FRA Extended Project

Bankruptcy Prediction

ANIL KUMAR PGBPDSBA AUG'2023

Contents

Problem Statement definition and EDA.....2

Data Pre-Processing 12

Model Building 15

Model Performance Improvement 17

Model Performance Comparison and Final Model Selection.....24

Actionable Insights and recommendations 28

1.0 Problem Statement – Bankruptcy Prediction

Bankruptcy prediction is a crucial component of financial risk management that protects the interests of creditors, investors, and other stakeholders. Predicting a company's impending bankruptcy can help with timely interventions and smart decision-making, which can reduce losses and promote stability in the economy. Predictive modelling can benefit from the abundance of financial data provided by US corporations listed on major exchanges such as the New York Stock Exchange (NYSE) and NASDAQ, which are subject to regulatory scrutiny and strict financial reporting requirements. A firm is considered bankrupt, according to the Securities Exchange Commission (SEC), if it files for bankruptcy under the Bankruptcy Code's Chapter 11 (reorganization) or Chapter 7 (liquidation) provisions.

Objective

A well-known financial analytics company wants to create a Bankruptcy Prediction Tool to help regulators, investors, and financial institutions assess the bankruptcy risk of US publicly traded corporations. The program will evaluate past financial data using cutting-edge machine learning algorithms to find important signs and trends related to bankruptcy. The following are this tool's main goals:

1. Bankruptcy Risk Assessment: Provide a probabilistic estimate of a company's likelihood of filing for bankruptcy within a specified time frame (e.g., one year), allowing stakeholders to make informed decisions and take preventive measures.
2. Early Warning System: Develop an early warning system that flags companies exhibiting financial distress signals, enabling proactive risk management and strategic planning.
3. Financial Health Analysis: Analyze various financial metrics to offer a comprehensive assessment of a company's financial health, highlighting areas of concern and potential vulnerabilities.

As part of the data science team in the firm, you have been provided with a dataset containing financial metrics of various companies. The task is to analyze the data and develop a predictive model using machine learning techniques to identify whether a given company is at risk of bankruptcy in the near future. The model will help the organization anticipate potential financial distress and enable proactive measures to manage risks effectively..

Data Dictionary:

The data consists of financial metrics from the balance sheets of different companies

1. **Company_id**: Unique identifier for each company
2. **Current_assets**: Total current assets (in millions)
3. **Cost_of_goods_sold**: Cost of goods sold (in millions)
4. **Depreciation_and_amortization**: Depreciation and amortization expenses (in millions)
5. **EBITDA**: Earnings Before Interest, Taxes, Depreciation, and Amortization (in millions)
6. **Inventory**: Value of inventory (in millions)
7. **Net_income**: Net income (profit or loss) (in millions)
8. **Total_receivables**: Total receivables (in millions)
9. **Market_value**: Market value of the company (in millions)
10. **Net_sales**: Net sales or revenue (in millions)
11. **Total_assets**: Total assets (in millions)
12. **Total_long_term_debt**: Total long-term debt (in millions)
13. **EBIT**: Earnings Before Interest and Taxes (in millions)
14. **Gross_profit**: Gross profit (in millions)
15. **Total_current_liabilities**: Total current liabilities (in millions)
16. **Retained_earnings**: Retained earnings (in millions)
17. **Total_revenue**: Total revenue (in millions)
18. **Total_liabilities**: Total liabilities (in millions)
19. **Total_operating_expenses**: Total operating expenses (in millions)
20. **Bankrupt**: Bankruptcy status (1 = Bankrupt, 0 = Not Bankrupt)

1 Exploratory Data Analysis

1a. Check Shapes

```
[ ]  df.shape ##  Complete the code to view dimensions of the data
```

```
(1983, 20)
```

1b. Data Type of each variable:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1983 entries, 0 to 1982
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Company_id                      1983 non-null   object
 1   Current_assets                  1983 non-null   float64
 2   Cost_of_goods_sold              1983 non-null   float64
 3   Depreciation_and_amortization   1983 non-null   float64
 4   EBITDA                          1983 non-null   float64
 5   Inventory                       1983 non-null   float64
 6   Net_income                      1983 non-null   float64
 7   Total_receivables               1983 non-null   float64
 8   Market_value                    1983 non-null   float64
 9   Net_sales                       1983 non-null   float64
 10  Total_assets                    1983 non-null   float64
 11  Total_long_term_debt            1983 non-null   float64
 12  EBIT                            1983 non-null   float64
 13  Gross_profit                    1983 non-null   float64
 14  Total_current_liabilities       1983 non-null   float64
 15  Retained_earnings               1983 non-null   float64
 16  Total_revenue                   1983 non-null   float64
 17  Total_liabilities               1983 non-null   float64
 18  Total_operating_expenses        1983 non-null   float64
 19  Bankrupt                        1983 non-null   int64
dtypes: float64(18), int64(1), object(1)
memory usage: 310.0+ KB
```

There are 1 object type,1 integer type and 18 float data type fields are available in the Dataset.

1c. Statistical Summary of dataset:

	count	mean	std	min	25%	50%	75%	max
Current_assets	1983.0	485.485038	1790.167597	0.0020	14.1920	63.2820	234.6860	36105.0000
Cost_of_goods_sold	1983.0	1024.313261	4316.517260	0.0000	14.8690	72.3840	338.1010	76809.0000
Depreciation_and_amortization	1983.0	86.553624	400.228725	0.0000	1.4920	6.4380	30.2085	9338.0000
EBITDA	1983.0	162.429048	880.011401	-5743.0000	-2.5165	5.3120	55.9780	18632.0000
Inventory	1983.0	119.278526	506.207801	0.0000	0.0000	5.7020	44.2140	8923.0000
Net_income	1983.0	-53.004835	1536.719467	-56121.9000	-14.9175	-0.7970	9.5450	8560.0000
Total_receivables	1983.0	171.701946	882.459285	0.0000	2.4410	13.8250	61.8435	28813.0000
Market_value	1983.0	1630.979822	8159.299586	0.0384	18.2261	96.3835	588.2171	180090.4065
Net_sales	1983.0	1426.305833	5627.381670	0.0020	23.9920	115.0030	535.4695	97863.0000
Total_assets	1983.0	1641.765901	6932.946945	0.0040	29.4450	131.4280	574.6155	165282.0000
Total_long_term_debt	1983.0	422.770059	1902.188540	0.0000	0.0415	5.6060	125.9820	45247.0000
EBIT	1983.0	75.875424	697.454528	-10537.0000	-7.7785	0.9800	28.2255	16295.0000
Gross_profit	1983.0	401.992572	1763.762767	-1317.0000	6.1790	35.9990	174.9465	41000.0000
Total_current_liabilities	1983.0	364.712545	1513.951417	0.0040	7.6135	27.4420	114.9555	25427.0000
Retained_earnings	1983.0	133.583326	2235.727309	-57158.0000	-56.0090	-3.8670	64.1825	33896.0000
Total_revenue	1983.0	1426.305833	5627.381670	0.0020	23.9920	115.0030	535.4695	97863.0000
Total_liabilities	1983.0	1030.597176	4515.317860	0.0040	11.0270	49.1320	315.2695	110042.0000
Total_operating_expenses	1983.0	1263.876785	4970.260681	0.0670	31.2950	111.3780	488.9880	81832.0000
Bankrupt	1983.0	0.208775	0.406535	0.0000	0.0000	0.0000	0.0000	1.0000

1d. Data Analysis

There are no duplicate values

```
# checking for duplicate values
df.duplicated().sum()
```

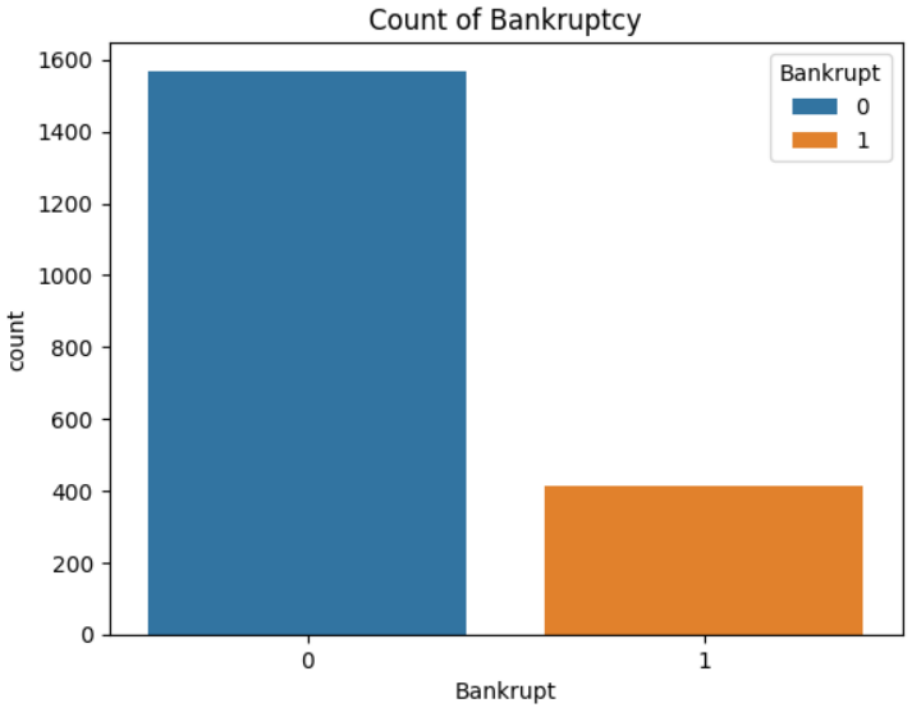
```
0
```

Unique entries in the dataset.

