HomeCredit ScoreCard Model

Ridwan Akmal
Home Credit Indonesia Data Scientist Virtual Internship Program Rakamin Academy

repo:
https://github.com/RidwendDev/HomeCredit_VIX_RidwanAkmal/blob/main/HomeCredit_Scorecard Model_Ridwan.ipynb

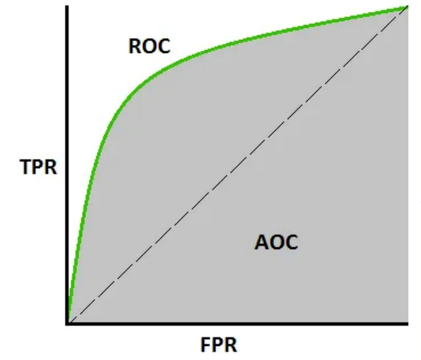
Problem Research

Home Credit Indonesia as a company provides financing services in shopping. This multipurpose financing makes it very easy to shop from online or offline. Here Home Credit seeks to classify clients who have difficulty paying loans and those who have no problems. In this case Home Credit uses a variety of alternative data including telecommunication and transactional information.

Main Goals

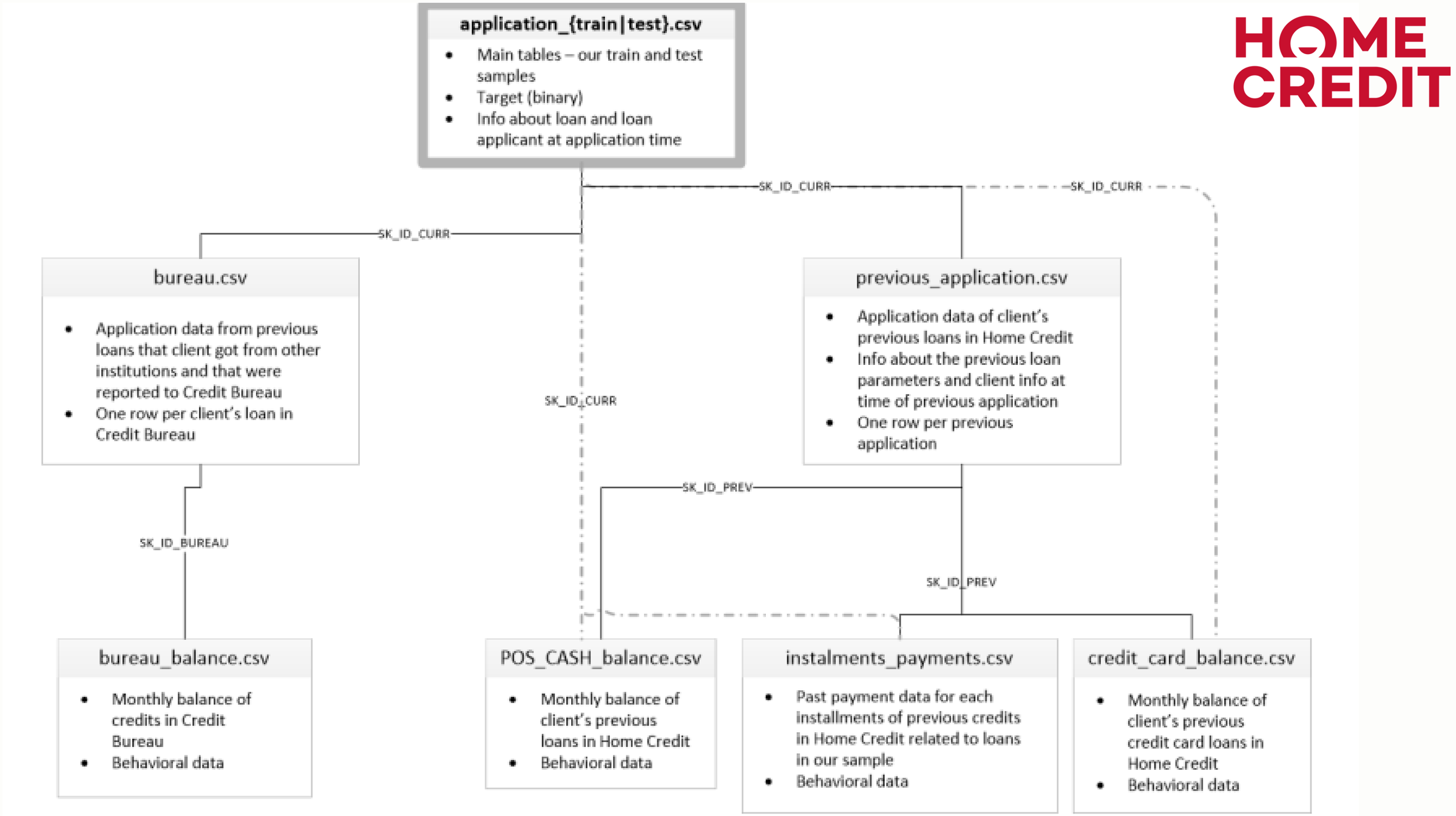
Here Home Credit want to minimize the numer of client who have a problem with repay the loan

