PROJECT REPORT

DSBA

By
Siddhanth Gupta

greatlearning
Learning for Life

INTRODUCTION

This report presents a classification model designed to assess the financial health of a company using Logistic Regression. The objective is to determine whether a company is in stable financial condition and whether it will maintain a positive net worth in the upcoming year. The implementation of this model has been carried out using Python. **Click here for the Jupyter Notebook.**

PROBLEM STATEMENT

Companies face financial risk if they fail to meet their debt obligations, which can lead to defaults. A default negatively impacts the company's credit rating, making it more difficult to secure future credit and potentially resulting in higher interest rates on both existing and new debts. From an investor's perspective, an ideal company is one that can effectively manage its financial responsibilities, sustain growth, and scale efficiently.

A balance sheet serves as a crucial financial statement, offering insights into a company's assets, liabilities, and shareholder investments. It plays a vital role in evaluating a business's overall performance.

The dataset used in this study includes financial statement details of companies from the year 2015. Additionally, the net worth of each company in the subsequent year (2016) is provided, enabling us to generate labeled data for model training.

Data Description :

- Networth Next Year: Net worth of the customer in the next year
- Total assets: Total assets of customer
- Net worth: Net worth of the customer of the present year
- Total income: Total income of the customer
- Change in stock: Difference between the current value of the stock and the value of stock in the last trading day
- Total expenses: Total expenses done by the customer
- Profit after tax: Profit after tax deduction
- PBDITA: Profit before depreciation, income tax, and amortization
- PBT: Profit before tax deduction
- Cash profit: Total Cash profit
- PBDITA as % of total income: PBDITA / Total income
- PBT as % of total income: PBT / Total income
- PAT as % of total income: PAT / Total income
- Cash profit as % of total income: Cash Profit / Total income
- PAT as % of net worth: PAT / Net worth
- Sales: Sales done by the customer
- Income from financial services: Income from financial services
- Other income: Income from other sources
- Total capital: Total capital of the customer

- Reserves and funds: Total reserves and funds of the customer
- Borrowings: Total amount borrowed by the customer
- Current liabilities & provisions: current liabilities of the customer
- Deferred tax liability: Future income tax customer will pay because of the current transaction
- Shareholders funds: Amount of equity in a company which belongs to shareholders
- Cumulative retained profits: Total cumulative profit retained by customer
- Capital employed: Current asset minus current liabilities
- TOL/TNW: Total liabilities of the customer divided by Total net worth
- Total term liabilities / tangible net worth: Short + long term liabilities divided by tangible net worth
- Contingent liabilities / Net worth (%): Contingent liabilities / Net worth
- Contingent liabilities: Liabilities because of uncertain events
- Net fixed assets: The purchase price of all fixed assets
- Investments: Total invested amount
- Current assets: Assets that are expected to be converted to cash within a year
- Net working capital: Difference between the current liabilities and current assets
- Quick ratio (times): Total cash divided by current liabilities
- Current ratio (times): Current assets divided by current liabilities
- Debt to equity ratio (times): Total liabilities divided by its shareholder equity
- Cash to current liabilities (times): Total liquid cash divided by current liabilities
- Cash to average cost of sales per day: Total cash divided by the average cost of the sales
- Creditors turnover: Net credit purchase divided by average trade creditors
- Debtors turnover: Net credit sales divided by average accounts receivable
- Finished goods turnover: Annual sales divided by average inventory
- WIP turnover: The cost of goods sold for a period divided by the average inventory for that period
- Raw material turnover: Cost of goods sold is divided by the average inventory for the same period
- Shares outstanding: Number of issued shares minus the number of shares held in the company
- Equity face value: cost of the equity at the time of issuing
- EPS: Net income divided by the total number of outstanding share
- Adjusted EPS: Adjusted net earnings divided by the weighted average number of common shares outstanding on a diluted basis during the plan year
- Total liabilities: Sum of all types of liabilities
- PE on BSE: Company's current stock price divided by its earnings per share

```
Number of Rows(Companies) = 4256
Number of Columns = 51
```

1. Missing Value Treatment

• The dataset contains less than 1% missing values.

• These missing values are imputed using the **K-Nearest Neighbors (KNN) Imputer** with **n_neighbors = 10**.

2. Handling Zero Values

• The dataset has a significant number of zero values, accounting for **15.1%** of the total data.

• Columns with more than **30% zero values** are dropped (**9 columns** removed).

• Analysis revealed that **164 out of 387** defaulting companies had more than **five zero values** in their records.
• **Conclusion**: A higher number of missing or zero values correlates with an increased likelihood of default.
• Remaining zero values are treated as missing values (**NaN**) and imputed using the **KNN Imputer (n_neighbors = 10)**.

3. Outlier Treatment
• **Interquartile Range (IQR) and Z-Score methods** were used separately to detect and manage outliers.
• Different **Logistic Regression** models were trained and tested using both techniques.
• The **Z-Score method** provided better results on the test dataset, so it was chosen for the final model.

4. Feature Scaling
• **Z-Score standardization** was applied to normalize the data.

5. Multi-Collinearity Handling
• Many variables in the dataset were found to be highly correlated.
• This **multi-collinearity** can affect model interpretability and predictive power.
• The **Variance Inflation Factor (VIF) method** was used to detect and remove highly correlated features.
• Columns with **VIF > 5** were **recursively dropped** one by one.

6. Feature Engineering
• Initially, **66 predictor variables** were considered.
• Steps for feature selection and refinement:
• **Unique identifiers** (e.g., Company Code, Name) were removed → **64 variables left**.
• **Columns with >30% zero values** were dropped → **55 variables left**.
• Variables with **VIF > 5** were eliminated:
• After **IQR-based outlier removal → 27 variables retained**.
• After **Z-Score-based outlier removal → 23 variables retained**.
• Additional refinement: Features with **p-values > 0.05** (at **95% confidence**) were excluded for some models.
• For **Model #7**, **Recursive Feature Elimination (RFE)** was used to select the **15 most important features**.

