Scorecard Model

Home Credit Indonesia Data Scientist Virtual Internship Program

Rakamin Academy

Milenia Nadia Afifah Puspitasari – DS 16 – JGP 1

https://github.com/milenianadia19/HomeCredit_ScorecardModel.git

1. Problem Statement

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to **make sure this underserved population has a positive loan experience**, Home Credit makes use of a variety of alternative data--including telco and transactional information. Home Credit is currently using various statistical and machine learning methods to **predict their client's repayment abilities.**

2. Goal

Minimize the number of clients who are **approved** but **actually defaulters**

3. Objective

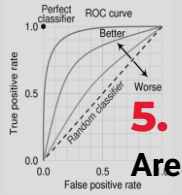
Create **predictive model** to determine potential client and default client.

4. Business Metrics

Decreased Loss Given Default (LGD).
The amount of money a financial institution loses when a borrower defaults on a loan, after taking into consideration any recovery, represented as a percentage of total exposure at the time of loss.
Source: https://bit.ly/36cNKsi

Assumption: losses due to default are only calculated from the client's total credit.

5. Model Evaluation



Area under the ROC curve

Dataset

Three datafiles are selected for modeling:

1. **application_train**: data for current credit application and binary TARGET variable. Every loan has its own row and is identified by the feature SK_ID_CURR.
   https://drive.google.com/file/d/1Ug3UX2YF70RKaoCqvd6AX_ERPzrt961_/view?usp=sharing

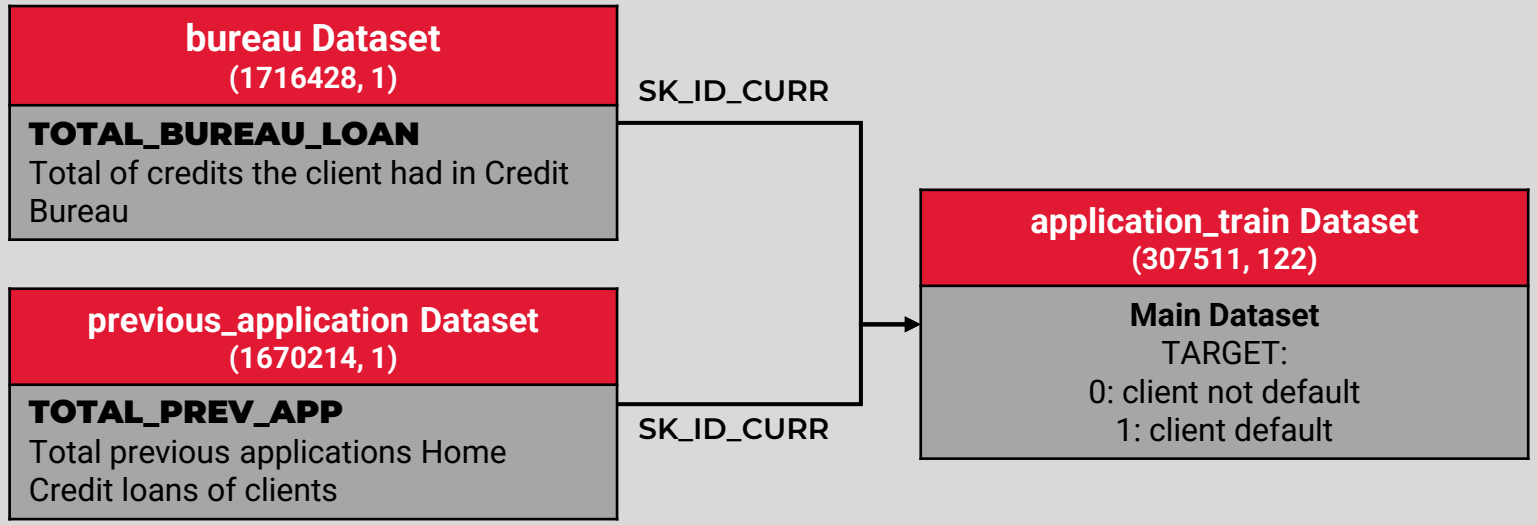
2. **bureau**: All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample). For every loan in the sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
   https://drive.google.com/file/d/1EZoUDTmClHQScNBHOFehn7AcceZ597IC/view?usp=sharing

