Rakamin Academy X id/x partners

Prediction Model

ID/X Partners - Data Scientist

Presented by
Anggun Dwi Lestari

ID/X Partners

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.

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.

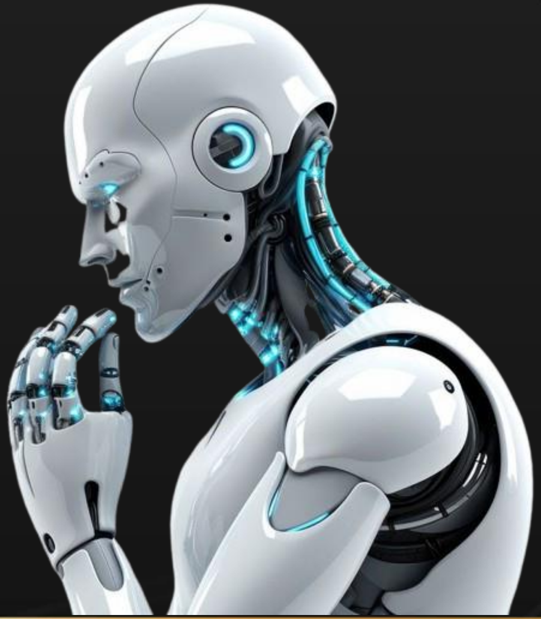
Developing a Model to Predict Credit Risk

Background

ID/X Partners, as a lending company (multifinance), aims to enhance accuracy in assessing and managing customer credit risk. This initiative is expected to optimize business decision-making while minimizing potential losses in the future.

Business Goal

Developing a machine learning model to predict credit risk.



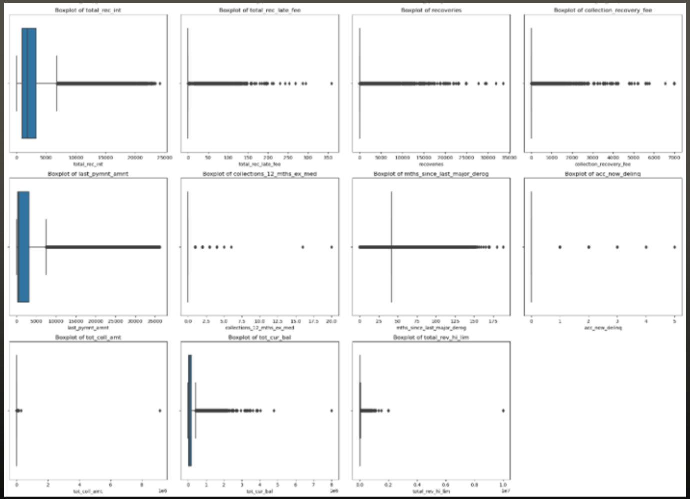
Missing Values

There are many features with missing values, including:

emp_title : 27588
emp_length : 21008
annual_inc : 4
desc : 340304
title : 21
delinq_2yrs : 29
earliest_cr_line : 29
mths_since_last_delinq : 250351
mths_since_last_record : 403647
open_acc : 29
pub_rec : 29
revol_util : 340
total_acc : 29
last_pymnt_d : 376
next_pymnt_d : 227214
last_credit_pull_d : 42
collections_12_mths_ex_med : 145
mths_since_last_major_derog : 367311
annual_inc_joint : 466285
dti_joint : 466285
verification_status_joint : 466285
mths_since_rcnt_il : 466285
tot_coll_amt : 70276
tot_cur_bal : 70276
open_acc_6m : 466285
open_il_6m : 466285
open_il_12m : 466285
open_il_24m : 466285
acc_now_delinq : 29
total_bal_il : 466285
il_util : 466285
open_rv_12m : 466285
open_rv_24m : 466285
max_bal_bc : 466285
all_util : 466285
total_rev_hi_lim : 70276
inq_fi : 466285
total_cu_tl : 466285
inq_last_12m : 466285
inq_last_6mths : 29

Outlier Values

There are many features in the dataset with outlier values. However, since this is historical or financial transaction data, it was decided **not to remove the outliers to avoid the potential loss of important information.**



