HOME CREDIT
Kamu Bisa!

VIRTUAL INTERNSHIP EXPERIENCE (VIX)

by Archie Citra Muhammad

PROBLEM RESEARCH

- Problem : How do you help the assessment team examine customer loans?
- Goal : Increase the speed of filing inspection without increasing costs
- Objective : Create a system to help loan assessments automatically
- Business Metrics : daily resolved applications & average resolved time

DATA PREPROCESSING

Data Cleaning

Check Data Duplicate
Check Missing Data

Feature Selection

Split Data Train (80:20)

Categorical (Chi Square)

Numerical (ANOVA)

Feature Engineering

Simple Imputer

OHE with dummy creation

Feature Engineering

WoE Binning

Information Value (IV)

https://towardsdatascience.com/feature-selection-and-eda-in-python-c6c4eb1058a3
https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03
https://medium.com/@finntanweelip/feature-selection-in-credit-scoring-b0eee604cd51

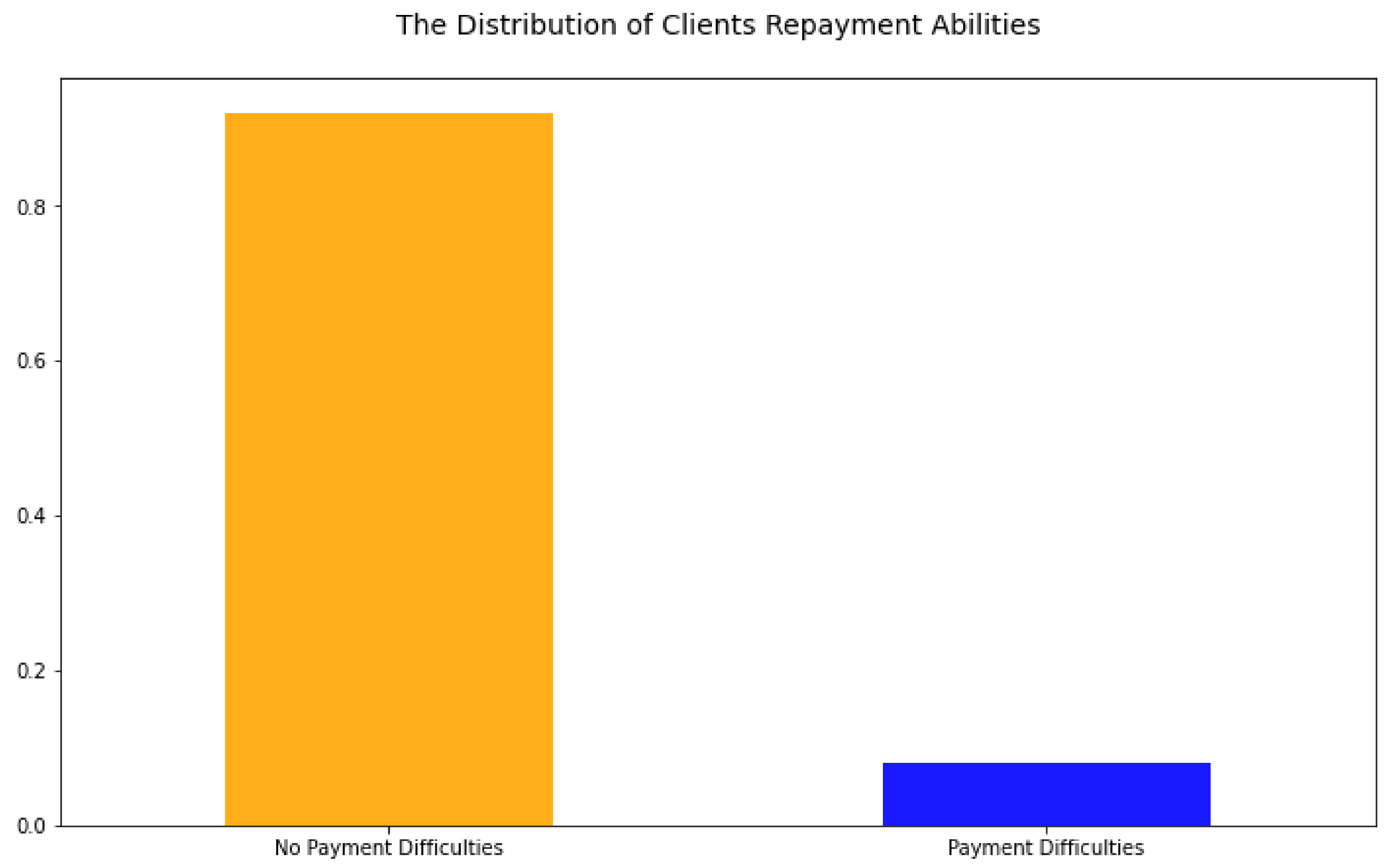
DATASET

01 application_train has 307511 rows and 122 columns

- Float64 : 65
- Int64 : 41
- Object : 16

02 application_test has 48744 rows and 121 columns

- Float64 : 65
- Int64 : 40
- Object : 16

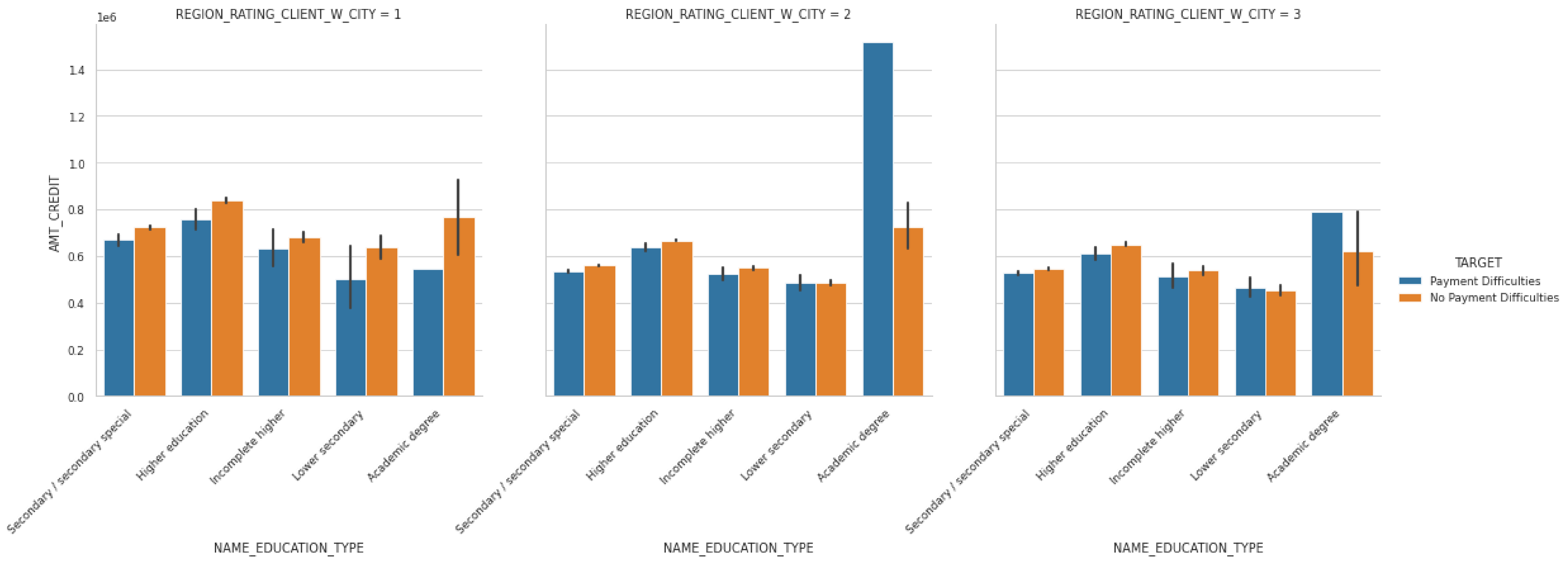


Column Target from application_train

- No Payment Difficulties : 92%
- Payment Difficulties : 8%

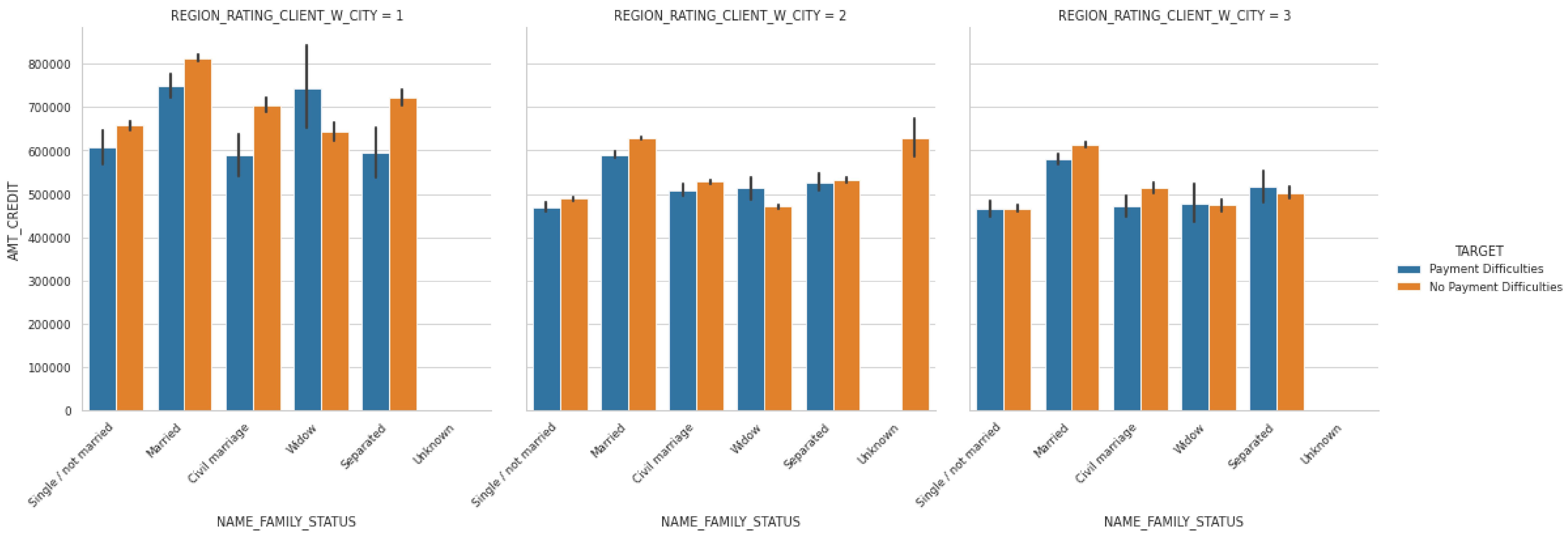
https://github.com/archie-cm/Credit_Score_Home_Credit_Indonesia

DATA VIZ AND BUSINESS INSIGHT



For clients who have an academic degree and live in an area with a rating of 2, have problems repaying loans for higher credit amounts. And, clients with the same degree but living in a region with a rating of 3 have problems repaying loans for moderate amounts of loan credit.

DATA VIZ AND BUSINESS INSIGHT



- Clients who are widowed, whether domiciled in areas with a rating of 1, 2, or 3, have difficulty paying off loans for moderate to high loan amounts.
- Clients who have separate family status, and live in areas rated 3, have problems repaying loans for a moderate amount of loan credit compared to clients who live in areas rated 1 or 2.

