CREDIT RISK LOAN

Application of Machine Learning for Credit Risk Prediction in Consumer Loans

by M. Razy Qarar Fairuzzabadi

OUTLINE

- Project Overview

- Exploratory Data Analysis (EDA)

- Machine Learning Modelling

- Recommendation

PROJECT OVERVIEW

Link Full Code on GitHub

Background

Credit risk is one of the most important things that can affect the **stability and profitability** of a financial institution. If the loan is not paid as agreed, the institution can suffer a huge loss. Therefore, it is important for financial institutions to be able to **recognize and manage credit risk** well, especially during **uncertain economic conditions**.

Traditional approaches to assessing creditworthiness, such as **fixed rule-based** credit scores, tend to lack the flexibility to handle the complexity of various customer data. It also tends to be difficult to recognize **hidden patterns** or **early signs** indicating **default risk**. Therefore, a more **modern and flexible approach** is needed to support a **more accurate credit scoring** process.

Machine learning provides a modern approach that enables **in-depth data analysis** by utilizing large amounts of **historical data**. These algorithms can detect **patterns in customer behavior**, estimate the **probability of default**, and improve **decision-making accuracy**. By building reliable **predictive models**, financial institutions can **reduce exposure** to **credit risk** and **strengthen loan management** systems.

Problem Statement 1

A 12,29% of the total 460 thousand customers are still in default, indicating a high level of credit risk in the loan portfolio.

Goal 2

The objective of this project is to reduce the number of customer defaults by improving the accuracy of credit risk assessment using machine learning models.

Objective 3

Develop machine learning models to identify potentially defaulting customers, analyze factors that affect credit-worthiness, and evaluate model performance using relevant metrics.

Business Metrics 4

The objective of this project is to reduce the number of customer defaults by improving the accuracy of credit risk assessment using machine learning models.

