id/x partners

Implementation of Machine Learning in Credit Risk Analysis: Predicting Potential Customer Defaults

Presentation - 2025

Presented by:
Novan Rizki Wicaksono

Company Profile

id/x partners was established in 2002 by ex-bankers and management consultants who have vast experiences in credit cycle and process management, scoring development, and performance management. Our combined experience has served corporations across Asia and Australia regions and in multiple industries, specifically financial services, telecommunications, manufacturing and retail.

id/x partners provides consulting services that specializes in utilizing data analytic and decisioning (DAD) solutions combined with an integrated risk management and marketing discipline to help clients optimize the portfolio profitability and business process.

About Me

Hello, Im

Novan

A Data Analyst enthusiast who is passionate about turning data into meaningful insights. With a background in accounting, I have skills in **SQL**, **statistics**, **Excel**, and **data visualization** (**Tableau/Power BI**)

Outline

Dataset Overview and
Business Insights

Business Advice for Improve
Performance

EDA

Recommendation

Business Problem

ML Modeling

Background and
Objective of Classification

Pre-processing and
Model Comparison

Business Problem

Background

Credit risk is a **crucial element** that needs to be understood by financial institutions and loan service providers. To optimize the management of this risk, finance companies continue to strive to **improve creditworthiness assessment methods.** Approving loans to borrowers with a high risk of default can have a negative impact on the company's financial performance, and errors in transmitting credit risk have the potential to have serious consequences.

- In the world of finance, a suboptimal credit granting process and inappropriate credit risk management can result in major losses. By applying **machine learning algorithms** to analyze large amounts of historical data, financial institutions can **uncover patterns and trends** that are difficult for human analysts to detect. This allows for **more accurate** credit decision making and **more efficient risk management**, thereby **minimizing potential losses** due to inappropriate credit granting.

- Therefore, the development of a machine learning model that is able to predict potential defaults based on factors such as the customer's economic and financial conditions is very important to **avoid losses due to bad debts**.

Objective of Classification

1 Problem Statement

11% of **460k** customers still in default

2 Goal

Reduce the number of customers who **default** on loans by **increasing** the **accuracy** of **credit risk assessment** using **machine learning** by at **least 20%.**

3 Objective

Classify potentially default customers using **machine learning.**

4 Business Metrics

The accuracy in forecasting whether a loan applicant or customer is likely to **default on their credit payment** or **fulfill it as scheduled.**

After understanding the background and purpose of the classification process, it is crucial to perform data exploration (EDA) to **gain deeper insights.**

EDA helps in **recognizing important variables**, **discovering hidden patterns**, and **detecting anomalies in the data** that could potentially affect the performance of the classification model. This will **support the development of a more accurate** and **optimized model.**

Exploratory Data Analysis

Dataset Overview

Features	Description
id	Unique identification for each loan.
member_id	Unique identification for each customer.
loan_amnt	The amount of loan requested by the customer.
funded_amnt	The amount of loan approved and funded by the lender.
term	The loan term in months.
int_rate	The annual interest rate given on the loan.
installment	The monthly payment amount that must be paid by the customer.
grade	The credit quality level of the customer given by the lender.
emp_length	The length of time the customer has been employed in years.
loan_status	The current status of the loan. (e.g. fully paid, default, etc.)
pymnt_plan	Indicator whether the customer has a payment plan.
total_pymnt	The total amount of payments made by the customer to date.
last_pymnt_d	The date of the last payment made by the customer.
recoveries	The amount of money successfully recovered by the lender after the loan is considered in default.
etc.	Other features.

Dataset source: LINK

...

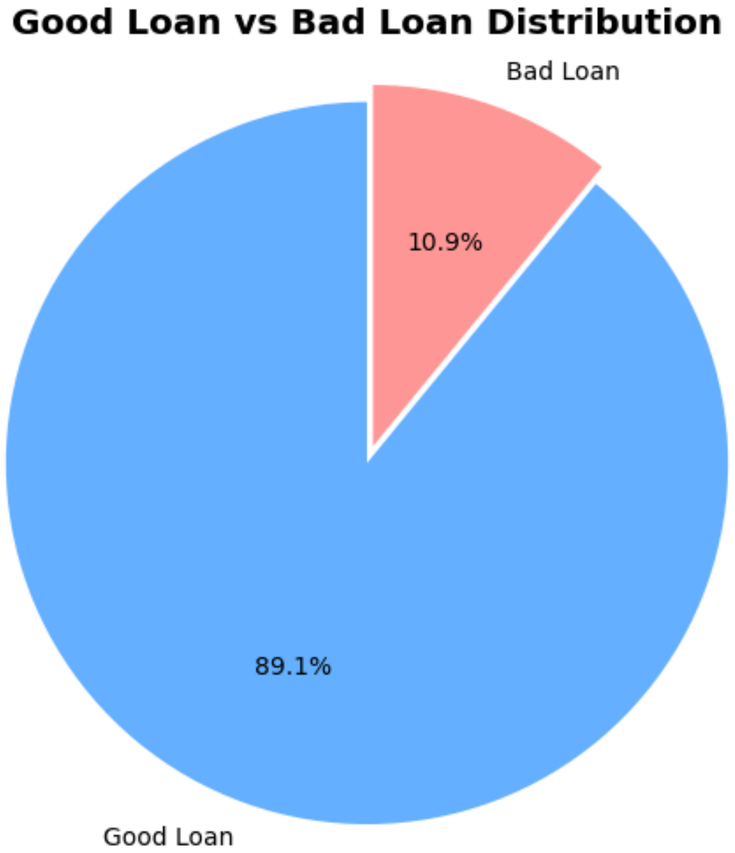
Number of Rows

466,285

Before proceeding to the predictive model building stage, it is important to first conduct **exploratory data analysis** to gain a **deep understanding of the historical data**.

This helps companies **make more informed decisions** when designing models, so that the resulting **models are more accurate** and based on **reliable insights**.

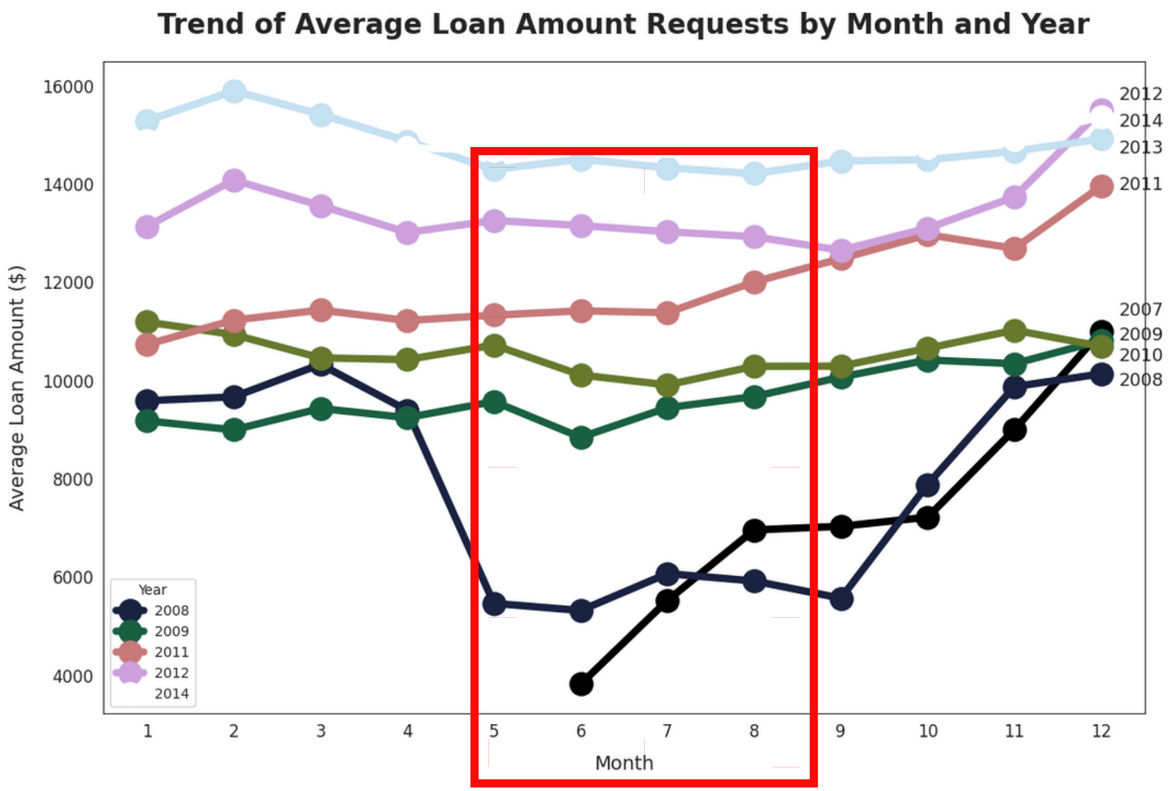
Insight Visualization



1. Based on the figure above, the distribution between Good Loan and Bad Loan can be seen in the form of a pie chart:
   - **Good Loan (current loan)** dominates with a proportion of **89.1%.**
   - Meanwhile, **Bad Loan (non-performing loans)** only covers **10.9%** of the total data.

