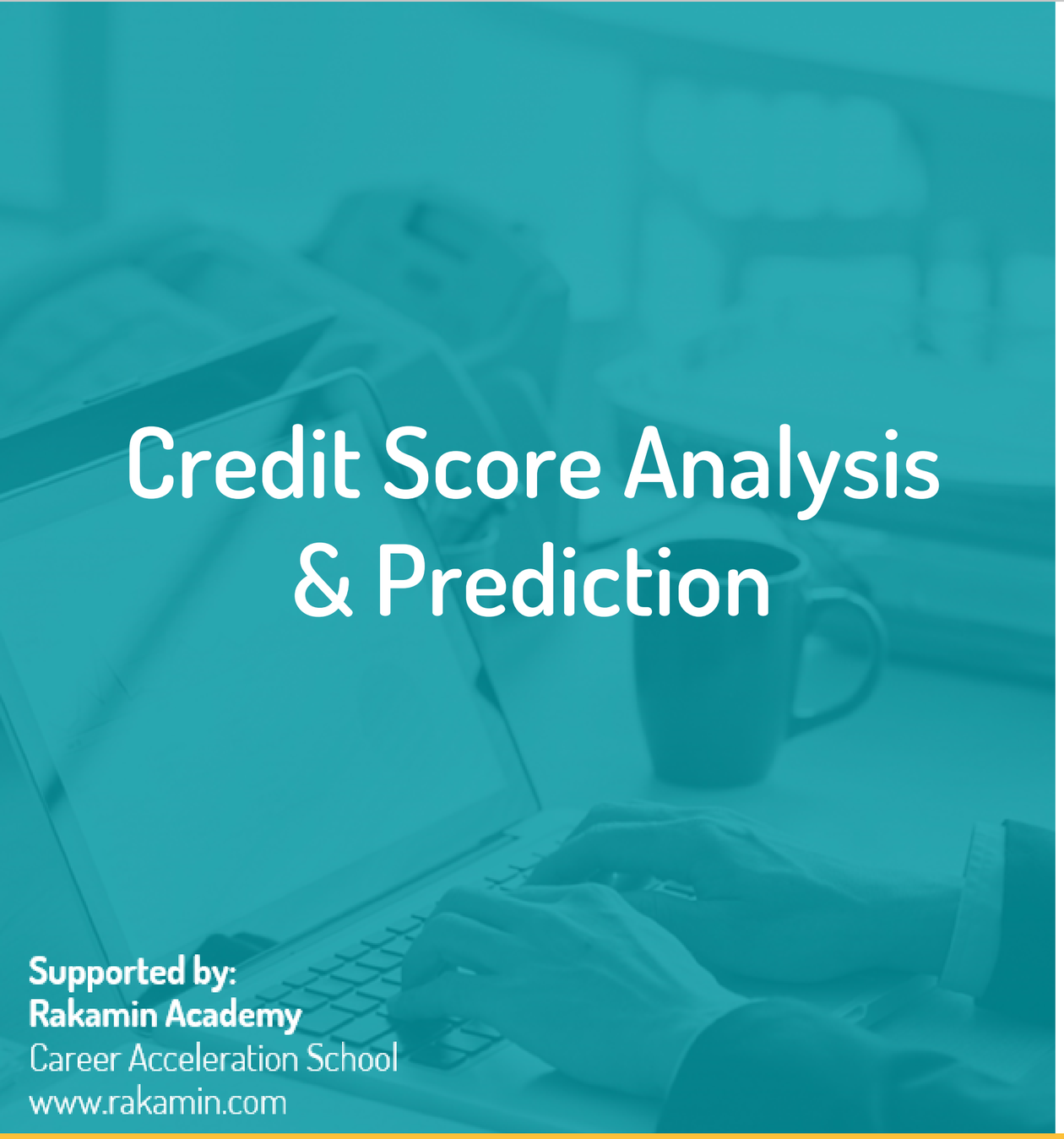


Credit Score Analysis & Prediction

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

HOME CREDIT

Home Credit Indonesia Data Scientist Virtual Internship Experience

Muhammad Randa Yandika

https://randayandika.github.io/
https://linkedin.com/in/muhammad-randa-yandika

Summary

Problem Statement:

- Home Credit Indonesia wants to accurately predict the repayment capability of customers based on their historical data. The objective is to develop a reliable model that can determine whether a customer is likely to repay a loan or not. By doing so, Home Credit Indonesia aims to minimize the risk of lending to customers who may default on their loans and ensure that credit is extended to deserving individuals who have the ability to repay.

