Group7_Final_Report

Kelly Tong, Rakeen Rouf, Lisa Wang, Javier Cervantes

Abstract

Introduction

In the complex landscape of finance, the management and assessment of credit risk play pivotal roles in shaping investment strategies and influencing market dynamics. The finance problem at the heart of this research revolves around the need for a deeper understanding of credit risk within the context of bonds issued by companies listed in the S&P 500. As investors seek avenues to optimize their portfolios and financial institutions strive for effective risk management, the questions of what factors contribute to a bond's credit rating and how credit spreads can be predicted become paramount. In an ever-evolving financial environment, where market conditions and sentiment can swiftly impact investment outcomes, addressing these questions is not merely an academic pursuit but a practical necessity. The outcomes of this analysis hold the potential to refine credit risk assessment methodologies, offering tangible benefits for investors, financial analysts, and the broader financial ecosystem.

The dataset we used to analyze the research problem is a subset of the holdings within an ETF, exclusively representing companies listed in the S&P 500. It comprises **2,341** rows, with each row corresponding to a specific bond issued by an S&P 500 company. Across the dataset, there are **34** variables, which can be grouped into four distinct categories:

1. **Bond information from iShare:** Information related to the bonds, including the issuer's name, industry sector, price, duration, yield to maturity, issuer's stock ticker, and market capitalization. Sourced from the USIG Ishares Credit Bond ETF[^1]

2. **Company fundamentals from Yahoo Finance:** Company fundamentals, including various financial ratios (e.g., revenue, debt). Sourced from Yahoo Finance[^2] using the yfinance package[^3].

3. **Credit ratings from Bloomberg:** Credit ratings from Fitch, Moody's, and S&P, and a composite credit rating. Sourced from the Bloomberg Terminal[^4]

4. **Social sentiment indicators from Finhubb API:** Social sentiment indicators including the number of positive and negative mentions on Reddit last year. Sourced from the Finnhub API[^5]

Methodology

Data Cleaning

...

Exploratory Data Analysis

...

Research Question 1

......

Technique

.....

Assumptions

.....

Assessments

.....

Research Question 2

This section will discuss the methodology for answer our research question 1. Recall that research question 2 is: "What factors influence a company's credit rating?"

Technique

Given the ordinal nature of the outcome variable (Bloomberg Composite Credit Ratings), Ordinal Logistic Regression was chosen to answer the research question. Ordinal logistic regression is a statistical model used to analyze ordinal data, which is categorical data with an ordered ranking. Unlike nominal logistic regression, which is used for un-ordered categorical data, ordinal logistic regression takes into account the ordering of the categories.

Assumptions

The assumptions for Ordinal Logistic Regression is very similar the Linear Regression model from research question 1, with the added assumption of Proportional Odds. The Proportional Odds assumption demands that the relationship between the predictors and the odds of being in a higher category is constant across all levels of the dependent variable.

Assessments
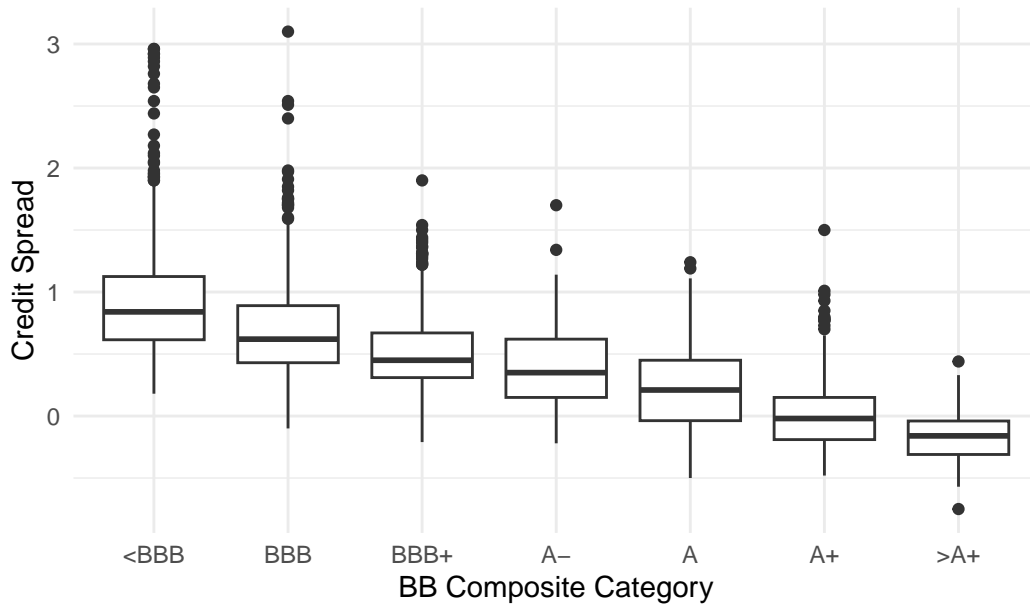
To assess the validity of the final model, ...

