CREDIT SCORE

Azhar Kanzu Arasy

https :// github.com/kanzu7/HCI_Credit_Scored

Latar Belakang

Credit Risk atau Risiko Kredit adalah kemungkinan kerugian yang bisa didapatkan oleh kreditur ketika peminjam gagal untuk membayar kembali pinjaman atau memenuhi kewajibannya sesuai kontrak yang disepakati sebelumnya. Risiko ini bias berupa:

- Gangguan aliran kas (cash flow) sehingga modal kerja terganggu.
- Meningkatkan biaya operasional untuk mengejar pembayaran tersebut (collection).

Untuk memperkecil risiko kredit, biasanya dilakukan proses penilaian risiko sebelum diberikan pinjaman yang disebut dengan *credit scoring* dan *credit rating* terhadap pihak peminjam.

Manfaat dari credit scoring ini adalah memperkecil risiko yang bisa terjadi pada lembaga peminjam, dimana berdasarkan hasil penilaian ini akan menjadi penentu apakah aplikasi pengajuan pinjaman diterima atau ditolak oleh lembaga finansial

Dataset

Bureau.csv

TOTAL_BUREAU_LOAN
Total of credits the client had in Credit Bureau

Previous_application.csv

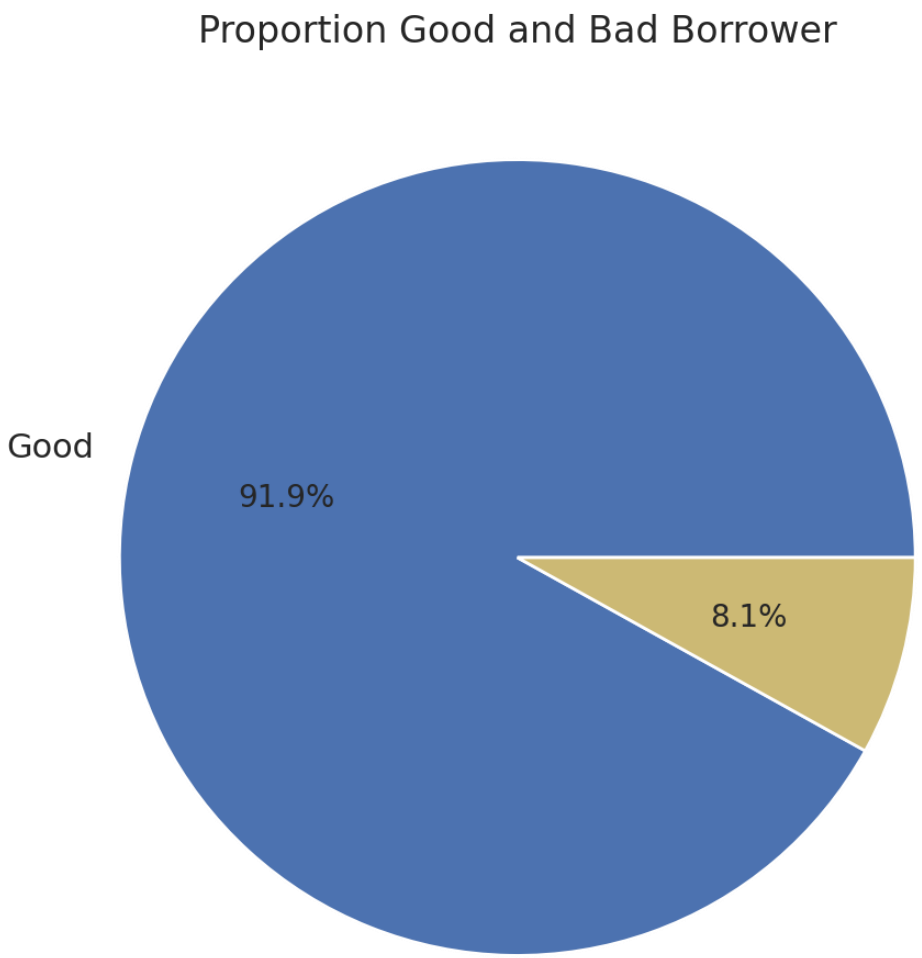
TOTAL_PREV_APP
Total previous applications Home Credit loans of clients

