CREDIT SCORE PREDICTION - HOME CREDIT INDONESIA

BY JIHAD AKBAR

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:

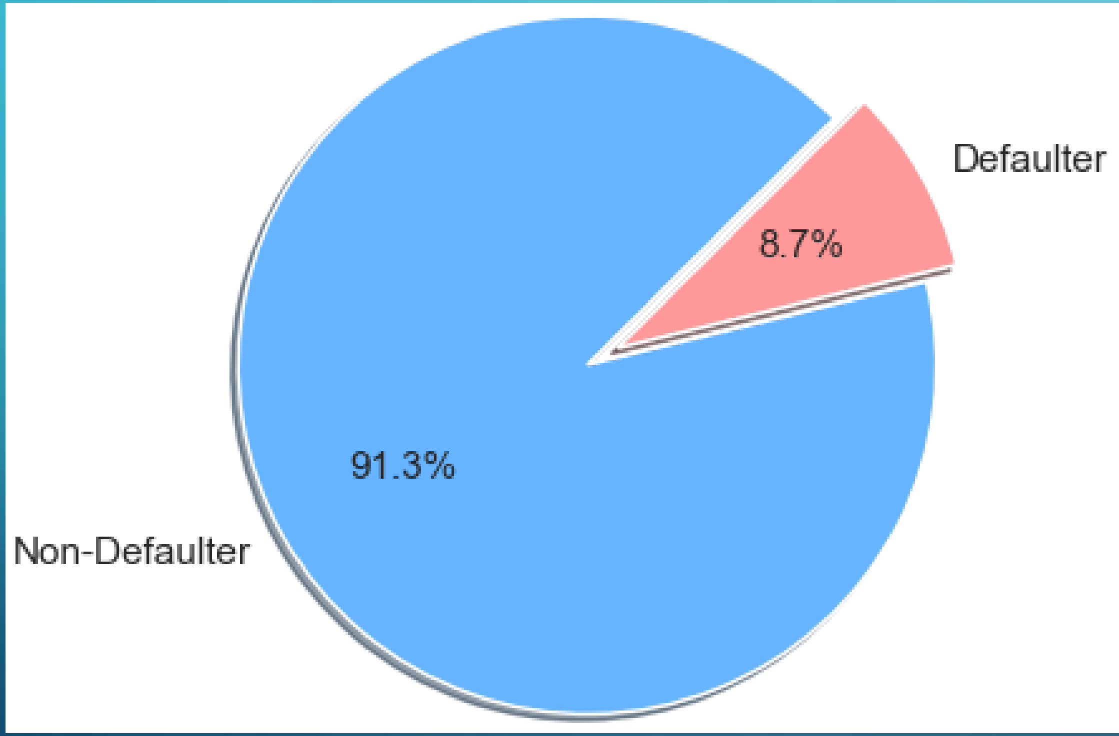
- Customers who are capable of repayment are not rejected when applying for a loan.
- Loans can be given with a principal, maturity, and repayment calendar that will motivate customers to succeed.

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

A machine learning model employing Logistic regression has been utilized, resulting in a mean AUROC of approximately 73.5%, Gini around 47.0%, and AUCPR of approximately 21.8%.

3. RESULTS – SCORECARD DEVELOPMENT

	SK_ID_CURR	Thresholds	y prob. prediction	Thresholds' Score	Test Sets' Score	AMT_CREDIT	Loan Status
0	100001	0.231985	0.463713	526.0	608.0	568800.0	Accepted
1	100005	0.231985	0.638471	526.0	660.0	222768.0	Accepted
2	100013	0.231985	0.228117	526.0	526.0	663264.0	Rejected
3	100028	0.231985	0.308561	526.0	556.0	1575000.0	Accepted
4	100038	0.231985	0.652965	526.0	664.0	625500.0	Accepted
5	100042	0.231985	0.311296	526.0	558.0	959688.0	Accepted
6	100057	0.231985	0.195946	526.0	512.0	499221.0	Rejected
7	100065	0.231985	0.530725	526.0	627.0	180000.0	Accepted
8	100066	0.231985	0.215306	526.0	519.0	364896.0	Rejected
9	100067	0.231985	0.680221	526.0	674.0	45000.0	Accepted
10	100074	0.231985	0.468593	526.0	608.0	675000.0	Accepted
11	100090	0.231985	0.422511	526.0	594.0	261621.0	Accepted
12	100091	0.231985	0.587694	526.0	644.0	296280.0	Accepted
13	100092	0.231985	0.660340	526.0	667.0	360000.0	Accepted
14	100106	0.231985	0.377728	526.0	580.0	157500.0	Accepted
15	100107	0.231985	0.411191	526.0	590.0	296280.0	Accepted
16	100109	0.231985	0.228688	526.0	526.0	407520.0	Rejected
17	100117	0.231985	0.144174	526.0	484.0	499221.0	Rejected
18	100128	0.231985	0.601328	526.0	649.0	431280.0	Accepted
19	100141	0.231985	0.230959	526.0	527.0	478498.5	Rejected

4. CONCLUSION – MONEY LOSSES AND SAVED

	Total Applicants	Total Accepted	Total Rejected	Acceptance Rate	Rejection Rate	Money Saved (IDR)	Money Losses (IDR)
0	48744	40900	7844	0.839078	0.160922	5.166947e+09	2.002105e+10

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

- If the machine predicts a True Positive (applicant is predicted to default and actually does default), the company stands to save approximately 5,000,000,000 IDR. Conversely, if a False Negative occurs (applicant is predicted not to default, but actually does), the company may lose approximately 20,000,000,000 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

The lower metrics can be attributed to the lack of Information Value (IV) between features. Additionally, there are several CSV files, such as

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

that 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.

THANK YOU!