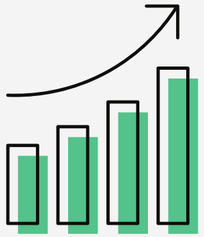
Credit Risk Prediction

Putu Arya Saputrawan

December 2022

Table of Contents

Background



Analyze & Modeling

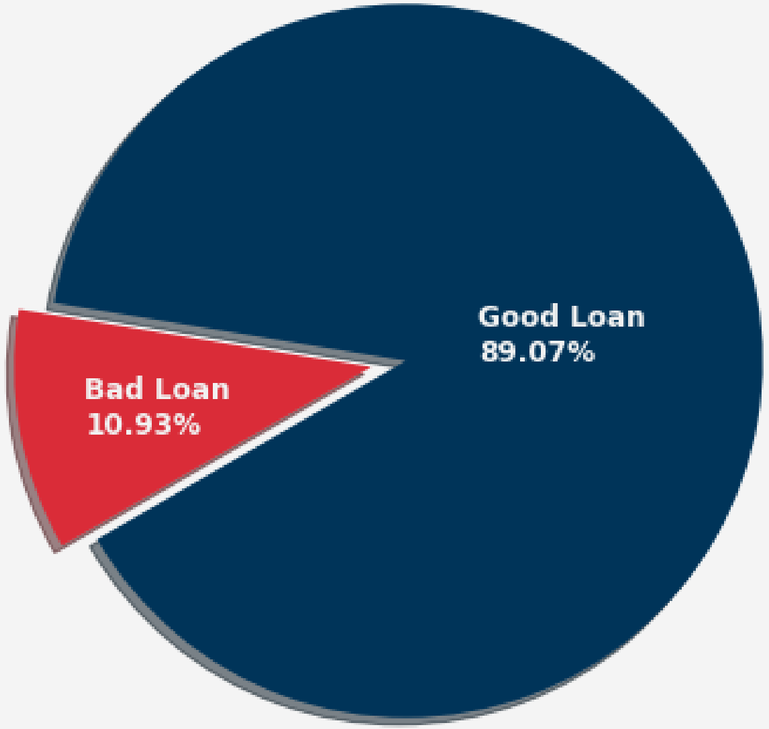
Results

Complete files for this project can be accessed here

Background

Problem

Our company has recorded a **Non Performing Loan Rate** of **10.93%**. This rate is high compared to the Asia average Bad Loan ratio, which is **5.11%**. The high Non Performing Loan rate causes a company loss. Our company needs a **model/system** to help **mitigate** the non performing loan rate.



Goal

Reduce non performing loan rate at least 5% next year.

