BUILD IDEAL MODEL FOR PREDICTING LOAN STATUS CUSTOMERS

Aldy Budhi Iskandar

ALDY BUDHI ISKANDAR

Data Scientist & Machine Learning

Projects

Build Ideal Model for **Predicting** Loan Status Customers •
E-Commerce General Customers **Clustering** •
Video Game Sales **Exploratory Data Analysis** •
Shipping Data **Exploratory Data Analysis** •
Python **Data Cleansing** •

Educational

Data Science
May – Sept 2022 at **Dibimbing.id**

Mechanical Engineering
2012 - 2016 at **Polman Bandung**

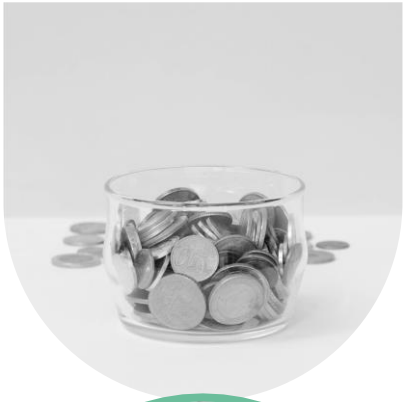
I'm a Data Scientist who interest in Math Logic and Statistics with a background as an Engineer in several manufacturing companies.

For programming skills, I prefer to use Python programming language, SQL query, & Tableau for visualization, because that's related to the software I always use for my projects.

CONTENTS

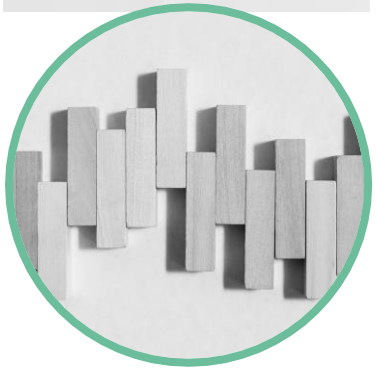
01

Background



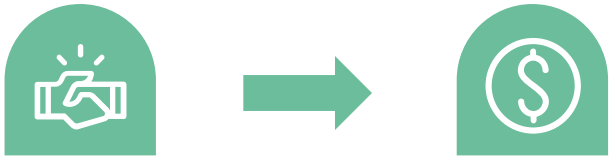
DATA SCIENCE

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.



LOAN INTRODUCTION

A loan is a financial product that allows a user to access a fixed amount of money at the outset of the transaction, with the condition that this amount, plus the agreed interest, be returned within a specified period. The loan is repaid in regular installments.



Banks keep the difference as a “Fee”.

CUSTOMERS LOAN STATUS

GOOD FLAG

Example status:
- Current,
- Fully Paid,
- In Grace Period,
- Late (16-30 days), etc.

BAD FLAG

Example status:
- Blacklist,
- Charged Off,
- Default,
- Late (31-120 days), etc.

BANK RISKS?