7. Best Model Selection & Deployment
• **Model #8** was identified as the best-performing model based on recall and precision for detecting default cases (**default = 1**).
• Model specifications:
• **Outlier treatment → Z-Score method**, with values capped at **±3 standard deviations**.
• **Feature selection → Recursive Feature Elimination (RFE)**, keeping the **top 15 features**.
• **Oversampling → SMOTE** applied to achieve a **50-50 balance** between defaulting and non-defaulting companies.
• **Optimal decision threshold** set at **0.5**.

Model Performance Metrics (for default = 1)
• **Recall**: **95%**
• **Precision**: **78%**
• **Accuracy**: **96%**
• **F1-Score**: **86%**

Importance of Outlier Treatment

Outlier treatment is essential in regression models to ensure accurate predictions. Outliers can significantly influence the regression line by pulling it towards extreme values, thereby affecting its slope and distorting the overall analysis. Addressing outliers helps improve model reliability and predictive accuracy.

Methods of Outlier Detection and Treatment

Two techniques are applied to detect and handle outliers in this case study:

1. **Interquartile Range (IQR) Treatment**

• The **25th percentile (Q1)** and **75th percentile (Q3)** are calculated.

• The **Interquartile Range (IQR)** is determined as **Q3 - Q1**.

• Values beyond **1.5 times the IQR** from Q1 and Q3 are considered outliers.

• Outliers are capped at their respective upper or lower limits to reduce their influence.

2. **Z-Score Treatment**

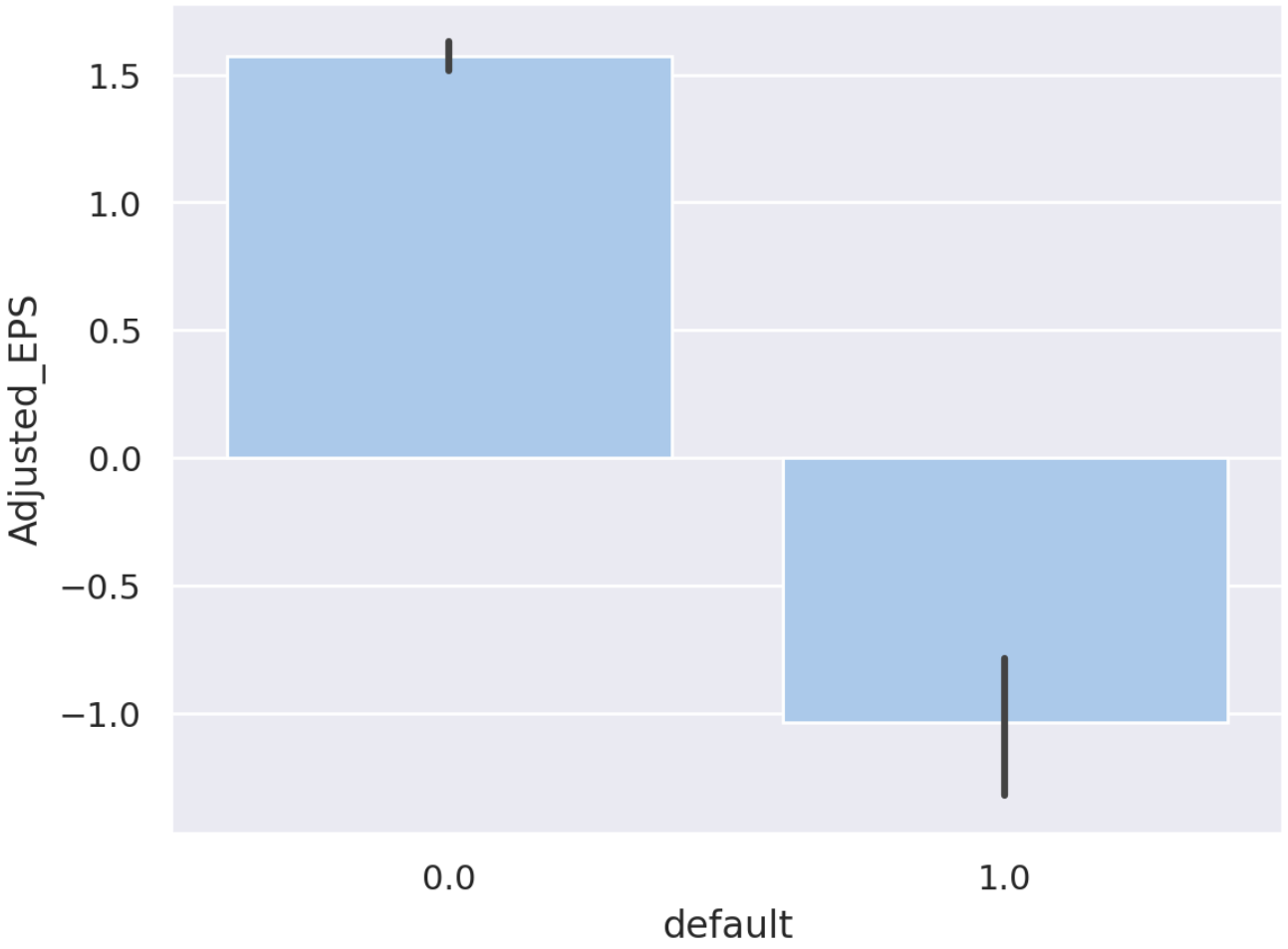
• The **Z-score method** standardizes data by measuring how many standard deviations a value deviates from the mean.

