Finance Risk & Analytics Project

`

Great Learning
Authored by: Shrinidhi CG

Contents

Part A - Define the problem and perform exploratory Data Analysis

Problem definition

The problem is to develop a Financial Health Assessment Tool that predicts whether a company will be tagged as a defaulter based on its Net worth Next Year using historical financial data. The prediction will be made by leveraging machine learning techniques, and a company will be classified as a defaulter if its Net worth Next Year is negative.

Check shape, Data types, statistical summary

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

Datatype: *The most columns are represented as float64, indicating numerical data, while the Num column is of type int64, serving as an identifier. The consistent use of numerical data types suggests the dataset is well-suited for quantitative analysis once the missing values are handled.*

Descriptive Statistics: *Based on the descriptive statistics, below are specific insights:*

1) *Net Worth Distribution:*

    Mean Net Worth Next Year: $2,128.50 (suggests a positive net worth projection overall).
    High Variability: Net worth ranges from a minimum of $1 to a maximum of $4,256, indicating diverse financial conditions across companies.

2) *Assets and Income:*

Total Assets: Mean is $1,344.74, with a maximum of $805,773.40. Significant differences in asset sizes may indicate varying scales of operations.
Total Income: Mean income is $1,351.95 with a high standard deviation, suggesting some companies have exceptionally high or low income.

3) *Profitability Ratios:*

Profit After Tax: Average is $295.05, but with high variability, indicating that some companies are significantly more profitable.
PBDITA as % of Total Income: Mean is 4.36%, showing variability in how much profit before depreciation, tax, and amortization contributes to total income.

4) *Debt and Liabilities:*

Debt-to-Equity Ratio: Companies show a wide range of leverage, with mean values indicating some companies are highly leveraged.

TOL/TNW Ratio: The mean is around 2.38, suggesting moderate levels of total liabilities compared to net worth.

5) *Liquidity Ratios:*

Current Ratio: Mean is 1.53, indicating that, on average, companies have more current assets than current liabilities, though some may be under significant liquidity pressure.

Quick Ratio: Average of 0.77 suggests that some companies may have liquidity issues, as the quick ratio is below 1.

6) *Turnover Ratios:*

Debtors Turnover: Mean of 17.93 suggests companies are relatively efficient in collecting receivables, but high variance indicates some companies may struggle in this area.

7) *Market Indicators:*

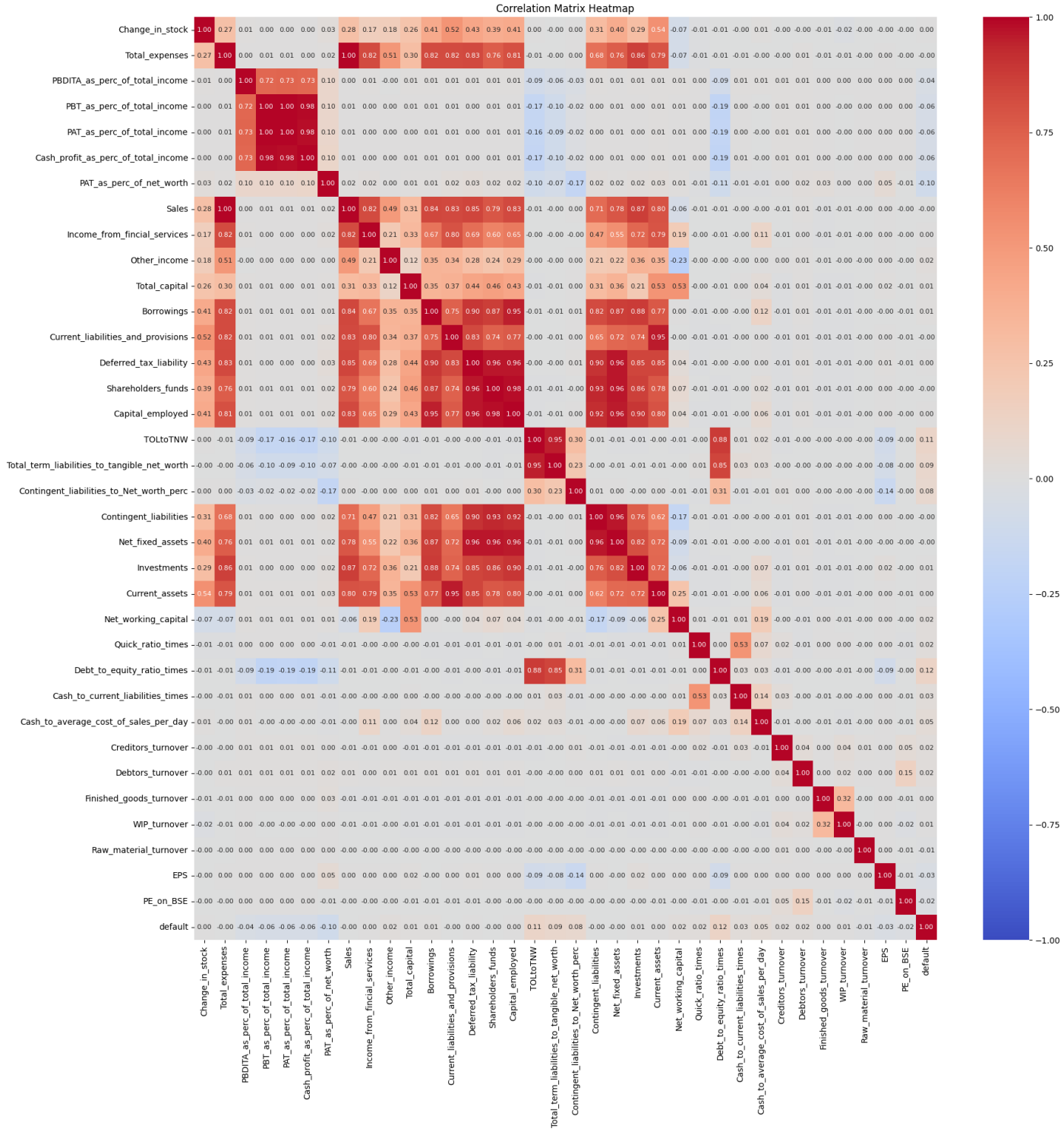
EPS: Mean is -1094.83, indicating that many companies may be experiencing losses or negative earnings.

8) *Contingent Liabilities:*

Mean Contingent Liabilities: High contingent liabilities relative to net worth may indicate potential future financial risks.

These insights reveal a broad range of financial health among companies, highlighting areas for potential improvement, risk assessment, and strategic planning.

Multivariate Analysis



Key meaningful observations on individual variables and the relationship between variables

The correlation matrix heatmap displays the relationships between multiple financial variables, allowing us to assess which variables are highly correlated (positively or negatively) with each other.

1. *Highly Positively Correlated Variables:*

Total_assets & Net_worth (0.93): This indicates that as total assets increase, net worth tends to increase significantly, which is expected as assets are a major component of net worth.

Total_income & Total_expenses (0.97): This strong positive correlation suggests that as a company's income increases, its expenses also rise, possibly due to variable costs scaling with revenue.

PBDITA & Total_income (0.94): Profit Before Depreciation, Interest, Tax, and Amortization (PBDITA) is closely tied to income, implying that revenue growth significantly impacts this profitability measure.

Capital_employed & Net_fixed_assets (0.99): The nearly perfect correlation suggests that fixed assets are a primary component of capital employed in the business.

2. *Highly Negatively Correlated Variables:*

Cash_to_current_liabilities_times & Total_liabilities (around -0.60): A strong negative correlation shows that companies with higher liabilities tend to have lower cash ratios, indicating potential liquidity issues.

Debt_to_equity_ratio_times & Net_worth (around -0.48): This suggests that higher leverage (debt) is typically associated with lower equity, impacting the company's financial structure and risk.

3. *Clusters of Interrelated Variables:*

Profitability Metrics (PBDITA, PBT, PAT, Cash Profit): These metrics show high positive correlations among themselves, indicating that a rise in one often signals an increase in others. For example, PBDITA_as_perc_of_total_income & PBT_as_perc_of_total_income (0.91) shows that operating profitability translates directly into pre-tax profitability.

Liquidity Measures (Current_ratio, Quick_ratio, Cash ratios): The correlations between different liquidity ratios suggest a close relationship, such as Current_ratio_times & Cash_to_current_liabilities_times (0.80), indicating that companies with better current ratios also tend to have better cash positions relative to liabilities.

