MGT 6203 Final Report

Harnessing Financial Ratios to Predict Corporate Bankruptcy

Team 119

Bhargava, Akhil

Ly, Trang

Troxler, Cyrill

Savage, Carsten

TABLE OF CONTENTS

BACKGROUND

Bankruptcy filings can have a substantial impact on the stock prices of public companies, with successfully avoiding bankruptcy leading to significant increases in share value. Moreover, investors have the potential to earn substantial profits by investing in companies that emerge from bankruptcy. A wide array of financial ratios offers insights into a company's financial health, collectively providing a comprehensive view of its market positioning.

Accurately predicting corporate bankruptcies offers numerous advantages for corporations, governments, and investors. Corporations can leverage bankruptcy analyses to assess industry competition, investors can utilize insights to trade equities or options, and governments can employ models to assess the likelihood of corporate collapse, especially for companies integral to the economy, enabling timely interventions.

Prior literature predominantly focuses on machine learning approaches to predict corporate bankruptcy. In their work, Liang et al. note that machine learning bankruptcy prediction models have typically outperformed statistical techniques in academia (Liang et al. 561). Liang et al. utilize five different machine learning algorithms to predict corporate bankruptcies based on financial ratio and corporate governance data (Liang et al. 561). This paper builds on these insights by implementing principal component analysis to facilitate feature selection in datasets with high quantities of features and deploys logistic regression and k-nearest neighbors (KNN) models for bankruptcy classification.

OBJECTIVE

This paper's primary objective is to assess the accuracy of bankruptcy prediction using logistic regression and KNN models. These models have been selected based on their capacity to accurately make classifications while avoiding underfitting by implementing cost functions.

Moreover, the paper recognizes the challenge posed by imbalanced bankruptcy data, where non-bankrupt observations typically outnumber bankrupt ones. This imbalance may skew models to predict all observations as non-bankrupt to maximize classification accuracy. Consequently, the paper emphasizes the importance of implementing measures to mitigate underfitting in these models.

In addition to assessing classification accuracy, the secondary objective of this research is to explore optimal feature selection methods for datasets containing numerous independent variables. Many of these variables may exhibit high correlation and can be categorized into distinct groups based on financial ratios. The paper investigates whether feature selection yields better results when performed on grouped or ungrouped variables, aiming to enhance the predictive performance of the models.

Furthermore, this study delves into the potential benefits of employing a strategy involving the construction of multiple models, each utilizing distinct groups of features. By leveraging this approach, the research aims to investigate whether amalgamating these individual models can lead to more robust and universally applicable forecasts. This methodology seeks to exploit the diverse insights provided by different sets of features, potentially enhancing the predictive power of the overall forecasting framework. Through this analysis, the paper endeavors to uncover whether the combination of models utilizing varied feature groups can yield more significant and reliable forecasts, thereby contributing to the advancement of bankruptcy prediction methodologies.

INITIAL HYPOTHESIS

The implementation of logistic regression and KNN models will demonstrate varying levels of accuracy in predicting corporate bankruptcy. These models, chosen for their classification capabilities and mitigation of underfitting through cost functions, are expected to perform differently due to the nature of the data and the inherent strengths and weaknesses of each algorithm.

Furthermore, it is hypothesized that due to the imbalance in the bankruptcy data, in which the non-bankrupt observations are the majority, the models will exhibit potential class bias issues. Addressing this imbalance through appropriate measures is expected to be crucial for improving the models' predictive performance.

Regarding feature selection, the hypothesis suggests that exploring optimal methods in datasets with numerous independent variables, including highly correlated ones, will enhance the models' accuracy. Whether feature selection is more effective when performed on grouped or ungrouped variables remains an open question, with potential implications for refining the models' predictive capabilities.

Additionally, it is hypothesized that diverse types of financial ratios, such as liquidity ratios, profitability ratios, and solvency ratios, may contribute differently to the prediction of corporate bankruptcy. Testing these diverse types of financial ratios will provide insights into which categories are most effective in accurately predicting bankruptcy in the future. This analysis aims to identify the key financial metrics that significantly influence bankruptcy prediction models, thereby enhancing their predictive accuracy and robustness.