Model Evaluation

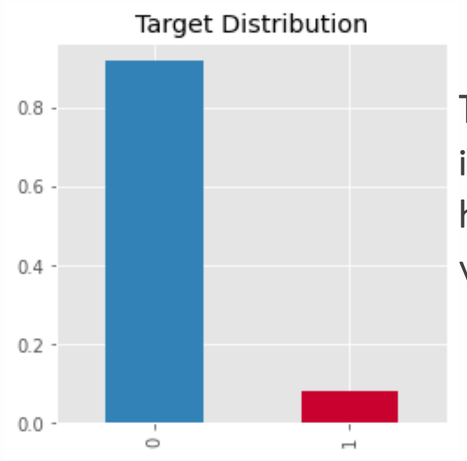


Area under ROC curve

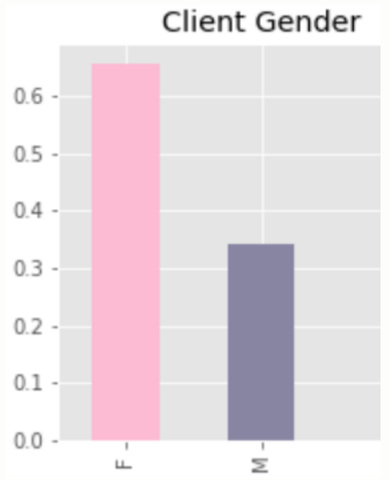
Dataset & Description


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
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_CURR
SK_ID_PREV
SK_ID_BUREAU
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

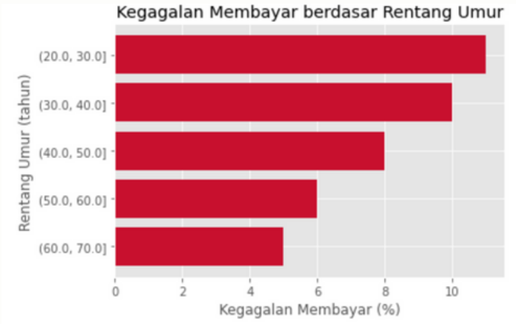
Exploratory Data Analysis

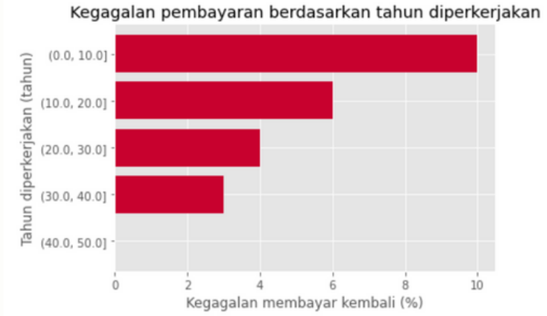


The visualization shows that there is an imbalance in the data where more clients have no problems paying off loans, then will do sampling before modeling



