Feature Engineering

After consideration, the key variable of interest will be the ‘Difference’ variable in Federal Funds Rates. If the actual policy action and degree can be forecast, this bundles into it the ‘Level’, ‘Increase’, and ‘Decrease’ values all into one.

If this is the purpose, and we see that Short-Term Treasury Bond Rates are almost lockstep with Federal Funds Rates, then it will benefit to include the difference in Short-Term Treasury Bond Rates as well for their predictive power. Thus, this variable was added to the dataset. Rolling Means are also added to the more relevant macroeconomic variables that are shown to have the strongest correlations with the outcome variable.

Data Augmentation

Some further data augmentation was done in previous notebooks so that previous processes can be applied to them as well. For example, the frequency of positive and negative words was included in a previous dataset. While this constitutes one form of word embedding, it can also become an oversimplification. As such, the words themselves are flagged as positive or negative, and the total list of positive and negative words as a string appear in each row extracted from the FOMC Minutes. This will eventually get the dataset ready to count the importance of specific words or perhaps group of words in predicting Federal Reserve monetary policy.

It was also noted that while macroeconomic variables should be strong predictors of monetary policy, they will not necessarily be available on the same day as the FOMC meetings. Further, even if they were, the point of this analysis is to be able to predict future monetary policy and not confirm concurrent policy through their Minutes and macroeconomic variables. Thus, the variables need to be observations from a long-enough frequency before the meeting for the model to have any sort of predictive capability. For this analysis, the delta will be 5 days. This results in observations that are at least 5 days prior to the meetings they predict.

Similarly, as the ‘Difference’ variable is going to be predicted and we would like to use Minutes sentiment scores for predictive purposes, the variables related to sentiment score such as word frequency and the net sentiment scores are “shifted up” in that they concur in row to the previous meeting. Thus, their predictive power for the next meeting will be analyzed through dimensionality reduction and univariate variable selection.

Dimensionality Reduction

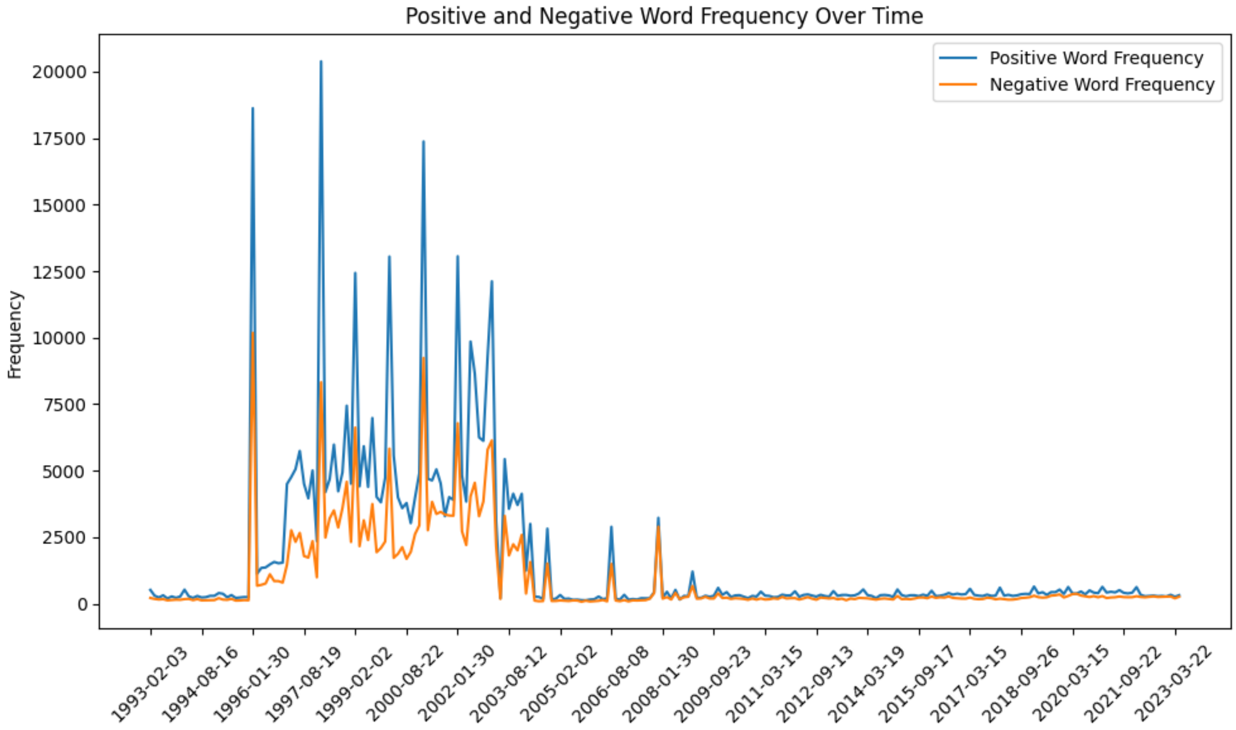
This is a key part of the next step, as the number of variables we have now and will have after feature engineering is truly sizeable, and would lack interpretability should someone attempt to explain any model based on these variables. As such, dimensionality reduction will help identify key variables that have the power to explain the dependent variable. In this case, it will be the ‘Difference’ variable, as being able to estimate this variable means we can estimate the change in the Federal Funds Rate, which captures both increases and decreases. That said, if univariate variable selection removes this variable as well, then it will not have as much predictive power as assumed.