Distribution of Loan Status

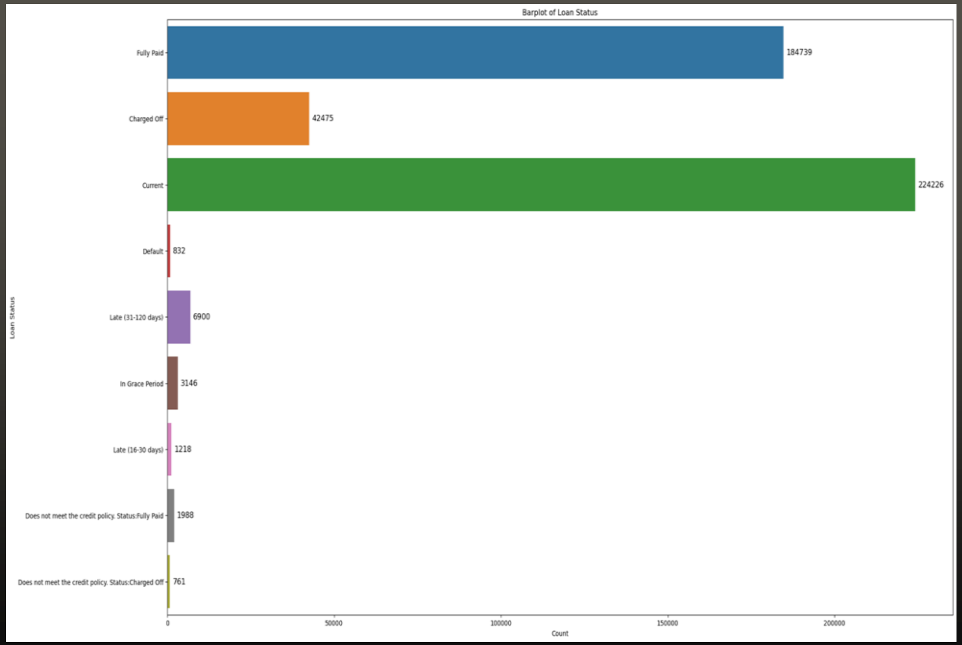
The feature that has the potential to be the target label of this dataset is the *Loan_status* column.

Based on the graph, it can be seen that debtors with :

- fully paid status (credit has been fully paid off)
- current status (still in the process of paying off credit)
- charge off status (credit is considered uncollectible)

have the highest distribution.

This indicates that the majority of debtors under ID/X Partners fall into the category of safe debtors.

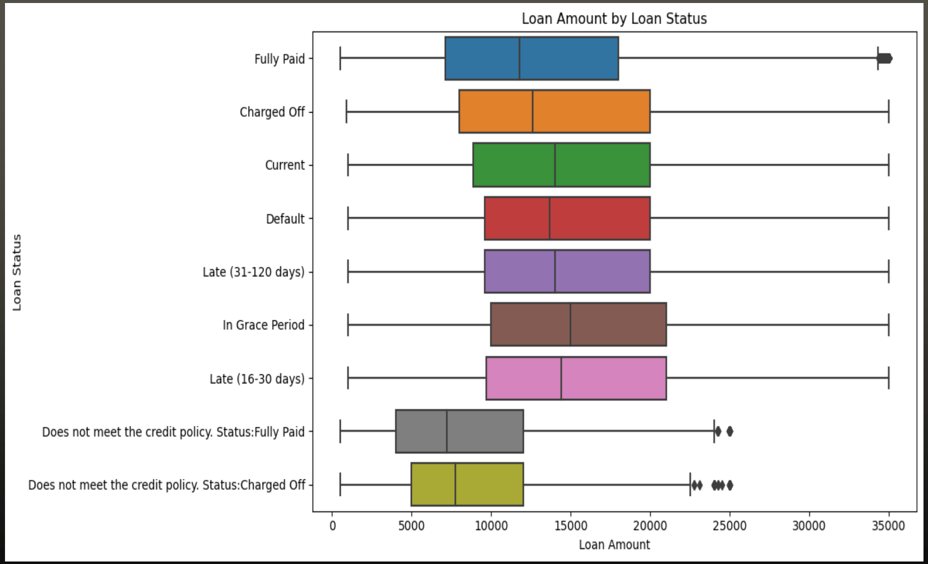


Distribution of Loan Amount by Loan Status

Based on the graph, it can be seen that the category of debtors with the lowest loan amount consists of those with a ‘doesn't meet the credit policy’ status.