3. **previous_application**: All previous applications for Home Credit loans of clients who have loans in the sample. There is one row for each previous application related to loans in the data sample.
   https://drive.google.com/file/d/1JaLi0L3T-JUeAXo8QtTyIqZP5H2rHDtW/view?usp=sharing

Then, for each SK_ID_CURR in the **application_test** set we will predict a probability for the TARGET variable.
https://drive.google.com/file/d/1Q-XhF7Zd-hXRTe9APFPZneZNc2DLPgo7/view?usp=sharing

---

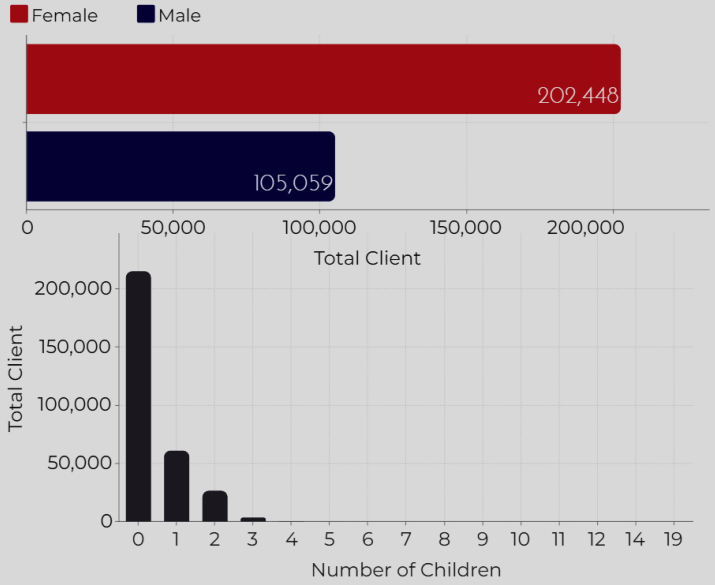
Define Single New Columns
from bureau dataset and previous_application dataset



