CREDIT RISK PREDICTION

Nova Zidane Ibrahim

DATA SCIENTIST

PROJECT BASED INTERSHIP PROGRAM

2023

Table of Content

01

Problem Research

Business Understanding

In this case, the lending company has to make a decision whether to approve or reject the loan application based on the applicant's profile.

1. **Good Risk** refers to a situation where the loan applicant has a high probability of repaying their loan.
2. **Bad Risk** refers to a situation where the loan applicant has a low probability of repaying their loan.

Problem Statement

Lending to applicants with Bad Risk is the biggest cause of financial loss. Credit losses are the amount of money lost by lenders when applicants refuse to pay or run away with money they should have paid.

Objectives

01. Identify Patterns Indicating Bad Risk

02. Implementation of Machine Learning Algorithms to Build Predictive Models

02

Data Processing

Loan Data

75 Features

466,285 Rows

Row Data

Data Understanding

Check Missing Values, Unique Values, Data Type, and Statistical Summary for Each Features

Data Preparation

EDA → **Data Cleansing**

Data Preprocessing

Encoding
using Manual Encoding, One-Hot-Encoding, dan Label Encoding

→ **Feature Selection**
using Pearson Correlation and visualize with Heatmap

→ **Handling Imbalanced Data**
using Oversampling

↓ **Splitting**
using Ratio 80:20

Modeling and Evaluation

Learning Algorithms:

1. Decision Tree Classifier
2. Random Forest Classifier
3. Logistic Regression
4. Gaussian Naive Bayes
5. XGBoost Classifier

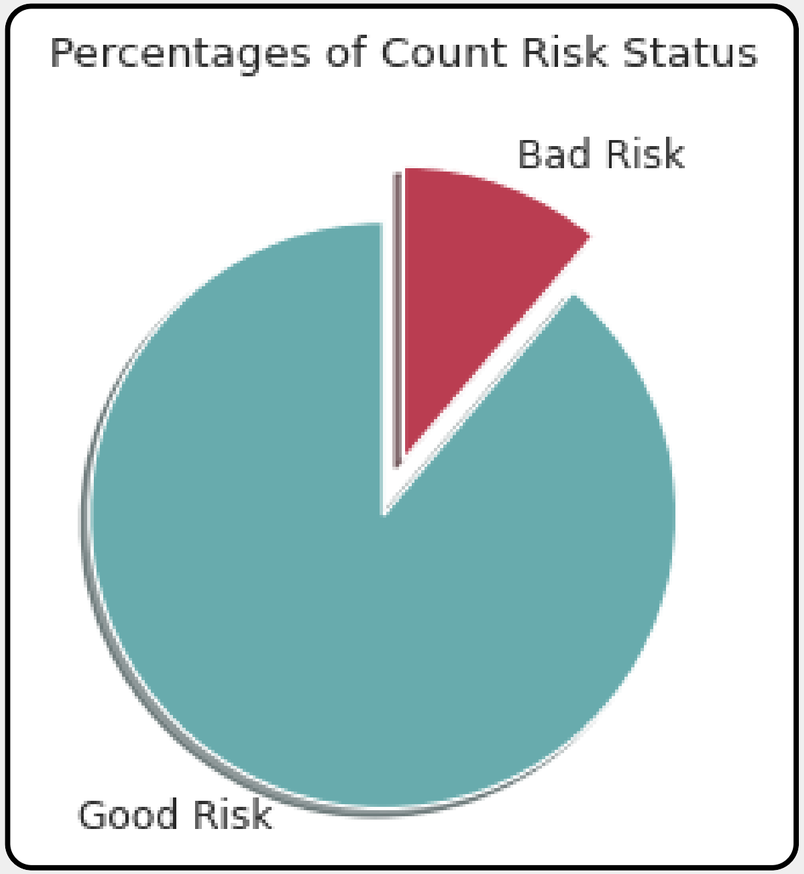
Evaluation Methods:

1. Confusion Matrix
2. Accuracy

03

Data Insight

Target Variable



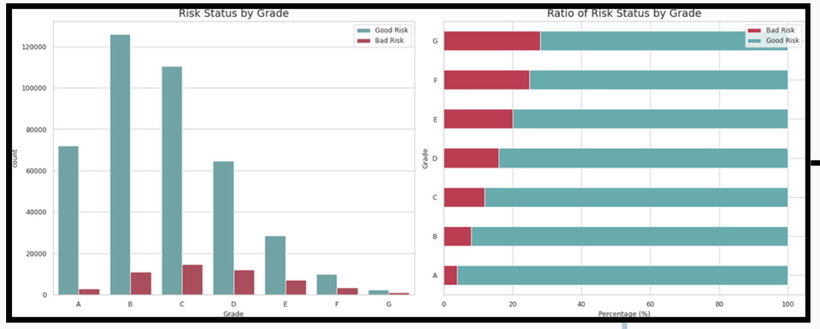
From loan_status, we can cluster the unique values into 2 risk statuses:

1. **Good Risk :** Fully Paid, Current, and In Grace Period.
2. **Bad Risk** **:** Late, Default, and Charged Off,

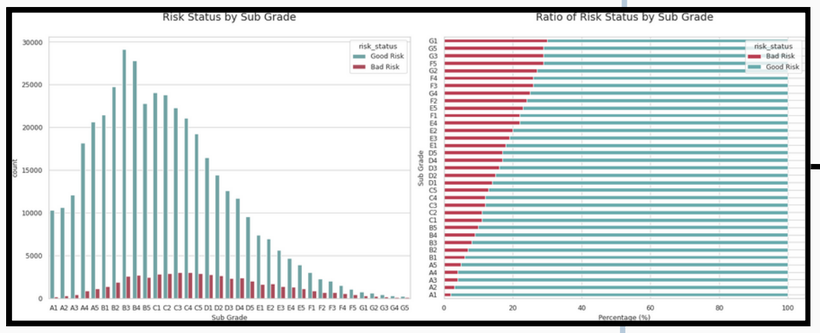
The dataset has significant data imbalance:

1. **Good Risk** : 414,099 (88.81%)
2. **Bad Risk** : 52,186 (11.19%)

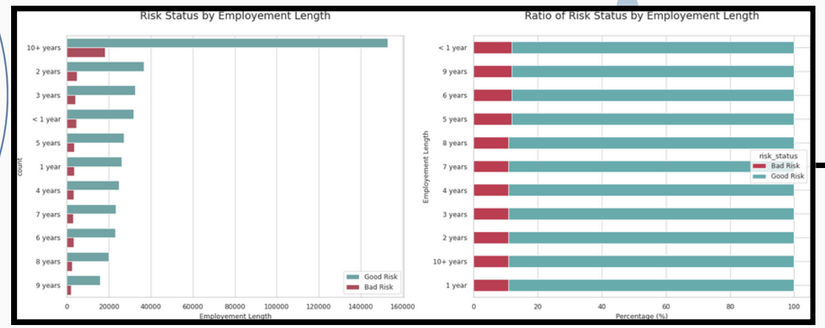
Categorical Variables



Applications with a **Low Grade** are more likely not to repay the loan.

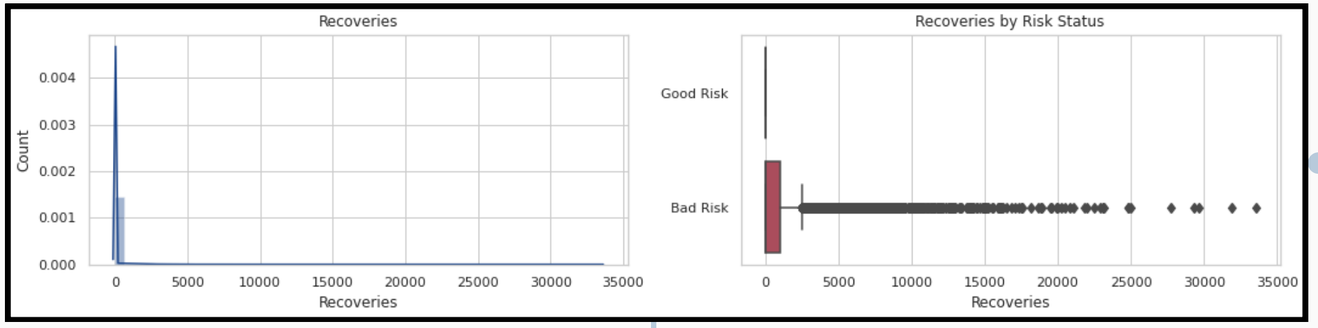


Applications with a **Low Sub Grade** are more likely not to repay the loan.

