“Can Forward Guidance Really Guide Rates Forward?”—FOMC Minutes Analysis Using NLP and ML

Alexander Bactat

Abstract

The goal of this project is to train a natural language processor using Machine Learning techniques to predict the actions of the Federal Reserve following FOMC (Federal Open Market Committee) meetings. The available data for training consists of Meeting Minutes, which can be freely accessed on the FOMC website at the Federal Reserve Website[2]. The degree and direction of each monetary policy action will be recorded alongside vectorized sentiment data, macroeconomic data, and various data augmentations. Using various supervised and unsupervised methods, this analysis intends to leverage FOMC forward guidance to allow market participants to make more informed decisions leading up to FOMC meetings. Work by Cynthia Royal Tori[1] cites that while the FOMC meeting dates account for only 4.42% of the trading days in a year, they make up over 13% of the cumulative returns over the same period. It concludes that the mean market return on the trading days contemporaneous with FOMC meetings was 5.7 times higher than non-FOMC meeting trading days during the span of 1980-2000. It also concludes that not investing on these given days results in an investor missing out on 16% of the cumulative returns between 1980 and 2000. Success in this analysis will most definitely lead to the potential for outsized gains in the market.

Key words: Natural Language Processor, FOMC, Federal Reserve, Machine Learning, Macroeconomic Data

Motivation

Given the substantial influence wielded by the Federal Funds Rate (FFR) on macroeconomic dynamics, its impact is challenging to precisely quantify. The FFR serves as the primary tool for the Federal Reserve to shape monetary policy, enabling control over the money supply by adjusting the interest rates among depository institutions for overnight uncollateralized transactions. This influence significantly resonates within the realm of FFR futures, commonly referred to as 30-day interest rate futures.

The Effective Federal Funds Volume, presently estimated at approximately $132 billion according to the Federal Reserve of St. Louis3, underscores the relevance of this endeavor. The Federal Reserve's ability to impact interest rates on these futures loans cascades through various sectors of the economy. For instance, an FFR increase triggers corresponding rises in credit card annual percentage rates (APRs), potentially leading to loan defaults. Furthermore, such FFR adjustments exert adverse effects on the stock market, heightening the cost of investment and potentially dampening the growth potential and earnings of select companies.

Anticipating the subsequent Federal Funds Target Rate has significant ramifications for a broad spectrum of investors. Institutional banks engaged in reserve trading, companies pursuing long-term investments, and homeowners contemplating second mortgages all stand to benefit from accurate predictions. Moreover, the potential to forecast unexpected rate fluctuations following Federal Open Market Committee (FOMC) meetings could yield substantial gains, boasting an alpha of 8.04% per FOMC meeting.

Literature Review

As mentioned, Tori[1] has conducted analysis quantifying the outsized of market activity during FOMC meeting dates, as well as the potential loss from not investing on these days. Gospodinov and Jamali (2012)[4] found an intriguing linkage between the FOMC and stock market returns. Specifically, they revealed that while alterations in Federal funds target rates have a positive, though statistically insignificant, impact on implied volatility, the decomposed movement of the target rates carries a statistically significant effect on volatility. This effect is particularly pronounced in the unanticipated, surprise component of rate changes, aligning with the principles of the Efficient Market Hypothesis as elucidated by Rajesh Kumar[5].

The noteworthy correlation between the surprise factor in rate changes and volatility metrics underscores investors' portfolio adjustments in response to information updates. Notably, Gospodinov and Jamali identified that a surprise percentage point upswing in the Federal Funds rate led to a volatility change ranging from 2.80% to 8.04%, contingent upon the utilized surprise metric.

This observation underscores the potential for robust portfolio returns through accurate prediction of surprise rate movements post FOMC meetings. Despite the acknowledged correlation between market alpha and FOMC meeting dates, the frequency of market misestimation of FOMC monetary policy actions remains an underexplored topic. One prevalent proxy for Federal Funds Target Rate prediction relies on short-term Treasury bond rates, holding potential to gauge market mispricings. However, the absence of granular hourly data from a stock trading API necessitates deferral of further investigation to subsequent research endeavors.