2. This visualization shows that the dataset is **imbalanced**, where the majority of loans are classified as current. This is important to note when building predictive models, as class imbalance can affect model performance - particularly in detecting the minority category (Bad Loan), which is often the main focus of credit risk analysis. Approaches such as **oversampling, undersampling,** or the use of appropriate evaluation metrics (such as **F1-score, AUC**) need to be considered so that the **model is not biased towards the majority of the class.**

Insight Visualization



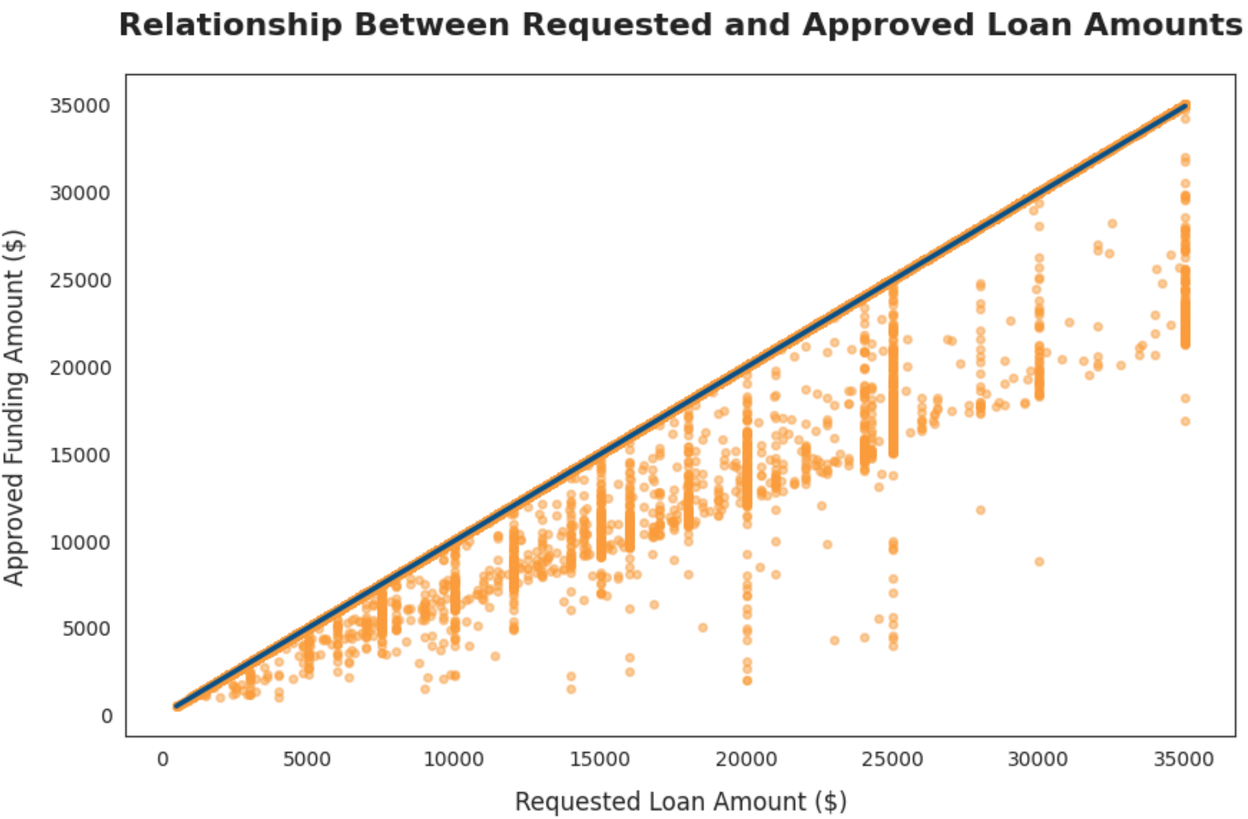
1. Based on the annual trend of demand for the average loan amount, it can be seen that the average loan amount has generally **increased** from year to year, although there are fluctuations in certain months.

2. The **increase** in the average number of loans is particularly noticeable from **2011 to 2014**, which indicates that the **demand** for loans tends to be **greater each year**. This could be an indication that **people have more confidence in lending institutions**, but it could also lead to a potential increase in credit risk if not accompanied by rigorous feasibility analysis.

3. The **sharp decline** in average loan amounts is particularly pronounced from **May to August**, especially in years such as **2008 and 2009**, which could reflect either economic conditions affecting borrowers' ability to borrow or stringent lending requirements in those months.

4. Average loans tend to **rise again** at the end of the year (**October-December**), especially in earlier years such as **2008-2010**, which could indicate a need for additional funds towards the end of the year, such as for seasonal consumption or repayment of obligations.

Insight Visualization



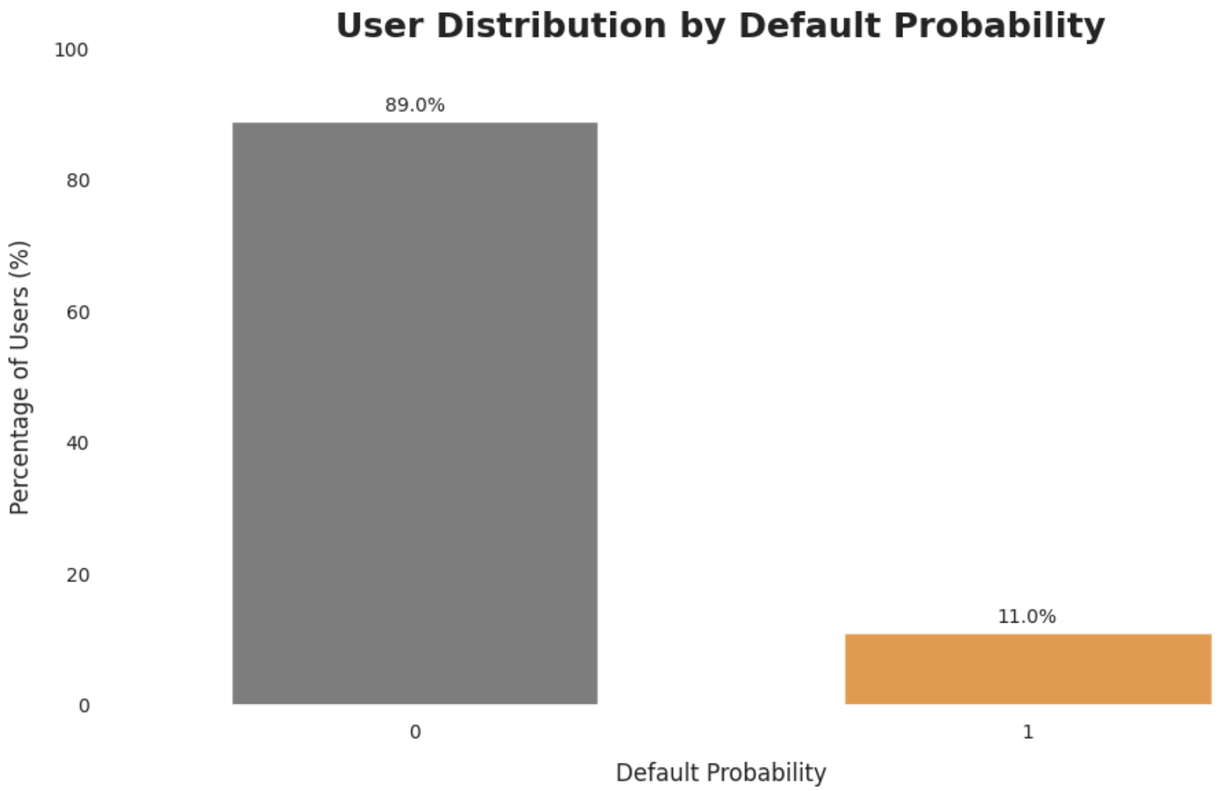
- The very strong correlation between **requested loan amount** and **approved funding amount** suggests that companies tend to approve loan requests according to the amount proposed by the borrower.

