Predicting Probability of Default for Banca Massiccia

TEAM TURQUOISE:

- Inderpreet Singh Walia
- Pradhyumn Bhale
- Shrey Jasuja

Loading..

Overview

Items	Start Time	Duration
Introduction	0:00	1:30
Definining Target Variable	1:30	1:00
Data Cleaning and Preparation	2:30	1:30
Financial Intuition & Ratios Considered	4:00	1:00
Modeling Objectives	5:00	1:00
Model Design & Interpretability	6:00	2:00
Walk Forward & Benchmarking	8:00	2:00
Results and Calibration	10:00	1:00
Conclusion	11:00	1:00

Introduction

Banca Massicia wants to understand how to reduce its portfolio default rate on individual loans.

Motivation is to calculate the probability of default for clients within a window of 12 months.

This is both a finance issue and a machine learning challenge.

Explainability of the model is of paramount importance because of the high cost associated with making bad decisions.

*Image generated by DALL-E. Notice the typos in the image. Even the most advanced ML algorithms aren't full proof yet!


DWIGHT K. SCHRUTE
BOISE
BOISE
BOISE

Firm Year

A firm year (sometimes referred to as the fiscal year) is the period that a company uses for accounting purposes and preparing financial statements.

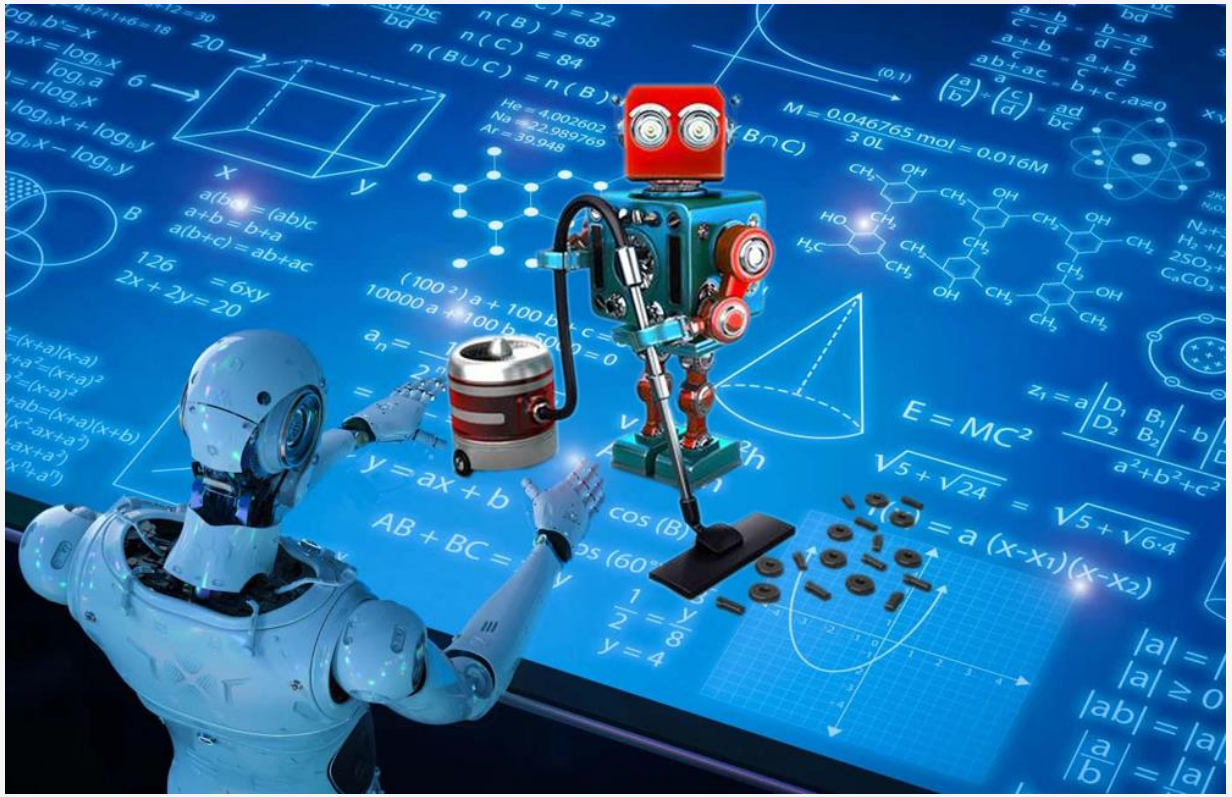
Why is a firm year important?

1. In Italy, banks release their financial statements around July (07)*.

2. We set up the problem by making the dependent variable (default flag = 1) if default date is within 7-18 months from the financial statements date.

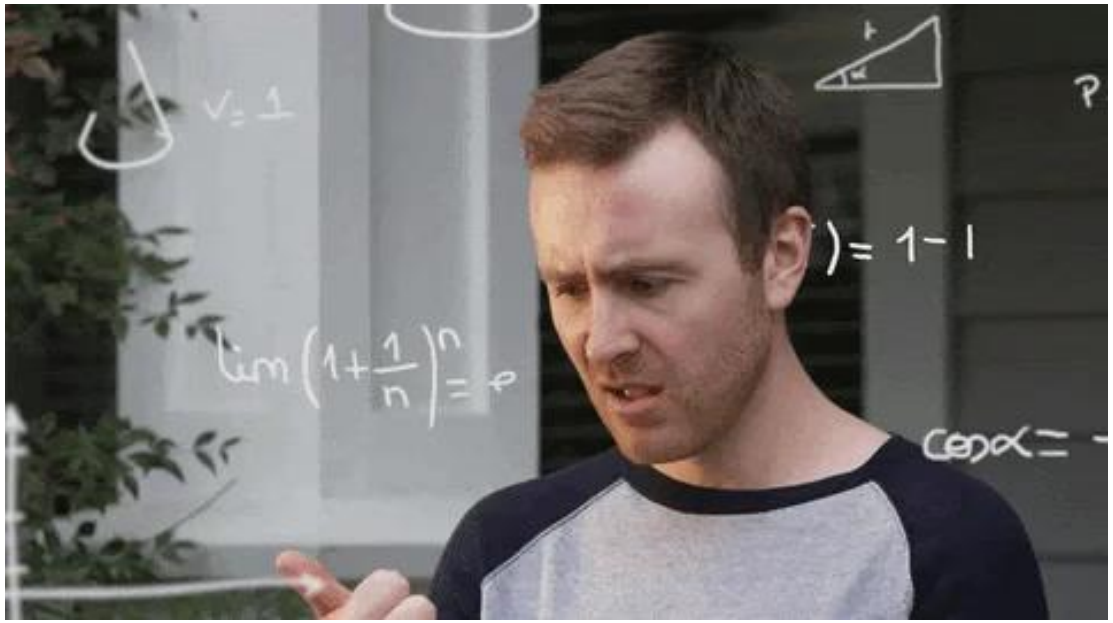
3. Model avoids future-peeking issue.

*https://italianbusinessregister.it/en/annual-accounts



Data Cleaning

- Calculated *net income* by multiplying *roa* with *total assets*.
- Filled in the missing *roe* values using the calculated *net income* & divided it by *total equity*.
- Calculated *non-current assets* by subtracting *total assets* with *current assets*.
- Calculated *operating revenue* by adding *COGS* and *operating profits*.
- Calculated *total liabilities* by subtracting *total equities* from *total assets*.
- Calculated *fixed assets* by adding *asset intangible fixed*, *asset tangible fixed* and *asset fixed fin*.
- Calculated *current liabilities* by subtracting *net working capital* from *current assets*.



Sanity Checks

- Current liabilities was greater than total liabilities (<0.01% times).
- Made the simplifying assumption of non-current liabilities to be long term debt.
- Current liabilities = short term debt with 99.9% match.
- Skewness and kurtosis of data is decreasing on cleaning and performing relevant transformations.
- Net income positive with negative taxes were removed as these were observed as erroneous data.

Data Preparation and Feature Selection for Default Risk Modeling

1. Our team has addressed the gaps in our dataset by utilizing the available financial ratios to estimate missing values.

2. To handle skewed data within our dataset, we've implemented a sophisticated transcendental transformation technique on the affected features.

3. The selection of model features was strategically conducted based on the Variance Inflation Factor (VIF) of each feature, ensuring optimal relevance and accuracy.

Incorporating Financial Intuition

We start by analyzing some key financial ratios and dividing them into "buckets".

These financial ratios were arrived at by studying their use in past literature as well as referring to additional study materials provided for the class and the summary sheets that are posted.

Liquidity

- Cash Ratio
- Current Ratio
- CFO Ratio

Profitability

- ROA
- ROE
- Gross Profit Margin

Debt Coverage

- Leverage
- CFO to Total Liabilities
- Interest Coverage Ratio
- Debt-Service Coverage Ratio

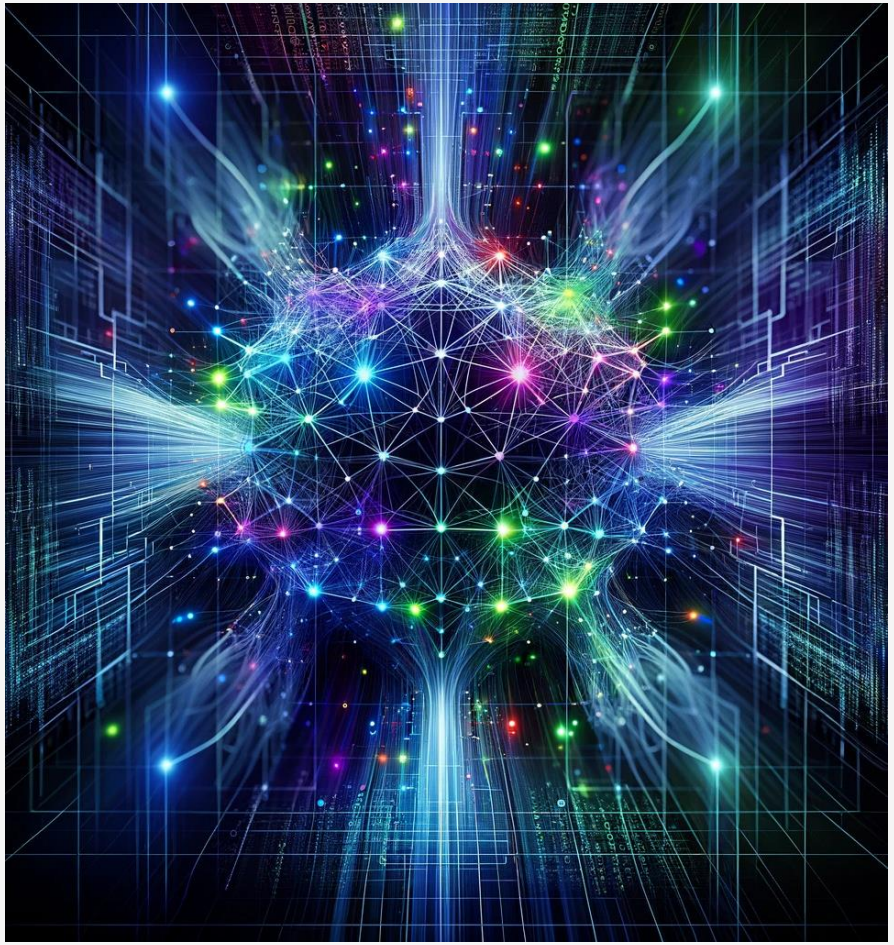
Asset Management

- Asset Coverage Ratio
- ROA
- Fixed Assets

Modeling Objectives

We developed 2 models - logit and XGBoost to predict PD. We also conducted a thorough ViF and Correlation analysis to finalize the variables we want to use.

