Bond Rating and Earning Calls

Corporate Credit Ratings Predictions using Financial Metrics and Earning Calls

Sean Zhou
M.A. in Statistics
University of California, Berkeley
Teammates: Owen Lin, Issac Liu, Zhengxing Chen

September 23, 2024

1 Introduction

1.1 Overview

Corporate credit ratings play a crucial role in the financial industry, providing an assessment of a company's creditworthiness and its ability to meet its financial obligations. These ratings are widely used by investors, lenders, and other stakeholders to evaluate the many risks associated with investing in or lending to a particular company. These ratings are typically issued by independent credit rating agencies, which are organizations that specialize in evaluating firms and assigning credit ratings. They continuously monitor the firms they rate and usually initiate a review process on a firm's rating under circumstances like changes in company financial metrics, major corporate events as well as other macroeconomic factors. However, that credit review and update process can sometimes be lengthy and slow. Thus, it becomes invaluable to design mathematical models that provide accurate predictions on credit ratings independent from professional agencies, offering an early insight into a firm's risks and financial wellbeing.

1.2 Problem Statement and Questions

Our primary objective is to develop a robust predictive model for corporate credit ratings by leveraging both financial metrics and information extracted from earning call transcripts using natural language processing (NLP) techniques. Specifically, the research questions we aim to address are:

1. How do we best define the prediction task, multiclass classification or binary credit change? What models perform best?
2. Can the combination of financial metrics and NLP features derived from earning call transcripts improve the accuracy of corporate credit rating predictions compared to using financial metrics alone?
3. What are the most significant financial and textual features that influence credit rating predictions?

1.3 Roadmap of Project

We begin by collecting comprehensive data on financial metrics, earning call transcripts and credit rating data. We will then conduct data wrangling and processing to reconcile the differences in the convoluted structures of our datasets. For modeling, we will begin by running Multinomial Logistic Regression on company financial metrics as baseline and experiment with adding NLP features. We then implement XGBoost and investigate model performance and feature importance. Finally we switch to predicting rating change and tackle the data imbalance issue via oversampling methods.

2 Data

2.1 Data Source and Description

We obtain long-term credit rating issuance (AAA, AA, A, BBB, BB, B, CCC, CC, C, D) from January 1, 2010, to December 31, 2016 from Kaggle and only select those rated by

S&P Rating Services [1]. The dataset is organized at the individual company level and matched to fixed quarterly dates (January 1, April 1, July 1, October 1 of each year). The finalized dataset has 429 U.S. firms and 5,509 firms by quarters. For each company and quarter, we add the most recent transcript of earnings call up to that quarter date as well as many financial statement variables. We include key financial metrics from the balance sheet, cash flow statement, and income statement totaling 124 variables such as Debt Ratio, Current Assets, and Operating Cash Flow.

2.2 Data Wrangling

Due to the complex structure of the raw datasets, significant efforts are directed towards processing and reconciling the credit ratings with financial metrics. We correct frequent mismatches between dates and quarters by standardizing data to fixed quarterly dates. For consistency, we drop items with filling date outside of fixed quarter date and more than 45 days after call date. Additionally, sanity checks are conducted to correct numerical values mismultiplied by 1,000 in the income and balance statements due to parsing errors.

We also carry out extensive data cleaning for earning call texts. Problematic calls are replaced using web scraping, and those not within a 250-day window of the respective quarter date are removed to ensure data relevance. To obtain our NLP features, we utilize the *Dask* library in Python to parallelize operations across 16 cores, ensuring each function handles a single earning call corpus. Calls with less than 500 words or missing significant speaker content are discarded to enhance data quality.

2.3 Feature Engineering

We construct six main NLP features:

1. **Net Positivity Score**: we implement FINBERT, a variant of the BERT model that has been pre-trained on financial text data to get sentence-level embedding [9]. We then use them for classifying sentence sentiments into *positive*, *neural*, or *negative* and calculate the net positivity score [10]:

   $$NetPositivityScore = \log_{10} \frac{CountPositive + 1}{CountNegative + 1}$$

   This score should capture the level of optimism expressed by company executives.

