technische universität
dortmund

BANKRUPTCY PREDICTION FOR TAIWANESE COMPANIES IN THE STOCK MARKET

DATA & AI IN ECONOMICS SS24

Search . . .

MEET THE TEAM

MOUHCINE LEYNE

AGYN ABDIMOMYN

TRUNG ANH HA

DUYEN TRUONG

Agenda

Search . . .

1. Introduction to our mission
2. Features selection & Exploratory data analysis
3. Data splliting + Modeling
4. Evaluation and results

Search . . .

1 . INTRODUCTION TO OUR MISSION

PREPARING, ANALYZING, AND LEVERAGING AN EXTENSIVE DATA SET TO PREDICT BANKRUPTCY

DEVELOPING CODE TO OPTIMALLY UTILIZE OUR DATA FOR TESTING VARIOUS MODELS

IDENTIFYING THE MOST EFFECTIVE MODEL FOR PREDICTING BANKRUPTCY THROUGH RESULT ANALYSIS

OUR DATASET

SAMPLE OF THE FEW FIRST COLUMNS AND ROWS OF OUR DATA

	Bankrupt?	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	...	Net Income to Total Assets	Total assets to GNP price	No-credit Interval
0	1	0.370594	0.424389	0.40575	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	...	0.716845	0.009219	0.622879
1	1	0.464291	0.538214	0.51673	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	...	0.795297	0.008323	0.623652

2 rows × 96 columns

> **BORDERLINE-SMOTE BOOSTS CLASSIFICATION ACCURACY BY CREATING SYNTHETIC SAMPLES NEAR THE DECISION BOUNDARY**

- We prepared and cleaned the data by checking for duplicates and handling missing values.
- We use Borderline-SMOTE to oversample bankruptcy companies and standardize the dataset (mean = 0, std = 1) to enhance model training and reliability.

2.1 FEATURES SELECTION

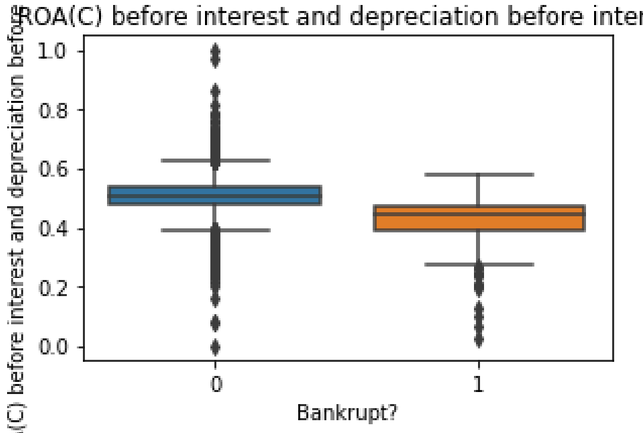
*SELECTKBEST IS A FEATURE SELECTION METHOD IN MACHINE LEARNING THAT SELECTS THE TOP K FEATURES BASED ON STATISTICAL TESTS. IT EVALUATES EACH FEATURE INDIVIDUALLY FOR ITS RELEVANCE TO THE TARGET VARIABLE, ALLOWING US TO FOCUS ON THE MOST SIGNIFICANT FEATURES AND REDUCE DIMENSIONALITY FOR IMPROVED MODEL PERFORMANCE AND EFFICIENCY.

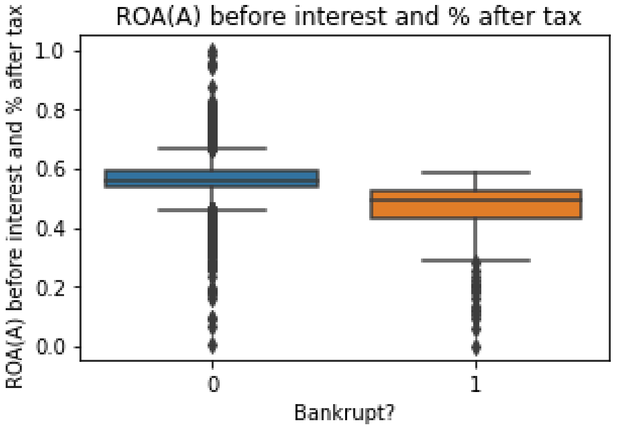
THE 18 FEATURES WE USED FOR EXPLORATORY DATA ANALYSIS