Removing Redundant variables

Here are potential redundancies in the dataset:

Net worth and Net worth Next Year: *Since we are already using Net worth Next Year as a dependent variable, keeping both could introduce multicollinearity.*

Net worth = Total assets - Total liabilities: *Net worth can be computed using Total assets and Total liabilities making Net worth redundant unless it's part of the analysis directly. Also, total assets and total liabilities are highly correlated with majority of the variables, we can remove them to avoid multicollinearity.*

PBDITA, PBDITA as % of total income, PBT, and PBT as % of total income, Cash Profit as % of total income: *Since dataset has percentage variables, the absolute values of PBDITA and PBT might be redundant.*

PAT, PAT as % of total income, and PAT as % of net worth: *These percentages could make one of the variables redundant if both percentage and absolute terms are included in the model.*

Current ratio and Quick ratio: *These ratios measure similar things, with the Quick ratio excluding inventory. If inventory is less relevant, choose one of these.*

Shares outstanding, Equity face value, and EPS/Adjusted EPS: *Depending on how equity is used in your model, having both Shares outstanding and EPS could be redundant. You may only need EPS or Adjusted EPS as they already account for equity and shares.*

Total_income and Sales *are highly correlated because sales directly contribute to total income, reflecting a common financial relationship. To avoid redundancy and potential multicollinearity in models, we can drop the less critical variable.*

Reserves and Funds, Shareholders' Funds, and Cumulative Retained Profits *represent different parts of the company's retained earnings or equity. Shareholders' Funds include both Reserves and Cumulative Retained Profits. Since these metrics overlap in representing a company's retained earnings, they will be removed to avoid multicollinearity.*

Part A - Data Pre-processing

Null values & Outliers detection & Treatment

Variables	*Null Values*	*Outliers*
Change_in_stock	*550*	*750*
Total_expenses	*165*	*518*
PBDITA_as_perc_of_total_income	*79*	*346*
PBT_as_perc_of_total_income	*79*	*546*
PAT_as_perc_of_total_income	*79*	*610*

Cash_profit_as_perc_of_total_income	*79*	*426*
PAT_as_perc_of_net_worth	*0*	*427*
Sales	*305*	*500*
Income_from_fincial_services	*1111*	*517*
Other_income	*1556*	*389*
Total_capital	*5*	*551*
Borrowings	*431*	*532*
Current_liabilities_and_provisions	*110*	*581*
Deferred_tax_liability	*1369*	*406*
Shareholders_funds	*0*	*588*
Capital_employed	*0*	*572*
TOLtoTNW	*0*	*414*
Total_term_liabilities_to_tangible_net_worth	*0*	*406*
Contingent_liabilities_to_Net_worth_perc	*0*	*478*
Contingent_liabilities	*1402*	*393*
Net_fixed_assets	*132*	*569*
Investments	*1715*	*451*
Current_assets	*80*	*532*
Net_working_capital	*37*	*806*
Quick_ratio_times	*105*	*371*
Debt_to_equity_ratio_times	*0*	*381*
Cash_to_current_liabilities_times	*105*	*539*
Cash_to_average_cost_of_sales_per_day	*100*	*583*
Creditors_turnover	*391*	*442*
Debtors_turnover	*385*	*408*
Finished_goods_turnover	*874*	*399*
WIP_turnover	*764*	*378*
Raw_material_turnover	*428*	*296*
EPS	*0*	*638*
PE_on_BSE	*2627*	*237*
default	*0*	

Missing Value Insights:

Data Gaps in Key Financial Areas: *The missing values in sales, expenses, and liquidity measures suggest significant data gaps that can affect overall financial health analysis, particularly when assessing profitability, liquidity, and capital structure.*

Impact on Risk Analysis: *Missing data in turnover and leverage-related ratios can undermine risk assessments, especially for companies where liquidity management and debt levels are critical factors.*

Potential Bias in Financial Metrics: *Missing values in profitability percentages (like PBDITA and PBT) may lead to biased performance evaluations, especially in comparative analysis across firms or over time.*

Effective imputation of missing data is crucial to ensure the reliability of financial insights and decision-making, especially in variables critical to understanding operational efficiency, liquidity status, and financial risk profiles.

Dropping columns with more than 30% missing values

The number of rows (observations) is 4256
The number of columns (variables) is 35

Imputing the remaining missing values with KNNImputer

Part A - Model Building & Model Performance Improvement

Model Building using Logistic Regression

```
Optimization terminated successfully.
         Current function value: 0.507571
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:                 4256
Model:                          Logit   Df Residuals:                     4227
Method:                           MLE   Df Model:                           28
Date:                Sat, 07 Sep 2024   Pseudo R-squ.:                0.01848
Time:                        17:43:48   Log-Likelihood:                -2160.2
converged:                       True   LL-Null:                       -2200.9
Covariance Type:            nonrobust   LLR p-value:                 4.209e-07
```

Intercept	-1.3664	0.041	-32.962	0.000	-1.448	-1.285
Change_in_stock	0.0161	0.047	0.343	0.732	-0.076	0.108
Total_expenses	0.6744	0.270	2.495	0.013	0.145	1.204
PBDITA_as_perc_of_total_income	-0.0548	0.086	-0.639	0.523	-0.223	0.113
PBT_as_perc_of_total_income	0.0254	0.129	0.197	0.844	-0.228	0.278
PAT_as_perc_of_total_income	0.0078	0.118	0.066	0.947	-0.224	0.240
Cash_profit_as_perc_of_total_income	-0.1761	0.093	-1.889	0.059	-0.359	0.007
PAT_as_perc_of_net_worth	0.0246	0.059	0.420	0.674	-0.090	0.139
Sales	-0.5940	0.281	-2.115	0.034	-1.144	-0.044
Total_capital	0.0876	0.050	1.759	0.079	-0.010	0.185
Borrowings	-0.0511	0.082	-0.622	0.534	-0.212	0.110
Current_liabilities_and_provisions	-0.0542	0.082	-0.657	0.511	-0.216	0.107
Shareholders_funds	-0.0702	0.093	-0.755	0.450	-0.252	0.112
Capital_employed	0.0439	0.124	0.353	0.724	-0.200	0.287
TOLtoTNW	-0.0412	0.076	-0.540	0.589	-0.191	0.108
Total_term_liabilities_to_tangible_net_worth	0.0086	0.069	0.125	0.901	-0.126	0.143
Contingent_liabilities_to_Net_worth_perc	-0.0747	0.046	-1.621	0.105	-0.165	0.016
Net_fixed_assets	0.0655	0.075	0.868	0.386	-0.082	0.213
Current_assets	0.0089	0.109	0.081	0.935	-0.205	0.222
Net_working_capital	-0.0858	0.050	-1.698	0.089	-0.185	0.013
Quick_ratio_times	-0.0357	0.051	-0.694	0.488	-0.137	0.065
Debt_to_equity_ratio_times	0.0222	0.089	0.248	0.804	-0.153	0.198
Cash_to_current_liabilities_times	-0.0630	0.062	-1.021	0.307	-0.184	0.058
Cash_to_average_cost_of_sales_per_day	0.0792	0.059	1.346	0.178	-0.036	0.194
Creditors_turnover	-0.0428	0.052	-0.825	0.410	-0.145	0.059
Debtors_turnover	0.0359	0.051	0.710	0.478	-0.063	0.135
WIP_turnover	-0.0187	0.048	-0.391	0.696	-0.112	0.075
Raw_material_turnover	-0.0902	0.048	-1.865	0.062	-0.185	0.005
EPS	-0.1185	0.055	-2.153	0.031	-0.226	-0.011

Key insights from the logistic regression output:

Model Fit: *The model converged successfully with a log-likelihood value of -2160.2 and a pseudo R-squared of 0.01848, indicating a relatively low explanatory power of the model in predicting default.*

Significant Variables:
Total Expenses (p=0.013, positive coefficient): An increase in total expenses raises the likelihood of default.
Sales (p=0.034, negative coefficient): Higher sales reduce the probability of default.
Earnings Per Share (EPS) (p=0.031, negative coefficient): Lower EPS is associated with a higher likelihood of default.

