University of St Gallen

School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science

Predicting Credit Ratings using Machine Learning

Lev Akhmerov
lev.akhmerov@student.unisg.ch
Matriculation Number: 22-616-072

Joshua Libon
joshua.libon@student.unisg.ch
Matriculation Number: 23-612-435

Noa Diego Frei
noadiego.frei@student.unisg.ch
Matriculation Number: 21-946-413

Julius Everwand
juliusvincenzbenedict.everwand@student.unisg.ch
Matriculation Number: 22-619-647

Leo Koch
leo.koch@student.unisg.ch
Matriculation Number: 22-611-578

Machine Learning in Finance
University of St. Gallen
Prof. Dr. Despoina Makariou
18 October 2024

Abstract

This paper develops an understanding of machine learning in the context of predicting Standard & Poor's (S&P) credit ratings of commercial entities by use of financial ratios. By describing the data collection process and research design the paper lays the foundation for the exploration of three learning models, namely random forest, linear regression, and neural networks. The paper has found that the random forest model performed best on the data set, followed by the linear regression and lastly the neural network. Factors such as qualitative decision-making processes at the crediting agency, imbalanced data, false implication of output linearity and lack of hyperparameter tuning impacted the performance of the models and further research is suggested to obtain an improved understanding of their impact on the algorithms.

Table of Contents

List of Figures

List of Tables

1. Introduction

Credit ratings are used by market participants to understand the degree of risk associated with investing in or lending money to an entity. The S&P rating, awarded by Standard & Poor's Global Rating, is among the most recognized credit rating systems in the financial world - it assesses the creditworthiness of governments, companies, and financial instruments, helping investors make informed decisions based on default probability. These ratings play a significant role in describing the reliability of financial activity as they affect borrowing costs and set the perceived financial stability of investment opportunities. Apart from qualitative processes, S&P Global uses financial data extensively to calculate risk and assign a rating. The exact distribution of factors used to generate these ratings is unknown and is likely an extensive process, presenting an excellent opportunity for research and conjecture (S&P Global, n.d.).

Machine learning employs various statistical approaches to explore relationships within datasets. With both explanatory and predictive capabilities it can be useful in supporting financial decision making. As credit ratings are often considered in such decisions, the ability to understand their appointment should prove useful for both investors and entities. Given the emphasis on financial topics in this course, the limitations arising from finding a dataset valid for supervised learning, and strong connection between financial information and credit ratings, a research topic emerges: *how can machine learning help understand and predict S&P credit ratings based on a commercial entity's financial ratios?* Through research, data analysis, and code work a predictive algorithm may be developed to observe and discuss how strongly standard financial ratios influence credit ratings. This is a regression problem with a significant potential for practical use in the financial industry – a superb topic to gain deeper insights into machine learning in finance.

2. Study Design and Financial Data Collection

The prediction of S&P credit ratings based on financial ratios was examined using three separate machine learning models: random forest, neural networks, and linear regression. The aim is to compare the performance of these models whilst concurrently gaining a better understanding of the data and the models themselves. The data set is used to explore the topic as the three models are developed and tested. After modeling, the results are compared using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R^2. The use of cross-validation ensures that the models are robust and have not been overfitted. The goal is to identify the model that most accurately predicts credit ratings using financial ratios.

The data was obtained from the University of Pennsylvania's Wharton Research Data Services (WRDS). The output consists of credit ratings, which were extracted from the Standard & Poor's Compustat

Capital IQ database within WRDS. This data set covers the period from December 2016 to February 2017. Based on that, the financial ratios (input data) were collected using the in-house WRDS suite for the end of the fiscal year 2016. The output data was appended to the input data using the company ticker symbols. The ratings were assigned a numerical score between 1-22 for processing in learning models. Lastly, a series of irrelevant predictors (financial ratio columns) were removed followed by the removal of all rows with blank values. This order of operations allowed for a larger cleaned dataset as removing unnecessary predictors decreased the number of blank cells. This left the data set with 573 observations and 52 predictors, a glimpse of which is seen in Table 1.

