Corporate Bankruptcy in the Manufacturing Sector:

Analysis on the Leading Financial Indicators

Mary Claire C. Doña
301323966

Bilal Hasanzadah
Project Advisor

David Parent
Program Director

Business Analytics Capstone Project

August 15, 2024

Table of Contents

Model Exploration

Model Recommendation

Key Insights, Conclusion, Recommendations

Executive Summary

1.0. Executive Introduction

Among all industries in Canada, the manufacturing sector has one of the highest rates of corporate bankruptcy. Weakness in financial management is one of the primary factors and oversight in early warning signs of financial distress is one of the underlying reasons. To address this challenge, an analysis was conducted to identify the leading financial indicators of corporate bankruptcy in manufacturing firms. This report serves as a documentation of how the analysis was executed, the methodologies employed, the key findings, recommended actions, and the corresponding business impact.

2.0. Executive Objective

The purpose of the aforementioned analysis is to identify which financial ratios along with their critical thresholds signal early indicators of financial distress, leading manufacturing firms towards bankruptcy. The insights gained from this analysis will help the financial management team of Canadian manufacturing firms to mitigate bankruptcy risk and ensure long-term financial stability.

3.0. Executive Model Description

Data analytics was leveraged in conducting this analysis. Using the financial ratios as data inputs, various predictive modeling techniques and machine learning algorithms were explored to classify bankrupt from operating firms. Basic models such as Decision Tree, Logistic Regression, and Neural Network were employed. Advance tree-based algorithms were also explored to address the limitations of basic models. The best model was selected based on its predictive and interpretability power; which meets the objectives of the analysis. Main evaluation metrics used are F1-score and ROC_AUC. **Logistic Regression** with F1-score of 17% and ROC_AUC of 74% came off as the best model complemented by *Random Forest* and *Recursive Feature Elimination* as feature selection techniques. A comprehensive analysis was further conducted comparing firms with highest and lowest probability of bankruptcy to determine the critical thresholds of the identified financial ratios.

4.0. Executive Recommendations

The results of the analysis revealed that the leading financial indicators of corporate bankruptcy in manufacturing firms are: *negative gross profit margin, negative retained earnings, and constant capital to*

asset below 50%. These findings suggest that manufacturing firms with high debt leverage, but with significant cost inefficiencies, and weak earnings capacity are at the highest risk of going bankrupt. The recommended solutions involved improving financial practices such as: proactive financial planning, robust cost monitoring, and optimal capital structure.

Introduction

1.0. Background

A formal legal process known as corporate bankruptcy is undertaken when a firm is experiencing financial distress, typically when it is unable to meet its financial obligations. In Canada, there are two forms of corporate bankruptcy proceedings: liquidation and reorganization (Smith, 2022). For the purpose of this study, bankruptcy is related to liquidation wherein the company permanently ceases operations, dissolves the business and the remaining assets are used to pay off obligations.

Corporate Bankruptcy filings in Canada are averaging 2,400 from 1990 to 2023 based on the historical records of the Office of Superintendent of Bankruptcy. In fact, the most filings in 25 years were recorded in 2023 at 3,107 (Appendix, Figure 1). Among all industries, the manufacturing sector has the highest bankruptcy rate; an average of 0.55% of its corporations filed bankruptcy from 2004 to 2009, the pre-pandemic period (Appendix, Figure 2).

Corporate Bankruptcy involves significant negative implications not just on the struggling firm, but also to its stakeholders. It entails administrative and legal costs associated to winding down the business and loss from the decline in asset value when debts are settled. Impact to company stakeholders is also detrimental: (1) low returns on investment or, worse, lost capital for shareholders, (2) job loss for employees, (3) disrupted operations and unsettled transactions for business partners.

2.0. Problem Statement

There is a multitude of underlying factors contributing to corporate bankruptcy, both internal and external to the firm. Baldwin et al. (1997) outlines that almost half of the firms in Canada go bankrupt due to deficiencies in internal management; in fact, the second key factor of business failure is weakness in financial management which accounts for 71% of companies filing for corporate bankruptcy (p.26). This includes weak financial planning and inability to provide sound recommendations with respect to how

management should improve operation to ensure sustainable business health. The inability to detect early indicators of financial distress is one of the underlying reasons of weakness in financial management.

3.0. Objectives& Measurement

3.0.1 Business Objective

The purpose of this study is to identify the leading financial indicators of corporate bankruptcy in manufacturing firms. The insights gained in this study can guide the financial management team of Canadian manufacturing firms as to what financial metrics they should closely monitor early on. They can then make informed decisions on how to mitigate risk of bankruptcy and ensure firms' long-term financial stability.

3.0.2 Analytics Objective

- **Main Research Question**

  This study aims to answer the question of: Which financial ratios signal early warning signs of financial distress that could lead manufacturing firms into corporate bankruptcy?

- **Question Design**
  1. Which financial ratios within a 5 year time horizon consistently indicates bankruptcy risk?
  2. What are the critical thresholds of financial ratios that strongly distinguish between bankrupt and operating firms?

- **Hypothesis Formulation**
  1. Financial ratios related to the firm's *profitability and solvency* are the key drivers of corporate bankruptcy in the manufacturing sector.
  2. Financial statements dating back up to five years can be used to detect bankruptcy risk.

4.0. Assumptions and Limitations

4.0.1 **Data Assumptions**

- The financial ratios derived from financial statements are accurate and reliable.
- The accounting practices employed by Polish and Canadian firms are comparable.

4.0.2 **Analytics Assumptions**

- Since the dataset under study is comprised of financial data from Polish companies, it is assumed that manufacturing firms track the same financial metrics regardless of their country of origin or operating environment. Hence, the findings in this study can be used to generalize the recommendations in relation to Canadian business context.

4.0.3 **Limitations**

Non-financial and Macro economic factors are beyond the scope of this study.

Data Sources

1.0. Data Set Introduction

The dataset was sourced from UCI Machine Learning repository. It is comprised of accounting ratios derived from financial statements of Polish manufacturing firms; compiled by Polish economists and originally obtained from Emerging Markets Information Service (EMIS).

The observation window for bankrupt firms is from 2000 to 2012 while operating firms is from 2007 to 2013. The dataset has 5 subsets based on the interval of financial data from status year. Each dataset has a total 65 numeric variables: 1 binary integer (target) and 64 continuous float (input).

Dataset	Years Interval from Observed Status	Rows and Columns
1st Year	5	7027, 65
2nd Year	4	10173, 65
3rd Year	3	10503, 65
4th Year	2	9792, 65
5th Year	1	5910, 65

Table 1. Dataset breakdown

2.0. Exclusions

Financial data that are deemed irrelevant and devoid of explanatory significance are excluded. It includes ratios that while theoretically sound, are not applied in real-life practice. To ensure comparability, these types of ratios were excluded.

3.0. Initial Data Cleansing or Preparation

3.0.1 Data Collection

The five datasets were loaded and converted from attribute-relation file format (arff) into pandas dataframe. After cross-checking their shapes to ensure they contain the same variables, year tagging was added respectively to be used as one of the input variables. The five data subsets were subsequently consolidated into one dataset.

3.0.2 Renaming Variables

The input variables are financial statement data expressed in varying units of measure: *ratios, indices, days, and absolute amounts*. To facilitate efficient data pre-processing and interpretation, the variables were renamed and their sequence were reordered. Naming conventions were structured according to metric classification and units of measure.

- Initial letter as metric class: *E: efficiency, L: liquidity, P: profitability, S: solvency, V: leverage*
- Last letter as unit of measure : *R: ratio, D: days, T: turnover, I:index*

3.0.3 Setting the right Data Type

The target variable was then converted from object into binary integers (0: operating and 1: bankrupt)

```
#Convert target variable into binary integers
conso['Bankrupt']=conso['Bankrupt'].astype('category').astype('int')
conso.head()
```

4.0. Data Dictionary

The consolidated and updated dataset has a total 43,405 observations and 66 variables. The detailed description of each variable and the corresponding formulas used in deriving its values are provided in the data dictionary below.

Variables	Data Source Description	Financial Description
Bankrupt	0 Operating; 1 Bankrupt	Operating and Bankrupt Companies
Year_Interval	Year of Financial Rates	Interval between financial reporting period vs. observed status
Log_Tasset	logarithm of total assets	Natural log of assets
E_Asset_TO_T	sales / total assets	Asset Turnover
E_Asset_TO2_T	total sales / total assets	Total Asset Turnover
E_Inventory_TO_T	sales / inventory	Inventory Turnover based on Sales
E_Receivable_TO_T	sales / receivables	Receivables Turnover
E_StLiab_TO_T	sales / short-term liabilities	Short-term Liabilities Turnover
E_FixAsset_TO_T	sales / fixed assets	Fixed Asset Turnover
E_DIO_Sales_D	(inventory * 365) / sales	Days Inventory Outstanding based on Sales
E_DIO_COGS_D	(inventory * 365) / cost of products sold	Days Inventory Outstanding
G_SlsG%Index_I	sales (n) / sales (n-1)	Sales Growth Index
L_WorkCap_Asset_R	working capital / total assets	Working Capital to Asset Ratio
L_Current_Ratio_R	current assets / short-term liabilities	Current Ratio
L_WorkCap_FixAsset_R	working capital / fixed assets	Working Capital to Fixed Asset Ratio
L_Opex_StLiab_R	operating expenses / short-term liabilities	Operating Expense to Short-term Liabilities Ratio
L_Opex_TotalLiab_R	operating expenses / total liabilities	Operating Expense to Total Liabilities Ratio
L_QuickAsset_LtLiab_R	(current assets - inventories) / long-term liabilities	Quick Asset to Long-term Liabilities Ratio
L_QuickAsset(ex-AR)_StLiab_R	(current assets - inventory - receivables) / short-term liabilities	Quick Asset(excluding Accounts Receivable) to Short-term Liabilities Ratio

L_QuickAsset_StLiab_R	(current assets - inventory) / short-term liabilities	Quick Asset to Short-term Liabilities Ratio
L_CurAsset_TotalLiab_R	current assets / total liabilities	Working Capital to Total Liabilities Ratio
L_ConstCap_FixedAsset_R	constant capital / fixed assets	Constant Capital to Fixed Asset Ratio
L_QuickWorkCap_NetCost_R	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)	Quick Working Capital to Net Cost Ratio
L_DLAO_D	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	Days Liquid Asset Outstanding
L_DLO_GP+Dep_D	(total liabilities * 365) / (gross profit + depreciation)	Days Total Liabilities Outstanding based on Gross Profit + Depreciation
L_DCPO_D	(current liabilities * 365) / cost of products sold	Days Current Payable Outstanding
L_CashConversion(inc.DPO)_D	rotation receivables + inventory turnover in days	Cash Conversion Cycle Inclusive of DPO
L_DSO_D	(receivables * 365) / sales	Days Sales Outstanding
L_DSPO_D	(short-term liabilities * 365) / cost of products sold)	Days Short-term Payable Outstanding based on Cost of Product
L_DPO_Sales_D	(short-term liabilities *365) / sales	Days Payable Outstanding based on Sales
L_WorkCap_A	working capital	Working Capital
P_ROA(NI)_R	net profit / total assets	Return on Assets based on Net Income
P_ROA(EBIT)_R	EBIT / total assets	Return on Assets based on Operating Income
P_ROA(GP+Eo+FinEx)_R	(gross profit + extraordinary items + financial expenses) / total assets	Return on Assets based on Gross Profit + Extraordinary Items+ Financial Expense
P_GP+Dep_Margin_R	(gross profit + depreciation) / sales	Gross Profit + Depreciation Margin
P_ROA(GP&IntEx)_R	(gross profit + interest) / total assets	Return on Assets based on Gross Profit + Interest Expense
P_ROA(GP)_R	gross profit / total assets	Return on Assets based on Gross Profit
P_GP_Margin_R	gross profit / sales	Gross Profit Margin

P_ROA(OP)_R	profit on operating activities / total assets	Return on Assets based on Operating Profit
P_NP_Margin_R	net profit / sales	Net Profit Margin
P_ROA(3yr_GP)_R	gross profit (in 3 years) / total assets	Return on Assets based on Gross Profit for 3 years
P_GP+IntEx_Margin_R	(gross profit + interest) / sales	Gross Profit+Interest Margin
P_ROA(GP2)_R	profit on sales / total assets	Return on Assets based on Sales Profit
P_GP_Margin2_R	profit on sales / sales	Sales Profit Margin
P_OP_Margin_R	profit on operating activities / sales	Operating Profit Margin
P_ROA(NI_Inv)_R	net profit / inventory	Return on Inventory
P_ROA(EBITDA)_R	EBITDA (profit on operating activities - depreciation) / total assets	Return on Assets based on EBITDA
P_EBITDA_Margin_R	EBITDA (profit on operating activities - depreciation) / sales	EBITDA Margin
P_GP_Margin3_R	(sales - cost of products sold) / sales	Gross Profit Margin
P_TotalCost_Sales_R	total costs /total sales	Total Cost to Sales Ratio
S_Debt_Asset_R	total liabilities / total assets	Debt to Asset Ratio
S_BVE_TotalLiab_R	book value of equity / total liabilities	Book Value of Equity (BE) to Total Liabilities
S_GP_StLiab_R	gross profit / short-term liabilities	Short-term Solvency Ratio
S_GP+Dep_TotalLiab_R	(gross profit + depreciation) / total liabilities	Long-term Solvency Ratio based on Gross Profit + Depreciation
S_NP+Dep_TotalLiab_R	(net profit + depreciation) / total liabilities	Long-term Solvency Ratio based on Net Profit + Depreciation
S_Financial_Coverage_R	profit on operating activities / financial expenses	Operating Income to Financial Expense Ratio
S_TotalLiab_DailyEBITDA_R	total liabilities / ((profit on operating activities + depreciation) * (12/365))	Total Liabilities per Daily EBITDA
V_RE_Asset_R	retained earnings / total assets	Retained Earnings to Asset Ratio

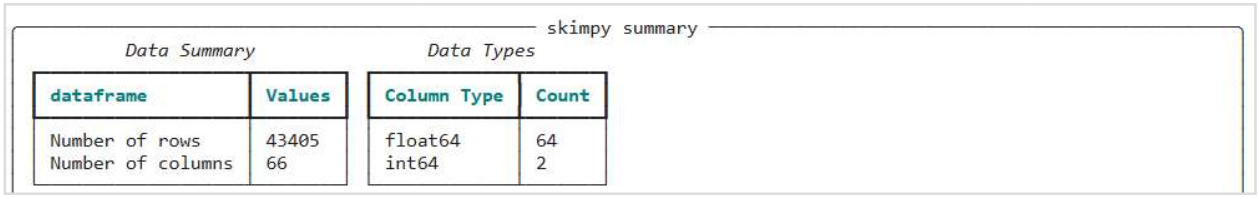
V_Equity_Asset_R	equity / total assets	Equity Ratio
V_Asset_Debt_R	total assets / total liabilities	Total Assets to Total Liabilities Ratio
V_Equity(ex-SC)_Asset_R	(equity - share capital) / total assets	Equity exluding Share Capital Ratio
V_TotalLiab(ex-Cash)_Sales_R	(total liabilities - cash) / sales	Total Liabilities (excluding-cash) to Sales Ratio
V_ConstCap_Asset_R	constant capital / total assets	Constant Capital to Total Asset Ratio
V_StLiab_Asset_R	short-term liabilities / total assets	Short-term Debt to Asset Ratio
V_Equity_FixAsset_R	equity / fixed assets	Equity Ratio based on Fixed Asset
V_LtLiab_Equity_R	long-term liabilities / equity	Long-term Debt to Equity Ratio

Table 2. Data Dictionary

Data Exploration

1.0. Data Structure

All variables are stored as numeric data types: 2 integers (binary target and year interval tagging) and the rest is 64 continuous float.



