CDAC-KHAGHAR

Project Presentation
On
“Bankruptcy **Prediction**”

Abhishek Hingmire 02

Guided by:
Prashant Bhonsle

OUTLINE

- Motivation
- Introduction
- Problem Statement
- Proposed Algorithms
- Statistical Model
- System Architecture Diagram
- Work Flow Diagram
- System Requirements
- Results

MOTIVATION

Advance data analytics is the field which most companies are adapting to make business decisions.

Statistical techniques have been widely employed to enhance bankruptcy prediction accuracy.

Advance statastical analysis used for risk of bankruptcy of company

INTRODUCTION

To Perform an in-depth Analysis of a Financial Dataset to Predict the Likelihood of a Company Going Bankrupt. The Analysis Involves Data Preprocessing, Exploratory Data Analysis (Eda), Hypothesis Testing, Feature Engineering and Selection, and Applying Machine Learning Techniques for Classification.

PROBLEM STATEMENT

- **Title:**
- Predicting corporate bankruptcy using financial and operational data

- **Background:**
- Corporate bankruptcy is a critical event that can have significant ramifications for stakeholders, including investors, employees, creditors, and the broader economy. Accurately predicting bankruptcy can help mitigate these impacts by allowing for early intervention and informed decision-making. This study aims to analyze various financial and operational factors to determine their influence on the likelihood of a company declaring bankruptcy.

- **Objective:**
- To develop a predictive model that identifies the key factors influencing corporate bankruptcy and accurately predicts whether a company will go bankrupt based on these factors.

- **Research questions:**

- What financial and operational factors are most strongly associated with corporate bankruptcy?
- How can these factors be quantified and modeled to predict the likelihood of bankruptcy?
- What is the relative importance of these factors in determining a company's financial health?

Scope:

- The dataset contains historical financial and operational data of companies, including a target variable indicating whether a company has gone bankrupt.
- The analysis will include exploratory data analysis (EDA), feature selection, model building, and evaluation.

Data Description:

- **Features:** A variety of financial ratios (e.g., liquidity ratios, profitability ratios, leverage ratios, activity ratios), operational metrics, and possibly non-financial data such as market sentiment or management effectiveness.
- **Target Variable:** A binary variable indicating whether the company has gone bankrupt (1) or not (0).

PROPOSED ALGORITHMS

- To develop the predicting model for this project we will be using supervised predictive statistics,.
- In our project we will be using supervised predictive statisticsas our dataset has As our dataset consist of continuous values and categorical we will be using regression algorithms to find relationship between **Bankurupt** and the features of the dataset. Algorithms that we will be using are:
  1. **Logistic Regression**

Overview of logistic regression

Logistic regression is a statistical method used for binary classification. It predicts the probability of a binary outcome based on one or more predictor variables. The outcome is typically coded as 0 or 1, where 1 indicates the occurrence of the event of interest (e.G., bankruptcy) and 0 indicates its absence.

Key concepts

Binary outcome: logistic regression is used when the dependent variable is binary. It predicts the probability that the outcome variable belongs to a particular category.

Logistic function (sigmoid function): the logistic function maps predicted values to probabilities:

P(y=1)=11+e−(β0+β1x1+β2x2+⋯+βnxn)p(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_n)}}p(y=1)=1+e−(β0 +β1 x1 +β2 x2 +⋯+βn xn)1

This ensures that the output probabilities range between 0 and 1.

Odds and log-odds:

Odds: the ratio of the probability of the event occurring to the probability of it not occurring: odds=p(y=1)1−p(y=1)\text{odds} = \frac{p(y=1)}{1 - p(y=1)}odds=1−p(y=1)p(y=1)

L

log-odds (logit): the natural logarithm of the odds: logit(p)=log(p(y=1)1−p(y=1))=β0+β1x1+β2x2+⋯+βnxn\text{logit}(p) = \log\left(\frac{p(y=1)}{1 - p(y=1)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_nlogit(p)=log(1−p(y=1)p(y=1))=β0 +β1 x1 +β2 x2 +⋯+βn xn

Maximum likelihood estimation (mle): the coefficients (βI\beta_iβI) are estimated using mle, which finds the values that maximize the likelihood of observing the given data.

Assumptions of logistic regression

Binary dependent variable: the outcome must be binary.

