CREDIT SCORE PREDICTION - HOME CREDIT INDONESIA

BY: JIHAD AKBAR

GITHUB: HTTPS://GITHUB.COM/JIHADAKBR/CREDIT-SCORE-PREDICTION

1. BUSINESS UNDERSTANDING

Home Credit Indonesia is currently using various statistical methods and Machine Learning to make credit score predictions. Now, we ask you to unlock the maximum potential of our data. By doing so, we can ensure that:

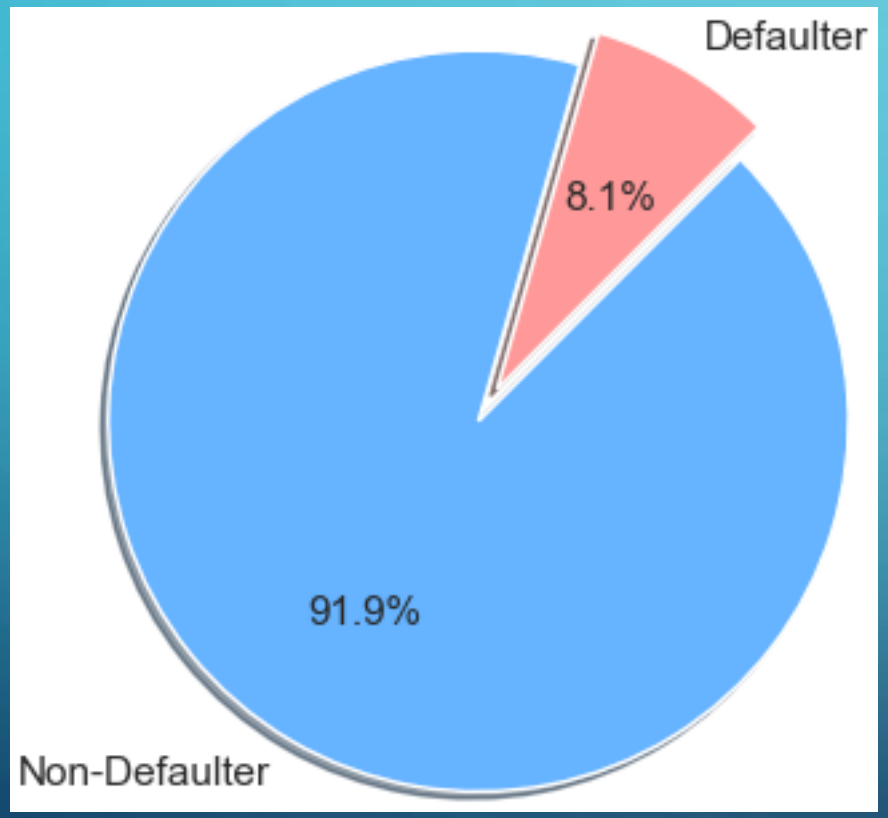
- Customers who are capable of repayment are not rejected when applying for a loan.
- Loans can be given with a principal, maturity, and repayment calendar that will motivate customers to succeed.

Evaluation will be done by checking how deep your understanding of the analysis is. Note that you need to use at least Logistic Regression to construct your machine learning models. After that, create a presentation slide containing end-to-end modeling analysis results along with business recommendations (maximum 10 pages).

2. THE PROJECT WORKFLOW

No.	Workflow	Weight
1	Problem Formulation	5%
2	Data Collecting	5%
3	Data Understanding	5%
4	Data preprocessing	20%
5	Exploratory Data Analysis (EDA) and Data Visualization	5%
6	Feature Selection and Engineering	30%
7	Model Selection and Building	15%
8	Scorecard Development	15%

3. RESULTS – TARGET VARIABLE



The target variables consist of 91.3% non-defaulters (accepted) and 8.7% defaulters (rejected).

3. RESULTS – ML METRICS

Logistic regression was employed in a machine learning model, yielding the following metrics: threshold ≈ 0.23, accuracy ≈ 0.90, precision ≈ 0.93, recall ≈ 0.96, F1 ≈ 0.94, AUROC ≈ 0.74, Gini ≈ 0.48, and AUCPR ≈ 0.97.

3. RESULTS – SCORECARD DEVELOPMENT

	index	Feature name	Coefficients	Original feature name	Score - Calculation	Score - Preliminary	Difference	Score - Final
0	0	Intercept	-0.320272	Intercept	536.624900	537.0	0.375100	537.0
1	1	NAME_CONTRACT_TYPE:Revolving loans	0.514260	NAME_CONTRACT_TYPE	28.038905	28.0	-0.038905	28.0
2	2	CODE_GENDER:M	-0.273369	CODE_GENDER	-14.904825	-15.0	-0.095175	-15.0
3	3	EXT_SOURCE_2:missing	-0.185798	EXT_SOURCE_2	-10.130231	-10.0	0.130231	-10.0
4	4	EXT_SOURCE_2:<=0.171	-0.857951	EXT_SOURCE_2	-46.777874	-47.0	-0.222126	-47.0
...	...	...	...	...	...	...	...	...
91	11	REG_CITY_NOT_LIVE_CITY:0	0.000000	REG_CITY_NOT_LIVE_CITY	0.000000	0.0	0.000000	0.0
92	12	DAYS_REGISTRATION:>0.0	0.000000	DAYS_REGISTRATION	0.000000	0.0	0.000000	0.0
93	13	AMT_GOODS_PRICE:>2254500.0	0.000000	AMT_GOODS_PRICE	0.000000	0.0	0.000000	0.0
94	14	REGION_POPULATION_RELATIVE:>0.0725	0.000000	REGION_POPULATION_RELATIVE	0.000000	0.0	0.000000	0.0
95	15	LOAN_DURATION:>39.702	0.000000	LOAN_DURATION	0.000000	0.0	0.000000	0.0

4. CONCLUSION – MONEY LOSSES AND SAVED

	Total Applicants	Total Accepted	Total Rejected	Acceptance Rate	Rejection Rate	TP+TN	FP+FN	Money Saved (IDR)	Money Losses (IDR)
0	61503	58082	3421	94.44%	5.56%	89.75%	10.25%	3.14e+10	1.72e+08

- True Positive (TP): If my machine predicts that the applicant will not default, and they actually do not default.
- True Negative (TN): If my machine predicts that the applicant will default, and they actually do default.
- False Positive (FP): If my machine predicts that the applicant will not default, but they actually do default.
- False Negative (FN): If my machine predicts that the applicant will default, but they actually do not default.

- Consequently, the company is expected to save around 30,000,000,000 IDR while incurring a loss of approximately 100,000,000 IDR.
- The high or low percentages of True Positive/Negative and False Positive/Negative depend on the metrics of the machine learning model mentioned above.

4. CONCLUSION – RECOMMENDATION (1)

We can enhance them further by incorporating features with higher information value (IV). Several CSV files encompassing such features possess significant IV potential, yet I was unable to merge them into the application_train.csv and application_test.csv datasets. These files comprise:

1. bureau.csv
2. bureau_balance.csv
3. credit_card_balance.csv
4. installments_payments.csv
5. POS_CASH_balance.csv
6. previous_application.csv

4. CONCLUSION – RECOMMENDATION (2)

- That files contain features with higher potential IV but couldn't be merged into application_train.csv and application_test.csv. This limitation is due to the current laptop (4GB RAM) experiencing crashes when attempting to merge these files.
- It is hoped that in the future, a more advanced laptop/computer can be acquired to successfully merge these files.

THANK YOU!