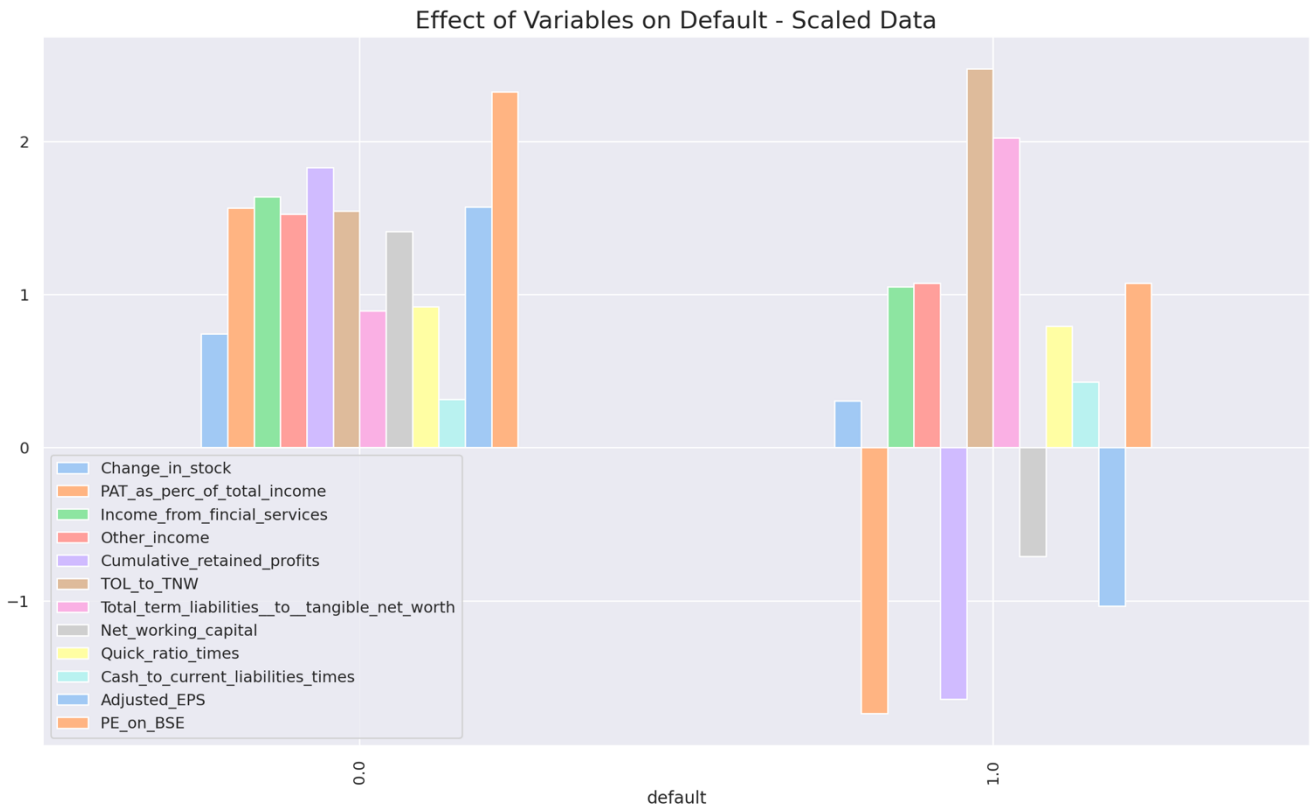
• Values exceeding **±3 standard deviations** are identified as outliers.

• Extreme values are capped within this range to maintain model stability.

Visualization and Feature Selection

Box plots of **15 selected variables** are presented before and after outlier treatment to demonstrate the impact of these methods. Scaling is applied to enhance comparison across variables. These **15 variables** are ultimately chosen as the best predictors for **Logistic Regression**.


1.5
1.0
0.5
0.0
−0.5
−1.0
Adjusted_EPS
0.0
1.0
default


Effect of Variables on Default - Scaled Data
2
1
0
−1
Change_in_stock
PAT_as_perc_of_total_income
Income_from_fincial_services
Other_income
Cumulative_retained_profits
TOL_to_TNW
Total_term_liabilities__to__tangible_net_worth
Net_working_capital
Quick_ratio_times
Cash_to_current_liabilities_times
Adjusted_EPS
PE_on_BSE
0.0
1.0
default

Missing values in the raw dataset are minimal, accounting for approximately **0.05%** of the total data. However, a significant proportion of zero values (**around 15%**) exist, which primarily serve as placeholders for missing values. These zero values contribute little to predictive performance and, in many cases, introduce computational issues.

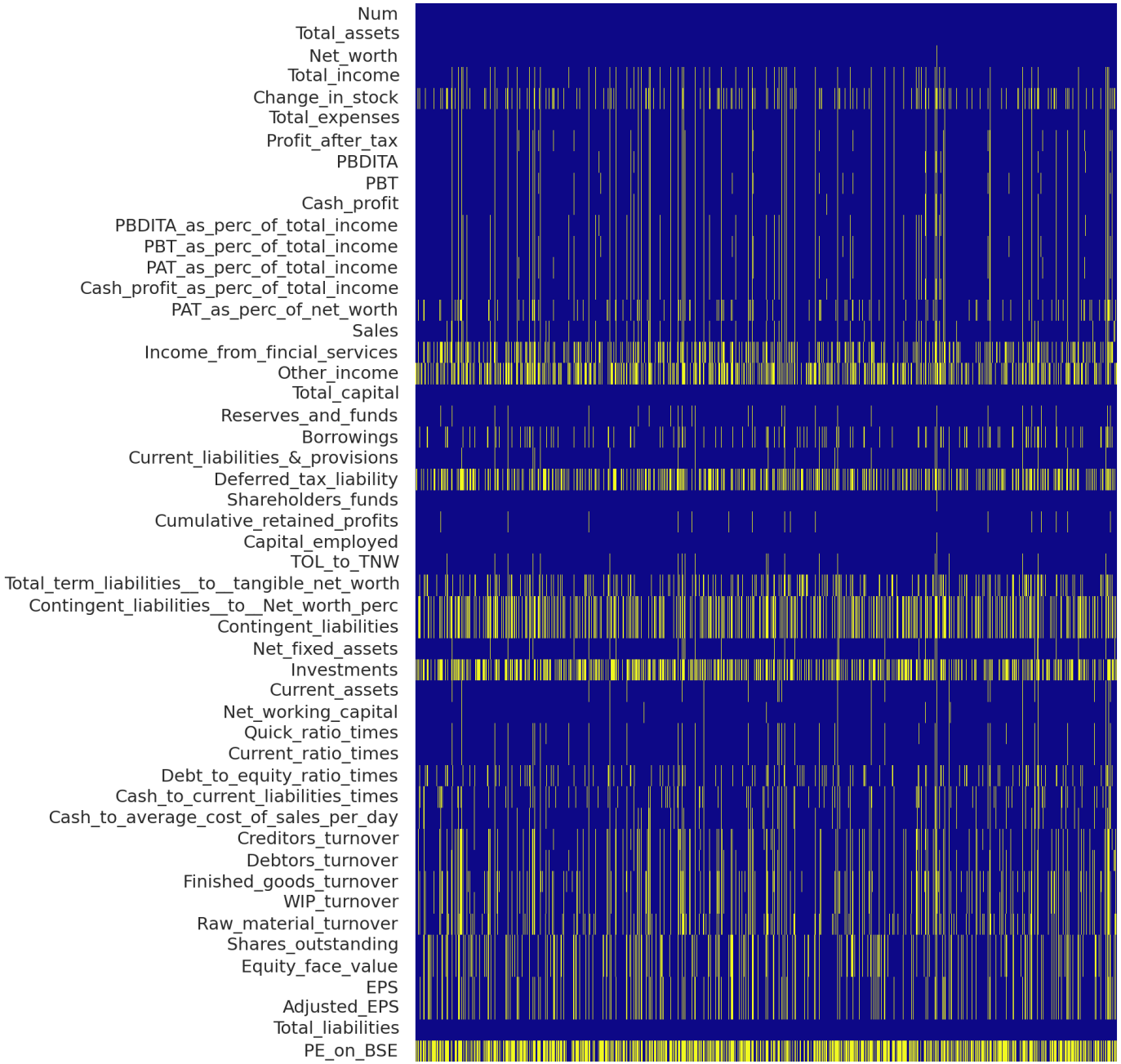
A high number of zero values in any feature can lead to **"Linear Algebra Error"** when using **StatsModel**, making it essential to handle them properly. To ensure model stability and accuracy, appropriate methods are applied to treat these zero values effectively.

