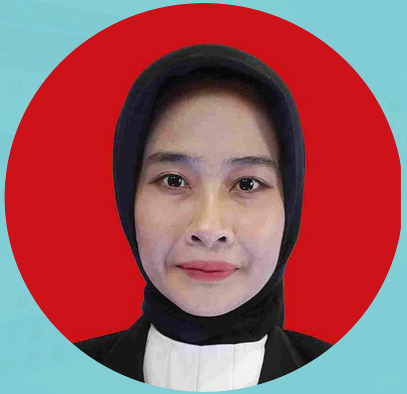
Rakamin Academy X HOME CREDIT JADI BISA!

Machine Learning Project

Home Credit Scorecard Model Using Logistic Regression

Home Credit Indonesia Data Scientist Project Based Internship Program

Presented by
Nicken Shidqia Nurahman



Nicken Shidqia Nurahman

About Me

Civil engineer graduate with some experience in administration and project management, who is interested in data science.
Detail oriented, and time management person, and familiar with Microsoft Office, Python, SQL and Jupyter.
Motivated to continue to learn and grow as a professional.

My Experience

Rakamin Academy

- Data Science Bootcamp Student – RAKAMIN ACADEMY
  Oct 2023 - Now
- Project Management Masters Degree Student – UNIVERSITAS INDONESIA
  Sep 2021 - Sep 2023
- Engineering Administration and Project Control Staff - PT. ISTAKA KARYA
  Aug 2019 - Sep 2021
- Project Control Intern - PT. ISTAKA KARYA
  Feb 2019 - Jul 2019
- Surveying Laboratory Assistant – UNIVERSITAS TRISAKTI
  Jul 2017- Agust 2019

Case Study

Problem

The main risk for loan companies is **failure** to **assess credit risk** accurately and efficiently

Disadvantage of Manual credit risk assessment

- **Subjectivity**

Subjectivity can introduce bias and inconsistency in decision-making.

- **Time-Consuming**

time-consuming especially when dealing with a large number of loan applications.

- **Risk of Error**

Humans errors, such as data entry mistakes, miscalculations, or oversight of important details.

Goal

- Predict client's repayment abilities
- Speed up inspection filing without spending more money

Challenges

Build a machine learning model that can automatically assess loans

Tool & Library Used

Data Preprocessing

A. Data Cleaning

Missing Values

index	total_null	data_type	percentage_missing
COMMONAREA_MEDI	214865	float64	69.872297
COMMONAREA_AVG	214865	float64	69.872297

There are **40 columns** that have **null** values

Handling missing value

- **Drop feature** that have missing value > 50%
- **Replace missing values** on numerical category with median & categorical with mode

Duplicated Values

No duplicated value

B. Feature Selection

Split Data Train (80:20)

```
x_train.shape, x_test.shape, y_train.shape, y_test.shape
((246008, 80), (61503, 80), (246008,), (61503,))
```

Categorical & Numerical Selection

Feature	p-value	count	unique
NAME_CONTRACT_TYPE	0.000000	246008	2
FLAG_OWN_CAR	0.000000	246008	2

- Low cardinality (unique)
- No null values
- p-value < 0.05 (using chi square for categorical & ANOVA for numerical)
- Correlation coefficient <= 0.7

- ***Before*** *Feature Selection = 122 columns*
- ***After*** *Feature Selection = 16 columns*

Data Preprocessing

C. Feature Engineering

Weight of Evidence (WOE) & Information Value (IV)

FLAG_OWN_CAR	good_distr	bad_distr	WOE	IV
N	0.657145	0.696928	-0.058778	0.007245
Y	0.342855	0.303072	0.123339	0.007245

- **WOE** generally described as a measure of the separation of good and bad customers
- **IV** helps to rank variables on the basis of their importance.

