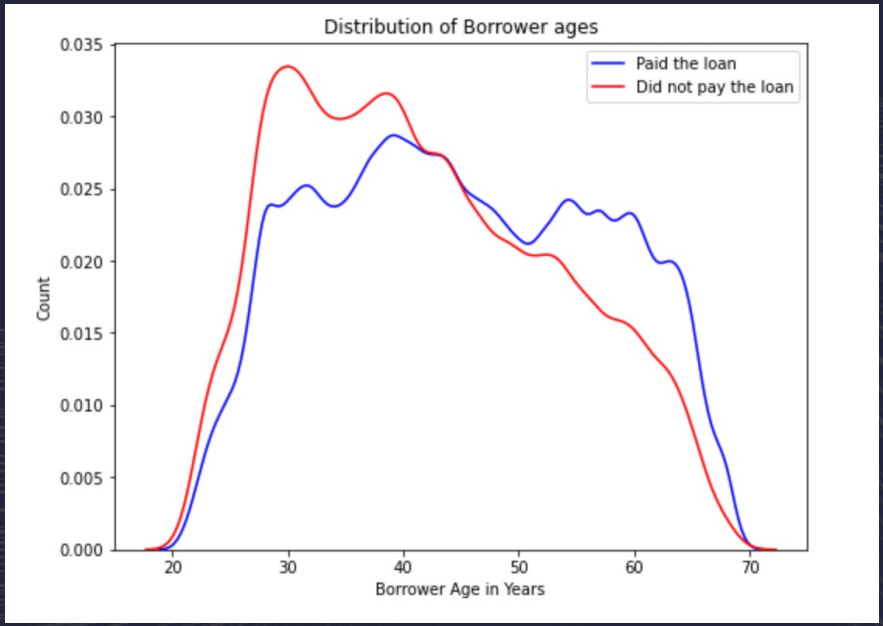




Loan amount obtained by various age groups in the past and present.

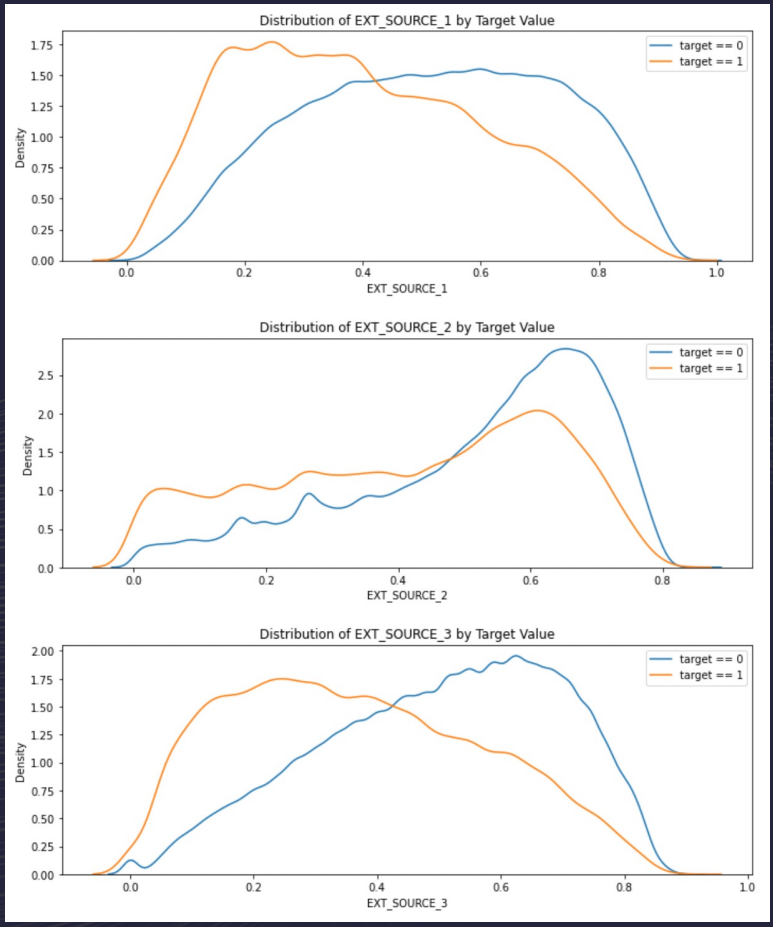
Age group 51-60 – Highest previously
Age group 41-50 – Present highest