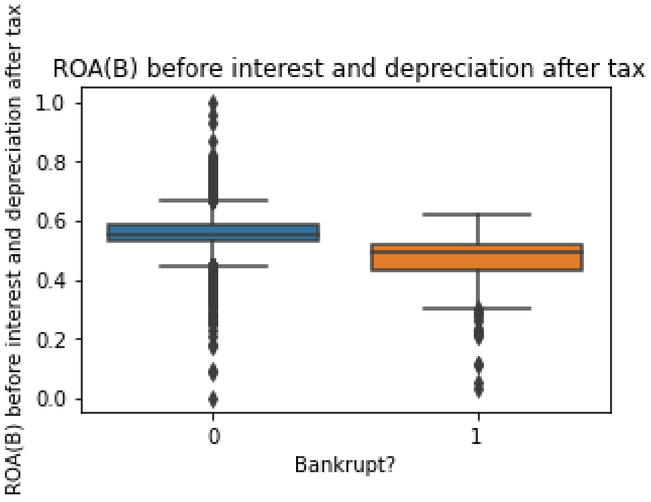
1. ROA(C) before interest and depreciation before interest
2. ROA(A) before interest and % after tax
3. ROA(B) before interest and depreciation after tax
4. Net Value Per Share (B)
5. Net Value Per Share (A)
6. Net Value Per Share (C)
7. Persistent EPS in the Last Four Seasons
8. Operating Profit Per Share (Yuan ¥)
9. Per Share Net profit before tax (Yuan ¥)
10. Debt ratio %
11. Net worth/Assets
12. Operating profit/Paid-in capital.
13. Net profit before tax/Paid-in capital
14. Working Capital to Total Assets
15. Current Liability to Assets
16. Retained Earnings to Total Assets
17. Current Liability to Current Assets
18. Net Income to Total Assets

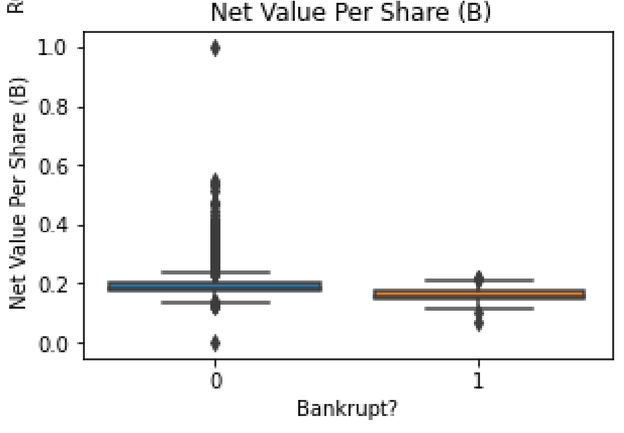
2.2 EXPLORATORY DATA ANALYSIS (EDA)

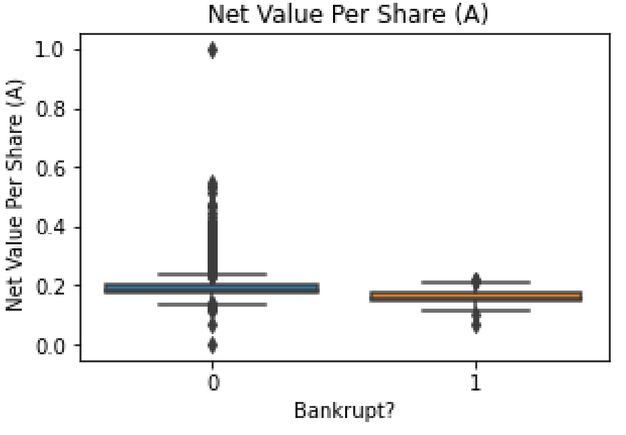
> TO CONDUCT EXPLORATORY DATA ANALYSIS (EDA), WE GENERATED BOX PLOTS FOR EACH FEATURE, HELPING US VISUALIZE AND UNDERSTAND THE DISTRIBUTION OF EACH FEATURE IN RELATION TO BANKRUPTCY STATUS.