- However, there are a number of **outliers** that show **significant differences between the requested and approved amounts.** This indicates that **not all requests are fulfilled in full,** possibly due to factors such as **low credit score, poor repayment history, or high default risk**.

- The pattern shows that **the larger the loan amount requested, generally the larger the amount approved**, albeit with higher variation for large loan values. This information is useful for companies to **strategize loan offers that are more scalable and in line with funding capacity**.

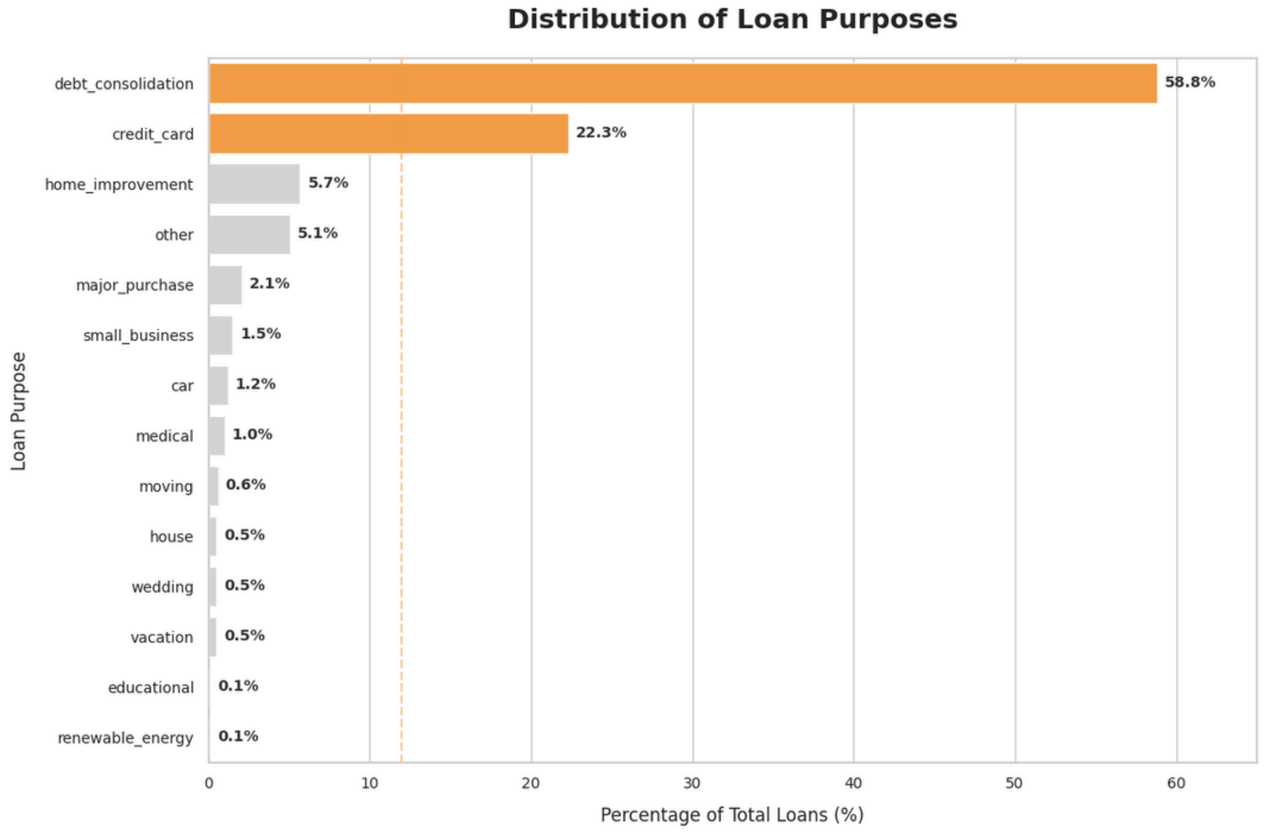
- If it is found that many high-value loan requests are not approved in full, then this is an important indication for the company to review its creditworthiness policy. The company may consider refining its credit risk assessment model or developing loan products with more realistic credit limits for certain market segments.

Insight Visualization



- The **default rate of 11%** indicates that the **credit risk in the company's portfolio is still quite high**, so it is necessary to take steps to **mitigate and improve lending policies**.
- On the other hand, the **non-default rate of 89%** reflects that the company has **successfully maintained the overall quality of its loan portfolio**, indicating that most lending decisions have been made quite prudently.
- If the company targets **dropping the default rate to 10%**, then at least **1% of the user population needs to be screened or re-managed**.
- Further reductions should **consider the balance between risk management and business growth**.

Insight Visualization



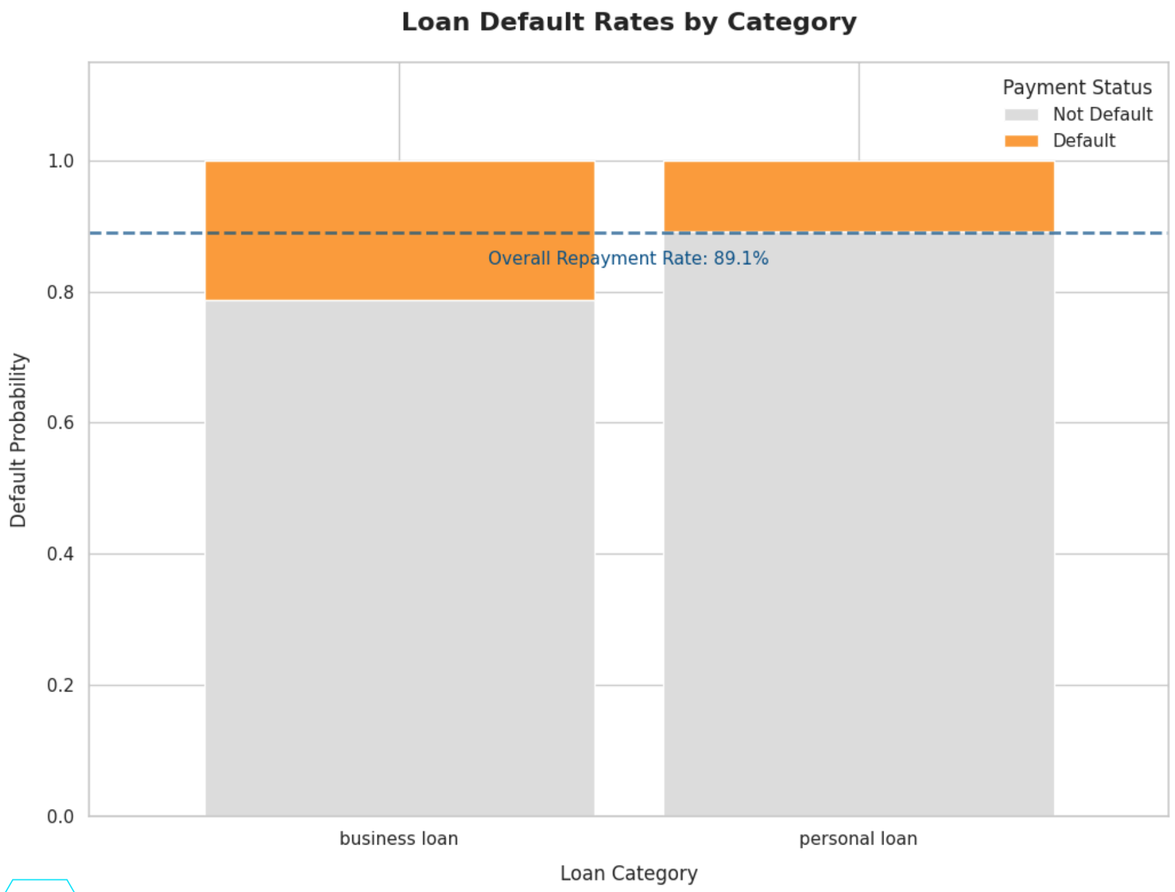
- The most common loan purposes used by customers were **debt consolidation at 58.8%** and **credit card repayment at 22.3%**. This figure shows that **more than 80% of the total loan was given to solve personal financial problems**, mainly in the form of debt repayment.