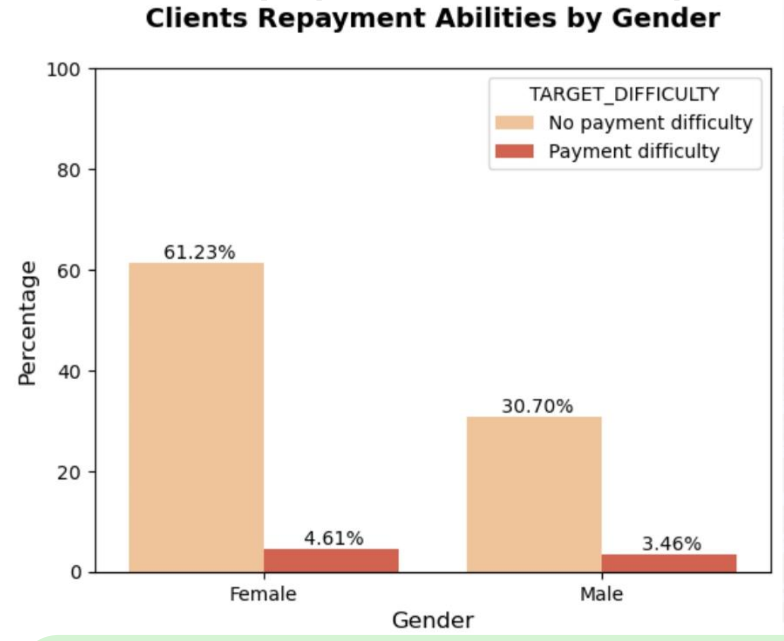
Drop Feature No Needed

- *IV < 0.02, The variable is Not useful for prediction*
- *IV> 0.5, The variable is Suspicious Predictive Power*

- ***Before*** *Feature Engineering = 16 columns*
- ***After*** *Feature Engineering = 14 columns*

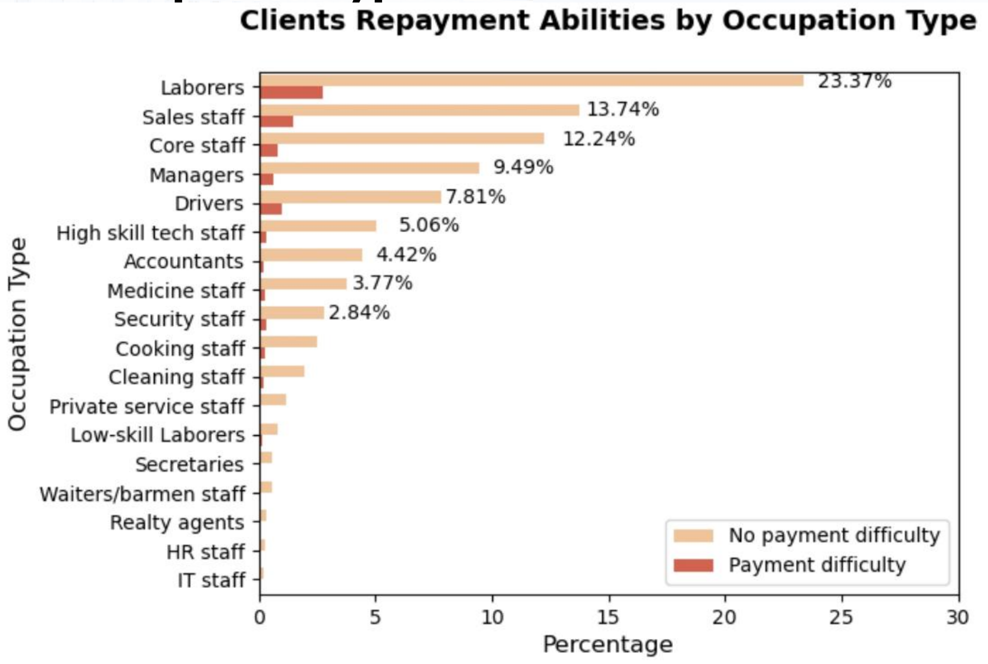
2 Top Data Visualization & Insight

A. Clients Repayment Abilities by Gender



- **61.23%** customers that do not have payment difficulty are **female,** and 30.70% are Male
- In **UK, Women** account for **65%** of the home credit industry's customers (Bermeo, 2018)
- **Recommendation** : Start a campaign to encourage more women to apply for credit