METHODOLOGY

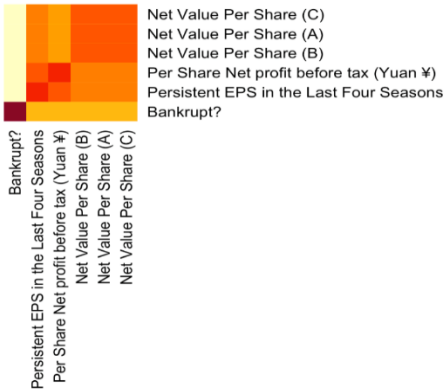
This paper uses a two-pronged approach. The bankruptcy data contains a wide array of variables, each of which can be grouped into financial ratio (FR) buckets. Consequently, the issue of whether the data should be ungrouped or grouped for the feature selection and model training and testing becomes a focal point of the analysis. The columns are first grouped into FR buckets, and subset dataframes are created for each FR group.

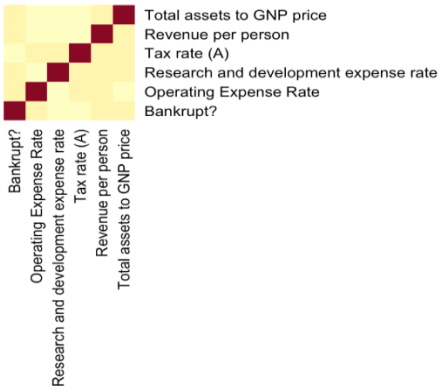
Correlation matrices are created and analyzed for the entire dataframe (ungrouped) and each FR subset dataframe. Then, principal component analysis (PCA) is conducted on the entire dataframe and FR subset dataframes. The resulting plots reveal that the first N principal components capture the most variance, while the next K principal components exhibit a steep drop off in variance. The number of principal components to retain is gauged by assessing the principal component plots. After selecting the first N principal components, the variables and their loading values are sorted by loadings, and the top 40% of unique variables across principal components with the highest loadings are selected from each FR group. This facilitates the identification of the variables whose variance can be explained by the top N principal components.

After these variables are identified and reconciled against the correlation matrices for multicollinearity, the variables' subsets dataframes are split into training and testing sets and fed into the logistic regression and KNN models. Confusion matrices are employed to assess the classification accuracy of the models.

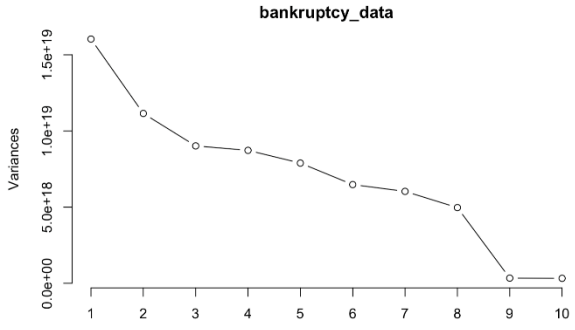
EXPLORATION AND INITIAL DISCOVERIES

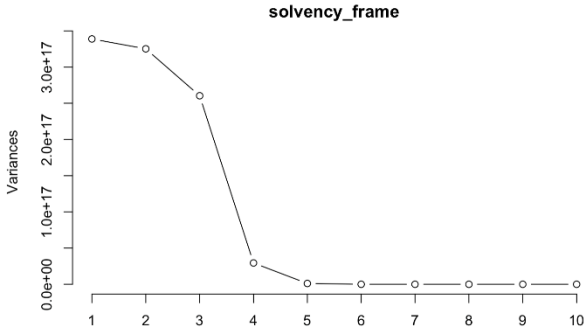
Implementing correlation matrices was integral to identifying potential multicollinearity issues. Within financial ratio groups, similar FRs tend to be highly correlated. For example, the Market Prospect group includes three different Net Value Per Share variables, which are both highly correlated with each other and with Per Share Net Profit Before Tax and Persistent EPS in the Last Four Seasons. Conversely, the Other Ratios frame is inherently a more diverse combination of FRs and are generally not highly correlated.





