Home Credit Indonesia Data Scientist Project Based Internship Program

Presented By

Philipus Dima Wira Pratomo

For details, you can see the notebook here

Hello!

My name is Philipus Dima Wira Pratomo, you could called me **Wira,** basically, I am Graduated from Pertamina University majoring in petroleum engineering with a GPA of 3.78. Currently I want to shift career to data scientist by attending a data science bootcamp. where I am developing expertise using data visualization tools such as Tableau and Looker Studio, SQL and several Python libraries to create machine learning. My objective is to become a professional data scientist and leverage my skills to contribute to data-driven decision-making in diverse industries.

Link CV :
https://docs.google.com/document/d/10t8-_gz2XVB8_aUQKpcN0iWJfHH9Rjl9zufunLR0nYo/edit?usp=sharing

For details, you can see the notebook here

Find Me

Whatsapp

081398758991

Email

pwirapratomo@gmail.com

LinkedIn

https://www.linkedin.com/in/philipus-dima-wira-pratomo-a221391ab/

GitHub

https://github.com/pwirap

For details, you can see the notebook here

Agenda

01. Background and Objective

02. Scope of problem

03. Data Selection

04. Exploratory Data Analysis

05. Data Preprocessing

06. Selected Model

07. Business Recommendation

For details, you can see the notebook here

Background And Objective

The target variable delves into identifying clients experiencing payment challenges, particularly those with delayed payments surpassing a set threshold (X days) on at least one initial installment (Y) of a loan. Crucial for financial institutions, this variable aids in assessing individual creditworthiness and risk. The primary objective involves crafting a predictive model or strategy to accurately pinpoint clients prone to payment difficulties. Understanding the root causes behind initial late payments enables the company to diminish default risks, enhance customer retention, curtail associated costs, and refine decision-making processes. Analyzing historical data encompassing payment behaviors, loan attributes, economic indicators, and potential external factors will pave the way for tailored interventions, personalized customer support, and targeted programs aimed at preemptively addressing payment challenges, thereby benefiting both the company and its clientele.

For details, you can see the notebook here

Scope of Problem

TARGET	Total Kredit	Percentage
0	164,509,700,000	92%
1	13,399,180,000	8%

Highest customer late payments could make company loss 13,399,180,000 (8%)

need a **strategy** to minimize this, so that losses **decrease**

It is necessary to indicate at the beginning whether the customer has the potential to be late in making payments or not in order to carry out treatment steps for the customer

machine learning can speed up this process

For details, you can see the notebook here

Data Selection

SK_ID_CURR	ID yang menjadi karakterisitik dari tiap customer
NAME_CONTRACT_TYPE	ID yang menjadi karakterisitik dari tiap customer
AMT_CREDIT	ID yang menjadi karakterisitik dari tiap customer
TARGET	ID yang menjadi karakterisitik dari tiap customer
CNT_FAM_MEMBERS	ID yang menjadi karakterisitik dari tiap customer
CODE_GENDER	ID yang menjadi karakterisitik dari tiap customer
DAYS_BIRTH	ID yang menjadi karakterisitik dari tiap customer
FLAG_OWN_REALTY, FLAG_OWN_CAR, NAME_HOUSING_TYPE, NAME_INCOME_TYPE, AMT_INCOME_TOTAL, AMT_ANNUITY	ID yang menjadi karakterisitik dari tiap customer
DEF_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE	ID yang menjadi karakterisitik dari tiap customer

Data Selection con’t

OCCUPATION_TYPE	ID yang menjadi karakterisitik dari tiap customer
NAME_FAMILY_STATUS	ID yang menjadi karakterisitik dari tiap customer
DAYS_BIRTH	ID yang menjadi karakterisitik dari tiap customer