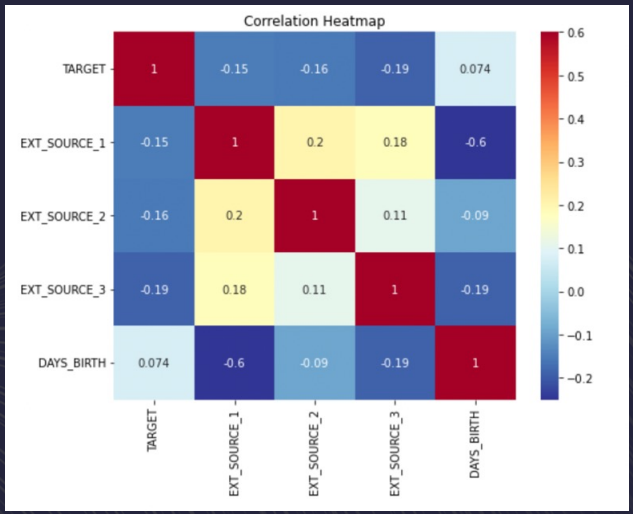
In general, the credit amount in the present period is 3x of the past credit amount sanctioned.



Visualizing the distribution of borrower ages with the target variable constraint, to check the trend between age and loan payment.

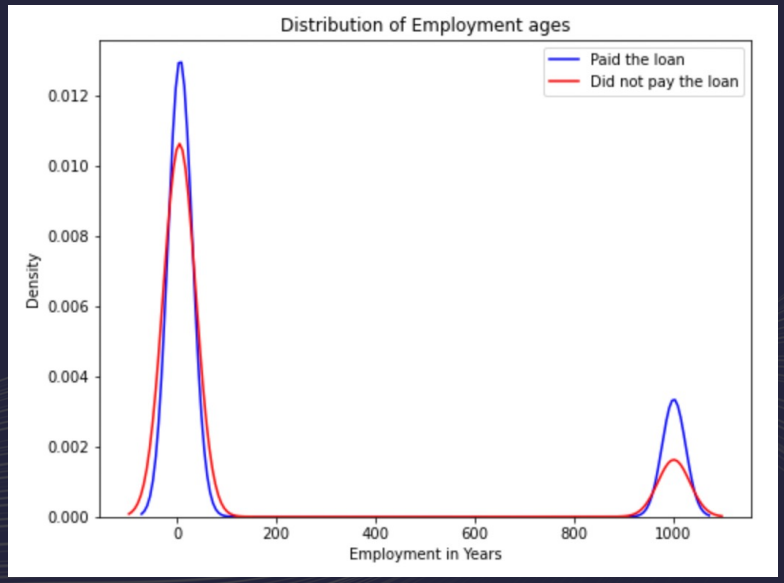
It shows that the people who took the credit in their 20’s are highly probable to be loan defaulters, and probability to be loan defaulters decreases with age and increases the probability to pay their loan.



