S&P CREDIT RATINGS

MACHINE LEARNING IN FINANCE

S&P CREDIT RATINGS

Introduction

Our team

Lev Akhmerov

lev.akhmerov@student.unisg.ch

Julius Everwand

juliusvincenzbenedict.everwand@student.unisg.ch

Noa Diego Frei

noadiego.frei@student.unisg.ch

Leo Koch

leo.koch@student.unisg.ch

Joshua Libon

joshua.libon@student.unisg.ch

Introduction to Credit Ratings

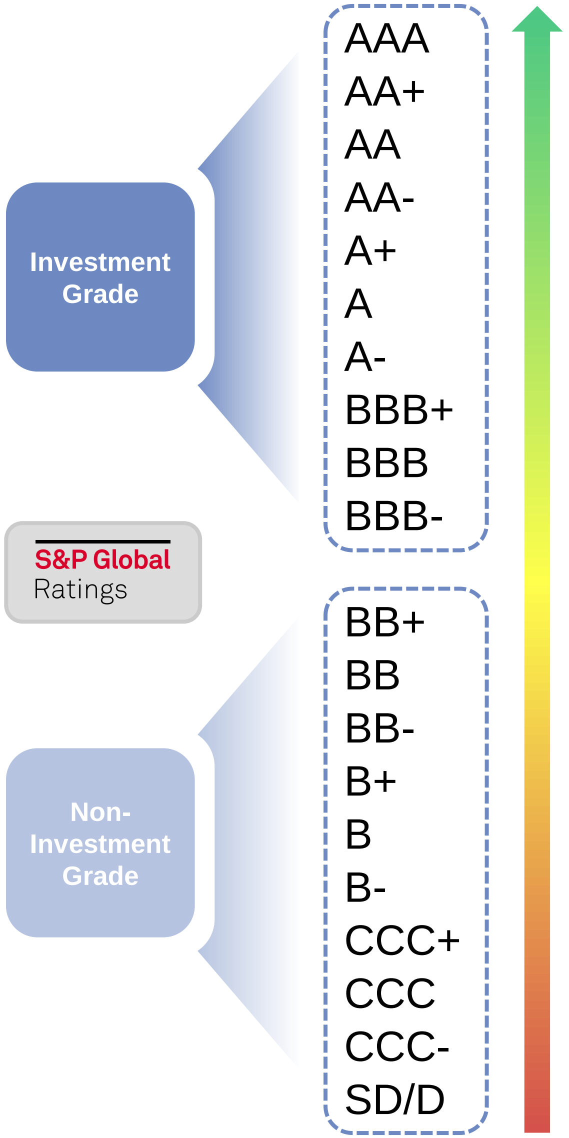
Credit Risk

- Risk that the borrower will default, resulting in a financial loss of the lender
- Affects the pricing of debt securities, interest rates and overall market stability

Credit Ratings

- Assessment of the credit risk by rating agencies to inform investors
- Scoring the likelihood of default
- Aid for pricing bonds

Accurate **bond ratings** and effective **credit risk management** are vital for ensuring **financial stability** and **minimizing potential losses**



S&P CREDIT RATINGS

Dataset

Data Collection

Data Collection

Wharton wrds
UNIVERSITY of PENNSYLVANIA
http://whartonwrds.com

We fetched data from University of Pennsylvania's Wharton Research Data Services (WRDS), accessible to HSG students.

Input

For the financial ratios, we used WRDS's native suite. Data was pulled for the end of CY 2016.

Output

For the credit ratings, we used Standard and Poor's Compustat - Capital IQ database. The data is from a time frame from 12/2016 to 02/2017.

Input and output are matched on the company ticker

A score is assigned linearly to each rating

Avertising	Labor Exp	Accruals/A	Ticker Sy	Rating	Score
0,00	0,00	-0,04	AIR	BB+	12
0,00	0,26	-0,01	AAL	BB-	10
[illegible]	[illegible]	[illegible]	PNW	A-	16
0,00	0,00	-0,05	ABT	BBB	14
0,00	0,08	-0,05	AET	A	17
0,00	0,00	-0,06	APD	A	17
0,01	0,25	-0,09	ALK	BB+	12
0,00	0,00	-0,01	HON	A	17
0,00	0,00	-0,05	ECOL	BB	11
0,00	0,00	-0,06	AEP	A-	16
0,00	0,00	-0,02	AME	BBB+	15
0,00	0,00	-0,02	AMGN	A	17
0,00	0,00	-0,06	ADI	BBB	14
0,00	0,00	-0,07	AAPL	AA+	21
0,00	0,00	-0,05	AMAT	A-	16
0,00	0,00	-0,02	ADM	A	17
0,00	0,00	-0,01	ARW	BBB-	13
0,01	0,00	-0,07	ASH	BB	11
0,00	0,00	-0,08	ATW	B-	7
0,00	0,00	-0,01	ADP	AA	20
0,00	0,00	-0,05	AVY	BBB	14
0,00	0,00	0,01	AVT	BBB-	13
0,00	0,00	0,01	BCR	A	17
0,00	0,00	0,18	BAX	A-	16
0,00	0,00	-0,06	BDX	BBB+	15
0,02	0,00	-0,06	VZ	BBB+	15
0,00	0,00	-0,07	BMS	BBB	14
0,02	0,00	-0,08	BBY	BBB-	13
0,00	0,00	-0,02	BIO	BBB	14
0,00	0,00	-0,04	BKH	BBB	14
0,10	0,35	-0,02	HRB	BBB	14
0,00	0,00	-0,07	BA	A	17
0,05	0,00	0,04	BMY	A+	18
0,03	0,00	-0,09	CAL	BB	11

