Textual Analysis and Financial Statements

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

May 8, 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] Features capturing call readability, transparency, and engagement join pre-trained language model representations of sentiment (Araci, 2019) and traditional tabular variables as inputs to a variety of supervised machine learning techniques for classification from logistic regression to tree-based methods. We also make use of advances in the study of graph neural networks to model linkages between firms implied by mentions in calls. (Das et al., 2023)

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 neural language models and a wider variety of algorithms for classification.

Data and Exploratory Data Analysis

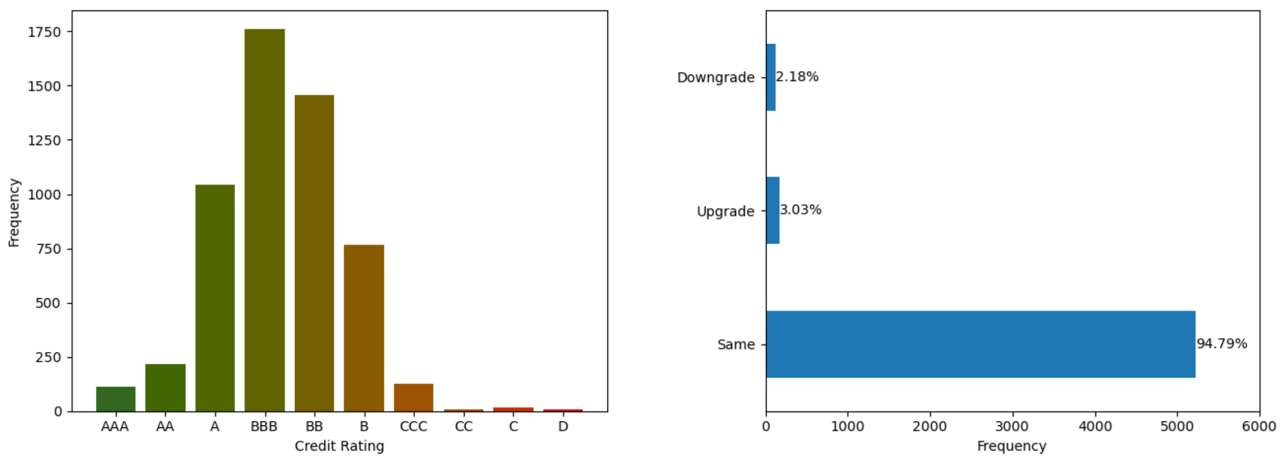
We combine a wide variety of data sources to support our predictions of credit ratings - merging rating data with company earnings calls, financial statement variables, and industry sector. In our combined dataset, each

[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, as well as some companies with only a few (3 or less) quarters. We identified one bankruptcy in our data - Peabody Energy on April 13, 2016 - and on further investigation, removed some quarters with incorrect ratings. In all, we have 5,509 quarters for 429 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 can 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 or end of data, 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.

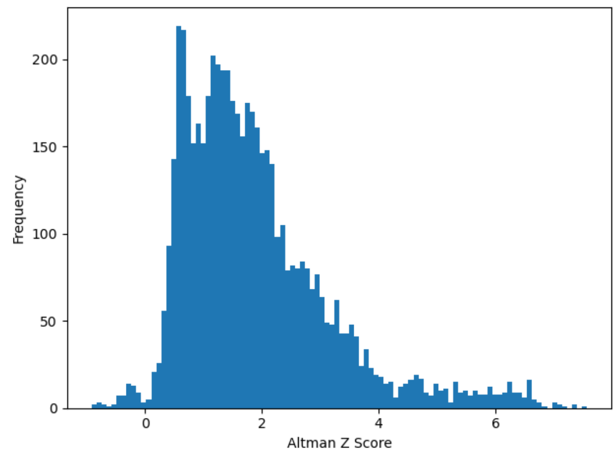
Figure 1 shows the distribution of rating grades used in our final dataset. Finer grades (AA+, CCC-, etc.) are sometimes assigned by agencies, but these grades were converted by dropping the +/- for this project. Ratings of BBB and above are considered investment grade - these bonds carry empirical one-year 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, 94.79% of ratings remain the same. Rating on the previous fixed quarter date can thus be an extremely strong predictor.

Earnings Calls

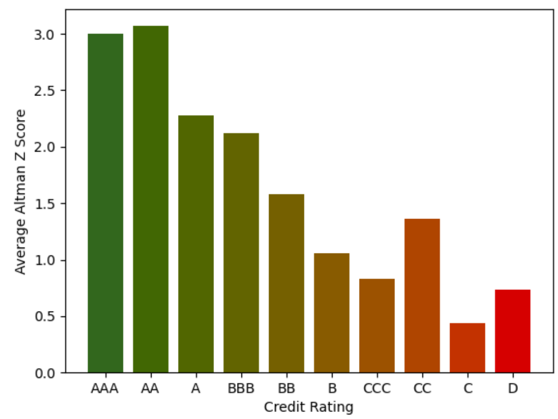
Our earnings call data comes from the Financial Modelling Prep API (Financial Modeling Prep, 2024), a trusted source widely used in industry. We remove all calls that happened more than 250 days prior and after the first day

Figure 2: Altman Z-Score

(a) Distribution



(b) Average by Rating



of the year and quarter they are supposed to discuss the results from, as well as calls for companies that provide them on an annual, rather than quarterly basis. Including both prepared remarks and analyst Q and A sessions, the overall average call length in our final data stands at 8,759.68 words.

