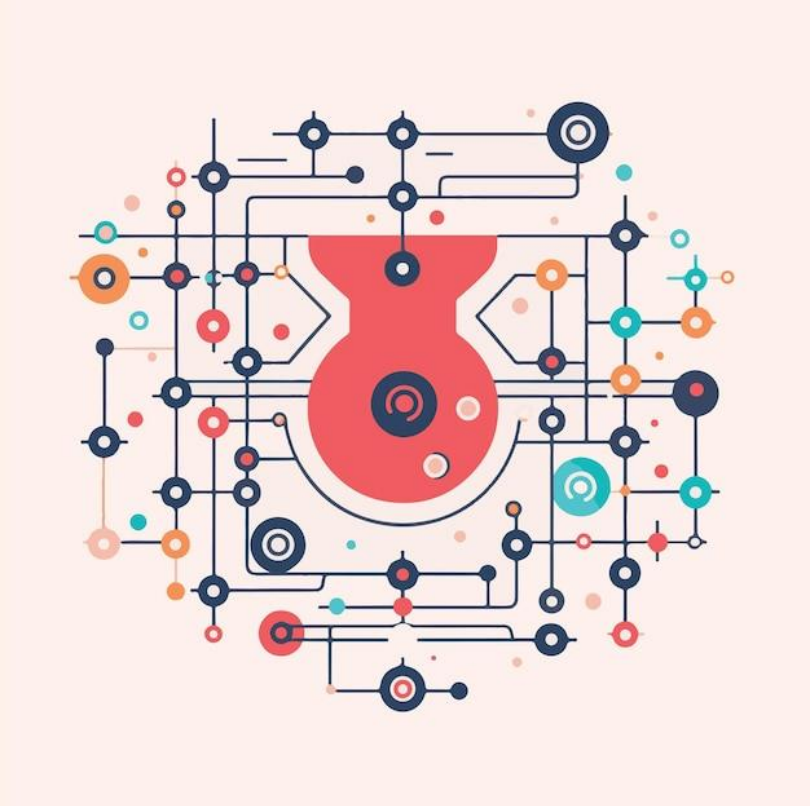

FINANCE RISK ANALYTICS PROJECT

By Kurt Warren Mario Gilby On November 24th 2024

Submitted to

Great Learning
POWER AHEAD

As a part of the requirements for completion of PGP-DSBA offered in affiliation with

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Table of Contents

List of Figures

List of Tables

No table of figures entries found.

List of Equations

No table of figures entries found.

Part A

Overview

The venture capitalists aim to build a Financial Health Assessment Tool to evaluate the financial stability and creditworthiness of companies. The tool will assist in:

Debt Management Analysis: Identifying trends and assessing companies' ability to manage and fulfil their financial obligations effectively.

Credit Risk Evaluation: Estimating the risk of default through financial metrics like liquidity ratios, debt-to-equity ratios, and other indicators.

As a Data Scientist, need to develop a predictive model using the given dataset, which includes detailed financial metrics of companies. The model's objective is to classify companies as defaulters or non-defaulters, based on whether their Net Worth Next Year is positive or negative.

Key Deliverables:

Target Variable: "Net Worth Next Year" – Positive values indicate a non-defaulter, while negative values indicate a defaulter.

Model Inputs: Financial metrics like total assets, total liabilities, debt-to-equity ratio, and other balance sheet indicators.

Outcome: A robust machine learning model that predicts the likelihood of default.

Enabling stakeholders to:

Identify companies at financial risk.

Make informed decisions about debt management and investments.

Business Impact: Facilitate proactive risk mitigation strategies and enhance the decision-making process for investors and businesses.

Questions Asked

Define the problem and perform Exploratory Data Analysis.

Problem definition

develop a predictive model using the given dataset, which includes detailed financial metrics of companies. The model's objective is to classify companies as defaulters or non-defaulters, based on whether their Net Worth Next Year is positive or negative.

Check shape

- Data has 4256 rows/observations
- Data has 51 columns/features.

Check Data Types

- 50 columns/features are of data type float

Statistical Summary

- There are substantial differences in scale among variables.
- Equity Face value has same value in the three percentiles, the value is 10, most of the values in this column as similar
- There are missing values for quite a few values and we will need treat them.

	count	mean	std	min	25%	50%	75%	max
Networth Next Year	4256	1345	15937	-74266	4	72	331	805773
Total assets	4256	3574	30074	0	91	316	1121	1176509
Net worth	4256	1352	12961	0	31	105	390	613152
Total income	4025	4688	53919	0	107	455	1485	2442828
Change in stock	3706	44	437	-3029	-2	2	18	14186
Total expenses	4091	4356	51398	-0	97	427	1396	2366035
Profit after tax	4102	295	3080	-3908	0	9	53	119439
PBDITA	4102	606	5646	-441	7	37	159	208576
PBT	4102	410	4217	-3895	1	13	74	145293
Cash profit	4102	408	4144	-2246	3	19	96	176912
PBDITA as % of total income	4177	3	172	-6400	5	10	16	100
PBT as % of total income	4177	-18	420	-21340	1	3	9	100
PAT as % of total income	4177	-20	424	-21340	0	2	6	150
Cash profit as % of total income	4177	-9	300	-15020	2	6	11	100
PAT as % of net worth	4256	10	62	-749	0	8	20	2467
Sales	3951	4646	53081	0	113	469	1481	2384984
Income from fincial services	3145	81	1043	0	0	2	10	51938
Other income	2700	56	1178	0	0	2	6	42857
Total capital	4251	225	1685	0	13	43	103	78273
Reserves and funds	4158	1211	12816	-6526	5	55	283	625138
Borrowings	3825	1176	8581	0	24	100	358	278257
Current liabilities & provisions	4146	961	9141	0	18	70	266	352240
Deferred tax liability	2887	234	2106	0	3	14	51	72797
Shareholders funds	4256	1376	13011	0	32	108	409	613152
Cumulative retained profits	4211	937	9853	-6534	1	37	206	390134
Capital employed	4256	2434	20496	0	61	221	790	891409
TOL/TNW	4256	4	21	-350	1	1	3	473
Total term liabilities / tangible net worth	4256	2	16	-326	0	0	1	456
Contingent liabilities / Net worth (%)	4256	56	369	0	0	5	31	14704
Contingent liabilities	2854	949	12057	0	6	38	195	559507
Net fixed assets	4124	1209	12502	0	26	94	353	636605
Investments	2541	722	6794	0	1	8	64	199979
Current assets	4176	1350	10156	0	37	148	515	354815
Net working capital	4219	163	3182	-63839	-1	17	86	85783
Quick ratio (times)	4151	1	9	0	0	1	1	341
Current ratio (times)	4151	2	12	0	1	1	2	505
Debt to equity ratio (times)	4256	3	16	0	0	1	2	456
Cash to current liabilities (times)	4151	1	5	0	0	0	0	165
Cash to average cost of sales per day	4156	145	2522	0	3	8	22	128041
Creditors turnover	3865	17	76	0	4	6	12	2401
Debtors turnover	3871	18	90	0	4	6	12	3135
Finished goods turnover	3382	84	563	-0	8	17	40	17948
WIP turnover	3492	29	170	-0	5	10	20	5651
Raw material turnover	3828	18	343	-2	3	6	12	21092
Shares outstanding	3446	23764910	170979041	-2147483647	1308382	4750000	10906020	4130400545
Equity face value	3446	-1095	34101	-999999	10	10	10	100000
EPS	4256	-196	13062	-843182	0	1	10	34523
Adjusted EPS	4256	-198	13062	-843182	0	1	8	34523
Total liabilities	4256	3574	30074	0	91	316	1121	1176509
PE on BSE	1629	55	1304	-1117	3	9	17	51008

Figure 1 Describe and Stats of the Data

Univariate Analysis

- All the columns have outliers, we would need to treat them

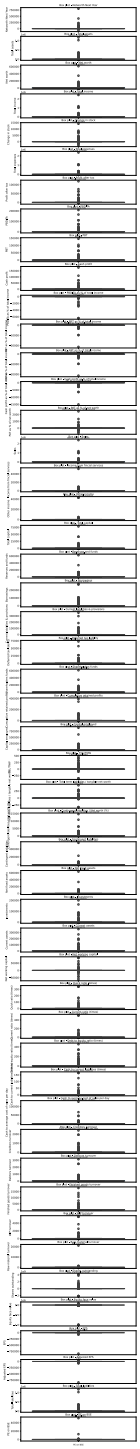

Figure 2 Boxplots before Outlier Treatment

Multivariate Analysis

- The airport shows weak and relations of "Net worth Next Year" with the non-ratio features.
- no relations of "Net worth Next Year" with the ratio features.
- relations of between the non-ratio and ratio features.