Based on summary statistics, almost all variables have maximum values that are significantly higher than the data points in interquartile range, indicating presence of outliers; subject for further examination in the succeeding exploration.

	Bankrupt	Year_Interval	Log_TAsset	E_Asset_TO_T	E_Asset_TO2_T	E_Inventory_TO_T	E_Receivable_TO_T	E_StLiab_TO_T	E_FixAsset_TO_T	E_DIO_Sales_D	E_DIO_COGS
count	43405.00	43405.00	43397.00	43396.00	43397.00	41253.00	43303.00	43271.00	42593.00	43278.00	43108
mean	0.05	3.06	4.01	2.65	2.91	448.09	17.03	9.34	72.79	243.02	357
std	0.21	1.28	0.83	62.93	62.98	32345.60	553.05	124.18	2369.34	37545.17	33146
min	0.00	1.00	-0.89	-3.50	-0.00	-12.44	-12.66	-1.54	-10677.00	-29.34	-96
25%	0.00	2.00	3.50	1.02	1.10	5.55	4.51	3.10	2.18	15.41	16
50%	0.00	3.00	4.01	1.20	1.64	9.79	6.64	5.09	4.28	35.15	38
75%	0.00	4.00	4.52	2.06	2.42	20.18	10.39	8.60	9.78	63.72	70
max	1.00	5.00	9.70	9742.30	9742.30	4818700.00	108000.00	23454.00	294770.00	7809200.00	6084200

2.0. Data Cleaning

2.0.1 Irrelevant Variables

Irrelevant financial ratios that are devoid of significant information or derived using erroneous calculations are excluded. There were 11 variables identified as irrelevant and removed from the dataset.

Variables	Formula Used	Reason for Exclusion
'E_DIO_Sales_D'	(inventory * 365) / sales	*Inventories* are valued at cost and *Short-term liabilities* mostly composed of *Accounts Payable* which are obligations to suppliers; hence, these are compared with Cost of Goods Sold instead of Sales.
'L_DPO_Sales_D'	(short-term liabilities *365) / sales	
'P_ROA(GP+Eo+FinEx)_R'	(gross profit + extraordinary items + financial expenses) / total assets	*Extraordinary Items, Financial Expenses*, *Interest Expense* and *Depreciation* are items not yet deducted from Gross Profit; thus adding them back overstates the numerator resulting to incorrect ratio.
'P_GP+Dep_Margin_R'	(gross profit + depreciation) / sales	
'P_ROA(GP&IntEx)_R'	(gross profit + interest) / total assets	
'L_DLO_GP+Dep_D'	(total liabilities * 365) / (gross profit + depreciation)	
'S_GP+Dep_TotalLiab_R'	(gross profit + depreciation) / total liabilities	
'P_GP+IntEx_Margin_R'	(gross profit + interest) / sales	
'P_ROA(EBITDA)_R'	EBITDA (profit on operating activities - depreciation) / total assets	Depreciation should be added and not deducted from operating profit to arrive at EBITDA
'P_EBITDA_Margin_R'	EBITDA (profit on operating activities - depreciation) / sales	
'V_TotalLiab(ex-Cash)_Sales_R'	(total liabilities - cash) / sales	Cash is not a component of Liabilities, hence subtracting it from the latter will not give any useful information.

Table 3. Summary of Excluded Variables

```
# Removing irrelevant columns

conso=conso.drop(columns=['E_DIO_Sales_D','L_DLO_GP+Dep_D','L_DPO_Sales_D','P_ROA(GP+Eo+FinEx)_R','P_GP+Dep_Margin_R',
                          'P_ROA(GP&IntEx)_R','P_GP+IntEx_Margin_R','P_ROA(EBITDA)_R','P_EBITDA_Margin_R',
                          'S_GP+Dep_TotalLiab_R','V_TotalLiab(ex-Cash)_Sales_R'])
```

After removing irrelevant variables, the updated dataset consists of 55 variables

2.0.2 Missing Values

There are 53 variables with missing values, 11 of them have 1% or more missing. To understand the nature of missing data and determine the appropriate approach to handle it, a summary of zero frequencies was also extracted to see patterns given that financial ratios were derived from interrelated accounting items.

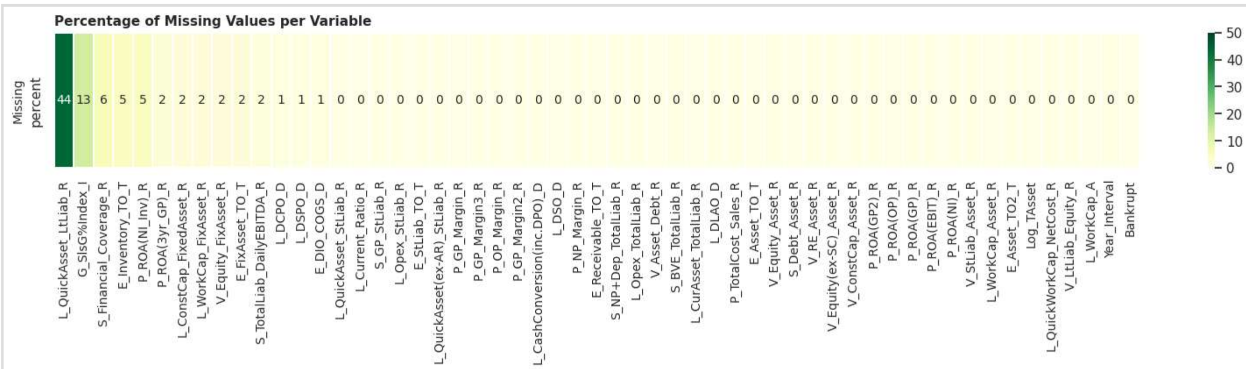


Figure4.Summary of Missing Values

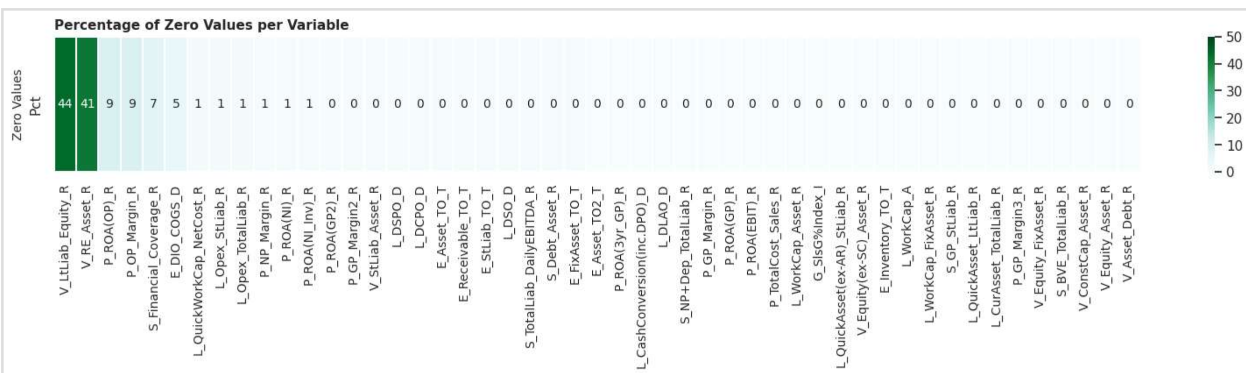


Figure5. Summary of Zero Values

Based on the comparisons made on the variables with highest number of missing and zero values, it was found out that *'L_QuickAsset_LtLiab_R' and 'V_LtLiab_Equity_R'*, are both related to Long-term Liability. This suggests that the reason for missing values is not random. It is possible that companies had no outstanding long-term liabilities in the reporting year resulting to zero *long-term liability to equity* ratio; and reason for null *quick asset/long-term liability* ratio. Subsequent analysis on other variables revealed the same pattern wherein variables derived from common denominator have nearly the same number of missing values.

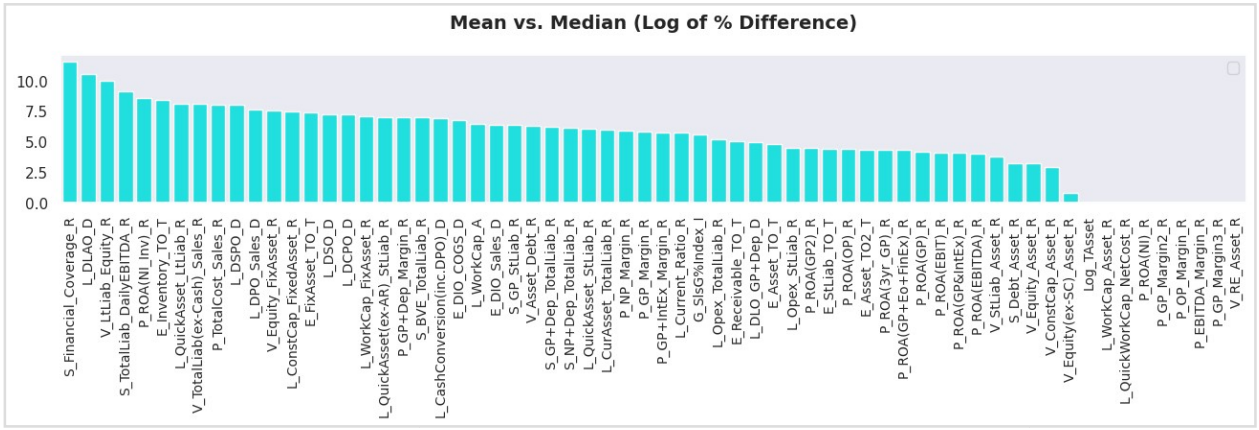
Handling Missing Values

Since the reason for missing values is not random and their proportion is significantly high, affected observations and variables cannot be censored to preserve any valuable information they may contain. Hence, imputation is the better approach.

Prior to imputation, additional 8 variables in the form of binary flags (0 and 1) were added. This is to account for the information that these zero denominator items might entail. This includes non-outstanding or absence of the following: *Long-term liability , Comparative previous year sale, Financial expenses incurred, Inventory, Incomplete 3year GP, Fixed asset, EBITDA, and Cost of Sales.*

```
#Created variables with binary flags (1(Yes), 0(No))-- Related to variables with atleast 1% missing

conso_cleaned['NoLtLiab'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['L_QuickAsset_LtLiab_R']) else 0, axis=1)
conso_cleaned['NoPrevYrSales'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['G_SlsG%Index_I']) else 0, axis=1)
conso_cleaned['NoFinEx'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['S_Financial_Coverage_R']) else 0, axis=1)
conso_cleaned['NoInventory'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['E_Inventory_TO_T']) and pd.isnull(row['P_ROA(NI_Inv)_R']) else 0, axis=1)
conso_cleaned['No3yrAsset'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['P_ROA(3yr_GP)_R']) else 0, axis=1)
conso_cleaned['NoFixedAsset'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['E_FixAsset_TO_T']) and pd.isnull(row['L_WorkCap_FixAsset_R'])
                                                    and pd.isnull(row['L_ConstCap_FixedAsset_R']) and pd.isnull(row['V_Equity_FixAsset_R']) else 0, axis=1)

conso_cleaned['NoEBITDA'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['S_TotalLiab_DailyEBITDA_R']) else 0, axis=1)
conso_cleaned['NoCOGS'] = conso_cleaned.apply(lambda row: 1 if pd.isnull(row['L_DCPO_D']) and pd.isnull(row['L_DSPO_D'])
                                                    and pd.isnull(row['E_DIO_COGS_D']) else 0, axis=1)
```

To gauge how extreme values affect the mean of each variable, a comparison between mean and median values was conducted. This is to appropriately select the imputation technique to be used.



The mean cannot be used due to the influence of extreme values; hence, missing values were imputed using median.

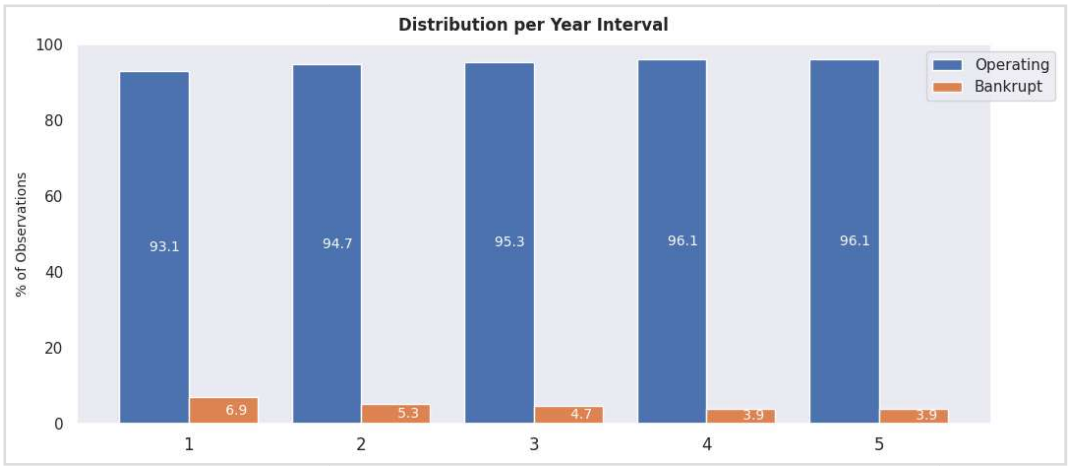
```
#Impute Using Median
conso_cleaned=conso_cleaned.fillna(conso_cleaned.median(numeric_only=True))
```

3.0. Data Distribution

3.0.1 Target Variables

The data is highly imbalanced with ~95%-5% proportion of operating vs. bankrupt firms. This will be subject to re-sampling to ensure unbiased classification in favor of the majority class (*operating*).