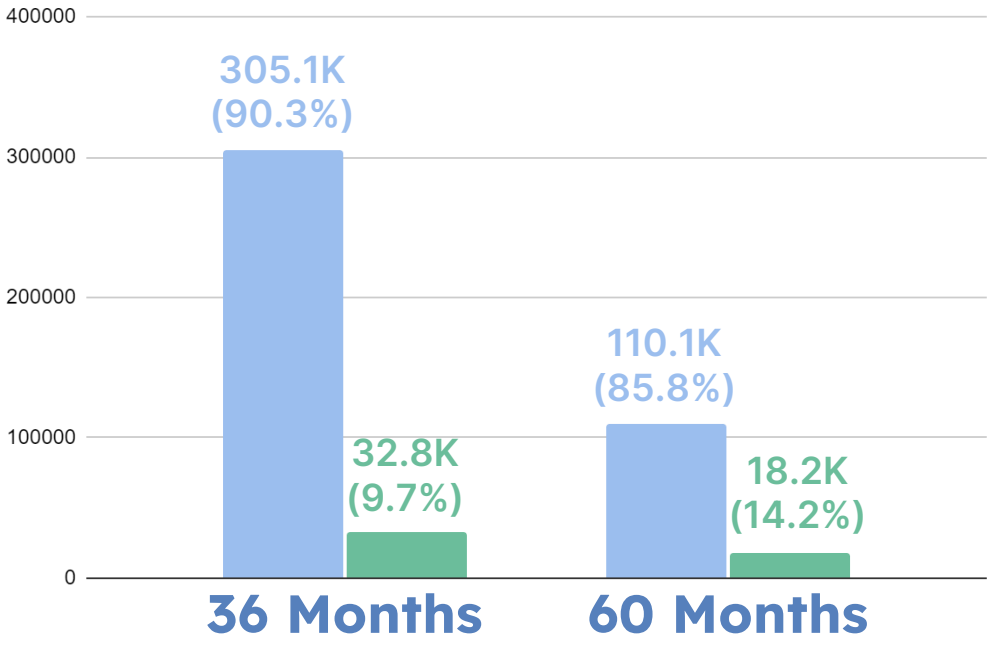
BAD CREDIT

The effectiveness of money will make bank limits reduced

AFFECTING THE COUNTRY'S ECONOMY

The bank will reduce the ability to automatically expand credit to the real sector.
So this will make it difficult for every industrial sector to borrow credit.