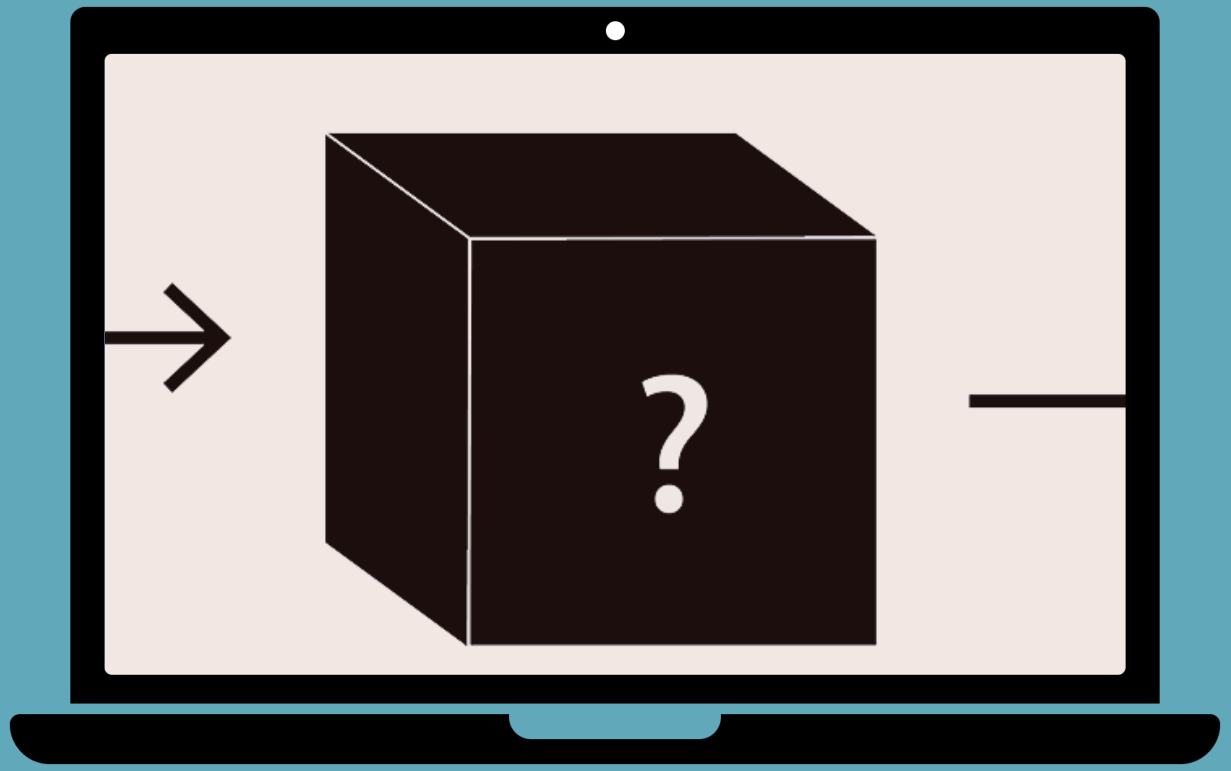
1. Univariate Analysis for each variable.
2. ViF and Correlation Analysis.
3. Walk Forward Analysis (training).



Model Input/Output:

Inputs:

- Legal Structure
- Current Ratio
- CFO Ratio
- Leverage
- ROA
- Fixed Assets



Outputs:

Probability of Default (PD)

Interpretability of Logit Model

Results from Logit Model

- Are the cofficients relevant?
- What about p-value?

Looks Great!

But , What about the last two columns?

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-10.0152	0.174	-57.470	0.000	-10.357	-9.674
legal_struct	0.0680	0.013	5.403	0.000	0.043	0.093
CFO_ratio	-0.9950	0.036	-27.886	0.000	-1.065	-0.925
current_ratio	-0.4096	0.032	-12.867	0.000	-0.472	-0.347
roa	-0.2941	0.009	-30.975	0.000	-0.313	-0.275
leverage	6.4752	0.162	39.990	0.000	6.158	6.793
fixed_assets	-7.339e-08	5.88e-09	-12.475	0.000	-8.49e-08	-6.19e-08

Feature Importance

We have calculated the feature importance by weight, gain and "shap" values for each input feature to the XGB model.

- ## Weight (frequency) Importance

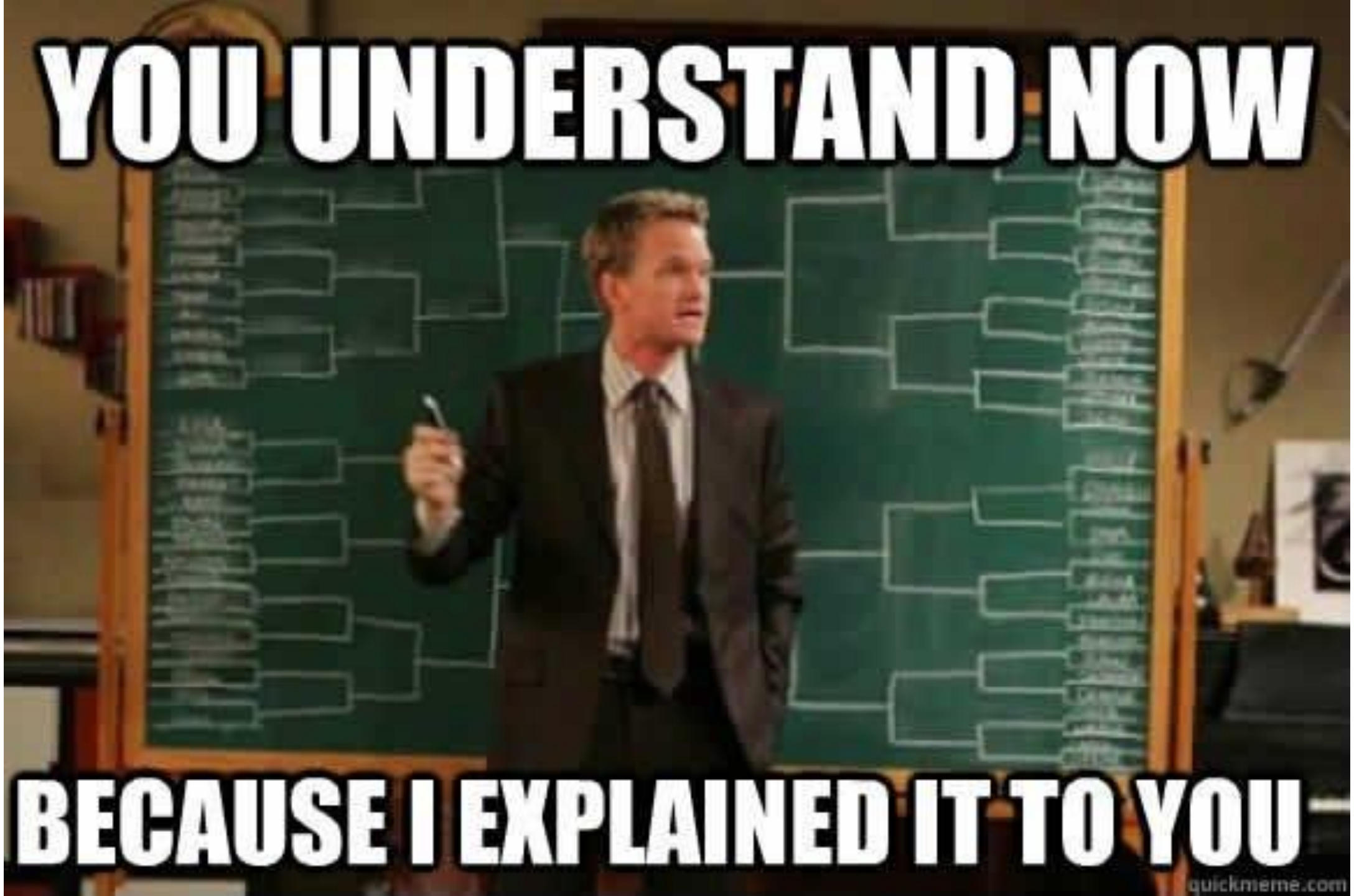
  This represents the number of times a feature is used to split the data across all trees. Legal Structure is least important (77 times) and CFO ratio and leverage are most important (272 times).

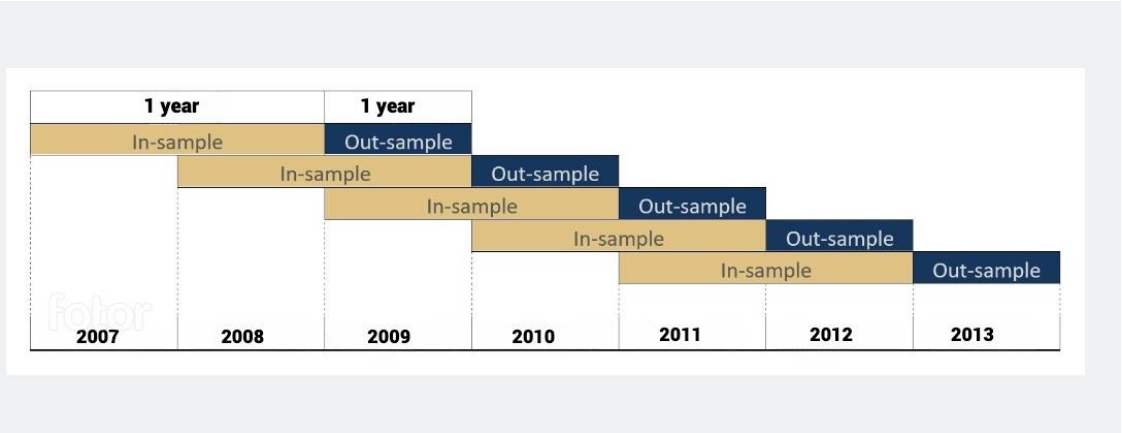
- ## Gain Importance

  It refers to the average gain of a feature when it is used in trees. It's a measure of the contribution of each feature to the model's performance. The most important is leverage (118.79) and legal structure is least important (12.93)

- ## Cover Importance

  It measures the average coverage of a feature when it is used in trees. It refers to the number of samples affected by the split based on that feature. The most important is leverage (8820.95) and least important is legal structure (2735.91)

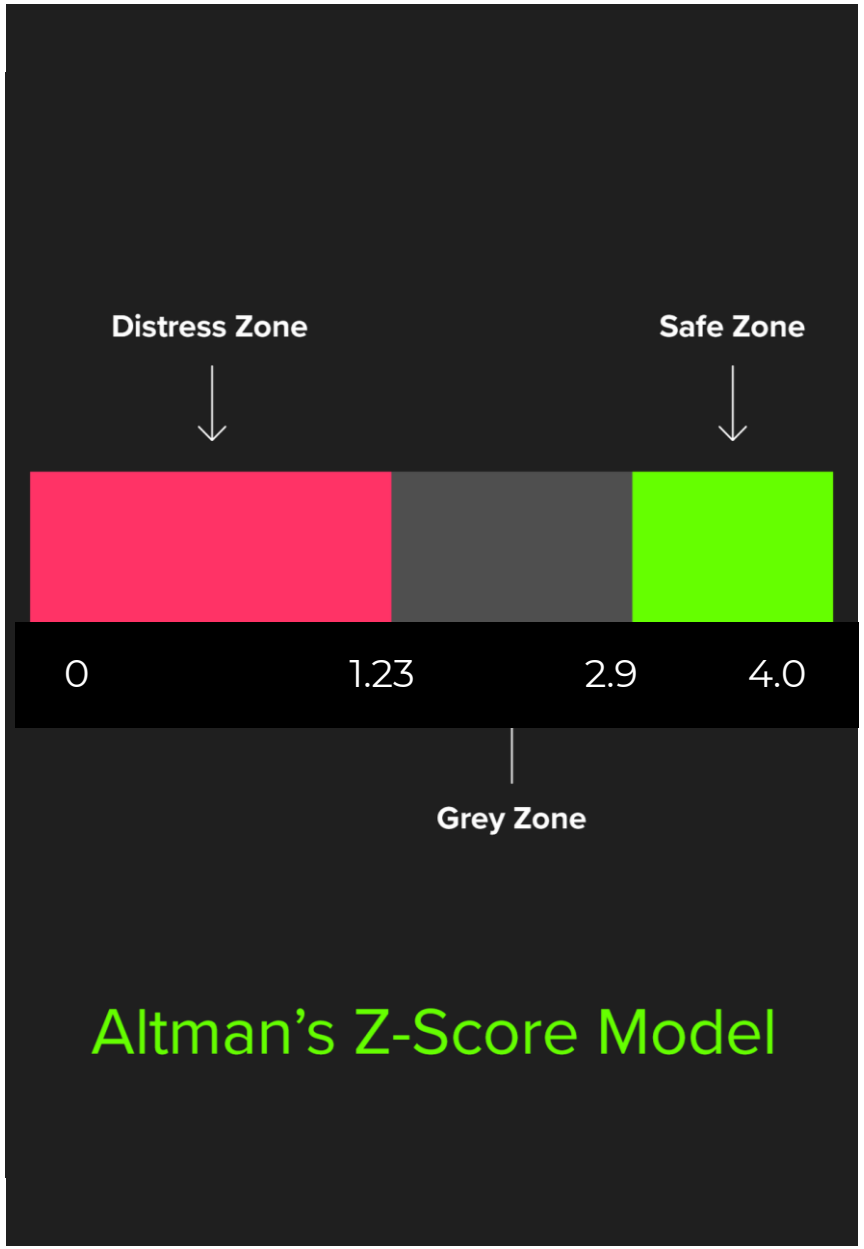

YOU UNDERSTAND NOW
BECAUSE I EXPLAINED IT TO YOU
quickmeme.com



Walk Forward Analysis

We performed a walk-forward analysis on the dataset to train our model. It involves a sequential, step-by-step approach where the model is continually adapted and validated against new data, mimicking real-world situations more closely than traditional backtesting methods.