Quick Ratio (Acid Test)	Current Ratio	Sales/Inv. Capital	Sales/Stock holders Equity	R&D/ Sales	Advertising Exp./ Sales	Labor Exp./ Sales	Accruals/ Avg. Assets	Ticker Symbol	Rating	Score
1,03	2,70	1,71	1,99	0,00	0,00	0,00	-0,04	AIR	BB+	12
0,67	0,74	1,61	8,40	0,00	0,00	0,26	-0,01	AAL	BB-	10
0,45	0,65	0,41	0,75	0,00	0,00	0,00	-0,04	PNW	A-	16
1,21	1,52	0,77	0,99	0,07	0,00	0,00	-0,05	ABT	BBB	14

Tab. 1: Extract of cleaned and joined dataset

The heatmap (Figure 1) depicts the covariance of the input variables. As expected, there is a high correlation between variables derived from similar positions in the financial statements. For instance, there is a correlation between different measures of debt, profit, and other corresponding margins. Multicollinearity makes it difficult to interpret the significance of individual predictors, as high correlations between variables can weaken the predictive power of models. In extreme cases, this can lead to the model not providing any meaningful or reliable results despite a high level of fit. To avoid this, it is important to identify variables that contain redundant information and perform a merging or removal thereof. In our example, most of the covariates are largely independent thanks to the initial data cleaning and there is no strong multicollinearity that could distort the results. This suggests that the predictor data set is well suited to train machine learning models.

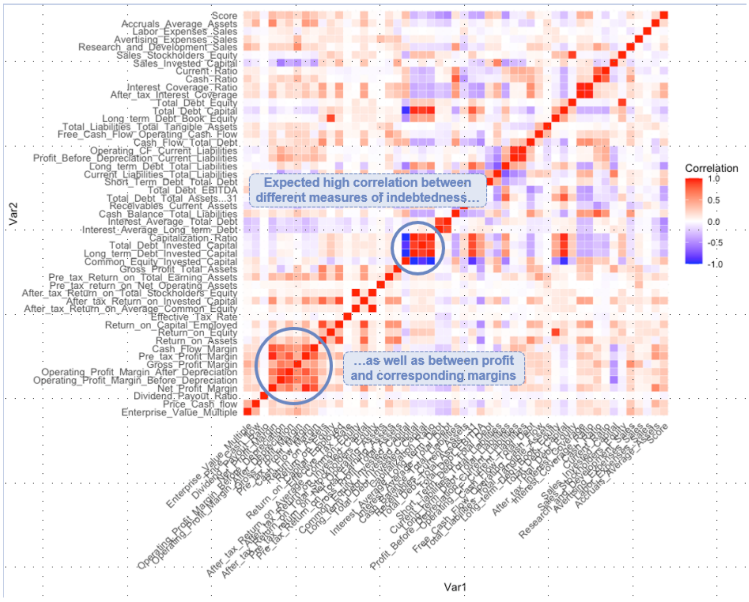

Fig. 1: Heatmap of variable covariances

The output data roughly mimics a normal distribution with certain exceptions. In particular, the ratings BBB- and BB+ are outliers and are found infrequently (Figure 2). Only several observations are recorded below B- with some data points in the default area. The imbalance in rating observations is a challenge for accurate modeling of the learning algorithm and literature suggests several approaches to address this problem and improve prediction accuracy. An example is to increase the number of underrepresented ratings using methods such as SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002). Another option is to assign higher weights to the underrepresented ratings. This approach helps to improve the model by simulating the expected normal distribution of the ratings. It is particularly effective for models such as random forest and neural networks that are sensitive to weights (Japkowicz & Stephen, 2002).

