Group 2 | CSCI 111

Going Bust:

A Machine Learning Model for Corporate Bankruptcy Prediction

Date: Dec 11, 2023

Prepared by: Lanz Fermin, Robert Leung, & Francis Velaque

Outline

- Background and Objectives
- Dataset
- Methodology
- Results
- Insights
- Conclusion and Recommendations

Introduction

In any economic environment, financial stability is a pillar of sustained growth.

Bankruptcy is a judicial recognition that a company can no **longer repay its debts.** Once a company is declared bankrupt, it typically **liquidates its assets**, **closes operations,** or **enters a loan restructuring agreement.**

Predicting corporate bankruptcies, especially from early warning signs, is a notoriously difficult task. **Thus, there is a pressing need to develop robust automated prediction systems for bankruptcy.**

Objective:

Develop and train **two supervised classifiers** on the Taiwanese Bankruptcy Prediction Dataset

- Achieve over 90% test accuracy for both classifiers

Dataset

Taiwanese Bankruptcy Prediction

6819
Instances

6599	**220**
Non Bankrupt	**Bankrupt**

Taiwanese Bankruptcy Prediction

- Uploaded in 2020 under a CC BY 4.0 License (for research use)
- 96 features, all financial indicators
- All feature values pre normalized (range from 0 to 1)
- Minimal need for cleaning, no missing/empty/problematic values

Taiwanese Bankruptcy Prediction

-1 Bankrupt?
0 ROA(C) before interest and depreciation before interest
1 ROA(A) before interest and % after tax
2 ROA(B) before interest and depreciation after tax
3 Operating Gross Margin
4 Realized Sales Gross Margin
5 Operating Profit Rate
6 Pre-tax net Interest Rate
7 After-tax net Interest Rate
8 Non-industry income and expenditure/revenue
9 Continuous interest rate (after tax)
10 Operating Expense Rate
11 Research and development expense rate
12 Cash flow rate
13 Interest-bearing debt interest rate
14 Tax rate (A)
15 Net Value Per Share (B)
16 Net Value Per Share (A)
17 Net Value Per Share (C)
18 Persistent EPS in the Last Four Seasons
19 Cash Flow Per Share
20 Revenue Per Share (Yuan ¥)
21 Operating Profit Per Share (Yuan ¥)
22 Per Share Net profit before tax (Yuan ¥)
23 Realized Sales Gross Profit Growth Rate
24 Operating Profit Growth Rate
25 After-tax Net Profit Growth Rate
26 Regular Net Profit Growth Rate
27 Continuous Net Profit Growth Rate
28 Total Asset Growth Rate
29 Net Value Growth Rate
30 Total Asset Return Growth Rate Ratio
31 Cash Reinvestment %
32 Current Ratio
33 Quick Ratio
34 Interest Expense Ratio
35 Total debt/Total net worth
36 Debt ratio %
37 Net worth/Assets
38 Long-term fund suitability ratio (A)
39 Borrowing dependency
40 Contingent liabilities/Net worth
41 Operating profit/Paid-in capital
42 Net profit before tax/Paid-in capital
43 Inventory and accounts receivable/Net value
44 Total Asset Turnover
45 Accounts Receivable Turnover
46 Average Collection Days
47 Inventory Turnover Rate (times)
48 Fixed Assets Turnover Frequency
49 Net Worth Turnover Rate (times)
50 Revenue per person
51 Operating profit per person
52 Allocation rate per person
53 Working Capital to Total Assets
54 Quick Assets/Total Assets
55 Current Assets/Total Assets
56 Cash/Total Assets
57 Quick Assets/Current Liability
58 Cash/Current Liability
59 Current Liability to Assets
60 Operating Funds to Liability
61 Inventory/Working Capital
62 Inventory/Current Liability
63 Current Liabilities/Liability
64 Working Capital/Equity
65 Current Liabilities/Equity
66 Long-term Liability to Current Assets
67 Retained Earnings to Total Assets
68 Total income/Total expense
69 Total expense/Assets
70 Current Asset Turnover Rate
71 Quick Asset Turnover Rate
72 Working capitcal Turnover Rate
73 Cash Turnover Rate
74 Cash Flow to Sales
75 Fixed Assets to Assets
76 Current Liability to Liability
77 Current Liability to Equity
78 Equity to Long-term Liability
79 Cash Flow to Total Assets
80 Cash Flow to Liability
81 CFO to Assets
82 Cash Flow to Equity
83 Current Liability to Current Assets
84 Liability-Assets Flag
85 Net Income to Total Assets
86 Total assets to GNP price
87 No-credit Interval
88 Gross Profit to Sales
89 Net Income to Stockholder's Equity
90 Liability to Equity
91 Degree of Financial Leverage (DFL)
92 Interest Coverage Ratio (Interest expense to EBIT)
93 Net Income Flag
94 Equity to Liability