SK_ID_CURR

SK_ID_CURR

Application_train.csv

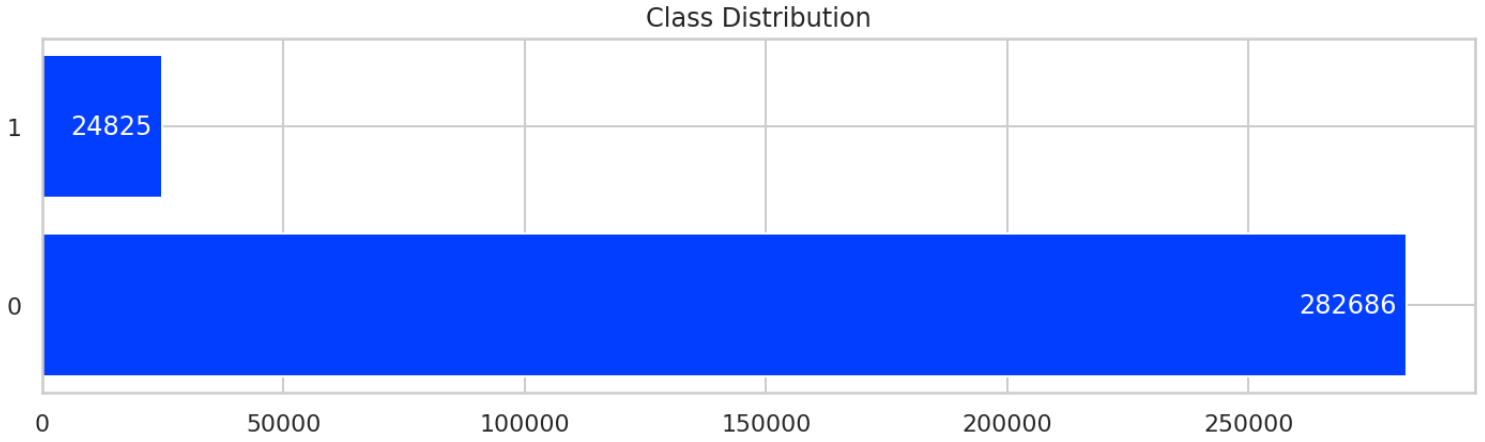
Main Dataset
TARGET:
0: client not default
1: client default



Explorary Data Analyst

Target Definition : *Good Debitur* adalah nasabah yang meminjam dengan kriteria *ontime payment* dan diberi *flag 0*, sedangkan untuk *Bad Debitur* adalah nasabah yang meminjam dengan kriteria *not ontime payment* diberi flag 1.

Model Definition: Klasifikasi nasabah peminjam *Good (Non-default)* dan *Bad (default)*



Data Prepocessing

application_train & application_test

1. Feature Engineering

- Mengubah DAYS_BIRTH ke tahun untuk dapatkan umur client
- Mengubah DAYS_EMPLOYED
- Menjumlahkan dokumen dari 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,
- 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.