DATA ANALYSIS



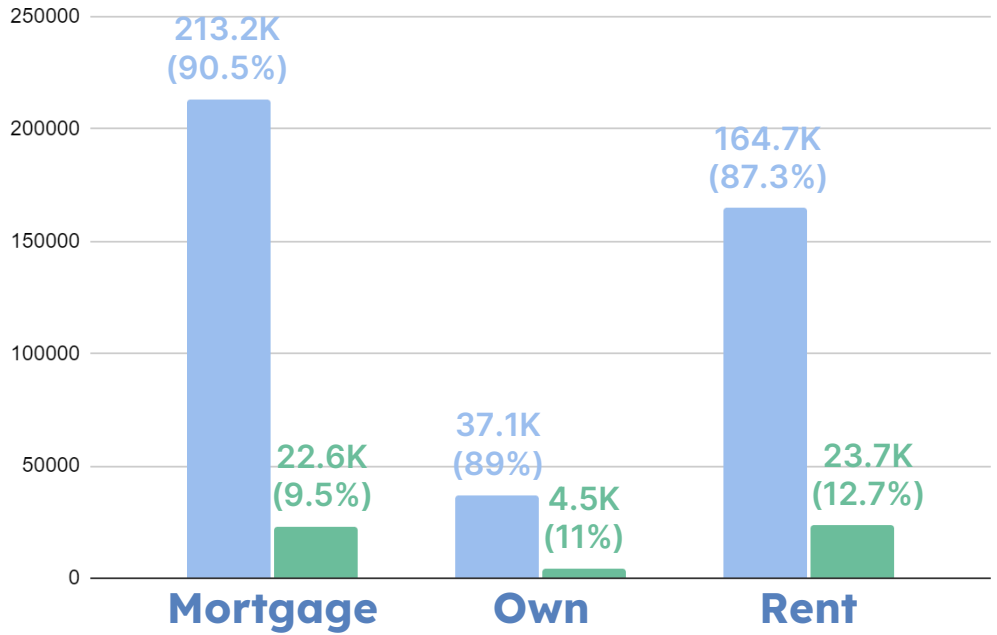
Bad customer

Good customer

Customers Status By Long-Term Customers Loan

The worst customer percentage is customers with 60 months term, which is 18.2K

DATA ANALYSIS



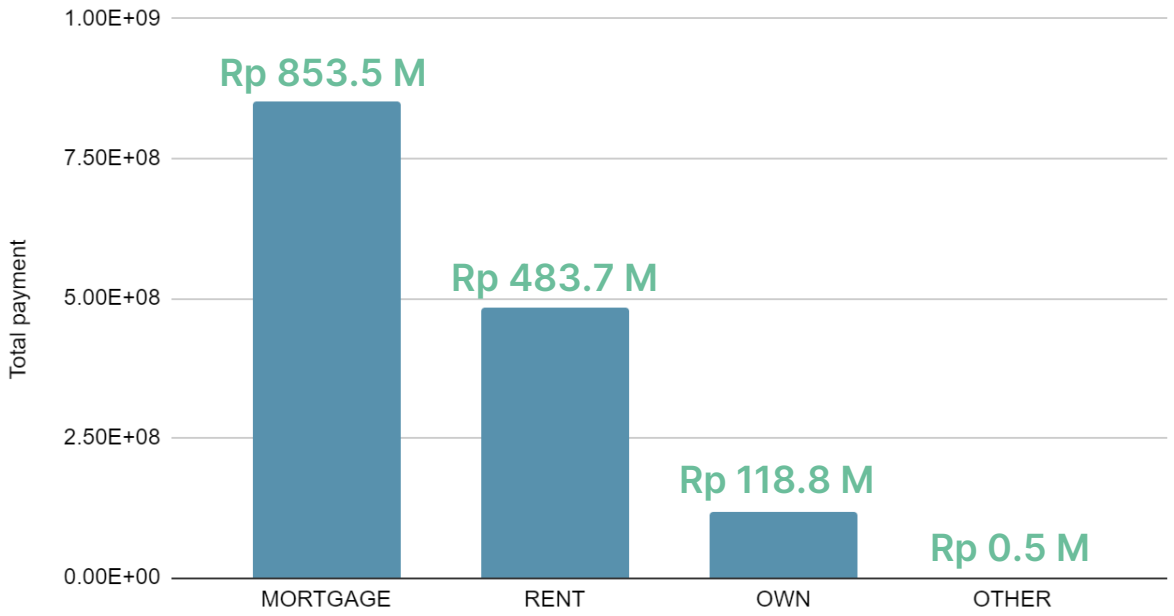
Bad customer

Good customer

Customers Status By Customer Home Ownership Status

Most customers who apply for loans are customers with mortgage home ownership status

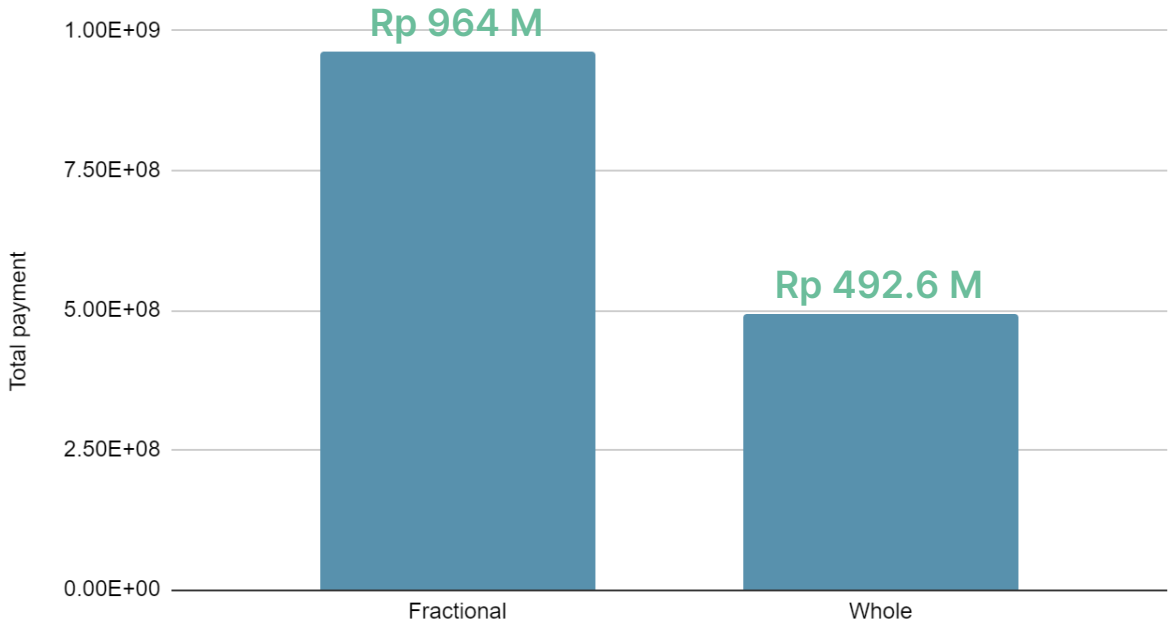
DATA ANALYSIS



Total Last Month Payment By Customers Home Ownership Status

The highest total payments came from customers with Mortgage home ownership status, which amounted to Rp 853.5 million

