Utilizing Economic Indicators to Predict AAA Bonds

Aarav Gambhir, Lucas Gay, Julen Marmol

Professor Eric Gerber

DS3000: Foundations of Data Science

November 29, 2023

Abstract

This model attempts to predict AAA rated bond prices using the economic indicators of, GDP, CPI, unemployment rate, and interest rate. It successfully fits each of the indicators with a polynomial regression of appropriate degree. [LINK] Our multiple polynomial regression model was successful in producing an r^2 of 0.91, meaning 91% of the variance in the dependent bond prices were explained by the economic indicators we chose to predict it with.

The resulting model can be expressed as a function in the following format:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$$

Where:
- y is the predicted AAA Bonds value
- x_1 is the GDP
- x_2 is the Consumer Price Index
- β_0 to β_5 are the model coefficients

Substituting in the coefficients:

$$y = 0 - 0.64815914 x_1 - 4.86084047 x_2 - 0.25775619 x_1^2 + 0.42069296 x_1 x_2 + 0.82769091 x_2^2$$

Introduction

With a higher market capitalization than the equities market, the bond market is the biggest sector in the stock market. Every business day, millions of dollars in bonds are traded mainly by institutional investors with the goal of preserving and acquiring money by exposing their assets to a lower risk than with equities. The most volatile subset of assets in the bond market are high yield corporate bonds. Bond prices often have a higher relationship to underlying economic conditions than equities. for this reason we will be assessing which economic indicators have a higher impact or relationship to high yield corporate bonds. Specifically we will be focusing on AAA-rated corporate bonds, representing the highest credit quality debt instruments issued by corporations. These bonds are considered very low risk and are given the highest credit rating by credit rating agencies like Moody's, Standard & Poor's, and Fitch.

This project explores how different economic indicators such as GDP, CPI, unemployment, and interest rates, play into the price of AAA rated corporate bonds. The goal is first to identify the factors that have the biggest impact on the price of AAA rated corporate bonds using regressions on each of the indicators individually. Then using this information we will incorporate the most influential factors into a multiple polynomial regression to try and create a model to predict the price of AAA corporate bonds.

Our two guiding questions are:

Which economic indicators have the most pronounced linear relationship with AAA rated corporate bond prices?

How can we create a model that predicts the price movements of AAA rated corporate bonds before they occur, utilizing economic indicators for assessment?

Data Description

The bond market is the biggest sector in the stock market in terms of market capitalization. Every business day, millions of dollars in bonds are traded mainly by institutional investors with the goal of preserving and acquiring money by exposing their assets to a lower risk than with equities.

Price movements in the bond market are more tied to Macro Economic Indicators Than Regular equities because Their value is tied to the financial situation of the entity that issues them. For these reasons, in this project we aimed to analyze the effects of macroeconomic indicators on AAA rated corporate bonds.

Data Source:

Our financial data consists of AAA rated corporate bond prices which is the dependent factor we are trying to predict. Our independent data include fundamental economic indicators including CPI, GDP, unemployment rates, and interest rates. Our data is being collected from St. Louis Fed Web Services: FRED® API (stlouisfed.org). We are using a 3rd party api 7astro7/full_fred: Full Python interface to Federal Reserve Economic Data (FRED) (github.com) to import this data.

Elements Analyzed:

- Moody's Seasoned Aaa Corporate Bond Yield (DAAA):
    - As our feature to predict, we choose Moody's Seasoned Aaa Corporate Bond Yield index, because it is a very good representation of the changes in the most traded subset of corporate bonds.
- Consumer Price Index (CPI):
    - CPI measures the average change over time in the prices paid by urban consumers for a basket of consumer goods and services. Inflation, as reflected by CPI, can impact the real returns on fixed-income securities like AAA-rated bonds.
    - If inflation is high, the purchasing power of the bond's future interest and principal payments decreases. This can erode the real value of the bond's cash flows, making it less attractive to investors.
    - Conversely, low inflation or deflation may enhance the attractiveness of AAA-rated bonds, as their fixed interest payments become relatively more valuable in a low inflationary environment.
- Gross Domestic Product (GDP):
    - GDP represents the total value of goods and services produced in a country. It is a key indicator of economic health. Strong GDP growth is generally positive for corporate bonds because it suggests a robust business environment.