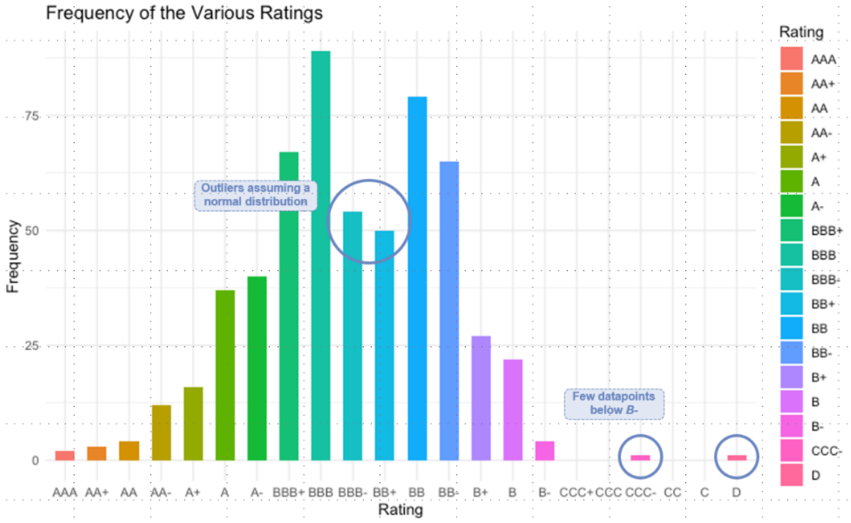

Fig. 2: Distribution of ratings

3. Random Forest

3.1 Architecture

The main model used in the research is random forest – an ensemble method which derives subset-based decision trees using bootstrap sampling, grows the trees using recursive partitioning, and aggregates these estimators to produce a prediction. The difference between bagging, the aforementioned process, and random forests lies in decorrelating the individual trees from one another by only conducting node splits using a subset of *m* randomly selected predictors where $m_{try} \subseteq p$. This accounts for a reduction in variance and bias. Since bootstrap sampling inherently renders observation in- and out of bag (OOB), a separate test set is not required as trees can use OOB observations. Lastly, the method's performance is evaluated with common methods including permutation importance, a process of removing variables and measuring the positive change in MSE, and minimal depth, where early contributions of a predictors within a respective tree indicate their predictive power.

3.2 Approach

To carry out the random forest, the data required a final clean predominantly in the realm of formatting. Given the linear transformation of the scores to numerical values with a linear-ordinal nature, it was important to distinguish the task as a regression. The first step now was to experiment with the default parameters set in R followed by a 10-fold cross validation with a 70-30 train to test split to tune our hyperparameters (Figure 3). The results revealed an optimal m_{try} of 16 which is very close to the generally regarded rule of thumb value *p/3* (52/3=17.33). A highly volatile outcome was observed in the n_{tree} cross validation, reaching an optimum at 700. As for node size, a default value of 5 was utilized. The final random forest was executed using these hyperparameters, the most significant predictors were exposed by the permutation importance and minimal depth tests, and the MSE and R^2 values were recorded for comparison with other models. As this was a regression task, no confusion matrix was produced and MSE was the main comparison metrics between the models.

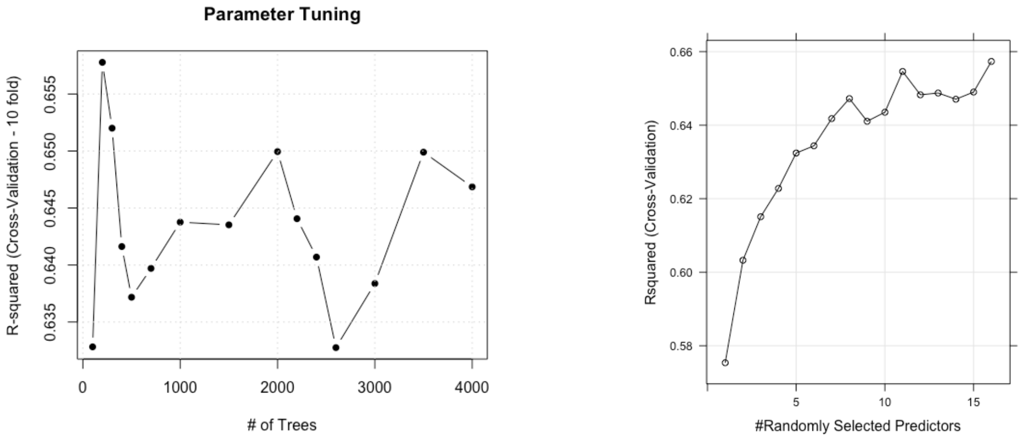

Fig. 3: Hyperparameter tuning using cross validation for tree numbers and predictor subset

3.3 Results & Analysis

The results demonstrate the prevalence of the random forest method in the research as the best performing model. With an MSE of ≈ 3.458, RMSE of ≈ 1.859, and a final R^2 value of ≈ 0.608, the model performed well with the data. The 60% explanatory power in variance (R^2) is a medium-to-high and the low RMSE suggests compared to the scale utilized suggests relatively accurate predictive capabilities of the model. Table 2 shows an extract of some of the predictions along with the real value and the delta (RMSE). Both the permutation importance and minimal depth seen Figure 4 concluded that the dividend payout ratio, interest coverage measures, and long-term debt ratios proved to be the most significant predictors in the model – a logical outcome when it comes to credit ratings.