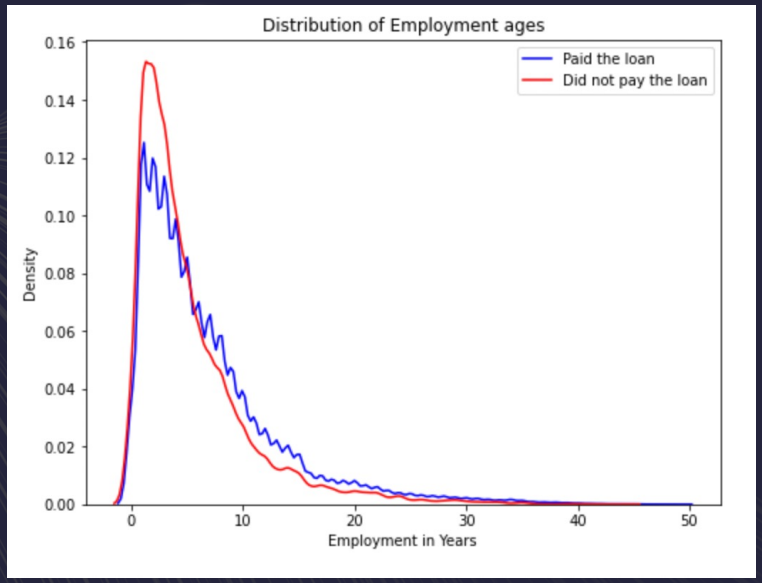


All 3 EXT_SOURCE variables display negative correlation with the target variable, as the important target is 0, indicating that as the value of the EXT_SOURCE increases, the client is more likely to repay the loan.

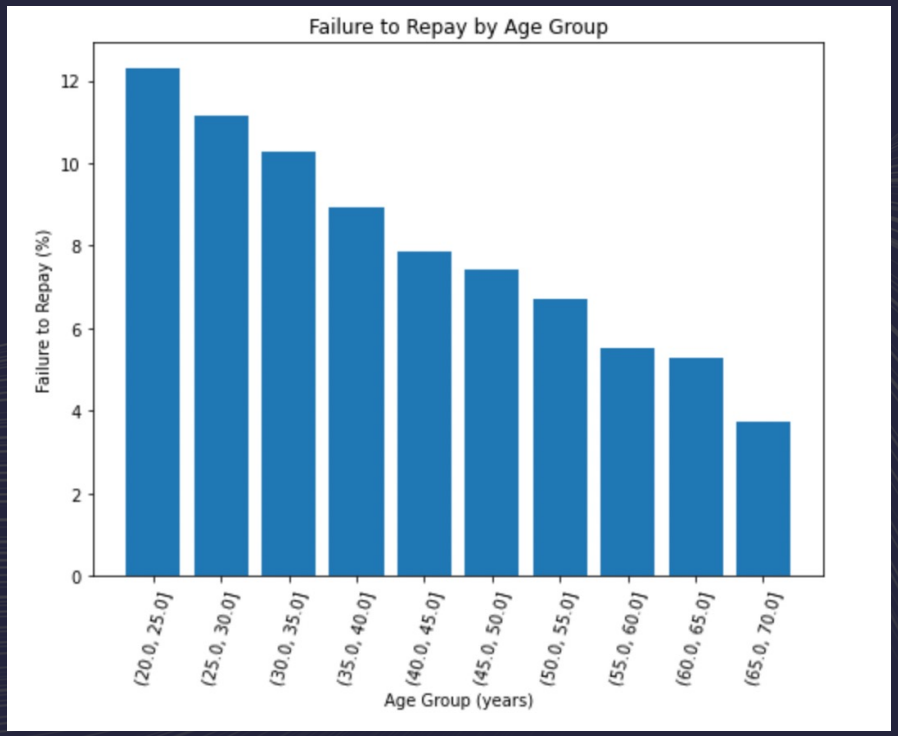
EXT_SOURCE_3 displays the greatest difference between the values of the target. We can clearly see that this feature has some relationship to the likelihood of an applicant to repay a loan. We will explore this relationship in model building.



Distribution plot of borrowers and their employment ages.
Here, we observed few anomalies and outliers.