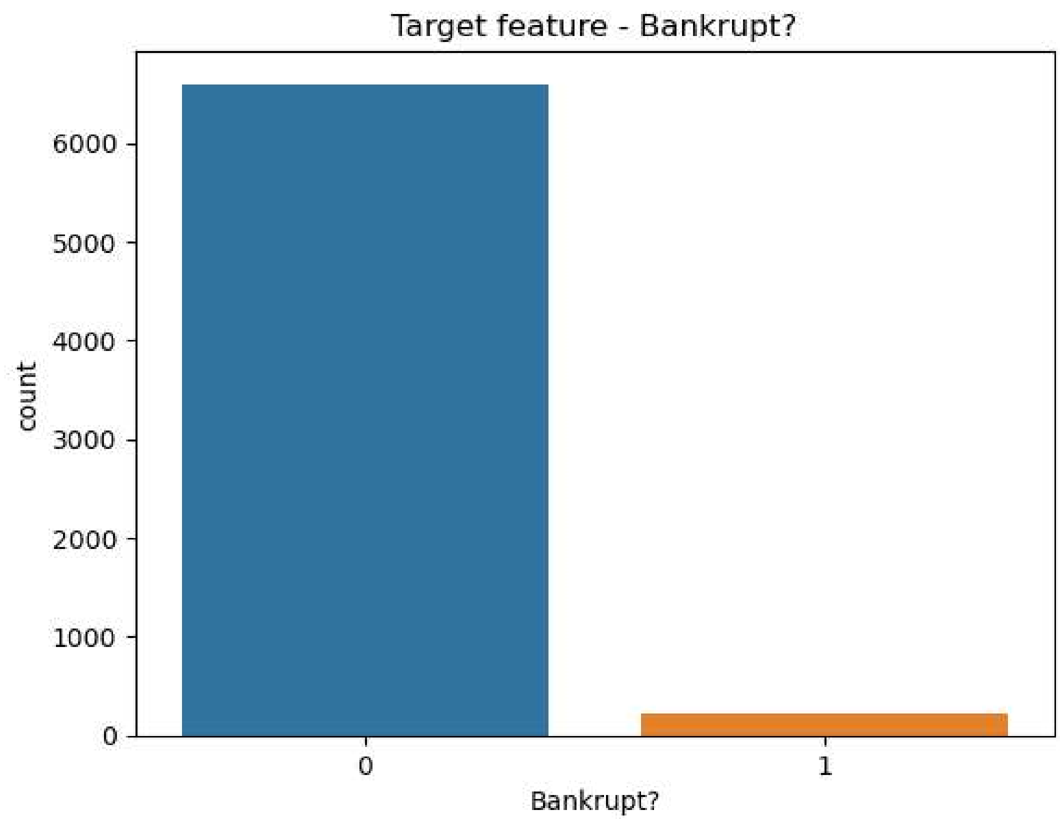
Linearity of independent variables and logit: the relationship between the independent variables and the log-odds of the dependent variable should be linear.

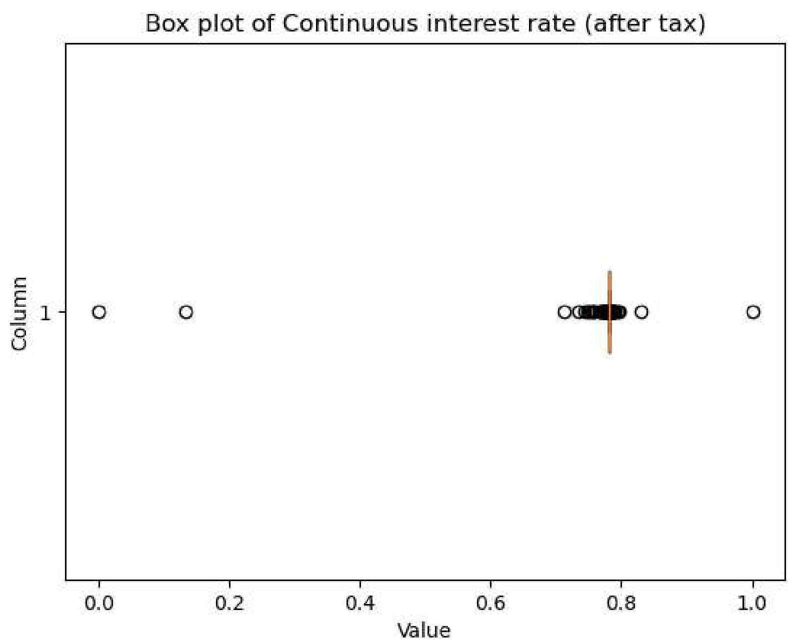
Independence of observations: observations should be independent of each other.

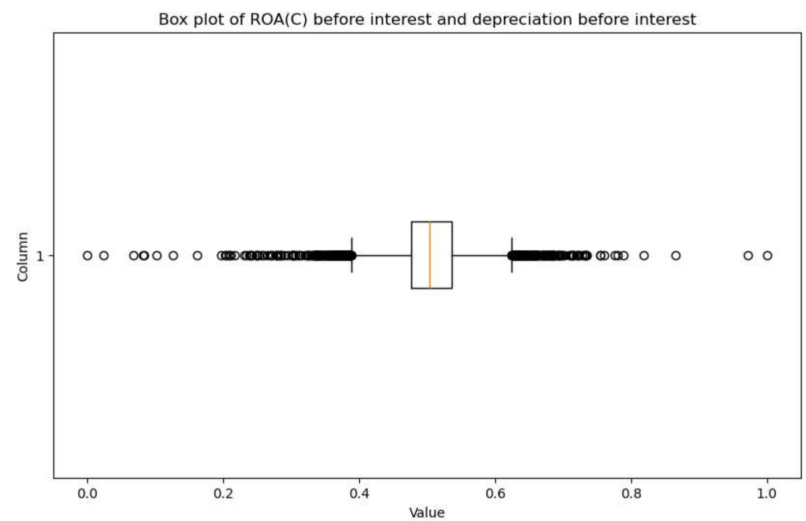
Absence of multicollinearity: independent variables should not be highly correlated with each other.

