Financial and Risk Analysis Business Report

DSBA

Vikash Kumar
Aug'22 Batch

Contents

List of Tables

List of figures and plots

Problem Statement:

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 interests 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 set for the Problem: Company

Read the data as an appropriate Time Series data and plot the data.

- To enable us to address the business issue, we have imported necessary libraries:
  - Basic - from Numpy and Pandas
  - Data visualization - Matplotlib, Seaborn,
  - General – Python standard warnings
  - Statistics - scipy and scikit learn for time series modelling and plotting time series specific plots
- The dataset in csv format 'Company' is loaded in python data and we have verified that the same is properly loaded. The Header and Tail of dataset is verified for completeness. (Due to large numbers of variables tables seems clutter)

Num	Num	worth Next	Total asset	Net worth	otal income	ange in sto	otal expens	ofit after t	PBDITA	PBT		Debtors turnov	d goods tu	WIP turnove	aterial turn	es outstan	ility face va	EPS	djusted EP	otal liabiliti	PE on BSE
0	1	395.3	827.6	336.5	534.1	13.5	508.7	38.9	124.4	64.6	...	5.65	3.99	3.37	14.87	8760056	10	4.44	4.44	827.6	NaN
1	2	36.2	67.7	24.3	137.9	-3.7	131	3.2	5.5	1	...	NaN	NaN	NaN	NaN	NaN	NaN	0	0	67.7	NaN
2	3	84	238.4	78.9	331.2	-18.1	309.2	3.9	25.8	10.5	...	2.51	17.67	8.76	8.35	NaN	NaN	0	0	238.4	NaN
3	4	2041.4	6883.5	1443.3	8448.5	212.2	8482.4	178.3	418.4	185.1	...	1.91	18.14	18.62	11.11	10000000	10	17.6	17.6	6883.5	NaN
4	5	41.8	90.9	47	388.6	3.4	392.7	-0.7	7.2	-0.6	...	68	45.87	28.67	19.93	107315	100	-6.52	-6.52	90.9	NaN
ws × 51 columns																					

TABLE 1 :HEADER OF DATASET

Data Information:

First Information of dataset about variable and their presence like , Data Type, Size, Shape,and number of variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4256 entries, 0 to 4255
Data columns (total 51 columns):
 #    Column                                      Non-Null Count  Dtype
---   ------                                      --------------  -----
 0    Num                                         4256 non-null   int64
 1    Networth_Next_Year                          4256 non-null   float64
 2    Total_assets                                4256 non-null   float64
 3    Net_worth                                   4256 non-null   float64
 4    Total_income                                4025 non-null   float64
 5    Change_in_stock                             3706 non-null   float64
 6    Total_expenses                              4091 non-null   float64
 7    Profit_after_tax                            4102 non-null   float64
 8    PBDITA                                      4102 non-null   float64
 9    PBT                                         4102 non-null   float64
 10   Cash_profit                                 4102 non-null   float64
 11   PBDITA_as_perc_of_total_income              4177 non-null   float64
 12   PBT_as_perc_of_total_income                 4177 non-null   float64
 13   PAT_as_perc_of_total_income                 4177 non-null   float64
 14   Cash_profit_as_perc_of_total_income         4177 non-null   float64
 15   PAT_as_perc_of_net_worth                    4256 non-null   float64
 16   Sales                                       3951 non-null   float64
 17   Income_from_fincial_services                3145 non-null   float64
 18   Other_income                                2700 non-null   float64
 19   Total_capital                               4251 non-null   float64
 20   Reserves_and_funds                          4158 non-null   float64
 21   Borrowings                                  3825 non-null   float64
 22   Current_liabilities_&_provisions            4146 non-null   float64
 23   Deferred_tax_liability                      2887 non-null   float64
 24   Shareholders_funds                          4256 non-null   float64
 25   Cumulative_retained_profits                 4211 non-null   float64
 26   Capital_employed                            4256 non-null   float64
 27   TOL_to_TNW                                  4256 non-null   float64
 28   Total_term_liabilities__to__tangible_net_worth 4256 non-null float64
 29   Contingent_liabilities__to__Net_worth_perc  4256 non-null   float64
 30   Contingent_liabilities                      2854 non-null   float64
 31   Net_fixed_assets                            4124 non-null   float64
 32   Investments                                 2541 non-null   float64
 33   Current_assets                              4176 non-null   float64
 34   Net_working_capital                         4219 non-null   float64
 35   Quick_ratio_times                           4151 non-null   float64
 36   Current_ratio_times                         4151 non-null   float64
 37   Debt_to_equity_ratio_times                  4256 non-null   float64
 38   Cash_to_current_liabilities_times           4151 non-null   float64
 39   Cash_to_average_cost_of_sales_per_day       4156 non-null   float64
 40   Creditors_turnover                          3865 non-null   float64
 41   Debtors_turnover                            3871 non-null   float64
 42   Finished_goods_turnover                     3382 non-null   float64
 43   WIP_turnover                                3492 non-null   float64
 44   Raw_material_turnover                       3828 non-null   float64
 45   Shares_outstanding                          3446 non-null   float64
 46   Equity_face_value                           3446 non-null   float64
 47   EPS                                         4256 non-null   float64
 48   Adjusted_EPS                                4256 non-null   float64
 49   Total_liabilities                           4256 non-null   float64
 50   PE_on_BSE                                   1629 non-null   float64
dtypes: float64(50), int64(1)
memory usage: 1.7 MB
```

The number of rows (observations) is 4256
The number of columns (variables) is 51

TABLE 2: SHAPE & INFORMATION OF THE DATASET

- The details of columns ie data dictionary is read for understanding of the data

Data Dictionary

	Variable Name	Description
0	Networth Next Year	Net worth of the customer in next year
1	Total assets	Total assets of customer
2	Net worth	Net worth of the customer of present year
3	Total income	Total income of the customer
4	Change in stock	difference between value of current stock and...
5	Total expenses	Total expense done by customer
6	Profit after tax	Profit after tax deduction
7	PBDITA	Profit before depreciation, income tax and am...
8	PBT	Profit before tax deduction
9	Cash profit	Total Cash profit
10	PBDITA as % of total income	PBDITA / Total income
11	PBT as % of total income	PBT / Total income
12	PAT as % of total income	PAT / Total income
13	Cash profit as % of total income	Cash Profit / Total income
14	PAT as % of net worth	PAT / Net worth
15	Sales	Sales done by customer
16	Income from financial services	Income from financial services
17	Other income	Income from other sources
18	Total capital	Total capital of the customer
19	Reserves and funds	Total reserves and funds of the customer
20	Deposits (accepted by commercial banks)	All blank values
21	Borrowings	Total amount borrowed by customer
22	Current liabilities & provisions	current liabilities of the customer
23	Deferred tax liability	Future income tax customer will pay because o...
24	Shareholders funds	Amount of equity in a company, which is belon...
25	Cumulative retained profits	Total cumulative profit retained by customer
26	Capital employed	Current asset minus current liabilities
27	TOL/TNW	Total liabilities of the customer divided by ...
28	Total term liabilities / tangible net worth	Short + long term liabilities divided by tang...
29	Contingent liabilities / Net worth (%)	Contingent liabilities / Net worth
30	Contingent liabilities	Liabilities because of uncertain events
31	Net fixed assets	purchase price of all fixed assets
32	Investments	Total invested amount
33	Current assets	Assets that are expected to be converted to c...
34	Net working capital	Difference of current liabilities and current...
35	Quick ratio (times)	Total cash divided by current liabilities
36	Current ratio (times)	Current assets divided by current liabilities
37	Debt to equity ratio (times)	Total liabilities divided by its shareholder ...

38	Cash to current liabilities (times)	Total liquid cash divided by current liabilities
39	Cash to average cost of sales per day	Total cash divided by average cost of the sales
40	Creditors turnover	Net credit purchase divided to average trade ...
41	Debtors turnover	Net credit sales divided by average accounts ...
42	Finished goods turnover	Annual sales divided by average inventory
43	WIP turnover	The cost of goods sold for a period divided b...
44	Raw material turnover	Cost of goods sold is divided by the average ...
45	Shares outstanding	Number of issued shares minus the number of s...
46	Equity face value	cost of the equity at the time of issuing
47	EPS	Net income divided by total number of outstan...
48	Adjusted EPS	Adjusted net earning divided by the weighted ...
49	Total liabilities	Sum of all type of liabilities
50	PE on BSE	Company current stock price divided by its ea...

TABLE 3: DATA DICTIONARY

- The shape of the dataset and the information of number of columns, number of records as well as data type, was verified

Exploratory data Analysis

- The **Duplicated value** check showed that the data do **NOT** have any Duplicate values
- Checking the variables and its distribution on the basis of plot of individual variables.
- Followed by table of Skewness of variables.: High Skewness Detected.

FIGURE 1 : DATA DISTRIBUTION OF VARIABLES

Perform Descriptive Summary Analysis to understand the data

- The descriptive statistics showing 5 point summary is verified

	count	mean	std	min	25%	50%	75%	max
Num	4256	2128.5	1228.75	1	1064.75	2128.5	3192.25	4256
Networth_Next_Year	4256	1344.74	15936.74	-74265.6	3.98	72.1	330.82	805773.4
Total_assets	4256	3573.62	30074.44	0.1	91.3	315.5	1120.8	1176509
Net_worth	4256	1351.95	12961.31	0	31.48	104.8	389.85	613151.6
Total_income	4025	4688.19	53918.95	0	107.1	455.1	1485	2442828
Change_in_stock	3706	43.7	436.92	-3029.4	-1.8	1.6	18.4	14185.5
Total_expenses	4091	4356.3	51398.09	-0.1	96.8	426.8	1395.7	2366035
Profit_after_tax	4102	295.05	3079.9	-3908.3	0.5	9	53.3	119439.1
PBDITA	4102	605.94	5646.23	-440.7	6.93	36.9	158.7	208576.5
PBT	4102	410.26	4217.42	-3894.8	0.8	12.6	74.17	145292.6
Cash_profit	4102	408.27	4143.93	-2245.7	2.9	19.4	96.25	176911.8
PBDITA_as_perc_of_total_income	4177	3.18	172.26	-6400	4.97	9.68	16.47	100
PBT_as_perc_of_total_income	4177	-18.2	419.91	-21340	0.56	3.34	8.94	100
PAT_as_perc_of_total_income	4177	-20.03	423.58	-21340	0.35	2.37	6.42	150
Cash_profit_as_perc_of_total_in	4177	-9.02	299.96	-15020	2	5.66	10.73	100
PAT_as_perc_of_net_worth	4256	10.17	61.53	-748.72	0	8.04	20.2	2466.67
Sales	3951	4645.68	53080.9	0.1	113.35	468.6	1481.2	2384984
Income_from_fincial_services	3145	81.36	1042.76	0	0.5	1.9	9.8	51938.2
Other_income	2700	55.95	1178.42	0	0.4	1.5	6.2	42856.7
Total_capital	4251	224.56	1684.95	0.1	13.2	42.6	103.15	78273.2
Reserves_and_funds	4158	1210.56	12816.23	-6525.9	5.3	55.15	282.52	625137.8
Borrowings	3825	1176.25	8581.25	0.1	24.4	99.8	358.3	278257.3
Current_liabilities_&_provision	4146	960.63	9140.54	0.1	17.5	70.3	265.92	352240.3
Deferred_tax_liability	2887	234.5	2106.25	0.1	3.2	13.5	51.3	72796.6
Shareholders_funds	4256	1376.49	13010.69	0	32.3	107.6	408.9	613151.6
Cumulative_retained_profits	4211	937.18	9853.1	-6534.3	1.1	37.4	206.2	390133.8
Capital_employed	4256	2433.62	20496.4	0	61.3	221.2	790.3	891408.9
TOL_to_TNW	4256	4.03	20.88	-350.48	0.6	1.42	2.83	473
Total_term_liabilities__to__tan	4256	1.85	15.88	-325.6	0.05	0.34	1	456
Contingent_liabilities__to__Net	4256	55.71	369.17	0	0	5.36	31.01	14704.27
Contingent_liabilities	2854	948.55	12056.74	0.1	6	37.85	195.32	559506.8
Net_fixed_assets	4124	1209.49	12502.4	0	26.2	93.85	352.82	636604.6
Investments	2541	721.87	6793.86	0	1	8.2	63.8	199978.6
Current_assets	4176	1350.36	10155.57	0.1	36.6	148.35	515	354815.2
Net_working_capital	4219	162.87	3182.03	-63839	-1.1	16.7	86.5	85782.8
Quick_ratio_times	4151	1.5	9.33	0	0.41	0.67	1.03	341
Current_ratio_times	4151	2.26	12.48	0	0.93	1.23	1.72	505
Debt_to_equity_ratio_times	4256	2.87	15.6	0	0.22	0.79	1.75	456
Cash_to_current_liabilities_tim	4151	0.53	4.8	0	0.02	0.07	0.19	165
Cash_to_average_cost_of_sales_p	4156	145.16	2521.99	0	2.88	8.04	21.97	128040.8
Creditors_turnover	3865	16.81	75.67	0	3.72	6.17	11.69	2401
Debtors_turnover	3871	17.93	90.16	0	3.81	6.47	11.85	3135.2
Finished_goods_turnover	3382	84.37	562.64	-0.09	8.19	17.32	40.01	17947.6
WIP_turnover	3492	28.68	169.65	-0.18	5.1	9.86	20.24	5651.4
Raw_material_turnover	3828	17.73	343.13	-2	3.02	6.41	11.82	21092
Shares_outstanding	3446	23764910	1.71E+08	-2.1E+09	1308383	4750000	10906020	4.13E+09
Equity_face_value	3446	-1094.83	34101.36	-999999	10	10	10	100000
EPS	4256	-196.22	13061.95	-843182	0	1.49	10	34522.53
Adjusted_EPS	4256	-197.53	13061.93	-843182	0	1.24	7.62	34522.53
Total_liabilities	4256	3573.62	30074.44	0.1	91.3	315.5	1120.8	1176509
PE_on_BSE	1629	55.46	1304.45	-1116.64	2.97	8.69	17	51002.74

TABLE 4 DATA DESCRIPTION

Insights of Key variables

Networth_Next_Year: The mean value of 1344.74 suggests that, on average, the net worth of the entities in the dataset for the next year is approximately 1344.74 units. However, the large standard deviation of 15936.74 indicates significant variability in net worth values, as reflected by the wide range between the minimum (-74265.60) and maximum (805773.40) values.
The large standard deviation of 15936.74 indicates significant variability in net worth values. This suggests that there is a wide range of net worth among the entities, with some entities having considerably higher or lower net worth compared to the mean. This variability could be influenced by various factors such as business performance, market conditions, and financial decisions.

Total_assets: With a standard deviation of 30074.44, the variable shows a substantial variation in total assets among the entities. This suggests that some entities have significantly higher or lower total assets compared to the average. The wide dispersion could be due to differences in business size, industry sector, asset composition, or financial strategies.

Net_worth: The standard deviation of 12961.31 indicates variability in net worth values. Some entities may have a relatively higher or lower net worth compared to the mean. The dispersion could be influenced by factors such as profitability, debt levels, capital investments, or changes in asset values.

Total_income: The large standard deviation of 53918.95 suggests a wide range of total income values among the entities. This indicates that some entities have significantly higher or lower total income compared to the average. The variation in total income could be attributed to differences in business models, revenue sources, market conditions, or operational efficiency.

Perform Variable Skewness Analysis

Variable	Skewness
Raw_material_turnover	60.61
Income_from_fincial_services	40.46
Cash_to_average_cost_of_sales_per_day	38.84
Contingent_liabilities	37.76
Net_fixed_assets	37.62
PE_on_BSE	37.2
Networth_Next_Year	36.38
Other_income	35.59
Reserves_and_funds	34.11
Current_ratio_times	33.28
Total_expenses	32.19
Net_worth	31.85

Shareholders_funds	31.55
Total_capital	31.49
Total_income	31.44
Sales	31.23
Capital_employed	28.28
Cumulative_retained_profits	27.82
Cash_profit	27.67
Quick_ratio_times	27.43
Current_liabilities_&_provisions	26.51
Cash_to_current_liabilities_times	26.46
Total_liabilities	26.42
Total_assets	26.42
WIP_turnover	25.69
Contingent_liabilities__to__Net_worth_perc	24.54
Profit_after_tax	24.29
PBDITA	24.12
Deferred_tax_liability	23.74
Debtors_turnover	22.91
PBT	22.28
Current_assets	21.33
Borrowings	20.89
Finished_goods_turnover	20.84
Creditors_turnover	19.72
Investments	19.44
Change_in_stock	18.02
PAT_as_perc_of_net_worth	17.76
Debt_to_equity_ratio_times	16.33
Shares_outstanding	11.03
Total_term_liabilities__to__tangible_net_worth	9.03
TOL_to_TNW	8.89
Net_working_capital	8.84
Num	0
PBDITA_as_perc_of_total_income	-29.03
Equity_face_value	-29.19
Cash_profit_as_perc_of_total_income	-36.02
PAT_as_perc_of_total_income	-37.17
PBT_as_perc_of_total_income	-37.94
EPS	-63.29
Adjusted_EPS	-63.29

TABLE 5 SKEWNESS OF DATA

Data Insights of Skewness:

Skewness is a measure of the asymmetry in the distribution of a variable. It provides insights into the shape and characteristics of the data distribution. Understanding the skewness can help in identifying potential data transformation needs or selecting appropriate statistical

methods for analysis. The skewness indicates the degree of skewness in each variable. Positive skewness indicates a right-skewed distribution, while negative skewness indicates a left-skewed distribution.Here are some insights based on the skewness values:

Positive Skewness (Right-skewed distribution):

The distribution of Data from 60.61 to 8.64 having Positive skewness. Which indicates a longer or fatter tail on the right side of the distribution. _

Negative Skewness (Left-skewed distribution):

The distribution of Data from 29.03 to 63.29 having Negative skewness This indicates that the majority of the data points are concentrated on the right side of the distribution, with a long tail on the left side. The skewness basically indicates the probable presence of Lower (Negative skewness) & higher (Positive Skewness) Outliers or Values in data set.

Dependent variable Creation on variable Net Worth Next Year

We need to create a default variable that should take a value of 1 when the net worth next year is negative and 0 when the net worth next year is positive.

Top 5 Data				Botton 5 Data		
	Default	Networth_Next_Year			Default	Networth_Next_Year
0	0	395.3		4251	0	0.2
1	0	36.2		4252	0	93.3
2	0	84		4253	0	932.2
3	0	2041.4		4254	0	64.6
4	0	41.8		4255	1	0

TABLE 6 DEFAULT VALUE DISTRIBUTION

Percentage Distribution of Target Valuable 'Default'

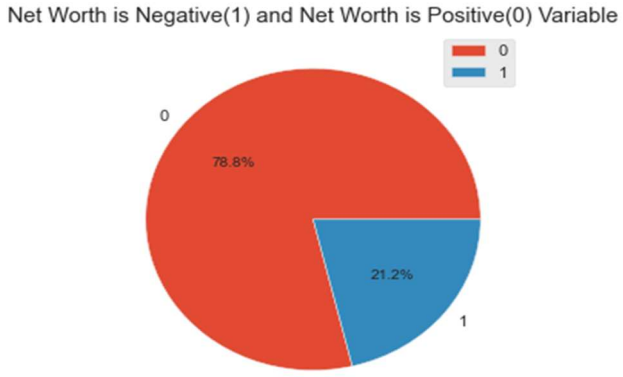

FIGURE 2 : PERCENT DISTRIBUTION OF TARGET VARIABLE

Univariate Analysis

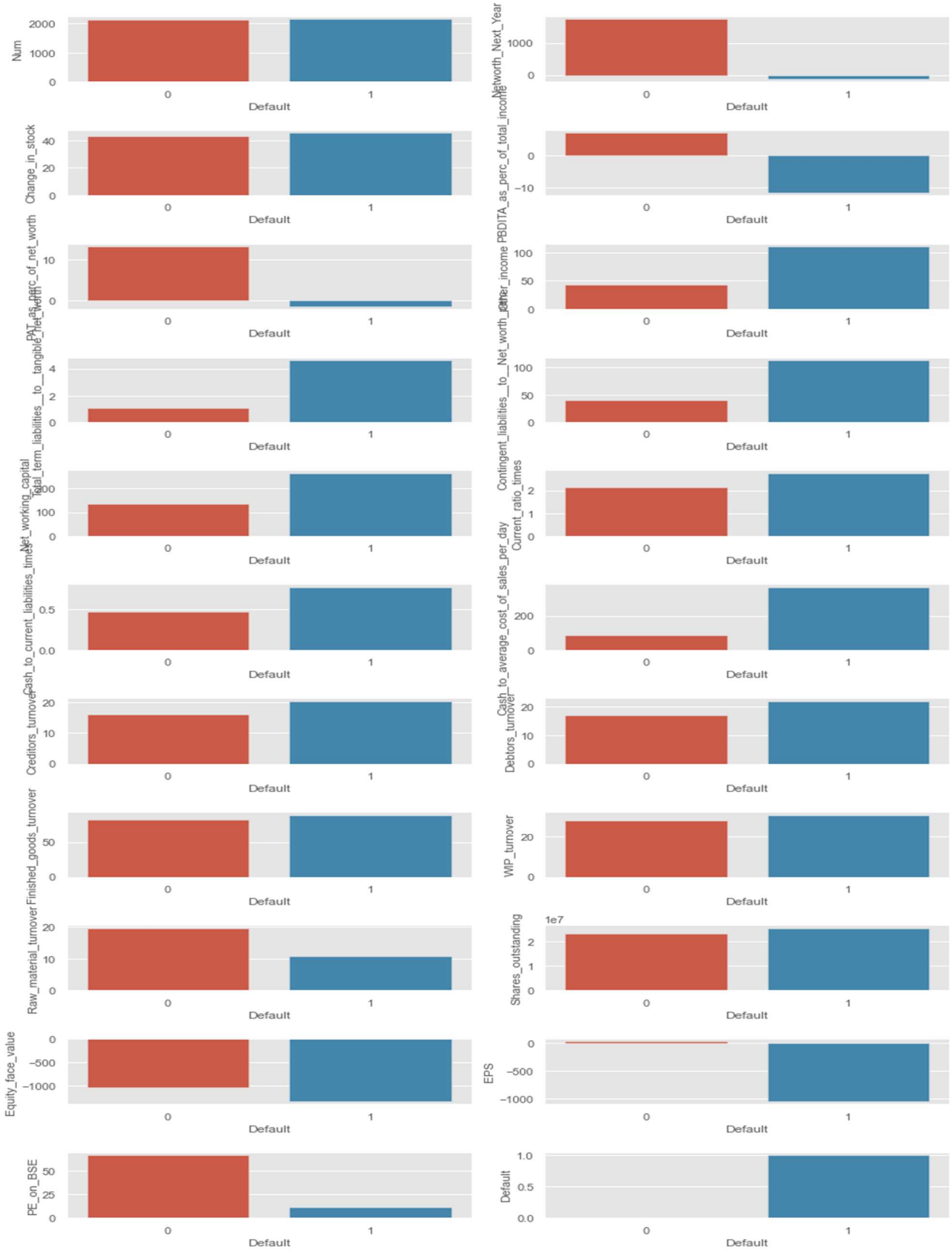

FIGURE 3 : UNIVARIATE ANALYSIS WITH DEFAULT

BIVARIATE ANALYSIS : ANALYSING NETWORTH_NEXT_YEAR VARIABLE, SINCE ON THE BASIS OF THIS VARIABLE WE WILL DEFINE THE DEPENDENT VARIABLE

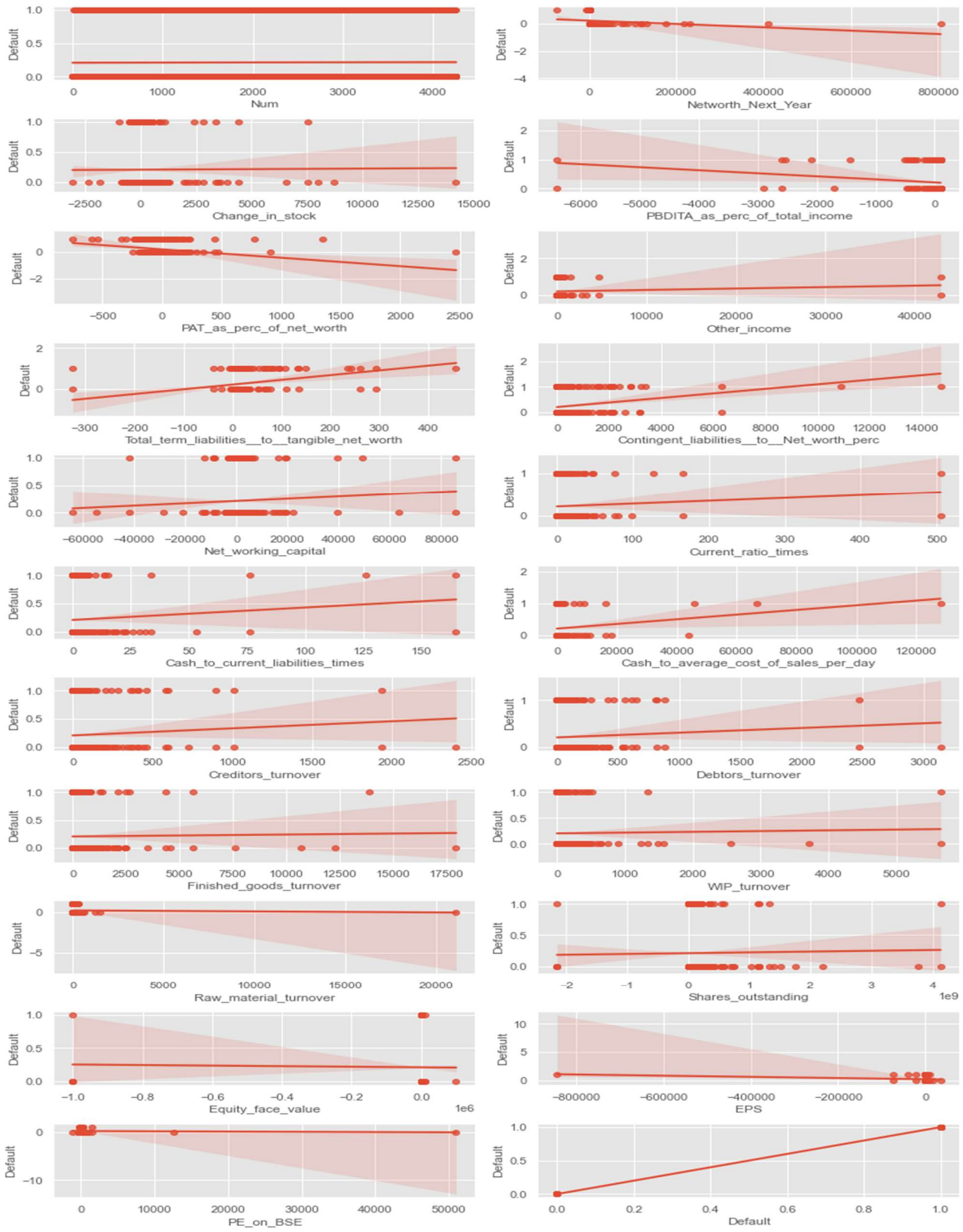

FIGURE 4 BIVARIATE ANALYSIS WITH DEFAULT

Multivariate Analysis

We also performed multi variate analysis on the data to see if there are any correlation that are observed within the data. Correlation function was used and seaborn heatmap is using to plot the correlations and to make better sense of the data.
We observed that net worth and net worth next year were highly correlated. Apart from this, we also found various variables were highly correlated. This analysis will illustrate the collinearity with the data set

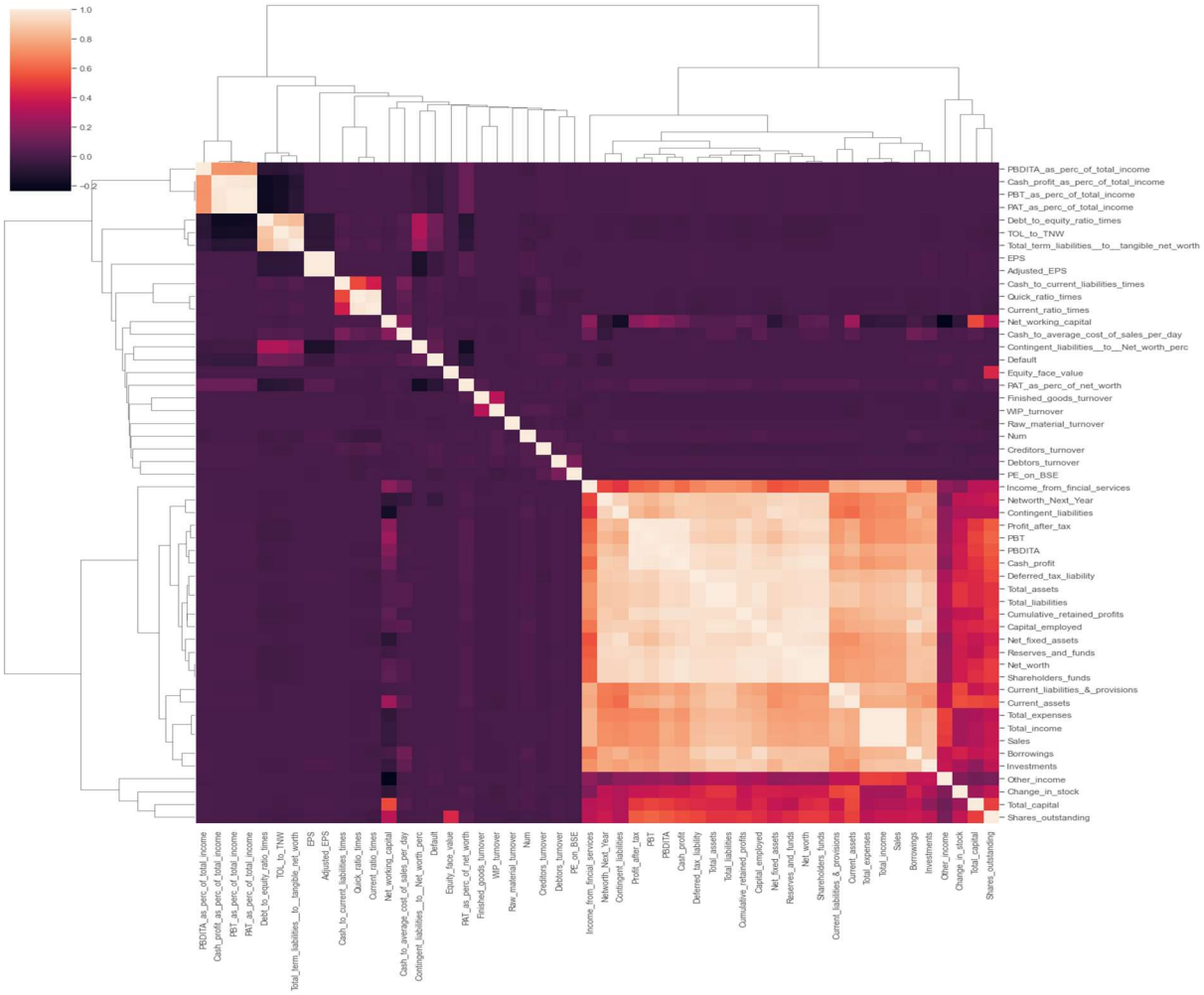

FIGURE 5 CLUSTER MAP ANALYSIS WITH DEFAULT

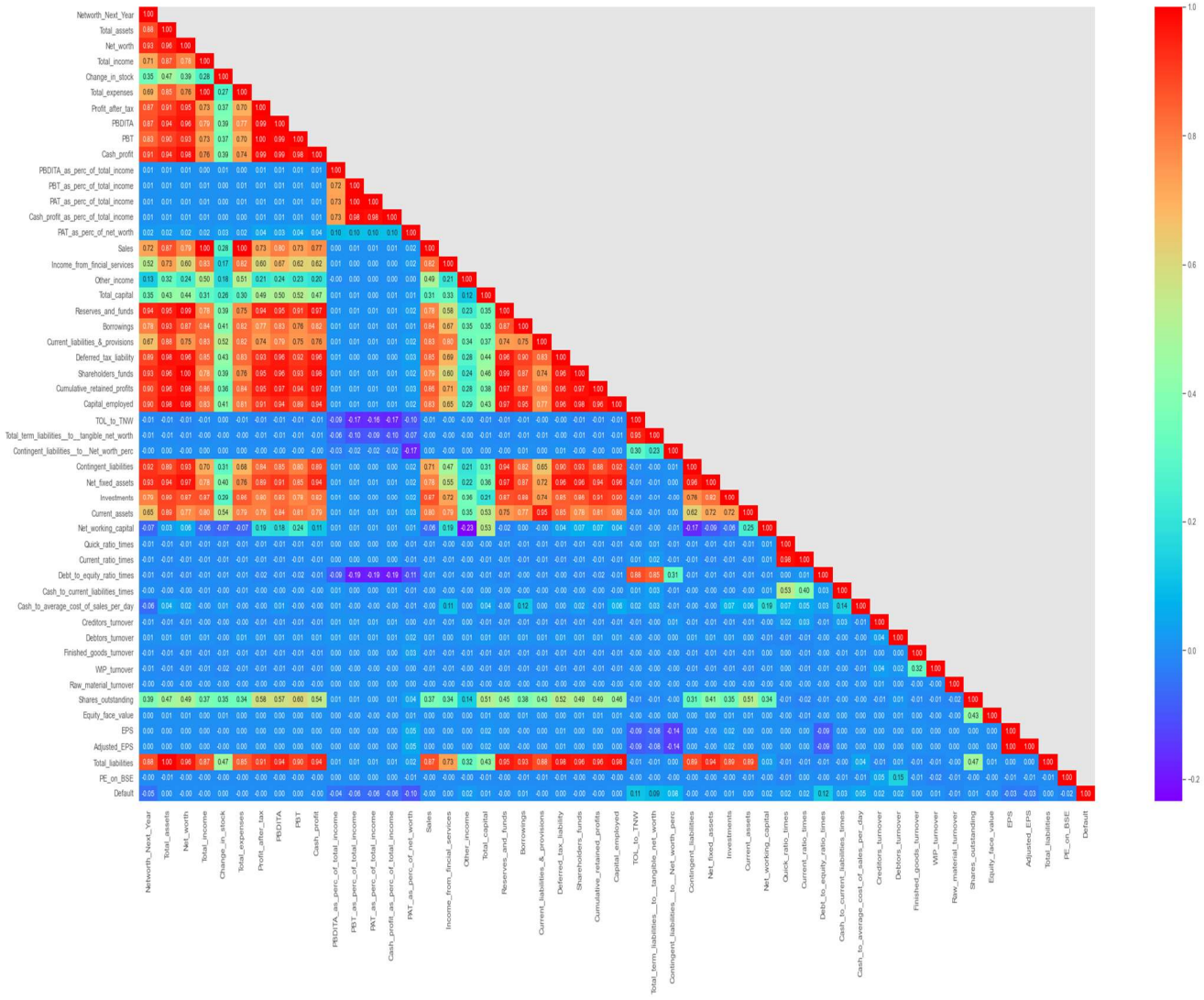

FIGURE 6 MULTIVARIATE ANALYSIS

Dropping the columns like from Dataset:

['Total_liabilities','Current_assets','Investments','Net_fixed_assets','Contingent_liabilities','Capital_employed','Cumulative_retained_profits','Shareholders_funds','Deferred_tax_liability','Current_liabilities_&_provisions','Borrowings','Reserves_and_funds','Income_from_fincial_services','Sales','Cash_profit_as_perc_of_total_income','PAT_as_perc_of_total_income','PBT_as_perc_of_total_income','PBDITA','PBT','Cash_profit','Profit_after_tax','Total_expenses','Total_income','Net_worth','Total_assets','Quick_ratio_times','Adjusted_EPS','Debt_to_equity_ratio_times','TOL_to_TNW','Total_capital']

Purpose: To clear the Multicollinearity of dataset

Data Information after dropping of variables the new data frame has been come in existence. Checking the shape and size of data with data types. Also analysing prima facie on missing values.

