Textual Analysis and Financial Statements

Isaac Liu with Owen Lin, Chengzheng Xing, and Sean Zhou

April 5, 2024

Introduction

Corporate credit ratings represent professional estimations of the default risk carried by company debt. These ratings represent critical information for investors - not just institutional investors and financially sophisticated bondholders, but also stockholders, who may be wiped out completely in the event of bankruptcy. Analyzing ways to predict ratings can offer substantial value to a variety of stakeholders. Predictive models may be useful for investors without access to data, companies or potential lenders that seek information about influential factors,[1] and by any parties seeking interpolated ratings for companies that do not have them.

In this project, we seek to fully leverage the text of earnings calls, along with traditional financial measures and variables, to improve predictions of corporate credit ratings for any given company and quarter and better understand the importance of various influences.[2] Textual features such as pre-trained language model vector embeddings (Araci, 2019) and analyses of sentiment join tabular variables as inputs to a variety of supervised machine learning techniques for classification from logistic regression to tree-based methods. If time allows, we will also make use of advances in the study of graph neural networks to incorporate additional embeddings modelling linkages between firms (Das et al., 2023) implied by calls.

To the best of our knowledge, the closest prior work to ours is Donovan et al. (2021), which leverages the textual content of earnings calls and financial statements to predict credit events such as bankruptcies, interest spread changes, and rating downgrades. Unigram and bigram word frequencies were used with the supervised machine learning techniques of Support Vector Regression, Latent Dirichlet Allocation, and Random Forests. The coefficient on a constructed textual measure of credit risk was found to be significant up the 1% level. In contrast to this approach, we focus on predicting the credit ratings themselves, and integrate more recent techniques such as pre-trained neural language models and a wider variety of algorithms for classification.

Data and Exploratory Data Analysis

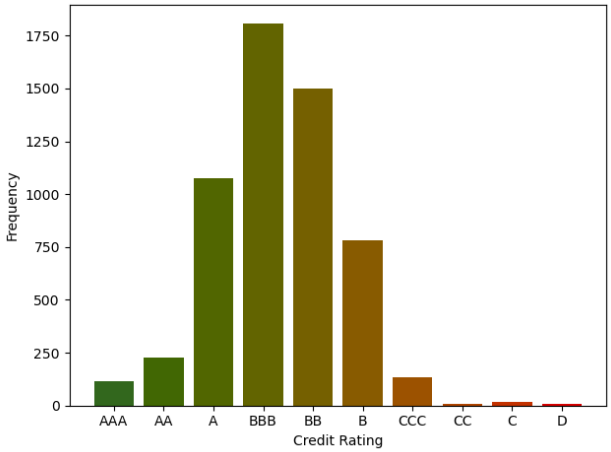
We combine a wide variety of data sources to support our predictions of credit ratings - merging rating data with company earnings calls, financial statement variables, and industry sector. In our combined dataset, each

[1] There is evidence suggesting financial factors and projections have a causal impact on ratings and are not manipulated by companies in response to forecasted rating changes (He, 2018).

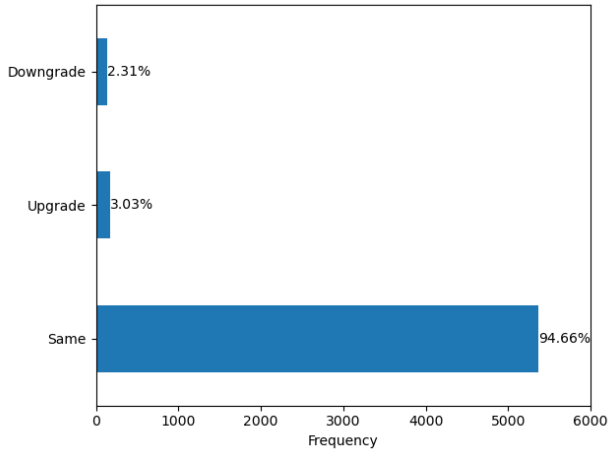
[2] Though much literature has focused on financial statements and reports and credit ratings (as just one example, see Makwana et al. (2022)), our paper takes a relatively underexplored approach, instead incorporating earnings call transcripts. We believe calls offer a richer picture of a firm's financial prospects because they include two-way conversation between company management and financial analysts in form of a Q and A section. This section incorporates the broader beliefs and concerns of the financial community into our predictions. Additionally, in contrast to financial statements, which must be (noisily) parsed to identify sections relevant to management analysis, earnings calls provide more directly valuable and readily available information.

Figure 1: Credit Ratings

(a) Distribution



(b) Change Between Fixed Quarter Dates



observation represents a fixed quarter date (1/1, 4/1, 7/1, 10/1) for a company, with the company's most recent credit rating, earnings call and associated financial statement variables, and sector attached.

Our scope of interest is publicly traded companies from 2010-2016 (a limitation due to the availability of credit rating data) - the distribution of call year and quarters can be found in Appendix Figure A.1. To ensure comparability, we drop items missing any predictor variable. In all, we have 5,670 quarters for 437 unique companies.

Credit Ratings

We make use of long-term credit rating issuances from S and P Rating Services, provided from a combination of two credit rating datasets downloaded in CSV and Excel format from Kaggle (Gewerc, 2020; Makwana, Bhatt and Delwadia, 2022). Each issuance be a change in rating (upgrade, downgrade) or reaffirmation - they occur at ad-hoc intervals. We reshape these rating issuances to a dataset of ratings for each company on each fixed quarter date by creating a rating end date variable that is the date of the next issuance, and joining a list of the fixed quarter dates on the condition that the fixed quarter date is between the issuance date and the end date.