```
Company_id                        1983
Current_assets                    1966
Cost_of_goods_sold                1969
Depreciation_and_amortization     1850
EBITDA                            1967
Inventory                         1452
Net_income                        1949
Total_receivables                 1847
Market_value                      1982
Net_sales                         1964
Total_assets                      1971
Total_long_term_debt              1452
EBIT                              1963
Gross_profit                      1968
Total_current_liabilities         1960
Retained_earnings                 1974
Total_revenue                     1964
Total_liabilities                 1962
Total_operating_expenses          1973
Bankrupt                             2
dtype: int64
```

1 - Univariate Analysis

The number of instances where bankruptcy did not occur (category '0') is significantly higher than the number of instances where default did occur (category '1'). This indicates that non-default cases are more prevalent than default cases within this dataset. A company will not be tagged as a defaulter if its net worth next year is positive, or else, it'll be tagged as a defaulter.

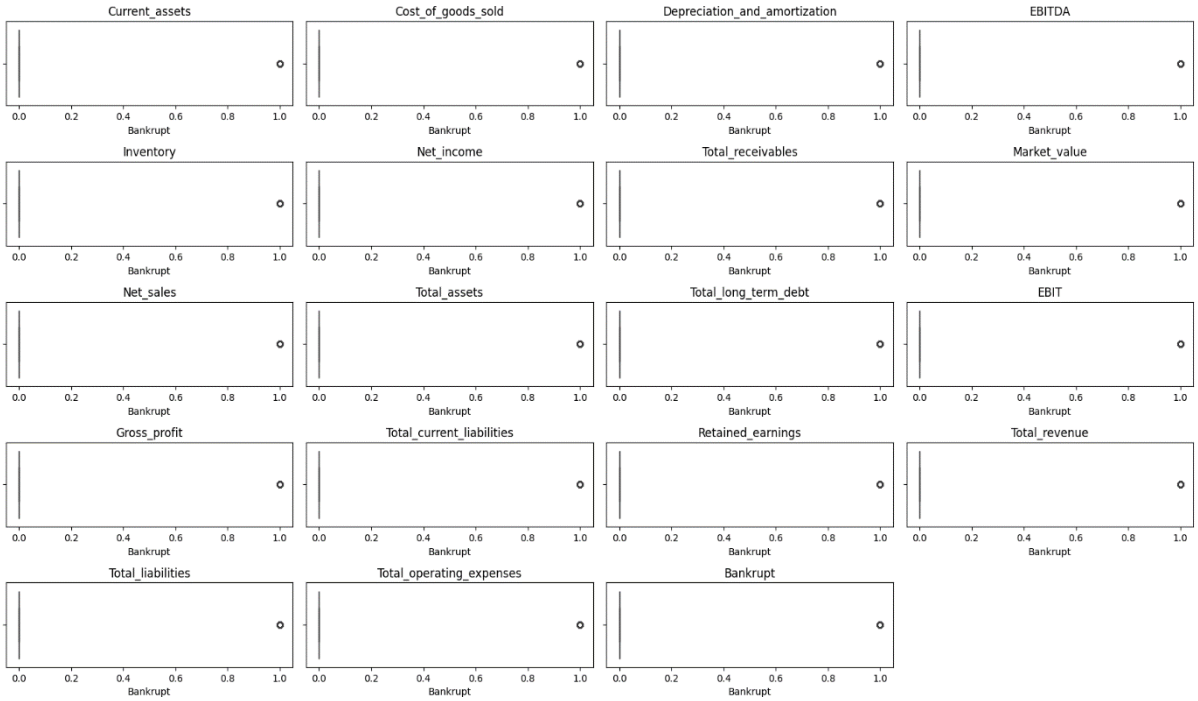


Percentage of Defaulters

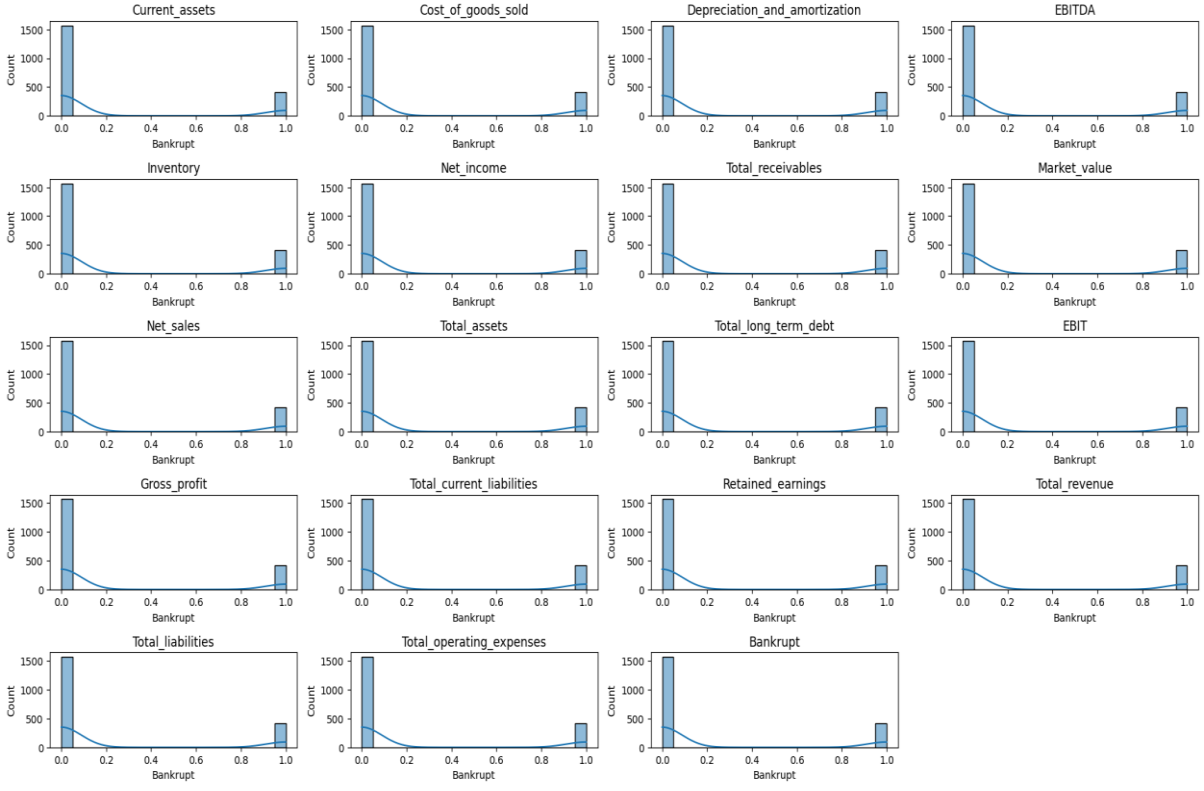
```
#Percentage of defaulters
(df.Bankrupt.sum()/len(df)) * 100
```