Actual Score	Predicted Score	Delta
16	15.40295	-0.59705
17	16.67764	-0.32236
11	12.22767	1.22767
21	16.70252	-4.29748
13	11.58979	-1.41021
14	14.01557	0.01557
13	13.26455	0.26455
15	13.04310	-1.95690
14	12.17752	-1.82248

Tab. 2: Extract of random forest prediction results

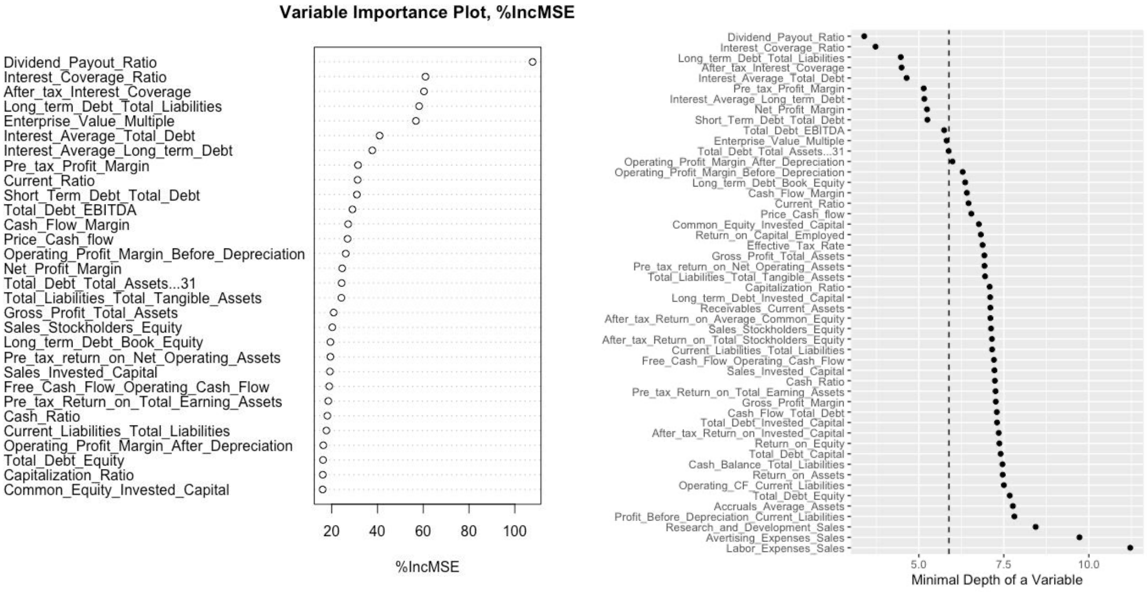

Fig. 4: Permutation importance and minimal depth test results

3.4 Limitations & Outlook

Random forest is a robust model and works well to counter variance, bias, overfitting, and correlation within the predictor space. Due to these qualities, the model is used extensively in the financial industry. The data in this research could be improved by introducing a uniform distribution of outputs – the scarcity of data for extremely highly or poorly rated commercial entities limits the proficiency of the model and results in higher variance as seen in row 4 of Table 2. For a real-world application, it would be beneficial to have more poorly rated examples to be able to predict credit defaults in advance. In practice, however, these cases occur highly infrequently causing a significant imbalance problem. The group attempted to resolve this issue by applying weightings to minority ratings and setting a threshold to observations to create a uniform data distribution. This resulted in a small dataset that posed complications pertaining to hyperparameter tuning and the imbalance issue was not examined further. Research on this characteristic is required as it is common in practice and would vastly improve the predictive capabilities of models testing normal distributions in real scenarios.

4. Linear Regression

4.1 Architecture

Linear regression is a widely used statistical technique for modeling the linear relationship between dependent and independent variables and is represented by the equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \text{ for } i = 1, \ldots, n.$$

The primary objective of linear regression is to identify the 'best fit' line that minimizes the sum of the squared differences between the observed data points and the predicted values (Sonkavde et al., 2023). This is achieved through various error metrics, including mean squared error (MSE), and R-squared (Dospinescu et al., 2019). These metrics are essential for assessing the accuracy of the model's fit to the data. While linear regression is effective for data sets with linear relationships, it can introduce bias when the underlying data patterns are non-linear (James et al., 2021).

4.2 Approach

The methodology for implementing linear regression in R follows a systematic approach. First, feature selection is performed using a random forest algorithm to identify the most relevant features that significantly increase the predictive power of the model. The linear regression model developed in this project includes the following seven predictors:

- Dividend Payout Ratio, Interest Coverage Ratio, After-tax Interest Coverage, Long-term Debt/Total Liabilities, Enterprise Value Multiple, Interest/Average Total Debt, Interest/Average Long-term Debt.

The linear regression model is trained using the selected features and the training data set. To evaluate the accuracy and predictive ability of the model, the Mean Squared Error (MSE) is calculated on the test dataset. In addition, a 10-fold CV is performed to calculate and reduce the variability of the resulting test MSE.