The number of principal components that capture the most variance varies significantly between the FR subsets. For example, below is a comparison between the dataset including all variables (bankruptcy_data) and the solvency FR subset (solvency_frame).





Without implementing class weights or stratified sampling, the models tend to predict a single class for all observations, especially because the original dataset contains more non-bankrupt observations than bankrupt. For example, the logistic model for the profit FRs without class weights predicts 12 observations as bankrupt, and 2,033 as non-bankrupt, as shown in the figure to the left. This model would have completely neglected 61 bankrupt observations. The tendency of the model to label observations as non-bankrupt becomes especially problematic in a production environment, such as one in which these decisions

are relied upon for financial decision-making. Its tendency to label observations as non-bankrupt increases the risk of predicting false negatives and makes relying on its decisions inherently risky.

```
[1] "PROFIT"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1972   61
         1    7    5

               Accuracy : 0.9667
                 95% CI : (0.958, 0.9741)
    No Information Rate : 0.9677
    P-Value [Acc > NIR] : 0.6294

                  Kappa : 0.1195

 Mcnemar's Test P-Value : 1.3e-10

            Sensitivity : 0.99646
            Specificity : 0.07576
         Pos Pred Value : 0.97000
         Neg Pred Value : 0.41667
             Prevalence : 0.96773
         Detection Rate : 0.96430
   Detection Prevalence : 0.99413
      Balanced Accuracy : 0.53611

       'Positive' Class : 0
```

MODEL SELECTION AND INTERPRETATION

The dataset is particularly suitable for classification tasks as it features a binary class variable, "Bankrupt?". Both logistic regression and KNN are well-suited for classification tasks and are utilized in this analysis. To address the class imbalance within the dataset, this study employs two methodologies: class weighting and resampling.

From a variable selection standpoint, the objective was to identify the most informative variables or features that contribute to the predictive power of the model while reducing redundancy and noise. In this study, two main approaches were employed to tackle this challenge:

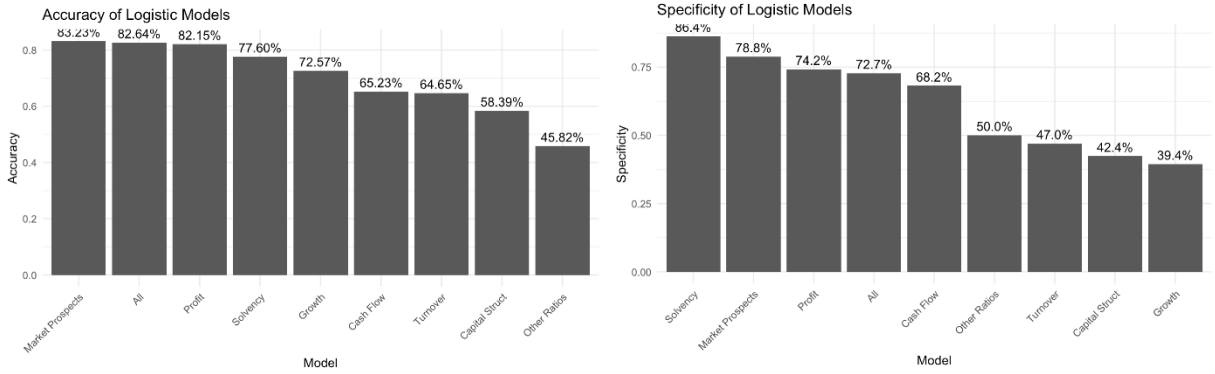
1. Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that transforms the original variables into a new set of orthogonal variables called principal components. By applying PCA, the dimensionality of the dataset can be reduced while retaining most of the information. This approach aids in simplifying the model and overcoming multicollinearity issues that may arise from highly correlated variables.
2. Consideration of Raw Variables and Subgroups: In addition to PCA, the study also examined the raw variables individually and grouped them into various subgroups based on financial ratio classification. Financial ratios, such as liquidity ratios, profitability ratios, and solvency ratios, provide insights into distinct aspects of a company's financial health and performance. By grouping the ratios into subgroups, the study aimed to capture the diverse dimensions of the company's financial position. This approach facilitated a more comprehensive analysis, allowing for the exploration of the unique contributions of each subgroup to predict bankruptcy.