For details, you can see the notebook here

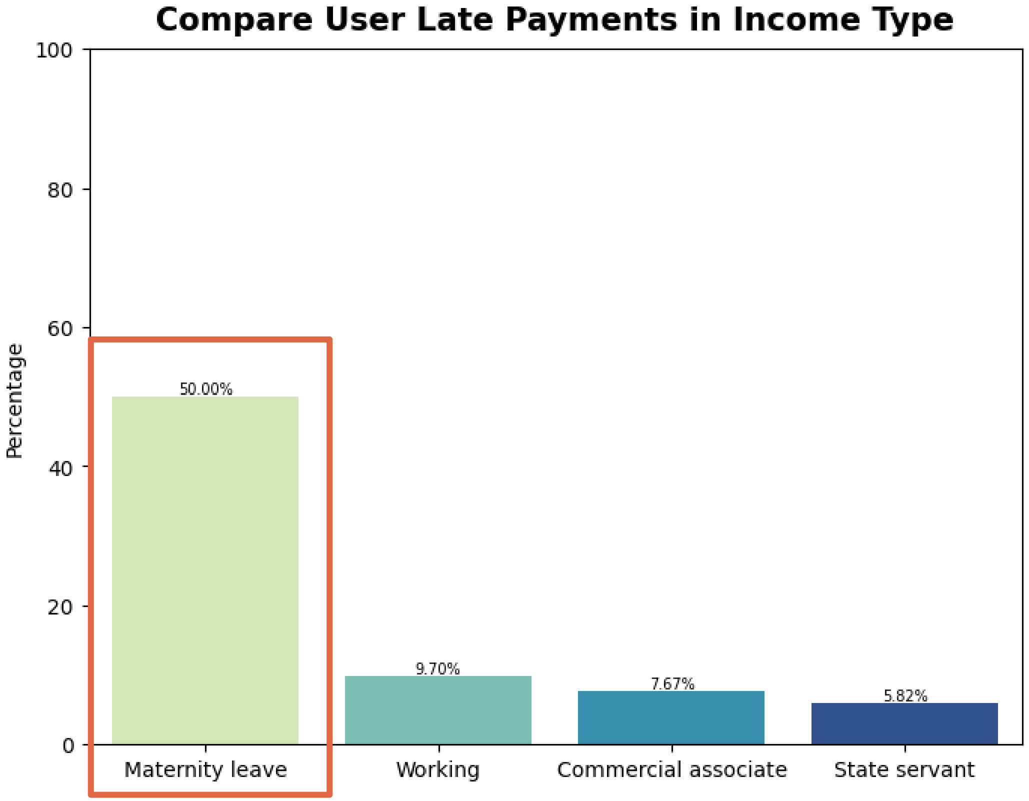
Exploratory Data Analysis

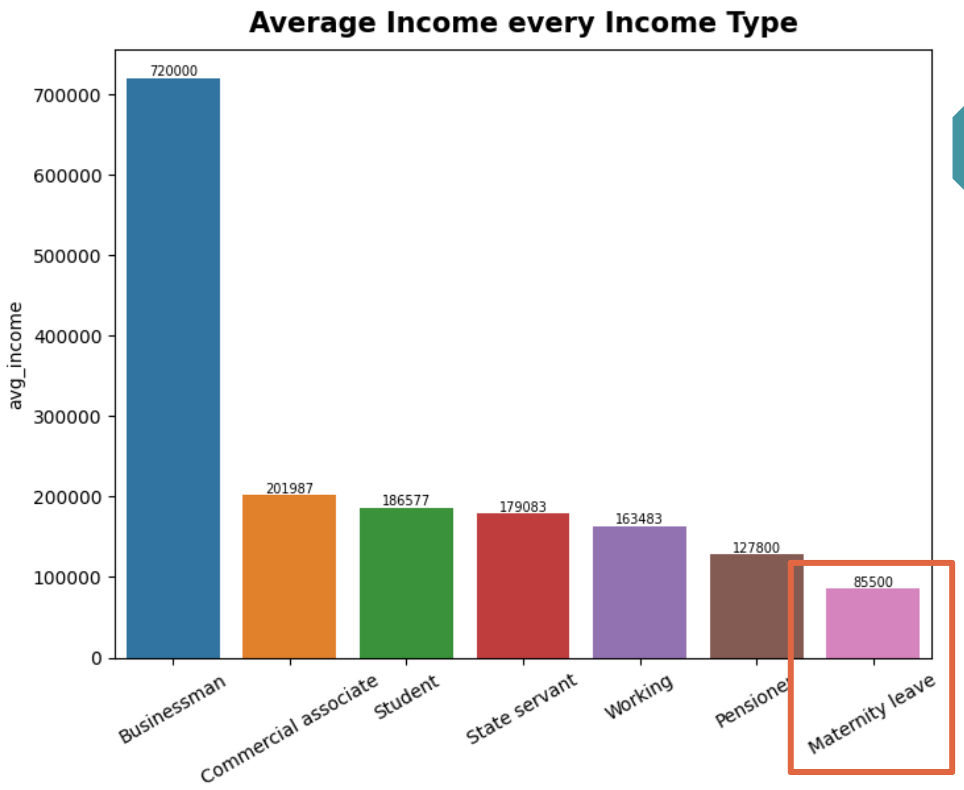
Compare User Late Payments in Income Type