B. Clients Repayment Abilities by Occupation Type



- **23.37%** customers that do not have payment difficulty are laborers, then followed by staff and managers.
- **Recommendation** : Start a campaign to encourage more laborers, staff, and managers to apply for credit

Machine Learning Implementation

A. Evaluation Score

Algorithm	Mean AUROC	GINI
Decision Tree	0.5384	0.0768
Logistic Regression	**0.7304**	**0.4608**

- **Mean AUROC** of **0.7304** is generally considered **good**, indicating that the logistic regression model is effective at distinguishing between the positive and negative classes.
- Based on (Trifonova, 2012) An AUC - ROC 0.7–0.8 is considered good.

- **Gini** coefficient of **0.4608** indicates a **relatively strong separation** between the model's performance and random chance.
- It suggests that the logistic regression model has a good discriminatory ability.
- Based on (Teng, 2011) Gini coefficient 0.4 - 0.5 considered big gap.

B. Score Card

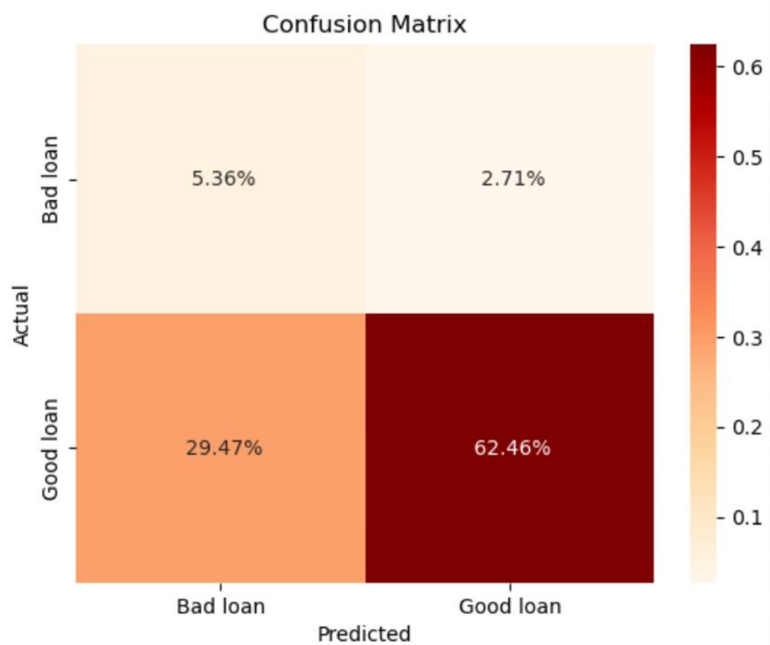
Ori_Feature_Name	Score_Calculation
intercept	555.0
CODE_GENDER	-8.0
CODE_GENDER	11.0
NAME_EDUCATION_TYPE	60.0

- Base (Intercept) = 555
- Min Score = 300 (FICO)
- Max Score = 850 (FICO)

C. Confusion Matrix with Threshold = 0.5

	precision	recall	f1-score	support
0	0.15	0.66	0.25	4965
1	0.96	0.68	0.80	56538
accuracy			0.68	61503

- **Precision** = Out of all the loan status that the model predicted would get good loan, only **96%** actually did.
- **Recall** = Out of all the loan status that actually did get good loan, the model only predicted this outcome correctly for **68%** ofthose loan status.
- **F1 Score** = **0.8**. F1 score of 0.7 or higher is often considered good (spotintelligence.com, 2023)
- The **accuracy** is not really good because we've got **0.68** out of 1