Link Full Code on GitHub

EXPLORATORY DATA ANALYSIS

EDA

Link Full Code on GitHub

Dataset Overview

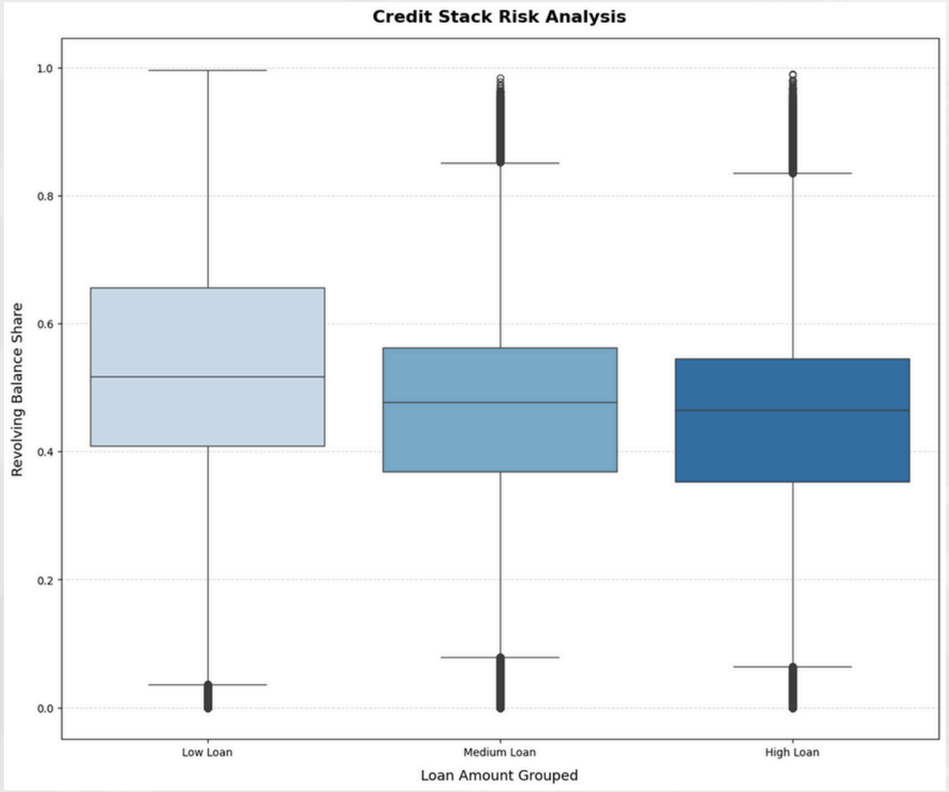
There are 466.285 rows and 70 columns

Target : 1 target variable

Features : 69 features

Link Full Code on GitHub

Credit Stack Risk Analysis



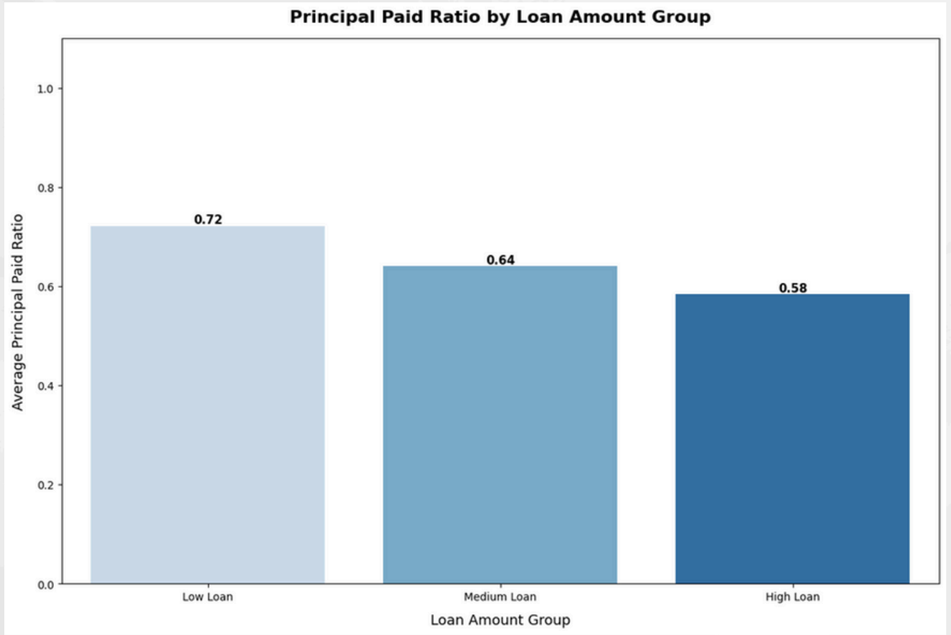
The analysis shows that the **larger the loan amount**, generally the **smaller the proportion of revolving debt** (RBS) in the borrower's total debt. This indicates that **borrowers with large loans** tend to have a **more controlled debt structure** and **rely less on consumptive debt facilities** such as credit cards.

While the **average RBS decreases as loan size increases**, there is a group of **borrowers with large loans** whose **RBS remains high**. This is an **indication of high risk**, as these borrowers are not only **carrying large loans**, but also have a **high reliance on short-term, high-interest debt**, which could be a sign of **liquidity stress** or **financial distress**.

On the other side, **borrowers with small loans show a wide variety of RBS**, so it is important **not to immediately assume that they are risk-free**. Therefore, a **thorough evaluation of their debt structure is still needed**.