```
20.87745839636914
```

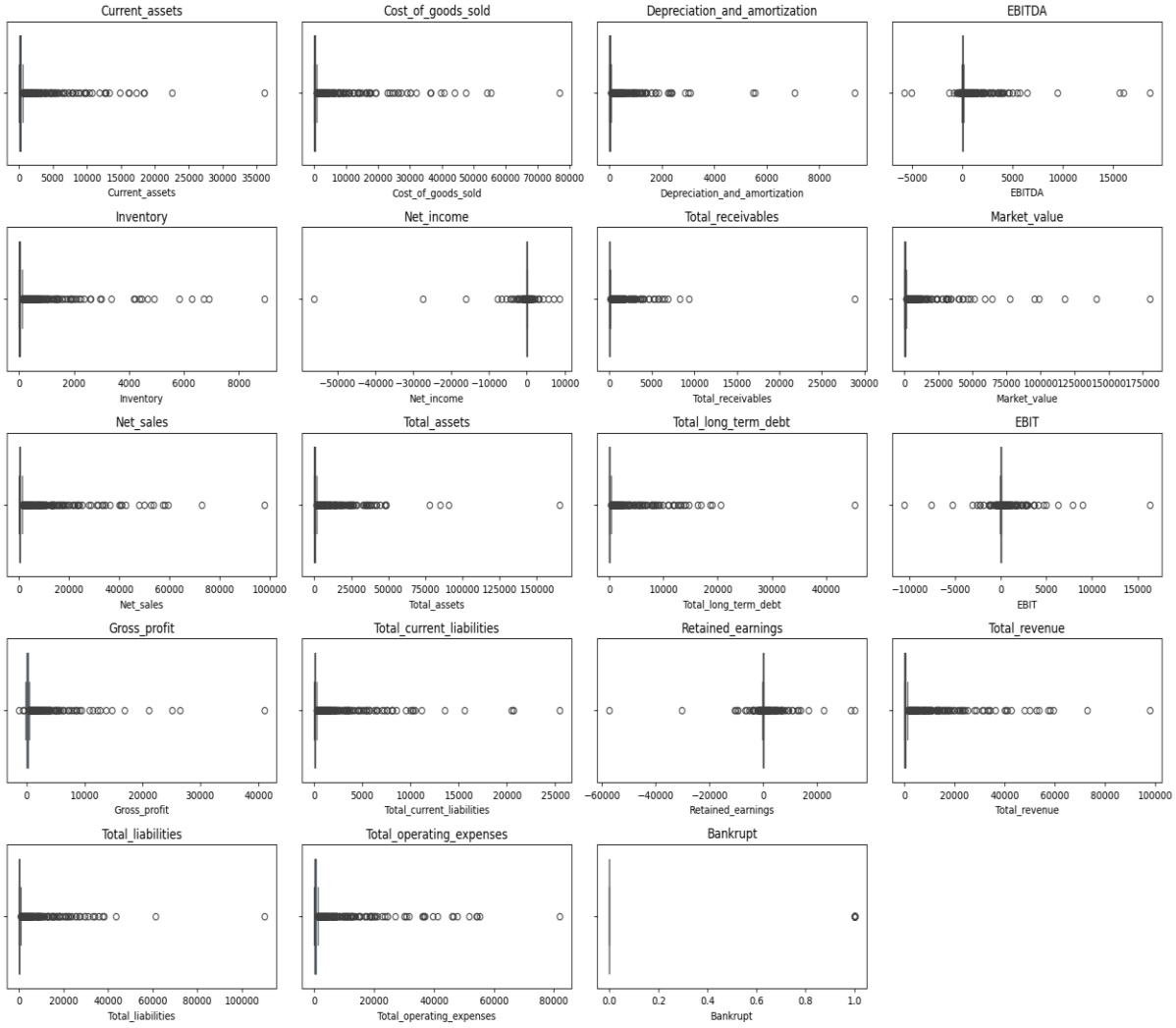
- **Boxplot**



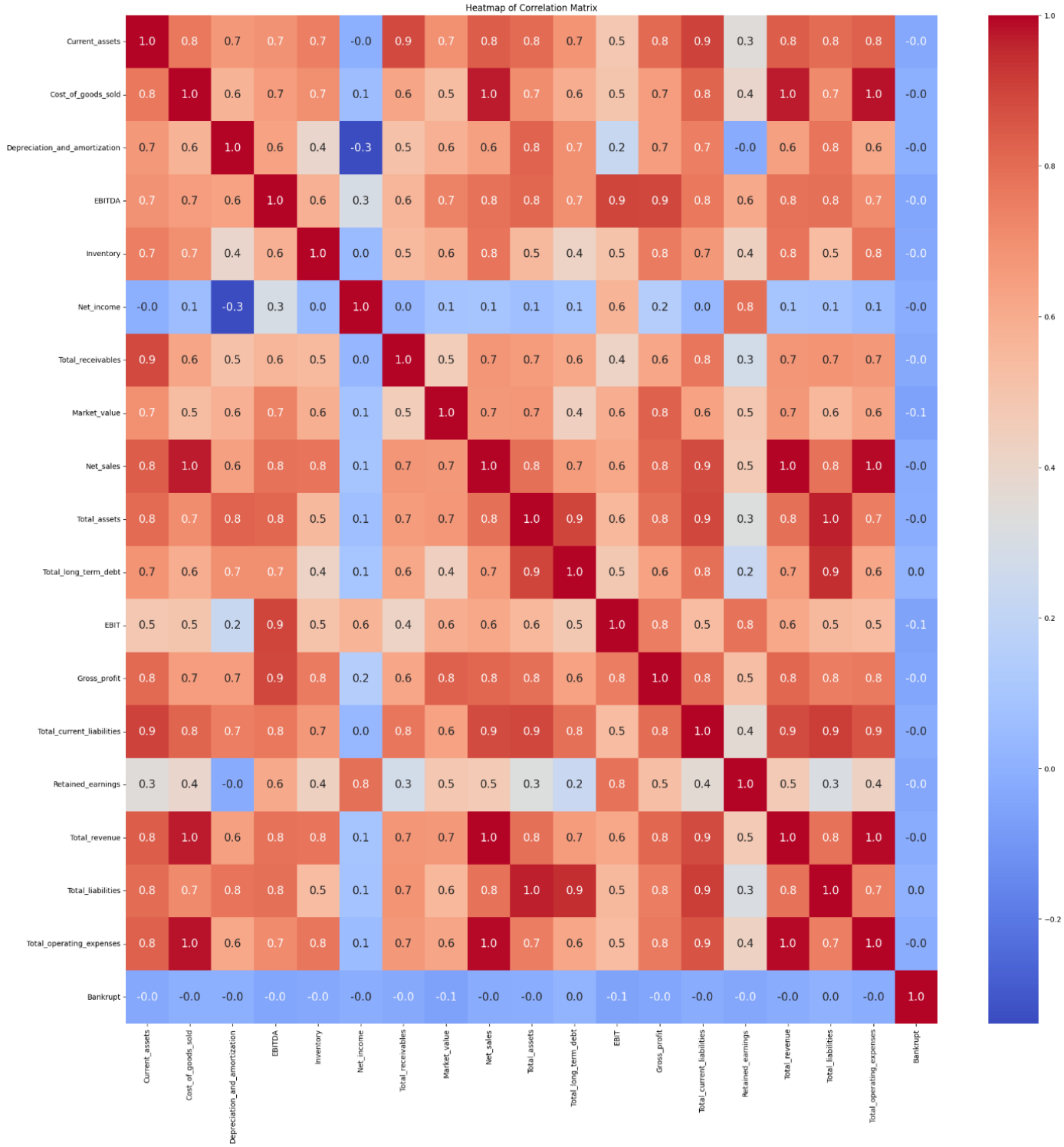
- **Histplot**



II - B


Current_assets
0 5000 10000 15000 20000 25000 30000 35000
Current_assets
Cost_of_goods_sold
0 10000 20000 30000 40000 50000 60000 70000 80000
Cost_of_goods_sold
Depreciation_and_amortization
0 2000 4000 6000 8000
Depreciation_and_amortization
EBITDA
−5000 0 5000 10000 15000
EBITDA
Inventory
0 2000 4000 6000 8000
Inventory
Net_income
−50000 −40000 −30000 −20000 −10000 0 10000
Net_income
Total_receivables
0 5000 10000 15000 20000 25000 30000
Total_receivables
Market_value
0 25000 50000 75000 100000 125000 150000 175000
Market_value
Net_sales
0 20000 40000 60000 80000 100000
Net_sales
Total_assets
0 25000 50000 75000 100000 125000 150000
Total_assets
Total_long_term_debt
0 10000 20000 30000 40000
Total_long_term_debt
EBIT
−10000 −5000 0 5000 10000 15000
EBIT
Gross_profit
0 10000 20000 30000 40000
Gross_profit
Total_current_liabilities
0 5000 10000 15000 20000 25000
Total_current_liabilities
Retained_earnings
−60000 −40000 −20000 0 20000
Retained_earnings
Total_revenue
0 20000 40000 60000 80000 100000
Total_revenue
Total_liabilities
0 20000 40000 60000 80000 100000
Total_liabilities
Total_operating_expenses
0 20000 40000 60000 80000
Total_operating_expenses
Bankrupt
0.0 0.2 0.4 0.6 0.8 1.0
Bankrupt

Heat Map


Heatmap of Correlation Matrix
Current_assets 1.0 0.8 0.7 0.7 0.7 -0.0 0.9 0.7 0.8 0.8 0.7 0.5 0.8 0.9 0.3 0.8 0.8 0.8 -0.0
Cost_of_goods_sold 0.8 1.0 0.6 0.7 0.7 0.1 0.6 0.5 1.0 0.7 0.6 0.5 0.7 0.8 0.4 1.0 0.7 1.0 -0.0
Depreciation_and_amortization 0.7 0.6 1.0 0.6 0.4 -0.3 0.5 0.6 0.6 0.8 0.7 0.2 0.7 0.7 -0.0 0.6 0.8 0.6 -0.0
EBITDA 0.7 0.7 0.6 1.0 0.6 0.3 0.6 0.7 0.8 0.8 0.7 0.9 0.9 0.8 0.6 0.8 0.8 0.7 -0.0
Inventory 0.7 0.7 0.4 0.6 1.0 0.0 0.5 0.6 0.8 0.5 0.4 0.5 0.8 0.7 0.4 0.8 0.5 0.8 -0.0
Net_income -0.0 0.1 -0.3 0.3 0.0 1.0 0.0 0.1 0.1 0.1 0.1 0.6 0.2 0.0 0.8 0.1 0.1 0.1 -0.0
Total_receivables 0.9 0.6 0.5 0.6 0.5 0.0 1.0 0.5 0.7 0.7 0.6 0.4 0.6 0.8 0.3 0.7 0.7 0.7 -0.0
Market_value 0.7 0.5 0.6 0.7 0.6 0.1 0.5 1.0 0.7 0.7 0.4 0.6 0.8 0.6 0.5 0.7 0.6 0.6 -0.1
Net_sales 0.8 1.0 0.6 0.8 0.8 0.1 0.7 0.7 1.0 0.8 0.7 0.6 0.8 0.9 0.5 1.0 0.8 1.0 -0.0
Total_assets 0.8 0.7 0.8 0.8 0.5 0.1 0.7 0.7 0.8 1.0 0.9 0.6 0.8 0.9 0.3 0.8 1.0 0.7 -0.0
Total_long_term_debt 0.7 0.6 0.7 0.7 0.4 0.1 0.6 0.4 0.7 0.9 1.0 0.5 0.6 0.8 0.2 0.7 0.9 0.6 0.0
EBIT 0.5 0.5 0.2 0.9 0.5 0.6 0.4 0.6 0.6 0.6 0.5 1.0 0.8 0.5 0.8 0.6 0.5 0.5 -0.1
Gross_profit 0.8 0.7 0.7 0.9 0.8 0.2 0.6 0.8 0.8 0.8 0.6 0.8 1.0 0.8 0.5 0.8 0.8 0.8 -0.0
Total_current_liabilities 0.9 0.8 0.7 0.8 0.7 0.0 0.8 0.6 0.9 0.9 0.8 0.5 0.8 1.0 0.4 0.9 0.9 0.9 -0.0
Retained_earnings 0.3 0.4 -0.0 0.6 0.4 0.8 0.3 0.5 0.5 0.3 0.2 0.8 0.5 0.4 1.0 0.5 0.3 0.4 -0.0
Total_revenue 0.8 1.0 0.6 0.8 0.8 0.1 0.7 0.7 1.0 0.8 0.7 0.6 0.8 0.9 0.5 1.0 0.8 1.0 -0.0
Total_liabilities 0.8 0.7 0.8 0.8 0.5 0.1 0.7 0.6 0.8 1.0 0.9 0.5 0.8 0.9 0.3 0.8 1.0 0.7 0.0
Total_operating_expenses 0.8 1.0 0.6 0.7 0.8 0.1 0.7 0.6 1.0 0.7 0.6 0.5 0.8 0.9 0.4 1.0 0.7 1.0 -0.0
Bankrupt -0.0 -0.0 -0.0 -0.0 -0.0 -0.0 -0.0 -0.1 -0.0 -0.0 0.0 -0.1 -0.0 -0.0 -0.0 -0.0 0.0 -0.0 1.0
Current_assets
Cost_of_goods_sold
Depreciation_and_amortization
EBITDA
Inventory
Net_income
Total_receivables
Market_value
Net_sales
Total_assets
Total_long_term_debt
EBIT
Gross_profit
Total_current_liabilities
Retained_earnings
Total_revenue
Total_liabilities
Total_operating_expenses
Bankrupt
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Correlation Matrix

	Current_assets	Cost_of_goods_sold	Depreciation_and_amortization	EBITDA	Inventory	Net_income	Total_receivables	Market_value	Net_sales	Total_assets	Total_long_te
Current_assets	1.000000	0.764695	0.666402	0.687359	0.744692	-0.000831	0.888061	0.693776	0.839917	0.805285	
Cost_of_goods_sold	0.764695	1.000000	0.571812	0.653193	0.701346	0.058542	0.628522	0.531706	0.971337	0.706130	
Depreciation_and_amortization	0.666402	0.571812	1.000000	0.636219	0.396143	-0.338753	0.511850	0.582133	0.643497	0.824857	
EBITDA	0.687359	0.653193	0.636219	1.000000	0.582503	0.295105	0.561993	0.736475	0.781311	0.812891	
Inventory	0.744692	0.701346	0.396143	0.582503	1.000000	0.043190	0.542657	0.588801	0.775437	0.546271	
