International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-5 (October 2020)
www.ijemr.net https://doi.org/10.31033/ijemr.10.5.15

Prediction of Corporate Bankruptcy using Financial Ratios and News

Isha Arora[1] and Navjot Singh[2]
[1]Software Engineer, Independent Researcher, INDIA
[2]Software Engineer, Independent Researcher, INDIA

[1]Corresponding Author: ishaa396@gmail.com

ABSTRACT

A corporate's insolvency can have catastrophic effects on not only the corporate but also on the returns of its lenders and investors. Predicting bankruptcy has been one of the most sought-after areas for researchers for many decades. This study involves predicting the bankruptcy of the United States corporates using financial ratios and news data. The financial ratios of the companies were extracted from yearly financial reports of the companies, and the news data of the companies was scrapped from online newspapers, reports and articles using Google News. The news data was analyzed for negative and positive sentiments. The sentiment scores, along with the financial ratios of the companies, were given as features to the machine learning models. Various models were analyzed for their results such as Random Forest, Logistic Regression and Support Vector Machines (SVM). The study finds the best results from the random forest model with an accuracy of 90%. Moreover, the significant feature importance of the sentiment score given by the model proves that unstructured data, such as news, can play a crucial part in predicting bankruptcy in conjunction with the structured data, such as financial ratios.

Keywords— Bankruptcy Prediction, Financial News, Financial Sentiment Analysis, News and Ratios

I. INTRODUCTION

O'Leary [17] proclaims that the bankruptcy of a firm affects the entire life span of a business. Thus, the prediction of bankruptcy probably is one of the most crucial decision-making problems for business. Due to frequent changes in the market and economy, bankruptcy has become an indispensable part of the economic ecosystem. Moreover, failure results in a high cost for society, the collaborators, and the country's economy [2]. A company's bankruptcy has disastrous effects on its various stakeholders, such as creditors, employees, suppliers, consumers, and the local community. Accurate assessment of the financial condition of a corporate is vital for its stakeholders as it influences the significant decisions concerning their engagement and relationship with the company. Bankruptcy prediction is a technique of projecting on company financial distress, and it can provide the early warning signals to a corporation's executives to take preventive measures. It also warns the credit lenders and institutional funds from investing in a company with high susceptibility to financial distress.

Beaver [4] created one of the first bankruptcy prediction models. He analyzed different financial ratios for their potential to predict the insolvency. Then, he built a univariate discriminant analysis. Altman(1968) proposed a discriminant analysis technique to classify corporates into bankrupt and non-bankrupt categories. He has improvised his model to predict the bankruptcy for both manufacturing and non-manufacturing companies. Since four decades, the insolvency prediction has evolved as a primary research domain in corporate finance. Researchers have employed many modelling techniques to get a better insight into the problem of predicting bankruptcy.

Ohlson [16] employed nine financial ratios into a logistic regression model to estimate the probabilities of bankruptcy. Prusak [18] used other techniques, such as linear programming, the method of recursive division, cluster analysis or classification trees. With the advent of more advanced analytical and statistical tools that provided the capability to process sophisticated and more massive datasets, the race to find a more efficient model started. Odom and Sharda [15] were one of the early researchers to propose the use of neural networks for bankruptcy classification. They used the same five financial ratios used by earlier mentioned Altman (1968). After that, a significant number of researchers applied various neural network models on financial ratios to predict bankruptcy. Although the neural network is a widespread technique, it cannot explain the causal relationships among variables. Gepp and Kumar [6] and Parker et al. (2002) implemented a survival analysis to predict corporate failure. Lee [9] also used the same technique on a few financial ratio variables and evaluated the method using the Cox Proportional Hazard Model. Lin and McClean [12] assess various techniques that predict corporate collapse and find that Decision Trees also provide good results. Most of these studies have followed the work of earlier mentioned Altman (1968), and they used the financial ratios used by him in Z-score as an input to the neural network, decision tree or other Machine Learning models.