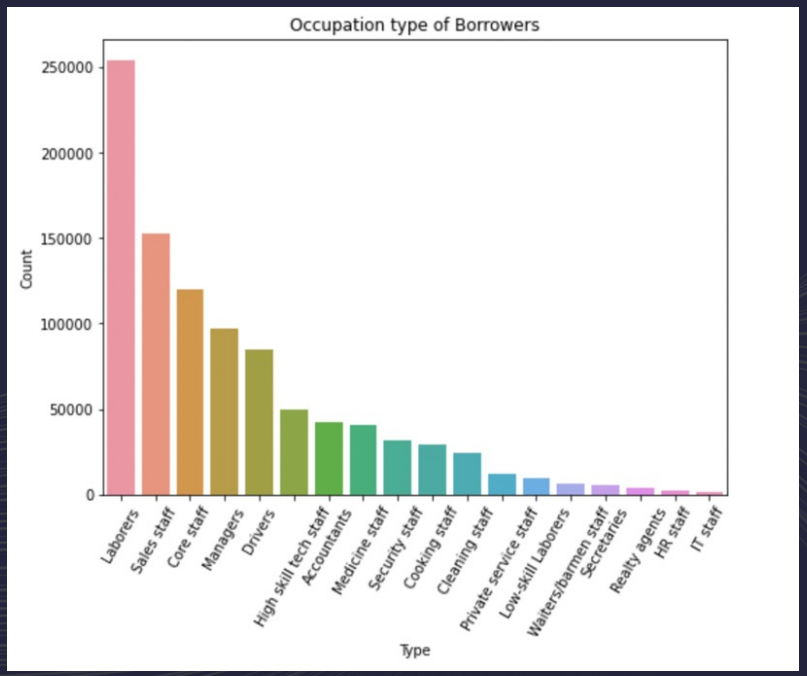


This is plotted after we delt with the anomalies.
In the early years of employment, the number of loan defaulters is higher.

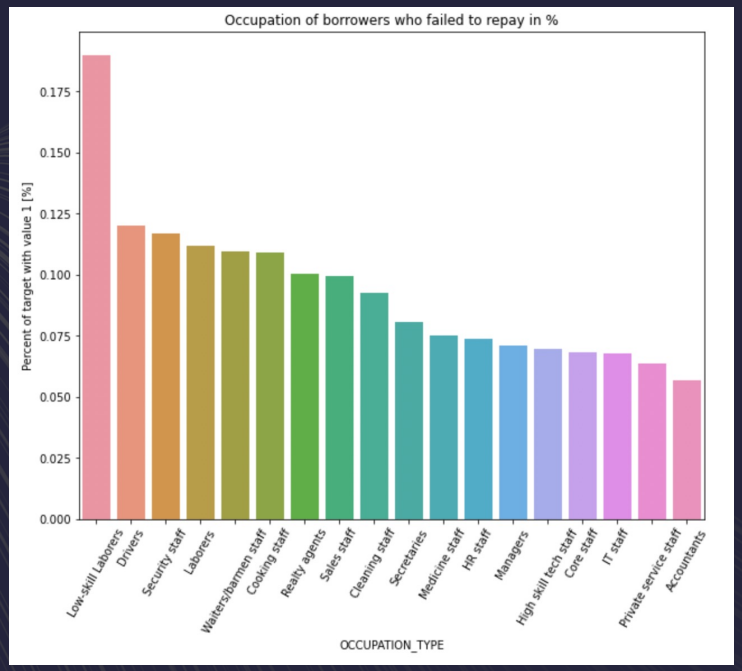


Checking the connection between age group and failure to repay the loan.