Compare User Late Payments in Social Circles

For details, you can see the notebook here

Compare User Late Payments in Income Type



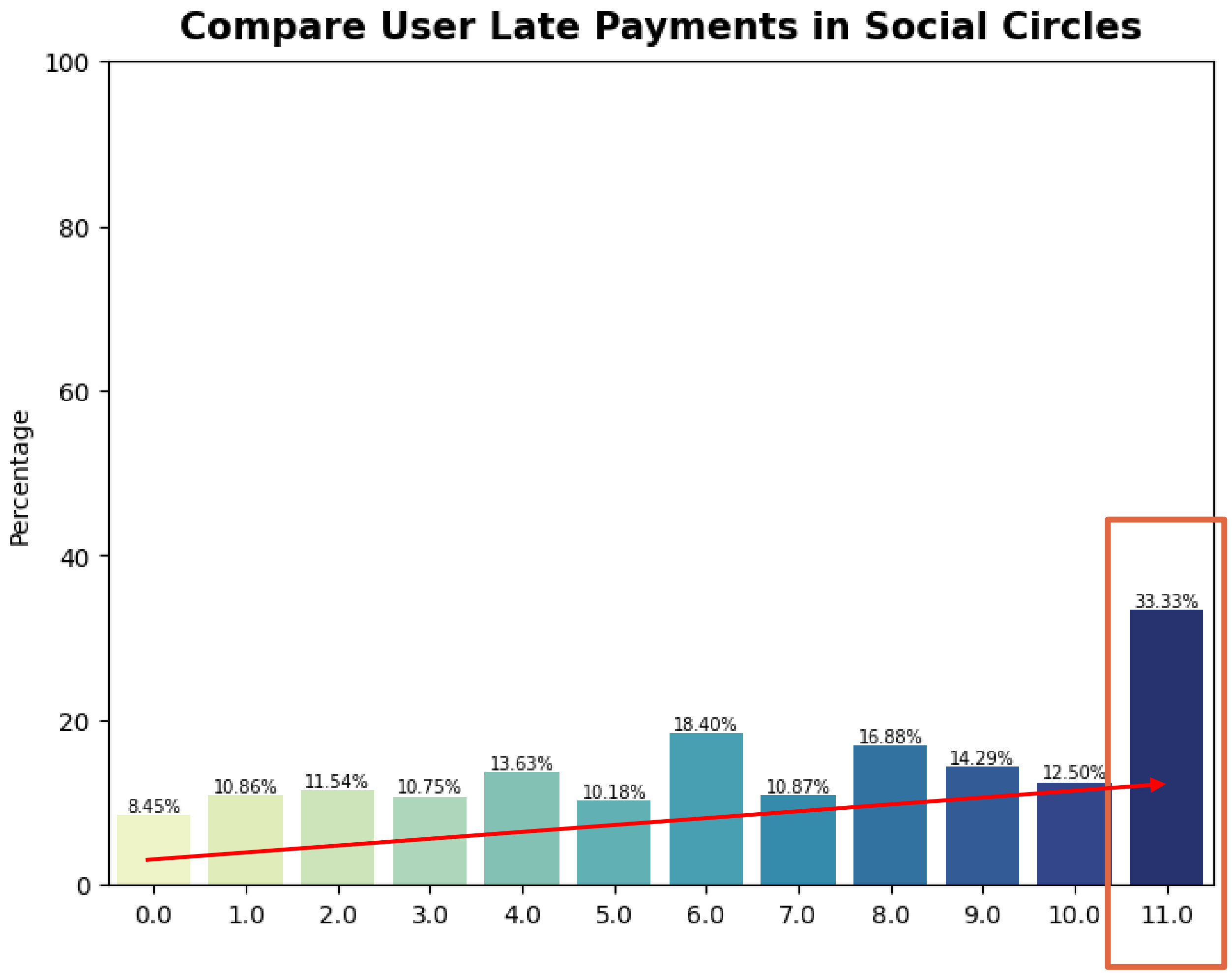


INSIGHT

- Maternity leave has the highest value of late payments, it has 50% customer that late payments
- If look at maternity leave, it has the lowest income compared to the others. Childbirth and pregnancy-related medical care can incur significant costs.

For details, you can see the notebook here

Compare User Late Payments in Social Circles



INSIGHT

- Social circles influence the culture in the environment whether payments are late or not
- It can be seen that, the more social circles are late in paying, the more customers are also late in making payments.
- Where it can be seen that there are 11 customers who are late paying, these customers also failed to pay

For details, you can see the notebook here

Data Preparation

Checking for Null Values, Data Types, Value in every Columns and Unique Value in ID	1. There are some null values in the OCCUPATION_TYPE,DEF_30_CNT_SOCIAL_CIRCLE,DEF_60_CNT_SOCIAL_CIRCLE 