On the other hand, debtors with other statuses have nearly the same maximum loan amounts, even though some of them have a *default* status.

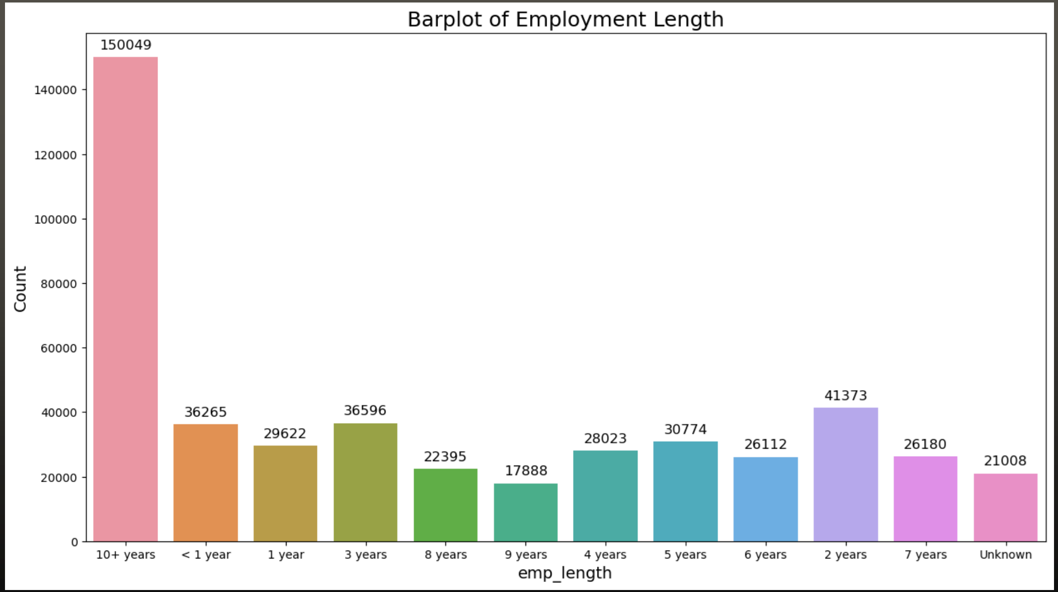
This can serve as a consideration to prioritize loan amounts for debtors with low-risk categories rather than solely based on whether they meet the credit policy requirements. This approach is expected to minimize default risks while improving the quality of the loan portfolio.



Distribution of Employment Length

The majority of debtors have an **employment length** of more than **10 years**, reflecting job stability and more secure income, as well as lower credit risk.