Data

The target variable under scrutiny is the Federal Funds Target Rate. This choice is driven by its comprehensive influence within our current capitalist system. This system hinges on credit availability, governing credit card usage, future planning, and national growth. The FOMC minutes provide a clear insight into the likely trajectory of the Federal Funds Rate. Incorporating these minutes into a model holds practical value, not only for stock selection but also for broader informational context. Presently, the Federal Reserve Chair, Jerome Powell, has taken steps to enhance transparency, reducing the potential for market surprises during FOMC meetings. This move aligns with the Federal Reserve's role in stabilization.

Federal Fund Target Rate Data is available, as the following macroeconomic variables are, on the St. Louis Federal Reserve's public API, where various data can be manipulated and downloaded. While the level of the Federal Funds Target Rate is available, its augmentations in the 'Difference' (magnitude up or down of monetary policy), 'Increase' (magnitude of increasing), and 'Decrease' (magnitude of decreasing) are generated within the file.

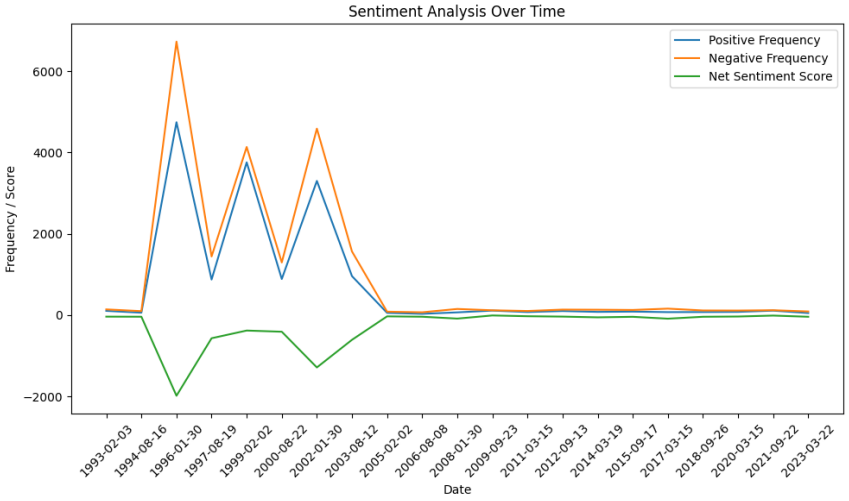
The predictive factors encompass the following: CPI, Consumer Sentiment, Retail Sales, Durable Goods Orders, short-term bond rates, long-term bond rates, bank reserves, and Housing sales. These variables can be broadly classified into two groups: indicators for inflation sentiment and indicators for macroeconomic health. Within the realm of inflation sentiment, we consider consumer sentiment, short-term, and long-term bond rates. Elevated bond rates often signify higher inflation expectations, prompting a potential need for the Fed to raise the Fed Funds rate to curb inflation. The remaining macroeconomic variables contribute to the fundamental assessment of the Federal Reserve's inflation metric. This is aligned with the dual mandates of monetary policy, namely, maintaining consistent inflation and stable employment. While other metrics like the unemployment rate, TIPS rate, bank reserves, and average hourly wages were desirable, their absence or data quality issues have regrettably precluded their inclusion. This omission significantly impairs the analysis, given their direct relevance to predicting the federal funds rate.

Exploratory Data Analysis

Prior to embarking on the Exploratory Data Analysis phase, a prerequisite involves quantifying the acquired Fed Minutes during the ingestion stage. Initially, the Loughran and McDonald Sentiment Word List is loaded. This compilation comprises domain-specific accounting and finance terms commonly featured in earnings calls. Each term is assigned a corresponding positive or negative sentiment. Merging this lexicon with the minutes facilitates the derivation of the cumulative count of positive and negative terms. By further integrating the terms from this list into the preprocessed text, a count of pertinent terms, their associated sentiment, and their occurrence frequency can be determined. The aggregation of positive and negative terms, along with their respective frequencies, results in a discernible Net Sentiment Score.