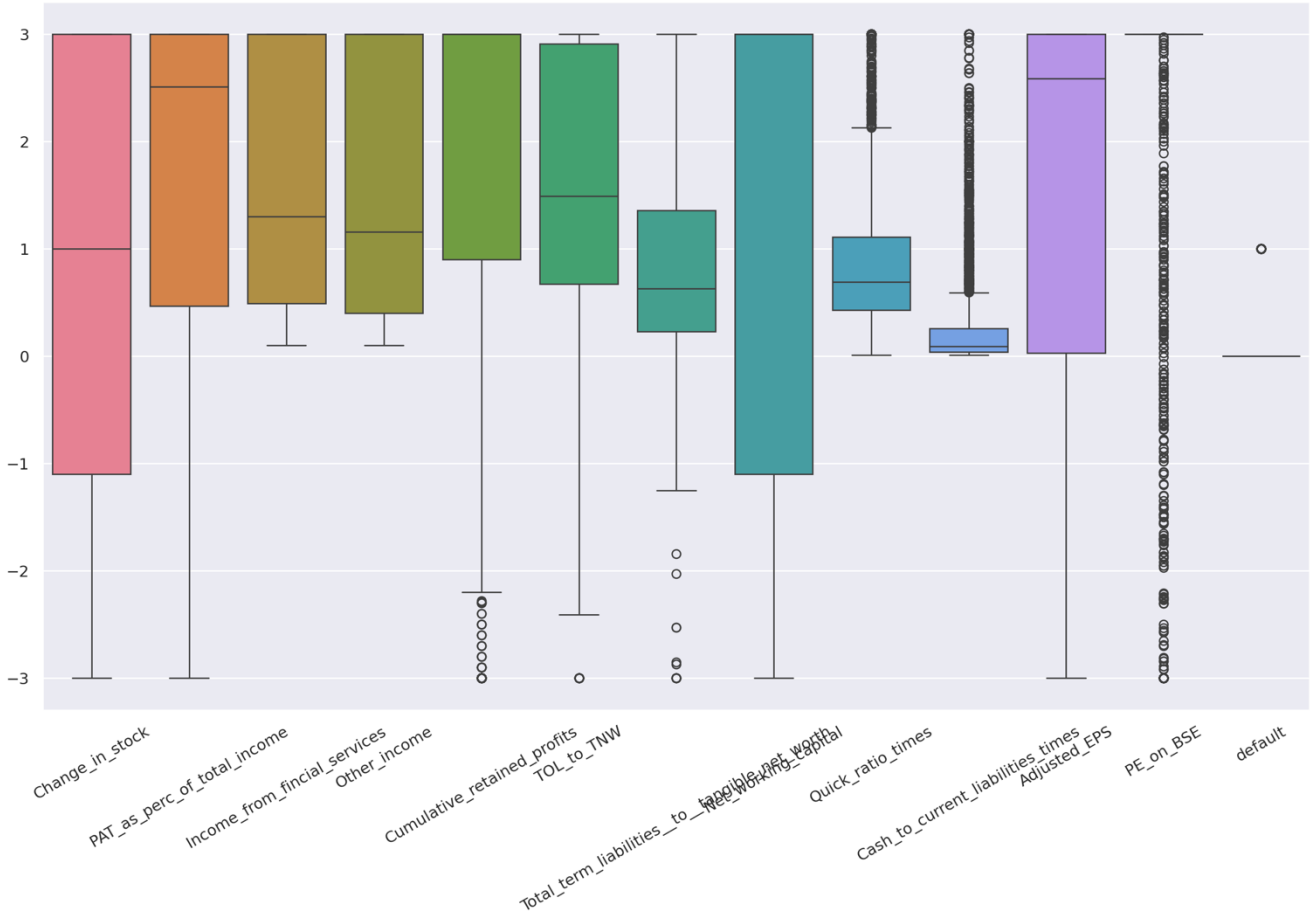
The first step in handling missing values involves dropping all features where **zero values exceed 30%** of the total observations. These features provide little predictive value and may introduce bias in the model.

Next, the remaining zero values are converted into **missing values (NaN)** to accurately represent data gaps. This transformation ensures that zero placeholders do not distort the dataset's structure.

Once all missing values, both original and transformed, are identified, **K-Nearest Neighbors (KNN) Imputer (n_neighbors = 10)** is applied to impute them. This method fills in missing values based on the nearest available data points, preserving underlying patterns in the dataset.

A visual representation of missing values after dropping the selected features is provided for better understanding.


