3/3/2024

Credit Risk Default Prediction Report

PGP-DSBA

Karthick Raj S

Table of Contents

List of Tables

List of Figures

Problem:

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interest on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Data that is available includes information from the financial statement of the companies for the previous year.

Dependent variable - No need to create any new variable, as the 'Default' variable is already provided in the dataset, which can be considered as the dependent variable.

Test Train Split - Split the data into train and test datasets in the ratio of 67:33 and use a random state of 42 (*random_state=42*). Model building is to be done on the train dataset and model validation is to be done on the test dataset.

Dataset: Credit Risk Dataset

Data Dictionary: Data Dictionary

Data Description:

S. No	Column Name	Description
1	Co_Code	Company Code
2	Co_Name	Company Name
3	_Operating_Expense_Rate	Operating Expense Rate: Operating Expenses/Net Sales. The operating expense ratio (OER) is the cost to operate a piece of property compared to the income the property brings in.
4	_Research_and_development_expense_rate	Research and development expense rate: (Research and Development Expenses)/Net Sales. Research and development (R&D) expenses are direct expenditures relating to a company's efforts to develop, design, and enhance its products, services, technologies, or processes.
5	_Cash_flow_rate	Cash flow rate: Cash Flow from Operating/Current Liabilities. Cash flow is a measure of how much cash a business brought in or spent in total over a period of time.
6	_Interest_bearing_debt_interest_rate	Interest-bearing debt interest rate: Interest-bearing Debt/Equity
7	_Tax_rate_A	Tax rate (A): Effective Tax Rate. Effective tax rate represents the percentage of their taxable income that individuals pay in taxes. For corporations, the effective corporate tax rate is the rate they pay on their pre-tax profits.
8	_Cash_Flow_Per_Share	Cash Flow Per Share. It is the after-tax earnings plus depreciation on a per-share basis that functions as a measure of a firm's financial strength
9	_Per_Share_Net_profit_before_tax_Yuan_	Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share. Pretax income, also known as earnings before tax or pretax earnings, is the net income earned by a business before taxes are subtracted/accounted for.
10	_Realized_Sales_Gross_Profit_Growth_Rate	Realized Sales Gross Profit Growth Rate.
11	_Operating_Profit_Growth_Rate	Operating Profit Growth Rate: Operating Income Growth. It is the rate of increase in operating income over the last year.
12	_Continuous_Net_Profit_Growth_Rate	Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
13	_Total_Asset_Growth_Rate	Total Asset Growth Rate: Total Asset Growth. It is the rate at which how quickly the company has been growing its Assets
14	_Net_Value_Growth_Rate	Net Value Growth Rate: Total Equity Growth

15	_Total_Asset_Return_Growth_Rate_Ratio	Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
16	_Cash_Reinvestment_perc	Cash Reinvestment %: Cash Reinvestment Ratio. It is the valuation ratio that is used to measure the percentage of annual cash flow that the company invests back into the business as a new investment.
17	_Current_Ratio	Current Ratio. The current ratio describes the relationship between a company's assets and liabilities
18	_Quick_Ratio	Quick Ratio: Acid Test. Acid-test ratio (also known as quick ratio) is a measure of a company's liquidity, which is its ability to pay its short-term obligations using only its most liquid assets.
19	_Interest_Expense_Ratio	Interest Expense Ratio: Interest Expenses/Total Revenue
20	_Total_debt_to_Total_net_worth	Total debt/Total net worth: Total Liability/Equity Ratio
21	_Long_term_fund_suitability_ratio_A	Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
22	_Net_profit_before_tax_to_Paid_in_capital	Net profit before tax/Paid-in capital: Pretax Income/Capital
23	_Total_Asset_Turnover	Total Asset Turnover. Net Sales/Average Total Assets
24	_Accounts_Receivable_Turnover	Accounts Receivable Turnover. The accounts receivable turnover ratio, or receivables turnover, is used in business accounting to quantify how well companies are managing the credit that they extend to their customers by evaluating how long it takes to collect the outstanding debt throughout the accounting period.
25	_Average_Collection_Days	Average Collection Days: Days Receivable Outstanding
26	_Inventory_Turnover_Rate_times	Inventory Turnover Rate (times). The inventory turnover ratio is the number of times a company has sold and replenished its inventory over a specific amount of time. The formula can also be used to calculate the number of days it will take to sell the inventory on hand.
27	_Fixed_Assets_Turnover_Frequency	Fixed Assets Turnover Frequency. Fixed Asset Turnover (FAT) is an efficiency ratio that indicates how well or efficiently a business uses fixed assets to generate sales. This ratio divides net sales by net fixed assets, calculated over an annual period.
28	_Net_Worth_Turnover_Rate_times	Net Worth Turnover Rate (times): Equity Turnover. Equity turnover is a ratio that measures

		the proportion of a company's sales to its stockholders' equity. The intent of the measurement is to determine the efficiency with which management is using equity to generate revenue.
29	_Operating_profit_per_person	Operating profit per person: Operation Income Per Employee
30	_Allocation_rate_per_person	Allocation rate per person: Fixed Assets Per Employee
31	_Quick_Assets_to_Total_Assets	Quick Assets/Total Assets
32	_Cash_to_Total_Assets	Cash/Total Assets
33	_Quick_Assets_to_Current_Liability	Quick Assets/Current Liability
34	_Cash_to_Current_Liability	Cash/Current Liability
35	_Operating_Funds_to_Liability	Operating Funds to Liability
36	_Inventory_to_Working_Capital	Inventory/Working Capital
37	_Inventory_to_Current_Liability	Inventory/Current Liability
38	_Long_term_Liability_to_Current_Assets	Long-term Liability to Current Assets
39	_Retained_Earnings_to_Total_Assets	Retained Earnings to Total Assets
40	_Total_income_to_Total_expense	Total income/Total expense
41	_Total_expense_to_Assets	Total expense/Assets
42	_Current_Asset_Turnover_Rate	Current Asset Turnover Rate: Current Assets to Sales. The current assets turnover ratio indicates how many times the current assets are turned over in the form of sales within a specific period of time. A higher asset turnover ratio means a better percentage of sales.
43	_Quick_Asset_Turnover_Rate	Quick Asset Turnover Rate: Quick Assets to Sales. The asset turnover ratio measures the efficiency of a company's assets in generating revenue or sales.
44	_Cash_Turnover_Rate	Cash Turnover Rate: Cash to Sales. The cash turnover ratio is an efficiency ratio that reveals the number of times that cash is turned over in an accounting period.
45	_Fixed_Assets_to_Assets	Fixed Assets to Assets. Fixed assets are also known as non-current assets—assets that can't be easily converted into cash.
46	_Cash_Flow_to_Total_Assets	Cash Flow to Total Assets. This ratio indicates the cash a company can generate in relation to its size.
47	_Cash_Flow_to_Liability	Cash Flow to Liability. The amount of money available to run business operations and complete transactions. This is calculated as current assets (cash or near-cash assets, like notes receivable) minus current liabilities

		(liabilities due during the upcoming accounting period)
48	_CFO_to_Assets	CFO to Assets. Cash flow on total assets is an efficiency ratio that rates cash flows to the company assets without being affected by income recognition or income measurements.
49	_Cash_Flow_to_Equity	Cash Flow to Equity. cash flow to equity is a measure of how much cash is available to the equity shareholders of a company after all expenses, reinvestment, and debt are paid.
50	_Current_Liability_to_Current_Assets	Current Liability to Current Assets. Current liabilities are a company's financial commitments that are due and payable within a year, Current assets are projected to be consumed, sold, or converted into cash within a year or within the operational cycle.
51	_Liability_Assets_Flag	Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
52	_Total_assets_to_GNP_price	Total assets to GNP price. Gross National Product (GNP) is the total value of all finished goods and services produced by a country's citizens in a given financial year, irrespective of their location.
53	_No_credit_Interval	No-credit Interval
54	_Degree_of_Financial_Leverage_DFL	Degree of Financial Leverage (DFL). The degree of financial leverage is a financial ratio that measures the sensitivity in fluctuations of a company's overall profitability to the volatility of its operating income caused by changes in its capital structure.
55	_Interest_Coverage_Ratio_Interest_expense_to_EBIT	Interest Coverage Ratio (Interest expense to EBIT). The interest coverage ratio is a debt and profitability ratio used to determine how easily a company can pay interest on its outstanding debt. The interest coverage ratio is calculated by dividing a company's earnings before interest and taxes (EBIT) by its interest expense during a given period.
56	_Net_Income_Flag	Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
57	_Equity_to_Liability	Equity to Liability Ratio.
58	Default	Whether the Company has Default (Bankrupted) or not? 1 - Defaulted, 0 - Not Defaulted.

Table 1 Data Description

There are 2058 rows and 58 columns in the dataset.

Dataset Info:

Most of the columns are in float64, Company name and Code are in Object and int64 respectively.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 58 columns):
 #   Column                                             Non-Null Count  Dtype
---  ------                                             --------------  -----
 0   Co_Code                                            2058 non-null   int64
 1   Co_Name                                            2058 non-null   object
 2   _Operating_Expense_Rate                            2058 non-null   float64
 3   _Research_and_development_expense_rate             2058 non-null   float64
 4   _Cash_flow_rate                                    2058 non-null   float64
 5   _Interest_bearing_debt_interest_rate               2058 non-null   float64
 6   _Tax_rate_A                                        2058 non-null   float64
 7   _Cash_Flow_Per_Share                               1891 non-null   float64
 8   _Per_Share_Net_profit_before_tax_Yuan_             2058 non-null   float64
 9   _Realized_Sales_Gross_Profit_Growth_Rate           2058 non-null   float64
 10  _Operating_Profit_Growth_Rate                      2058 non-null   float64
 11  _Continuous_Net_Profit_Growth_Rate                 2058 non-null   float64
 12  _Total_Asset_Growth_Rate                           2058 non-null   float64
 13  _Net_Value_Growth_Rate                             2058 non-null   float64
 14  _Total_Asset_Return_Growth_Rate_Ratio              2058 non-null   float64
 15  _Cash_Reinvestment_perc                            2058 non-null   float64
 16  _Current_Ratio                                     2058 non-null   float64
 17  _Quick_Ratio                                       2058 non-null   float64
 18  _Interest_Expense_Ratio                            2058 non-null   float64
 19  _Total_debt_to_Total_net_worth                     2037 non-null   float64
 20  _Long_term_fund_suitability_ratio_A                2058 non-null   float64
 21  _Net_profit_before_tax_to_Paid_in_capital          2058 non-null   float64
 22  _Total_Asset_Turnover                              2058 non-null   float64
 23  _Accounts_Receivable_Turnover                      2058 non-null   float64
 24  _Average_Collection_Days                           2058 non-null   float64
 25  _Inventory_Turnover_Rate_times                     2058 non-null   float64
 26  _Fixed_Assets_Turnover_Frequency                   2058 non-null   float64
 27  _Net_Worth_Turnover_Rate_times                     2058 non-null   float64
 28  _Operating_profit_per_person                       2058 non-null   float64
 29  _Allocation_rate_per_person                        2058 non-null   float64
 30  _Quick_Assets_to_Total_Assets                      2058 non-null   float64
 31  _Cash_to_Total_Assets                              1962 non-null   float64
 32  _Quick_Assets_to_Current_Liability                 2058 non-null   float64
 33  _Cash_to_Current_Liability                         2058 non-null   float64
 34  _Operating_Funds_to_Liability                      2058 non-null   float64
 35  _Inventory_to_Working_Capital                      2058 non-null   float64
 36  _Inventory_to_Current_Liability                    2058 non-null   float64
 37  _Long_term_Liability_to_Current_Assets             2058 non-null   float64
 38  _Retained_Earnings_to_Total_Assets                 2058 non-null   float64
 39  _Total_income_to_Total_expense                     2058 non-null   float64
 40  _Total_expense_to_Assets                           2058 non-null   float64
 41  _Current_Asset_Turnover_Rate                       2058 non-null   float64
 42  _Quick_Asset_Turnover_Rate                         2058 non-null   float64
 43  _Cash_Turnover_Rate                                2058 non-null   float64
 44  _Fixed_Assets_to_Assets                            2058 non-null   float64
 45  _Cash_Flow_to_Total_Assets                         2058 non-null   float64
 46  _Cash_Flow_to_Liability                            2058 non-null   float64
 47  _CFO_to_Assets                                     2058 non-null   float64
 48  _Cash_Flow_to_Equity                               2058 non-null   float64
 49  _Current_Liability_to_Current_Assets               2044 non-null   float64
 50  _Liability_Assets_Flag                             2058 non-null   int64
 51  _Total_assets_to_GNP_price                         2058 non-null   float64
 52  _No_credit_Interval                                2058 non-null   float64
 53  _Degree_of_Financial_Leverage_DFL                    2058 non-null   float64
 54  _Interest_Coverage_Ratio_Interest_expense_to_EBIT  2058 non-null   float64
 55  _Net_Income_Flag                                     2058 non-null   int64
 56  _Equity_to_Liability                                 2058 non-null   float64
 57  Default                                              2058 non-null   int64
dtypes: float64(53), int64(4), object(1)
memory usage: 932.7+ KB
```

The Default, Liability Assets Flag and Net Income Flag columns are changed to categorical datatype.

Descriptive Statistics:

Variables	count	mean	std	min	25%	50%	75%	max
_Operating_Expense_Rate	2058	2.05E+09	3.25E+09	0	0	0	4.11E+09	9.98E+09
_Research_and_development_expense_rate	2058	1.21E+09	2.14E+09	0	0	0	1.55E+09	9.98E+09
_Cash_flow_rate	2058	0.47	0.02	0	0.46	0.46	0.47	1
_Interest_bearing_debt_interest_rate	2058	11130224	90425949	0	0	0	0	9.9E+08
_Tax_rate_A	2058	0.11	0.15	0	0	0.04	0.22	1
_Cash_Flow_Per_Share	1891	0.32	0.02	0.17	0.31	0.32	0.33	0.46
_Per_Share_Net_profit_before_tax_Yuan_	2058	0.18	0.03	0	0.17	0.18	0.19	0.79
_Realized_Sales_Gross_Profit_Growth_Rate	2058	0.02	0.02	0	0.02	0.02	0.02	1
_Operating_Profit_Growth_Rate	2058	0.85	0	0.74	0.85	0.85	0.85	1
_Continuous_Net_Profit_Growth_Rate	2058	0.22	0.01	0	0.22	0.22	0.22	0.23
