Predicting Bankruptcy Risk in Public Companies

ISYE 7406 – Data Mining and Statistical Learning

Submitted by

Agrawal, Abhinav – 080– aagrawal372@gatech.edu

Aparicio, Valentina – 343 – vaparicio@gatech.edu

Hales, Morgan – 309 – mhales8@gatech.edu

Tuahivaatetonohiti, Guillaume – 933 – get3@gatech.edu

Singh, Pradhyumn – 476 – psingh344@gatech.edu

Date

04/25/2023

Abstract

In this project, we propose, compare, and evaluate various machine learning methods to detect bankruptcy risk in public companies. The key results show that non-linear methods can explain a significant portion of the cross-sectional variation in corporate defaults compared to benchmarking techniques such as logistic regression and the once-popular Altman Z score. Also, the empirical evidence shows that financial variables such as leverage, return on equity, etc., play a substantial role in default predictions.

Our results provide essential insights to various market participants, such as investment & commercial banks, rating agencies, and institutional investors.

1. Introduction

1.1. Problem Description and Motivation

The purpose of this project is to develop an accurate predictive model for identifying bankruptcy risk in public companies listed on the New York Stock Exchange and NASDAQ. Bankruptcy is defined as a company filing Chapter 11 or Chapter 7 of the Bankruptcy Code. Chapter 11 bankruptcy refers to "reorganization" bankruptcy, in which business may continue to operate, but a plan of reorganization is proposed and voted upon. Chapter 7 bankruptcy provides for liquidation in which creditors are compensated.

The motivation for this project stems from the insight it can provide investors, financial analysts and policy makers. The ability to understand the financial health of a company with ease allows investors to make informed decisions, with less risk overall.

1.2. Data Mining Challenges

The primary obstacle in this project is identifying the most effective machine learning model. To accomplish this, we employed seven distinct models, each of which is thoroughly outlined below.

However, before we could run the models, we had to cleanse the data. The dataset, which originated from three distinct sources, was intricate, containing imbalanced, negative data and numerous outliers, as well as the existence of multicollinearity.

Determining which machine learning model is the most efficient is the main challenge of this project. We ran our data through seven different models, which are explained in detail below. Before the models could be run, however, the data needed to be cleaned. The data set, coming from three different sources, was complex, with unbalanced, negative data that contained numerous outliers, multicollinearity was present as well.

1.3. Problem Solving Strategies

A number of methods were implemented on the data set to deal with these challenges. These include, but are not limited to, Synthetic Minority Oversampling Technique and Robust Sampling. These methods are described in detail in the following sections.