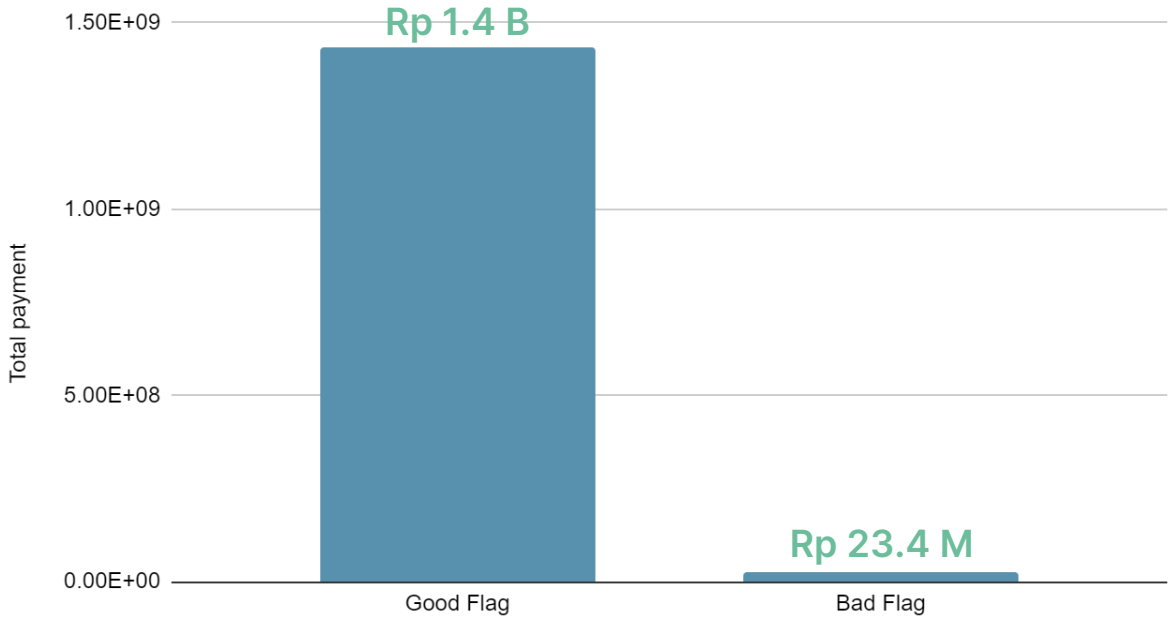
DATA ANALYSIS



Total Last Month Payment By Customers Initial Listing Status of The Loan

The highest total payments came from customers with Fractional type of payment, which amounted to Rp 964 million

DATA ANALYSIS



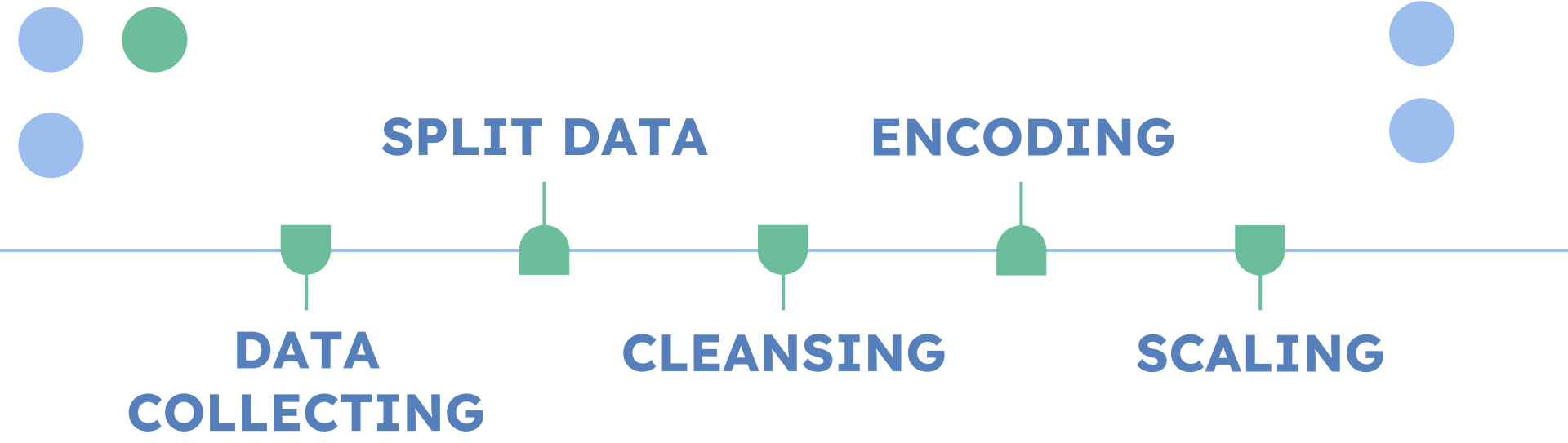
Total Last Month Payment By Customers Status

The highest total payments came from customers who have a good track record, which amounted to Rp 1.4 billion

