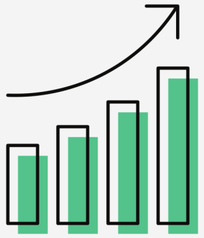
Credit Risk Prediction

Putu Arya Saputrawan

December 2022

Table of Contents

Background



Analyze & Modeling

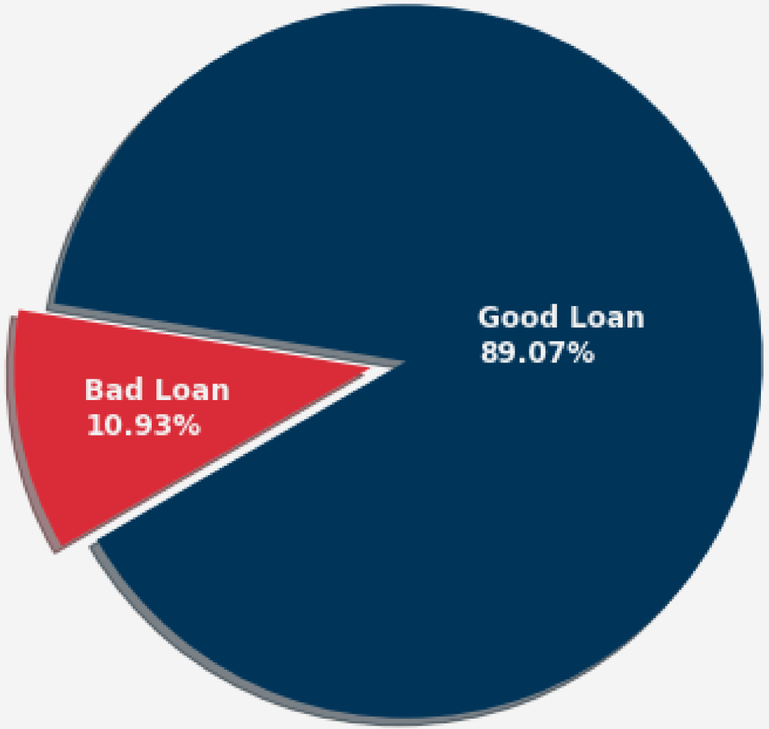
Results

Complete files for this project can be accessed here

Background

Problem

Our company has recorded a **Bad Loan Rate** of **10.93%**. This rate is above the tolerated rate **3.8%**. The high Bad Loan rate causes a company loss. Our company needs a **model/system** to help **mitigate** the bad loan rate.



Goal

Reduce bad loan rate at least 5% next year.

Objective

Make a machine learning model which able to predict that a loan will become a bad loan or good loan.

Analyze and Modeling

Dataset

The dataset that we use for this project is **loan record dataset**. The dataset consists of **466,285 rows** and **75 columns**. The description about every columns can be accessed here.

Exploratory Data Analysis

The purpose of this step is to understand about data **distribution** for each features. In this step also investigated about **data integrity** and **correlation** between features.

Preprocessing

Useless Features

30 useless features are dropped. 45 columns remaining

Feature Extraction

Extracting values from existing features. Total of 8 features are extracted

Missing Values

Dropping 4 features with more than 40% missing values. Replacing missing values for other features with 0 or the medians.

Feature Encoding

Applying Weight of Evidence encoding for categorical features.

Preprocessing

Feature Selection

Dropping features with high correlation wit other features. The dropped features is the features with smaller informational value.

Feature Transformation

Yeo Johnson transformation for numerical features. Then, all features are scaled by MinMaxScaler

Imbalance Handling

Oversampling by SMOTE method with a sampling strategy of 0.5

Modeling

Logistic Regression

Max. Iteration = 500

The table beside shows that the trained Logistic Regression model has a good performance (AUC > 85%). The model is good enough to be implemented for business.

Model Performance		
Metric	Train Set	Test Set
Accuracy	0.94	0.97
Precison	0.97	0.89
Recall	0.84	0.84
AUC*	0.96	0.96