```
class 'pandas.core.frame.DataFrame'>
RangeIndex: 4256 entries, 0 to 4255
Data columns (total 22 columns):
 #   Column                                    Non-Null Count        Dtype
---  ------                                    --------------  -----
 0   Num                                       4256 non-null   int64
 1   Networth_Next_Year                        4256 non-null   float64
 2   Change_in_stock                           3706 non-null   float64
 3   PBDITA_as_perc_of_total_income            4177 non-null   float64
 4   PAT_as_perc_of_net_worth                  4256 non-null   float64
 5   Other_income                              2700 non-null   float64
 6   Total_term_liabilities__to__tangible_net_worth  4256 non-null   float64
 7   Contingent_liabilities__to__Net_worth_perc      4256 non-null   float64
 8   Net_working_capital                       4219 non-null   float64
 9   Current_ratio_times                       4151 non-null   float64
 10  Cash_to_current_liabilities_times         4151 non-null   float64
 11  Cash_to_average_cost_of_sales_per_day     4156 non-null   float64
 12  Creditors_turnover                        3865 non-null   float64
 13  Debtors_turnover                    3     871 non-null   float64
 14  Finished_goods_turnover                   3382 non-null   float64
 15  WIP_turnover                              3492 non-null   float64
 16  Raw_material_turnover                     3828 non-null   float64
 17  Shares_outstanding                        3446 non-null   float64
 18  Equity_face_value                         3446 non-null   float64
 19  EPS                                       4256 non-null   float64
 20  PE_on_BSE                                 1629 non-null   float64
 21  Default                                   4 256 non-null   int32
dtypes: float64(20), int32(1), int64(1)
memory usage: 715.0 KB
```

TABLE 7 NEW DATA SET WITH DROPPED VARIABLES

Purpose of creating the Heat map again to ensure the clearance of Multicollinearity for further analysis and modelling.

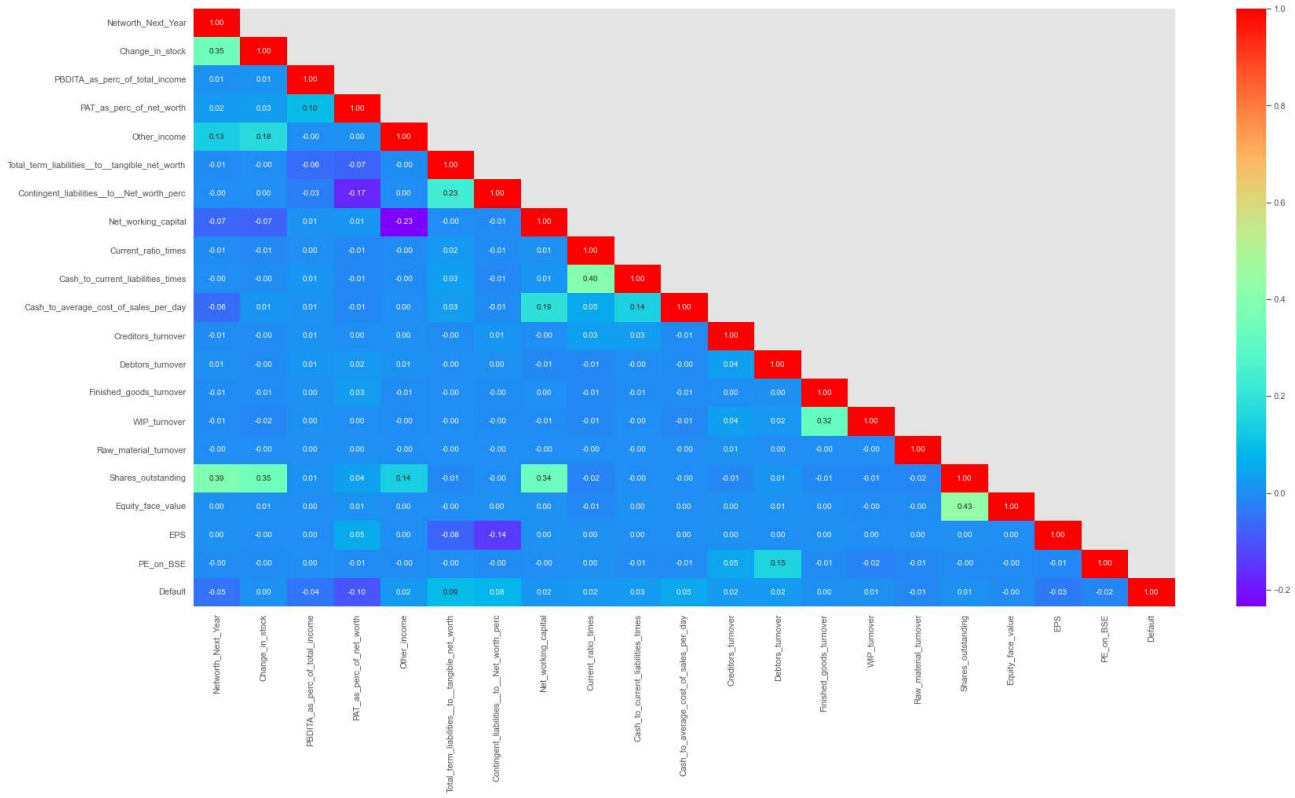

FIGURE 7 HEAT MAP AFTER CORRELATION VARIABLE DROP

Correlation of Data with Default

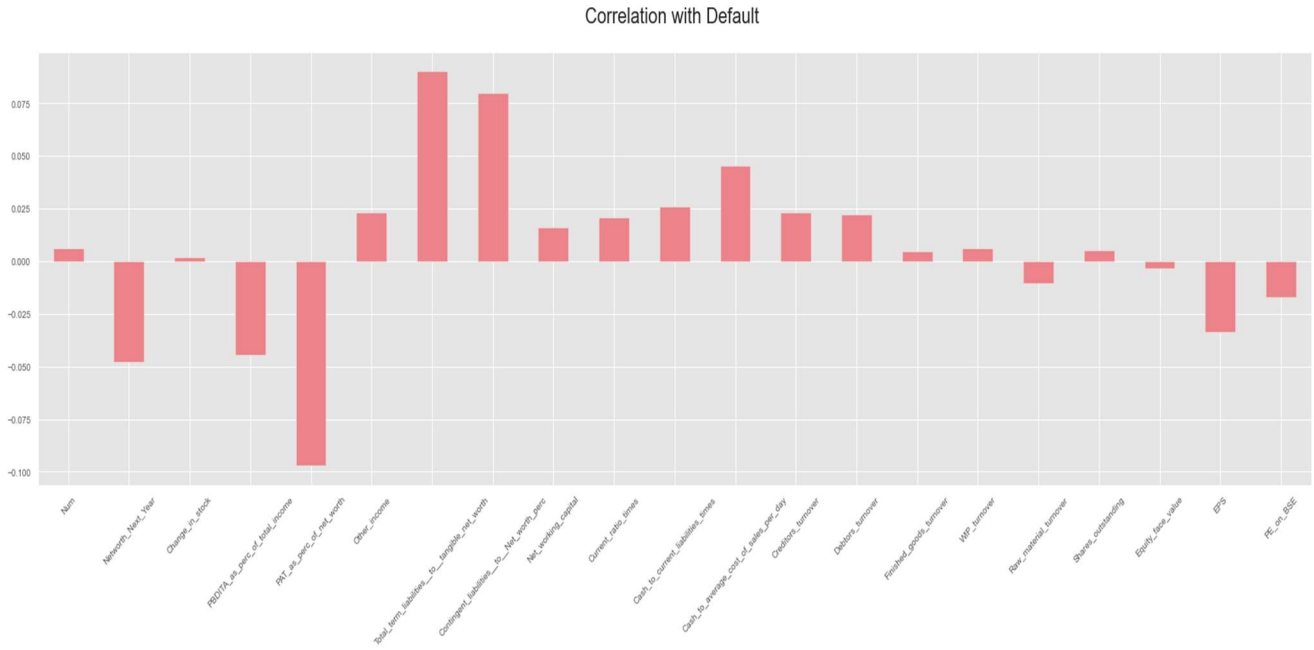

TABLE 8 CORRELATION WITH DEFAULT

Outlier Detection:

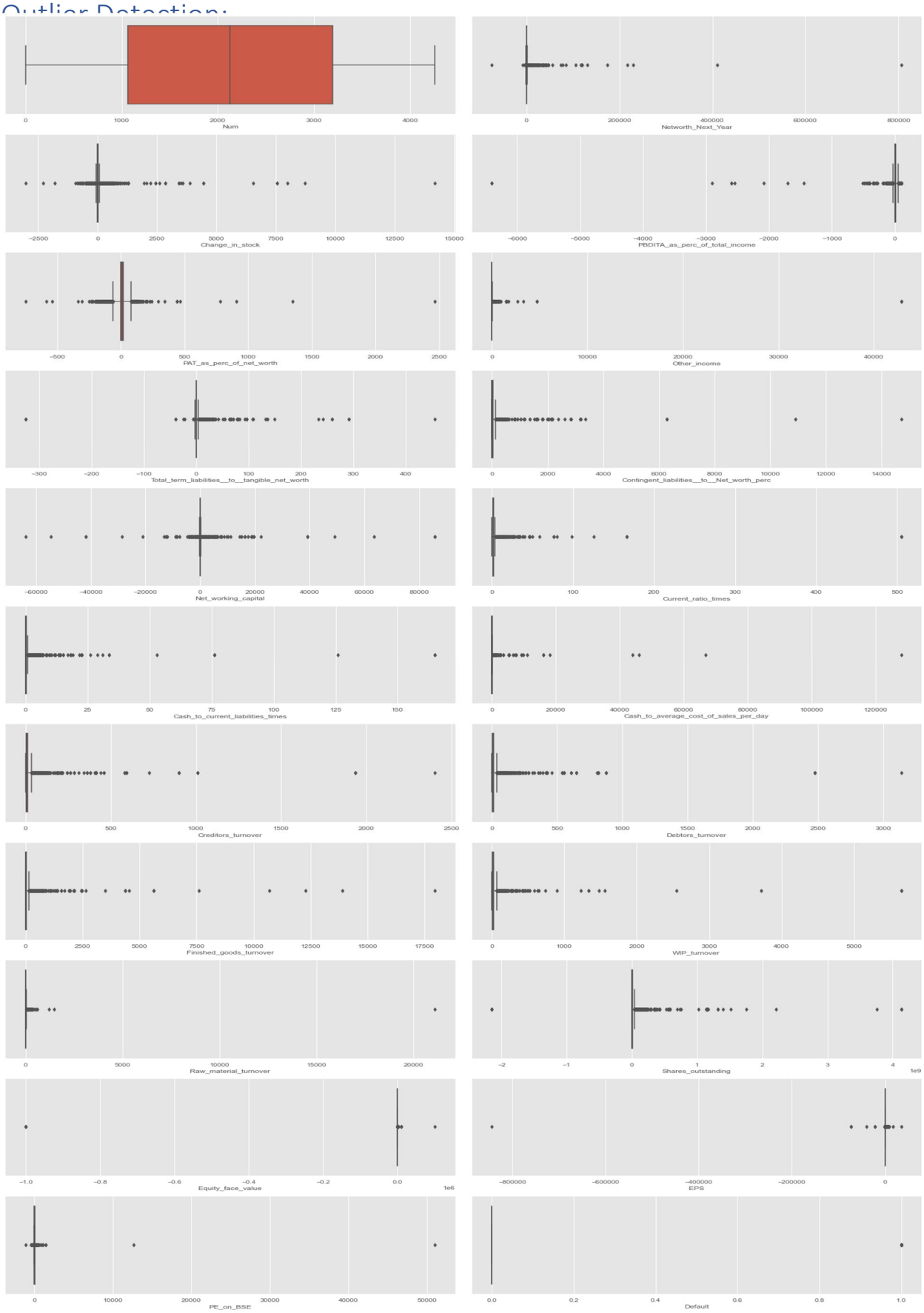

FIGURE 8 CORRELATION WITH DEFAULT

Insight on Outliers:

There is presence of both lower and upper outlier with each variable. Some variables like current ratio times, cost to avarage cost of sales per day,cash to current liability times,Creditors turn over,finished goods turnover,WIP turnover,Share outstandings,Debitor turnover,Contigent libility to networth perc

Missing Value Analysis

PE_on_BSE	0.62
Other_income	0.37
Finished_goods_turnover	0.21
Equity_face_value	0.19
Shares_outstanding	0.19
WIP_turnover	0.18
Change_in_stock	0.13
Raw_material_turnover	0.10
Creditors_turnover	0.09
Debtors_turnover	0.09
Cash_to_current_liabilities_times	0.02
Current_ratio_times	0.02
Cash_to_average_cost_of_sales_per_day	0.02
PBDITA_as_perc_of_total_income	0.02
Net_working_capital	0.01
Networth_Next_Year	0.00
Contingent_liabilities__to__Net_worth_perc	0.00
Total_term_liabilities__to__tangible_net_worth	0.00
PAT_as_perc_of_net_worth	0.00
EPS	0.00
Num	0.00

dtype: float64

TABLE 9 MISSING VALUE PERCENTAGE

Outlier analysis

Q1 = df_X.quantile(0.05)
Q3 = df_X.quantile(0.95)
IQR = Q3 - Q1
UL = Q3 + 3*IQR
LL = Q1 - 3*IQR

Insights:

Q1: It represents the first quartile, which is the value below which 25% of the data falls. It is calculated by finding the value that separates the lowest 25% of the data from the upper 75% of the data.

Q3: It represents the third quartile, which is the value below which 75% of the data falls. It is calculated by finding the value that separates the lowest 75% of the data from the upper 25% of the data.

IQR: It stands for the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). It provides a measure of the spread or dispersion of the data.

UL: It represents the upper limit for outlier detection. It is calculated by adding 3 times the IQR to the third quartile (Q3). Any data point above this upper limit is considered a potential outlier.

LL: It represents the lower limit for outlier detection. It is calculated by subtracting 3 times the IQR from the first quartile (Q1). Any data point below this lower limit is considered a potential outlier.

By using the more stringent threshold of 3 times the IQR, values that are even further from the central distribution of the data are flagged as potential outliers. Using this approach because in this case the stricter outlier detection is desired, datasets having the extreme values or high variability.

OUTLIERS