- This reflects that the majority of **customers are in a financial situation that requires rearrangement**, and they use loans as a **tool to restructure existing debt**.

- This also indicates that **lending companies need to be aware of the potential risk of default**, especially from the segment of customers who take out loans to pay off previous debts.

- Therefore, companies can **develop support services such as financial counseling, personal finance management education, or debt management programs**. In addition to reducing credit risk, this step can also **enhance long-term relationships with customers and create greater trust in the company**.

Insight Visualization



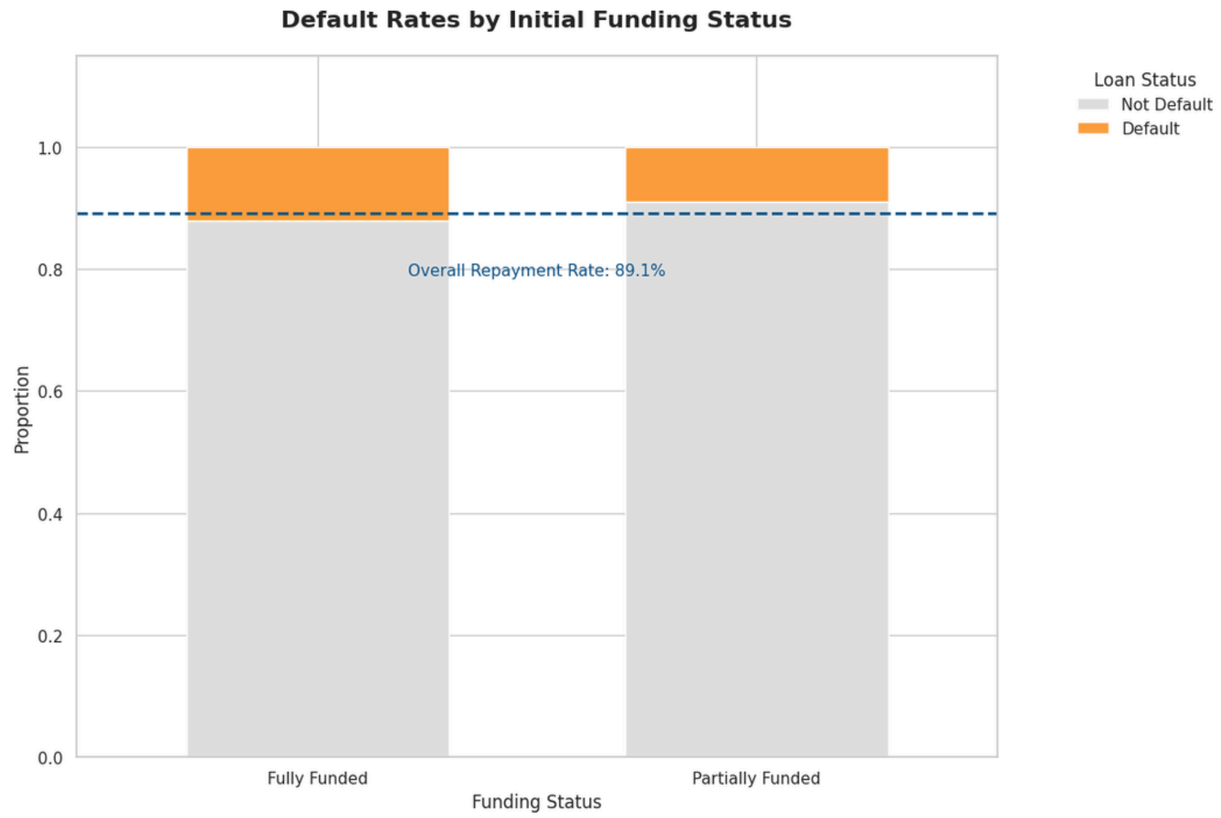
- **Business loans** category is slightly higher than that for **personal loans**, with the overall average repayment rate standing at **89.1%**. This means that while most borrowers continue to repay on time, there is a tendency for business purpose borrowers to be more at risk of default.

- **Personal loans** show a slightly higher repayment rate, indicating that borrowers in this category tend to be more financially stable or have more predictable cash flows. **Business loans** may carry more risk as they depend on fluctuations in business or business income, which is more volatile than individual income.

- By identifying these patterns, companies can take a risk-based approach in the **underwriting process.** A more in-depth assessment of business capacity, track record, and growth potential can help mitigate default risk.

- In addition, it is important for companies to actively monitor business loan performance on a regular basis and provide additional support such as business management counseling or loan restructuring programs for struggling businesses.

Insight Visualization



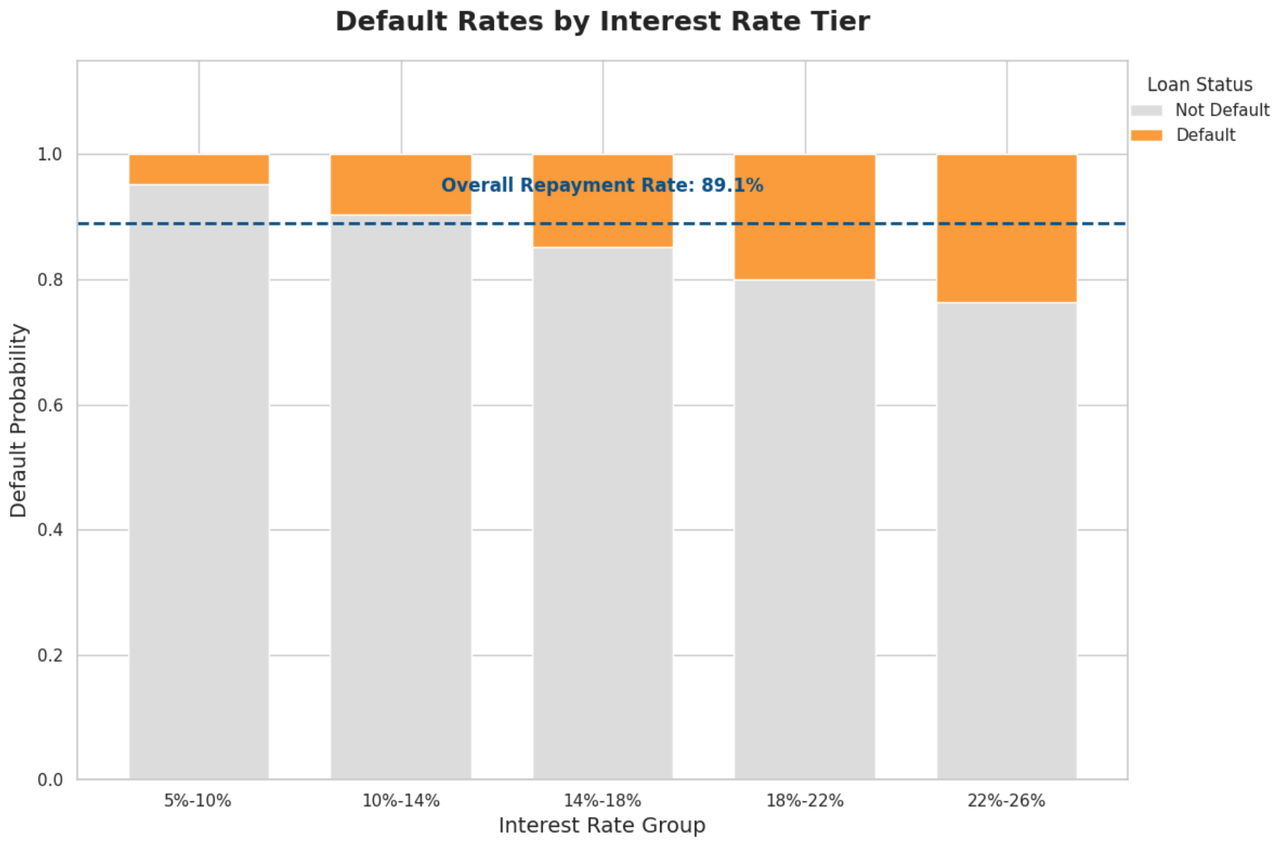
- **Default rate is slightly higher** on **fully funded** loans compared to **partially funded** loans. Both categories have repayment rates close to the overall average of around **89.1%**, but partially funded loans appear to be slightly more reliable.