- Walk forward analysis is used to avoid overfitting.
- We train our model on out-of-sample, out-of-time, and out-of-universe samples to ensure the best performance for our model on unseen data.
- The model is trained for 1st year of data with a holdout sample based on firm year.
- The model is validated on unseen data along with the next year's data and this analysis is carried forward till the last year to train and validate the model.



Introduction to Altman Z-Score

The Altman Z-Score is a financial formula used to predict the likelihood of a company filing for bankruptcy within two years. It was developed by Edward I. Altman in the 1960s and combines multiple financial ratios into a single score that can indicate if a company is headed for potential financial distress.

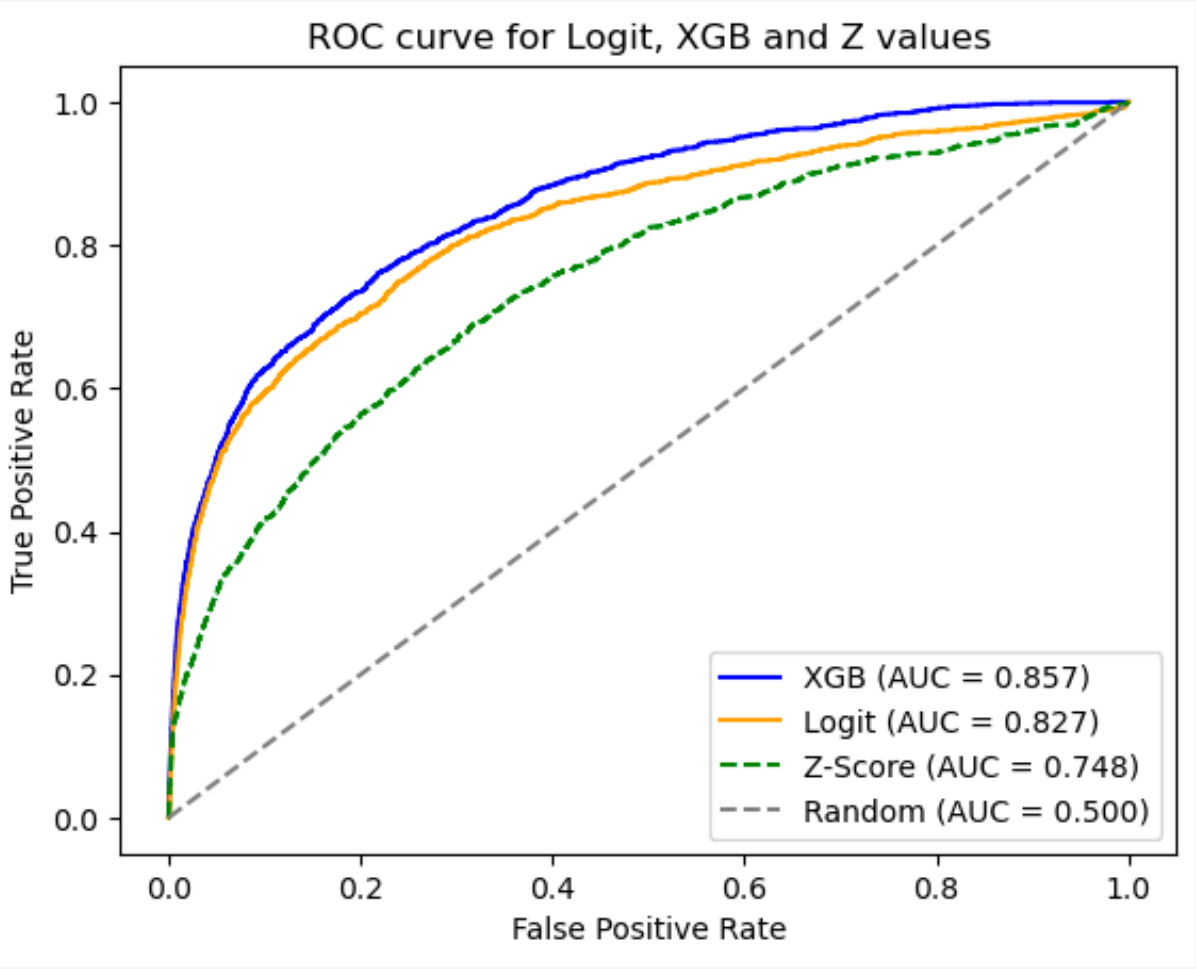
Calculating the Altman Z-Score

The Altman Z-Score provides an objective measure of bankruptcy risk based on established financial ratios. We use it to benchmark our model scores to understand how well or worse our model performs in different setting compared to a well-established and understood objective score.

Percentage ranges from 0 to 100. It is calculated as percentage of unitary change of each ratio.

Results:

Model	AUC Score
Logit	0.827
XGBoost	**0.857**
Z-Score	0.748



Calibration

Model Score

Sort firms by model score

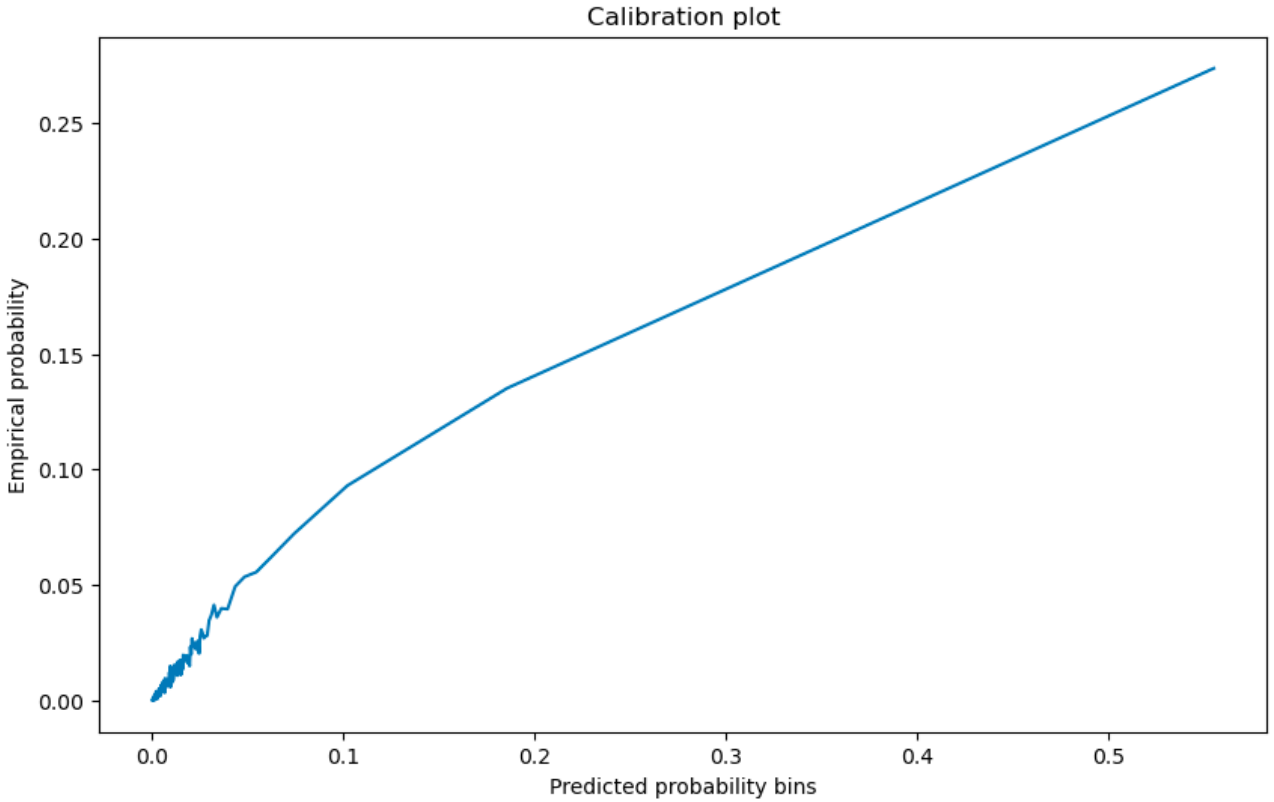
Bucket Size

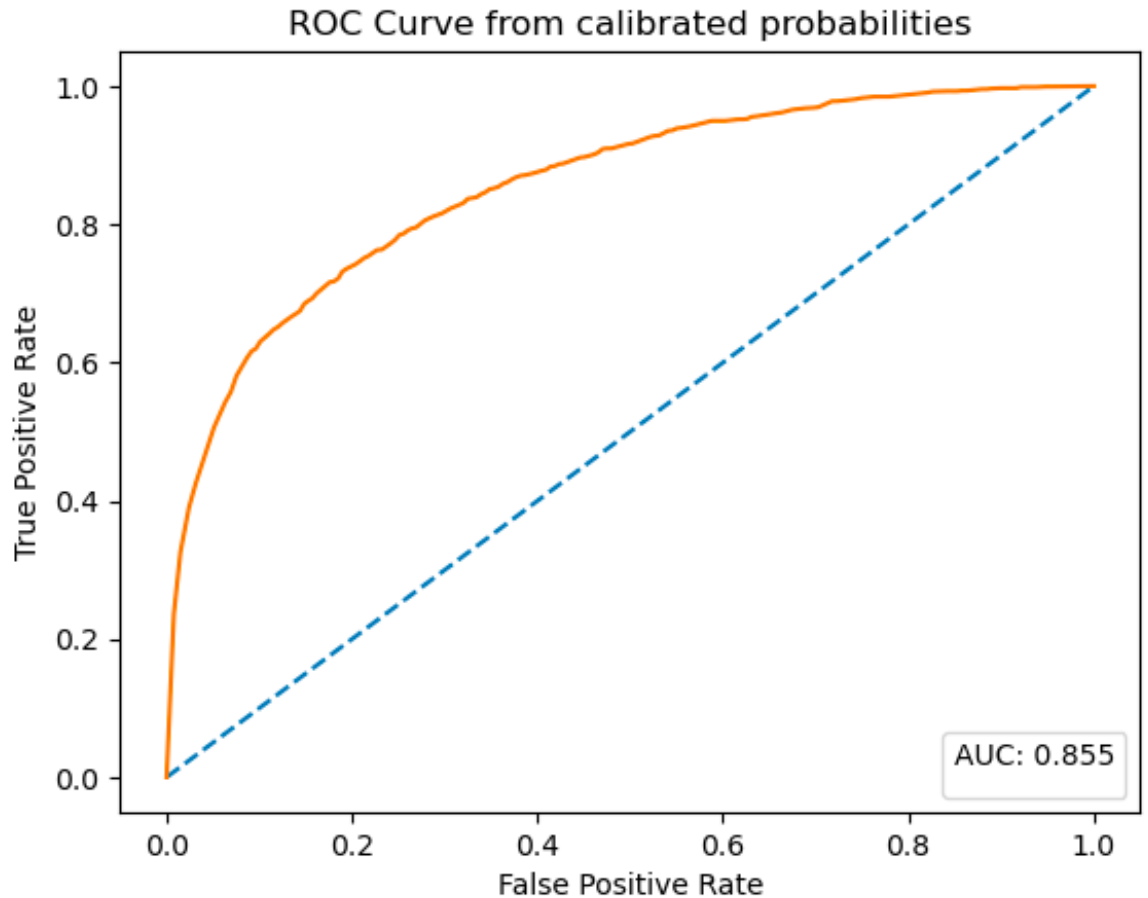
Create *k* buckets, each having *N/k* firms and calculate number of defaulting firms in each bucket.