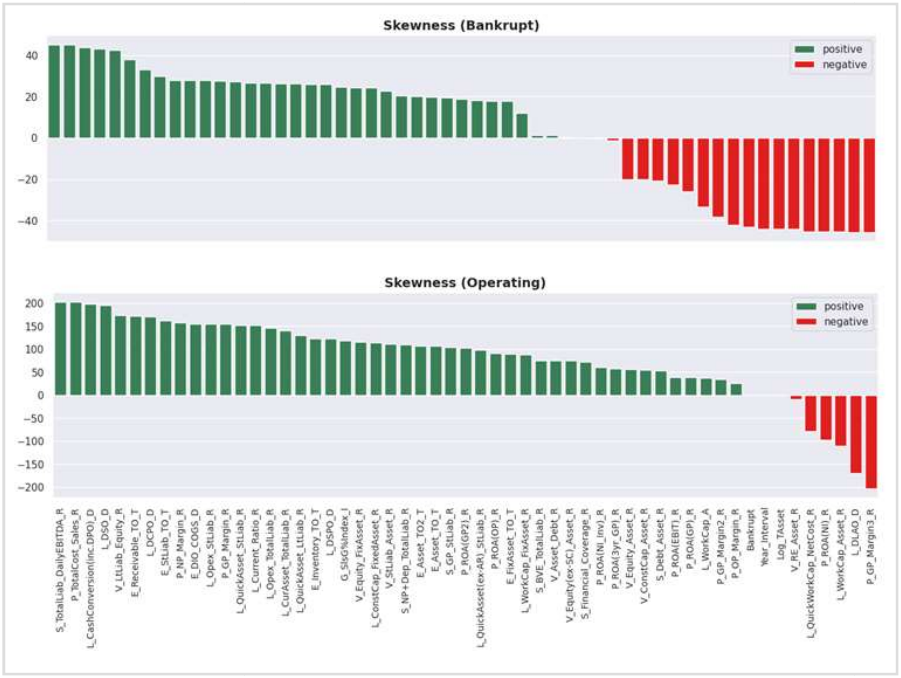
The data proportion is approximately the same across all years. A higher proportion of bankrupt firms is found in the financial reporting period closer to the status year.



3.0.2 Input Variables

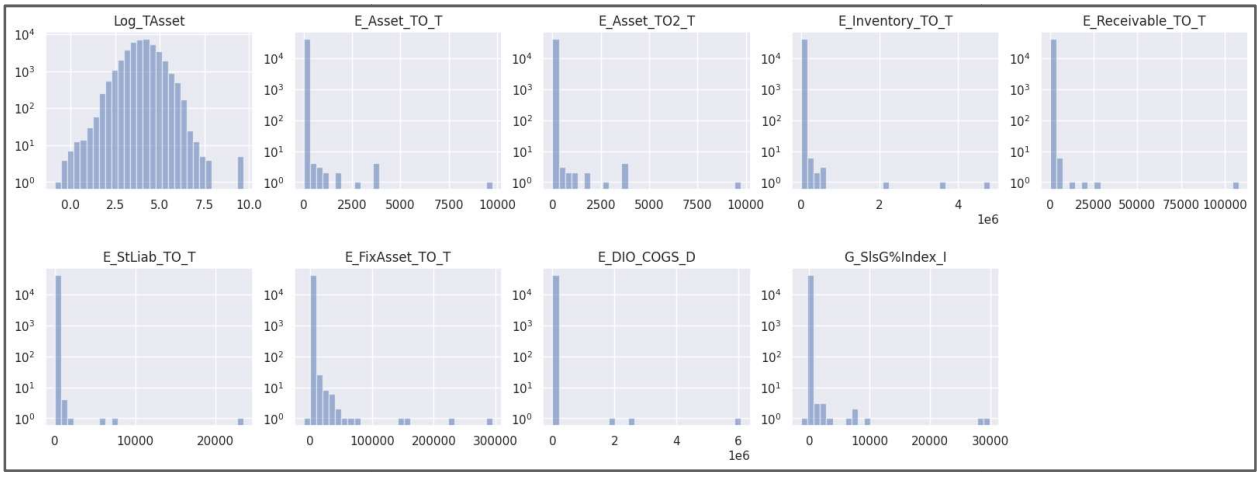
To get a general idea on the overall data distribution, a summary of skewness was generated. The data distribution is highly skewed overall, ranging from -200 to 200 which significantly deviate from -1 to 1 cut-off. 45 positively skewed while 7 are negatively skewed. When comparing the target classes, the

majority of operating firms' financial ratios is positively skewed, whereas bankrupt firms are negatively skewed.

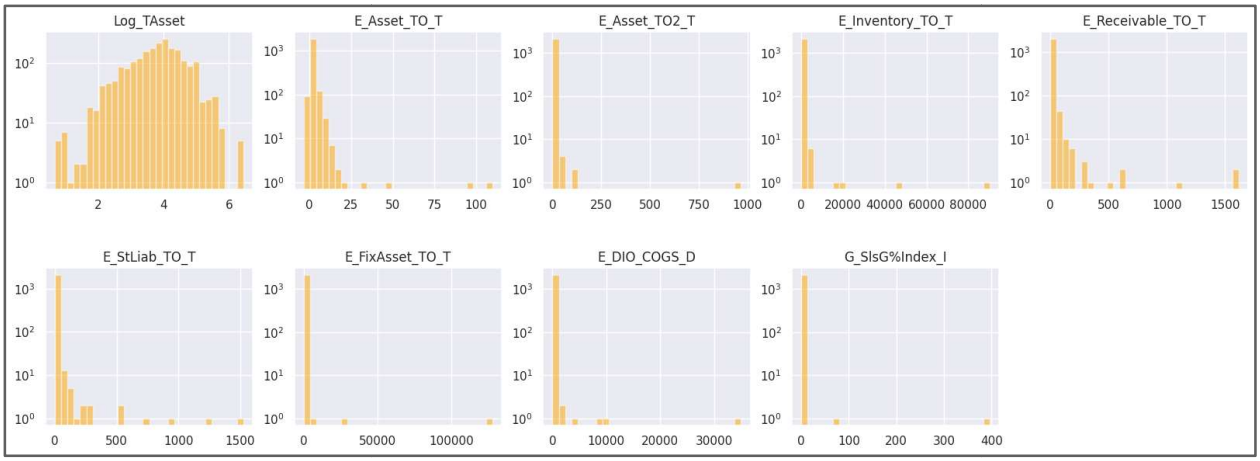


Through histograms, distribution of each input variable per target class was compared. To allow a more effective analysis flow, visualizations were grouped per financial metric class.

a. Efficiency & Growth



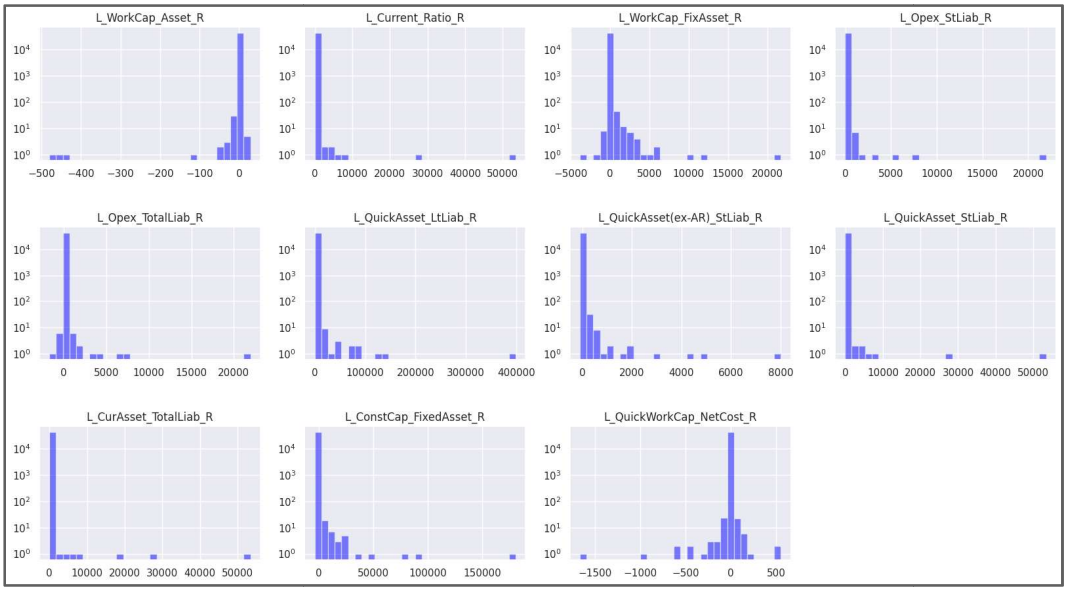
Operating Firms



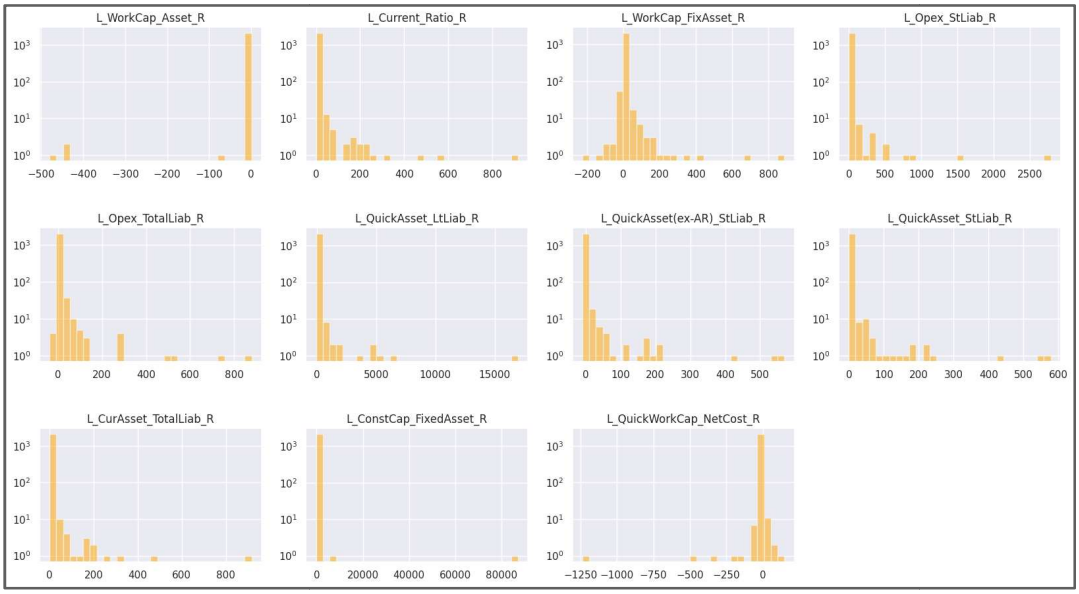
Bankrupt Firms

- The Logarithm of Asset (Log_TAsset) is the only normally distributed variable. The majority of bankrupt firms have log of asset between 2 and 6.
- Distribution shape of efficiency ratios for both classes is fairly the same wherein all are positively skewed, unimodal, and commonly peaks nearly 0. There are also isolated bars that are extremely far from the center and appeared to be outliers.
- However, operating firms have higher range of efficiency ratios compared to bankrupt firms
- The two asset turnover ratios (E_Asset_TO_T and E_Asset_TO2_T) have identical distribution.

b. Liquidity (in Ratios)



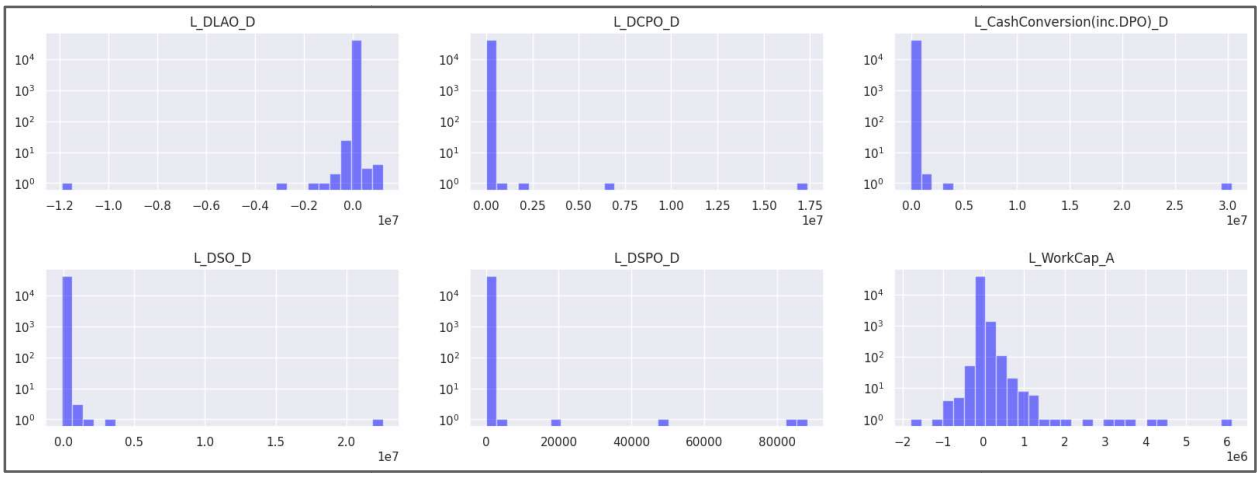
Operating Firms



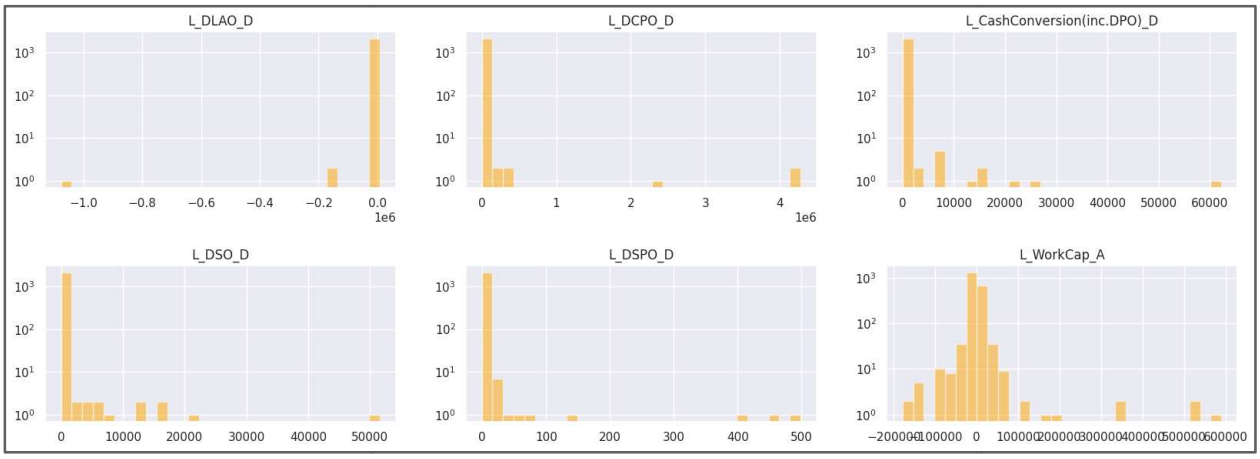
Bankrupt Firms

- Distribution shape of Liquidity ratios for both classes is also fairly the same wherein Working Capital to Asset and Quick Working Capital to Net Cost (L_WorkCap _Asset, L_QuickWorkCap _Net Cost) appears to be negatively skewed indicating high current liabilities exceeding current assets.
- The rest are positively skewed, unimodal, and commonly peaks nearly 0. There are also isolated bars that are extremely far from the center and appeared to be outliers.
- Operating firms have higher range of liquidity ratios compared to bankrupt firms

c. Liquidity (in Days)