```
EPS                                                104
Net_working_capital                                 83
Total_term_liabilities__to__tangible_net_worth      71
Cash_to_average_cost_of_sales_per_day               61
Networth_Next_Year                                  57
Current_ratio_times                                 53
Cash_to_current_liabilities_times                   52
Contingent_liabilities__to__Net_worth_perc          50
Finished_goods_turnover                             47
Shares_outstanding                                  47
Other_income                                        46
Equity_face_value                                   45
Debtors_turnover                                    45
Change_in_stock                                     41
Creditors_turnover                                  38
PBDITA_as_perc_of_total_income                      36
WIP_turnover                                        30
Raw_material_turnover                               30
PE_on_BSE                                           20
PAT_as_perc_of_net_worth                            13
Num                                                  0
dtype: int64
```

TABLE 10 OUTLIER PER VARIABLE

We are converting all outlier as nan values , the variables having nan values are:

PE_on_BSE	0.62
Other_income	0.38
Finished_goods_turnover	0.22
Shares_outstanding	0.20
Equity_face_value	0.20
WIP_turnover	0.19
Change_in_stock	0.14
Raw_material_turnover	0.11
Debtors_turnover	0.10
Creditors_turnover	0.10
Cash_to_average_cost_of_sales_per_day	0.04
Current_ratio_times	0.04
Cash_to_current_liabilities_times	0.04
Net_working_capital	0.03
PBDITA_as_perc_of_total_income	0.03
EPS	0.02
Total_term_liabilities__to__tangible_net_worth	0.02
Networth_Next_Year	0.01
Contingent_liabilities__to__Net_worth_perc	0.01
PAT_as_perc_of_net_worth	0.00
Num	0.00

TABLE 11 PERCENT NULL OF OUTLIERS

The Plot of Variables with consolidated missing values

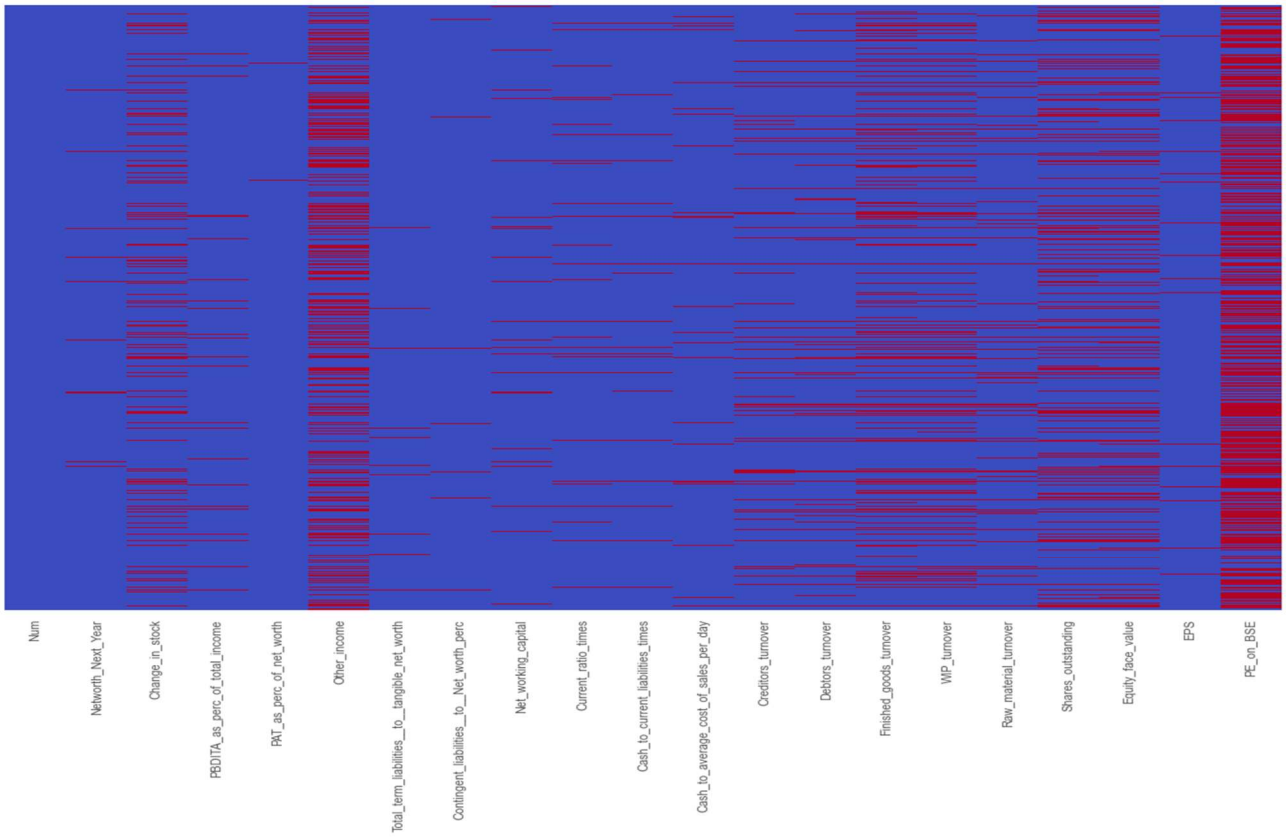

FIGURE 9 VISUALISATION OF MISSING VALUES

Variables After Dropping ['Num','PE_on_BSE','Other_income','Networth_Next_Year',]

```
<class 'pandas.core.frame.DataFrame'>
Range Index: 4256 entries, 0 to 4255
Data columns (total 17 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Change_in_stock                               3665 non-null   float64
 1   PBDITA_as_perc_of_total_income                4141 non-null   float64
 2   PAT_as_perc_of_net_worth                      4243 non-null   float64
 3   Total_term_liabilities__to__tangible_net_worth  4185 non-null   float64
 4   Contingent_liabilities__to__Net_worth_perc    4206 non-null   float64
 5   Net_working_capital                           4136 non-null   float64
 6   Current_ratio_times                           4098 non-null   float64
 7   Cash_to_current_liabilities_times             4099 non-null   float64
 8   Cash_to_average_cost_of_sales_per_day         4095 non-null   float64
 9   Creditors_turnover                            3827 non-null   float64
 10  Debtors_turnover                              3826 non-null   float64
 11  Finished_goods_turnover                       3335 non-null   float64
 12  WIP_turnover                                  3462 non-null   float64
 13  Raw_material_turnover                         3798 non-null   float64
 14  Shares_outstanding                            3399 non-null   float64
 15  Equity_face_value                             3401 non-null   float64
 16  EPS                                           4152 non-null   float64
dtypes: float64(17)
memory usage: 565.4 KB
```

TABLE 12 DATA INFO AFTER DROPPING NET WORTH NEXT YEAR

The consolidated (Outlier+Actual Missing) missing value in percentage

Finished_goods_turnover	0.22
Shares_outstanding	0.20
Equity_face_value	0.20
WIP_turnover	0.19
Change_in_stock	0.14
Raw_material_turnover	0.11
Debtors_turnover	0.10
Creditors_turnover	0.10
Cash_to_average_cost_of_sales_per_day	0.04
Current_ratio_times	0.04
Cash_to_current_liabilities_times	0.04
Net_working_capital	0.03
PBDITA_as_perc_of_total_income	0.03
EPS	0.02
Total_term_liabilities__to__tangible_net_worth	0.02
Contingent_liabilities__to__Net_worth_perc	0.01
PAT_as_perc_of_net_worth	0.00
Default	0.00

TABLE 13 CONSOLIDATED (OUTLIER +ACTUAL) NULL PERCENTAGE

Insights

Since the outliers are too large in the number.it will affect the model. But Also given the fact that this is a financial data and the outliers might very well reflect the information which is genuine in nature. Since data captured from different size of companies

Although most outliers have nan values which is a missing data which should be treated with missing data imputation method. here KNN imputation method is used

Scale the predictors

It can also be a good idea to scale the target variable for regression predictive modelling problems to make the problem easier to learn. A target variable with a large spread of values, in turn, may result in large error gradient values causing weight values to change dramatically, making the learning process unstable.

Scaling input and output variables is a critical step in regression

Here StandardScaler is used for preprocessing the data. We will use the default configuration and scale values to subtract the mean to centre them on 0.0 and divide by the standard deviation to give the standard deviation of 1.0. First, a StandardScaler instance is defined with default hyperparameters.

Once defined, we can call the fit transform function and pass it to our dataset to create a transformed version of our dataset.

Imputing the remaining missing values by KNN

For this dataset, converted the outlier to null values so that both can be treat together using KNN imputer. KNN method is to identify 'k' samples in the dataset that are similar to other data points. Then using these 'k' samples to estimate the value of the missing data. Each missing values are imputed using the mean value of the k-neighbours found in the dataset.

Change_in_stock	0
PBDITA_as_perc_of_total_income	0
PAT_as_perc_of_net_worth	0
Total_term_liabilities__to__tangible_net_worth	0
Contingent_liabilities__to__Net_worth_perc	0
Net_working_capital	0
Current_ratio_times	0
Cash_to_current_liabilities_times	0
Cash_to_average_cost_of_sales_per_day	0
Creditors_turnover	0
Debtors_turnover	0
Finished_goods_turnover	0
WIP_turnover	0
Raw_material_turnover	0
Shares_outstanding	0
Equity_face_value	0
EPS	0
Default	0

TABLE 14 NULL VALUES AFTER KNN IMPUTATION

Now the dataset is clear with all type of nan values across the variables.

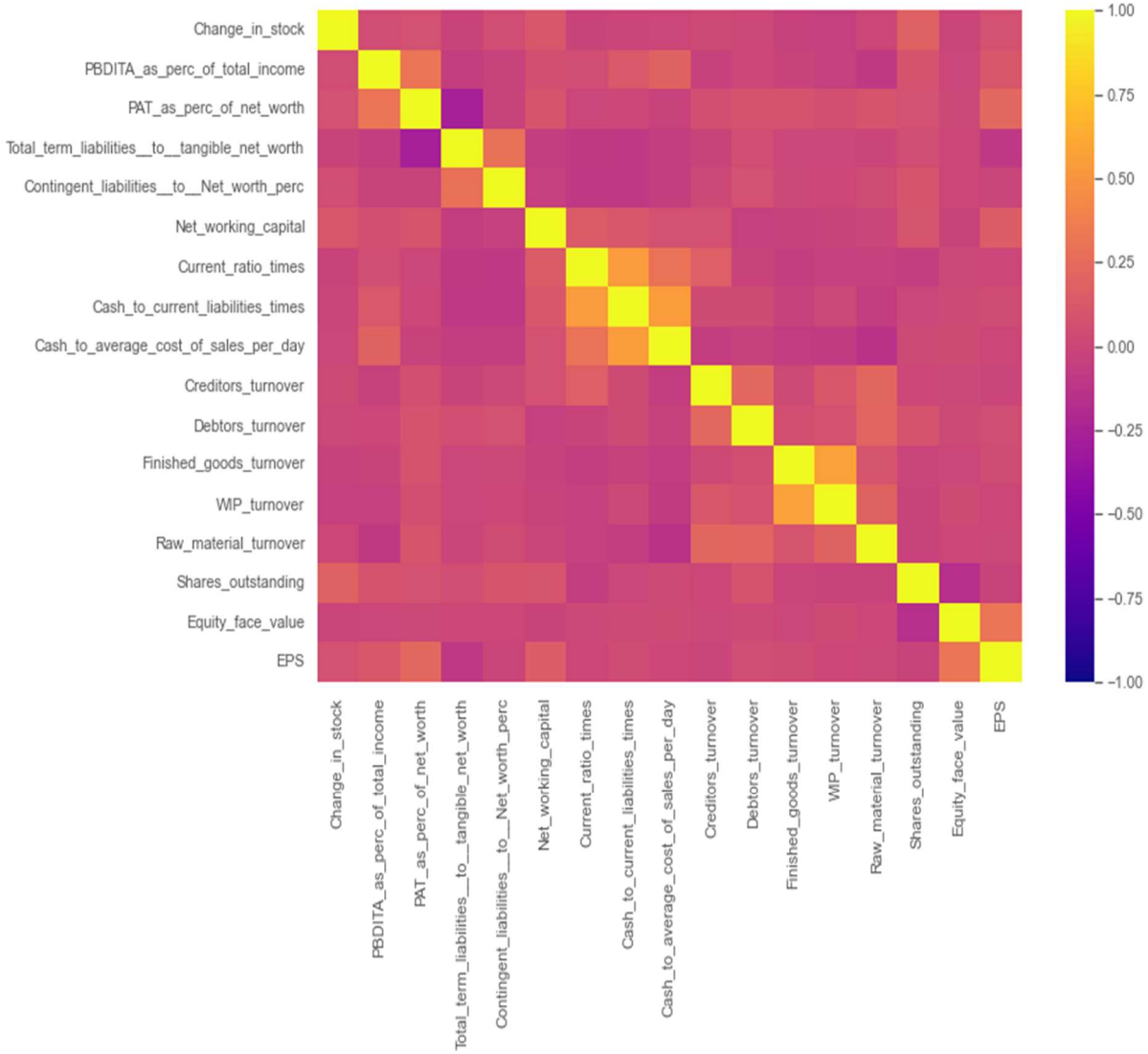

FIGURE 10 CORRELATION PLOT WITH DEFAULT

Feature Selection

Since there are too many columns, we need to determine the columns which are related and eliminate them if possible. We will use VIF to determine the collinearity and eliminate using a threshold of 5.

variables		VIF
7	Cash_to_current_liabilities_times	1.87
12	WIP_turnover	1.55
6	Current_ratio_times	1.51
11	Finished_goods_turnover	1.50
8	Cash_to_average_cost_of_sales_per_day	1.49
2	PAT_as_perc_of_net_worth	1.30
16	EPS	1.21
3	Total_term_liabilities__to__tangible_net_worth	1.20
1	PBDITA_as_perc_of_total_income	1.18
9	Creditors_turnover	1.16
15	Equity_face_value	1.14
13	Raw_material_turnover	1.14
10	Debtors_turnover	1.12
4	Contingent_liabilities__to__Net_worth_perc	1.11
14	Shares_outstanding	1.10
5	Net_working_capital	1.09
0	Change_in_stock	1.05
17	Default	1.04

TABLE 15 FEATURES WITH VIF VALUES

Train Test Split (67:33 ratio)

For the given business problem, 'default' is the target variable since the problem is to come up with a model to predict whether a particular company will default or not.