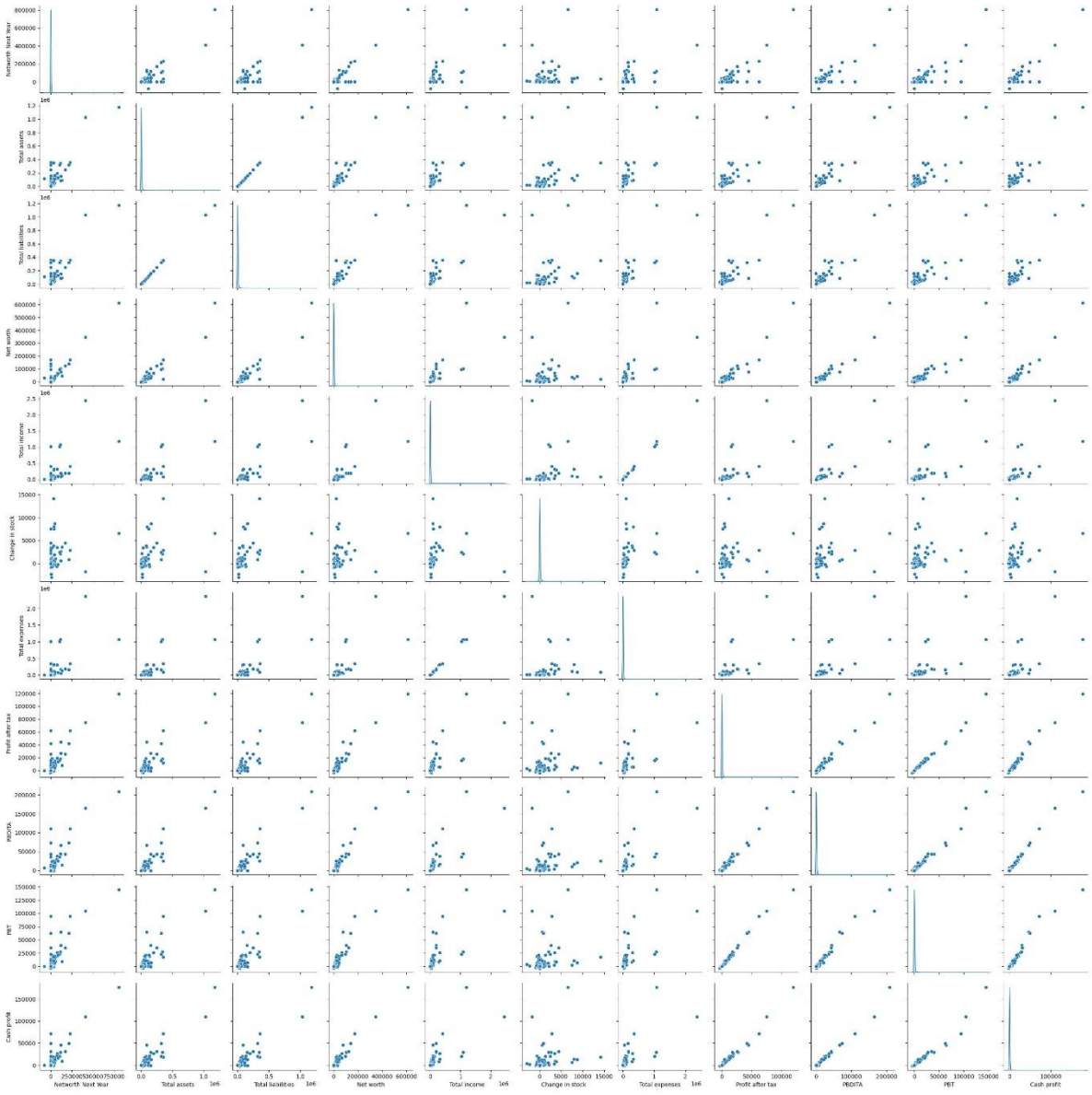

Figure 3 Pair Plot Column set 1

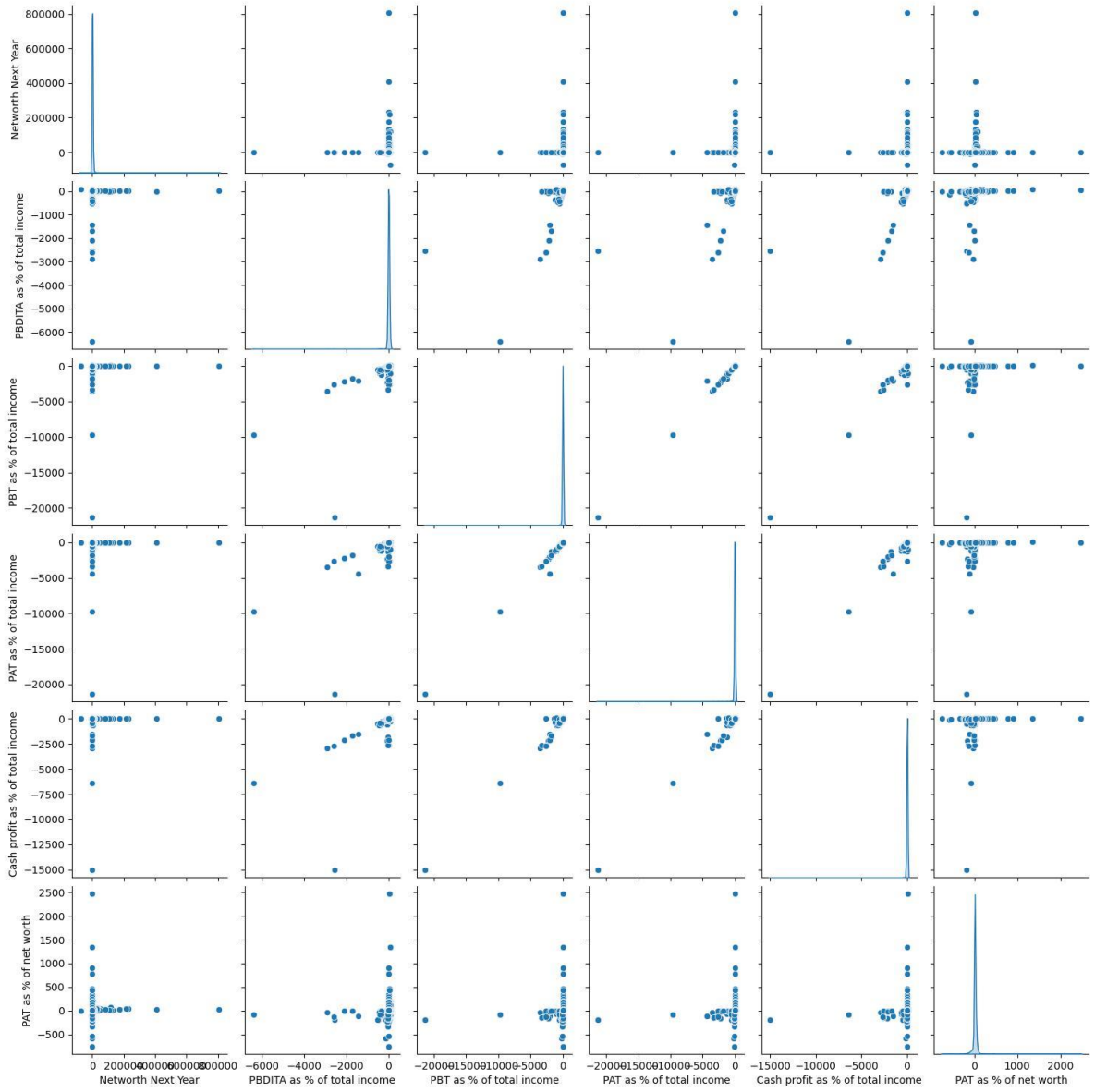

Figure 4 Pair plot Column set 2

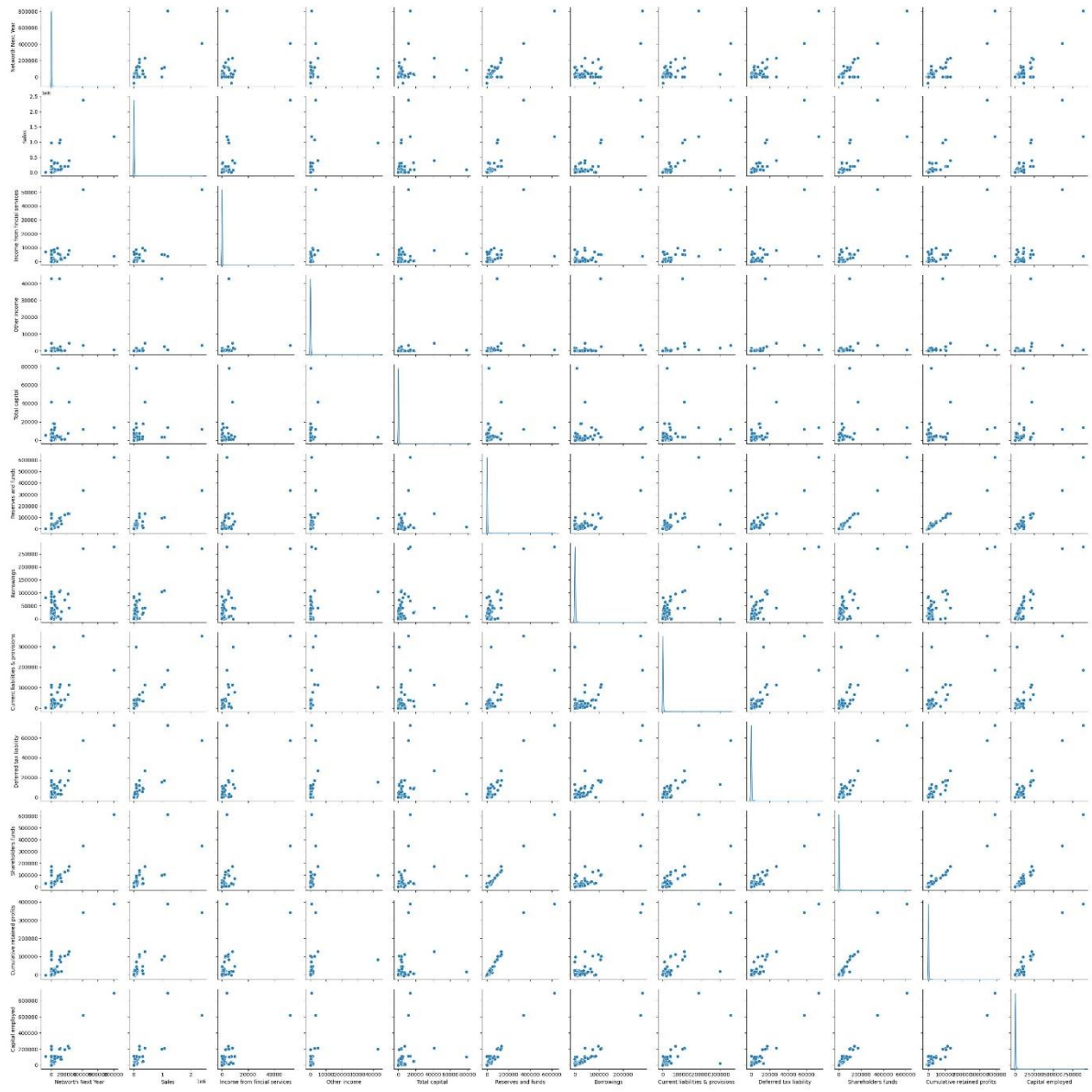

Figure 5 Pair plots Column set 3

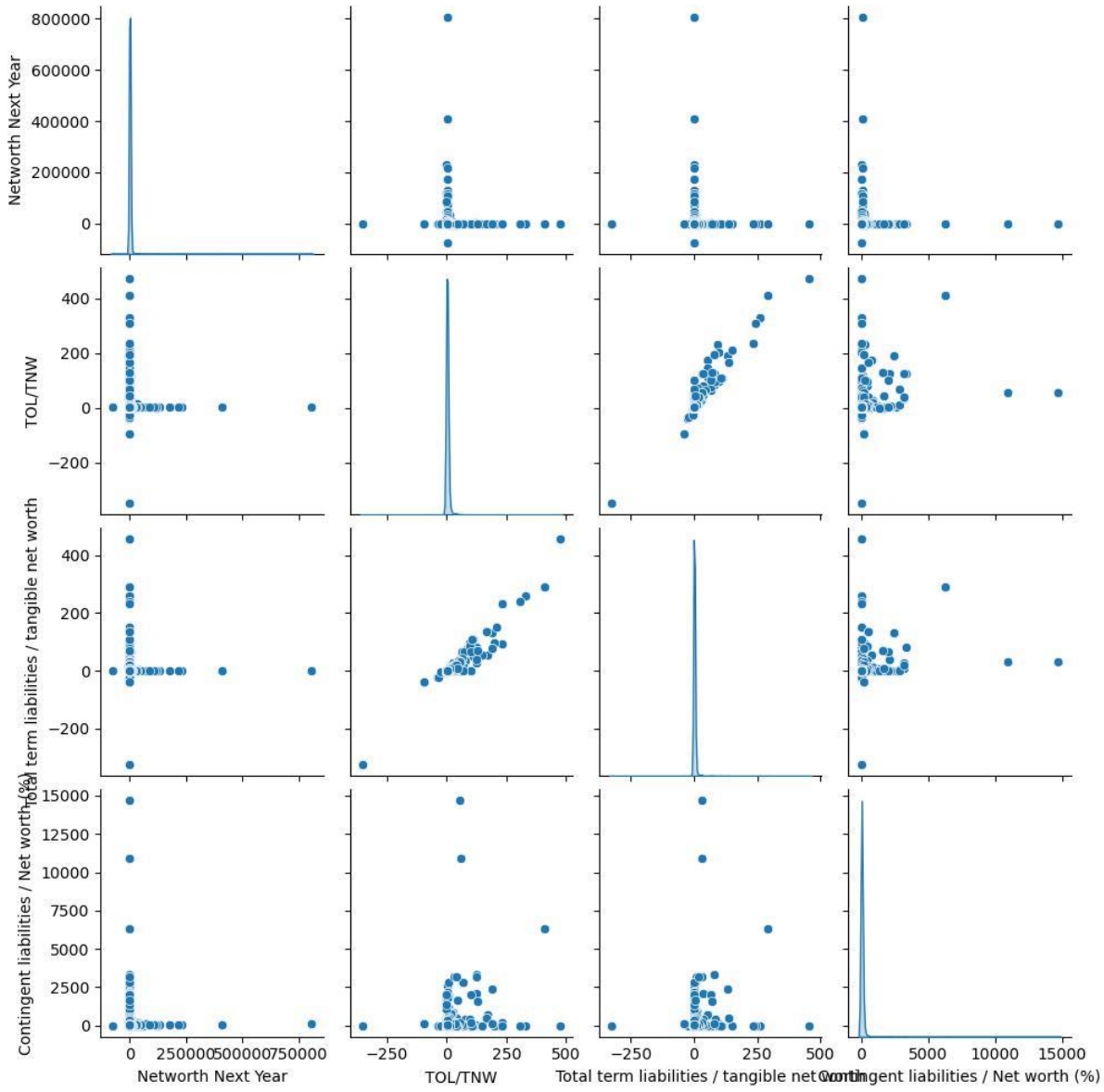

