Credit Risk modeling - ID/X Partners

By: Jihad Akbar

GitHub: https://github.com/jihadakbr/credit-risk-modeling

1. Business Understanding

As the final task of your internship as a Data Scientist at **ID/X Partners**, this time you will be involved in a project for a lending company. You will collaborate with various other departments in this project to provide technological solutions for the company. You are asked to build a model that can predict credit risk using a dataset provided by the company, which consists of loan data that has been accepted and rejected.

Additionally, you also need to prepare visual media to present the solution to the client. Make sure the visual media you create is clear, easy to read, and communicative. You can carry out this end-to-end solution development in your preferred programming language while adhering to the framework/methodology of Data Science.

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
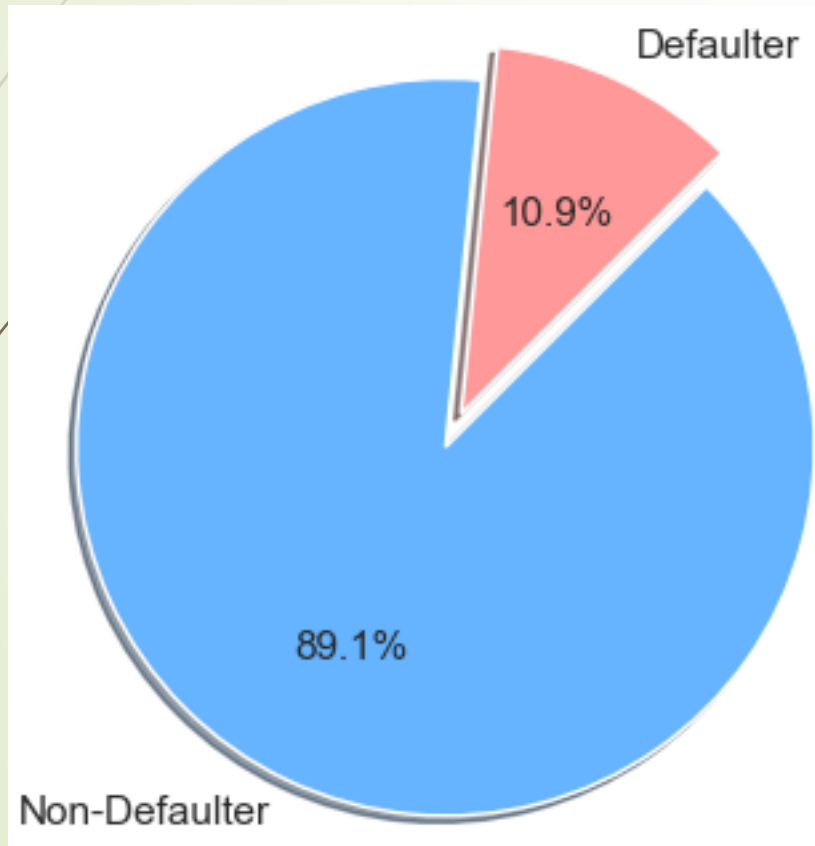
The target variables consist of 89.1% non-defaulters (accepted) and 10.9% defaulters (rejected).

3. Results – ML Metrics

Logistic regression was employed in a machine learning model, yielding the following metrics: threshold ≈ 0.22, accuracy ≈ 0.90, precision ≈ 0.93, recall ≈ 96.0, F1 ≈ 0.95, AUROC ≈ 0.84, Gini ≈ 0.67, and AUCPR ≈ 0.97.

These metrics are very good for credit risk modeling.

3. Results – Scorecard Development

	index	Feature name	Coefficients	Original feature name	Score - Calculation	Score - Preliminary	Difference	Score - Final
0	0	Intercept	-2.160175	Intercept	498.199433	498.0	-0.199433	498.0
1	1	grade:A	0.995542	grade	26.475784	26.0	-0.475784	26.0
2	2	grade:B	0.865267	grade	23.011224	23.0	-0.011224	23.0
3	3	grade:C	0.652893	grade	17.363255	17.0	-0.363255	17.0
4	4	grade:D	0.476261	grade	12.665859	13.0	0.334141	13.0
...	...	...	...	...	...	...	...	...
93	12	dti:>33.992	0.000000	dti	0.000000	0.0	0.000000	0.0
94	13	tot_cur_bal:>1597300.0	0.000000	tot_cur_bal	0.000000	0.0	0.000000	0.0
95	14	mths_since_earliest_cr_line:>524.5	0.000000	mths_since_earliest_cr_line	0.000000	0.0	0.000000	0.0
96	15	total_rev_hi_lim:>290106.667	0.000000	total_rev_hi_lim	0.000000	0.0	0.000000	0.0
97	16	total_rec_int:>17409.168	0.000000	total_rec_int	0.000000	0.0	0.000000	0.0

4. Conclusion – Money Losses & Saved

	Total Applicants	Total Accepted	Total Rejected	Acceptance Rate	Rejection Rate	TP+TN	FP+FN	Money Saved (USD)	Money Losses (USD)
0	93257	86520	6737	92.78%	7.22%	90.12%	9.88%	1.12e+09	9.05e+06

- 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 1,000,000,000 USD while incurring a loss of approximately 9,000,000 USD.

➡ The high or low percentages of True Positive/Negative and False Positive/Negative depend on the metrics of the machine learning model mentioned above.

THANK YOU!