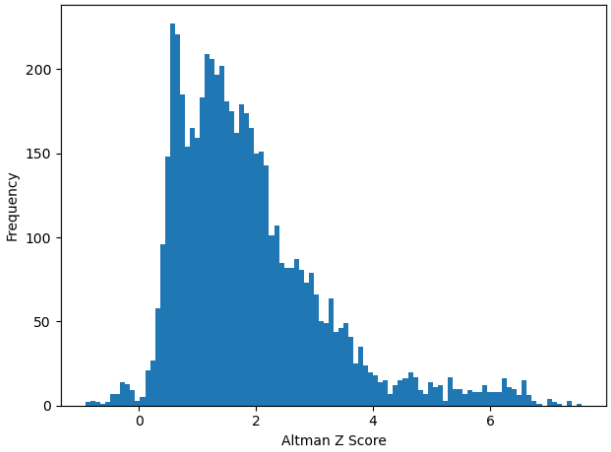
Figure 1 shows the distribution of rating grades used in our final dataset. Finer grades (+, -) are sometimes assigned by agencies, but these grades were removed for this project. Ratings of BBB and above are considered investment grade - these bonds carry empirical one-year default rates of 0 to 1%. Ratings below that are classified as junk, with default rates from 1 to 30, 40, or even 50% for some years (S and P Global Ratings, 2024). Most company-quarters have ratings around the BBB threshold, with very few cases on the extreme ends of the spectrum. Ratings also tend to be constant over time. Relative to the previous fixed quarter date, 94.66% of ratings remain the same. Rating on the previous fixed quarter date can thus be an extremely strong predictor.

Earnings Calls

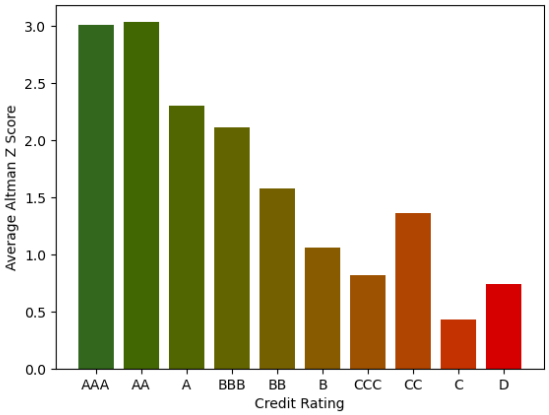
Our earnings call data comes from the Financial Modelling Prep API (Financial Modeling Prep, 2024), a trusted source widely used in industry. We remove all calls that happened more than 250 days prior and after the year and quarter they are supposed to discuss the results from. Including both prepared remarks and analyst Q and A sessions, the overall average call length in our final data stands at 8,754.25 words.

Figure 2: Altman Z-Score

(a) Distribution



(b) Average by Rating



Financial Statements

Our financial statement variables are also retrieved using the Financial Modelling Prep API. We make use of items from company balance sheets, cash flow statements, and income statements, as well as company market capitalization. To prepare the data, we limit our observations to items reported in USD, check for and correct items off by a factor of 1,000 as a result of parsing (if last few digits are 000.00 and the item is above or below 2.5% and 97.5% quantile, divide by 1,000), and check some accounting identities in Das et al. (2023), setting failing variables to missing. We also discard observations where statement filing dates do not agree between the three types of statements, where the filing date falls outside of the fixed quarter matched on via earnings call date, and where the filing date is more than 45 days after the earnings call date.

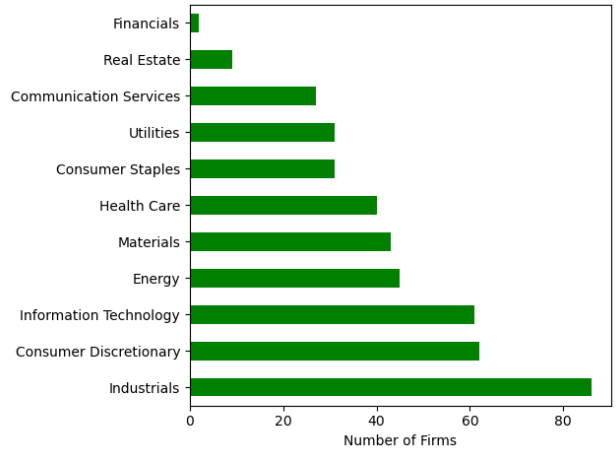
In some of our models, we make use of Altman's Z-score, a traditional measure of bankruptcy risk that accounts for company earnings, equity, and assets and liabilities (Altman, 1968) (for details on the construction of the score, see Appendix section A.3). Figure 2 shows the distribution of adjusted Z-scores in our dataset. Traditionally, values above 3.0 have been considered safe, while those below 1.8 are considered to have a high chance of bankruptcy. The average scores for each rating in our data seem to align well with this interpretation, with high scores being associated with higher ratings in a linear manner. Aside from a few quirks on the lower end of the rating spectrum (where not many companies and ratings are available), Z-Score is likely to be highly useful as a predictor.

Sector

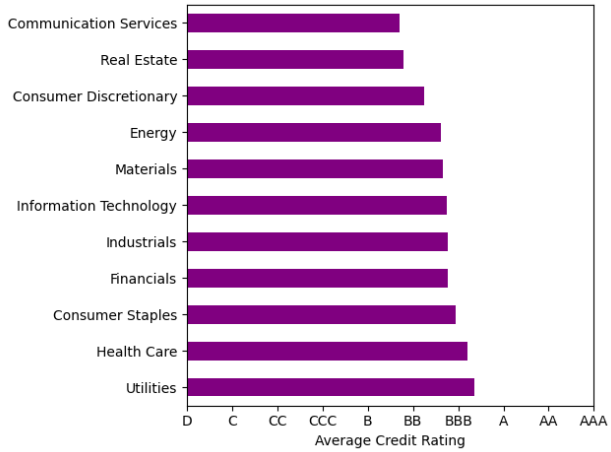
The GCIS industry classification standard divides companies into 11 major industry sectors (there are finer groupings as well, but this data was not easily obtainable for our project) (S and P and MSCI, 2024). It is widely used in the financial community, and was developed in part by S and P, the same company responsible for our credit ratings. We obtained classifications from Kaggle in CSV format and supplemented them with manual lookup (Kozlov, 2022). Figure 3 shows the unfortunate sectoral imbalance present in our data, with a large share of firms in consumer, industrial, and technology sectors, relative to very few in the distinctly different financials and real estate sectors.