Net_income	-0.000831	0.058542	-0.338753	0.295105	0.043190	1.000000	0.030464	0.089075	0.092535	0.067055	
Total_receivables	0.888061	0.628522	0.511850	0.561993	0.542657	0.030464	1.000000	0.452692	0.676703	0.659380	
Market_value	0.693776	0.531706	0.582133	0.736475	0.588801	0.089075	0.452692	1.000000	0.667264	0.674433	
Net_sales	0.839917	0.971337	0.643497	0.781311	0.775437	0.092535	0.676703	0.667264	1.000000	0.788975	
Total_assets	0.805285	0.706130	0.824857	0.812891	0.546271	0.067055	0.659380	0.674433	0.788975	1.000000	
Total_long_term_debt	0.684081	0.595583	0.701741	0.698284	0.418482	0.095780	0.614947	0.435396	0.655816	0.896229	
EBIT	0.484864	0.496034	0.228906	0.896658	0.507648	0.566739	0.415372	0.595193	0.616551	0.552326	
Gross_profit	0.808335	0.651769	0.653697	0.894235	0.757644	0.151965	0.620854	0.827678	0.813369	0.789130	
Total_current_liabilities	0.902681	0.821968	0.744939	0.765496	0.669996	0.037280	0.772982	0.638919	0.881054	0.891135	
Retained_earnings	0.346057	0.397545	-0.021363	0.610048	0.431274	0.790055	0.277845	0.482765	0.469773	0.329346	
Total_revenue	0.839917	0.971337	0.643497	0.781311	0.775437	0.092535	0.676703	0.667264	1.000000	0.788975	
Total_liabilities	0.800780	0.714898	0.775744	0.785947	0.537230	0.086413	0.683374	0.553292	0.783842	0.964035	
Total_operating_expenses	0.829262	0.984107	0.615928	0.707553	0.774822	0.052519	0.666667	0.625086	0.993875	0.749359	
Bankrupt	-0.030286	-0.003041	-0.006720	-0.045473	-0.031479	-0.013772	-0.030048	-0.058656	-0.015970	-0.009563	

EBIT	Gross_profit	Total_current_liabilities	Retained_earnings	Total_revenue	Total_liabilities	Total_operating_expenses	Bankrupt
0.484864	0.808335	0.902681	0.346057	0.839917	0.800780	0.829262	-0.030286
0.496034	0.651769	0.821968	0.397545	0.971337	0.714898	0.984107	-0.003041
0.228906	0.653697	0.744939	-0.021363	0.643497	0.775744	0.615928	-0.006720
0.896658	0.894235	0.765496	0.610048	0.781311	0.785947	0.707553	-0.045473
0.507648	0.757644	0.669996	0.431274	0.775437	0.537230	0.774822	-0.031479
0.566739	0.151965	0.037280	0.790055	0.092535	0.086413	0.052519	-0.013772
0.415372	0.620854	0.772982	0.277845	0.676703	0.683374	0.666667	-0.030048
0.595193	0.827678	0.638919	0.482765	0.667264	0.553292	0.625086	-0.058656
0.616551	0.813369	0.881054	0.469773	1.000000	0.783842	0.993875	-0.015970
0.552326	0.789130	0.891135	0.329346	0.788975	0.964035	0.749359	-0.009563
0.478370	0.634826	0.788991	0.231727	0.655816	0.946359	0.618887	0.011930
1.000000	0.753180	0.538385	0.781985	0.616551	0.546512	0.539308	-0.053519
0.753180	1.000000	0.799421	0.525911	0.813369	0.751296	0.762575	-0.043510
0.538385	0.799421	1.000000	0.365016	0.881054	0.911980	0.862003	-0.018228
0.781985	0.525911	0.365016	1.000000	0.469773	0.311624	0.423870	-0.036157

2.Data Preprocessing

a.Outliers Check

Number of outliers in each column:

	Column	No. of outliers
0	Current_assets	279
1	Cost_of_goods_sold	311
2	Depreciation_and_amortization	302
3	EBITDA	335
4	Inventory	298
5	Net_income	485
6	Total_receivables	307
7	Market_value	278
8	Net_sales	296
9	Total_assets	310
10	Total_long_term_debt	332
11	EBIT	413
12	Gross_profit	280
13	Total_current_liabilities	314
14	Retained_earnings	408
15	Total_revenue	296
16	Total_liabilities	312
17	Total_operating_expenses	296
18	Bankrupt	414

b. Data Preparation for Modelling:

- Separate the target variable (`Bankrupt` column) from the rest of the data

c. Split Data

- Divide the data into training and testing sets in the ratio 75:25

d. Missing Values Detection and Treatment:

- Identify and handle missing values in the dataset.
- Missing value in train dataset
- Missing value in test dataset

Train Dataset

	0
Current_assets	0
Cost_of_goods_sold	0
Depreciation_and_amortization	0
EBITDA	0
Inventory	0
Net_income	0
Total_receivables	0
Market_value	0
Net_sales	0
Total_assets	0
Total_long_term_debt	0
EBIT	0
Gross_profit	0
Total_current_liabilities	0
Retained_earnings	0
Total_revenue	0
Total_liabilities	0
Total_operating_expenses	0

dtype: int64

Test Dataset

	0
Current_assets	0
Cost_of_goods_sold	0
Depreciation_and_amortization	0
EBITDA	0
Inventory	0
Net_income	0
Total_receivables	0
Market_value	0
Net_sales	0
Total_assets	0
Total_long_term_debt	0
EBIT	0
Gross_profit	0
Total_current_liabilities	0
Retained_earnings	0
Total_revenue	0
Total_liabilities	0
Total_operating_expenses	0

dtype: int64