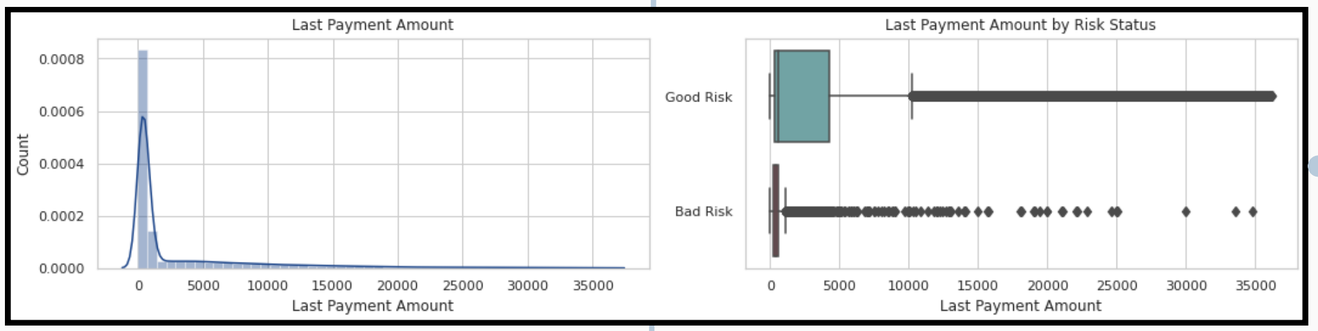


Employement Length does not influence the Applications for risk.

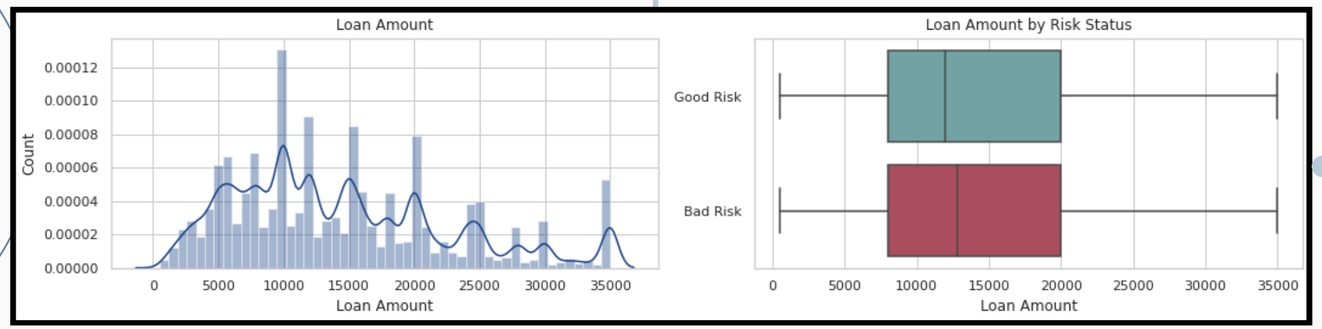
Numerical Variables



Applications **without Recoveries** are most likely to pay off the loan.



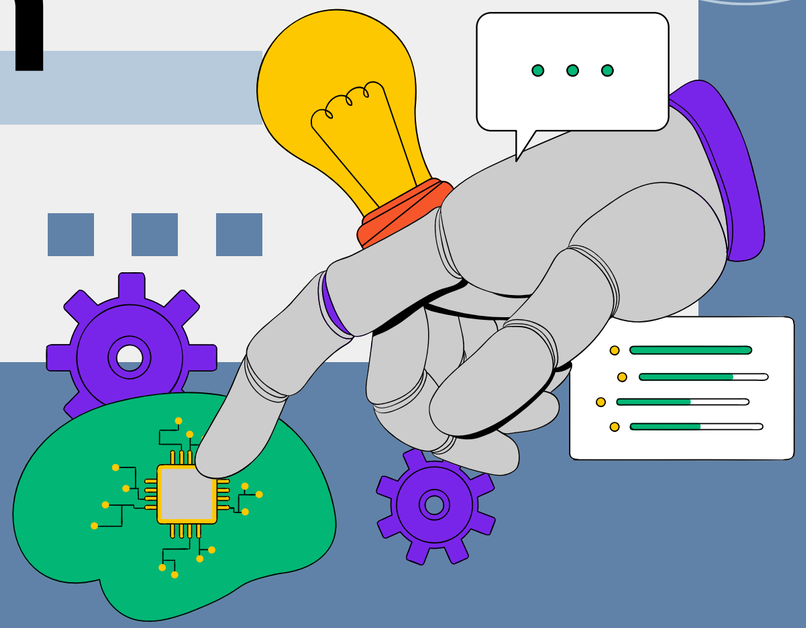
Applications with a **Low Last Payment Amount** are more likely not to repay the loan.