2. **Tone**: Tone is a more comprehensive indicator of speaker sentiments than Net Positivity Score. In addition to *Positive* and *Negative*, we count the number of occurances for words in categories: *Active*, *Passive*, *Strong*, *Weak*, *Overstated*, and *Understated*. The categories are determined by referring to the Harvard IV-4 Psychosocial Dictionary, a list of opinionated words [14]. Then the overall tone is constructed by taking the first principle component of the matrix $A_{n \times p}$ where n is the number of calls and $p = 4$ are the features $[\frac{Positive_i}{Negative_i}, \frac{Active_i}{Passive_i}, \frac{Strong_i}{Weak_i}, \frac{Overstated_i}{Understated_i}]$ with $i = 1, 2, ..., n$ [11]. In later analysis, we also include the individual tone components in our models along with the overall tone.

3. **Numeric Transparency**: we tokenize the texts, remove punctuation and count the number of numerical occurrences as well as the unique word types. Numeric Transparency is simply the proportion of numerical occurrences to word occurrences [12]. Speakers are more likely to use statistics to support their statements in calls with high Numeric Transparency score.

4. **Analyst Engagement**: S&P 500 firms with less engagement from sell-side analysts in Q2 2017 saw an average performance drop of 2.14% in the following months [13]. Thus, we count the number of question marks adjusted by the length of call.

5. **Readability**: We use the Gunning-Fog Index to calculate how well an earning call can be interpreted [8]. The GF score is calculated as $0.4 \left[\left(\frac{\text{words}}{\text{sentences}} \right) + 100 \left(\frac{\text{complex words}}{\text{words}} \right) \right]$, where *complex words* have three or more

syllables. The higher the score, the less readable a text is. CEOs who use longer sentences and more complex words in earnings calls often signal potential declines in earnings and stock prices [13].

6. **Word Count**: The total number of tokens in a call.

Additionally, we use **Altman's Z-Score** as a summary measure of financial strength and bankruptcy risk [2]. It combines five financial ratios, represented as:

$$Z = 3.3A + 0.99B + 0.6C + 1.2D + 1.4E$$

where:

- $A = \frac{\text{EBIT}}{\text{Total Assets}}$
- $B = \frac{\text{Net Sales}}{\text{Total Assets}}$
- $C = \frac{\text{Market Value of Equity}}{\text{Total Liabilities}}$
- $D = \frac{\text{Working Capital}}{\text{Total Assets}}$
- $E = \frac{\text{Retained Earnings}}{\text{Total Assets}}$.

3 Exploratory Data Analysis (EDA)

3.1 Visualizations and Insights

From Figure 1, we observe that our dataset consist of slightly more firms with investment-grade ratings (AAA to BBB) than junk-grade ratings (BB - D). However, lower A- to B-grade ratings make up the majority of the observations. We note that the lack of observations for firms with low-grade ratings might pose a class imbalance issue, where our sampling distributions for those classes are not representative and the few samples might even be randomly omitted during the train-test split.

We do not provide visualizations on our tabular financial metrics due to having over 100 features. Figure 2 shows the mean Altman Z score by credit rating. The scores mostly drop as credit rating worsens except for the minority classes in the right tail, providing further evidence for the imbalance issue.

