ANALYZE AND PREDICT CREDIT RISK

ID/X PARTNER DATA SCIENTIST PROJECT BASED INTERNSHIP

CREATED BY WANDA ARMADIANTI

Business Understanding

Credit Risk

Credit Risk is the possibility that the borrower fails to fulfill their payment obligations in accordance with the terms of the loan agreement. This includes the potential inability of the borrower to pay loan installments, interest, or other financial obligations according to a predetermined schedule. Credit risk analysis aims to increase efficiency, accuracy and responsiveness in facing challenges related to credit risk management. In this case, credit risk prediction will be carried out using Machine Learning modeling for effectiveness.

Exploring Data

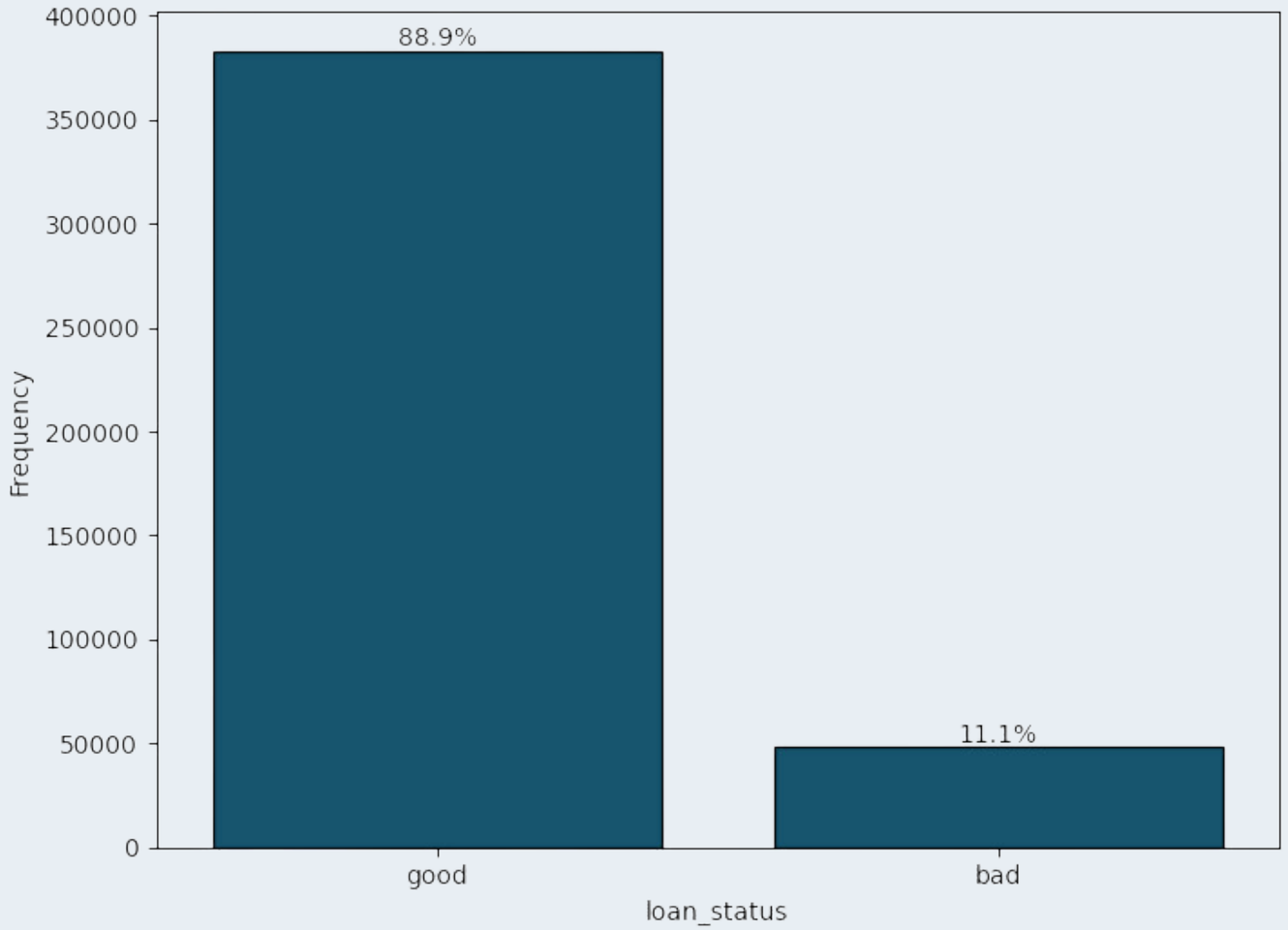
The data used is loan data provided by the company for 2007 to 2014 with 74 columns and 466.285 rows of data (before preprocessing data)