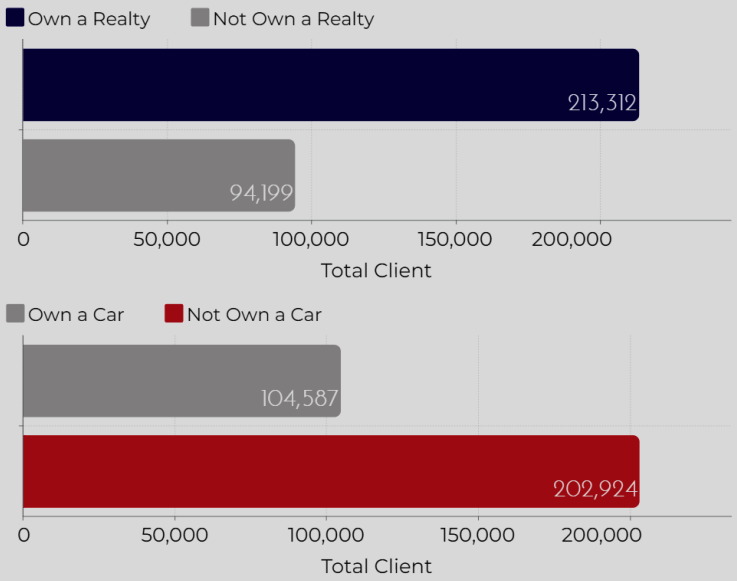
Exploratory Data Analysis

Basic Info of The Clients

Most clients were female and without any children

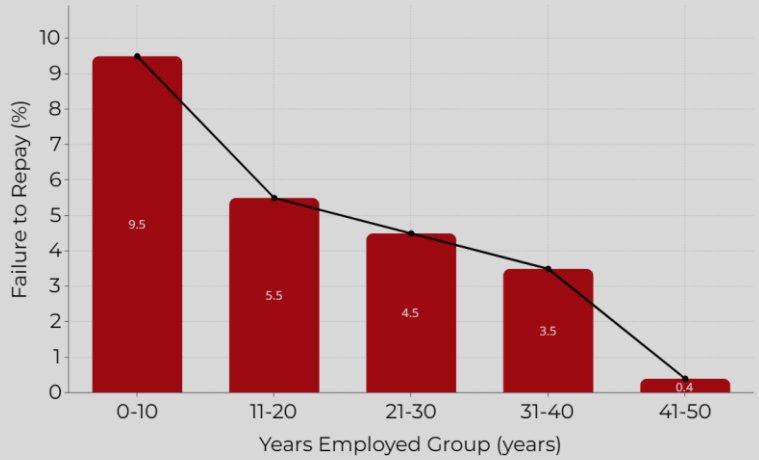


Most of the clients owned a realty but not a car



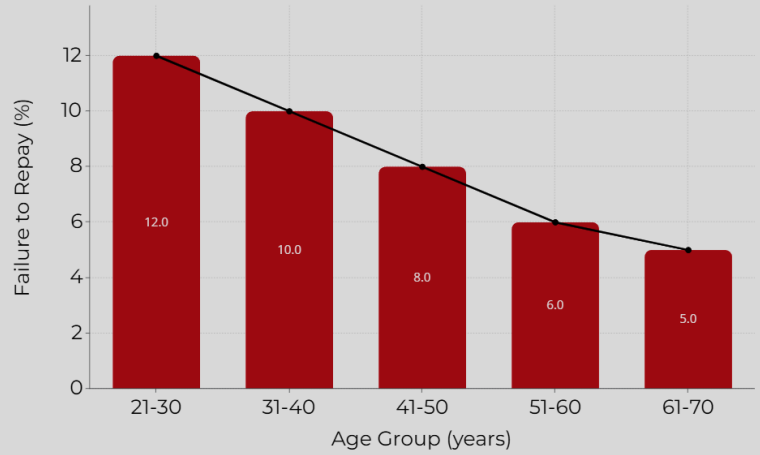
Failure to Repay by Years Employed Group

The longer the current employment of the clients are more likely to repay the loan



Failure to Repay by Age Group

Younger clients are more likely to not repay the loan



Data Preprocessing

application_train & application_test

1. Feature Engineering

- Converting DAYS_BIRTH to years to get client age
- Converting DAYS_EMPLOYED, DAYS_REGISTRATION, and DAYS_ID_PUBLISH to years
- Calculating number of documents provided by a customer
- Calculating Income Annuity Percentage (Annuity/Total Income)
  Source: https://www.blueprintincome.com/resources/income-annuities/
- Calculating Earned Income Tax Credit (Credit/Total Income)
  Source: https://sgp.fas.org/crs/misc/R43805.pdf

2. Replace XNA values with NaN

- application_train for **Training**: CODE_GENDER, ORGANIZATION_TYPE
- application_train for **Test**: ORGANIZATION_TYPE

3. Handling Missing Values

- Keeping columns that include **less than equal to 60% of missing values**
- Imputation categorical features with **mode** and numerical features with **median**

4. Scaling Numerical Features

MinMaxScaler()

5. Feature Encoding

For categorical variable with **2 unique categories** will use **label encoding** and for any categorical variable with **more than 2 unique categories** will use **one-hot encoding**.

6. Aligning Data Train and Data Test

One-hot encoding has created more columns in the training data because there were some categorical variables with categories not represented in the testing data. To remove the columns in the training data that are not in the testing data, we align the dataframe.