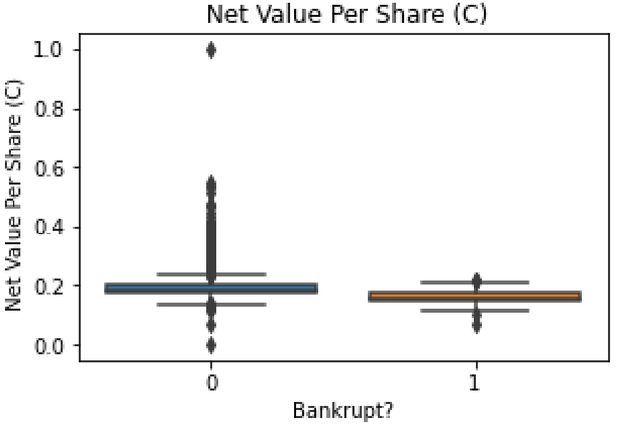

ROA(C) before interest and depreciation before interest
ROA(C) before interest and depreciation before interest
Bankrupt?

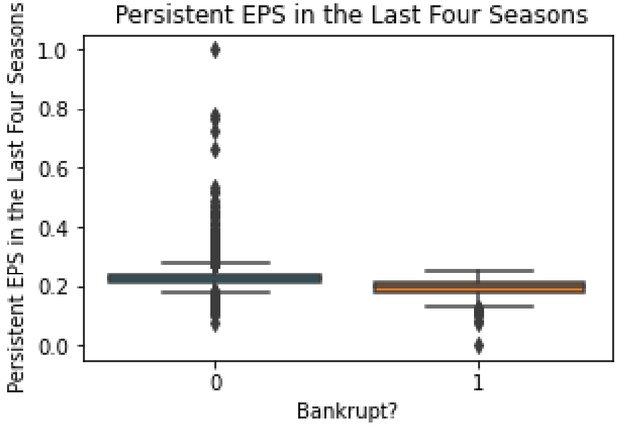

ROA(A) before interest and % after tax
ROA(A) before interest and % after tax
Bankrupt?

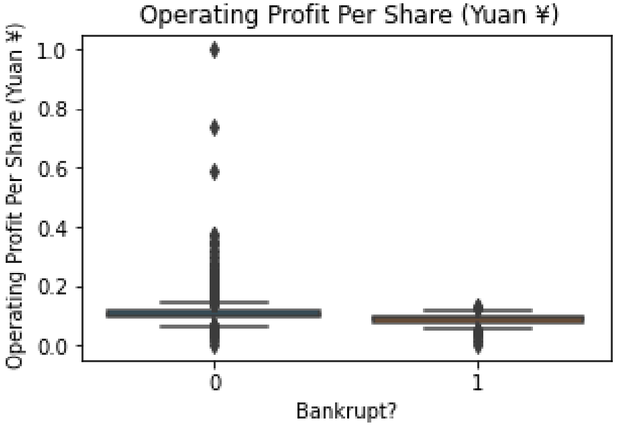

ROA(B) before interest and depreciation after tax
ROA(B) before interest and depreciation after tax
Bankrupt?

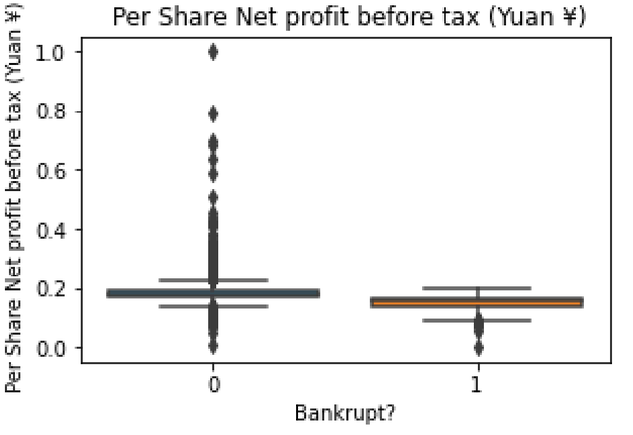

Net Value Per Share (B)
Net Value Per Share (B)
Bankrupt?