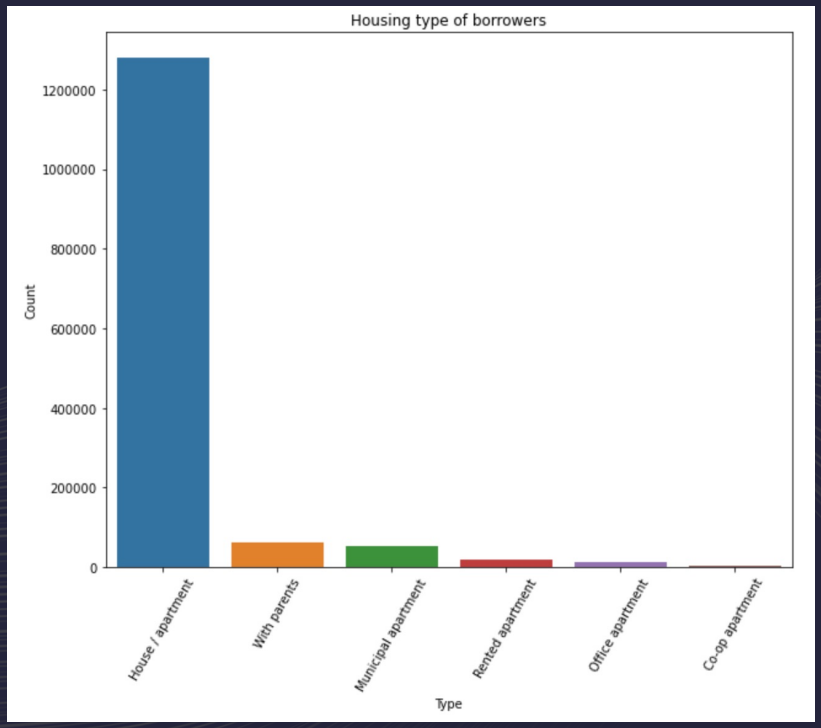
We can see that, as the age increases the percentage of people who are failing to repay the loan are decreasing.



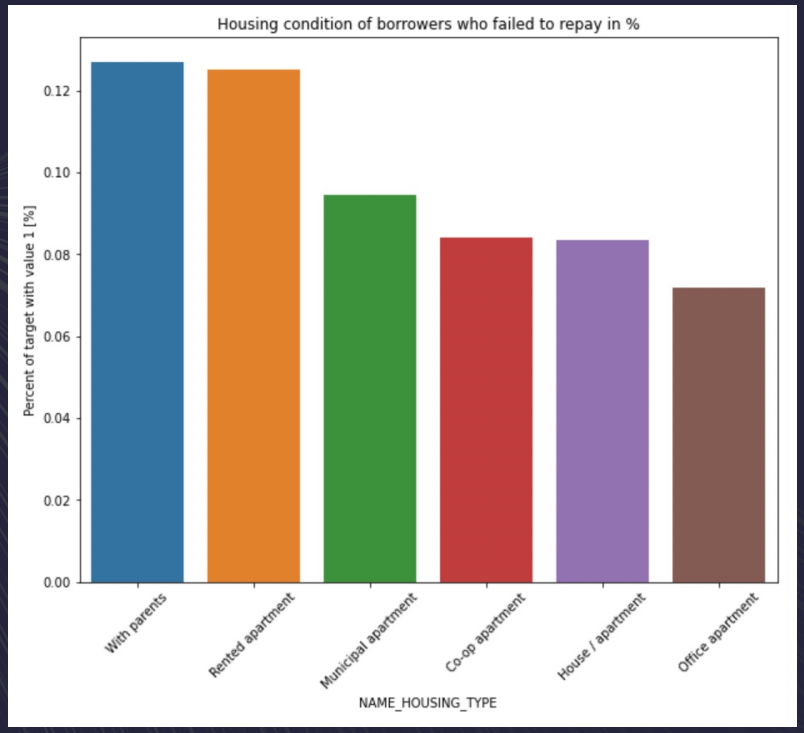
Occupation Type of borrowers.