2. Data Sources & Preparation

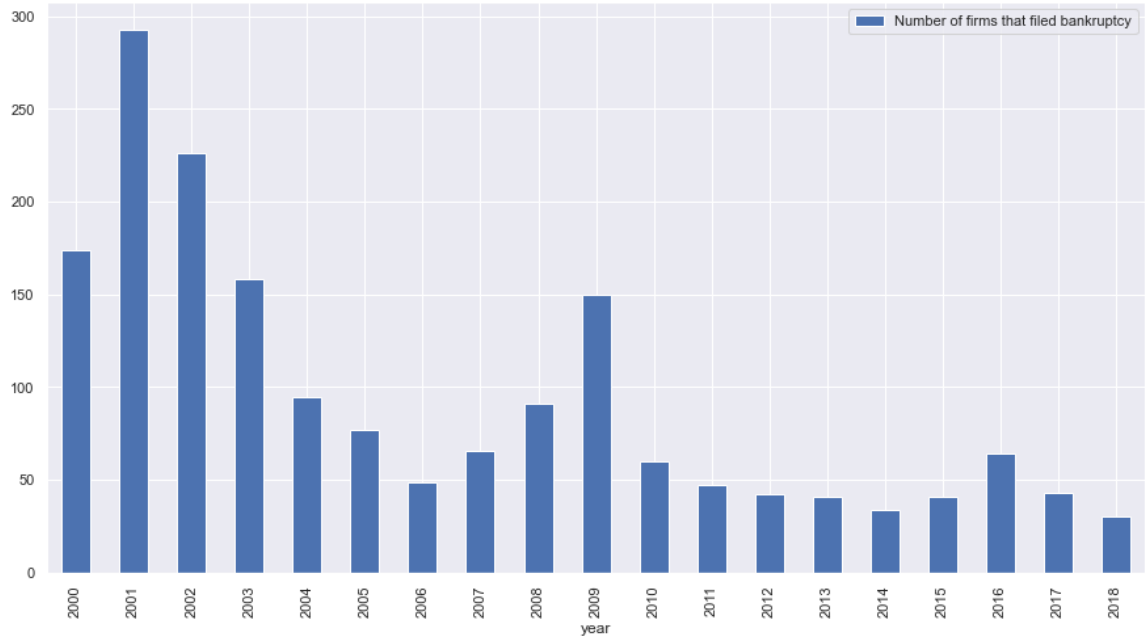
2.1. Data Extraction

The project was built on raw data extracted from the COMPUSTAT database (explanatory variables) and predicted binary variables created from Moody's Bankruptcy data. COMPUSTAT has financial information for an extensive range of active and inactive global companies on annual & quarterly frequency.

2.2. Data Cleaning & Feature Creation

The raw version of the extracted data had several issues, such as datatype mismatches, missing values, and erroneous negative entries for some of the critical variables. After correcting for these things, below specified changes were made to the extracted data

Categorical indicator for Bankruptcy: Moody's Bankruptcy data was used to find the companies that declared bankruptcy in a particular year. The explained variable (y) for our analysis was assigned a value of '0' for the year it declared bankruptcy and '1' for other years.



Feature Creation: Financial variables provide valuable information about a company's financial health and its ability to meet its debt obligations, which can be used to predict the likelihood of default. However, when used in absolute terms, financial information such as Revenue, Total Debt merely have a story to tell in cross-sectional analysis.

Hence, the predicting features/variables were created using comparable ratios for Profitability (such as Gross Margin, EBITDA Margin), Liquidity (such as Days of Sales Outstanding, Current Ratio), Leverage, etc. Moreover, rather than looking at these ratios in isolation for a particular year, we used rolling averages which are a stronger measure of the deteriorating health of any company.

2.3. Exploratory Data Analysis & Modifications

Outlier Detection

The descriptive stats of the data indicate that our explanatory factors, such as Gross Margin, EBITDA Margin, Return on Equity, etc., had values that lay several standard deviations away from the mean. This depicted the presence of outliers in the dataset. The presence of outliers was further validated by investigating the Box Plots (figure xxx) of each independent variable.

Since working with outliers can induce bias in the model, and in our case study, the outliers may carry some valuable insights as well (variables tend to have extreme values during the recessionary phase, etc.), appropriate scaling measures were taken. As the factors are not normally distributed, we used scaling under the interquartile range, which is a more robust measure of spread.

Multicollinearity Analysis

Multicollinearity refers to the occurrence of a high correlation between two or more independent variables (or a linear combination of those) in a regression model. It can pose a significant problem as it undermines the statistical significance of individual variables and weakens the model's reliability.

To investigate the presence of multicollinearity in our dataset, we used correlation scores (figure xxx) backed up by Variance Inflation Factor (VIF). As VIF measures the degree of multicollinearity present in a regression model, the threshold tested for it was below 10 for all the independent variables. As we observed (figure xxx) for our dataset, features such as Gross Margin and EBITDA Margin have a high VIF score. This fundamentally makes sense as well because EBIDTA flows from gross profit.

As multicollinearity violates the assumptions of linear models, we used elastic net in our logistic regression model to introduce penalty terms.