7. Feature Selection

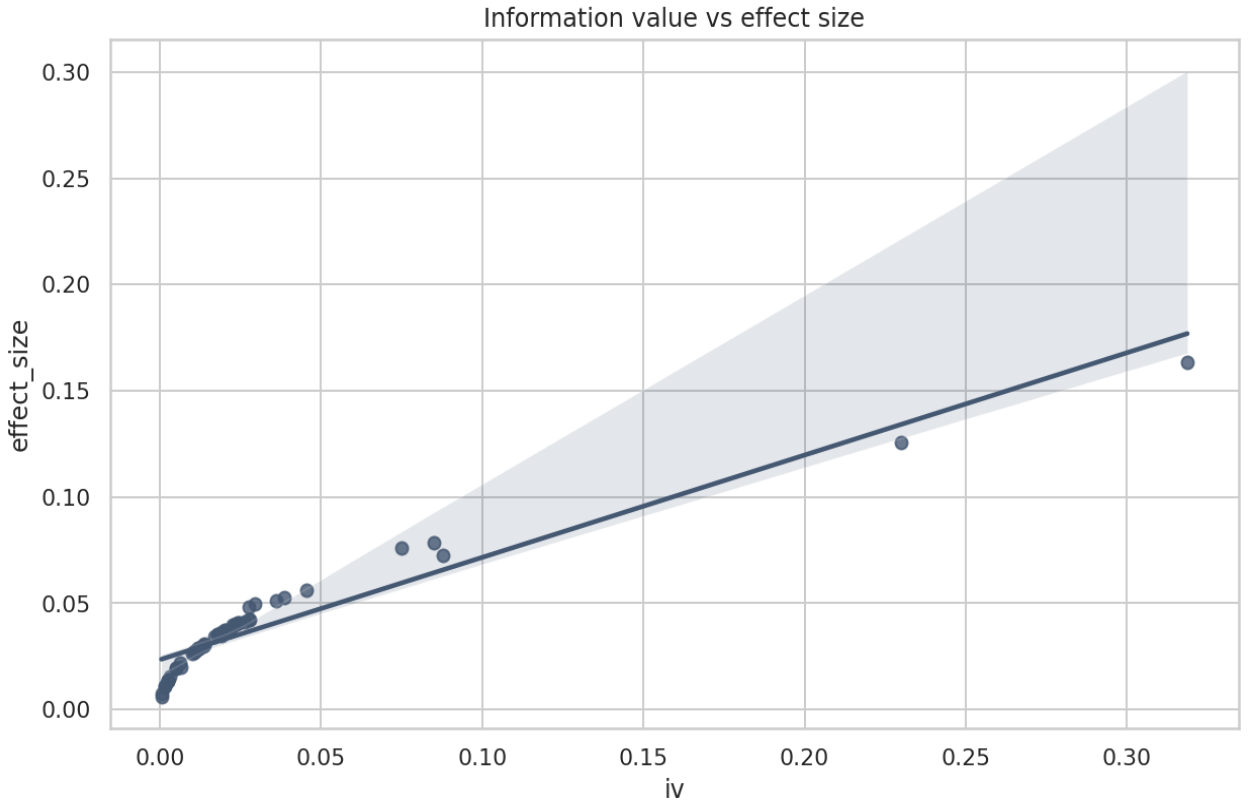
- WOE (Weight of Evidance)
- IV (Information Value)

WOE (Weight of Evidance)
IV (Information Value)

$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

$$IV = \sum(DistributionGood_i - DistributionBad_i) * WoE_i$$

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power



Variable yang digunakan yaitu variable yang memiliki value diatas nol (0)