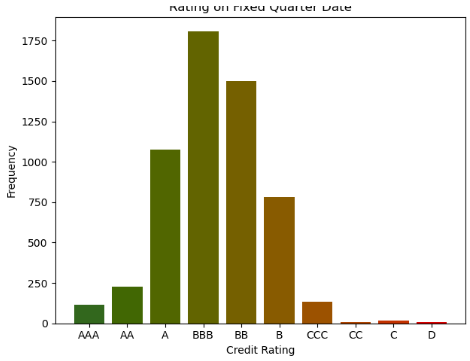

Figure 1: Distribution of Rating Issuances

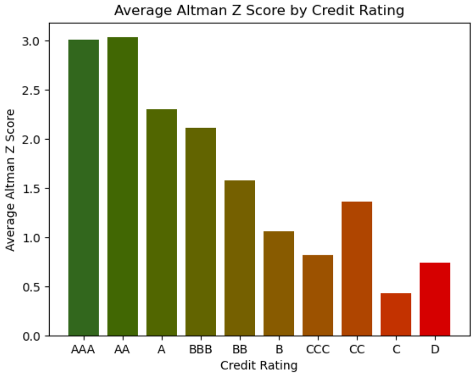

Figure 2: Mean Altman Z by Credit Rating

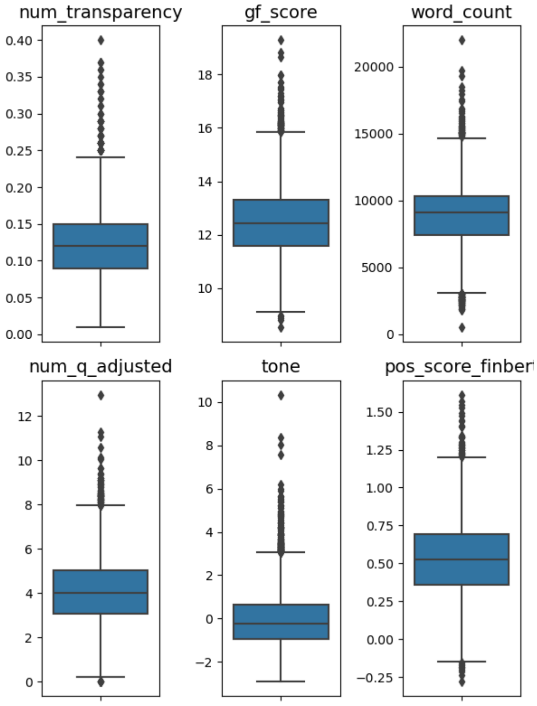

Figure 3: Boxplot for NLP Features

Figure 3 reveals right-skewed distributions for most NLP features, signaling more outliers with higher-than-average values. After inspection, we decide to keep most outlier entries as many of them provide strong indication of the underlying semantic structure of the calls. We also visualize the same histogram but separated by credit ratings in 5. It reveals that poorly rated firms tend to have shorter calls with little to no analyst engagement and high level of numeric transparency.

Additionally, we discover moderately strong negative association between the number of questions and how easy-to-interpret a given call is. The tone of a text also appears to be negatively associated with the number of questions and positively associated with the positivity score. Finally, Figure 4 reveals an interesting relationship: better-rated firms tend to have earning calls that are generally longer in length yet more readable to the audience. Therefore, it is plausible that high-quality firms prefer to convey rich information while keeping the language and delivery as simple as possible. Also, they may have more analyst engagement in the form of concise question and answer sequences, hence shorter sentences and words.

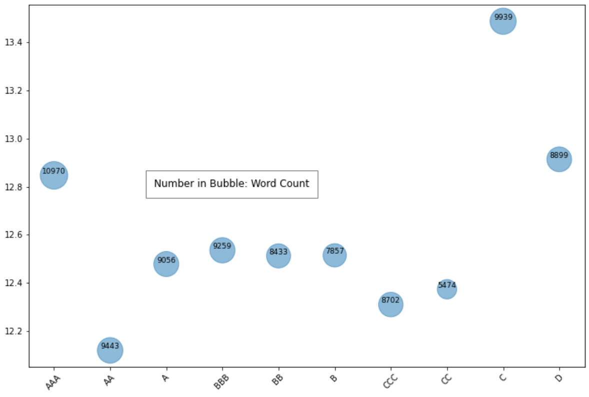

Figure 4: Scatter Plot of Readability by Credit Rating with Word Count Sizes

4 Modeling

4.1 General Model Architecture

The model is defined as:

$$Y = f\left(\sum_i \beta_i X_i + \sum_j \alpha_j Z_j\right)$$

where:

- X_i: Quantitative variables from financial statements and categorical variables like sectors.
- Z_j: NLP features extracted from earning calls transcripts.

