

CORPORATE CREDIT RATING PREDICTION

Author: Nguyen Ngoc Tue Minh

2024

Table of Contents

Table of Figures	3
Introduction and project purposes.....	4
Dataset overview.....	4
Insights from Exploratory Data Analysis	5
Categorical variables	6
Continuous variables.....	7
Features' relationships.....	8
Model selections and performance	10
Evaluation principle	10
Model performance	11
Enhance model performance with Feature Selections and Cross Validation.....	13
Feature importance	14
Conclusion and recommendations	15

Table of Figures

Figure 1	5
Figure 2	6
Figure 3	6
Figure 4	7
Figure 5	7
Figure 6	8
Figure 7	8
Figure 8	9
Figure 9	9
Figure 10	9
Figure 11	10
Figure 12	11
Figure 13	12
Figure 14	12
Figure 15	13
Figure 16	14
Figure 17	15

Introduction and project purposes

Corporate credit ratings provide an assessment of a company's creditworthiness. They give investors a concrete idea of the risk associated with a company's credit investment returns. This notebook presents an analysis of corporate credit ratings based on a dataset obtained from Kaggle:

[link here](#)

The purpose of this project are:

- Explore features' relationship
- Determine the most influential indicators.
- Build classification machine learning models to predict companys' credit rating
- Make recommendations on how to improve credit ratings

Dataset overview

The dataset has 7805 rows which contain information on financial performance and credit rating of 678 companies from 2010 to 2016. Credit ratings here are based solely on financial data from the dataset. The target column for modelling will be Binary rating.