```
X – Independent variable (Removing 'default' variable)

y – Dependent/ Target variable (Having only 'default' variable)
```

Next step is to Split the data into training and testing test. Splitting the data as 67% training and 33% testing with a random state = 42 and stratify = y.
Output of this step will be: Training independent variable (X_train),
Testing independent variable (X-test), Training dependent variable (y_train) and testing dependent variable (y_test).

Description of Train Test Split

Train_test_split function from scikit-learn to split the data into training and testing sets. The test size parameter is set to 0.33, which means that approximately 33% of the data will be allocated for testing, and the remaining 67% will be used for training.

The stratify parameter is set to y, which means that the data will be stratified based on the values of the y variable. This is useful due to dealing with imbalanced datasets, as it helps maintain the distribution of classes in both the training and testing sets. By setting stratify=y, we ensure that the proportion of different classes in the original dataset is preserved in the training and testing sets.

The random_state parameter is set to 42, which is used to ensure reproducibility. By setting a specific random state, the data will be split in the same way each time the code is run, making the results consistent

Resampling using SMOTE

Dataset Shape and Target variables analysis before SMOTE

```
Before Oversampling the shape of X: (4256, 17)
Before oversampling the shape of y: (4256,)
Before oversampling, counts of label '1': 904
Before oversampling, counts of label '0': 3352
```

TABLE 16 DATA SHAPE BEFORE RESAMPLING

```
# Dataset Shape and Target variables analysis After SMOTE

After oversampling the shape of X: (6704, 17)
After oversampling the shape of y: (6704,)
After oversampling, counts of label '1': 3352
After oversampling, counts of label '0': 3352
```

TABLE 17 DATASHAPE AFTER RESAMPLING

Checking shape of X and y

Shape of df train and Test

```
The shape of X: (2851, 18)
The shape of y: (1405, 18)
```

TABLE 18 DATASHAPE X,Y BEFORE RESAMPLING

Shape of Resampled train and Test data

```
The shape of X: (4491, 18)
The shape of y: (2213, 18)
```

TABLE 19 DATASHAPE X,Y AFTER RESAMPLING

Modelling and tuning of Models

Model Number 1 with top 5 p value variables

```
Optimization terminated successfully.
         Current function value: 0.495117
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 2851
Model:                          Logit   Df Residuals:                     2833
Method:                           MLE   Df Model:                           17
Date:                Sat, 17 Jun 2023   Pseudo R-squ.:                0.04293
Time:                        16:29:35   Log-Likelihood:                -1411.6
converged:                       True   LL-Null:                       -1474.9
Covariance Type:            nonrobust   LLR p-value:                 8.321e-19
================================================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                     -1.3697      0.048    -28.505      0.000      -1.464      -1.276
Change_in_stock                                0.0636      0.050      1.263      0.207      -0.035       0.162
PBDITA_as_perc_of_total_income                -0.0423      0.049     -0.859      0.390      -0.139       0.054
PAT_as_perc_of_net_worth                      -0.3377      0.056     -6.031      0.000      -0.447      -0.228
Total_term_liabilities__to__tangible_net_worth 0.2043      0.047      4.336      0.000       0.112       0.297
Contingent_liabilities__to__Net_worth_perc    -0.0241      0.048     -0.505      0.614      -0.118       0.070
Net_working_capital                           -0.0540      0.050     -1.075      0.282      -0.152       0.044
Current_ratio_times                           -0.1547      0.063     -2.469      0.014      -0.277      -0.032
Cash_to_current_liabilities_times              0.1105      0.063      1.755      0.079      -0.013       0.234
Cash_to_average_cost_of_sales_per_day          0.0148      0.051      0.290      0.772      -0.086       0.115
Creditors_turnover                             0.0202      0.051      0.392      0.695      -0.081       0.121
Debtors_turnover                               0.0112      0.050      0.224      0.823      -0.087       0.109
Finished_goods_turnover                       -0.1223      0.075     -1.625      0.104      -0.270       0.025
WIP_turnover                                   0.0825      0.066      1.241      0.215      -0.048       0.213
Raw_material_turnover                         -0.0806      0.056     -1.429      0.153      -0.191       0.030
Shares_outstanding                            -0.0312      0.055     -0.564      0.573      -0.140       0.077
Equity_face_value                             -0.0156      0.054     -0.289      0.773      -0.121       0.090
EPS                                            0.0583      0.051      1.152      0.249      -0.041       0.158
================================================================================================================
```

Out[1273]:

```
Debtors_turnover                          0.82
Equity_face_value                         0.77
Cash_to_average_cost_of_sales_per_day     0.77
Creditors_turnover                        0.69
Contingent_liabilities__to__Net_worth_perc 0.61
dtype: float64
```

TABLE 20 STATS MODEL NUMBER 1

Model Number 2

Optimization terminated successfully.
Current function value: 0.495573
Iterations 6

Logit Regression Results

Dep. Variable:	Default	No. Observations:	2851
Model:	Logit	Df Residuals:	2837
Method:	MLE	Df Model:	13
Date:	Sat, 17 Jun 2023	Pseudo R-squ.:	0.04205
Time:	16:29:38	Log-Likelihood:	-1412.9
converged:	True	LL-Null:	-1474.9
Covariance Type:	nonrobust	LLR p-value:	3.206e-20

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3668	0.048	-28.523	0.000	-1.461	-1.273
Change_in_stock	0.0567	0.048	1.172	0.241	-0.038	0.151
PBDITA_as_perc_of_total_income	-0.0413	0.049	-0.841	0.400	-0.138	0.055
PAT_as_perc_of_net_worth	-0.3303	0.055	-6.050	0.000	-0.437	-0.223
Total_term_liabilities__to__tangible_net_worth	0.2022	0.047	4.317	0.000	0.110	0.294
Contingent_liabilities__to__Net_worth_perc	-0.0244	0.048	-0.511	0.610	-0.118	0.069
Current_ratio_times	-0.1606	0.063	-2.563	0.010	-0.283	-0.038
Cash_to_current_liabilities_times	0.1102	0.063	1.754	0.079	-0.013	0.233
Cash_to_average_cost_of_sales_per_day	0.0119	0.051	0.231	0.817	-0.089	0.113
Creditors_turnover	0.0183	0.050	0.365	0.715	-0.080	0.117
Finished_goods_turnover	-0.1182	0.075	-1.568	0.117	-0.266	0.030
WIP_turnover	0.0798	0.066	1.201	0.230	-0.050	0.210
Raw_material_turnover	-0.0772	0.056	-1.384	0.166	-0.186	0.032
Equity_face_value	0.0076	0.051	0.151	0.880	-0.092	0.107

TABLE 21 STATS MODEL NUMBER 2

```
Equity_face_value                          0.88
Cash_to_average_cost_of_sales_per_day      0.82
Creditors_turnover                         0.72
Contingent_liabilities__to__Net_worth_perc 0.61
PBDITA_as_perc_of_total_income             0.40
dtype: float64
```

TABLE 22 TOP 5 P VALUE VARIABLE OF MODEL 2

Model Summary Insights:

Optimization terminated successfully: This indicates that the optimization algorithm used to estimate the model parameters converged without any issues.
Current function value: 0.501690: This represents the value of the objective function (log-likelihood) at the estimated parameter values. In this case, the value is 0.501690.
Iterations 6: It took 6 iterations for the optimization algorithm to converge and find the estimated parameter values.
Logit Regression Results: This section provides a summary of the logistic regression model.
Dep. Variable: Default: This is the dependent variable or the target variable in the logistic regression model.
No. Observations: 2851: This is the number of observations used to estimate the model.
Model: Logit: Indicates that the model used is a logistic regression model.
Df Residuals: 2839: Represents the degrees of freedom of the residuals.
Method: MLE (Maximum Likelihood Estimation): Indicates that the model was estimated using maximum likelihood estimation.
converged: True: Indicates that the optimization algorithm converged and found a solution.
LL-Null: -1491.7: Log-likelihood value of the null model (a model with no predictors). The lower the value, the better the model fit.
Pseudo R-square.: 0.04118: Represents the pseudo-R-squared value, which measures the proportion of variance explained by the model. In this case, the model explains about 4.12% of the variance.
Log-Likelihood: -1430.3: Log-likelihood value of the current model. The higher the value, the better the model fit.
Covariance Type: no robust: Indicates the type of covariance used in the model estimation.
LLR p-value: 4.843e-21: Represents the p-value of the likelihood ratio test, which compares the current model to the null model. A low p-value suggests that the current model significantly improves upon the null model.
Coefficients: This table provides the estimated coefficients for each predictor variable in the logistic regression model. Here's how to interpret the columns:
coefficients: The estimated coefficient for each predictor variable.
std err: The standard error associated with each coefficient estimate.
z: The z-statistic, which is the coefficient divided by its standard error. It measures the number of standard deviations the coefficient is away from zero.
P>|z|: The p-value associated with the z-statistic. It indicates the statistical significance of each coefficient. A p-value less than the chosen significance level (e.g., 0.05) suggests that the predictor variable is significantly related to the outcome variable.
[0.025, 0.975]: The 95% confidence interval for each coefficient estimate. It provides a range of plausible values for the true population coefficient.
Each predictor variable in the model is listed along with its estimated coefficient, standard error, z-statistic, p-value, and confidence interval. These values help assess the significance and direction of the relationship between each predictor variable and the probability of the "Default" outcome.

Model number 3

Optimization terminated successfully.
Current function value: 0.496178
Iterations 6

Logit Regression Results

Dep. Variable:	Default	No. Observations:	2851
Model:	Logit	Df Residuals:	2840
Method:	MLE	Df Model:	10
Date:	Sat, 17 Jun 2023	Pseudo R-squ.:	0.04088
Time:	16:29:40	Log-Likelihood:	-1414.6
converged:	True	LL-Null:	-1474.9
Covariance Type:	nonrobust	LLR p-value:	3.863e-21

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3639	0.048	-28.536	0.000	-1.458	-1.270
PBDITA_as_perc_of_total_income	-0.0376	0.049	-0.767	0.443	-0.134	0.058
PAT_as_perc_of_net_worth	-0.3425	0.056	-6.151	0.000	-0.452	-0.233
Total_term_liabilities__to__tangible_net_worth	0.1951	0.044	4.392	0.000	0.108	0.282
Current_ratio_times	-0.1566	0.062	-2.534	0.011	-0.278	-0.035
Cash_to_current_liabilities_times	0.1195	0.055	2.173	0.030	0.012	0.227
Debtors_turnover	0.0128	0.048	0.264	0.791	-0.082	0.108
WIP_turnover	0.0124	0.053	0.233	0.815	-0.092	0.117
Raw_material_turnover	-0.0764	0.055	-1.385	0.166	-0.185	0.032
Shares_outstanding	-0.0188	0.053	-0.355	0.723	-0.122	0.085
EPS	0.0470	0.047	0.992	0.321	-0.046	0.140

```
WIP_turnover                    0.82
Debtors_turnover                0.79
Shares_outstanding              0.72
PBDITA_as_perc_of_total_income  0.44
EPS                             0.32
Raw_material_turnover           0.17
dtype: float64
```

TABLE 23 MODEL 3 AND TOP 5 P VALUE VARIABLE

Model Number 4

Optimization terminated successfully.
Current function value: 0.496497
Iterations 6

Logit Regression Results

Dep. Variable:	Default	No. Observations:	2851
Model:	Logit	Df Residuals:	2845
Method:	MLE	Df Model:	5
Date:	Sat, 17 Jun 2023	Pseudo R-squ.:	0.04026
Time:	16:29:41	Log-Likelihood:	-1415.5
converged:	True	LL-Null:	-1474.9
Covariance Type:	nonrobust	LLR p-value:	5.740e-24

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3619	0.048	-28.592	0.000	-1.455	-1.269
PAT_as_perc_of_net_worth	-0.3452	0.051	-6.785	0.000	-0.445	-0.245
Total_term_liabilities__to__tangible_net_worth	0.1913	0.044	4.344	0.000	0.105	0.278
Current_ratio_times	-0.1557	0.061	-2.539	0.011	-0.276	-0.035
Cash_to_current_liabilities_times	0.1163	0.054	2.140	0.032	0.010	0.223
Raw_material_turnover	-0.0669	0.053	-1.269	0.205	-0.170	0.036

```
Raw_material_turnover                              0.20
Cash_to_current_liabilities_times                  0.03
Current_ratio_times                                0.01
Total_term_liabilities__to__tangible_net_worth     0.00
PAT_as_perc_of_net_worth                           0.00
dtype: float64
```

TABLE 24 MODEL 4 AND TOP 5 P VALUE VARIABLE

Confusion Matrix on train dataset for Model 4

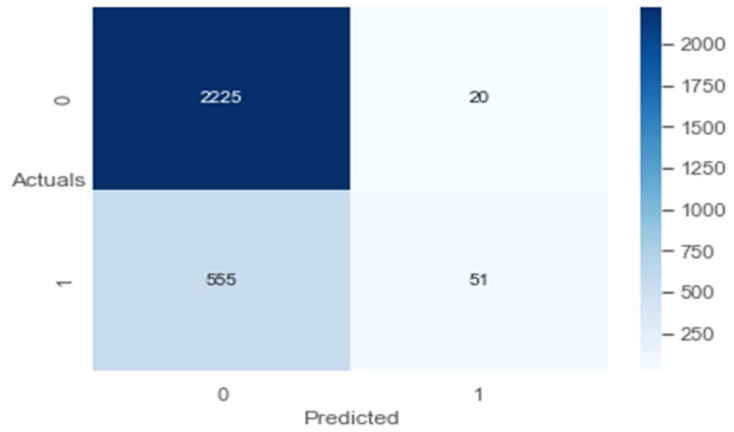

FIGURE 11 CONFUSION MATRIX ON TRAIN DATASET MODEL 4

Classification Report on Train dataset of Model Number 4

	precision	recall	f1-score	support
0.0	0.800	0.991	0.886	2245
1.0	0.718	0.084	0.151	606
accuracy			0.798	2851
macro avg	0.759	0.538	0.518	2851
weighted avg	0.783	0.798	0.729	2851

TABLE 25 CLASSIFICATION REPORT OF MODEL 4

Visualization of target 'Default' on Model Number 4

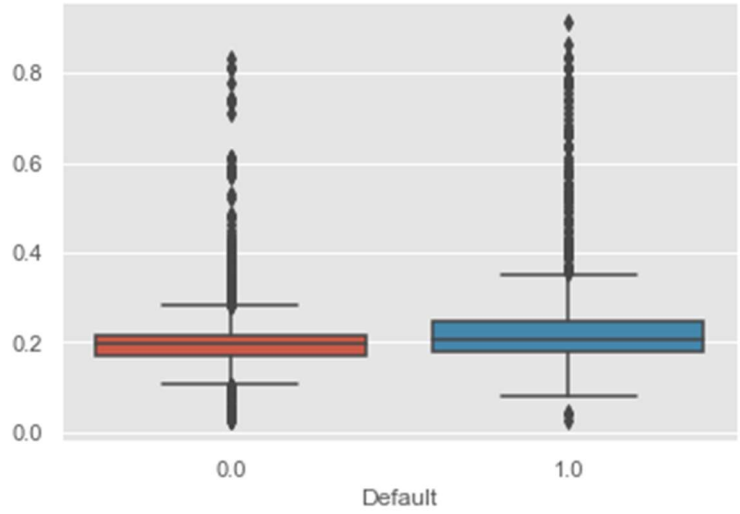

FIGURE 12 TARGET VARIABLE TRAIN DATASET MODEL 4

Confusion Matrix on train dataset

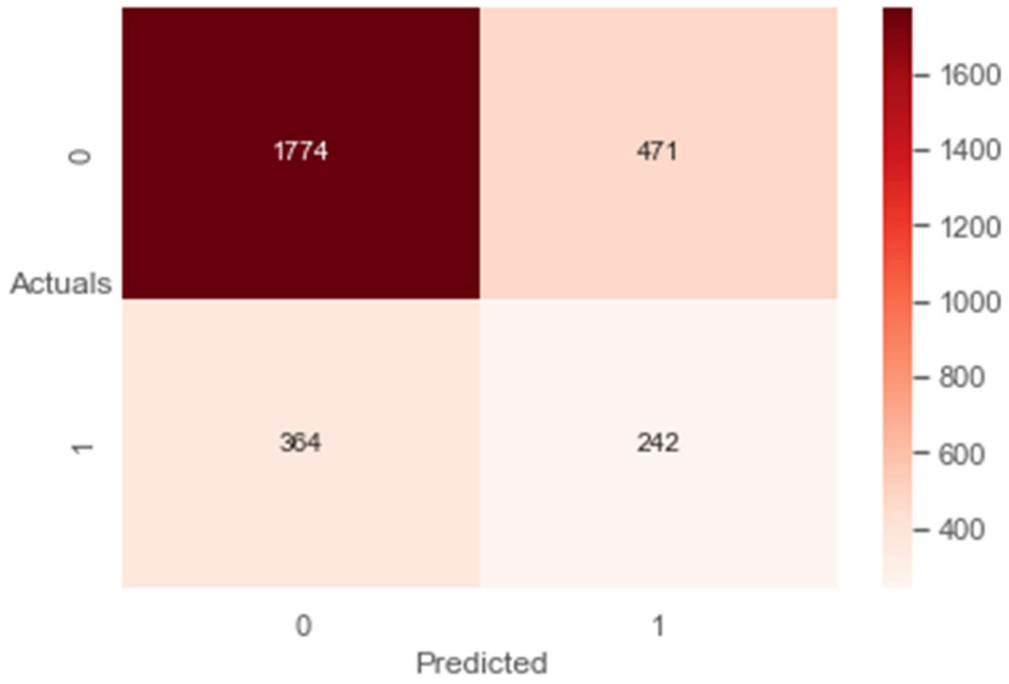

FIGURE 13 CONFUSION MATRIX ON TRAIN DATASET MODEL 4

Classification Report on Train dataset

	precision	recall	f1-score	support
0.0	0.830	0.790	0.809	2245
1.0	0.339	0.399	0.367	606
accuracy			0.707	2851
macro avg	0.585	0.595	0.588	2851
weighted avg	0.726	0.707	0.715	2851

TABLE 26 CLASSIFICATION REPORT OF MODEL 4

Confusion Matrix on test dataset

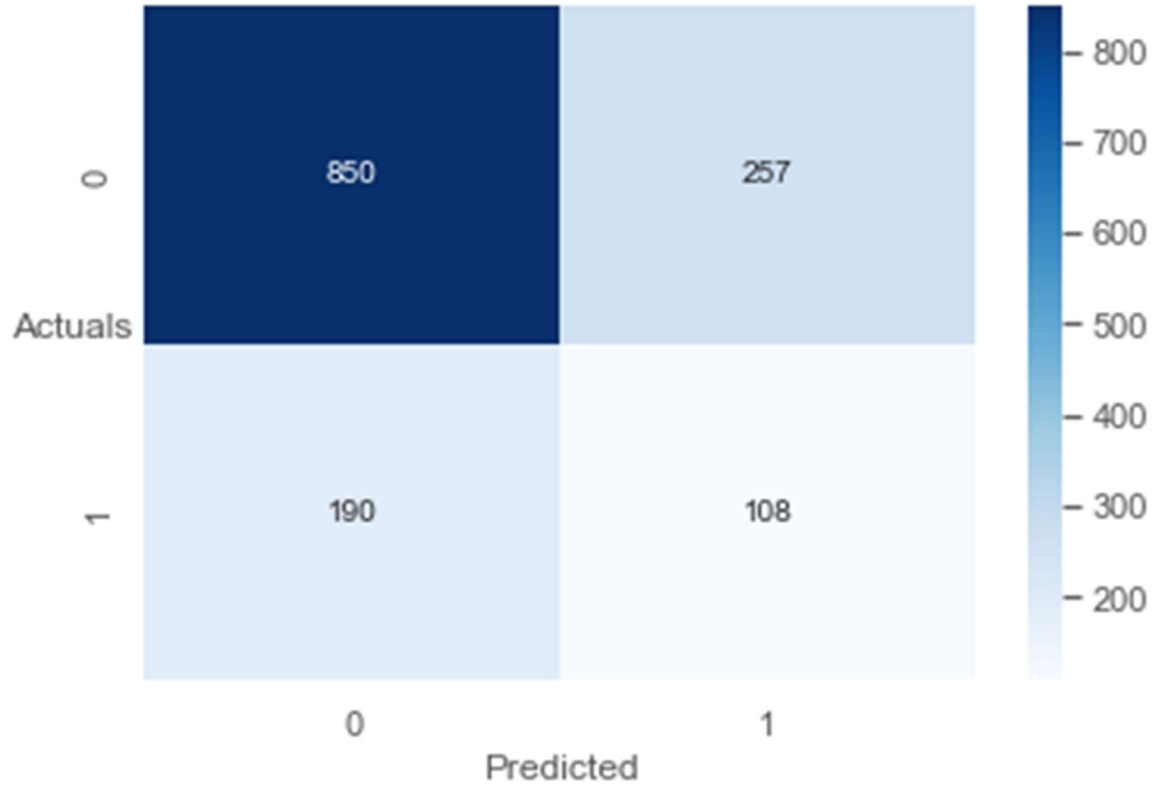

FIGURE 14 CONFUSION MATRIX ON TEST DATASET MODEL 4

Classification Report on Test dataset

	precision	recall	f1-score	support
0.0	0.817	0.768	0.792	1107
1.0	0.296	0.362	0.326	298
accuracy			0.682	1405
macro avg	0.557	0.565	0.559	1405
weighted avg	0.707	0.682	0.693	1405

TABLE 27 CLASSIFICATION REPORT ON TRAIN DATA SET ON OPTIMAL THRESHOLD

Confusion Matrix Insights:

Precision: The precision for the "1.0" class is 0.368, which means that when the model predicts the class as "1.0" (default), it is correct approximately 36.8% of the time. This suggests that the model has a relatively high rate of false positives for the "1.0" class.

Recall: The recall for the "1.0" class is 0.368, indicating that the model correctly identifies approximately 36.8% of the actual "1.0" instances. This means that the model has a relatively high rate of false negatives for the "1.0" class.

F1-score: The F1-score for the "1.0" class is 0.316, which is the harmonic mean of precision and recall. It indicates the balance between precision and recall for the "1.0" class. The F1-score suggests that the model's performance for the "1.0" class is relatively low.

Accuracy: The overall accuracy of the model is 0.676, which means that it correctly predicts the class for approximately 67.6% of the instances. However, accuracy alone may not be sufficient to assess the model's performance, especially when dealing with imbalanced datasets.

Support: The support represents the number of instances of each class in the dataset.

Modelling on SMOTE Dataset