Data Modelling

application_train

Split Train & Test
80 : 20

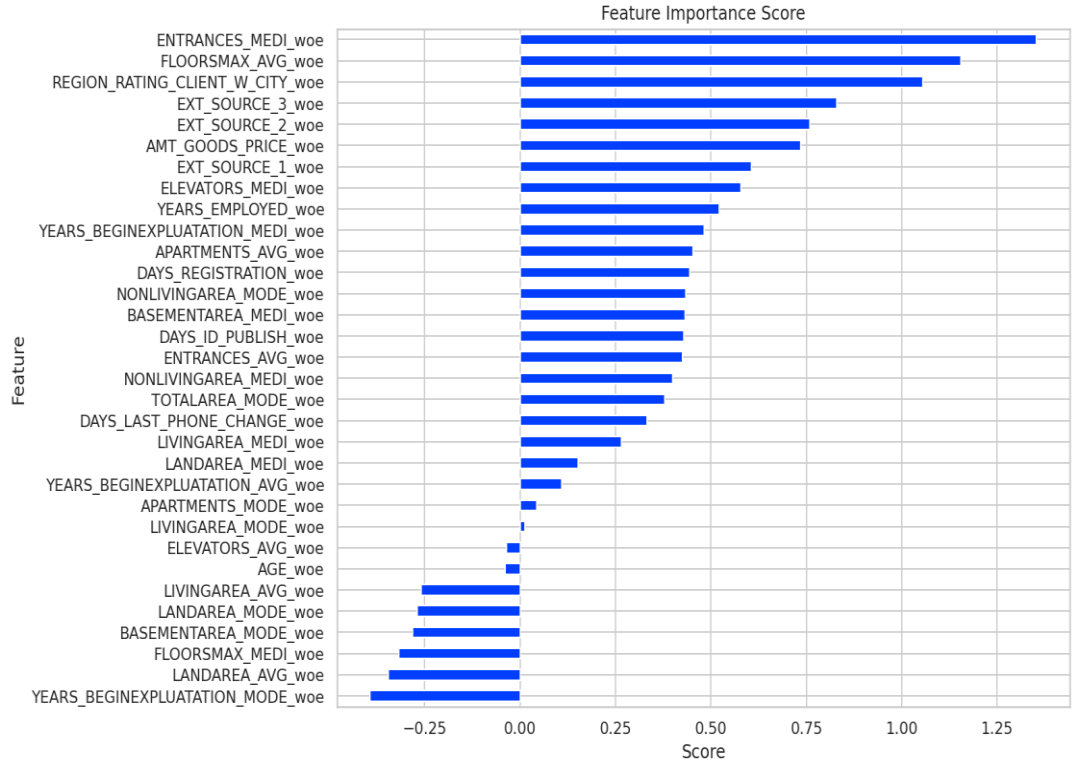
Imbalancing
Oversampling SMOTE with ratio 2 : 1

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)

Model	AUC	Confusion Metrix			
		Predit T Actual T	Predit T Actual F	Predit F Actual T	Predit F Actual F
Logistic Regression	0,73	40650	25039	201093	72416
Random Forest	0,64	2758	13203	43351	2191