Dataset	Before	After
application_train	(307511, **205**)	(307511, **201**)
application_test	(48744, 200)	(48744, 200)

7. Feature Selection

- **VIF (Variance Inflation Factor)**
  drop Features with VIF > 10 (indicator of multicollinearity)
- **Correlation with Target**
  after trying several combinations, decide to pick 15 features which most correlated with the TARGET
- **Domain Knowledge**

Dataset	After Selection
application_train	(307511, **30**)
application_test	(48744, **29**)

Modeling
application_train

Splitting Train and Test
80 : 20

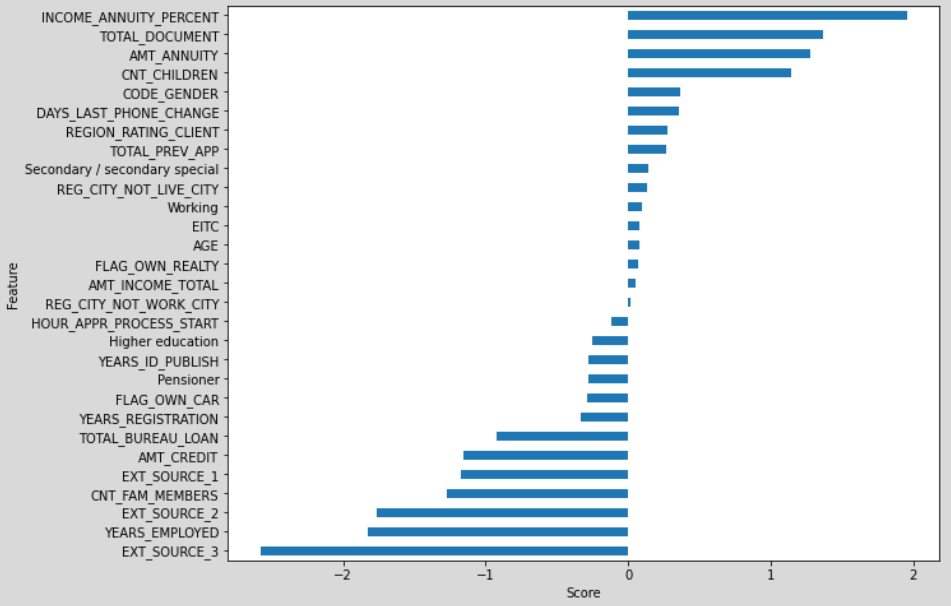
Imbalancing
Oversampling SMOTE with ratio 2 : 1

Dataset		Before Imbalancing	After Imbalancing
application_train	X_train y_train	(**246008**, 29) (**246008**,1)	(**339198**, 29) (**339198**,1)
	X_test y_test	(61503, 29) (61503, 1)	(61503, 29) (61503, 1)

Modeling

MODEL	MODEL PERFORMANCE	CONFUSION METRICS			
	AUC	PREDICT T ACTUAL T	PREDICT F ACTUAL F	PREDICT T ACTUAL F	PREDICT F ACTUAL T
Logistic Regression	0.63	1910	49937	6617	3039
Logistic Regression Tuned	0.63	1833	50370	6184	3116
Decision Tree	0.53	893	50163	6391	4056
Decision Tree Tuned	0.54	768	52832	3722	4181
Random Forest	0.51	161	56256	298	4788
Random Forest Tuned	0.53	377	55645	909	4572
XGBoost	0.51	97	56392	162	4852
XGBoost Tuned	0.51	131	56383	171	4818

Feature Importance
Logistic Regression



For the **application_test** dataset **default predict probability result**, you can check here

Business Metrics

Decreased Loss Given Default (LGD).