_Total_Asset_Growth_Rate	2058	5.29E+09	2.91E+09	0	4.32E+09	6.23E+09	7.22E+09	9.98E+09
_Net_Value_Growth_Rate	2058	5189504	2.08E+08	0	0	0	0	9.33E+09
_Total_Asset_Return_Growth_Rate_Ratio	2058	0.26	0	0.25	0.26	0.26	0.26	0.36
_Cash_Reinvestment_perc	2058	0.38	0.03	0.03	0.37	0.38	0.39	1
_Current_Ratio	2058	1336249	60619173	0	0.01	0.01	0.01	2.75E+09
_Quick_Ratio	2058	27755102	4.45E+08	0	0	0.01	0.01	9.23E+09
_Interest_Expense_Ratio	2058	0.63	0.01	0.53	0.63	0.63	0.63	0.81
_Total_debt_to_Total_net_worth	2037	10714286	2.7E+08	0	0	0.01	0.01	9.94E+09
_Long_term_fund_suitability_ratio_A	2058	0.01	0.03	0	0.01	0.01	0.01	1
_Net_profit_before_tax_to_Paid_in_capital	2058	0.18	0.03	0	0.17	0.17	0.18	0.79
_Total_Asset_Turnover	2058	0.13	0.1	0	0.06	0.1	0.17	0.92
_Accounts_Receivable_Turnover	2058	41598639	5.05E+08	0	0	0	0	9.74E+09
_Average_Collection_Days	2058	26297862	4.11E+08	0	0	0.01	0.01	8.8E+09

_Inventory_Turnover_Rate_times	2058	2.03E+09	3.08E+09	0	0	19100000	3.82E+09	9.99E+09
_Fixed_Assets_Turnover_Frequency	2058	1.23E+09	2.65E+09	0	0	0	0.01	9.99E+09
_Net_Worth_Turnover_Rate_times	2058	0.04	0.04	0.01	0.02	0.03	0.04	1
_Operating_profit_per_person	2058	0.4	0.05	0	0.39	0.4	0.4	1
_Allocation_rate_per_person	2058	5725559	1.98E+08	0	0	0.01	0.02	8.28E+09
_Quick_Assets_to_Total_Assets	2058	0.34	0.21	0	0.17	0.31	0.48	0.99
_Cash_to_Total_Assets	1962	0.08	0.1	0	0.02	0.05	0.1	0.93
_Quick_Assets_to_Current_Liability	2058	11904762	3.12E+08	0	0	0.01	0.01	8.82E+09
_Cash_to_Current_Liability	2058	92825073	7.85E+08	0	0	0	0.01	9.17E+09
_Operating_Funds_to_Liability	2058	0.35	0.04	0.03	0.34	0.35	0.35	1
_Inventory_to_Working_Capital	2058	0.28	0.02	0	0.28	0.28	0.28	1
_Inventory_to_Current_Liability	2058	57863460	6.28E+08	0	0	0.01	0.01	9.6E+09
_Long_term_Liability_to_Current_Assets	2058	73401069	6.69E+08	0	0	0	0.01	9.31E+09
_Retained_Earnings_to_Total_Assets	2058	0.93	0.03	0	0.93	0.94	0.94	0.97
_Total_income_to_Total_expense	2058	0	0	0	0	0	0	0.01
_Total_expense_to_Assets	2058	0.03	0.04	0	0.01	0.02	0.04	1
_Current_Asset_Turnover_Rate	2058	1.27E+09	2.84E+09	0	0	0	0	9.99E+09
_Quick_Asset_Turnover_Rate	2058	2.57E+09	3.45E+09	0	0	0	5.79E+09	1E+10
_Cash_Turnover_Rate	2058	2.65E+09	2.82E+09	0	0	1.73E+09	4.55E+09	9.99E+09
_Fixed_Assets_to_Assets	2058	4042760	1.83E+08	0	0.1	0.21	0.42	8.32E+09
_Cash_Flow_to_Total_Assets	2058	0.64	0.05	0	0.63	0.64	0.65	1
_Cash_Flow_to_Liability	2058	0.46	0.03	0.03	0.46	0.46	0.46	0.91
_CFO_to_Assets	2058	0.58	0.06	0	0.55	0.58	0.61	0.98
_Cash_Flow_to_Equity	2058	0.31	0.01	0	0.31	0.31	0.32	0.57
_Current_Liability_to_Current_Assets	2044	0.04	0.05	0	0.02	0.03	0.04	1
_Liability_Assets_Flag	2058	NaN	NaN	NaN	NaN	NaN	NaN	NaN
_Total_assets_to_GNP_price	2058	27793975	4.72E+08	0	0	0	0.01	9.82E+09
_No_credit_Interval	2058	0.62	0.01	0.41	0.62	0.62	0.62	0.96
_Degree_of_Financial_Leverage_DFL	2058	0.03	0.01	0.01	0.03	0.03	0.03	0.46
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	2058	0.57	0.01	0.17	0.57	0.57	0.57	0.67
_Net_Income_Flag	2058	NaN	NaN	NaN	NaN	NaN	NaN	NaN
_Equity_to_Liability	2058	0.04	0.06	0	0.02	0.03	0.04	1
Default	2058	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2 Descriptive Statistics

The Company Name and Company Code are dropped as there is no significant information for the model.

The Net Income Flag has Unique Value with only one category and 99% of the Liability Assets Flag is also unique with one category.

These are also not bringing any significant information for the model. These two are dropped.

EDA

Count Plot:

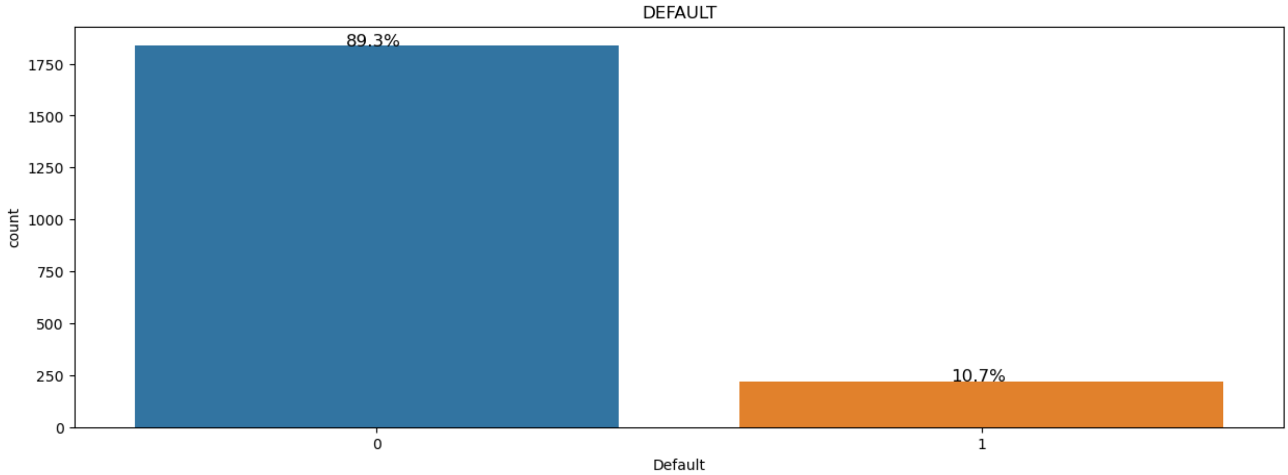


Figure A Count Plot - Default

There are 89.3% of Non-defaulters and only 10.7% of Defaulters in the dataset.

The dataset is not balanced.

Box Plot:

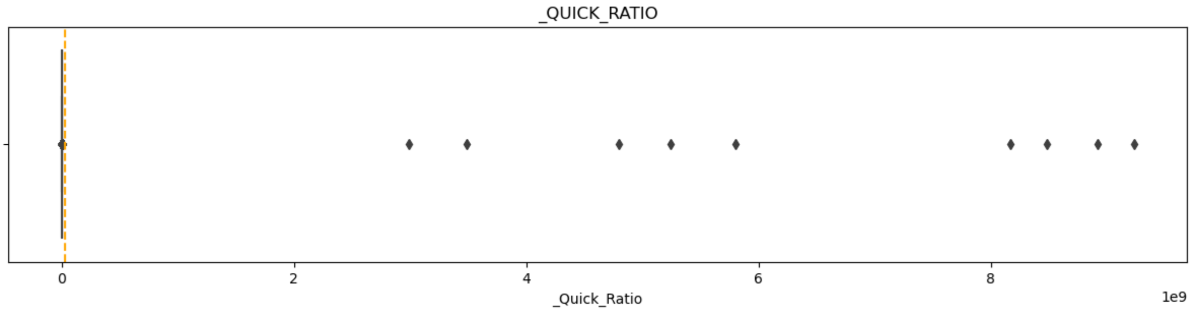


Figure B Boxplot - Quick ratio

The Quick Ratio is mostly 0.01 and only some companies are outliers which has quick ratio greater than 2.

Other than outliers 99% of the company don't have that much liquidity with them.

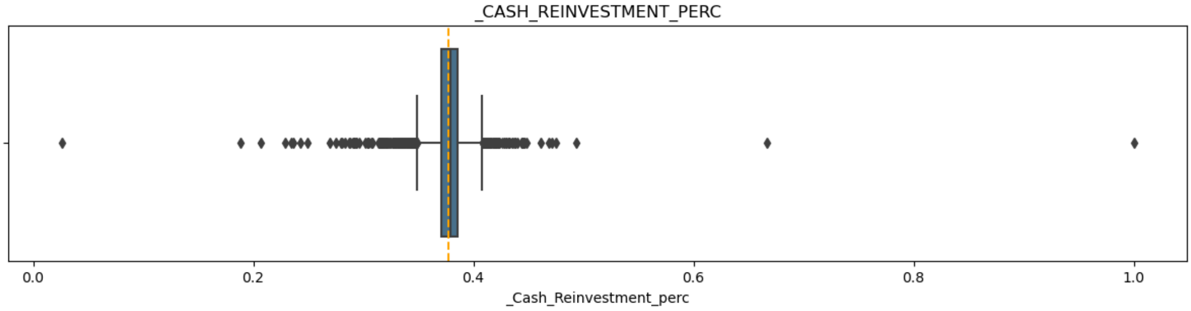


Figure C Boxplot - Cash Reinvestment %

But looking into the cash reinvestment percentage, 50% of the companies are Re investing around half of the cash reserves.

This is a good indicator even though they have less cash.

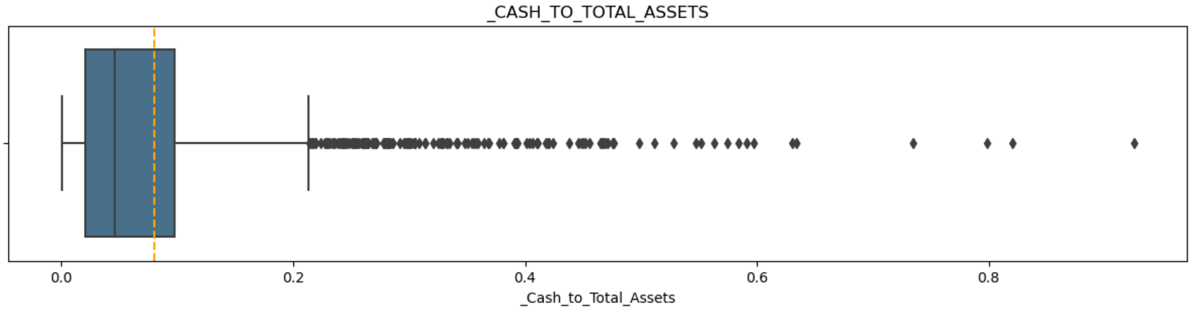


Figure D Boxplot - Cash to Total Assets

Because of Reinvestment, the cash to Total Asset is also low.

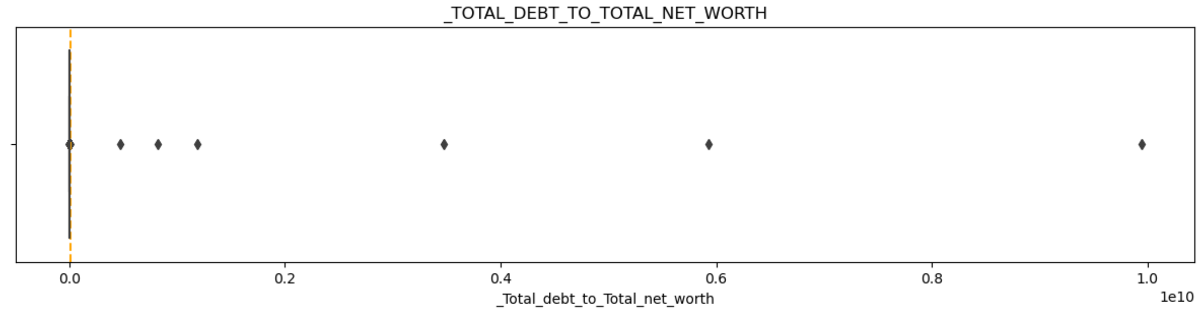


Figure E BoxPlot - Total Debt to Total Net Worth

Only few companies are having total debt to total net worth greater than 60%.

Those companies are mostly running on their debts.

This is a good indicator that most of the companies are having a ratio of 0.01 which means they are not solely doing business on the debts.

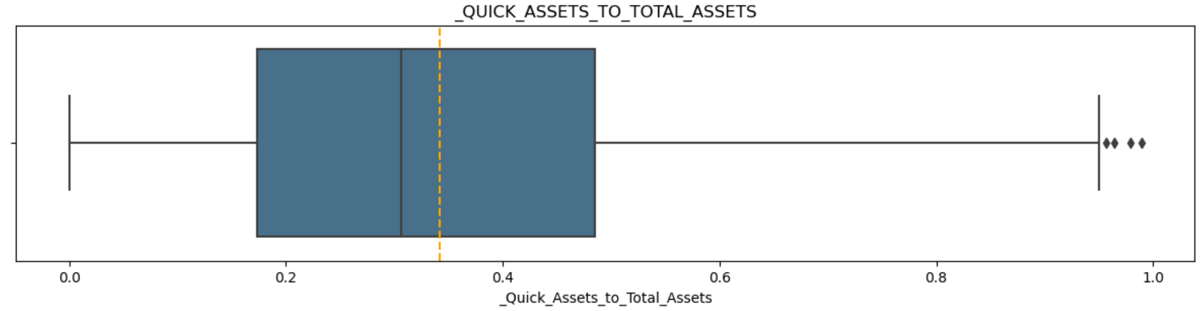


Figure F Boxplot - Quick Assets to Total Assets

Around 75% of the companies has 50% of their assets as current assets.

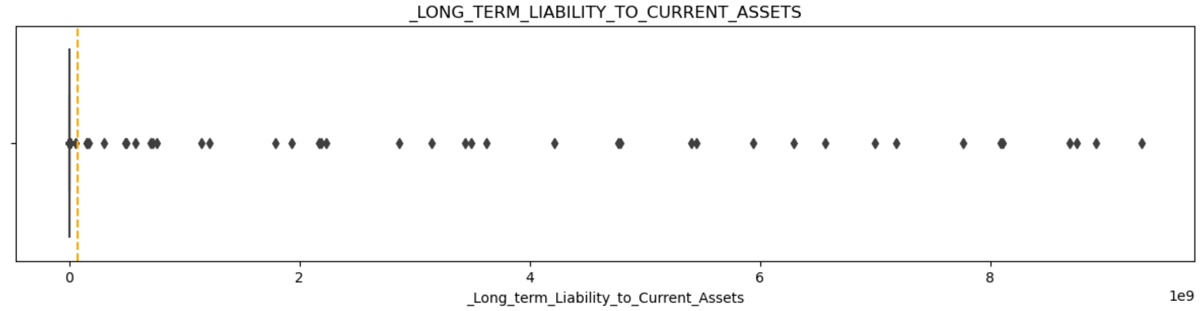


Figure G Boxplot - Long term liability to Current Assets

Even the companies goes into loss, most of the companies can payoff their liabilities with their current assets.

Some of the companies which has long term liability to current assets as more than 1 will struggle to pay of its debts with current assets.

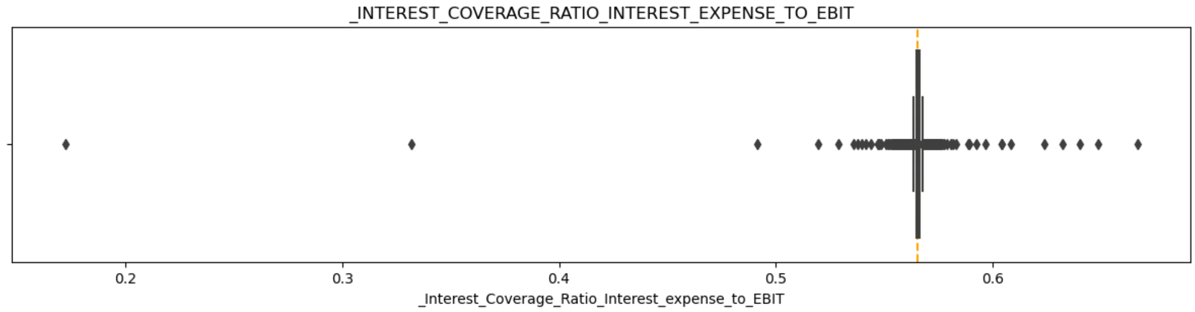


Figure H Boxplot - Interest Coverage Ratio Interest Expense to EBIT

The Interest Coverage Ratio is good for many companies, they can pay off the interest from their EBIT without any problem and have enough profit to utilise in other activities.

Box Plot by Default:

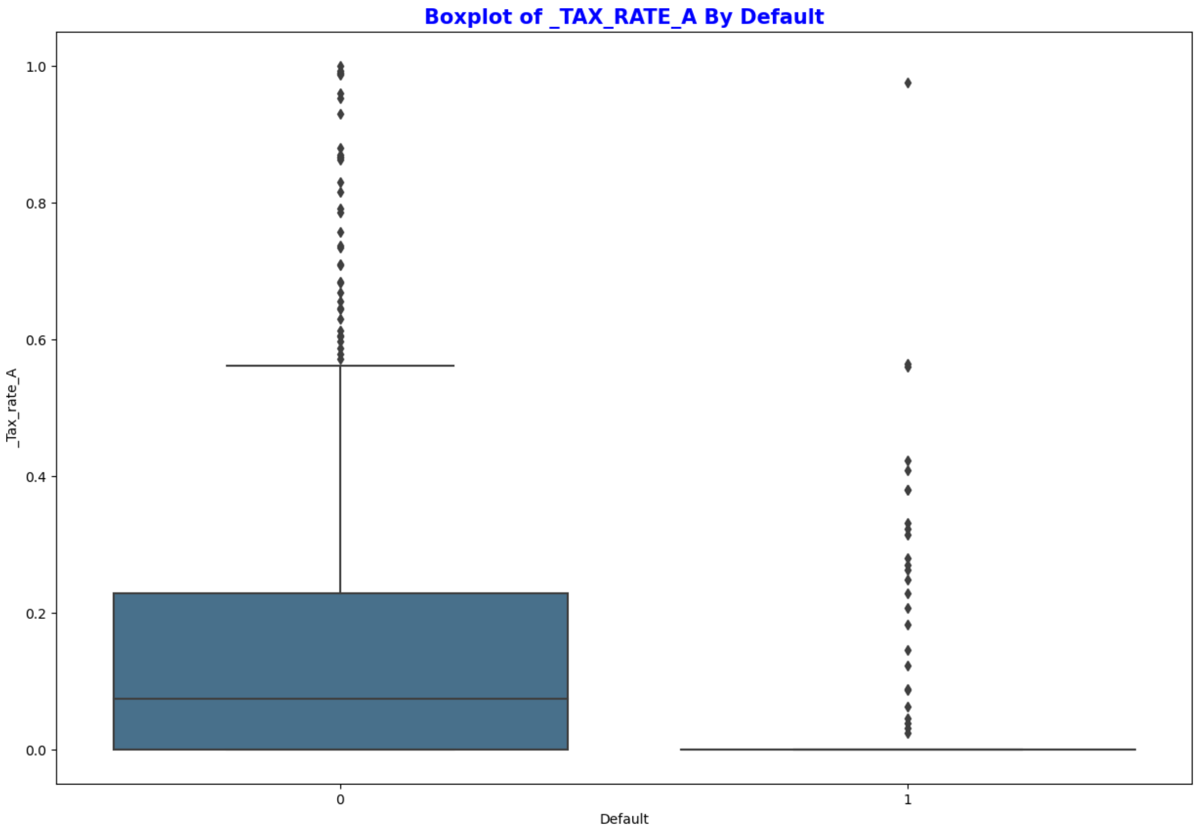


Figure I Boxplot - Taxrate by Default

The Tax Rate is less for the defaulters thus defaulters won't be paying their taxes effectively.

Non defaulters have a good tax rate.

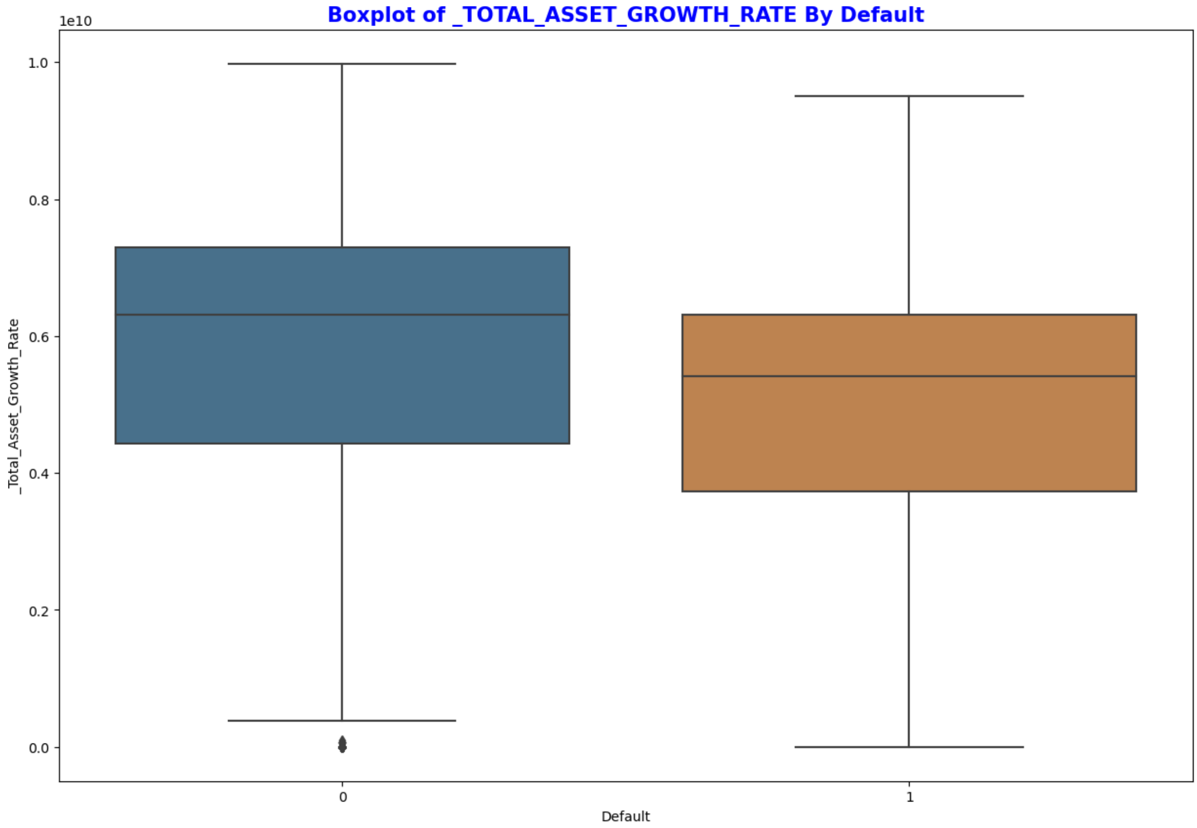


Figure J Boxplot - Total Asset Growth rate by default

