Rakamin Academy X id/x partners

Prediction Model for Credit Risk

ID/X Partners Data Scientist

Project Based Internship Batch Mei 2024

Presented by
Chianti Suksmarani Ridhwan

Chianti Suksmarani Ridhwan

About You

A fresh graduate with a bachelor's degree in management, I possess hands-on experience as a marketing intern and administrator. Keen on learning and highly committed, excel in time management. I am particularly interested Data Science/Analytics. Currently, I am focusing on developing my skills in data analytics and have gained proficiency in utilizing tools like Microsoft Excel and SQL to support my learning process.

Insert Your Experience

- SkytreeDGTL, 2023
  Event Manager
- PT Miskat Alam, 2022-2023
  Professional Conference Organizer (PCO)
- PT Innerindo Dinamika, 2022-2023
  Professional Conference Organizer (PCO)

Case Study

Mengembangkan *model machine learning* yang dapat memprediksi risiko kredit (*credit risk*) pada perusahaan pemberi pinjaman (multifinance)

Goals

Meningkatkan keakuratan dalam menilai dan mengelola risiko kredit, sehingga dapat mengoptimalkan keputusan bisnis mereka dan mengurangi potensi kerugian.

Objective

Mengembangkan *model machine learning* yang dapat memprediksi risiko kredit (*credit risk*) berdasarkan dataset yang disediakan.

Rakamin Academy

Data Understanding

Melakukan eksplorasi awal

Ringkasan Struktur Dataset

Mengidentifikasi Atribut Data

After understanding the data, it was found that:

- The dataset has 466285 rows and 75 columns
- There are several columns that have missing values
- issue_d, earliest_cr_line, last_pymnt_d, next_pymnt_d, and last_credit_pull_d, will be converted into datetime
- No duplicate data

Exploratory Data Analysis

Membuat visualisasi data dan menganalisis korelasi antar fitur

Descriptive Statistics

Observation from both numerical and categorical descriptive statistics:

- Unnecessary features such as features that have only one or equal to the number of rows unique values will be removed
- Features that have high cardinality such as emp_title, title, desc features will be removed as well as zip_code

Target Variable

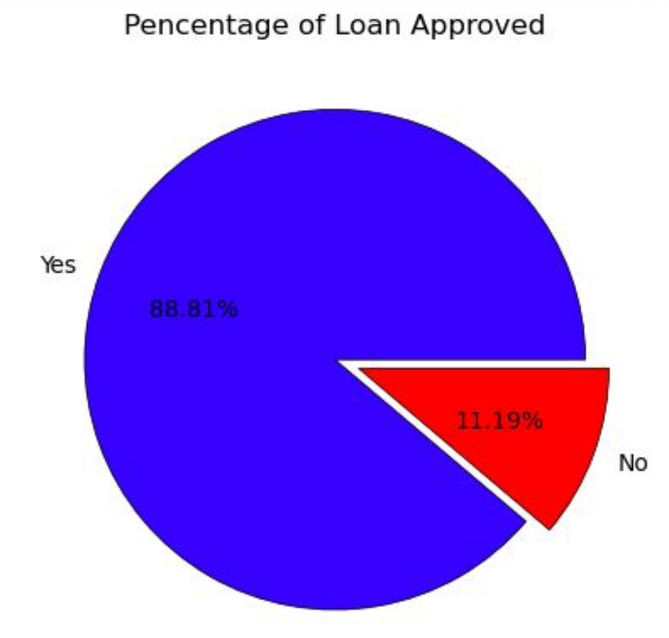
To predict credit risk, loan_status column will be divided into two category:

Bad Loan :