Column	Non-Null Count	Dtype
Unnamed: 0	466285	int64
id	466285	int64
member_id	466285	int64
loan_amnt	466285	int64
funded_amnt	466285	int64
funded_amnt_inv	466285	float64
term	466285	object
int_rate	466285	float64
installment	466285	float64
etc	etc	etc

Defining Target Column

The column targeted for building the model is 'loan_status' because it can show the credit performance of the borrower. This column consists of 9 values which will be relabeled to determine which are good loans and bad loans.

1. Good: Current, Fully Paid, In Grace Period, Does not meet the credit policy. Status: Fully Paid
2. Bad: Charged Off, Late (31-120 days), Late (16-30 days), Default, Does not meet the credit policy. Status: Charged Off



Preprocessing Data

Before the data is used in further statistical analysis or machine learning modeling, data preprocessing is carried out to prepare the data so that it can be used effectively by the model or algorithm that will be used.

01

Handle Missing Value

Fill in column mode values in 17 columns with missing values <50% and delete 22 columns with missing values >50%

02

Handle Unique Column

Removed 27 columns with number of unique values=1 or too many unique values

03

Delete Useless Column

Delete 3 columns that are not needed for further analysis

Preprocessing Data

Before the data is used in further statistical analysis or machine learning modeling, data preprocessing is carried out to prepare the data so that it can be used effectively by the model or algorithm that will be used.

04

Change Data Type

2 columns are changed to numeric and 3 columns are changed to datetime

05

Handle Duplicated Data

There is no duplicated data

06

Handle Outlier

Removing outliers using Z-score with threshold=3

Feature Analysis

Feature analysis helps understand the characteristics and interactions between features in a dataset.