Estimating non-linear curve

Estimate a (nonlinear) curve that maps quantiles to their default rates and record the value of model output associated with each quantile.

Can I trust my model's probabilities?


Calibration plot
Empirical probability
Predicted probability bins
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.1
0.2
0.3
0.4
0.5


ROC Curve from calibrated probabilities
True Positive Rate
False Positive Rate
AUC: 0.855
0.0
0.2
0.4
0.6
0.8
1.0

Conclusions

- ## Benchmarking Results

  Our model significantly outperforms a good baseline model of Altman Z-Score.

- ## XGB vs Logit

  While a logit model is more explainable, we have preferred XGB as it gave slightly more powerful results. We performed sensitivity analysis through counterfactuals for better explainability along with a univariate and multivariate analysis.

- ## Accuracy

  Our final XGB model has an AUC of 0.86 on the validation set in a walk forward analysis.

- ## Feature Importance

  Based on our analysis, our onput features in descending order of their importance are: Leverage, CFO ratio, Return on Assets, Current Ratio, Fixed Assets, Legal Structure.

- ## Calibration

  After calibration, our model gives probabilities which are more aligned with real-world probabilities while not affecting the AUC score significantly.

Contributions

The contributions of the three of us are briefly outlined below.

Pradhyumn Bhale

Major contributions in reports, slide deck, and problem formulation. Minor contribution in coding.

Shrey Jasuja

Major contributions in coding, and problem formulation. Minor contributions in report and slide deck.

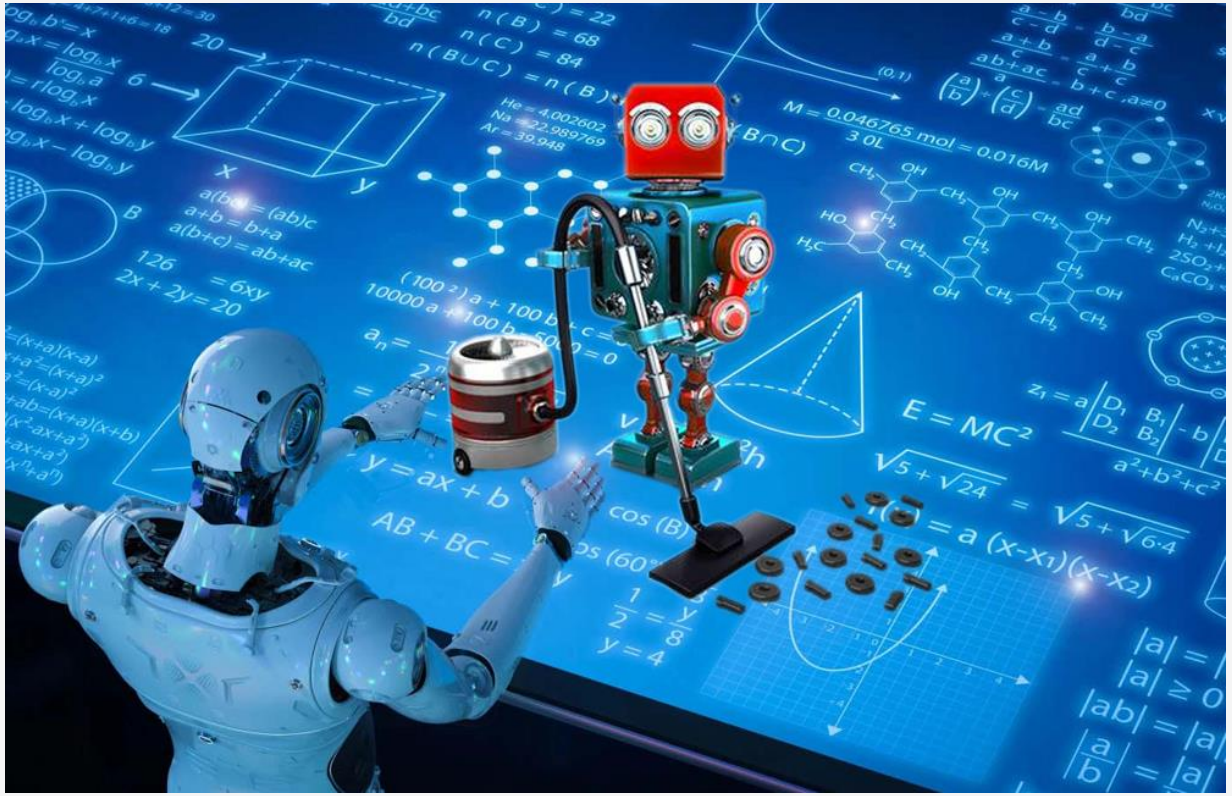
Inderpreet Singh Walia

Major contributions in coding, and problem formulation. Minor contributions in report and slide deck.


ASK ME ANYTHING.

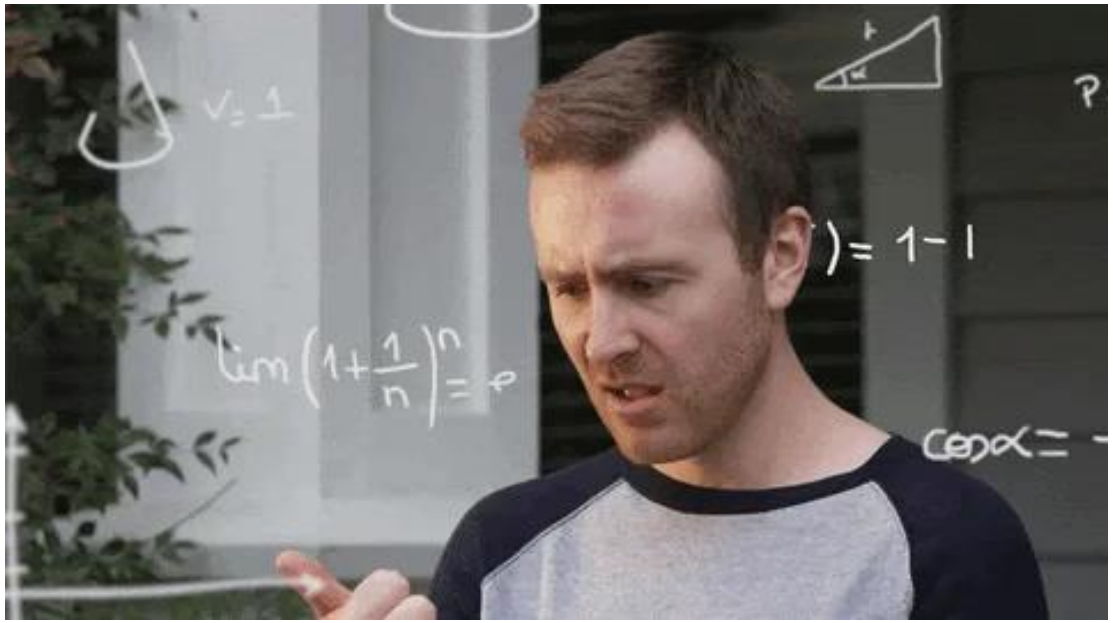
Appendix

An appendix providing additional details on selected slides



Data Cleaning

- Removed HQ_city, fs_year, eqty_corp_fam_tot, and other unnecessary columns from the dataset.
- Removed NaNs from cash and equiv and Cf_operations as we weren't able to impute them using financial know-how.
- Imputed missing values using finance knowledge as explained below.
- Calculated *net income* by multiplying *roa* with *total assets*.
- Filled in the missing *roe* values using the calculated *net income* & divided it by *total equity*.
- Calculated *non-current assets* by subtracting *total assets* with *current assets*.
- Calculated *operating revenue* by adding *COGS* and *operating profits*.
- Calculated *total liabilities* by subtracting *total equities* from *total assets*.
- Calculated *fixed assets* by adding *asset intangible fixed*, *asset tangible fixed* and *asset fixed fin*.
- Calculated *current liabilities* by subtracting *net working capital* from *current assets*.
- Outliers outside 3-97% interval were removed.

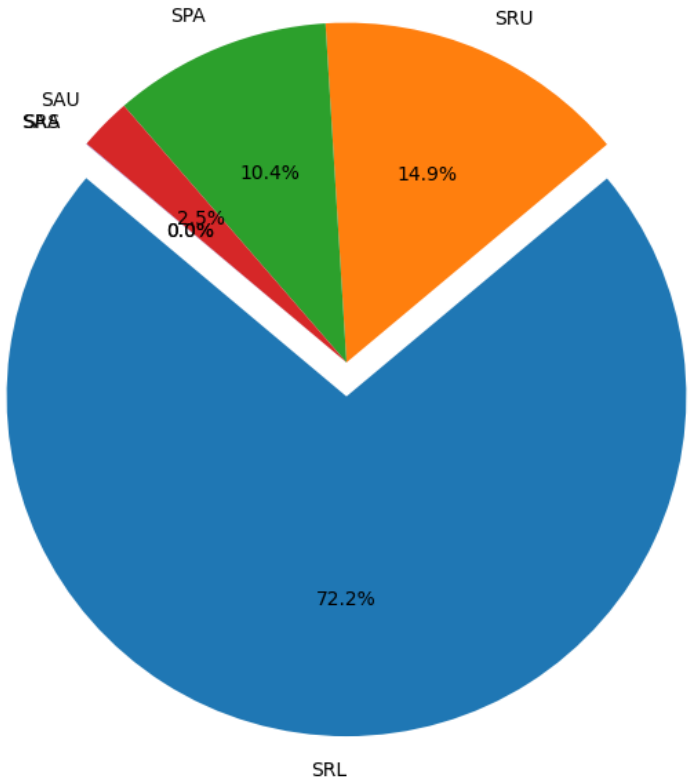


Sanity Checks

- Current liabilities was greater than total liabilities (<0.01% times). These rows were removed.
- Rev_operating - COGS = Prof_operation (>99% match).
- Current liabilities = short term debt with 99.9% match.
- Sum of liability and equity after imputations equals assets. (>99%)
- We observed that <1% data had negative COGS. After discussion with professor, we didn't remove these.
- Negative COGS and negative income were observed. This could be accounted for corrections on overstated costs.
- Made the simplifying assumption of non-current liabilities to be long term debt.
- Skewness and kurtosis of data is decreasing on cleaning and performing relevant transformations.
- Net income positive with negative taxes were removed as these were observed as erroneous data.
- While calculating financial ratios, we made sure that division by zero wasn't encountered.