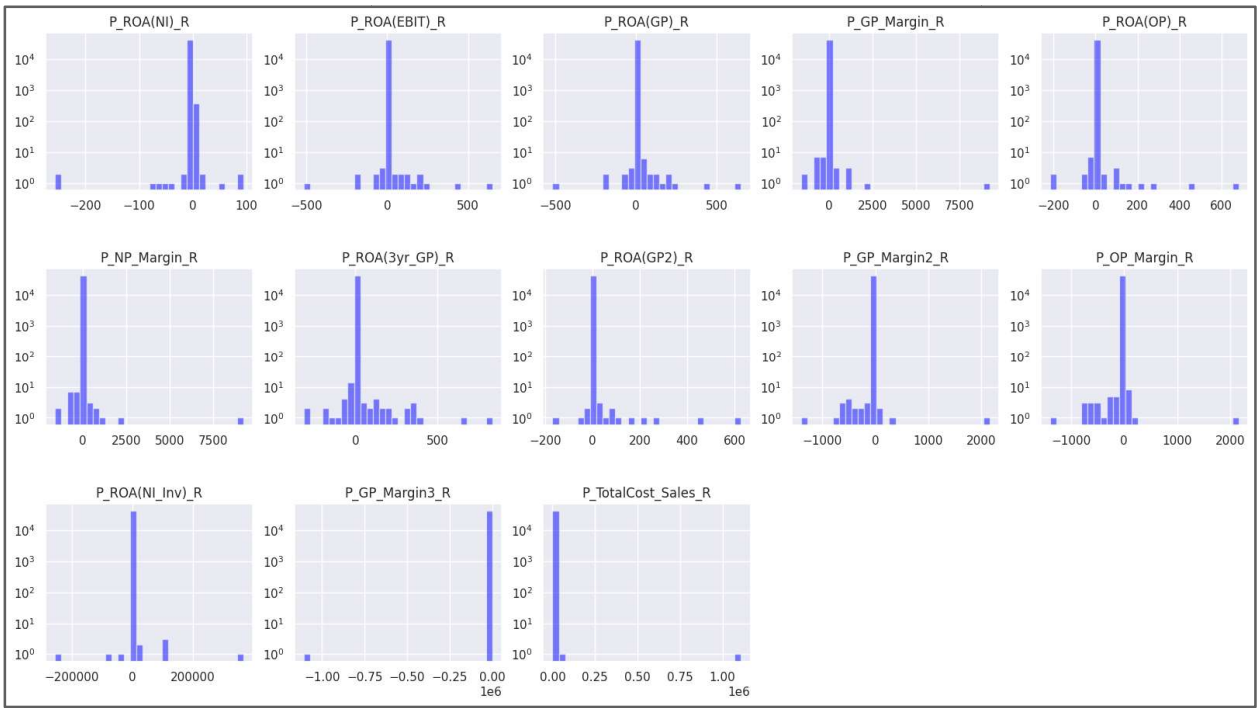
Operating Firms



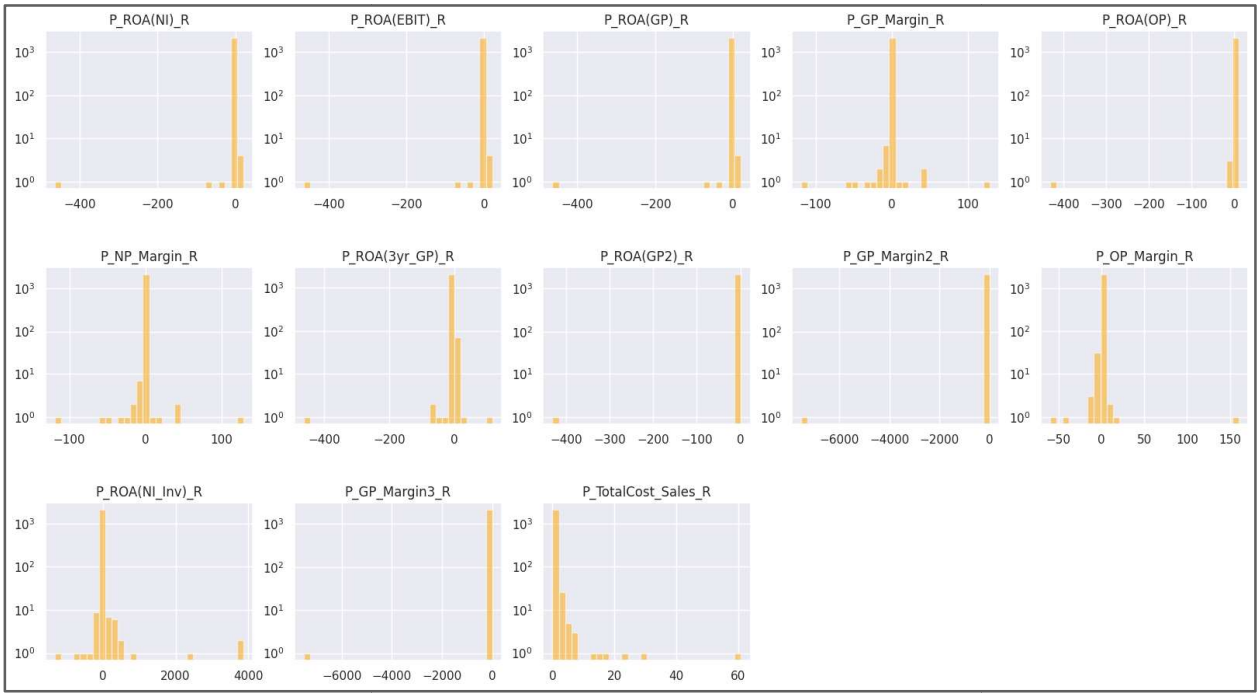
Bankrupt Firms

- Distribution shape for liquidity in days for both classes is fairly the same wherein Days in Liquid Asset Outstanding (L_DLAO_D) appears to be negatively skewed. Working Capital (L_WorkCap_A) is slightly normally distributed. The rest are positively skewed, and unimodal.

d. Profitability



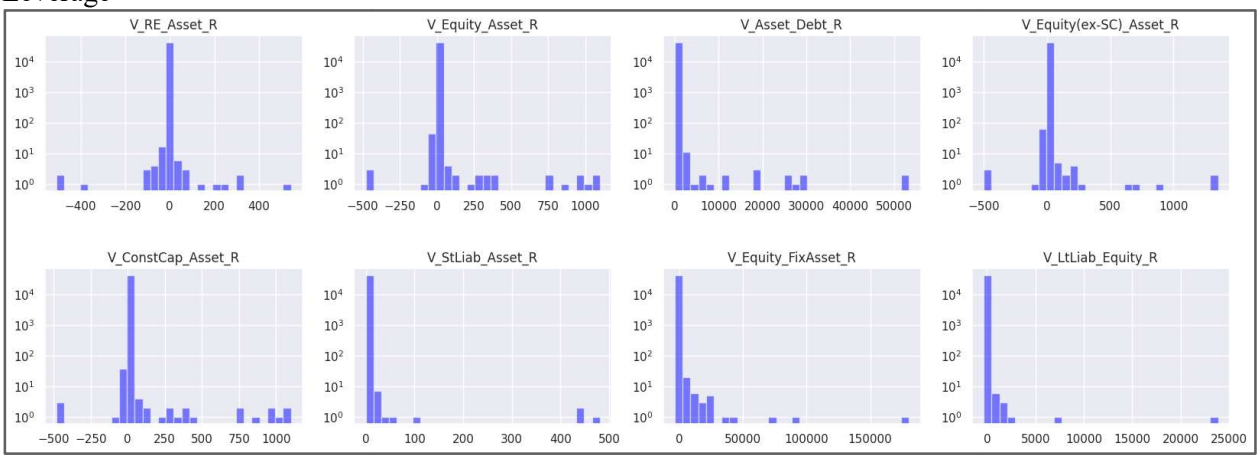
Operating Firms



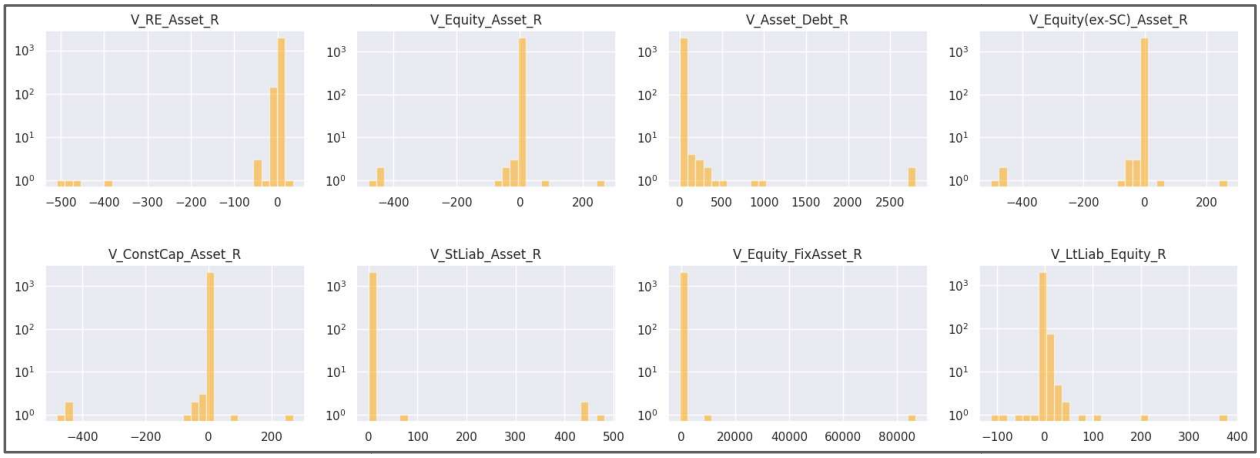
Bankrupt Firms

- The distribution shape of profitability ratios for operating firms are slightly normally distributed with data points and slight skew in both positive and negatives sides.
- However, profitability ratios for bankrupt firms are almost negatively skewed especially those related to return on asset based on operating income (EBIT) and Gross Profit (GP)
- Percentage of total cost to sales (P_TotalCost_Sales_R) is higher in bankrupt firms with (values >1), compared to operating firms (values <=0).