Model Results and Discussion

Data Cleaning

Exploratory Data Analysis

Credit Spread vs Credit Rating (BB Composite)

Figure 1: Box Plot of Credit Rating vs Credit Spread

Research Question 1

This section will discuss the results from linear regression model and discuss them based on research question 1. Recall that research question 1 is: "Can we predict a certain bond's credit spread based on a company's fundamentals and the market's sentiment related to that company?" and have credit spread as the outcome variable.

Table A includes all the statistically significant predictor results from the linear regression model output. Their statistical significance is demonstrated by a small p-value (smaller than 0.001). Understanding the financial concepts associated with the predictors and their correlation with the outcome variable credit spread (from data exploration section) support interpreting these output thoroughly.

Duration, which measures the sensitivity of bond price to interest rates variation, is often used to reflect the bond's interest rate risk. The model result demonstrates that for every unit increase in duration, credit spread increases by 0.0285684 units. This positive relationship can be understood through interest rate risk and default time risk. Longer-duration bonds are more sensitive to changes in interest rates. When interest rates rise, the prices of these bonds fall more sharply compared to short-duration bonds. Additionally, longer the duration of a bond, the longer the period over which the issuer must maintain its financial health to avoid default. These all lead to increase uncertainty, risk and risk premium, which are reflected by increased credit spread.

4

Market capitalization is defined as the total market value of a company's outstanding shares of stock. This could indicate the size of a company. Market capitalization is transformed with a log trasformation for the selected variable model, hence the estimated coefficient has to be interpretted with an exponentiation. Here, everyone one unit of log of market capitalization will lead to a decrease of 0.1175721 of credit spread. As suggested by the model output, market capitalization is negatively correlated with credit spread. This might be due to the fact that larger company tend to have higher stock liquidity and stability, which contributes to lower investor risk and thus lower credit spread.

BB composite, which measures credit rating, have all positive estimated coefficients. However, it actually holds a negative correlation with credit spread as we have set the default reference level to "larger than A+ (>A+)." The positive coefficients resulted from the fact that all the displayed ratings are lower or equal to than A+. This inverse relationship is also demonstrated by **Figure 1** in data exploration, as lower rating hold higher median credit spread. Financially, this also matches with our expectation, as credit spread tends to widen when credit rating drops due to increase potential risk and less liquidity with lower rated companies.

Editda margin measures a company's operating profitability as a percentage of its revenue. The estimated coefficient suggests that as Editda margin increases by one unit, credit spread will decrease by 0.2333926 units. This inverse correlation is supported by its financial implications. A higher EBITDA margin indicates that a company is generating substantial earnings from its operations relative to its revenue, suggesting better financial health and efficiency. Thus companies with higher EBITDA margins are generally seen more capable of covering their interest expenses and other financial obligations. It can also increase investor confidence, which leads to lower yields demanded by investors. These all translate into lower credit spreads.

Table A: Credit Spread Selected Variable Linear Regression Model Output (Partial)

Variables	Estimated Coefficient	Standard Error	t-Value	p-Value
Duration	0.0285684	0.0013259	21.547	< 2e-16
Market Capitalization (new)	-0.1175721	0.0097427	-12.068	< 2e-16
BB COMPOS-ITEBBB+	0.3068471	0.0383963	7.992	2.28e-15
BB COMPOS-ITEBBB	0.4835675	0.0377419	12.812	< 2e-16
BB COMPOS-ITE<BBB	0.8059875	0.0434127	18.566	< 2e-16
Ebitda margin	-0.2333926	0.0493347	-4.731	2.40e-06

Model Assessment

From **Table B**, we find that the root mean squared error (RMSE) from selected variable model is smaller than that of the priori model (0.2676 < 0.2818). This suggests that the selected variable model performs better. This might be due to the fact that interaction terms are added for the selected variable model. Moreover, the RMSE values from cross validation are similar to RMSE from model output, which suggests that overfitting is avoided pretty well.

R-squared from model output is 0.6965, which claims that 69.65% of the variation in dependent variable can be explained by the model. This is a good enough r-squared in most cases. F-statistics can be used to test the overall significance of the model. While better F-statistics generally suggests higher significance, it needs to be interpreted in conjunction with the p-values. Though the second model has a lower F-statistics than the priori model (118.7 < 125.4), it has a higher model complexity as it included interaction terms.