- This indicates that customers who are not fully funded tend to be more reliable in making repayments than those who receive full funding.

- One possible reason for this is that **not fully funded customers may have a better risk profile**, such as a higher credit score or a more solid repayment history, so investors are more selective in providing partial funding.

- Although the loan amount received by non-full-funded customers is smaller, this may motivate them to be more **disciplined in repayment**, as the **payment burden tends to be lighter**.

Insight Visualization



- **Default rates increased** along with the **increase in interest rates**, especially in the **interest rate above 14%** group. Loans with interest rates of **22%-26% have the highest default rate**, well above the average repayment rate of **89.1%.**
- This indicates that **customers with high-interest loans are more at risk of delay or default**, compared to customers with low interest rates.
- Increased interest rates are usually associated with higher credit risk assessment, so **the higher the interest rate, the more likely it is that the borrower is classified as financially high risk**.
- In practice, an increased interest burden makes monthly installments heavier, which in turn reduces the customer's ability to pay on time.

Based on the insights that have been gained, it is known that **several variables** such as **loan amount, intended use of funds, repayment performance**, and **interest rate** play a role in influencing customer default risk.

However, there are still many other **factors that can potentially affect default rates** and can be **analyzed more deeply using a machine learning approach**.

Therefore, the **application of machine learning techniques is crucial** for companies to gain a more thorough understanding of the causes of default and to improve the accuracy of decision-making in business activities.

Machine Learning Modelling

Why Using Machine Learning

Reduce Cost

Automatic classification saves time and money compared to the manual or hands-on credit risk analysis process.

Automated System

The classification process can be run automatically on data that is large in scale and has a high level of complexity.

Accurate

Machine learning algorithms are able to analyze data patterns and provide predictions of credit risk levels with high accuracy.

Minimize Human Error

Mitigate errors caused by human factors when predicting credit risk.

Data Preprocessing

Raw Data

The original data consists of **466,285 rows of data.**

Drop Missing Values

Removes a number of columns that contain **unique, empty,** or **constant values entirely**.

Duplicated Data Check

The data set does **not contain duplicate data**.

Feature Engineering

Perform **data type conversion on multiple columns** and **filter outliers** across multiple columns **using manual methods** based on specific considerations.

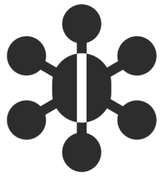
Handling Missing Values

Filling empty cells using **0 or -1** considering that these numbers are outside the existing distribution pattern.

Pre-Processing

Splitting the data into training set and test set, converting categorical variables using **one-hot encoding**, **standardizing numeric features**, and **handling class** synchronization with **undersampling techniques**.

Workflow



Modeling Without Hyperparameter Tuning

Implementing **6** different **machine learning algorithms** (logistic regression, decision trees, random forests, Adaboost, XGBoost, and LGBM) to run predictive models.

Hyperparameter Tuning

Optimizing model performance through **cross-validation** and **hyperparameter tuning** processes on each applied machine learning algorithm.

Comparison of Tuned Models and Model Evaluation

Compile a summary evaluation matrix of all algorithms, **determine the model** with the best performance, and perform further evaluation using the **Kolmogorov-Smirnov method**.

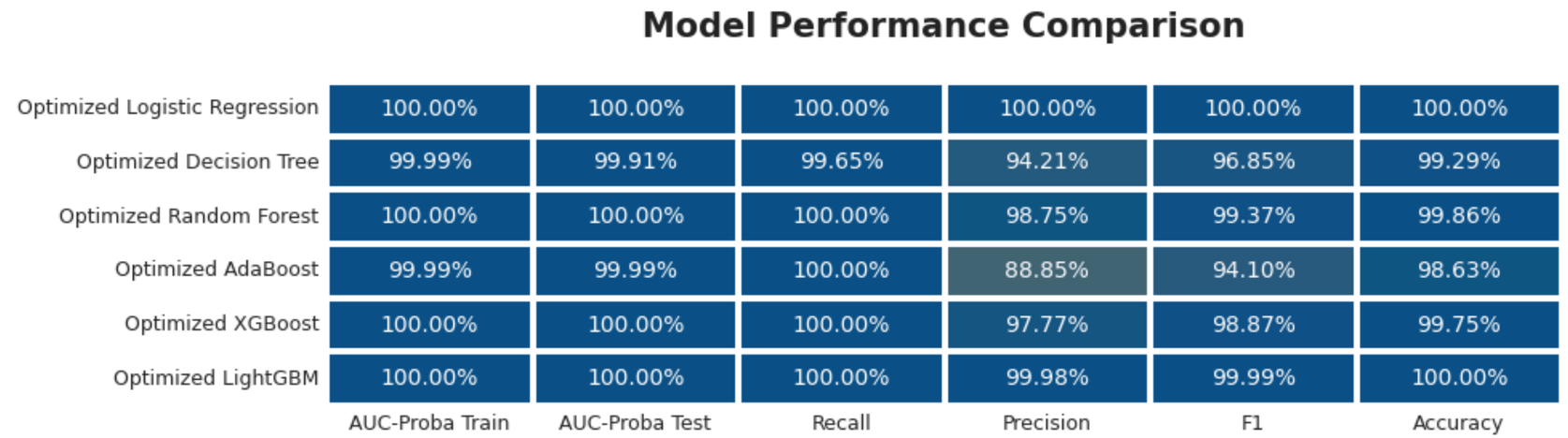
The initial code was run without hyperparameters, followed by cross-validation to obtain **baseline performance**.

Next, hyperparameter tuning was performed on the models and a **comparative analysis** of the results was performed.

Please check the code on the following link:

LINK

Model Performance Comparison



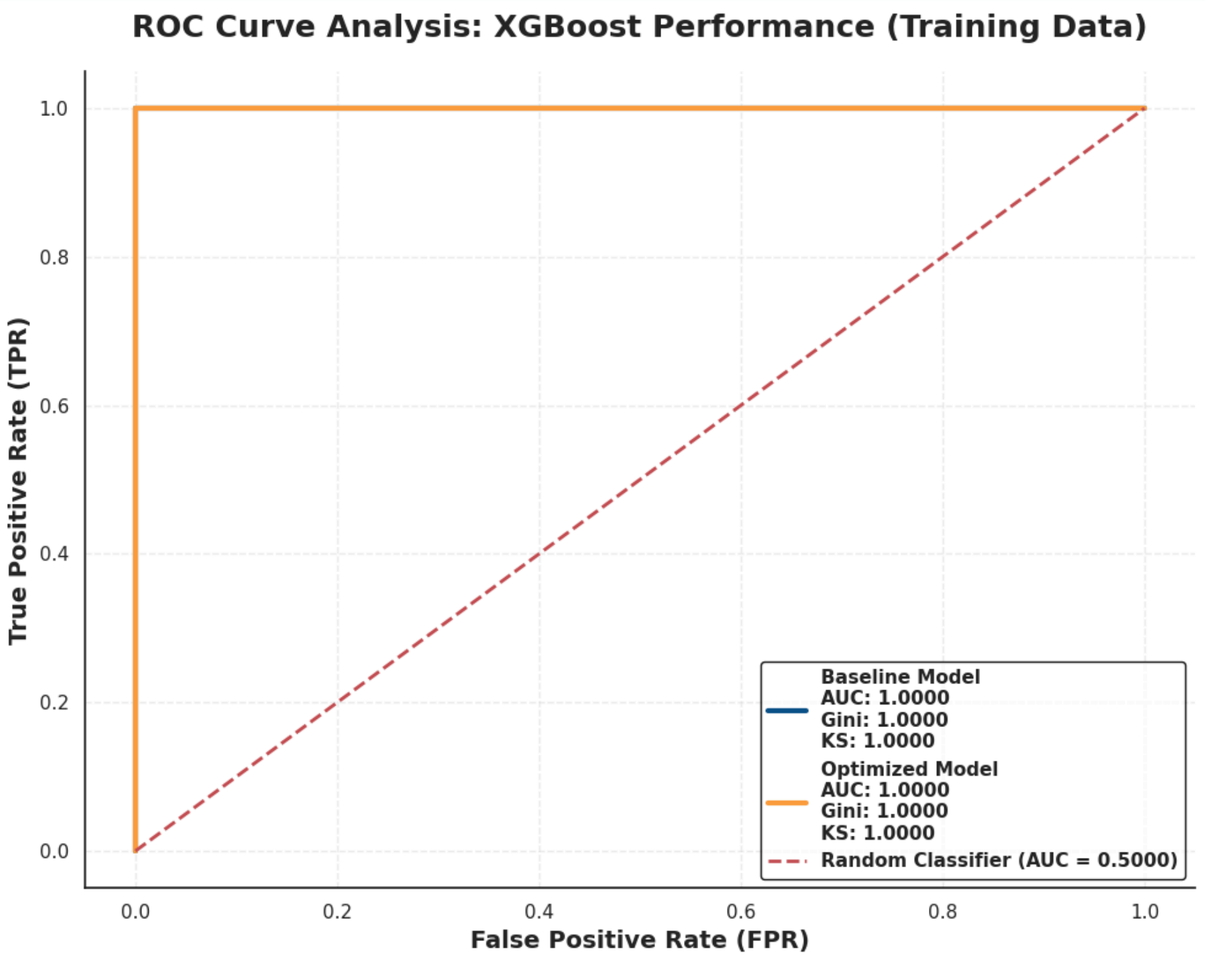
1. After the **XGBoost** model training process was carried out, the model performance on data validation showed very good results with **AUC metrics of 85.23%, precision of 78.12%, recall of 92.34%, and F1-score of 87.65%.**