One of the major problems with most of the existing models is that they mainly use backwards-looking financial ratios, and thus, their predicting capabilities are

limited. In 2018, Ahmadi et al. [1] utilized deep sentiment analysis on business management reports and inputted the results to a neural network model, along with Z-score financial ratios. Hajek et al. [7] suggest that financial ratios have a non-linear relationship with the sentiment analysis score of annual reports of companies. Thus, suggesting that a more accurate model for predicting the collapse can be created by adding more features.

However, far too little attention has been paid to analyze the sentiment of the companies' news. The objective of our study is to provide a framework to predict the bankruptcy by using financial ratios incorporated with the news data of the companies. News articles project the latest operational condition of a company, both at micro as well as at a macro level and can help in the evaluation of the market's sentiment for the company. We hypothesize that the bankruptcy prediction model can be improved significantly by using financial news data along with financial ratios data. We use Random Forest, Support Vector Machines and Logistic Regression classifiers to test this hypothesis.

The rest of this paper is organized as follows. First, we briefly review previous literature on sentiment analysis of companies' news. Next, the dataset used for the study is described. In section 4, we present the methods used for the sentiment analysis and the prediction of financial collapse. Finally, the results of the experiments are provided and analyzed.

II. LITERATURE SURVEY

Often, investors use news articles and financial data in conjunction to gauge the financial health of the company. The news articles tell them the recent happenings of a company – positive, negative or neutral – whereas, financial data gives a sneak-peek of how competent a company is to handle the adversity. This study aims to automate and simplify this analyzing process by leveraging machine learning models.

Altman Z-Score Model

According to Siddiqui [19], Altman Z-score [3] can be safely and effectively used to predict bankruptcy two or three years before the collapse. Z-score is based on five key financial ratios. This multivariate discriminant model was the first real breakthrough into the problem of predicting bankruptcy and is still one of the most sought after methods of predicting bankruptcy among financial analysts. The Z score for public firms is calculated using the formula given below –

$$Z\text{-Score} = 1.2\,X1 + 1.4\,X2 + 3.3\,X3 + 0.6\,X4 + X5 \quad (1)$$

where,

X1 = Working capital/Total assets
X2 = Retained earnings/Total assets
X3 = Earnings before interest and tax(EBIT)/Total assets
X4 = Market value of equity/Total liabilities
X5 = Net sales/Total assets

If Z-score is less than 1.8, then, there are high chances that the company is heading towards bankruptcy. However, a Z-score above 3 indicates the financial stability, and a score between 1.8 and 3 is considered as a grey area. In other words, the lower the z-score, the higher the probability that the company is going to collapse. The accuracy of Z score two years before the collapse of a company is 72-80%. Z-score can predict the bankruptcy for two and one years before the failure of a firm.

Sentiment Analysis and Loughran and McDonald Dictionary

Sentiment analysis is the process of "computationally" categorizing and identifying the sentiment of the author of a text or a speech. With the ever-increasing amount of data, sentiment analysis has become a popular field. Sentiment analysis can be done in two different ways – machine learning-based approach or lexicon-based approach [10]. In the machine learning-based approach, first, the model is trained using the training set, which is a set of labelled phrases. Then, the testing set, which is an unlabeled set, is classified using the trained model. While, in the lexicon-based approach, a dictionary, which contains words classified in different categories, is used. Chaovalit and Zhou [5] used both the sentiment analysis methods to classify the reviews of movies and reported that both the approaches were challenging. Loughran and McDonald [13] suggest a dictionary which can be beneficial for financial text analysis. The dictionary has four different types of categories: Positive, Negative, Litigious and Uncertainty. Hájek and Olej [8] use this dictionary to find the correlation between the sentiment of financial reports and stock returns and report the positive results. Furthermore, Li et al. [11] also reported promising results when he used the Loughran-McDonald dictionary to explore the relationship between financial news and stock prices.