The amount of money a financial institution loses when a borrower defaults on a loan, after taking into consideration any recovery, represented as a percentage of total exposure at the time of loss.
Source: https://bit.ly/36cNKsi

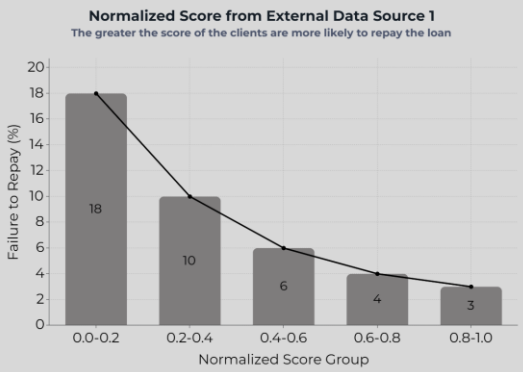
Assumption: losses due to default **are only calculated from the client's total credit**.

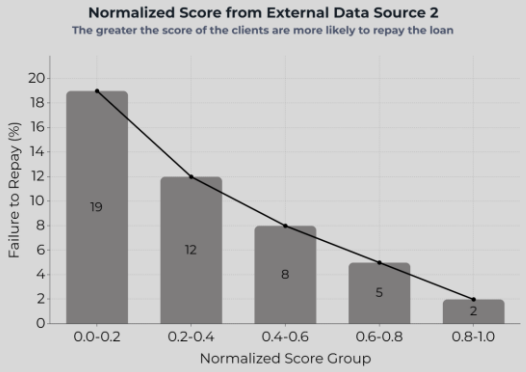
Logistic Regression Model	**Before** Data: X_test 61503 rows	**After** Data: X_test 61503 rows
Total Defaulters	4949 (Total Clients Target = 1)	3039 (Total False Negative)
Total Loss Given Default (Total Credit of Defaulters)	2,780,554,153.5	1,816,980,601.5
LGD Decrease	**- 963,573,552** **- 34.65%**	

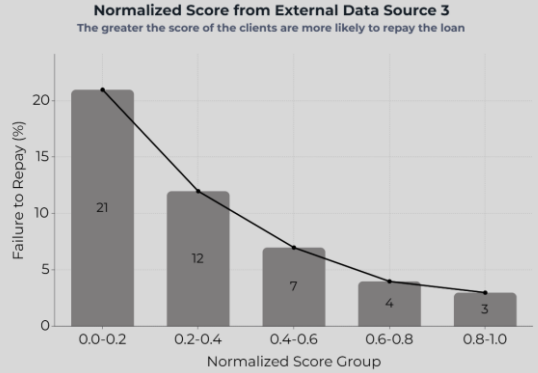
Business Recommendation

Top 4 Feature Importance	Value
EXT_SOURCE_3	-2.58
INCOME_ANNUITY_PERCENT	+1.96
YEARS_EMPLOYED	-1.83
EXT_SOURCE_2	-1.76

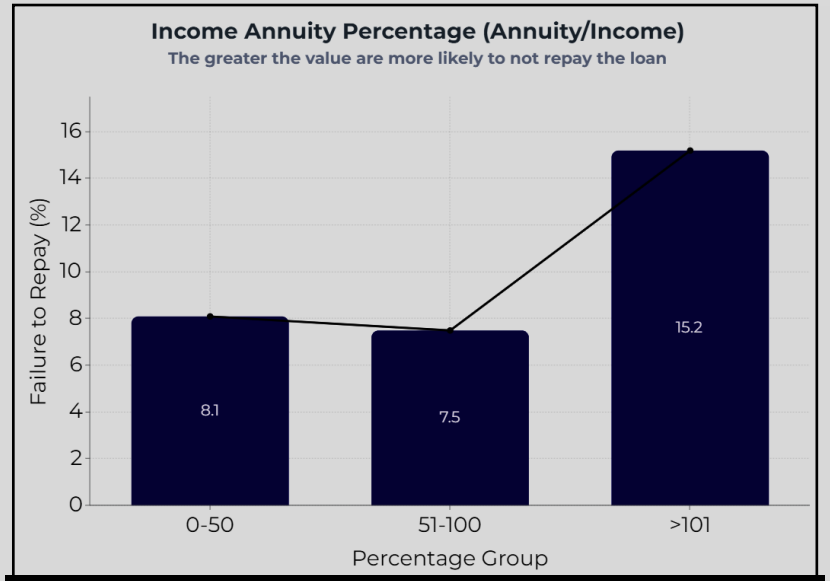
Each data source has a great influence on the possibility of the client to default, especially EXT_SOURCE_3