02

Data PreProcessing

STEPS



DATA COLLECTING

466,285

466285 Customers data in one of banks in Indonesia

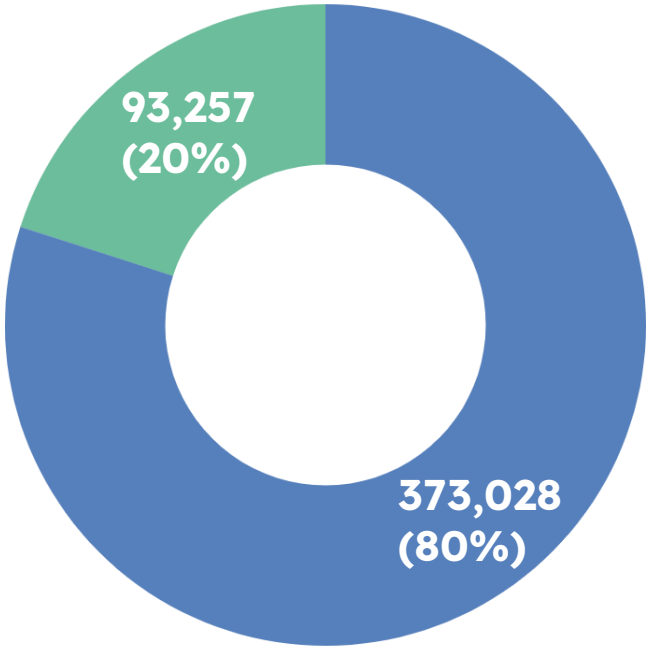
75

75 columns of data as Features in dataset

2007-2014

Customer data periode 2007 - 2014

SPLIT DATA



TRAIN

TEST

SPLIT DATA

Splitting data (80% train and 20% test) before data preparation is to avoid data leakage.

DATA CLEANSING

DATASET	MISSING VALUE	DUPLICATE DATA
DROP COLUMNS	48	0

From the number of columns that contain Null or Missing Values, column drop is carried out for columns that have a missing value of >20%.

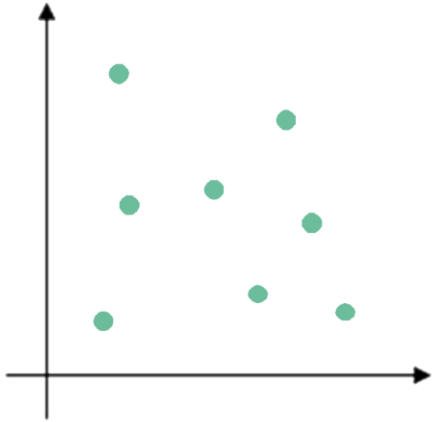
DATA ENCODING

VERIFICATION_STATUS	FREQUENCY
Not_verified	114806
Verified	211153

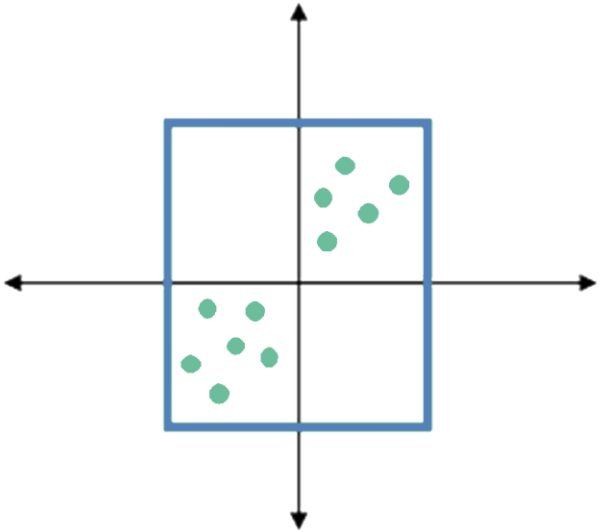
VERIFICATION_STATUS	FREQUENCY
0	114806
1	211153