MACHINE LEARNING & EVALUATION

Modelling

Class_Weight = Balance
Woe_transform

Pipeline

Pipeline(steps=
[('woe', woe_transform),
('model', dt)])

Cross Validation

RepeatedStrafieldKfold
N_split = 5
N_repeats = 3
Random_state = 42

Evaluation

Models	MEAN AUROC	GINI
Decision Tree	0.54	0.07
Logistic Regression	0.73	0.46

I think the feature we selected isn't the best feature to model a credit scorecard. But, 0.73 is acceptable for the baseline. (Hosmer & Lemeshow (2013). Applied logistic regression. p.177)

SCORECARD

General

BASE(intercept) = 569

Min score = 300

Max score = 850

YEAR_LAST_PHONE_CHANGE	SCORE
<2	-7
2-4	-3
4-6	1
6-8	4
8-10	5
>10	0

YEAR_ID_PUBLISH	SCORE
<4	-11
4-8	-7
8-12	-4
12-16	4
>16	18

CODE_GENDER	SCORE
M	-10
F or XNA	10

YEAR_REGISTRATION	SCORE
<17	-4
17-34	-1
34-51	-4
>51	8

YEAR_BIRTH (AGE)	SCORE
<30	-2
30-40	-10
40-50	-2
50-60	6
>60	8

REGION_RATING_CLIENT_W_CITY	SCORE
0	0
1	18
2	9

REGION_POPULATION_RELATIVE	SCORE
<0.0147	2
0.0147-0.0292	1
0.0292-0.0436	-2
0.0436-0.0581	1
>0.0581	-1

AMT_CREDIT	SCORE
<846000	-3
846000-1647000	1
1647000-2448000	10
2448000-3249000	3
>3249000	-11

NAME_EDUCATION_TYPE	SCORE
Academic degree	53
Higher education	4
Incomplete higher	-8
Lower secondary	-30
Secondary / secondary special	-19

NAME_FAMILY_STATUS	SCORE
Single or Unknown	-2
Civil marriage	-4
Married	6
Separated	-3
Widow	3

NAME_INCOME_TYPE	SCORE
Businessman or Commercial Associate	10
Pensioner or maternity leave	9
Student or unemployed	-48
State servant	24
Working	5

FLAG_DOCUMENT_3	SCORE
0	8
1	-8

REG_CITY_NOT_LIVE_CITY	SCORE
0	6
1	-6

EXT_SOURCE_2	SCORE
<0.0855	-52
0.0855-0.171	-34
0.171-0.256	-24
0.256-0.342	-13
0.342-0.427	-5
0.427-0.513	3
0.513-0.598	10
0.598-0.684	20
0.684-0.769	37
>0.769	56

EXT_SOURCE_3	SCORE
<0.0901	-65
0.0901-0.18	-48
0.18-0.269	-32
0.269-0.359	-14
0.359-0.448	-1
0.448-0.538	5
0.538-0.627	25
0.627-0.717	35
0.717-0.806	46
>0.806	50

https://github.com/archie-cm/Credit_Score_Home_Credit_Indonesia

PREDICTION

General
BASE(intercept) = 569
Min score = 300
Max score = 850

Application_test
48744 applicants

Model
Logistic Regression
AUC 0.73
Recall 0.96

Threshold
0.5

Best Threshold
0.29957

Threshold = 0.5

Accept Score	N Approved	N Rejected	Approval Rate	Rejection Rate
569.0	40097	21406	0.651952	0.348048

Best Threshold

Accept Score	N Approved	N Rejected	Approval Rate	Rejection Rate
524.0	55529	5974	0.902867	0.097133

Threshold 0.5 would result in a very high rejection rate with a corresponding loss of business.
Accordingly, we will stick with our ideal threshold and the corresponding Credit Score of 524

https://github.com/archie-cm/Credit_Score_Home_Credit_Indonesia

BUSINESS RECOMMENDATION

Possible Scenarios:

- Full automation

  Submissions are instantly accepted/rejected based on the output of the model
- Auto-reject

  Submissions that may be bad immediately rejected.

  If not, it needs to be checked manually first by the assessment team
- Partial Auto-reject & Auto-approve

  Submissions that may be bad immediately rejected.

  Submissions that are highly likely to be good are immediately accepted.

  If it's still 'gray', just checked manually by the assessment team

Metrics Impact

Business Metrics	Before	After
Daily resolved applications	10.000	50.000
Average resolved time	50 hours	1 hour

https://www.observe.ai/blog/what-is-average-resolution-time-art-for-live-chat-conversations