Link Full Code on GitHub

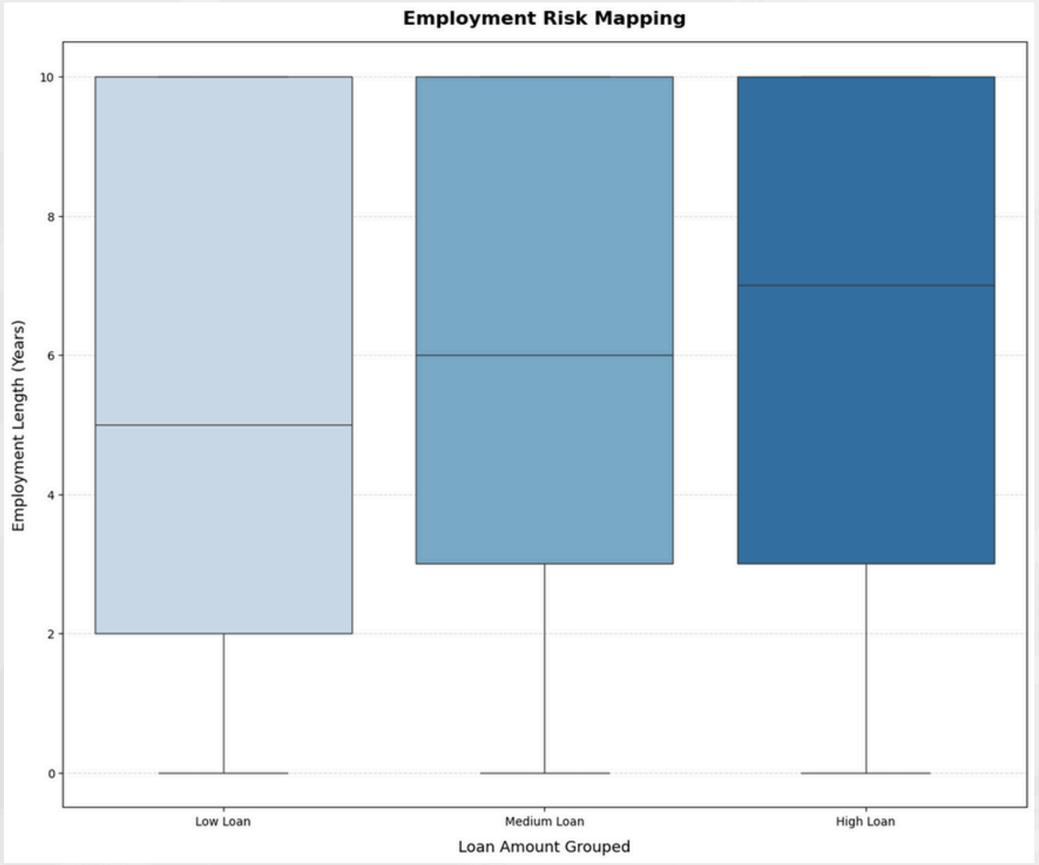
Principal Paid Ratio



The analysis shows that there is a decreasing trend in repayment progress as the loan amount increases. The **Low Loan** group had the **highest average principal repayment of 0,72**, followed by the **Medium Loan group with a ratio of 0,64**, and the **High Loan group recorded the lowest ratio of 0,58. This indicates that the larger the loan amount, the slower the repayment progress.**

Link Full Code on GitHub

Employment Risk Mapping



The analysis shows that borrowers with larger loan amounts (High Loan) tend to have longer tenure. This is reflected in the **higher median length of employment compared to the medium and low loan groups.** This trend indicates that, **in general, borrowers with large loans have a better level of employment stability**, which can be a **positive signal in evaluating creditworthiness.**

Although the large loan group appears to be generally stable, there are still **individuals in this category who have a low length of employment, even zero years. This condition is of particular concern** because **large loans granted to individuals with low employment stability** have the **potential to increase the risk of default.**

Link Full Code on GitHub

MACHINE LEARNING MODELLING

Link Full Code on GitHub

Data Preprocessing Workflow

1 Raw Data

The original dataset contains **466.285 rows and 70 columns of data**.

2 Check Duplicated Values

There is no duplicated data in the dataset.

3 Drop Missing Values

Delete columns that have entirely **unique values, are empty, have constant values, are redundant, or do not provide relevant information**.

4 Feature Engineering

Creating new columns, transforming the values of some columns, and **adjusting data types** to suit analysis and modeling needs.

5 Filling Missing Values

Fill the **missing values** in some columns using the **median value**, considering that the **median is more robust to outliers** than **the average**.

6 Pre-processing

Splitting the data into training and testing sets, **categorical feature transformation** with Label Encoding and One-Hot Encoding, **numerical feature standardization, relevant feature selection**, and **handling class imbalance** with oversampling techniques using SMOTE.

Modeling Workflow

1

Modeling without Hyperparameter Tuning

Running 5 machine learning algorithms, which are **Logistic Regression**, Extra Trees **Classifier**, **Random Forest**, **AdaBoost**, and **XGBoost**

2

Tuning Hyperparameter

Perform **hyperparameter tuning** using **cross-validation** of the **one best performing algorithm** to optimize the modeling results.

3

Tuned Model Comparison Model Evaluation

Perform a thorough in-person evaluation of the **selected model based on the best performance**.

Base Model Comparison

Model	Accuracy	AUC	Recall	Precision	F1-Score
Logistic Regression	82,60%	82,90%	89,80%	94,20%	91,90%
Extra Trees Classifier	91,50%	83,00%	97,10%	93,40%	95,20%
Random Forest	91,80%	83,80%	97,30%	93,60%	95,40%
AdaBoost	85,30%	82,10%	88,90%	94,00%	91,40%
XGBoost	92,40%	84,10%	98,10%	93,40%	95,70%

Of the **five algorithms used**, **XGBoost was chosen** because it **showed the highest accuracy of 92,4%**.

In addition, the **Recall value** of the model used is also high, which is an **important aspect in the context of credit risk**, given that **misclassifying a bad loan** as a good loan can lead to a **significant risk of loss**.