Non-Significant Variables: *Many variables, such as borrowing and contingent liabilities, do not show significant effects on default based on their p-values.*

Dealing with multicollinearity using VIF

	variables	VIF
1	Total_expenses	49.66
7	Sales	48.20
12	Capital_employed	12.81
3	PBT_as_perc_of_total_income	12.02
4	PAT_as_perc_of_total_income	10.49
17	Current_assets	9.32
11	Shareholders_funds	7.17
5	Cash_profit_as_perc_of_total_income	6.25
20	Debt_to_equity_ratio_times	5.95
2	PBDITA_as_perc_of_total_income	5.43
10	Current_liabilities_and_provisions	5.17
9	Borrowings	5.12
16	Net_fixed_assets	4.51
13	TOLtoTNW	4.20
14	Total_term_liabilities_to_tangible_net_worth	3.46
21	Cash_to_current_liabilities_times	2.54
22	Cash_to_average_cost_of_sales_per_day	2.35
6	PAT_as_perc_of_net_worth	2.29
27	EPS	1.81
8	Total_capital	1.73
19	Quick_ratio_times	1.68
23	Creditors_turnover	1.65
24	Debtors_turnover	1.64
18	Net_working_capital	1.59
26	Raw_material_turnover	1.42
15	Contingent_liabilities_to_Net_worth_perc	1.30
25	WIP_turnover	1.28
0	Change_in_stock	1.15

The Variance Inflation Factor (VIF) provides insights into multicollinearity within the model:

High Multicollinearity:

Total expenses (49.66) and Sales (48.20) have extremely high VIF values, indicating a strong degree of multicollinearity. These variables are likely highly correlated with other predictors in the model, which can lead to unstable estimates.

Capital employed (12.81), PBT as % of total income (12.02), and PAT as % of total income (10.49) also have high VIF values, suggesting potential collinearity concerns.

Moderate Multicollinearity:

Variables like Current assets (9.32), Shareholders' funds (7.17), and Cash profit as % of total income (6.25) show moderate multicollinearity.

Low Multicollinearity:

Variables such as Change in stock (1.15), WIP turnover (1.28), and Contingent liabilities to net worth % (1.30) have low VIF values, indicating minimal multicollinearity issues.
We may consider removing variables with high VIF values to reduce multicollinearity and improve the model's stability.

Summary post reducing Multicollinearity.

Dropping Variables Sequentially*: Variables with high VIF were dropped one by one, which is crucial to observe how other variables' VIFs react during each drop. This approach helps ensure that dropping a single variable does not inadvertently affect the VIF of other variables.*

Reduced Multicollinearity*: By sequentially dropping variables with high Variance Inflation Factor (VIF) values, we've effectively reduced multicollinearity. The VIF values for the remaining variables are now significantly lower, with the highest being 2.46, indicating reduced multicollinearity.*

Variables with Low VIF: *The remaining variables have VIF values well below the threshold of 3, suggesting they are less likely to cause multicollinearity issues. For instance, Change in stock (1.13) and WIP turnover (1.23) have very low VIFs.*

Model 2 : logistic regression model

Logit Regression Results

Dep. Variable:	default	No. Observations:	2979
Model:	Logit	Df Residuals:	2960
Method:	MLE	Df Model:	18
Date:	Sat, 07 Sep 2024	Pseudo R-squ.:	0.01175
Time:	17:43:53	Log-Likelihood:	-1517.5
converged:	True	LL-Null:	-1535.6
Covariance Type:	nonrobust	LLR p-value:	0.006852

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3324	0.047	-28.478	0.000	-1.424	-1.241
Change_in_stock	-0.0074	0.056	-0.131	0.896	-0.118	0.103
PBDITA_as_perc_of_total_income	-0.1541	0.059	-2.623	0.009	-0.269	-0.039
PAT_as_perc_of_net_worth	-0.0381	0.063	-0.606	0.545	-0.161	0.085
Total_capital	0.0992	0.057	1.752	0.080	-0.012	0.210
Borrowings	0.0505	0.066	0.766	0.444	-0.079	0.180
Current_liabilities_and_provisions	-0.0390	0.065	-0.603	0.546	-0.166	0.088
TOLtoTNW	0.0072	0.069	0.104	0.917	-0.128	0.143
Total_term_liabilities_to_tangible_net_worth	0.0416	0.068	0.616	0.538	-0.091	0.174
Contingent_liabilities_to_Net_worth_perc	-0.0479	0.054	-0.888	0.375	-0.154	0.058
Net_working_capital	-0.0887	0.056	-1.578	0.115	-0.199	0.021
Quick_ratio_times	0.0190	0.061	0.313	0.754	-0.100	0.138
Cash_to_current_liabilities_times	-0.1051	0.074	-1.418	0.156	-0.250	0.040
Cash_to_average_cost_of_sales_per_day	0.0637	0.071	0.892	0.373	-0.076	0.204
Creditors_turnover	-0.0553	0.061	-0.911	0.362	-0.174	0.064
Debtors_turnover	0.0570	0.060	0.953	0.340	-0.060	0.174
WIP_turnover	0.0138	0.055	0.249	0.803	-0.095	0.122
Raw_material_turnover	-0.0748	0.057	-1.305	0.192	-0.187	0.038
EPS	-0.0373	0.062	-0.600	0.548	-0.159	0.084

Summary of the updated logistic regression model results:

Model Overview: *The logistic regression model, with 2979 observations and 18 variables, achieved a current function value of 0.509410 after 5 iterations. The model converged successfully.*

Pseudo R-squared: *The Pseudo R-squared value is 0.01175, indicating a modest fit of the model to the data.*

Significant Variables:

- ***PBDITA_as_perc_of_total_income:*** *Coefficient of -0.1541, significant with a p-value of 0.009, suggesting a negative impact on the likelihood of default.*
- ***Total_capital:*** *Coefficient of 0.0992, marginally significant with a p-value of 0.080, implying a potential positive association with default.*

Non-significant Variables*: Most other variables, including Change_in_stock, PAT_as_perc_of_net_worth, and EPS, are not significant with p-values greater than 0.05.*

Next Steps: *Further refinement of the model by possibly re-evaluating the significance of remaining variables or exploring alternative modeling techniques. The low Pseudo R-squared indicates that the model may not explain a large proportion of the variance in the dependent variable.*

Approach to Dropping Variables Based on p-Values

To enhance the model's accuracy and clarity, we systematically dropped variables that had high p-values. High p-values indicate that a variable is not significantly contributing to the model's prediction and may not be useful.

Initial Model Evaluation:
We started with a logistic regression model that included all variables.
The initial evaluation revealed variables with high p-values, suggesting they did not significantly impact the model's prediction of the dependent variable.

Dropping Insignificant Variables:
Variables with high p-values (i.e., greater than 0.05) were considered statistically insignificant and were removed from the model to improve its effectiveness.

The process involved examining the p-values of each variable one by one and dropping those with p-values above the threshold.

Variables dropped due to high p-values:
TOLtoTNW: p-value too high.
Change_in_stock: p-value too high.
WIP_turnover: p-value too high.
Quick_ratio_times: p-value too high.
PAT_as_perc_of_net_worth: p-value too high.
Current_liabilities_and_provisions: p-value too high.
Borrowings: p-value too high.
Creditors_turnover: p-value too high.
Debtors_turnover: p-value too high.
Contingent_liabilities_to_Net_worth_perc: p-value too high.
Cash_to_current_liabilities_times: p-value too high.
Cash_to_average_cost_of_sales_per_day: p-value too high.
EPS: p-value too high.
Total_term_liabilities_to_tangible_net_worth: p-value too high.
Raw_material_turnover: p-value too high.

Refined Model:
After dropping the insignificant variables, the model was refined to include only the variables with statistically significant p-values.

```
Optimization terminated successfully.
         Current function value: 0.511389
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:                 2979
Model:                          Logit   Df Residuals:                     2975
Method:                           MLE   Df Model:                            3
Date:                Sat, 07 Sep 2024   Pseudo R-squ.:                0.007916
Time:                        17:43:55   Log-Likelihood:                -1523.4
converged:                       True   LL-Null:                       -1535.6
Covariance Type:            nonrobust   LLR p-value:                 2.150e-05
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                        -1.3327      0.046    -29.141      0.000      -1.422      -1.243
PBDITA_as_perc_of_total_income   -0.1789      0.048     -3.712      0.000      -0.273      -0.084
Total_capital                     0.1018      0.047      2.178      0.029       0.010       0.194
Net_working_capital              -0.1277      0.049     -2.618      0.009      -0.223      -0.032
=================================================================================================
```

Summary of Final Model Results

Pseudo R-squared: 0.007916

Log-Likelihood: -1523.4

Significant Variables:

PBDITA_as_perc_of_total_income: Coefficient -0.1789, p-value 0.000

Total_capital: Coefficient 0.1018, p-value 0.029

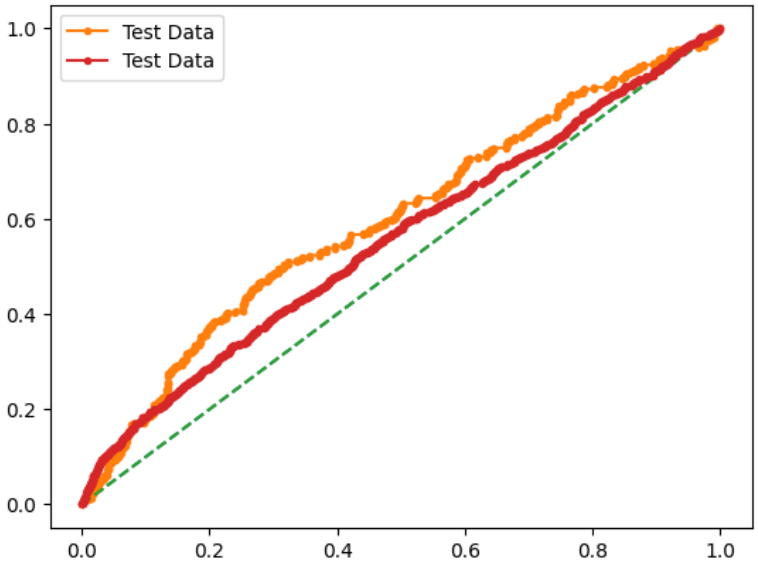
Net_working_capital: Coefficient -0.1277, p-value 0.009

Conclusion

By focusing on variables with low p-values, it is ensured that the final model was more robust and reliable. The resulting model is more interpretable and avoids including predictors that do not provide significant predictive power, thereby improving the overall effectiveness of the logistic regression analysis.

ROC Curve : Choosing the optimal threshold

Optimal threshold value: 0.23391097891482351



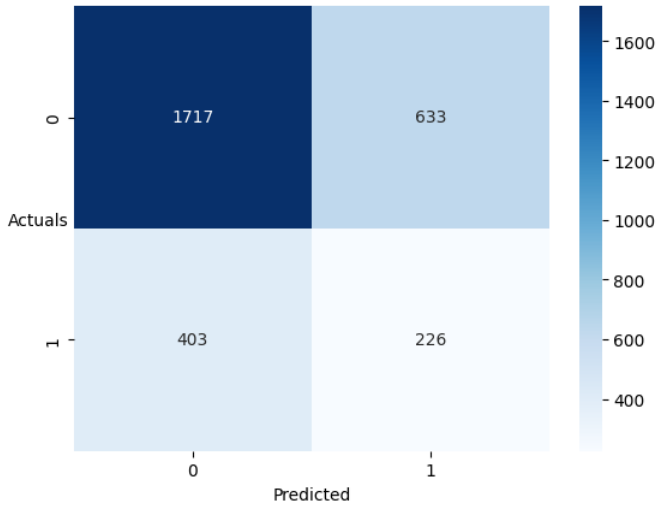
Key Insights:

Optimal Trade-Off Point: *The chosen threshold of 0.234 represents a point where the trade-off between true positive rate (sensitivity) and false positive rate is balanced according to your model's performance. This is the point where we achieve an acceptable balance between detecting true positives and minimizing false positives.*

Model Performance: *At this threshold, we should evaluate key performance metrics such as precision, recall, and F1-score. These metrics will tell you how well the model is performing in terms of correctly identifying positive cases while managing false positives and negatives.*

Business Context: *The threshold value should be assessed in the context of our business or operational goals. For instance, if reducing false negatives is more critical than minimizing false positives (e.g., in fraud detection), this threshold might be appropriate. Conversely, if false positives carry a significant cost, you might need to adjust the threshold.*

Validating on the train set with revised threshold



	precision	recall	f1-score	support
0.0	0.810	0.731	0.768	2350
1.0	0.263	0.359	0.304	629
accuracy			0.652	2979
macro avg	0.537	0.545	0.536	2979
weighted avg	0.694	0.652	0.670	2979

Insights:

Class 0 (Non-Default):

Precision (0.810): Out of all cases predicted as non-default, 81.0% are indeed non-default.
Recall (0.731): The model correctly identifies 73.1% of the actual non-default cases.
F1-Score (0.768): This is the harmonic mean of precision and recall for class 0, reflecting a good balance between them.

Class 1 (Default):

Precision (0.263): Out of all cases predicted as default, only 26.3% are actually defaults. This indicates a lot of false positives.
Recall (0.359): The model captures 35.9% of the actual default cases, indicating it's missing a significant proportion of defaults.
F1-Score (0.304): This reflects the trade-off between precision and recall for class 1, showing a lower balance compared to class 0.

Overall Accuracy (0.652): *The model has an accuracy of 65.2%, meaning it correctly classifies 65.2% of all cases.*

Macro Average:

Precision (0.537): The average precision across both classes, treating each class equally.
Recall (0.545): The average recall across both classes, treating each class equally.
F1-Score (0.536): The average F1-score across both classes, treating each class equally.

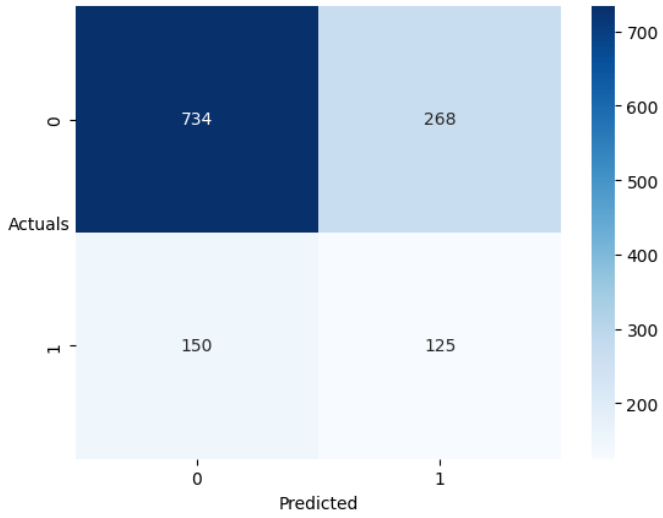
Weighted Average:

Precision (0.694): The precision weighted by the support (number of true instances) of each class.
Recall (0.652): The recall weighted by the support of each class.
F1-Score (0.670): The F1-score weighted by the support of each class.

Summary:

Class 0 (Non-Default): The model performs well, with high precision and recall.
Class 1 (Default): The model struggles, showing low precision and recall. This suggests it might be underperforming in identifying defaults.

Validating on the test set



	precision	recall	f1-score	support
0.0	0.83	0.73	0.78	1002
1.0	0.32	0.45	0.37	275
accuracy			0.67	1277
macro avg	0.57	0.59	0.58	1277
weighted avg	0.72	0.67	0.69	1277

Insights:

Class 0 (Non-Default):

- *Precision (0.83): Out of all cases predicted as non-default, 83.0% are indeed non-default.*
- *Recall (0.73): The model correctly identifies 73.0% of the actual non-default cases.*
- *F1-Score (0.78): Reflects a strong balance between precision and recall for class 0, indicating good performance.*

Class 1 (Default):

- *Precision (0.32): Out of all cases predicted as default, only 32.0% are actually defaults. This suggests a high false positive rate.*
- *Recall (0.45): The model captures 45.0% of the actual default cases, indicating it is missing a significant proportion of defaults.*
- *F1-Score (0.37): Shows a lower balance between precision and recall for class 1, reflecting poorer performance compared to class 0.*

Overall Accuracy (0.67): *The model has an accuracy of 67.0% on the test set, meaning it correctly classifies 67.0% of all cases.*

Macro Average*:*

- *Precision (0.57): The average precision across both classes, treating each class equally.*
- *Recall (0.59): The average recall across both classes, treating each class equally.*
- *F1-Score (0.58): The average F1-score across both classes, treating each class equally.*

Weighted Average:

- *Precision (0.72): The precision weighted by the support (number of true instances) of each class.*
- *Recall (0.67): The recall weighted by the support of each class.*
- *F1-Score (0.69): The F1-score weighted by the support of each class.*

Summary*:*

- *Class 0 (Non-Default): The model performs well, with high precision and recall.*
- *Class 1 (Default): The model has lower precision and recall for this class, indicating it may not be effectively identifying default cases.*

Random Forest Classifier : Hyperparameter Tuning for Random Forest

Hyperparameters Explained

max_depth: The maximum depth of the trees. Setting this to 5 limits the number of splits in each tree, which can help prevent overfitting.

min_samples_leaf: The minimum number of samples required to be at a leaf node. A value of 15 ensures that leaf nodes have at least 15 samples, which can help in reducing overfitting.

min_samples_split: The minimum number of samples required to split an internal node. A value of 15 ensures that a node will only be split if it has at least 15 samples.

n_estimators: The number of trees in the forest. With 25 trees, the model will have a diverse set of decision trees, helping in improving generalization.