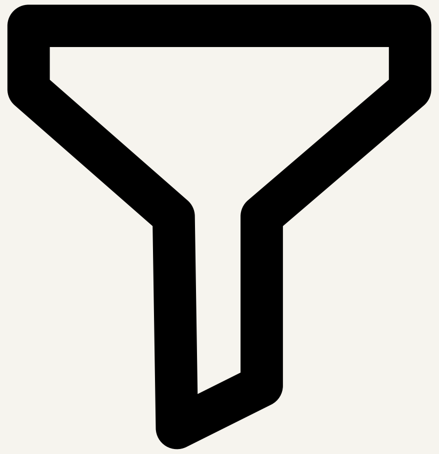
Methodology

I. Feature Selection & Engineering
II. Model Development
III. Model Evaluation

I. Feature Selection and Engineering

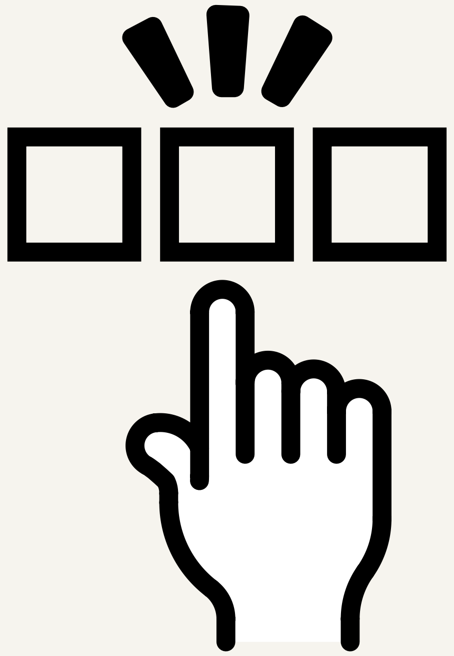
Problem: Too many features delays training times and may lead to less accurate results/failed convergence (ran into this problem for Linear Kernel SVM)

Solution: 3 Phase Preprocessing



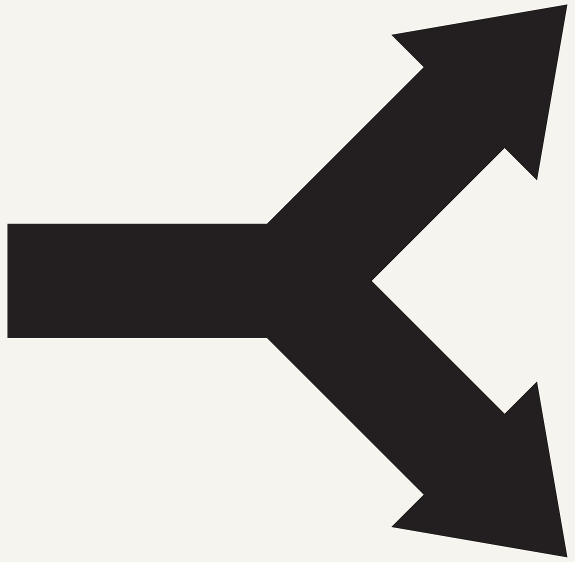
1. Filtering

Highly correlated feature pairs were dropped out, leaving **79 features**



2. Selection

The **7 best fit** features were chosen using SKLearn's in-built feature selection package



3. Split

Data was split into test and train sets in a **7:3 ratio ~ 4773 training instances and 2046 test instances**

I. Feature Selection and Engineering

Features Selected:

- **Operating profit per share**
- **After Tax Net Profit Growth Rate**
- **Net Assets**
- **Borrowing Dependency**
- **Inventory Turnover Rate**
- **Working Capital**
- **Working Capital Turnover Rate**

II. Model Development

Two models were developed for this project: a **Support Vector Classifier (SVC)** and a **Random Forest Classifier (RF)**