```
Int64Index: 4491 entries, 1029 to 6637
Data columns (total 17 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Change_in_stock                               4491 non-null   float64
 1   PBDITA_as_perc_of_total_income                4491 non-null   float64
 2   PAT_as_perc_of_net_worth                      4491 non-null   float64
 3   Total_term_liabilities__to__tangible_net_worth  4491 non-null   float64
 4   Contingent_liabilities__to__Net_worth_perc    4491 non-null   float64
 5   Net_working_capital                           4491 non-null   float64
 6   Current_ratio_times                           4491 non-null   float64
 7   Cash_to_current_liabilities_times             4491 non-null   float64
 8   Cash_to_average_cost_of_sales_per_day         4491 non-null   float64
 9   Creditors_turnover                            4491 non-null   float64
 10  Debtors_turnover                              4491 non-null   float64
 11  Finished_goods_turnover                       4491 non-null   float64
 12  WIP_turnover                                  4491 non-null   float64
 13  Raw_material_turnover                         4491 non-null   float64
 14  Shares_outstanding                            4491 non-null   float64
 15  Equity_face_value                             4491 non-null   float64
 16  EPS                                           4491 non-null   float64
dtypes: float64(17)
memory usage: 631.5 KB
```

TABLE 28 DATA INFORMATION AFTER RESAMPLING

SMOTE MODEL 1 and Top 5 P values variables

Optimization terminated successfully.
Current function value: 0.663847
Iterations 5

Logit Regression Results

Dep. Variable:	Default	No. Observations:	4491
Model:	Logit	Df Residuals:	4473
Method:	MLE	Df Model:	17
Date:	Fri, 16 Jun 2023	Pseudo R-squ.:	0.04227
Time:	16:21:58	Log-Likelihood:	-2981.3
converged:	True	LL-Null:	-3112.9
Covariance Type:	nonrobust	LLR p-value:	4.215e-46

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0632	0.031	-2.034	0.042	-0.124	-0.002
Change_in_stock	-0.0209	0.037	-0.571	0.568	-0.092	0.051
PBDITA_as_perc_of_total_income	-0.0308	0.034	-0.902	0.367	-0.098	0.036
PAT_as_perc_of_net_worth	-0.2718	0.038	-7.131	0.000	-0.347	-0.197
Total_term_liabilities__to__tangible_net_worth	0.1669	0.033	5.098	0.000	0.103	0.231
Contingent_liabilities__to__Net_worth_perc	-0.0218	0.031	-0.707	0.479	-0.082	0.039
Net_working_capital	-0.0474	0.034	-1.378	0.168	-0.115	0.020
Current_ratio_times	-0.2735	0.044	-6.173	0.000	-0.360	-0.187
Cash_to_current_liabilities_times	0.0636	0.048	1.339	0.180	-0.029	0.157
Cash_to_average_cost_of_sales_per_day	0.0105	0.038	0.278	0.781	-0.064	0.085
Creditors_turnover	0.0677	0.036	1.893	0.058	-0.002	0.138
Debtors_turnover	-0.0647	0.035	-1.854	0.064	-0.133	0.004
Finished_goods_turnover	-0.0223	0.041	-0.540	0.589	-0.103	0.059
WIP_turnover	0.0087	0.043	0.204	0.839	-0.075	0.093
Raw_material_turnover	-0.0948	0.037	-2.561	0.010	-0.167	-0.022
Shares_outstanding	-0.0524	0.037	-1.405	0.160	-0.126	0.021
Equity_face_value	-0.0968	0.037	-2.616	0.009	-0.169	-0.024
EPS	-0.0186	0.037	-0.500	0.617	-0.092	0.054

WIP_turnover	0.84
Cash_to_average_cost_of_sales_per_day	0.78
EPS	0.62
Finished_goods_turnover	0.59
Change_in_stock	0.57

dtype: float64

TABLE 29 VARIABLES WITH HIGHER P VALUES

SMOTE MODEL 2 and Top 5 P values variables

Optimization terminated successfully.
Current function value: 0.664257
Iterations 5

Logit Regression Results

Dep. Variable:	Default	No. Observations:	4491
Model:	Logit	Df Residuals:	4478
Method:	MLE	Df Model:	12
Date:	Fri, 16 Jun 2023	Pseudo R-squ.:	0.04168
Time:	16:22:00	Log-Likelihood:	-2983.2
converged:	True	LL-Null:	-3112.9
Covariance Type:	nonrobust	LLR p-value:	1.431e-48

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0611	0.031	-1.970	0.049	-0.122	-0.000
PAT_as_perc_of_net_worth	-0.2941	0.035	-8.405	0.000	-0.363	-0.226
Total_term_liabilities__to__tangible_net_worth	0.1644	0.033	5.039	0.000	0.100	0.228
Contingent_liabilities__to__Net_worth_perc	-0.0189	0.031	-0.616	0.538	-0.079	0.041
Current_ratio_times	-0.2823	0.044	-6.427	0.000	-0.368	-0.196
Cash_to_current_liabilities_times	0.0648	0.040	1.601	0.109	-0.015	0.144
Creditors_turnover	0.0646	0.036	1.816	0.069	-0.005	0.134
Debtors_turnover	-0.0611	0.035	-1.761	0.078	-0.129	0.007
Finished_goods_turnover	-0.0209	0.041	-0.506	0.613	-0.102	0.060
WIP_turnover	0.0113	0.043	0.265	0.791	-0.072	0.095
Raw_material_turnover	-0.0920	0.037	-2.502	0.012	-0.164	-0.020
Shares_outstanding	-0.0613	0.036	-1.683	0.092	-0.133	0.010
Equity_face_value	-0.1001	0.036	-2.769	0.006	-0.171	-0.029

```
WIP_turnover                                0.79
Finished_goods_turnover                     0.61
Contingent_liabilities__to__Net_worth_perc  0.54
Cash_to_current_liabilities_times           0.11
Shares_outstanding                          0.09
dtype: float64
```

TABLE 30 SMOTE MODEL 2 WITH HIGH P VALUES VARIABLES

SMOTE MODEL 3 and Top 5 P values variables

```
Optimization terminated successfully.
         Current function value: 0.665764
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 4491
Model:                          Logit   Df Residuals:                     4482
Method:                           MLE   Df Model:                            8
Date:                Fri, 16 Jun 2023   Pseudo R-squ.:                0.03951
Time:                        16:22:01   Log-Likelihood:                -2989.9
converged:                       True   LL-Null:                       -3112.9
Covariance Type:            nonrobust   LLR p-value:                 1.239e-48
================================================================================================================
                                                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                       -0.0531      0.031     -1.721      0.085      -0.114       0.007
PAT_as_perc_of_net_worth                        -0.2992      0.035     -8.582      0.000      -0.367      -0.231
Total_term_liabilities__to__tangible_net_worth   0.1559      0.032      4.923      0.000       0.094       0.218
Current_ratio_times                             -0.2581      0.042     -6.108      0.000      -0.341      -0.175
Cash_to_current_liabilities_times                0.0525      0.040      1.309      0.191      -0.026       0.131
Debtors_turnover                                -0.0509      0.033     -1.527      0.127      -0.116       0.014
Finished_goods_turnover                         -0.0198      0.041     -0.481      0.630      -0.100       0.061
WIP_turnover                                     0.0137      0.042      0.323      0.747      -0.069       0.097
Raw_material_turnover                           -0.0744      0.036     -2.088      0.037      -0.144      -0.005
================================================================================================================
```

```
WIP_turnover                          0.75
Finished_goods_turnover               0.63
Cash_to_current_liabilities_times     0.19
Debtors_turnover                      0.13
Intercept                             0.09
dtype: float64
```

TABLE 31 SMOTE MODEL 3 WITH HIGH P VALUES VARIABLES

SMOTE MODEL 4 and Top 5 P values variables

```
Optimization terminated successfully.
         Current function value: 0.666047
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 4491
Model:                          Logit   Df Residuals:                     4484
Method:                           MLE   Df Model:                            6
Date:                Fri, 16 Jun 2023   Pseudo R-squ.:                0.03910
Time:                        16:22:02   Log-Likelihood:                -2991.2
converged:                       True   LL-Null:                       -3112.9
Covariance Type:            nonrobust   LLR p-value:                 1.046e-49
================================================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                     -0.0514      0.031     -1.669      0.095      -0.112       0.009
PAT_as_perc_of_net_worth                      -0.3043      0.035     -8.774      0.000      -0.372      -0.236
Total_term_liabilities__to__tangible_net_worth     0.1509      0.031      4.794      0.000       0.089       0.213
Current_ratio_times                           -0.2547      0.042     -6.054      0.000      -0.337      -0.172
Cash_to_current_liabilities_times              0.0501      0.040      1.251      0.211      -0.028       0.129
WIP_turnover                                  -0.0004      0.034     -0.013      0.990      -0.066       0.066
Raw_material_turnover                         -0.0822      0.035     -2.327      0.020      -0.151      -0.013
================================================================================================================
```

```
WIP_turnover                                   0.99
Cash_to_current_liabilities_times              0.21
Intercept                                      0.10
Raw_material_turnover                          0.02
Total_term_liabilities__to__tangible_net_worth   0.00
dtype: float64
```

TABLE 32 SMOTE MODEL 4 WITH HIGH P VALUES VARIABLES

SMOTE MODEL 5 and Top 5 P values variables

```
Optimization terminated successfully.
         Current function value: 0.666047
         Iterations 5
                          Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 4491
Model:                          Logit   Df Residuals:                     4485
Method:                           MLE   Df Model:                            5
Date:                Fri, 16 Jun 2023   Pseudo R-squ.:                0.03910
Time:                        16:22:03   Log-Likelihood:               -2991.2
converged:                       True   LL-Null:                      -3112.9
Covariance Type:            nonrobust   LLR p-value:                1.420e-50
==========================================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------
Intercept                                        -0.0514      0.031     -1.669      0.095      -0.112       0.009
PAT_as_perc_of_net_worth                         -0.3043      0.035     -8.786      0.000      -0.372      -0.236
Total_term_liabilities__to__tangible_net_worth    0.1509      0.031      4.794      0.000       0.089       0.213
Current_ratio_times                              -0.2547      0.042     -6.064      0.000      -0.337      -0.172
Cash_to_current_liabilities_times                 0.0500      0.040      1.252      0.210      -0.028       0.128
Raw_material_turnover                            -0.0823      0.035     -2.360      0.018      -0.151      -0.014
==========================================================================================================================
```