Table B: Cross Validation and Model Matrices

	Priori Model	Model (selected)	Cross Validation(Priori)	Cross Validation (selected)
RMSE	0.2818	0.2676	0.2863259	0.269756
R-squared	0.6725	0.6965	0.6610463	0.6860763
MAE	NA	NA	0.2033042	0.1938545
F-statistics	125.4	118.7	NA	NA

VIF is also processed for testing collinearity. All modified GVIF values (shown in **Table C**) for selected predictors are less than 2. This shows that no serious collinearity exists for the model.

```
GVIFs computed for predictors
```

Table C: VIF Output and GVIF Values

Variables	GVIF	Degree of freedom	$\text{GVIF}^{\wedge}(1/(2*\text{Df}))$
Sector	2.182726e+01	26	1.061085
Duration	1.055998e+00	1	1.027618
Market Capitalization	2.762059e+00	1	1.661944
BB COMPOSITE	1.383921e+01	6	1.244780
roa	2.527011e+00	1	1.589658
ebitda margin	2.544020e+00	1	1.594998
debt to assets	3.081602e+07	17	1.660574
operating profit margin	1.623157e+00	1	1.274032
score	2.580289e+09	17	1.891532

To verify that the assumptions for linear regression are met by the model, diagnostic plots on residuals are plotted (**Figure 2**). The plot on residuals and fitted values confirms that the linearity assumption is met, as the residuals are randomly distributed around the horizontal axis (at zero). The Q-Q residuals plot verifies the normality of residual assumption as the points fall approximately in a straight line. The scale location plot have residuals spread evenly across all levels of fitted value, which confirms that the homoscedasticity assumption is hold. The last plot on residual and leverage shows that there are no influential points as there are no points outside of the cook's distance boundaries.

Figure 2: Diagnostic Plots For Model Output
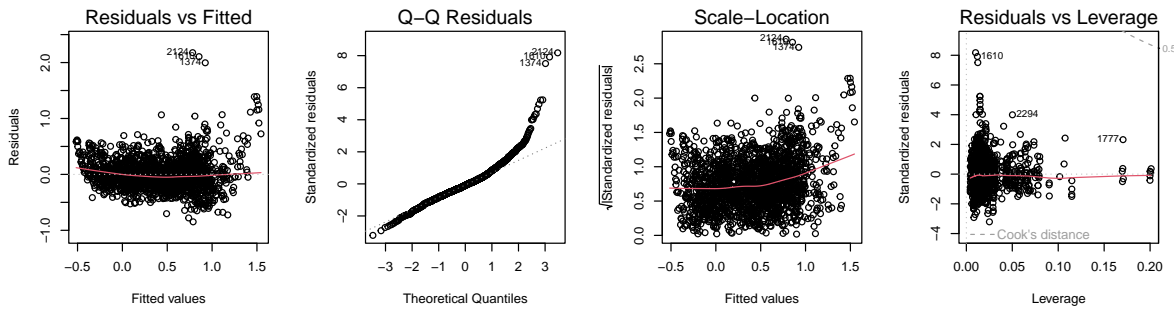
Research Question 2

In this section, we will delve into the outcomes derived from our research question 2. Recall that research question 2 is: "What factors influence a company's credit rating?" The goal of this analysis is to find out which factors contribute to the credit rating of a company. This analysis is based on a subset of the data described in the data overview section. The number of observations are relatively evenly separated into7 ordered categories of credit ratings. Therefore, a Ordinal Logistic Regression model is used to answer the research question.

Modeling

Full Model

A Priori variable selection was done to make the initial model. The A priori variables were selected based on our EDA and domain knowledge. One variable "Market Capitalization" underwent a log transformation. In financial contexts, it is common for the impact of large values to be disproportionately influential. For a metric like Market Capitalization, where a few companies might have extremely large values, taking the log helps to mitigate this effect, making the relationship between Market Capitalization and credit rating more interpretable. Below you can see a representation of our initial model.

7

Reduced Model

With the goal of making the final model easier to interpret and understand, less susceptible to over-fitting, and increased generality , Likelihood Ratio Tests were performed to prune statistically insignificant predictors.