Goal:

- Developing an accurate credit score prediction model to ensure that customers capable of repayment are not rejected when applying for a loan.

Objective:

- Identify the variables that have the most influence on credit scores.
- Build a predictive model for credit scores based on relevant variables.
- Test and evaluate the performance of the credit score prediction model using relevant metrics.
- Use the model to make informed decisions or recommendations related to creditworthiness and risk assessment.

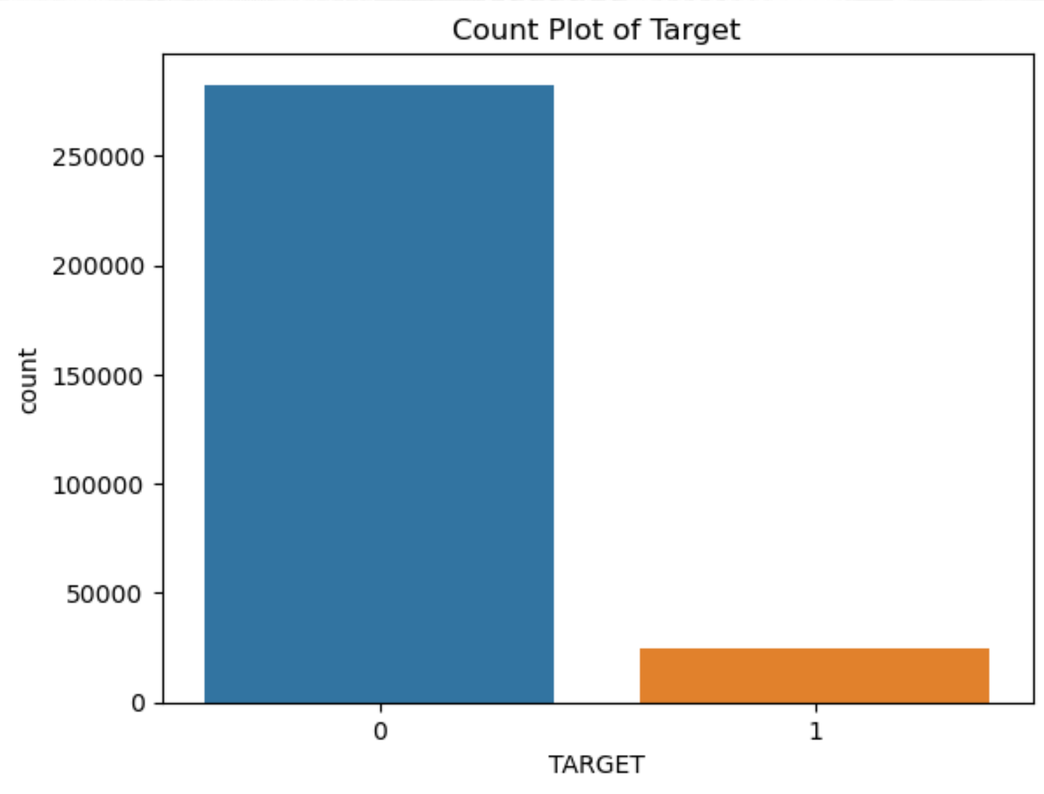
Business Understanding

- Credit Risk Prediction: In the lending business, the primary goal is to effectively manage credit risk. Using this dataset, businesses can develop accurate prediction models to determine the credit risk of customers. By analyzing variables such as employment history, total income, and other financial information, businesses can gain a deeper understanding of the factors that contribute to credit risk. This enables businesses to make better decisions in granting loans to customers who potentially have low credit risk and reduce the risk of loan defaults.