    - In an expanding economy, companies are more likely to generate higher revenues and profits, reducing the likelihood of default. This, in turn, enhances the creditworthiness of corporate bond issuers, including those with AAA ratings.
- Unemployment Rates:
    - Unemployment rates reflect the health of the labor market. High unemployment can signal economic distress, potentially leading to increased default risk for corporate bonds.
    - AAA-rated bonds are generally associated with companies that have strong financial positions. However, sustained high unemployment across the broader economy could still pose risks, as it may impact consumer spending and corporate earnings.
- Interest Rates:
    - Changes in interest rates can significantly impact the value of existing bonds. When interest rates rise, the market value of existing bonds tends to fall, and vice versa.
    - AAA-rated corporate bonds, being fixed-income securities, are particularly sensitive to interest rate movements. If rates rise after the issuance of the bond, the existing fixed interest payments become less attractive compared to newly issued bonds with higher coupon rates.

<u>Data Processing Pipeline:</u>

From the FRED API we obtained The previously mentioned data sets. There were a lot of Adjustments needed to process the data. The data processing was done separately and independently for all the models due to several reasons and in the following pipeline:

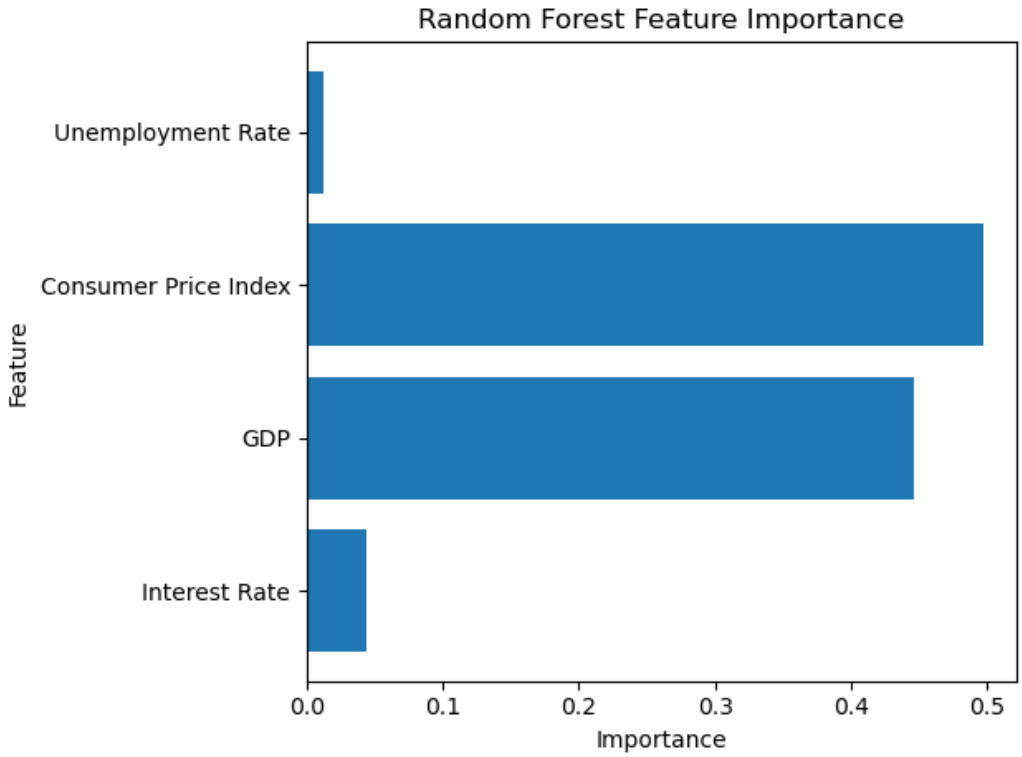
1. Missing values: Some of the Data sets Dated Back to the 1950's. For this reason, there were some data points missing at random dates and expressed as ".", So for each model we had to drop the values at that dates for all the dependent and independent features. When combining the models together in part 2 this adjustment led to the deletion of approximately 25% of the data points.
2. DIfferent Start dates: The data sets had different start dates, So in order to work with the highest amount of data possible, we had to drop all rows for which the other data sets didn't date back to.
3. Data Covered Different Time Frames: The data was obtained in different timeframes. Data for the AAA rated bonds was obtained daily, so in the beginning of the analysis, the first trading day of each month was selected as a representative value for that month. Regarding the economic indicators, CPI, unemployment, and interest rates were given monthly. In contrast, GDP was quarterly, so in the individual models for gdp, the prices for the 3 months of each quarter were averaged, but in the models with multiple features, the value for gdp was replicated for 3 months after the data was released
4. Data Merging: Before fitting each model and after completing the previous 3 steps, the data was merged with an inner join on the months with valid data remaining for all data sets used in each respective model.

