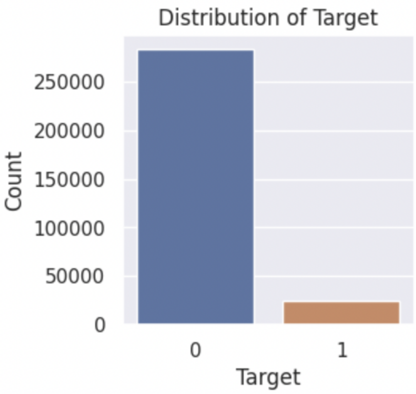
Home Credit Default Risk Group 27

Problem:

Taking loans has been a crucial part of many of our lives. Whether it's for education, or buying a new house or supporting a business, people prefer taking a loan over spending liquid cash. However, sometimes people have a disadvantage in obtaining loans due to nonexistent or low credit history. Home Credit is a financial provider that is trying to make borrowing a positive experience and ensure that clients capable of repayment are not rejected. In this project, we try to utilize alternative data such as transactional history in order to predict the client's repaying ability and thus allow people with repayment ability to obtain the loan they need.

Exploratory Data Analysis and Data Preprocessing:

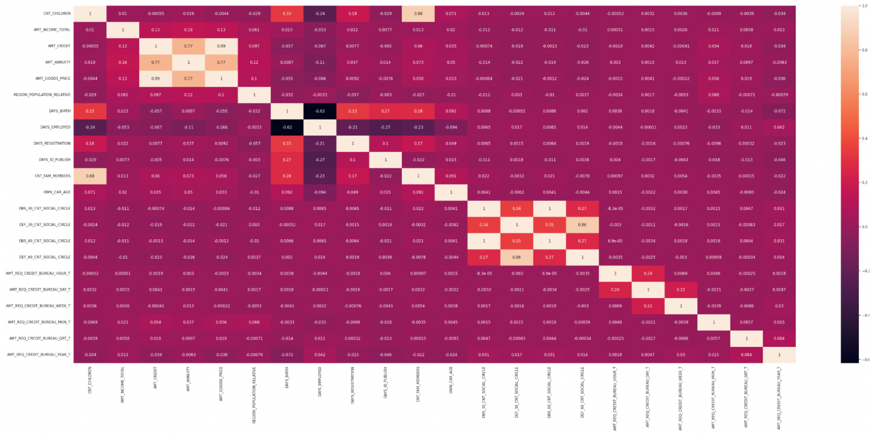
1. Imputed missing values with -1 and median.
2. Dropped columns with more than 35% missing data.
3. Analyzed data distributions and relationships between features and the target variable.
4. Removed multicollinearity by removing highly correlated data features (more than 85%).
5. Scaled numerical features using StandardScaler.
6. One Hot Encoding on Categorical features and Target Encoding on Occupation Type and Organization Type.



Observations from data preprocessing:

1. Highly imbalanced dataset.
2. Highly correlated pairs:

- *AMT_CREDIT* and *AMT_GOODS_PRICE*: Credit is likely to be equivalent or close to the price the house is valued at.
- *CNT_CHILDREN* and *CNT_FAM_MEMBERS*: Family members are proportional to the number of children.
- *DEF_30_CNT_SOCIAL* and *DEF_60_CNT_SOCIAL*: A person with a payment due for 60 days implicitly falls into the category of payment due for more than 30 days also.



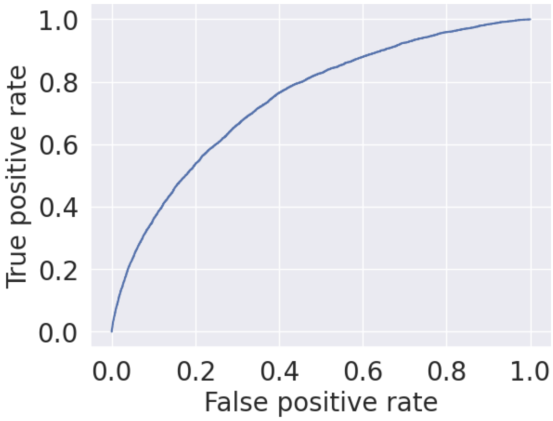
Splitting Methodology and Sampling:

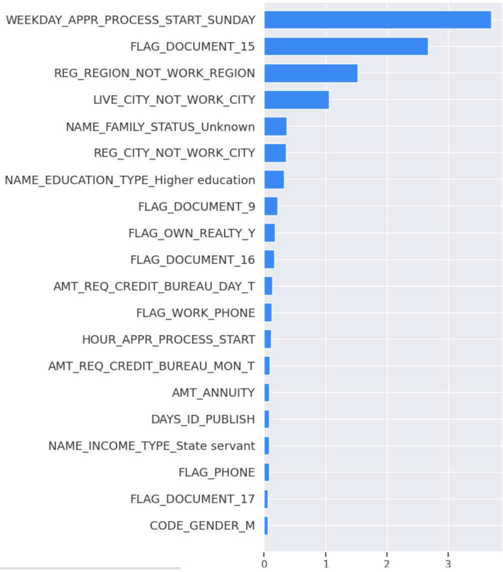
1. Used stratified splitting to handle class imbalance.
2. Split the training set such that 20% of it is allocated for testing.

Models:

1. Logistic Regression:

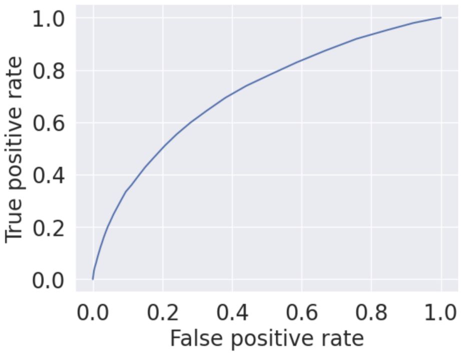
- We start by using Logistic Regression.
- UsingGridSearch for hyperparameter tuning, we got the best model with a 68% accuracy score.
- Using SHAP we got the most important features: WEEKDAY_APPR_PROCESS_START_SUNDAY, FLAG_DOCUMENT_15, REG_REGION_NOT_WORK_REGION and LIVE_CITY_NOT_WORK_CITY.





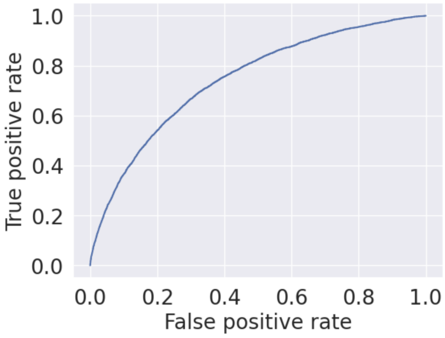
2. Random Forest Classifier:

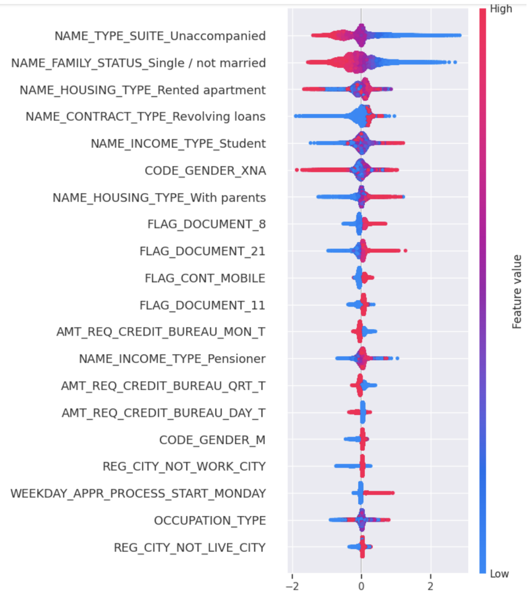
- Next we try out an ensemble technique.
- Due to time limitations, we run it with default parameters and get a precision score of 85%.