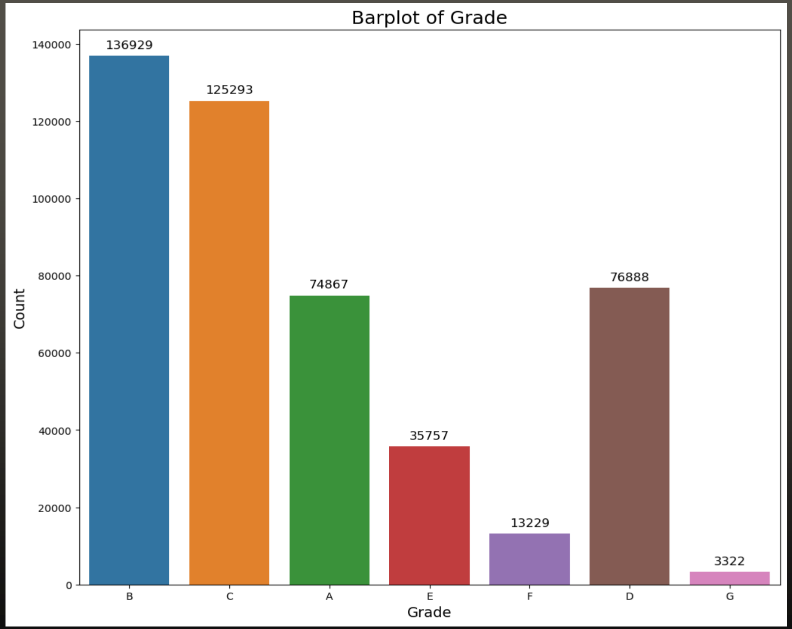
Additionally, there are debtors with an **employment length** of around **2 years** and **3 years**, which may indicate debtors who are newer to the workforce. While a shorter employment duration can suggest higher risk, debtors with 2 to 3 years of experience still show potential to meet their credit obligations if their financial management is sound.



Distribution of Grade

ID/X Partners classifies debtors based on a **grade** category, ranging from **A to G**, with **category A** representing debtors with the best risk profiles.

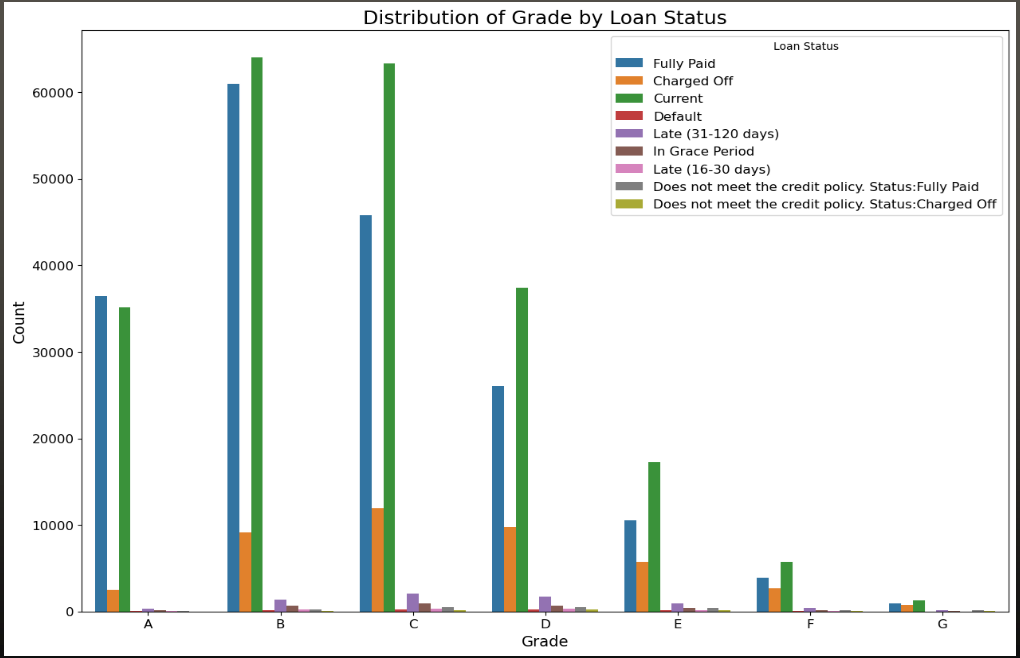
Based on the chart obtained, most debtors are classified into **categories B and C**, indicating that the majority of debtors fall into the **mid-level** category. The number of debtors in **category B** is **136,929**, while **category A** consists of **74,867** debtors.



Distribution of Grade by Loan Status

Based on the chart obtained, the majority of debtors have a **Fully Paid** status (loan settled) and **Current** status (still in the process of loan settlement), with the highest **grades** found in **B** and **C** categories.

This is consistent with the previously presented distribution of **grades** and **loan statuses**.