There is also difference in the Total Asset Growth rate between defaulters and Non-defaulters.

Missing Value & Outlier Treatment

There is no Duplicates in the dataset.

Only 0.26% of the data is missing from the total data size.

Variables	Missing Data%
_Cash_Flow_Per_Share	**8.11**
_Total_debt_to_Total_net_worth	**1.02**
_Cash_to_Total_Assets	**4.66**
_Current_Liability_to_Current_Assets	**0.68**

Table 3 Missing Data %

The Missing Data is less than 10% for each column.

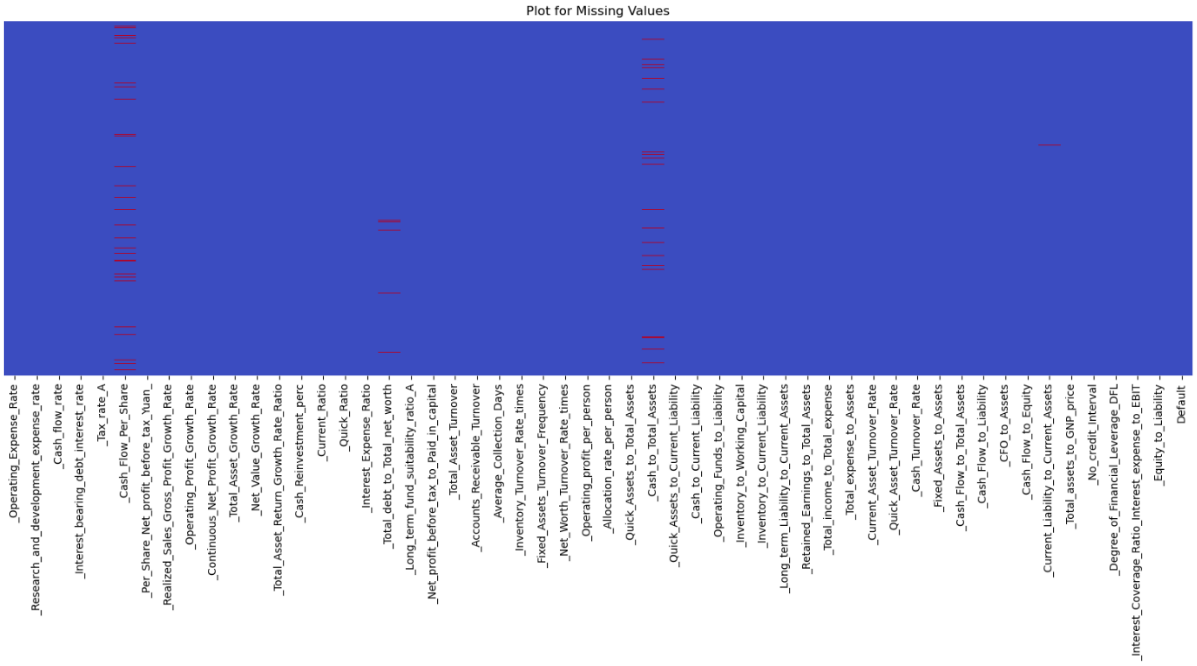


Figure K Missing Values Plot

The Red Coloured Boxes are Missing data. They are only a few.

In the case of outlier around 10% of the data is outlier from the total data size.

After imputing the outliers as NULL, the missing value also increased to 10%. This can be seen clearly in the below plot of missing values.

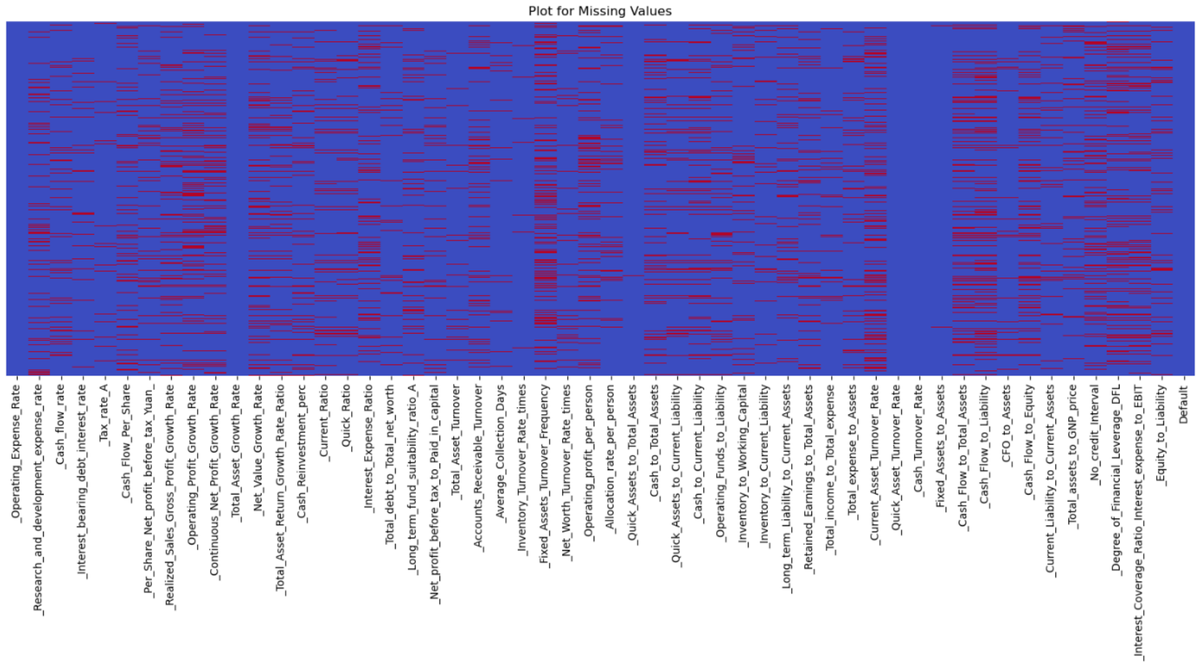


Figure L Missing Values Plot - After Outlier Imputation

If the data is removed based on row wise significance, around half of the rows has been removed and the defaulters data is reduced from 11% to 7%.

After imputing outliers as Nan, the variables are not exceeding the 30% of missing data. so, none of the columns will be dropped. They will be imputed through KNN Imputer.

The data has been scaled as each of the financial variables are in its own set of units.

The KNN Imputer with neighbour as 5 has been used for imputing the missing values with their nearest values.

Train Test Split

The Dataset has been split with test size of 0.33 and random state as 42.

The data has been split with a proportion of 66.95% of data to train and 33.04% of data to test.

Logistic Regression Model

Correlation Plot:

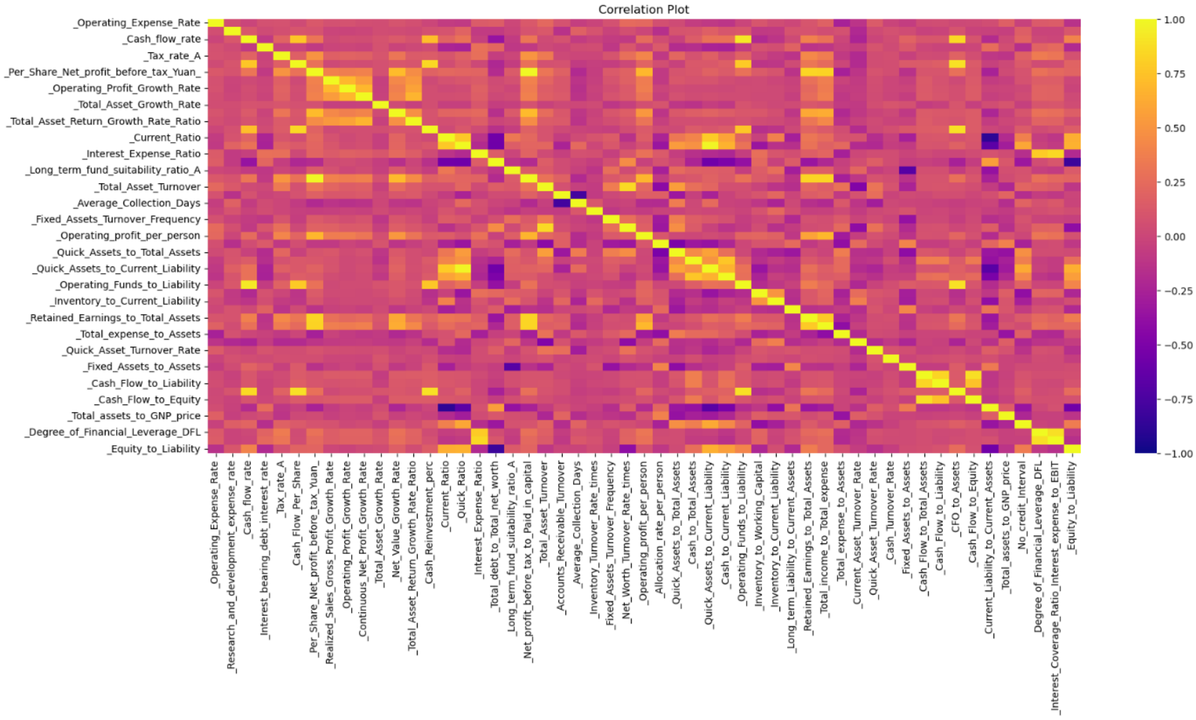


Figure M Correlation Plot

Most of the variables has a correlation of 0.25 to -0.25. Not many variables are too much positively or negatively correlated.

These can be removed in the VIF Analysis.

Model 1:

Model Formula:

Default = _Operating_Expense_Rate + _Research_and_development_expense_rate + _Cash_flow_rate + _Interest_bearing_debt_interest_rate + _Tax_rate_A + _Cash_Flow_Per_Share + _Per_Share_Net_profit_before_tax_Yuan_ + _Realized_Sales_Gross_Profit_Growth_Rate + _Operating_Profit_Growth_Rate + _Continuous_Net_Profit_Growth_Rate + _Total_Asset_Growth_Rate + _Net_Value_Growth_Rate + _Total_Asset_Return_Growth_Rate_Ratio + _Cash_Reinvestment_perc + _Current_Ratio + _Quick_Ratio + _Interest_Expense_Ratio + _Total_debt_to_Total_net_worth + _Long_term_fund_suitability_ratio_A + _Net_profit_before_tax_to_Paid_in_capital + _Total_Asset_Turnover + _Accounts_Receivable_Turnover + _Average_Collection_Days + _Inventory_Turnover_Rate_times + _Fixed_Assets_Turnover_Frequency + _Net_Worth_Turnover_Rate_times + _Operating_profit_per_person + _Allocation_rate_per_person + _Quick_Assets_to_Total_Assets + _Cash_to_Total_Assets +

_Quick_Assets_to_Current_Liability + _Cash_to_Current_Liability + _Operating_Funds_to_Liability + _Inventory_to_Working_Capital + _Inventory_to_Current_Liability + _Long_term_Liability_to_Current_Assets + _Retained_Earnings_to_Total_Assets + _Total_income_to_Total_expense + _Total_expense_to_Assets + _Current_Asset_Turnover_Rate + _Quick_Asset_Turnover_Rate + _Cash_Turnover_Rate + _Fixed_Assets_to_Assets + _Cash_Flow_to_Total_Assets + _Cash_Flow_to_Liability + _CFO_to_Assets + _Cash_Flow_to_Equity + _Current_Liability_to_Current_Assets + _Total_assets_to_GNP_price + _No_credit_Interval + _Degree_of_Financial_Leverage_DFL + _Interest_Coverage_Ratio_Interest_expense_to_EBIT + _Equity_to_Liability.

Most of the variables are not significant and the variables are correlated in this model.

After removing the correlated variables with VIF Method. The total independent variable has been reduced from 53 to 43.

S.No	variables	VIF
1	_Quick_Assets_to_Total_Assets	4.7
2	_Cash_Flow_Per_Share	4.58
3	_Cash_Reinvestment_perc	4.22
4	_Cash_to_Current_Liability	4.13
5	_Total_income_to_Total_expense	4.07
6	_Total_debt_to_Total_net_worth	4.02
7	_Equity_to_Liability	3.9
8	_Current_Liability_to_Current_Assets	3.81
9	_Cash_flow_rate	3.53
10	_Cash_to_Total_Assets	3.49
11	_Retained_Earnings_to_Total_Assets	3.44
12	_Fixed_Assets_to_Assets	3.4
13	_Net_Worth_Turnover_Rate_times	3.33
14	_Interest_Expense_Ratio	3.23
15	_Degree_of_Financial_Leverage_DFL	3.15
16	_Average_Collection_Days	3.13
17	_Accounts_Receivable_Turnover	2.85
18	_Inventory_to_Current_Liability	2.8
19	_Operating_Profit_Growth_Rate	2.72
20	_Operating_profit_per_person	2.56
21	_Continuous_Net_Profit_Growth_Rate	2.44
22	_Cash_Flow_to_Liability	2.39
23	_Cash_Flow_to_Equity	2.36
24	_Net_Value_Growth_Rate	2.28
25	_Long_term_fund_suitability_ratio_A	2.24
26	_Total_Asset_Return_Growth_Rate_Ratio	2.18
27	_Realized_Sales_Gross_Profit_Growth_Rate	2.16

28	_Allocation_rate_per_person	2.11
29	_Current_Asset_Turnover_Rate	2.07
30	_Total_expense_to_Assets	2
31	_Inventory_to_Working_Capital	1.94
32	_No_credit_Interval	1.91
33	_Tax_rate_A	1.59
34	_Long_term_Liability_to_Current_Assets	1.57
35	_Total_assets_to_GNP_price	1.52
36	_Fixed_Assets_Turnover_Frequency	1.37
37	_Quick_Asset_Turnover_Rate	1.35
38	_Interest_bearing_debt_interest_rate	1.29
39	_Operating_Expense_Rate	1.27
40	_Inventory_Turnover_Rate_times	1.25
41	_Research_and_development_expense_rate	1.16
42	_Total_Asset_Growth_Rate	1.1
43	_Cash_Turnover_Rate	1.08

Table 4 VIF for Variables

The Variables in the top table has VIF less than 5.

These variables are used for the Next Model.

Model 2:

Model Formula:

Default = _Operating_Expense_Rate + _Research_and_development_expense_rate + _Cash_flow_rate + _Interest_bearing_debt_interest_rate + _Tax_rate_A + _Cash_Flow_Per_Share + _Realized_Sales_Gross_Profit_Growth_Rate + _Operating_Profit_Growth_Rate + _Continuous_Net_Profit_Growth_Rate + _Total_Asset_Growth_Rate + _Net_Value_Growth_Rate + _Total_Asset_Return_Growth_Rate_Ratio + _Cash_Reinvestment_perc + _Interest_Expense_Ratio + _Total_debt_to_Total_net_worth + _Long_term_fund_suitability_ratio_A + _Accounts_Receivable_Turnover + _Average_Collection_Days + _Inventory_Turnover_Rate_times + _Fixed_Assets_Turnover_Frequency + _Net_Worth_Turnover_Rate_times + _Operating_profit_per_person + _Allocation_rate_per_person + _Quick_Assets_to_Total_Assets + _Cash_to_Total_Assets + _Cash_to_Current_Liability + _Inventory_to_Working_Capital + _Inventory_to_Current_Liability + _Long_term_Liability_to_Current_Assets + _Retained_Earnings_to_Total_Assets + _Total_income_to_Total_expense + _Total_expense_to_Assets + _Current_Asset_Turnover_Rate + _Quick_Asset_Turnover_Rate + _Cash_Turnover_Rate + _Fixed_Assets_to_Assets + _Cash_Flow_to_Liability + _Cash_Flow_to_Equity +

_Current_Liability_to_Current_Assets + _Total_assets_to_GNP_price + _No_credit_Interval + _Degree_of_Financial_Leverage_DFL + _Equity_to_Liability.

After removing the correlated variables, most of the independent variables are not significant.

Not Significant Variables:

S.No	Variable Name
1	Operating_Expense_Rate
2	Cash_flow_rate
3	Tax_rate_A
4	Cash_Flow_Per_Share
5	Realized_Sales_Gross_Profit_Growth_Rate
6	Operating_Profit_Growth_Rate
7	Continuous_Net_Profit_Growth_Rate
8	Total_Asset_Growth_Rate
9	Net_Value_Growth_Rate
10	Total_Asset_Return_Growth_Rate_Ratio
11	Cash_Reinvestment_perc
12	Interest_Expense_Ratio
13	Long_term_fund_suitability_ratio_A
14	Average_Collection_Days
15	Inventory_Turnover_Rate_times
16	Fixed_Assets_Turnover_Frequency
17	Net_Worth_Turnover_Rate_times
18	Operating_profit_per_person
19	Quick_Assets_to_Total_Assets
20	Cash_to_Total_Assets
21	Cash_to_Current_Liability
22	Inventory_to_Current_Liability
23	Long_term_Liability_to_Current_Assets
24	Retained_Earnings_to_Total_Assets
25	Current_Asset_Turnover_Rate
26	Quick_Asset_Turnover_Rate
27	Fixed_Assets_to_Assets
28	Cash_Flow_to_Equity
29	Current_Liability_to_Current_Assets
30	Total_assets_to_GNP_price
31	Degree_of_Financial_Leverage_DFL

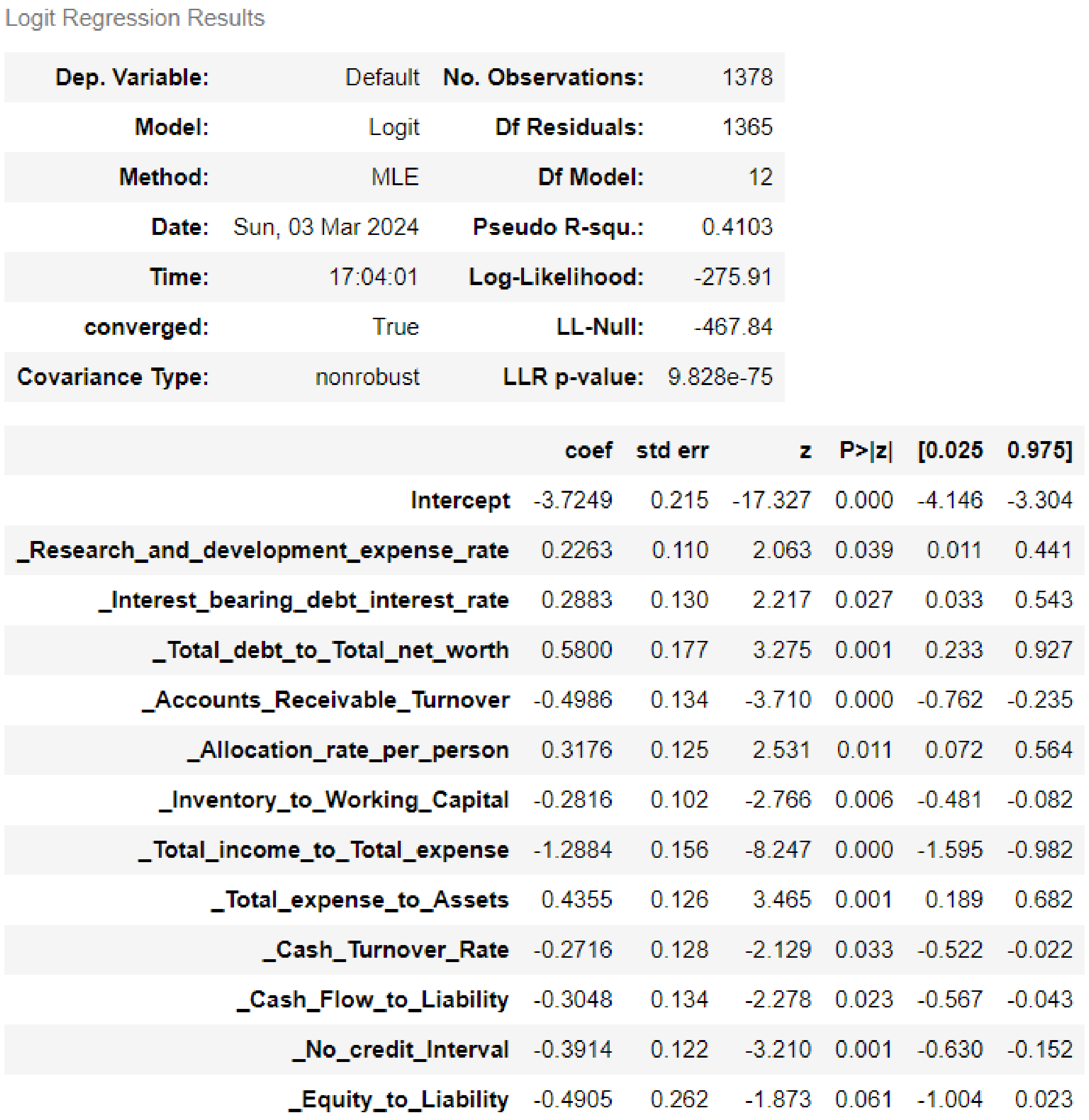
Table 5 Non Significant Variables

These variables are removed in the next model.

Model3 – Final Model:

Model Formula:

Default = _Research_and_development_expense_rate + _Interest_bearing_debt_interest_rate + _Total_debt_to_Total_net_worth + _Accounts_Receivable_Turnover + _Allocation_rate_per_person + _Inventory_to_Working_Capital + _Total_income_to_Total_expense + _Total_expense_to_Assets + _Cash_Turnover_Rate + _Cash_Flow_to_Liability + _No_credit_Interval + _Equity_to_Liability.



Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1365
Method:	MLE	**Df Model:**	12
Date:	Sun, 03 Mar 2024	**Pseudo R-squ.:**	0.4103
Time:	17:04:01	**Log-Likelihood:**	-275.91
converged:	True	**LL-Null:**	-467.84
Covariance Type:	nonrobust	**LLR p-value:**	9.828e-75

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-3.7249	0.215	-17.327	0.000	-4.146	-3.304
_Research_and_development_expense_rate	0.2263	0.110	2.063	0.039	0.011	0.441
_Interest_bearing_debt_interest_rate	0.2883	0.130	2.217	0.027	0.033	0.543
_Total_debt_to_Total_net_worth	0.5800	0.177	3.275	0.001	0.233	0.927
_Accounts_Receivable_Turnover	-0.4986	0.134	-3.710	0.000	-0.762	-0.235
_Allocation_rate_per_person	0.3176	0.125	2.531	0.011	0.072	0.564
_Inventory_to_Working_Capital	-0.2816	0.102	-2.766	0.006	-0.481	-0.082
_Total_income_to_Total_expense	-1.2884	0.156	-8.247	0.000	-1.595	-0.982
_Total_expense_to_Assets	0.4355	0.126	3.465	0.001	0.189	0.682
_Cash_Turnover_Rate	-0.2716	0.128	-2.129	0.033	-0.522	-0.022
_Cash_Flow_to_Liability	-0.3048	0.134	-2.278	0.023	-0.567	-0.043
_No_credit_Interval	-0.3914	0.122	-3.210	0.001	-0.630	-0.152
_Equity_to_Liability	-0.4905	0.262	-1.873	0.061	-1.004	0.023

The Model Equation is

Default = (-3.72) * Intercept + (0.23) * _Research_and_development_expense_rate + (0.29) * _Interest_bearing_debt_interest_rate + (0.58) * _Total_debt_to_Total_net_worth + (-0.5) * _Accounts_Receivable_Turnover + (0.32) * _Allocation_rate_per_person + (-0.28) * _Inventory_to_Working_Capital + (-1.29) * _Total_income_to_Total_expense + (0.44) * _Total_expense_to_Assets + (-0.27) * _Cash_Turnover_Rate + (-0.3) * _Cash_Flow_to_Liability + (-0.39) * _No_credit_Interval + (-0.49) * _Equity_to_Liability.

The Significant Variables are

- _Research_and_development_expense_rate
- _Interest_bearing_debt_interest_rate
- _Total_debt_to_Total_net_worth
- _Accounts_Receivable_Turnover
- _Allocation_rate_per_person
- _Inventory_to_Working_Capital
- _Total_income_to_Total_expense
- _Total_expense_to_Assets
- _Cash_Turnover_Rate
- _Cash_Flow_to_Liability
- _No_credit_Interval
- _Equity_to_Liability

The Chance of Default will be reduced if the below ratios are increased.

- _Accounts_Receivable_Turnover
- _Inventory_to_Working_Capital
- _Total_income_to_Total_expense
- _Cash_Turnover_Rate
- _Cash_Flow_to_Liability
- _No_credit_Interval
- _Equity_to_Liability.

These ratios are related to Working capital and Cash Flows. High value in these are a good indicator that the Total Income, Total Cash and Working capital are higher and debt and debt credit interval are lower.

The Chance of Default will be reduced if the below ratios are decreased in the expense and debt side.

- _Research_and_development_expense_rate
- _Interest_bearing_debt_interest_rate
- _Total_debt_to_Total_net_worth
- _Allocation_rate_per_person

The Vice versa of the above two combinations will increase the chance of default, which the model predicts.

Optimum Threshold:

With 0.5 as threshold, the model recall is not good.

The Recall is important as the prediction is for defaulters.

The cost of predicting Non defaulters as defaulters is less instead of predicting defaulters as Non defaulters.

The Confusion Matrix and Classification report for 0.5 threshold

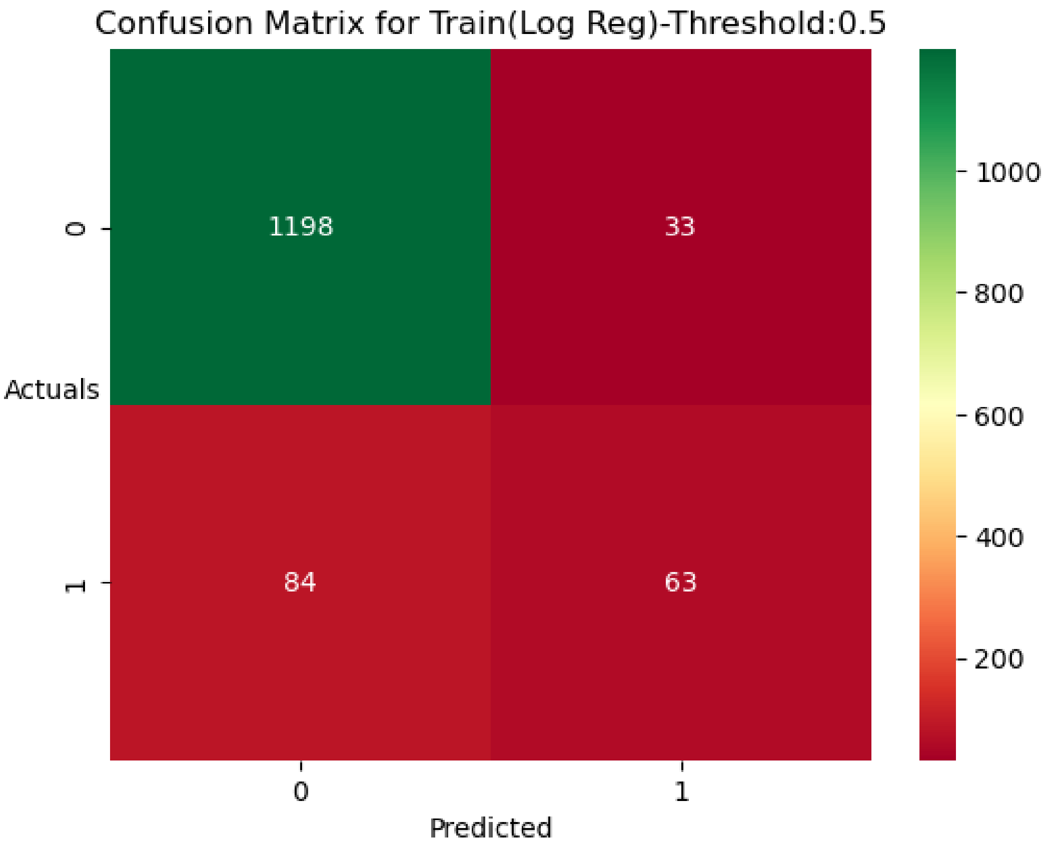
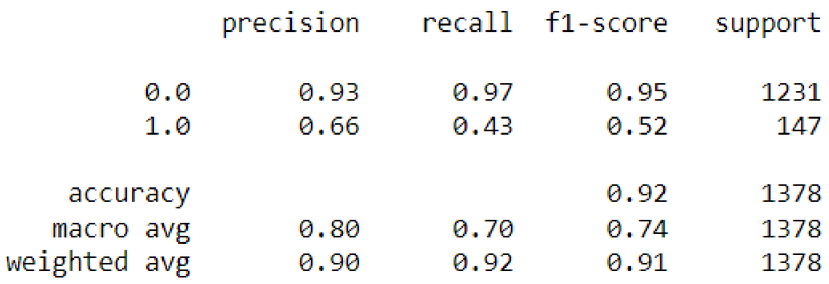


Figure N Confusion Matrix for Train(Log Reg)