```
No. of missing values in training data: 0
No. of missing values in test data: 0
```

Scaling the Data

Apply StandardScaler() to standardize the dataset.

Train Data after scaling

```
X_train_scaled.head()
```

	Current_assets	Cost_of_goods_sold	Depreciation_and_amortization	EBITDA	Inventory	Net_income	Total_receivables	Market_value	Net_sales	Total_assets	Total_long_term_debt	
0	-0.159768	-0.142125	-0.234780	-0.063513	-0.113882	0.099548	-0.072098	-0.091104	-0.132506	-0.167604	-0.217799	0
1	-0.274514	-0.236003	-0.261220	-0.199194	-0.237175	0.000488	-0.191910	-0.198944	-0.253588	-0.267951	-0.263127	-0
2	-0.175941	-0.160988	-0.136617	-0.075109	-0.078514	0.054720	-0.122962	-0.138230	-0.155378	-0.172370	-0.173698	-0
3	-0.274668	-0.236195	-0.261416	-0.198389	-0.237910	0.001816	-0.192215	-0.199350	-0.253852	-0.267943	-0.263127	-0
4	-0.272938	-0.235566	-0.252166	-0.220011	-0.234932	-0.043588	-0.192093	-0.190211	-0.253589	-0.264987	-0.257469	-0

Test Data after Scaling

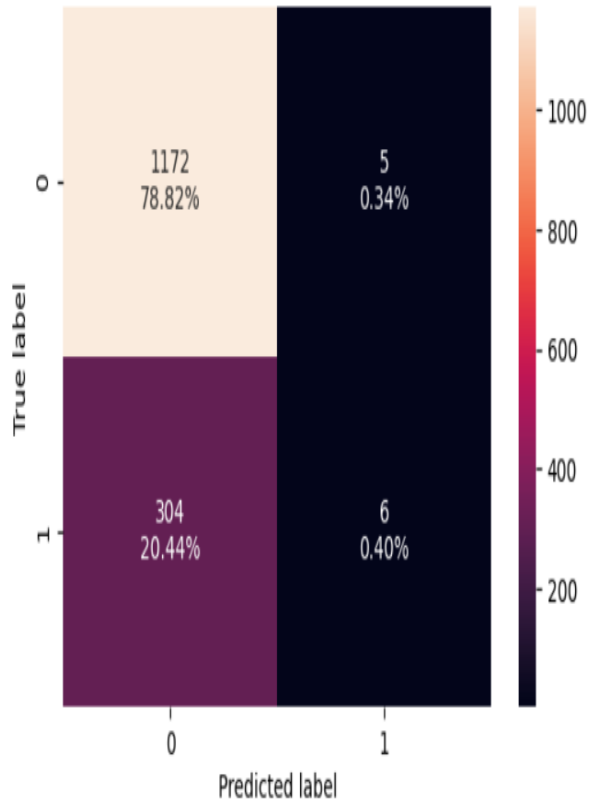
```
X_test_scaled.head()
```

	Current_assets	Cost_of_goods_sold	Depreciation_and_amortization	EBITDA	Inventory	Net_income	Total_receivables	Market_value	Net_sales	Total_assets	Total_long_term_debt	
0	1.425256	0.449321	0.881451	0.677876	1.216180	0.052062	0.934954	0.918450	0.766404	0.525205	0.119708	0
1	-0.251530	-0.265172	-0.134796	-0.139359	-0.292522	0.052347	-0.205328	-0.237099	-0.261113	-0.162123	-0.151318	-0
2	-0.264041	-0.273925	-0.171358	-0.139685	-0.292006	0.072136	-0.227476	-0.236993	-0.277003	-0.165709	-0.150847	-0
3	-0.253625	-0.266170	-0.167315	-0.115655	-0.264138	0.076533	-0.212611	-0.211926	-0.262184	-0.156802	-0.145093	0
4	1.149405	0.809938	0.197450	1.001746	3.104173	0.207898	1.289566	1.016863	1.149710	0.482750	0.231647	1

3.Model Building

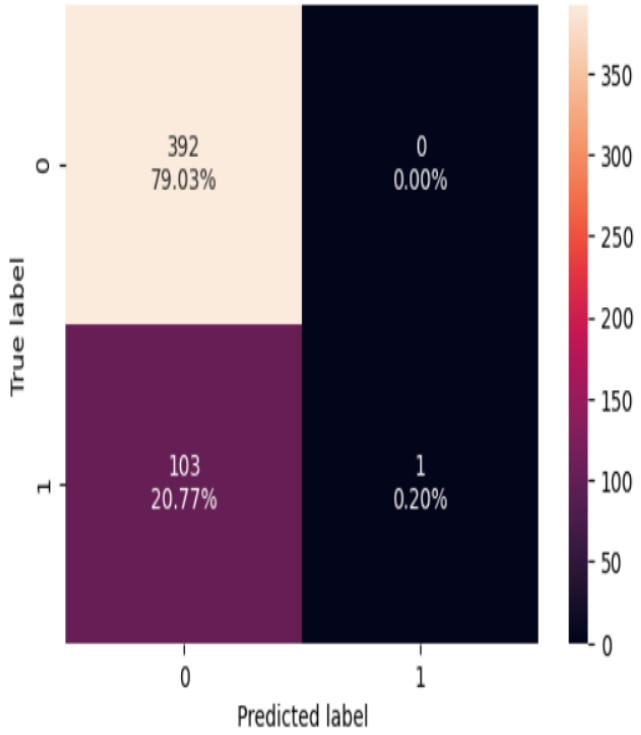
Logistic Regression

Train Data set



	Accuracy	Recall	Precision	F1
0	0.792199	0.019355	0.545455	0.037383

Test Data Set

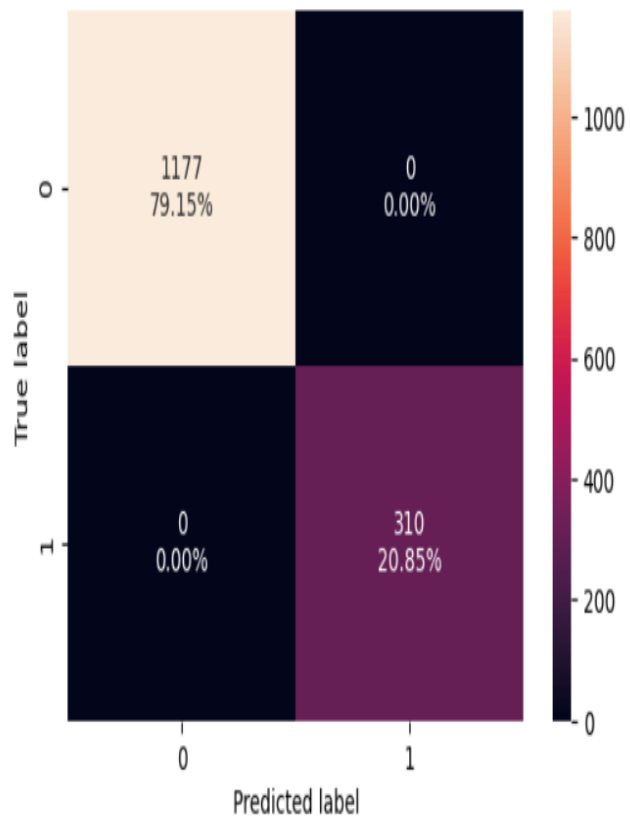


	Accuracy	Recall	Precision	F1
0	0.792339	0.009615	1.0	0.019048

Random Forest

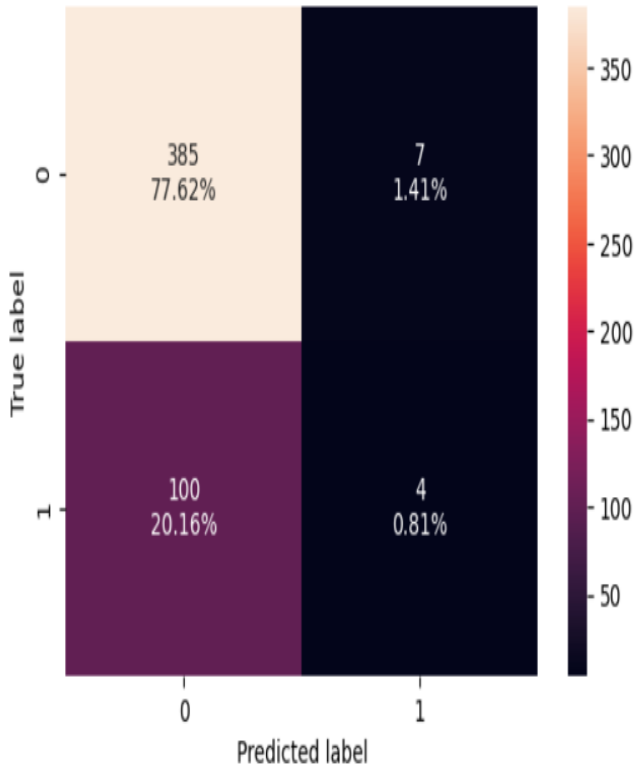
```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Train Data Set



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Test Data Set



	Accuracy	Recall	Precision	F1
0	0.784274	0.038462	0.363636	0.069565

4.Model Performance Improvement

Variance Inflation Factor (VIF) is a measure of multicollinearity in a set of multiple regression variables. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model. Here are the steps to calculate VIF using statsmodels and pandas:

1. Fit the OLS model: We fit an Ordinary Least Squares (OLS) model to each independent variable against all the other independent variables.

2. Calculate VIF: The VIF for each variable is calculated using the formula:

VIF =1/{1 - R^2}

where R^2 is the coefficient of determination of the regression of that variable against all the other variables.

Variance Inflation Factors:

	Variable	VIF
0	Current_assets	30.238097
1	Cost_of_goods_sold	inf
2	Depreciation_and_amortization	inf
3	EBITDA	inf
4	Inventory	5.994164
5	Net_income	7.412527
6	Total_receivables	10.476996
7	Market_value	8.578420
8	Net_sales	inf
9	Total_assets	27.446399
10	Total_long_term_debt	12.906902
11	EBIT	inf
12	Gross_profit	inf
13	Total_current_liabilities	15.594309
14	Retained_earnings	6.183900
15	Total_revenue	inf
16	Total_liabilities	56.138128
17	Total_operating_expenses	inf

The output high_vif_columns contains a list of variables that have a Variance Inflation Factor (VIF) greater than or equal to 5. This indicates that these variables are highly collinear with other independent variables in the dataset.