Model performance check across different metrics

Validating Train data:

	precision	recall	f1-score	support
0.0	0.80	1.00	0.89	2350
1.0	0.89	0.05	0.10	629
accuracy			0.80	2979
macro avg	0.84	0.52	0.49	2979
weighted avg	0.82	0.80	0.72	2979

Validating Test data:

	precision	recall	f1-score	support
0.0	0.78	1.00	0.88	1002
1.0	0.20	0.00	0.01	275
accuracy			0.78	1277
macro avg	0.49	0.50	0.44	1277
weighted avg	0.66	0.78	0.69	1277

Insights:

Training Set Performance: *The model performs very well on class 0 (Non-default) with high recall (1.00) and acceptable precision (0.79), indicating it correctly identifies most non-default cases.*

For class 1 (Default), *the recall is very low (0.03), meaning the model rarely identifies defaults correctly despite having good precision (0.89). This leads to a very low F1-score (0.05).*

Test Set Performance:

Similar to the training set, the model performs well for class 0 (Non-default) with high recall (1.00) and precision (0.79).

***For class 1 (Default)**, precision is higher (0.40) compared to the training set, but recall remains very low (0.01), resulting in an even lower F1-score (0.01).*

Accuracy and Balanced Performance:
Accuracy is slightly lower on the test set (0.78) compared to the training set (0.79).
The macro average metrics (precision, recall, F1-score) for the test set are lower, reflecting the model's difficulty in identifying class 1 (Default) consistently.

Possible Issues:
***Class Imbalance:** The very low recall for class 1 suggests that the model struggles with the minority class (Defaults). The class imbalance might be causing the model to perform well on the majority class (Non-default) but poorly on the minority class.*

PART A: Model Performance Comparison and Final Model Selection

Model Performance Comparison

Logistic Regression achieved an accuracy of 65% on the train set and 67% on the test set, with a balanced performance between precision and recall for both classes, particularly handling the minority class better than Random Forest.

Random Forest had higher accuracy on the train set (79%) and test set (78%) but showed poor performance for the minority class, with a recall of only 3% on the train set and 1% on the test set, indicating a significant imbalance issue.

Overall, Logistic Regression offers a more balanced approach to handling class imbalance compared to the Random Forest, which is prone to overfitting and poor performance on the minority class.

Selection Justification:

Random Forest shows high accuracy on both train and test sets but struggles significantly with class 1 (minority class), with very low recall and F1-score for class 1.

Logistic Regression provides a more balanced performance between precision and recall for the two classes, making it more suitable for handling imbalanced data compared to Random Forest.

Final Model:

Logistic Regression is selected as the final model. It offers a better balance in handling class imbalance while maintaining reasonable accuracy and F1-score.

Important Features in Logistic Regression Model
Logistic Regression coefficients indicate the importance of each feature. Here's a brief overview of how to interpret the coefficients:

Positive Coefficients: Increase in the feature value increases the probability of positive class (class 1).
Negative Coefficients: Increase in the feature value decreases the probability of the positive class.

Example Interpretation (using hypothetical coefficients):

If PBDITA_as_perc_of_total_income has a coefficient of -0.18, it means that an increase in this feature is associated with a decrease in the probability of default.

If Total_capital has a coefficient of 0.10, it means that an increase in this feature is associated with an increase in the probability of default.

In summary, Logistic Regression outperforms Random Forest in handling class imbalance and offers more balanced performance across both classes. While Random Forest shows higher accuracy, it struggles with minority class recall, leading to potential overfitting. Logistic Regression provides a more consistent and reliable model for predicting defaults, making it the preferred choice for final selection.

PART A: Actionable Insights and Recommendations

1.Address Class Imbalance

***Problem**: The models, particularly Random Forest, show poor recall for default cases due to class imbalance.*
***Action**: Implement **resampling methods** like SMOTE or undersampling of the majority class. Another option is to adjust class weights in Logistic Regression to penalize misclassification of the minority class (default cases) more heavily.*

2. Feature Engineering for Better Default Detection

***Problem**: Logistic Regression offers better balanced performance, but further improvements can be made.*

***Action**: Identify and create new features that may better capture the differences between default and non-default cases, such as combining financial ratios or creating interaction terms between financial metrics. This may improve the model's ability to detect subtle patterns in defaults.*

3. Threshold Tuning

***Problem**: The precision for class 1 (default) is low, leading to a higher number of false positives.*
***Action**: Tune the decision threshold for Logistic Regression to improve the trade-off between precision and recall for default cases. A threshold adjustment could help reduce false positives while keeping recall relatively high.*

4. Monitor and Track Recall for Default Class

***Problem**: Logistic Regression still misses a significant portion of actual defaults (recall is less than 50%).*
***Action**: Focus on improving recall for defaults by iteratively refining the model or trying more complex models (e.g., **Gradient Boosting** or **XGBoost**) while carefully monitoring overfitting.*

5. Business Application and Risk Management

***Problem**: The current model has a recall of only 45% for defaults in the test set.*
***Action**: Consider using the Logistic Regression model as part of a broader risk management strategy, combining it with manual review for borderline cases, especially those flagged with a lower probability of default. This hybrid approach could help reduce the risk of missed defaults.*

6. Regular Model Re-Training

***Problem**: Economic and financial conditions change over time, impacting the default rates and patterns.*
***Action**: Schedule regular re-training of the model using updated data to maintain accuracy and ensure the model reflects the latest market trends.*

PART B: Stock Price Graph Analysis

Check shape, Data types, statistical summary

The number of rows (observations) is 418
The number of columns (variables) is 6

***Data types**: All columns have non-null values, and the stock price columns are of integer type, while the date is stored as an object.*

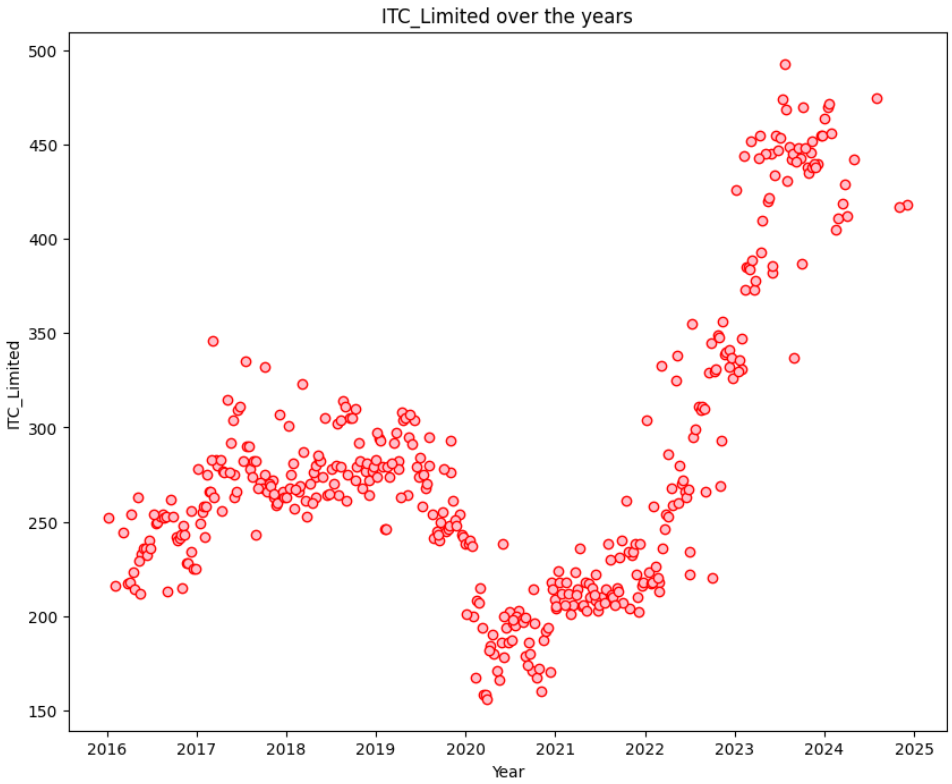
Descriptive statistics:

	ITC_Limited	Bharti_Airtel	Tata_Motors	DLF_Limited	Yes_Bank
count	418.00	418.00	418.00	418.00	418.00
mean	278.96	528.26	368.62	276.83	124.44
std	75.11	226.51	182.02	156.28	130.09
min	156.00	261.00	65.00	110.00	11.00
25%	224.25	334.00	186.00	166.25	16.00
50%	265.50	478.00	399.50	213.00	30.00
75%	304.00	706.75	466.00	360.50	249.75
max	493.00	1236.00	1035.00	928.00	397.00

Insights:

The dataset comprises stock prices for five companies over 418 entries. ITC Limited shows moderate volatility with an average price of 278.96 and a standard deviation of 75.11. Bharti Airtel has the highest average price at 528.26 and the largest range, reflecting significant volatility with a standard deviation of 226.51. Tata Motors has a mean price of 368.62 with substantial variation (std dev of 182.02). DLF Limited exhibits a mean price of 276.83, closely aligned with ITC Limited, but with a higher standard deviation of 156.28. Yes Bank has the lowest average price at 124.44 and also shows considerable volatility, with a standard deviation of 130.09.