```
Cash_to_current_liabilities_times               0.21
Intercept                                       0.10
Raw_material_turnover                           0.02
Total_term_liabilities__to__tangible_net_worth  0.00
Current_ratio_times                             0.00
dtype: float64
```

TABLE 33 SMOTE MODEL 5 WITH HIGH P VALUES VARIABLES

Smote Model 5 Summary Insight

Intercept: The intercept term represents the estimated log-odds of the default event when all predictor variables are zero.

Coefficients: The coefficients represent the change in the log-odds of the default event for a one-unit increase in each respective predictor variable, holding other variables constant.

Interpretation of the coefficients:

PAT_as_perc_of_net_worth: For a one-unit increase in the PAT_as_perc_of_net_worth variable, the log-odds of the default event decrease by 0.3043, assuming all other variables are held constant.

Total_term_liabilities__to__tangible_net_worth: For a one-unit increase in the Total_term_liabilities__to__tangible_net_worth variable, the log-odds of the default event increase by 0.1509, assuming all other variables are held constant.

Current_ratio_times: For a one-unit increase in the Current_ratio_times variable, the log-odds of the default event decrease by 0.2547, assuming all other variables are held constant.

Cash_to_current_liabilities_times: The coefficient is not statistically significant (p-value > 0.05) at the 5% significance level. Therefore, there is not enough evidence to conclude that this variable has a significant impact on the log-odds of the default event.

Raw_material_turnover: For a one-unit increase in the Raw_material_turnover variable, the log-odds of the default event decrease by 0.0823, assuming all other variables are held constant.

The pseudo R-squared value of 0.0391 indicates that the model explains about 3.91% of the variation in the default event. The LLR p-value of 1.420e-50 suggests that the model, as a whole, is statistically significant in explaining the default event.

It's important to note that the interpretation of coefficients depends on the specific context and the assumptions of the logistic regression model.

Confusion Matrix on train SMOTE Data

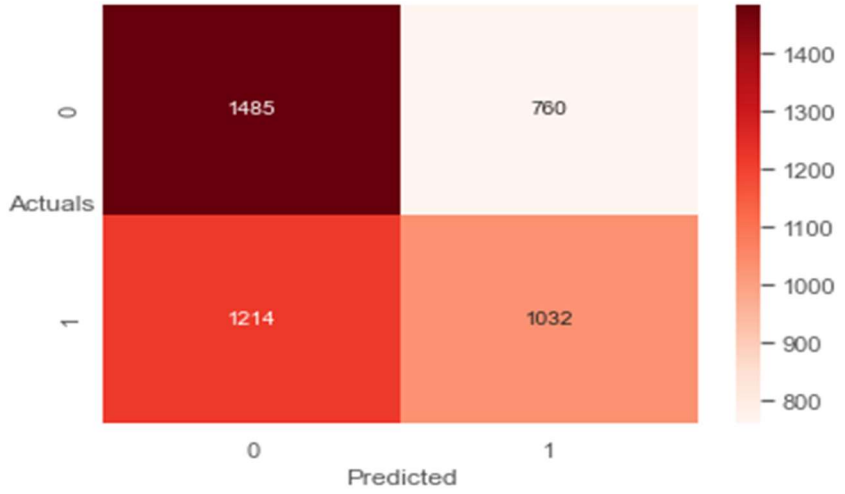

FIGURE 15 : CONFUSION MATRIX ON SMOTE TRAIN DATASET

Classification report on train SMOTE Data

	precision	recall	f1-score	support
0.0	0.550	0.661	0.601	2245
1.0	0.576	0.459	0.511	2246
accuracy			0.560	4491
macro avg	0.563	0.560	0.556	4491
weighted avg	0.563	0.560	0.556	4491

TABLE 34 CLASSIFICATION REPORT ON TRAIN SMOTE DATASET

Visualization of target Default on Smote Dataset

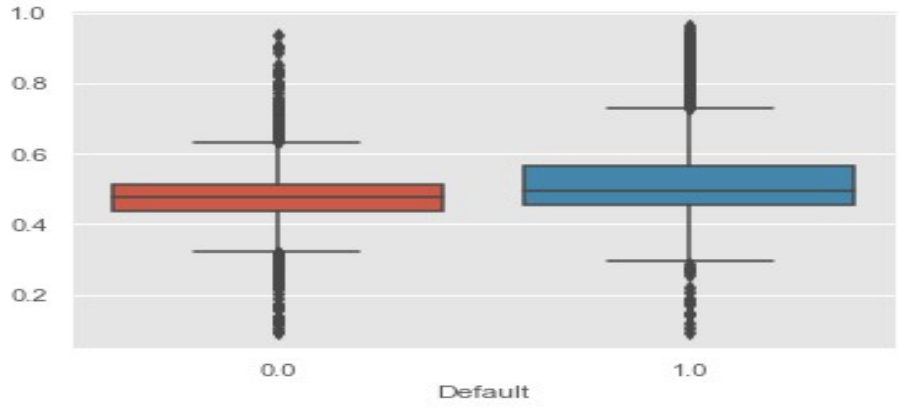

FIGURE 16 TARGET VARIABLE ON SMOTE DATA SET

Confusion Matrix on train SMOTE Data on Optimum threshold

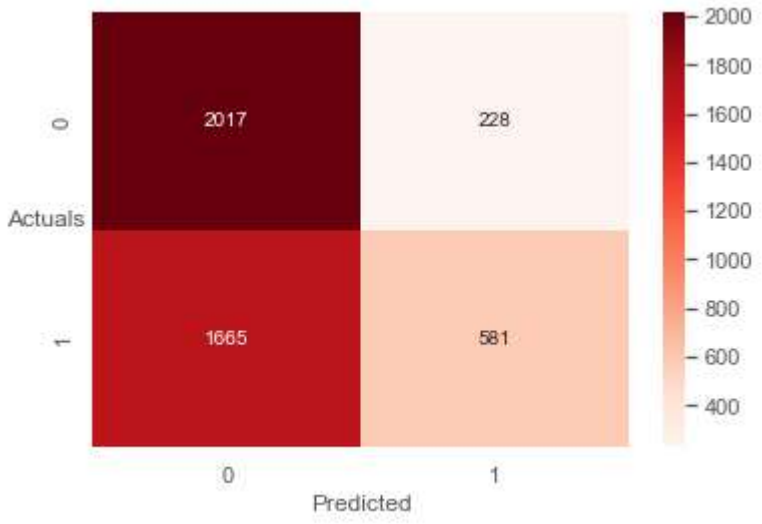

FIGURE 17 : CONFUSION MATRIX ON TRAIN SMOTE OPTIMUM THRESHOLD DATA SET

Classification report on Smote Train Dataset on optimum Threshold

	precision	recall	f1-score	support
0.0	0.548	0.898	0.681	2245
1.0	0.718	0.259	0.380	2246
accuracy			0.578	4491
macro avg	0.633	0.579	0.530	4491
weighted avg	0.633	0.578	0.530	4491

TABLE 35 CLASSIFICATION REPORT ON SMOTE TRAIN ON OPTIMUM THRESHOLD

Confusion Matrix on test SMOTE Data on Optimum threshold

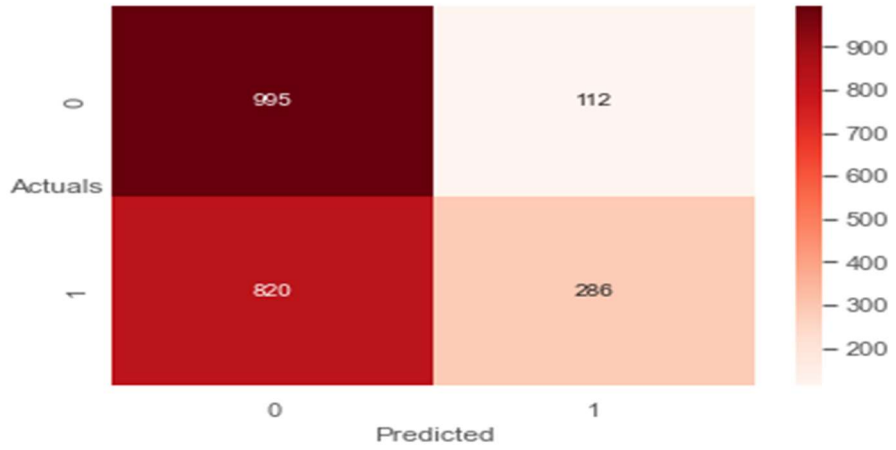

FIGURE 18 CONFUSION MATRIX ON TEST SMOTE OPTIMUM THRESHOLD DATA SET

Test Data Confusion Matrix Insights:

confusion matrix that shows the counts of true positive, true negative, false positive, and false negative predictions made by model. The interpretation of the confusion matrix:

True Positives (TP): 277
This indicates the number of instances that were correctly predicted as positive (class 1) by the model.

True Negatives (TN): 974
This indicates the number of instances that were correctly predicted as negative (class 0) by the model.

False Positives (FP): 164
This indicates the number of instances that were incorrectly predicted as positive (class 1) by the model.

False Negatives (FN): 798
This indicates the number of instances that were incorrectly predicted as negative (class 0) by the model.

In summary, model correctly predicted 277 instances as positive and 974 instances as negative. However, it incorrectly predicted 164 instances as positive and 798 instances as negative.

The confusion matrix provides important information for evaluating the performance of a classification model.

Classification report on smote Test data on optimal threshold

	precision	recall	f1-score	support
0.0	0.548	0.899	0.681	1107
1.0	0.719	0.259	0.380	1106
accuracy			0.579	2213
macro avg	0.633	0.579	0.531	2213
weighted avg	0.633	0.579	0.531	2213

TABLE 36 : CLASSIFICATION REPORT ON SMOTE TEST ON OPTIMUM THRESHOLD

Test Data Classification Report Insights

These metrics provide information about the performance of the model in predicting both classes (0.0 and 1.0). Precision represents the ability of the model to correctly identify true positives, while recall represents the proportion of actual positives that are correctly identified. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of model performance. The support indicates the number of instances in each class.

The recall for class 1 (1.0) is 0.262. Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances (class 1) that are correctly identified by the model. In this case, the model has a recall of 0.262 for class 1, indicating that it only correctly identifies approximately 25.8% of the actual positive instances.

The model shows moderate performance with relatively higher precision and recall for class 0.0 compared to class 1.0. The accuracy of the model is 0.564, indicating that it correctly predicts the class labels for 56.4% of the instances.

A low recall value for class 1 suggests that the model is not effectively capturing all instances of the positive class, potentially leading to a significant number of false negatives. This means that the model is missing a substantial portion of the positive cases, which may have adverse consequences depending on problem being addressed.

Model Building on Random Forest

The Random forests classifier is a supervised learning algorithm and is known as the most flexible and easy to use algorithm. A forest is comprised of decision trees. The more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.

Advantages of random forest:

Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process. It does not suffer (that much) from the overfitting problem. The main reason is that it takes the average of all the predictions and therefore cancels out the biases. Disadvantages of random forest:

The model is more difficult to interpret compared to a decision tree. Random forests process is time-consuming. This is because the model is slow in generating predictions due to the multiple decision trees.

Random Forest Model 1: RandomForestClassifier (n_estimators=1000,class_weight = "balanced")

Classification report on Train data

	precision	recall	f1-score	support
0.0	1.000	0.922	0.959	2245
1.0	0.775	1.000	0.873	606
accuracy			0.938	2851
macro avg	0.887	0.961	0.916	2851
weighted avg	0.952	0.938	0.941	2851

TABLE 37 CLASSIFICATION REPORT ON RF TRAIN

Confusion matrix on Train data:

```
[[2069  176]
 [   0  606]]
```

TABLE 38 CONFUSION MATRIX ON RF TRAIN

Confusion matrix on Test Data

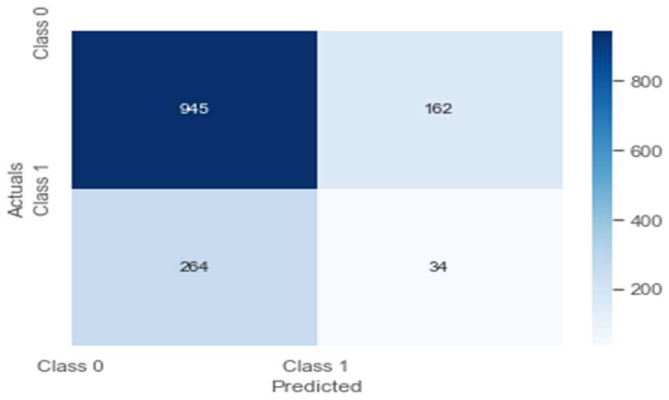

FIGURE 19 CONFUSION MATRIX ON TEST DATA SET OF RF

Confusion matrix on Test Data Random forest

Classification report on Test data

	precision	recall	f1-score	support
0.0	0.782	0.854	0.816	1107
1.0	0.173	0.114	0.138	298
accuracy			0.697	1405
macro avg	0.478	0.484	0.477	1405
weighted avg	0.653	0.697	0.672	1405

TABLE 39 CLASSIFICATION REPORT ON RF TEST DATASET

Random Forest Model Number 2