III. DATA AND METHODOLOGY

The complete study was divided into four phases. In phase one, names and financial data of the companies was collected. In the next step, we carried out the scraping of relevant news articles for both bankrupt and non-bankrupt companies. In the third phase, sentiment analysis was done on the news articles to get the sentiment scores. Finally, a Machine Learning model was trained and tested on the combined financial ratios data and sentiment scores.

Data Overview

For bankrupt companies' dataset, the companies which had filed for bankruptcy under Chapter 11 between 2013 and 2018 were selected for two reasons. The number of internet users has, approximately, doubled from 2.3

billion in 2013 to 4.3 billion in 2018 [14]. The number of online newspapers and articles have also increased substantially in this period. Thus, using this period guarantees that we can get a significant amount of relevant online news articles. Moreover, after the 2008 financial crisis, many companies were financially distressed because of the prevalent market conditions. The companies which filed bankruptcy post-2013 were selected to avoid any anomaly which might exist as a result of the 2008 crisis.

The list of the companies which filed bankruptcy was extracted from Wikipedia Open Source U.S. bankrupt companies list. Non-bankrupt companies with similar market size and sector were selected using the Standard Industrial Classification (SIC) code. SIC code is a standard industry categorization method used in the U.S.

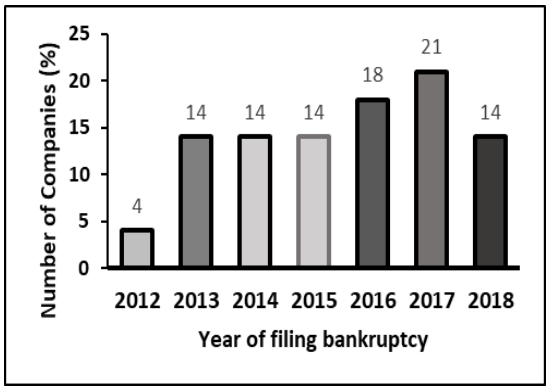


Figure 1: Distribution of the percentage of companies that filed bankruptcy with the filing year in the dataset

A. Financial Ratios and News Data Overview

For this study, the five financial ratios used in the Altman Z-score were taken as the base. (1)

For bankrupt companies, we extracted the data for two years – one year(t1) and two years(t2) before the collapse. In the dataset, the maximum number of companies filed for bankruptcy in the year 2017 (Figure 1). Therefore, for non-bankrupt companies, 2017 was taken as the reference year with t1 and t2 as 2016 and 2015 respectively. For both bankrupt and non-bankrupt companies, the needed financial attributes were extracted from yearly balance sheets and income statements.

News data of both bankrupt and non-bankrupt companies were also extracted for the same years, t1 and t2, as of financial ratios data. For all the companies, the data of online newspapers, articles and reports was scraped - both in automated and manual procedure. The scraped news dataset was divided into two parts based on the duration before the filing of the bankruptcy (t1 or t2).

Modelling

A. Pre-Processing

Binary classification, a supervised learning method of Machine Learning, divides the data into two categories. Bankruptcy prediction is a binary classification problem. Therefore, the companies in the dataset which filed for bankruptcy were labelled as 1, and non-bankrupt companies were labelled as 0.

All the collected financial news data of a company for a year were aggregated into a single text file. The raw text files were loaded into data frames of Pandas, a fast data analysis and manipulation library for Python.

As part of pre-processing, the news text was separated into single smaller units called tokens, which are the building block of Natural Language Processing(NLP) based analysis.

In the next step, the removal of stop words, words that do not add any significant meaning to the sentence, was carried out. Finally, lemmatization was carried out on the tokens left after the stop words removal. Lemmatization, a text normalization procedure, is the process of extracting the root form of a word from its different inflected forms. As lemmatization internally uses the linguistic dictionary, it is more accurate than other such techniques. In this study, WordNet lemmatizer, a part of the Python's Natural Language Toolkit (NLTK) library, was used.