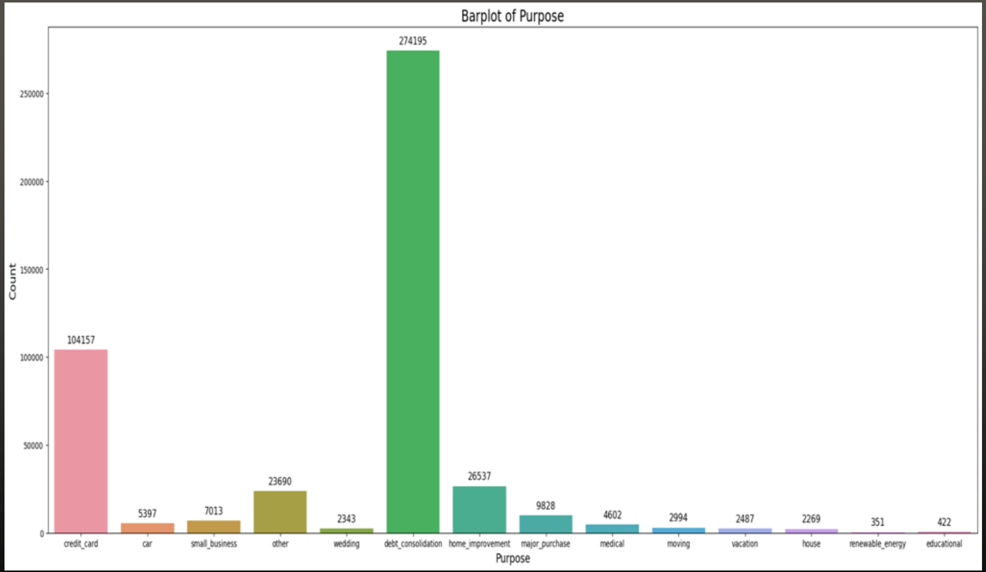


Distribution of Purpose

The most common purpose for credit applications is **debt consolidation**, which indicates that many debtors apply for credit to consolidate multiple debts into a single loan with a lower interest rate or more favorable terms.

The purpose of **credit card** also ranks highly, suggesting that many debtors use credit to meet daily expenses.

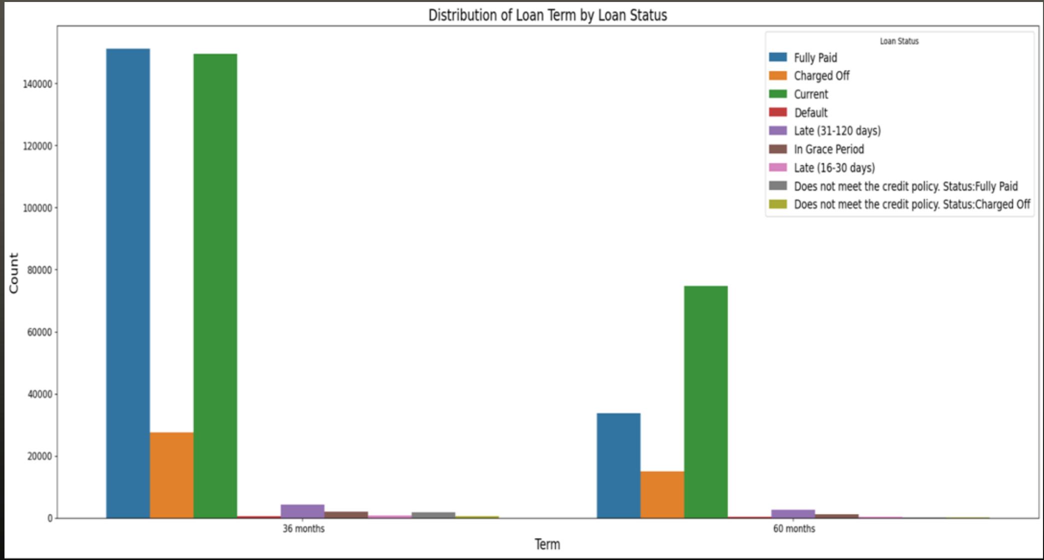
Additionally, **home improvement** is also among the top categories, meaning that a significant number of debtors apply for credit to improve or renovate their property, indicating a focus on long-term asset enhancement.