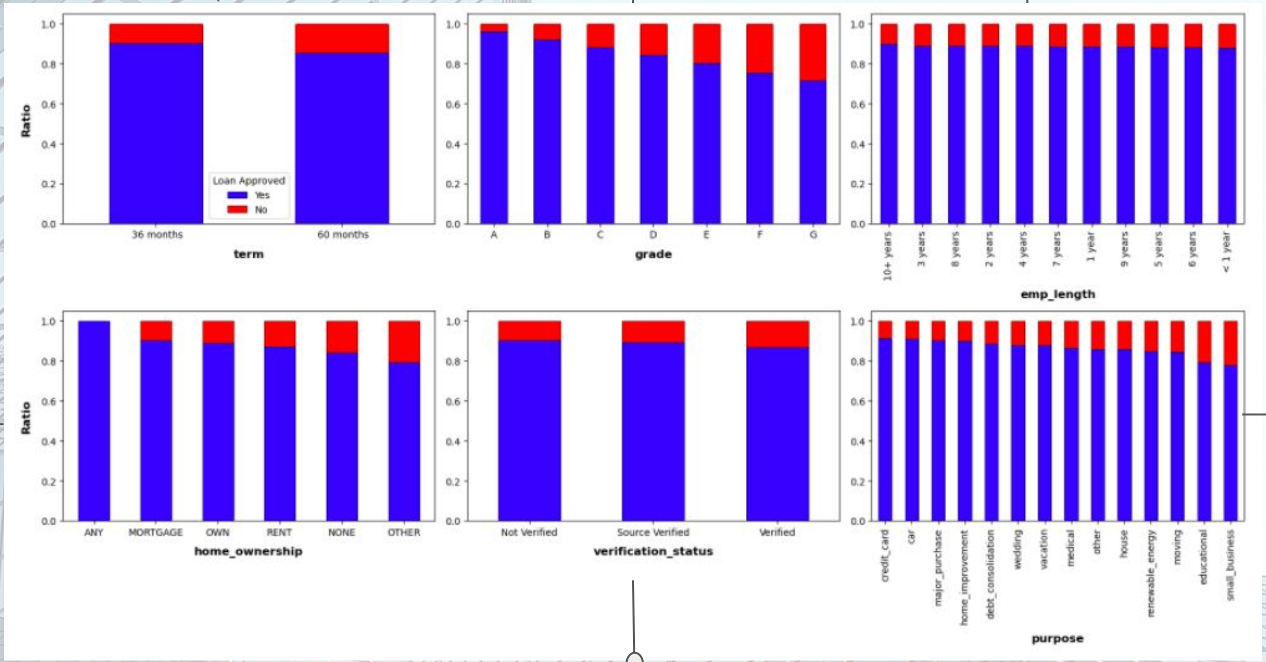
- Default
- Charged Off
- Late (31-120 days)
- Late (16-30 days)
- Does not meet the credit policy. Status:Charged Off

Good Loan :

- Fully Paid
- Current
- In Grace Period
- Does not meet the credit policy. Status: Fully Paid



The Ratio of Loan Approved Based on Term, Grade, Employment Length,Home Ownership, Verification Status, & Purpose



Grade G merupakan grade dengan bad credit tertinggi.

Semakin lama jangka waktunya (term) semakin tinggi bad credit.

Masa kerja > 10 tahun memiliki bad credit yang semakin rendah.

