HOME CREDIT INDONESIA

Score Card Model

Muhammad Hanafi

Problem Research

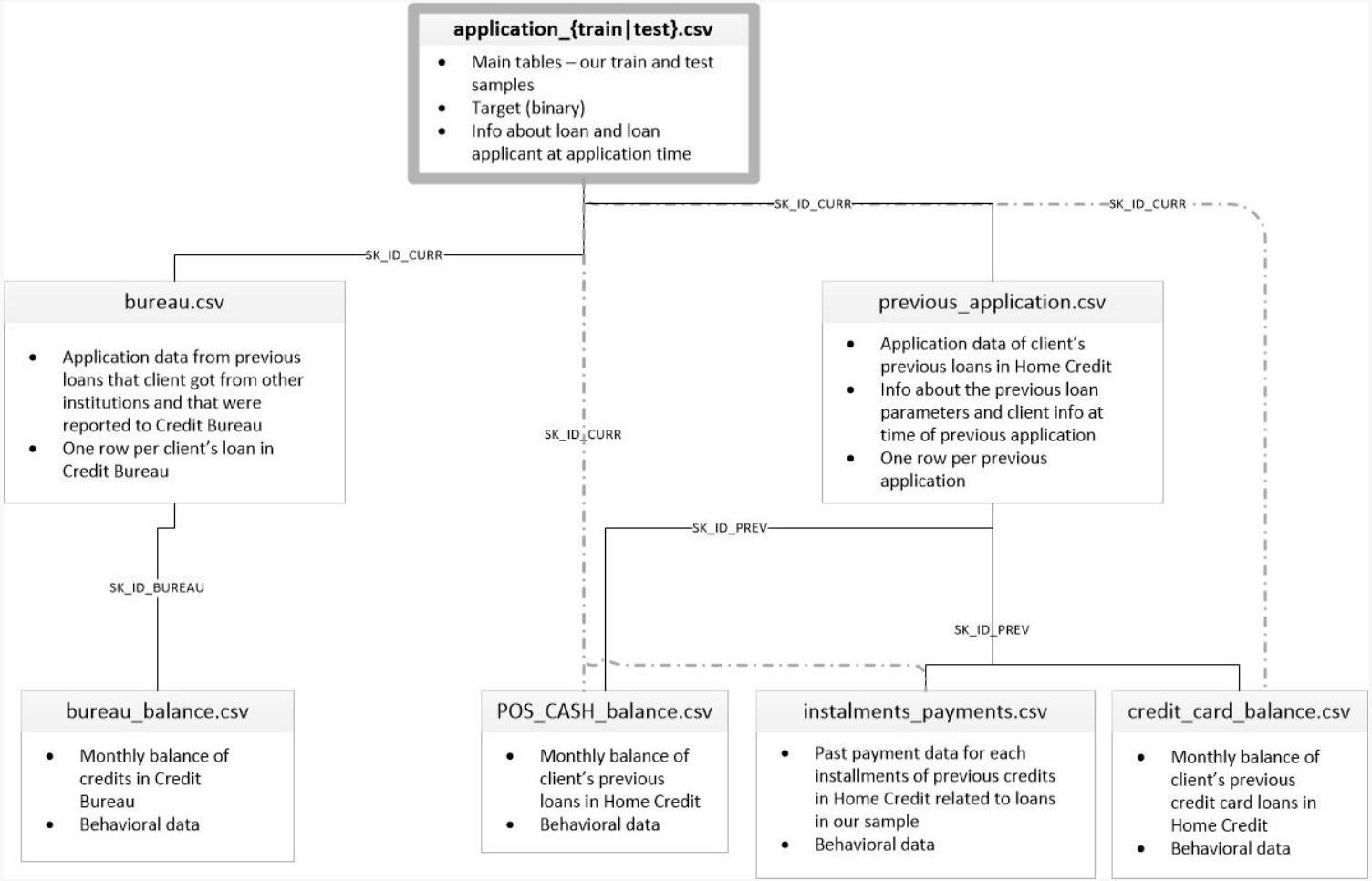
Home Credit Indonesia strives to make predictions through alternative data using various statistical methods and Machine Learning.

Objective

- Identify customers who have difficulty in repaying loans.
- Predict customers who have the capacity to repay their loans.

Goals

- Perform data cleaning, data visualization and build prediction models with machine learning.
- Predict customer repayment ability with the best machine learning model.
- Provide recommendations to the company to increase customer success in the next loan application.


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_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