Figure 3: Sector

(a) Firms by Sector



(b) Average Rating by Sector



Quality Control

Our data preparation was subject to rigorous quality control standards. We extensively code reviewed all data cleaning code. Our exploratory analyses identified data quality issues such as extreme values in financial statement variables, which we handled by winsorizing, and date gaps between quarters, earnings calls, and financial statements, which we dropped in the case of aggregiously mismatched observations.

NLP Features

Our NLP features capture the sentiment of calls, the transparency of discussion, and the level of analyst engagement.

- Net Positivity Score - In line with Kantos et al. (2022), we use the Harvard IV-4 Psychosocial Dictionary (n.d.) to count positive and negative words and compute

$$\text{Net Positivity Score} = log_{10}\frac{\text{Count Positive} + 1}{\text{Count Negative} + 1}$$

- Tone - Following Price et al. (2012), we use the Harvard dictionary to count words falling in various categories. Then we construct tone using the first principal component of the matrix with each call as a row and each column as one of the following:

$$\frac{\text{Positive}}{\text{Negative}}, \frac{\text{Active}}{\text{Passive}}, \frac{\text{Strong}}{\text{Weak}}, \frac{\text{Overstated}}{\text{Understated}}$$

- Numeric Transparency - ratio of numbers to words in the word-tokenized call
- Readability - We construct the Gunning-Fog readability score (Gunning, 1952) as

$$0.4 \times (\frac{\text{Words}}{\text{Sentences}} + 100 \times \frac{\text{3+ Syllable Words}}{\text{Words}})$$

- Number of Questions - count of question marks

We also add the word count of each call, as this appears to be highly predictive. The distribution of each NLP feature by rating is shown in Figure XXX below.

INSERT FIGURE

We are working on preparing FinBERT[3] (Araci, 2019) embeddings and have created Doc2Vec[4] (Le and Mikolov, 2014) embeddings to represent sentences and calls. These may be used to improve estimations of sentiment or a direct input to our classifier.

Modelling

Our overall model architecture is of the form

$$\text{Predicted Credit Rating} = f(\text{Previous Rating}, \text{Metadata}, \text{Financial Variables}, \text{Sector}, \text{NLP Features})$$

where Metadata includes relevant date variables from the data sources and the identity of the company and the other variables are as described above.

Logistic Regression

Table 1: Model Comparison

Model/Baseline	Accuracy	Weighted Average Precision	Weighted Average Recall	F1 Score	Share ≤ 1 Rating From Actual
Rating Model 1	0.36	0.30	0.36	0.26	0.82
Rating Model 2	0.51	0.49	0.51	0.46	0.89
Rating Model 3	0.95	0.95	0.95	0.95	0.99
Rating Model 4	0.95	0.95	0.95	0.95	0.99
Majority Baseline	0.32				

Table 1 shows prediction statistics for logistic regression models aiming to predict ratings (for predicting changes in rating, see Appendix Section A.5). Rating Model 1 includes only Altman's Z-Score as a predictor - its overall accuracy is not much better than the majority baseline, though predictions are generally close to true ratings. Rating Model 2 adds a full suite of financial statement variables (for a list, see items marked as Variable Type 'Financial Statements' and 'Market Capitalization' in Table A.1) and leads to improvements across the board. Rating Model 3 adds industry sector and the previous rating as predictors, and achieves a very high level of accuracy which we are not currently able to improve upon by adding the NLP features in Rating Model 4.

Table 2 generally shows that our most complex model (Rating Model 4) generally performs well across all classes. This is in large part due to our use of balanced class weighting to handle rare classes. We performed grid search 5-fold cross validation to arrive at these balanced weights. We also found via grid search that an Elastic Net penalty (which collapses to entirely a LASSO penalty) with a slight amount of regularization (C) effectively handles the large number of variables present in our data (for details, see Appendix Section A.4).

[3]BERT is a pretrained transformer-based language model that encodes text into embedding vectors using surrounding context. FinBERT is a version of BERT finetuned for tasks in the financial domain (language model embedding performance can vary greatly by domain).

[4]This method involves constructing representations of each call based on the bag-of-words and skipgram tasks - a neural network is trained to either a word or a word's context while accounting for a vector identifying the document.

Table 2: Classification Report - Most Complex Model

Rating	Precision	Recall	F1-Score	Support
AAA	0.80	0.84	0.82	19
AA	0.86	0.88	0.87	43
A	0.93	0.92	0.92	219
BBB	0.96	0.97	0.97	356
BB	0.98	0.98	0.98	313
B	0.97	0.95	0.96	144
CCC	0.93	0.93	0.93	27
CC	0.50	1.00	0.67	1
C	1.00	0.67	0.80	3
D	1.00	1.00	1.00	2

Table 3: Permutation Importance - Most Complex Model

Feature	Mean	Standard Deviation
Rating on Previous Fixed Quarter Date BBB	0.281622	0.010257
Rating on Previous Fixed Quarter Date BB	0.230148	0.009061
Rating on Previous Fixed Quarter Date A	0.107111	0.005848
Rating on Previous Fixed Quarter Date B	0.079304	0.004862
Rating on Previous Fixed Quarter Date AA	0.013898	0.002093
Rating on Previous Fixed Quarter Date CCC	0.012949	0.001282
Total Non-Current Liabilities	0.000867	0.000612
Total Stockholders' Equity	0.000809	0.000366
Research and Development Expenses	0.000789	0.000284
Net Receivables	0.000779	0.000314
Gunning-Fog Score	0.000764	0.000307
Dividends Paid	0.000663	0.000504
Other Current Liabilities	0.000659	0.000388
Debt Repayment	0.000631	0.000402
Numeric Transparency	0.000543	0.000469

Conclusion and Next Steps

We have seen above that textual and NLP features are contributing to our predictions...