II. Model Development - **Support Vector Classifier**

- One of the most common and reliable classification algorithms
- Works by plotting instances in a multi-dimensional feature space and finding a dividing “hyperplane” to sort them
- **Parameters for this project:**
  - **Polynomial kernel** - used when data is not linearly separable (earlier attempts with Linear kernel would often fail to converge or have poor accuracy)
  - **Balance class weight** - since instances of bankrupt and non bankrupt companies are unequal, we adjust model parameters to consider balance
  - **Scaled gamma** - weights features equally

II. Model Development - **Random Forest**

- Fits a number of decision trees on data based on available features then takes aggregate results
- Commonly used in financial classification tasks such as credit card fraud detection, risk assessment, and options pricing determination
- Default sklearn parameters used
  - n_estimators = 100 :: 100 forests taken in aggregate

III. Model Evaluation

A series of metrics were used to evaluate model performance

01. Test Accuracy

Raw accuracy of models evaluated on testing set, given by number of correct predictions over total number of predictions

02. Confusion Matrix

Plot of model predictions vs. actual labels per category. Used to compute:

- **Precision = TP/(TP+FP)**
- **Recall = TP/(TP+FN)**
- **F1 = 2*(Precision*Recall)/(Precision+Recall)**

03. Cross-fold Validation

Measures test accuracy over a number of resamples of the data to more accurately assess performance on unseen instances. In this project, we used **5-fold validation**

Results

Test Accuracy | Confusion Matrix | Cross-Fold

Test Accuracy

91.45%

SVC Test Accuracy

97.07%

RF Accuracy on Test Set

While both models performed well (over 90%), on the surface it seems like RF performed better. To validate this, we need to analyze deeper metrics