Commencing with the recently derived Net Sentiment Score, the subsequent step involves the evaluation of a time series plot depicting the occurrence frequency of positive and negative terms, juxtaposed with the Net Sentiment Score. Within this plot, a conspicuous skewness in the sentiment distribution is evident, a matter that will be expounded upon in subsequent sections of this report. A notable observation is the elevated "verbal activity" during the 1990s and early 2000s, characterized by

an increased frequency of words per minute aligning with the Loughran and McDonald Sentiment Word List. This observation calls for consideration, as the existing skewness potentially introduces an overly sentimental bias within the training dataset. Addressing this concern necessitates either skewness mitigation or data standardization.



Revisiting the matter of standardizing the occurrence frequency of positive and negative terms, a viable approach involves computing the proportion of positive and negative terms per minute, while concurrently incorporating data regarding the length of published minutes. This augmentation is significant since lengthier speeches could inherently convey a measure of sentimentality. The summary statistics align with earlier findings: the mean frequency of negative terms exceeds that of positive terms, culminating in a standardized sentiment score reflective of this pattern. The data is now primed for analysis alongside the remaining variables that have been ingested.

Calculating correlations across variables and identifying the highest correlations gleans some insight that may aid the analysis in terms of data selection and dimensionality reduction.

Level	Short-Term Treasury Bond Rate	0.996611
Average Hourly Earnings	CPI	0.996386
Bank Reserves	Treasury Deposits	0.994763
Positive Frequency	Word Count	0.986351
Average Hourly Earnings	Retail Sales	0.984795
CPI	Retail Sales	0.981418
Negative Frequency	Word Count	0.961431
CEI	LAG	0.955394
Negative Frequency	Positive Frequency	0.949909
Nonfarm Payroll	Retail Sales	0.923750
CPI	Nonfarm Payroll	0.915047
Average Hourly Earnings	Nonfarm Payroll	0.909158
Durable Goods Orders	Retail Sales	0.886853

Primarily, a striking correlation exists between the Federal Funds Rate and the Short-Term Treasury Bond Rate, with their movements appearing closely intertwined. This alignment poses a potential challenge for the study, raising the prospect of redundancy. Essentially, the trajectory of the Federal Funds rate can seemingly be inferred by observing Short-Term Treasury Bonds. This alignment also finds substantiation within macroeconomic logic.

The Short-Term Bond rate signifies the interest rate at which the Treasury extends loans to corporations and banks. It is determined by market dynamics governing these bonds. This rate can be likened to a "risk-free" rate due to the exceedingly remote likelihood of federal government default. Conversely, the Federal Funds Rate pertains to the rate at which institutional banks engage in interbank trading. The Federal Reserve exerts control over this rate through open market operations, influencing the reserves volume within the banking system and thereby facilitating the risk management of interbank trades. The Federal Reserve's execution of open market operations serves as an indicator of their perspective on short-term rates. Consequently, institutional investors tend to align with the notion that Short-Term Treasury Bonds should approximate the same rate.

Progressing, it becomes evident that a measure of redundancy is discernible among the most highly correlated variables. Notably, Nonfarm Payroll, Retail Sales, and Average Hourly Earnings exhibit robust intercorrelations, prompting the consideration of dimensionality reduction. This analytical step holds substantial promise for subsequent exploration. On a related note, autocorrelations, while informative, do not yield the same level of nuanced insight. Across the spectrum of macroeconomic factors, consistent patterns in autocorrelations are generally observed.