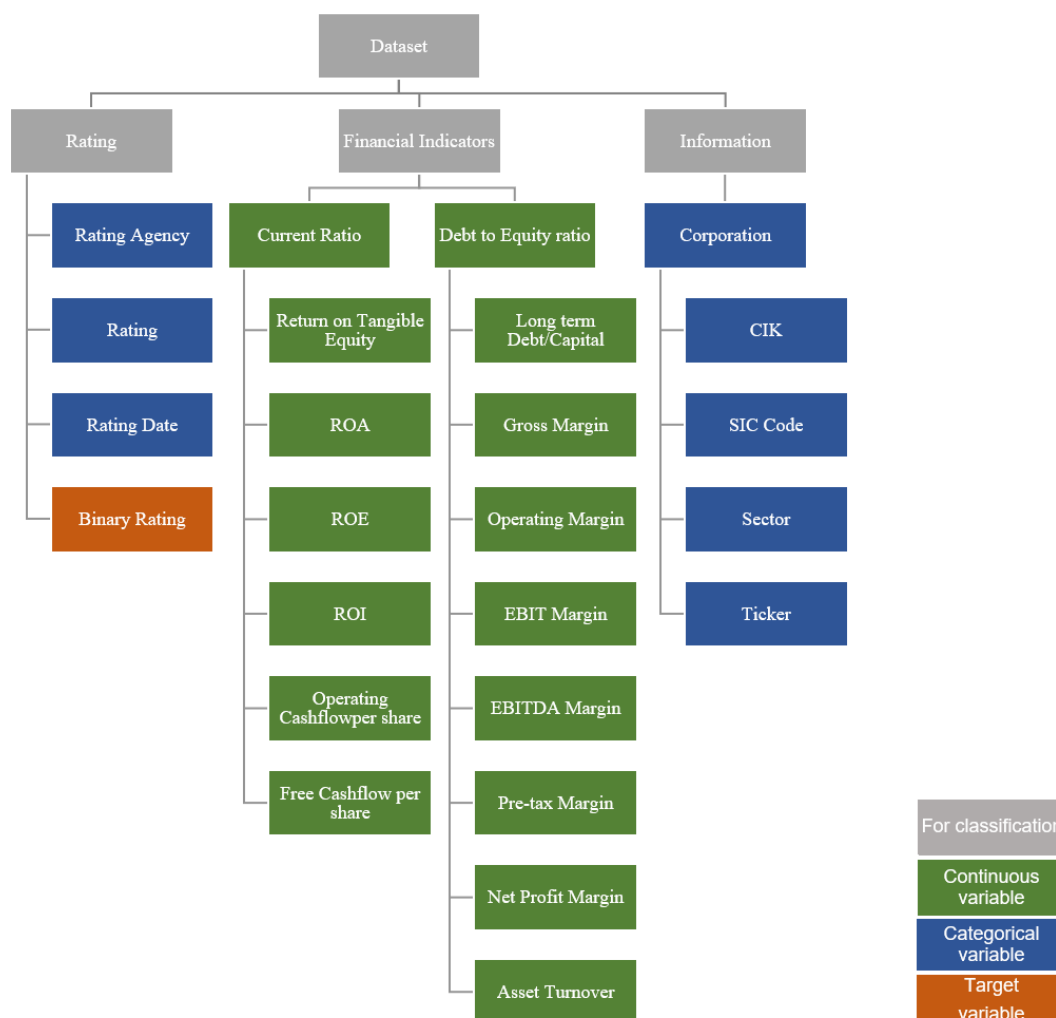


Figure 1

Insights from Exploratory Data Analysis

The dataset includes continuous and categorical variables like company ratings, sector, contracts, and financial ratios. Key insights from exploratory data analysis and visualization are outlined below.

Figure 3

Binary rating where '0' is for junk grade, '1' is for investment grade companies. Approximately 5200 are investment grade, under 2700 are junk. Over 1500 companies are in other sectors, the manufacturing sector is in the second position following by business and utility sectors.

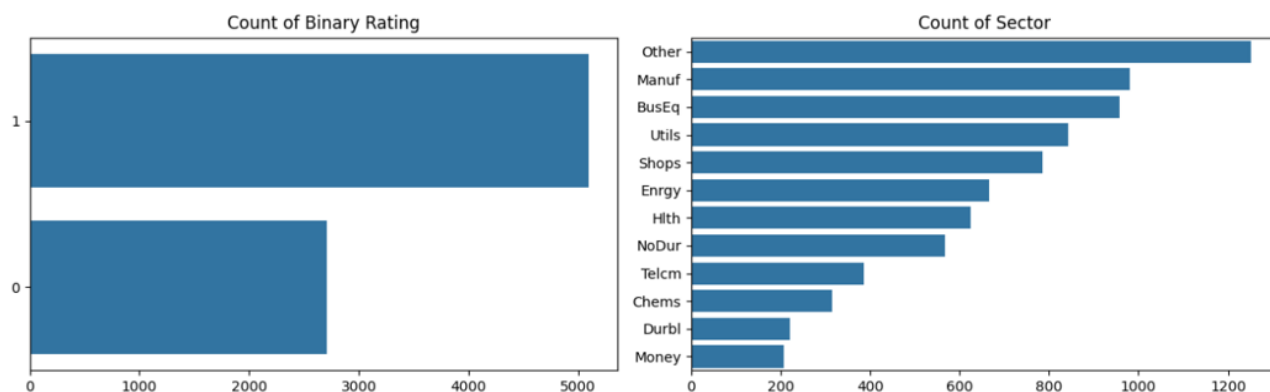


Figure 4

Continuous variables

50% of the companies have less than 40% gross profit margin, 75% of the companies have gross profit margin less than 60% and there are some exceptions where the gross profit margin is near to 100%.