Objective

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

Analyze and Modeling

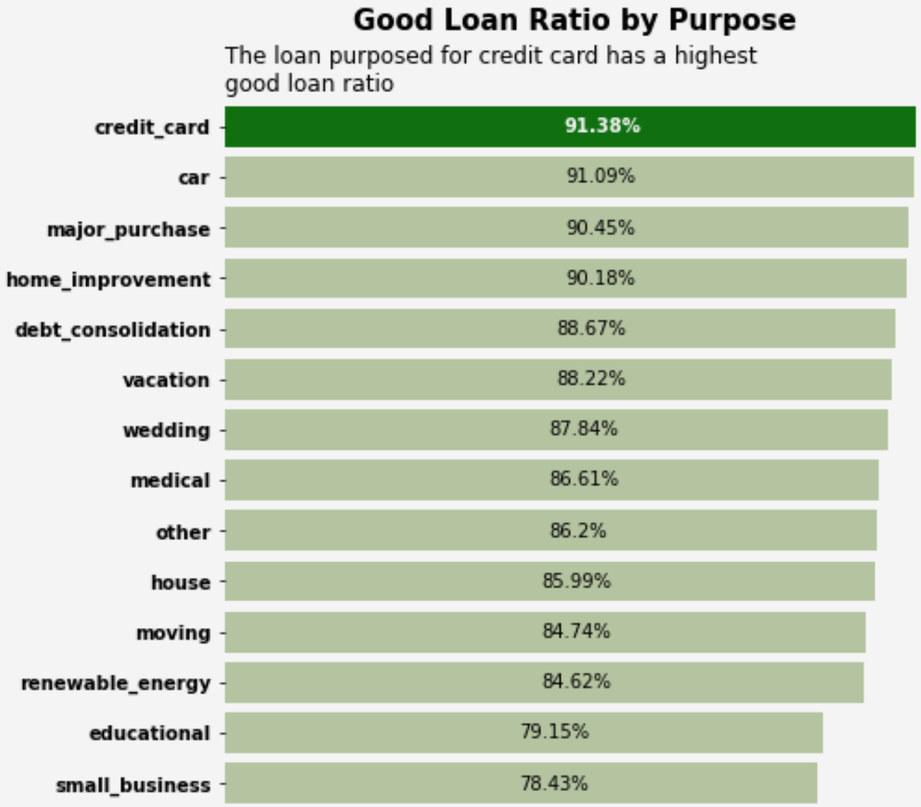
Dataset

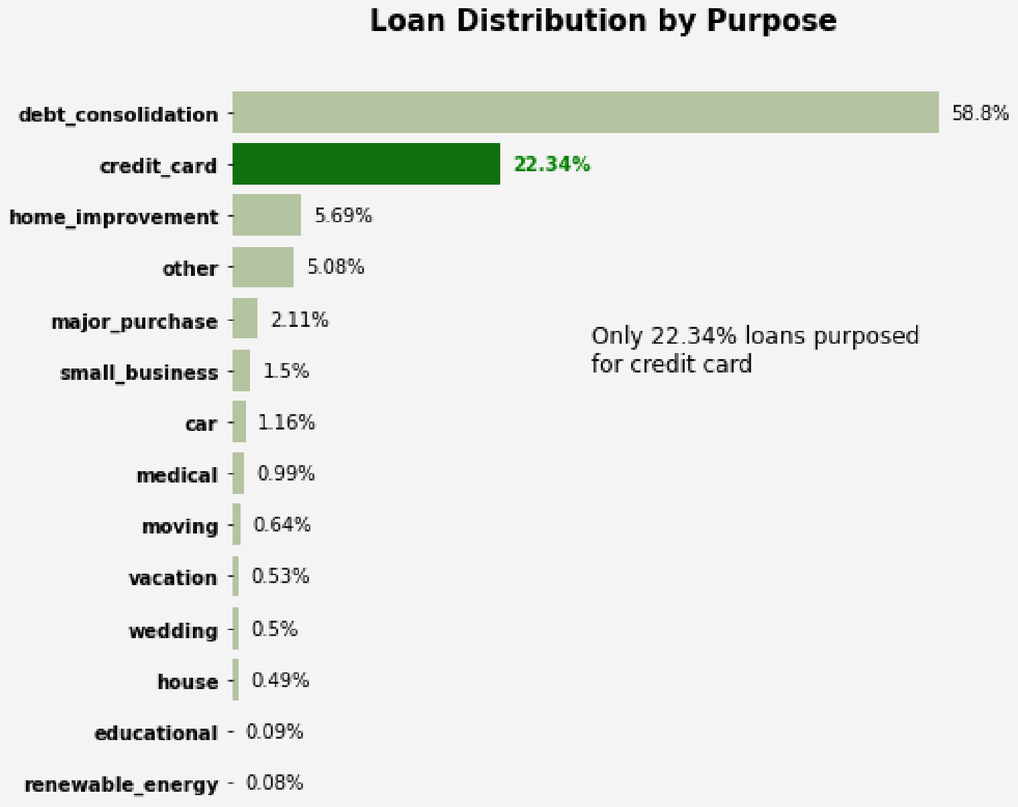
The dataset that we use for this project is **loan record dataset**. The dataset consists of **466,285 rows** and **75 columns**. The description about every columns can be accessed here.

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.

Business Insight





Recommendation : The loans purposed for credit cards has the highest good loan ratio (91.38%), but only 22.34% loans are purposed for credit cards. We need to attract more customers who are interested in applying loans for credit card.

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 too low or too high informational value. Features with high correlation to other features are also dropped.

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

For this project, I experimented with two machine learning models, namely Logistic Regression and Decision Tree Classifier. After evaluating their performance, I decided to implement the Decision Tree Classifier as it showed better results compared to Logistic Regression.

Performance Comparation (Test Set)		
Metric	Logistic Regression	Decision Tree
Accuracy	0.95	0.97
Precison	0.78	0.83
Recall	0.8	0.95
F1	0.79	0.89
AUC*	0.95	0.99