Distribution of Loan Term by Loan Status

The majority of debtors have a **term of 36 months**, with a **Fully Paid** credit status, indicating that they have completed their loan repayment obligations, and a **Current** status, meaning they are still in the process of repaying their loan according to the agreed schedule.

This reflects that most debtors within this term are in good standing, either having settled their debts or still actively involved in the repayment process.



Data Pre-Processing

Drop Column

➢ **Drop Columns with 100% Missing Values**

inq_last_12m
total_cu_tl
inq_fi
all_util
max_bal_bc
open_rv_24m
open_rv_12m
il_util
total_bal_il
open_il_24m
open_il_12m
open_il_6m
open_acc_6m
verification_status_joint
dti_joint
annual_inc_joint
mths_since_rcnt_il

➢ **Drop Columns with No Valuable Information**

application_type
mths_since_last_record
Unnamed:0
desc
next_pymnt_d
policy_code
id
member_id
application_type

Data Pre-Processing

Handling Missing Values

Fill with Median

annual_inc
inq_last_6mths
mths_since_last_delinqopen_acc,
pub_rec
revol_util, total_acc
collections_12_mths_ex_med
mths_since_last_major_derog
acc_now_delinq
tot_coll_amt, tot_cur_bal
total_rev_hi_lim

Fill with 0 & Unknown

deliq_2years,
empth_tittle, empth_lengh, title

Fill with Mode

earliest_cr_line, last_payment_d,
last_credit_pull_d

Data Pre-Processing

Loan_Status

The target is created based on the **Loan_Status** column, as follows:

1. Current
2. Fully Paid
3. Charged Off
4. Late (31-120 days)
5. In Grace Period
6. Does not meet the credit policy. Status:Fully Paid
7. Late (16-30 days)
8. Default
9. Does not meet the credit policy. Status:Charged Off

NEW FEATURE

TARGET

Credit_Label

The target labels **GOOD** and **BAD** are assigned based on the debtor's loan status, as follows:

'GOOD' refers to:

1. Fully Paid
2. Current
3. In Grace Period
4. Does not meet the credit policy. Status:Fully Paid

'BAD' refers to:

1. Does not meet the credit policy. Status:Charged Off
2. Charged Off
3. Default
4. Late (16-30 days)
5. Late (31-120 days)

Data Pre-Processing

With Hierarchy

Transformation of Object Data Type to Numerical (With Hierarchy):

1. **Grade Column**
   A= 1, B=2, C=3... G=7
1. **Sub_Grade**
2. **Emp_length**
   Unknown=0, < 1 year=1, 1 year=2...
   10+ year = 11
1. **Loan_Status**
2. **Payment_Plan**
   no=0, yes=1
1. **Credit_status**
   Good = 1, BAD = 0

FEATURE TRANSFORMATION

Without Hierarchy

Transformation of Object Data Type to Numerical (Without Hierarchy):
term, home_ownership, verification_status, purpose, initial_status, earliest_cr_line

YearMonth Format

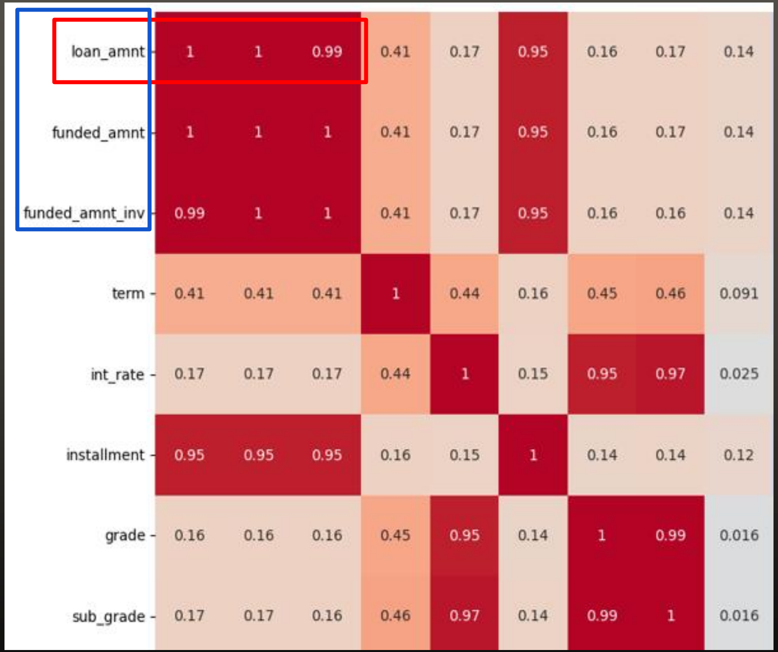
Transformation of Object Data Type to Numerical (YearMonth Format: YYMM):
issue_d, last_pymnt_d, last_credit_pull_d