Methods

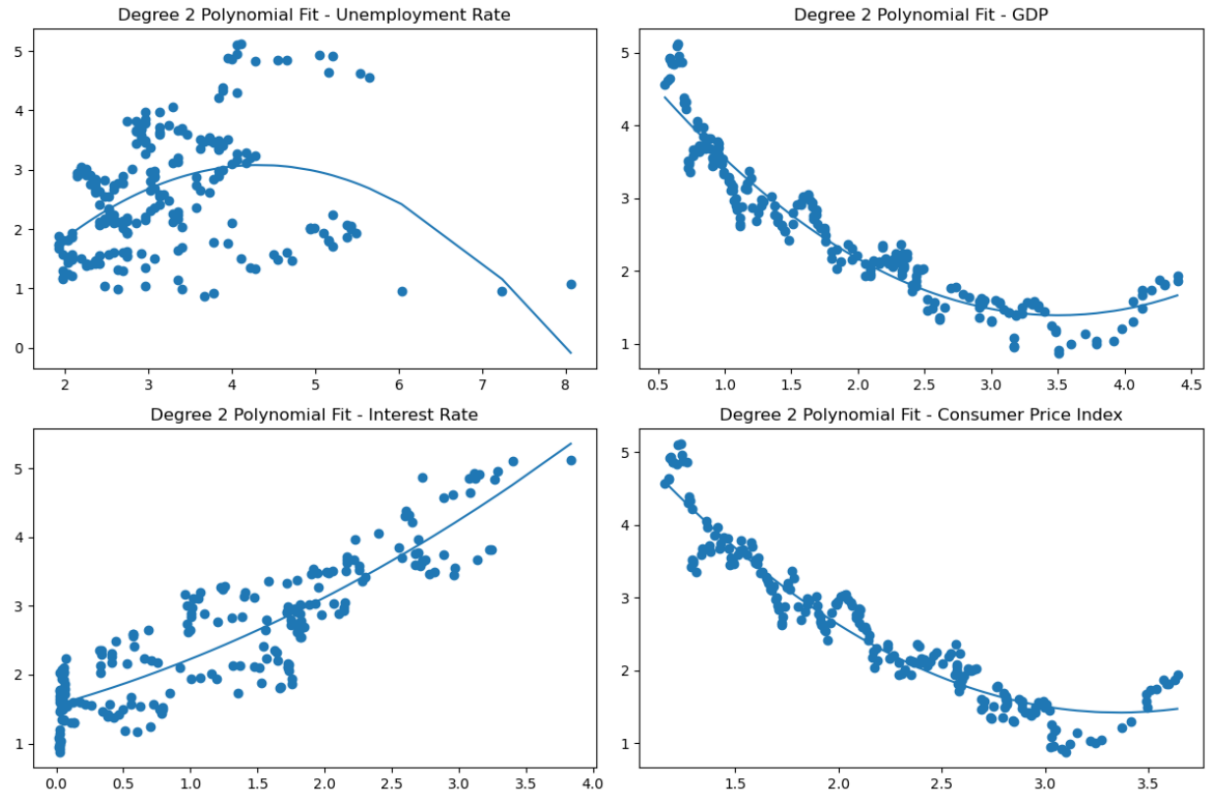
Our investigation will involve applying multiple polynomial regression with factors such as GDP, CPI, interest rate, and unemployment rate to assess whether we can predict the prices of AAA corporate bonds. This is appropriate because in order to create the most accurate model for predicting bond prices we must incorporate multiple of these indicators each of which may have a polynomial degree which best suits their individual regressions. We must first understand the most optimal degree of each of the features in order to create the most accurate model for each of the features without overfitting. Next we must explore which of the features we will use in the final multiple polynomial regression. Prior to model implementation, we will ensure that our numeric features are appropriately scaled, and we will incorporate cross-validation as we explore various models.