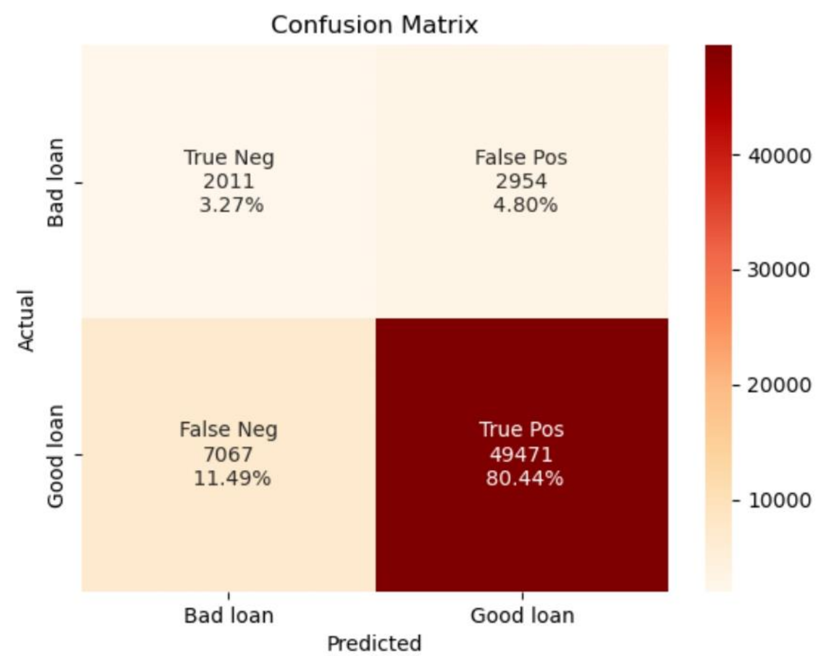
62.46% got correct variable of good loan

D. Confusion Matrix with Best Threshold

- **Best threshold** = **0.353918** (Using Youden J-Statistic)
- Best threshold is used to minimized the False Positive Rate and **maximize** the **True Positive Rate**

	precision	recall	f1-score	support
0	0.22	0.41	0.29	4965
1	0.94	0.88	0.91	56538
accuracy			0.84	61503

- **Precision** = Out of all the loan status that the model predicted would get good loan, only **94%** actually did.
- **Recall** = Out of all the loan status that actually did get good loan, the model only predicted this outcome correctly for **88%** ofthose loan status.
- **F1 Score** = **0.9**. F1 score of 0.7 or higher is often considered good (spotintelligence.com, 2023)
- The **accuracy** increased significantly from 0.68 to 0.84



80.44% got correct variable of good loan

E. Approval & Rejection Rate

(Threshold = 0.5)

threshold	score	n_approved	n_rejected	approval_rate	rejection_rate
0.500012	551.0	40082	21421	0.651708	0.348292

- Choosing a **0.5** threshold might mean rejecting a lot of applicants with **rejection rate 34%,** which could lead to losing business

(Best Threshold = 0.353918)

threshold	score	n_approved	n_rejected	approval_rate	rejection_rate
0.353918	516.0	52426	9077	0.852414	0.147586

- With **best threshold**, we've got **rejection rate 14%**
- So, we've decided to keep our preferred threshold = 0.353918 and Credit Score of 516

Business Recommendation

Partial Auto Reject & Auto Approve

- If a submission seems bad, it is rejected right away.
- If a submission appears to be very good, it is accepted immediately.
- If there's uncertainty, it is manually checked by the assessment team.

Create targeted campaign

- We should launch additional campaigns targeting women, laborers, staff, and managers to encourage them to apply for credit.

Rakamin Academy

Link Portolio On Github :

https://github.com/nickenshidqia/Credit_Scorecard_Model_Home_Credit_Indonesia

LinkedIn:

https://www.linkedin.com/in/nickenshidqia/

Thank You

Rakamin Academy X id/x partners