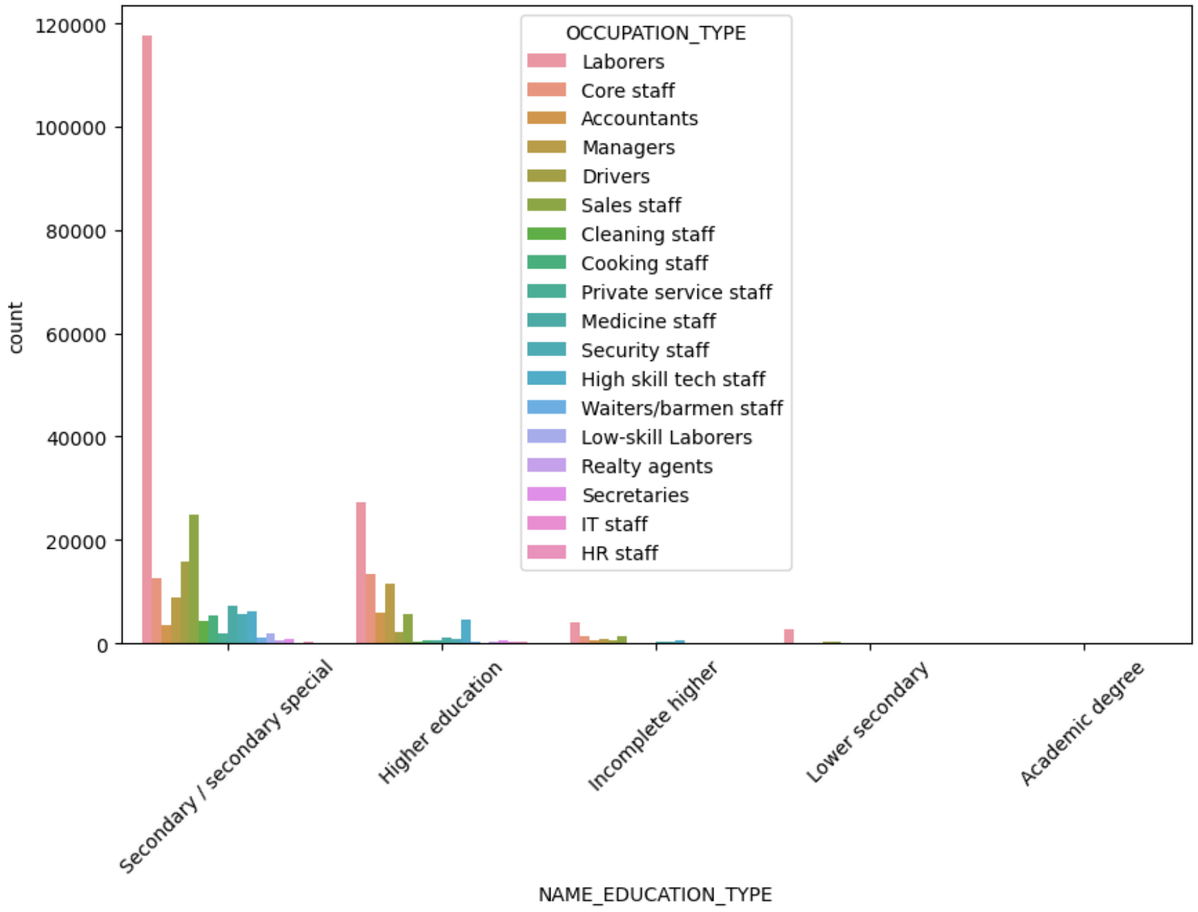
- Loan Approval Optimization: In the lending business, it is crucial to ensure that customers who are capable of repayment are not rejected when applying for a loan. Using this dataset, businesses can develop models to identify customers with good financial capability to repay loans. This enables businesses to expedite loan approval and reduce unnecessary rejections, thereby enhancing customer satisfaction and loan approval efficiency.