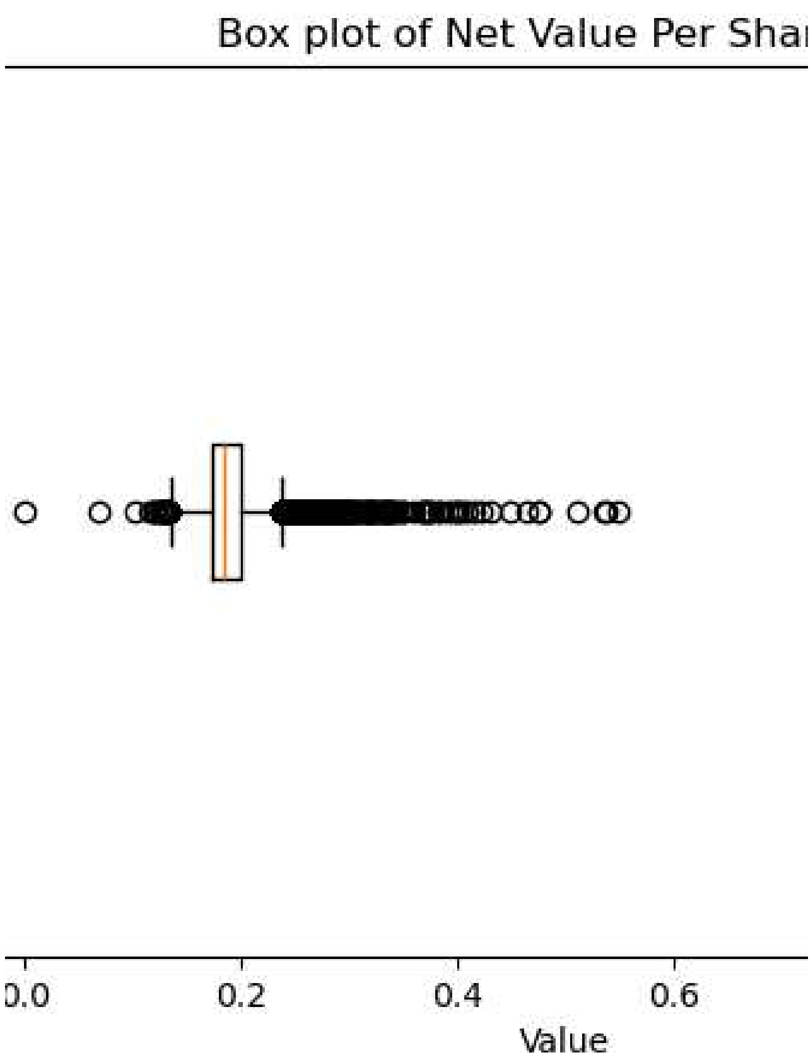
Large sample size: logistic regression requires a large sample size to produce reliable results.