In terms of degree, it's possible that certain numeric features will produce non-linear relationships, prompting us to explore polynomial regression for each of the economic indicators. This is important to know because in the multiple polynomial regression we need to know the most predictive degree to use for each of the features. - *Refer to the polynomial fit graph*

Using random forest regression we will determine which of the factors most influence bond prices. This will allow us to identify which of the four factors to include in the final multiple polynomial regression. Initial analysis of the results suggests that CPI might be the most influential factor, closely followed by GDP. Lagging behind significantly is the unemployment rate and the interest rate which we probably will not include in the multiple regression.



Results



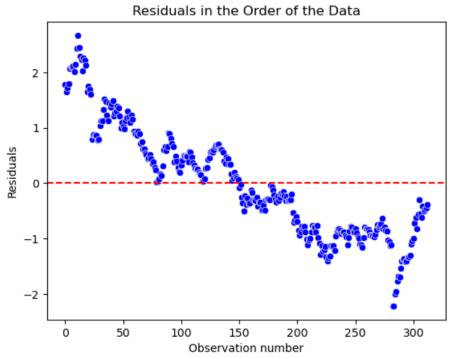
In simple terms, these graphs show the 4 dependent features graph against the dependent feature with a quadratic polynomial model fitted to that data. As we can see, Unemployment rate is not a very good predictor of AAA Rated Bonds, Interest Rate is a decent predictor, but GDP and CPI are better fitted by the model (higher R2 scores), so they can predict AAA rated bonds in an accurate way.
As a result of this analysis we decided to create a multiple polynomial regression model with degree 2 that predicts AAA rated bonds. This model had an R^2 to of 0.91 approximately which means 91% of the variance in bond prices can be explained by changes in GDP and CPI.

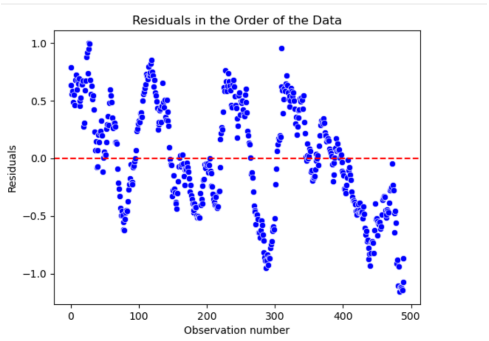
In [58]:
```
#Fit the whole model
# Extract GDP and CPI as features
X = df_final[['GDP', 'Consumer Price Index']]

# AAA Bonds as target
y = df_final['DAAA']

# Polynomial transformation
poly = PolynomialFeatures(degree=2)

X = poly.fit_transform(X)

# Polynomial regression
model = LinearRegression()
model.fit(X, y)

# Evaluation
r2 = model.score(X, y)

n = len(y_test)
p = X_test_poly.shape[1]
r2_adj = 1 - (1 - r2)*(n - 1)/(n - p - 1)
coefficients = model.coef_

y_pred = model.predict(X)

print('Polynomial Regression Results (Degree 2)')
print('Adjusted R-Squared:', r2_adj)
print('model Coefficients', coefficients)
```