2. The SK_ID_CURR feature should be converted to a object data type 3. There is no duplicated data in SK_ID_CURR indicated that the customer is unique
Checking for Data Duplicates	1. There are no duplicate data

For details, you can see the notebook here

Data Preprocessing

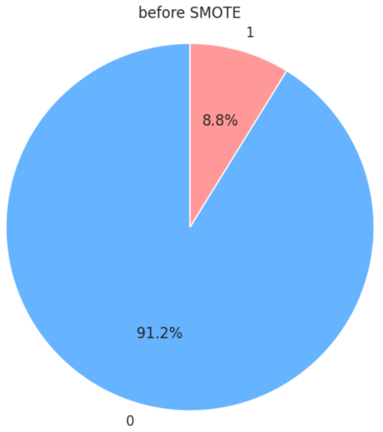
Handling Type Data	Perform in SK_ID_CURR
Handling Missing Value	Data with missing values will be taken out because there is still enough data
Feature Engineering	1. YEARS : calculate how old the customer is (DAYS_BIRTH/365) 2. TOTAL_LATE_PAYMENT_SOCIAL_CIRCLE : 3. calculate the total number of customers in the neighborhood who are late in making payments (DEF_30_CNT_SOCIAL_CIRCLE + DEF_60_CNT_SOCIAL_CIRCLE) **features that have been carried out feature engineering will be taken out**
Label Encoding	Label encoding will be carried out on the following features : 1. NAME_CONTRACT_TYPE 2. CODE_GENDER 3. FLAG_OWN_CAR 4. FLAG_OWN_REALTY 5. NAME_HOUSING_TYPE 6. NAME_FAMILY_STATUS 7. NAME_INCOME_TYPE 8. OCCUPATION_TYPE 9. NAME_EDUCATION_TYPE