e. Leverage



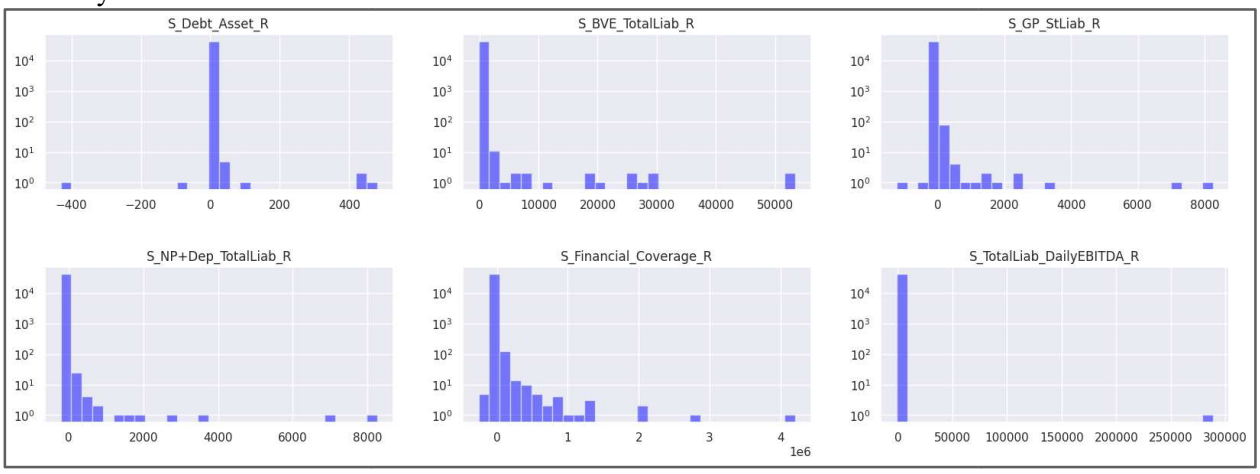
Operating Firms



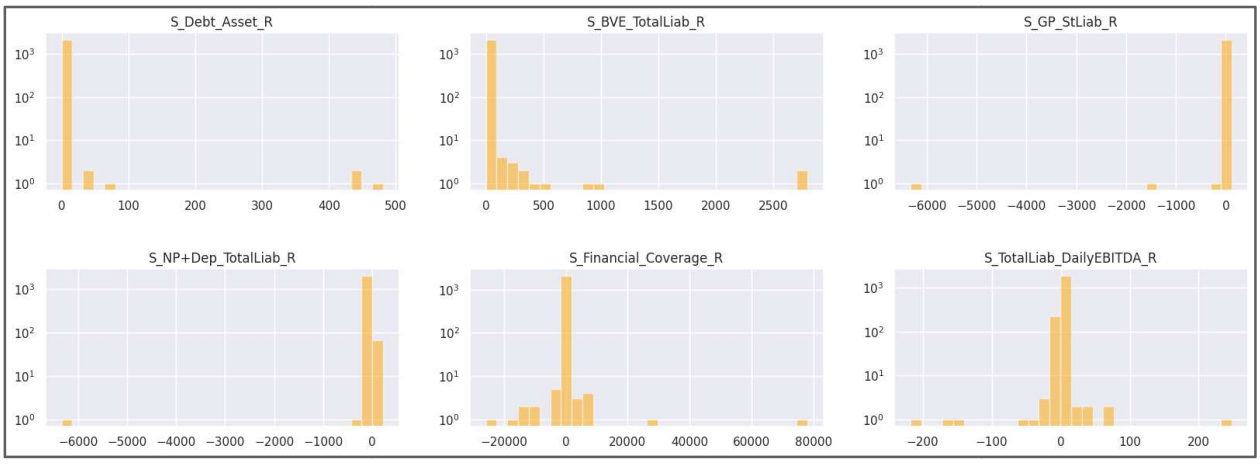
Bankrupt Firms

- The distribution shape of Retained Earnings to Asset (V_RE_Asset_R) for operating firms are slightly normally distributed with data points and slight skew in both positive and negatives sides. However, for bankrupt firms it is negatively skewed
- The Equity to Asset ratio (V_Equity_Asset_R) is positively skewed for operating firms and negatively skewed for bankrupt firms
- The ranges of Asset to Debt ratio (V_Asset_Debt_R) for operating firms are significantly higher than bankrupt firms
- Ratios to asset, such as Equity (V_Equity (ex-SC) _Asset_R) and Constant Capital (V_ConstCap_Asset_R) are positively skewed for operating firms and negatively skewed for bankrupt firms.
- For both classes, the distribution shape of Short-term Liability to Asset (V_StLiab_Asset_R) and Equity to Fixed Asset (V_Equity_FixAsset_R) are positively skewed and fairly the same.
- The Long-term Liability to Equity ratio (V_LtLiab_Equity_R) is positively skewed for operating firms and negatively skewed for bankrupt firm.

f. Solvency



Operating Firms



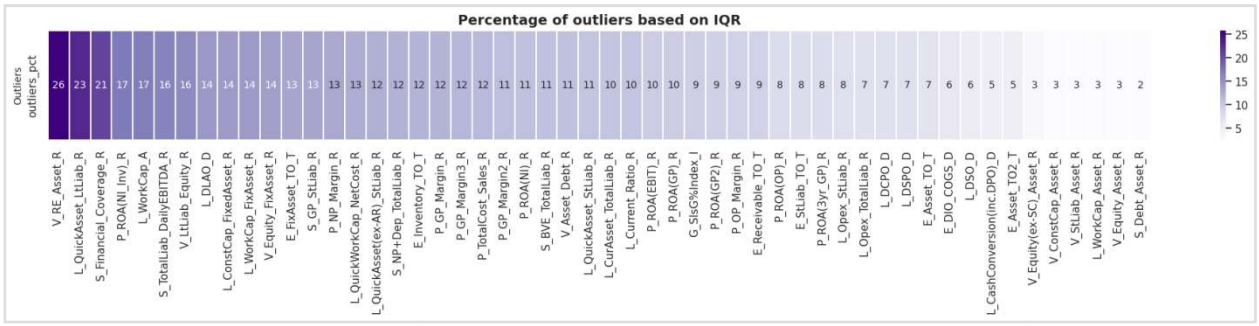
Bankrupt Firms

- The range of Debt to Asset ratio (S_Debt_Asset_R) for bankrupt firms is higher than operating firms
- The range of Book Value of Equity to Liability ratio (S_BVE_TotalLiab_R) for bankrupt firms is lower than operating firms
- The Gross Profit to Short-term Liability ratio (S_GP_StLiab_R) is heavy on the positive side for operating firms while negative for bankrupt firms.
- The EBITDA to Total Liability ratio (S_NP+Dep_TotalLiab_R) is positively skewed for operating firms and negatively skewed for bankrupt firms.
- The Financial Coverage ratio (S_Financial_Coverage_R) is heavy on the positive side for operating firms while negative for bankrupt firms.

- The Total Liability to Daily EBITDA ratio (S_TotalLiab_DailyEBITDA_R) is positively skewed for operating firms and negatively skewed for bankrupt firms.

3.0.3 Outliers

The highly skewed distribution indicates presence of extreme values or outliers. To examine the distribution of outliers per variable, the IQR method was used.



A total of 52 numerical variables have outliers, with *"Retained Earnings to Asset"*, *"Quick Asset to Long-term Liability"*, and *"Financial Coverage Ratio"* having the most with at least 20% of their values is outliers, a proportion deemed material. These extreme values may represent significant information; for instance, ***Retained Earnings to Asset*** is typically not below 0. Extreme low ratios may indicate that the firm has sustained loss over extended period of time.

Handling Outliers

To retain the valuable insights that these outliers may contain, binary flags were created to tag respective records as outliers. Capping and floor then performed using the IQR method. Z-score is deemed not appropriate because of the highly skewed distribution, which impacts mean value.

```
# Flag the variables with above 20% outliers
out=list(outliers.sort_values(by='outliers_pct', ascending=False).head(3).index)

for i in out:
  conso_cleaned[i+'_out']= conso_cleaned[i].apply(lambda x: 1 if x < outliers.loc[i,'lower_bound']
                                        or x > outliers.loc[i,'upper_bound'] else 0)
```

```
#Cap and Floor Outliers

feat=conso_cleaned.iloc[:,3:55]
for i in feat:
  conso_cleaned[i]= conso_cleaned[i].apply(lambda x: outliers.loc[i,'lower_bound'] if x < outliers.loc[i,'lower_bound']
                                        else outliers.loc[i,'upper_bound'] if x > outliers.loc[i,'upper_bound'] else x)
```

Skewness was re-evaluated after capping. Since degree of skewness was reduced to nearly -1 to 1 after capping, log transformation is deemed not necessary.

4.0. Exploratory Data Analysis

4.0.1 Correlation

Pearson's correlation coefficient was used to examine linear relationship among continuous independent variables. As evident in correlation matrix, a high degree of multicollinearity exists among variables; particularly among *liquidity, profitability, and solvency* ratios.

To gauge the severity of multicollinearity and how it can impact the coeffient estimates of variables, Variance Inflation Factor (VFI) analysis was employed. The commonly followed VIF thresholds are as follows:

- 1: No correlation with other independent variables
- >1 <= 5: Moderate multicollinearity
- >5: Severe multicollinearity that impacts reliability of coefficient estimates and p-values

Based on the results of analysis, majority of variables have VFI score above 10, indicating severe degree of multicollinearity. The fact that these are financial ratios derived from common components of financial statement, explains the strong multicollinearity.

Feature	VIF
P_ROA(EBIT)_R	2487.5
P_ROA(GP)_R	2271.4
V_Asset_Debt_R	587.5
P_TotalCost_Sales_R	490.4
S_BVE_TotalLiab_R	364.5
V_Equity_Asset_R	276.3
L_DSPO_D	254.0
L_DCPO_D	249.7
V_ConstCap_Asset_R	241.9
S_Debt_Asset_R	234.8
L_ConstCap_FixedAsset_R	220.6
P_GP_Margin_R	145.5
P_ROA(NI)_R	138.5
P_NP_Margin_R	133.2
V_StLiab_Asset_R	125.3
L_WorkCap_FixAsset_R	79.6
L_Current_Ratio_R	78.4
V_Equity_FixAsset_R	75.7
E_StLiab_TO_T	64.3
L_CurAsset_TotalLiab_R	60.5
L_Opex_StLiab_R	57.0
L_CashConversion(inc.DPO)_D	55.8
L_QuickAsset_StLiab_R	53.7
P_ROA(OP)_R	50.4
L_DSO_D	40.7
E_Asset_TO2_T	38.6

Feature	VIF
P_OP_Margin_R	35.8
P_ROA(GP2)_R	35.2
P_GP_Margin2_R	29.1
G_SlsG%Index_I	27.2
L_WorkCap_Asset_R	22.7
E_Asset_TO_T	21.3
E_Receivable_TO_T	15.9
E_DIO_COGS_D	15.2
S_GP_StLiab_R	13.2
P_GP_Margin3_R	13.1
S_NP+Dep_TotalLiab_R	11.1
E_FixAsset_TO_T	11.1
L_Opex_TotalLiab_R	9.7
L_QuickAsset(ex-AR)_StLiab_R	9.6
L_QuickAsset_LtLiab_R	7.9
V_Equity(ex-SC)_Asset_R	7.5
E_Inventory_TO_T	6.9
P_ROA(NI_Inv)_R	5.4
P_ROA(3yr_GP)_R	4.7
S_Financial_Coverage_R	3.2
L_QuickWorkCap_NetCost_R	2.9
L_DLAO_D	2.6
L_WorkCap_A	2.4
V_LtLiab_Equity_R	2.4
V_RE_Asset_R	2.4
S_TotalLiab_DailyEBITDA_R	1.8

Table 4. VIF Scores before removal of correlated variables

Handling Multicollinearity

High degree of multicollinearity needs to be addressed when regression models are employed. Moreover, since interpretation of leading financial indicators is one of the primary objectives of this study, generating reliable coefficient estimates and p-values is critical.

This problem was addressed by employing a combination of correlation and VIF analysis. Variables with correlation of <=-0.70 or >=0.70 were evaluated. Those with highest correlation with the target are retained and the rest needs to be dropped. A total of 29 variables were considered for removal.

Strongest Corr to Target	Correlation	Removed Variables
L_WorkCap_Asset_R	0.81	L_Current_Ratio_R
	0.81	L_WorkCap_FixAsset_R
	0.76	L_QuickAsset_StLiab_R
	0.79	L_CurAsset_TotalLiab_R
	0.80	L_ConstCap_FixedAsset_R
	0.78	V_Equity_FixAsset_R
V_ConstCap_Asset_R	0.74	S_BVE_TotalLiab_R
	(0.83)	S_Debt_Asset_R
	0.85	V_Equity_Asset_R
	0.73	V_Asset_Debt_R
	(0.96)	V_StLiab_Asset_R
P_GP_Margin2_R	0.77	P_NP_Margin_R
	0.86	P_ROA(GP2)_R
	0.89	P_OP_Margin_R
P_GP_Margin_R	0.87	P_ROA(NI)_R
	0.87	P_ROA(GP)_R
	0.77	P_ROA(OP)_R
L_DCPO_D	(0.80)	L_Opex_StLiab_R
	0.99	L_DSPO_D
E_Receivable_TO_T	(0.85)	L_DSO_D
P_ROA(NI)_R	0.99	P_ROA(EBIT)_R
	0.76	P_ROA(3yr_GP)_R
	0.76	S_NP+Dep_TotalLiab_R
E_Asset_TO2_T	0.79	E_Asset_TO_T
L_Current_Ratio_R	0.71	E_StLiab_TO_T
P_NP_Margin_R	0.99	P_GP_Margin_R
S_GP_StLiab_R	0.71	P_ROA(NI_Inv)_R
P_GP_Margin3_R	(0.92)	P_TotalCost_Sales_R
S_NP+Dep_TotalLiab_R	0.88	S_GP_StLiab_R

Table 5. Summary of removed variables due to high correlation

```
#Remove Highly Correlated Variables

conso_cleaned=conso_cleaned.drop(columns=['E_Asset_TO_T','E_StLiab_TO_T','L_Current_Ratio_R','L_WorkCap_FixAsset_R','L_Opex_StLiab_R',
                                          'L_QuickAsset_StLiab_R','L_CurAsset_TotalLiab_R','L_ConstCap_FixedAsset_R','L_DSO_D','L_DSPO_D',
                                          'P_ROA(NI)_R','P_ROA(EBIT)_R','P_ROA(GP)_R','P_GP_Margin_R','P_ROA(OP)_R','P_NP_Margin_R',
                                          'P_ROA(3yr_GP)_R','P_ROA(GP2)_R','P_OP_Margin_R','P_ROA(NI_Inv)_R','P_TotalCost_Sales_R',
                                          'S_BVE_TotalLiab_R','S_Debt_Asset_R','S_GP_StLiab_R','S_NP+Dep_TotalLiab_R','V_Equity_Asset_R',
                                          'V_Asset_Debt_R','V_StLiab_Asset_R','V_Equity_FixAsset_R'])
```

After removing highly correlated variables, VFI scores have improved with majority of variables reduced to below 10.

Feature	VIF
V_ConstCap_Asset_R	22.7
G_SlsG%Index_I	21.2
L_CashConversion(inc.DPO)_D	19.3
E_Asset_TO2_T	15.7
L_DCPO_D	11.7
L_WorkCap_Asset_R	9.3
E_DIO_COGS_D	9.0
E_Receivable_TO_T	8.6
E_FixAsset_TO_T	6.9
L_QuickAsset_LtLiab_R	6.3
V_Equity(ex-SC)_Asset_R	5.6
E_Inventory_TO_T	5.2

Feature	VIF
L_Opex_TotalLiab_R	4.5
L_QuickAsset(ex-AR)_StLiab_R	4.3
P_GP_Margin2_R	4.0
P_GP_Margin3_R	3.4
S_Financial_Coverage_R	2.6
L_DLAO_D	2.4
L_WorkCap_A	2.4
V_LtLiab_Equity_R	2.0
V_RE_Asset_R	2.0
L_QuickWorkCap_NetCost_R	2.0
S_TotalLiab_DailyEBITDA_R	1.7

Table 6. Summary of VIF Scores after removing correlated variables

5.0. Summary

The key findings on the data exploration are as follows:

- Distribution Patterns – for efficiency and liquidity ratios, the shape of data distribution between target classes are almost the same. However, there is notable difference with respect to profitability, leverage and solvency ratios. The majority of operating firms have positively skewed ratios, whereas bankrupt firms have negatively skewed ratios.

- Correlation – there is high degree of multicollinearity among variables which was expected since some financial ratios are inherently related to each other. Despite removing highly correlated variables using 0.7 thresholds, some variables still have VIF scores above 10.