Machine Learning Workflow

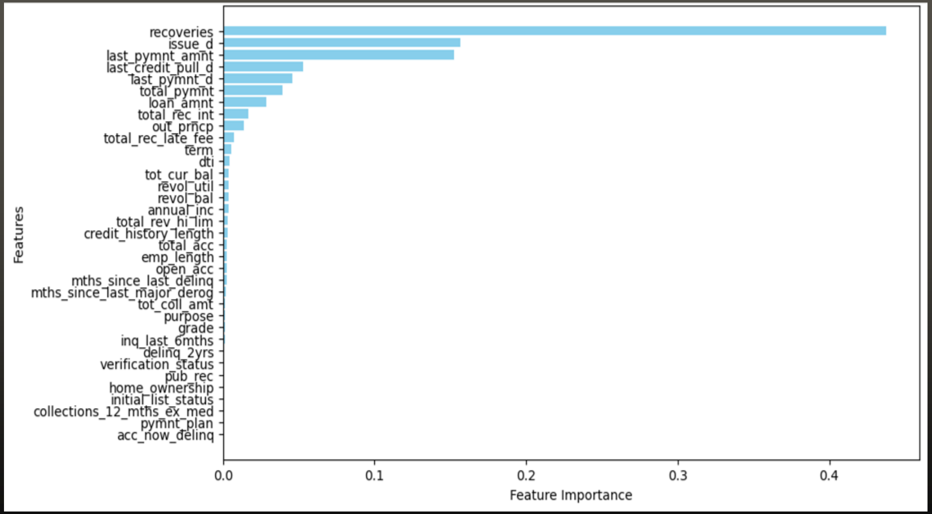
Machine Learning Preparation

1. **Only features with numerical data types are selected:**
   Features with an object data type (a total of 6) will not be processed in the machine learning model.

1. **Remove Features with Multicollinearity:**
   Features with multicollinearity, or those with a correlation value above 0.8, will be removed. These include: funded_amnt, funded_amnt_inv, installment, int_rate, sub_grade, out_prncp_inv, total_pymnt_inv, total_rec_prncp, collection_recovery_fee, loan_status



Machine Learning Workflow



3. **Performing feature importance analysis.**
   Based on the feature importance process, the top five features that significantly influence the target (label) are as follows:

Feature	Importance
recoveries	0.437564
issue_d	0.156408
last_pymnt_amnt	0.152591
last_credit_pull_d	0.052840
last_pymnt_d	0.045524

4. **Splitting the data**
   into training and test sets, with 80% of the data used for training and 20% for testing.

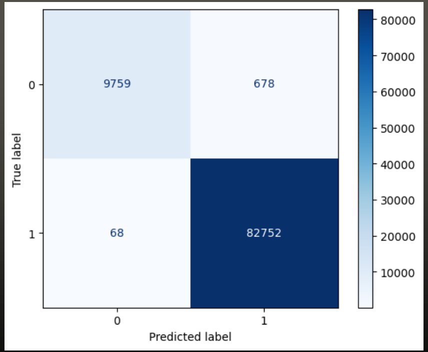
4. **Applying SMOTE**
   to address the class imbalance issue in the training data by generating synthetic samples for the minority class.

