Corporate Credit Rating Prediction Using Multi-class Support Vector Machines, Random Forest, and Neural Network

Liyuan Chen, SCPD, liyuan@stanford.edu

1 Introduction

Corporate credit rating reflects the risk of lending money to the company. Thus, it has been an important assessment and indicator for investors (Wang & Ku, 2021). Meanwhile, the credit rating of a company usually does not deteriorate or improve in a sudden. Instead, there are usually signs indicating if the company is doing well or not. Therefore, it would be beneficial to investors if corporate credit ratings can be forecast using companies' performance in the past with the help of machine learning.

The input to the algorithm is a group of companies' credit ratings and other features collected in the past. A multi-class Support Vector Machines (SVM) model is used to predict the corporate credit ratings. The model is then enhanced using Random Forest. A feed-forward neural network model is also created to compare with the SVM model on the accuracy of the prediction results, which is calculated by the number of correct predictions divided by the total number of samples.

2 Related Work

Prediction on corporate credit rating has been a popular topic among researchers. The forecast used to be conducted using statistical models, which are gradually being replaced by machine learning models (Lee, 2007). Among all the machine techniques, neural networks, SMV, and random forest have been observed to be able to achieve better performance than the rest in credit rating forecast (Wallis et al., 2019). Although it is mostly believed that finding features of greater significance will improve machine learning algorithms, it was found that using all features might be optimal when it comes to credit rating prediction (Golbayani et al., 2020). In the past, backpropagation neural network was found to be able to predict company credit ratings with an accuracy rate of 80% (Huang et al., 2004). A shift towards qualitative analysis on company reports and interviews to forecast company's credit performance has also been observed in recent years (Choi et al., 2020).

In this project, I am also going to examine how machine learning algorithms such as SVM, random forest, and neural network could be used for corporate credit rating prediction. However, the focus of the project is whether random forest could improve SVM model and also compare it with feed-forward neural network. The project also intends to compare the model results using aggregated data from various rating agencies with that from individual rating agency, so as to explore if different rating agencies have an impact on the modelling result.

3 Dataset and Features

A set of data on corporate credit ratings is downloaded from Kaggle (Delwadia, 2022). The dataset includes various companies, their credit ratings, respective rating agencies, and values of companies' various indicators and performances used by rating agencies to generate ratings. There are 7805 records and 23 features in total. All the columns and their explanations are summarized in Table 1 which helps to determine the features that could potentially affect companies' credit ratings.

Table 1: Features Given in the Database

Column	Feature	Description
1	Rating Agency	The institution who evaluates the corporation
2	Corporation	The entity that is being evaluated

Column	Feature	Description
3	Rating	The company's credit score
4	Rating Date	The date when the company is evaluated
5	CIK	Central Index Key, a key representing the entity used by Securities and Exchange Commission's (SEC's) computer systems
6	Binary Rating	The rating given to the company, which is either zero or 1
7	SIC Code	Standard Industrial Classification (SIC) codes by U.S. government
8	Sector	The industry focus of the company
9	Ticker	Company's stock market information / reports
10	Current Ratio	A liquidity ratio that indicates a company's ability to pay back in a short period of time
11	Long-term Debt / Capital	An indicator used to measure how much asset and debt a company has
12	Debt/Equity Ratio	A ratio between company's debt and its value of shareholders' equity
13	Gross Margin	The surplus when the cost is deducted from net sales
14	Operating Margin	The surplus when the cost is deducted from net sales per dollar
15	EBIT Margin	The ratio between company's total cost and total sales
16	EBITDA Margin	The ratio between a company's profit and its revenue
17	Pre-Tax Profit Margin	Company's operating revenue
18	Net Profit Margin	The surplus of company's earnings after all costs, tax expenses, etc. are deducted from the revenue
19	Asset Turnover	The ratio between company's revenue and assets
20	ROE - Return On Equity	The ratio between profit and shareholders' equity
21	Return On Tangible Equity	The earning of investment into the company
22	ROA - Return On Assets	An indicator on the profitability of the company; ratio between EBIT and the company's assets
23	ROI - Return On Investment	Earning from the investment
24	Operating Cash Flow Per Share	Net profit per share where inflation is also counted
25	Free Cash Flow Per Share	The amount of cash generated per company share in the past 12 months

The features, "Rating Agency", "Corporation", "Rating", "Rating Date", "CIK", "SIC Code", "Sector", and "Ticker", are properties of the company or rating agency, Therefore, they are not included in the experiment since they are not indicators on companies' performance.