3. XGBoost:

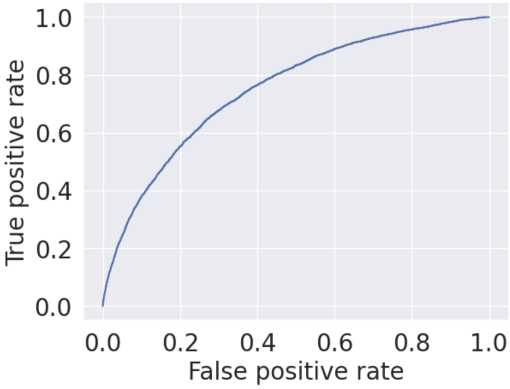
- Since the training set is close to 307K, we thought that boosting might improve the predictions.
- GridSearch resulted in a precision score of 50%.
- A similar SHAP analysis was performed for XGBoost.

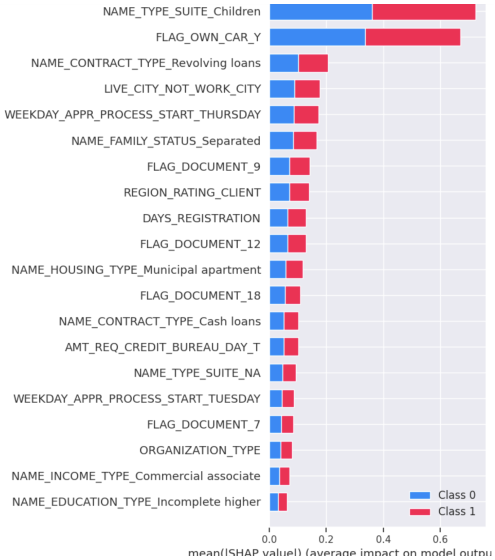




4. LightGBM:

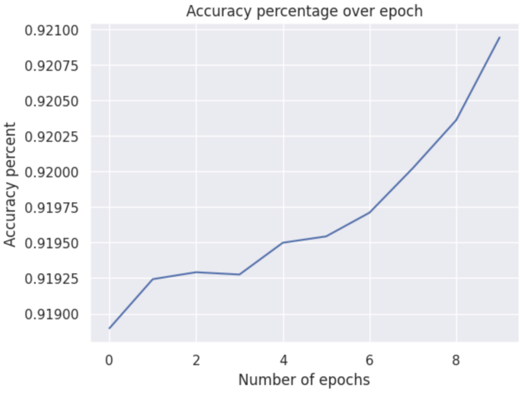
- For a faster execution, we tried using LightGBM.
- GridSearch resulted in a precision score of 56.5%.

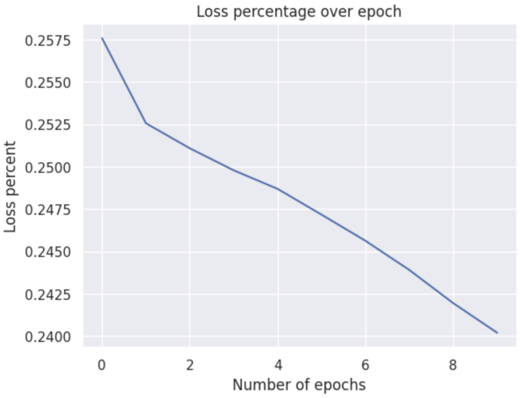




5. Neural Network:

- To capture patterns and correlations that other classifiers might have missed, we tried Neural Network.
- On training for 10 epochs, a test loss of 25.6% and a test accuracy of 91.8% was achieved.





Observations:

Of these ML models, Neural Networks performed the best with the highest precision. Using SHAP we see from the different models that factors such as Marital status, Owning a car, Type of housing, Residential city not the same as work city impact the model performance the most. Therefore, we can conclude that our model performance is good enough.

Conclusions:

Neural Network was the best model with the highest accuracy of 91.8%. It is capable of learning complex data patterns which is good when the input data is dynamic or can change over time, such as ours. We believe that the ML models could have a better performance. With more time available, we could use better tuning methods like Random Search or Bayesian Optimization. We could also explore some feature engineering techniques such as SMOTE.

Shruti Agarwal (sa4136)
Parth Batra (pb2882)
Bora Elci (be2246)
Mark Wu (rw2921)