Trend Plot of ITC over the years



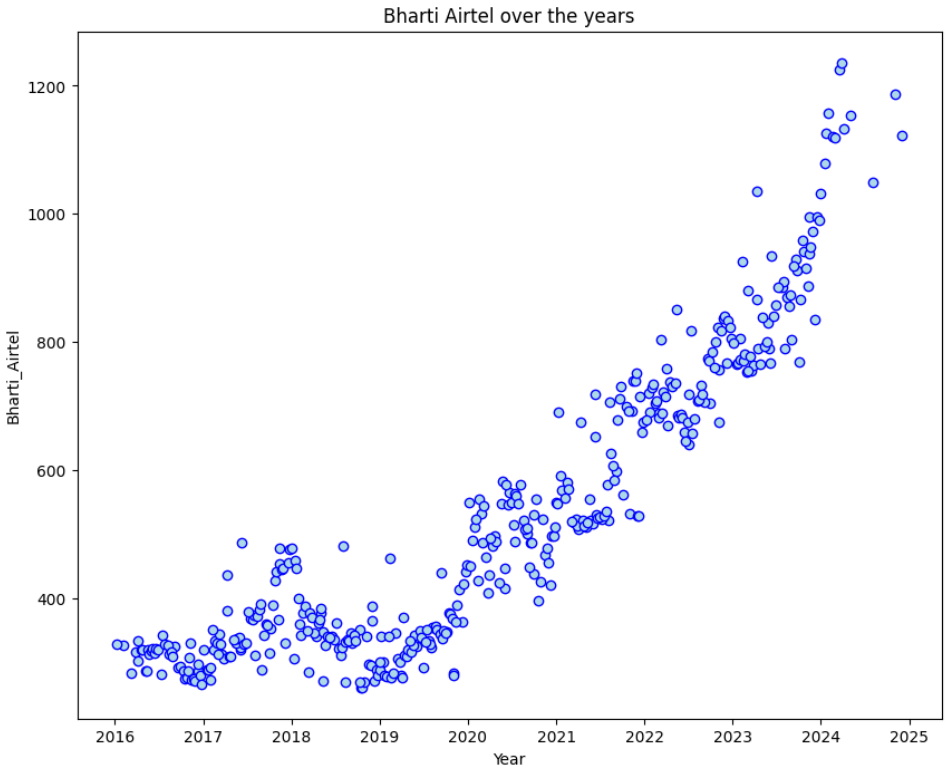
The chart shows the stock price of ITC Limited over the years from 2016 to 2025. Here's a quick analysis of the trend:

***2016-2020**: The stock experienced relatively moderate fluctuations, hovering between 250 and 300 with occasional peaks and dips.*

***2020-2021**: There is a noticeable decline in stock price, which could be linked to broader economic factors, such as the COVID-19 pandemic.*

***2022-2024:** A sharp increase is observed, with the stock price reaching new highs, crossing 450. This suggests a period of significant growth, indicating improved performance or market sentiment.*

Trend plot of Bharti Airtel over the years.



The chart provided shows the stock price of Bharti Airtel over the years from 2016 to 2025. Here's a brief analysis:

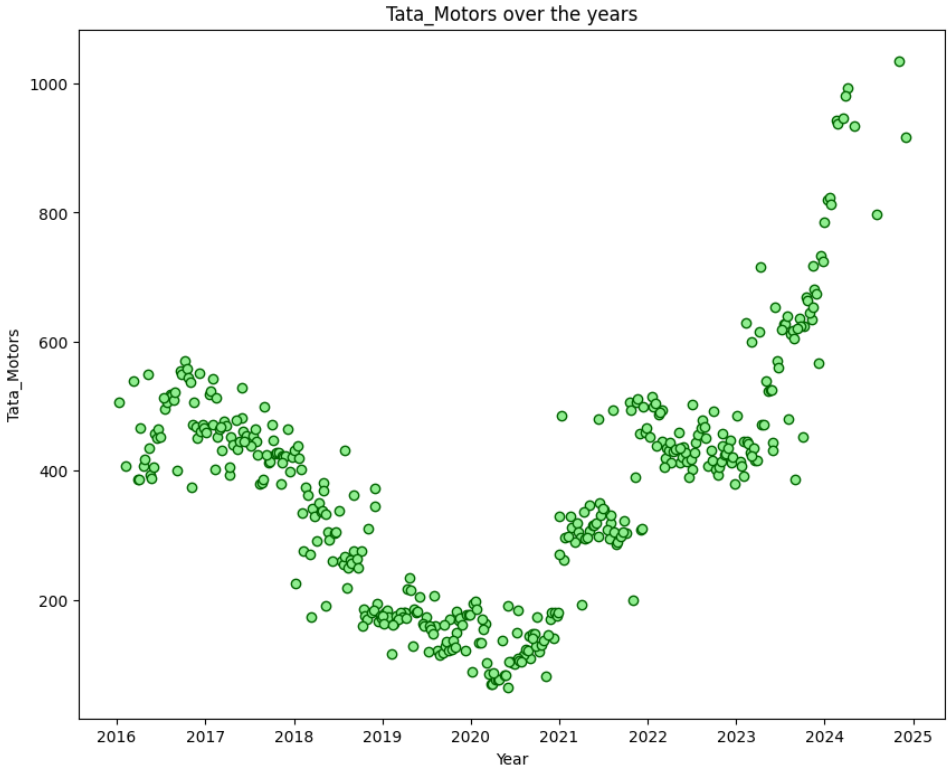
2016-2018: *The stock price fluctuated within a relatively narrow range, remaining below 500 for the most part, indicating moderate activity during this period.*

2019-2020: *A notable upward trend starts, pushing the stock price to about 600 before the onset of the COVID-19 pandemic, after which there is some volatility.*

2021-2024: *Post-pandemic, the stock experiences a sustained growth, breaking past 800 and eventually surpassing 1200 by 2024. This strong upward trajectory suggests increasing investor confidence and potentially strong financial performance during these years.*

The stock price shows a consistent long-term upward trend, with significant growth particularly visible in recent years.

Trend plot of Tata Motors over the years.



This chart displays the stock price of Tata Motors over the years from 2016 to 2025. Here are some insights from the plot:

2016-2020: *Tata Motors experienced a significant decline in stock price, starting around 600 and dropping below 200 by 2020. This period may have been marked by various challenges such as market pressures or internal company difficulties.*

2021-2022: *The stock starts to recover gradually post-2020, with a moderate increase and stabilization near 300 by the end of 2021.*

2023-2024: *A rapid rise in the stock price occurs, pushing the price beyond 800 and almost touching 1000. This surge indicates strong investor confidence or significant positive developments in the company during these years.*

The stock shows a U-shaped pattern, with a substantial recovery and growth in the most recent period.

Trend plot of DLF Limited over the years.