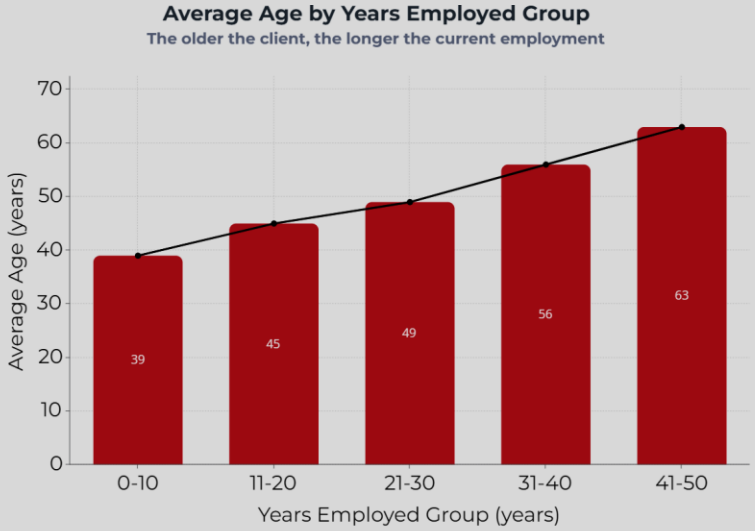


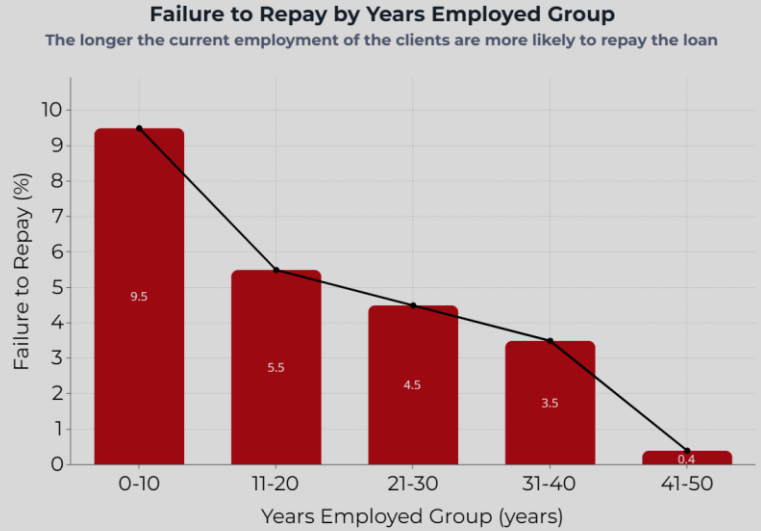


EXT_SOURCE3 can be used as one of the most important indicators in determining client's repayment abilities as it is the most important feature from the model



It is important to ensure the client's ability, especially for clients who purchase a larger income annuity





Young clients, which means that their current employment is not long, they have more potential to not repay the loan. Maybe they should be provided with more guidance or financial planning tips

Suggestion

- The best model obtained actually still does not have a very good performance. Even though it has been able to reduce LGD by 34.71%, it should still be possible for the performance of this model to be better **by adding other important features from other datasets** other than those selected in the current modeling. However, due to the limitation of computation, only three datasets were selected. It should still be able to use other important datasets.

- From the modeling results, several important features are obtained from the feature engineering. Thus, **feature engineering play an important role**, especially with a large number of dataset features. So that features engineering features can also be done for other features.