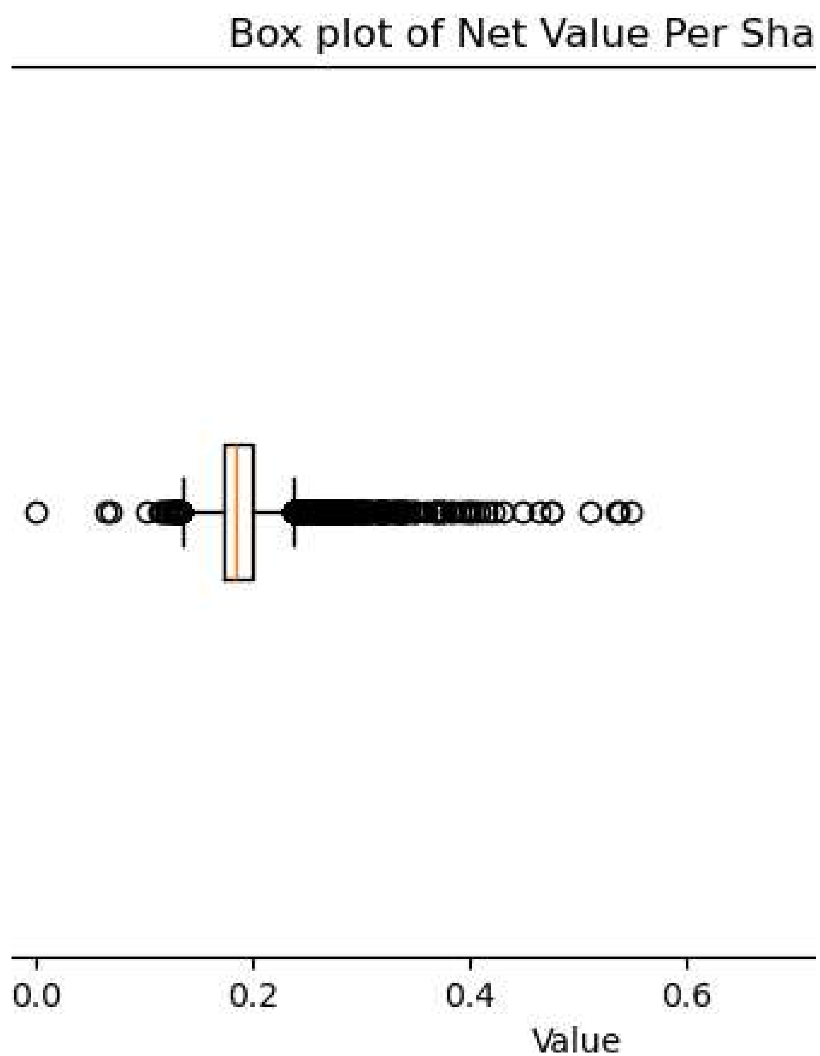
- **Exploratory Data Analysis (EDA):**

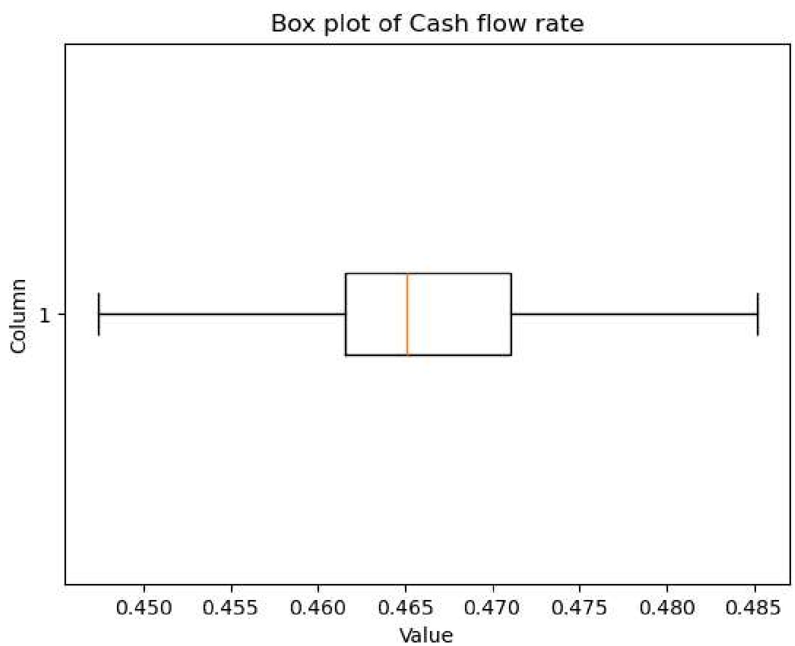


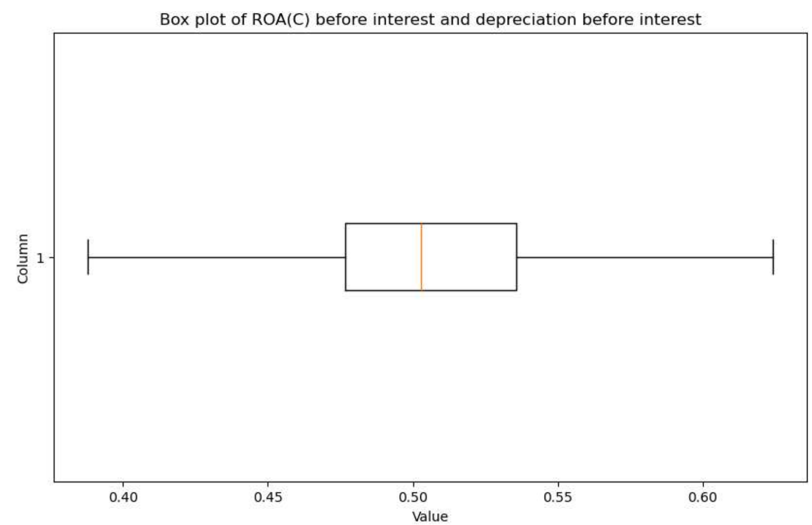

Box plot of Continuous interest rate (after tax)
Column
1
0.0
0.2
0.4
0.6
0.8
1.0
Value


Box plot of ROA(C) before interest and depreciation before interest
Column
1
0.0
0.2
0.4
0.6
0.8
1.0
Value