5-Cross Fold Validation

91%

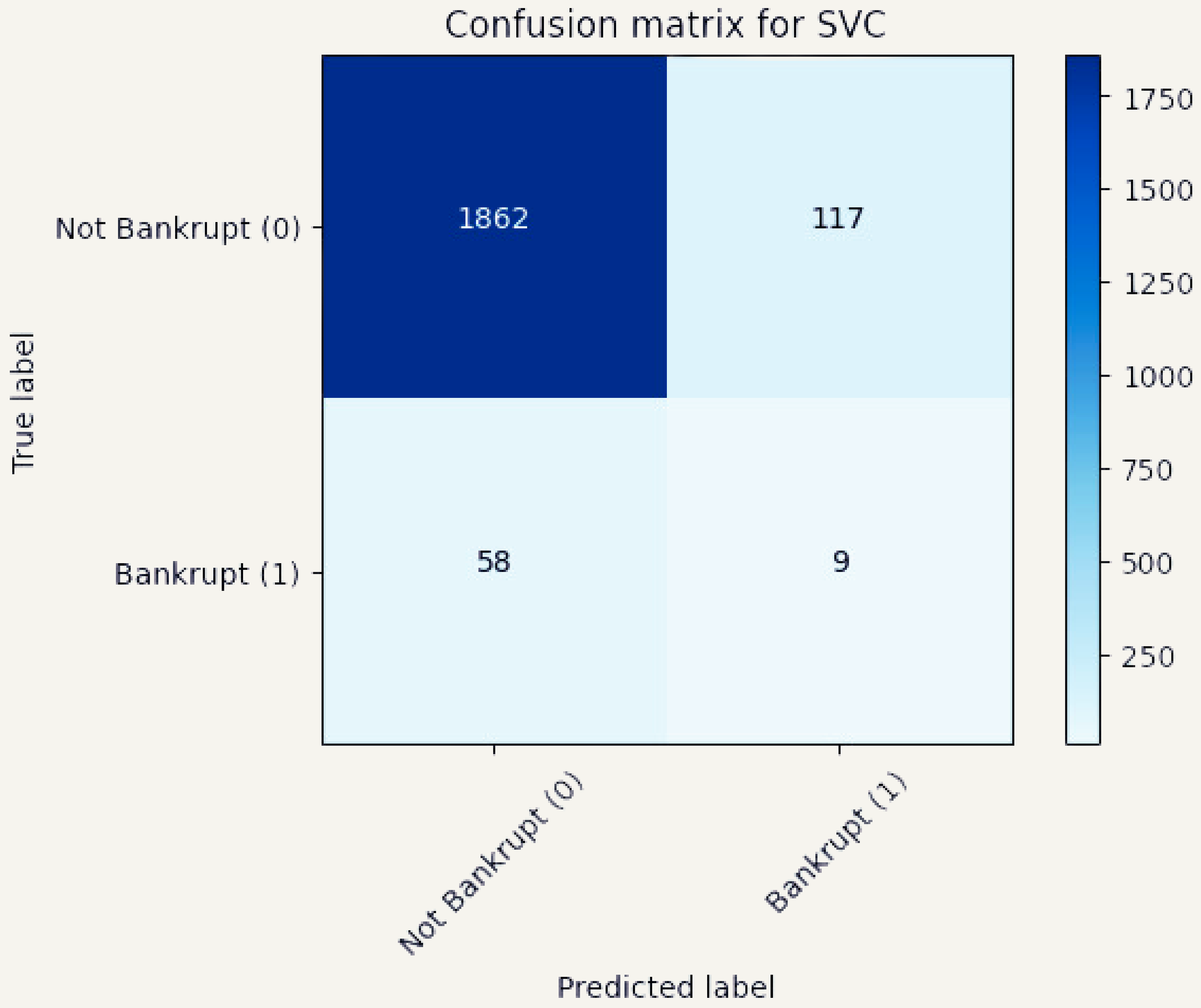
SVC Accuracy
w/ 2% std. dev

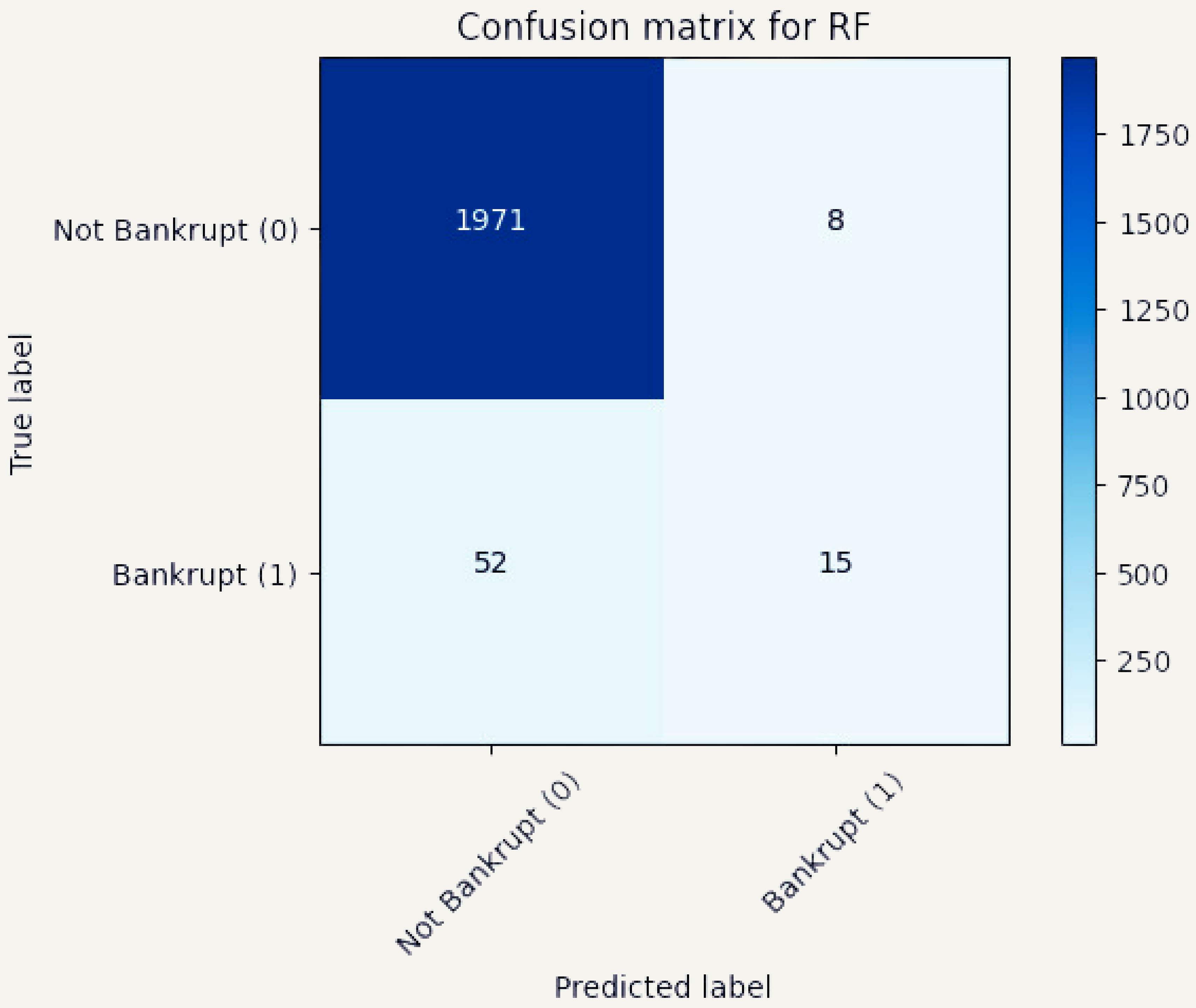
97%

RF Accuracy
w/ 0 std. dev.

The results of the 5-Fold Cross validation support the RF model being more accurate than the SVC, suggesting that the data (like other financial datasets) is more suited to RF classification.

Non-normalized Confusion Matrices





Precision-Recall-F1

SVC

	Non Bankrupt	Bankrupt
Precision	0.97	0.07
Recall	0.94	0.13
F1	0.96	0.09

RF

	Non Bankrupt	Bankrupt
Precision	0.97	0.65
Recall	1	0.22
F1	0.99	0.33

Both models performed very well at correctly identifying non bankrupt companies, but struggled to identify bankrupt ones, with the RF performing notably better. **But why?**

Insights

Interpreting Results

Possible reasons why models struggled to classify bankrupt companies:

- Imbalance in dataset. Comparatively few instances of bankrupt companies to base predictions on. Both models were **underpredicting** bankruptcy
- Soon-to-be-bankrupt companies are very very difficult to distinguish from non-bankrupt ones (if shareholders knew, they would have already sold their stock!)
  - Financial and economic conditions can change rapidly
- Lack of time series data
  - Company financials need to be looked at over a period of time rather than just one slice as in the bankruptcy dataset

Interpreting Results

Why RF Outperformed

- The ensemble nature of Random Forest may provide better resilience against overfitting and contribute to its superior performance.

Interpretability vs. Performance

- Understanding how each tree contributes to the performance can be challenging making it harder to pinpoint the exact features and their interactions influencing the prediction

Potential Overfitting

- SVC's relatively higher recall but lower precision may suggest a propensity for overfitting, capturing more bankrupt instances but at the cost of increased false positives.