Dataset Statistics

Variable	Mean	Median	Standard Dev.	Minimum	Maximum
Enterprise_Value_Multiple	11.54	10.74	4.84	-23.58	44.86
Price_Cash_flow	12.32	11.26	7.90	-49.60	71.21
Dividend_Payout_Ratio	0.79	0.30	4.90	0.00	102.12
Net_Profit_Margin	0.08	0.07	0.12	-1.69	0.49
Operating_Profit_Margin_Before_Depreciation	0.21	0.19	0.12	-0.52	0.70
Operating_Profit_Margin_After_Depreciation	0.14	0.14	0.12	-1.37	0.66
Gross_Profit_Margin	0.39	0.36	0.19	-0.36	0.97
Pre_tax_Profit_Margin	0.11	0.10	0.14	-1.50	0.72
Cash_Flow_Margin	0.15	0.13	0.13	-1.58	0.66
Return_on_Assets	0.14	0.13	0.06	-0.07	0.42
Return_on_Equity	0.22	0.12	0.73	-3.30	15.50
Return_on_Capital_Employed	0.16	0.13	0.10	-0.21	1.00
Effective_Tax_Rate	0.14	0.31	2.16	-48.73	3.85
After_tax_Return_on_Average_Common_Equity	0.44	0.14	5.27	-39.28	106.71
After_tax_Return_on_Invested_Capital	0.11	0.09	0.09	-0.55	0.51
After_tax_Return_on_Total_Stockholders_Equity	0.44	0.14	5.28	-40.20	106.71
Pre_tax_return_on_Net_Operating_Assets	0.49	0.25	4.64	-51.74	92.33
Pre_tax_Return_on_Total_Earning_Assets	0.19	0.17	0.15	-0.18	1.12
Gross_Profit_Total_Assets	0.30	0.27	0.18	-0.16	1.06
Common_Equity_Invested_Capital	0.51	0.53	0.20	-0.20	0.98
Long_term_Debt_Invested_Capital	0.48	0.45	0.20	0.02	1.38
Total_Debt_Invested_Capital	0.53	0.50	0.23	0.04	1.62
Capitalization_Ratio	0.49	0.46	0.21	0.02	1.81
Interest_Average_Long_term_Debt	0.05	0.05	0.02	0.01	0.18
Interest_Average_Total_Debt	0.05	0.04	0.02	0.01	0.13
Cash_Balance_Total_Liabilities	0.18	0.09	0.26	0.00	2.94

A **large variety of differing regressors** are being tested, which are key indicators from corporate finance and can possibly predict credit risk associated with the corresponding firm.

Dataset Statistics (cont‘d)

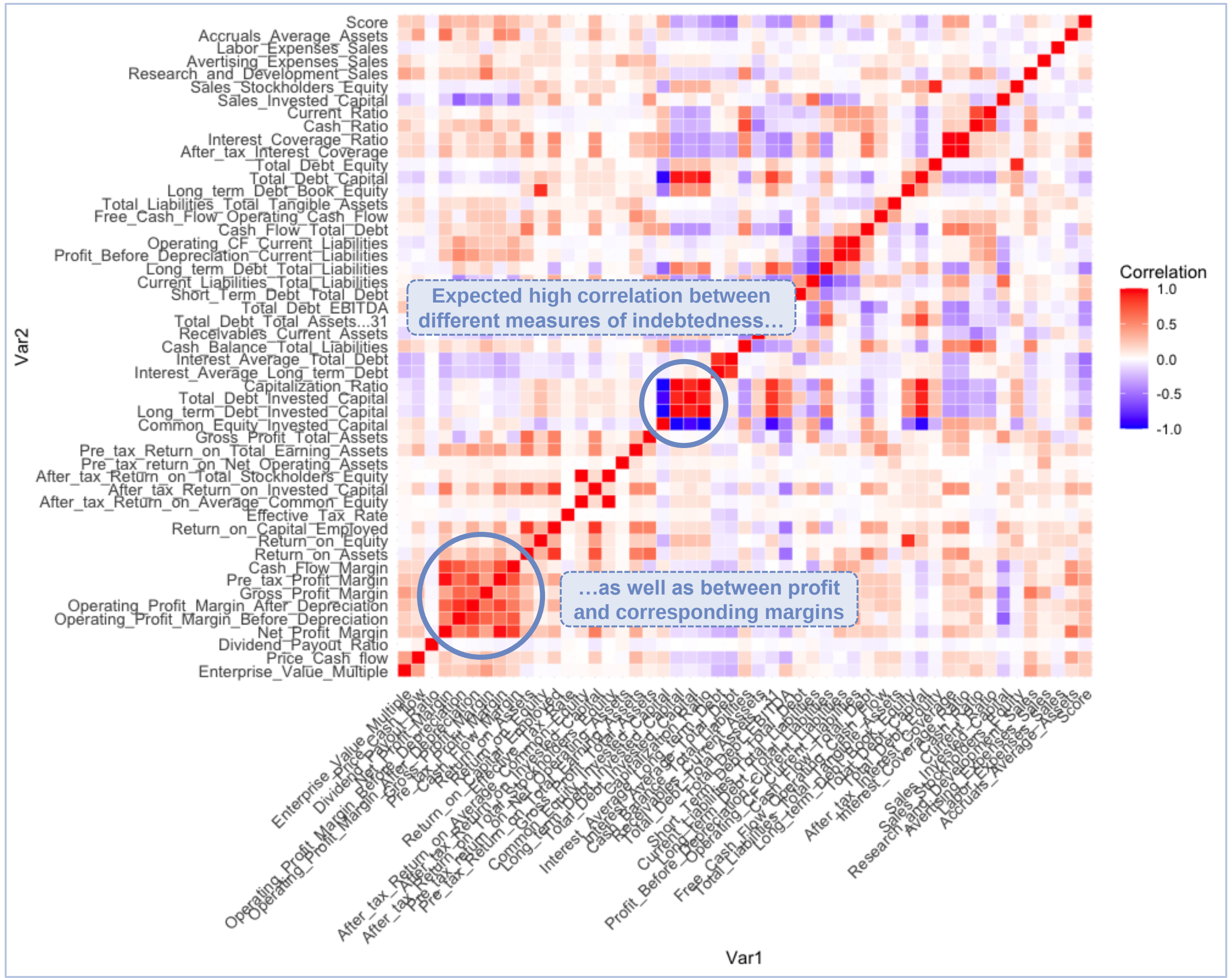
Variable	Mean	Median	Standard Dev.	Minimum	Maximum
Receivables_Current_Assets	0.36	0.36	0.19	0.00	0.91
Total_Debt_Total_Assets	0.35	0.33	0.15	0.03	0.90
Total_Debt_EBITDA	2.92	2.57	1.77	-4.10	18.71
Short_Term_Debt_Total_Debt	0.09	0.05	0.11	0.00	0.73
Current_Liabilities_Total_Liabilities	0.31	0.28	0.17	0.03	0.92
Long_term_Debt_Total_Liabilities	0.48	0.48	0.17	0.05	0.94
Profit_Before_Depreciation_Current_Liabilities	0.92	0.80	0.72	-1.32	11.41
Operating_CF_Current_Liabilities	0.72	0.60	0.68	-0.10	12.66
Cash_Flow_Total_Debt	0.18	0.15	0.11	0.00	1.23
Free_Cash_Flow_Operating_Cash_Flow	0.54	0.67	0.39	-1.33	1.00
Total_Liabilities_Total_Tangible_Assets	6.17	3.26	9.51	0.32	101.48
Long_term_Debt_Book_Equity	1.48	0.73	3.96	0.02	82.11
Total_Debt_Capital	0.56	0.54	0.19	0.05	1.52
Total_Debt_Equity	5.00	1.89	27.40	-52.91	468.71
After_tax_Interest_Coverage	7.05	4.42	11.01	-9.83	163.97
Interest_Coverage_Ratio	9.66	6.23	13.79	-12.74	191.94
Cash_Ratio	0.60	0.34	0.78	0.00	5.58
Current_Ratio	1.83	1.60	1.06	0.38	7.96
Sales_Invested_Capital	1.50	1.08	1.40	0.11	11.31
Sales_Stockholders_Equity	6.29	2.28	27.28	0.17	538.60
Research_and_Development_Sales	0.03	0.00	0.06	0.00	0.40
Avertising_Expenses_Sales	0.01	0.00	0.03	0.00	0.32
Labor_Expenses_Sales	0.03	0.00	0.09	0.00	0.64
Accruals_Average_Assets	-0.05	-0.05	0.04	-0.54	0.18
Score	13.08	13.00	2.95	1.00	22.00