4.3 Results & Analysis

The results presented in Table 3 show that the linear regression model effectively predicts the S&P credit ratings, as evidenced by a low mean squared error (MSE) of 5.03, obtained from 10-fold cross-validation, and a root mean squared error (RMSE) of 2.24. These metrics indicate that the model has good accuracy and fit, suggesting that a significant portion of the variance in credit ratings can be explained by the selected financial ratios. The RMSE shows that the approximate average distance between the predicted value and the true value is 2.24, which is low considering that we have 22 different credit scores.

10-fold CV Test MSE	**5.030737**
RMSE	**2.242930**

Tab. 3: Results of linear regression

4.4 Limitations and outlook

Despite the strengths of the analysis, there are some limitations. Linear regression inherently assumes a linear relationship between the independent variables (financial ratios) and the dependent variable (credit ratings), which may not adequately reflect the complexity of financial markets. While the model provided reasonable predictions, its simplicity may miss non-linear patterns that could improve accuracy in more complex financial environments.

In addition, newer machine learning techniques, including random forests and neural networks, are increasingly used in financial forecasting. These methods are particularly adept at capturing non-linear relationships between variables and offer potential improvements over linear models. However, their effectiveness depends heavily on appropriate tuning and the availability of sufficient training data.

5. Neural Network

5.1 Architecture

As an additional approach, a neural network is employed as an advanced method for data analysis. Inspired by the functioning of the human brain, neural networks are composed of interconnected neurons (also referred to as nodes), which are structured in layers to process information (Di Franco & Santurro, 2021). This architectural configuration enables computers to analyse data in a manner that emulates human cognitive processes.

The neurons are organized into multiple layers. The initial layer, designated the input layer, is tasked with receiving the unprocessed data. The information is then conveyed through one or more hidden layers, where patterns and relationships are identified, before reaching the output layer. This layered structure enables neural networks to learn from data.

5.2 Approach

We employed a structured approach to develop and evaluate a neural network model using R. The initial stage involved data preprocessing, whereby the features were scaled to a common range to enhance the model's training efficiency and performance. This normalization step is of critical importance, as it serves to mitigate the impact of varying scales across different features. Subsequently, the data was partitioned into a training and testing subset, with 70% of the data allocated for training and 30% for testing.

The neural network model was constructed using a specified architecture with multiple layers (Figure 5). The input layer consisted of 50 neurons and served as the starting point for data processing. Two hidden layers were then incorporated into the model: the first hidden layer contained five neurons, while the second hidden layer contained three neurons. The final layer of the network was a single neuron responsible for generating the predicted output. This configuration, with distinct input, hidden, and output layers, enabled the model to effectively process complex data patterns and make accurate predictions.