Legal Structure of the Firm

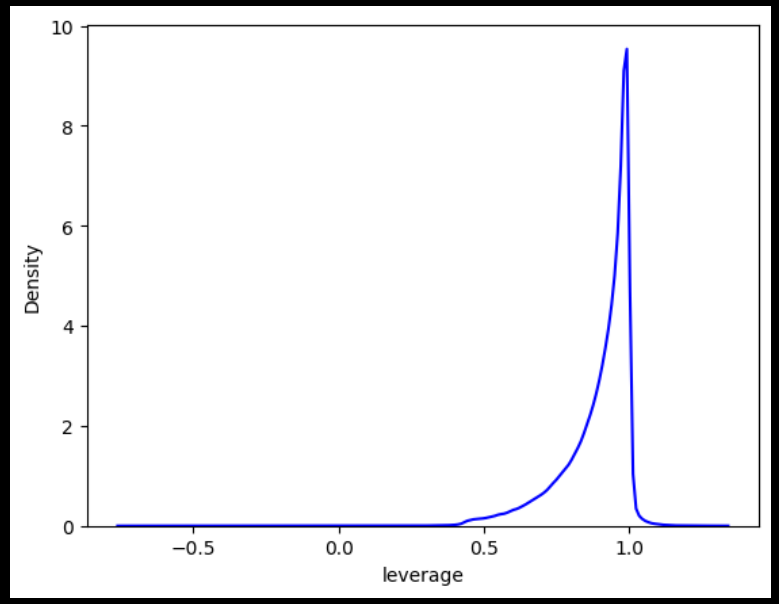
Previous studies have often highlighted size of the firm as a significant factor in their outcomes. To account for this, we employed target encoding for the legal structure variable. It appears logical that the base default rate is higher for single-unit entities (SRUs) on average, while public-sector companies (SPAs) possess a rate lower than average. This could be attributed to the size advantage that SPAs usually possess in comparison to SRUs.



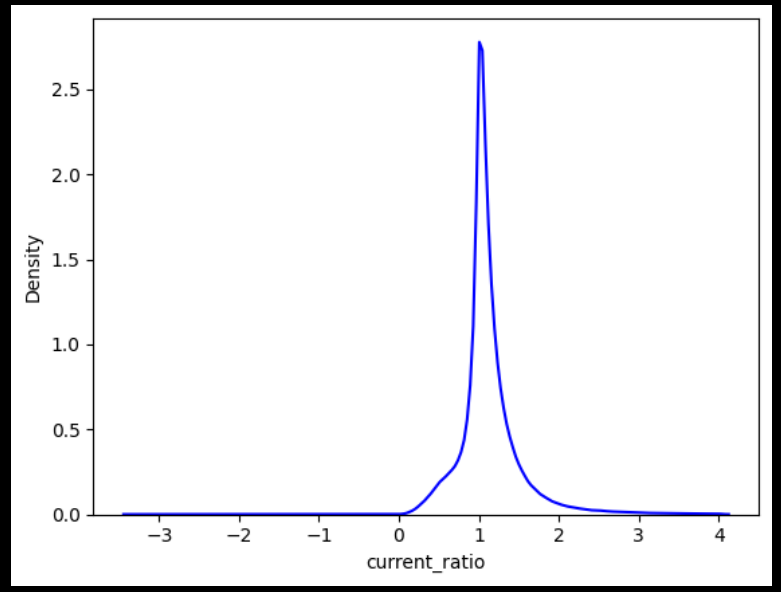
Mean Default rate: 1.29

1. SRL: 1.24
2. SPA: 0.91
3. SRU: 1.85
4. SRS: ~~10.00~~ 1.0
5. SAA: 0.81
6. SAU: 0.86

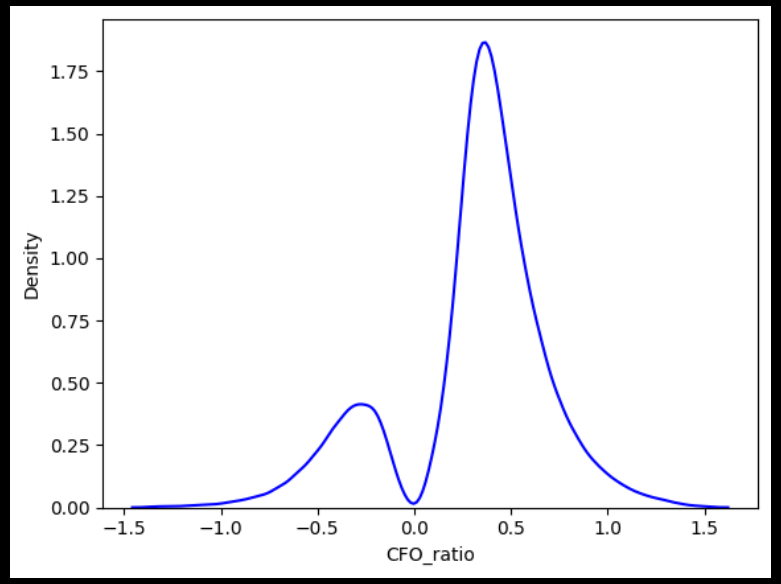
Exploratory Data Analysis



Leverage



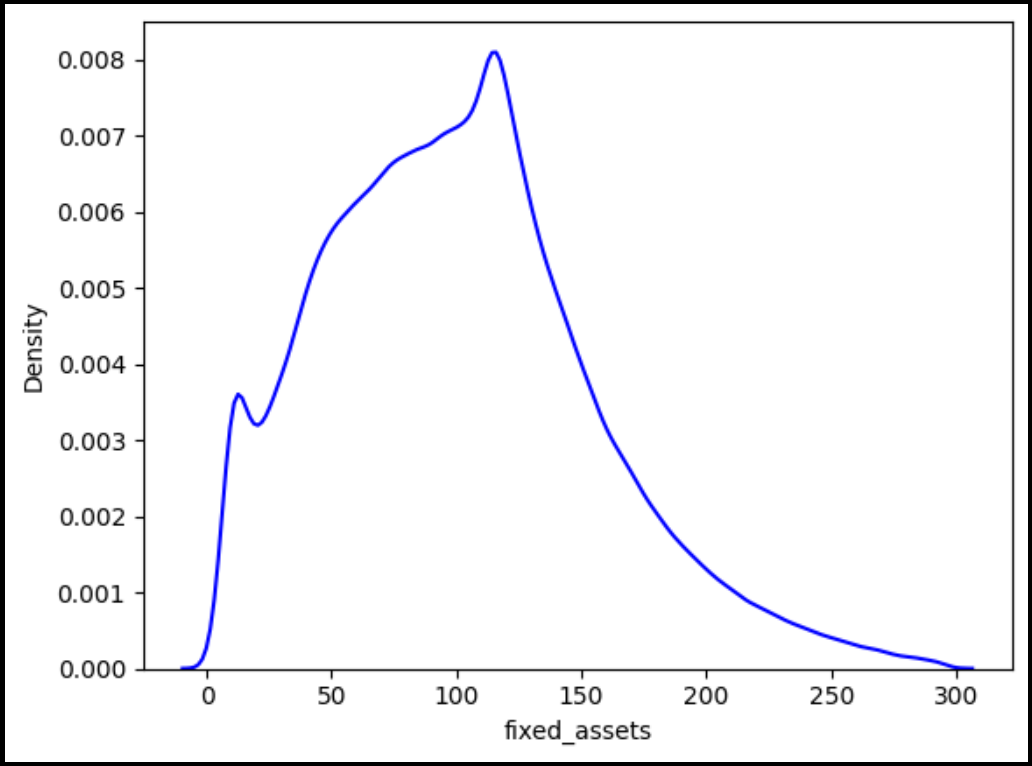
Current Ratio



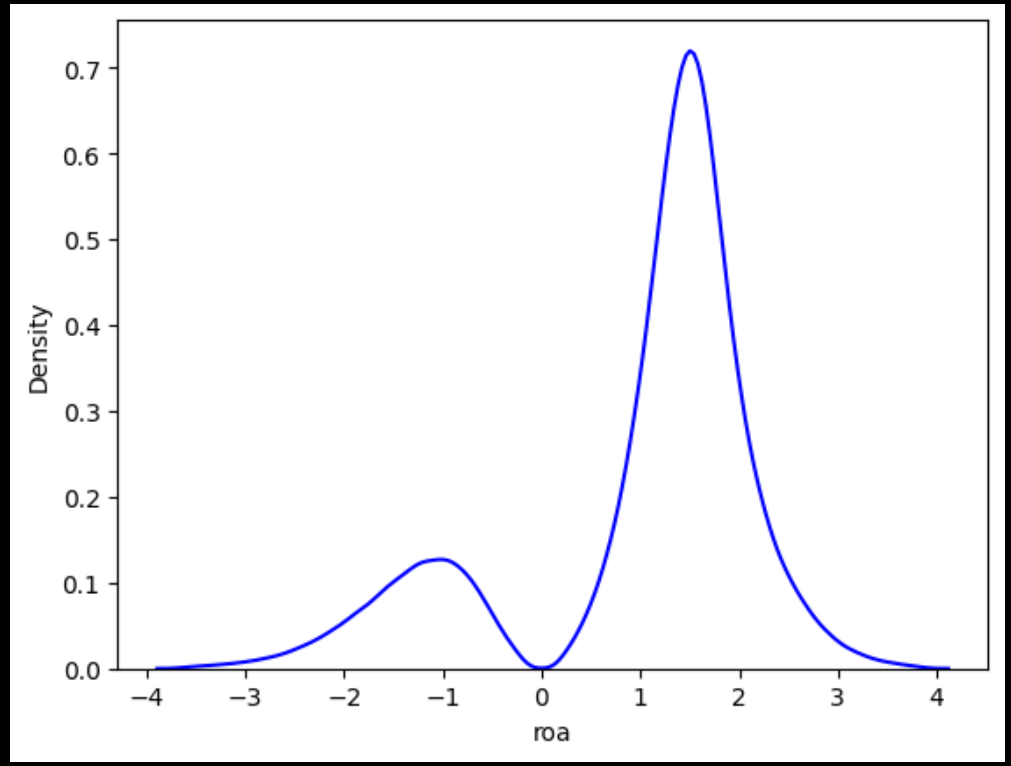
CFO Ratio

Visualizations for the independent variables.

Exploratory Data Analysis



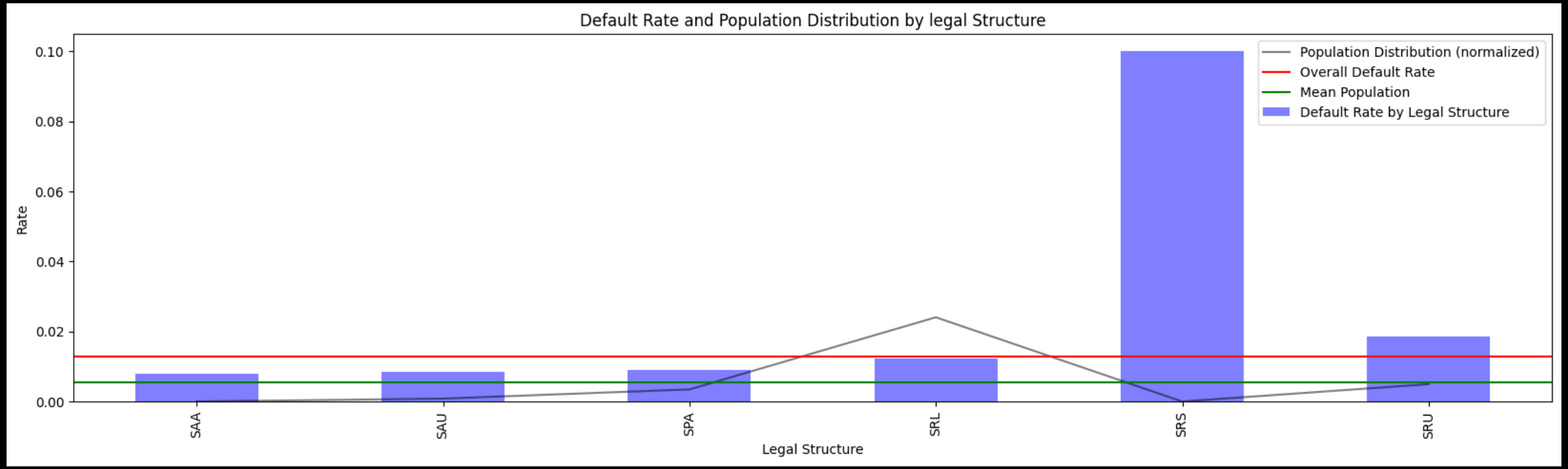
Fixed Assets



ROA

Visualizations for the independent variables.