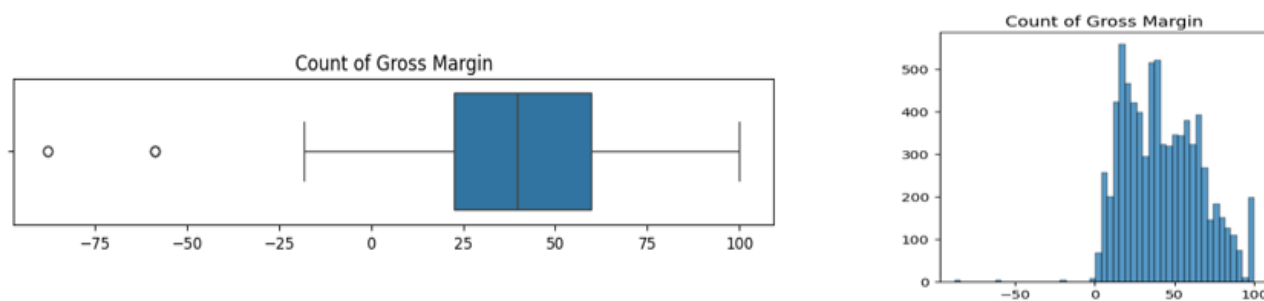


Figure 5

75% of the companies have less than 1 in their asset turnover ratio. There are many outliers which have asset turnover ratio above 2 and the maximum ratio is around 8.

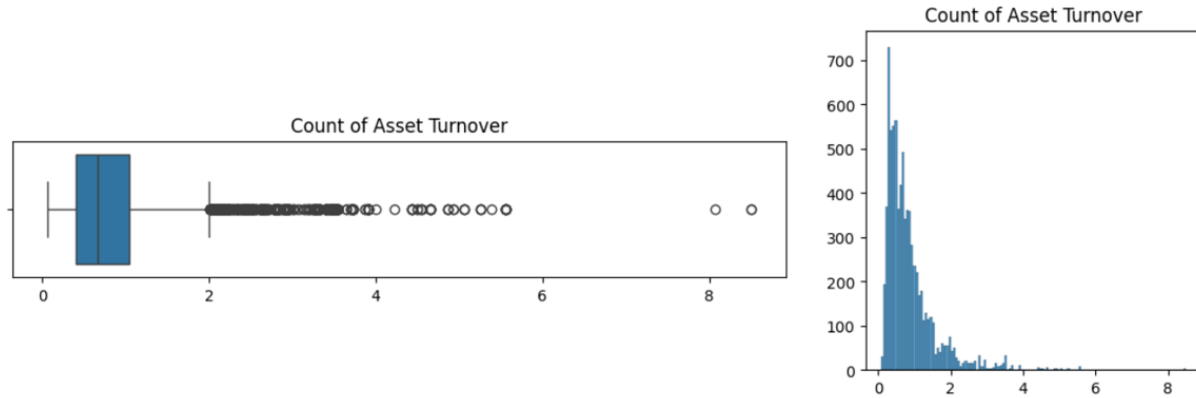


Figure 6

Features' relationships

The heatmap shows the strong relationship between income statement financial ratios such as Operating margin, EBITDA margin, pre-tax margin and net profit margin, the strongest relationship is between Operating profit margin and EBIT margin.