The rationale behind these approaches lies in the fact that many financial ratios may measure similar underlying concepts and exhibit high correlations with each other. Therefore, by reducing the dimensions through PCA and exploring different subgroups, the study aimed to address multicollinearity and identify a more concise set of features that effectively capture the variability in the data. This process not only enhances the interpretability of the model but also improves its predictive performance by focusing on the most relevant variables.

A logistic regression model which implements class weights solves the model's bias towards labelling observations as non-bankrupt. The logistic regression model is trained and tested on each FR subset dataframe with class weights. The class weights are equal to 1 for observations where Bankrupt == 0 and are equal to the proportion of non-bankrupt to bankrupt observations (Count of Bankrupt == 0) / Count of Bankrupt == 1) for observations where Bankrupt == 1.

The accuracy of the logistic models in predicting bankruptcy is starkly different for each FR subset. The Market Prospects subset model has an accuracy rate of 83.23%, for example, while the Other Ratios subset model has an accuracy rate of only 45.82%. The model trained on the variables selected by principal component analysis on all columns ("All" in the figure below) had the second highest accuracy rate of the models at 82.64%.

The proportions of actual negatives that the models correctly classified as negative, or the specificity of the logistic models, are shown below on the right. Specificity of the Solvency model was the highest at 86.4%, followed by the Market Prospects and Profit models. The Growth model had the lowest specificity. There was no clear tradeoff between accuracy and specificity for the models.



In the following logistic regression results for the solvency logistic model, predictors such as Long-term Liability to Current Assets are statistically significant at the 99% confidence level, while predictors such as Quick Assets / Current Liability are not statistically significant. All else equal, a one-unit increase in the Long-term Liability to Current Assets ratio is associated with a decrease in the log odds of bankruptcy by approximately -1.958e-10. The Solvency model's confusion matrix shown below demonstrates that the model correctly predicted 1,530 non-bankrupt and 57 bankrupt observations while misclassifying 449 non-bankrupt as bankrupt and 9 bankrupt as non-bankrupt.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1383  -1.1119  -0.7863  -0.4642  15.7012

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             1.794e-02  1.372e-01   0.131 0.895961
`Inventory/Current Liability`          -8.738e-11  6.075e-11  -1.438 0.150316
`Long-term Liability to Current Assets` -1.958e-10  5.915e-11  -3.310 0.000933 ***
`Cash/Current Liability`                1.230e-10  3.775e-11   3.260 0.001116 **
`Quick Assets/Current Liability`       -1.697e-09  1.084e-07  -0.016 0.987515
`Current Ratio`                         9.489e+00  2.363e+00   4.015 5.94e-05 ***
`Quick Assets/Total Assets`            -5.916e-01  1.970e-01  -3.003 0.002673 **
`Current Assets/Total Assets`          -2.335e+00  2.815e-01  -8.294  < 2e-16 ***
`Cash/Total Assets`                    -4.056e-02  3.734e-01  -0.109 0.913488
`Current Liabilities/Liability`        -3.588e+00  1.829e-01 -19.614  < 2e-16 ***
`Current Liability to Assets`           3.094e+01  1.186e+00  26.098  < 2e-16 ***
`Liability-Assets Flag`                 1.222e+01  1.800e+02   0.068 0.945878
`Current Liability to Current Assets`   7.620e+00  1.997e+00   3.816 0.000136 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12809.4  on 4773  degrees of freedom
Residual deviance:  8979.1  on 4761  degrees of freedom
AIC: 9005.1

Number of Fisher Scoring iterations: 13
```

```
[1] "SOLVENCY"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1530    9
         1  449   57