B. Sentiment Analysis Modelling

After the pre-processing, VADER (Valence Aware Dictionary and Sentiment Reasoner) model was used. VADER is a rule-based and lexicon sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains. VADER analyzer outputs a dictionary of scores divided into four categories - negative, neutral, positive and compound (normalized score of the previous three scores). Since this is financial analysis, the corpus of the model was updated with the words from the Loughran and McDonald Sentiment Word Lists 2018 to increase the accuracy of the analyzer. VADER model was used on the cleaned text for both t1 and t2, and scores for negative, positive and uncertain sentiments were obtained.

C. Bankruptcy Prediction

To test our null hypothesis that bankruptcy can be predicted more accurately by using news data along with financial ratios, both the resultant datasets –

- The sentiment analysis score of financial news
- The financial ratios of companies

were given as an input to three Machine Learning models - random forest classifier, logistic regression and support vector machines(SVM). For each of the models, the data was divided into 70% as training and 30% as the test set.

IV. RESULTS AND DISCUSSIONS

Models Performance Comparison

Three machine learning models were applied to prove the hypothesis of the study – SVM, Random Forest

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-5 (October 2020)
https://doi.org/10.31033/ijemr.10.5.15

and Logistic Regression. The models were measured on three factors - accuracy, recoil and precision. Accuracy quantifies how many times the model was correct in its predictions. Precision tells us how many of the positive predicted cases turned out to be correct. Recall answers about how many of the actual positive instances we were predicted correctly.

TABLE 1: Results of all models for t1

Classifier	Precision(%)	Recall(%)	Accuracy(%)
Random Forest	87.5	87.5	86.7
Logistic Regression	81.2	81.2	80.0
SVM	81.2	81.2	80.0

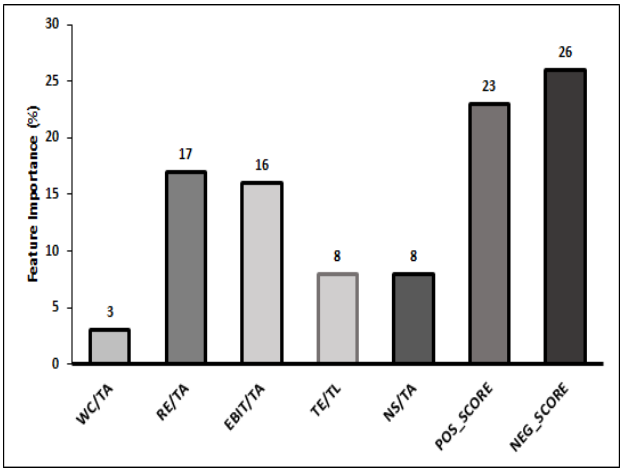
TABLE 2: Results of all models for t2

Classifier	Precision(%)	Recall(%)	Accuracy(%)
Random Forest	88.2	93.8	90.0
Logistic Regression	90.2	62.5	76.7
SVM	91.7	68.8	80.0

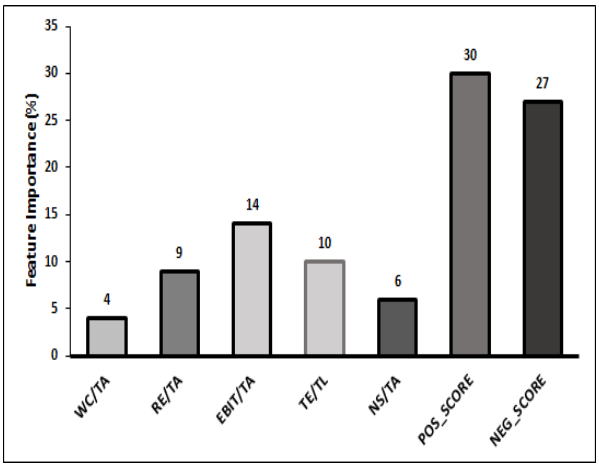
The Altman Z-score [3] gives the accuracy of 70-80% for t1 and t2. From the above results, we can confirm that the Random forest outperforms the Altman Z-score [3] model. Thus, the results prove the hypothesis of this study that news articles can significantly contribute to the bankruptcy prediction model. Random Forest can be confidently relied on for bankruptcy prediction.