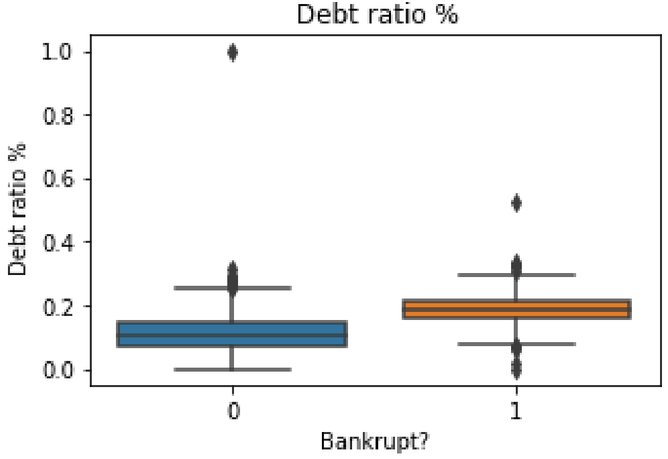

Net Value Per Share (A)
Net Value Per Share (A)
Bankrupt?

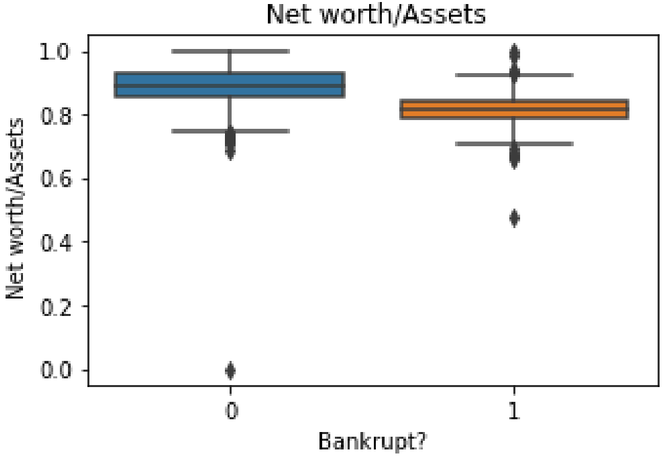

Net Value Per Share (C)
Net Value Per Share (C)
Bankrupt?

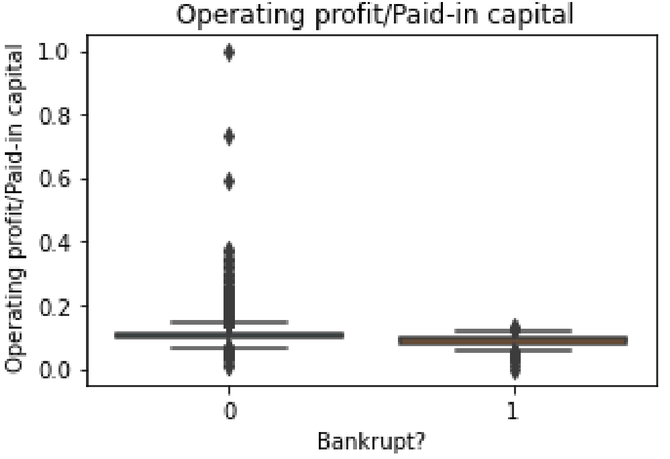

Persistent EPS in the Last Four Seasons
Persistent EPS in the Last Four Seasons
Bankrupt?

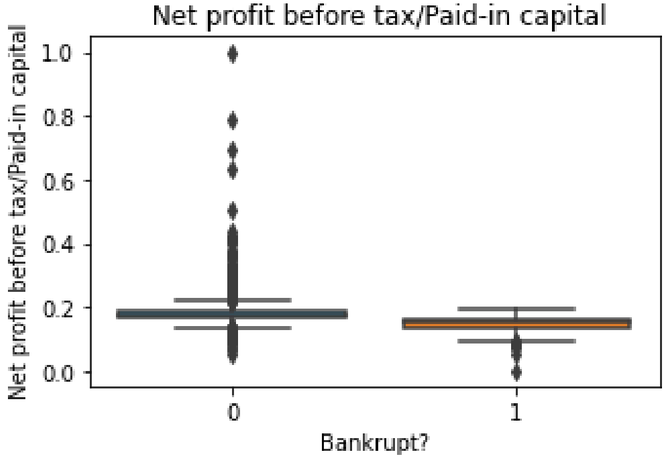

Operating Profit Per Share (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Bankrupt?