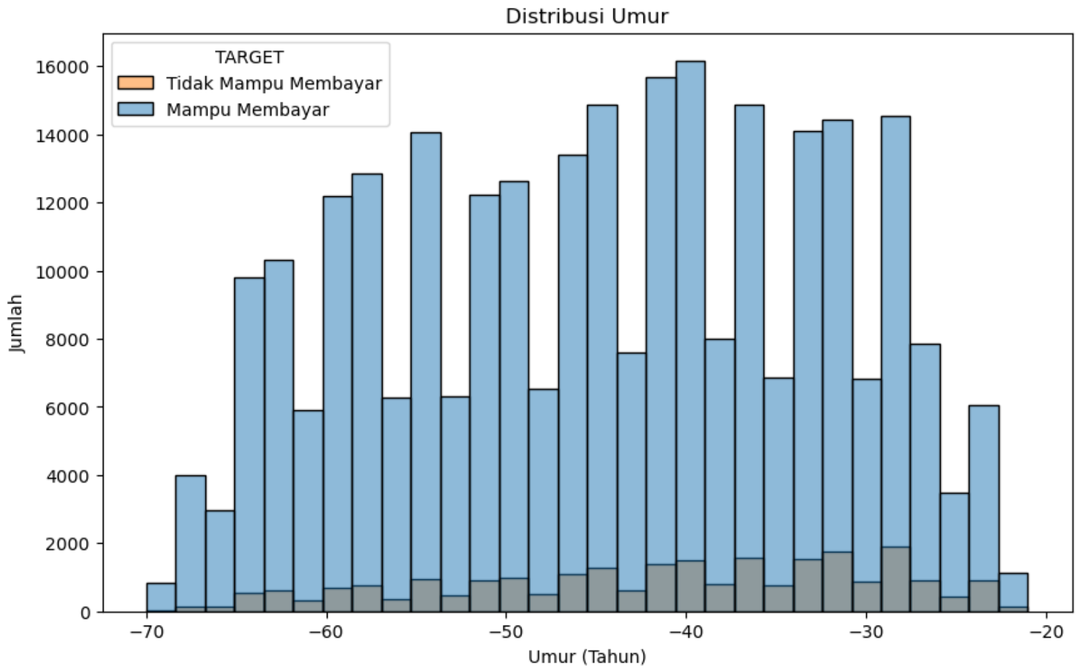
Data Understanding

- Dataset have have 122 columns with 307511 entries
- Have 16 categorical feature and 105 numerical feature
- Have a data type like int64, float64, and object
- TARGET distribution is 282686 not difficult to pay and 24825 difficult to pay