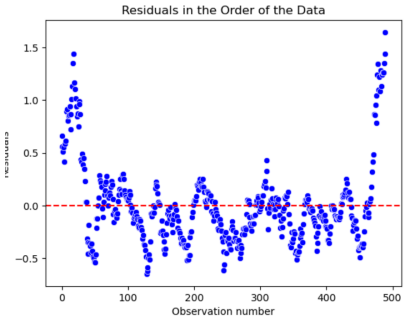
```
Polynomial Regression Results (Degree 2)
Adjusted R-Squared: 0.9113029248479723
model Coefficients [ 0.         -0.64815914 -4.86084047 -0.25775619  0.42069296  0.82769091]
```

Discussion

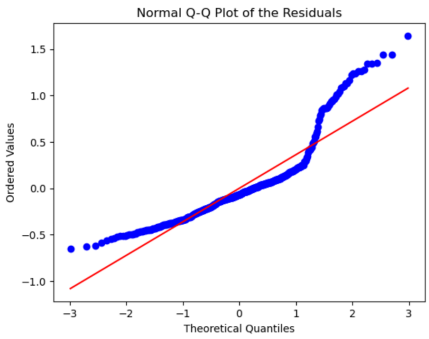
To answer our first question, we performed simple linear regressions between each indicator and AAA triple bonds. We found that CPI had both the highest R^2 value and lowest MSE (0.84 and 0.15). out of the 4 indicators, displaying that it has the most pronounced linear relationship with AAA Bonds. Out of the indicators, unemployment had the worst R^2 value and the highest MSE (-0.004 and 0.98), signifying that it does not have a linear relationship with AAA Bonds. It is critical to understand the values in the context that it has been scaled to match the other values. When considering the validity of the results, we must hesitate in accepting these at face value, due to the three critical assumptions not being passed. To recap, our three assumptions are Independence, Homoscedasticity, and Normality. When we analyze the residual plots, we see that, for all the indicators, the homoscedasticity (constant variance) assumption does not pass; there is a clear trend within the points. For the reference of the order of the images, the images are in the order [GDP, Unemployment, CPI, Interest Rate].

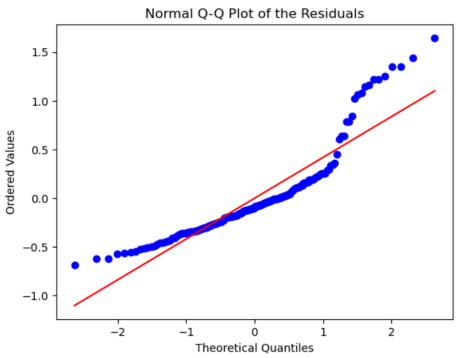


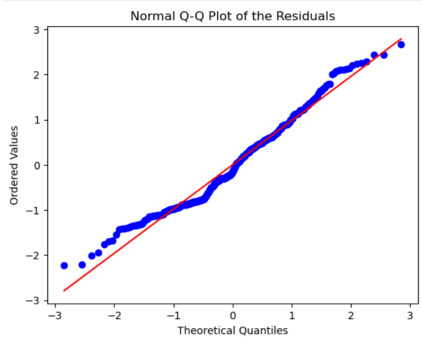


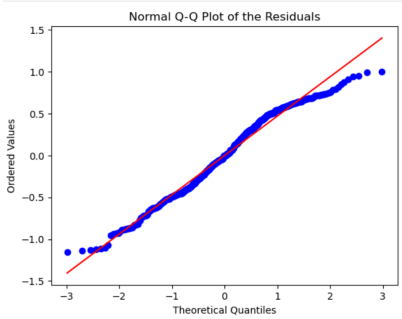


It is perhaps understandable that unemployment had the lowest R^2 value, as the distribution of the data clearly passed the assumption to the least extent. For the independence assumption, all the indicators fulfilled that assumption. Finally, in regards to the normality assumption, we see that not all the indicators uphold the assumption. For the reference of the order of the images, the images are in the order [CPI, GDP, Unemployment, Interest Rate].