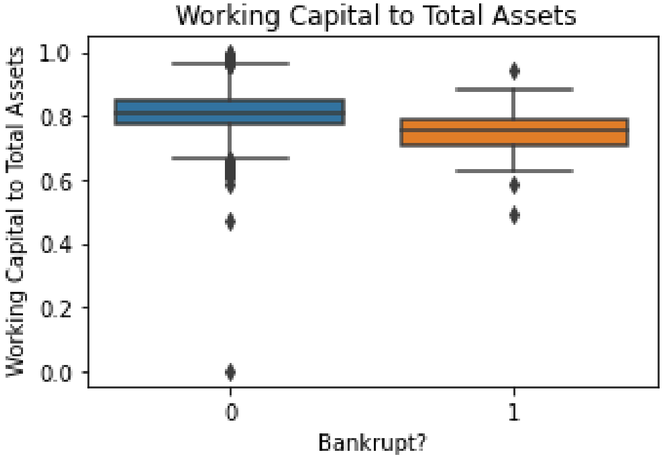

Per Share Net profit before tax (Yuan ¥)
Per Share Net profit before tax (Yuan ¥)
Bankrupt?

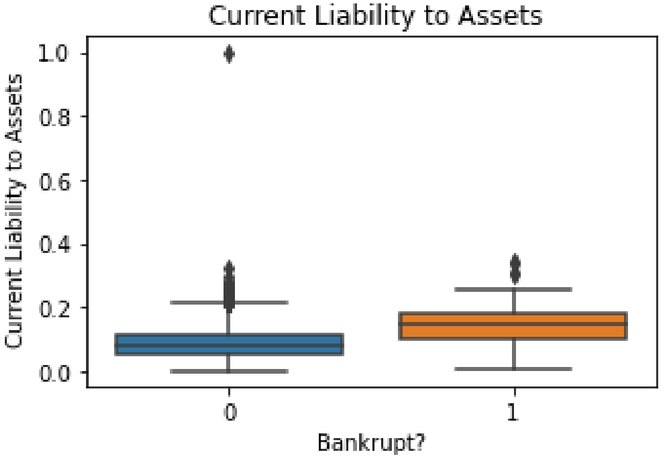

Debt ratio %
Debt ratio %
Bankrupt?

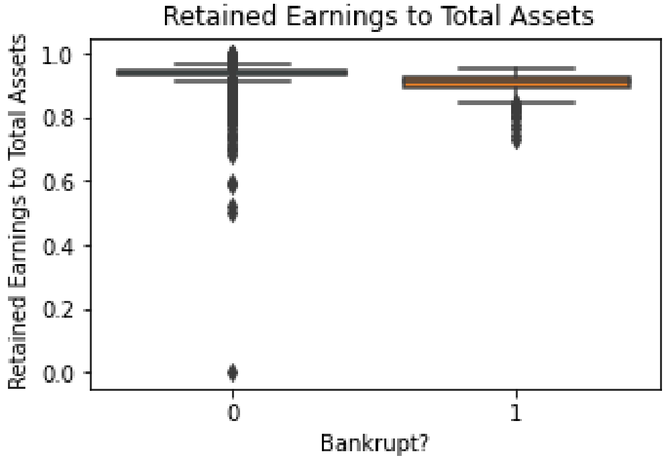

Net worth/Assets
Net worth/Assets
Bankrupt?

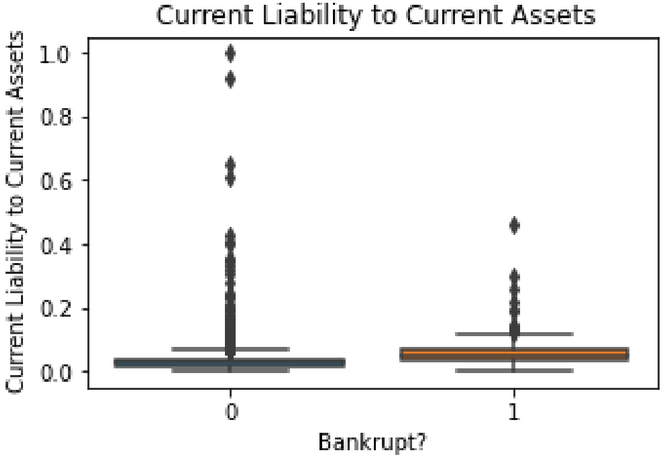

Operating profit/Paid-in capital
Operating profit/Paid-in capital
Bankrupt?