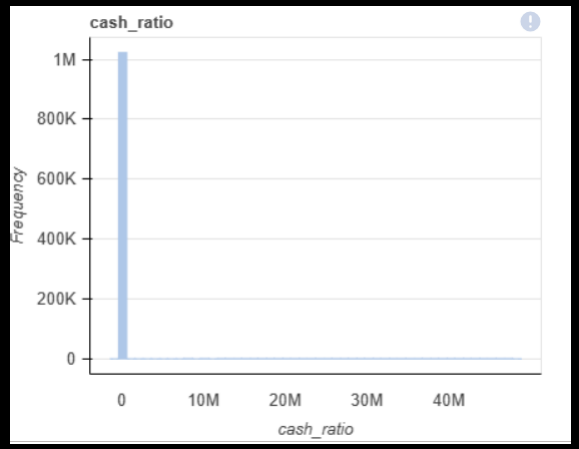
Exploratory Data Analysis



Legal Struct

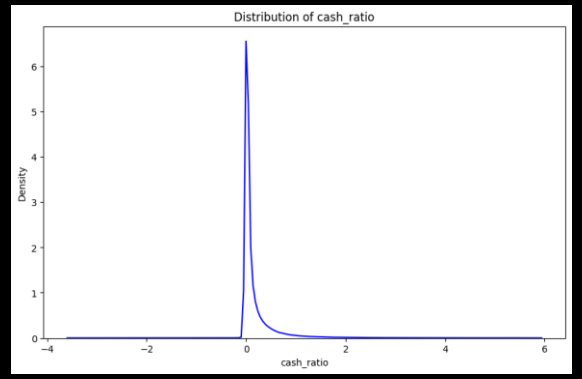
Visualizations for the independent variables.

Exploratory Data Analysis



- Original (Uncleaned) data

Raw data is hard to observe and infer for conclusions.

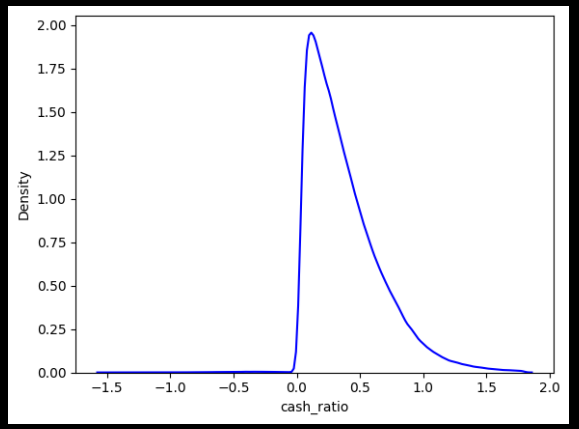


- Data on cleaning

Observe the skew in the distribution

- Transcendental transformations

Transformations (like log(x), $\sqrt[3]{x}$) are important to make the distribution more normally distributed.



- After Transformation

The skew and kurtosis in the data has reduced after $\sqrt[3]{x}$ transformation.

These were found to be more essential for logit model but not for Tree-based models (XGB).

Univariate Analysis Scores

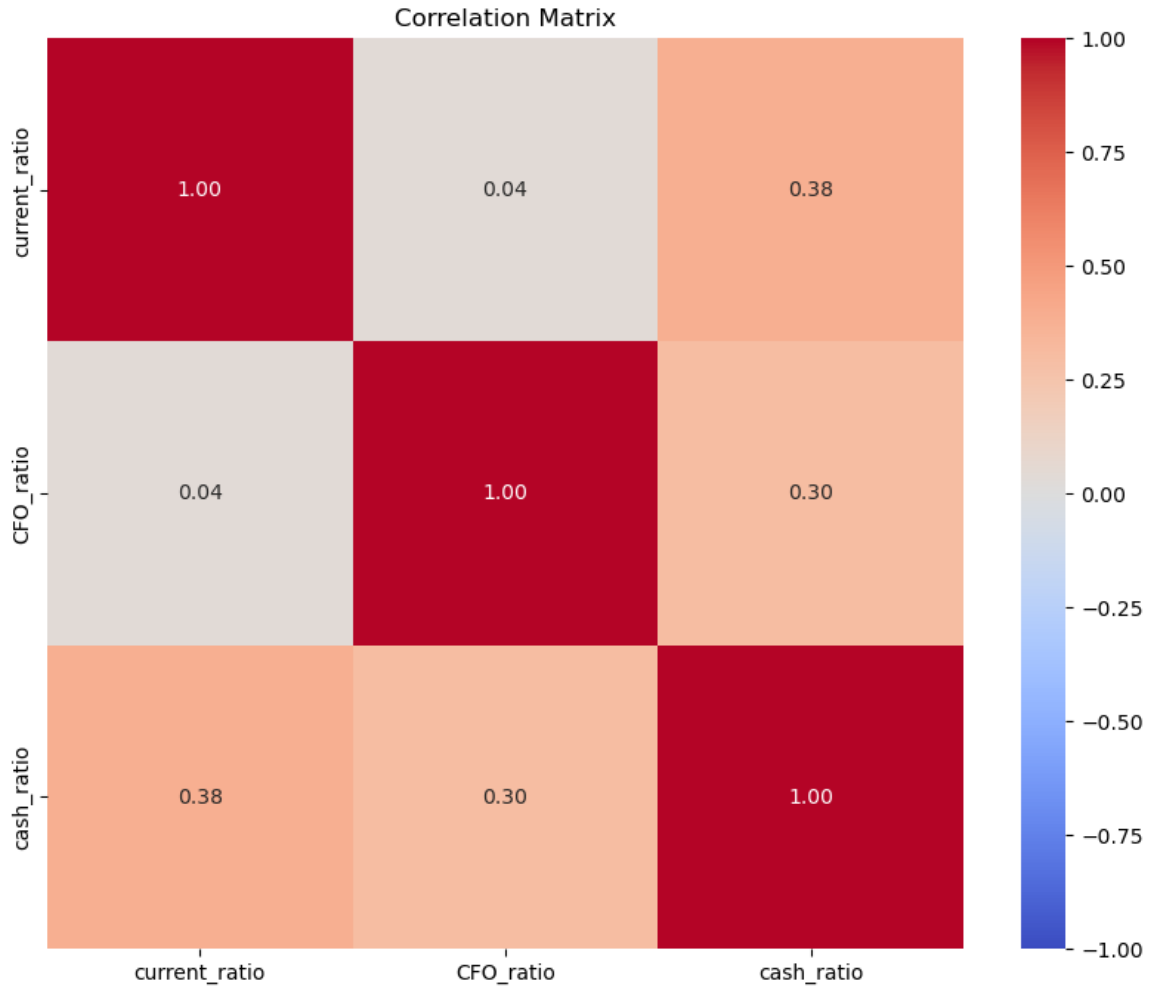
These scores are generated using mutual information score which is a measure of dependency between the target variable and independent variable.

1. Gross Profit Margin (0.003)
2. Legal Struct (0.12)
3. CFO to total liab (0.006)
4. CFO ratio (0.005)
5. Asset Cov ratio (0.002)
6. DSCR (0.005)

1. Current Ratio (0.001)
2. Leverage (0.006)
3. Cash Ratio (0.001)
4. ROA (0.005)
5. Fixed Assets (0.001)

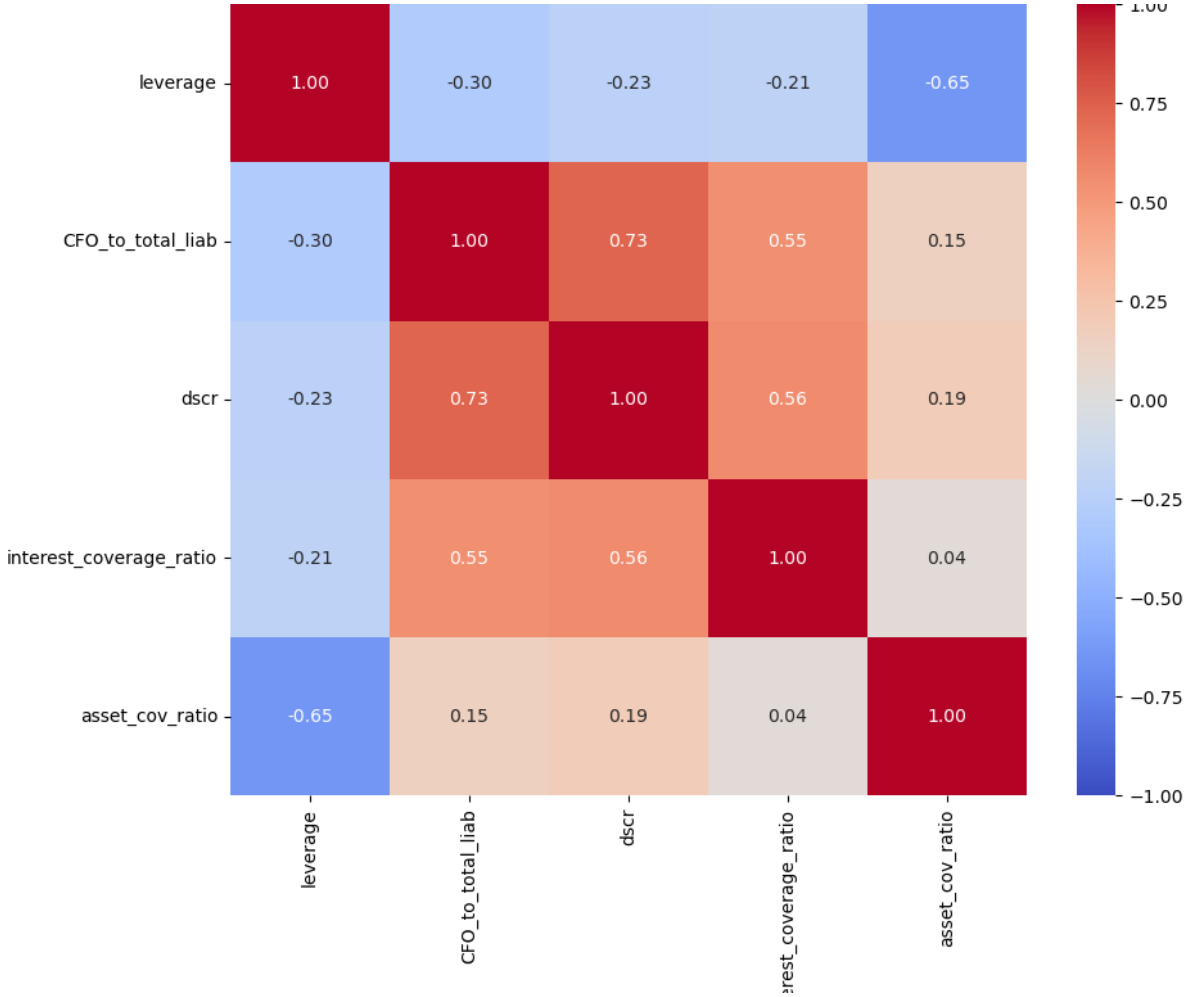
Variance Inflation Factor (ViF) - Liquidity

