Rakamin Academy

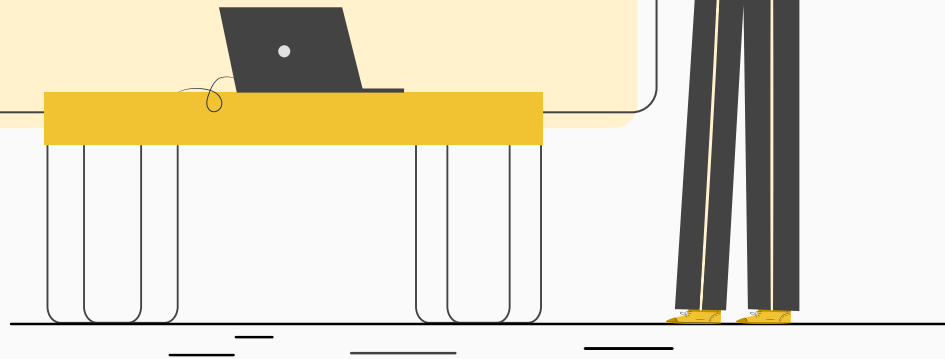
id/x partners

LOAN PREDICTION

Presented by
Salsabila Mardhiyah

BACKGROUND

As my final assignment as an internal Data Scientist at ID/X Partners, this time I will be involved in a project from a lending company. My objective is to build a model that can predict whether a loan considered good or bad using a dataset provided by the company.



DATA OVERVIEW

- Data is a credit history of the customers from a financial institution, from a personal loan marketplace that connects borrowers looking for credit with investors eager to lend money and earn a profit.
- Each borrower completes a thorough application, outlining their current financial situation, the reason for the loan, and other informations.

466K, 75

Data Size
(Rows, Cols)

53, 22

Numerical, Categorical
Features

DATA OVERVIEW

17
FEATURES DROPPED

Originally have 100% missing values.

14
FEATURES DROPPED

- Containing free text
- All values are unique
- High cardinality
- One unique value
- Redundant
- One dominant category
- Indicates leakage from the future data

NO
ROWS

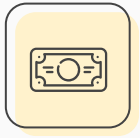
Originally have 100% duplicated values.

TARGET LABEL

CURRENT

FULLY PAID

CHARGED OFF



DEFAULT

DOES NOT MEET CREDIT POLICY.
STATUS: PAID OFF

DOES NOT MEET CREDIT POLICY.
STATUS: CHARGED OFF

IN GRACE PERIOD

LATE (31-120 DAYS)

LATE (16-30 DAYS)

Considerations:

- There is 9 different values in 'loan_status' feature.
- Our objective is building machine learning to classify whether a loan is bad and good using the final status of the loan. There is no certainty of the outcome of the ongoing loan.
- Therefore, Fully Paid and Charged-Off means the final outcome of a loan and will be encoded to 0 and 1.
- The other categories will be dropped.

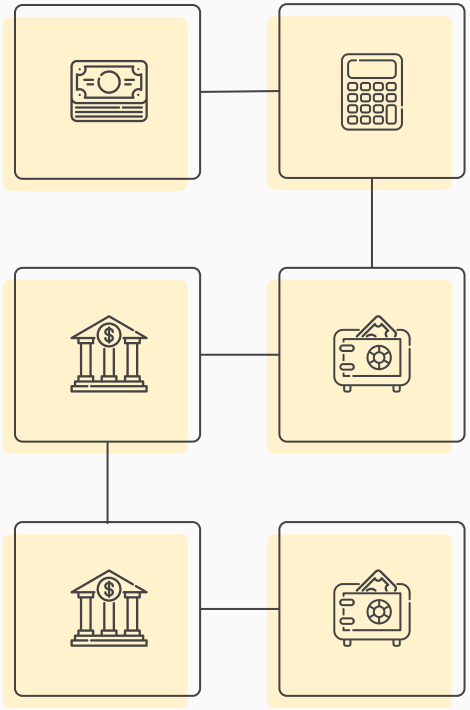
DATA PRE-PROCESSING

01 - MISSING VALUES

Handling missing values with the Imputation Method, which imputes the mode value for categorical data and the median value for numerical data.