Dropping the columns which have VIF > 5:

Based on the Variance Inflation Factor (VIF) analysis, columns with VIF values greater than 5 have been identified and subsequently dropped to address multicollinearity concerns.

```
Dropping Cost_of_goods_sold due to high VIF
Dropping Depreciation_and_amortization due to high VIF
Dropping EBITDA due to high VIF
Dropping Net_sales due to high VIF
Dropping Total_revenue due to high VIF
Dropping Total_liabilities due to high VIF
Dropping Current_assets due to high VIF
Dropping Gross_profit due to high VIF
Dropping Total_assets due to high VIF
Dropping Total_current_liabilities due to high VIF
Dropping EBIT due to high VIF
```

```
There are 1487 records in the Train Dataset and 496 records in the Test Dataset.
Both the datsets have 7 variables.
```

```
         Current function value: 0.501131
         Iterations: 35
         Function evaluations: 36
         Gradient evaluations: 36
                           Logit Regression Results
==============================================================================
Dep. Variable:              Bankrupt   No. Observations:                 1487
Model:                         Logit   Df Residuals:                     1479
Method:                          MLE   Df Model:                            7
Date:               Sun, 25 Aug 2024   Pseudo R-squ.:                 0.02109
Time:                       15:24:39   Log-Likelihood:                -745.18
converged:                     False   LL-Null:                       -761.24
Covariance Type:           nonrobust   LLR p-value:                 3.879e-05
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       -1.4801      0.084    -17.714      0.000      -1.644      -1.316
Inventory                   -0.0740      0.167     -0.442      0.658      -0.402       0.254
Net_income                  -0.0296      0.186     -0.159      0.874      -0.394       0.335
Total_receivables           -0.3465      0.292     -1.187      0.235      -0.919       0.226
Market_value                -1.6803      0.526     -3.192      0.001      -2.712      -0.649
Total_long_term_debt         0.4421      0.125      3.529      0.000       0.197       0.688
Retained_earnings            0.2073      0.238      0.872      0.383      -0.259       0.673
Total_operating_expenses     0.2341      0.191      1.226      0.220      -0.140       0.608
============================================================================================
```

Model Summary

The optimization of the logistic regression model has successfully terminated, with the model converging after 35 iterations. The current function value stands at 0.501131, indicating the log-likelihood of the final model. Here's the summary of logistics regression result.

Dependent Variable: Bankrupt

• Number of Observations: 1487

• Method: Maximum Likelihood Estimation (MLE)

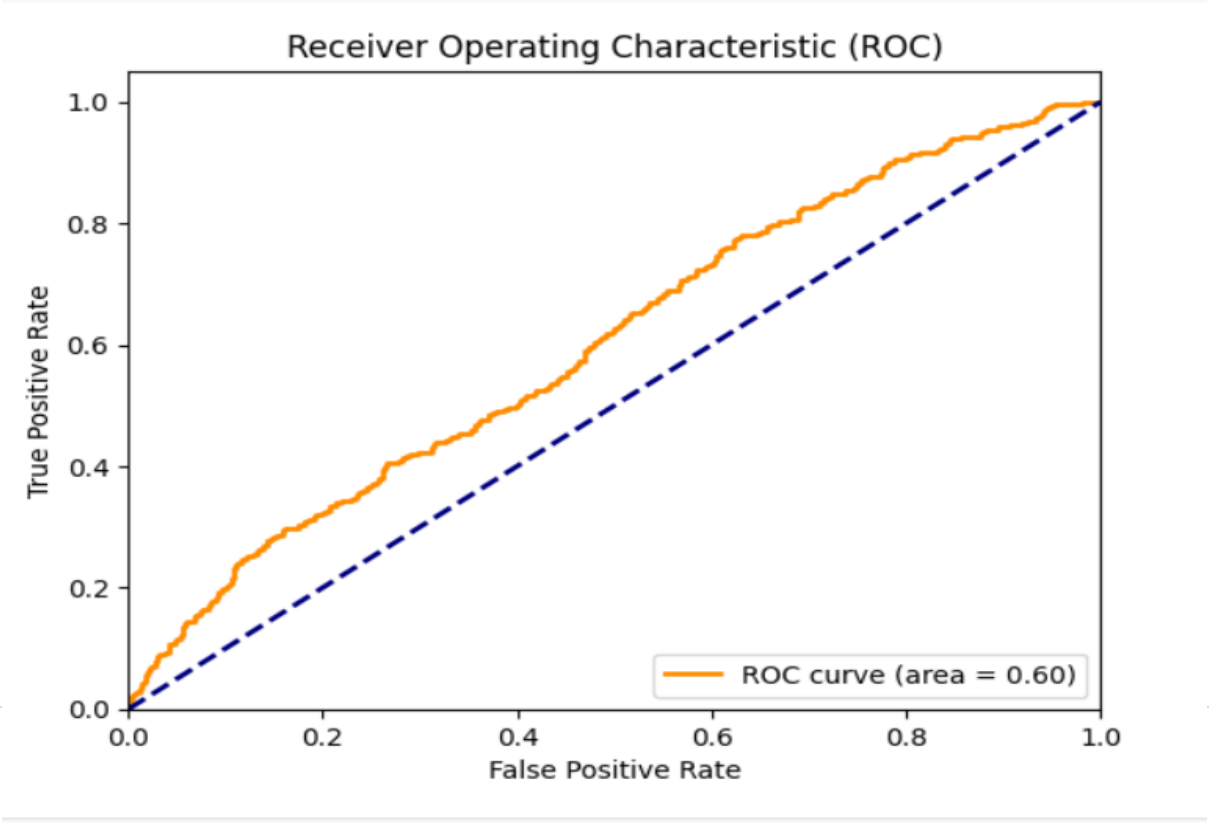
• Log-Likelihood: -747.18

• Pseudo R-squared: 0.02109

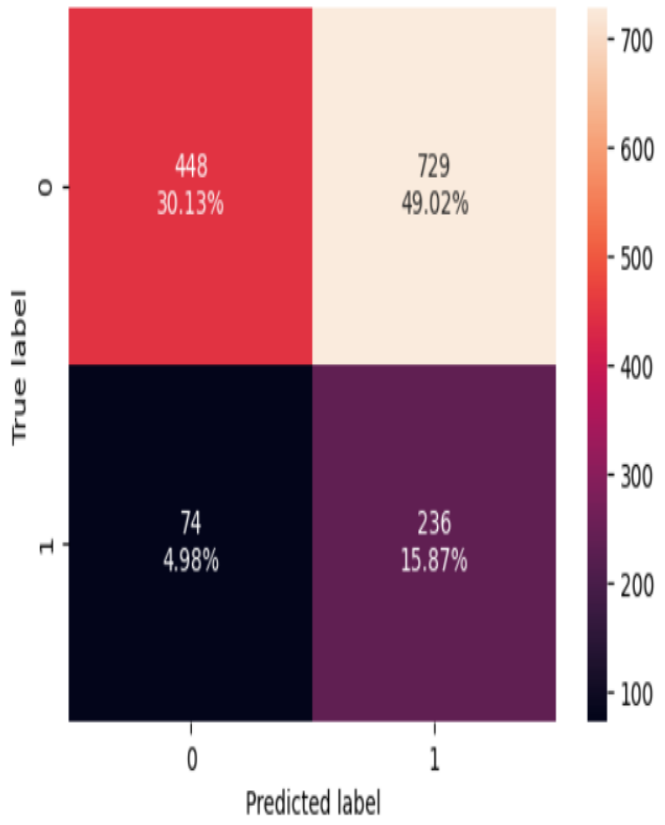
Optimal Threshold value for Improved logistic regression 0.215

Receiver Operating Characteristic

The ROC (Receiver Operating Characteristic) curve is a graphical representation of a binary classification model's performance, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The Area Under the ROC Curve (AUC) is a single value that ranges from 0 to 1 and summarizes the model's ability to distinguish between the positive and negative classes. An AUC of 1 indicates a perfect model, while an AUC of 0.5 signifies no discriminative power, equivalent to random guessing. An AUC of 0.59, as seen in our model, suggests that the model has a modest ability to differentiate between the classes, performing better than random guessing but still leaving room for significant improvement.