One major next step is continuing to improve the construction of our NLP features and methods. Much more work could be done to improve the construction of our sentiment scores and analysis - seeking out better pretrained models for earnings call sentiment, or improving the methods through which embedding representations are converted to sentiment. Some members of the group have also been working on a separate class project involving annotating earnings calls with topics discussed. This work could be integrated into our project to provide additional features for our models - we could identity topics and then connect them to credit ratings. Finally, we could try building an end-to-end transformer classifier that takes in earnings calls and outputs credit ratings (perhaps finetuning and adding a classifier on top of the longformer (Beltagy, Peters and Cohan, 2020) transformer encoder model[5]).

We've also begun using the AutoML library Autogluon to explore a wider variety of classifiers. Autogluon runs a wide variety of state-of-the-art prediction algorithms and performs hyperparameter tuning. The results, shown in Table 4 for all allowed predictors in our dataset (metadata about call and statement dates, the rating on the previous fixed quarter date, all the financial statement variables, all the constructed NLP features, and sector) can provide a good starting point for our further modelling choices.

Tree-based bagging and boosting methods (boosting in particular) seem to perform extraordinarily well on our modelling task (ExtraTrees is a somewhat simplified variant of random forest; GBM stands for Gradient Boosting Machine). Each of the trained Autogluon models also comes with a set of optimized hyperparameters - after

[5]Longformer or other techniques are ideal for document, rather than word or sentence embedding creation

Table 4: Autogluon Leaderboard

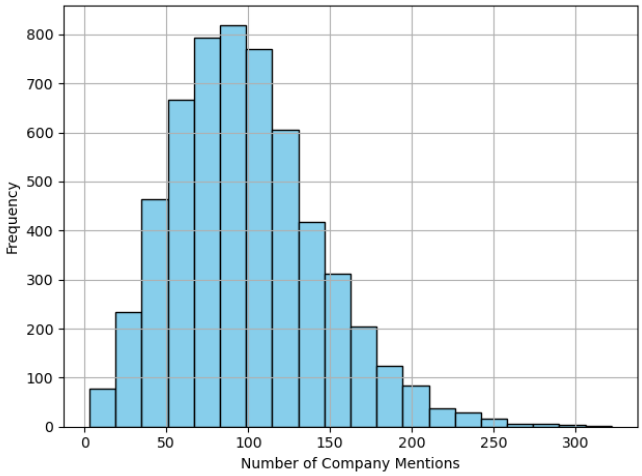
Model	Test Accuracy
XGBoost	0.95
LightGBM	0.95
LightGBMXT	0.95
WeightedEnsembleL2	0.95
CatBoost	0.95
NeuralNetTorch	0.95
LightGBMLarge	0.94
ExtraTreesGini	0.94
ExtraTreesEntr	0.94
RandomForestGini	0.94
NeuralNetFastAI	0.94
RandomForestEntr	0.93
KNeighborsDist	0.87
KNeighborsUnif	0.85

selecting a model, we will further tune these hyperparameters to improve performance. The library also provides feature importance measures, computed by permuting each feature and measuring the drop in accuracy. Though these tests are not of high quality (and many results are not reported or are not significant), they confirm that previous ratings, earnings call word counts and tone variables, a few financial variables, and several date metadate variables are all very important.

CATEGORICAL PREVIOUS RATING IS THE MOST IMPORTANT PREDICTOR

Another area of interest for us is continuing to pursue the approach in Das et al. (2023), which uses graph neural networks to model the relationships between companies in combination with tabular financial data and NLP features. Prerequisite to this approach is the construction of an undirected graph showing linkages between companies based on the earnings call data. We've pursued begun work on this step from two angles. First, following the original paper, Doc2Vec embeddings representing calls can be averaged for each company. A graph can then be constructed with connecting edges added for cases when the vectors for each company have cosine similarity above a certain threshhold. As a second approach, which also opens more opportunities for exploration even without a neural network, we have used transformer-based Named Entity Recognition to identify mentions of any company in each earnings call. On average, each earnings call has 98.66 company mentions - Figure 4 shows the distribution.

Figure 4: Company Mentions



Though the vast majority of these mentions are likely to be of the company whose call is being discussed, a casual

glance at the data does suggest there are a fair number of mentions of partners, suppliers, and competitors within some calls. Our next step involves the use of entity resolution algorithms (trigram matching, supervised learning) to link these mentions to firm tickers in order to construct a graph of relationships.

Acknowledgements

Special thanks to the UC Berkeley Stats Department Statistical Computing Facility (SCF). Other acknowledgements: Libor Pospisil, Robert Thompson. GitHub Co-Pilot was used for python code generation (mostly for plotting and table creation/parsing).

References

Altman, Edward I. 1968. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *The Journal of Finance*, 23(4): 589–609. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1968.tb00843.x.

Araci, Dogu. 2019. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." arXiv:1908.10063 [cs].

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. "Longformer: The Long-Document Transformer." arXiv:2004.05150 [cs].

Das, Sanjiv, Xin Huang, Soji Adeshina, Patrick Yang, and Leonardo Bachega. 2023. "Credit Risk Modeling with Graph Machine Learning." *INFORMS Journal on Data Science*, 2(2): 197–217. Publisher: INFORMS.

Donovan, John, Jared Jennings, Kevin Koharki, and Joshua Lee. 2021. "Measuring credit risk using qualitative disclosure." *Review of Accounting Studies*, 26(2): 815–863.