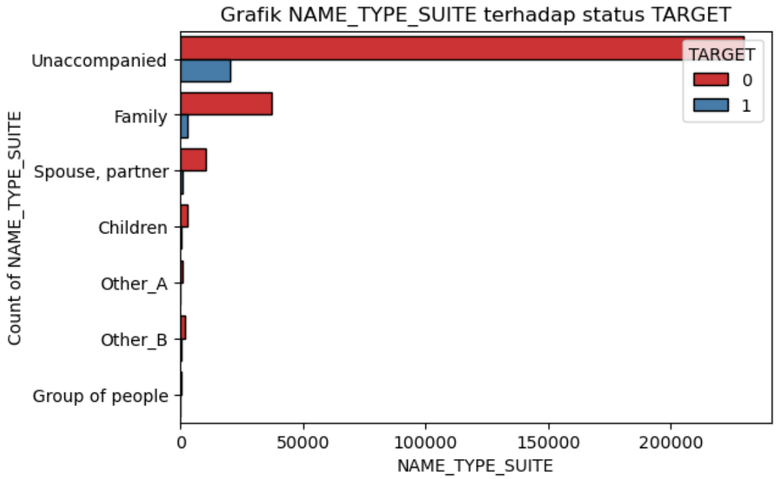


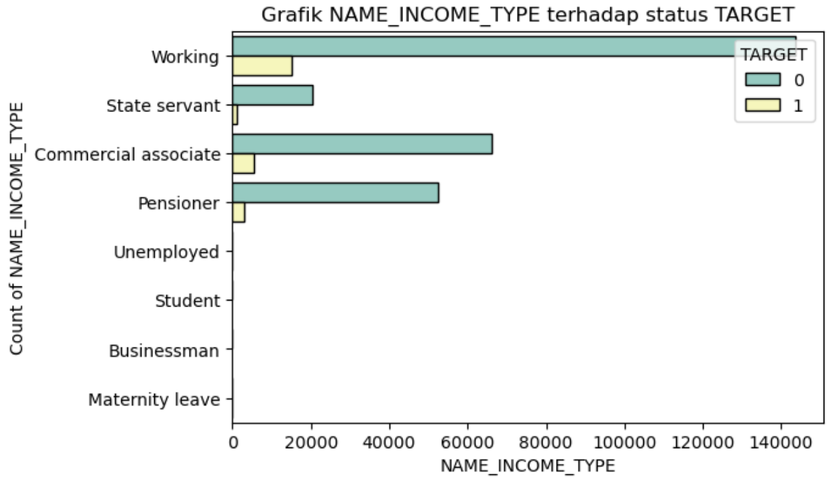
Data Visualization & Insight

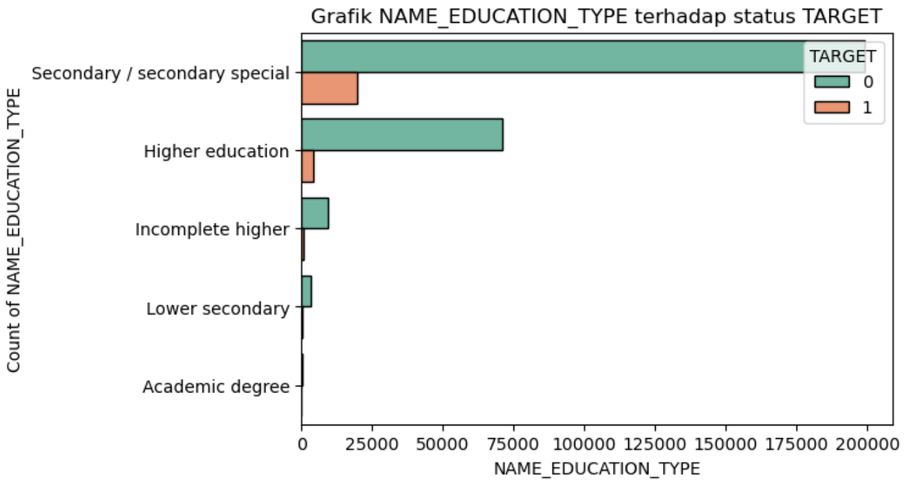


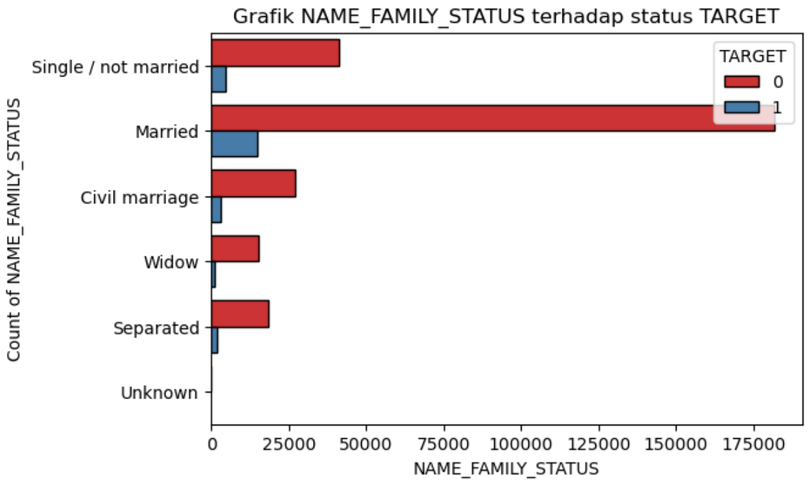


Data Visualization & Insight









Data Preprocessing

Data Cleaning:

- Handling negative value in some columns.
- Handling missing value, drop columns and imputation with mode & median
- Check Outliers from data
- Check Duplicated from data

Data Encoding

- Label Encoding (Column with 2 unique value)
- One Hot Encoding (Column with more than 2 unique value)
- Frequency Encoding (Column with many unique value)

Modelling Process

Feature Selection,

- We Choose top 20 best feature for modelling

Scaling Data

- MinMax Scaler

Handling Imbalanced Data

- SMOTE

In Modelling process, we use some of Machine Learning Algorithm.

- Logistic Regression, Logistic Regression
- Random Forest, Naive Bayes
- Decision Tree, XGBoost

For Modelling process, we split data with ratio 80 data train :20 data test, and we check shape of data train and test

Model Result

Based on our evaluation, the Random Forest model has emerged as the best-performing choice compared to other models. It has demonstrated high accuracy in prediction, providing consistent and reliable results. One of the key advantages of the Random Forest model is its ability to address the issue of overfitting, which can occur when a model becomes too specialized to the training data and performs poorly on new data.

	Model	Akurasi Train	Akurasi Test	Akurasi Val
0	DecisionTree	100.00%	90.35%	83.36%
1	RandomForest	100.00%	95.43%	91.72%
2	Naive Bayes	64.84%	64.95%	75.50%
3	LogisticRegression	66.44%	66.58%	67.61%
4	XGBoost	95.62%	95.39%	91.86%
5	KNN	88.38%	84.05%	66.55%

	Model	Precision	Recall	F1 score	ROC AUC
0	DecisionTree	89.42%	91.51%	90.45%	90.35%
1	RandomForest	99.49%	91.33%	95.23%	95.43%
2	Naive Bayes	70.35%	51.62%	59.54%	64.94%
3	LogisticRegression	67.01%	65.27%	66.13%	66.58%
4	XGBoost	99.57%	91.16%	95.18%	95.38%
5	KNN	75.84%	99.92%	86.23%	84.06%

Conclusion & Recommendation

the recommendation is to use the Random Forest model as the preferred model for credit score prediction in Home Credit Indonesia. This model can be implemented in the loan approval process to assess the creditworthiness of customers accurately. By leveraging the Random Forest model, Home Credit Indonesia can minimize the risk of default and make informed decisions when approving loan applications.

Additionally, it is crucial to continuously monitor and update the model's performance over time. This can be done by regularly retraining the model with new data and assessing its performance using relevant evaluation metrics. By doing so, Home Credit Indonesia can ensure the model remains robust and reliable in predicting credit scores and identifying customers with a high likelihood of repayment.


Thank You