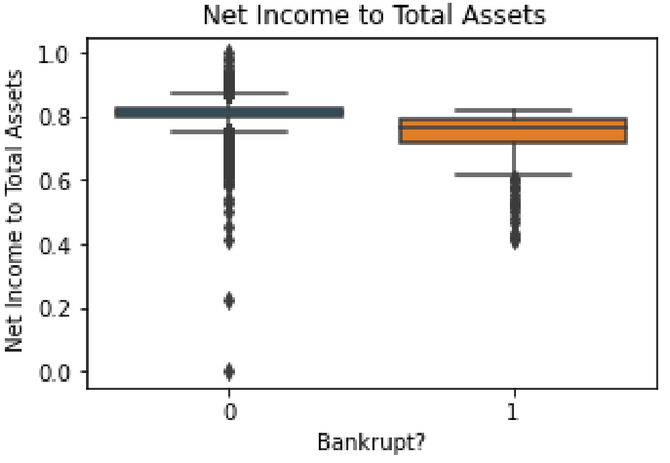

Net profit before tax/Paid-in capital
Net profit before tax/Paid-in capital
Bankrupt?


Working Capital to Total Assets
Working Capital to Total Assets
Bankrupt?

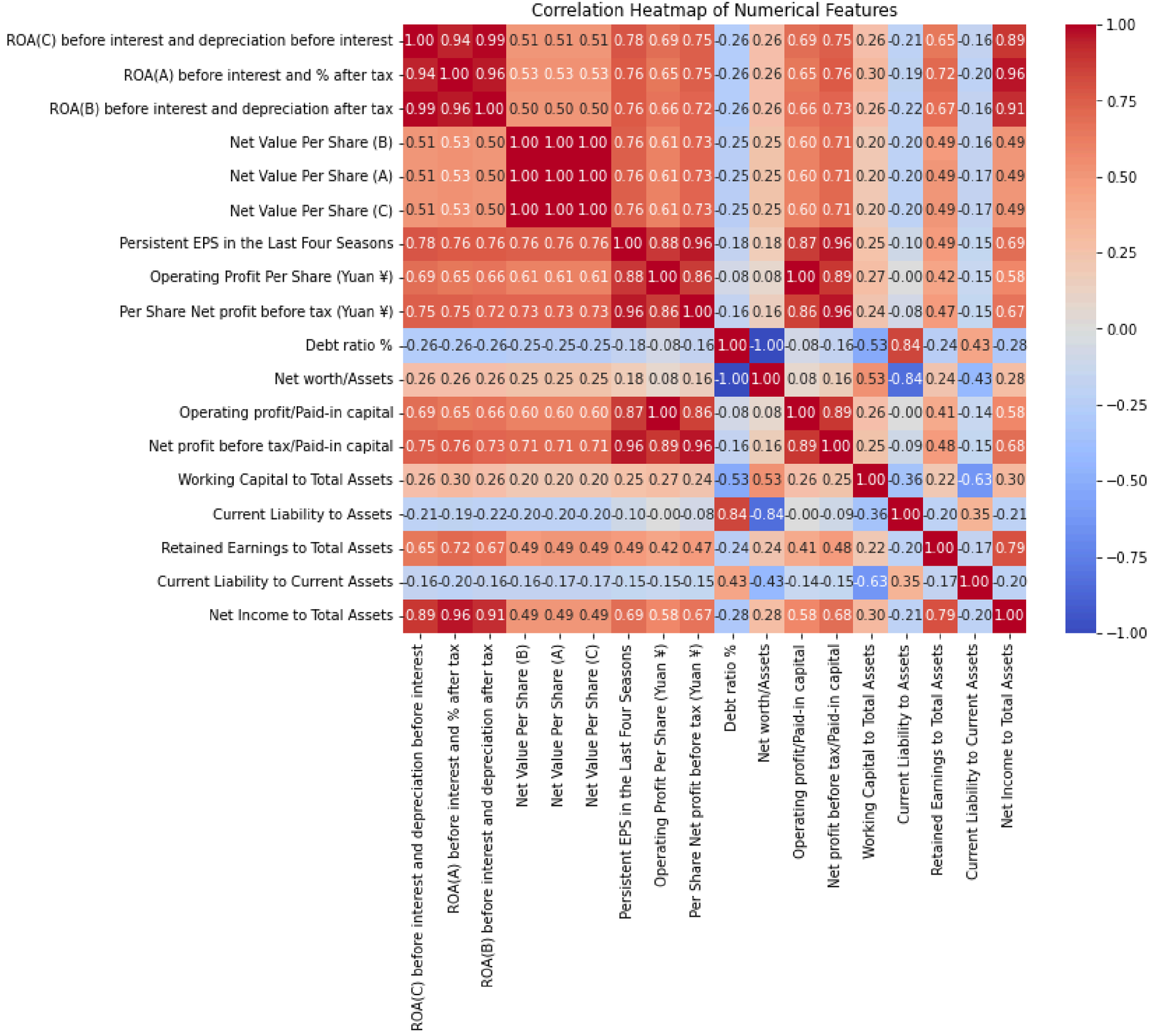

Current Liability to Assets
Current Liability to Assets
Bankrupt?


Retained Earnings to Total Assets
Retained Earnings to Total Assets
Bankrupt?


Current Liability to Current Assets
Current Liability to Current Assets
Bankrupt?


Net Income to Total Assets
Net Income to Total Assets
Bankrupt?

CORRELATION HEATMAP OF NUMERICAL FEATURES



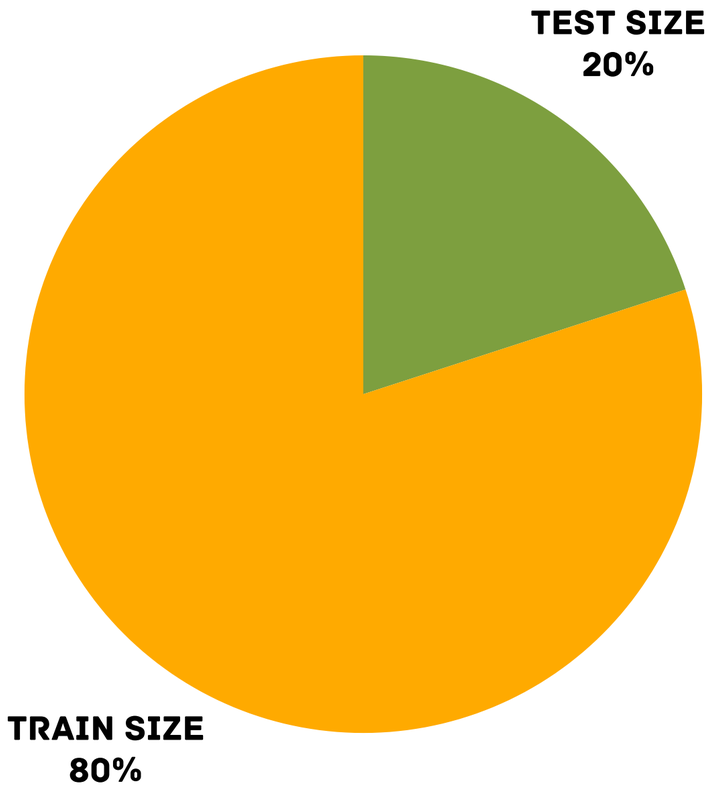
3.1 DATA SPLITTING

Method Used:

- Library: **sklearn.model_selectionRandom**
- Function: **train_test_split**

Parameters:

- Test Size: 20% (for testing)
- Train Size: 80% (for training)



EVALUATION METRICS

- Accuracy: - Ratio of correct predictions to total predictions
- F1 Score: - Harmonic mean of precision and recall
- Why These Metrics: - Accuracy provides a general performance measure- F1 Score balances precision and recall, important for imbalanced data

3.2 MODELS WE TESTED

- Basic Logistic Classifier
- Stochastic Gradient Descent Classifier
- Random Forest
- Decision Tree
- Neural Network
- Support Vector Machine
- K-Nearest Neighbors Classifier
- Extreme Gradient Boosting Classifier