The response variable Y can represent:

1. Actual ratings (AAA, AA, A, BBB, BB, B, CCC, CC, C, D)
2. Changes from quarter to quarter (Upgrade, No change, Downgrade)

The function f is the prediction model, such as Logistic regression and XGBoost.

4.2 Model Selection

We apply stratified sampling to ensure our train, validation and test sets contain enough samples from the minority classes. The prediction task is first defined as a 10-class classification problem where each class is the firm's credit rating. Then we attempt to model quarterly change in rating, predicting Upgrade, No change, or Downgrade. Here we show results of two sets of models: Multinomial Logistic Regression and XGBoost. We apply MLR as it is a popular model in credit modeling, and XGBoost is the best-performing model after running Autogluon [7].

4.3 Results - Rating Classification

For each set of covariates we iteratively run our models with grid search to tune our hyperparameters. We report model accuracy with distinct set of features in Table 1 and Table 2. It seems most of the models' predictive power comes from previous quarter's rating. This is due to the fact that credit ratings are

Table 1: Logistic Regression Performance with/without Including Previous Credit Rating

Model/Baseline	**Accuracy**	
	Include Previous	**Exclude Previous**
Altman's Z	0.7442	0.1923
Financial Variables and Sector	0.9508	0.6225
Financial Variables, Sector, and NLP Features	0.9508	0.6333
Majority Baseline	0.3247	0.3247

Table 2: XGBoost Performance with/without Including Previous Credit Rating

Model/Baseline	**Accuracy**	
	Include Previous	**Exclude Previous**
Altman's Z	0.9517	0.3855
Financial Variables and Sector	0.9535	0.7630
Financial Variables, Sector, and NLP Features	0.9535	0.9034

not modified often. The overall accuracy improvements from adding NLP features appear to be marginal. However, if we exclude last-quarter credit ratings, using NLP features in XGBoost provides an impressive 90% accuracy. This suggests that features from earning calls could still play a crucial role in enhancing predictive capabilities, particularly when historical rating data is absent. In general, our most complex models with all available features outperform the others.

4.4 Results - Change in Rating

To address the issue of data imbalance caused by the lack of rating changes, we apply the Synthetic Minority Over-sampling Technique (SMOTE) during training [4]. SMOTE mitigates the imbalance by generating synthetic examples of the minority class using KNN [5], thereby enhancing the model's ability to learn from underrepresented classes. Furthermore, we incorporate differences in financial ratios into the models to capture essential variations and dynamics in financial performance over time.

Under the new prediction task, all XGBoost configurations still offer better predictive accuracy than Logistic Regression. However, XGBoost still favors "No change" in rating even after SMOTE.

4.5 Feature Importance

We measure the contribution of each feature to fitted models' performance by randomly permuting out feature values. As evidenced in 4.3, we still see credit ratings from previous quarter dominate feature importance if included in models. Thus, we mainly explore importance of features in Logistic Regression models excluding previous ratings.

Table 3 makes clear that certain financial and NLP features significantly influence model accuracy. The Debt Ratio feature shows the largest drop in mean accuracy when permuted out, indicating its critical role in predicting credit quality. Notably, among the NLP features, Passive Tone ranks as the third most impactful, and Overstated Tone ranks thirteenth, underscoring the relevance of speaker attitude and word choices. Lastly, Word Count, another NLP feature, also appears significant but slightly less so, ranking sixteenth.