Data Pre-processing

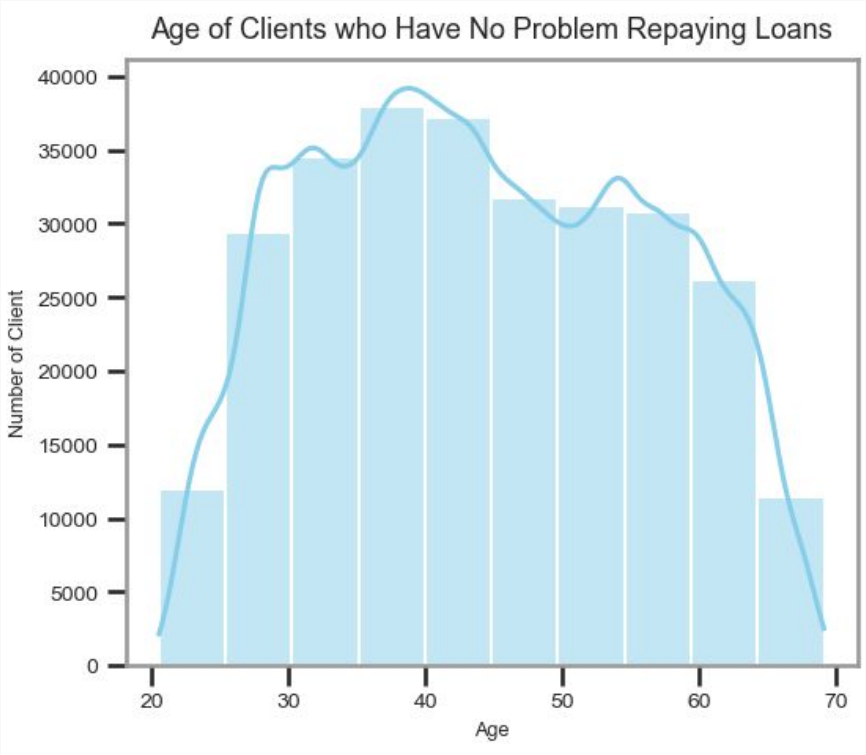
Merge Data

The process of merging data begins with reducing data by changing the data type based on the value of each feature with the aim of using less memory and preprocessing runs more smoothly.
Data merging based on the ERD Diagram of the previous dataset, with the ID value of each table being the relationship between tables

Cleaning Data

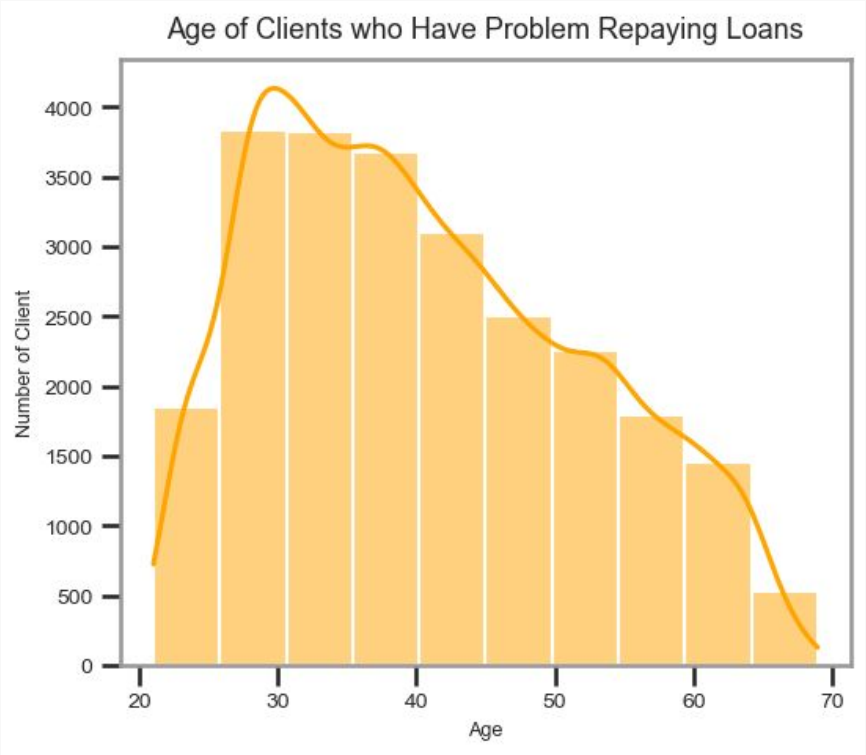
The data cleaning process is carried out by deleting several columns in each dataframe that have null values with the condition that the column has a null value > 70%. The remaining null columns are filled with the median value in each column.

Visualization & Business Insight



While clients who have no payments difficulties are client in the range of **35 - 45 years.**

Visualization & Business Insight



Clients who have payments difficulties are client in the range of **25 - 35 years.**

Based on the data information, I can provide recommendations to provide an age limit for clients, between **30 - 50 years** old so that loan repayment can be smoother. If there are clients under the age range making loans, provide special conditions so that loan payments can be completed immediately.

Machine Learning Model

Feature Selection

Feature selection ⇒ Select Kbest method.
Top 30 score of feature

Scaling Data

Scaling Data ⇒ Standar Scaler

Balancing Data

Oversampling ⇒ Random Over Sampling
Undersampling ⇒ Random Under Sampling

Modeling

- Logistic Regresion
- LightGBM
- Decision Tree

Evaluation

	Accuracy Train	Accuracy Test	ROC Score
Decision Tree oversampling	100%	95%	95%
Decision Tree undersampling	100%	55%	55%
LightGBM undersampling	71%	64%	64%
LightGBM oversampling	68%	67%	67%
Log Reg undersampling	62%	61%	61%
Log Reg undersampling	62%	62&	62%

Recommendation

1. Put a limit on the age that can take a loan, between 30 - 50 so that loan repayment can be smoother
2. If there are underage clients making loans, provide special conditions so that loan payments can be made immediately.
3. For clients who have income only as casual workers, please give the option to make a loan but with a certain loan limit.
4. Based on the prediction model I used, the LightGBM model can be implemented in the next customer prediction.

Visit My Repo!

https://github.com/hannz0/home-credit-scorecard-model