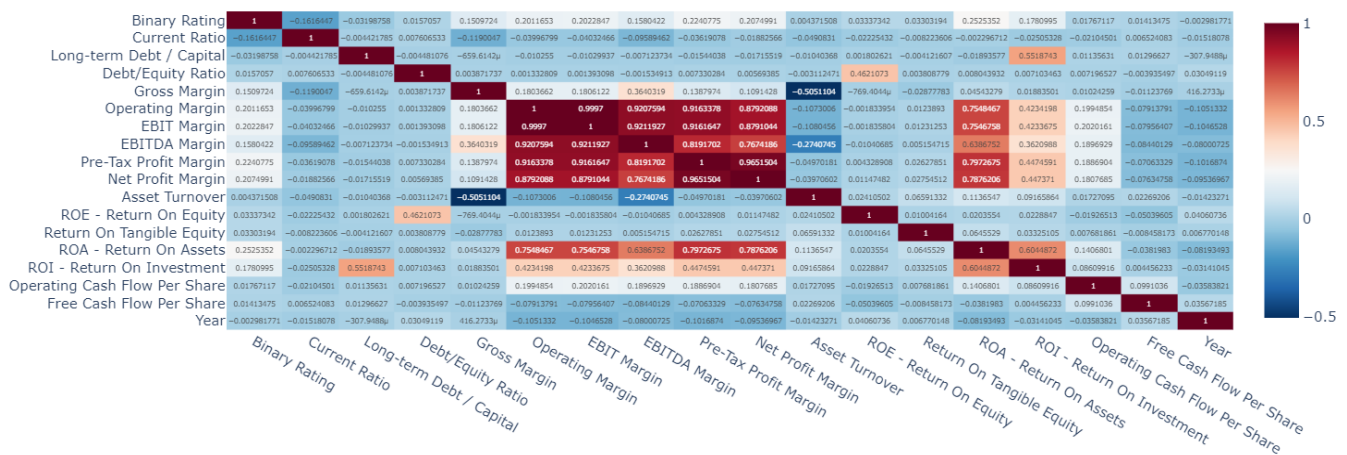


Figure 7

The more popular credit rating agencies for investment companies are Eager and Jones and Standard & Poor rating's agency. All A and BBB rated companies are investment companies.

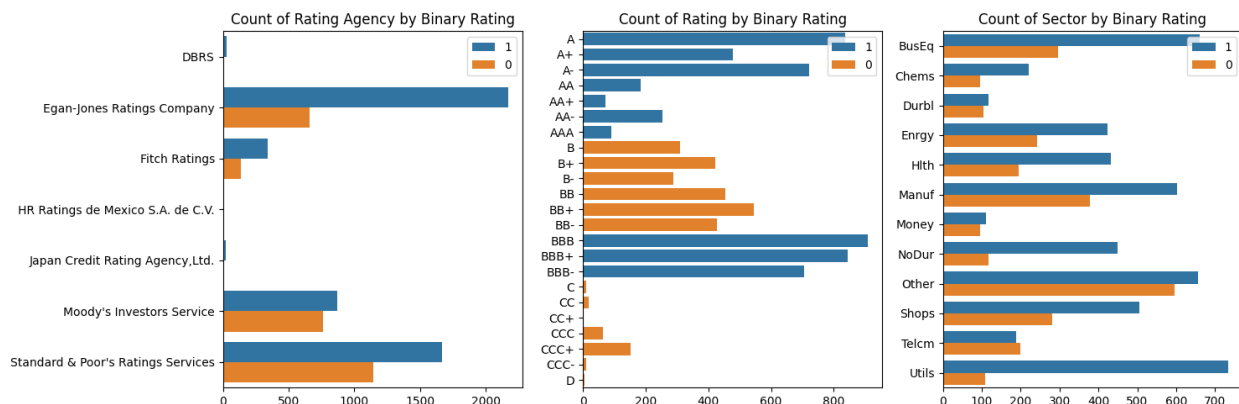


Figure 8

Half of the investment companies have gross margin of around 40%, 75% companies have ratio of 60%. Investment companies have higher gross margin than Junk companies.

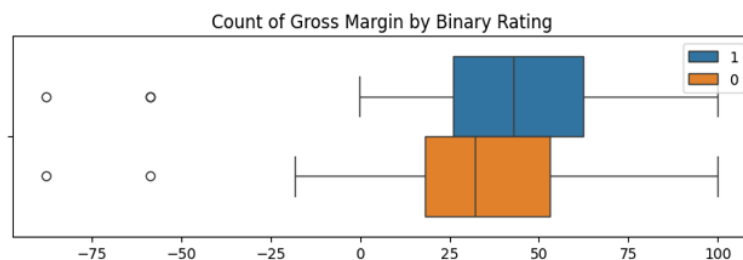


Figure 9

Most of the Junk companies has slightly higher asset turnover than Investment companies.

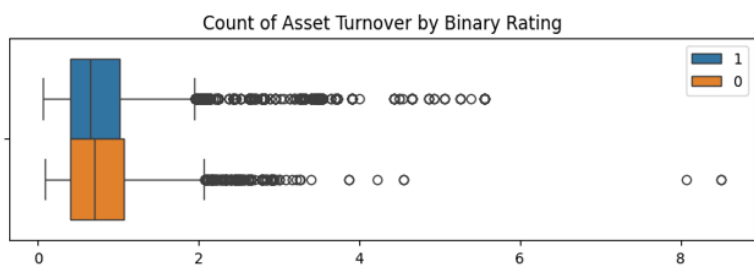


Figure 10

Energy sector has the highest gross margin is followed by health and money sectors.