Figure 6 Pair plots Column set 4

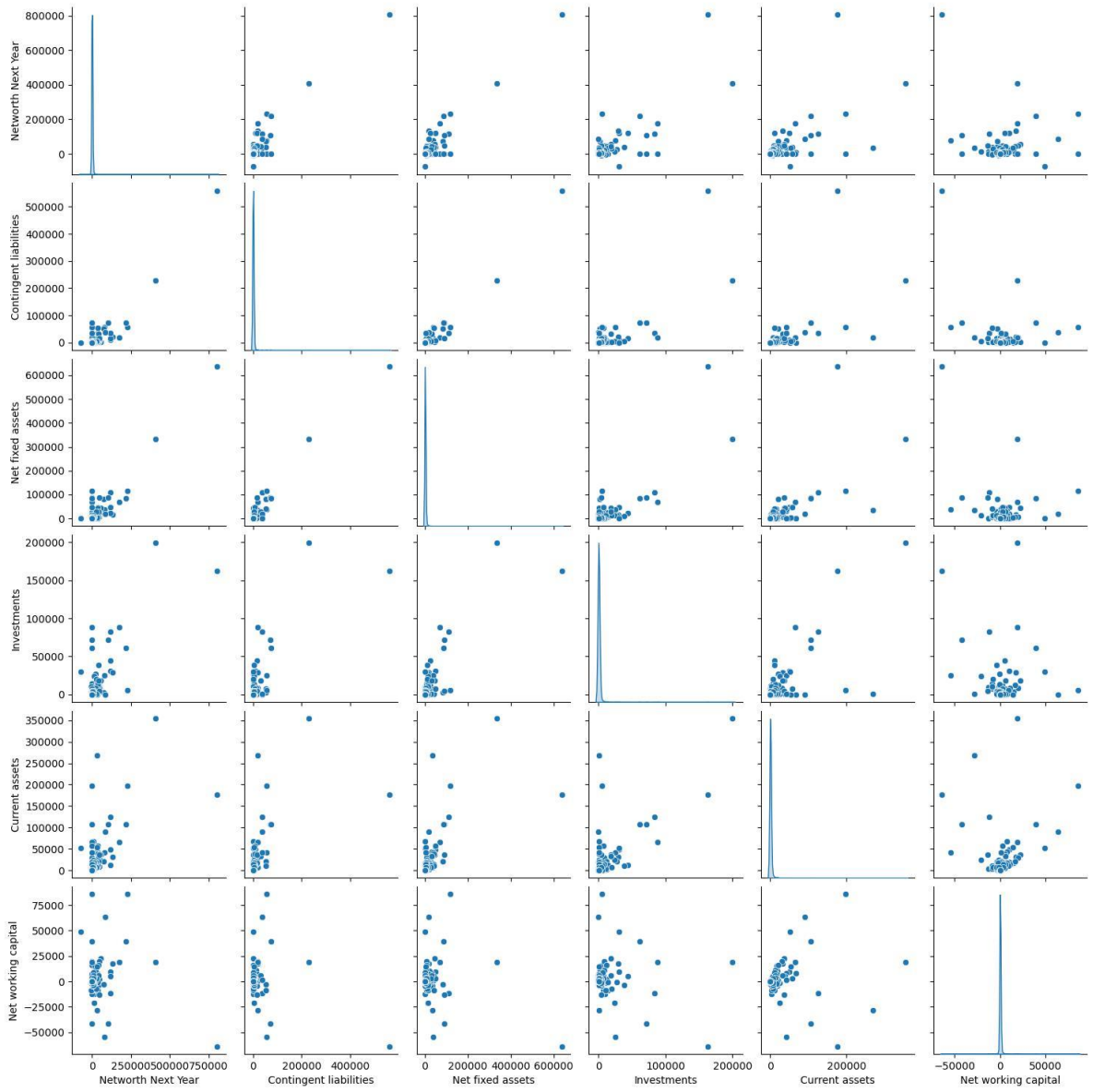

Figure 7 Pair plots Cloudnet 5

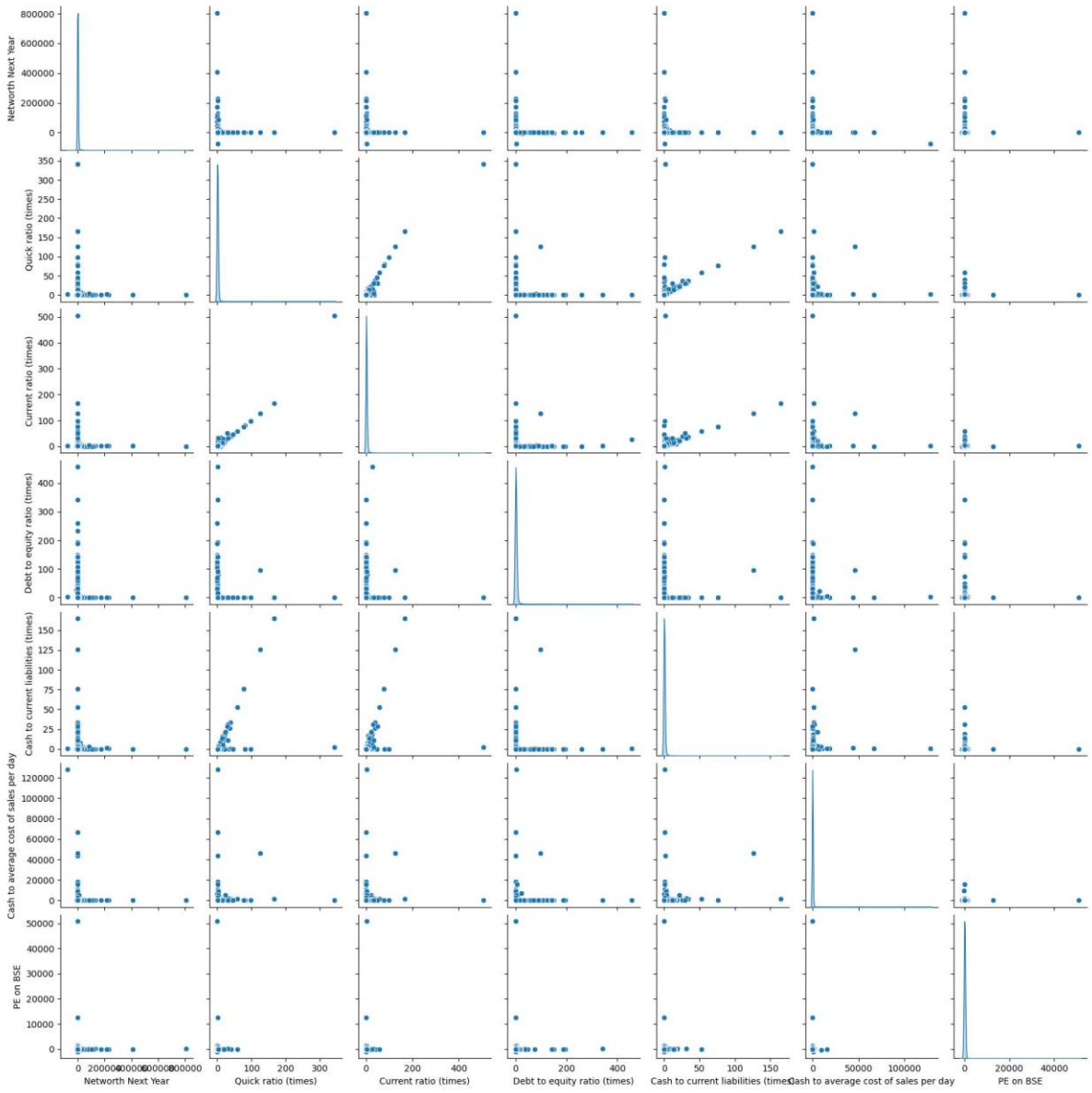

Figure 8 Pair plots Column set 6

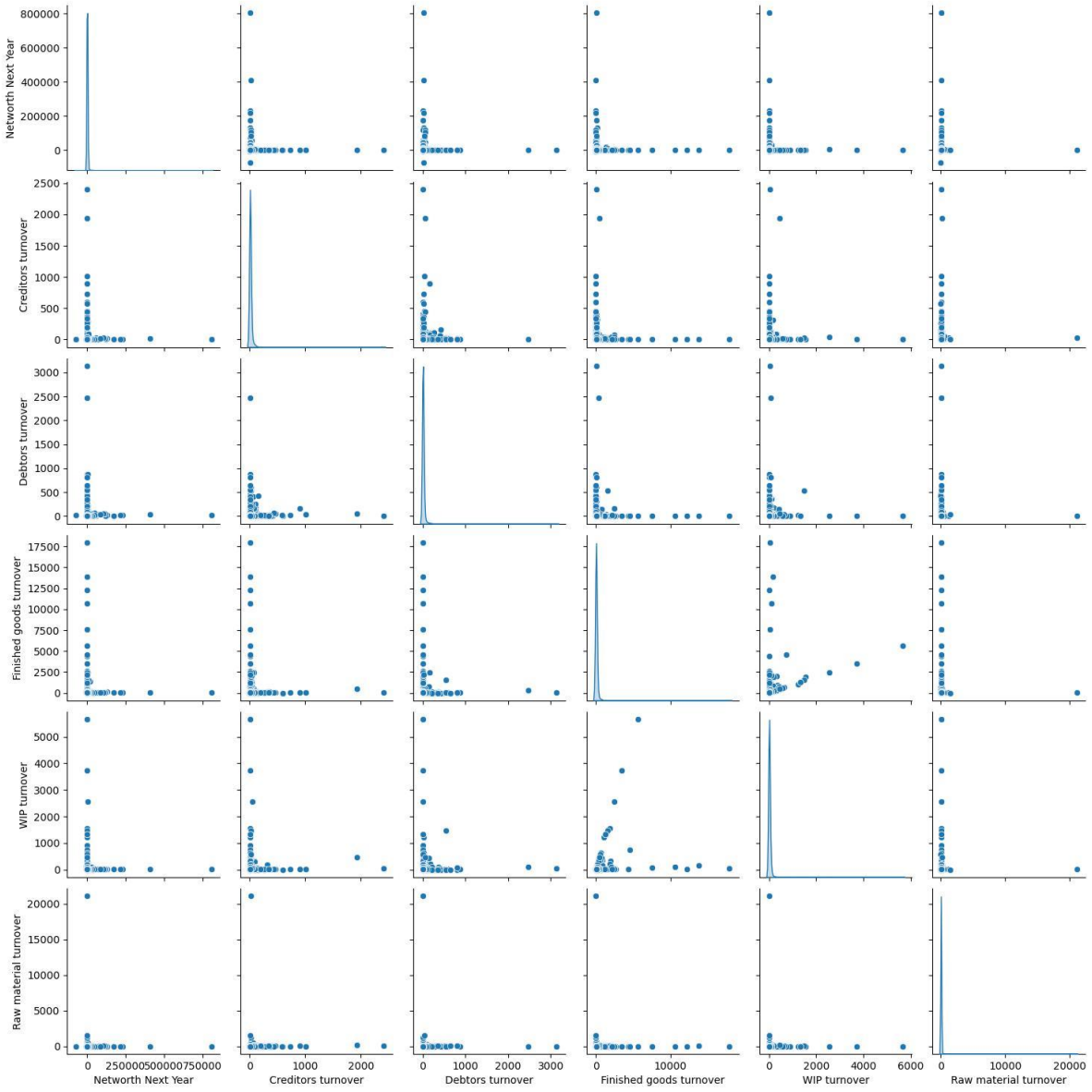

Figure 9 Pair plots Column set 7

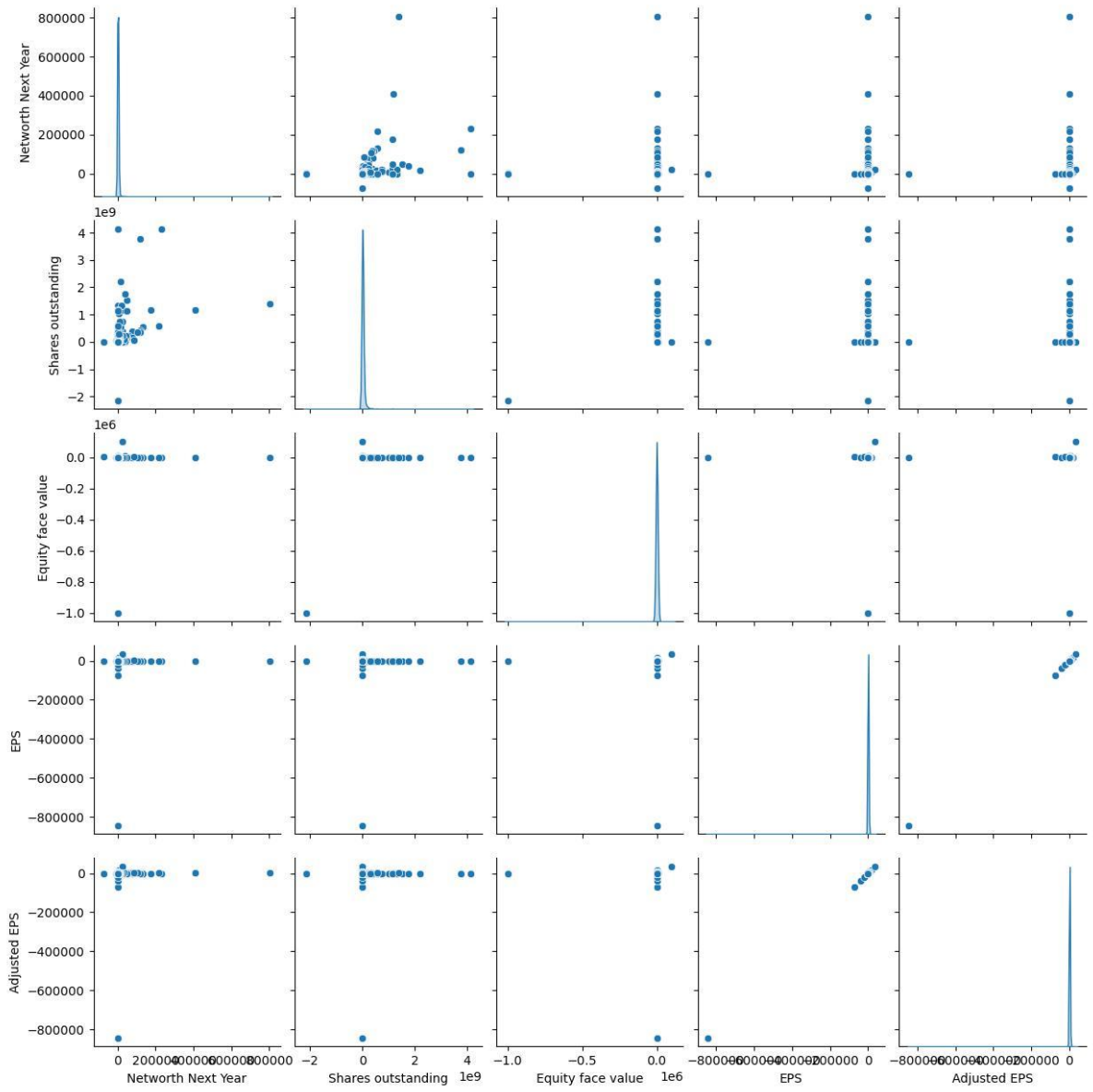

Figure 10 Pair plots Column set 8

- The heatmap shows high corelation of "Net worth Next Year", with ~19 features.
- shows high corelation between a feature which points us in the direction of the need for feature elimination using technique like pvalue or VIF