Box plot of Net Value Per Shai
0.0
0.2
0.4
0.6
Value

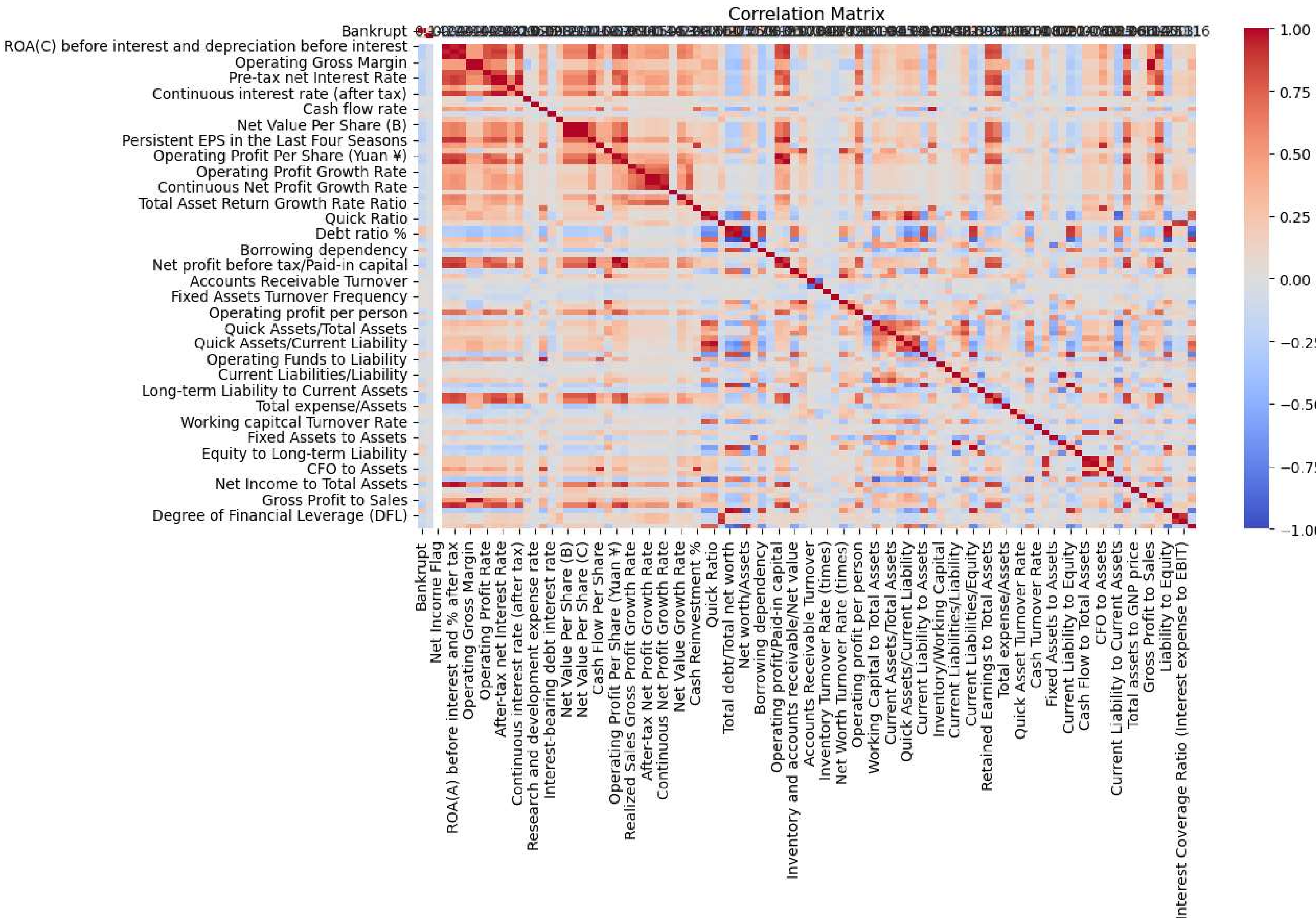

Box plot of Net Value Per Sha
0.0
0.2
0.4
0.6
Value


Box plot of Cash flow rate
Column
1
0.450
0.455
0.460
0.465
0.470
0.475
0.480
0.485
Value


Box plot of ROA(C) before interest and depreciation before interest
Column
1
0.40
0.45
0.50
0.55
0.60
Value