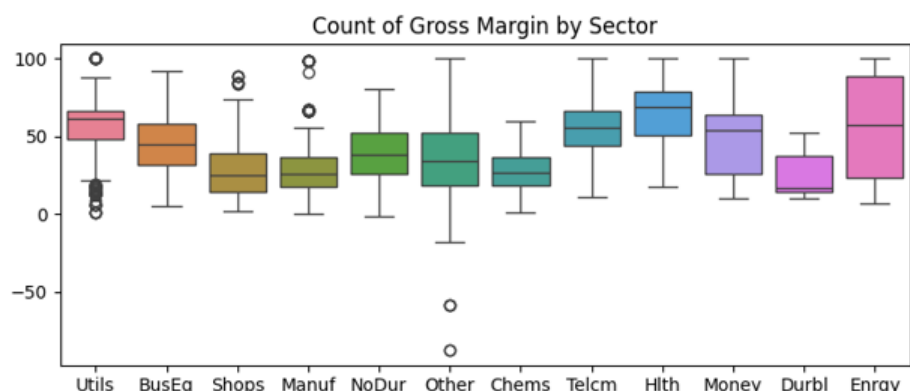


Figure 11

Model selections and performance

Evaluation principle

In credit ratings, minimizing risk is crucial for investors, emphasizing the need to maximize Recall for Class 0 (Junk Companies) to accurately identify them and prevent costly defaults.

While there's a trade-off between precision and recall, balancing both is key. Low recall for Junk raises default risks, and low precision may cause missed investment opportunities. However, the cost of a default loan typically outweighs missing a good investment.

Therefore, models favored for balanced risk management and identifying investment opportunities have **high Junk recall, acceptable Junk precision, and balanced accuracy**.

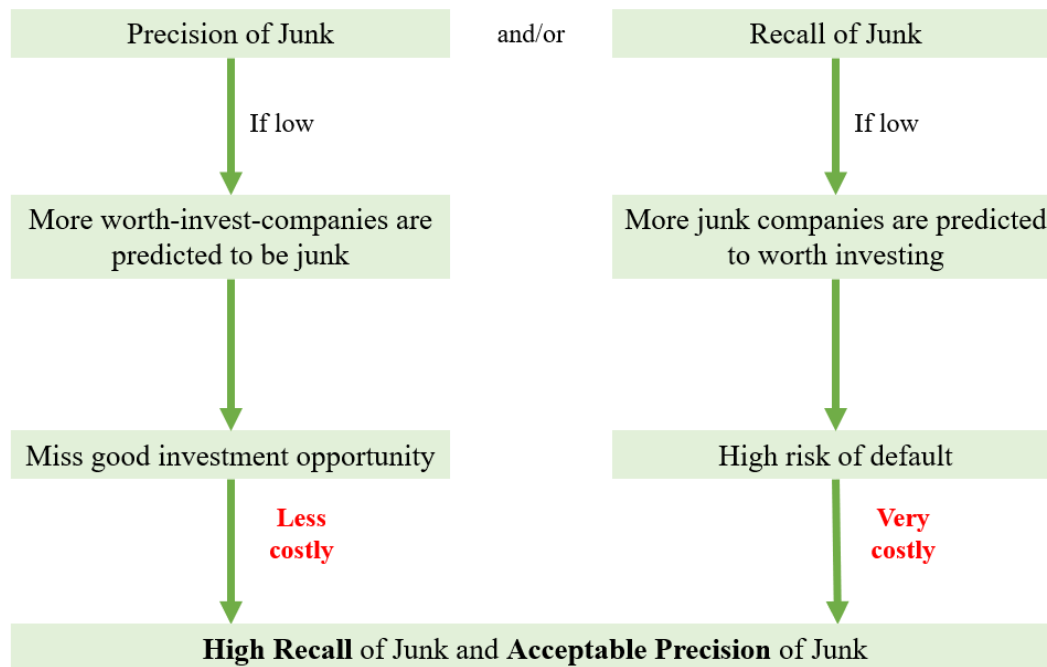


Figure 12

Model performance

Evaluating model performance:

- **KNN** excels with a 91% recall_of_Junk, 83% precision_of_Junk and 91% balanced_accuracy_score, indicating it effectively identifies most junk companies and accurately labels them.
- **Random Forest** shows strong performance with an 80% recall_of_Junk and 87% precision_of_Junk and 87% balanced_accuracy_score, suggesting it's slightly more precise but less sensitive than KNN.
- **Gradient Boosting** offers a balance with an 81% recall_of_Junk and 76% precision_of_Junk and 84% balanced_accuracy_score, being nearly as sensitive as Random Forest but less precise.

KNN seems to be the best model as it has the highest recall for junk and also a very high precision for junk. However, we also consider the Random Forest because **interpretability is also important** when we want to find what most important indicators are, Random Forest can provide insights into feature importance, while KNN is more of a black-box model.

	Model	accuracy_score	balance_accuracy_score	Precision of Junk	Recall of Junk	F1
0	LogisticRegression	0.684177	0.677100	0.530395	0.654784	0.586062
1	K Nearest Neighbours	0.907111	0.907793	0.833333	0.909944	0.869955
2	Decision Tree	0.666240	0.700973	0.507042	0.810507	0.623827
3	Gaussian NB	0.689942	0.563131	0.696000	0.163227	0.264438
4	Random Forest	0.891095	0.869886	0.868154	0.803002	0.834308
5	Gradient Boosting	0.845612	0.836256	0.757042	0.806754	0.781108
6	Support Vector Classifier	0.753363	0.744536	0.620130	0.716698	0.664926

Figure 13

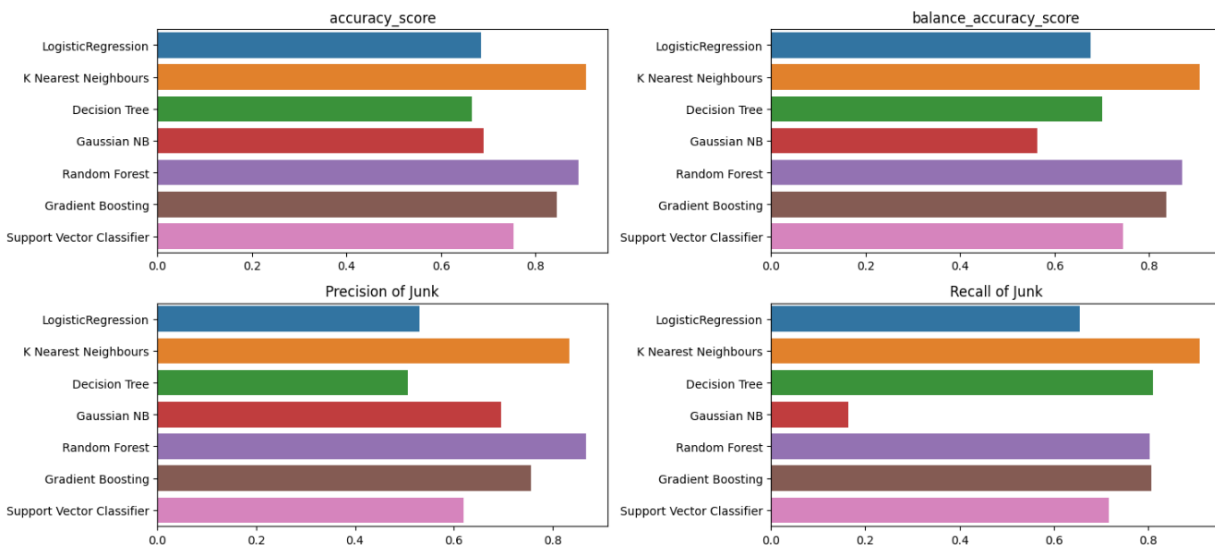


Figure 14

Enhance model performance with Feature Selections and Cross Validation

Below is the table of models' performance with Feature Selection and Cross Validation (FS&CV) and graphs show comparison of performance across models before and after FS&CV.