The data was split into training and test sets where 70% data were used for training and 30% were used for testing. Company features in the test dataset was fed into the trained model and predicted the corporate's credit rating, which was compared to their actual credit ratings.

4 Methodology

An SVM model is selected since it is able to map samples to multi-dimensional space to classify them based on different features (Baeldung, 2021). To transform the input data into forms that can be processed by SVM algorithms, kernels are used to complete the conversion (Data Flair, 2021). In this experiment, a radial basis function (RBF) kernel is used since it has proven to be the most effective kernel in most cases (Savas & Dovis, 2019). The equation of RBF kernel is shown as follow (Sreenivasa, 2020):

$$K(X_1, X_2) = \exp\left(-\frac{\| X_1 - X_2 \|^2}{2\sigma^2}\right)$$

Random forest is another machine learning technique which uses randomly generated decision trees to predict one outcome (IBM, 2020). The random forest model calculates the importance of each feature using the number of samples reaching the end of each branch divided by the total number of samples. Thus, a greater feature importance indicates that the feature has a stronger influence on the predicted results (Ronaghan, 2019).

The next step of the experiment is to enhance the model using random forest. The random forest model is used to fit the training data and find the top features that have the greatest importance on the target value. The training dataset is then reconstructed with the important features only, which are imported into the SVM model for prediction on corporate credit rating. The model is evaluated on its accuracy score, which is the ratio of the correct prediction (true positive and true negative) to the total number of samples (Baeldung, 2021).

A neural network refers to a model with multiple layers of processors where each layer of processors taking the output of the previous layer as input (Burns & Burke, 2021). In this project, a feed-forward neural network which does not form a cycle is selected due to its implicitly (DeepAI, 2019). In neural network, the target value does not have to be a single integer and hence the credit ratings were encoded into lists of values. Two layers of processors were selected where the first activation function is rectified linear unit (ReLU) and the second activation function is softmax. ReLU is one of the most popular neural network activation functions as it is relatively computationally cheap (Chaudhary, 2020). ReLU is also not affected by vanishing gradient problem compared to sigmoid function as illustrated in Figure 1.

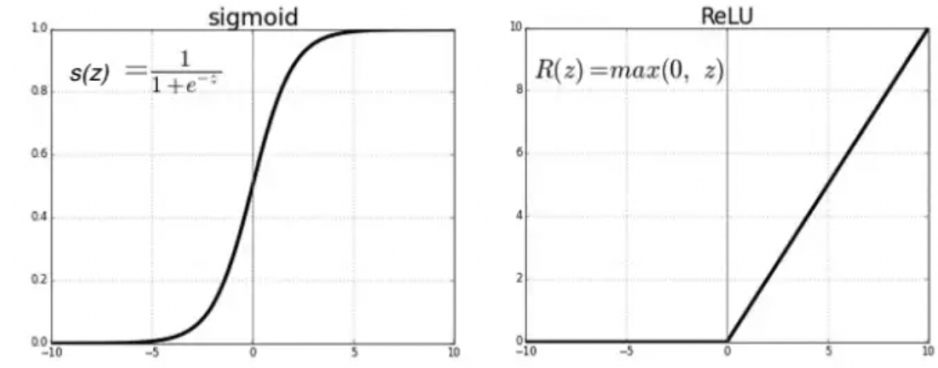

Figure 1: Sigmoid Function vs. ReLU Function (Chaudhary, 2020)

Softmax function is selected since it calculates the probability distribution of one instance over the rest different instances, which allows it to tackle problem with making classification among multiple classes (Chaudhary, 2020). The equation of softmax function is shown below (Chaudhary, 2020):

$$softmax(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \; for \; j = 1, \dots, k$$

The neural network model is evaluated using categorical cross-entropy loss, which is calculated as follow (Anand, 2021):

$$Loss = -\sum_{i=1}^{i=N} y_{true_i} \cdot \log\left(y_{pred_i}\right) = -\sum_{i=1}^{i=N} y_i \cdot \log(\hat{y}_i)$$

5 Experiment / Results / Discussion

The corporate credit ratings are given in letter's format and converted into numeric values for the SVM model. The experiment started with creating an SVM model to forecast corporate credit ratings. By comparing the predicted values and the actual values, the accuracy score of the model was found to be 12.1%.