Bankrupt
Liability-Assets Flag
Net Income Flag
ROA(C) before interest and depreciation before interest
ROA(A) before interest and % after tax
ROA(B) before interest and depreciation after tax
Operating Gross Margin
Realized Sales Gross Margin
Operating Profit Rate
Pre-tax net Interest Rate
After-tax net Interest Rate
Non-industry income and expenditure/revenue
Continuous interest rate (after tax)
Operating Expense Rate
Research and development expense rate
Cash flow rate
Interest-bearing debt interest rate
Tax rate (A)
Net Value Per Share (B)
Net Value Per Share (A)
Net Value Per Share (C)
Persistent EPS in the Last Four Seasons
Cash Flow Per Share
Revenue Per Share (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Per Share Net profit before tax (Yuan ¥)
Realized Sales Gross Profit Growth Rate
Operating Profit Growth Rate
After-tax Net Profit Growth Rate
Regular Net Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Growth Rate
Net Value Growth Rate
Total Asset Return Growth Rate Ratio
Cash Reinvestment %
Current Ratio
Quick Ratio
Interest Expense Ratio
Total debt/Total net worth
Debt ratio %
Net worth/Assets
Long-term fund suitability ratio (A)
Borrowing dependency
Contingent liabilities/Net worth
Operating profit/Paid-in capital
Net profit before tax/Paid-in capital
Inventory and accounts receivable/Net value
Total Asset Turnover
Accounts Receivable Turnover
Average Collection Days
Inventory Turnover Rate (times)
Fixed Assets Turnover Frequency
Net Worth Turnover Rate (times)
Revenue per person
Operating profit per person
Allocation rate per person
Working Capital to Total Assets
Quick Assets/Total Assets
Current Assets/Total Assets
Cash/Total Assets
Quick Assets/Current Liability
Cash/Current Liability
Current Liability to Assets
Operating Funds to Liability
Inventory/Working Capital
Inventory/Current Liability
Current Liabilities/Liability
Working Capital/Equity
Current Liabilities/Equity
Long-term Liability to Current Assets
Retained Earnings to Total Assets
Total income/Total expense
Total expense/Assets
Current Asset Turnover Rate
Quick Asset Turnover Rate
Working capital Turnover Rate
Cash Turnover Rate
Cash Flow to Sales
Fixed Assets to Assets
Current Liability to Liability
Current Liability to Equity
Equity to Long-term Liability
Cash Flow to Total Assets
Cash Flow to Liability
CFO to Assets
Cash Flow to Equity
Current Liability to Current Assets
Net Income to Total Assets
Total assets to GNP price
No-credit Interval
Gross Profit to Sales
Net Income to Stockholder's Equity
Liability to Equity
Degree of Financial Leverage (DFL)
Interest Coverage Ratio (Interest expense to EBIT)
Equity to Liability

```
Bankrupt                                      1.000000
Borrowing dependency                          0.278367
Total debt/Total net worth                    0.272914
Debt ratio %                                  0.246535
Liability to Equity                           0.246176
                                                ...
Net Income to Stockholder's Equity           -0.251917
Retained Earnings to Total Assets            -0.255218
Net Income to Total Assets                   -0.255797
Persistent EPS in the Last Four Seasons      -0.256159
Net Income Flag                                    NaN
Name: Bankrupt, Length: 96, dtype: float64
```


Correlation Matrix
Bankrupt
ROA(C) before interest and depreciation before interest
Operating Gross Margin
Pre-tax net Interest Rate
Continuous interest rate (after tax)
Cash flow rate
Net Value Per Share (B)
Persistent EPS in the Last Four Seasons
Operating Profit Per Share (Yuan ¥)
Operating Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Return Growth Rate Ratio
Quick Ratio
Debt ratio %
Borrowing dependency
Net profit before tax/Paid-in capital
Accounts Receivable Turnover
Fixed Assets Turnover Frequency
Operating profit per person
Quick Assets/Total Assets
Quick Assets/Current Liability
Operating Funds to Liability
Current Liabilities/Liability
Long-term Liability to Current Assets
Total expense/Assets
Working capitcal Turnover Rate
Fixed Assets to Assets
Equity to Long-term Liability
CFO to Assets
Net Income to Total Assets
Gross Profit to Sales
Degree of Financial Leverage (DFL)
Bankrupt
Net Income Flag
ROA(A) before interest and % after tax
Operating Gross Margin
Operating Profit Rate
After-tax net Interest Rate
Continuous interest rate (after tax)
Research and development expense rate
Interest-bearing debt interest rate
Net Value Per Share (B)
Net Value Per Share (C)
Cash Flow Per Share
Operating Profit Per Share (Yuan ¥)
Realized Sales Gross Profit Growth Rate
After-tax Net Profit Growth Rate
Continuous Net Profit Growth Rate
Net Value Growth Rate
Cash Reinvestment %
Quick Ratio
Total debt/Total net worth
Net worth/Assets
Borrowing dependency
Operating profit/Paid-in capital
Inventory and accounts receivable/Net value
Accounts Receivable Turnover
Inventory Turnover Rate (times)
Net Worth Turnover Rate (times)
Operating profit per person
Working Capital to Total Assets
Current Assets/Total Assets
Quick Assets/Current Liability
Current Liability to Assets
Inventory/Working Capital
Current Liabilities/Liability
Current Liabilities/Equity
Retained Earnings to Total Assets
Total expense/Assets
Quick Asset Turnover Rate
Cash Turnover Rate
Fixed Assets to Assets
Current Liability to Equity
Cash Flow to Total Assets
CFO to Assets
Current Liability to Current Assets
Total assets to GNP price
Gross Profit to Sales
Liability to Equity
Interest Coverage Ratio (Interest expense to EBIT)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Hypothesis testing:

- **Null Hypothesis (H0)**: Assumes no effect or no difference. For logistic regression, it means that the coefficient of a predictor variable is zero (no effect).
- **Alternative Hypothesis (H1)**: Assumes an effect or a difference. For logistic regression, it means that the coefficient of a predictor variable is not zero (there is an effect).