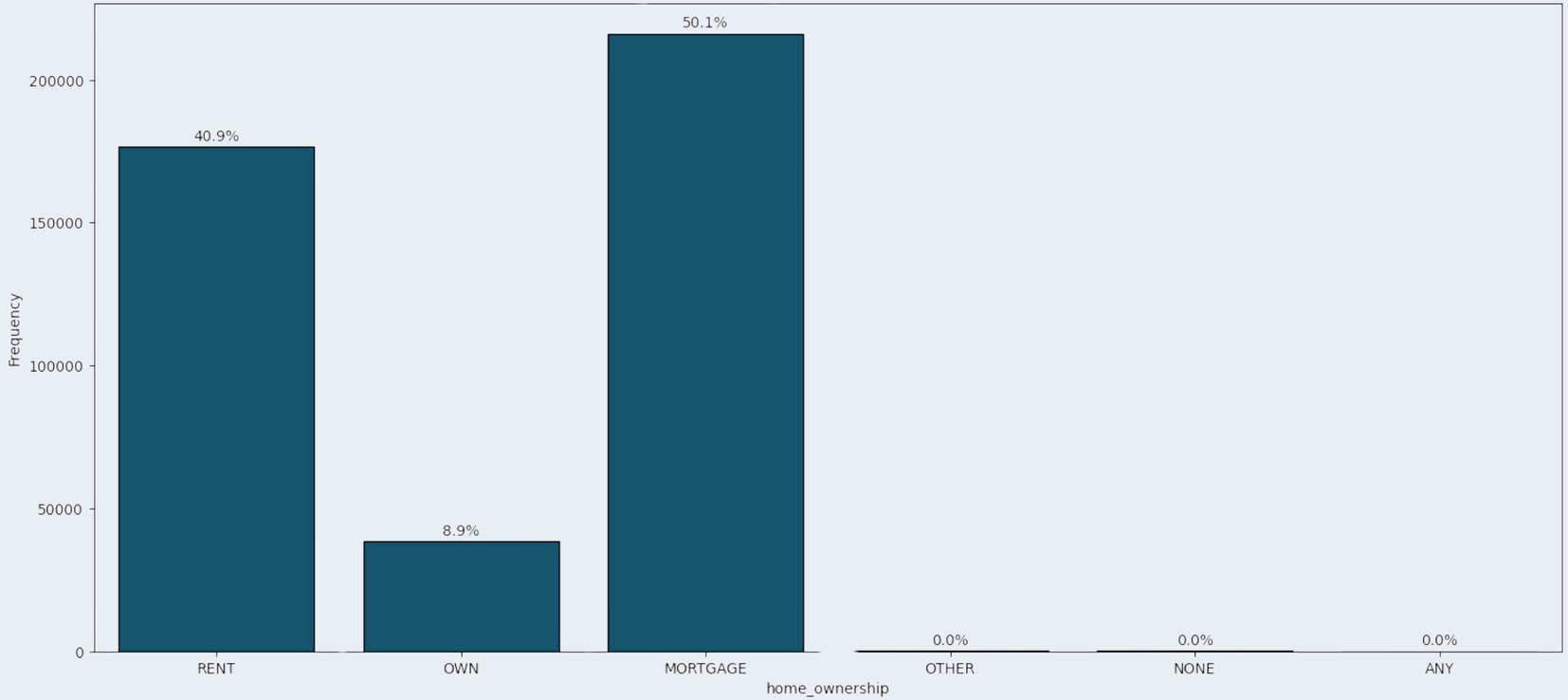
- Displays a bar chart for each categorical column
- Displays histogram bars for each numeric column
- Displays a line chart for each date type column

Univariate Analysis

- Displays the correlation of independent column pairs with a correlation plot and heatmap matrix
- Displays the correlation of independent and dependent columns using a stacked bar chart