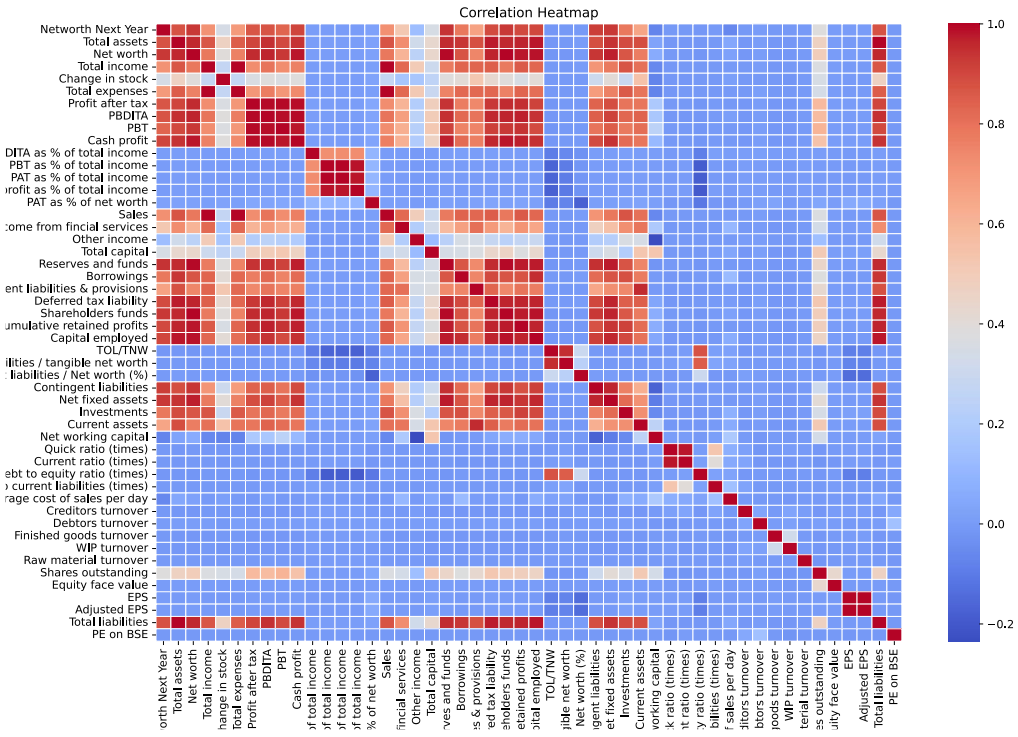

Figure 11 Heatmap

Key meaningful observations

- There are substantial differences in scale among variables.
- Equity Face value has same value in the three percentiles, the value is 10, most of the values in this column as similar.
- There are missing values for quite a few values and we will need treat them.
- All the columns have outliers, we would need to treat them.
- The pair plot shows weak and relations of "Net worth Next Year" with the non-ratio features.
- The pairpolt no relations of "Networth Next Year" with the ratio features.
- The pairpolt relations of between the non-ratio and ratio features.
- The heatmap shows high corelation of "Net worth Next Year", with ~19 features.
- The heatmap shows high corelation between a feature which points us in the direction of the need for feature elimination using technique like pvalue or VIF.
-

Question 2: Data Preprocessing

Missing Values Treatment

- There are 17778 missing values of total 212800 values.
- Percentage of missing = 8.354323308270677
- There are 1062 rows with great than 5 missing values.
- Percentage of rows missing more than 5 values = 24.953007518796994
- A lot of values are missing we will use KNN to impute missing values

-

Outlier Treatment

- All Columns/Features have outliers We will treat as we want to go Logistic regression in addition to other classification models.
- Outliers treated with threshold values of
    - lower range= Q1-(1.5 * IQR)
    - upper range= Q3+(1.5 * IQR)

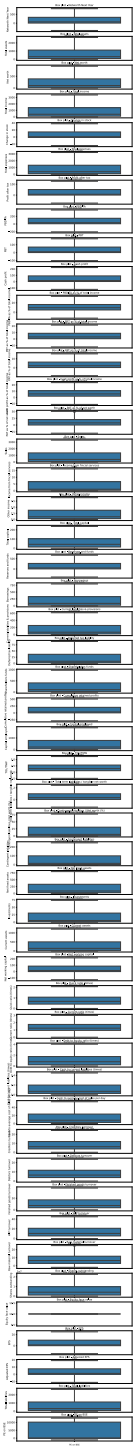

Figure 12 Boxplots after outlier treatment

Train-Test Split

- -We split the dataset in a 70-30.
- Create the Defaulter Target column Split the data
- The Defaulter column is less than 30% we will use stratify

Scale Data

- Scaling data as we want to do Logistic regression in addition to other classification models.

Question 3: Model Building - Original Data

- We will build two models Logistic Regression and Radom Forest.
- We will check the performance of the two models using metrics of Recall primary as this is Defaulter Classification
- Compare the other metrics like accuracy, precision etc. once a good recall is found.

Logistic Regression Model without feature selection

Logistic Regression Train Results

	precision	recall	f1-score	support
0	0.79	1.00	0.88	2356
1	0.62	0.01	0.02	623
accuracy			0.79	2979
macro avg	0.71	0.50	0.45	2979
weighted avg	0.76	0.79	0.70	2979

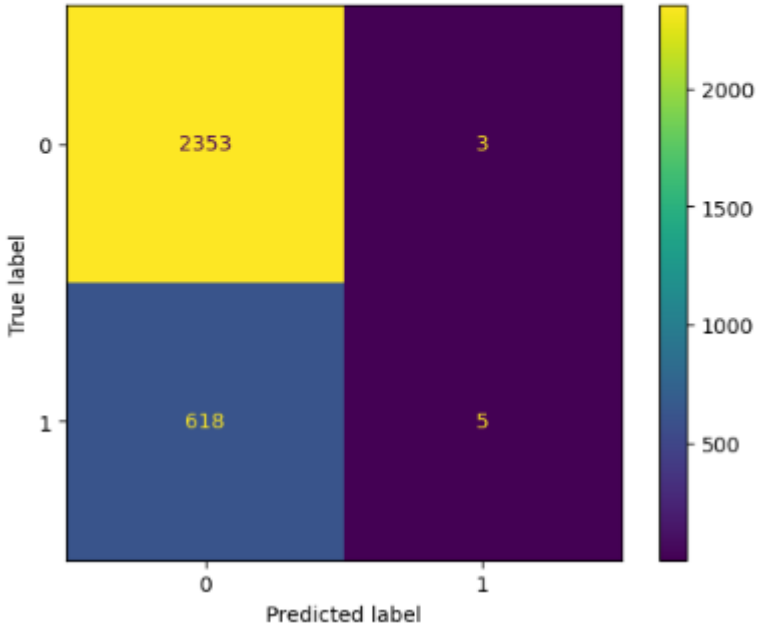

Figure 13 LR Model wo Feature selection Performance Train

Logistic Regression Test Results

	precision	recall	f1-score	support
0	0.79	1.00	0.88	1010
1	0.75	0.01	0.02	267
accuracy			0.79	1277
macro avg	0.77	0.51	0.45	1277
weighted avg	0.78	0.79	0.70	1277

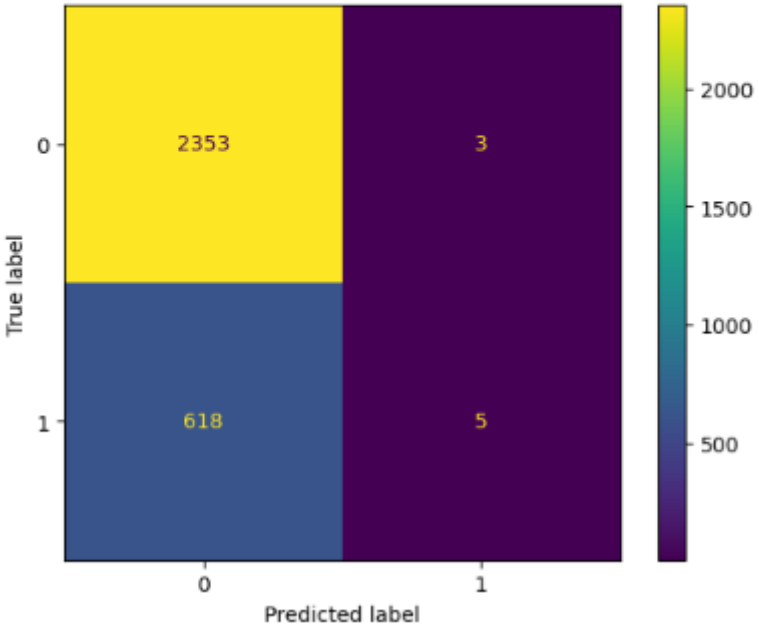

Figure 14 LR Model wo Feature Selection Performance Test

Random Forest Model without feature selection

Random Forest Train Results

	precision	recall	f1-score	support
0	0.94	0.98	0.96	2356
1	0.92	0.74	0.82	623
accuracy			0.93	2979
macro avg	0.93	0.86	0.89	2979
weighted avg	0.93	0.93	0.93	2979

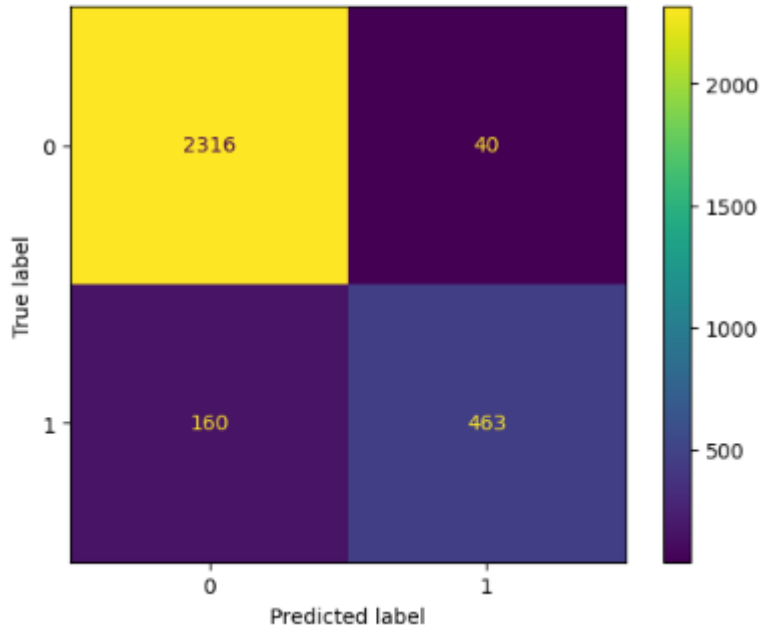

Figure 15 RF Model wo Feature selection Train

Random Forest Test Results

	precision	recall	f1-score	support
0	0.78	0.85	0.82	1010
1	0.15	0.10	0.12	267
accuracy			0.70	1277
macro avg	0.47	0.48	0.47	1277
weighted avg	0.65	0.70	0.67	1277