Here we also get insight where clients are mostly dominated by women and clients who dont have children





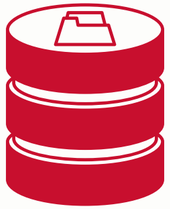
From here we can see that a group in a range youngest old & group range shortest year employed have a problem which is failure to repay the loan

Most of the clients owned realty but dont have a car

HOME CREDIT

Data Preprocessing

RAW DATA



Data Cleaning

1. Replace XNA values with NaN
2. Detect & Handling Missing values
3. Detect redundant data

Feature Engineering

1. Convert day birth to years and get client age
2. Calculating number of documents, IAP, & EITC
3. Scalling the numerical features
4. Encoding the categorical features
5. Drop features with VIF>10

Model Building

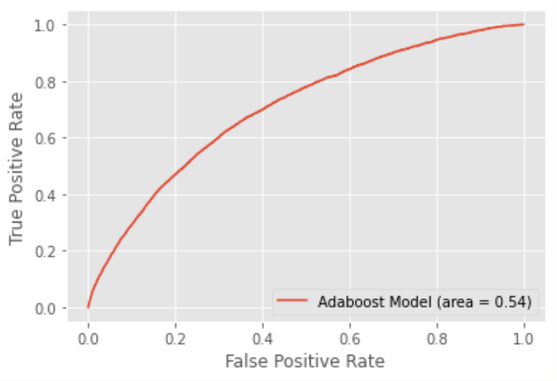
1. Oversampling smote with ratio 2:1
2. Build model with various algorithm like Logistic Regression, Adaboost, & KNN
3. Model evaluation, compare all model which one is the best then all