A **large variety of differing regressors** are being tested, which are key indicators from corporate finance and can possibly predict credit risk associated with the corresponding firm.

Covariance of Input Variables

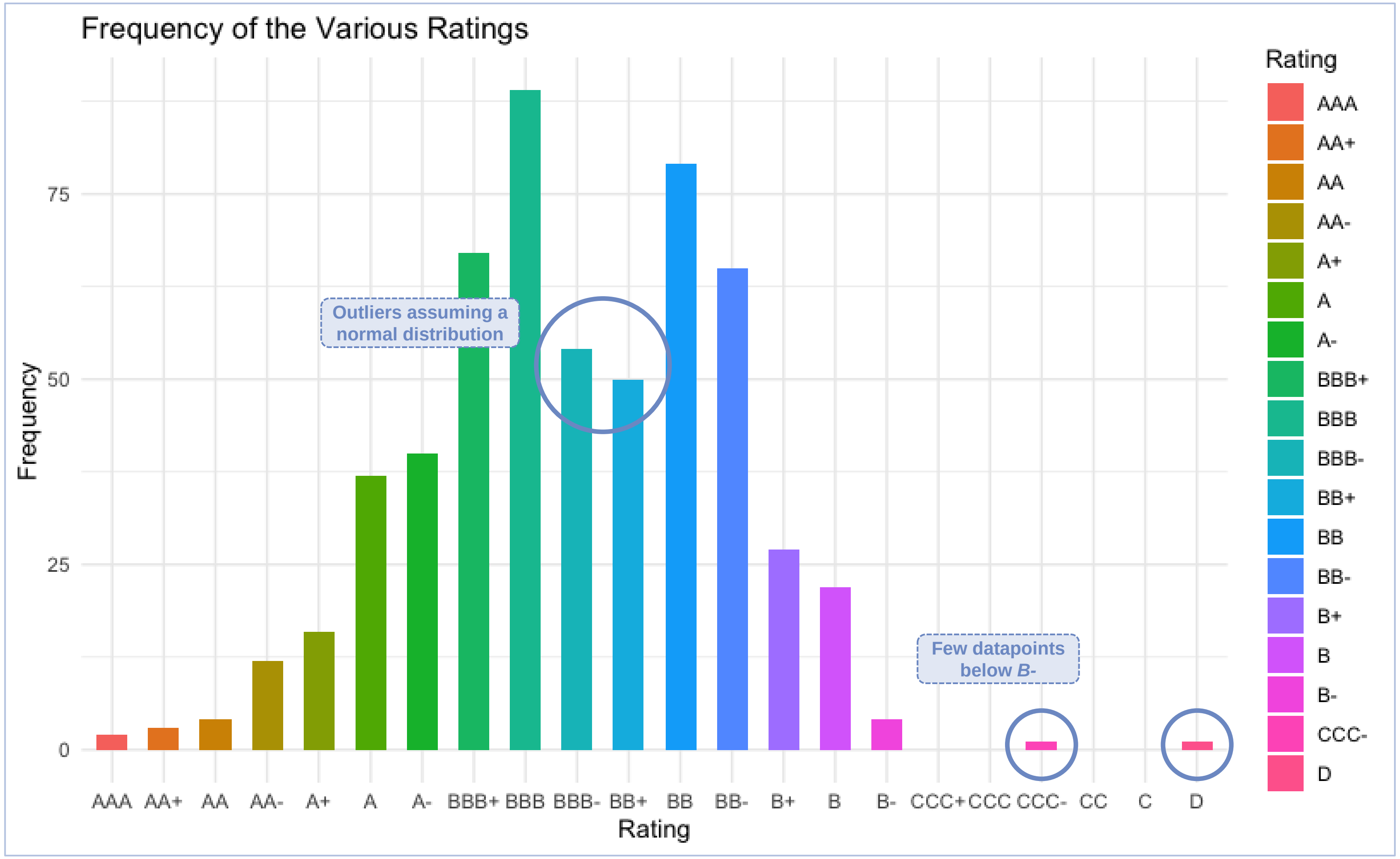
Covariance Analysis

- As expected, a high correlation between variables, that are calculated from similar balance sheet / income statement items persists
- Besides that, fairly **good independence of covariates**, allowing the dataset to be **used for various ML tasks**



For the most part, predictors are independent from each other allowing for training of a ML model

Frequency of Various Ratings



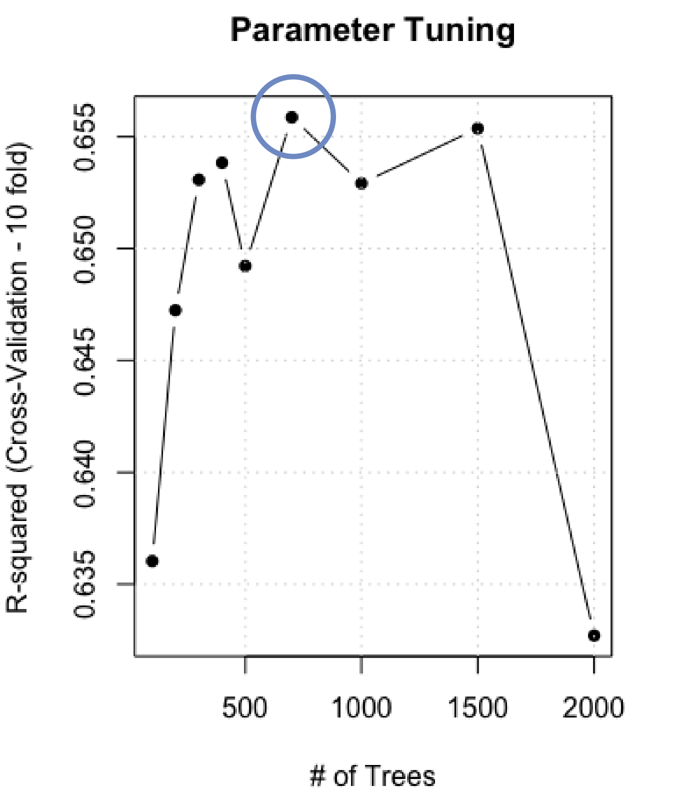
- The distribution of frequency of the various rating in the dataset used **mimics a normal distribution**
- Specifically, the **ratings *BBB-* and *BB+* act as outliers**, with a prevalence lower than expected
- Only a few outliers below *B-*, although datapoints in default exist as well

S&P CREDIT RATINGS

Random Forest

Tuning of Hyperparameters

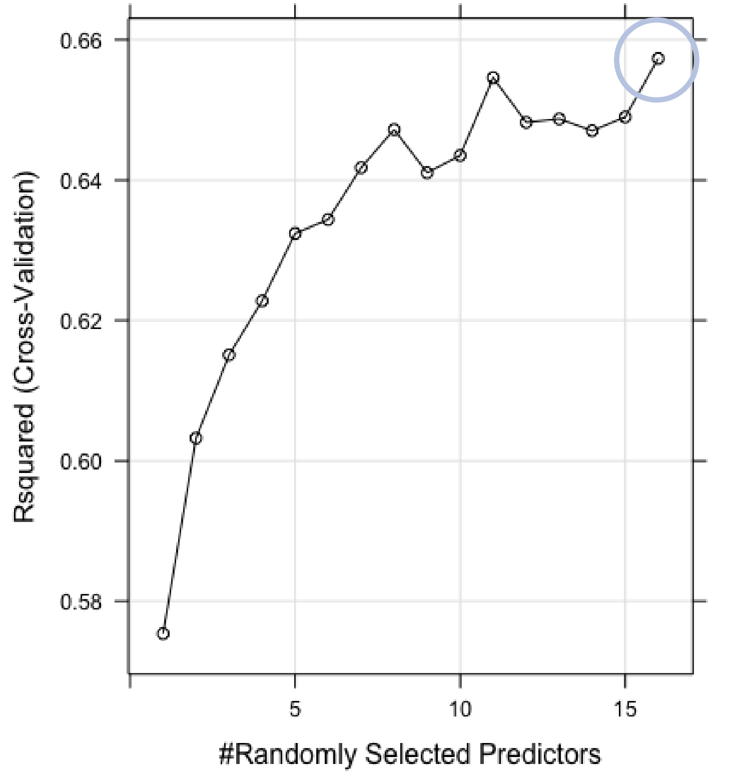
Number of Trees n_{tree}



- The chart shows the relationship between the number of trees and R^2
- For calculation of R^2, a cross-validation is utilised with $K = 10$
- The resulting relationship is **very volatile** with optimal results being reached in the **mid-range near 700 trees**

m_{try}

- The number of variables randomly sampled at each level of the recursive partitioning process to construct a random forest is called m_{try}
- For increasing m_{try}, the $\boldsymbol{R^2}$ **of the random forest model is increasing** at first
- However, the **growth rates are decreasing** with basically optimal results being reached for $m_{try} = 16$