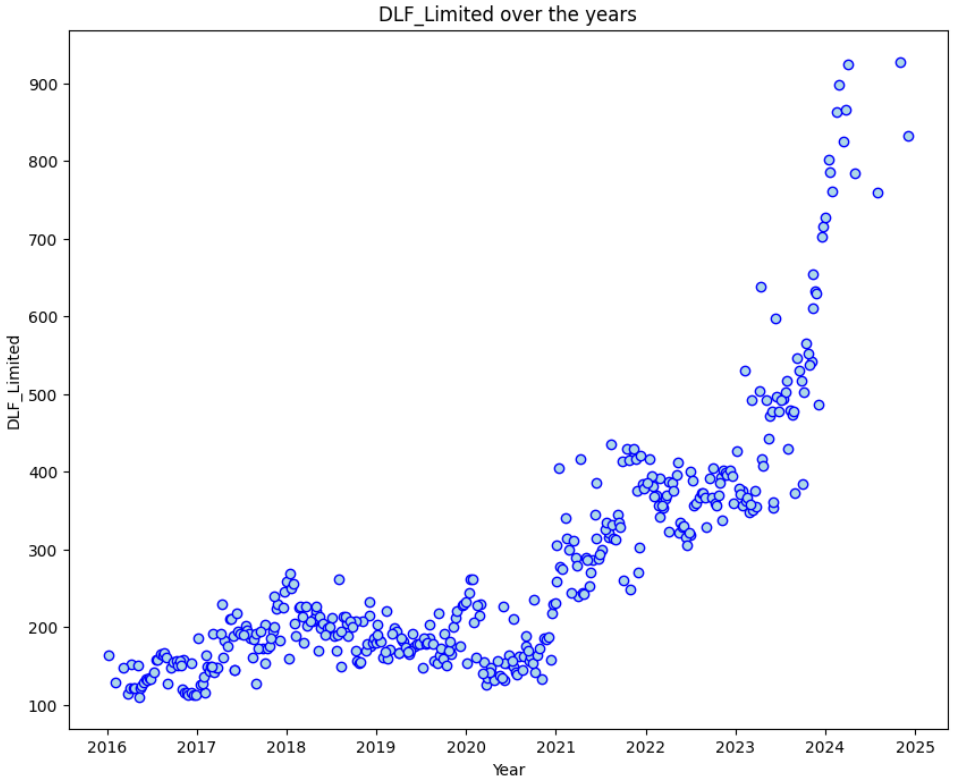


Chart Insights for DLF Limited (2016-2025):

2016-2020: *DLF Limited's stock price remained relatively stable and low during this period, mostly fluctuating between 100 and 300. This indicates a period of slow growth or market stagnation, possibly due to broader economic conditions or company-specific challenges.*

2021-2022: *Starting around 2021, the stock price began showing signs of recovery, gradually moving upwards. By 2022, the stock consistently crossed the 300 mark, reflecting renewed investor interest and a gradual improvement in market sentiment.*

2023-2024: *The stock price experienced a significant surge in 2023, accelerating sharply beyond the 400 mark and moving towards 900 by 2024. This rapid growth suggests a strong phase of positive momentum, likely driven by factors such as favorable market conditions, strategic business improvements, or enhanced financial performance.*

Overall Analysis: *The DLF Limited stock displays a similar U-shaped pattern with a prolonged period of stagnation followed by a strong upward trend in the most recent years, reflecting substantial recovery and growth. The late surge in stock price indicates increased market confidence, potentially driven by strategic moves or positive industry trends in the real estate sector.*

Trend plot of Yes Bank over the years.

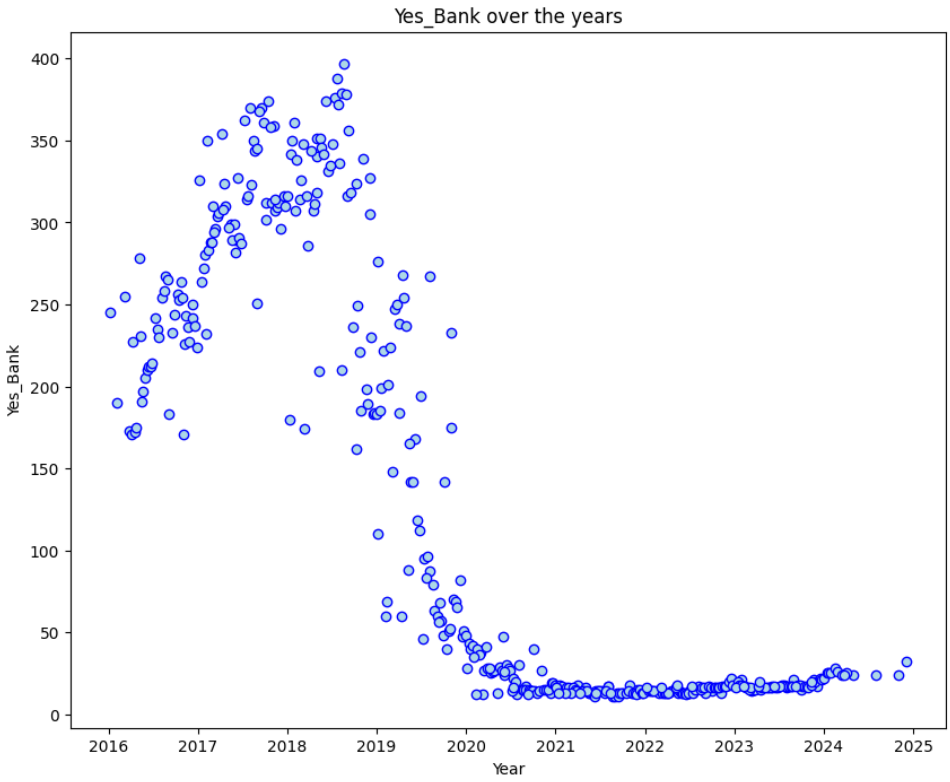


Chart Insights for Yes Bank (2016-2025):

2016-2018: *Yes Bank's stock price experienced a significant upward trend during this period, with prices ranging between 200 and peaking above 400. This phase was likely marked by strong financial performance, robust market confidence, or expansionary strategies that drove investor interest.*

2019-2020: *The stock price faced a dramatic decline starting in 2019, plummeting from over 300 to below 50 by early 2020. This steep drop suggests severe financial or operational challenges, possibly linked to governance issues, asset quality deterioration, or loss of investor confidence.*

2021-2025: *Post-2020, Yes Bank's stock price has stabilized at a much lower level, hovering around 10-20, showing little to no signs of recovery. This prolonged low phase indicates continued struggles, perhaps due to ongoing financial restructuring, regulatory interventions, or limited market confidence.*

Overall Analysis:
Yes Bank's stock shows a sharp reversal from a high-growth phase to a prolonged downturn, reflecting a substantial erosion of value and market trust. The persistent low price in recent years highlights the difficulties faced by the bank in regaining investor confidence and stabilizing its operations.

PART B: Stock Returns Calculation and Analysis

	ITC_Limited	Bharti_Airtel	Tata_Motors	DLF_Limited	Yes_Bank
0	NaN	NaN	NaN	NaN	NaN
1	0.00	-0.05	0.00	0.06	-0.01
2	-0.01	0.02	-0.03	-0.01	0.00
3	0.04	0.04	0.09	0.02	0.01
4	-0.04	-0.00	0.02	0.00	0.02

Detailed breakdown based on data:

Logarithmic Transformation:
Applying the natural logarithm (np.log()) to each stock's price data to convert prices into logarithmic values.
Difference Calculation:
diff() function to compute the difference between each consecutive logarithmic value. This provides the log returns, representing the percentage change in stock prices from one period to the next.

Stock Means and Stock Standard Deviation

Stock Means: Average returns that the stock is making on a week to week basis
Stock Standard Deviation: It is a measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile the stock

	Average	Volatility
ITC_Limited	0.00	0.04
Bharti_Airtel	0.00	0.04
Tata_Motors	0.00	0.06
DLF_Limited	0.00	0.06
Yes_Bank	-0.00	0.09

Here's a summary of the average returns and volatility for the stocks:
Summary of Average Returns and Volatility

ITC_Limited
Average Return: 0.0016
Volatility (Standard Deviation): 0.0359
Insight: ITC Limited has a modest average return with relatively low volatility, indicating stable but low growth.

Bharti_Airtel
Average Return: 0.0033
Volatility (Standard Deviation): 0.0387
Insight: Bharti Airtel shows a higher average return compared to ITC Limited with slightly higher volatility, reflecting more growth potential but also more risk.