A random forest regressor was then used to select the top features that have the greatest influence on corporate credit ratings. Since the data uses 16 features, 100 estimators were used to avoid overfitting but also cover a wide range of estimation. The results showed that features with greatest importance are "Current Ratio", "Long-term Debt/Capital", "Pre-Tax Profit Margin", and "ROI – Return On Investment". The importance of respective features is shown in Figure 2.

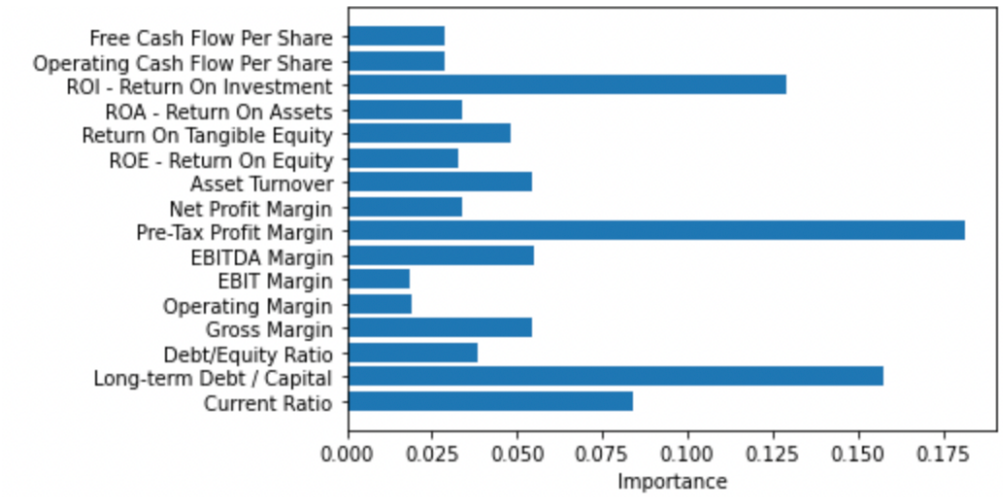

Figure 2: Importance of Various Features Calculated Using Random Forest

A new dataset is structured using the top four features identified using random forest and a multi-class SVM model is created using the newly filtered training data. When evaluating the model using the test data, it was found that the accuracy score increased to 16.2%.

As shown above, although random forests were able to increase the model's accuracy, the model's performance is still undesirable. To figure out what could have negatively impacted the model's prediction, same experiment was conducted on different rating agencies since different rating agencies may have different rating standards. The accuracy score of each rating agency is summarized in Table 2.

Table 2: Accuracies on SVM and SVM + Random Forest (RF) Models on Each Rating Agency