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. We also calculated and included a wide variety of ratios, levels, and changes in variables. (for a list, see variables marked as 'Financial Statements' in Table A.1)

To prepare the data, we limit our observations to items reported in USD, check for and correct values off by a factor of 1,000 as a result of parsing,[3] and check some accounting identities in Das et al. (2023),[4] 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 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 ends of the rating spectrum (where not many companies and ratings are available), Z-Score is likely to be highly useful as a predictor.

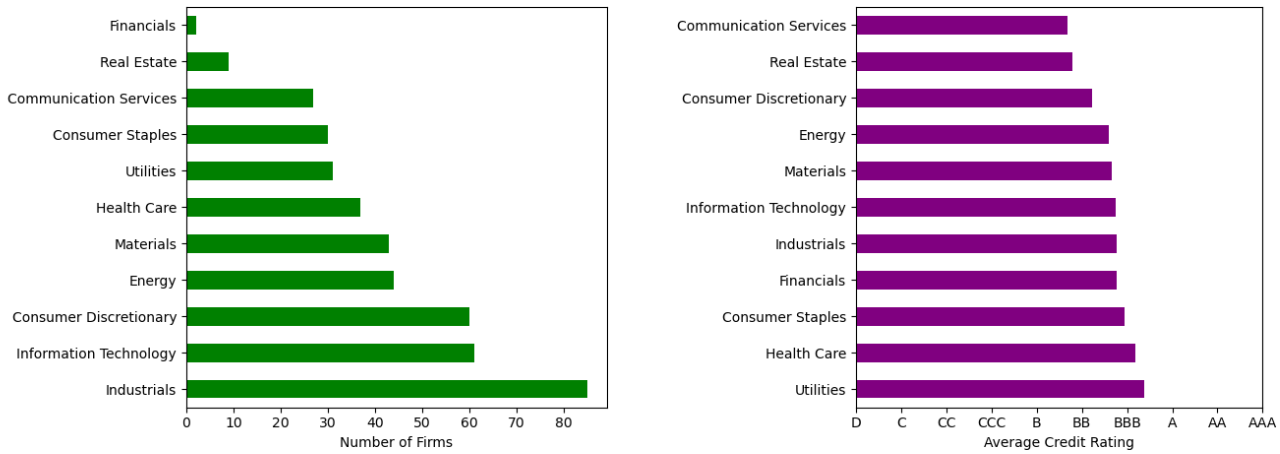
[3] If the last few digits are 000.00 and the item is above or below the 2.5% and 97.5% quantile, we divide by 1,000.

[4] We check total liabilities are greated than current liabilities, total assets are greater than total current assets, and net sales (revenue) is greated than EBIT. We originally also checked that total assets were greater than or equal to total equity + retained earnings + total liabilities, but this proved to be too restrictive.

Figure 3: Sector

(a) Firms by Sector

(b) Average Rating by Sector



Sector

The GCIS industry classification standard divides companies into 11 major industry sectors (S and P and MSCI, 2024).[5] 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 (Kozlov, 2022) and supplemented them with manual lookup. Figure 3 shows the sectoral imbalance present in our data, with a large share of firms in consumer, industrial, and technology sectors. However, when we quantize ratings and compute average values by sector, we do not see large differences, suggesting our results still may provide some generalizability. Though it is not yet clear that sector provides enough useful variation in rating to be a useful predictor, we still include it in our models, particularly as it may improve models including interactions (such as tree-based methods).

NLP Features

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

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

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

- Word Count
- Number of Questions - Count of question marks - Normalized by call length/word count
- Tone - Following Price et al. (2012), we use the Harvard dictionary to count words falling in various categories (Positive, Negative, Active, Passive, etc.). Then we construct tone using the first principal component of the

[5]There are finer groupings as well, but this data was not easily obtainable for our project.

matrix with each call as a row and each column as one of the following:

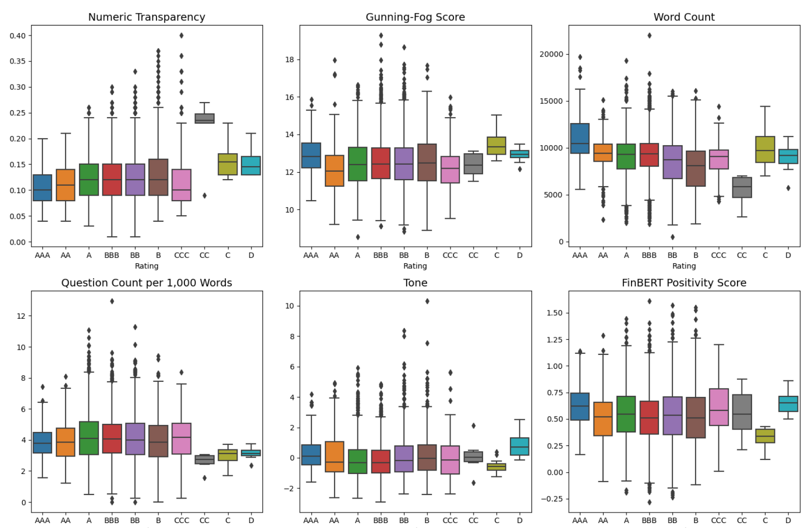
$$\frac{\text{Positive}}{\text{Negative}}, \frac{\text{Active}}{\text{Passive}}, \frac{\text{Strong}}{\text{Weak}}, \frac{\text{Overstated}}{\text{Understated}}$$

- FinBERT Positivity Score - [6]

We removed observations with outliers for these features, produced, for example, as a result of zeroes or low values in denominators.