Considering the above findings, decision tree models will be explored and utilized first to select the most important features. Tree-based algorithms are generally robust to multicollinearity and skewed data. Afterwards, logistic regression will be employed using the selected features.

Data Preparation and Feature Engineering

1.0. Feature Transformation

The “logarithm of total assets” was binned into discrete values to reduce dimensionality.

```
#BIN Logarithm of Asset
labels=[1,2,3,4,5,6,7,8,9,10]
range=(pd.qcut(conso_cleaned['Log_TAsset'], q=10, labels=labels)).astype('int')
labels=list(np.unique(range))

conso_cleaned['Log_TAsset'] = (pd.qcut(conso_cleaned['Log_TAsset'], q=10, labels=labels)).astype('int')
```

2.0. Data Partition

The cleaned dataset was partitioned into training and validation set using 60-40 split. Stratified train_test_split was employed to mitigate the impact of highly imbalanced data.

```
# Partition Dataset based on 60-40 split

X = conso_cleaned.iloc[:,1:]
y = conso_cleaned['Bankrupt']

train_X, valid_X, train_y, valid_y = train_test_split(X,y, test_size=0.4,stratify=y, random_state=1)
```

3.0. Upsampling through SMOTE

After data partitioning, the train set was subject to re-sampling. To ensure accurate predictions, the model needed to be trained on a balanced dataset to avoid biased results in favor of the dominant class. Since our dataset is primarily comprised of operating firms, outcomes might be skewed and bankrupt class might be underrepresented. To balance the data, oversampling was chosen over undersampling to avoid losing significant amount of information. Synthetic Minority Over-sampling Technique (SMOTE) was employed in the training set. It works by generating synthetic data by randomly selecting k nearest neighbor of minority class.

```
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state = 2)

train_Xsm, train_ysm = sm.fit_resample(train_X, train_y.ravel())
```

Validation set, on the other hand remained to be in the original data distribution, for it must reflect the real data imbalance. This is to ensure that the model will perform fairly in the unseen data regardless of distribution.

Modeling

1.0. Modeling Approach/Introduction

Various classification models were built and evaluated to predict whether the sampled manufacturing firms will go bankrupt based on their financial characteristics. Three main modeling techniques were employed namely: *Decision tree classifier, Logistic Regression*, and *Neural Network MLP classifier.* There were series of hyper parameter tuning using *exhaustive grid search* in order to improve model performance. Advanced algorithms for the aforementioned techniques were also explored such as:

- Ensembling models for Decision Tree like *Random Forest and XGboost.*
- Sequential selection for logistic regression : *Forward Stepwise, and Recursive Feature Elimination(RFE)*

2.0. Model Selection Criteria

In line with the objectives of this study, which is to identify the leading financial indicators and how they influence the exposure of the firm to bankruptcy risk, the criteria for selecting the best model is the one with high predictive power and high interpretability.

The evaluation metrics used are the following:

- **Precision :** model is confident that the identified bankrupt firms are bankrupt in actual
- **Recall :** actual bankrupt firms were correctly identified as bankrupt
- **F1-score:** right balance of precision and recall. It minimizes the chance of incorrectly identifying non-bankrupt firms as bankrupt and missing on actual bankrupt firms
- **Accuracy** : overall accuracy of the model in identifying both bankrupt and non-bankrupt firms
- **ROC_AUC:** model's ability to effectively distinguish bankrupt from non-bankrupt firms.

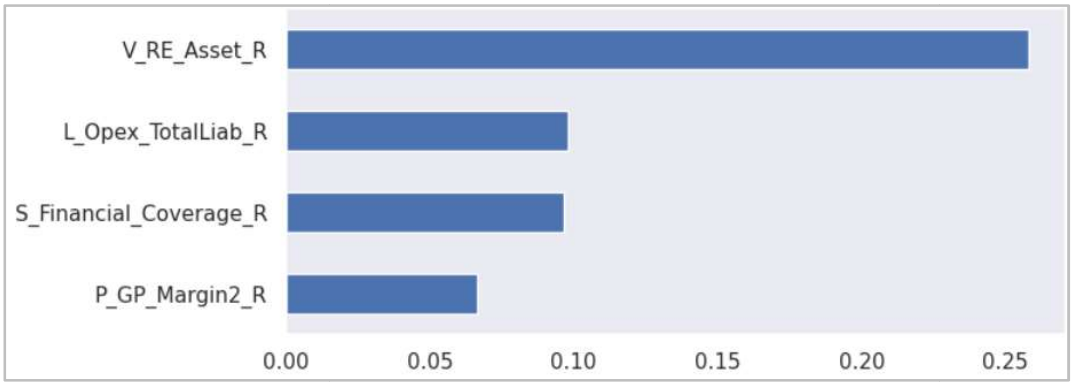
3.0. Model Exploration

3.0.1 Modeling Technique #1: Basic Full Decision Tree

The first modeling approach is full decision tree using CART or binary trees. Based on the result of plotted full tree, there are numerous splits that indicate high complexity.

Figure 1. Plotted Full Decision Tree

The most important features at .05 thresholds were extracted to identify the most important variables of the full tree. These metrics are*: Retained Earnings to Asset, Opex to Total Liabilities, Financial Coverage, and Gross Profit Margin.*



The performance of the full tree in scoring validation observations was examined, and classification summary revealed that it is a fully-fit tree, for the training accuracy is at 100% while validation performance is only at 90%, an indication of overfitted model. This needs to be addressed since it might be unstable and not generalized well when predicting unseen data. Hence, further configuration and tuning was performed to reduce the noise and improve model performance.

Full Class Tree

```
Confusion Matrix (Accuracy 1.0000)

       Prediction
Actual     0     1
     0 24788     0
     1     0 24788
Train None

Confusion Matrix (Accuracy 0.8979)

       Prediction
Actual     0     1
     0 15206  1320
     1   453   383
Valid: None
```

Classification Report: Full Class Tree

	precision	recall	f1-score	support
Operating	0.97	0.92	0.94	16526
Bankrupt	0.22	0.46	0.30	836
accuracy			0.90	17362
macro avg	0.60	0.69	0.62	17362
weighted avg	0.94	0.90	0.91	17362

ROC_AUC:
0.6891

3.0.2 Modeling Technique #2 : Basic Tuned Tree (Grid Search)

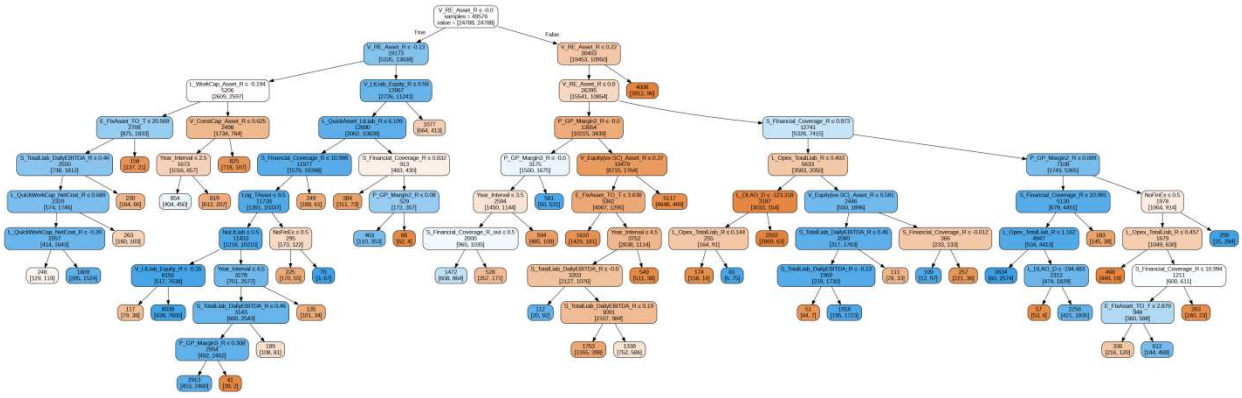
Provided that the full tree is an over fitted model, an exhaustive grid search with cross validation was performed to identify the best set of parameters that leads to highest accuracy both training and validation. Parameter grid was set comprising different thresholds for number of splits, minimum number of samples per split, and minimum reduction in impurity.

The best set of parameter selected by grid search is: 12 numbers of splits, at least 0.1% records per split, and at least 0.5% decrease in impurity to warrant further splits.

```
#Print Best Score and Parameters

print("Best Score:", gridsearch.best_score_)
print("Best Parameters:",gridsearch.best_params_)
```

```
Best Score: 0.8493431013147859
Best Parameters: {'max_depth': 12, 'min_impurity_decrease': 0.001, 'min_samples_split': 0.005}
```

The tuned decision tree was less complex compared to the full decision tree, but still with significant number of repeating splits that makes it challenging to interpret.



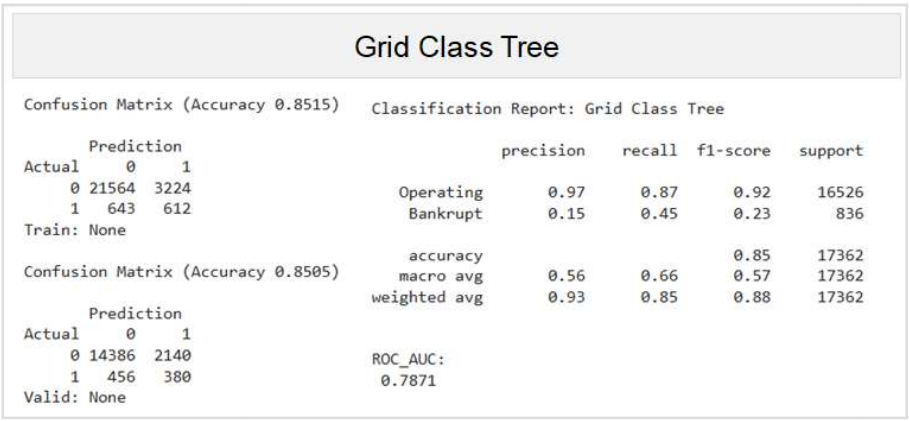
The grid class tree revealed that the feature that best distinguishes bankrupt from operating firms is *Retained Earnings to Asset at <= -0.0 threshold*, which accounts for the best split.

Interpretation of the Tuned Tree's rules based on top split points:

- If the *Retained Earnings to Asset* is less than -0.13, *Working Capital to Asset* is below -0.19, Fixed Asset Turnover of less than 20.57, *Total Liability per daily EBITDA* is less than 0.46, and *Quick*

Working Capital to Net Cost is between -0.39 and 0.69, then there is 84% chance of bankruptcy since out of 1,809 firms, 1524 were classified by the tree as bankrupt.

Grid class tree was less overfit and with higher ROC_AUC by .10 (0.79 vs. 0.69) compared to full class tree; however, with lower f1-score and overall accuracy by 0.07 and 0.5 respectively. It implies that reducing the complexity of the full tree through hyper parameter tuning did not improve model performance, for significant patterns might have been lost during the tuning process.



3.0.3 Modeling Technique #3: Random Forest

Given the limitations of basic decision tree, it calls for exploring more advanced tree-based algorithms. Random Forest was thus explored. In contrast to basic decision tree that trains on entire set of features and observations, random forest trains on multiple subsets of trees with different random combination of features and observations. Features are then ranked with respect to their importance of reducing error or noise in the model.

Random Forest was fitted into training data and the results revealed that indeed it has exceptional model performance. Overall accuracy is at 95%. With respect to its accuracy of predicting bankrupt instances, its performance is significantly better than the previously built and tuned basic tree. F1-score is at 41% while ROC_AUC is at 0.867

Random Forest

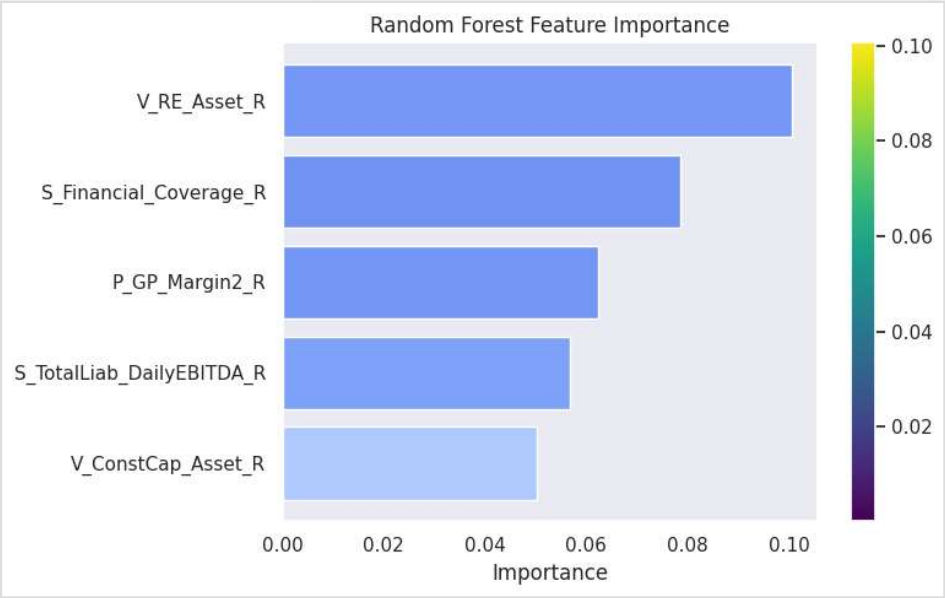
```
Confusion Matrix (Accuracy 1.0000)

       Prediction
Actual     0     1
     0 24788     0
     1     0 24788
Train: None

Confusion Matrix (Accuracy 0.9477)

       Prediction
Actual     0     1
     0 16137   389
     1   519   317
Valid: None
```

Classification Report: Random Forest

	precision	recall	f1-score	support
Operating	0.97	0.98	0.97	16526
Bankrupt	0.45	0.38	0.41	836
accuracy			0.95	17362
macro avg	0.71	0.68	0.69	17362
weighted avg	0.94	0.95	0.95	17362

ROC_AUC:
0.867

Features with at least .05 importance as ranked by random forest were then extracted. The most important features ordered from highest to lowest are: *"Retained Earnings to Asset", "Financial Coverage Ratio", "Gross Profit Margin", "Total Liability to Daily EBITDA",* and *" Constant Capital to Asset"*. Based on the features revealed by the random forest, leading indicators of bankruptcy are related to the firm's **profitability** (earnings' capacity), **leverage** (level of debt financing), and **solvency** (ability to fulfill debt obligations)



Although random forest is a powerful model in identifying the leading indicators, it has limitations with respect to interpretability. Understanding how financial indicators influence the risk of bankruptcy is critical to gaining valuable insights on how to manage them. Hence, a complementing modeling technique was explored to overcome these limitations, and this is logistic regression.