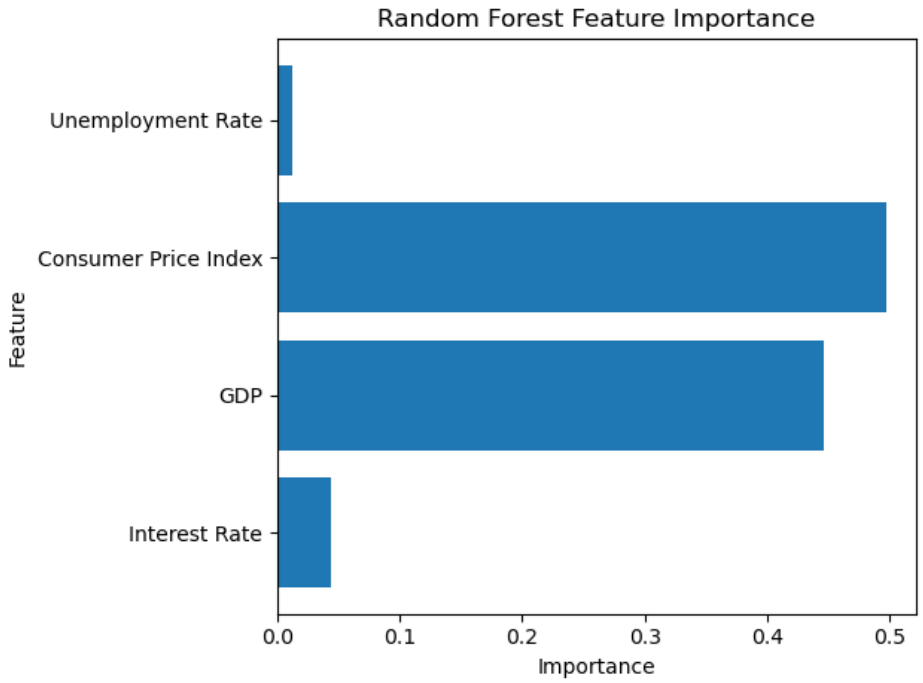




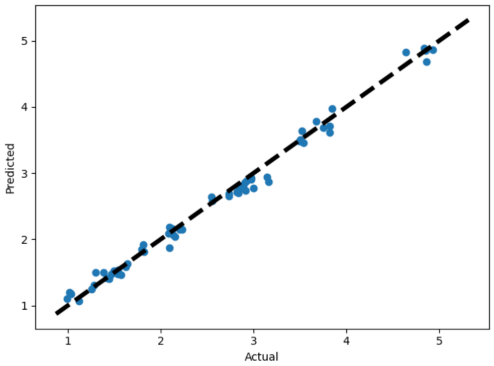
We can see that not all the indicators pass the normality test; this suggests that our data does not follow a normally distributed pattern, which means we need to be aware that outliers and skewness could impact our data in any further analysis. Because all the assumptions didn't pass for every indicator, we cannot be

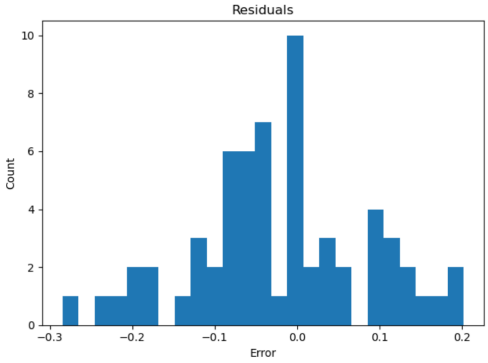
sure that the simple linear model is the most appropriate in predicting AAA Bonds. We can say, however, that CPI held the strongest linear relationship with AAA Bonds.

As we determined that we cannot confirm that the linear model is the most appropriate for our data, we wanted to implement a polynomial regression. However, we noticed that some of the features increased the R^2 by a negligible amount and we wanted to exclude those features. We decided to first implement a Random Tree Regression. Random Tree Regressions are beneficial as the R^2 value within it measures how well the forest model fits the training data. Their R^2 is higher than the cross validated R^2 because the trees are based on all the factors and thus are overfitted. We did not directly utilize the R^2 value; instead, we wanted to use the Feature Importance. Feature Importance displays how much each feature would contribute to the R^2 for the random forest regression. However, feature importance will help indicate which features are causing adding little to the regression, which will make our polynomial regression far more accurate. In our feature importance, we see that the two most prominent indicators are GDP and CPI. This matches our data from the linear regression, as those were the two values that had the highest R^2 values. We thought that interest rates would affect the bonds prices the most, because the prices of bonds are directly derived from the interest rate (Price of bond = Risk Free Interest Rate + Risk Premium), but this was not the case. Interest rates did affect bond prices but they were not the main predictors of movement. This is because, most of the time interest rates change an almost negligible amount, which causes other economic indicators such as CPI or GDP to explain a higher portion of the variance in bond prices.