The distribution of each NLP feature by rating is shown in Figure 4 below. Lower quality companies seem to provide more numbers with less commentary and also have less readable calls (higher Gunning-Fog grade level). It appears to be the case that higher quality companies tend to have longer calls. Though somewhat noisy, our FinBERT positivity score does seem to correlate with higher ratings.

Figure 4: Distribution of NLP Features by Rating



Network of Firms

In addition to our standard NLP features, which already capture a rich representation of calls, we also created a network graph representing the connections between firms based on mentions within calls. We deployed transformer-based Named-Entity Recognition (NER) (spaCy, 2024) to identify company names in the text, then matched these names to standardized versions. An interative visualization of our entire network of firms (aggregating mentions up from the call level - where we also have a network) can be found at https://sites.google.com/view/isaac-liu/company-mentions-network?authuser=0, and a 50% sample of nodes (faster load time) can be found at https://sites.google.com/view/isaac-liu/co-mentions-50-node-sample?authuser=0.

Modelling

Our overall model architecture is of the form

[6] We originally considered directly incorporating FinBERT embeddings into our models, or creating an end-to-end classifier making use of a BERT model. Our calls, however, are too long for readily available transformer embeddings or models to efficiently and effectively represent.

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

Logistic Regression

Table 1: Logistic Regression Model Comparison

Model/Baseline	Accuracy
Altman's Z	0.7442
Financial Variables and Sector	0.9508
Financial Variables, Sector, and NLP Features	0.9508
Majority Baseline	0.3247

Include Previous Rating

Model/Baseline	Accuracy
Altman's Z	0.1923
Financial Variables and Sector	0.6225
Financial Variables, Sector, and NLP Features	0.6333
Majority Baseline	0.3247

Exclude Previous Rating

Table **??** shows prediction statistics for our initial set of classifiers - simple and interpretable logistic regression models aiming to predict ratings (for predicting changes in rating, see Appendix Section A.6). 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 a wide variety of metrics. 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.

The left side of Table **??** 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 inform our use of these 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).

The right side of Table **??** shows the 15 most important features as determined by the average drop in test accuracy when the feature is permuted 1,000 times (we are also working on assessing coefficient significance). It is clear that previous rating is driving success for our predictions, without much clear contribution from NLP features at the moment.

Table 2: Most Complex Logistic Regression Model - Permutation Importance

Permuted Feature	Mean Accuracy Drop	Standard Deviation
Rating on Previous Fixed Quarter Date BB	0.256178	0.009675
Rating on Previous Fixed Quarter Date BBB	0.233306	0.008979
Rating on Previous Fixed Quarter Date A	0.111181	0.006236
Rating on Previous Fixed Quarter Date B	0.064464	0.003919
Rating on Previous Fixed Quarter Date CCC	0.013557	0.001143
Rating on Previous Fixed Quarter Date AA	0.010714	0.001722
Rating on Previous Fixed Quarter Date D	0.001829	0.000050
Ratio D	0.000866	0.000799
Weighted Average Shares Outstanding (Diluted)	0.000849	0.000249
Other Expenses	0.000840	0.000262
Net Income Ratio	0.000713	0.000779
Numeric Transparency	0.000703	0.000566
EBITDA	0.000681	0.000428
Ratio C	0.000412	0.000538
Ratio B	0.000130	0.000340

Include Previous Rating

Permuted Feature	Mean Accuracy Drop	Standard Deviation
Ratio E	0.070625	0.009156
Passive Tone	0.056786	0.007741
Sector: Utilities	0.043208	0.005661
Interest Expense	0.043019	0.007802
Ratio D	0.041765	0.007607
Ratio C	0.040578	0.008042
Depreciation and Amortization (Income Statement)	0.038593	0.007163
Net Receivables	0.036435	0.007130
Word Count	0.035743	0.007747
Long-Term Debt	0.035463	0.007198
Market Capitalization	0.031103	0.006867
Goodwill and Intangible Assets	0.030084	0.007430
Gross Profit	0.027059	0.006837
Total Debt	0.026485	0.007413
Net Debt	0.025156	0.007661

Exclude Previous Rating

Checking the sign of coefficients

Table 3: XGBoost Model Comparison

Model/Baseline	Accuracy
Altman's Z	0.9517
Financial Variables and Sector	0.9535
Financial Variables, Sector, and NLP Features	0.9535
Majority Baseline	0.3247

Include Previous Rating

Model/Baseline	Accuracy
Altman's Z	0.3855
Financial Variables and Sector	0.7630
Financial Variables, Sector, and NLP Features	0.9034
Majority Baseline	0.3247

Exclude Previous Rating

Table 4: Most Complex XGBoost Model - Permutation Importance

Permuted Feature	Mean Accuracy Drop	Standard Deviation
Rating on Previous Fixed Quarter Date BB	0.276554	0.010192
Rating on Previous Fixed Quarter Date BBB	0.257352	0.010267
Rating on Previous Fixed Quarter Date B	0.080826	0.004940
Rating on Previous Fixed Quarter Date A	0.047979	0.004233
Rating on Previous Fixed Quarter Date AA	0.036817	0.001890
Rating on Previous Fixed Quarter Date CCC	0.025477	0.002348
Rating on Previous Fixed Quarter Date AAA	0.021269	0.002349
Net Property Plant Equipment	0.001779	0.000093
Rating on Previous Fixed Quarter Date C	0.000900	0.000098
Cash Per Share	0.000834	0.000225
Return on Capital Employed	0.000024	0.000150
Market Capitalization	0.000022	0.000140
Operating Cash Flow to Sales	0.000020	0.000131
Cash at Beginning of Period	0.000000	0.000000
Interest Income	0.000000	0.000000