Change the value from categorical to code / numerical

DATA SCALING



Actual Data

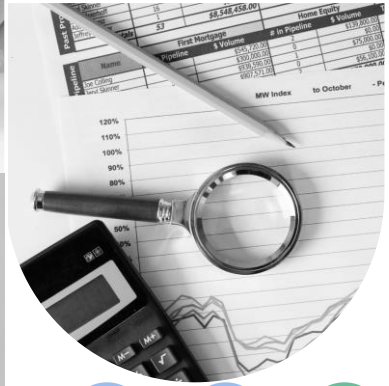


After Standardization

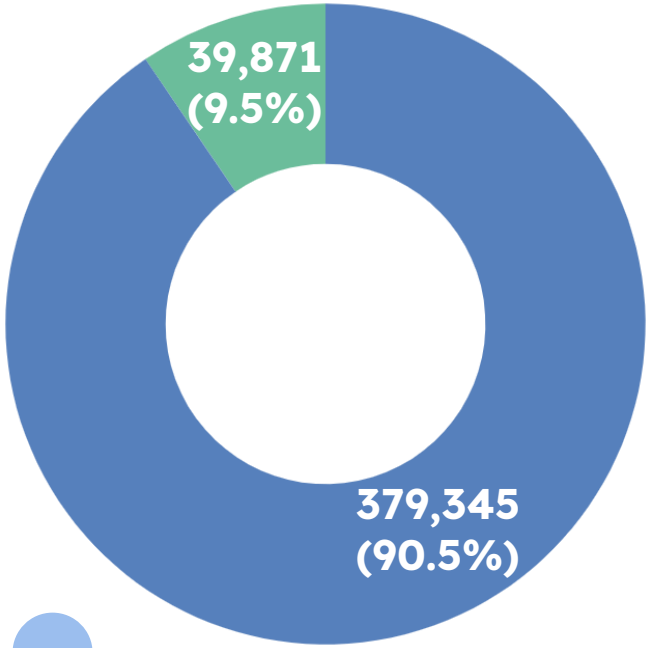
StandardScaler to normalize the data so that the data used does not have large deviations.

03

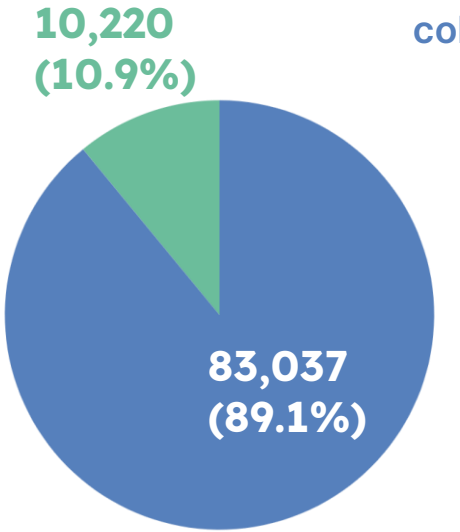
Modeling



SET TARGET



overall



y_test

is_bad_flag

Set `is_bad_flag` column as a target

- 0 (False)
- 1 (True)

Because the distribution is not balanced, we use **PRECISION** for the assessment

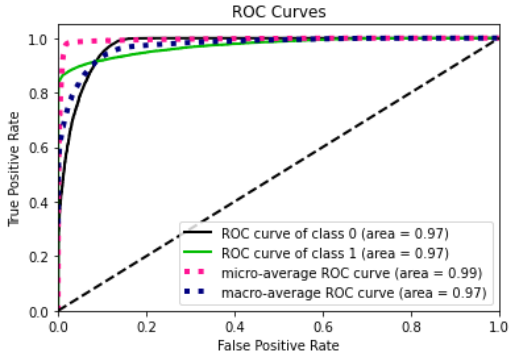
MODEL COMPARISON

MODEL	Accuracy	Recall	Precision	F1-Score
Logistic Regression	0.98	0.83	0.99	0.90
KNN	0.96	0.65	0.97	0.78
XGBoost	0.98	0.82	1.00	0.90
Decision Tree	0.97	0.88	0.87	0.88
Naive Bayes	0.94	0.51	0.93	0.66
Random Forest	0.98	0.86	0.99	0.92

All models use Baseline Modeling. No need to add Hyperparameter tuning.