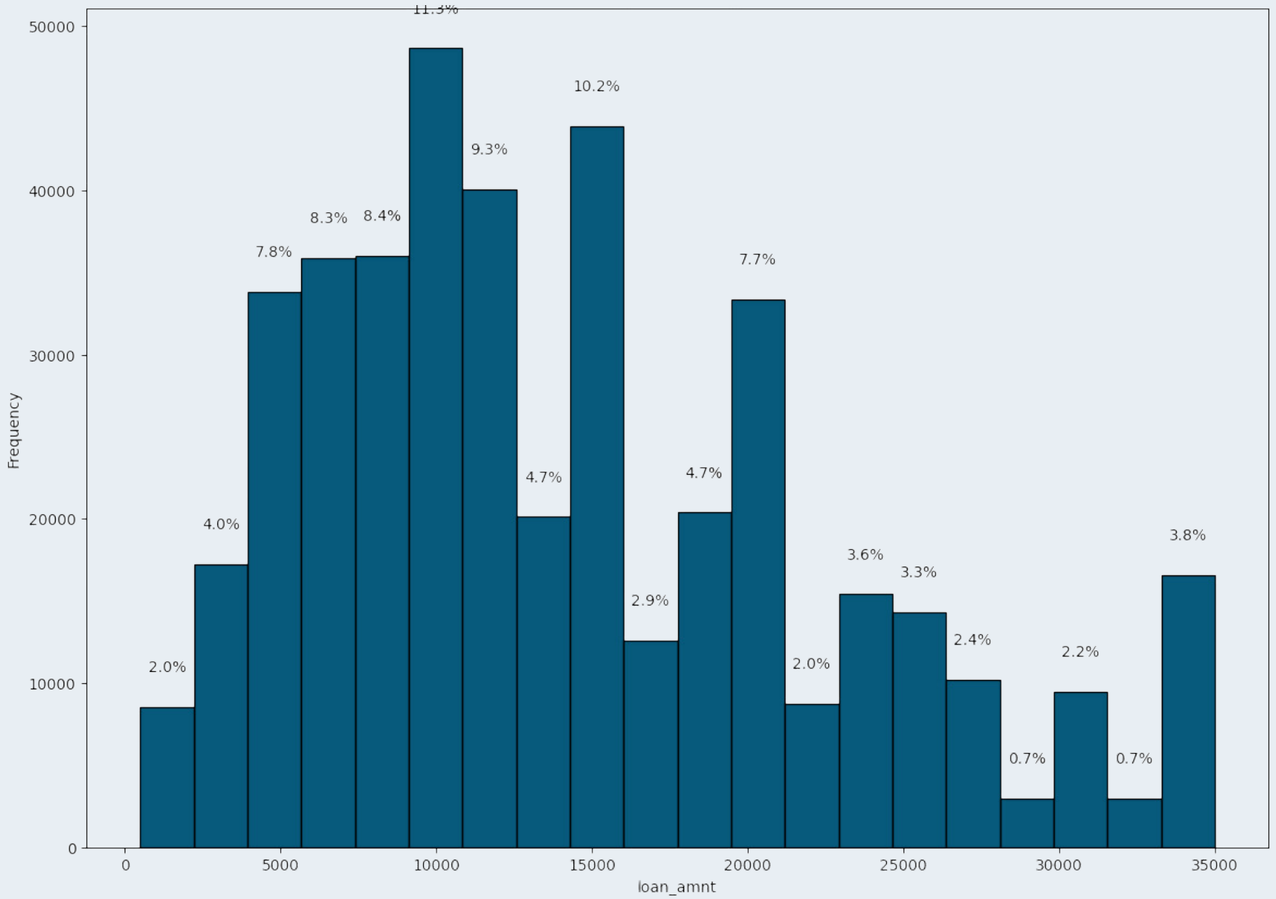
Bivariate Analysis

Home Ownership Distribution



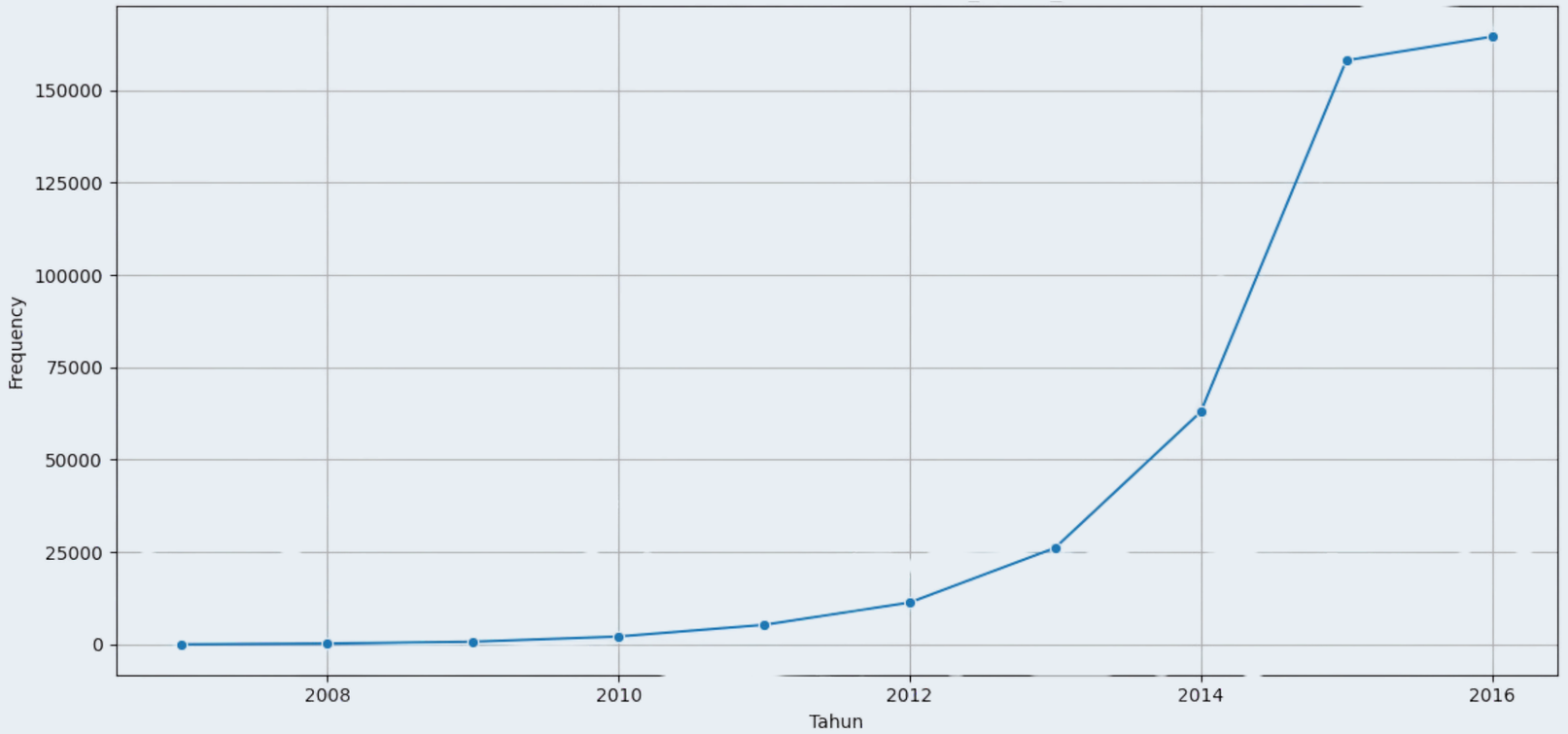
The distribution of home ownership statuses emerges as a critical factor influencing credit risk assessment. Notably, the data reveals that mortgage holders represent the majority, comprising 50.1% of total loan applicants. This finding underscores the significance of mortgage ownership in the lending landscape, suggesting a correlation between home ownership through mortgages and the propensity to seek financial assistance.