Agency	No. of Samples	Accuracy (%)			Selected Features
		SVM	SVM + RF	Increase in Accuracy	
Standard & Poor's Ratings Services	2813	11.5	12.9	+1.4	'Current Ratio', 'Long-term Debt / Capital', 'Pre-Tax Profit Margin'
DBRS	26	25.0	0.0	-25.0	'Current Ratio', 'Debt/Equity Ratio', 'Gross Margin', 'Asset Turnover', 'Return On Tangible Equity', 'ROI - Return On Investment'
Moody's Investors Service	1636	15.7	12.6	-3.1	'Current Ratio', 'Long-term Debt / Capital', 'EBITDA Margin', 'Pre-Tax Profit Margin', 'Net Profit Margin', 'ROI - Return On Investment'
Fitch Ratings	477	16.0	19.4	+3.4	'Current Ratio', 'Long-term Debt / Capital', 'Gross Margin', 'EBITDA Margin', 'Pre-Tax Profit Margin'
Japan Credit Rating Agency Ltd.	22	28.6	57.1	+28.5	'Current Ratio', 'EBIT Margin', 'Asset Turnover', 'Free Cash Flow Per Share'
HR Ratings de Mexico S.A. de C.V.	5	0.0	50.0	+50.0	'Long-term Debt / Capital', 'Debt/Equity Ratio', 'Gross Margin', 'Operating Margin', 'EBIT Margin', 'EBITDA Margin', 'Pre-Tax Profit Margin', 'Net Profit Margin', 'Return On Tangible Equity'
Egan-Jones Ratings Company	2826	12.1	17.0	+4.9	'Current Ratio', 'Long-term Debt / Capital', 'ROE - Return On Equity', 'ROI - Return On Investment'

From Table 2, it can be seen that corporate credit ratings conducted by all agencies except Standard & Poor's Ratings Services and HR Ratings de Mexico S.A. de C.V. had higher validation accuracy than the overall accuracy of 12.1%. By comparing predictions made by SVM model and SVM model with random forest, HR Ratings de Mexico S.A. de C.V. showed the greatest improvement in accuracy on test data, followed by Japan Credit Rating Agency Ltd. However, credit ratings conducted by DBRS and Moody's Investors Service worsened with random forest. The volatility in test accuracy could be due to the small sample size, which cannot guarantee if the model result is representative (Clancy, 2020). Moreover, some features are correlated. For example, "Current Ratio" and "Pre-Tax Profit Margin" are both related to the company's revenue. Correlated features may result in lower importance compared to the same tree without the correlated features (Dubey, 2018).

Since SVM and random forest did not provide promising forecast results, a neural network model was structured and tested on its performance. Although 8000 epochs were set to be used to determine the best parameters, the model would stop iterations when there is no improvement in the validation loss for 10 consecutive epochs. The model stopped after the 44th epoch with a training accuracy of 59.2% and a validation accuracy of 55.2% as illustrated in Figure 3.

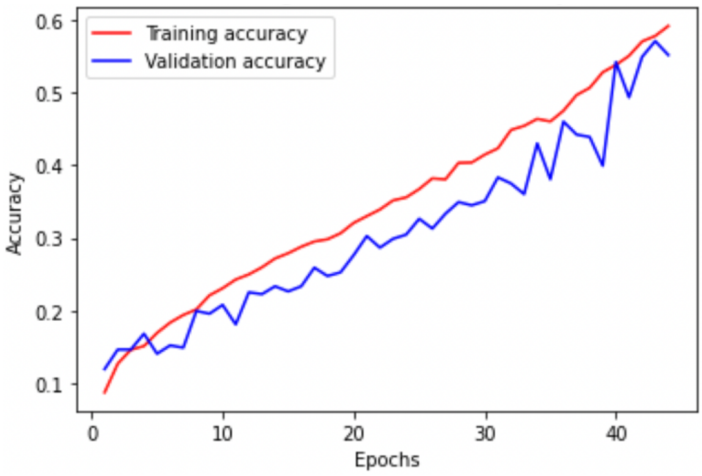

Figure 3: Training and Validation Accuracy of Neural Network Model

6 Conclusion / Future Work

The experiment shows that random forest is able to improve SVM model on corporate credit rating forecast in most cases. However, a large sample size is required to build the model. Feed-forward neural network is proved to be more accurate in predicting corporate credit rating than SVM and random forest. The model results showed volatility between different rating agencies, which could be caused by correlation between features. It is recommended to repeat the experiment using a larger sample size and independent features.

In the future, since corporate credit rating varies with time, a Long Short-Term Memory model could be used and compared with SVM, random forest and neural network models. A white reality check could also be performed to evaluate the accuracy of the algorithm to make sure the predictions do not match the actual values by chance.