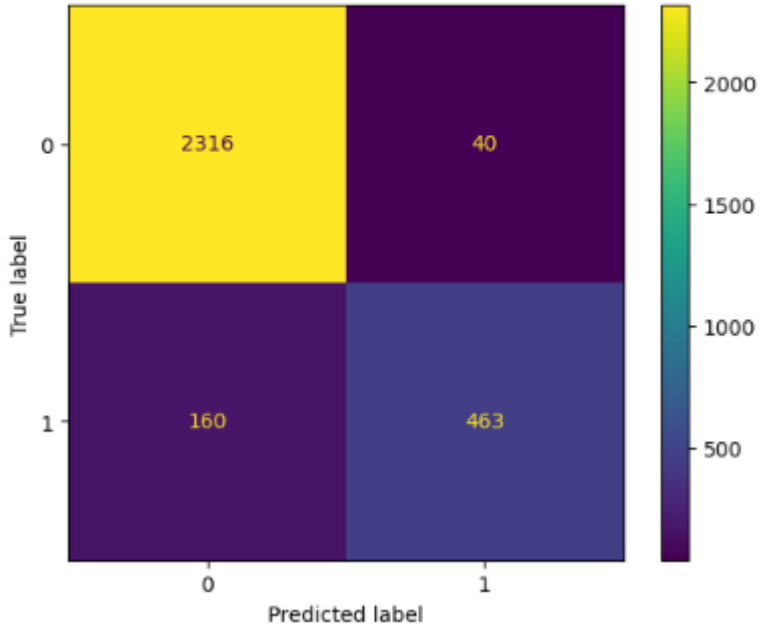

Figure 16 RF Model wo Feature Selection Test

Observations

- The Logistic Regression model does not perform good with both the train and test data set with the recall for "Defaulters" only at 0.01 for both.
- The Random Forest model does better performance on train with recall for "Defaulters" only at 0.74 but fails in test with a recall of only 0.10, which tells us it is over fitting

Question 4: Model Performance Improvement

- We will try to improve the both the models with using methods like
  - VIF
  - Pvalue method
  - Optimal Threshold for Logistic Regression and parameter tunning for Random Forest

VIF(Variance Inflation Factor)

- The VIF helps drop 11 Features for a threshold of 5,
  - Equity Face values
  - Total Assets
  - Sales
  - Total Income
  - Net Worth
  - Total Liabilities
  - PBT
  - -PBDITA
  - PBT as a % of Total Income
  - Capital Employed
  - Current Assets

Pvalue based selection:

- We used the Logit Model to get the pvalue
- Features with pvalue > 0.05 were removed concurrently.
- After Pvalue Selection we have 5 Features
  - Total expenses
  - Cash profit as % of total income
  - Reserves and funds
  - Current ratio (times)
  - Raw material turnover

LR Model Performance post feature reduction

Logistic Regression Train Results

	precision	recall	f1-score	support
0	0.79	1.00	0.88	2356
1	0.50	0.00	0.00	623
accuracy			0.79	2979
macro avg	0.65	0.50	0.44	2979
weighted avg	0.73	0.79	0.70	2979

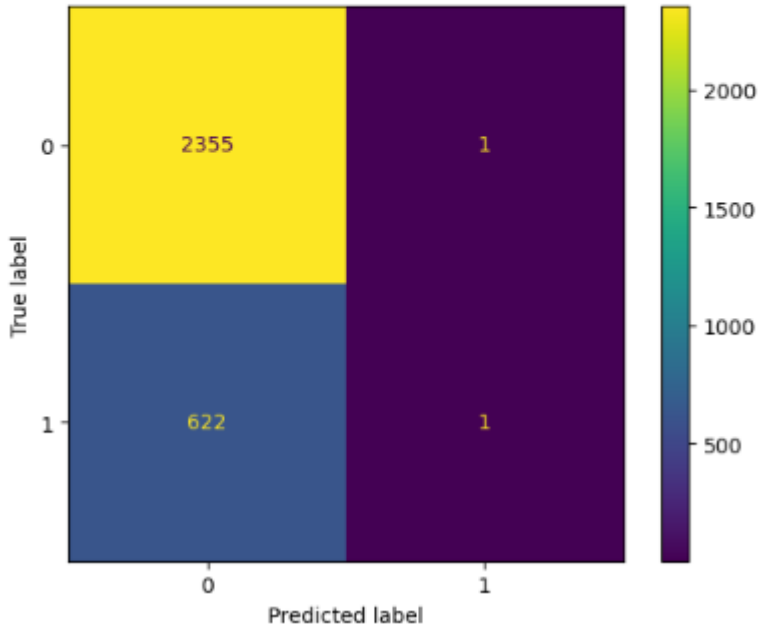

Figure 17 LR Model Performance on Train post Feature selection

Logistic Regression Test Results

	precision	recall	f1-score	support
0	0.79	1.00	0.88	1010
1	1.00	0.00	0.01	267
accuracy			0.79	1277
macro avg	0.90	0.50	0.45	1277
weighted avg	0.84	0.79	0.70	1277

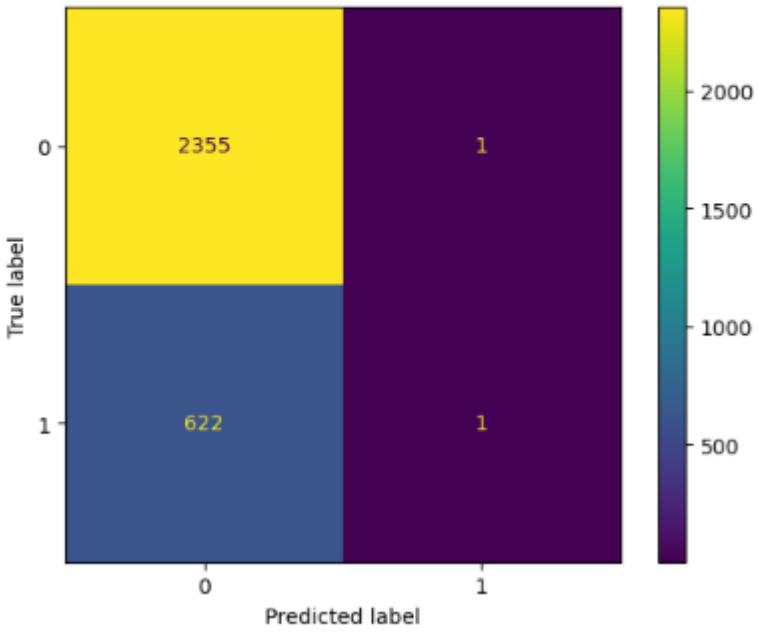

Figure 18 LR Model Performance on Test post Feature selection

LR Model Performance post feature reduction reduced so we plot ROC Curve to get optimal threshold

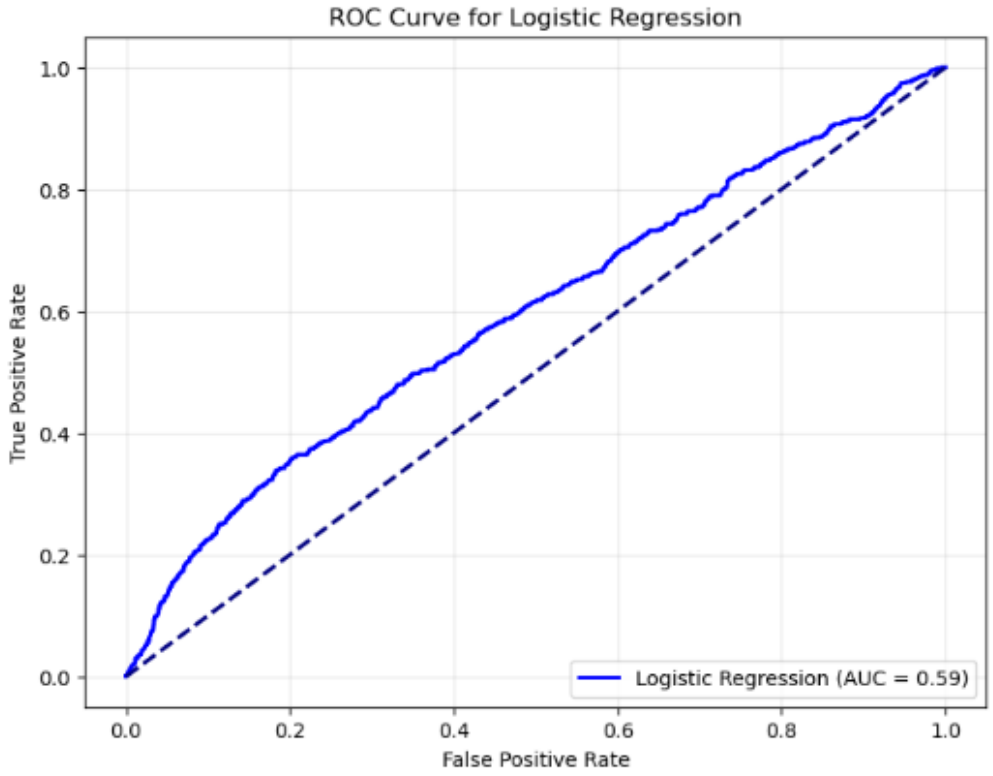

Figure 19 LR ROC Curve post feature reduction

LR Model Performance post feature reduction and using optimal Threshold

Logistic Regression Train Results as of Threshold

	precision	recall	f1-score	support
0	0.82	0.80	0.81	2356
1	0.32	0.36	0.34	623
accuracy			0.71	2979
macro avg	0.57	0.58	0.57	2979
weighted avg	0.72	0.71	0.71	2979