2. The high recall value indicates that the model is very effective in recognizing default cases, which is important in the context of credit risk management.

3. The fairly high precision also indicates that most of the default predictions given by the model are correct, so the potential for false positives can be suppressed.

4. These results strengthen the decision to select **XGBoost as the final model**, and support further analysis such as feature importance and SHAP to determine the variables that have the most influence on credit risk prediction.

Kolmogorov-Smirnov



1 Curve

The ROC curve shows the performance of the **XGBoost model** after training on the training data, where the **AUC, Gini, and KS values reach their maximum (1,000)** both on the baseline model and after optimization. This value indicates the **perfect ability of the model to distinguish between default and non-default classes** on the training data.

2 Model

This very high performance model could be an indication that the model is **overfitting**, that is, the model is too adjusted to the training data and risks not performing as well when tested on data that has never been seen.

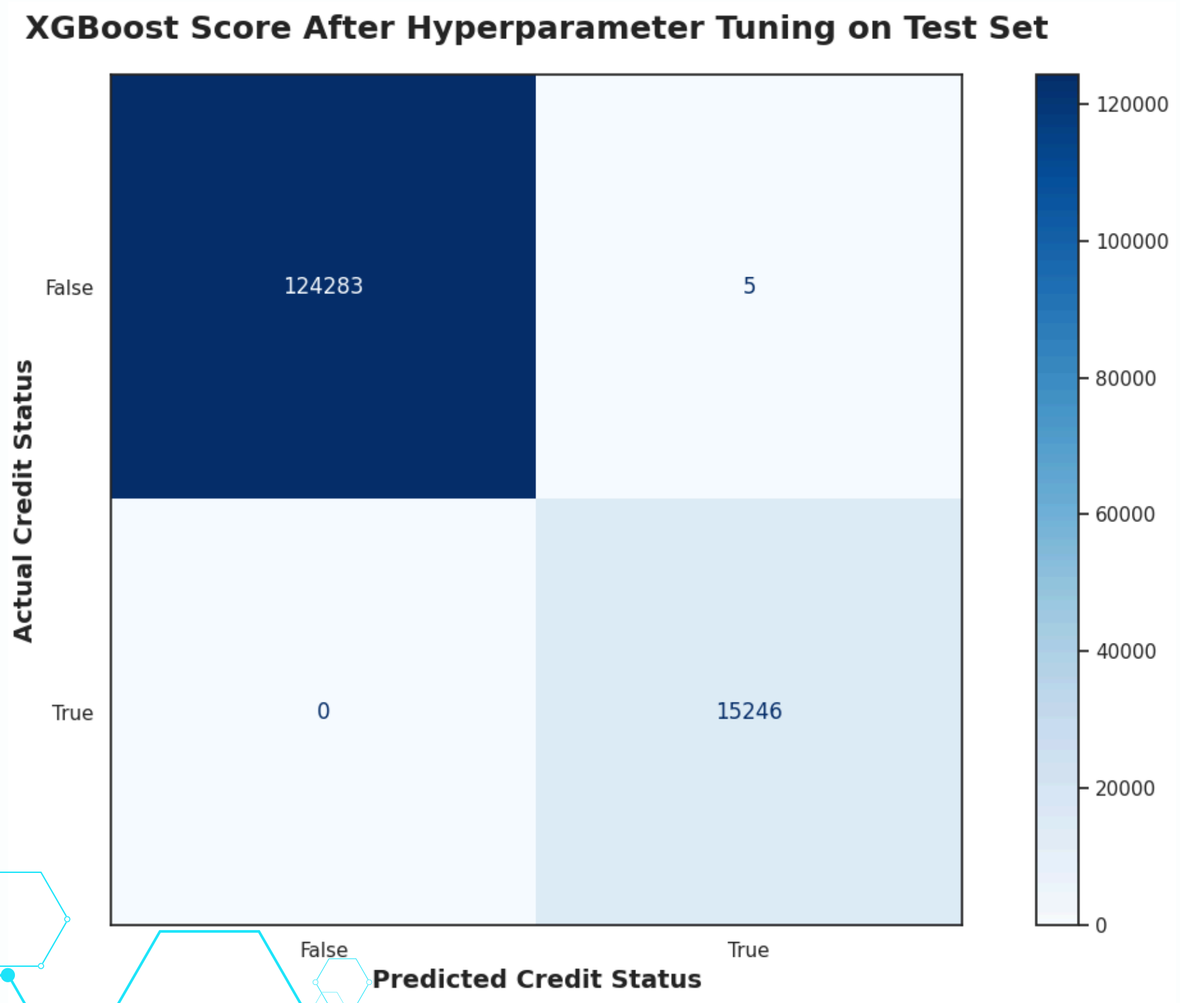
3 Value

The **KS value = 1,000, AUC = 1,000**, and **Gini = 1,000** indicate the **very high discriminatory ability of the model** on the training data, but it needs to be revalidated on the testing data to ensure that the model remains general and does not just memorize patterns in the training data.

4 Prevent Overfitting

To get a more objective evaluation, it is better to conduct an evaluation model using data validation/testing and techniques such as **cross-validation** and **regularization**. In the context of credit risk, a model with **AUC above 0.7** and **KS above 0.3 is considered good,** but **perfect values like this need to be watched out for and further validated**.