Performance

Actual	Pred.	Delta	Actual	Pred.	Delta	Actual	Pred.	Delta	Actual	Pred.	Delta	Actual	Pred.	Delta
16	15.40295	-0.59705	16	14.64576	-1.35424	18	17.21671	-0.78329	11	11.06460	0.06460	11	12.93981	1.93981
17	16.67764	-0.32236	14	13.90090	-0.09910	18	16.28895	-1.71105	11	10.98555	-0.01445	4	9.60400	5.60400
11	12.22767	1.22767	9	9.15769	0.15769	15	11.37886	-3.62114	9	11.85517	2.85517	17	16.75695	-0.24305
21	16.70252	-4.29748	17	16.34536	-0.65464	14	12.67740	-1.32260	12	11.85086	-0.14914	14	13.82343	-0.17657
13	11.58979	-1.41021	17	16.24171	-0.75829	20	13.62479	-6.37521	11	12.14102	1.14102	12	13.95002	1.95002
14	14.01557	0.01557	16	12.83900	-3.16100	12	12.47581	0.47581	11	13.95431	2.95431	13	12.62629	-0.37371
13	13.26455	0.26455	14	14.40831	0.40831	13	13.00048	0.00048	10	9.63107	-0.36893	9	8.60195	-0.39805
15	13.04310	-1.95690	15	16.07336	1.07336	14	13.82581	-0.17419	10	11.92457	1.92457	17	16.54838	-0.45162
14	12.17752	-1.82248	14	15.16800	1.16800	11	11.46131	0.46131	11	11.59979	0.59979	10	11.18955	1.18955
14	12.05095	-1.94905	15	13.30621	-1.69379	10	10.17962	0.17962	13	11.40998	-1.59002	11	10.40488	-0.59512
13	12.82612	-0.17388	13	14.93176	1.93176	11	11.47095	0.47095	14	16.21240	2.21240	10	10.69417	0.69417
11	11.25212	0.25212	14	13.46793	-0.53207	15	13.62421	-1.37579	11	9.81926	-1.18074	12	13.07926	1.07926
13	11.92714	-1.07286	15	13.17536	-1.82464	15	14.78290	-0.21710	18	15.24636	-2.75364	10	13.19679	3.19679
15	16.33655	1.33655	11	11.52924	0.52924	10	10.51343	0.51343	13	14.96429	1.96429	14	12.43212	-1.56788
10	10.52410	0.52410	15	13.85271	-1.14729	10	9.97200	-0.02800	14	14.02707	0.02707	15	13.84726	-1.15274
12	11.47043	-0.52957	16	16.07145	0.07145	13	13.09886	0.09886	10	11.74826	1.74826	10	14.66860	4.66860
11	10.12614	-0.87386	15	14.23183	-0.76817	10	10.36169	0.36169	14	14.48286	0.48286	8	10.15733	2.15733
17	15.53643	-1.46357	20	17.21414	-2.78586	13	16.11948	3.11948	11	10.68276	-0.31724	15	15.87117	0.87117
12	14.46798	2.46798	17	15.74145	-1.25855	10	9.57848	-0.42152	10	10.13524	0.13524	9	11.45610	2.45610
15	15.48445	0.48445	14	10.37010	-3.62990	11	9.77943	-1.22057	12	13.64000	1.64000	12	10.10293	-1.89707
18	14.53879	-3.46121	16	14.94083	-1.05917	8	9.67848	1.67848	12	11.69283	-0.30717	12	13.24293	1.24293
8	11.79317	3.79317	15	13.51083	-1.48917	10	11.49390	1.49390	11	13.16764	2.16764	18	15.83443	-2.16557
12	10.81762	-1.18238	16	15.28071	-0.71929	11	13.43664	2.43664	9	10.04836	1.04836	15	15.67798	0.67798
17	14.67969	-2.32031	19	17.34483	-1.65517	13	13.66179	0.66179	15	13.76757	-1.23243	13	14.02581	1.02581
17	13.53817	-3.46183	14	12.62929	-1.37071	8	9.41238	1.41238	16	14.09110	-1.90890	10	9.75479	-0.24521
12	12.66733	0.66733	17	15.64421	-1.35579	19	17.69595	-1.30405	19	13.31429	-5.68571	14	11.88929	-2.11071
11	15.34867	4.34867	17	16.47350	-0.52650	12	15.00576	3.00576	15	13.78910	-1.21090	10	14.11471	4.11471
13	11.02143	-1.97857	15	14.89724	-0.10276	10	11.40474	1.40474	11	11.30231	0.30231	8	9.48390	1.48390
17	15.85893	-1.14107	17	14.82779	-2.17221	9	9.47021	0.47021	15	13.46638	-1.53362	10	11.30381	1.30381
14	15.07505	1.07505	15	13.72902	-1.27098	8	8.49357	0.49357	12	14.04410	2.04410	10	12.70281	2.70281
17	16.05579	-0.94421	16	15.14143	-0.85857	11	12.40598	1.40598	14	15.26267	1.26267	10	9.44112	-0.55888
14	11.10088	-2.89912	14	12.97269	-1.02731	10	10.17988	0.17988	13	16.39338	3.39338	11	10.06833	-0.93167
12	11.90345	-0.09655	15	13.86045	-1.13955	14	14.67962	0.67962	10	10.62929	0.62929	14	12.64048	-1.35952
21	15.13648	-5.86352	11	10.61074	-0.38926	13	13.54064	0.54064	14	13.93679	-0.06321	14	15.08383	1.08383