Num
Total_assets
Net_worth
Total_income
Change_in_stock
Total_expenses
Profit_after_tax
PBDITA
PBT
Cash_profit
PBDITA_as_perc_of_total_income
PBT_as_perc_of_total_income
PAT_as_perc_of_total_income
Cash_profit_as_perc_of_total_income
PAT_as_perc_of_net_worth
Sales
Income_from_fincial_services
Other_income
Total_capital
Reserves_and_funds
Borrowings
Current_liabilities_&_provisions
Deferred_tax_liability
Shareholders_funds
Cumulative_retained_profits
Capital_employed
TOL_to_TNW
Total_term_liabilities__to__tangible_net_worth
Contingent_liabilities__to__Net_worth_perc
Contingent_liabilities
Net_fixed_assets
Investments
Current_assets
Net_working_capital
Quick_ratio_times
Current_ratio_times
Debt_to_equity_ratio_times
Cash_to_current_liabilities_times
Cash_to_average_cost_of_sales_per_day
Creditors_turnover
Debtors_turnover
Finished_goods_turnover
WIP_turnover
Raw_material_turnover
Shares_outstanding
Equity_face_value
EPS
Adjusted_EPS
Total_liabilities
PE_on_BSE


3
2
1
0
−1
−2
−3
Change_in_stock
PAT_as_perc_of_total_income
Income_from_fincial_services
Other_income
Cumulative_retained_profits
TOL_to_TNW
Total_term_liabilities__to__tangible_net_worth
Net_working_capital
Quick_ratio_times
Cash_to_current_liabilities_times
Adjusted_EPS
PE_on_BSE
default

The target variable is transformed into a **binary classification format (0 and 1)** to facilitate logistic regression modeling. The financial health of companies is assessed based on whether they will have a **positive or negative net worth** in the following year.

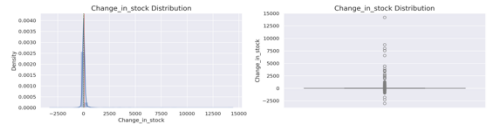
To define the target variable:

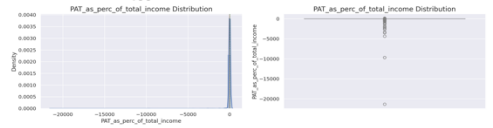
• The **'Networth Next Year'** column is selected as the **Default Variable**.

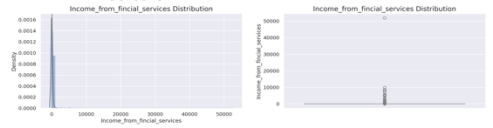
• Companies with **negative net worth** are labeled as **Default = 1**, indicating financial distress.

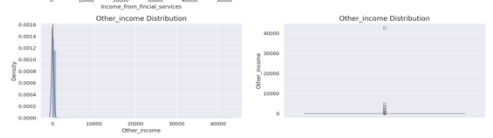
• Companies with **zero or positive net worth** are labeled as **Default = 0**, signifying financial stability.

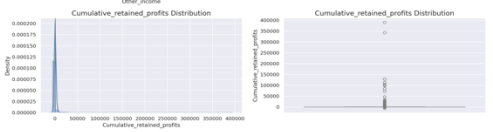
UNIVARIATE ANALYSIS

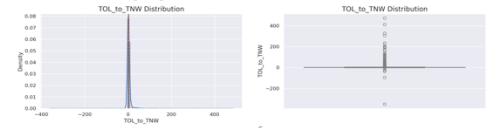


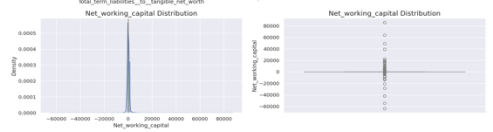


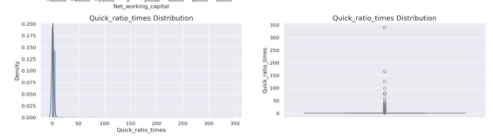


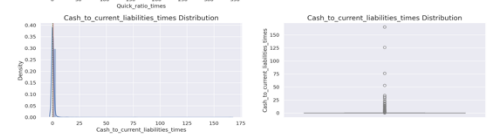


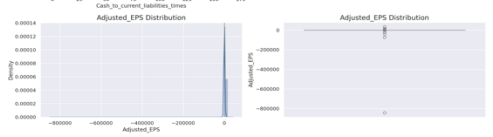


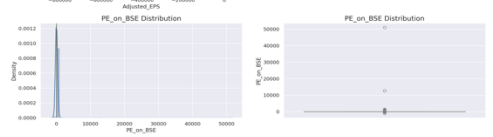


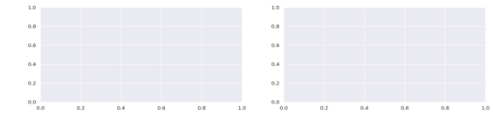












The distribution of the **top 15 predictor variables** after **Z-score scaling** is analyzed to understand their central tendencies and skewness. The visualization includes **coloured vertical lines** indicating:

• **Mean (Red)**

• **Median (Black)**

• **Mode (Green)**

Key Observations:

1. **Selling Cost**

• The majority of companies cluster around the **mean**.

• The distribution exhibits **right skewness**, with outliers present on the higher side.

2. **PBIDT (Profit Before Interest, Depreciation, and Tax)**

• Most companies lie close to the **mean** but show a strong **right skew**.

• The presence of companies with high **PBIDT** suggests strong initial financial performance before additional costs.

3. **Cash Flow from Operating Activities**

• Follows a **normal distribution**, with the majority of companies concentrated near the **mean**.

4. **Other Financial Ratios (ROG Networth, ROG Capital Employed, ROG Total Assets, ROG PBIDT, ROG PBT, ROG CP, Current Ratio, Interest Cover Ratio, Value of Output to Total Assets, Net Working Capital, Book Value Adjusted)**

• Most companies are densely packed around the **mean**.

• The distributions exhibit **right skewness**, indicating that some companies have significantly higher values, acting as outliers.

5. **APATM (After Tax Profit Margin)**

• The distribution is **left-skewed**, meaning a considerable number of companies have lower **net profit** after tax.

• This trend suggests a potential **default risk**, as companies struggle to maintain profitability after all costs are accounted for.