Include Previous Rating

Permuted Feature	Mean Accuracy Drop	Standard Deviation
Retained Earnings	0.043819	0.005761
Market Capitalization	0.035169	0.005735
Dividends Paid	0.021455	0.004481
Debt Ratio	0.009987	0.003413
Common Stock	0.009693	0.002248
Ratio E	0.009535	0.003463
Other Total Stockholders' Equity	0.009288	0.003028
Total Current Liabilities	0.006888	0.002785
Inventory (Balance Sheet)	0.006802	0.002994
Total Current Assets	0.006684	0.003243
Selling General and Administrative Expenses	0.006031	0.002395
Interest Expense	0.005973	0.002740
Net Property Plant Equipment	0.005729	0.001915
Ratio C	0.005677	0.002476
Total Non-Current Assets	0.005589	0.003420

Exclude Previous Rating

XGBoost

Graph Neural Network

Conclusion

Overall, we have seen that we are able to predict credit ratings with a high degree of accuracy, but at the moment our results are largely driven by inclusion of the previous rating as a predictor. Our current NLP and textual features are unable to contribute much to improve our predictions.

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].

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, 16: 321–357. arXiv:1106.1813 [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.

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

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."

spaCy. 2024. "spaCy · Industrial-strength Natural Language Processing in Python."