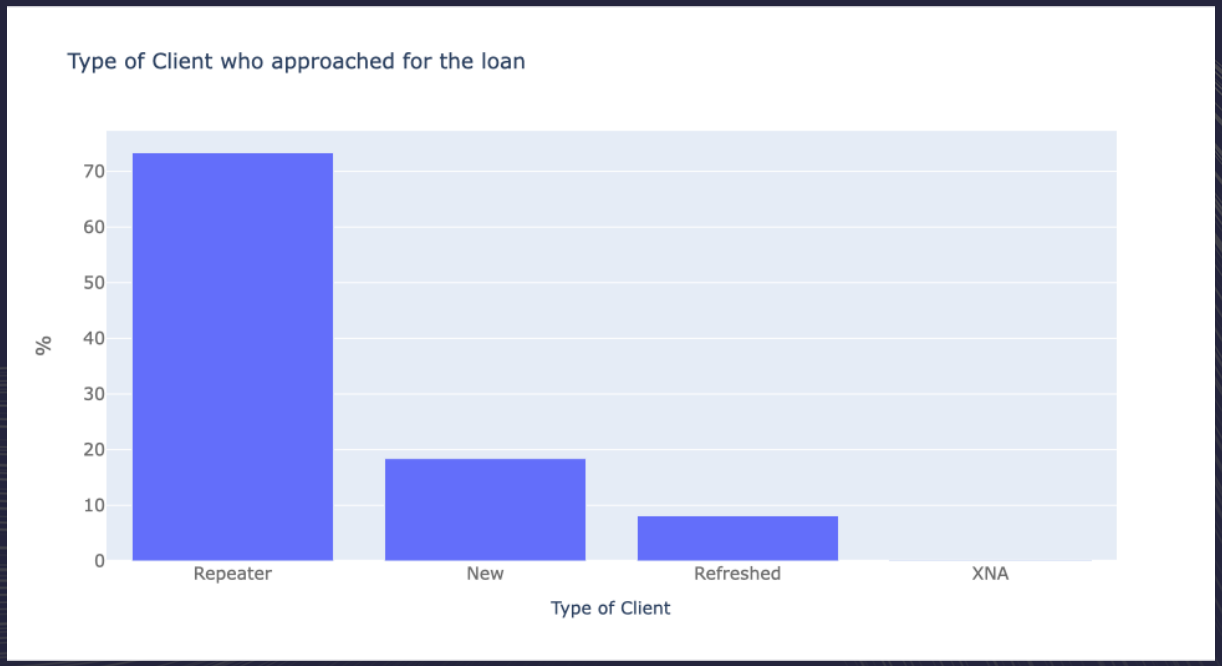
Plot of loan defaulters according to their occupation type.
Low Skill laborers are 6th lowest to borrow, yet highest to default the loan.
Also, the least percent of defaulters are working at highly skilled jobs such as IT, Accountants, etc.



Housing Type of borrowers.



Plot of loan defaulters according to their housing type.
Highest percent of defaulters live with parents.



Plot of clients who approached for the loan.
We can see that the highest percentage of clients are repeaters.

Feature Engineering

Domain Features

- 3 Flag features
- 27 Numerical Features
- Credit to annuity ratio
- Credit to goods price ratio
- Annuity to income ratio
- Interest Paid

Group by Features

For better Merging of Datasets
For every Dataset:

- Create numeric features by grouping on SK_ID_CURR and finding group means.
- One-hot encode the categorical variables.
- Create categorical features by grouping on SK_ID_CURR and taking the group means.

Merging

- Merging bureau_balance into bureau using SK_ID_BUREAU key, then merging bureau into train/test using SK_ID_CURR key.
- All other tables will be merged directly into train/test using SK_ID_CURR key
- Created group features in all the datasets for dealing with multiple records of a single person issue.

After Preprocessing and Merging

10 Files		310,000 Records
288 Numerical	→	398 Numeric
51 Categorical		15 Categorical

Train & Test set Preparation

Train/Test – 75/25

#Folds - 10

Pipeline for Numerical features

- Imputer – Median
- Scaler – Standard Scaler

Pipeline for Categorical features

- Imputer – Constant('Unknown')
- One-Hot Encoder

Train and validation sets are processed through both the pipelines before used for modeling.

MODELING TECHNIQUES

Logistic Regression

Random Forest

Decision Tree

XG Boost

Light GBM

Extra Trees

Modeling

Logistic Regression

Tried both lbfgs and saga solvers and got faster performance (and slightly higher scoring) out of lbfgs.

I experimented with max_iter as low as 200, but did not see any degradation of accuracy below 400.

AUC Train – 0.763948
AUC Test – 0.768523
Hyperparameters-
{'C':0.005, 'penalty':'l2'}

Decision Tree

The best parameters that I could find were max_depth = 8 and min_samples_leaf = 10.

This model is not going to be our best model, but it becomes the benchmark for Random forest and Extra Trees Models.