Loan Amount does not really affect the Applications for risk.

04

Modeling and Evaluation



Model Comparison

Algorithms	Training Accuracy	Testing Accuracy	Error Margin
Decision Tree	99.98%	98.82%	1.16%
Random Forest	99.98%	99.35%	0.63%
Logistic Regression	87.92%	88.00%	0.08%
Gaussian Naive Bayes	75.19%	75.22%	0.03%
XGBoost Classifier	95.90%	95.94%	0.04%

- The best model to predict the risk status of loan applications is **Random Forest**.
- Although Decision Tree and XGBoost also have very high accuracy, they have quite a small difference value from Random Forest. It is seen from the training and testing results that the Random Forest model is better than both in general.

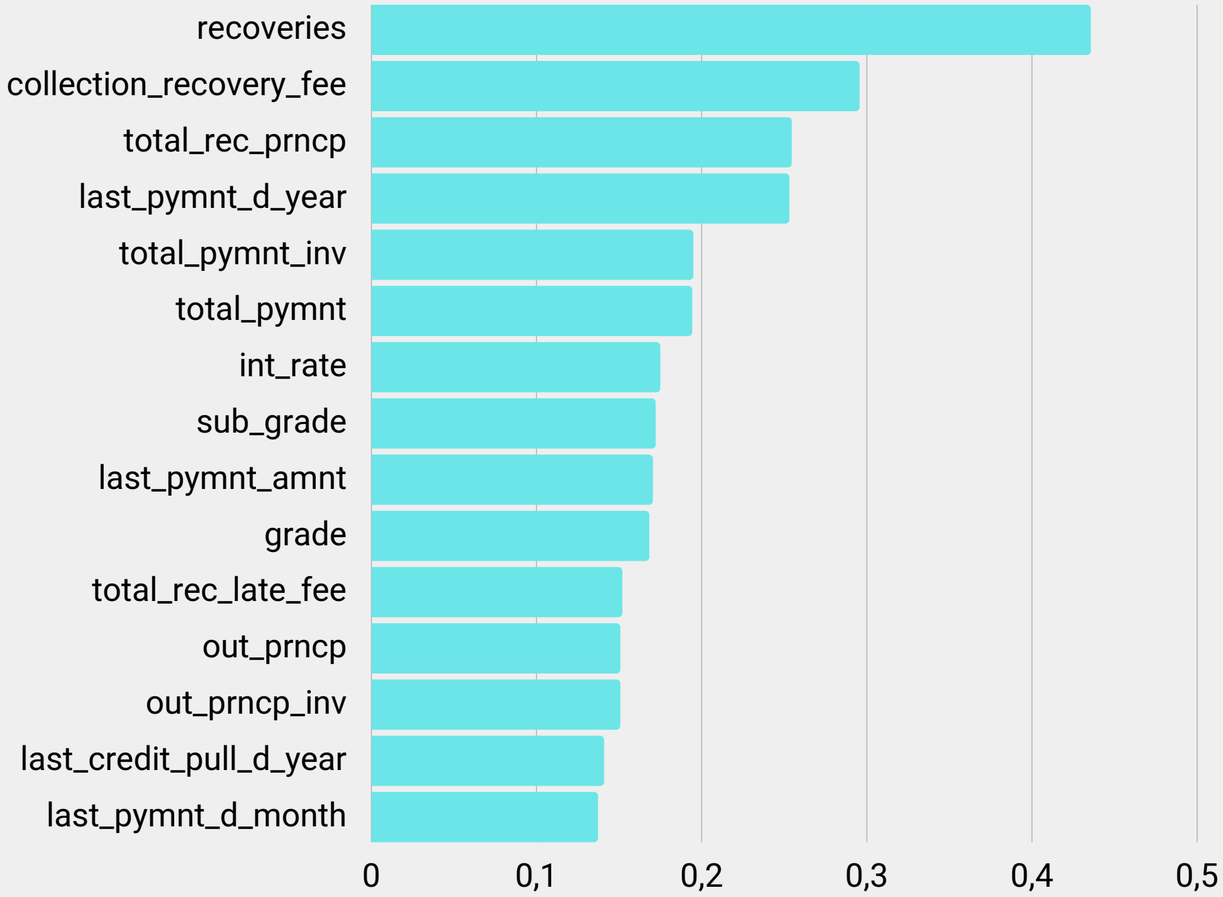
Features

This model only uses 15 features that correlate to risk_status.