MSE 3.45770203230959

R^2 0.608313

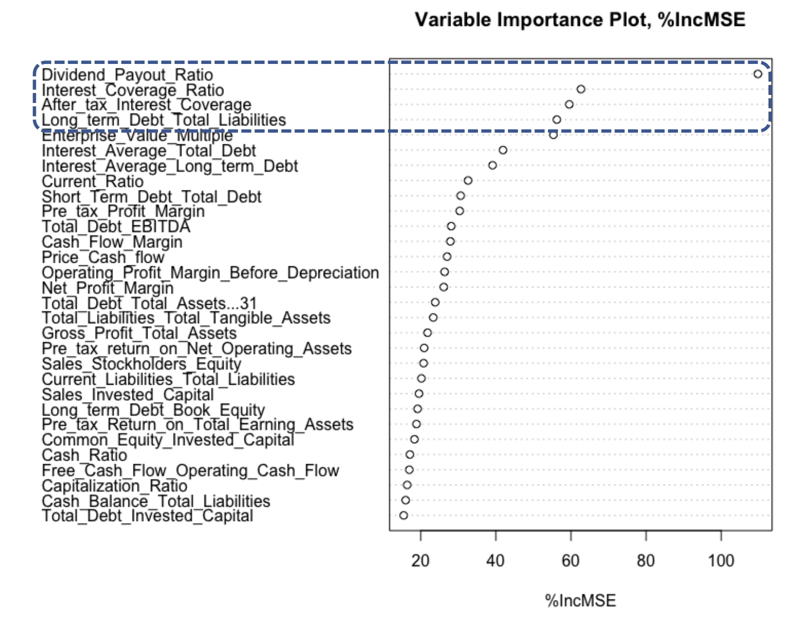
Conclusion

Random forest model delivers **high explanatory power** as seen by R^2

Variable Importance

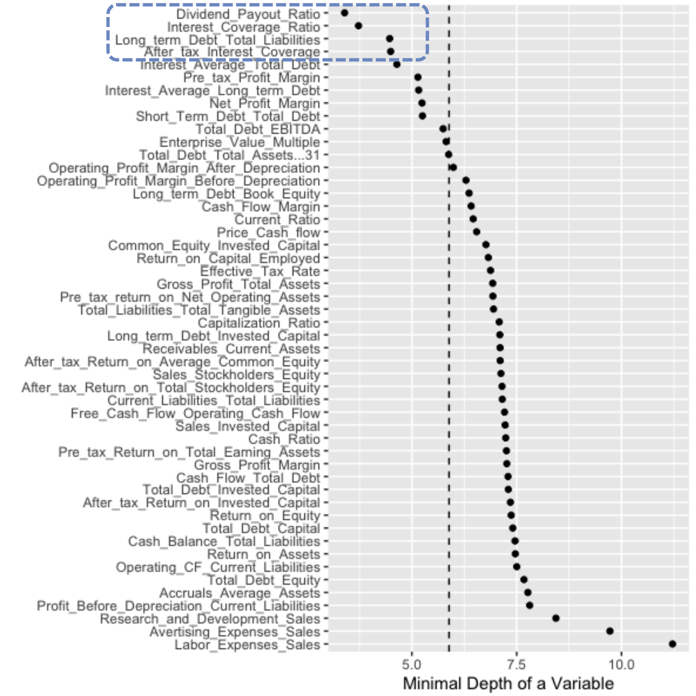
MSE Increase: Methodology

We measured variable importance by the increase in mean square error (MSE) if the corresponding variable is removed from the model. Therefore, variables that **explain more output** by themself, **increase MSE** the most.



Minimal Depth: Methodology

- The minimal depth shows the distance between the root of the tree and the node where the input variable was used
- **Early contribution** of a variable signifies its **power**



Conclusion

Both methods for determining variable importance highlight **Dividend Payout Ratio** to be the most important variable, followed by **interest coverage measures** and the **long-term debt ratio**

S&P CREDIT RATINGS

Linear Regression

Approach for Linear Regression in R

Select 7 Most Relevant Features
(from Random Forest)