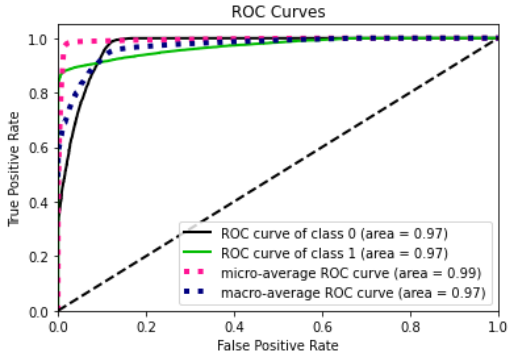
ROC CURVES

Logistic Regression



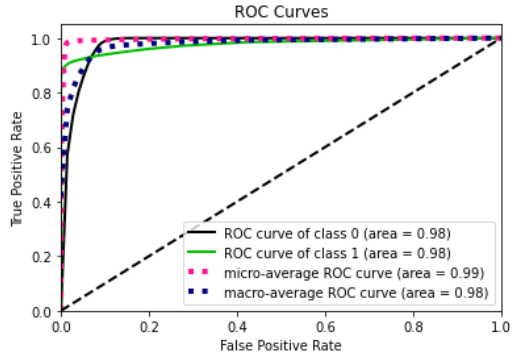
97%

XGBoost



97%

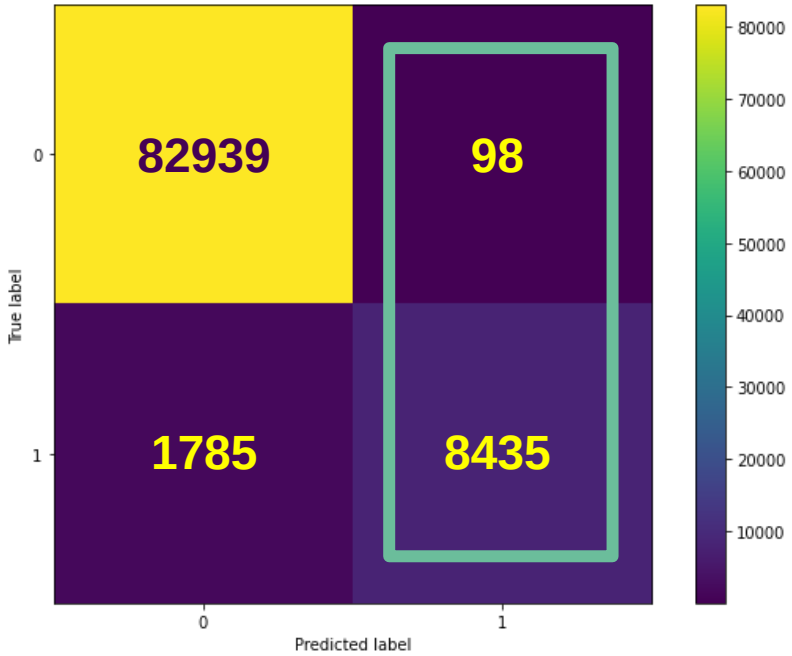
Random Forest



98%

The level of accuracy of the model that can distinguish between customers is bad_flag or good_flag

LOGISTIC REGRESSION



Precision

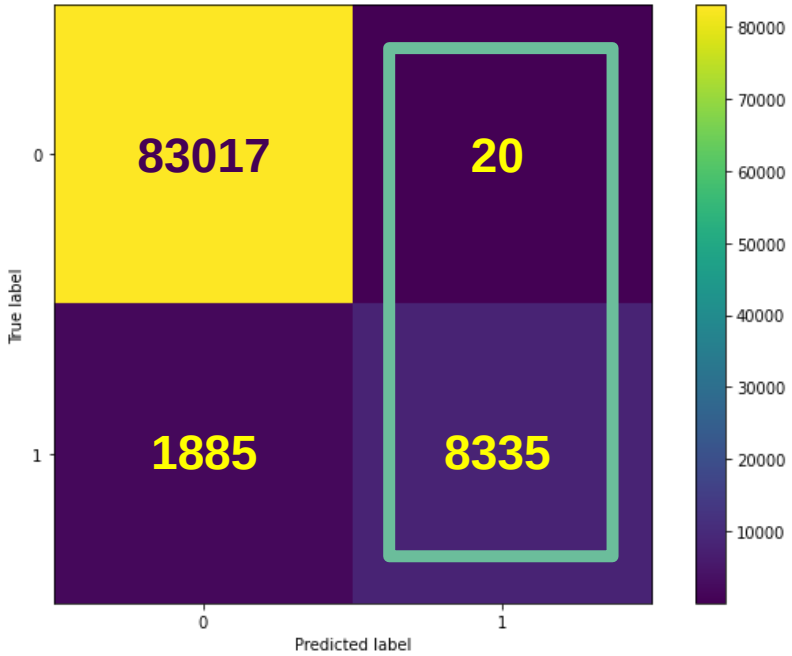
The assumption is that out of **8,533** customers who are predicted to be “bad” customers, there are **98** customers whose predictions turn out to be wrong (in fact, are “good” customers). So this model correctly predicts **8,435** customers who have a bad track record.

ROC

The assumption of the accuracy of the model predicts **8,435** customers who have a bad track record is 97% or as many as **8,182** customers.

In other words, from **8,533** customers, this model can accurately predict the minimum number of bad customers as many as **8,182**, and the maximum of **8,435** customers.