Loan Amount Distribution



The visualization of loan amount against frequency reveals a notable trend. Particularly, the highest frequency occurs within the loan amount range of $8000 to $12000, indicating a significant concentration of loan applications within this interval. This insight underscores the importance of understanding the distribution of loan amounts in assessing credit risk, allowing lenders to tailor their risk assessment models and strategies accordingly to mitigate potential financial vulnerabilities effectively.

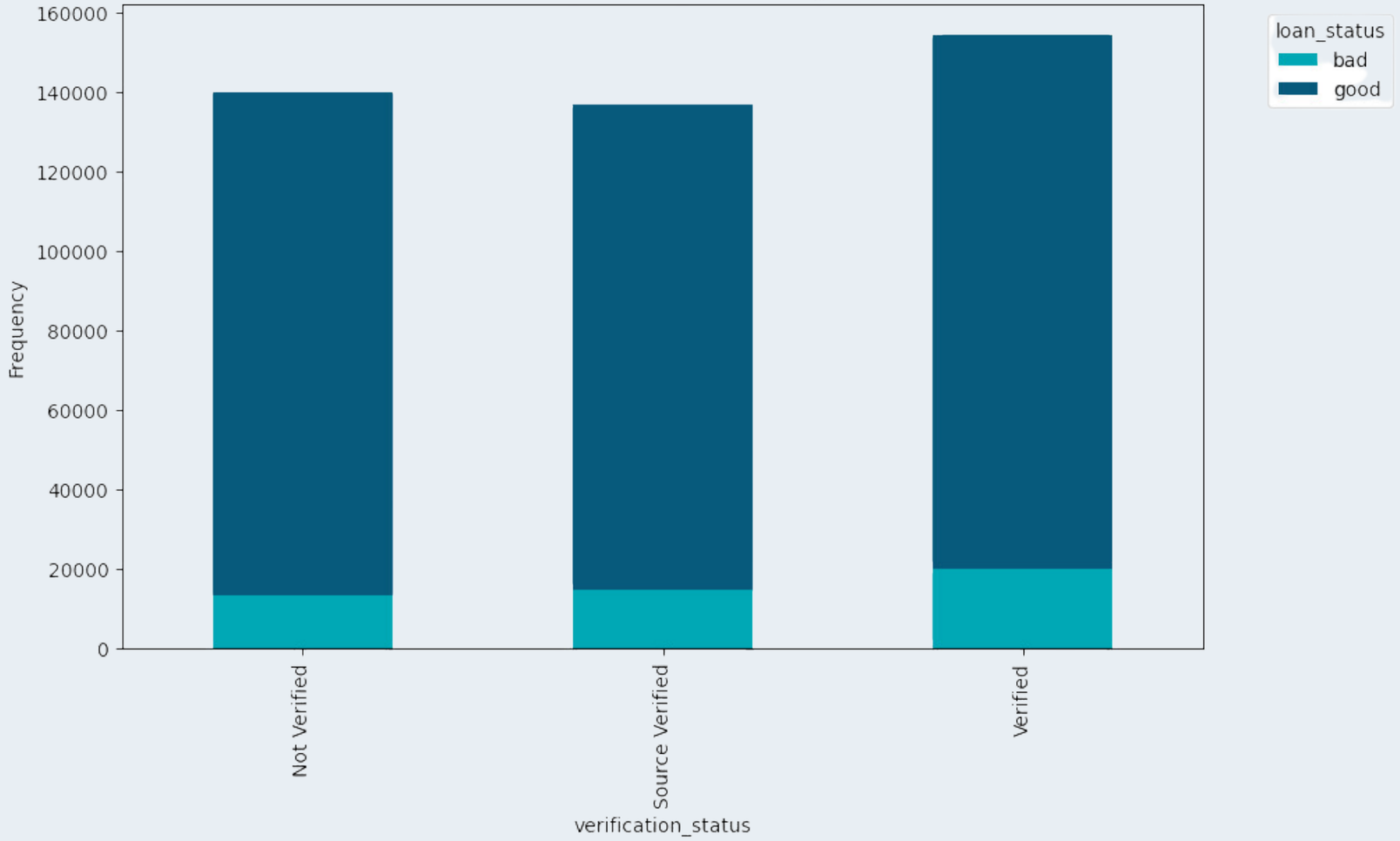
Loan Payment Trends from 2008-2016



The line chart depicts a notable upward trend, particularly showcasing a significant surge from 2014 to 2015. This drastic increase suggests a heightened frequency of loan payments in the latter year, indicating a potential improvement in borrowers' financial stability or repayment behaviors.

Loan status verification based on loan status

The data indicates that the highest frequency is associated with verified loan applications, particularly those categorized as "good" status, comprising approximately 85%. This finding underscores the significance of thorough verification processes in mitigating credit risk, suggesting that verified applicants with favorable loan statuses contribute significantly to a more robust lending portfolio.



Labeling

Label Encoding Overview

	loan_amnt	int_rate	grade	home_ownership	verification_status	loan_status	pymnt_plan	purpose	addr_state	delinq_2yrs	inq_last_6mths	open_acc	p
0	5000	10.65	1	5	2	1	0	1	3	0.0	1.0	3.0	
2	2400	15.96	2	5	0	1	0	11	14	0.0	2.0	2.0	
3	10000	13.49	2	5	1	1	0	9	4	0.0	1.0	10.0	
4	3000	12.69	1	5	1	1	0	9	36	0.0	0.0	15.0	
5	5000	7.90	0	5	1	1	0	13	3	0.0	3.0	9.0	

Data Balancing

Oversampling

Before

1: 383259 data

0: 48000 data

After

1: 383259 data

0: 383259 data

Classification Model

The classification model is used to predict whether the loan is classified as good or not by considering the borrower's profile. Before creating the model, the scale of the numerical data is equalized using a standard scaler. The data is split into **70% train data** and **30% test data**. Among the 5 classification models used, the best performance for making predictions was shown by the **Random Forest** model.

f-1 score: 0,73
ROC-AUC: 0,73

Logistic Regression

f-1 score: 0,80
ROC-AUC: 0,80

Neural Network

f-1 score: 0,96
ROC-AUC: 0,96

Decision Tree

f-1 score: 0,97
ROC-AUC: 0,97

Random Forest

f-1 score: 0,89
ROC-AUC: 0,89

KNN

THANK YOU