Split Data into Training and Test Sets

70% Training Set	**30%** Test Set

Train Linear Regression Model → **Calculate MSE**

10-fold Cross-Validation for Prediction

Evaluate Model Performance
Compute the Mean Squared Error (MSE) on the test data

Results for Linear Regression in R

10-fold CV Test MSE: 5.030737

RMSE: 2.242930

Interpretation

To **reduce variability of the test MSE** by using an average over 10 folds

Interpretation

Approximately the **difference between our predicted value and the true value**

S&P CREDIT RATINGS

Neural Network

Approach for Neural Networks in R

1. Data Scaling

Scale the data to the same range

2. Split Data into Training and Test Sets

70%	30%
Training Set	Test Set

3. Create Model

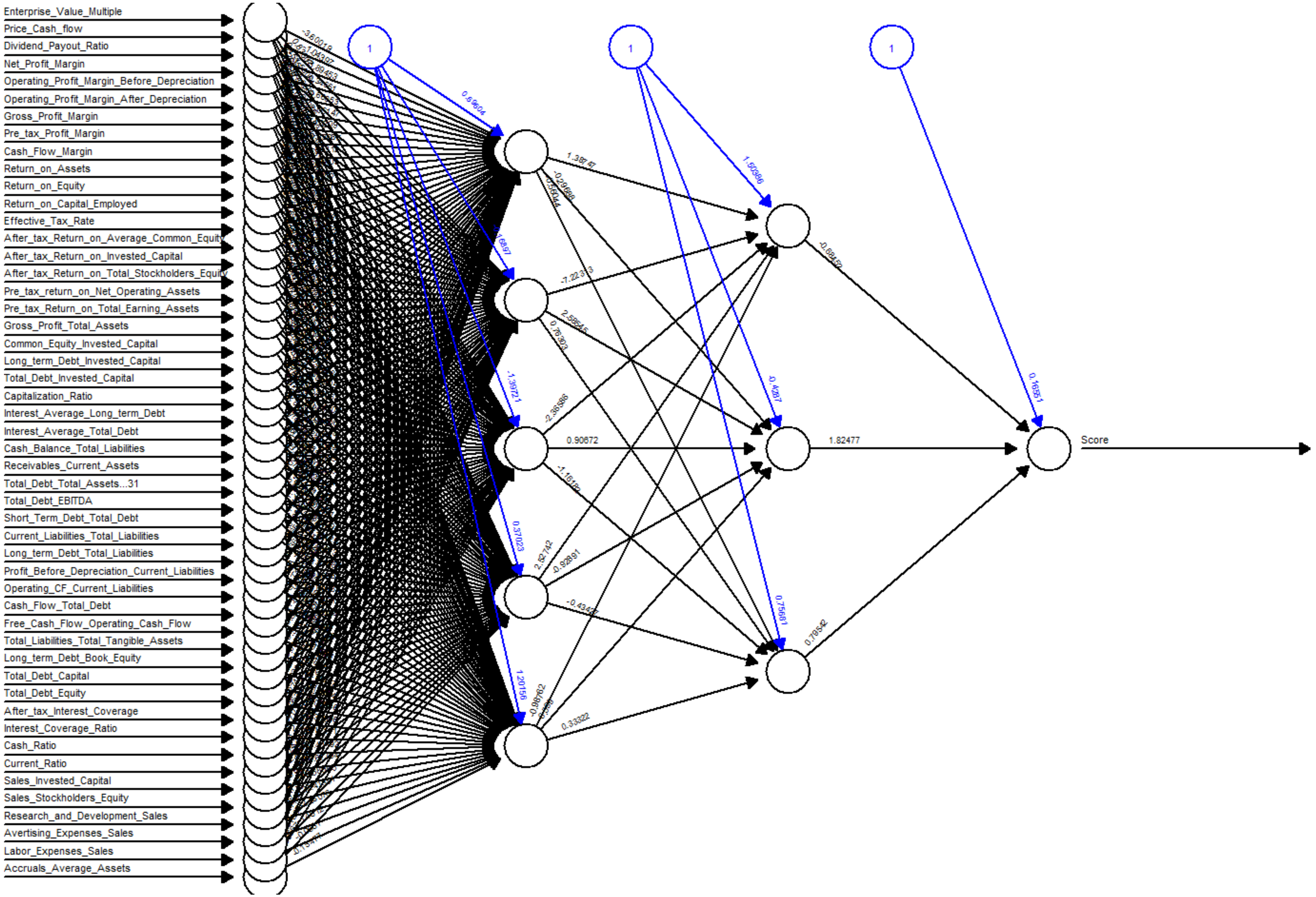
Create a neural network with 5 neurons in the first hidden layer and 3 neurons in the second.

- Train the model
- Visualize the Neural Network
- Calculate the MSE on the test data
- Plot Predicted vs. True Score (70% Train, 30% Test)

4. Perform 10-fold cross-validation, calculate average MSE across 10 iterations.

Plot Predicted vs. True Score (for all 10-folds)