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
Difference in Cash Per Share from prior fixed quarter	-0.01	-69.19	0.00	69.02	4.54	Additional Change Ratios
Difference in Cash Ratio from prior fixed quarter	0.05	-53.11	0.00	53.11	3.98	Additional Change Ratios
Difference in Debt Ratio (Alternative) from prior fixed quarter	0.00	-0.76	0.00	0.78	0.05	Additional Change Ratios
Difference in Debt Ratio from prior fixed quarter	0.00	-0.82	0.00	0.83	0.05	Additional Change Ratios
Difference in Debt to Equity Ratio from prior fixed quarter	-1.90	-1,915.81	0.00	1,892.37	113.71	Additional Change Ratios
Difference in EBIT to Revenue from prior fixed quarter	-0.00	-0.66	0.00	0.59	0.09	Additional Change Ratios
Difference in Enterprise Value Multiplier from prior fixed quarter	0.25	-1,036.95	0.00	1,036.95	121.22	Additional Change Ratios
Difference in Equity Multiplier from prior fixed quarter	-1.33	-1,292.75	0.00	1,292.75	80.06	Additional Change Ratios
Difference in Free Cash Flow Per Share from prior fixed quarter	0.01	-10.68	0.01	10.68	1.83	Additional Change Ratios
Difference in Free Cash Flow to Operating Cash Flow from prior fixed quarter	0.01	-13.40	0.00	13.40	2.40	Additional Change Ratios
Difference in Operating Cash Flow Per Share from prior fixed quarter	0.01	-12.80	0.02	12.80	1.79	Additional Change Ratios
Difference in Operating Cash Flow to Sales from prior fixed quarter	0.00	-0.79	0.01	0.79	0.14	Additional Change Ratios
Difference in Quick Ratio from prior fixed quarter	-0.00	-5.30	0.00	5.16	0.51	Additional Change Ratios
Difference in Return on Assets from prior fixed quarter	-0.00	-0.10	0.00	0.10	0.01	Additional Change Ratios
Difference in Return on Capital Employed from prior fixed quarter	-0.00	-0.14	0.00	0.13	0.02	Additional Change Ratios
Difference in Return on Equity from prior fixed quarter	-0.00	-2.11	0.00	2.11	0.23	Additional Change Ratios
Differnce in Current Ratio from prior fixed quarter	-0.00	-6.87	0.00	6.97	0.61	Additional Change Ratios
Cash Per Share	4.57	0.00	2.13	69.91	9.64	Additional Ratios
Cash Ratio	1.21	0.00	0.28	53.17	6.23	Additional Ratios
Current Ratio	1.93	0.35	1.58	7.93	1.33	Additional Ratios
Debt Ratio	0.35	0.00	0.32	0.94	0.19	Additional Ratios
Debt Ratio (Alternative Definition)	0.65	0.28	0.64	1.22	0.17	Additional Ratios
Debt to Equity Ratio	-34.63	-1,890.41	1.70	25.40	256.02	Additional Ratios
EBIT to Revenue	0.12	-0.26	0.11	0.47	0.12	Additional Ratios
Enterprise Value Multiplier	59.08	-309.75	40.67	727.20	125.93	Additional Ratios
Equity Multiplier	-22.79	-1,270.10	2.71	22.64	175.08	Additional Ratios
Free Cash Flow Per Share	0.57	-2.98	0.39	7.70	1.52	Additional Ratios
Free Cash Flow to Operating Cash Flow	0.72	-2.42	0.66	10.98	1.86	Additional Ratios
Operating Cash Flow Per Share	1.48	-0.98	1.04	11.82	1.92	Additional Ratios
Operating Cash Flow to Sales	0.16	-0.15	0.14	0.64	0.15	Additional Ratios
Quick Ratio	1.36	0.00	1.15	6.12	0.98	Additional Ratios
Return on Assets	0.01	-0.03	0.01	0.06	0.02	Additional Ratios
Return on Capital Employed	0.03	-0.03	0.02	0.11	0.03	Additional Ratios
Return on Equity	0.01	-1.32	0.03	0.78	0.25	Additional Ratios
Altman's Z Score	1.88	-0.91	1.61	7.56	1.28	Altman's Z Score
Difference in Altman's Z from prior fixed quarter	-0.01	-4.84	0.01	4.41	0.39	Change Ratios
Difference in EBITDA Ratio from prior fixed quarter	-0.00	-3.09	0.00	5.20	0.16	Change Ratios
Difference in Gross Profit Ratio from prior fixed quarter	-0.00	-3.16	0.00	5.23	0.15	Change Ratios
Difference in Income Before Tax Ratio from prior fixed quarter	-0.00	-6.69	0.00	6.43	0.33	Change Ratios
Difference in Net Income Ratio from prior fixed quarter	-0.00	-7.13	0.00	5.45	0.28	Change Ratios
Difference in Operating Income Ratio from prior fixed quarter	-0.00	-7.36	0.00	5.20	0.27	Change Ratios
Difference in Ratio A from prior fixed quarter	-0.00	-0.10	0.00	0.10	0.01	Change Ratios
Difference in Ratio B from prior fixed quarter	-0.00	-0.56	0.00	0.57	0.04	Change Ratios
Difference in Ratio C from prior fixed quarter	-0.01	-7.77	0.01	7.15	0.57	Change Ratios
Difference in Ratio D from prior fixed quarter	-0.00	-0.57	0.00	0.60	0.05	Change Ratios
Difference in Ratio E from prior fixed quarter	0.00	-0.80	0.00	0.98	0.07	Change Ratios
Accounts Payable (Balance Sheet)	957,290,323.93	-237,651,171.00	356,700,000.00	11,433,000,000.00	1,551,108,353.02	Financial Statements
Accounts Payable (Cash Flow Statement)	5,154,565.15	-321,769,000.00	0.00	1,789,652,000.00	82,110,968.91	Financial Statements
Accounts Receivables	-11,478,236.25	-544,000,000.00	0.00	325,000,000.00	91,535,961.30	Financial Statements
Accumulated Other Comprehensive Income (Loss)	-404,483,300.22	-5,290,000,000.00	-77,514,000.00	431,595,000.00	874,353,108.41	Financial Statements
Capital Expenditure	-192,514,484.47	-1,867,000,000.00	-60,129,000.00	412,700.00	310,057,440.27	Financial Statements
Capital Lease Obligations	24,642,498.79	0.00	0.00	9,056,234,000.00	228,328,885.18	Financial Statements
Cash and Cash Equivalents	862,135,865.07	0.00	333,000,000.00	9,223,000,000.00	1,366,595,243.17	Financial Statements
Cash and Short Term Investments	1,060,086,810.64	0.00	363,008,000.00	15,601,000,000.00	1,890,682,420.93	Financial Statements
Cash at Beginning of Period	867,410,489.82	-2,556,000.00	334,000,000.00	9,610,000,000.00	1,388,834,800.13	Financial Statements
Cash at End of Period	871,017,693.39	-154,400.00	335,469,000.00	9,743,000,000.00	1,394,641,397.30	Financial Statements
Change in Working Capital	-17,557,103.20	-870,000,000.00	-2,384,000.00	753,000,000.00	183,788,257.05	Financial Statements
Common Stock	329,277,684.36	-539,800.00	3,800,000.00	9,817,134,000.00	925,626,949.20	Financial Statements
Common Stock Issued	44,672,509.36	-3,572,000.00	43,000.00	1,111,490,728.00	124,027,450.20	Financial Statements
Common Stock Repurchased	-78,527,033.90	-2,086,545,366.00	-773,000.00	545,656,614.52	188,219,352.34	Financial Statements
Cost and Expenses	2,317,513,877.07	-2,495,000.00	1,121,064,000.00	22,769,000,000.00	3,357,899,606.58	Financial Statements
Cost of Revenue	1,624,233,369.18	-3,094,000.00	787,700,000.00	18,303,000,000.00	2,405,765,370.43	Financial Statements
Debt Repayment	-247,880,234.24	-3,001,000,000.00	-33,400,000.00	200.00	471,724,050.37	Financial Statements
Deferred Income Tax	6,154,669.54	-253,000,000.00	64,000.00	1,850,454,000.00	58,927,713.28	Financial Statements
Deferred Revenue	310,000,739.66	-116,912,000.00	50,066,000.00	4,918,100,000.00	642,489,899.31	Financial Statements
Depreciation and Amortization (Cash Flow Statement)	141,811,048.14	-675,312.00	53,551,000.00	1,529,000,000.00	210,315,836.18	Financial Statements
Depreciation and Amortization (Income Statement)	140,571,212.83	-1,550,000.00	54,507,000.00	1,371,000,000.00	203,167,331.44	Financial Statements
Diluted EPS	0.51	-156.36	0.51	49.73	3.31	Financial Statements
Dividends Paid	-91,357,096.76	-1,233,000,000.00	-21,054,000.00	0.00	182,429,714.55	Financial Statements
EBITDA	444,995,396.82	-66,200,000.00	193,000,000.00	4,410,000,000.00	644,706,471.62	Financial Statements