3. Methodology

3.1. Bankruptcy Prediction Techniques

In the literature there are two main approaches to measuring bankruptcy risk which revolve around structural or/and statistical models. While each methodology has expanded over time, they continue to be actively used today.

Structural models

Structural models for default prediction are a class of financial models that attempt to predict the probability of default of a company based on its underlying financial structure. One common type of structural model for default prediction is the Merton model.

Reduced form models

From a statistical perspective, predicting corporate bankruptcy is a common classification problem where a company is categorized as either non-bankrupt or bankrupt based on its features. There are three primary methodologies for forecasting corporate default using statistical models, which can be classified into three generations; representative studies for each generation are provided in the below table

Statistical Approaches	References
Discriminant Analyses	Beaver (1966); Altman (1968); Mare et al. (2017)
Binary Response Models	Ohlson (1980); Zmijewski (1984); Foreman (2003); Campbell et al. (2008); Kukuk and Rönnberg (2013); Aretz et al. (2018)
Hazard Models	Shumway (2001); Chava and Jarrow (2004); Nam et al. (2008); Bonfim (2009); Dakovic et al. (2010); Duan et al. (2012); Figlewski et al. (2012); Tian et al. (2015); Traczynski (2017)

For the forthcoming analysis, we focus on Binary Response Models that bring into play several Machine Learning techniques to identify the relationship between company's financial information from the most recent fiscal year (or rolling period) before its bankruptcy and its default in the subsequent year.

3.2. Model Implementation

Logistic Regression

Logistic regression is a statistical technique utilized for binary classification, which aims to estimate the likelihood of an event occurring based on specific input characteristics. For instance,

in predicting bankruptcy, the objective is to determine whether a company will fail or not based on the company fundamentals. To accomplish this, the logistic regression model computes the linear combination of input characteristics and a group of coefficients acquired through training. Subsequently, this linear combination is fed into a sigmoid function, which translates the result into a value between 0 and 1, signifying the event's probability. The logistic regression method is a prevalent alternative for binary classification tasks because of its simplicity, explainability, and efficiency. It performs notably well when the decision boundary between the two classes is linear or can be approximated by a linear function.

In the context of bankruptcy prediction, logistic regression can be a suitable option since the dataset contains a binary target variable and some features that may have a linear correlation with the target variable. Additionally, logistic regression produces understandable coefficients, which can assist in identifying the most important features in predicting default.

In our model implementation we have used the penalty term to be elastic net which solves the problem of overfitting. In addition to this, we have used the solver to be “saga” and the λ in the range of {0.05, 0.02, 0.2, 0.5, 1}. These hyperparameters are tuned on the basis of 3-fold cross validation and the best performance by the cross-validation suggests the use of $\lambda = 0.1$.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a non-parametric supervised learning model which serves both classification and regression purposes. The algorithm works by finding the K closest training datapoints to the new data point being predicted and then taking a weighted average of their labels to predict the label of the new data point.

In context to our dataset, we will use it to classify whether a company is considered bankrupt or not based on the fundamental features. We have used 3-fold cross validation along with the hyperparameter tuning of the number of neighbors.

Random Forest

Random Forest is an ensemble learning algorithm used for both classification and regression tasks. The algorithm creates a set of decision trees on random subsets of the input data and combines them to make a final prediction. The idea behind the Random Forest algorithm is that a group of weak learners (the individual decision trees) can come together to form a strong learner that can make more accurate predictions.

Since random forest is a very good algorithm when we deal with large and complex non-linear data and performs very well when we have both numeric and categorical data, we use it for our dataset to predict bankruptcy. We use 3-fold cross validation to tune hyperparameters such as number of trees and maximum tree depth.

Gradient Boosting Machine

Gradient Boosting Machine (GBM) is a popular machine learning technique that uses an ensemble of weak prediction models, such as decision trees, to make accurate predictions on complex data sets. The main idea behind using GBM is to reduce the bias error along with hyperparameter tuning and also minimize the error in each iteration of variable selection.