On the model ingested from Exploratory Data Analysis, the selected variables and their relevant feature scores are as follows:

Selected Features:	Feature Scores:
Retail Sales	-0.304
Durable Goods Orders	-0.303
LEI	-0.244
CEI	-0.206
LAG	-0.204
Net Sentiment Score	-0.316
Negative Frequency	-0.322
Word Count	-0.326
Standardized Sentiment Score	-0.316
Short-Term Treasury Diff	3.736
LEI_RollingMean	-0.186
CEI_RollingMean	-0.186
LAG_RollingMean	-0.183
Retail Sales_RollingMean	-0.310
Durable Goods Orders_RollingMean	-0.331

We see that sentiment score was one of the top 15 coefficients for predicting the 'Difference' variable. While the number of variables is set beforehand, we certainly see that even here we have the ability to reduce dimensionality. LEI, CEI, and LAG are strong enough in terms of predictability that their rolling averages. Thus, they can be excluded from our subsetted model. The opposite is true, though, of the rolling means for Retail Sales and Durable goods, both of which have greater predictability that their raw values. Finally, a surprising inclusion is Word Count. Given the negative sign to its relevance score, it looks like longer Minutes result in greater decreases in Federal Funds Rates.

While the Loughran-McDonald Master Dictionary is used in similar research online, there are other sentiment dictionaries associated with economic variables as well. The Joint Research Centre Data Catalogue[1] provides an economic lexicon that attaches a range of sentiment scores to economic vocabulary that is most frequently used. One of the limitations of the Loughran-McDonald Master Dictionary is that it only assigns positivity and negativity in equal scale (1 for positive, -1 for negative). The Economic Lexicon provides varying scores for many words commonly associated with economic discussion. This may prove to be more capable of capturing randomness than a simple binary scoring system. Rerunning the data ingestion process with this new lexicon and the processes prior to this one, we see different values associated to net sentiment score. Looking at the time series plot:



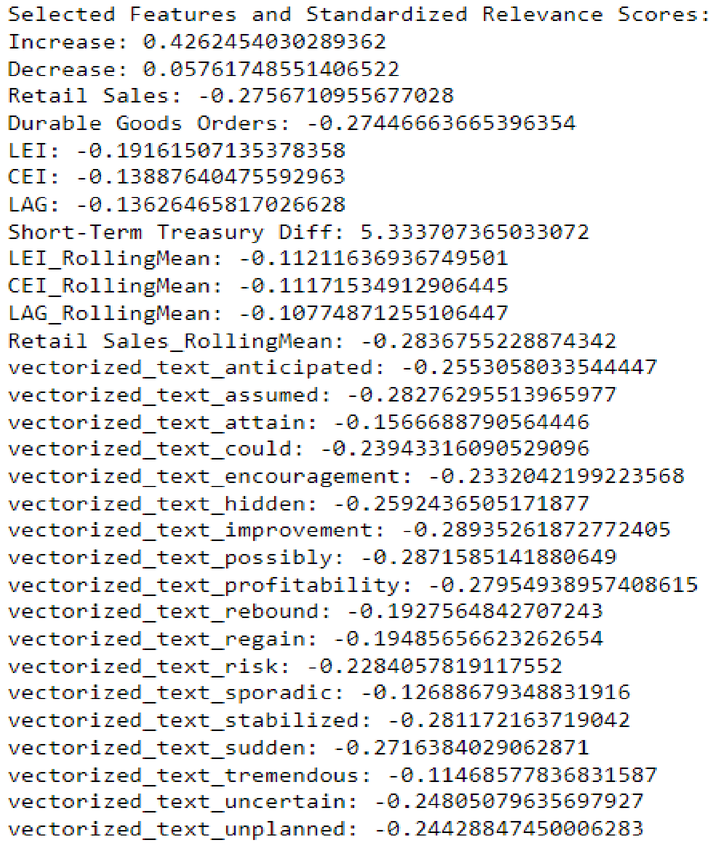
It similarly shows more activity in the 1990's and early 2000's, but the magnitude of the positive word frequency is much higher, and the resulting net sentiment score should reflect consistently positive. Running all the processes in this report, we also get different selected features from univariate variable selection:

Selected Features and Standardized Relevance Scores:	
Increase	0.142
Decrease	-0.127
Retail Sales	-0.370
Durable Goods Orders	-0.370
LEI	-0.310
CEI	-0.269
LAG	-0.268
Total Negative Sentiment	-0.383
Negative Word Frequency	-0.350
Proportion Negative Words	-0.285
Short-Term Treasury Diff	3.711
LEI_RollingMean	-0.250
CEI_RollingMean	-0.250
LAG_RollingMean	-0.247
Retail Sales_RollingMean	-0.375

The ‘Increase’ and ‘Decrease’ variables have been dropped, showing that this model calculates that the most variability in policy action can be explained even without the previous policy action.

Feature Engineering – Word Embedding

One last option for utilizing the Minutes as extensively as possible comes from word vectorization. This take the minutes from a string of characters to vectors with numerical application. Some of this was done with calculating net sentiment score, but it would also be beneficial to calculate the frequency of individual words. First, the CountVectorizer package is used to take the processed text and convert each unique word into its own column vector. It iterates over each row and calculates the number of instances of that word. This gives a numerical representation to the predictive power of a single word. Taking this a step further, applying the sentiment index from Loughran McDonald gives each word’s vector a positive or negative value. Thus, if a negative word appears 28 times in a certain meeting, then the row value is -28. Running the univariate variable selection again with a great number of variables provides this list:



```
Selected Features and Standardized Relevance Scores:
Increase: 0.4262454030289362
Decrease: 0.05761748551406522
Retail Sales: -0.2756710955677028
Durable Goods Orders: -0.27446663665396354
LEI: -0.19161507135378358
CEI: -0.13887640475592963
LAG: -0.13626465817026628
Short-Term Treasury Diff: 5.333707365033072
LEI_RollingMean: -0.11211636936749501
CEI_RollingMean: -0.11171534912906445
LAG_RollingMean: -0.10774871255106447
Retail Sales_RollingMean: -0.2836755228874342
vectorized_text_anticipated: -0.2553058033544447
vectorized_text_assumed: -0.2827629551396597
vectorized_text_attain: -0.1566688790564446
vectorized_text_could: -0.23943316090529096
vectorized_text_encouragement: -0.2332042199223568
vectorized_text_hidden: -0.2592436505171877
vectorized_text_improvement: -0.28935261872772405
vectorized_text_possibly: -0.2871585141880649
vectorized_text_profitability: -0.27954938957408615
vectorized_text_rebound: -0.1927564842707243
vectorized_text_regain: -0.19485656623262654
vectorized_text_risk: -0.2284057819117552
vectorized_text_sporadic: -0.12688679348831916
vectorized_text_stabilized: -0.281172163719042
vectorized_text_sudden: -0.2716384029062871
vectorized_text_tremendous: -0.11468577836831587
vectorized_text_uncertain: -0.24805079635697927
vectorized_text_unplanned: -0.24428847450006283
```

Given that our selected subset has a mixture of macroeconomic variables—not just that, but variables that were expected to have strong predictive power—and statistically significant unique words, we can now move on to model building.

Citations

1: Luca Barbaglia; Luca Tiozzo Pezzoli; Sergio Consoli; Elisa Tosetti; Sebastiano Manzan (2021): Economic Lexicon. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/1c054ef4-561a-464a-9077-3f6b09630da2