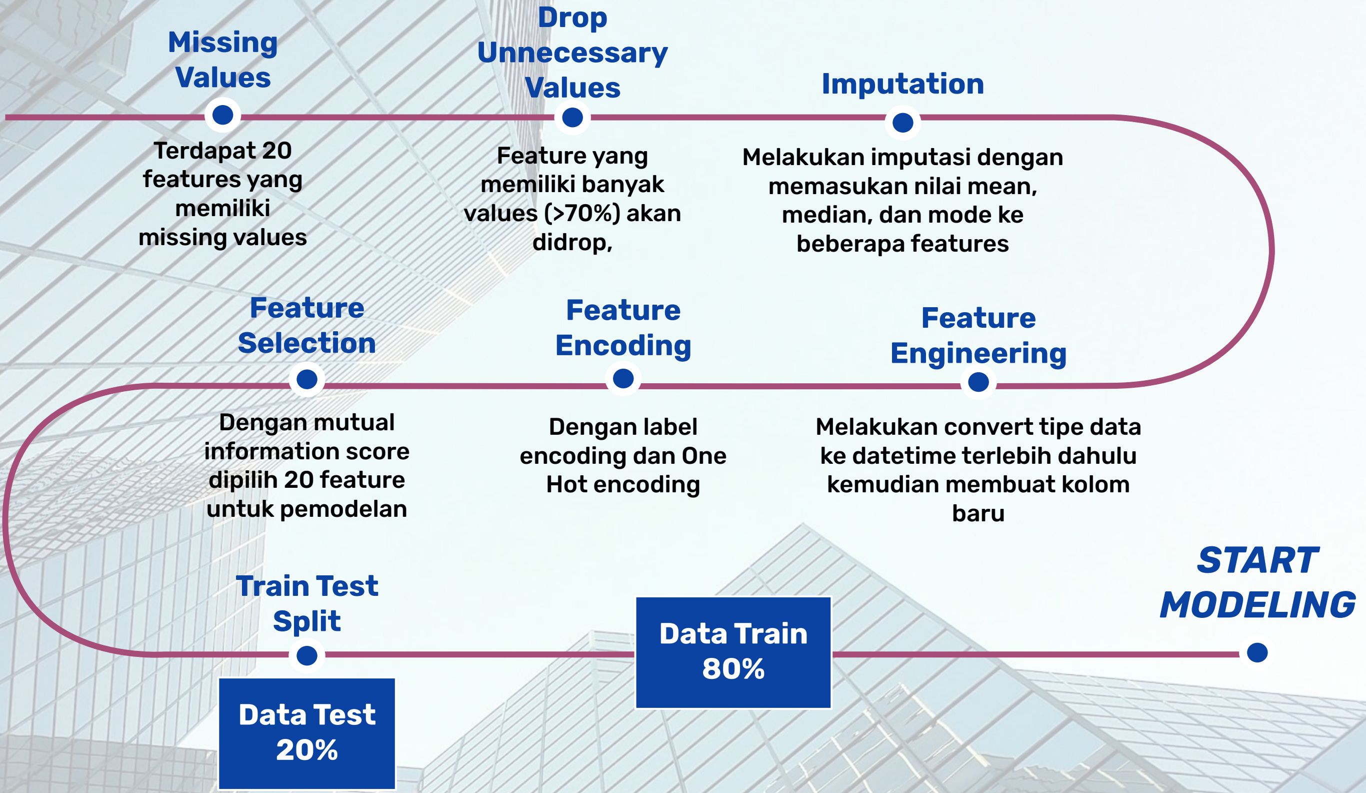
Kepemilikan rumah KPR dan Sendiri menunjukkan good credit yang paling tinggi.

Small business memiliki probabilitas bad credit yang paling tinggi.

Pendapatan dengan status Terverifikasi justru memiliki rasio kredit macet paling tinggi.

Pre-Processing Flow

Rakamin
Academy



Missing Values

Terdapat 20 features yang memiliki missing values

Drop Unnecessary Values

Feature yang memiliki banyak values (>70%) akan didrop,

Imputation

Melakukan imputasi dengan memasukan nilai mean, median, dan mode ke beberapa features

Feature Engineering

Melakukan convert tipe data ke datetime terlebih dahulu kemudian membuat kolom baru

Feature Encoding

Dengan label encoding dan One Hot encoding

Feature Selection

Dengan mutual information score dipilih 20 feature untuk pemodelan

Train Test Split

Data Test 20%

Data Train 80%

START MODELING

Data Modelling

Rakamin Academy

Metrik evaluasi Recall sangat penting dalam konteks risiko kredit karena perusahaan ingin menjaring sebanyak mungkin pelanggan yang berisiko tinggi (yang mungkin gagal bayar).

Model	Precision (Train)	Precision (Test)	Recall (Train)	Recall (Test)
Gradient Boosting	0.97808	0.97718	0.99963	0.99955
Random Forest	0.99998	0.98288	1.00000	0.99941
Logistic Regression	0.97370	0.97314	0.99801	0.99778
Decision Tree	1.00000	0.99132	0.99999	0.99202

Hasil

Dari nilai recall di atas, **Gradient Boosting memiliki nilai recall tertinggi** pada data pengujian dengan nilai sebesar 0.99955 yang berarti model ini menangkap hampir seluruh true positif pada data pengujian.

Rakamin Academy

Evaluation

Melakukan evaluasi kinerja model

Evaluasi

Gradient Boosting adalah model terbaik berdasarkan metrik recall pada data pengujian. Meskipun perbedaannya kecil dibandingkan Random Forest, Gradient Boosting juga menunjukkan performa yang lebih seimbang dan potensi overfitting yang lebih rendah. Oleh karena itu, Gradient Boosting tidak hanya memberikan recall tertinggi namun juga memiliki generalisasi yang baik.

Thank You

Rakamin Academy X id/x partners