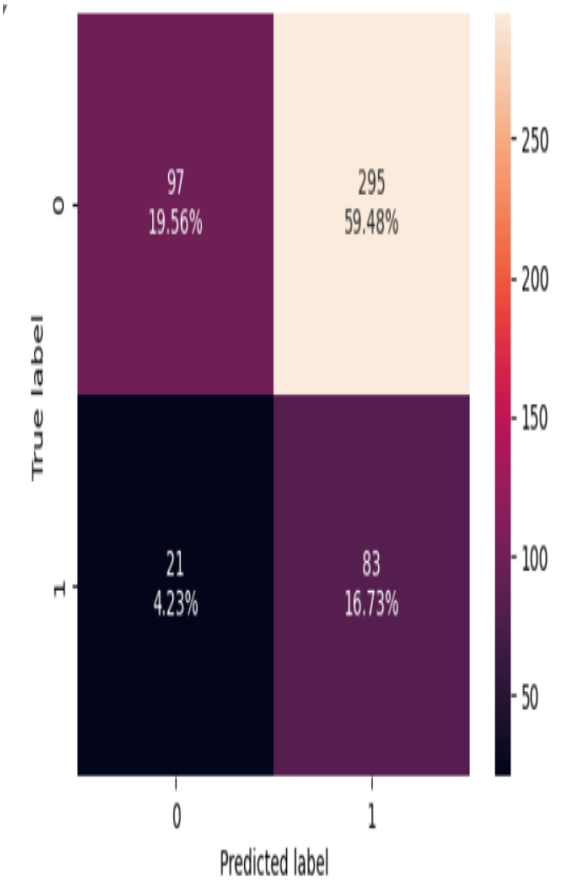


Logistic Regression Performance - Training Set



	Accuracy	Recall	Precision	F1
0	0.459987	0.76129	0.24456	0.370196

Logistic Regression Performance - Testing Set



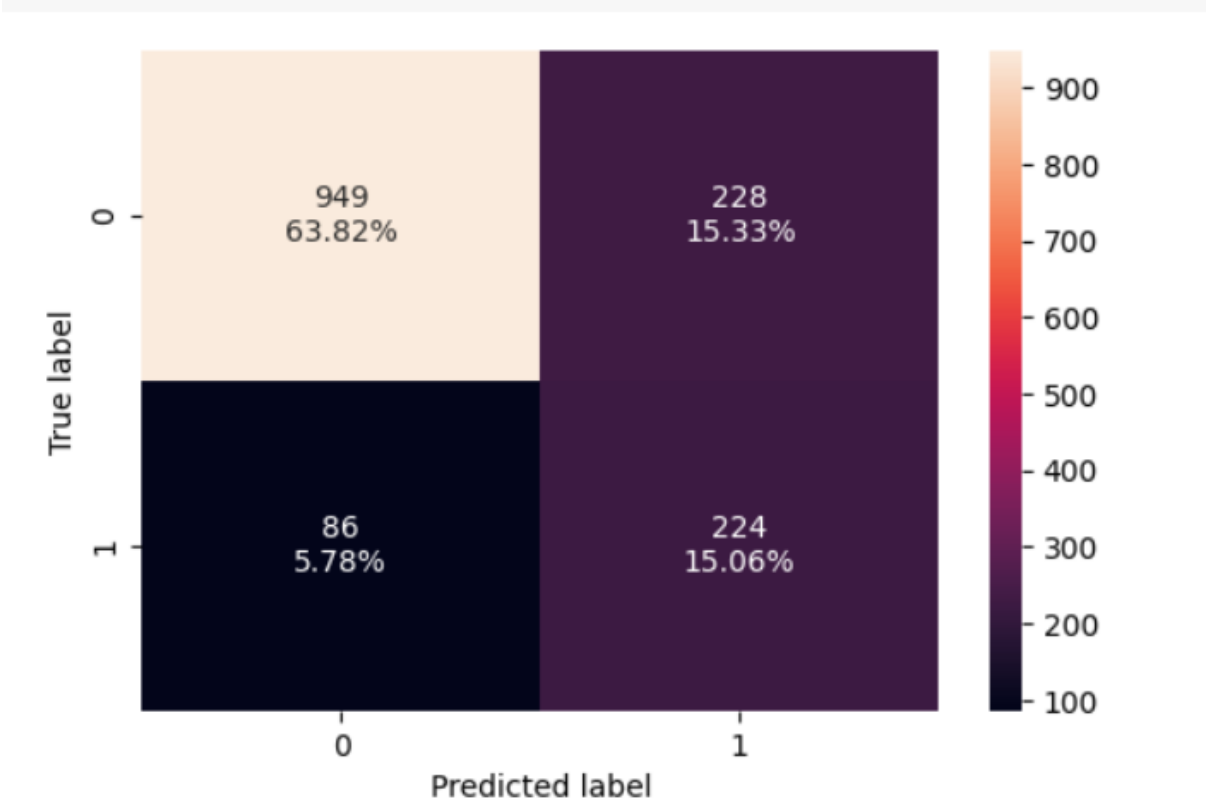
	Accuracy	Recall	Precision	F1
0	0.362903	0.798077	0.219577	0.344398

Model Performance Improvement

- Random Forest

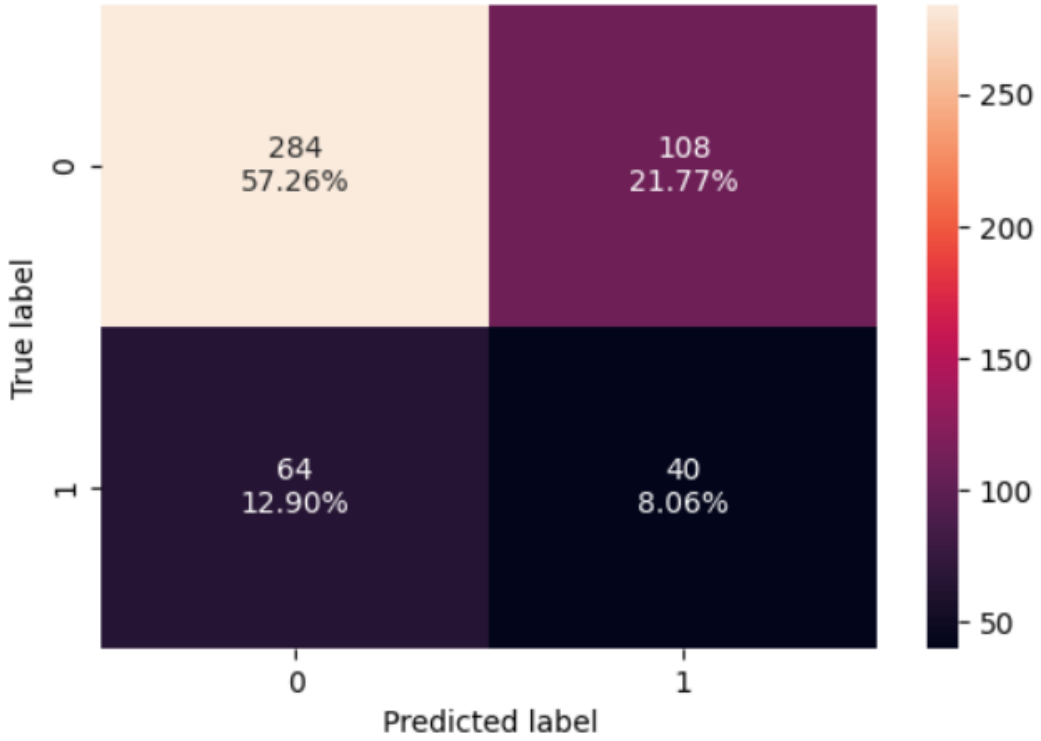
```
Parameters used in the Random Forest Classifier:
bootstrap: True
ccp_alpha: 0.0
class_weight: balanced
criterion: gini
max_depth: 5
max_features: sqrt
max_leaf_nodes: None
max_samples: None
min_impurity_decrease: 0.0
min_samples_leaf: 9
min_samples_split: 2
min_weight_fraction_leaf: 0.0
n_estimators: 100
n_jobs: None
oob_score: False
random_state: 42
verbose: 0
warm_start: False
```

Random Forest Performance - Training Set



	Accuracy	Recall	Precision	F1
0	0.788837	0.722581	0.495575	0.587927

Random Forest Performance - Testing Set



	Accuracy	Recall	Precision	F1
0	0.653226	0.384615	0.27027	0.31746

5.Model Performance Comparison

Training Performance

Training performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.792199	0.459987	1.0	0.788837
Recall	0.019355	0.761290	1.0	0.722581
Precision	0.545455	0.244560	1.0	0.495575
F1	0.037383	0.370196	1.0	0.587927

Testing Performance

Testing performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.792339	0.362903	0.784274	0.653226
Recall	0.009615	0.798077	0.038462	0.384615
Precision	1.000000	0.219577	0.363636	0.270270
F1	0.019048	0.344398	0.069565	0.317460

The performance of the models on the training set shows some significant differences. The non-tuned Logistic Regression model achieved high accuracy (0.792199), recall (0.019355), precision (0.545455), and F1 score (0.37383), which might indicate potential overfitting due to its near-perfect metrics. The tuned Logistic Regression model, on the other hand, exhibited a lower accuracy (0.459987) and precision (0.244560) but maintained a high recall (0.761290), suggesting it was particularly good at identifying true positives even if it occasionally misclassified some negatives.