As Logistic Regression provides greater model interpretability, we analyze the coefficients of significant NLP features for different classes. The analysis reveals that companies with higher ratings tend to exhibit a more passive tone in their calls, use less exaggerations or overstatements, and generally have a greater word count. The evidence points to a more detailed and measured strategy of communication during earnings calls. Therefore, investors and analysts may benefit from paying greater

Table 3: Mean Accuracy Drop for Most Important Permuted Features in MLR

Dropped Feature	Accuracy Drop
Debt Ratio	0.06
Ratio E	0.05
Passive Tone	0.05
Sector: Utilities	0.05
Ratio C	0.04
Ratio D	0.03
Depr. and Amort.	0.03
Total Debt	0.02
Gross Profit	0.02
Retained Earnings	0.02
Market Capitalization	0.02
Deferred Tax Liab.	0.02
Overstated Tone	0.02
Gross Profit Ratio	0.01
Stockholders' Equity	0.01

attention to the linguistic style and length of earnings calls when evaluating corporate credit risks.

5 Conclusion

Comparing our models, XGBoost consistently outperforms Logistic Regression, demonstrating its ability to handle complex, non-linear relationships and interactions between financial and NLP variables. Without previous ratings, XGBoost achieves an accuracy of 90% in 10-class rating prediction. Das et al. achieves similar accuracy levels but with a simplified binary prediction problem - only distinguishing between investment-grade and non-investment grade [6]. While oversampling techniques like SMOTE decrease overall accuracy, they enhance the model's sensitivity towards predicting the minority class, underlining the trade-offs involved in achieving a more balanced model performance.

Overall, the inclusion of NLP features across different models and specifications results in a slight increase in accuracy. However, in XGBoost with previous ratings excluded, the impact of NLP features is significant. Specifically, the pattern observed in EDA where word count decreases with lower ratings is confirmed by model results. However, the increase in numerical transparency associated with lower ratings is not captured by the models, likely obscured by the effects of SMOTE oversampling. Interestingly, despite the absence of clear patterns in Figure 5, the tone of earnings calls shows a strong contribution to the predictive power in both Logistic Regression and XGBoost models.

5.1 Gaps and Limitations

Our approach has many limitations that need to be addressed. For one, the class imbalance issue presents an insufficient representation of the low-rating firms, decreasing our model accuracy and robustness. By using a set of manually-determined dates rather than the original dates, we also introduce noise to our financial metrics data which is often time-sensitive. We frequently utilize libraries like *nltk* in Python when conducting tokenization, lemmatization and named-entity recognition on the earning call texts. This makes the accuracy of our NLP features highly dependent on the quality of these intermediate tasks. Lastly, we do not group calls by firm size in our analysis. Yet, firm size should impact the validity of NLP values, as larger firms may dedicate more time and resources to their earning calls compared to boutique firms.

5.2 Roadmap for Future Work

Additionally, we could segment earnings calls by topic, which could be integrated to our model with new features linking discussed topics to credit ratings. Moreover, we are exploring the potential of developing an end-to-end transformer-based classifier, specifically utilizing a transformer encoder model like the Longformer [3], to directly predict credit ratings from earnings calls.

Looking forward, we're also inspired by recent research to construct a network graph representing inter-company relationships derived from earnings call data. We intend to explore and map out firm-to-firm connections based on company mentions, which could provide deeper insights into the influence and relational dynamics between entities. We experiment with our data and provide a visualization of the graph here. We can use the graph to construct

Graph Neural Networks, which show promising performance in the work by Das et al [6].

References

[1] Agewerc. Corporate credit rating. https://www.kaggle.com/datasets/agewerc/corporate-credit-rating/data, 2017.

[2] Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):189–209, 1968.

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] Thomas M Cover and Peter E Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[6] Sanjiv Das, Xin Huang, Soji Adeshina, Patrick Yang, and Leonardo Bachega. Credit risk modeling with graph machine learning. *INFORMS Journal on Data Science*, 2(2):197–217, 2023.

[7] Nick Erickson et al. AutoGluon: Automl for text, image, and tabular data. https://github.com/awslabs/autogluon, 2020. Accessed: 2024-05-12.

[8] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[9] Allen H. Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 2022.

[10] Christopher Kantos, Dan Joldzic, Gautam Mitra, and Kieu Thi Hoang. Comparative analysis of nlp approaches for earnings calls. Available at SSRN: https://ssrn.com/abstract=4210529 or http://dx.doi.org/10.2139/ssrn.4210529, 9 2022.