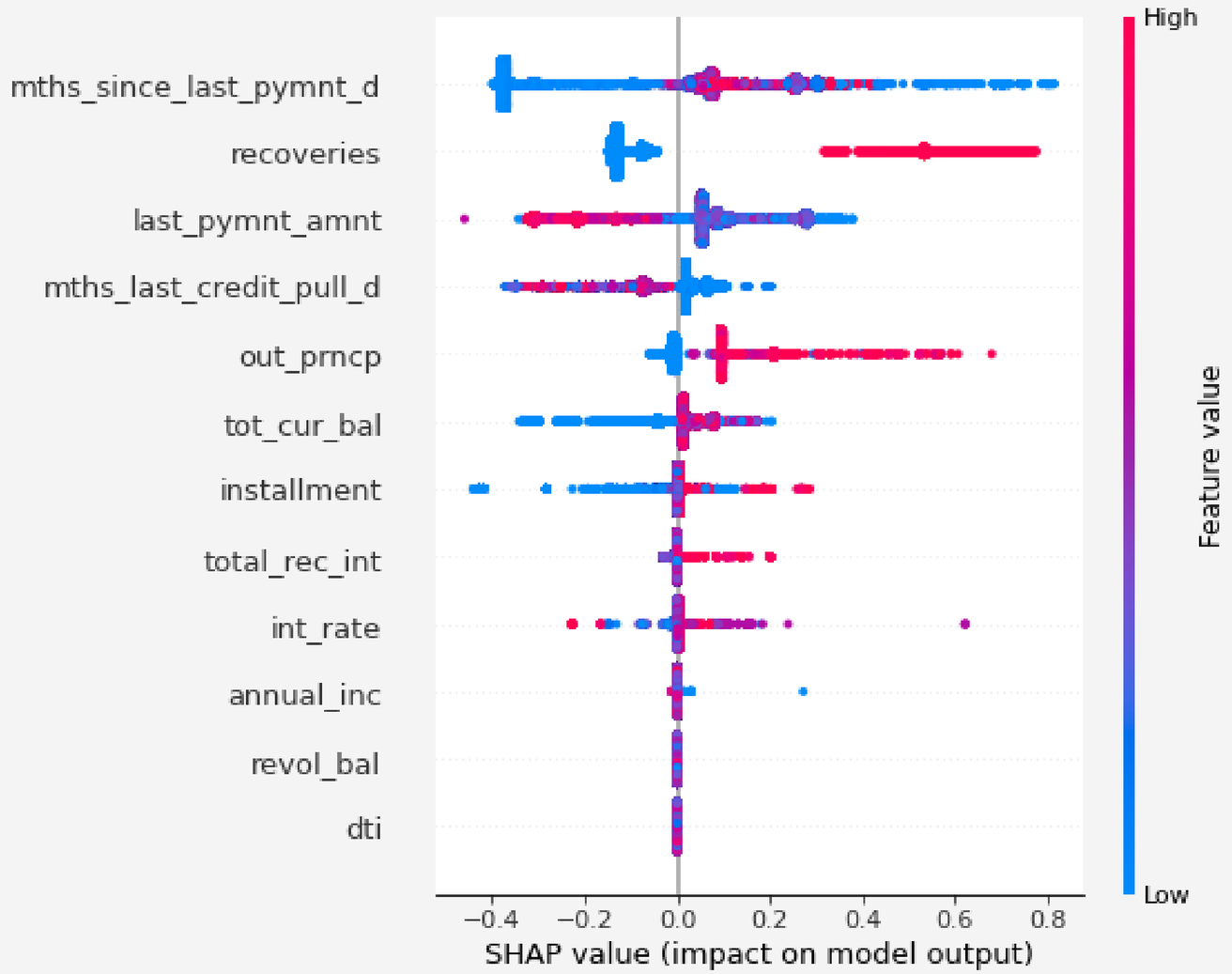
*Since the test set is imbalance, we will use the AUC score as main metric

Results

Prediction Results

The prediction results above shows that the model predicted correctly about **95%** of all actual **non performing loans**.

Feature Importance

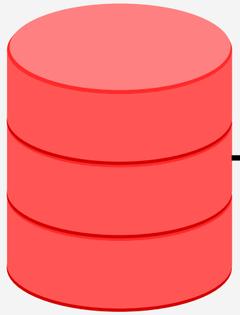


Recomendation Based on Feature Importance

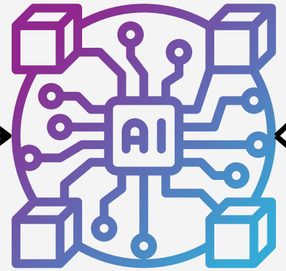
Be cautious of borrowers with:

- a long last payment date,
- a high recoveries amount,
- a low amount of last total payment received,
- a more recent last credit pull date, or
- a high remaining outstanding principal for the total amount funded

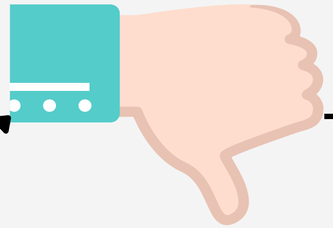
Model Simulation
(Workflow)



Loan Datasets



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



Predicted as Non Performing Loan

Take preventing actions

Predicted as Good Loan

Model Simulation
(Results)

If we can **successfully prevent** at least **65%** of the **9802** non performing loans that are **correctly predicted**, the non performing loan rate will **decrease to 4.2%**. This indicates a **reduction** of **6.73%** in our company's non performing loan rate.

Conclusion

The trained Decision Tree Classifier model for this project has a very good performance. The model has a AUC score of 99% that means the model can correctly predict 99% from total prediction. The model can correctly predicted about 95% of entire non performing loans.

Thank you!