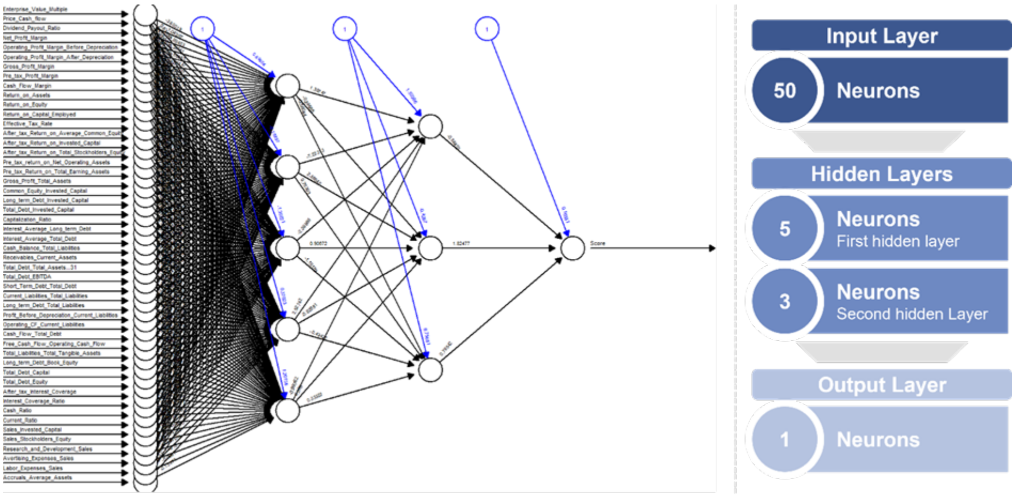
Enterprise_Value_Multiple
Price_Cash_flow
Dividend_Payout_Ratio
Net_Profit_Margin
Operating_Profit_Margin_Before_Depreciation
Operating_Profit_Margin_After_Depreciation
Gross_Profit_Margin
Pre_tax_Profit_Margin
Cash_Flow_Margin
Return_on_Assets
Return_on_Equity
Return_on_Capital_Employed
Effective_Tax_Rate
After_tax_Return_on_Average_Common_Equity
After_tax_Return_on_Invested_Capital
After_tax_Return_on_Total_Stockholders_Equity
Pre_tax_return_on_Net_Operating_Assets
Pre_tax_Return_on_Total_Earning_Assets
Gross_Profit_Total_Assets
Common_Equity_Invested_Capital
Long_term_Debt_Invested_Capital
Total_Debt_Invested_Capital
Capitalization_Ratio
Interest_Average_Long_term_Debt
Interest_Average_Total_Debt
Cash_Balance_Total_Liabilities
Receivables_Current_Assets
Total_Debt_Total_Assets...31
Total_Debt_EBITDA
Short_Term_Debt_Total_Debt
Current_Liabilities_Total_Liabilities
Long_term_Debt_Total_Liabilities
Profit_Before_Depreciation_Current_Liabilities
Operating_CF_Current_Liabilities
Cash_Flow_Total_Debt
Free_Cash_Flow_Operating_Cash_Flow
Total_Liabilities_Total_Tangible_Assets
Long_term_Debt_Book_Equity
Total_Debt_Capital
Total_Debt_Equity
After_tax_Interest_Coverage
Interest_Coverage_Ratio
Cash_Ratio
Current_Ratio
Sales_Invested_Capital
Sales_Stockholders_Equity
Research_and_Development_Sales
Avertising_Expenses_Sales
Labor_Expenses_Sales
Accruals_Average_Assets
1
1
1
Score
Input Layer
50
Neurons
Hidden Layers
5
Neurons
First hidden layer
3
Neurons
Second hidden Layer
Output Layer
1
Neurons

Fig. 5: Neural Network

Once the model had been trained, an assessment of its predictive accuracy was conducted by calculating the mean squared error (MSE) on the test dataset.

Moreover, a plot of the predicted values against the true values was constructed to conduct a visual assessment of the model's performance (Figure 6).

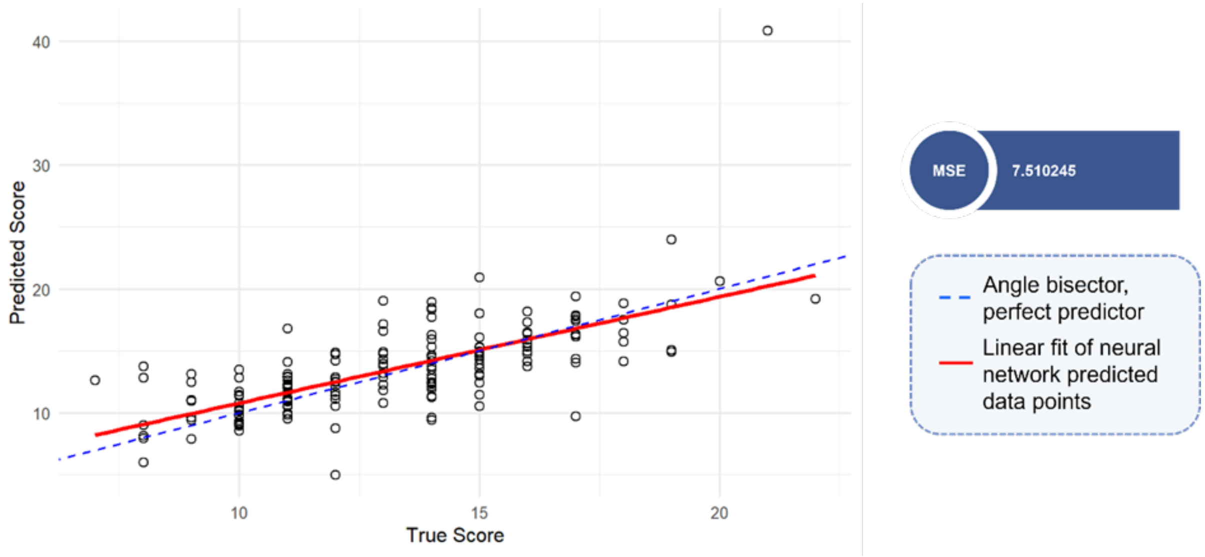

Fig. 6: True vs Predicted Values (70% Train and 30% Test)

To enhance the robustness of our findings, we implemented 10-fold cross-validation. This method entailed repeatedly training the model on different subsets of the data, thereby reducing the variability of the 10-fold test MSE.