XGBOOST



Precision

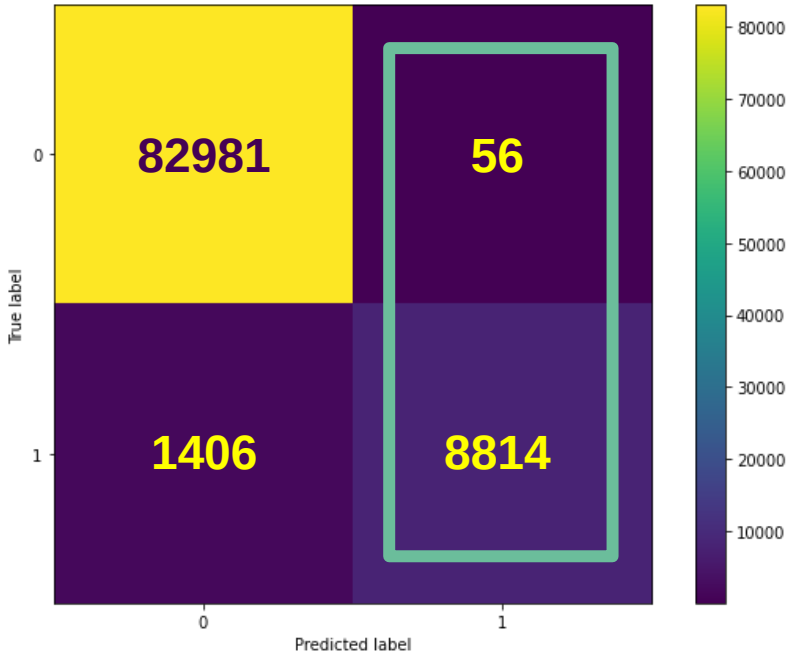
The assumption is that out of **8,355** customers who are predicted to be “bad” customers, there are **20** customers whose predictions turn out to be wrong (in fact, are “good” customers). So this model correctly predicts **8,335** customers who have a bad track record.

ROC

The assumption of the accuracy of the model predicts **8,335** customers who have a bad track record is 97% or as many as **8,085** customers.

In other words, from **8,355** customers, this model can accurately predict the minimum number of bad customers as many as **8,085**, and the maximum of **8,335** customers.

RANDOM FOREST



Precision

The assumption is that out of **8,870** customers who are predicted to be “bad” customers, there are **56** customers whose predictions turn out to be wrong (in fact, are “good” customers). So this model correctly predicts **8,814** customers who have a bad track record.

ROC

The assumption of the accuracy of the model predicts **8,814** customers who have a bad track record is 98% or as many as **8,638** customers.

In other words, from **8,870** customers, this model can accurately predict the minimum number of bad customers as many as **8,638**, and the maximum of **8,814** customers.

IDEAL MODEL COMPARISON

IDEAL MODEL	Minimum True Positive	Maximum True Positive	False Positive
Logistic Regression	8182	8435	98
XGBoost	8085	8335	20
Random Forest	8638	8814	56

Assuming that each customer makes a loan of 10 Million IDR and the bank gets a fee of 5 Million IDR for each loan, then:

Logistic Regression → Bank saves 81.82 – 84.35 Billion IDR & loss of 0.49 Billion IDR.
XGBoost → Bank saves 80.85 – 83.35 Billion IDR & loss of 0.1 Billion IDR.
Random Forest → Bank saves 86.38 – 88.14 Billion IDR & loss of 0.28 Billion IDR.

04

Evaluation

IDEAL MODEL

Random Forest model (baseline).

Assuming each customer makes a loan of **10 million IDR**, this model has saved **86-88 Billion IDR** of bank money from customers labeled as bad customers.

BUSINESS INSIGHT

- With this model, it's hoped that It'll make it easier for the bank to predict whether the customer applying for a loan is a good customer or not, so that the bank has no difficulty in assessing customers manually.
- Because the model can predict good customers with high Precision, Recall, F1-Score, and AUC-ROC, the bank can also carry out promotions to offer loans back to customers who have a good track record after paying off their loans.

The less error in model prediction → **the more** bank money saved → **the more** capital to turn to the right customers and **the more profits** for the bank

REFERENCES

- https://towardsdatascience.com/hyperparameters-of-decision-trees-explained-with-visualizations-1a6ef2f67edf#:~:text=The%20importance%20of%20hyperparameters%20in,forests%2C%20GBDT%2C%20and%20XGBOOST.
- https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/
- https://medium.com/@adiptamartulandi/tuning-hyperparameters-logistic-regression-menggunakan-grid-search-ucupstory-fb1ab9db082a
- https://accurate.id/ekonomi-keuangan/kredit-macet/
- https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f
- https://www.freecodecamp.org/news/how-naive-bayes-classifiers-work/
- https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/#:~:text=XGBoost%20is%20an%20implementation%20of,first%20XGBoost%20model%20in%20Python.

THANK YOU!

Have any questions?
aldybudhi003@gmail.com
+62 896 5729 1021
www.linkedin.com/in/aldybudhi/
www.github.com/qodym

Let's Check out my python code Jupyter notebook!
Don't hesitate to contact me if you want to do some corrections or discuss!

#DataScience #ClassificationModeling