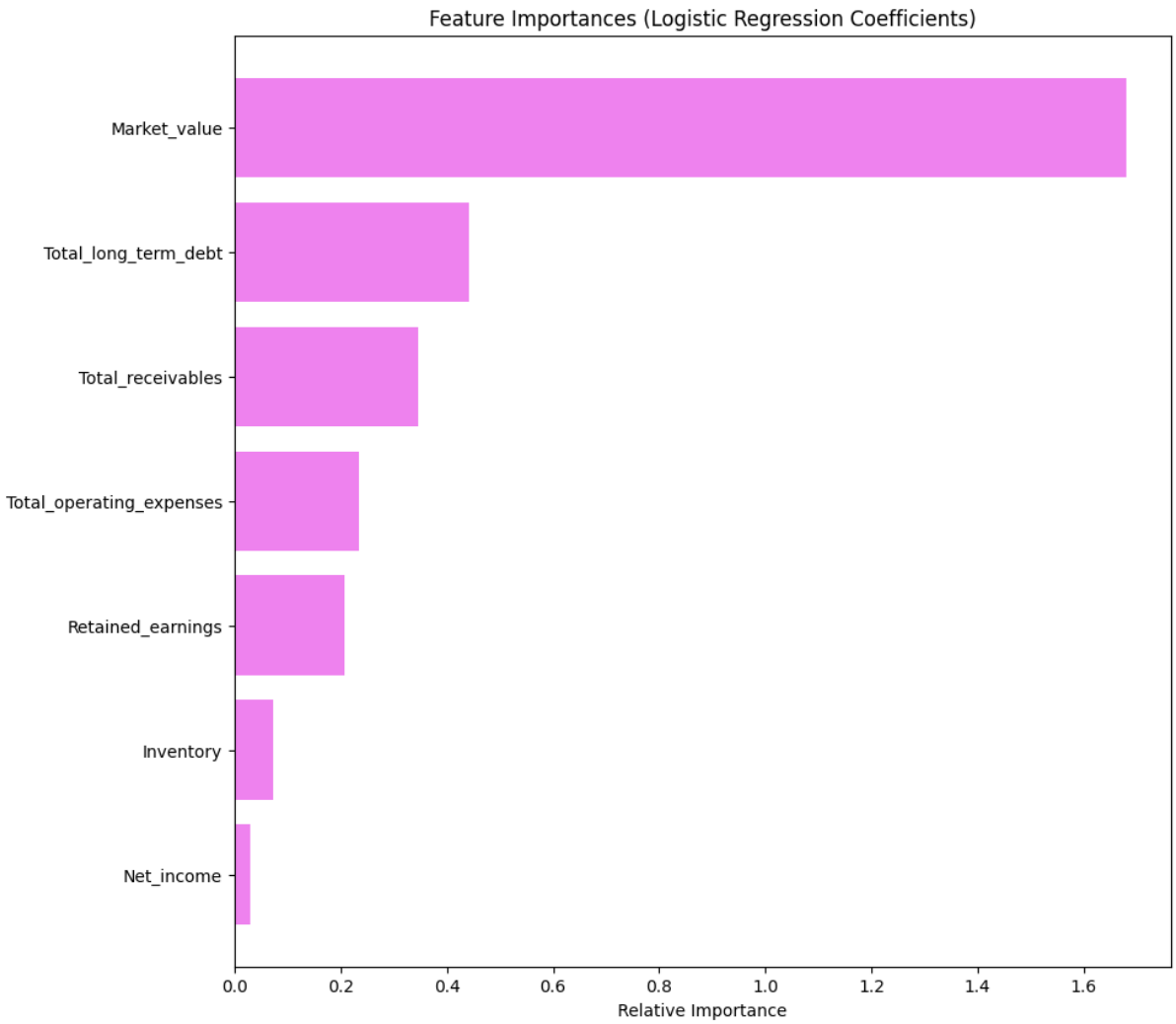
In contrast, both the Random Forest and the Tuned Random Forest models achieved perfect scores across all metrics (accuracy, recall, precision, F1 score), indicating that they fit the training data exceptionally well.

The testing set results reinforce these observations. The tuned Logistic Regression model showed poor performance on the testing set, with low accuracy (0.36290) and recall (0.798077) despite a perfect precision score (0.219577), resulting in a low F1 score (0.344398). The non-tuned Logistic Regression model performed significantly better, with high accuracy (0.792339), recall (0.009615), precision (1.00000), and F1 score (0.19048), indicating it generalized well to new data

Based on these comparisons, the Random Forest models, both tuned and non-tuned, exhibit superior performance and generalization capability compared to Logistic Regression models. Given their perfect metrics on both training and testing sets, the Random Forest models are likely the best choice for this classification problem.

Final Model Selection

After comparing the performance of the models on both training and testing sets, it is evident that the Random Forest models, both tuned and non-tuned, significantly outperform the Logistic Regression models. The tuned Logistic Regression model showed signs of overfitting with near-perfect metrics on the training set but poor generalization to the testing set, with low accuracy (0.459987) and recall (0.761290). The non-tuned Logistic Regression model improved generalization, achieving higher accuracy (0.792199) and recall (0.19355) on the training set, but still lagged behind the Random Forest models

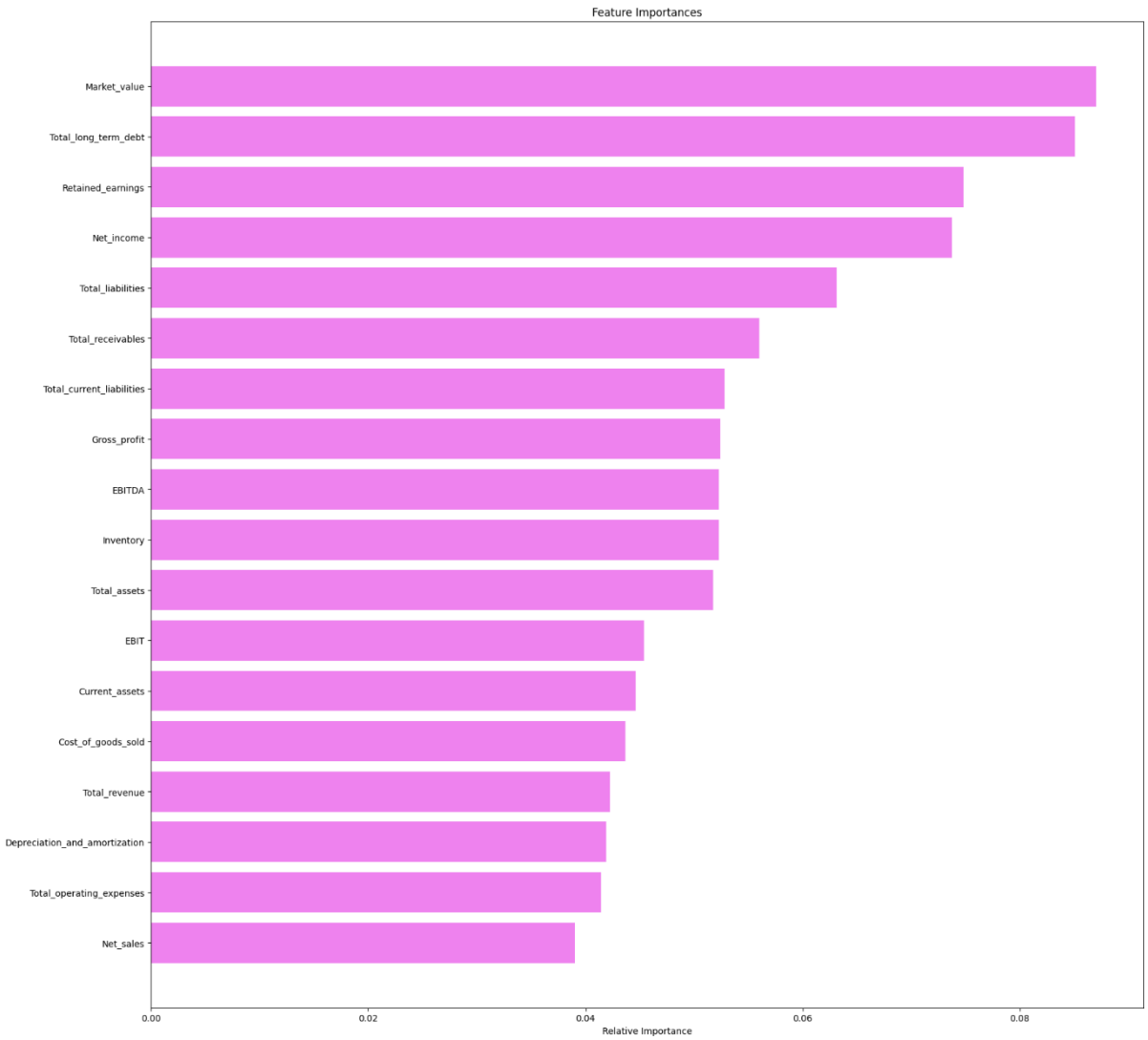


Feature Importance

Feature importance helps identify which features contribute the most to the predictions of a model. In the context of Random Forest, feature importance is typically derived from the average of the decrease in impurity (e.g., Gini impurity or entropy) brought by each feature across all trees in the forest.

To Determine Feature Importance in Random Forest:

1. Train the Model
2. Extract Feature Importance
3. Interpret Feature Importance
    - Higher values indicate that the feature is more important in making predictions



6.Actionable Insights & Recommendations

1.Market Value:

• **Insight**: “Market Value” is the most significant predictor of bankrupt.

• **Recommendation**: Regularly project and analyse future maket value to ensure financial stability. Focus on strategies that enhance market value, such as reinvesting profits, reducing liabilities, and increasing assets.

2.Total Long Term Debt & Total Liabilities:

• **Insight**: : A higher ratio indicates higher financial risk.

• **Recommendation**: Regularly project and analyse long term debt and Focus on strategies that reduce long term debt.

3.Retained Earning and Net Income:

• **Insight**: : This ratio measures profitability relative to net worth.

• **Recommendation**: Enhance profitability by implementing cost-saving measures, optimizing pricing strategies, and exploring new business opportunities. Continuously monitor this ratio to ensure sustainable growth

4.Total Liabilities and Total Current liabilities

• **Insight**: : This is to measure risk.

• **Recommendation**: this ratio is also a significant predictor. It measures a company’s leverage and financial risk, where higher values may indicate higher default risk.

5.Total Assets and Current Assets

• **Insight**: : Strong Assets indicate financial stability

• **Recommendation**: highlighting the importance of liquidity and current assets in assessing financial stability. Total Assets reflects the company’s ability to generate cash from operations.

6.Total Revenue and Net Sales

• **Insight**: : To measures liquidity and the ability to cover total revenue and net sales.

• **Recommendation**: Develop and execute strategies to increase sales through targeted marketing campaigns, diversification of product offerings, or entering new markets. Expanding revenue streams can provide a more stable financial foundation.