```
              precision    recall  f1-score   support

         0.0       0.93      0.97      0.95      1231
         1.0       0.66      0.43      0.52       147

    accuracy                           0.92      1378
   macro avg       0.80      0.70      0.74      1378
weighted avg       0.90      0.92      0.91      1378
```

Table 6 Classification report for Train(Log Reg)

Even though the accuracy is 92%, as the dataset is not balanced. The Recall for defaulters is important for the model performance.

The Optimum threshold value is 0.11.

The Confusion Matrix and Classification report for optimum threshold

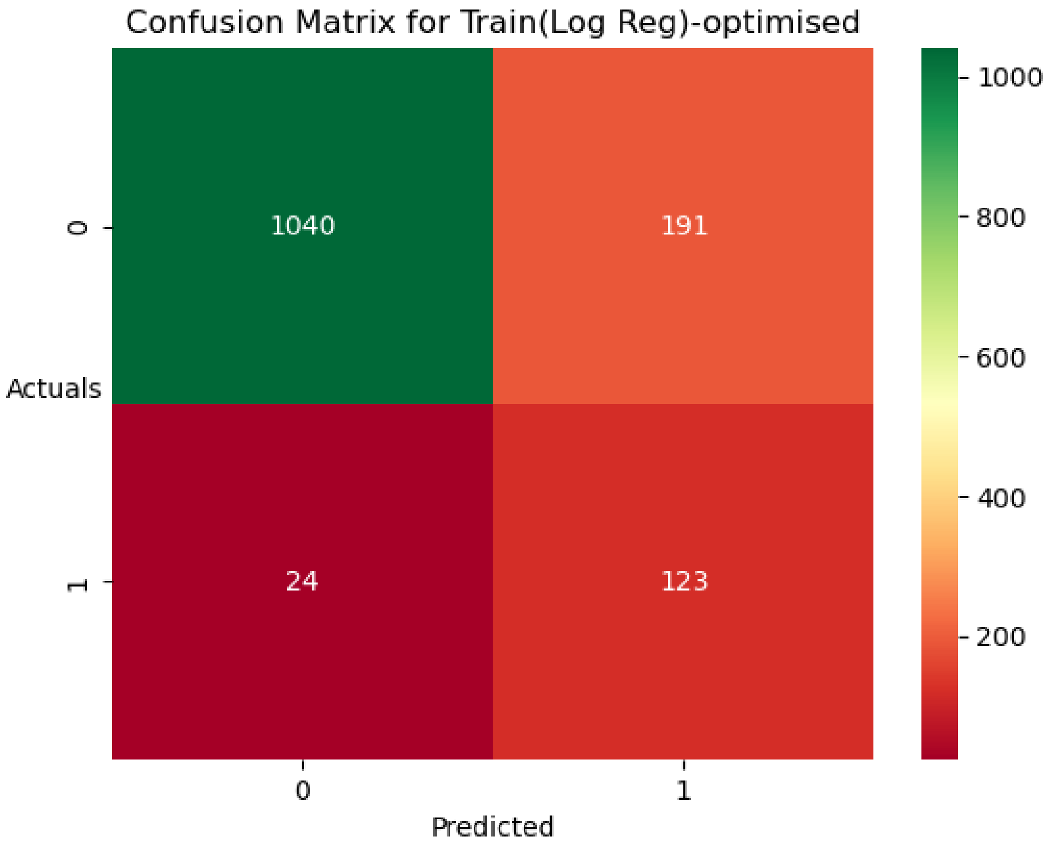
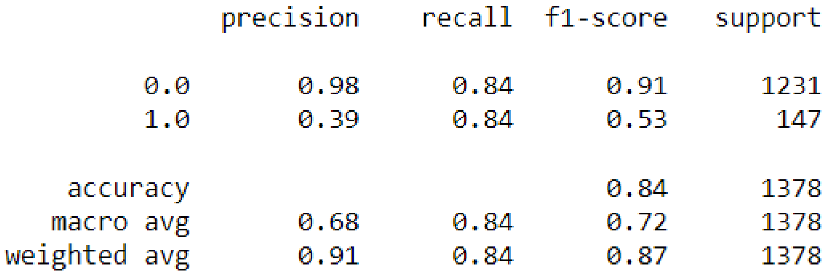


Figure O Confusion Matrix for Train(Log Reg)-Optimised



```
              precision    recall  f1-score   support

         0.0       0.98      0.84      0.91      1231
         1.0       0.39      0.84      0.53       147

    accuracy                           0.84      1378
   macro avg       0.68      0.84      0.72      1378
weighted avg       0.91      0.84      0.87      1378
```

Table 7 Classification Report for Train(Log Reg)-Optimised

The recall is improved a lot in this model.

This optimum threshold is preferred as the model predicts 84% of the actual defaulters correctly.

The Confusion Matrix and Classification report for optimum threshold in Test data

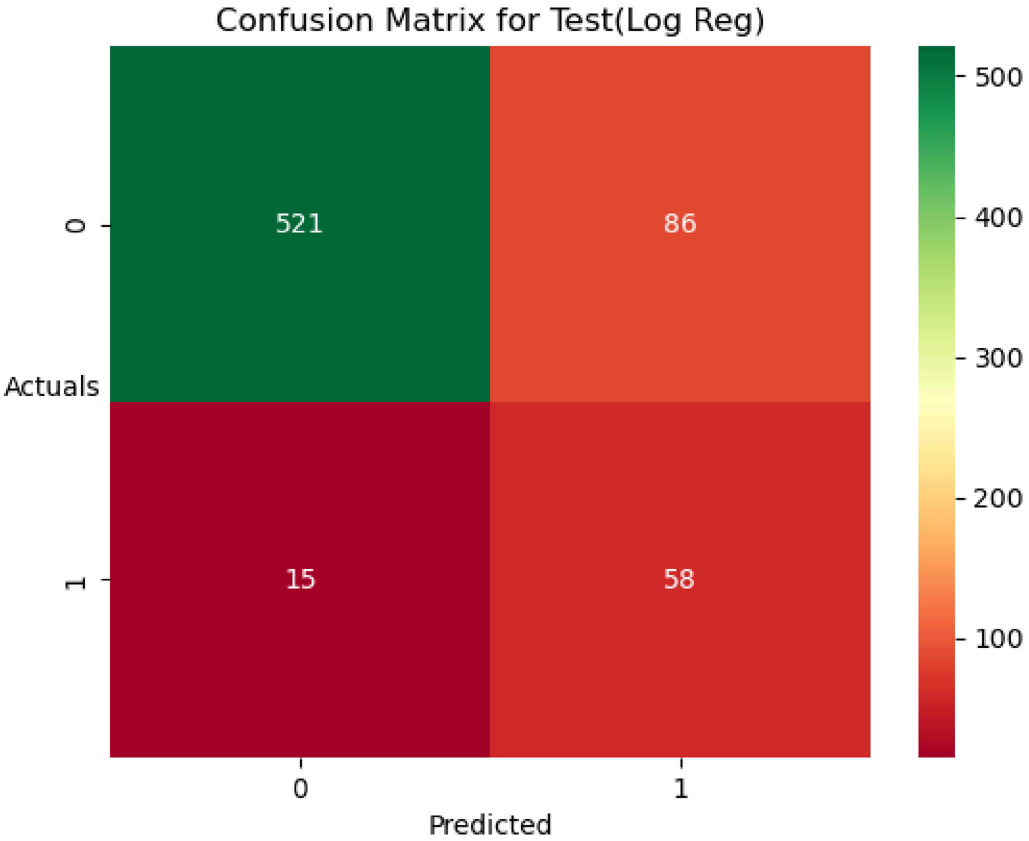
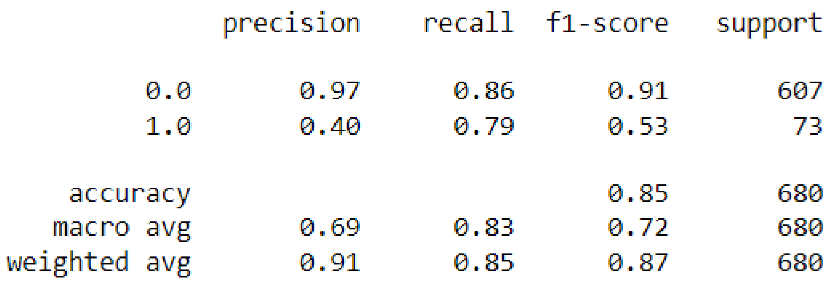


Figure P Confusion Matrix for Test(Log Reg)-Optimised