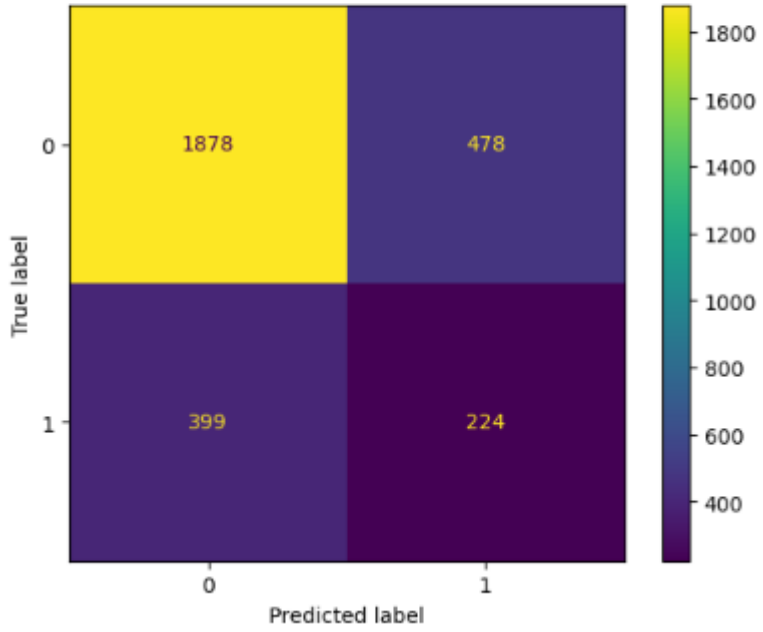

Figure 20 LR Model Performance with Feature Reduction and Optimal Threshold on Train

Logistic Regression Test Results as of Threshold

	precision	recall	f1-score	support
0	0.82	0.80	0.81	2356
1	0.32	0.36	0.34	623
accuracy			0.71	2979
macro avg	0.57	0.58	0.57	2979
weighted avg	0.72	0.71	0.71	2979

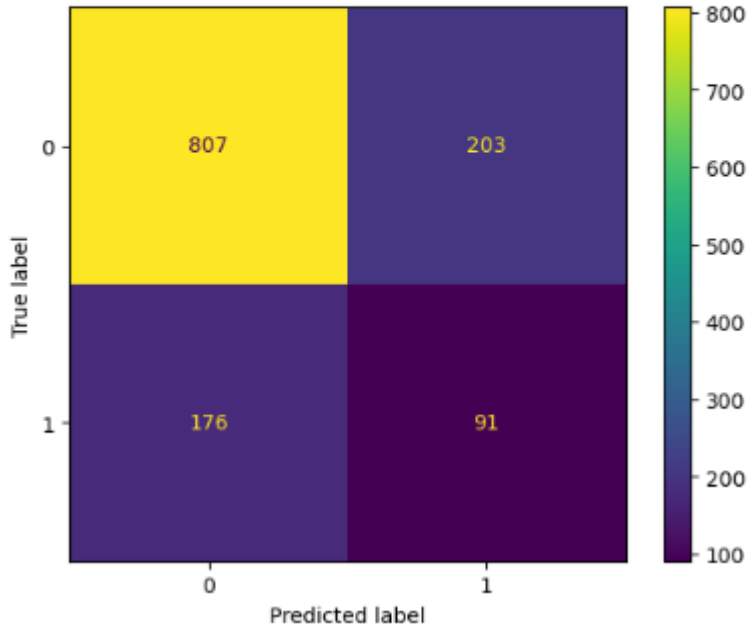

Figure 21 LR Model Performance with Feature Reduction and Optimal Threshold on Test

- Still The Logistic Model is not performing good based on the recall which is only in the 30%¶

Hyperparameter Tuning for RF Model

- Best Parameters: {'max_depth': None, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}
- Best Recall: 0.07703225806451612
- This is still not very good.

Random Forest Train Results

	precision	recall	f1-score	support
0	0.82	0.99	0.90	2356
1	0.86	0.20	0.32	623
accuracy			0.83	2979
macro avg	0.84	0.59	0.61	2979
weighted avg	0.83	0.83	0.78	2979

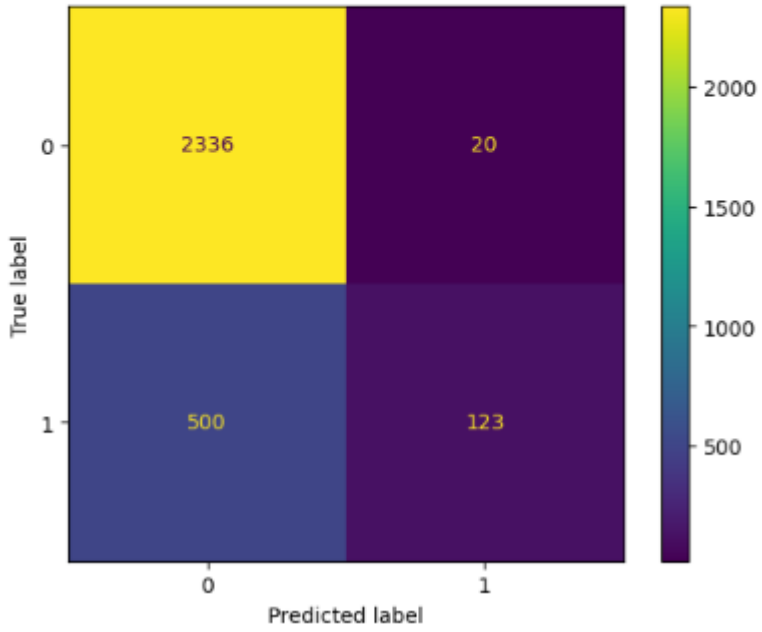

Figure 22 RF Model Performance with Feature reduction and Hyperparameter Tuning on Train

Random Forest Test Results

	precision	recall	f1-score	support
0	0.80	0.96	0.87	1010
1	0.35	0.09	0.14	267
accuracy			0.78	1277
macro avg	0.58	0.52	0.51	1277
weighted avg	0.71	0.78	0.72	1277

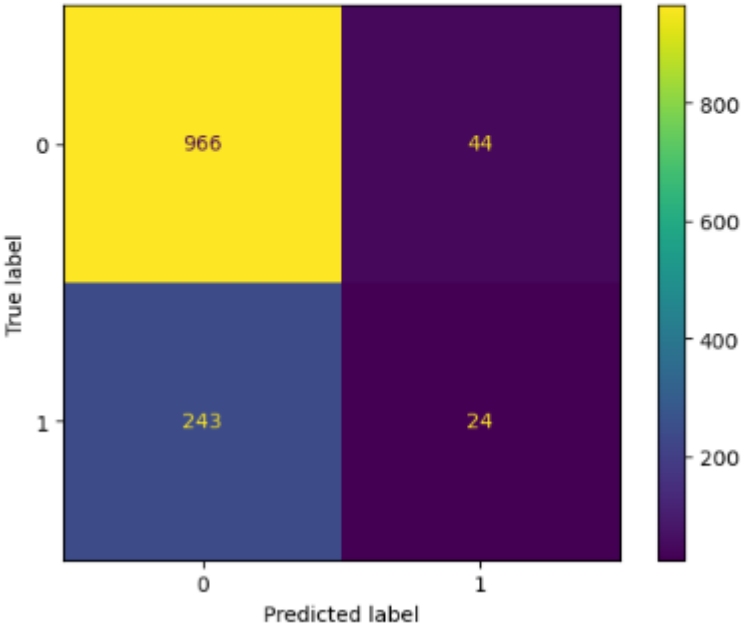

Figure 23 RF Model Performance with Feature reduction and Hyperparameter Tuning on Test

Since the models are still not good use SMOTE to oversample Defaults
Logistic regression Post SMOTE with Feature reduction and Optimal Threshold

Logistic Regression Train Results as of Threshold

	precision	recall	f1-score	support
0	0.61	0.88	0.72	2356
1	0.59	0.24	0.34	1767
accuracy			0.60	4123
macro avg	0.60	0.56	0.53	4123
weighted avg	0.60	0.60	0.56	4123

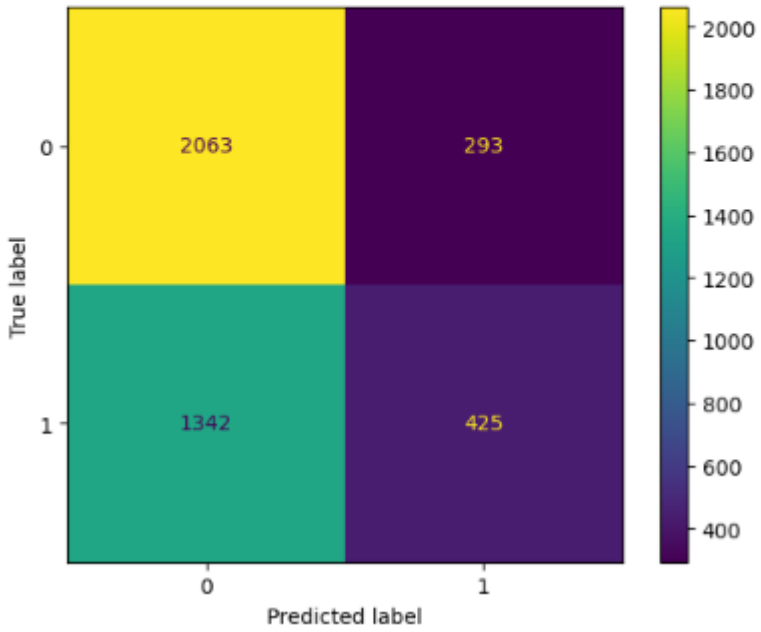

Figure 24 LR Model Performance with Feature Reduction, Optimal Threshold and SMOTE on Train