Conclusion

Conclusion

The two models developed were able to **successfully** classify bankrupt vs. non-bankrupt companies at **above 90% test accuracy (both raw and cross fold validated).**

Across all evaluation metrics used, the **RF classifier performed better than the SVC**, however both struggled with **underpredicting** bankrupt instances.

Recommendations

Addressing Data Imbalances

- Wang and Liu (2021) presents a three-step framework.
- Consider undersampling techniques to address data imbalance.
- Mix and match unersampling techniques and machine learning models to achieve optimality.

Time Series Analysis of Features

- Relevant trends and patterns may be revealed when considering time series data.
- Temporal dynamics between financial and economic conditions can be observed.

Feature Importance, Ensemble Techniques

- Demystify the complexity of random forest decision-making process.
- Strike a balance between model interpretability and performance.
- Leverage domain knowledge as a guide for feature selection and engineering.

References

References

- Brownlee, J. (2018, May 22). A Gentle Introduction to k-fold Cross-Validation. MachineLearningMastery.Com. https://machinelearningmastery.com/k-fold-cross-validation/
- IBM. (2023). What is Random Forest? | IBM. https://www.ibm.com/topics/random-forest
- Shung, K. P. (2020, April 10). Accuracy, Precision, Recall or F1? Medium. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
- Sklearn.svm.SVC — scikit-learn 1.3.2 documentation. (n.d.). Retrieved December 9, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- Soriano, F. (2020). Company Bankruptcy Prediction. Kaggle. https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction
- Wang, H., & Liu, X. (2021). Undersampling bankruptcy prediction: Taiwan bankruptcy data. *PLOS ONE, 16*(7), e0254030–e0254030. https://doi.org/10.1371/journal.pone.0254030