Continued on next page

Table A.1: Numeric Summary Statistics

Variable Name	Mean	Minimum	Median	Maximum	Standard Deviation	Variable Type
EBITDA Ratio	0.20	-5.77	0.17	2.16	0.22	Financial Statements
EPS	0.52	-156.36	0.52	53.75	3.33	Financial Statements
Effect of Foreign Exchange Changes on Cash	-1,697,085.83	-65,000,000.00	0.00	52,000,000.00	11,200,007.88	Financial Statements
Free Cash Flow	156,892,657.81	-541,000,000.00	51,691,000.00	2,683,000,000.00	389,666,937.19	Financial Statements
General and Administrative Expenses	153,933,016.99	-2,738,500.00	33,768,000.00	2,007,000,000.00	303,900,948.38	Financial Statements
Goodwill	2,009,260,205.06	-202,702,100.00	636,039,000.00	23,389,000,000.00	3,554,057,246.39	Financial Statements
Goodwill and Intangible Assets	3,102,882,804.88	-1,618,944,000.00	970,000,000.00	37,123,000,000.00	5,639,038,312.52	Financial Statements
Gross Profit	861,821,178.07	-7,195,000.00	378,500,000.00	9,223,000,000.00	1,365,410,717.45	Financial Statements
Gross Profit Ratio	0.37	-5.65	0.34	2.32	0.26	Financial Statements
Income Before Tax	255,351,974.53	-353,153,000.00	91,900,000.00	2,951,000,000.00	434,623,029.43	Financial Statements
Income Before Tax Ratio	0.07	-9.38	0.09	2.68	0.35	Financial Statements
Income Tax Expense	69,444,774.33	-119,131,000.00	22,100,000.00	736,000,000.00	121,681,731.43	Financial Statements
Intangible Assets	835,940,509.51	-421,000.00	170,197,000.00	14,110,100,000.00	1,785,542,119.17	Financial Statements
Interest Expense	46,568,508.69	-16,400,000.00	23,000,000.00	386,000,000.00	61,712,161.15	Financial Statements
Interest Income	2,372,725.23	-62,900.00	0.00	69,000,000.00	6,859,086.75	Financial Statements
Inventory (Balance Sheet)	933,043,177.40	-19,626,000.00	403,789,000.00	8,328,000,000.00	1,398,934,358.21	Financial Statements
Inventory (Cash Flow Statement)	-10,302,495.14	-420,000,000.00	0.00	289,000,000.00	70,374,129.32	Financial Statements
Investments in Property, Plants, and Equipment	-193,897,744.95	-1,921,864,000.00	-60,373,000.00	412,700.00	313,436,441.14	Financial Statements
Long-Term Debt	4,159,473,460.27	-651,718.00	1,822,139,000.00	31,359,000,000.00	5,574,538,232.32	Financial Statements
Long-Term Investments	494,196,440.41	-490,677,000.00	12,449,000.00	10,981,000,000.00	1,359,571,399.50	Financial Statements
Minority Interest	90,043,651.07	-20,252,654.04	1,600,000.00	2,316,406,000.00	268,200,905.93	Financial Statements
Net Acquisitions	-32,878,764.18	-805,960,000.00	0.00	249,000,000.00	116,107,004.20	Financial Statements
Net Cash Provided by Operating Activities	352,446,106.81	-179,404,000.00	143,626,000.00	3,870,000,000.00	545,602,564.63	Financial Statements
Net Cash Used for Investing Activities	-252,575,304.44	-2,840,033,000.00	-71,100,000.00	325,900,000.00	443,647,871.52	Financial Statements
Net Cash Used or Provided by Financing Activities	-114,570,062.00	-2,444,000,000.00	-29,157,000.00	1,094,000,000.00	399,330,481.52	Financial Statements
Net Change in Cash	3,933,018.18	-1,161,000,000.00	573,000.00	1,401,000,000.00	269,005,283.68	Financial Statements
Net Debt	3,597,141,664.59	-1,044,500,000.00	1,508,594,000.00	30,761,000,000.00	5,338,457,121.62	Financial Statements
Net Income (Cash Flow Statement)	189,122,176.12	-327,000,000.00	66,190,000.00	2,402,000,000.00	336,635,167.35	Financial Statements
Net Income (Income Statement)	185,944,828.27	-329,864,000.00	66,389,000.00	2,340,000,000.00	330,952,161.49	Financial Statements
Net Income Ratio	0.05	-8.88	0.07	2.72	0.29	Financial Statements
Net Property Plant Equipment	4,931,687,321.78	0.00	1,389,600,000.00	44,441,000,000.00	7,885,938,319.99	Financial Statements
Net Receivables	1,276,905,848.63	-4,199,600.00	570,338,000.00	12,116,000,000.00	1,776,578,353.43	Financial Statements
Non-Current Deferred Revenue	248,840,448.23	-500,933,000.00	0.00	5,778,000,000.00	723,186,467.01	Financial Statements
Non-Current Deferred Tax Liabilities	702,874,797.74	-3,818,507.00	135,597,000.00	8,306,000,000.00	1,400,029,509.57	Financial Statements
Operating Cash Flow	352,446,106.81	-179,404,000.00	143,626,000.00	3,870,000,000.00	545,602,564.63	Financial Statements
Operating Expenses	538,189,512.49	-13,530,000.00	221,700,000.00	6,252,000,000.00	918,426,909.60	Financial Statements
Operating Income	302,231,079.76	-208,377,000.00	122,000,000.00	3,294,000,000.00	475,077,278.15	Financial Statements
Operating Income Ratio	0.11	-9.71	0.12	2.86	0.31	Financial Statements
Other Assets	5,662.39	-19,834,700.00	0.00	8,948,000.00	421,776.93	Financial Statements
Other Current Assets	370,526,390.88	-98,000.00	119,600,000.00	4,968,950,000.00	664,643,317.21	Financial Statements
Other Current Liabilities	955,075,890.93	-48,317,000.00	322,800,000.00	12,137,000,000.00	1,782,231,297.37	Financial Statements
Other Expenses	50,749,806.82	-64,000,000.00	585,000.00	16,189,674,590.00	342,110,629.66	Financial Statements