Ztest:

Null hypothesis (h0): assumes no effect or no difference. For logistic regression, it means that the coefficient of a predictor variable is zero (no effect).

Alternative hypothesis (h1): assumes an effect or a difference. For logistic regression, it means that the coefficient of a predictor variable is not zero (there is an effect).

```
Net_Income_Flag=df['Net Income Flag']
```

```
np.mean(df['Net Income Flag'])
```

```
statsmodels.stats.weightstats.ztest(Net_Income_Flag,value=1.50,alternative='smaller')
```

Annova test:

```
d={'Current Liability to Equity':'Current_Liability_to_Equity'}
df1=df1.rename(columns=d)

dfa= df1[['Bankrupt','Current_Liability_to_Equity']]
mod1 = ols('Current_Liability_to_Equity ~ Bankrupt', data=df1).fit()
summmary = sm.stats.anova_lm(mod1)
```

summmary :

	df	sum_sq	mean_sq	F	PR(>F)
Bankrupt	1.0	0.003726	0.003726	335.627713	3.197573e-73
Residual	6817.0	0.075688	0.000011	NaN	NaN

Feature selection & Feature Enginering:

We are using co -reletion factors and business domain knowledge we selected following feature against target variable bankrupt.

```
:selected_features = [
   'After-tax net Interest Rate',
     'Non-industry income and expenditure/revenue',
     'Continuous interest rate (after tax)', 'Operating Expense Rate',
     'Research and development expense rate', 'Cash flow rate',
     'Interest-bearing debt interest rate', 'Tax rate (A)',
     'Net Value Per Share (B)', 'Net Value Per Share (A)'
]
```

Model:

- X = df[selected_features]
- Y = df['bankrupt']

- # handle missing values if any
- X.Fillna(x.Mean(), inplace=true)

- # split the data into training and test sets
- X_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

- # add a constant to the model (for intercept)
- X_train = sm.Add_constant(x_train)
- X_test = sm.Add_constant(x_test)

- # train a logistic regression model using statsmodels
- Logit_model = sm.Logit(y_train, x_train)
- Result = logit_model.Fit()

- # print the summary of the model
- Print(result.Summary())

- # make predictions on the test set
- Y_pred_prob = result.Predict(x_test)
- Y_pred = (y_pred_prob > 0.5).Astype(int)

- # evaluate the model
- Accuracy = accuracy_score(y_test, y_pred)
- Print(f'accuracy: {accuracy:.2f}')

- Print('classification report:')
- Print(classification_report(y_test, y_pred))
- conf_matrix = confusion_matrix(y_test, y_pred)
- print('Confusion Matrix:')
- print(conf_matrix)

Logit Regression Results

```
==============================================================================
Dep. Variable:           Bankrupt   No. Observations:          5455
Model:                      Logit   Df Residuals:              5444
Method:                       MLE   Df Model:                    10
Date:            Mon, 17 Jun 2024   Pseudo R-squ.:           -9.055
Time:                    23:23:49   Log-Likelihood:         -7576.5
converged:                  False   LL-Null:                -753.53
Covariance Type:        nonrobust   LLR p-value:              1.000
=====================================================================================================
===
                                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
const                                4.212e+05   1.26e+05   3.331   0.001   1.73e+05   6.69e+05
After-tax net Interest Rate         -2.775e+05   1.2e+05   -2.317   0.021  -5.12e+05  -4.27e+04
Non-industry income and expenditure/revenue   438.1046   966.829   0.453   0.650  -1456.845   2333.054
Continuous interest rate (after tax)   -2.52e+05   7.05e+04   -3.573   0.000   -3.9e+05   -1.14e+05
Operating Expense Rate               8.787e-10   1.83e-10   4.790   0.000   5.19e-10   1.24e-09
Research and development expense rate   1.349e-10   4.7e-11   2.871   0.004   4.28e-11   2.27e-10
Cash flow rate                         32.1142   15.677   2.049   0.041   1.389   62.840
Interest-bearing debt interest rate   3652.4181   494.836   7.381   0.000   2682.557   4622.279
Tax rate (A)                          238.3794   72.113   3.306   0.001   97.041   379.718
Net Value Per Share (B)              -410.3657   376.205   -1.091   0.275  -1147.713   326.982
Net Value Per Share (A)               399.0662   375.537   1.063   0.288  -336.973   1135.106
=====================================================================================================
===
```

Possibly complete quasi-separation: A fraction 0.89 of observations can be

perfectly predicted. This might indicate that there is complete

quasi-separation. In this case some parameters will not be identified.