Since here we are working on a classification problem and as seen in SVM that we have non-linear GBM along with log-likelihood as our loss function.

We perform 5-fold cross validation to tune the hyperparameters such as number of trees, learning rate and the regularization parameter.

Extreme Gradient Boosting

XGBoost is an advanced machine learning algorithm that is used for both regression and classification problems. It is an optimized version of the gradient boosting algorithm which works by iteratively adding decision trees to a model to minimize the loss function. It uses various hyperparameters like regularization, tree pruning and parallel processing.

For our dataset, we have used XGBoost to forecast the possibility of bankruptcy of a company. It is used particularly because of its characteristic of handling high dimensional and sparse data. In addition to this, it can very well handle the non-linear relationship amongst the features and the outcome. As we saw in the Logistic regression that the results are not very good particularly because of the non-linearity in the dataset, XGBoost performs very well.

To tune the hyperparameters, we used 3-fold cross validations and tuned the number of trees, learning rate (shrinkage) and the regularization parameter. Our best hyperparameters include depth of 10, learning rate of 0.01 and number of features to be 55%. Once we have the hyperparameters, we fit the model using these parameters.

Support Vector Machine (SVM)

Support vector machines (SVMs) are a popular class of supervised learning algorithms used for classification, regression, and outlier detection. The fundamental idea behind SVMs is to find the optimal hyperplane that maximizes the margin between the different classes in the data. It can handle both linear and non-linear data by transforming the data into a higher-dimensional space where the classes can be separated by a hyperplane.

For our dataset, we used SVMs to classify the binary variable i.e., bankruptcy using the independent features. Since the model performs good with both linear and non-linear data, we start with a single dimensional feature set and subsequently increase it to get the best decision boundary.

To tune the hyperparameter, we use 3-fold cross validation and found the best degree of the kernel function that would increase our accuracy and ROC.

Neural Network

Neural Networks (NN) are a type of machine learning algorithm that uses interconnected nodes to process and transform data. By adjusting weights and biases through feed forward or backpropagation, the algorithm can learn from the data.

Neural networks are particularly effective in predicting non-linear and complex time-series data. Considering the potential non-linear relationships and time-series dependencies of our features, a neural network is a suitable model to consider. To avoid overfitting, we can adjust hyperparameters such as the number of hidden layers, nodes, and network density during tuning.