Neural Network



Input Layer

50 Neurons

Hidden Layers

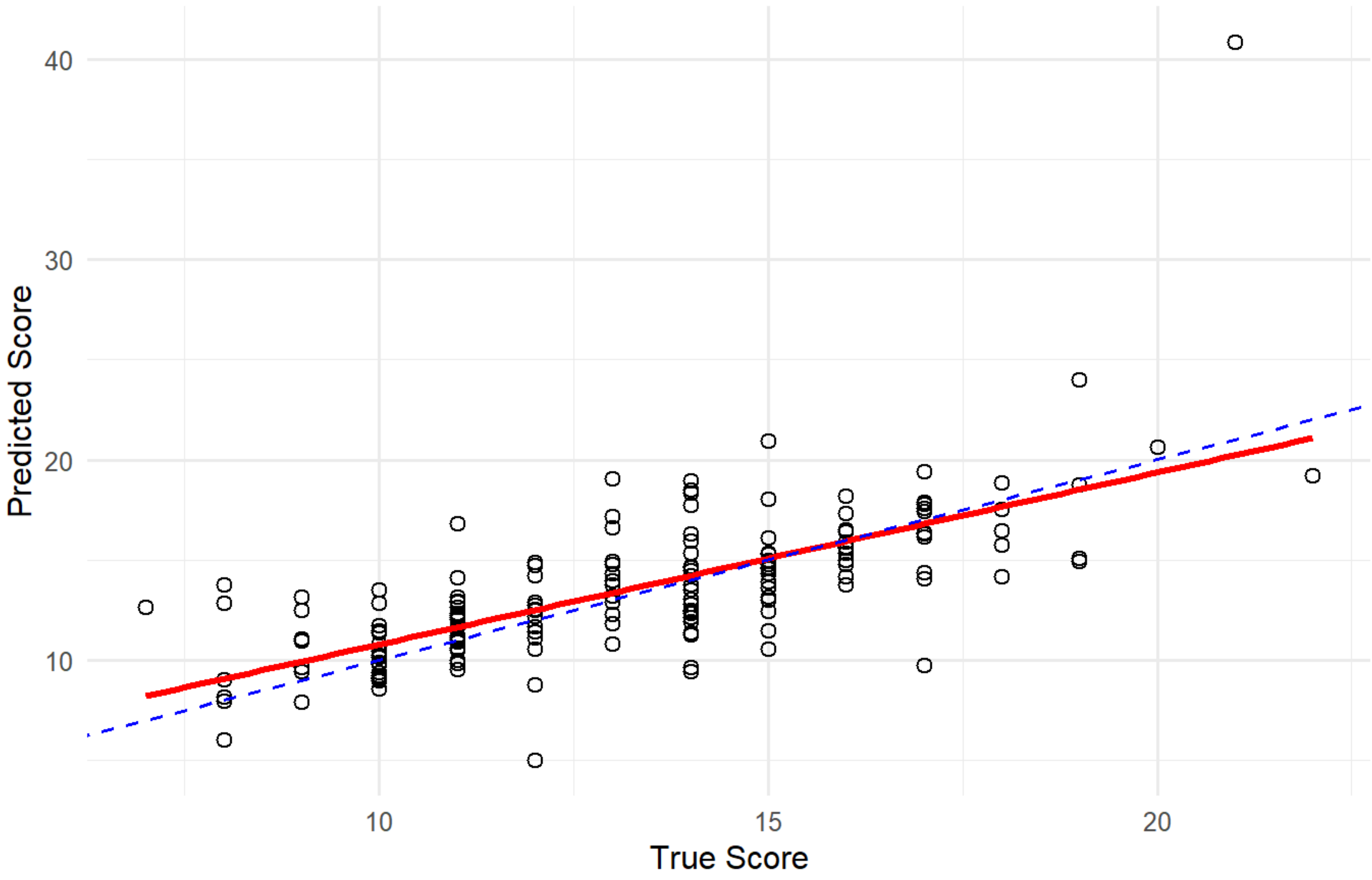
5 Neurons
First hidden layer

3 Neurons
Second hidden Layer

Output Layer

1 Neurons

True vs Predicted Values (70% Train and 30% Test)

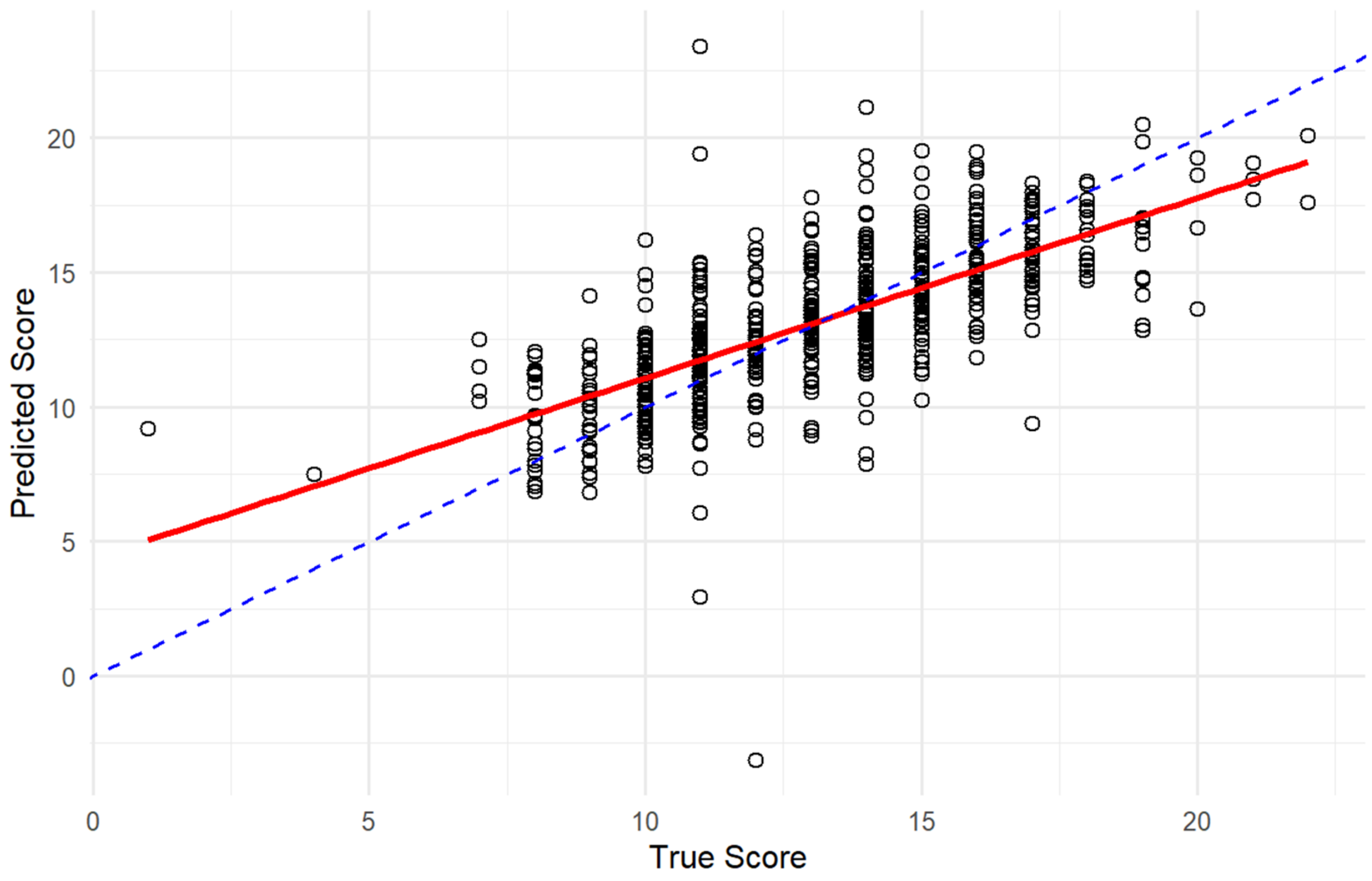


MSE 7.510245

Angle bisector, perfect predictor

Linear fit of neural network predicted data points

True vs Predicted Values (10-Fold CV)



MSE 5.203134

Angle bisector, perfect predictor

Linear fit of neural network predicted data points

Conclusion and Limitations

Conclusion

- With a **more complex** model the **accuracy** should **increase** (trade-off between model interpretability and performance)
- In the case: **random forest** is **more accurate** than **linear regression**
- «Rule» does not hold for neural networks

Limitations

- Assigned **numerical values** to the **credit ratings**, which means there is an **exact difference** of one **between consecutive ratings** (e.g., AA is one unit better than A)
- In reality: the **difference** in creditworthiness **between ratings** is **not** clearly **defined**
- Qualitative decision process by rating agency

Further Research

- **Address numerical representation** of **ordinal ratings**: ordinal regression (ordered ratings without equal spacing), group ratings into categories, expert input to weight the differences between the ratings – non-linear scaling
- **Rounding** could lead to an improvement in the regression results – however the question arises how to round values effectively