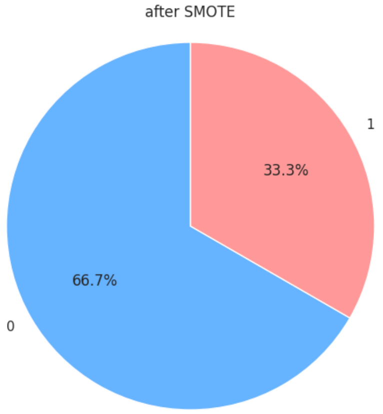
For details, you can see the notebook [here]

Preparing Data Train and Data Test

Class Imbalanced

because the differences in labels 1 and 0 are very different, class imbalanced will be carried out using SMOTE





Split Data

The data will be split, where 70% is data train and 30% is data test

Standardization

because some features, such as 'AMT_CREDIT','AMT_ANNUITY','AMT_INCOME_TOTAL','YEARS' are not normally distributed and have a long range, standardization will be carried out to increase accuracy during experiments using several machine learning models

For details, you can see the notebook here

Preparing Data Train and Data Test, con’t

Outlier Handling	Because there are some feature is having a big large data, Outlier handling in data Train must be doing

For details, you can see the notebook here

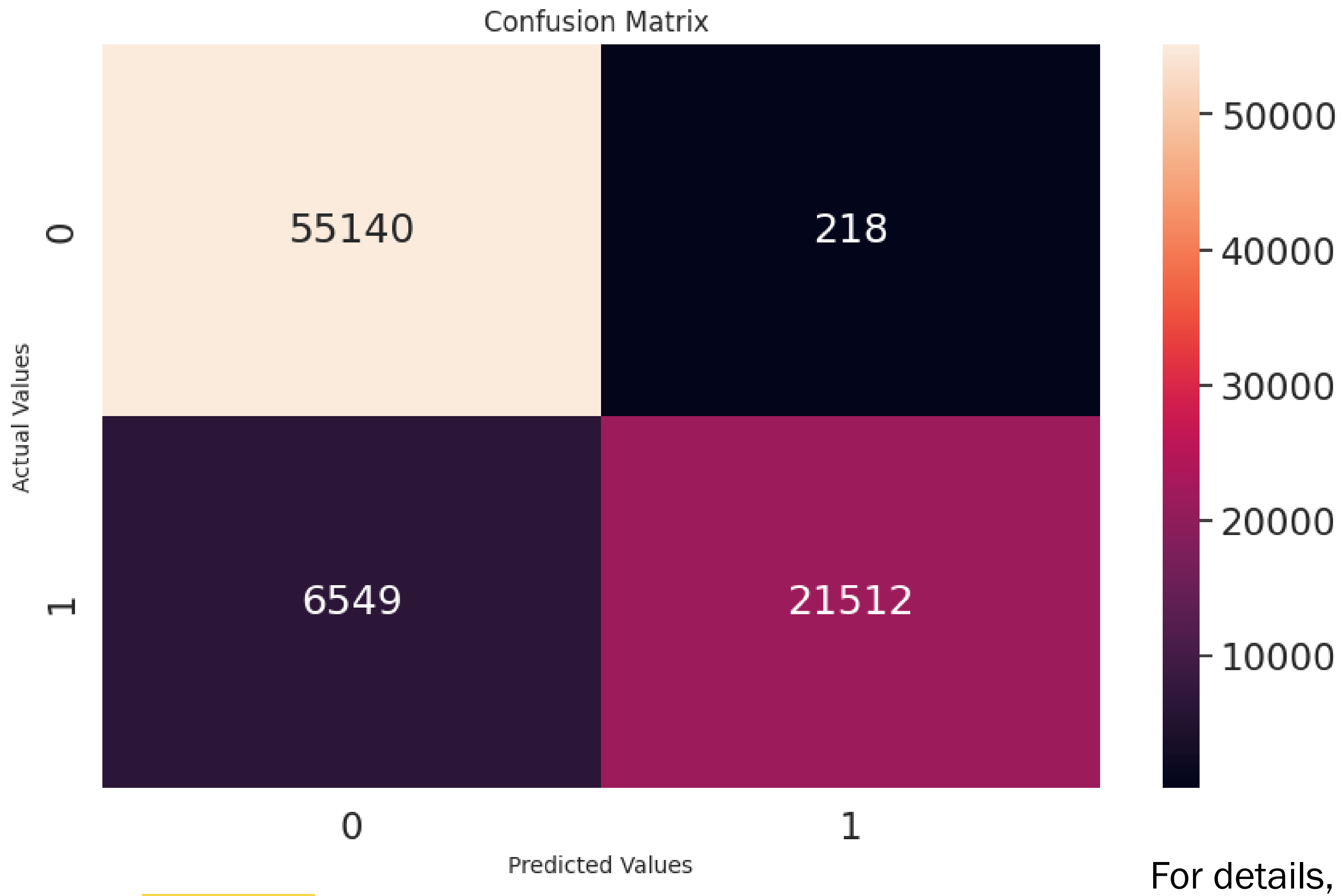
Selected Model

Model	Gradient Boosting	Random Forest	Decision Trees	XGBoost
Accuracy Train (%)	88.49%	99.99%	99.99%	92.23%
Accuracy Test (%)	88.35%	91.04%	84.89%	91.89%
Precision Train (%)	99.32%	100%	100%	99.31%
Precision Test (%)	99.41%	97.13%	76.37%	99%
Recall Train (%)	65.8%	99.97%	99.98%	77.14%
Recall Test (%)	65.77%	75.61%	79.77%	76.66%
F1 Score Train (%)	79.16%	99.98%	99.99%	86.83%
F1 Score Test (%)	79.16%	85.03%	78.03%	86.41%
ROC AUC Train (%)	82.79%	99.99%	99.99%	88.44%
ROC AUC Test (%)	82.79%	87.24%	83.63%	88.13%
CV Accuracy (%)	88.63%	90.67%	84.01%	91.89%
CV Precision (%)	99.37%	96.67%	74.81%	98.88%
CV Recall Test(%)	66.18%	74.47%	78.18%	76.45%
CV Recall Train(%)	66.22%	99.98%	99.98%	76.99%
CV F1 Score (%)	79.44%	84.13%	76.45%	86.23%
CV ROC AUC (%)	90.1%	92.39%	82.54%	92.45%

- Among the various machine learning models that have been explored, XGBoost has been selected as the top-performing model. Following this, hyperparameter tuning will be conducted after standardization to mitigate the risk of overfitting.
- because the existing data has empty values, and the data does not have a normal distribution, and has good performance in accuracy and recall which does not indicate over fitting, XGBoost is the machine learning model used

For details, you can see the notebook here

Confusion Matrix



For details, you can see the notebook here

Business Recommendation

Collaboration with Healthcare Providers or Insurers

- Consider collaborating with a health care provider or insurance company to provide specific information or solutions regarding the costs of medical care related to pregnancy and birth. This could be providing information about health insurance programs, affordable health services, or payment options related to medical care.

Improved Risk Evaluation

- Update the risk evaluation process to identify potential customers who are at risk of failure to pay, especially in areas that have the potential to make late payments

For details, you can see the notebook here

Business Simulation

About	Total Kredit
0	23,032,810,000
1	2,155,181,000

- By applying the machine learning model that has been created, it will be predicted that customers who are late in making payments, which are marked with (1), will carry out certain treatments, so that customers are not late in making their payments. With this machine learning model, company will get a profit of 2,155,181,000.

For details, you can see the notebook here