Analysis of Random Forest model results

A. Feature Importance



(a)



(b)

Figure 2. Feature importance by Random Forest Classifier (a) For one year before filing (t1) (b) For two years before filing (t2)

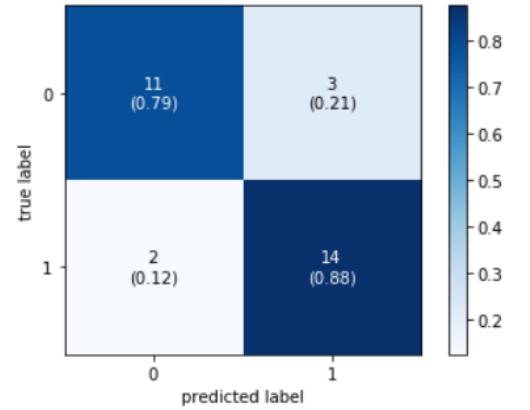
where,
WC/TA = Working capital/Total assets
RE/TA = Retained earnings/Total assets
EBIT/TA=Earnings before interest and tax(EBIT)/ Total assets
TE/TL = Market value of equity/Total liabilities
NS/TA = Net sales/Total assets
POS_SCORE = positive sentiment score
NEG_SCORE = negative sentiment score

According to the figure, positive and negative sentiment score had cumulative feature importance of 49% and 57% for t1 and t2 respectively. From this significant feature importance, it can be concluded that the financial news of the companies is an influential factor for forecasting their downfall.

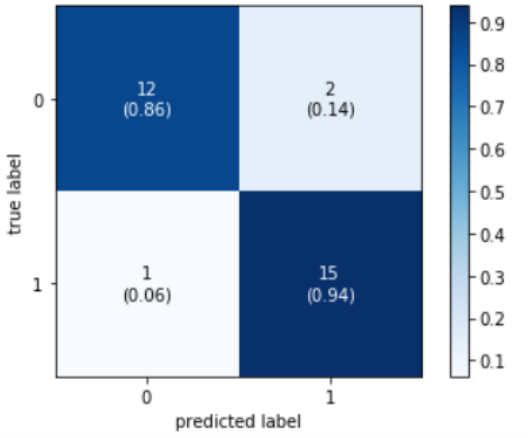
B. Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification algorithm. The confusion matrix gives the number of companies which were predicted correctly (cell 00 and 11) by the model. It also provides the number of companies which filed for bankruptcy but were not detected by the model, formally known as False Negatives (cell 10), and the number of companies which did not file for bankruptcy but were labelled as bankrupt by the model, also called as False Positives (cell 01). The false negatives are harmful to the credibility of an algorithm and thus, should be as low as possible for a model to be reliable.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-5 (October 2020)
https://doi.org/10.31033/ijemr.10.5.15



(a)



(b)

Figure 3. The confusion matrix for Random Forest Classifier (a) For one year before filing (t1) (b) For two years before filing (t2)

For t1, there were 6.66% false negatives, and for t2, there were 3.33% false negatives. As the false-negative percentages are minimal, this shows that the study's working approach is both reliable and robust.

V. CONCLUSION

The main objective of the study was to investigate if financial news can improve the accuracy of a bankruptcy prediction model. The dataset contained U.S. companies that filed for bankruptcy between 2013 and 2018, and an equal number of the US non-bankrupt companies belonging to a similar SIC code. The financial ratios and news of the bankrupt companies were extracted for two years (t1 and t2) before the collapse. For non-bankrupt companies, the data was collected for 2015 and 2016.