AUC Train – 0.722701
AUC Test – 0.730909
Hyperparameters-
{'max_depth':8, 'min_samples_leaf':10}

Random Forest

We used Optuna to tune the hyperparameters (max_depth and min_samples_leaf).

39 and 37 as final hyperparameters, but the model is not very sensitive over a wide range of choices.

AUC Train – 0.758414
AUC Test – 0.764844
Hyperparameters-
{'max_depth':39, 'min_samples_leaf':37}

Modeling

Extra Trees

Difference is that Extra Trees model uses the whole original sample, while Random Forest takes subsamples with replacement.

Our Extra Trees model produced about the same performance as the Random Forest

AUC Train – 0.753552
AUC Test – 0.760236
Hyperparameters-
{‘min_samples_split’:8,
‘min_samples_leaf’:18,
‘n_estimators’:300}

XG Boost

XGBoost runs quickly on GPU. Produced better accuracy than logistic regression. It is also easier to tune than Light GBM.

Although it can be susceptible to overfitting if you set the max depth or learning rate parameters too high.

AUC Train – 0.773414
AUC Test – 0.780347
Hyperparameters-
{‘gamma’:0, ‘learning_rate’:0.05,
‘n_estimators’:300,‘max_depth’:6}

Light GBM

LGBM runs fast and produces our best model. Performance starts to decline if the number of estimators is below 1400.

Optimized LGBM is the strongest model. Adjusting the alpha and gamma helped to boost the model's accuracy.

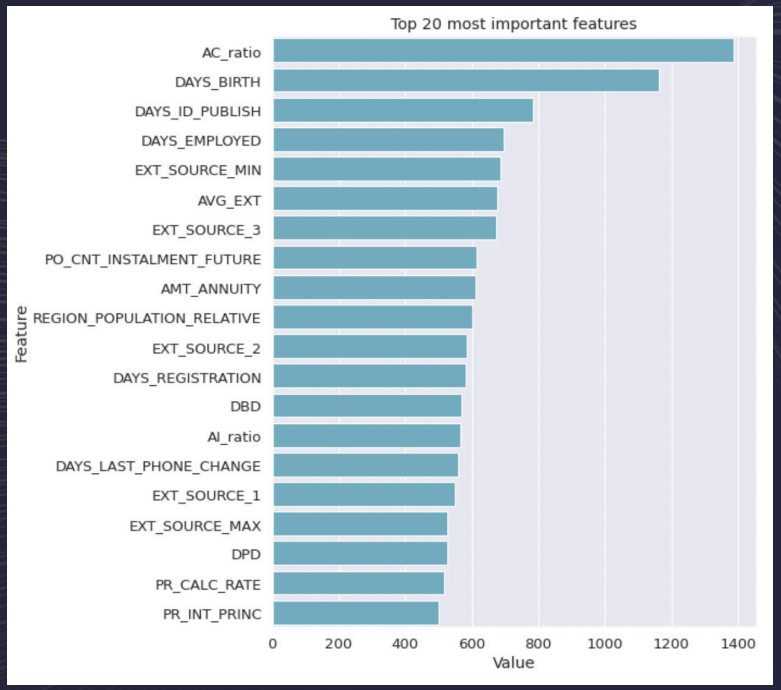
AUC Train – 0.773414
AUC Test – 0.780347
Hyperparameters-
{‘learning_rate’:0.02}