               Accuracy : 0.776
                 95% CI : (0.7573, 0.7939)
    No Information Rate : 0.9677
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1508

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7731
            Specificity : 0.8636
         Pos Pred Value : 0.9942
         Neg Pred Value : 0.1126
             Prevalence : 0.9677
         Detection Rate : 0.7482
   Detection Prevalence : 0.7526
      Balanced Accuracy : 0.8184

       'Positive' Class : 0
```

K-nearest neighbor models were implemented for each FR subset. Before running the KNN-algorithms, however, resampling was done on the training data to address class imbalance. The resampling technique used was simultaneous oversampling (of the minority class = bankrupt) and undersampling (of the majority class = not bankrupt). The test data was not changed to ensure that evaluating model performance reflets how well the models perform on new, unseen data. The function used in R to perform simultaneous over- and under sampling (ovun.sample) allows the user to specify the desired size of the resampled training data (N) and the probability of resampling from the minority class proportion (p).

Regarding the choice of N, there are several approaches: Equalize Classes, Proportional Oversampling, and Custom Sampling. The approach chosen was Proportional Oversampling, which sets N to a multiple of the minority class size to bring it closer to the majority class size. We chose a five times multiple of the instances of the minority class (259), thus resulting in N = (5 x 259) = 1’259. p was set to 0.4, i.e., 40% of the resampled data is of the minority class (bankrupt). Below is a comparison of the training data before and after resampling:

	Bankrupt	Not bankrupt	Sample size
Training data, original	259 (= 5%)	4’515 (=95%)	4’774
Training data, resampled	509 (= 40%)	750 (60%)	1’259

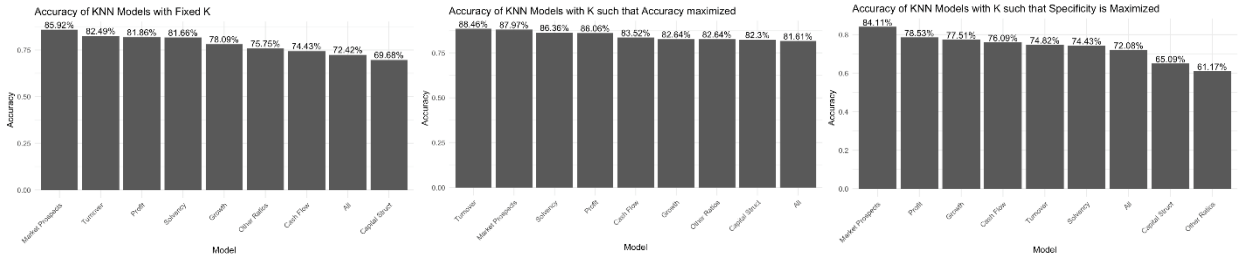
As a second step, the value for “K” was chosen (i.e., the number of nearest neighbors to consider for classification of an observation) using three different approaches:

a. Fixed K for all models: K = sqrt(n); with n = number of observations in training dataset[1]
b. K that maximizes accuracy
c. K that maximizes specificity

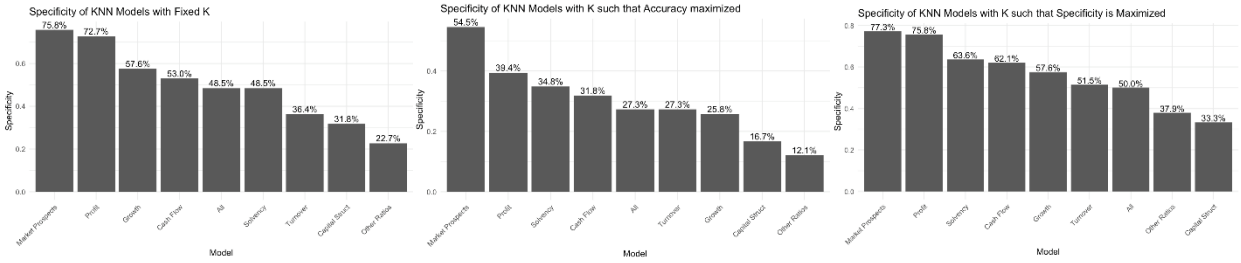
[1] This is a common starting point / rule of thumb for choosing K.

As a third step, model performance was compared using the criteria of accuracy and specificity.

Performance of KNN models: Accuracy



Performance of KNN models: Specificity



The analysis reveals that certain subsets of financial ratios consistently perform well in predicting bankruptcy, regardless of the modeling approach or performance measure used. Specifically, the "Market Prospects" and "Profit" subgroups yield the best results across different configurations, while the "Capital Structure" subgroup performs poorly.

Interestingly, the findings suggest that using more independent variables does not necessarily lead to better performance. In fact, the "All" dataset showed only average performance across different modeling approaches and performance measures. The highest accuracy achieved was 88.46%, obtained using the "Turnover" financial ratios, while the highest specificity reached was 77.3%, achieved with the "Market Prospects" financial ratios. Regarding the trade-off between accuracy and specificity, it appears that some KNN models perform well in both metrics. For instance, using the "Market Prospects" financial ratios with a fixed K produced an accuracy of 85.92% and a specificity of 75.8%.