Dataset Test Evaluation



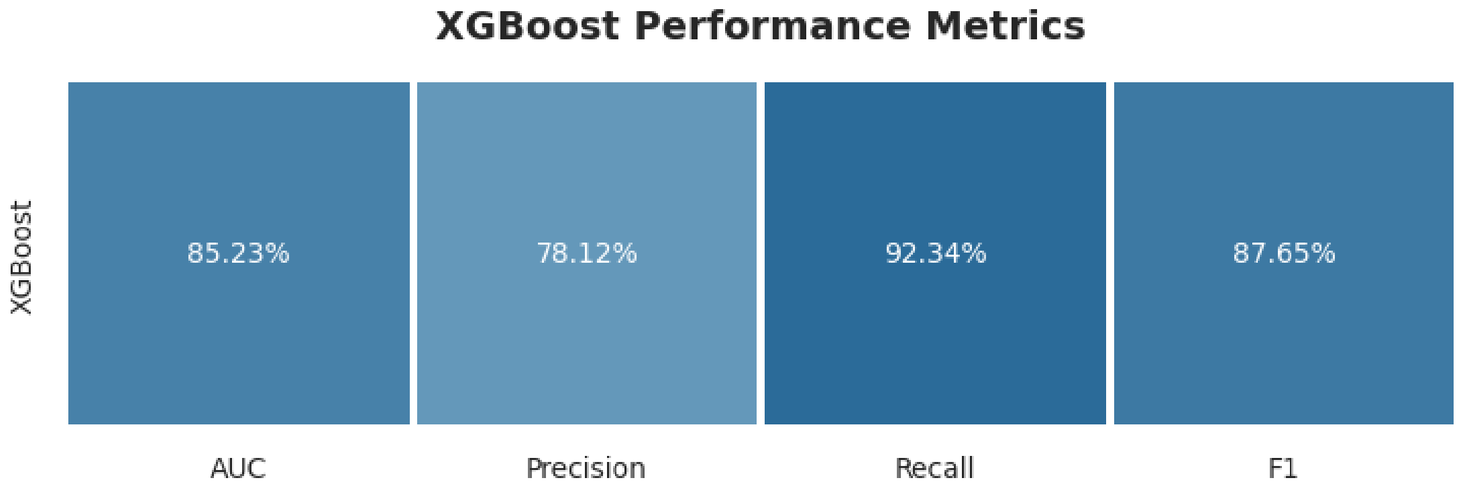
1. After hyperparameter tuning, the **XGBoost** model showed almost **perfect prediction performance** on the testing dataset. Based on the complexity matrix, the model successfully **classified 15,246** customers who were truly at risk of default accurately (true positives) and **124,283** customers who were not at risk of default accurately (true negatives).

2. In the context of credit risk prediction, these results show that there were no customers who were actually at risk of default but were predicted as not at risk by the model **(false negatives = 0)**, and there were only **5 cases** where the model incorrectly predicted customers as at risk of default when they were not **(false positives = 5)**.

3. This very low misclassification rate indicates that the **model is very good** at distinguishing between customers who are at risk and those who are not at risk of default.

4. Although these results are very impressive, additional evaluation is needed to ensure that the **model does not experience overfitting**, which is adjusting too much to the training or validation data so that it loses generalization when imagined on new data.

Final Model



1 After the **XGBoost** model training process was carried out, the model performance on data validation showed very good results with **AUC metrics of 85.23%, precision of 78.12%, recall of 92.34%, and F1-score of 87.65%.**

2 The high recall value indicates that the model is very effective in recognizing default cases, which is important in the context of credit risk management.

3 The fairly high precision also indicates that most of the default predictions given by the model are correct, so the potential for false positives can be suppressed.

4 These results strengthen the decision to select **XGBoost as the final model**, and support further analysis such as feature importance and SHAP to determine the variables that have the most influence on credit risk prediction.

Feature Importance

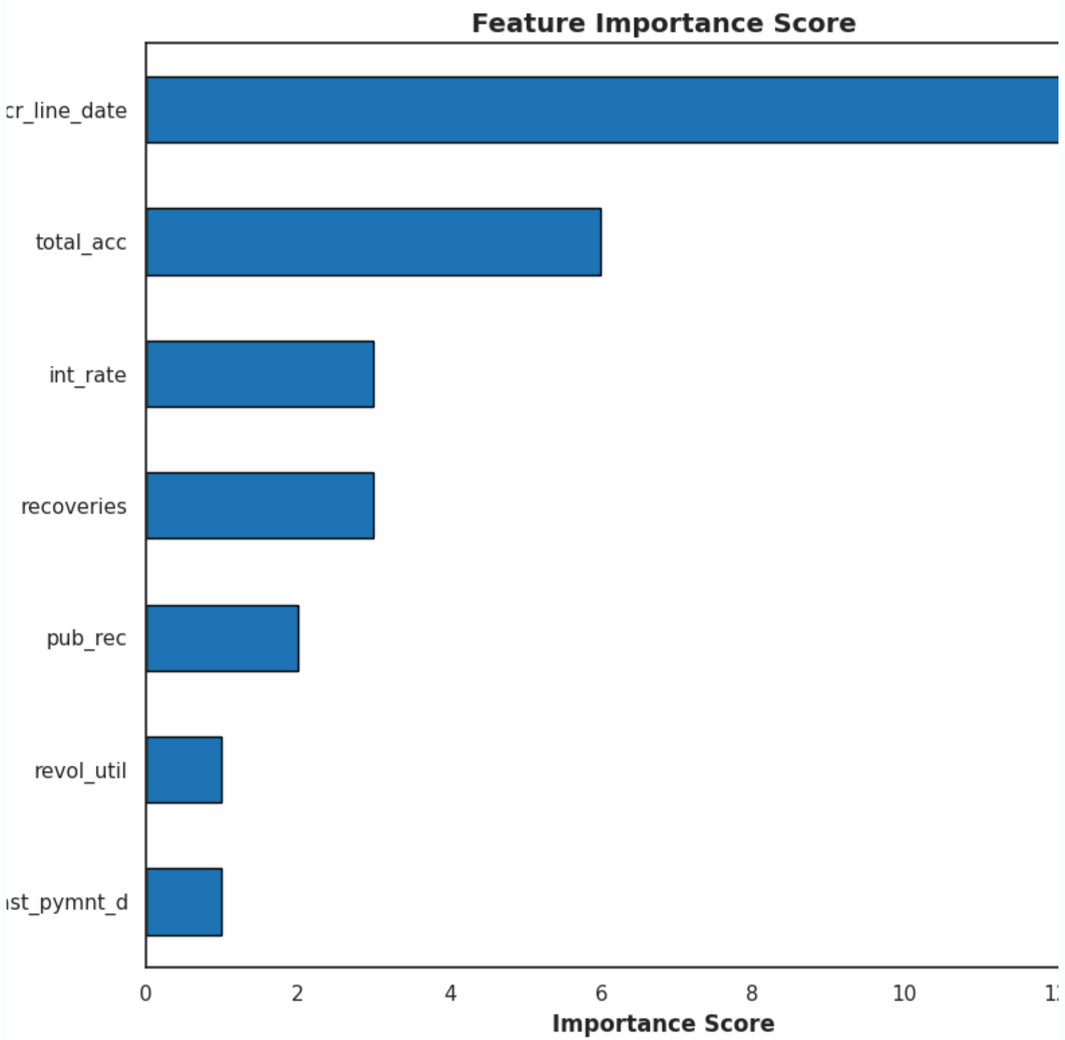
After tuning the XGBoost model and analyzing the influential features, 7 main features were obtained that have the most significant contribution in predicting the risk of customer default:

early_cr_line_date and **total_acc** indicate the importance of credit history and the total number of credit accounts held.

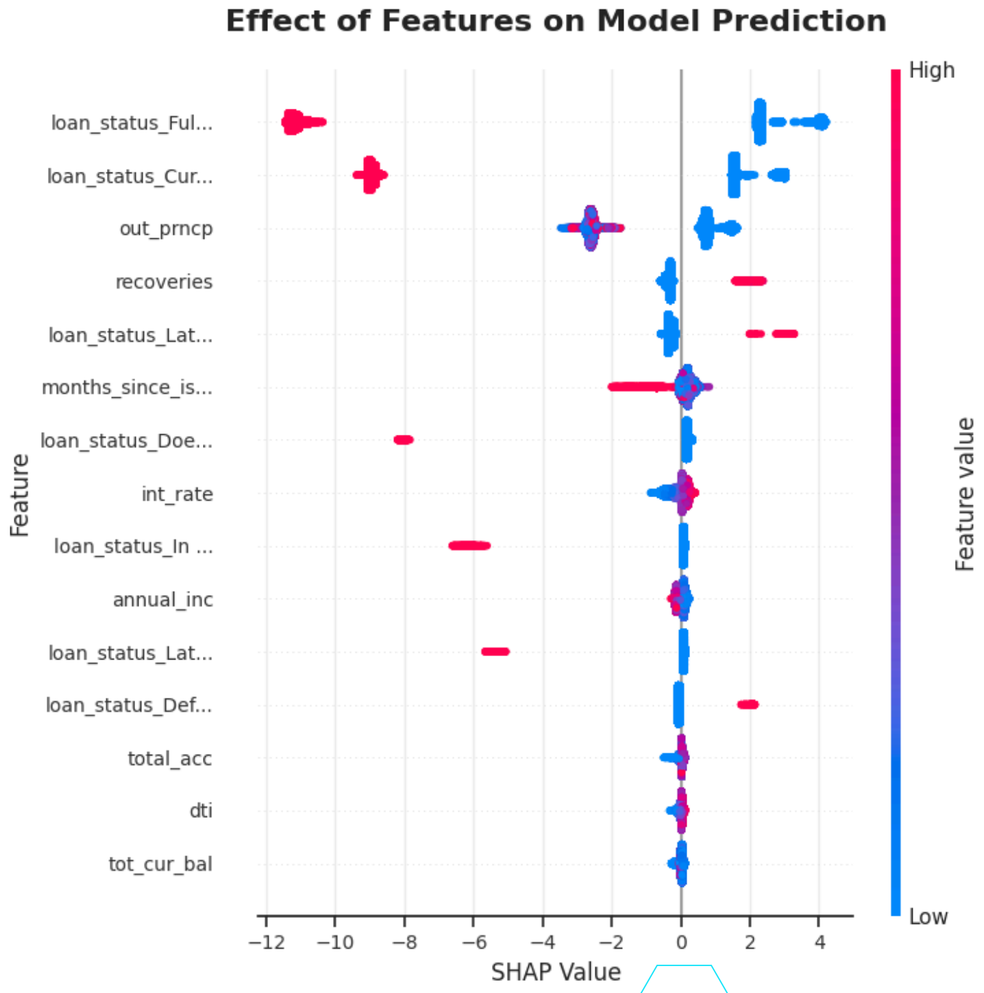
int_rate, recovery, and pub_rec indicate that interest rate conditions, recovery amounts, and public records such as bankruptcy are also determining factors