Table 3) Result of Likelihood ratio test on Debt to Equity Ratio

Log Liklihood	Df	Chi-Square	Pr(>Chi-Square)
-2300.8	-1	0.3017	0.5828

The table above shows the results of the only significant Likilihood Ratio test. The null hypothesis (H0) for this test is that removing `debt to equity ratio` does not significantly worsen the model fit. The chi-squared statistic is noted to be 0.3017 with 1 degree of freedom, resulting in a p-value of 0.5828. The p-value is noted to be greater than the typical significance level of 0.05, suggesting that there is not enough evidence to reject the null hypothesis. Therefore, based on the likelihood ratio test, removing the `debt to equity rtatio` variable does not significantly impact the model fit, and Model 2 may be preferred due to its simplicity. Therefore the `debt to equity ratio` was dropped from the model.

Note on Interaction term

Despite having identified Sector as a potential interaction term in our analysis plan, the interaction term was not included in the final model. The chosen interaction variable was introducing perfect separability in the model. The "Sector" variable was therefore dropped to address the numerical instability and convergence problems that can arise when estimating the model parameters. By removing the variable, we can avoid the situation where certain levels of the "Sector" variable perfectly predict specific credit score categories, making the model more stable and avoiding potential issues like infinite coefficients or standard errors.

Model Assessment

To evaluate the validity of the model, our team conducted a thorough examination, mirroring the assumption verification steps employed for research question 1. This scrutiny revealed that none of the conditions were violated. Below you can see the results to assess the final model's multicollinearity, proportional odds assumption and overall performance.

Table 4) Variance Inflation Factors (VIF) for the final model

Predictor	VIF
Sentiment Score	1.15
Duration	1.26

Predictor	VIF
Credit Spread	1.31
Debt to Assets	1.09
Operating Profit Margin	1.23
EBIDTA Margin	1.38
Return on Assets (ROA)	1.55
Log of Market Capitalization	1.19

In general, VIF values around 1 indicate low multicollinearity, and values above 5 or 10 are often considered concerning. The results from the table above suggest that the predictors in our final model are not highly correlated with each other.

Table 6: Predicted Probabilities for Each Category (Multinomial Regression)

<BBB	BBB	BBB+	A-	A	A+	>A+
0.224	0.605	0.140	0.028	0.002	0	0
0.244	0.557	0.157	0.039	0.004	0	0
0.144	0.367	0.272	0.163	0.054	0	0
0.299	0.587	0.098	0.015	0.001	0	0
0.408	0.413	0.130	0.043	0.006	0	0
0.119	0.332	0.285	0.191	0.073	0	0
0.194	0.487	0.220	0.083	0.016	0	0
0.154	0.407	0.263	0.137	0.038	0	0
0.168	0.424	0.253	0.123	0.032	0	0
0.223	0.398	0.231	0.118	0.031	0	0

Table 7: Predicted Probabilities for Each Category (Ordinal Regression)

<BBB	BBB	BBB+	A-	A	A+	>A+
0.080	0.651	0.185	0.058	0.021	0.004	0.001
0.063	0.613	0.218	0.073	0.027	0.005	0.002
0.017	0.333	0.333	0.198	0.094	0.019	0.007
0.121	0.690	0.134	0.038	0.013	0.002	0.001
0.054	0.587	0.236	0.083	0.031	0.006	0.002
0.014	0.292	0.333	0.220	0.111	0.022	0.008
0.033	0.484	0.294	0.125	0.051	0.009	0.003
0.020	0.374	0.329	0.176	0.079	0.015	0.006
0.023	0.399	0.323	0.164	0.072	0.014	0.005
0.024	0.409	0.321	0.159	0.069	0.013	0.005

To check the proportional odds assumption the group fit a multinational logistic regression model with the same variables as the final ordinal regression model. The two tables above show predictions from each of the models, on 10 randomly selected observations from out data set. The predicted probabilities for all the categories were similar enough to conclude that the proportional odds assumption was not violated.
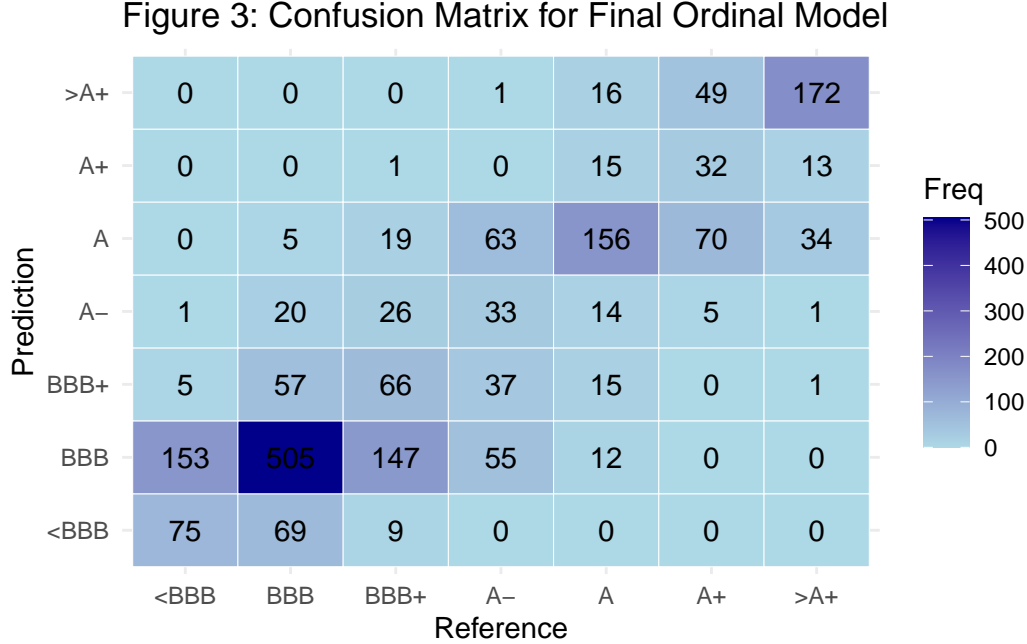
Figure 3: Confusion Matrix for Final Ordinal Model