4.1 EVALUATION AND RESULTS

	Model	Accuracy	Precision	Recall	F1	ROC-AUC
0	RandomizedSearchCV_RandomForestClassifier	0.970076	0.970695	0.970293	0.970073	0.970293
1	RandomizedSearchCV_GradientBoostingClassifier	0.968939	0.969695	0.969177	0.968935	0.969177
2	XGBoostingClassifier	0.966288	0.967128	0.966537	0.966282	0.966537
3	RandomizedSearchCV_KNeighborsClassifier	0.963636	0.965226	0.963970	0.963620	0.963970
4	RandomForestClassifier	0.962879	0.963571	0.963108	0.962874	0.963108
5	VotingClassifier	0.956061	0.957246	0.956353	0.956047	0.956353
6	KNeighborsClassifier	0.937879	0.941884	0.938402	0.937782	0.938402
7	RandomizedSearchCV_DecisionTreeClassifier	0.934470	0.934510	0.934547	0.934469	0.934547
8	DecisionTreeClassifier	0.933712	0.933852	0.933829	0.933712	0.933829
9	GradientBoostingClassifier	0.927652	0.928496	0.927905	0.927637	0.927905
10	XGBRFClassifier	0.926894	0.927424	0.927099	0.926887	0.927099
11	RandomizedSearchCV_SVC	0.925000	0.927067	0.925385	0.924947	0.925385
12	Support Vector Machine	0.908333	0.909295	0.908604	0.908310	0.908604
13	RandomizedSearchCV_XGBClassifier	0.905303	0.905476	0.905429	0.905303	0.905429
14	LogisticRegression	0.904545	0.904807	0.904695	0.904543	0.904695
15	RandomizedSearchCV_LogisticRegression	0.897727	0.898670	0.897997	0.897701	0.897997
16	SGDClassifier	0.892424	0.892424	0.892468	0.892421	0.892468

4.1 EVALUATION AND RESULTS

```
[ ]  train_and_evaluate_model(SGDClassifier(), 'SGDClassifier')
```

```
[[1185  149]
 [ 135 1171]]
Number of wrong classifiers: 284
Accuracy score of 'SGDClassifier': 89.2424%
```

Stochastic Gradient Descent Classifier
performed the worst producing Accuracy score of:
89.2424%

4.1 EVALUATION AND RESULTS

	Model	Accuracy	Precision	Recall	F1	ROC-AUC
0	RandomizedSearchCV_RandomForestClassifier	0.970076	0.970695	0.970293	0.970073	0.970293
1	RandomizedSearchCV_GradientBoostingClassifier	0.968939	0.969695	0.969177	0.968935	0.969177
2	XGBoostingClassifier	0.966288	0.967128	0.966537	0.966282	0.966537
3	RandomizedSearchCV_KNeighborsClassifier	0.963636	0.965226	0.963970	0.963620	0.963970
4	RandomForestClassifier	0.962879	0.963571	0.963108	0.962874	0.963108
5	VotingClassifier	0.956061	0.957246	0.956353	0.956047	0.956353

RandomizedSearchCV_RandomForestClassifier produced the most accurate results as a performing model, even after completing hyperparameter tuning for all the models producing Accuracy score of:
97.0076%

WHY RANDOMIZEDSEARCHCV WITH RANDOMFORESTCLASSIFIER PERFORMED BEST?

- Efficient Exploration: Covers a wide range of hyperparameters.
- Performance Improvement: Optimizes key parameters.
- Computational Efficiency: More efficient than grid search.

CONCLUSION AND FUTURE WORK

The RandomizedSearchCV with RandomForestClassifier emerged as the best-performing model for bankruptcy prediction in Taiwan due to several key factors:

- **Ensemble Learning**
- **Feature Importance**
- **Robustness and Flexibility**

CONCLUSION AND FUTURE WORK

Practical Implications:

- **The ability to accurately predict bankruptcy has significant implications for financial institutions and businesses:**
  - **Risk Management: Financial institutions can better manage credit risk and make informed lending decisions.**
  - **Proactive Measures: Companies can take proactive measures to address financial distress and potentially avoid bankruptcy.**
  - **Strategic Planning: Businesses can use these insights for strategic planning and improving their financial health.**

Search . . .

THANK YOU

Q&A