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

[12] S&P Global Market Intelligence. Analyzing Sentiment in Quarterly Earnings Calls - Q2 2022. https://www.spglobal.com/marketintelligence/en/news-insights/podcasts/masters-of-risk-episode-11, 2021. [Online; accessed 11-May-2024].

[13] S&P Global Market Intelligence. Hanging on Every Word: Natural Language Processing Unlocks New Frontier in Corporate Earnings Sentiment Analysis. https://www.spglobal.com/marketintelligence/en/news-insights/blog/hanging-on-every-word-natural-language-processi 2024. Accessed: 2024-05-11.

[14] Philip Stone. Home category – harvard inquirer, 2021.

A Appendix

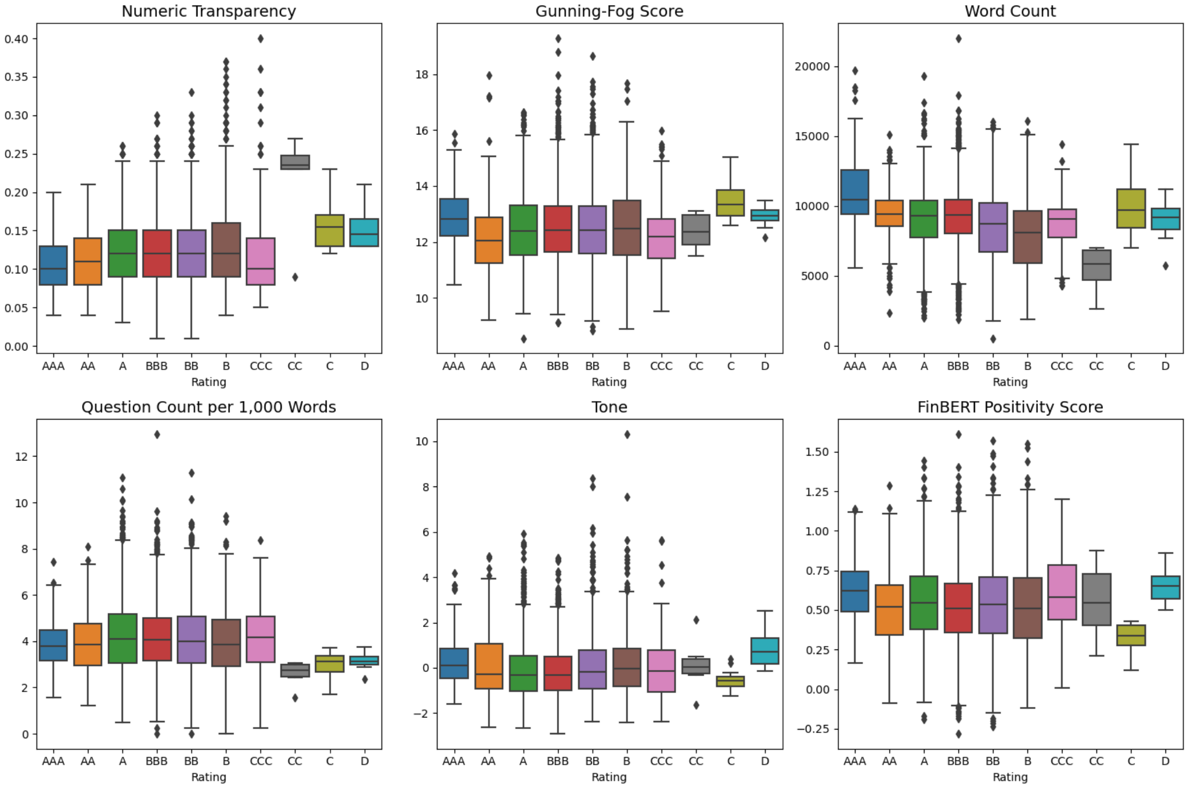

Figure 5: Histogram of NLP Features by Rating