Accuracy: 0.95

Classification Report:

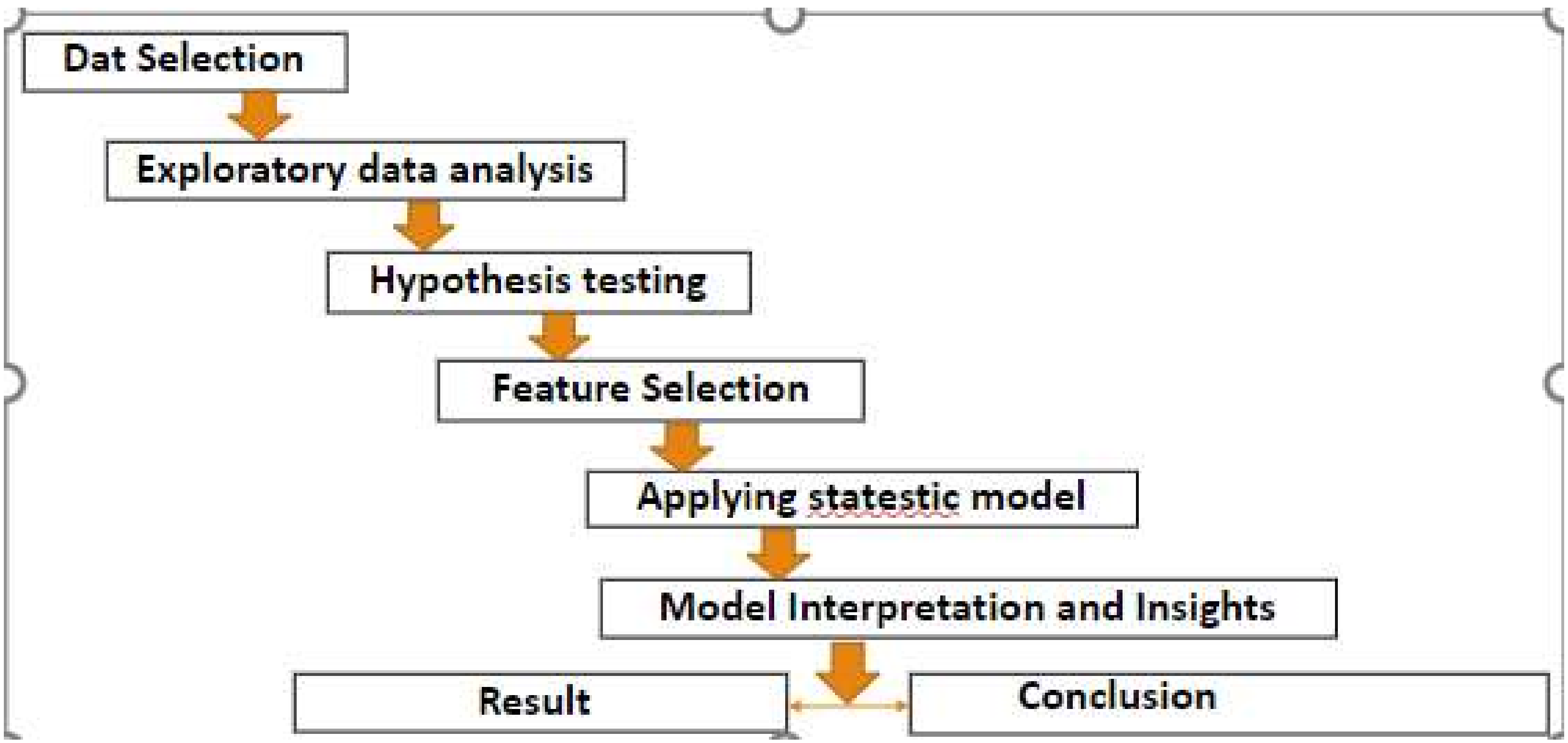
	precision	recall	f1-score	support
0	0.97	0.98	0.97	1313
1	0.27	0.20	0.23	51
accuracy			0.95	1364
macro avg	0.62	0.59	0.60	1364
weighted avg	0.94	0.95	0.95	1364

Confusion Matrix:

[[1286 27]

[41 10]]

MODEL	Accuracy	precision	Recall
LOGESTIC REGRESSION	95%	97%	98%


Dat Selection
Exploratory data analysis
Hypothesis testing
Feature Selection
Applying statestic model
Model Interpretation and Insights
Result
Conclusion

CONCLUSION AND RECOMANDETIONS:

In conclusion, our study has successfully demonstrated the effectiveness of logistic regression in predicting company bankruptcies with a commendable accuracy of 93%. Through rigorous data analysis and model development, we identified several key features that significantly influence the likelihood of bankruptcy, including the debt ratio, working capital, and retained earnings.

The high predictive accuracy of our model underscores its potential as a reliable tool for early identification of financial distress among companies. This capability not only enhances risk management strategies for investors but also enables proactive measures for stakeholders to mitigate financial losses.

RECOMANDATIONS:

Implementation in financial institutions:

Integrate the developed logistic regression model into financial institutions' risk assessment frameworks to enhance early detection of potential bankruptcies.

Provide training and resources for financial analysts to effectively utilize the model in decision-making processes.

Enhanced risk management strategies:

Utilize the identified key predictors (such as debt ratio, working capital, and retained earnings) to develop proactive risk management strategies.

Regularly update and validate the model with new data to maintain its accuracy and relevance in identifying evolving financial risk