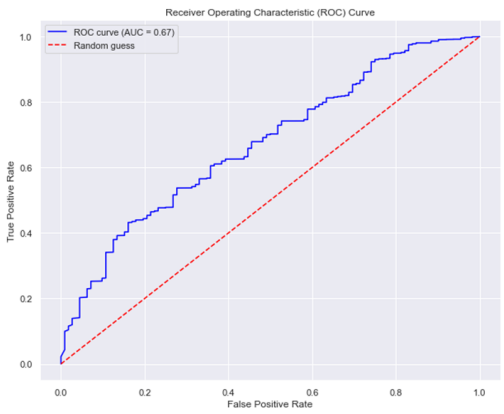
4. Analysis & Results

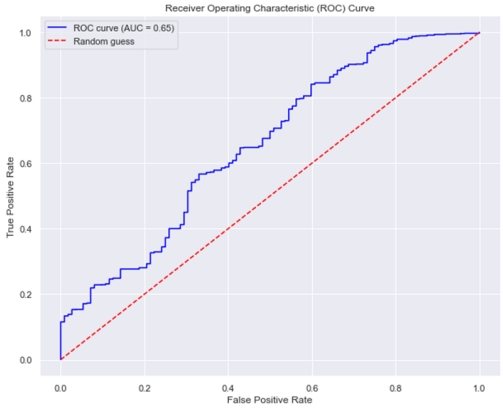
4.1. Model Results Summary

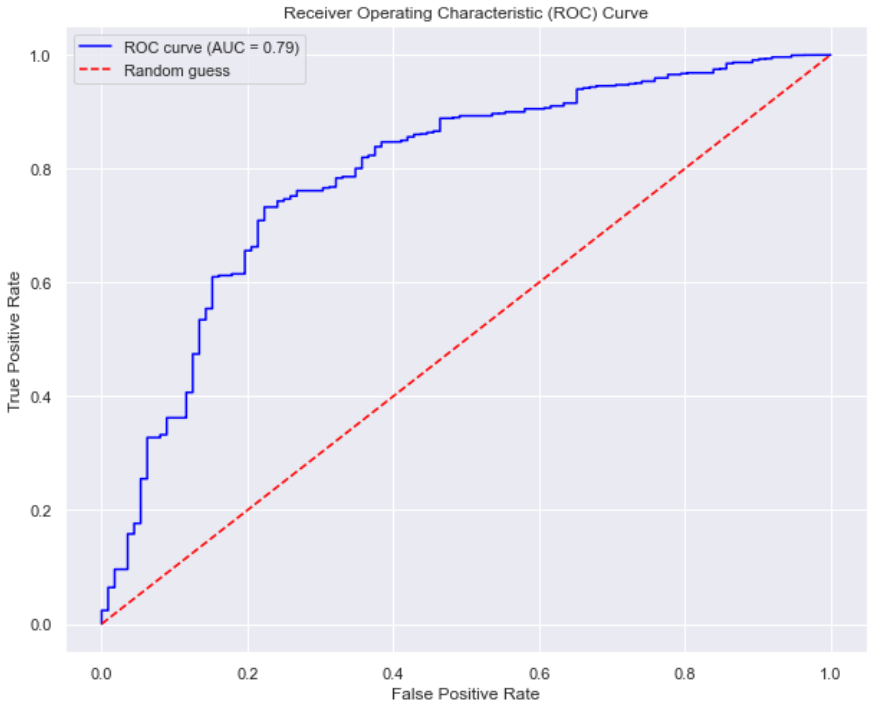
The table below represents the evaluation metrics (Table xxx) of above-described machine learning models trained on our dataset.

	Accuracy	Precision	Recall	F1 Score	AUC Score
Neural Network	0.796	0.524	0.608	0.562	0.794
Gradient Boosting	0.879	0.511	0.708	0.492	0.790
Random Forest	0.990	0.511	0.507	0.508	0.715
XG Boosting	0.981	0.521	0.551	0.529	0.672
Logistic Regression	0.519	0.502	0.603	0.349	0.654
SVM	0.962	0.501	0.587	0.540	0.589
KNN	0.956	0.505	0.534	0.503	0.583