02 - CATEGORIZATION

Grouping 5 continuous or numerical features into discrete bins or categories with Custom Binning Discretization Methods.



04 - ENCODING

Conduct Label Encoding to features that have ordinal values and One Hot Encoding to nominal features.

03 - TRANSFORMATION

Transform numerical data with Box-Cox, Logarithmic and Yeo Johnson Method to address skewness. Also scale it to 0-1 using MinMaxScaler.

05 - FEATURE SELECTION

Select important features with Spearman Correlation, ANOVA test, and Recursive Feature Elimination (RFE).

06 - IMBALANCE CLASS

Using one of oversampling technique SMOTE, where the synthetic samples are generated for the minority class.

DATA OVERVIEW AFTER PRE-PROCESSING

296K, 20

Train Data Size
(Rows, Cols)

50 : 50

Good:Bad Loan Ratio

148K : 148K

Good:Bad Loan Size

45K, 20

Test Data Size
(Rows, Cols)

80 : 20

Good:Bad Loan Ratio

37K : 8K

Good:Bad Loan Size

EVALUATION METRIC: FI-SCORE

RISK CONSIDERATION

Rejecting a creditworthy applicant (false negative) could lead to a loss of potential business and customer dissatisfaction. Approving a high-risk applicant (false positive) can result in financial losses due to loan defaults and increased credit risk.

OUR OBJECTIVE

The goal is to make reasonably accurate loan approval decisions while minimizing both false positives and false negatives and aim for a balanced approach between precision and recall. So the F1 score, could be a suitable evaluation metric.

MODEL EVALUATION

	F1 TRAIN	MEAN F1 CV	STD F1 CV
DECISION TREE	1.00	0.73	0.20
RANDOM FOREST	1.00	0.82	0.13
GRADIENT BOOSTING	0.81	0.74	0.28
ADABOOST	0.72	0.71	0.10

- First we build base model using oversampled training data and conducted cross-validation with 3 fold.
- We obtained F1-Score from training data, also mean F1-Score and standard deviation from cross validation as seen beside.
- High standard deviations in CV F1 scores may exhibit variability in performance across different data splits.
- Then we try to build two hyperparameter tuned models, Adaboost-which has lowest standard deviation showing more stable performance and Random Forest-which shows highest mean cross validation F1-Score.

MODEL EVALUATION

	BAD LOAN F1	GOOD LOAN F1	WEIGHTED AVG F1
RANDOM FOREST	0.40	0.76	0.70
ADABOOST	0.30	0.88	0.77

- A high F1-score for Good Loan indicates that the model is effective at correctly identifying loans that are more likely to be paid back on time.

- But this model also has low F1-Score Bad Loans suggesting that it struggles to correctly classify instances with Label 1. This label may be more challenging to predict accurately.

- The weighted average F1 score is a single metric that provides an overall measure of model performance across all classes, taking class imbalance into account (this is important because our data has severe class imbalance). It considers Label 0 and Label 1, which are weighted based on the number of instances in each class.

- The weighted average F1 score is a more comprehensive evaluation of model performance across an imbalance dataset. So we chose AdaBoost as our classifier because it shows better result on this metric.