Financial Modeling Prep. 2024. "Financial Modeling Prep - FinancialModelingPrep."

Gewerc, Alan. 2020. "Corporate Credit Rating with Financial Ratios."

Gunning, Robert. 1952. *The Technique of Clear Writing*. McGraw-Hill. Google-Books-ID: ofI0AAAAMAAJ.

He, Guanming. 2018. "The Impact of Impending Credit Rating Changes on Management Earnings Forecasts." *Global Journal of Management and Business Research*, 18: 1–18.

Inquirer Home Page. n.d..

Kantos, Christopher, Dan Joldzic, Gautam Mitra, and Kieu Thi Hoang. 2022. "Comparative Analysis of NLP Approaches for Earnings Calls."

Kozlov, Alex. 2022. "US public companies classification."

Le, Quoc V., and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." arXiv:1405.4053 [cs].

Makwana, Ravi, Dhruvil Bhatt, and Kirtan Delwadia. 2022. "Corporate Credit Rating."

Makwana, Ravi, Dhruvil Bhatt, Kirtan Delwadia, Agam Shah, and Bhaskar Chaudhury. 2022. "Understanding and Attaining an Investment Grade Rating in the Age of Explainable AI."

Price, S. McKay, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2012. "Earnings conference calls and stock returns: The incremental informativeness of textual tone." *Journal of Banking & Finance*, 36(4): 992–1011.

S and P, and MSCI. 2024. "GICS® - Global Industry Classification Standard."

S and P Global Ratings. 2024. "S and P Global Ratings."

A Appendix

A.1 Summary Statistics for Numeric Variables

Table A.1 shows summary statistics for all numeric variables in our dataset. Important numeric and categorical variables are explained in the main text. We also have numerous date variables, which we may use in future predictions.