3.0.4 Modeling Technique #4: Logistic Regression

The logistic regression was trained using the features selected by the random forest. To tune the model's performance, exhaustive grid search cross validation was also used. Additionally, for further refinement, sequential selection techniques such as **Stepwise** and **Recursive Feature Elimination (RFE)** were employed to identify those best set of leading indicators that strongly influence the likelihood of bankruptcy.

Below is an overview of how the aforementioned logistic regression models work:

- Grid Class Logistic Regression :
  Parameter grid was set comprising different optimization algorithm, type of regularization penalties, strength of inverse regularization, and maximum number of iterations.

```
#Tune the logistic regression using grid search CV
parameters = [{'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']},
              {'penalty':['none', 'elasticnet', 'l1', 'l2']},
              {'C':[0.001, 0.01, 0.1, 1, 10, 100]},
              {'max_iter':[1,10,100,1000]}]
```

- Stepwise Logistic:
  Forward Stepwise selection was trained to search for the three statistically significant features using accuracy as the performance metric. Forward Stepwise works by starting with no variables and iteratively adding features if they meet the p-value cut-off of at least 0.05. The three statistically significant features selected by stepwise regression are: *Contant Capital to Asset, Gross Profit Margin, and Retained Earnings to Asset*

```
# Perform stepwise regression
sfs = SequentialFeatureSelector(linear_model.LogisticRegression(),
                                k_features=3,
                                forward=True,
                                scoring='accuracy',
                                cv=None)
selected_features = sfs.fit(train_XBR, train_ysm)
```

```
#Check the features selected by Stepwise
selected=list((selected_features.k_feature_names_))
selected
```

```
['V_ConstCap_Asset_R', 'P_GP_Margin2_R', 'V_RE_Asset_R']
```

- RFE Logistic:

  RFE selection was trained in combination with AdaBoost regressor as an estimator. RFE works by starting with all features and iteratively removing the least significant ones based on their contribution to model performance. One of the advantages of RFE over other sequential regressions is its systematic approach of identifying important variables. It has feature importance ranking attribute. Based on the results of RFE, all features were ranked as 1, indicating that all features selected by random forest are all statistically significant.

```
#RFE selection using AdaBoost as the estimator

estimator = AdaBoostRegressor(random_state=0, n_estimators=100)
rfe_reg = RFE(estimator, n_features_to_select=5, step=1)
rfe_reg = rfe_reg.fit(train_XBR, train_ysm)

filter = rfe_reg.support_
ranking = rfe_reg.ranking_

ranking

array([1, 1, 1, 1, 1])
```

The comparative results of logistic regression models show that they perform fairly with each other.

Model	Precision	Recall	F1_Score	Accuracy	ROC_AUC
Simple Logistic	10	66	17	69	74
Grid Class Logistic	10	66	17	70	74
Stepwise Logistic	10	67	17	68	74
RFE Logistic	10	66	17	69	74

However, considering the advantages of Recursive Feature Elimination (RFE), it was selected to be the best input selection technique for logistic regression model; mainly due to its robustness in handling complex data relationships.

RFE Logistic Regression

The odds ratios were generated to determine the extent to which every increase in respective variables impact the likelihood of bankruptcy. The reverse odds ratios were also generated to determine the impact if the variables decrease.

```
# Crate Dataframe for coef and odds ratio

odds_rev=pd.DataFrame({'variable': ada_impfeat,'odds': np.e**rfe_reg.coef_[0].round(2),
                       'reverse_odds': (1/(np.e**rfe_reg.coef_[0])).round(2)})

print(odds_rev.sort_values(by='reverse_odds', ascending=False))
```

```
                   variable      odds  reverse_odds
2             P_GP_Margin2_R  0.002221        449.38
4               V_RE_Asset_R  0.029305         34.11
0         V_ConstCap_Asset_R  0.170333          5.88
1  S_TotalLiab_DailyEBITDA_R  0.697676          1.44
3     S_Financial_Coverage_R  0.990050          1.01
```

Variable	For every increase, lower likelihood of Bankruptcy by:	For every decrease, higher likelihood of Bankruptcy by:
Gross Profit Margin	99.8%	449 times
Retained Earnings_Asset	96.9%	34 times
Constant Capital_Asset	83.0%	6 times
Liability to Daily EBITDA	30.9%	44%
Financial Coverage	0.7%	1%

Based on the results of the odds ratios generated from RFE Logistic Regression, the financial metrics with strongest influence using at least 80% threshold are: *Gross Profit Margin, Retained Earnings to Asset, and Constant Capital to Asset.* The Interpretation of odds ratios are as follows:

- **Gross Profit Margin:** The risk of bankruptcy decreases by 99.8% for every percentage increase in gross profit margin; on the other hand, bankruptcy risk increases by 449 times for every percentage decrease.
- **Retained Earnings to Asset:** The risk of bankruptcy decreases by 96.9% for every percentage increase in retained earnings to asset; on the other hand, bankruptcy risk increases by 34 times for every percentage decrease.
- **Constant Capital to Asset:** The risk of bankruptcy decreases by 83% for every percentage increase in constant capital to asset; on the other hand, bankruptcy risk increases by 6 times for every percentage decrease.

3.0.5 Modeling Technique #5: Neural Network

To evaluate the effectiveness of logistic regression in comparison to a more sophisticated model, Neural Network was also explored. The results show that Neural Network outperforms logistic regression with respect to identifying bankrupt firms. But in terms of overall accuracy and ROC_AUC, they are fairly the same.

Neural Network

Confusion Matrix (Accuracy 0.7411)

Actual \ Prediction	0	1
0	17333	7455
1	5382	19406

Train: None

Confusion Matrix (Accuracy 0.6997)

Actual \ Prediction	0	1
0	11513	5013
1	201	635

Valid: None

Classification Report: Neural Network

	precision	recall	f1-score	support
Operating	0.98	0.70	0.82	16526
Bankrupt	0.11	0.76	0.20	836
accuracy			0.70	17362
macro avg	0.55	0.73	0.51	17362
weighted avg	0.94	0.70	0.79	17362

ROC_AUC:
0.7865

3.0.6 Modeling Technique #6: Advanced Ensemble Trees

Advanced Ensemble Trees were also explored to examine their predictive power in comparison to the previously explored models. Compared to the Random Forest, Bagging Classifier seems to underperform. However, XGBoost delivered a very promising result with F1-score of 0.52 and ROC_AUC of 90.16, indicating a high predictive power considering the degree of data imbalance.

Bagging Classifier

Confusion Matrix (Accuracy 1.0000)

Actual \ Prediction	0	1
0	24788	0
1	0	24788

Train: None

Confusion Matrix (Accuracy 0.9386)

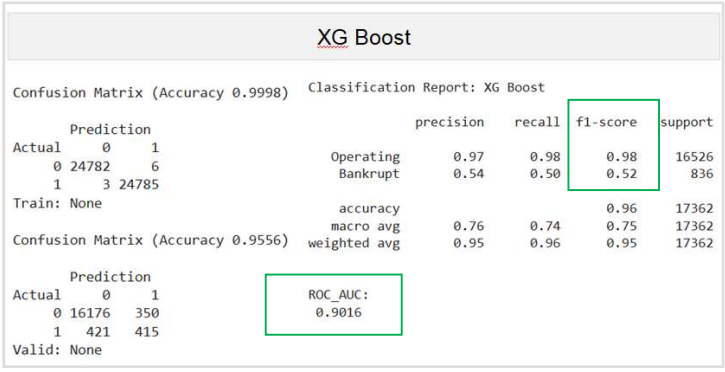
Actual \ Prediction	0	1
0	15942	584
1	482	354

Valid: None

Classification Report: Bagging Classifier

	precision	recall	f1-score	support
Operating	0.97	0.96	0.97	16526
Bankrupt	0.38	0.42	0.40	836
accuracy			0.94	17362
macro avg	0.67	0.69	0.68	17362
weighted avg	0.94	0.94	0.94	17362

ROC_AUC:
0.8643



Model Recommendation

1.0. Model Selection

Based on the results of model explorations, the best model is the one with the highest predictive power based on F1-score and ROC_AUC. However, aside from predictive capability, interpretability of results is also an equally important criterion. Although Random Forest and XG Boost appeared to be very promising with respect to predicting high number of bankrupt firms, they have limitations with respect to interpretability. Therefore, the best model selected is **Logistic Regression** in which its inputs undergone two stages of selection process; feature importance in *Random Forest* and subsequently variable significance through *Recursive Feature Elimination (RFE).*

Modeling Technique	Model	Precision	Recall	F1_Score	Accuracy	ROC_AUC
Decision Tree	FullClassTree	22	46	30	90	69
	GridClassTree	15	45	23	85	79
	RandomForest	45	38	41	95	87
	Bagging Classifier	38	42	40	94	86
	XGBoost	54	50	52	96	90
Logistic Regression	Simple Logistic	10	66	17	69	74
	GridClass Logistic	10	66	17	70	74
	Stepwise Logistic	10	67	17	68	74
	RFE Logistic	**10**	**66**	**17**	**69**	**74**
Neural Network	NeuralNetwork	11	76	20	70	79

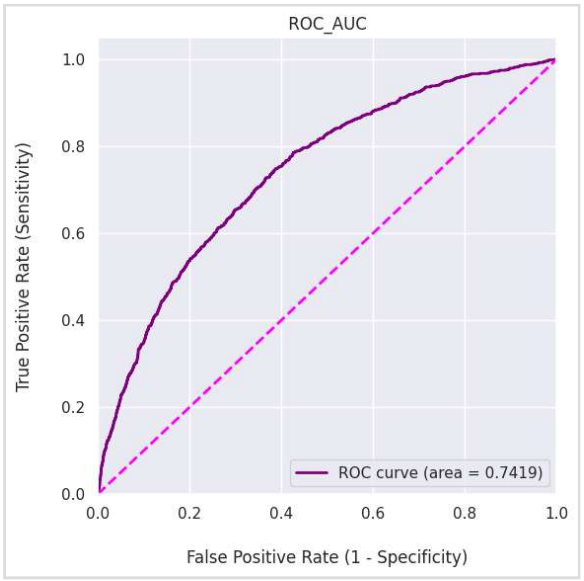
*** Model Performance scores are in %*

*** Precision, Recall, and F1-score are based on result for positive class ("Bankrupt")*

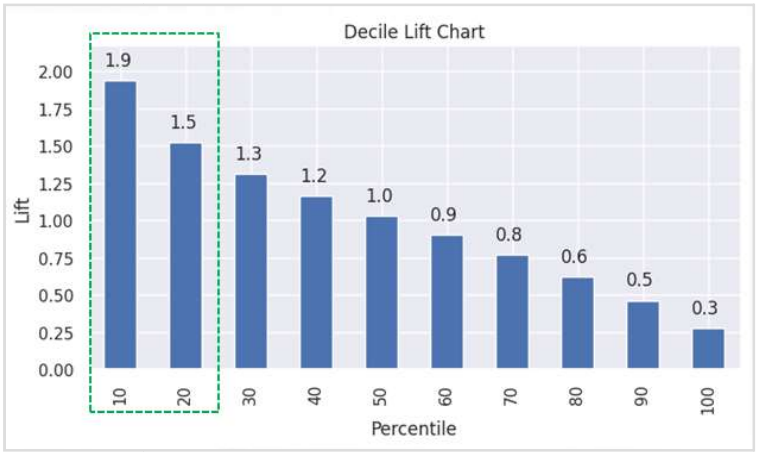
It is worthy to note that there is significant gap between the result of F1-score and overall accuracy. This was mainly driven by highly imbalanced validation set that is dominated by 95% non-bankrupt class. Hence, a secondary metric which is **ROC_AUC** was used**.** Unlike the F1-score, ROC_AUC is not sensitive to imbalanced distribution since it focuses on the model's ability to effectively distinguish bankrupt from non-bankrupt firms.

Based on the results of the best model which is RFE logistic regression, the financial ratios that signal early indicators of corporate bankruptcy in manufacturing firms are: ***Gross Profit Margin, Retained Earnings to Asset,*** *and* ***Constant Capital to Asset***

Efficacy of the Best Model: RFE Logistic Regression



RFE Logistic Regression has an ROC_AUC of .7419 which means that the model has 74.19% effectivity in distinguishing bankrupt from operating firms.



Furthermore, the model is better than just random guessing or no model at all by 1.9 and 1.5 times for the top 10% and 20%, respectively.

2.0. Model Theory

Logistic regression is a widely used machine learning technique for predictions specifically binary classification problems such as bankruptcy prediction. It was explored and selected in this study because of its potential of estimating how a unit movement in the financial leading indicators impact likelihood of bankruptcy, which is one of the objectives of this study.

Logistic regression is an extension of linear regression; however, instead of numerical predictions, its outcome is in the form of probabilities; in the case of this study, the probability of a particular firm to belong into bankrupt or non-bankrupt class. Based on the values of input features, it uses logit as its mathematical function which maps predictions into probabilities within the range of 0 and 1.