In conclusion, the examination of z-scores for both skewness and kurtosis spans the entire dataset. While this analysis is codified, it's worth noting that skewness, although primarily suited for cross-sectional data, can still unveil underlying trends within time series datasets. Notably, a scrutiny of individual time series reveals pronounced skewness in certain segments—specifically Net Sentiment Score and the frequency of positive and negative terms. The primary driver underlying these tendencies is the skewness inherent in Word Count. During the 1990s and early 2000s, the FOMC minutes were notably longer compared to contemporary sessions. Consequently, opting to compute the proportion of positive and negative terms played a pivotal role in normalizing the data and eliminating any potential outliers. This may prove an issue during model training, especially because this seeming outlier will have to be a part of the training model while the validation and test sets, which coincidentally fall right around the start of Jay Powell’s tenure. This could lead to significant bias which will need to be factored out.

Feature Engineering, Data Augmentation

The focal variable of significance will be the "Difference" variable in Federal Funds Rates. Given this objective, and observing the near-lockstep alignment of Short-Term Treasury Bond Rates with Federal Funds Rates, an advantage lies in incorporating the difference in Short-Term Treasury Bond Rates as an additional predictive element. Accordingly, this variable was introduced to the dataset. Additionally, Rolling Means are introduced to the pertinent macroeconomic variables that exhibit the most robust correlations with the outcome variable. These also help convey the notion of time trends being a key factor in monetary policy.

Furthermore, it has been recognized that while macroeconomic variables possess potential as robust predictors of monetary policy, their availability might not align with the same day as FOMC meetings. Even if synchronous data were attainable, the central aim of this analysis remains the anticipation of future monetary policy shifts, rather than the confirmation of ongoing policy alignment through the convergence of Minutes and macroeconomic indicators. Consequently, to equip the model with meaningful predictive capabilities, these variables must encompass observations from a sufficiently extended timeframe prior to the meetings.

For the scope of this analysis, a time delta of 5 days has been adopted. This approach ensures that the observations are situated at least 5 days preceding the meetings they aim to predict. Similarly, in the context of predicting the "Difference" variable and leveraging Minutes sentiment scores for predictive purposes, the variables associated with sentiment—such as word frequency and net sentiment scores—are adjusted upward. This alignment coincides with the previous meeting's data. As a result, their predictive efficacy for the upcoming meeting will be subjected to scrutiny through methods involving dimensionality reduction and univariate variable selection.

A final avenue to fully leverage the Minutes involves the application of word vectorization. This process transforms the textual Minutes from character strings into numerical vectors suitable for analysis. While a portion of this approach has been employed in calculating the net sentiment score, extending it to compute the frequency of individual words offers additional benefits. the process initiates by utilizing the CountVectorizer package, which translates the preprocessed text into distinct column vectors for each unique word. Through row-wise iteration, the occurrences of each word are computed, furnishing a numerical depiction of the predictive capacity associated with individual words. Taking this a step further, incorporating the sentiment index from the Loughran McDonald lexicon assigns each word vector a positive or negative value. Consequently, should a negative word recur 28 times within a specific meeting, the corresponding row value would be -28.

This results in an extensive list of variables—much more than would be ideal in terms of reproducibility and complexity. As such, univariate variable selection assists in selection features that have the highest importance and relevance to the desired outcome variable.

Model Building

Ordinary Least Squares

The chosen model approach for this analysis revolves around establishing a straightforward Ordinary Least Squares (OLS) model. Initiating the modeling process with OLS serves a dual purpose—it aids in the identification of influential variables while also highlighting potential challenges for more intricate models. As such, this methodology encompasses a progression through various stages, commencing with elementary linear models employing a predefined subset of hyperparameters for variable inclusion. This progression entails initial modeling with all variables, addressing concerns of overfitting, performing reverse variable selection, and culminating in the application of ridge regression.

Moving ahead, the primary criterion for evaluating each model's performance will be accuracy in both the training and validation sets. This measure is readily interpretable and adaptable across diverse model types. Furthermore, it enables the assessment of model bias by comparing the discrepancies between accuracy in training and validation sets.