Table A.1: Numeric Summary Statistics

Variable Name	Mean	Minimum	Median	Maximum	Standard Deviation	Variable Type
Altman's Z Score	1.88	-0.91	1.61	7.56	1.28	Altman's Z Score
Accounts Payable (Balance Sheet)	966,256,179.24	-237,651,171.00	358,675,500.00	11,433,000,000.00	1,564,143,869.54	Financial Statements
Accounts Payable (Cash Flow Statement)	5,115,097.34	-321,769,000.00	0.00	1,789,652,000.00	82,020,828.64	Financial Statements
Accounts Receivables	-11,276,972.01	-544,000,000.00	0.00	325,000,000.00	91,302,900.98	Financial Statements
Accumulated Other Comprehensive Income (Loss)	-404,904,065.77	-5,290,000,000.00	-74,923,500.00	431,595,000.00	880,562,727.62	Financial Statements
Capital Expenditure	-192,838,993.57	-1,867,000,000.00	-61,000,000.00	412,700.00	309,755,067.36	Financial Statements
Capital Lease Obligations	25,062,333.48	0.00	0.00	9,056,234,000.00	228,139,330.33	Financial Statements
Cash and Cash Equivalents	869,139,501.47	0.00	334,747,500.00	9,223,000,000.00	1,373,785,187.87	Financial Statements
Cash and Short Term Investments	1,064,264,890.38	0.00	365,150,000.00	15,601,000,000.00	1,892,893,943.18	Financial Statements
Cash at Beginning of Period	875,066,189.86	-2,556,000.00	336,188,500.00	9,610,000,000.00	1,399,673,823.88	Financial Statements
Cash at End of Period	881,467,356.71	-154,400.00	337,800,000.00	9,743,000,000.00	1,410,805,303.47	Financial Statements
Change in Working Capital	-17,622,464.72	-870,000,000.00	-2,509,500.00	756,000,000.00	183,386,926.92	Financial Statements
Common Stock	333,701,836.30	-539,800.00	3,962,000.00	9,817,134,000.00	936,073,406.58	Financial Statements
Common Stock Issued	44,261,851.94	-3,572,000.00	43,000.00	1,111,490,728.00	123,285,381.77	Financial Statements
Common Stock Repurchased	-78,688,233.14	-2,086,545,366.00	-797,000.00	545,656,614.52	187,838,839.66	Financial Statements
Cost and Expenses	2,339,756,005.83	-2,495,000.00	1,130,100,000.00	22,769,000,000.00	3,398,867,286.33	Financial Statements
Cost of Revenue	1,639,068,160.07	-3,094,000.00	791,749,000.00	18,303,000,000.00	2,432,969,396.59	Financial Statements
Debt Repayment	-246,924,032.20	-3,001,000,000.00	-32,044,500.00	200.00	471,308,141.50	Financial Statements
Deferred Income Tax	6,021,092.62	-253,000,000.00	64,500.00	1,850,454,000.00	59,026,597.62	Financial Statements
Deferred Revenue	309,581,598.24	-116,912,000.00	48,000,000.00	4,918,100,000.00	646,673,378.88	Financial Statements
Depreciation and Amortization (Cash Flow Statement)	142,785,944.04	-675,312.00	53,814,500.00	1,529,000,000.00	212,219,854.14	Financial Statements
Depreciation and Amortization (Income Statement)	141,427,421.90	-1,550,000.00	54,944,500.00	1,371,000,000.00	204,346,754.15	Financial Statements
Diluted EPS	0.52	-156.36	0.51	49.73	3.27	Financial Statements
Dividends Paid	-91,477,627.48	-1,233,000,000.00	-21,000,500.00	0.00	182,607,289.06	Financial Statements
EBITDA	446,138,301.60	-66,200,000.00	194,050,000.00	4,410,000,000.00	643,419,252.02	Financial Statements
EBITDA Ratio	0.20	-5.77	0.17	2.16	0.22	Financial Statements
EPS	0.53	-156.36	0.52	53.75	3.29	Financial Statements
Effect of Foreign Exchange Changes on Cash	-1,653,388.25	-65,000,000.00	0.00	52,000,000.00	11,180,607.85	Financial Statements
Free Cash Flow	157,084,973.55	-541,000,000.00	51,192,500.00	2,683,000,000.00	388,980,015.53	Financial Statements
General and Administrative Expenses	153,358,448.12	-2,738,500.00	32,647,000.00	2,007,000,000.00	301,738,888.27	Financial Statements
Goodwill	2,002,690,830.53	-202,702,100.00	630,531,000.00	23,389,000,000.00	3,533,369,156.82	Financial Statements
Goodwill and Intangible Assets	3,124,212,237.48	-1,618,944,000.00	966,716,000.00	37,123,000,000.00	5,678,672,562.17	Financial Statements
Gross Profit	868,732,901.33	-7,195,000.00	379,801,000.00	9,223,000,000.00	1,372,085,330.98	Financial Statements
Gross Profit Ratio	0.37	-5.65	0.34	2.32	0.26	Financial Statements
Income Before Tax	256,118,983.25	-353,153,000.00	92,535,500.00	2,951,000,000.00	434,784,626.43	Financial Statements
Income Before Tax Ratio	0.07	-9.38	0.09	2.68	0.35	Financial Statements
Income Tax Expense	69,610,448.41	-119,131,000.00	22,400,000.00	736,000,000.00	121,357,794.26	Financial Statements
Intangible Assets	838,566,674.64	-421,000.00	169,050,000.00	14,110,100,000.00	1,786,919,275.21	Financial Statements
Interest Expense	46,670,431.96	-16,400,000.00	23,050,000.00	386,000,000.00	61,968,392.63	Financial Statements
Interest Income	2,377,179.84	-62,900.00	0.00	69,000,000.00	6,882,052.66	Financial Statements
Inventory (Balance Sheet)	941,316,423.17	-19,626,000.00	407,512,500.00	8,328,000,000.00	1,407,393,286.13	Financial Statements
Inventory (Cash Flow Statement)	-10,383,742.26	-420,000,000.00	0.00	289,000,000.00	70,575,014.44	Financial Statements
Investments in Property, Plants, and Equipment	-194,182,976.28	-1,921,864,000.00	-61,381,500.00	412,700.00	313,040,469.62	Financial Statements
Long-Term Debt	4,157,384,374.86	-651,718.00	1,828,522,000.00	31,359,000,000.00	5,563,500,112.28	Financial Statements
Long-Term Investments	493,613,595.97	-490,677,000.00	12,732,500.00	10,981,000,000.00	1,369,480,812.68	Financial Statements
Minority Interest	92,264,442.06	-20,252,654.04	1,759,000.00	2,316,406,000.00	270,956,746.41	Financial Statements
Net Acquisitions	-32,625,967.26	-805,960,000.00	0.00	249,000,000.00	115,701,304.21	Financial Statements
Net Cash Provided by Operating Activities	354,893,994.40	-179,404,000.00	144,174,000.00	3,870,000,000.00	549,880,984.75	Financial Statements
Net Cash Used for Investing Activities	-253,339,898.47	-2,840,033,000.00	-71,524,000.00	325,900,000.00	444,328,516.72	Financial Statements
Net Cash Used or Provided by Financing Activities	-116,189,055.63	-2,444,000,000.00	-29,317,500.00	1,094,000,000.00	399,791,510.60	Financial Statements
Net Change in Cash	4,919,539.31	-1,161,000,000.00	439,500.00	1,401,000,000.00	269,603,027.07	Financial Statements
Net Debt	3,582,100,670.97	-1,044,500,000.00	1,498,679,500.00	30,761,000,000.00	5,313,477,844.99	Financial Statements
Net Income (Cash Flow Statement)	189,838,210.19	-327,000,000.00	66,869,500.00	2,402,000,000.00	337,034,710.22	Financial Statements
Net Income (Income Statement)	186,964,309.93	-329,864,000.00	67,050,000.00	2,340,000,000.00	332,573,670.18	Financial Statements
Net Income Ratio	0.05	-8.88	0.07	2.72	0.29	Financial Statements
Net Property Plant Equipment	4,946,156,314.31	0.00	1,413,197,500.00	44,441,000,000.00	7,878,581,995.75	Financial Statements
Net Receivables	1,283,363,100.04	-4,199,600.00	569,700,000.00	12,146,000,000.00	1,794,448,777.75	Financial Statements
Non-Current Deferred Revenue	248,223,425.63	-500,933,000.00	0.00	5,778,000,000.00	725,439,952.22	Financial Statements
Non-Current Deferred Tax Liabilities	704,046,356.92	-3,818,507.00	137,248,500.00	8,306,000,000.00	1,394,360,238.89	Financial Statements
Operating Cash Flow	354,893,994.40	-179,404,000.00	144,174,000.00	3,870,000,000.00	549,880,984.75	Financial Statements
Operating Expenses	546,048,778.26	-13,530,000.00	222,695,500.00	6,252,000,000.00	932,339,076.14	Financial Statements
Operating Income	302,408,763.01	-208,377,000.00	123,446,500.00	3,294,000,000.00	472,764,842.78	Financial Statements
Operating Income Ratio	0.11	-9.71	0.12	2.86	0.31	Financial Statements
Other Assets	6,824.88	-19,834,700.00	0.00	8,948,000.00	427,499.44	Financial Statements
Other Current Assets	371,170,937.98	-98,000.00	120,313,000.00	4,968,950,000.00	665,138,617.21	Financial Statements
Other Current Liabilities	971,625,467.12	-48,317,000.00	326,047,500.00	12,137,000,000.00	1,812,385,139.72	Financial Statements
Other Expenses	50,748,275.86	-64,000,000.00	571,500.00	16,189,674,590.00	338,468,185.65	Financial Statements
Other Financing Activities	216,082,622.68	-975,168,999.00	8,000,000.00	3,297,501,000.00	513,442,205.81	Financial Statements
Other Investing Activities	4,846,309.09	-448,000,000.00	108,000.00	3,060,433,659.00	96,863,864.12	Financial Statements
Other Liabilities	101,452.43	-3,063,000.00	0.00	51,076,000.00	2,036,383.74	Financial Statements