Below is the translated mathematical function using the intercept and coefficients of the financial ratios identified through the RFE logistic regression.

$$P(Y=1|X)= \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)}}$$

Where:

$P(Y=1|X)$: probability of bankruptcy

B_0 : intercept

$\beta_1, \beta_2, \beta_3$ = coefficients of leading financial indicator

e = natural logarithm

variable	intercept	coef
P_GP_Margin2_R	1.08	-6.11
V_RE_Asset_R	1.08	-3.53
V_ConstCap_Asset_R	1.08	-1.77
S_TotalLiab_DailyEBITDA_R	1.08	-0.36
S_Financial_Coverage_R	1.08	-0.01

$$P(Y=1|X)= \frac{1}{1+e^{-((1.08+(-6.11 * \text{Gross Profit Margin})+(-3.53 * \text{Retained Earnings to Asset})+(-1.77 * \text{Constant Capital to Asset}))}}$$

3.0. Model Assumptions and Limitations

3.0.1. Model Assumptions

In order for a logistic regression model to generate reliable results in bankruptcy prediction, the following assumptions must be satisfied:

- Target must be binary variable expressed as 0 and 1
- The firms under observation are independent of each other
- Predictor variables are not highly correlated
- No extreme outliers
- Existence of linear relationship between predictors and logit of target variables
- The sample size must be large

3.0.2. Model Limitation

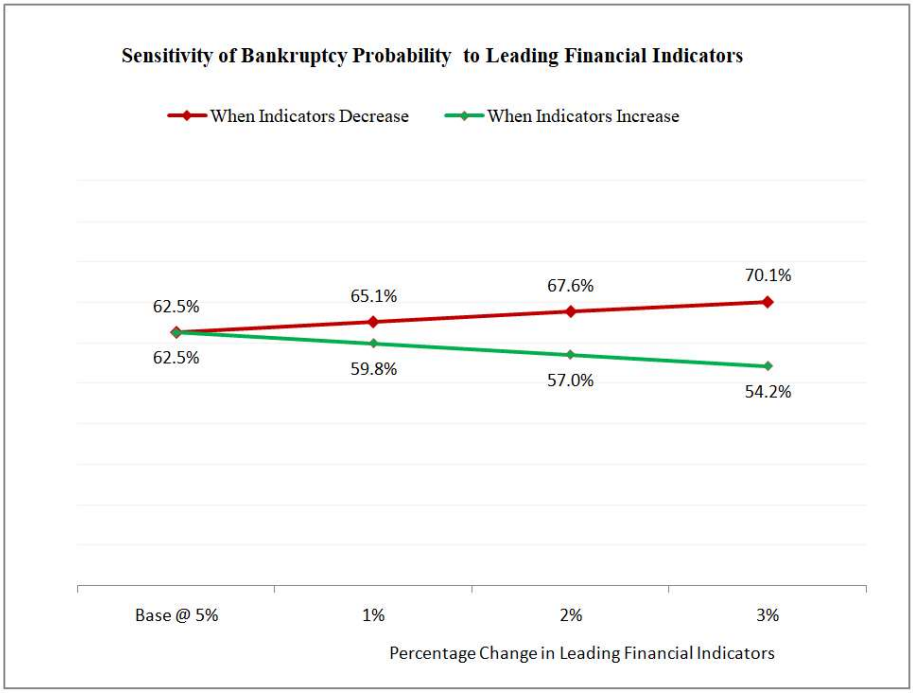
Considering the above assumptions, the logistic regression cannot be trained with set of input variables that have high degree of multicollinearity. This is the challenge of using financial ratios as input variables. This is due to the fact that some financial ratios, which were derived from common components of financial statements, are inherently related with each other. This is evident in the outcome VIF scoring that was performed earlier during the data pre-processing phase. Despite of carrying out necessary techniques to reduce multicollinearity such as removing highly correlated variables using 0.7 thresholds, some variables still have VIF scores above 10.

Hence, the financial ratios that must be fed into the model should have no or low degree of multicollinearity in order to maximize its predictive and interpretability power. This was addressed by employing tree-based algorithms such as *Random Forest* in selecting features. The robustness of random forest to various data quality issues such as missing values, outliers, and multicollinearity make it an ideal complement to boost reliability of logistic regression.

4.0. Model Sensitivity to Key Drivers

A simulation was performed to test the sensitivity of bankruptcy probability to change in the identified financial ratios. With a base ratio of 5%, the corresponding bankruptcy probability is at 62.5%. As their ratio increases, the probability of bankruptcy decreases. On the other hand, as their ratio decreases the

probability of bankruptcy increases. Given the sensitivity of the model to change in the identified financial ratios, it can be gleaned that it is effective in predicting bankruptcy.



Key Insights, Conclusion, and Recommendations

15.0. Additional Insights

To further understand the critical thresholds that distinguish financially sound firms, from those at-risk, comparative analysis was performed. Those manufacturing firms that were correctly predicted by the best model were compared. Firms with >=90% probability (highest bankruptcy risk) were compared to firms with <=10% probability (lowest bankruptcy risk) with respect to the three leading financial indicators (Appendix, Figures 3-5). The results of the analysis are as follows:

- **Gross Profit Margin**: while operating firms have at least 7% gross profit margin, bankrupt firms have negative over the five-year horizon. This indicates that direct manufacturing costs exceed revenue and suggests either operational inefficiency (i.e. wastage) or poor pricing strategy.

- **Retained Earnings to Asset**: while operating firms maintain at least 5%, bankrupt firms have negative retained earnings to asset ratio over the five-year horizon. This indicates that bankrupt

firms have accumulated and sustained loss over the years, suggesting struggle of maintaining sustainable earnings.

- **Constant Capital to Asset**: while operating firms maintain at least 65% of their asset financed by constant capital such as share capital, bankrupt firms have over 30% over the five-year horizon. This indicates high leverage on external financing such as interest-bearing debt.

16.0. Conclusion

Based on the results of the predictive modeling and comparative analysis performed, it is concluded that the Leading Financial Indicators of Corporate Bankruptcy for manufacturing firms are:

- Negative Gross Profit Margin - Production cost exceeding revenue that suggests operational inefficiencies
- Negative Retained Earnings – Inability to generate sustainable profit over years
- Constant Capital to Asset of <50% – Minimal stable long-term capital and high reliance on external funding such as interest-bearing debts

Therefore, the manufacturing firms with highest risk of going bankrupt are those with **high debt leverage**, but with significant **cost inefficiencies**, and **weak earnings capacity**.

17.0. Impacts on Business Problem

Through predictive modeling, the leading financial indicators of bankruptcy were identified. Additional analysis further revealed the critical thresholds that distinguish operating from bankrupt firms. The insights obtained from this study will help the financial management team of Canadian manufacturing firms to mitigate bankruptcy risk by acting on the warnings signs revealed by financial indicators. This will then lead to long-term financial stability and business viability.

18.0. Recommendation

18.0.1 Business Recommendation: Internal Finance Team of Canadian Firms

- Proactive Financial Planning
  - Regularly monitor the identified leading indicators and their related metrics.
  - Consider the critical thresholds as benchmarks in establishing 5-year financial forecast and budget plans

- Robust Cost Monitoring
  - Analyze thoroughly the areas of production with high cost inefficiencies
  - Clearly communicate the findings to operations team
  - Propose them actionable solutions to improve production process
- Optimal Capital Structure
  - Determine the right balance of equity and debt financing considering company's earnings capacity

18.0.2. Model Recommendation: Business Analytics Team

- Usage of the same model employed in this analysis
  - Test the model using most recent dataset of Canadian manufacturing firms
  - If the outcomes differ from those generated in this analysis, re-train the model employing the methodologies discussed in this paper.
  - Data inputs should be of high quality such as no missing values and outliers to ensure better model performance
  - Model calibration to handle highly imbalanced dataset
- Utilize more advanced models
    - If the objective is to solely predict risk of bankruptcy, it is recommended to use more sophisticated models such as XG Boost due to its superior predictive ability.

References

Baldwin, J., Gray, T., Johnson, J., Proctor, J., Rafiquzzaman, M., & Sabourin, D. (1997). *Failing Concerns : Business Bankruptcy in Canada.* Statistics Canada. https://publications.gc.ca/site/eng/9.688052/publication.html

Cooke, A., & McDougall A. (2023). *A profile of corporate exits and insolvencies.* Statistics Canada. https://www150.statcan.gc.ca/n1/pub/36-28-0001/2023010/article/00005-eng.htm

Statistics Canada. (2023). *Historic Insolvency Statistics.* https://open.canada.ca/data/en/dataset/ed4d48f3-750f-4eeb-93f9-cb62be138264/resource/179a7ca2-c922-4ca0-8f3c-1489cae8a2cb

Smith, B. (2023). *The Complete Corporate Bankruptcy in Canada Guide: What Every Business Owner Needs to Know.* IRA Smith. https://irasmithinc.com/blog/corporate-bankruptcy-in-canada-2/

Tomczak, S. (2016). *Polish Companies Bankruptcy*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/365/polish+companies+bankruptcy+data

Appendix

Figure1. Corporate Bankruptcy filings in Canada from 1990 to 2023

Sources: Statistics Canada; Office of the Superintendent of Bankruptcy; and authors' tabulations.

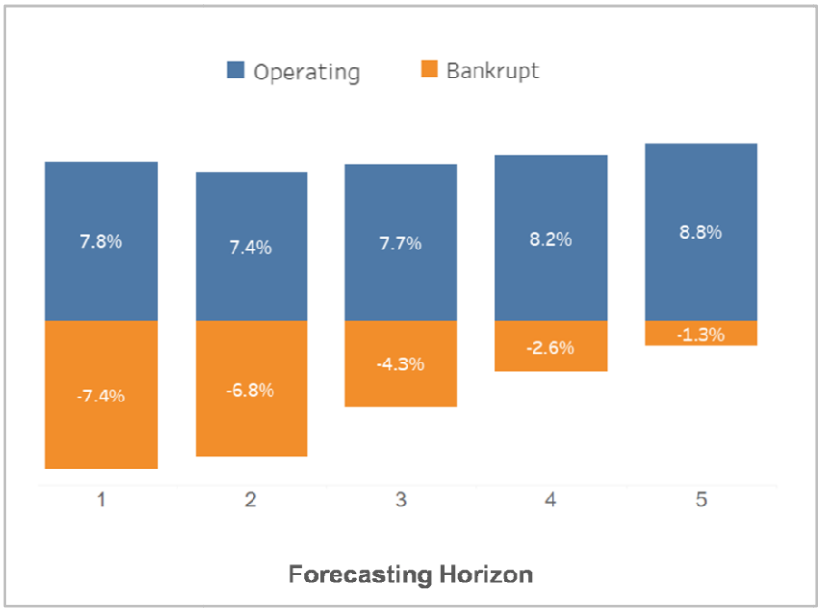
Figure 1

Figure2. Bankruptcy rates by industry, average over each period

Industry	Bankruptcy rates	
	2004 to 2009	2010 to 2019
	percent	
Agriculture, forestry, fishing and hunting	0.16	0.08
Mining, quarrying, and oil and gas extraction	x	x
Utilities	x	x
Construction	0.2	0.23
Manufacturing	0.64	0.45
Wholesale trade	0.33	0.26
Retail trade	0.31	0.27
Transportation and warehousing	0.29	0.15
Information and cultural industries	0.22	0.18
Real estate and rental and leasing	0.11	0.09
Professional, scientific and technical	0.11	0.09
Administrative and support, waste	0.2	0.2
Finance and insurance, and management of	0.08	0.08
Arts, entertainment and recreation	0.21	0.2
Accommodation and food services	0.46	0.48
Other services (except public	0.15	0.13

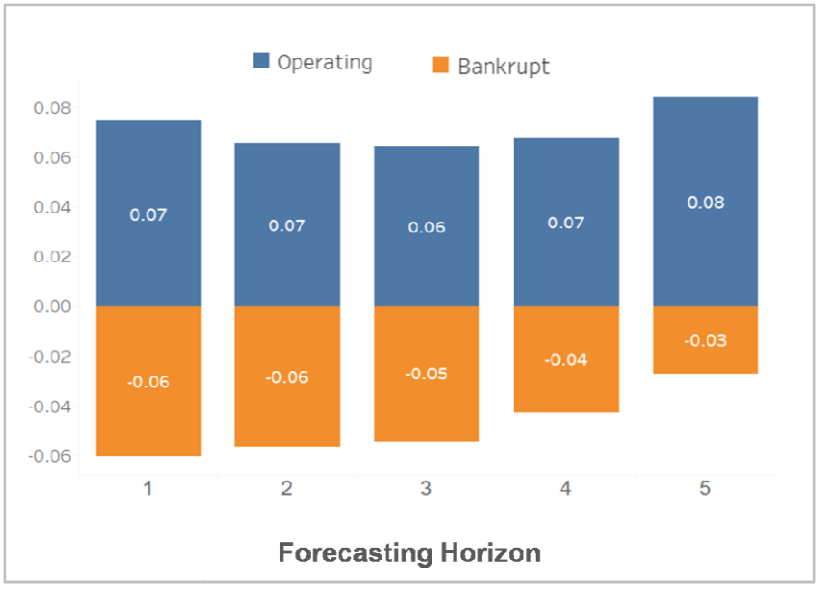
Sources: Statistics Canada, National Accounts Longitudinal Microdata File; Office of the Superintendent of Bankruptcy; and authors' tabulations.

Figure 2



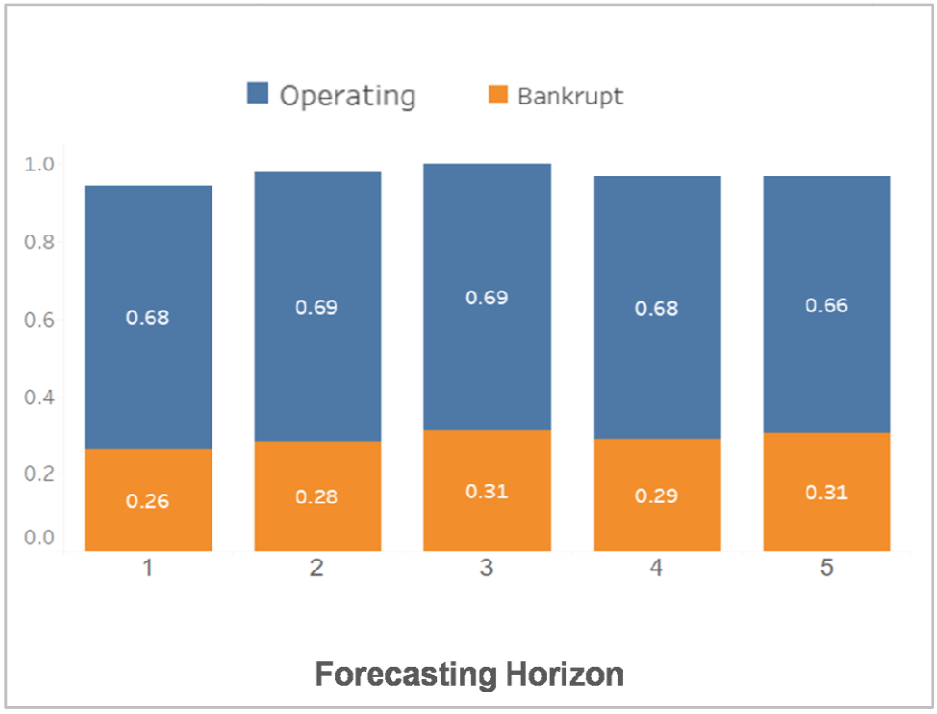
Gross Profit Margin

Figure 3



Retained Earnings to Asset

Figure 4



Constant Capital to Asset

Figure 5