4.2. Comparative Analysis with benchmarking techniques/baseline models

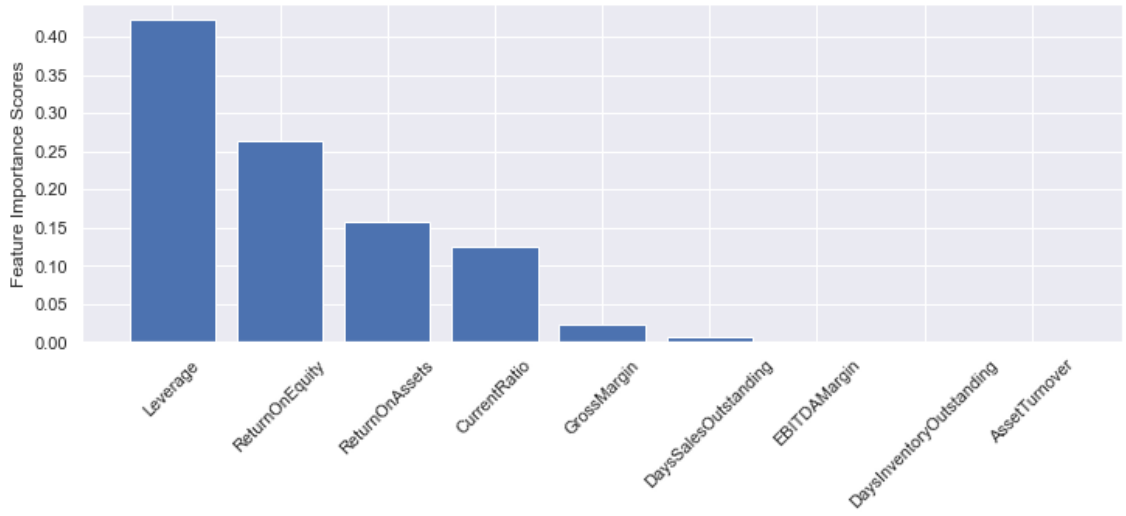






4.3. Feature Importance

The feature importance score in GBM which is typically based on the number of times a feature is selected for splitting, weighted by the improvement to the model's performance resulting from that split.



4.4. Key Findings

- We can see from the results that Neural Network performs the best in terms of the AUC Score which is probably the best evaluation parameter for an imbalanced binary class classification problem. Since the data that worked on had a non-linear relationship with some time dependency, it makes sense for neural networks to perform best because of its ability to find patterns in complex and high dimensional data and the hyperparameter tuning of the density of the hidden layers.
- As observed from the ROC curves we can infer that though our best performing interpretable model GBM is able to add value to our analysis, Logistic regression performs more or less in line with the Altman Z score predictions.
- In general, a feature with a higher importance score is considered more important in predicting the target variable.

5. Conclusions

In our study on predicting the risk of bankruptcy in public companies, we found that Neural Network was the most effective algorithm. However, there are several limitations to our study that should be considered when interpreting our results.

One major limitation is that we did not take into account the cyclic nature of the market across the 11 existing sectors. Some sectors may have more chances of receiving financial aid or bailouts than others, depending on the economic climate. For example, during a recession, companies in the finance and real estate sectors may be more likely to receive assistance than companies in the retail or hospitality sectors. This can affect the overall risk of bankruptcy for companies in different sectors and may impact the accuracy of our predictions.

Another limitation is that we did not consider the possibility of companies on the verge of bankruptcy being bought up or absorbed by larger competitors. This can prevent a company from actually going bankrupt and can make our predictions less accurate. Additionally, political agendas can play a significant role in the survival of public companies, with larger companies being less likely to be allowed to fail by governments due to their economic impact. This can also impact the overall risk of bankruptcy for public companies.

While our study found neural network to be effective in predicting the risk of bankruptcy in public companies, these limitations suggest that further research is needed to fully understand the factors that contribute to the financial health and bankruptcy risk of public companies. Future studies should take into account the cyclic nature of the market, the possibility of buyouts or absorption, and the role of political agendas in determining the survival of public companies.