ADABOOST RESULTS ANALYSIS: CONFUSION MATRIX

45K

Test Data Size

80 : 20

Good:Bad Loan Ratio

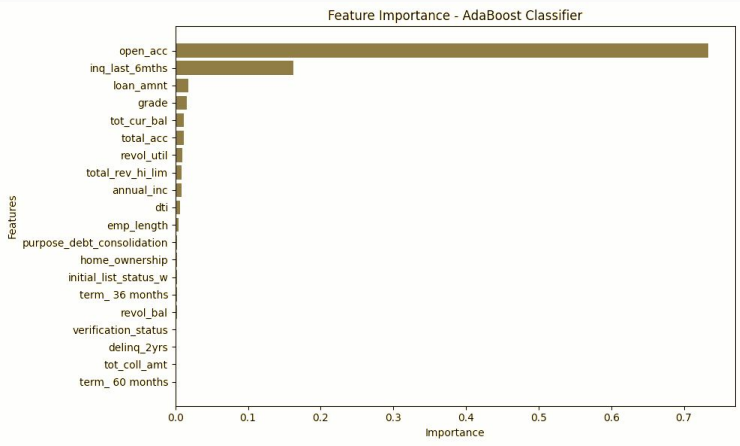
37K : 8K

Good:Bad Loan Size

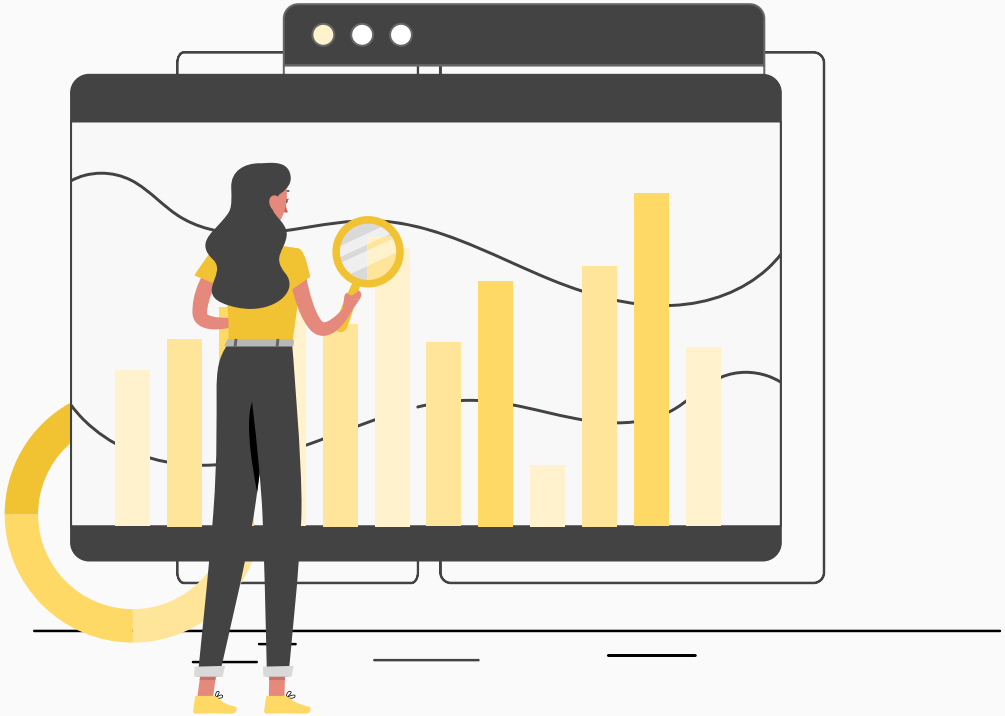
ADABOOST RESULTS ANALYSIS: FEATURE IMPORTANCE

`open_acc` (Open Credit Lines) feature represents the number of open credit lines (e.g., credit cards, loans) a borrower has. A higher number of open credit lines may indicate a borrower's creditworthiness and ability to manage multiple lines of credit.

'inq_last_6mths' (Inquiries in the Last 6 Months) signifies the number of credit inquiries made by the borrower in the last 6 months. Too many recent inquiries might suggest increased credit risk.



BUSINESS RECOMMENDATION



Establish clear lending criteria based on the importance of these features. Create guidelines for evaluating applicants' creditworthiness.

Use loan grades or credit scores to categorize applicants and determine interest rates.

Implement automated credit risk assessment systems that consider these features to make consistent and data-driven lending decisions.

Continuously monitor and improve lending criteria based on evolving market conditions and borrower behaviors.

THANK YOU

Passionate to uncover meaningful insights from data and make data-driven decisions. Recently completed a Data Science program, gaining proficiency in Python, SQL, and data visualization. Experienced in projects involving predictive modeling and data exploration to inform business decision-making.

salsabila.mardhiyah@gmail.com