It's noteworthy that not all models will inherently produce outputs in the customary format of the typical range of monetary policy, which spans from -1 to 1 in increments of 0.25. For instance, linear models may not easily align with the ordinal logistic nature of the outcome variable. To address this, a post-prediction rounding process is applied, enhancing the model's capacity to accurately gauge its precision.

The data underwent three iterations of fitting using the OLS model: manual subset selection, selection involving all variables, and reverse variable selection. This overarching approach will serve as the

blueprint for most subsequent models. The analytical trajectory typically commences with the inclusion of as many variables as feasible, initially prioritizing baseline accuracy over the potential risk of overfitting. Subsequently, this phase transitions into a subset selection process, where variables are refined based on sorting criteria, achieved through techniques such as univariate variable selection or hyperparameter optimization. The final stage generally encompasses a composite methodology, combining the aforementioned sorting strategies.

Of the three, the reverse variable selection model performed the best in terms of training and validation set accuracy for the OLS model. The benefit of the reverse variable selection model is the retention of explanatory data while also keeping model complexity in mind. The goal is to remove variables that minimize the change in explanatory p-value one by one until the model is as strong as it can be with the variables remaining. That said, it clearly leaves room for improvement, necessitating a more complex approach.

Metric	Training Set	Validation Set
Accuracy	0.822917	0.666667

Random Forest

The Random Forest Classifier presents an alternative to multi-level logistic regression. This method offers several advantages within the realm of machine learning. Functioning as an ensemble learning technique, it amalgamates numerous decision trees to formulate predictions, resulting in a robust and accurate model. Notably, it excels in handling high-dimensional datasets comprising a substantial number of features, as it selectively employs a subset of features during each division, mitigating the risk of overfitting. Moreover, Random Forests exhibit resistance to outliers and the ability to accommodate missing data without necessitating imputation. It's worth noting that compared to multinomial logistic regression, Random Forests are relatively more intricate, with each tree possessing its distinct subset of features and samples, culminating in the amalgamation of multiple decision trees.

The best of the three models in this case was one that optimized hyperparameter selection. Iterating until it received the best model possible, this model constructed a grid of possible configurations. The grid encompasses essential hyperparameters such as the number of decision trees, the depth of these trees, and the minimum samples required to split nodes. Employing GridSearchCV, the algorithm systematically searches this parameter space, utilizing 5-fold cross-validation for robust evaluation. The model is then fitted to the training data, allowing the identification of the optimal hyperparameter combination that yields the highest predictive accuracy. The final model integrates these optimized hyperparameters, consequently enhancing the model's performance.

Metric	Training Set	Validation Set
Accuracy	0.942708	0.708333

While a strong model in the training set, it suffers from overfitting in that the accuracy in the validation set is much weaker. This means that it has trained too specifically for the data in the training set

and is unable to adapt as well to new, unobserved data. That said, it has the highest accuracy measures of any model so far, and thus represents the current winning model.

XGBoost

XGBoost Models introduce advantageous characteristics when predicting multi-level ordinal logistic variables such as the 'Difference' variable. Firstly, the concept of boosting proves invaluable by amalgamating predictions from multiple weak learners to construct a robust predictive model. This framework also facilitates the identification of feature importance, elucidating the pivotal variables influencing the outcome prediction. This facet becomes particularly useful during the hyperparameter selection process. Moreover, XGBoost Models incorporate regularization techniques, effectively mitigating the risk of overfitting.

Although these models entail increased complexity compared to prior approaches, their enhanced accuracy justifies their intricacy. Notably, XGBoost leverages gradient boosting to optimize model performance. This mechanism involves iteratively adjusting the weights of each training instance based on the errors made by the preceding model. This approach essentially assigns higher significance to the data points that were inaccurately predicted by the preceding model, thereby refining the model's predictive prowess.