Model Evaluation

As the dataset is imbalanced, AUC score is used as metric

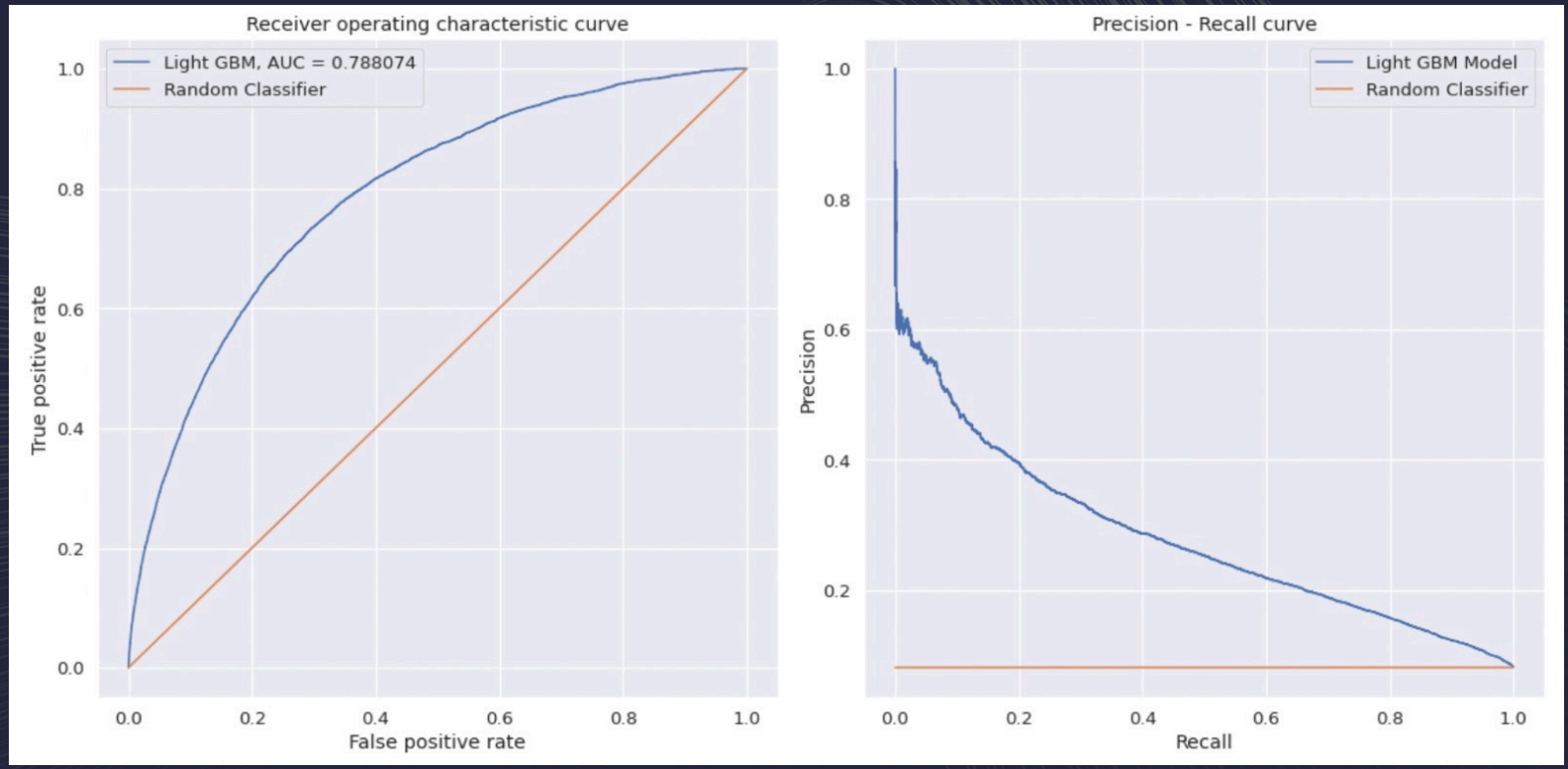
Model Name	AUC- Train	AUC- Test
Logistic Regression	0.763948	0.768523
Decision Tree	0.722701	0.730909
Random Forest	0.758414	0.764844
Extra Tress	0.753552	0.760236
XG Boost	0.773414	0.780347
Light GBM	0.779771	0.788100

Feature Importance’s



- AC ratio is the debt percentage of a client, it is calculated by the total annuity divided by total credit amount.
- Calculating this AC Ratio determines the debt-to-income ratio.

Results


Receiver operating characteristic curve
Light GBM, AUC = 0.788074
Random Classifier
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Precision - Recall curve
Light GBM Model
Random Classifier
Precision
Recall


thank you!