Variables	ViF
Current Ratio	3.26
CFO Ratio	1.71
Cash Ratio	3.54



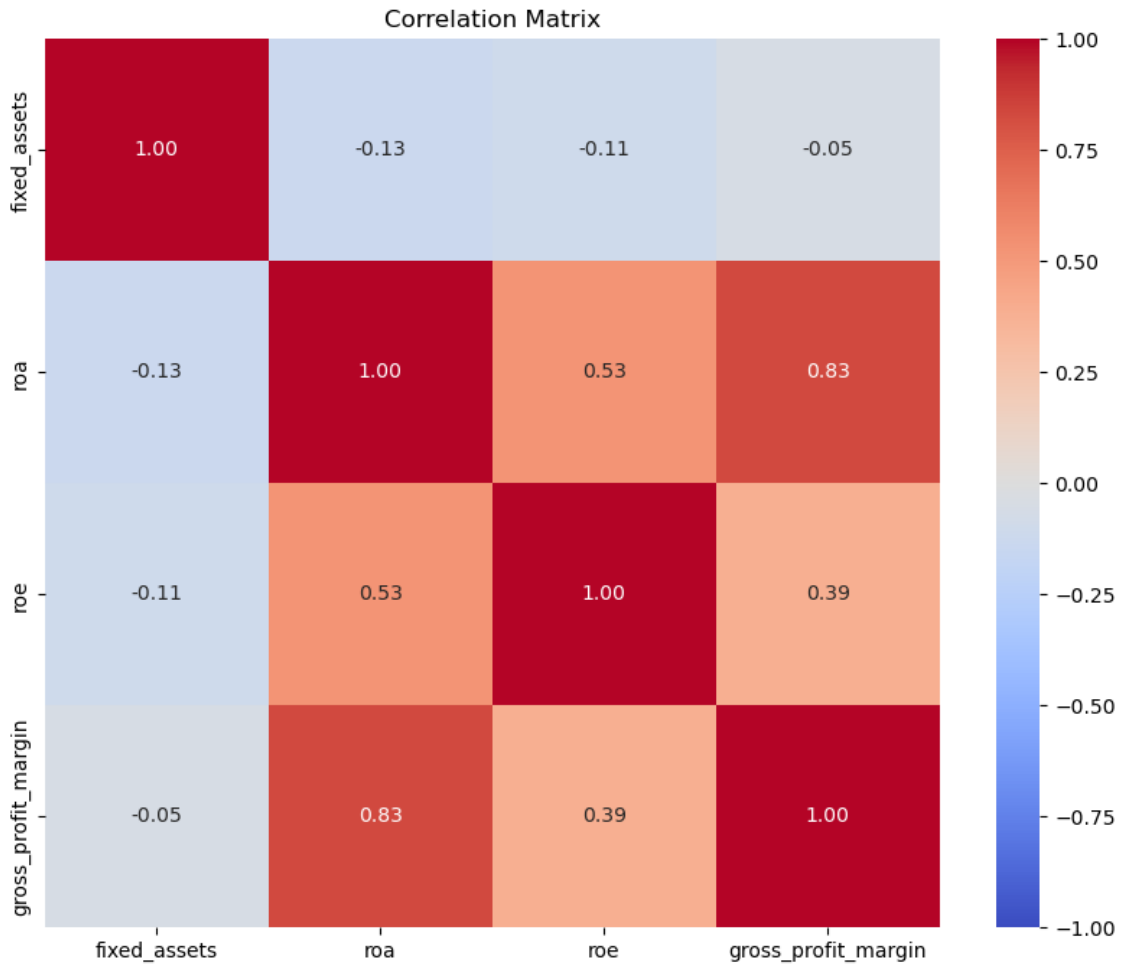
Variance Inflation Factor (ViF)- Debt Coverage

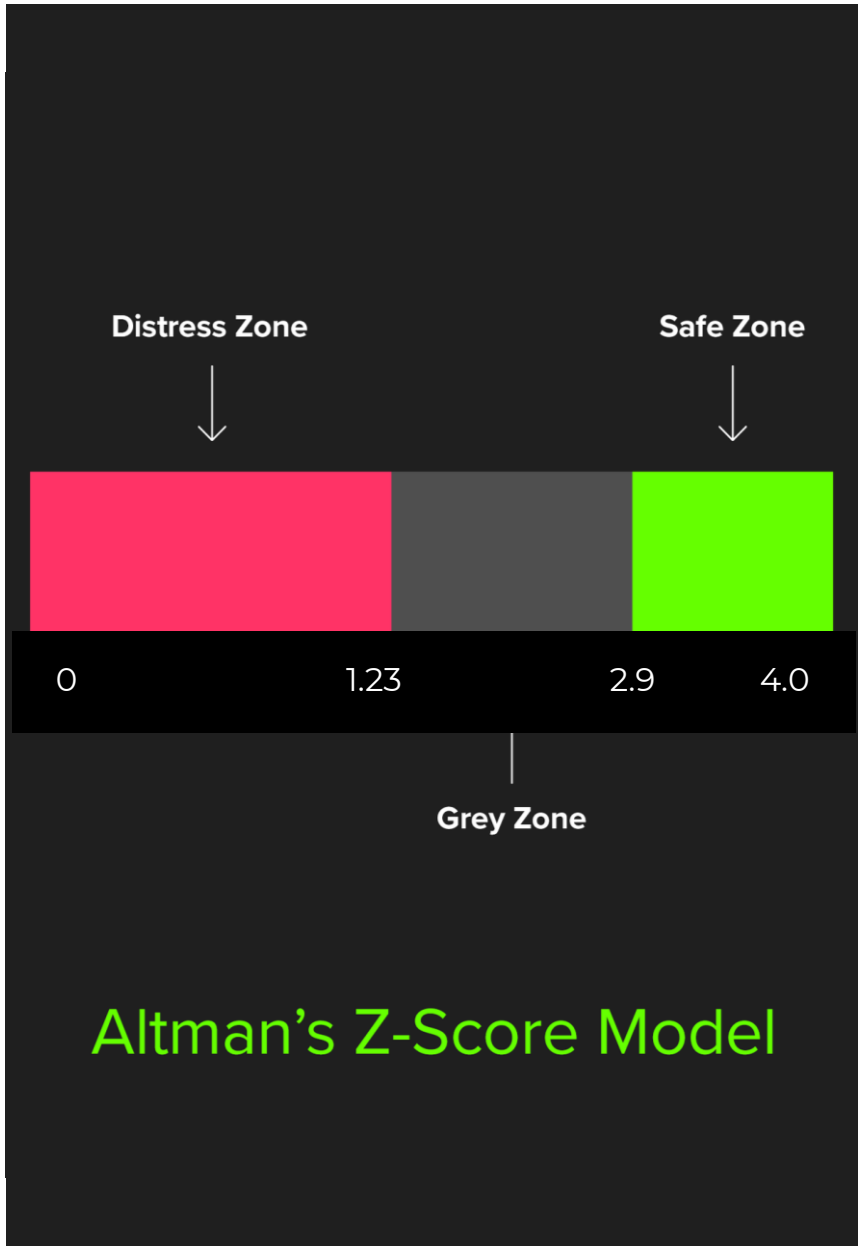
Variables	ViF
Leverage	3.74
CFO to total liabilites	4.07
DSCR	5.72
Interest Coverage Ratio	2.37
Asset Coverage Ratio	3.67



Variance Inflation Factor (ViF)- Profitability

Variables	ViF
Fixed Assets	1.37
ROA	1.56
ROE	5.61
Gross Profit Margin	4.24





Calculation of Altman Z-Score

Generally, the Altman Z-score for private companies would give you the likelihood that the company would default within the next 2 years.

Altman Z-Score = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E

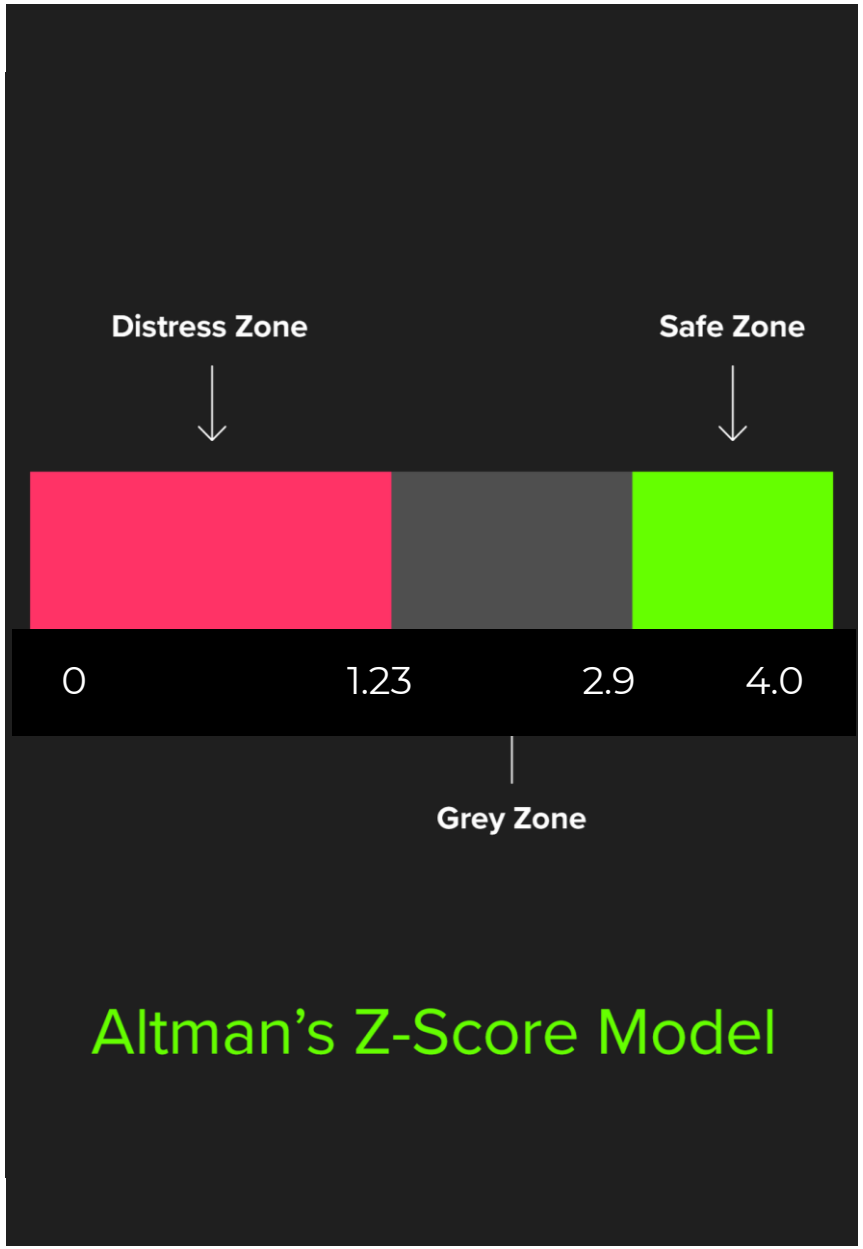
Where:

- A = working capital / total assets
- B = retained earnings / total assets
- C = earnings before interest and tax / total assets
- D = market value of equity / total liabilities
- E = sales / total assets

Under following assumptions:

1. net_income as proxy for retained earnings
2. rev_operatings as proxy for sales