Other Financing Activities	217,421,866.42	-975,168,999.00	8,000,000.00	3,297,501,000.00	515,334,960.45	Financial Statements
Other Investing Activities	4,573,739.09	-448,000,000.00	106,000.00	3,060,433,659.00	96,736,267.62	Financial Statements
Other Liabilities	95,902.58	-3,063,000.00	0.00	51,076,000.00	1,967,227.53	Financial Statements
Other Non-Cash Items	15,325,139.75	-1,848,719,007.00	1,621,000.00	703,000,000.00	109,294,805.79	Financial Statements
Other Non-Current Assets	506,778,121.04	-75,012,534,818.00	158,696,000.00	8,037,000,000.00	1,778,143,597.09	Financial Statements
Other Non-Current Liabilities	975,892,048.39	-286,041,895.00	327,700,000.00	11,890,564,000.00	1,686,827,873.95	Financial Statements
Other Total Stockholders' Equity	1,135,331,510.72	-12,393,000,000.00	427,000,000.00	34,030,400,000.00	3,586,435,863.55	Financial Statements
Other Working Capital	21,414,823.22	-1,788,851,160.00	0.00	40,341,689,407.00	786,599,061.35	Financial Statements
Preferred Stock	9,475,146.22	0.00	0.00	401,500,000.00	42,785,110.93	Financial Statements
Purchases of Investments	-104,151,034.82	-11,997,654,000.00	0.00	81,823,000.00	346,711,949.30	Financial Statements
Research and Development Expenses	28,169,938.85	-214,000.00	0.00	893,000,000.00	94,071,513.75	Financial Statements
Retained Earnings	3,628,393,969.72	-4,839,000,000.00	1,293,100,000.00	37,899,000,000.00	6,424,744,717.89	Financial Statements
Revenue	2,728,749,857.76	-4,273,000.00	1,297,700,000.00	25,420,000,000.00	3,959,362,594.26	Financial Statements
Sales and Maturities of Investments	99,796,411.86	-9,409,000.00	0.00	8,936,406,000.00	311,292,561.88	Financial Statements
Selling General and Administrative Expenses	296,899,615.00	-5,054,000.00	119,600,000.00	3,343,000,000.00	486,131,457.73	Financial Statements
Selling and Marketing Expenses	25,431,647.83	-3,003,000.00	0.00	876,761,000.00	97,367,023.08	Financial Statements
Short Term Investments	182,988,242.55	-515,000.00	0.00	6,178,000,000.00	599,747,024.65	Financial Statements
Short-Term Debt	465,870,869.02	-655,561.00	83,800,000.00	5,363,000,000.00	885,210,679.51	Financial Statements
Stock-Based Compensation	14,496,292.55	-36,000,000.00	5,106,000.00	254,000,000.00	29,968,462.79	Financial Statements
Tax Assets	378,132,518.58	-2,310,712,000.00	48,963,000.00	6,535,000,000.00	909,237,680.35	Financial Statements
Tax Payable	60,670,669.07	-87,400.00	2,810,000.00	1,187,000,000.00	150,628,980.40	Financial Statements
Total Assets	15,592,495,985.55	123,279.00	7,048,475,000.00	131,119,000,000.00	21,911,032,910.64	Financial Statements
Total Current Assets	3,937,085,272.11	29,954.00	1,933,750,000.00	41,276,000,000.00	5,729,273,613.69	Financial Statements
Total Current Liabilities	2,811,976,684.34	24,083.00	1,138,200,000.00	29,919,000,000.00	4,247,045,840.39	Financial Statements
Total Debt	4,593,265,532.66	0.00	2,019,244,000.00	37,124,000,000.00	6,254,194,800.16	Financial Statements
Total Equity	4,968,502,543.29	-501,467,000.00	2,095,000,000.00	49,975,000,000.00	7,272,421,518.55	Financial Statements
Total Investments	729,199,594.64	-334,673,000.00	43,275,000.00	19,331,000,000.00	1,944,649,108.26	Financial Statements
Total Liabilities	9,817,545,124.72	79,283.00	4,308,693,000.00	87,293,000,000.00	13,527,062,565.42	Financial Statements
Total Liabilities and Stockholders' Equity	15,556,696,866.65	123,279.00	7,043,426,000.00	131,119,000,000.00	21,905,884,302.05	Financial Statements
Total Liabilities and Total Equity	15,556,696,866.65	123,279.00	7,043,426,000.00	131,119,000,000.00	21,905,884,302.05	Financial Statements
Total Non-Current Assets	11,011,964,229.49	49,861.00	4,119,200,000.00	104,263,000,000.00	15,994,777,583.25	Financial Statements
Total Non-Current Liabilities	6,639,451,321.63	53,696.00	2,809,300,000.00	54,300,000,000.00	9,424,654,097.47	Financial Statements
Total Other Income Expenses Net	-13,134,652.92	-503,976,000.00	-920,000.00	286,000,000.00	72,414,124.07	Financial Statements
Total Stockholders' Equity	4,933,321,107.00	-526,491,000.00	2,088,608,000.00	49,269,000,000.00	7,194,176,771.15	Financial Statements
Weighted Average Shares Outstanding	352,790,171.17	0.00	146,000,000.00	13,751,391,147.00	720,460,888.99	Financial Statements
Weighted Average Shares Outstanding (Diluted)	316,630,108.94	0.00	145,951,913.00	13,986,214,405.00	547,337,219.46	Financial Statements
Market Capitalization	18,996,749,034.57	106,422.00	6,409,459,125.00	726,320,349,360.00	44,246,873,159.19	Market Capitalization
Days Since Call	58.39	0.00	61.00	91.00	13.05	Metadata
FinBERT Positivity Score	0.53	-0.28	0.52	1.61	0.25	NLP Feature
First Principal Component of Tone	-0.03	-2.91	-0.22	10.33	1.28	NLP Feature
Gunning-Fog Score	12.50	8.55	12.41	19.29	1.31	NLP Feature
Number of Questions	36.50	0.00	35.00	107.00	16.38	NLP Feature
Number of Questions Divided By Call Word Count	0.00	0.00	0.00	0.01	0.00	NLP Feature