5.3 Results & Analysis

The neural network model achieved a 10-fold test mean-squared error (MSE) of ≈ 5.203, as shown in Figure 7. The comparison of predicted and true scores shows a general upward trend, suggesting that the model is capturing some underlying patterns in the data. However, there are notable deviations from the ideal prediction line (represented by the blue dashed line), particularly at the extremes of the rating spectrum.

In addition, the red linear fit line in the graph further illustrates the performance of the model. It shows that while the predictions are generally in line with the true scores, they tend to flatten out at higher rating levels. This flattening indicates that the model has difficulty distinguishing between higher ratings, such as AA and AAA, which is a common challenge due to the ordinal nature of credit rating.

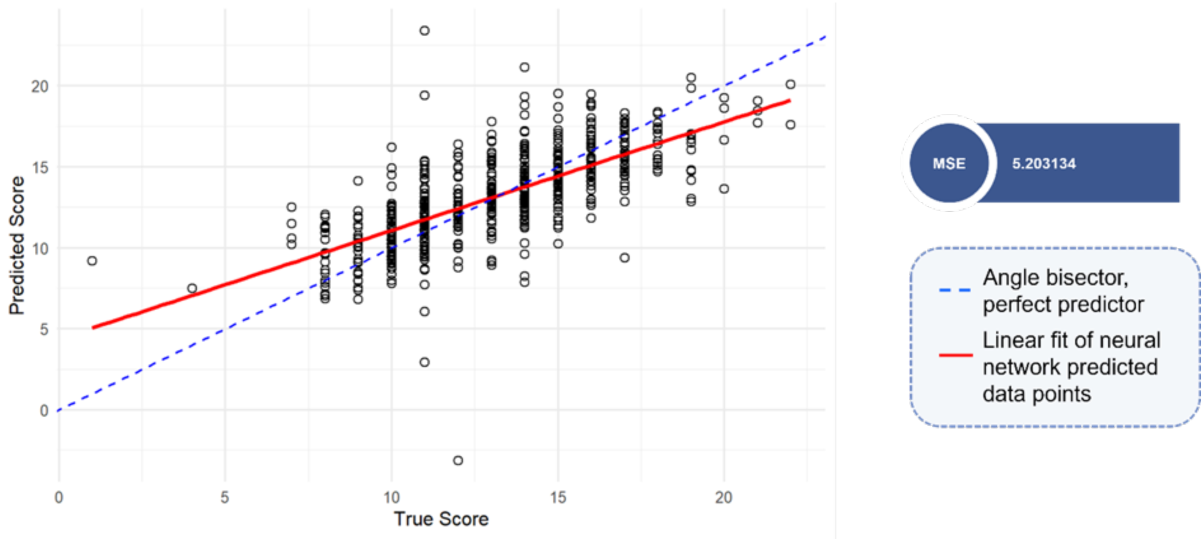

Fig. 7: True vs Predicted Values (10-Fold CV)

5.4 Limitations & Outlook

One shortcoming of the existing neural network model is the absence of hyperparameter tuning. In this study, the model architecture, and parameters, including the number of neurons, learning rate, and activation functions, were predefined without systematic optimization. This approach may have constrained the capacity of the model to achieve its full potential. Applying hyperparameter tuning techniques, such as Bayesian optimization, could enhance model performance by identifying the optimal settings for the network architecture and training process.

The architecture of the neural network used in this study is relatively simple, comprising only two hidden layers. While this configuration is sufficient for basic tasks, more complex architectures could better capture non-linearities in the data. However, such models also carry the risk of overfitting, particularly when trained on smaller datasets.

Moreover, one of the core limitations of neural networks, which is also applicable to this study, is their intrinsic "black box" nature, which impairs interpretability. Consequently, it is challenging to grasp the rationale behind the model's predictions, which can be problematic in sensitive domains such as credit rating, where it is vital for stakeholders to comprehend the rationale behind each prediction.

6. Conclusion

Regarding the performance of the various machine learning models tested to predict the *S&P Credit Ratings* of commercial entities, the random forest model fared best in terms of accuracy. In the end, the model had an MSE of c. 3.457 with a R^2 of c. 60.8%, inferring high explanatory power (Chapter 3). In comparison, the MSE of the linear regression and neural network were ≈ 5.031 and ≈ 5.203 respectively, both obtained by 10-fold CV (Chapter 4 & 5). Despite good overall performance, the neural network shows notable deviations from the ideal prediction line.