Overall, the application of FS&CV has led to improvements in the accuracy of some models, notably Logistic Regression and Decision Tree, while others have seen a slight reduction in accuracy, like KNN and Random Forest. The application of FS&CV also tends to increase the precision of the models, making them more reliable when they predict a company as junk. However, this often comes at the cost of recall, especially for Logistic Regression, which now misses many junk companies it used to identify.

	Model	accuracy_score_cv	balance_accuracy_score_cv	Precision of Junk_cv	Recall of Junk_cv	F1_cv
0	LogisticRegression	0.702498	0.586476	0.758760	0.208056	0.326566
1	K Nearest Neighbours	0.883536	0.885370	0.796829	0.891353	0.841444
2	Decision Tree	0.739654	0.700848	0.638455	0.574279	0.604669
3	Gaussian NB	0.692249	0.633002	0.573218	0.439763	0.497700
4	Random Forest	0.884561	0.850254	0.911912	0.738359	0.816010
5	Gradient Boosting	0.851634	0.815775	0.846464	0.698817	0.765587

Figure 15

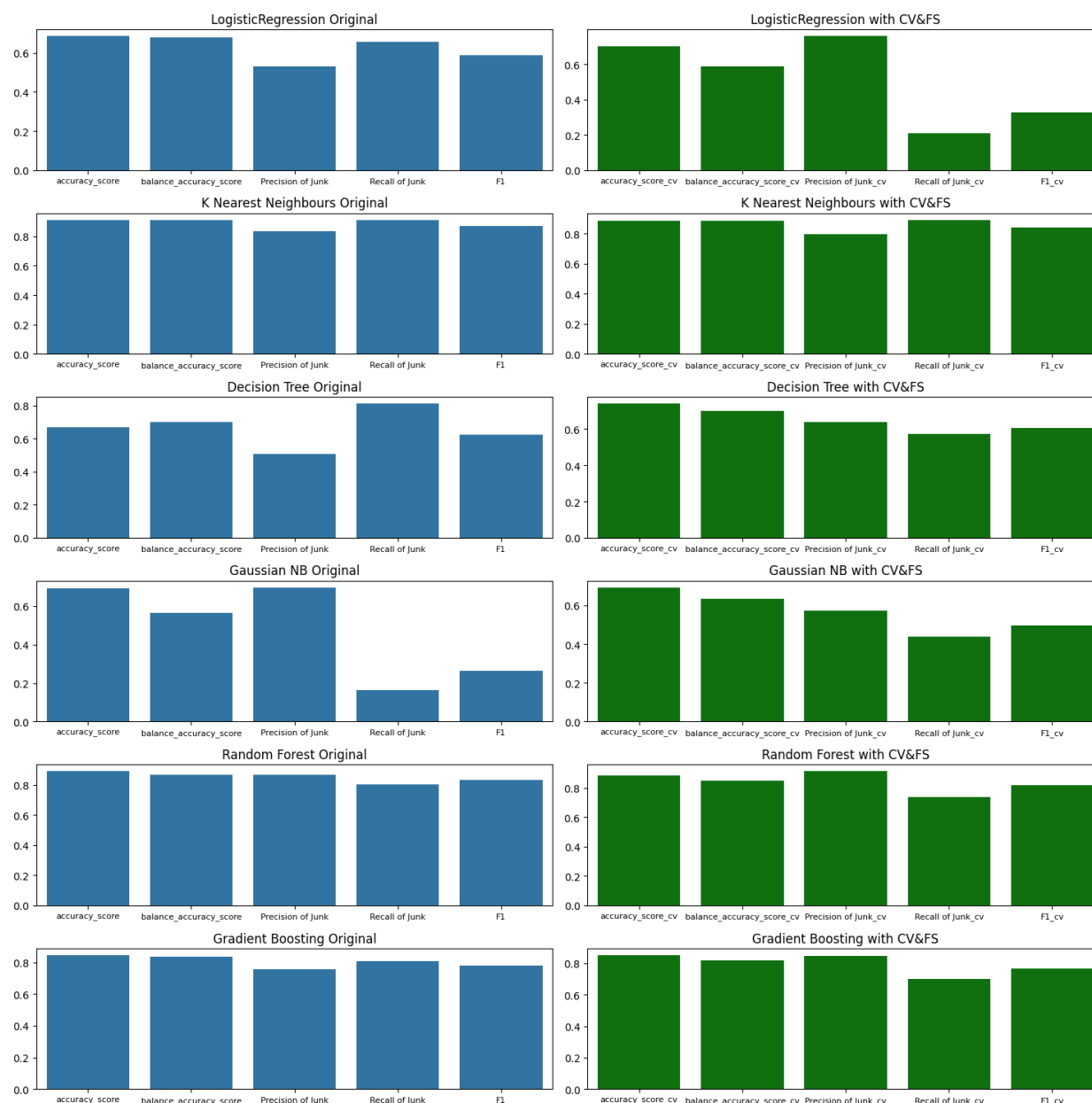


Figure 16

Feature importance