Overall Insight:

Many companies demonstrate **strong financial performance** before tax and additional costs. However, after incorporating all expenses, a significant portion **shifts towards the lower end** of profitability. This pattern highlights the need for businesses to **optimize costs and improve bottom-line management** to maintain financial stability.

BIVARIATE ANALYSIS

- Correlation Heat-map is given below of 55 variables (after dropping vars with zeros>30%)

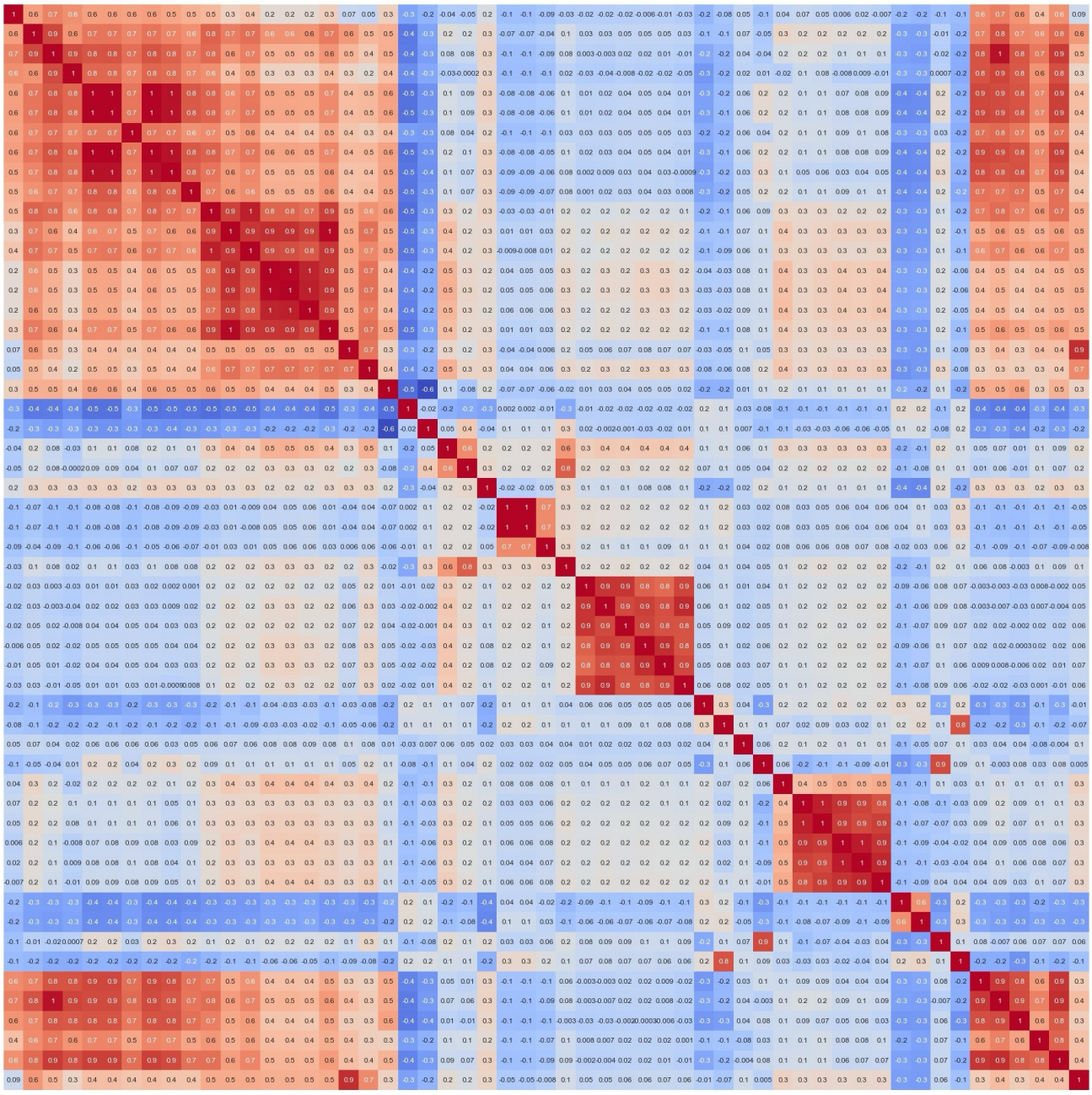
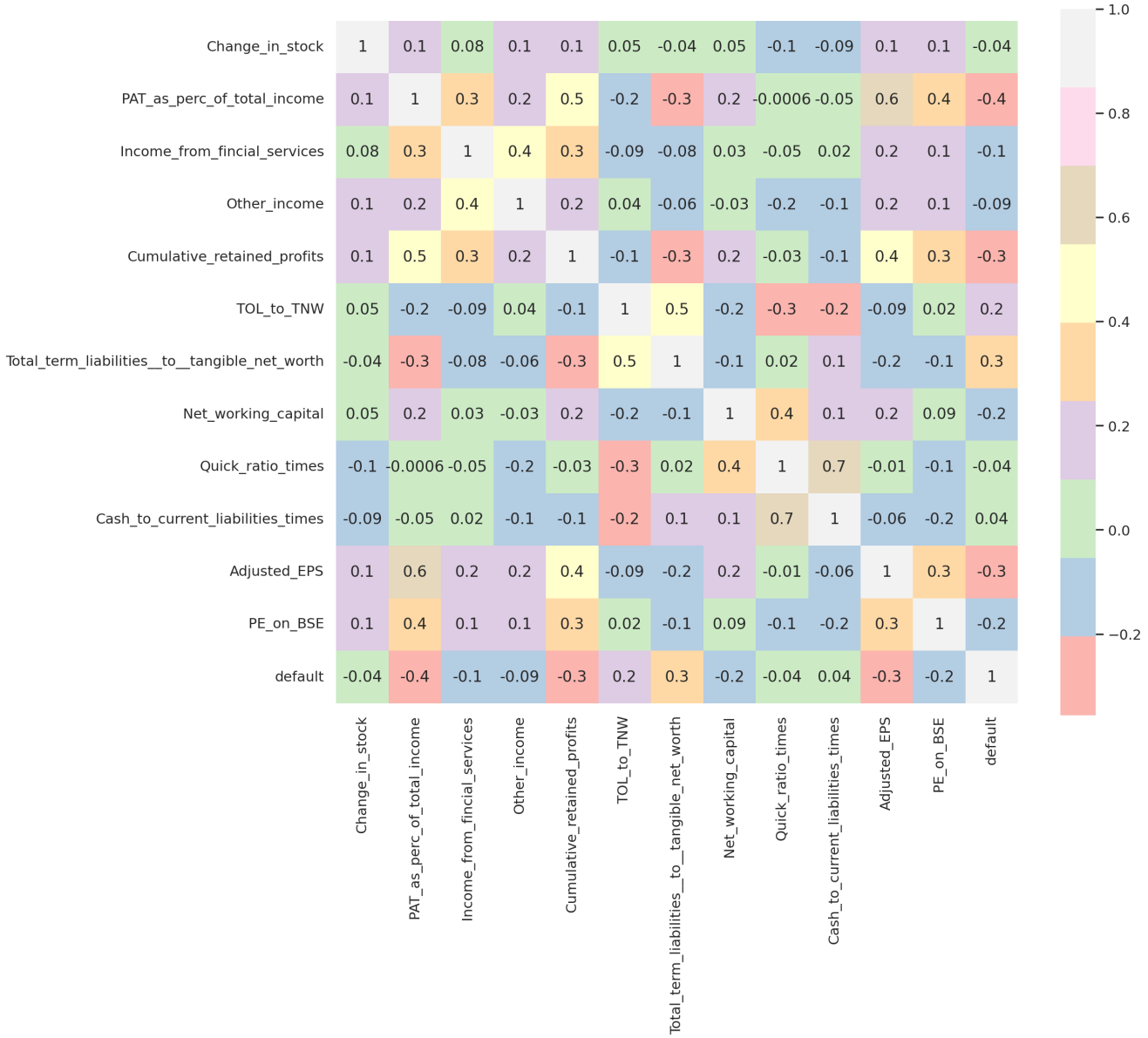