5.1. Lessons Learned

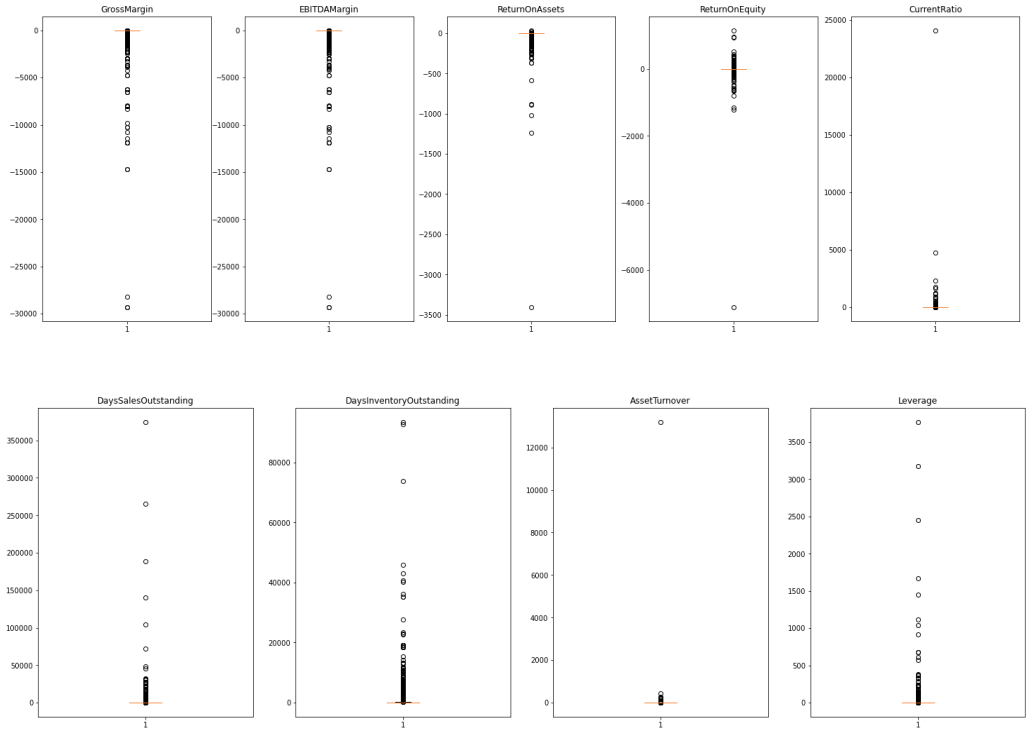
This project, as well as the overall course, has provided a valuable opportunity to enhance our data management and model development skills, while also refining our proficiency in R and Python. The problem we tackled in this project is highly relevant in current times, enabling us to compare our findings with real-world scenarios. Throughout the project, we recognized the significance of Exploratory Data Analysis and feature engineering, which are crucial elements of any successful machine learning project.

Furthermore, the project allowed us to explore the importance of hyperparameter tuning and cross-validation for achieving stable results.

Not only did we advance our technical capabilities, but the project presentation also helped us develop our public speaking skills and foster a sense of teamwork. Engaging in healthy discussions within our group and receiving constructive feedback from our peers and mentors helped us identify our weaknesses and work on them. Overall, we found this project to be an enriching learning experience.

6. Appendix

To account for imbalanced dataset, we have used “macro” F1 score to have a better understanding of the model in terms of precision and recall. This result makes sense as we can see in case of Random Forest and XGBoost where we have an inline precision and recall with given level of AUC Score despite very high accuracy


GrossMargin
EBITDAMargin
ReturnOnAssets
ReturnOnEquity
CurrentRatio
DaysSalesOutstanding
DaysInventoryOutstanding
AssetTurnover
Leverage

Evaluation Metric

Accuracy	Accuracy is a measure of how well a classification model correctly predicts the classes of a dataset. It is the ratio of the total number of correct predictions to the total number of predictions made.
Precision	Precision is a measure of the model's ability to correctly predict the positive class. **It is the ratio of the true positive predictions to the sum of true positive and false positive predictions**. In other words, precision is the proportion of predicted positive cases that are actually positive.
Recall	Recall is a measure of the model's ability to correctly identify the positive class. **It is the ratio of the true positive predictions to the sum of true positive and false negative predictions**. In other words, recall is the proportion of actual positive cases that are correctly identified by the model.

F1 Score	F1 score is a harmonic mean of precision and recall, and it provides a single score that balances the tradeoff between precision and recall. It is calculated as **2 * (precision * recall) / (precision + recall)**.
AUC Score	AUC score (Area Under the Curve) is a performance metric that measures the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1 - specificity) for different threshold values. AUC score measures the overall performance of the model across all possible classification thresholds. A higher AUC score indicates a better-performing model

7. Bibliography & Credits