Continued on next page

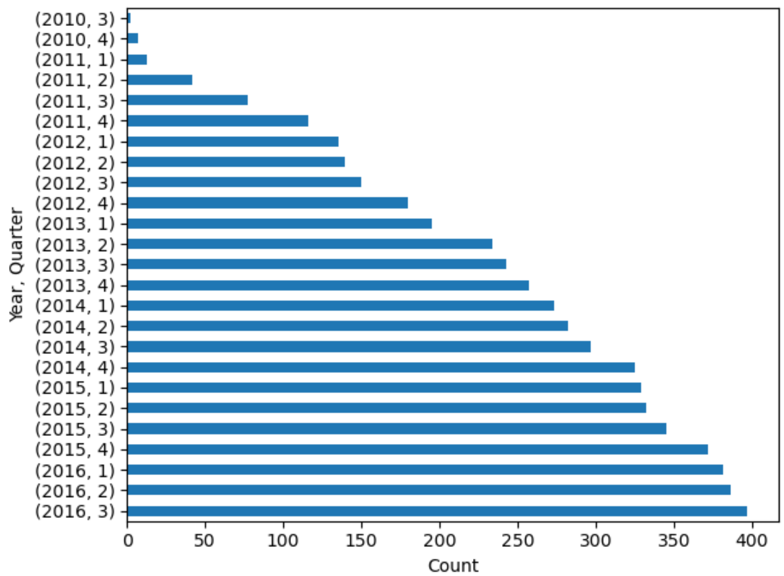
Table A.1: Numeric Summary Statistics

Variable Name	Mean	Minimum	Median	Maximum	Standard Deviation	Variable Type
Numeric Transparency	0.12	0.01	0.12	0.40	0.05	NLP Feature
Word Count	8,834.15	525.00	9,083.00	22,006.00	2,471.87	NLP Feature
Change Since Last Fixed Quarter Date	0.01	-2.00	0.00	2.00	0.26	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 **??** and Figure **??** show the high level of accuracy we are able to attain even for sparse classes when including all available features with an L1 penalty (elastic net with fully L1), balanced class weighting, and a simple one versus rest multiclass prediction setup (a binary is/is not logistic regression probability is estimated for each class, and class with the highest score is taken).

A.5 XGBoost - Most Complex Model - Additional Details

A.6 Predicting Changes in Rating

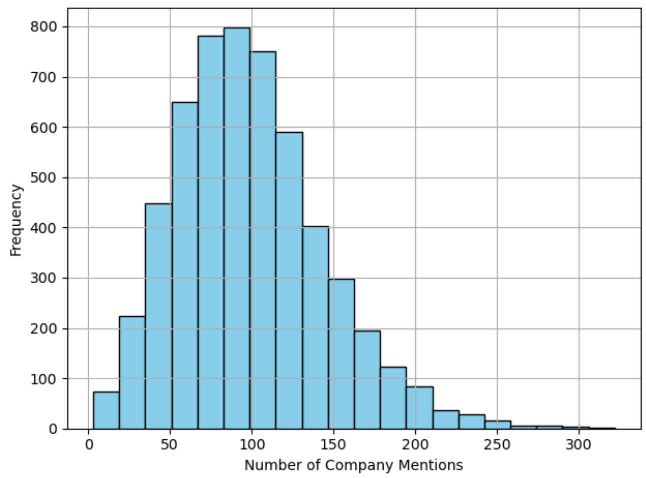
As shown in figure 1, 94.79% of ratings remain the same of ratings remain the same from one fixed quarter date to the next. This poses a serious challenge for classification, which is easily dominated by the majority class. We implemented SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) to oversample the minority classes in the trainig data and balance the dataset.

Table **??** 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 **??** displays the confusion matrix. We fine-tuned our hyperparameters for this model with an accuracy objective, and so grid search was allowed to completely ignore the non-majority classes and not perform balanced class weighting. More work is needed to either force balanced weighting or change the grid search objective.

A.7 Company Mentions

On average, each earnings call has 98.63 company mentions. Figure A.2 shows the distribution.

Figure A.2: Company Mentions



Though the vast majority of these mentions are likely to be of the company presenting the call, a casual glance at the data does suggest there are a fair number of mentions of partners, suppliers, and competitors. 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.