4. **Hyperparameter tuning**
   to identify the optimal combination of parameters that can enhance model performance, Grid Search CV is used.

4. **Cross-validation**
   to evaluate the model more robustly using 3-fold cross-validation.

Model Machine Learning

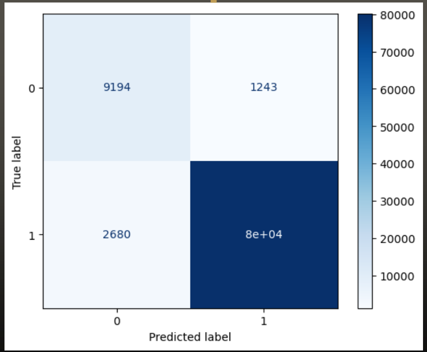
Random Forest



Model	Train	Test
Accuracy	0,9578	0,9579
Precision	0,9847	0,9847
Recall	0,9676	0,9676
F1-Score	0,9761	0,9761
ROC-AUC	0,9803	0,9800

Model Machine Learning

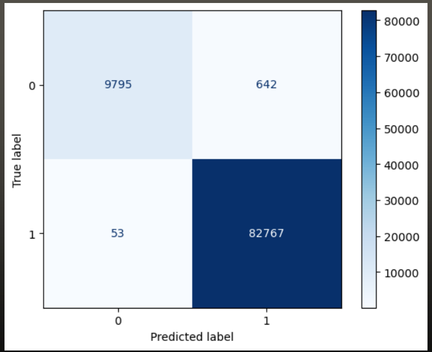
Logistic Regression



Model	Train	Test
Accuracy	0,9961	0,9920
Precision	0,9956	0,9919
Recall	1,0000	0,9992
F1-Score	0,9978	0,9955
ROC-AUC	1,0000	0,9931

Model Machine Learning

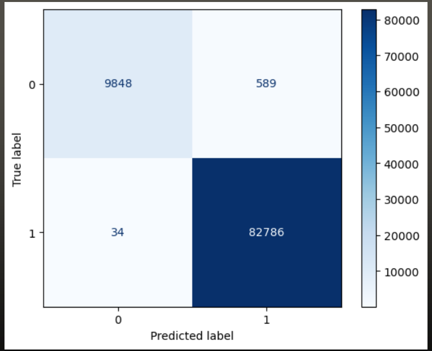
LGBM (Light GBM)



Model	Train	Test
Accuracy	0,9926	0,9925
Precision	0,9923	0,9923
Recall	0,9995	0,9994
F1-Score	0,9959	0,9958
ROC-AUC	0,9953	0,9950

Model Machine Learning

XGBoost



Model	Train	Test
Accuracy	0,9936	0,9933
Precision	0,9931	0,9929
Recall	0,9998	0,9996
F1-Score	0,9964	0,9963
ROC-AUC	0,9967	0,9956

Business Recommendation

Model ML

XGBoost stands out as the most effective machine learning model for credit prediction due to its ability to achieve an optimal balance between precision and recall. The results are as follows:

- **Accuracy:** 99.33%
- **F1 Score:** 99.63%
- **ROC-AUC:** 99.56%

These results have been rigorously validated using 3-fold cross-validation, meaning that the high scores reflect the model's strong predictive ability.

Business and Data Features

- ➢ The company needs to pay attention to data quality, especially for features that have a high correlation with the target, such as *recoveries*, *issue_d*, *last_pymnt_amnt*, and other related features. The data imputation process can be focused on these features to improve the quality of the analysis.
- ➢ Based on the analysis, it was found that the loan amount is only determined by whether the borrower meets the credit policy or not. This needs to be improved, as the determination of the loan amount should take other factors into account, so that the potential losses to the company can be minimized.
- ➢ The model developed is capable of handling credit prediction based on the available data. With the model in place, the company can maintain data quality to ensure it remains relevant to the company’s conditions.

Thank You !

Link Code Here !

Visit Me

Supported By : Rakamin Academy id/x partners slidesgo