```
RandomForestClassifier (class_weight='balanced',max_depth=3,max_featur
                        es=18,min_samples_leaf=2, min_samples_split=30,
                         n_estimators=1000, random_state=42)
```

Confusion matrix on Train data

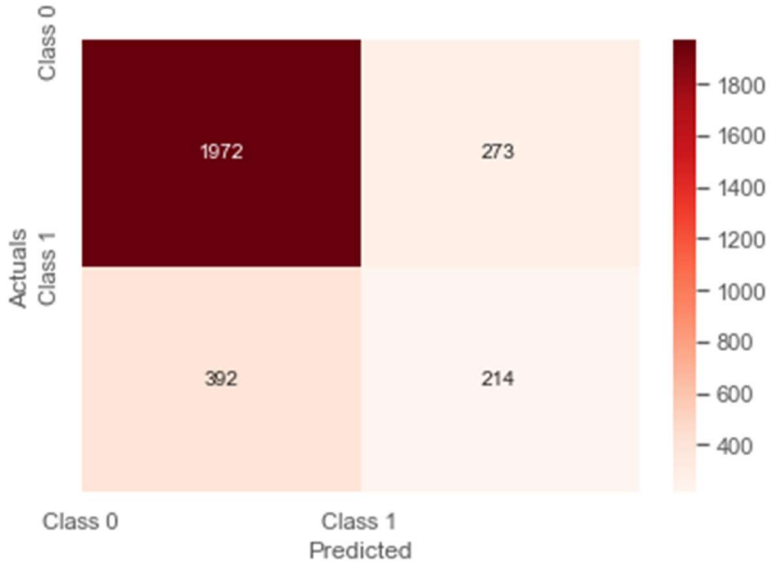

FIGURE 20: CONFUSION MATRIX ON TRAIN DATA SET OF RFM2

Classification report on Train data

	precision	recall	f1-score	support
0.0	0.834	0.878	0.856	2245
1.0	0.439	0.353	0.392	606
accuracy			0.767	2851
macro avg	0.637	0.616	0.624	2851
weighted avg	0.750	0.767	0.757	2851

TABLE 40 CLASSIFICATION REPORT ON RF TRAIN DATASET MODEL2

Confusion matrix on Test data

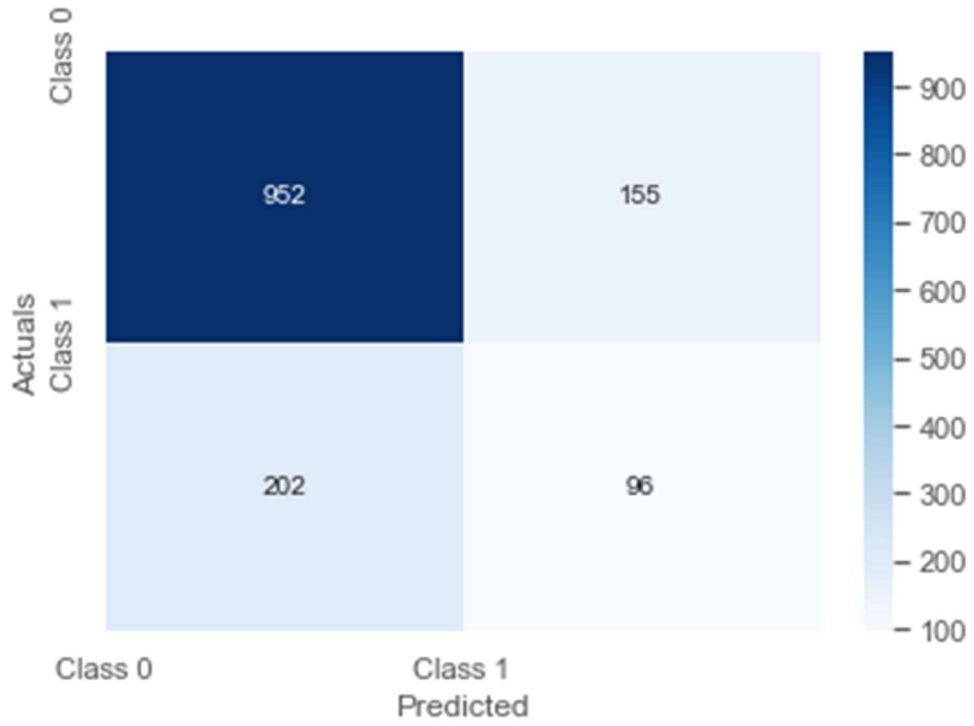

FIGURE 20 CONFUSION MATRIX ON TEST DATA SET OF RFM2

Classification report on Train data Model 2

	precision	recall	f1-score	support
0.0	0.825	0.860	0.842	1107
1.0	0.382	0.322	0.350	298
accuracy			0.746	1405
macro avg	0.604	0.591	0.596	1405
weighted avg	0.731	0.746	0.738	1405

TABLE 41 : CLASSIFICATION REPORT ON RF TRAIN DATASET MODEL2

Insight on Classification on Best grid

Precision: The precision for class 1 is 0.382, which means that out of all the instances predicted as class 1, approximately 38.2% of them are actually true positives. This indicates that the model has a relatively low accuracy in identifying true positive cases for class 1.

Recall: The recall for class 1 is 0.322, which means that only about 32.2% of the actual instances of class 1 are correctly identified by the model. This suggests that the model has a relatively high number of false negatives, meaning it fails to capture a significant portion of actual positive cases for class 1.

F1-score: The F1-score for class 1 is 0.350, which is the harmonic mean of precision and recall. The low F1-score indicates that the model's performance for class 1 is relatively poor, and there is room for improvement in capturing true positives and reducing false negatives.

Accuracy: The overall accuracy of the model is 0.746, which represents the proportion of correctly predicted instances out of the total instances.

The macro avg and weighted avg values provide aggregated metrics considering all classes. The macro avg calculates the average performance across all classes, giving equal weight to each class. The weighted avg calculates the average performance, but weights each class by its support (number of instances), providing a more accurate representation of overall performance when classes are imbalanced.

In summary, the model's performance for class 1 is characterized by low precision (relatively low proportion of true positives among predicted positives) and low recall (low ability to correctly identify actual positives). This suggests that the model may struggle in correctly identifying instances of class 1, potentially leading to a higher number of false negatives.

Feature Importance

	Imp
PAT_as_perc_of_net_worth	0.52
Current_ratio_times	0.07
PBDITA_as_perc_of_total_income	0.07
Change_in_stock	0.06
Total_term_liabilities__to__tangible_net_worth	0.04
WIP_turnover	0.04
EPS	0.03
Shares_outstanding	0.03
Finished_goods_turnover	0.02
Raw_material_turnover	0.02
Creditors_turnover	0.02
Net_working_capital	0.02
Debtors_turnover	0.02
Cash_to_current_liabilities_times	0.01
Cash_to_average_cost_of_sales_per_day	0.01
Equity_face_value	0.01
Contingent_liabilities__to__Net_worth_perc	0.01

TABLE 42 FEATURE IMPORTANCE VIF

Random Forest Model Number 3

RandomForestClassifier (class_weight='balanced', max_depth= 10, max_features=30,
min_samples_leaf=10, min_samples_split=40,
n_estimators=1500, random_state=42)

Confusion matrix on Train data

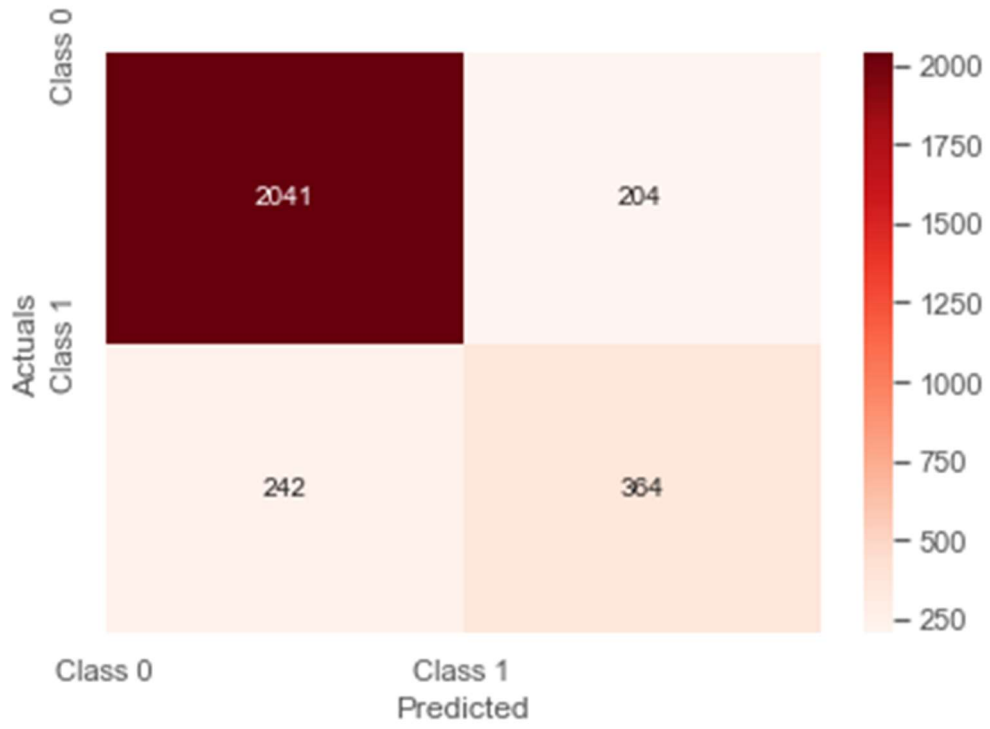

FIGURE 21 CONFUSION MATRIX ON TRAIN DATA SET OF RFM3

Classification report on Train data

	precision	recall	f1-score	support
0.0	0.894	0.909	0.902	2245
1.0	0.641	0.601	0.620	606
accuracy			0.844	2851
macro avg	0.767	0.755	0.761	2851
weighted avg	0.840	0.844	0.842	2851

TABLE 43 CLASSIFICATION REPORT ON RF TRAIN DATASET MODEL3

Confusion matrix on Train data

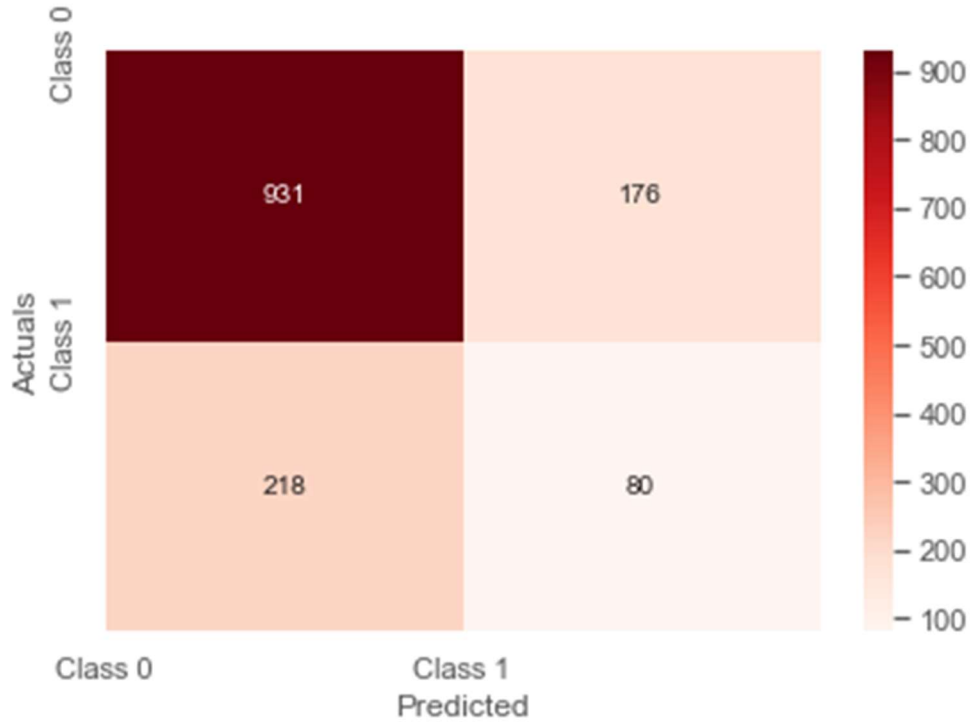

FIGURE 22 CONFUSION MATRIX ON TEST DATA SET OF RFM3

Classification report on Test data

	precision	recall	f1-score	support
0.0	0.810	0.841	0.825	1107
1.0	0.312	0.268	0.289	298
accuracy			0.720	1405
macro avg	0.561	0.555	0.557	1405
weighted avg	0.705	0.720	0.712	1405

TABLE 44 CLASSIFICATION REPORT ON RF TEST DATASET MODEL3

Insight on Classification Test data

Precision: It represents the ability of the classifier to correctly identify the positive class (1.0) among the samples it predicted as positive. In this case, the precision for the positive class is 0.312, indicating that out of all the samples predicted as positive, only 31.2% were actually positive.

Recall: It represents the ability of the classifier to correctly identify the positive class (1.0) among all the actual positive samples. In this case, the recall for the positive class is 0.268, indicating that the classifier correctly identified only 26.8% of the actual positive samples.

F1-score: It is the harmonic mean of precision and recall, providing a balance between the two metrics. In this case, the F1-score for the positive class is 0.289.

Support: It represents the number of samples in each class. In this case, there are 1107 samples of the negative class (0.0) and 298 samples of the positive class (1.0).

Accuracy: It represents the overall accuracy of the classifier in predicting the correct labels for all the samples. In this case, the accuracy is 0.720, indicating that the classifier correctly predicted the labels for 72.0% of the samples.

The macro average and weighted average metrics provide an overall summary of the classification performance across both classes. The macro average takes the unweighted mean of the metrics for each class, while the weighted average takes the weighted mean based on the number of samples in each class.

Conclusion

These metrics indicate that the classifier's performance for class 1 is relatively low. It has a low precision, recall, and F1-score, suggesting that it struggles to accurately identify and classify instances of class 1.

MODEL COMPARISION

AUC and ROC for the training data of Stats Model

```
AUC for the Training Data: 0.604
AUC for the Test Data:     0.569
```

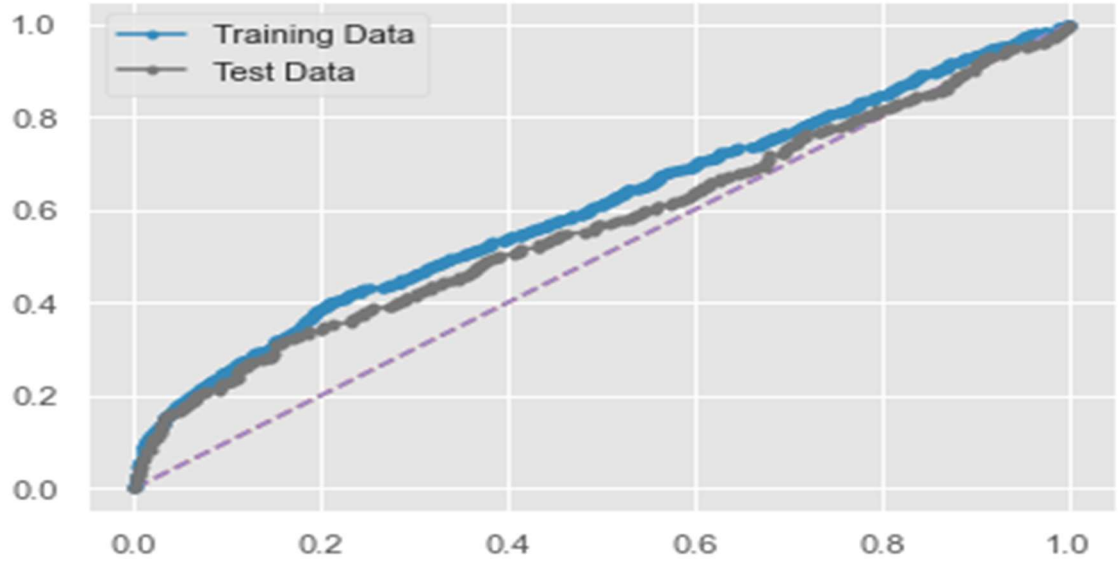

FIGURE 23 : AUC FOR STATS MODEL

AUC and ROC for the training data of Random forest

TABLE 39: AUC FOR TRAINING AND TEST DATA OF STATS MODEL

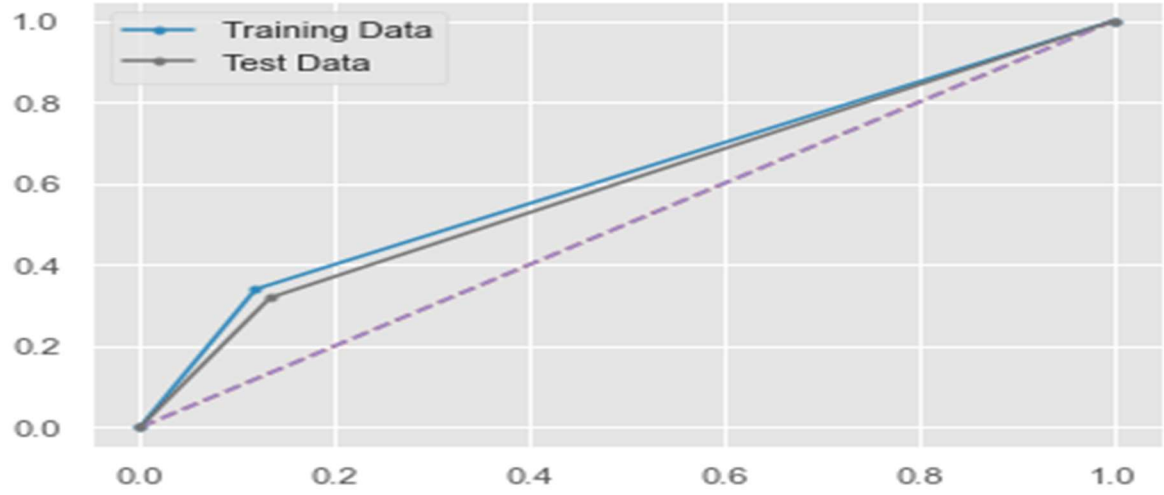

FIGURE 24 AUC FOR RANDOM FOREST MODEL

```
AUC for the Training Data: 0.611
AUC for the Test Data:     0.593
```

TABLE 45 AUC FOR TRAINING AND TEST DATA OF RANDOM FOREST MODEL

AUC comparison insight:

The Random Forest model has a slightly higher AUC on the test data than the Stats model . This suggests that the Random Forest model is better at generalizing to new data.

The Random Forest model might be better than the Stats model:

Random Forest is a non-parametric model, which means that it does not make any assumptions about the distribution of the data. This makes it more robust to changes in the data distribution. Random Forest is an ensemble model, which means that it is made up of multiple decision trees. This makes it more resistant to overfitting than a single decision tree. However, it is important to note that the difference in AUC between the two models is small. It is possible that the Random Forest model is only slightly better than the Stats model.

Comparison of Classification report for Stats Model and random Forest

	Stats Model				Random Forest		
	precision	recall	f1-score	support	precision	recall	f1-score
0.0	0.82	0.77	0.79	1107.00	0.81	0.84	0.83
1.0	0.30	0.36	0.33	298.00	0.31	0.27	0.29
accuracy	0.68	0.68	0.68	0.68	0.72	0.72	0.72
macro avg	0.56	0.57	0.56	1405.00	0.56	0.55	0.56
weighted avg	0.71	0.68	0.69	1405.00	0.70	0.72	0.71

	support
0.0	1107.00
1.0	298.00
accuracy	0.72
macro avg	1405.00
weighted avg	1405.00

TABLE 46 CLASSIFICATION TABLE COMPARISON TEST DATA

Conclusion of classification of test Models:

compares the performance of Best Stats Model and Random Forest with reference to Class 1,

Precision (Class 1): Stats Model has a precision of 0.30 for Class 1, while Random Forest has a slightly higher precision of 0.31 for Class 1. Both models have relatively low precision, indicating a higher rate of false positives.

Recall (Class 1): Stats Model has a recall of 0.36 for Class 1, whereas Random Forest has a lower recall of 0.27 for Class 1. Stats Model performs better in terms of capturing the true positives, i.e., Class 1 instances.

F1-score (Class 1): Stats Model has an F1-score of 0.33 for Class 1, and Random Forest has an F1-score of 0.29 for Class 1. The F1-score considers both precision and recall, and in this case, Stats Model has a slightly higher F1-score for Class 1.

Based on these metrics, Stats Model appears to perform slightly better than Random Forest with reference to Class 1. It has a higher recall and F1-score, indicating a better ability to identify positive instances (Class 1).

However, both models have relatively low precision, suggesting a higher rate of false positives.

Comparing Confusion matrix for both the models on test data

	Stats Model		Random Forest	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0	850	257	931	176
Actual 1	190	108	218	80

TABLE 47 COMPARISON OF CONFUSION MATRIX OF TEST DATA

Conclusion of Confusion Matrix of test Models:

By comparing the performance of Stats Model and Random Forest with reference to Class 1, we can follow metrics from the classification report:

Precision: Precision is the ratio of true positives (Class 1 correctly predicted) to the total predicted positives (Class 1 predicted). Higher precision indicates a lower false positive rate.

Recall: Recall is the ratio of true positives to the total actual positives (Class 1 in the dataset). Higher recall indicates a lower false negative rate.

F1-score: F1-score is the harmonic mean of precision and recall. It provides a balanced measure of model performance, considering both precision and recall.

Based on the provided classification report, here's a comparison of the metrics for Best Stats Model and Random Forest:

Stats Model

Precision (Class 1): 0.312 Recall (Class 1): 0.268 F1-score (Class 1): 0.289

Random forest

Precision (Class 1): 0.268 Recall (Class 1): 0.256 F1-score (Class 1): 0.262

From these metrics, we can observe that Best Stats Model performs slightly better than Random Forest with reference to Class 1. It has higher precision, recall, and F1-score for Class 1.

Recommendations:

Based on the analysed information, the business recommendation would be to focus on Stats Model for predicting defaults in companies. Here's why:

Higher Precision: Stats Model has a slightly higher precision for identifying default cases (Class 1). This means that when the model predicts a default, it is more likely to be correct, reducing the risk of false alarms.

Balanced F1-score: Stats Model has a higher F1-score for default cases, indicating a better balance between precision and recall. This means that the model is able to identify defaults while minimizing both false positives and false negatives.

Importance of Identifying Defaults: Given the consequences of defaults, such as lower credit ratings and increased costs for the company, it is crucial to accurately identify potential default cases. Stats Model, with its better precision and F1-score, provides a more reliable prediction of defaults, allowing businesses or investors to make informed decisions about credit exposure and investment opportunities.

Consider Other Factors: While model performance is important, it's also necessary to consider other factors, such as the specific industry, market conditions, and additional data that might be relevant to predicting defaults. These factors should be taken into account when implementing the model in a real-world scenario.

Overall, Stats Model is recommended due to its better performance metrics for identifying defaults.

End of the Report