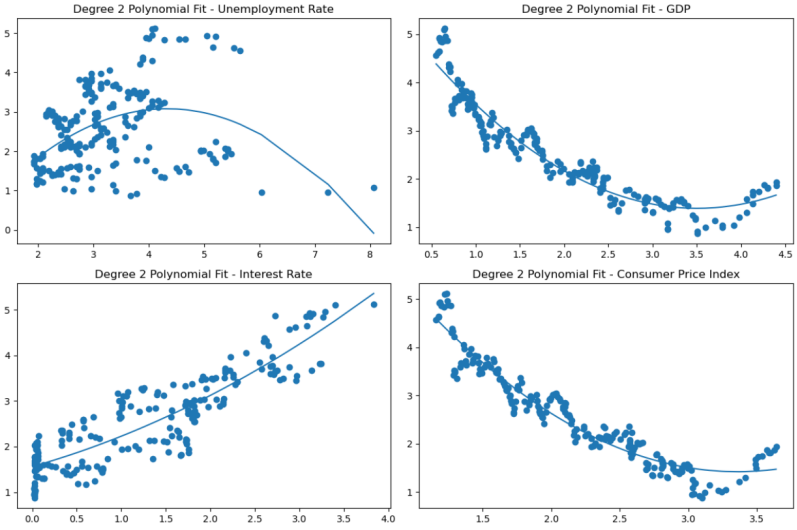
It is important to check whether the values from the random forest regression are reliable or not. Included below is the histogram of the errors and the actual vs predicted.





In both these graphs, we can see the distribution of the errors is normally distributed; this is key as we now know that the model does not tend to overestimate or underestimate the values as often. Additionally, the actual vs linear plot displays the accuracy of the model in its predictions. With these, we have shown that the random forest is an adequate model, and that the most important features it displayed can be utilized for future insights, while removing noise from less useful variables.

In regards to the multiple polynomial regression, one of the first things we had to do was determine a degree for the regression. We did this manually, by setting up a loop and seeing the individual polynomial regressions for each feature. We found that degree 2 worked best for this.



With the features chosen and the degree decided on, we implemented the 5-fold cross-validated polynomial regression which outputted an adjusted R^2 value of 0.9114. We also implemented a regular regression to compare it with the cross-validated, to see if we overfitted the data. The adjusted R^2 value for the non-cross-validated regression was 0.9113. When looking at these values, we can see that both the R^2 values indicate a strong, positive relationship between GDP, CPI, and AAA Bonds. Furthermore, because

both values are near-identical, we can be sure that we have not overfitted the data, which is a large concern with implementing polynomial regression. With these values, we can be confident that the polynomial regression model is an adequate model for predicting AAA Bonds. Below is the polynomial equation we came up with:

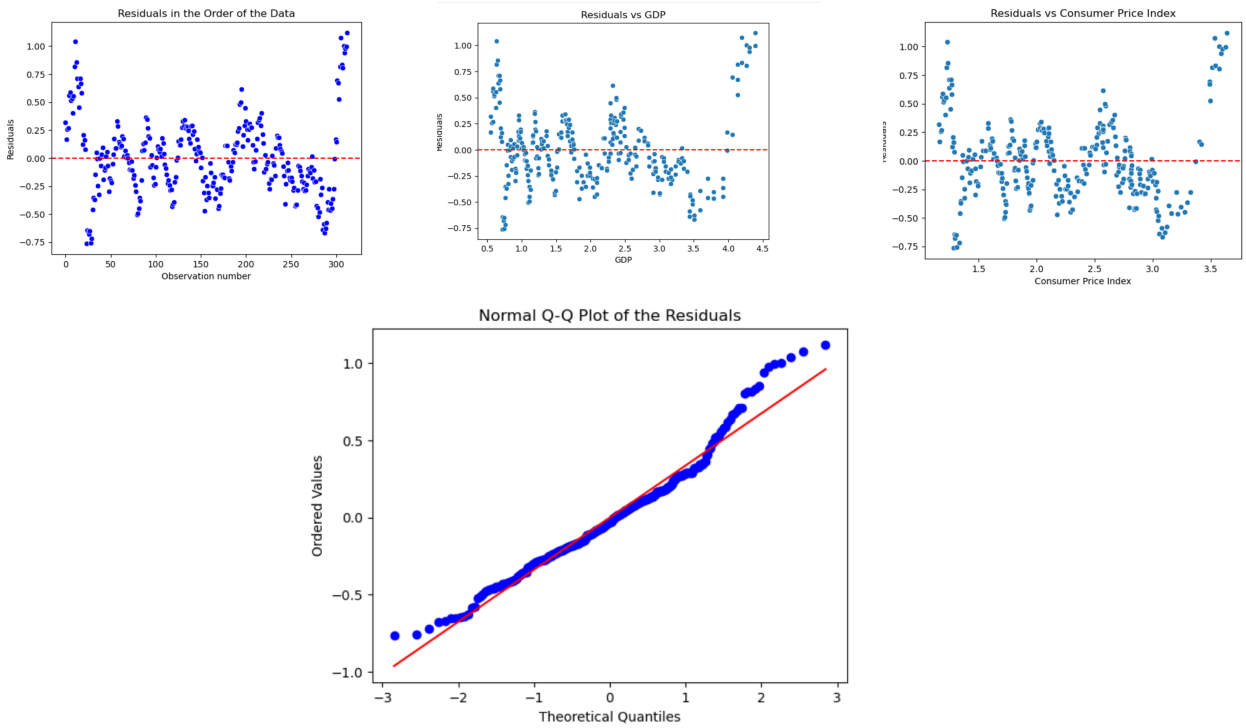
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$$

Where:
- y is the predicted AAA Bonds value
- x_1 is the GDP
- x_2 is the Consumer Price Index
- β_0 to β_5 are the model coefficients

Substituting in the coefficients:

$$y = 0 - 0.64815914x_1 - 4.86084047x_2 - 0.25775619x_1^2 + 0.42069296x_1x_2 + 0.82769091x_2^2$$

Looking at the validity of the polynomial model, we also created the residual plots to follow the same assumptions as mentioned prior. For the assumptions of constant variance, it follows a similar trend as the linear regressions, in the sense that it doesn't pass.



For the independence assumption, we can say that the assumption passes for GDP and CPI both. Where the difference matters most is in the QQ Plot for the normality assumption. It indicates that the distribution of the data is normally distributed, which is a significant change from the linear regression

models. Despite some assumptions still not passing, we can say that the polynomial model is a more adequate model.

When we consider any future improvements with this analysis, one of the major ones that stands out is the data cleaning methodology. For every individual linear regression, we had to reclean and reformat the data so that the two features would be able to align with each other. This created a lot of inconsistencies within the data, but it is understandable due to the varying date ranges for each economic indicator. GDP is only measured quarterly, while the rest were monthly. When we were collating the data for the polynomial regression, we had to be even more strict, as we were combining all the indicators into one dataframe. This could have resulted in inaccurate data for the polynomial regression. Realistically, we couldn't have done much to fix the missing data; the only considerable solution would be to analyze a larger pool of data sources. For GDP, we tried to extrapolate data in between to make it yearly (by making the middle months the same as the measured one). This could have also impacted the validity of the data. We are still confused at how interest rate had such a low importance in the random forest regression. We would need a further exploration into the random forest regression model to be able to determine why this was the case. Despite these, to a large extent, we can say that a polynomial model of GDP and CPI is the best model in predicting AAA Bonds.