Link Full Code on GitHub

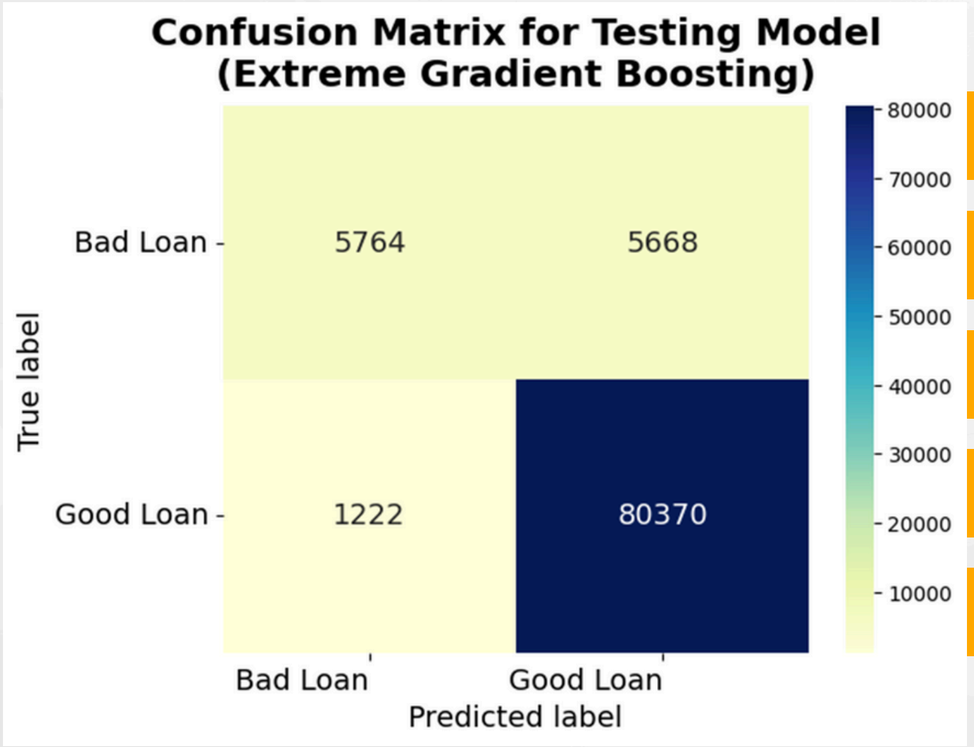
Hyperparameter Tuning

Model	Accuracy	AUC	Recall	Precision	F1-Score
Base XGBoost	92,40%	84,10%	98,10%	93,40%	95,70%
Tuned XGBoost	92,60%	84,20%	98,50%	93,40%	95,90%

Hyperparameter tuning improved model performance, although the improvement was **not significant**.