Logistic Regression Test Results as of Threshold

	precision	recall	f1-score	support
0	0.82	0.87	0.84	1010
1	0.35	0.25	0.29	267
accuracy			0.74	1277
macro avg	0.58	0.56	0.57	1277
weighted avg	0.72	0.74	0.73	1277

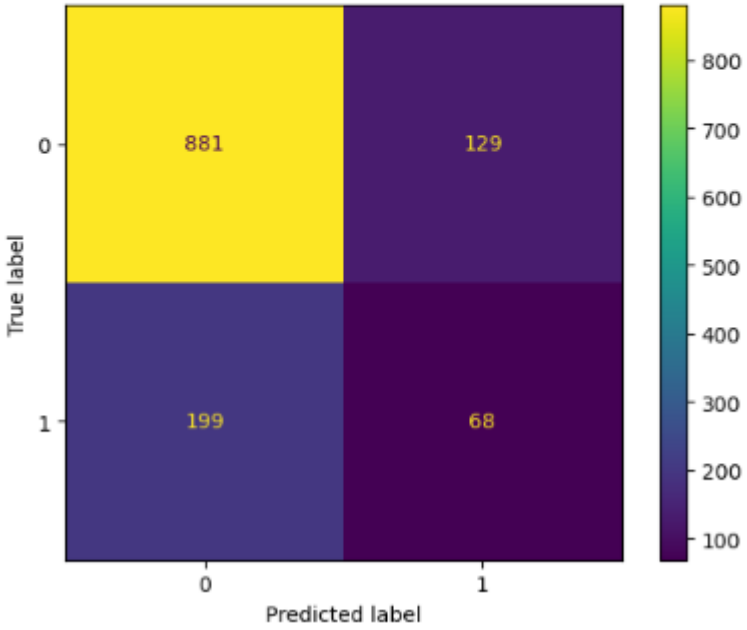

Figure 25 LR Model Performance with Feature Reduction, Optimal Threshold and SMOTE on Test

RF Post SMOTE with Feature reduction and Best Hyperparameters

Random Forest Train Results

	precision	recall	f1-score	support
0	0.88	0.94	0.91	2356
1	0.91	0.83	0.86	1767
accuracy			0.89	4123
macro avg	0.89	0.88	0.88	4123
weighted avg	0.89	0.89	0.89	4123

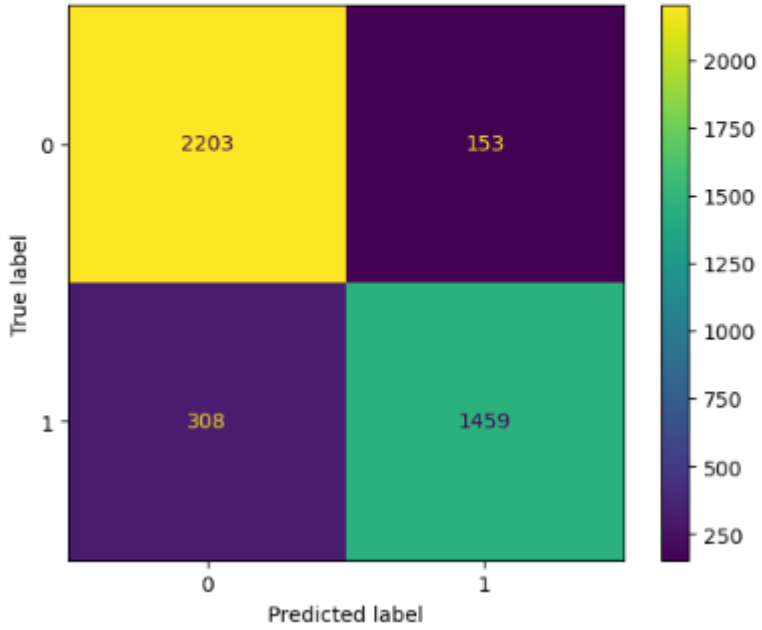

Figure 26 RF Model Performance post SMOTE, Feature reduction and using best estimator on Train

Random Forest Test Results

	precision	recall	f1-score	support
0	0.77	0.71	0.74	1010
1	0.17	0.22	0.19	267
accuracy			0.61	1277
macro avg	0.47	0.46	0.47	1277
weighted avg	0.65	0.61	0.63	1277

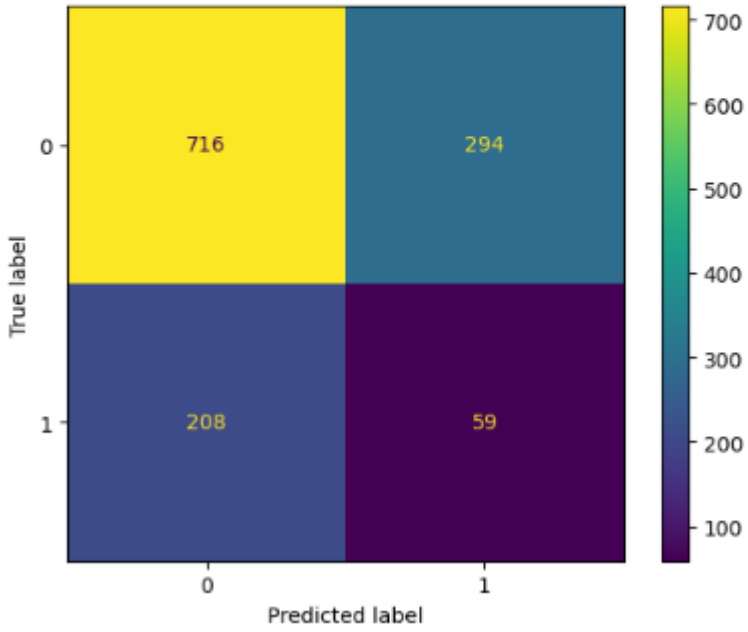

Figure 27 RF Model Performance post SMOTE, Feature reduction, using best estimator on Test

Question 5: Model Performance Comparison

Logistic regression and Random Forest are not the most suitable models for this classification problem

- We did get better and between in each step
- From base to VIF and P value Treatment to SMOTE
- We would need to look at other methods like PCA or other model like boosting/bagging to see if we can get a better model
- The final recall on Test by the models were 0.25 for LR and 0.22 for Random Forest which was much better than the 0.01 for LR and 0.10 for RF what we started with, but both these two models will not work in this case.
- important Feature list post VIF and Pvalue elimination are ordered as:
  - Raw material turnover 0.022
  - Total expenses 0.020
  - Current ratio (times) 0.012
  - Cash profit as % of total income 0.005
  - Reserves and funds 0.000

Question 6: Actionable Insights & Recommendations

- We would need to look at other methods like PCA or other model like boosting/bagging to see if we can get a better model

Part B

Question 1: Draw a Stock Price Graph (Stock Price vs Time) for the given stocks - Write observations

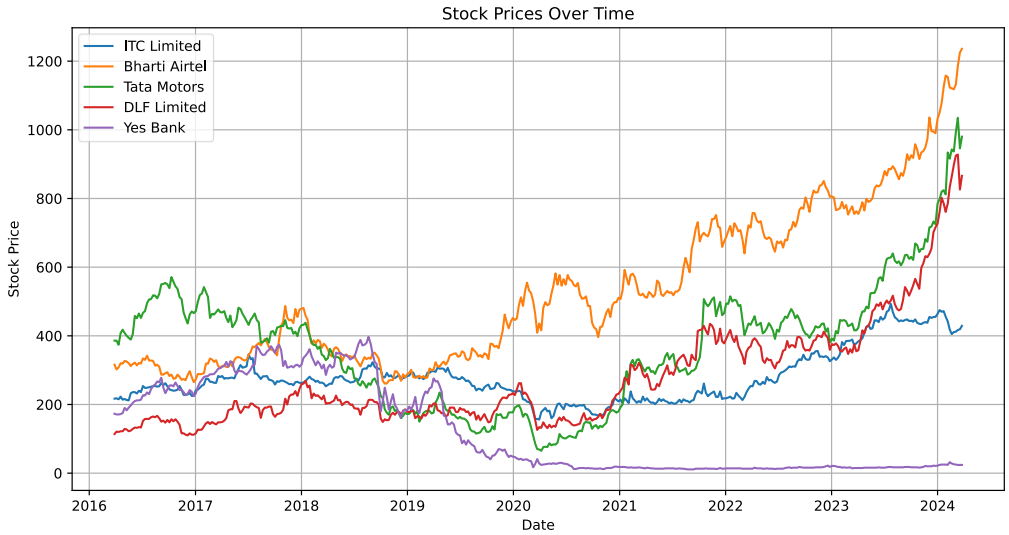

Figure 28 Stock Price Trends

Observations:

- The graph shows the stock price trends for all five stocks over the given period.
- Some stocks show consistent growth, while others exhibit high volatility or downward trends.
- Bharti Airtel has the highest gains vs Yes Bank who has the least

Question 2: Stock Returns Calculation and Analysis

- Calculate Returns for all stocks
    - Calculated as the % Change from one week to the next
- Calculate the Mean and Standard Deviation for the returns of all stocks
    - Calculated the Mean and STD for each Stock for the Series of Dates.
- Draw a plot of Mean vs Standard Deviation for all stock returns

Figure 29 Stocks Mean vs STD Plot

Observations and Inferences:

- The mean returns indicate the average performance of each stock.
- The standard deviation of returns shows the volatility (risk) associated with each stock.
- The Mean vs Standard Deviation plot highlights which stocks offer higher returns for a given level of risk.
- Bharti Airtel and ITC Limited give mid-range returns for the least vitality, DLF and Tata have higher returns than them but have mid-range volatility, Yes Bank has the least returns and the most volatility

Question 3: Actionable Insights & Recommendations

- High-return stocks with low volatility are ideal for risk-averse investors.(Airtel/ITC)
- Stocks with high volatility but significant returns might be suitable for risk-tolerant investors.(DLF/Tata)
- Portfolio diversification can balance high-return and low-risk stocks to achieve an optimal risk-adjusted return.(Mix of the above)