Under normal circumstances, the higher a model's complexity, the greater its performance should be, leading one to expect the best performance from the more complex neural network. However, this assumption does not hold true, likely because the random forest model is most robust against the imbalanced dataset, making it the most accurate model. To resolve this issue, further research could train the models on a uniform distribution by applying weightings and setting thresholds.

A critical aspect of our project's limitations is the numerical representation of the ordinal ratings. In the current implementation, credit ratings are assigned integer values, implying that the difference between consecutive ratings is always exactly one. In practice, however, the differences between ratings and, thus, between different creditworthiness levels are not clearly defined and, therefore, not always exactly one. This allocation of numbers to ratings can affect the accuracy of the predictions because the model does not know the actual differences between the various ratings.

In the broader context of credit rating prediction, alternative statistical methods such as ordered logistic regressions and ordered probit models have often performed better. These methods consider the ordinal nature of creditworthiness, which is structured hierarchically from higher to lower scores, which can lead to more accurate predictions (Gutierrez et al., 2016; Hwang, 2013; Hwang et al., 2010). In addition, a heteroscedasticity adjustment could be made to account for the different variances of the ratings and better adapt the model to the actual distribution of creditworthiness, which would enable the model to react more flexibly to the non-linear differences between the ratings (Steyerberg, 2016).

Another major limitation of the project is that the algorithm is based only on numerical and quantitative financial indicators. In practice, *S&P Credit Ratings* also include qualitative data, such as a company's market position, the quality of its management, industry-specific risks, and macroeconomic developments. These qualitative assessments often influence the final credit rating but are difficult to represent. One possible solution would be to use natural language processing to integrate qualitative information from reports, company analyses and business news into the model. Such an extension could make the model more accurate, but it remains difficult to model the expertise and subjective assessments of rating agencies fully.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *The Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Di Franco, G., & Santurro, M. (2021). Machine learning, artificial neural networks and social research. *Quality & Quantity, 55*(3), 1007-1025.

Dospinescu, N., & Dospinescu, O. (2019). A profitability regression model in financial communication of Romanian stock exchange's companies. *Ecoforum Journal, 8*(1).

Gutierrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., & Hervas-Martinez, C. (2016). Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering, 28*(1), 127–146. https://doi.org/10.1109/TKDE.2015.245791

Hwang, R. C. (2013). Forecasting credit ratings with the varying-coefficient model. *Quantitative Finance, 13*(12), 1947-1965.

Hwang, R. C., Chung, H., & Chu, C. K. (2010). Predicting issuer credit ratings using a semiparametric method. *Journal of Empirical Finance, 17*(1), 120-137.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Linear regression. In *An introduction to statistical learning*. New York: Springer.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449. https://doi.org/10.3233/ida-2002-6504

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies, 11*(3), 94.

S&P Global. (n.d.). Understanding credit ratings. https://www.spglobal.com/ratings/en/about/understanding-credit-ratings

Steyerberg, E. W. (2016). Frank E. Harrell, *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Heidelberg: Springer. *Biometrics, 72*(3), 1006–1007. https://doi.org/10.1111/biom.12569

Declaration of Authorship

I hereby declare

- that I have written this term paper without any help from others and without the use of documents or aids other than those stated above;
- that I have mentioned all the sources used and that I have cited them correctl y according to established academic citation rules;
- that I have acquired any immaterial rights to materials I may have used, such as images or graphs, or that I have produced such materials myself;
- that the topic or parts of it are not already the object of any work or examination of another course unless this has been explicitly agreed to with the faculty member in advance and is referred to in the term paper;
- that I will not pass on copies of this work to third parties or publish them without the university's written consent if a direct connection can be established with the University of St.Gallen or its faculty members;
- that I am aware that my work can be electronically checked for plagiarism and that I hereby grant the University of St.Gallen copyright in accordance with the Examination Regulations insofar as this is required for administrative action;
- that I am aware that the university will prosecute any infringement of this declaration of authorship and, in particular, the employment of a ghostwriter, and that any such infringement may result in disciplinary and criminal consequences which may result in my expulsion from the university or my being stripped of my degree

By uploading this academic term paper, I confirm through my conclusive action that I am submitting the Declaration of Authorship, that I have read and understood it, and that it is true.

Julius V. B. Everwand J. Libon

Lev Akhmerov, Julius Everwand, Joshua Libon, Leo Koch, Noa Diego Frei