Continued on next page

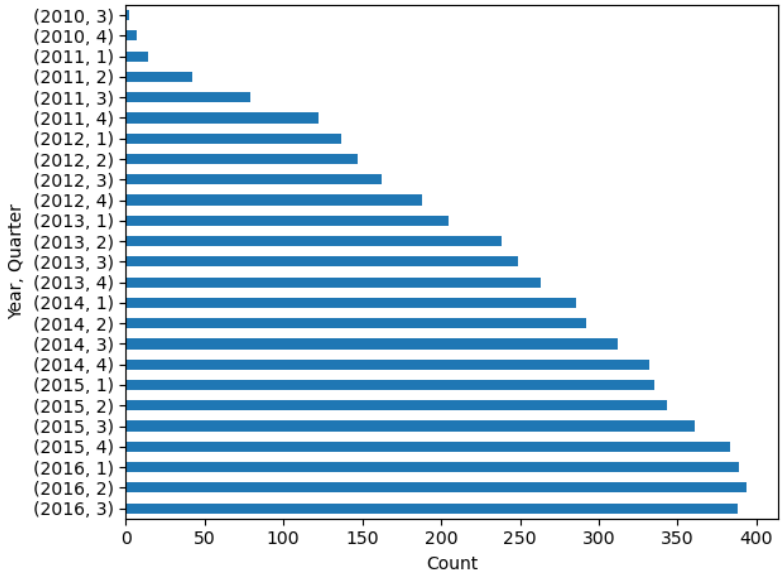
Table A.1: Numeric Summary Statistics

Variable Name	Mean	Minimum	Median	Maximum	Standard Deviation	Variable Type
Other Non-Cash Items	15,426,331.87	-1,848,719,007.00	1,600,000.00	703,000,000.00	109,962,680.11	Financial Statements
Other Non-Current Assets	506,812,134.37	-75,012,534,818.00	160,004,000.00	8,037,000,000.00	1,760,996,004.39	Financial Statements
Other Non-Current Liabilities	977,014,033.76	-286,041,895.00	328,324,500.00	11,890,564,000.00	1,685,843,225.83	Financial Statements
Other Total Stockholders' Equity	1,130,983,775.67	-12,393,000,000.00	421,317,000.00	34,030,400,000.00	3,589,780,224.87	Financial Statements
Other Working Capital	21,147,653.98	-1,788,851,160.00	0.00	40,341,689,407.00	775,619,463.26	Financial Statements
Preferred Stock	9,365,360.70	0.00	0.00	401,500,000.00	42,432,348.37	Financial Statements
Purchases of Investments	-102,973,220.34	-11,997,654,000.00	0.00	81,823,000.00	344,612,110.05	Financial Statements
Research and Development Expenses	27,530,996.50	-214,000.00	0.00	893,000,000.00	93,165,301.96	Financial Statements
Retained Earnings	3,643,670,803.88	-4,839,000,000.00	1,300,015,500.00	37,899,000,000.00	6,417,848,831.20	Financial Statements
Revenue	2,763,740,179.56	-4,273,000.00	1,308,000,000.00	25,420,000,000.00	4,026,578,837.74	Financial Statements
Sales and Maturities of Investments	97,797,529.17	-9,409,000.00	0.00	8,936,406,000.00	307,975,538.51	Financial Statements
Selling General and Administrative Expenses	300,347,623.86	-5,054,000.00	120,995,500.00	3,343,000,000.00	491,028,139.53	Financial Statements
Selling and Marketing Expenses	25,644,369.54	-3,003,000.00	0.00	876,761,000.00	99,025,764.56	Financial Statements
Short Term Investments	178,641,675.15	-515,000.00	0.00	6,178,000,000.00	590,076,249.16	Financial Statements
Short-Term Debt	464,412,829.97	-655,561.00	83,408,000.00	5,363,000,000.00	881,907,750.85	Financial Statements
Stock-Based Compensation	14,543,170.39	-36,000,000.00	5,100,000.00	254,000,000.00	30,176,788.95	Financial Statements
Tax Assets	376,499,884.70	-2,310,712,000.00	49,195,000.00	6,535,000,000.00	902,752,313.04	Financial Statements
Tax Payable	61,484,524.50	-87,400.00	2,828,000.00	1,187,000,000.00	152,142,261.85	Financial Statements
Total Assets	15,623,432,251.63	123,279.00	7,059,132,500.00	131,556,000,000.00	21,907,186,479.15	Financial Statements
Total Current Assets	3,959,899,836.06	29,954.00	1,936,936,500.00	41,276,000,000.00	5,741,692,814.20	Financial Statements
Total Current Liabilities	2,835,208,388.08	12,178.00	1,142,000,000.00	29,919,000,000.00	4,276,706,355.30	Financial Statements
Total Debt	4,591,360,324.25	0.00	2,025,288,000.00	37,124,000,000.00	6,242,100,940.74	Financial Statements
Total Equity	4,987,498,022.07	-501,467,000.00	2,115,000,000.00	49,975,000,000.00	7,273,418,780.53	Financial Statements
Total Investments	728,580,576.01	-334,673,000.00	42,569,500.00	19,331,000,000.00	1,963,853,124.76	Financial Statements
Total Liabilities	9,810,477,769.61	79,283.00	4,315,644,500.00	87,293,000,000.00	13,464,804,883.88	Financial Statements
Total Liabilities and Stockholders' Equity	15,588,449,533.11	123,279.00	7,057,469,000.00	131,556,000,000.00	21,902,353,741.35	Financial Statements
Total Liabilities and Total Equity	15,588,449,533.11	123,279.00	7,057,469,000.00	131,556,000,000.00	21,902,353,741.35	Financial Statements
Total Non-Current Assets	11,003,567,097.67	49,861.00	4,170,300,000.00	104,263,000,000.00	15,962,173,630.31	Financial Statements
Total Non-Current Liabilities	6,631,926,419.44	53,696.00	2,823,750,000.00	54,300,000,000.00	9,401,594,024.20	Financial Statements
Total Other Income Expenses Net	-13,084,263.76	-503,976,000.00	-900,000.00	286,000,000.00	72,543,230.98	Financial Statements
Total Stockholders' Equity	4,949,232,739.44	-526,491,000.00	2,110,773,500.00	49,269,000,000.00	7,190,925,518.03	Financial Statements
Weighted Average Shares Outstanding	355,222,897.10	0.00	146,624,411.00	13,751,391,147.00	727,556,653.14	Financial Statements
Weighted Average Shares Outstanding (Diluted)	318,383,151.79	0.00	146,042,500.00	13,986,214,405.00	548,657,214.64	Financial Statements
Market Capitalization	19,114,488,572.95	106,422.00	6,415,265,242.50	726,320,349,360.00	44,481,854,214.23	Market Capitalization
Days Since Call	58.13	0.00	61.00	91.00	13.42	Metadata
First Principal Component of Tone	-0.03	-2.91	-0.22	25.35	1.32	NLP Feature
Gunning-Fog Score	12.51	8.55	12.42	19.29	1.32	NLP Feature
Number of Questions	36.47	0.00	35.00	107.00	16.38	NLP Feature
Numeric Transparency	0.12	0.01	0.12	0.40	0.05	NLP Feature
Positivity Score	0.19	-0.20	0.18	0.67	0.10	NLP Feature
Word Count	8,829.04	525.00	9,080.00	22,006.00	2,479.44	NLP Feature
Change Since Last Fixed Quarter Date	0.01	-2.00	0.00	2.00	0.27	Predicted - Change