Out of all the models tested, the Random Forest classifier was able to provide the best accuracy, recoil and precision. The accuracy of 90.0%, obtained from the random forest model, suggests that the financial news can significantly contribute to the bankruptcy prediction model. The feature importance of combined (positive and negative) sentiment score of 57% indicates that the daily financial and operational news can give a better and reliable overview of the company's future performance. This study promotes the use of financial news along with the financial ratios to solve the long-sought-after problem of bankruptcy prediction.

The accuracy of the model can be further enhanced by training the models on a larger dataset. The financial institutions can leverage the paid news subscriptions to extract more relevant and accurate financial news. The future study can evaluate the impact of the macroeconomic features like GDP, industry growth etc. on the company's future performance.

REFERENCES

[1] Ahmadi, Z., Martens, P., Koch, C., Gottron, T., & Kramer, S. (2018). Towards bankruptcy prediction: Deep sentiment mining to detect financial distress from business management reports. *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy*, pp. 293-302. Available at: https://doi.org/10.1109/DSAA.2018.00040.

[2] Ahn, B.S., Cho, S.S., & Kim, C.Y. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications 18*, 65-74.

[3] Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

[4] Beaver, W. (1966). Financial ratios as predictors of failure, empirical research in accounting, selected studies. *Journal of Accounting Research, 4*, 71-111.

[5] Chaovalit, Pimwadee & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 112c-112c.

[6] Gepp, Adrian & Kumar, Kuldeep. (2008). The role of survival analysis in financial distress prediction. *International Research Journal of Finance and Economics 16*, 13-34.

[7] Hajek, Petr, Olej, Vladimir & Myskova, Renata. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, *20*(4), 721-738. Available at: https://doi.org/10.3846/20294913.2014.979456.

[8] Hájek, Petr & Olej, Vladimír. (2015). Word categorization of corporate annual reports for bankruptcy

prediction by machine learning methods. *Proceedings of the 18th International Conference on Text, Speech, and Dialogue 9302*. *Springer-Verlag, Berlin, Heidelberg*, pp. 122–130. Available at: https://doi.org/10.1007/978-3-319-24033-6_14.

[9] Lee, Ming-Chang. (2014). Business bankruptcy prediction based on survival analysis approach. *International Journal of Computer Science and Information Technology, 6*, 103-119. Available at: https://doi.org/10.5121/ijcsit.2014.6207.

[10] Li, N. & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*(2), 354–368.

[11] Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14–23.

[12] Lin, F.Y. McClean, & S.I. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems 14*, (3-4), 189-195. Available at: https://doi.org/10.1016/S0950-7051(01)00096-X.

[13] Loughran, T. and McDonald, B. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance 66*, 35-65. Available at: https://doi.org/10.1111/j.1540-6261.2010.01625.x.

[14] Miniwatts Marketing Group. (2020). *Internet world stats: Usage and population statistics*. Viewed 19 October, 2010. Available at: https://www.internetworldstats.com/emarketing.htm.

[15] Odom, Marcus & Sharda, Ramesh. (1990). A neural network model for bankruptcy prediction. *IEEE International Joint Conference on Neural Networks, 2*, 163 – 168. Available at: https://doi.org/10.1109/ IJCNN.1990.137710.

[16] Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*, 109-131.

[17] O'Leary, E. G. (2000). *Business failure prediction and the efficient market hypothesis*. Simon Fraser University. (Parker et al. 2002) Parker, S., Peters, G. F., and Turetsky, H. F. (2002). Corporate governance and corporate failure: a survival analysis. *Corporate Governance. The International Journal of Business in Society, 2*(2), 4–12.

[18] Prusak, Błaz˙ej. (2005). *Modern methods of forecasting financial risk of enterprises*. Warszawa: Difin.

[19] Siddiqui, Sanobar. (2012). Business bankruptcy prediction models: A significant study of the Altman's Z-score model. *SSRN Electronic Journal*. Available at: https://doi.org/10.2139/ssrn.2128475.