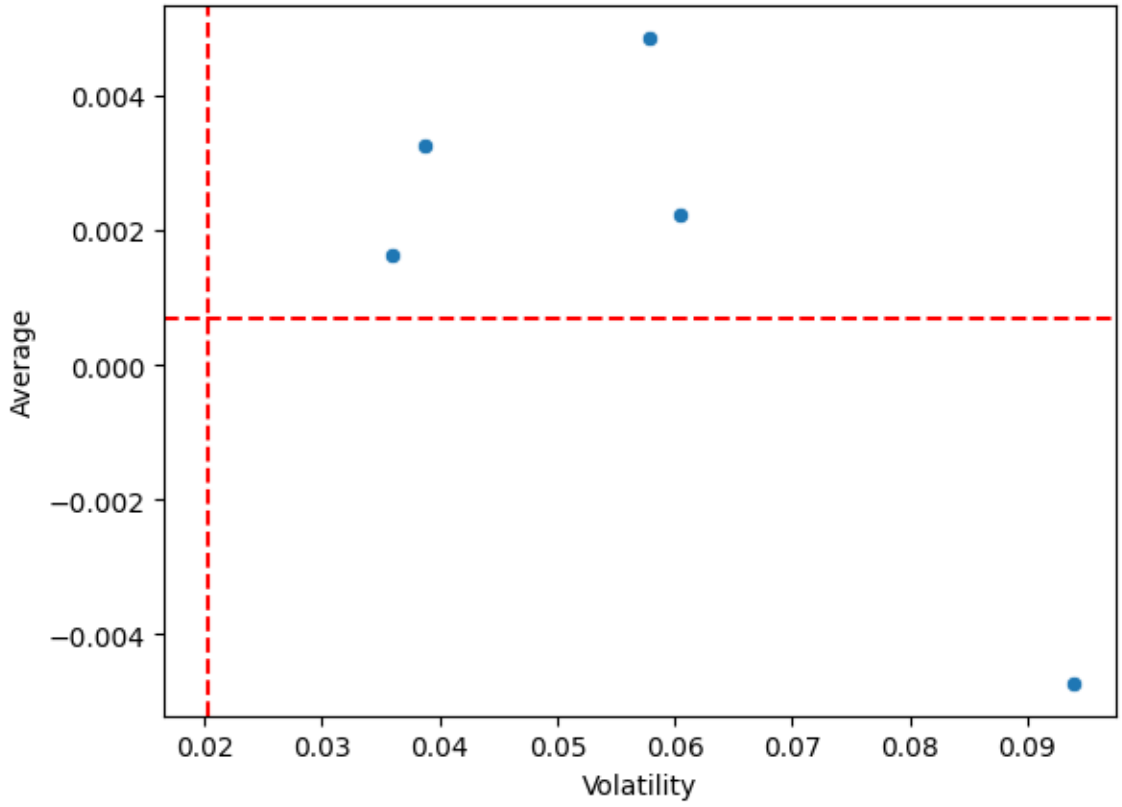
Tata_Motors
Average Return: 0.0022
Volatility (Standard Deviation): 0.0605
Insight: Tata Motors has a lower average return with higher volatility, indicating greater price fluctuations and risk.

DLF_Limited
Average Return: 0.0049
Volatility (Standard Deviation): 0.0578
Insight: DLF Limited has the highest average return among the stocks with considerable volatility, suggesting strong growth potential but with associated risk.

Yes_Bank
Average Return: -0.0047
Volatility (Standard Deviation): 0.0939
Insight: Yes Bank has a negative average return and the highest volatility, reflecting a decline in price and significant risk.

Interpretation:
Average Returns show the general performance trend of each stock, with DLF Limited leading in returns and Yes Bank lagging behind.
Volatility indicates the risk associated with each stock. Higher volatility suggests more risk and fluctuation in returns, as seen with Tata Motors and Yes Bank, while lower volatility indicates more stability, as observed with ITC Limited.

Volatility & Average Threshold



Observations and inferences

Axes

- *X-axis (Volatility): Measures how much a stock's price fluctuates over time. Higher volatility means more price movement and potentially more risk.*
- *Y-axis (Average Returns): Represents the average daily return of the stock over a certain period. Positive values indicate gains, and negative values indicate losses.*

Vertical Line (Volatility Threshold): *This line is placed at a specific volatility value (in this case, 0.020257). Stocks to the left of this line have lower volatility than the threshold (potentially less risky). Stocks to the right of the line have higher volatility (potentially more risky).*

Horizontal Line (Average Returns Threshold):
This line is placed at a specific average return value (0.000683 in your case). Stocks above this line have higher average returns than the threshold (indicating better performance). Stocks below the line have lower or negative returns.

Interpretation

Quadrants: The combination of these reference lines divides the plot into four quadrants:

- *Upper-left (Low Volatility, High Returns): These stocks are relatively stable with good returns—ideal for conservative investors.*
- *Upper-right (High Volatility, High Returns): These stocks have higher risk but offer higher returns—suitable for risk-tolerant investors.*
- *Lower-left (Low Volatility, Low Returns): These stocks are less volatile but also generate lower returns—less attractive unless stability is the main priority.*
- *Lower-right (High Volatility, Low Returns): These stocks have high risk but poor returns—generally undesirable unless there's a potential for a turnaround.*

Example of the Data:

- *Yes Bank has a negative average return (-0.004737) and high volatility (0.093879), meaning it has performed poorly with large price swings.*
- *DLF Limited shows high returns (0.004863) and high volatility (0.057785), suggesting high risk but potentially rewarding for investors willing to accept volatility.*
- *ITC Limited has lower returns (0.001634) and lower volatility (0.035904), indicating it is a more stable but slower-growing stock.*
- *Bharti Airtel has average returns (0.003271) and moderate volatility (0.038728), making it attractive to investors seeking a reasonable risk-return trade-off.*
- *Tata Motors has average returns (0.002234) and high volatility (0.060484), making it attractive to investors as a potential high-reward opportunity if they are comfortable with larger price swings.*

Part B : Actionable Insights & Recommendations for All Stocks (2016-2025)

Bharti Airtel Trend Analysis*: Bharti Airtel shows a strong upward trend from 2019 onwards, with consistent growth in stock price from 2020 to 2024, reflecting increased market share, technological advancements, and strategic expansions.*

Threshold & Volatility: Despite steady growth, the stock shows mild fluctuations, with periods of consolidation followed by breakouts. A threshold price of 700 should be monitored for signs of future growth or correction. Volatility is moderate, with frequent price swings offering opportunities for both short- and long-term investors.

Recommendations:

- *Investors: Consider Bharti Airtel for long-term holdings due to its steady upward trajectory.*
- *Short-term traders should capitalize on volatility by monitoring resistance levels around 700-800.*
- *management: Continue to focus on digital expansion and 5G rollouts, which are driving growth.*

Tata Motors Trend Analysis: *Tata Motors follows a U-shaped pattern. After a major decline between 2017-2020, the stock shows strong recovery, surpassing 1000 by 2024, indicating a comeback driven by electric vehicle (EV) development and recovery in global demand.*

*Threshold & Volatility**:** The stock is highly volatile with sharp price changes, particularly during the recovery phase. A threshold around 600 can be considered a critical level, indicating investor confidence or correction points.*

Recommendations:

- *Investors: Tata Motors offers long-term potential, especially with its EV focus.*
- *Short-term traders can leverage price volatility, but cautious stops around 600 are recommended.*
- *Management: Continue leveraging the EV market and sustainability trends to maintain growth momentum.*

DLF Limited Trend Analysis: *DLF shows a prolonged stagnation period from 2016 to 2020, followed by rapid growth from 2021 onwards. The surge indicates renewed interest in real estate, likely driven by favorable government policies or economic recovery.*

Threshold & Volatility: DLF exhibits sharp growth, surpassing 800 by 2024. A key threshold to monitor is 500, which could indicate market pullbacks or a phase of consolidation. Volatility is high, particularly during the upward trend.

Recommendations:

- *Investors: DLF is a solid buy for long-term growth, particularly with the real estate market improving.*
- *Short-term traders should take advantage of volatility with stop-losses around 500.*
- *Management: Focus on expanding the luxury segment and affordable housing projects to continue capitalizing on the real estate boom.*

Yes Bank Trend Analysis*: Yes Bank experienced a rapid decline starting in 2019, with prices dropping from over 300 to below 50 in early 2020. Post-2020, the stock price stabilized at a much lower level, around 10-20, with little signs of recovery.*

Threshold & Volatility: The stock has low volatility post-2020, hovering around the 10-20 level. Any price movement above 30 could indicate early recovery signs. Volatility remains low, suggesting little investor confidence.

Recommendations:

- *Investors: Yes Bank remains a high-risk stock due to its weak recovery. It may be suitable for speculative investors looking for high-risk, high-reward opportunities if restructuring gains momentum.*
- *Management: Focus on rebuilding investor trust by addressing governance and asset quality issues. Strengthen the balance sheet through strategic partnerships or capital infusions.*