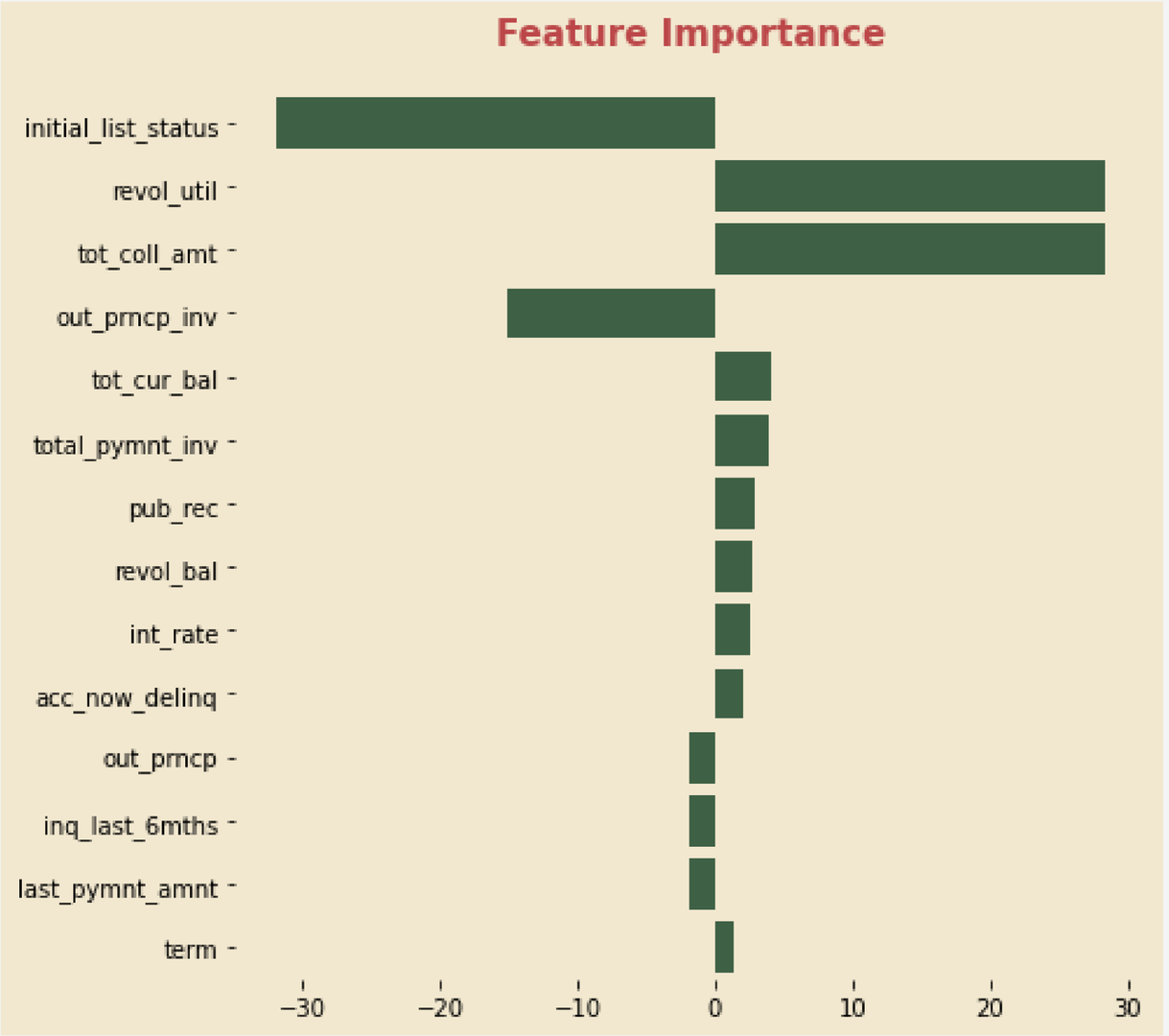
***Since the test set is imbalance, we will use the AUC score as main metric**

Results

Prediction Results

The prediction results above shows that the model predicted correctly about **84%** of all actual **bad loans**.

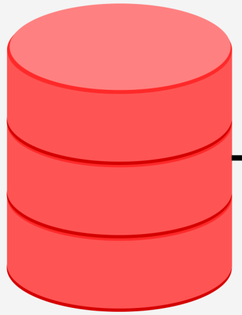
Feature Importance



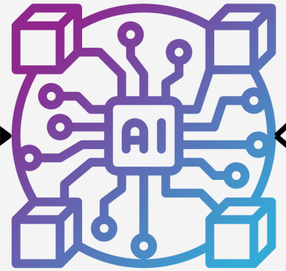
Feature Importance Interpretation

- The loan with initial status **'Whole (w)'** is more likely to become a bad loan
- The loan with **higher revolving utility** amount is more likely to become a good loan
- The loan with **higher collected** amount is more likely to become a good loan
- The loan with **higher outstanding principal** amount is more likely to become a bad loan

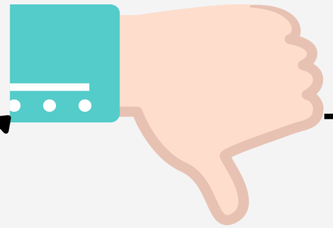
Model Simulation
(Workflow)



Loan Datasets



Model predict that the loan will be a bad or good loan



Predicted as Bad Loan

Take preventing actions

Predicted as Good Loan

Model Simulation
(Results)

If at least **60%** from **8642** of **correctly predicted bad loan** successfully **prevented**, then the **bad loan rate** will become **5,5%**. That means our company bad loan rate **reduced** about **5,43%**.

Conclusion

The trained Logistic Regression model for this project has a very good performance. The model has a AUC score of 96% that means the model can correctly predict 96% from total prediction. The model can correctly predicted about 84% of entire bad loans.

Thank you!