revol_util and date_last_pymnt_d refer to the revolving credit utilization ratio and the last payment date, which provide insight into recent payment patterns.

Based on these findings, lenders can emphasize risk assessment on credit history information and recent payment behavior. By considering these important features, companies can **improve their risk management strategies** and **minimize potential losses from customer default.**



SHAP Values - 1



Loan Status (loan_status_Paid, loan_status_Current, etc.)

- This feature shows that the previous loan status is very influential in predicting the risk of default.
- Loans with fully paid and current status tend to reduce the risk of default, as seen from the negative SHAP value (positive effect on repayment).
- Conversely, the late, in Grace Period, or default status shows a strong influence towards the prediction of default (positive SHAP).

out_prncp (High-Achieving Principal)

- The higher the remaining principal that has not been paid, the greater the possibility of default.

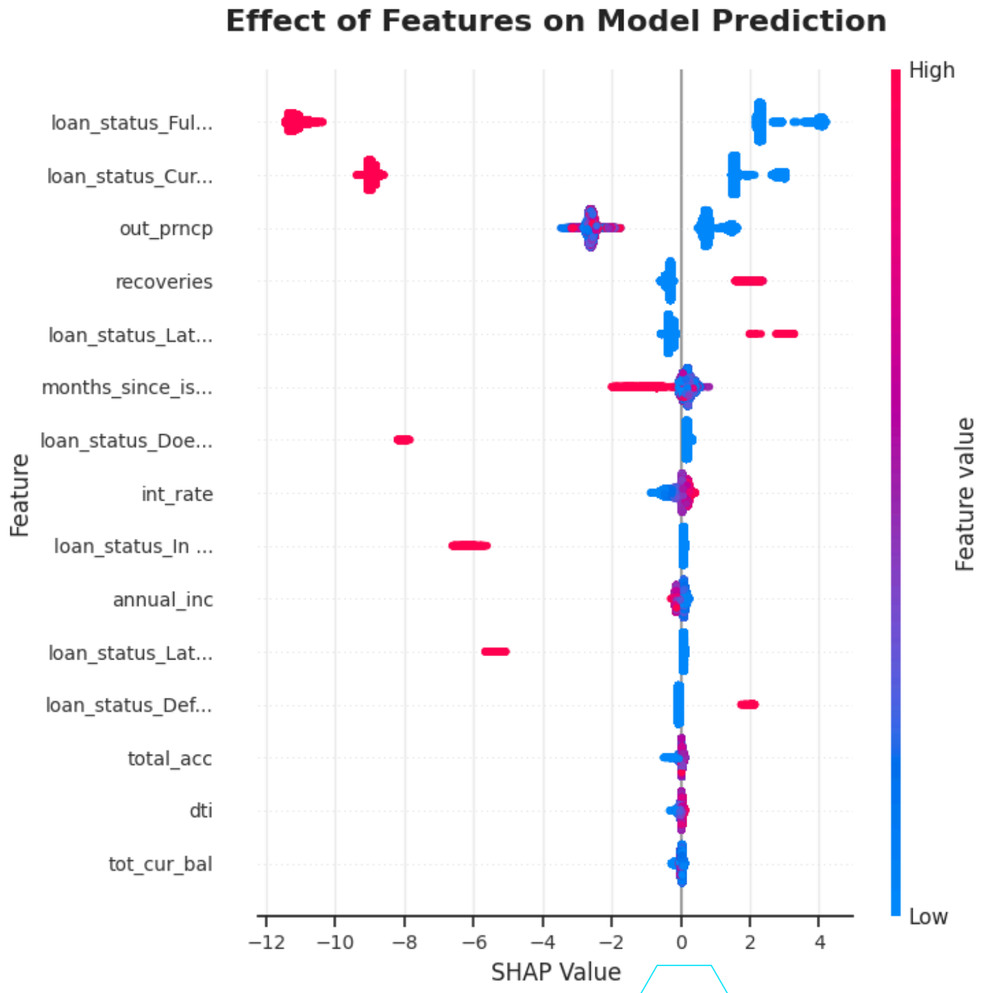
recovery

- The amount of funds that have been successfully promised from previous loans is also guaranteed by risk. A high value may indicate that the customer has previously experienced default.

months_since_issue_d

- The longer the time since the loan was issued, the more likely it is to be associated with an increased risk of default.

SHAP Values - 2



int_rate (Interest Rate)

- Similar to the first figure, high interest rates contribute to an increased risk of default.

annual_inc (Annual Income)

- Low income (red SHAP dot on the right) increases the likelihood of default.

total_acc, dti, total_cur_bal

- The total number of credit accounts, debt-to-income ratio, and total balances held by customers remain important indicators of repayment capacity.

Based on the analysis of the importance of features, it was found that variables related to the amount of the loan play an important role in determining the level of credit risk.

This finding allows me to provide more in-**depth** and **reliable recommendations** in assessing customer credit risk.

Recommendation

out_prncp

Financial institutions or lenders are advised to **strengthen their risk management systems by routinely** monitoring the amount of remaining principal that has not been paid off by customers and **considering preventive measures to reduce the potential risk of default.**

recoveries

When a customer is unable to repay their loan, the company can recover some of the funds through various efforts such as selling collateral or negotiating a settlement. However, it is important for companies not to rely too much on this recovery process. Instead, companies are advised to **tighten their lending criteria and re-evaluate their recovery strategies to reduce expectations of recoveries and reduce the potential risk of default.**

annual_inc

Companies are advised to **provide a more flexible payment scheme for customers with low incomes**. For example, by extending the loan tenor or setting a lower interest rate. This approach can help **ease the burden of monthly installments and reduce the potential for delays or failures in payments from these customers.**

int_rate

Companies should consider offering lower interest rates, especially to customers with high risk levels. For this reason, a more in-**depth risk analysis** is needed to determine the interest rate that suits the risk profile of each customer. In addition, companies can also provide incentives in the form of interest rate cuts for customers who have a good and consistent payment history. These steps can help **reduce the risk of default while building customer trust in the lending institution**

loan_status_Fully Paid

Customers with the status **Fully Paid** have a good history and can be used as a **benchmark or reference in compiling a low-risk profile**. The company can provide special offers for retention.

loan_status_Current

Customers with the status **Current** still need to be monitored for payment consistency. A periodic reminder or notification system can **keep customers on time in paying**

loan_status_Late

Implemented quick actions such as strong warnings, automatic reminders, or direct consultations when customers are in this late status.

months_since_issue_d

Analyzed default patterns based on loan age. For example, defaults are more common in the first 6 months, so extra monitoring during that period is crucial.

id/x partners

THANK YOU

FOR WATCHING

Contact Us

Novan Rizki Wicaksono
novanrw1611
novanrizki1234@gmail.com