Prediction \ Reference	<BBB	BBB	BBB+	A−	A	A+	>A+
>A+	0	0	0	1	16	49	172
A+	0	0	1	0	15	32	13
A	0	5	19	63	156	70	34
A−	1	20	26	33	14	5	1
BBB+	5	57	66	37	15	0	1
BBB	153	505	147	55	12	0	0
<BBB	75	69	9	0	0	0	0

Freq: 500, 400, 300, 200, 100, 0

Table 8: Overall Performance Statistics from the final model

Metric	Score
Accuracy	53.23%
95% Confidence Interval	50.98% to 55.46%
Kappa	0.4047
No Information Rate	33.61 %

To check the overall performance of the final model, a confusion matrix shown in the figure above was produced. The overall statistics from the confusion matrx is showed in the table table above. The overall accuracy indicates that the model correctly predicted the credit ratings for approximately 53.23% of the observations. The Kappa statistic measures the agreement between the predicted and actual ratings, considering the agreement occurring by chance. A Kappa value of 0.4047 suggests a moderate level of agreement beyond chance. The No Information Rate represents the accuracy achieved by always predicting the most frequent class. The model outperforms this baseline accuracy significantly. In summary, the model demonstrates

a moderate level of accuracy and agreement, outperforming a baseline approach that predicts the most frequent class.

Model Results

Table 9: Final Model Results

Predictor Variables	Odds Ratio	2.5%	97.5%	P-Value
Sentiment Score	0.999	0.995	1.003	0.598
Duration	1.132	1.101	1.155	<0.01
Credit Spread	0.012	0.008	0.017	<0.01
Debt to Assets Ratio	0.078	0.042	0.142	<0.01
Operating Profit Margin	0.303	0.189	0.483	<0.01
EBIDTA Margin	0.779	0.460	1.318	0.353
Return on Assets (ROA)	1.131	1.087	1.178	<0.01
Log of Market Capitalization	2.960	2.635	3.330	<0.01

The table above answers our research question and shows us the statistical relationship between each of our predictor variables and credit rating of a company.

Some of the most compelling relationships found are as follows.

The model shows a high degree of correlation between Credit Spread and Credit Rating. Intuitively, the higher the credit spread for a certain bond, the more likely that the bond has a lower credit rating. As we can see from the table, keeping all else constant, increasing credit spread by 1 unit drops the odds of moving to the next higher rating by 98.8%. This confirms our initial intuition.

The model shows a high degree of correlation between log of Market Capitalization and Credit Rating. Intuitively, larger market capitalization generally implies a larger and more stable company. In line with our intuintion, the final model suggests, keeping all else constant, increasing the log of Market Capitalization by on unit, leads to 196% increase in the odds of moving to the next higher credit rating.

The final model shows a high degree of correlation between Debt to Asset ratio and Credit Rating. As the amount of debt that a company holds relative to the assets it holds increases, we would expect the company's riskiness to increase. Our model suggests that, keeping all else constant, for a 1 unit increase in the debt to assets ratio, the odds of moving to the next higher category decreases by 92.2%.

According to the final model, Sentiment Score and EBIDTA Margin were not statistically significant predictors for credit rating. This was concluded due to their p-values being above the chosen significance level of 0.05.

Investors often rely on credit ratings to assess the creditworthiness of companies before making investment decisions. The insights gained from these results can guide investors in evaluating the risk associated with companies' financial profiles, helping them make more informed and prudent investment choices.

Conclusion

…