7 References

Anand, H. (2021, December 26). *Categorical cross-entropy loss - the most important loss function*. Medium. Retrieved December 4, 2022, from https://neuralthreads.medium.com/categorical-cross-entropy-loss-the-most-important-loss-function-d3792151d05b

Baeldung. (2021, August 25). Multiclass classification using support Vector Machines. Baeldung on Computer Science. Retrieved October 28, 2022, from https://www.baeldung.com/cs/svm-multiclass-classification

Burns, E., & Burke, J. (2021, March 26). *What is a neural network? explanation and examples*. Enterprise AI. Retrieved December 4, 2022, from https://www.techtarget.com/searchenterpriseai/definition/neural-network

Chaudhary, M. (2020, August 28). *Activation functions: SIGMOID, Tanh, Relu, leaky relu, softmax*. Medium. Retrieved December 4, 2022, from https://medium.com/@cmukesh8688/activation-functions-sigmoid-tanh-relu-leaky-relu-softmax-50d3778dcea5#:~:text=As%20per%20our%20business%20requirement,use%20in%20last%20output%20layer%20.

Clancy, L. (2020, April 16). *Discussing your study's limitations - international science editing examples*. International Science Editing. Retrieved December 4, 2022, from https://www.internationalscienceediting.com/study-limitations/#:~:text=Sample%20size%20limitations,the%20study%20groups%20is%20reported.

Choi, J., Suh, Y., & Jung, N. (2020). Predicting corporate credit rating based on qualitative information of MD&A transformed using document vectorization techniques. *Data Technologies and Applications*, *54*(2), 151–168. https://doi.org/10.1108/dta-08-2019-0127

Data Flair. (2021, March 8). *Kernel functions-introduction to SVM Kernel & Examples*. Retrieved December 4, 2022, from https://data-flair.training/blogs/svm-kernel-functions/

DeepAI. (2019, May 17). *Feed Forward Neural Network*. Retrieved December 4, 2022, from https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network

Delwadia, K. (2022, June 18). *Corporate credit rating with financial ratios*. Kaggle. Retrieved October 22, 2022, from https://www.kaggle.com/datasets/kirtandelwadia/corporate-credit-rating-with-financial-ratios

Dubey, A. (2018, December 15). *Feature selection using Random Forest*. Medium. Retrieved December 4, 2022, from https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

Golbayani, P., Wang, D., & Florescu, I. (2020, March). *Application of deep neural networks to assess corporate credit rating ...* ResearchGate. Retrieved December 4, 2022, from https://www.researchgate.net/publication/339737199_Application_of_Deep_Neural_Networks_to_assess_corporate_Credit_Rating

Han, S. (2019, January 1). Iowa State University Digital Repository. Retrieved October 22, 2022, from https://dr.lib.iastate.edu/handle/20.500.12876/16948/

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support Vector Machines and neural networks: A market comparative study. *Decision Support Systems*, *37*(4), 543–558. https://doi.org/10.1016/s0167-9236(03)00086-1

IBM Cloud Education. (2020, December 7). *What is Random Forest?* IBM. Retrieved October 28, 2022, from https://www.ibm.com/cloud/learn/random-forest

Lee, Y.-C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, *33*(1), 67–74. https://doi.org/10.1016/j.eswa.2006.04.018

Ronaghan, S. (2019, November 1). *The mathematics of decision trees, random forest and feature importance in Scikit-learn and Spark*. Medium. Retrieved December 4, 2022, from https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Savas, C., & Dovis, F. (2019). The impact of different kernel functions on the performance of scintillation detection based on support Vector Machines. *Sensors*, *19*(23), 5219. https://doi.org/10.3390/s19235219

Sreenivasa, S. (2020, October 12). *Radial basis function (RBF) kernel: The go-to kernel*. Medium. Retrieved December 4, 2022, from https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a

Wang, M., & Ku, H. (2021). Utilizing historical data for Corporate Credit Rating Assessment. *Expert Systems with Applications*, *165*, 113925. https://doi.org/10.1016/j.eswa.2020.113925

Wallis, M., Kumar, K., & Gepp, A. (2019). Credit rating forecasting using Machine Learning Techniques. *Advances in Data Mining and Database Management*, 180–198. https://doi.org/10.4018/978-1-5225-7277-0.ch010