```
              precision    recall  f1-score   support

         0.0       0.97      0.86      0.91       607
         1.0       0.40      0.79      0.53        73

    accuracy                           0.85       680
   macro avg       0.69      0.83      0.72       680
weighted avg       0.91      0.85      0.87       680
```

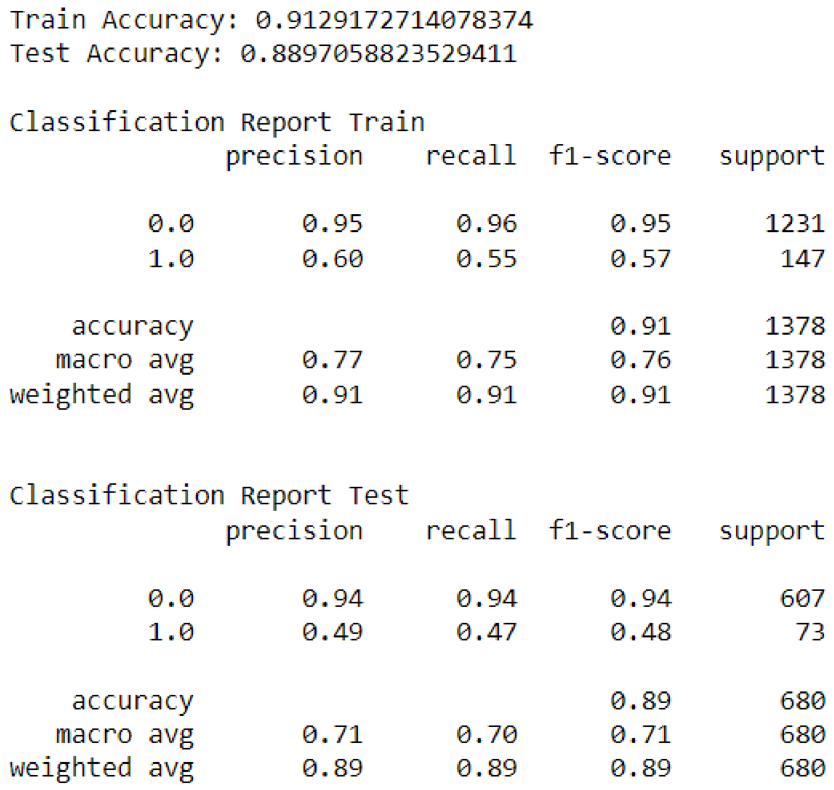
Table 8 Classification report for Test(Log Reg)-Optimised

The Model performs good in the test data also.

LDA Model

The LDA is another model for prediction of defaulters.

Model 1: Basic LDA Model



```
Train Accuracy: 0.9129172714078374
Test Accuracy: 0.8897058823529411

Classification Report Train
              precision    recall  f1-score   support

         0.0       0.95      0.96      0.95      1231
         1.0       0.60      0.55      0.57       147

    accuracy                           0.91      1378
   macro avg       0.77      0.75      0.76      1378
weighted avg       0.91      0.91      0.91      1378


Classification Report Test
              precision    recall  f1-score   support

         0.0       0.94      0.94      0.94       607
         1.0       0.49      0.47      0.48        73

    accuracy                           0.89       680
   macro avg       0.71      0.70      0.71       680
weighted avg       0.89      0.89      0.89       680
```

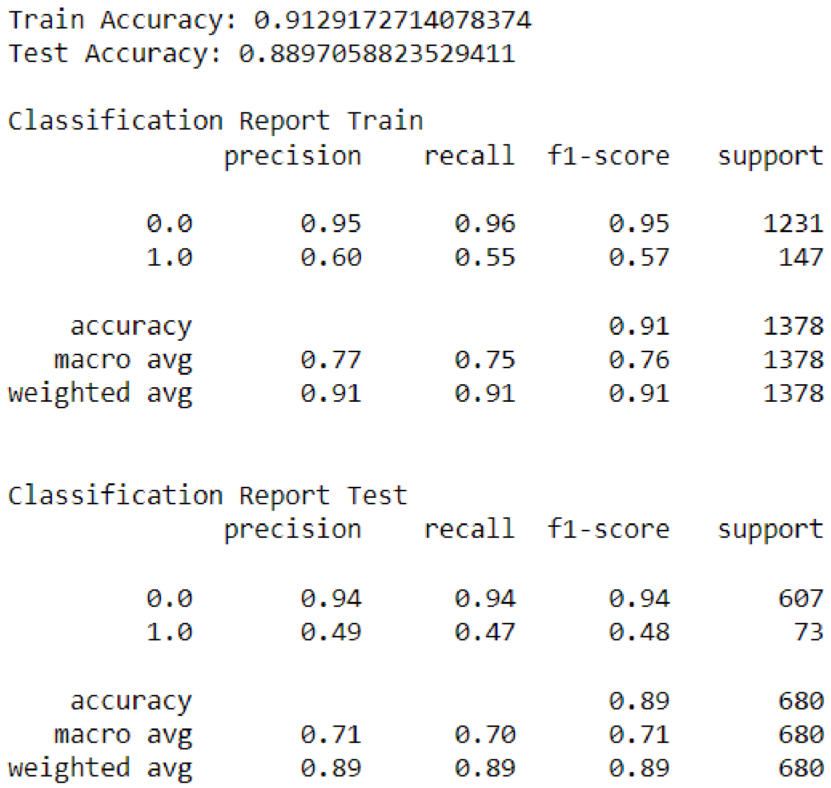
Table 9 Classification Report For LDA Model 1

The LDA model is not performing better than the Log Reg.

After Tuning, the model performance remains the same.

Model 2: LDA Tuning

The Solver parameters for tuning are 'solver':['svd','lsqr','eigen'].



```
Train Accuracy: 0.9129172714078374
Test Accuracy: 0.8897058823529411

Classification Report Train
              precision    recall  f1-score   support

         0.0       0.95      0.96      0.95      1231
         1.0       0.60      0.55      0.57       147

    accuracy                           0.91      1378
   macro avg       0.77      0.75      0.76      1378
weighted avg       0.91      0.91      0.91      1378


Classification Report Test
              precision    recall  f1-score   support

         0.0       0.94      0.94      0.94       607
         1.0       0.49      0.47      0.48        73

    accuracy                           0.89       680
   macro avg       0.71      0.70      0.71       680
weighted avg       0.89      0.89      0.89       680
```

Table 10 Classification report for LDA Model 2

Optimum Threshold:

The Model performance is not better than the Log Reg even after tuning also. So, the optimum threshold is calculated to improve the model.

The Optimum threshold value is 0.03.

The Confusion Matrix and Classification report for optimum threshold

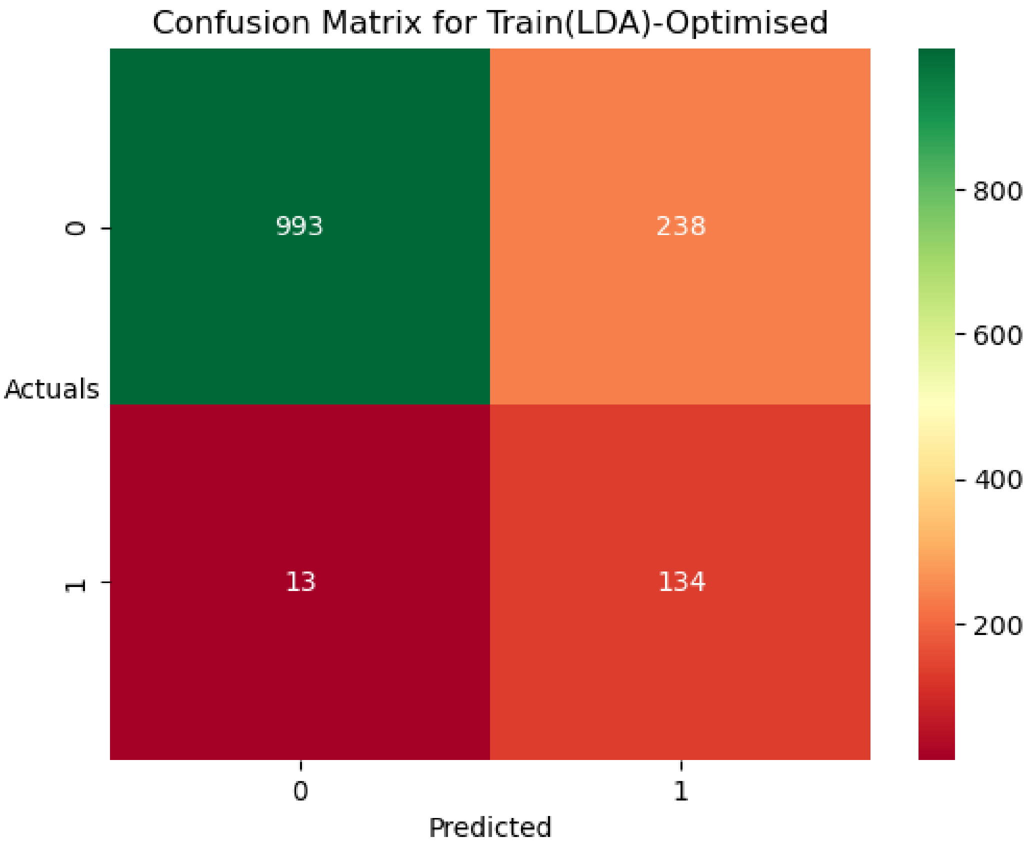
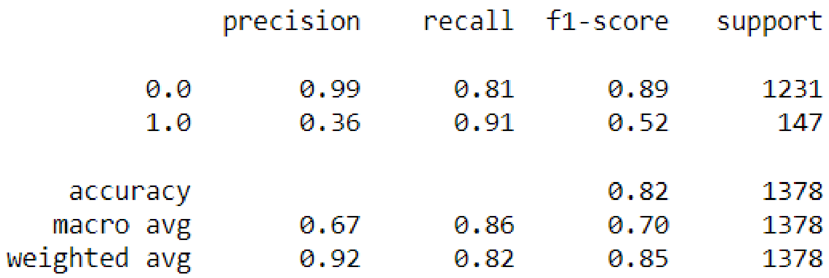


Figure Q Confusion Matrix for Train(LDA)-Optimised