The chart shows feature importance generated from Random Forest model. The model places the importance on long-term financial stability and profitability indicators. This suggests that from a financial perspective, managing debt levels relative to capital and maintaining profitability are key to a good credit rating.

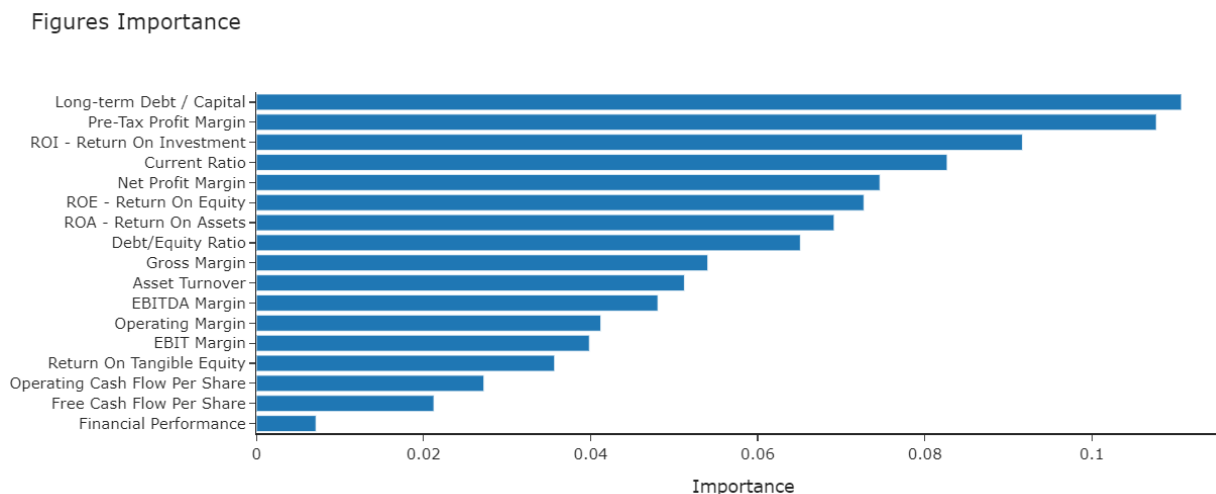


Figure 17

Conclusion and recommendations

KNN gives the best performance with a 91% recall and 83% precision for Junk, indicating high accuracy in identifying at-risk companies. Feature importance from the Random Forest model shows that long-term financial stability and profitability are crucial for improving credit ratings.

Our recommendation focuses on utilizing these models to enhance risk assessment and investment strategies, and improve financial ratios for better creditworthiness. Investment companies face risks from sensitive financial indicators, where minor negative changes can downgrade their ratings. Investors should use both financial and non-financial factors, including the company's risk tolerance, in their decisions.