Rakamin Academy X id/x partners

# Prediction Model

## ID/X Partners - Data Scientist

Presented by
Hafizh Fadhl Muhammad

Rakamin Academy

# Hafizh Fadhl Muhammad

I am a third-semester Informatics Engineering student at Padjadjaran University (Unpad) with a strong interest in Artificial Intelligence and Data Science. My academic journey is driven by a passion for leveraging technology to solve real-world problems, particularly through the innovative application of AI and data-driven insights.


Indonesia, West Java

hafizhfadhl22@gmail.com

https://www.linkedin.com/in/hafizhfadhlm/

# Project Portfolio

**This project serves as final task for data scientist project based internship program hosted by Rakamin x ID/X Partners. This project aims to create a machine learning model from loan data provided by companies consisting of accepted and rejected loan data which is expected to be able to predict credit risk. The dataset is collected by ID/X Partners from a company**

**Link code here!**

# 1. Data Understanding

What should we check :

- The number of rows and columns

  The dataset has 466,285 rows and 74 columns
- The data type for each columns

  The dataset has 22 categorical columns and 52 numerical columns
- Sample data
- Number of duplicated data

  The dataset don’t have duplicated rows

This step aims to make us understand more about the dataset that we will process.

# 2. Feature Engineering

At this stage, unimportant columns will be cleaned in the hope that it will reduce modeling time later. The following are the criteria for selecting columns :

Active columns (not 100% missing values)

Not a categorical data with high unique values (e.g emp_title, title, zip_code)

Not a highly unbalanced class (e.g pymnt_plan with a ratio of 99.9 : 0.1)

Not a column containing text (e.g url, desc)

Not a column with high missing value (e.g mths_since_last_delinq), we limit to columns with < 50% missing values.

Not a numerical data with high correlation (above 0.7)

Right now, we should have cleaned 42 columns and leaving 32 columns to worked on.

# 3. Exploratory Data Analysis

## Explain Your Result Here

<You should **explain your strategy** on this page and you can add **image or link** result. You can add an explanation of how you got the result also. Duplicate this slide if you need more space to provide explanations>

<Please complete it with the visualization of the data you have created.>

# 4. Data Preparation

Rakamin Academy

Data Transformation

Mostly, our data transformation steps will be applied to categorical data. There are some data that can be transformed in the same way. term, grade, sub_grade, emp_length would be transformed into numerical data by classifying some values into number or by removing the string. earliest_cr_line, issue_d, last_pymnt_d, next_pymnt_d, last_credit_pull_d would be transformed into regular date format. Then we will add 2 new features, pymnt_time and credit_duration. pymnt_time : this is the distance in months from the last_pymnt_d to the next_pymnt_d. credit_duration : this is the number of year between last_credit_pull_d and earliest_cr_line.

# 4. Data Preparation

Data Transformation

home_ownership values will be merged ("ANY" and "NONE" to "OTHER") to reduce the result of one hot encoding later. For loan_status, it will be left for now because it will be defined as our target variable later. With this, our remaining categorical data are : home_ownership verification_status purpose initial_list_status

# 4. Data Preparation

Defining Target Variable

We need to process loan_status to binary values for our model training later. So let's first classify the values into two groups (good loan and bad loan) based on wheter it is risky or safe. Current : Safe Fully Paid : Safe Charged Off : Risk Late (31-120 days) : Risk In Grace Period : Safe Does not meet the credit policy. Status:Fully Paid : Safe Late (16-30 days) : Risk Default : Risk Does not meet the credit policy. Status:Charged Off : Risk Then we labeled good loan as 1 and bad loan as 0.

# 4. Data Preparation

Handling Missing Values

Since some of the categorical data have transformed into numerical data and have the missing values filled, we will now just focusing on numerical data. There are 27 columns with numerical data type that still have missing values. We will fill the missing values with mean value from each columns. With the help of SimpleImputer from library sklearn we will have the missing values filled in no time.

# 4. Data Preparation

Rakamin Academy

Data Encoding

All the encoding applied to categorical columns using one hot encoding which are applied to: home_ownership verification_status purpose initial_list_status This is the end for data preparation steps, and right now we have 50 columns for training models.

# 5. Data Modeling

To start the model training, we need to determine what classification algorithm we will use to make the model later. To do this, we will need to test several classification algorithm and see which one get the accuracy better.

We choose 6 different algorithm to be tested for our model. All classification model are trained using preprocessed data and default setting for hyperparameter.

We examined which algorithm has the best performance using our preprocessed data (without balancing classes). We choose accuracy, number of mislabeled, Receiver Operating Characteristics (ROC), and Kolmogorov-Smirnov (KS) as a variable for our consideration.

# 6. Evaluation

More exploration and more understanding can be provided if there is more time to work this project on

The model still have room for improvement by using ensemble method as comparison.

The hyperparameter tuning should be done by detailed observations and lots of experiments

Detailed business understanding should help understanding the data acquisition

# 7. Conclusion

So in conclusion, our selected model is XGBoost Classifier. The important features of our model are :

* recoveries : Indicates if a payment plan has been put in place for the loan
* pymnt_time : Number of months between last_pymnt_d and next_pymnt_d
* out_prncp : Remaining outstanding principal for total amount funded
* total_rec_late_fee : Late fees received to this date
* term : The number of payments on the loan. Values are in months and can be either 36 or 60.
* initial_list_status_f : The initial listing status of the loan. Possible values are – Whole, Fractional (in this case it is Fractional)
* int_rate : Indicates if income was verified by LC, not verified, or if the income source was verified

# Thank You