```
              precision    recall  f1-score   support

         0.0       0.99      0.81      0.89      1231
         1.0       0.36      0.91      0.52       147

    accuracy                           0.82      1378
   macro avg       0.67      0.86      0.70      1378
weighted avg       0.92      0.82      0.85      1378
```

Table 11 Classification report for Train(LDA)-Optimised

The Recall is 91%, which is better than the Log Reg Train Recall. The LDA Model with optimised threshold is preferred.

The Confusion Matrix and Classification report for optimum threshold in Test data

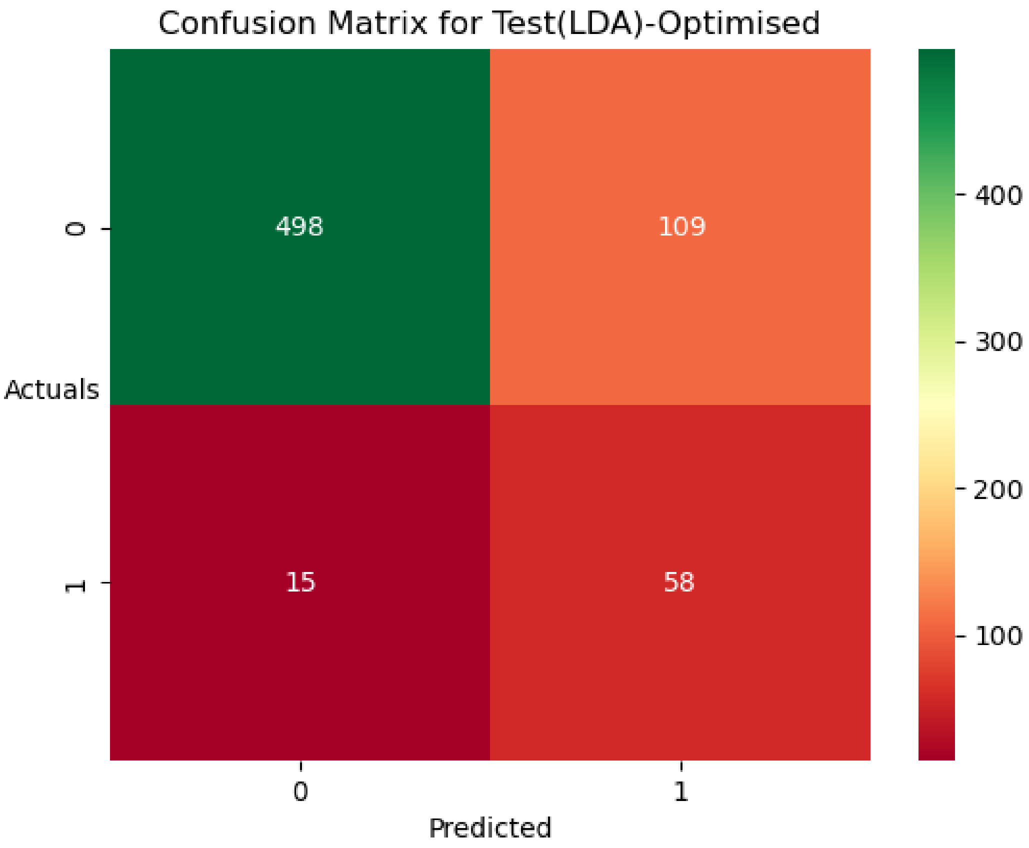
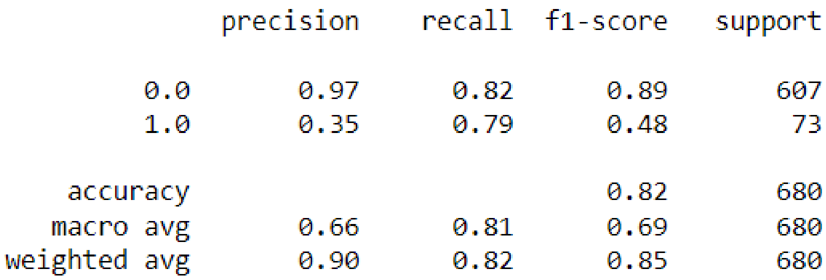


Figure R Confusion Matrix for Test(LDA)-Optimised



```
              precision    recall  f1-score   support

         0.0       0.97      0.82      0.89       607
         1.0       0.35      0.79      0.48        73

    accuracy                           0.82       680
   macro avg       0.66      0.81      0.69       680
weighted avg       0.90      0.82      0.85       680
```

Table 12 Classification Report for Test(LDA)-Optimised

The LDA Model performs similar to the Log Reg Model.

Random Forest Model

The Random Forest Model is the third model for prediction of defaulters.

The Model Tuning Parameters for RF are

{'criterion':['gini','entropy'],

'n_estimators':list(range(100,1000,2)),

'min_samples_leaf':list(range(1,10)),

'max_features':list(range(1,15)),

'max_samples':list(np.arange(0.1,1))}

The Confusion Matrix and Classification Report for Train and Test

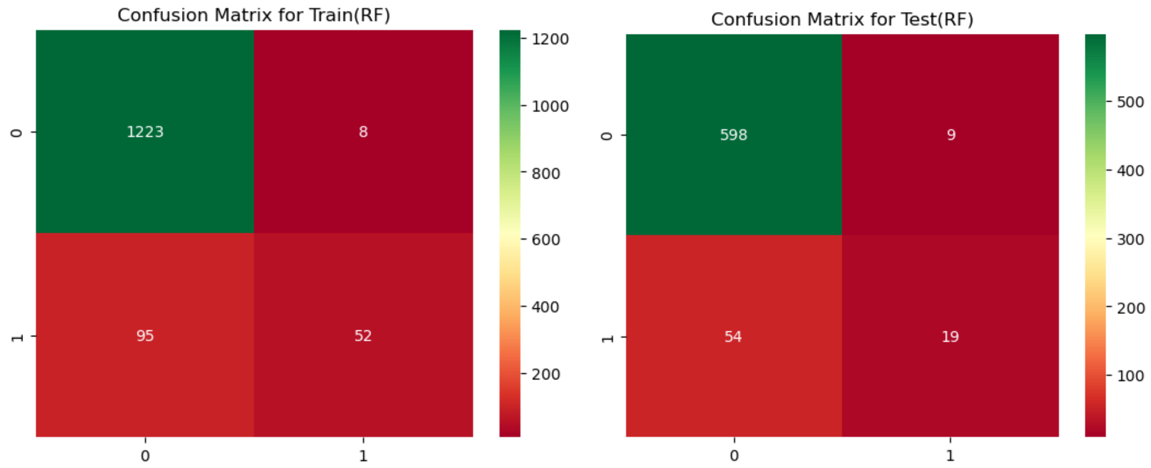
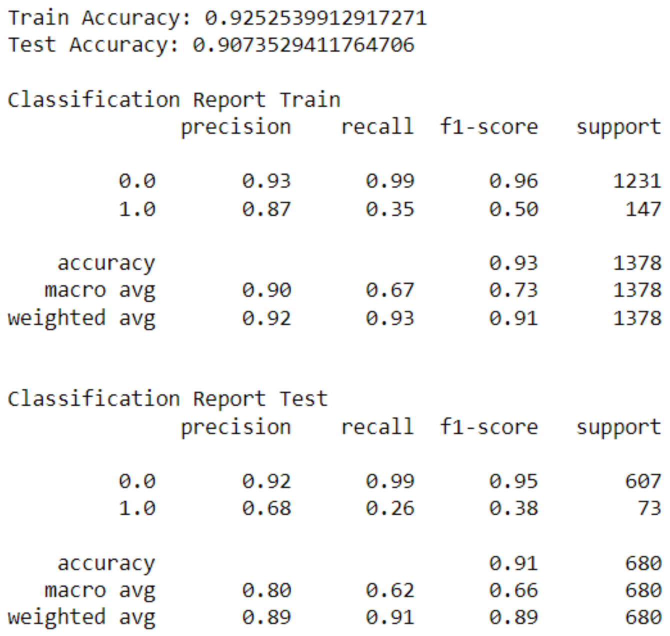


Figure S Confusion Matrix for RF



```
Train Accuracy: 0.9252539912917271
Test Accuracy: 0.9073529411764706

Classification Report Train
              precision    recall  f1-score   support

         0.0       0.93      0.99      0.96      1231
         1.0       0.87      0.35      0.50       147

    accuracy                           0.93      1378
   macro avg       0.90      0.67      0.73      1378
weighted avg       0.92      0.93      0.91      1378


Classification Report Test
              precision    recall  f1-score   support

         0.0       0.92      0.99      0.95       607
         1.0       0.68      0.26      0.38        73

    accuracy                           0.91       680
   macro avg       0.80      0.62      0.66       680
weighted avg       0.89      0.91      0.89       680
```

Table 13 Classification Report for RF

The RF is the worst model among the three for the prediction of defaulters and it is also over fitting the train data.

It has very low Recall with high accuracy and average Precision.

This model can’t be used for the prediction.

Model Comparison

By Comparing the three model’s performance based on the classification report and ROC Curve, the best model is considered for defaulters prediction.

	Log Reg	LDA	RF
AUC:	0.90	0.87	0.62

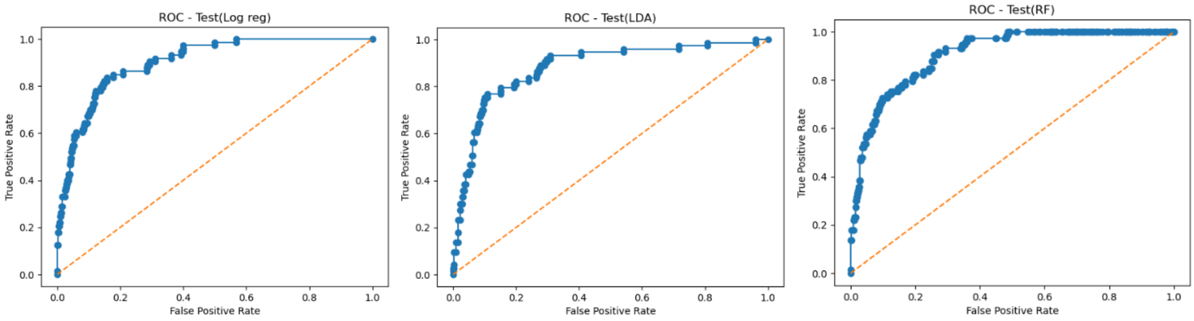
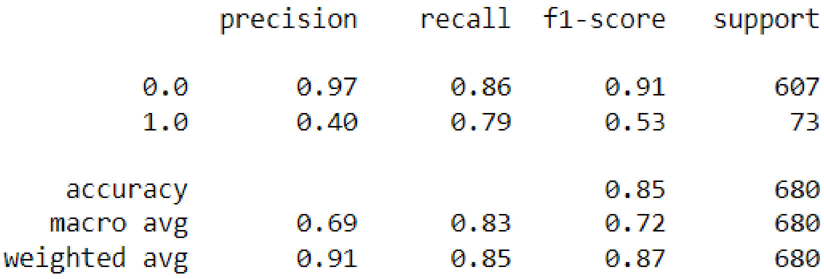


Figure T ROC Curve - Log Reg,LDA,RF

From the classification report, RF has the worst Recall and here it has low AUC. So it can be removed from the comparison.

The Log Reg has the Higher AUC value.

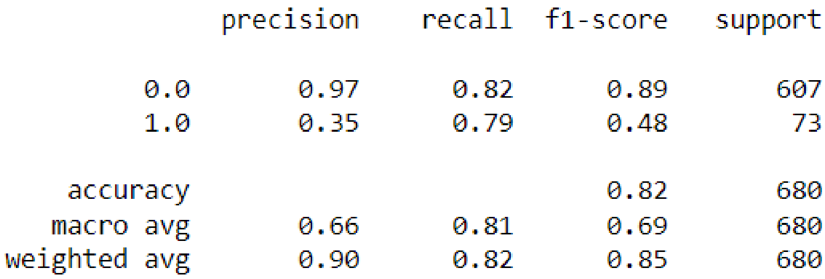


```
              precision    recall  f1-score   support

         0.0       0.97      0.86      0.91       607
         1.0       0.40      0.79      0.53        73

    accuracy                           0.85       680
   macro avg       0.69      0.83      0.72       680
weighted avg       0.91      0.85      0.87       680
```

Table 14 Classification Report for Log Reg



```
              precision    recall  f1-score   support

         0.0       0.97      0.82      0.89       607
         1.0       0.35      0.79      0.48        73

    accuracy                           0.82       680
   macro avg       0.66      0.81      0.69       680
weighted avg       0.90      0.82      0.85       680
```

Table 15 Classification Report for LDA

The Log Reg has better Precision and Accuracy as the recall is same for both models and Log Reg has higher AUC. The Logistic Regression is the best model among the three.

Conclusions and Recommendations

The Defaulters and Non-Defaulters has mostly same characteristics in their ratio. But there are some indicators that can be used for the prediction of defaulters.

These ratios are related to the debt, cash and capital of the company.

The defaulters’ characteristics can be analysed from these below ratios

- _Research_and_development_expense_rate
- _Interest_bearing_debt_interest_rate
- _Total_debt_to_Total_net_worth
- _Accounts_Receivable_Turnover
- _Allocation_rate_per_person
- _Inventory_to_Working_Capital
- _Total_income_to_Total_expense
- _Total_expense_to_Assets
- _Cash_Turnover_Rate
- _Cash_Flow_to_Liability
- _No_credit_Interval
- _Equity_to_Liability

If the Cash turnover and cashflow to liability are low, they are mostly likely running low on cash and if the debts ratios are high with interest bearing debt, they are most likely won’t be able to pay the interest and will have increased chance of default.

These ratios indicators and model prediction can be recommended for the prediction of defaulters.

The worst case if the loan is ongoing and they are predicted as defaulters, the asset ratio can be analysed and loan amount recovery analyses from those assets can be done.