Fig 7 - Correlation Heatmap - 55 Variables

- There are a lot of patches seen. This indicates high correlation between many variables
- Highly correlated features cause Multi-Collinearity which affect the interpretability of Logistic Regression model. They are best removed.

- We use Variance Inflation factor method and remove all variables with VIF > 5. This is done recursively, one-by-one
- Correlation heat-map of top 15 predictors and 1 Target, used to get the best model is given below-



Train Test Split

- We use train_test_splitfunction from scikit-learn library to split the data into train and validation sets
- We split in the ratio of 67-33 - 67% in Training Set and 33% in Testing (Validation) Set
- We seed this split at random_state=42
- So after split, Out of Total 3586 —> Train Set has 2402 observations

  Test Set has 1184 observations

Build Logistic Regression Model (using stats model library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach

PREPROCESSING DATA FOR ALL MODELS

- We had large number of 66 predictor variables in the raw data
- 2 unique identifiers - Company Code and Name - were dropped
- There are large number of zero values in the data. We drop all variables with zeros > 30% (9 vars)
- Highly correlated features exist in the data, which cause Multi-Collinearity. This indicates existence of redundant features
- We perform Outlier treatments using IQR and Z-score methods
- We drop all features recursively one-by-one with VIF > 5 (VIF - Variance Inflation Factor)
- We are left with - 27 variables after IQR Outlier treatment

  - 23 variables after Z-score treatment

LOGISTIC REGRESSION MODELS

- We build multiple Logistic Regression models with different approaches and strategies. We test each model on Test set and fine-tune to improve Recall and Precision of default=1
- We use StatsModel and SciKitLearn libraries to build and test these models

- Model 1 :
    - 27 vars, IQR Outlier treated
- Model 2 :
    - From 27 vars - insignificant vars dropped with p-values > 0.05 (final 10 vars)
    - IQR Outlier treatment
- Model 3 :
    - 23 vars, Z-score Outlier treatment
- Model 4 :
    - From 23 vars - insignificant vars dropped with p-values > 0.05 (final 9 vars)
    - Z-score Outlier treatment
- Model 5 :
    - Above 9 vars, Z Score treatment
    - Regularising the model by Hyper-parameter tuning with GridSearch over 10 folds over following -

      {'penalty':['l2','none', 'l1'],

      'solver':['lbfgs', 'liblinear', 'sag', 'saga', 'newton-cg'], 'tol':[0.0001,0.00001]}
    - Best Parameters were found as follows -

      {'penalty': 'none', 'solver': 'lbfgs', 'tol': 0.0001}
- Model 6 :
    - From 23 vars - insignificant vars dropped with p-values > 0.05 (final 9 vars)
    - Z-score Outlier treatment
    - Check for optimum threshold to get max Recall for default=1
    - This is obtained by maximising the difference between True Positivity rate and False Positivity rate (tpr - fpr)
    - Optimum Threshold = 0.084
- Model 7 :
    - 23 vars, Z-score Outlier treatment
    - Extracting top 15 features using Recursive Feature Elimination (RFE)
- Model 8 :
    - 23 vars, Z-score Outlier treatment
    - Extracting top 15 features using Recursive Feature Elimination (RFE)

    - This model gave the best metrics on Test Set
    - StatsModel report of Model 9 given below -
        - highest positive impact on Probability of Default

- Model 9 :
    - 23 vars, Z-score Outlier treatment
    - Extracting top 15 features using Recursive Feature Elimination (RFE)

- Check for optimum threshold to get max Recall for default=1
- This is obtained by maximising the difference between True Positivity rate and False Positivity rate (tpr - fpr)
- Optimum Threshold = 0.4246

- Model 10 :
  - Z-Score Outlier treatment, Top 15 features through RFE
  - Class balancing 50-50 using SMOTE
  - Optimum Threshold = 0.5
  - Dropping insignificant vars with p-values > 0.05