Link Full Code on GitHub

Model Evaluation



Accuracy : 92,6%

AUC : 84,2%

Recall : 98,5%

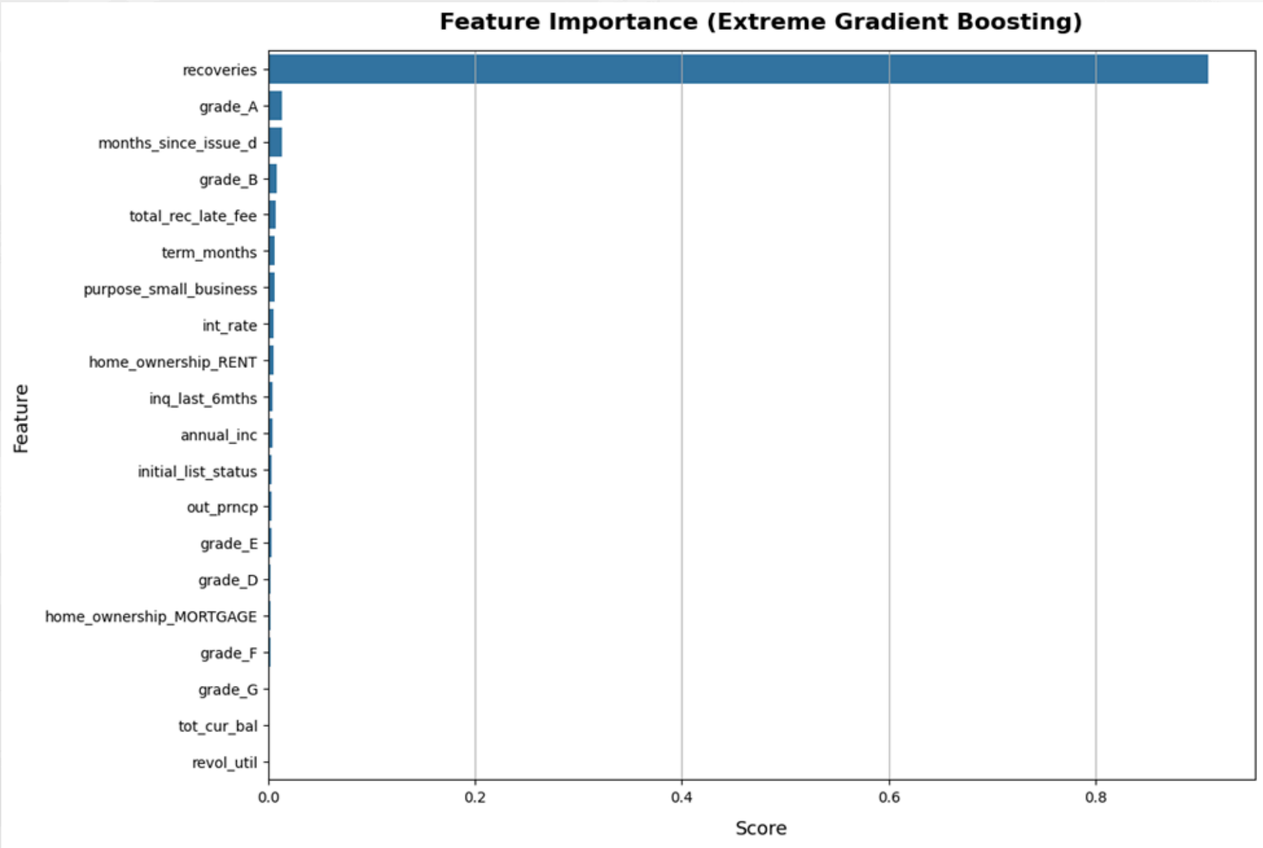
Precision : 93,4%

F1-Score : 95,9%

Link Full Code on GitHub

RECOMMENDATION

How Model Interpret the Data



Based on the **classification modeling results**, a number of **features** are obtained that have a **significant influence** on the **borrower's loan status**. From all available features, the **most influential features** were selected to support the **interpretation of results** and **business decision-making**. The **top five features** selected based on **their importance in predicting loan status are recoveries, grade, months_since_issue_d, total_rec_late_fee**, and **term**.

Link Full Code on GitHub

Business Recomendation

Recoveries

Recoveries provide an overview of the amount of funds recovered from non-performing loans. To **minimize potential losses due to bad debts**, companies need to build a **responsive and integrated collection system**. Therefore, it is recommended that companies **identify the characteristics of borrowers** who tend to be cooperative in the recovery process and **focus collection efforts more intensively on these groups**.

Grade

The grade assigned by the financial institution reflects the risk level of the borrower based on an internal system assessment. To strengthen risk control, **companies should adjust loan terms based on this grade**. Borrowers with low grades (e.g. A or B) can be given more competitive interest rates, while borrowers with high grades (e.g. F or G) should be charged higher interest rates or even rejected if they do not meet additional requirements such as collateral or guarantees. **This strategy will help balance the profit potential and default risk**.

Business Recomendation

months_since_issue_d

New loans generally do not show a consistent payment pattern, which **increases the level of uncertainty of default risk**. Therefore, companies need to implement a **strict monitoring system for loans less than six months old**. This enables **quick decisions** to be made if there are **indications of delays or violations**, such as through sending **early warnings** or providing **financial education sessions** to borrowers.

total_rec_late_fee

The higher the late fees paid by borrowers, the greater the risk they exhibit. This information can be utilized by companies to assess subsequent creditworthiness and **categorize borrowers into appropriate risk categories. For active loans, borrowers with a history of repeated delinquency should be given special attention**, such as through **sending automated alerts, providing financial counseling**, or **limiting future loan ceilings**.

Business Recomendation

term

Term, whether 36 months or 60 months, has a **direct impact on the level of credit risk**. Loans with **longer tenors tend to have a higher probability of default** due to the **increased uncertainty of economic conditions** over a longer period of time. Therefore, companies are advised to provide **incentives**, such as **lower interest rates**, to borrowers who choose **short tenors**. Meanwhile, **for loans with long tenors**, additional verification processes need to be carried out, such as assessment of **employment history**, **income stability**, and the borrower's **ability to meet long-term payment obligations**.

THANK YOU

Please reach me via:

qararfairuzzabadi@gmail.com

https://www.linkedin.com/in/m-razy-qarar-fairuzzabadi/

https://github.com/QararFairuzzabadi