Feature Important

Logistic Regression



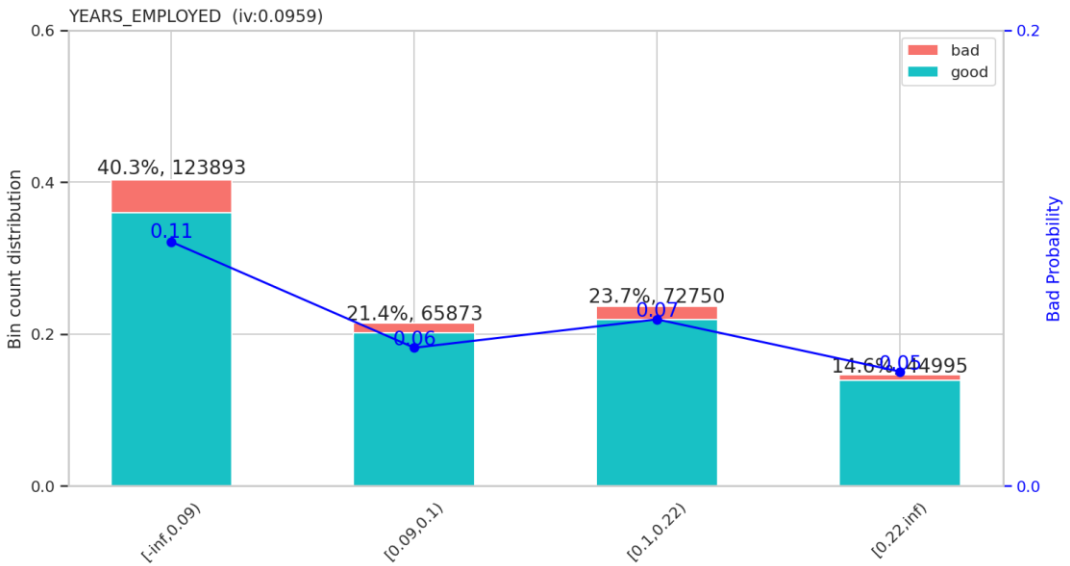
Business Metric

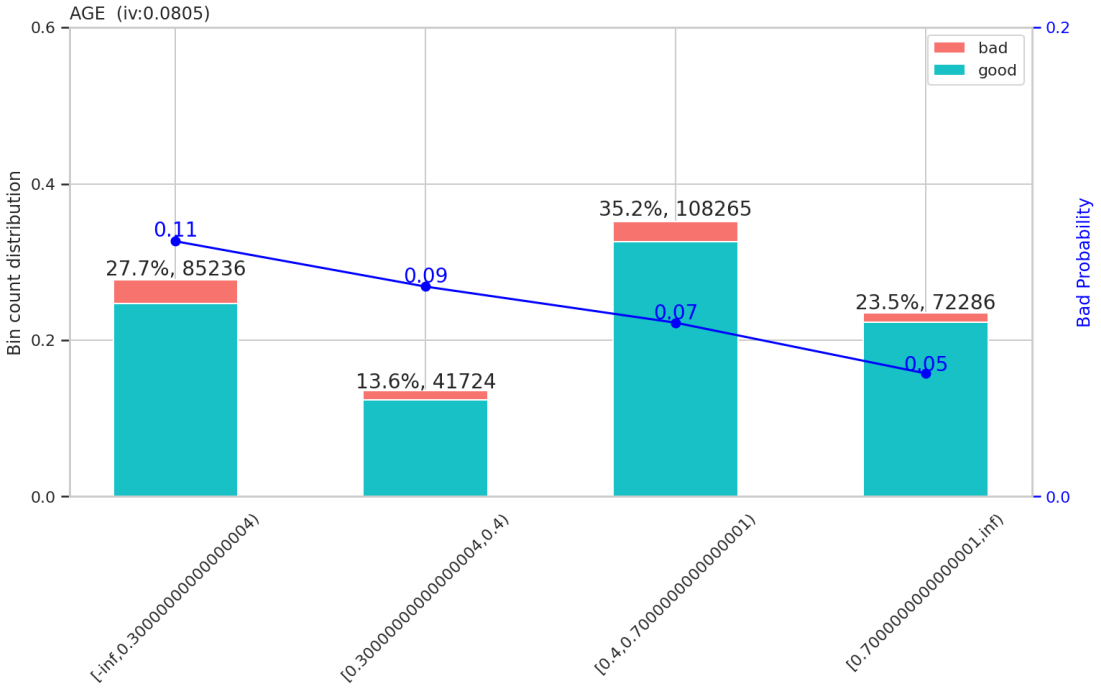
Decreased Loss Given Default (LGD)
The amount of money a financial institution loses when a borrower defaults on a loan, after taking into consideration any recovery, represented as a percentage of total exposure at the time of loss.

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)	3206 (Total False Negative)
Total Loss Given Default (Total Credit of Defaulters)	2,780,554,153.5	1,951,274,416.5
LGD Decrease	-829,279,737 **-29,82%**	

Business Recomendation





Client dengan usia muda memiliki kerentanan untuk tidak membayar yang cukup tinggi,
hal ini dikarenakan memiliki pekerjaan yang tidak lama bisa jadi suka berpindah-pindah tempat kerja.
Untuk mengatasi ini, munkin kita bisa mengajarkan financial planning kepada mereka yang berusia muda
Agar memiliki pengetahuan yang cukup.

THANKYOU