Performance

The Random Forest model has a testing accuracy of 99.35%.

The Random Forest model has an error margin of 0.63%.



05

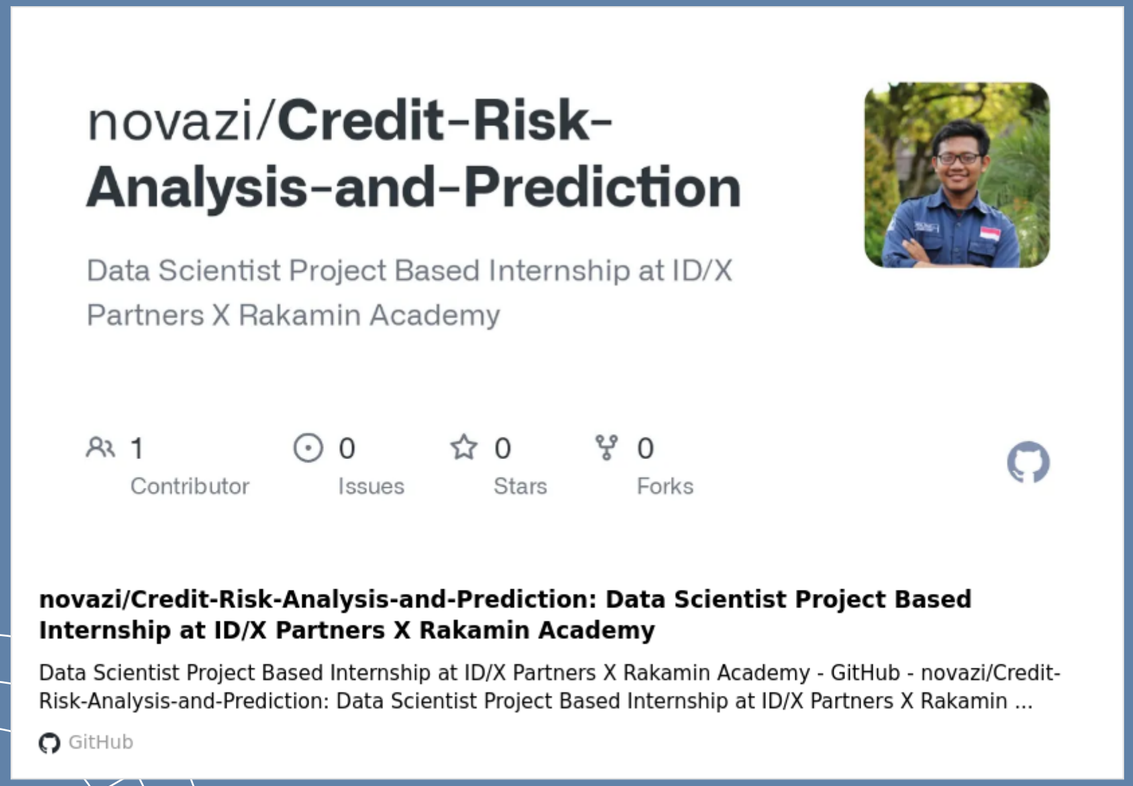
Business Recommendation

Conclusion

- The five most correlated features in determining the likelihood of loan repayment include `**recoveries**`, `**collection_recovery_fee**`, `**total_resc_prncp**`, `**last_pymnt_d_year**`, and `**total_pymnt_inv**`. These features are the most correlated in assessing loan repayment risk.

- The recommended strategy for dealing with applicants who exhibit high-risk indicators is that if an applicant has characteristics associated with a high risk of not repaying the loan, the company should consider actions such as rejecting their loan application, reducing the loan amount, or charging a higher interest rate.

My Project File



https://github.com/novazi/Credit-Risk-Analysis-and-Prediction

Terima Kasih