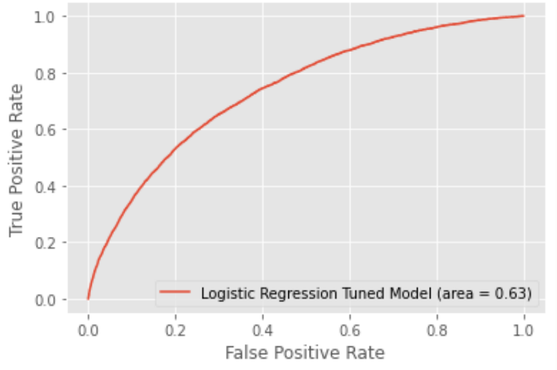
Model Comparasion

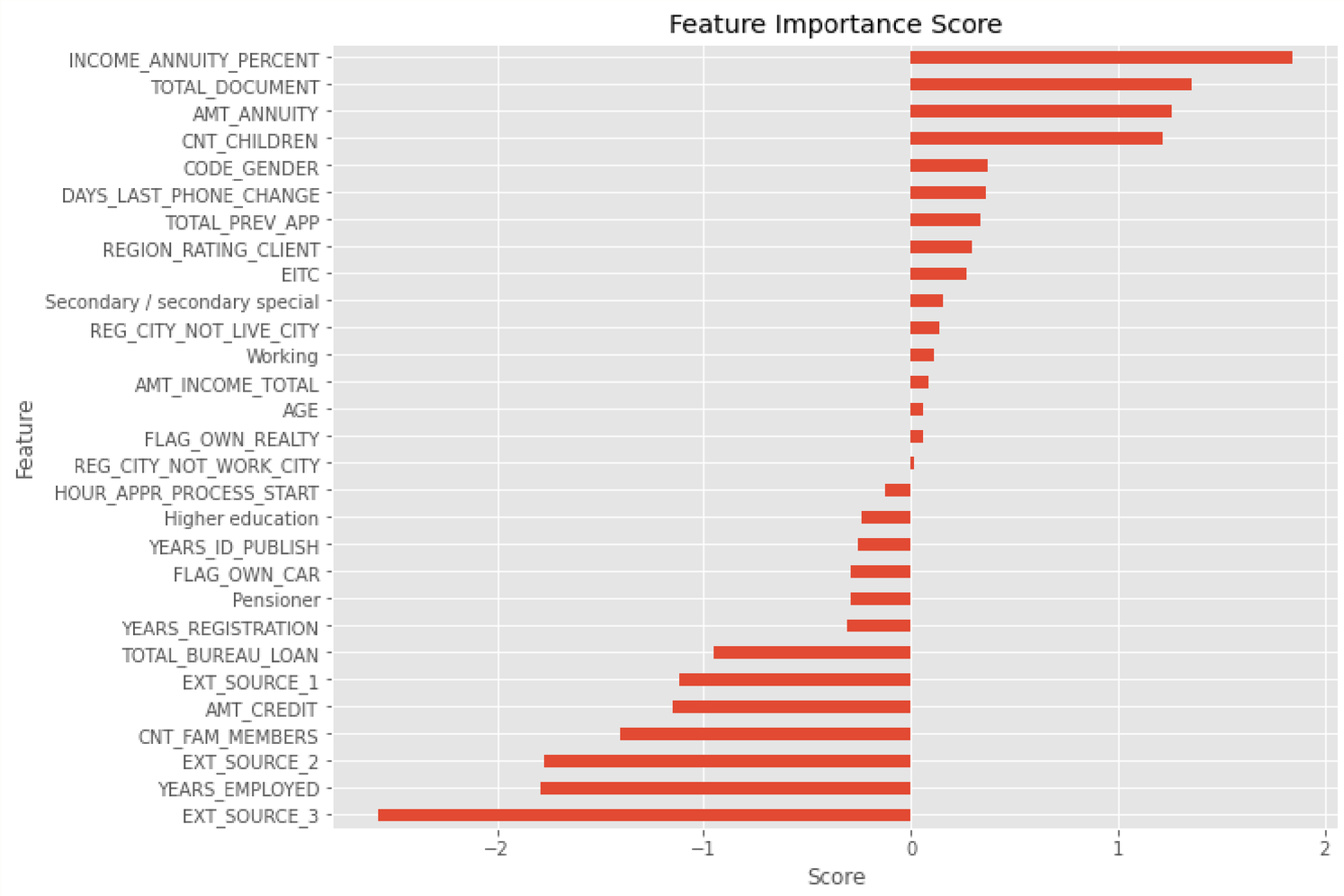
	Training acc score	Testing acc score	Area under ROC
Logistic Regression	0.715	0.843	0.63
XGBOOST	0.936	0.919	0.5
Adaptive Boosting	0.830	0.907	0.54

Best Algo

As previously mentioned we make a area under ROC as a model eval then with the highest score it can be ascertained that the winner is logistics regression, but i feel there is an overfit model







We can see here that five most important features is INCOME_ANNUITY_PERCENT, TOTAL_DOCUMENT, AMT_ANNUITY, CNT_CHILDREN, CODE_GENDER

Business Recomendation

1. From the visualizations we know that the feature NAME_INCOME_TYPE with values stuedent and Business Man have no problems to repay the loan(100% application approved) and next one is feature OCCUPATION_TYPE with values Managers, High skill tech staff, and Accountants so we can create campaign for Student, Business Man, Managers High skill tech staff, and Accountants so that more can be interested & apply for loan.
2. Next, we can see that clients with maternity leave and unemployed status have a difficult tendency to repay loans, where both reach a percentage of more than 30%, for that we can recommend other types of loan contracts that are suitable for both clients like that.

backenddeviwan@gmail.com

Get in touch and let's have a chat!

https://github.com/RidwendDev

Ridwan Akmal