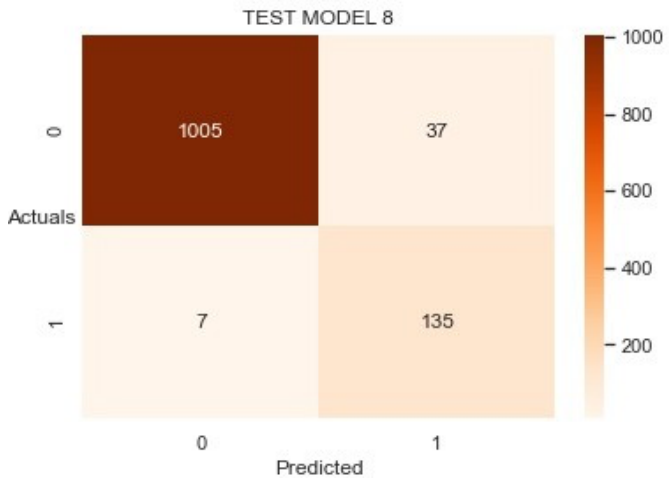
Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model

- Performance Metrics of all models on Test Dataset is given below

MODEL NAME	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
Model 1	99	23	61	38
Model 2	99	23	60	37
Model 3	88	90	97	89
Model 4	88	89	97	88
Model 5	88	89	97	88
Model 6	95	71	95	81
Model 7	87	89	97	88
Model 8	95	78	96	86

	precision	recall	f1-score	support
0	0.99	0.96	0.98	1042.00
1	0.78	0.95	0.86	142.00
accuracy	0.96	0.96	0.96	0.96
macro avg	0.89	0.96	0.92	1184.00
weighted avg	0.97	0.96	0.96	1184.00

Table 6 - Classification Report of Model 8



Model Comparison and Selection

• Models **1 and 2** achieved a **Recall of 99%**, but their **Precision was very low**, making them unsuitable for deployment.

• **Model 8** was selected as the final model due to its **optimal balance between Recall and Precision**.

Performance Metrics of Model 8

• **Recall = 95% → 95% of actual defaulting companies were correctly predicted** as defaults.

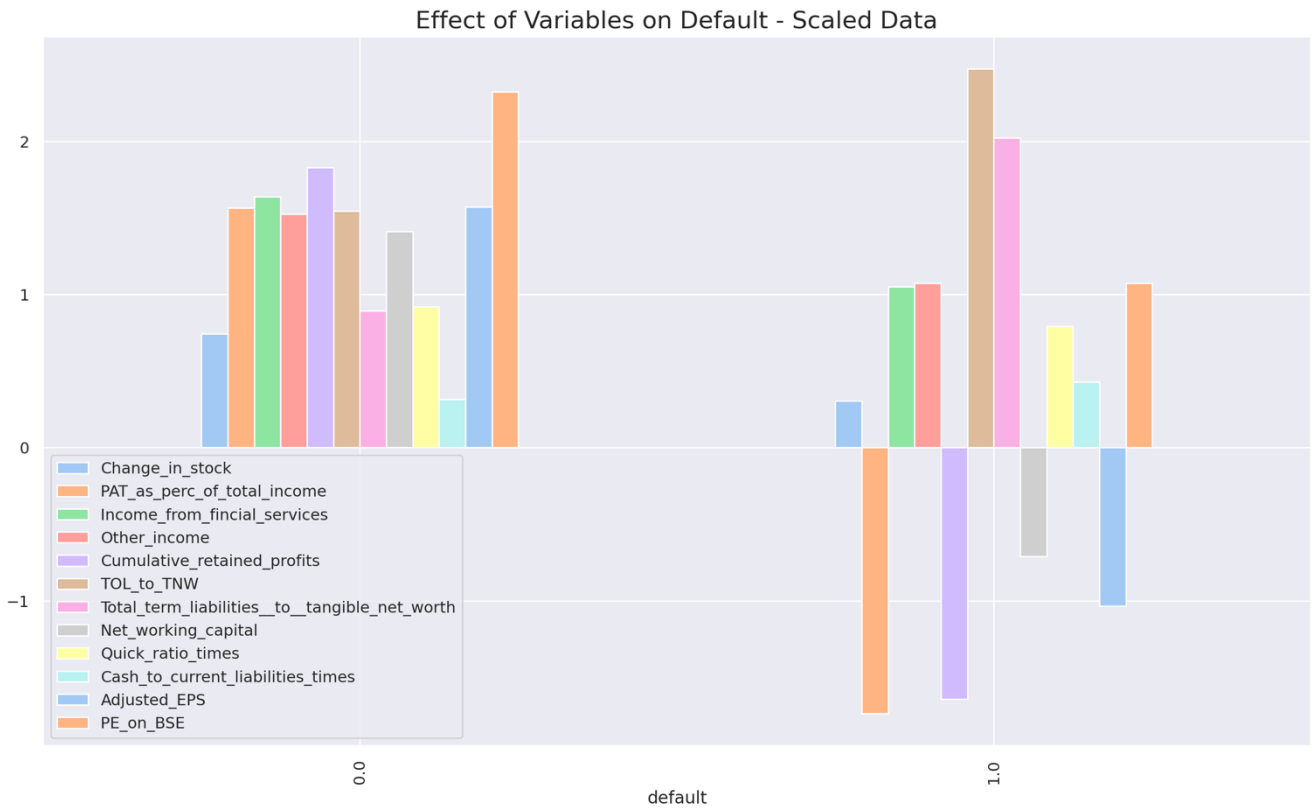
• **Precision = 78% → 78% of predicted defaults were actual defaults**.

Interpretation and Justification

• The primary goal of this model was to **identify as many actual defaults as possible**, minimizing **Type II errors** (false negatives).

• **Recall** was prioritized first to **capture the maximum number of actual defaults**, followed by **Precision** to reduce false alarms.

• Table **4** presents the **coefficients** of all variables, indicating their respective importance in predicting default risk.



The coefficients of the logistic regression model provide insights into the impact of different financial variables on the likelihood of default:

• **Positive Coefficient**: A higher value of this variable increases the probability of default, assuming all other factors remain constant.

• **Negative Coefficient**: A higher value of this variable decreases the probability of default, assuming all other factors remain constant.

Key Observations from Multi-Variate Analysis

• Figure **9** illustrates the influence of various financial features on default probability.

• Many companies exhibit **strong profit margins before accounting for taxes, interest, and other costs**.

• However, after incorporating **taxes, depreciation, and additional expenses**, a significant number of companies **shift to the lower half of the profitability spectrum**.

• This pattern underscores the importance of **cost optimization and efficient financial management** to sustain long-term profitability.