CONCLUSION

The selection of financial ratio (FR) categories utilized in logistic regression and KNN models plays a crucial role in determining the models' performance. Within the chosen FR categories, prioritizing feature selection to mitigate multicollinearity is essential.

KNN models consistently demonstrate higher accuracy rates compared to logistic models across almost all FR groups. Notably, the KNN model optimized for accuracy using Turnover ratios achieves the highest accuracy among all models, reaching 88.46%, surpassing the logistic Market Prospects model's accuracy of 83.23%. While logistic models exhibit significant drops in accuracy between FR groups, the accuracy rates among FR groups for KNN models are more uniform. However, logistic models outperform

KNN models in terms of specificity for most FR groups, even surpassing specificity-optimized KNN models.

Incorporating the class weights and simultaneous oversampling of the minority class (bankrupt) and undersampling of the majority class (non-bankrupt) are indispensable components of the logistic and KNN models, respectively. Without these methods, both models would have higher accuracy but would be much more likely to misclassify true positives as false negatives, making the models riskier to deploy in production.

Expanding the analysis may involve expert evaluation to validate the findings against domain-specific knowledge. Expert judgment can provide insights into the coherence of the results with established principles in finance and econometrics, enhancing the credibility of the study's conclusions. Furthermore, comparing the performance of the models with data from other robust markets, such as those in North America or Europe, can offer valuable insights. By assessing the models' performance across different market conditions and regulatory environments, we can better evaluate the universal applicability of the selected financial ratios and modeling methodologies. This comparative analysis would provide a broader perspective on the effectiveness and generalizability of the models, contributing to their robustness and reliability in real-world applications.

WORKS CITED

Beaver, William H. "Financial Ratios As Predictors of Failure." *Journal of Accounting Research*, vol. 4, 1966, pp. 71–111. *JSTOR*, https://doi.org/10.2307/2490171. Accessed 17 Mar. 2024. http://www.jstor.org/stable/40226072. Accessed 17 Mar. 2024.

Liang, Deron, et al. "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study." *European journal of operational research* 252.2 (2016): 561-572.

Sidanius, James. "Principal Components Analysis." *Advanced Research Computing - Statistical Methods and Data Analytics*, UCLA, stats.oarc.ucla.edu/spss/output/principal_components/. Accessed 12 Apr. 2024.

Tamari, Meir. "Financial Ratios as a Means of Forecasting Bankruptcy." *Management International Review*, vol. 6, no. 4, 1966, pp. 15–21. *JSTOR*, http://www.jstor.org/stable/40226072. Accessed 17 Mar. 2024.

DATA SOURCES

Liang et al. "Taiwanese Bankruptcy Prediction." *Multivariate Datasets*, 27 June 2020, https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction. Accessed 12 Apr. 2024.

README/DOCUMENTATION FOR GITHUB

https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-119