A.2 Observations by Quarter and Year

Figure A.1 demonstrates that the data is temporally unbalanced, with many companies entering the dataset in later years, after they first receive an observable credit rating.

Figure A.1: Observations by Quarter and Year



A.3 Altman's Z-Score

As in Das et al. (2023), the components of the Z-score are as follows:

- A: EBIT / Total Assets
- B: Net Sales / Total Assets
- C: Market Capitalization / Total Liabilities
- D: Working Capital / Total Assets
- E: Retained Earnings / Total Assets

We Winsorize extreme values of Ratio A, B, D, and E by setting the top and bottom 2.5% of values to the 97.5 and 2.5 percentile, respectively. Due to the presence of additional outliers and the sourcing of market capitalization from a different dataset than the rest of the variables, Ratio C is instead Winsorized over the top and bottom 5% of values.

The ratios are combined via the following equation:

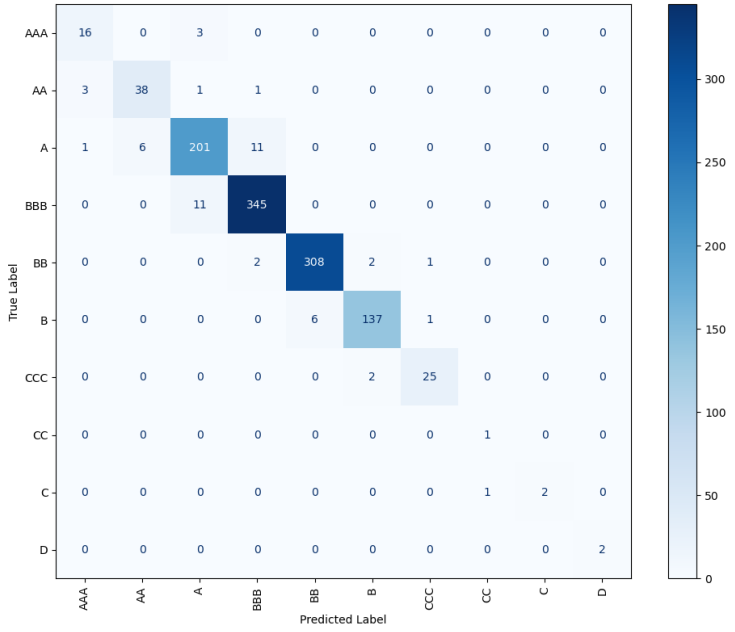
$$\text{Z-Score} = 3.3A + 0.99B + 0.6C + 1.2D + 1.4E$$

A.4 Logistic Regression - Most Complex Model - Additional Details

Table A.2: Best Hyperparameters - Most Complex Model

C	Class Weighting Strategy	L1 Ratio	Multi-Class Strategy	Penalty	Solver
0.10	Balanced	1.00	One vs Rest	Elastic Net	SAGA

Figure A.2: Confusion Matrix - Most Complex Model



A.5 Logistic Regression - Predicting Changes in Rating

Table A.3 shows that our most complex model (with the same variables as Rating Model 4) is able to predict changes in rating with a high degree of accuracy, and the weighted average statistics are as expected. Figure A.3 displays the confusion matrix. We fine-tuned our hyperHyperparameters for this model with an accuracy objective, and so grid search was allowed to completely ignores the non-majority classes and not perform balanced class weighting. More work is needed to either force balance weighting or change the grid search objective.

Table A.3: Classification Report - Change Prediction

Accuracy	Weighted Average Precision	Weighted Average Recall	F1 Score	Majority Baseline
0.96	0.91	0.96	0.93	0.96

Figure A.3: Confusion Matrix - Changes in Rating