Source: https://www.investopedia.com/terms/a/altman.asp

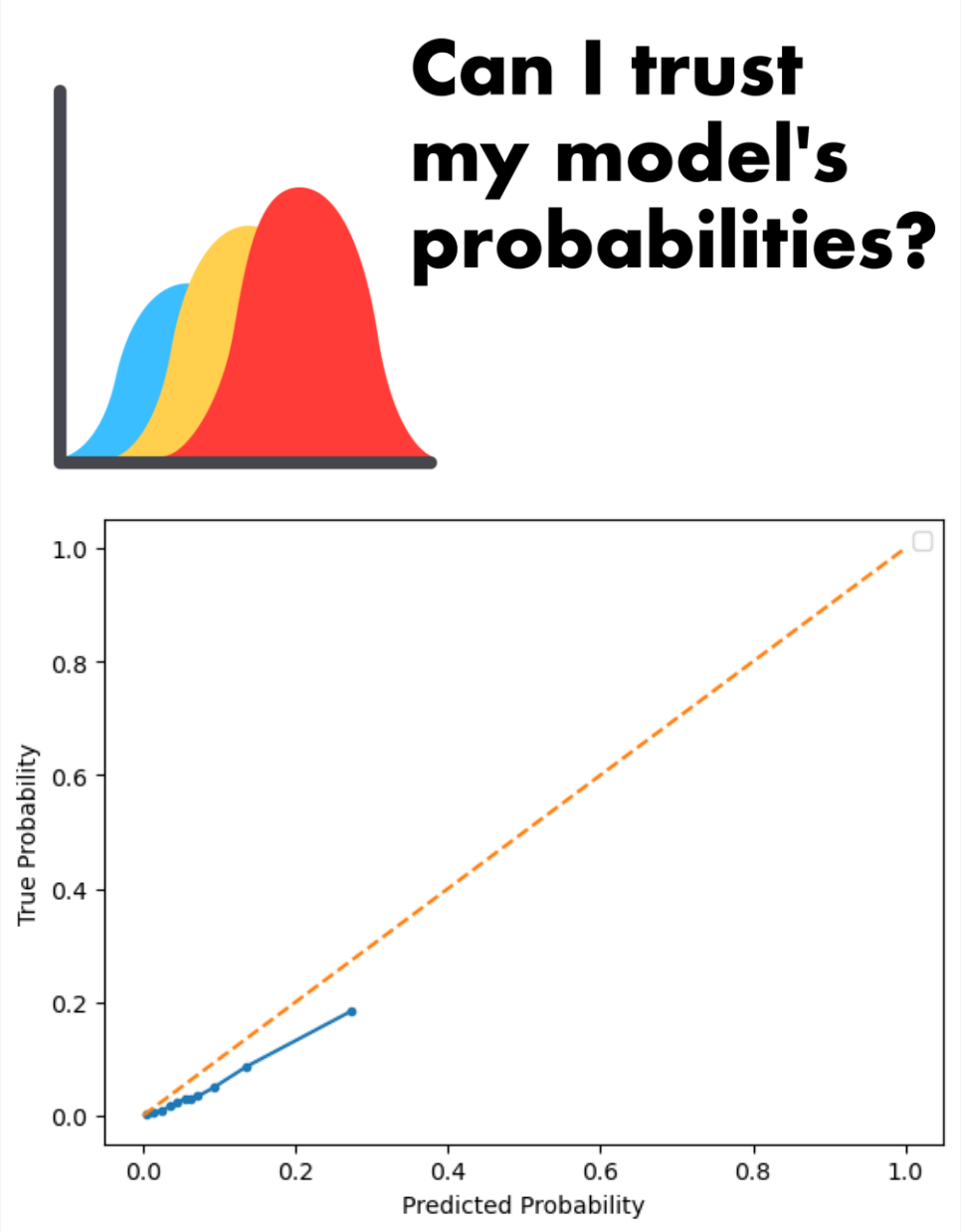


Calculation of Altman Z-Score

Generally, the Altman Z-score for private companies would give you the likelihood that the company would default within the next 2 years.

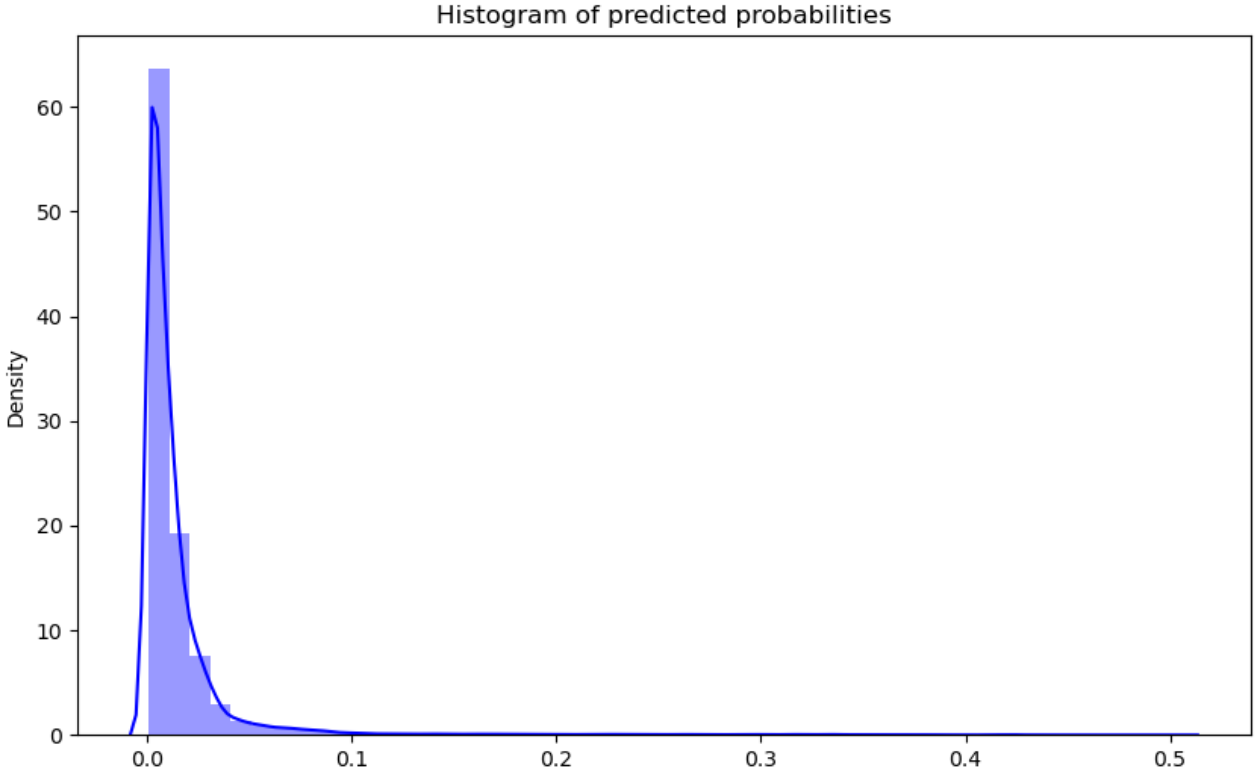
How to map it to 12 months?

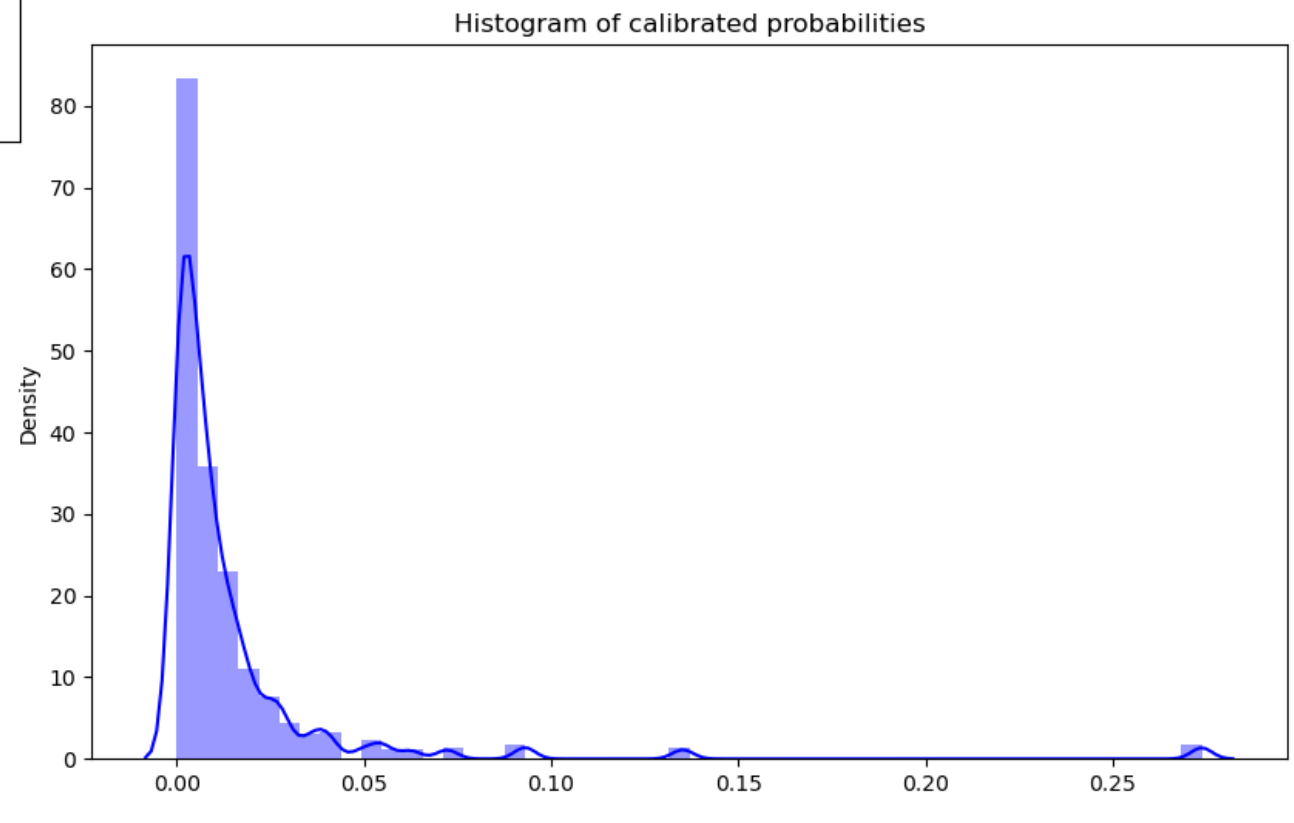
- Here, we didn't use the provided bands as is.
- Instead, we trained a logit model over the observed Z-score to predict the probability of defaulting in 12 months.



Calibration

- Sort all firms by model output and create *k* buckets, each having *N/k* firms.
- Calculate the number of defaulting firms in each bucket and divide this by N/k to obtain the default rate in each bucket.
- Estimate a (nonlinear) curve that maps quantiles to their default rates and record the value of model output associated with each quantile.


Histogram of predicted probabilities
Density
0
10
20
30
40
50
60
0.0
0.1
0.2
0.3
0.4
0.5


Histogram of calibrated probabilities
Density
0
10
20
30
40
50
60
70
80
0.00
0.05
0.10
0.15
0.20
0.25

Counterfactual Analysis (on XGB)

Counterfactuals are "what-if" scenarios where a model is tested by systematically generating synthetic data by modifying features of the factual data to understand the impact of different variable on outcomes (probability of default) in our case.

- ## Decision to explain

  We begin by setting our target variable (PD) as the model decision to be explained.

- ## Defining Causes

  Used standard machine learning techniques to train a model to predict default risk based on the factual data

- ## Counterfactual data generation

  Programmatically generated synthetic counterfactual data by modifying features of the factual data

- ## Develop Counterfactual Scenarios

  Evaluated the trained model on the counterfactual data to analyze its sensitivity on independent variables.

- ## Drawing Insights and conclusions

  Our analysis concluded with CFO ratio and current ratio having the highest sensitivity while fixed assets having the lowest sensitivity.

Summary

- Most frequently used for Splits:

  CFO ratio and leverage

- Most Impactful (Gain)

  Leverage, followed by CFO ratio

- Wildest Influence (Cover)

  Leverage, indicating its splits involve the most observations.

- Least Impactful

  legal structure, both in terms of frequency and gain.

References

1. Bank of Slovenia. (2014). Estimating the probability of default and comparing it to credit rating classification by banks.

2. Bohn, J. R., & Stein, R. M. (2010). Active Credit Portfolio Management in Practice. John Wiley & Sons.

3. Investopedia.com

4. Stein, R. (1993). Preprocessing data for neural networks.

5. Tasche, D. The art of probability-of-default curve calibration.

6. Stein, R. (1993). Selecting data for neural networks.

7. Amel-Zadeh, A., Glaum, M., & Sellhorn, T. (2023). Empirical Goodwill Research: Insights, Issues, and Implications for Standard Setting and Future Research. European Accounting Review, 32(2), 415-446. https://doi.org/10.1080/09638180.2021.1983854