Metric	Training Set	Validation Set
Accuracy	1	0.625

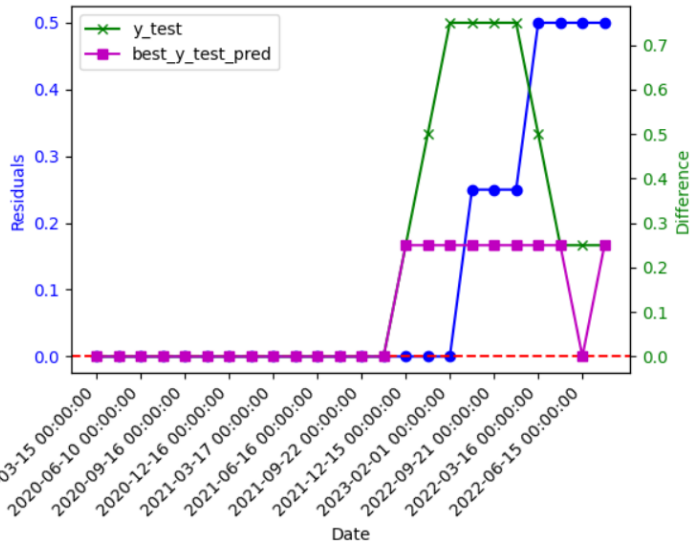
Unfortunately, these models clearly suffer from gross overfitting. Variations in accuracy to this degree generally indicate that the model will be weak to generalization. In fact, running a prediction against the test set returns an accuracy of zero. As such, despite some level of overfitting, the optimized random forest model is chosen as the best model for prediction.

Best Model Selection

Further analysis can be gleaned from analyzing the optimized random forest model.

Metric	Training Set	Validation Set	Test Set
MSE	0.00358073	0.0182292	0.0494792
RMSE	0.0598392	0.135015	0.222439
R^2	0.910565	0.719533	0.393617
Adjusted R^2	0.816322	1.08601	1.18596
Accuracy	0.942708	0.708333	0.708333

Looking at the test set, the model seems to have performed admirably, especially in relation to the validation set. A key issue with the model was its overfitting issue, but consistent accuracy between validation and test sets is encouraging, though the score itself is wanting.



Residual analysis shows some trend in residuals, which is often the cause of underlying trends that are not captures. Looking at the actual observations shows further how these residuals came to be. It seemed that the model was able to react and show increases to the 'Difference' variable, but did not respond strongly enough. This may be a weakness of the model, where previous values are not taken well enough into account. Earlier iterations of this analysis included a stepwise approach to analyzing monetary policy action. Namely, the first step would be to implement a logistic model to predict the direction of monetary policy ("Increase", "Decrease", or "Hold"). Then, a regression model would predict the degree to which monetary policy would "Increase" or "Decrease". In the first step, this model succeeds perfectly. Training a model specifically for the second step could prove promising.

Data-Centric AI

The outlined methodology signifies a shift in the prediction generation framework. In contrast to the prior methodology, which focused on identifying optimal models using pre-processed training, validation, and test datasets, this approach embraces a best-model strategy while incorporating adjustments to the training dataset to improve model performance. Subsequent stages will involve creating supplementary data constructs at a later phase.

To begin, more data is gleaned from the Economic Dictionary. While all unique words are included in the original dataset, filtering by those only in the Dictionary for which sentiment is necessarily provided reduces dimensionality substantially. Further, many of the actual words provide data themselves, in that there are those that are more prevalent in regular FOMC minutes than others, which reduces the instances of outliers. Running the model but only including words that have an average of 3 uses per minutes reduces dimensionality while also receiving competitively strong accuracy in the validation set, though it lags behind the previous model in generalizing to the test set. Further data augmentation was placed on leveraging feature importance from this new model in an attempt to reduce dimensionality and potentially increase generalizability. Finally, considering that these scripts do not arrive in a vacuum and in fact often describe much of the macroeconomic data at hand, interaction effects multiplying the vectorized words and the macroeconomic variables that had the most explaining power over the 'Difference' variable added further depth to the model. While a model arose that could provide similar levels of explanation over the validation set relative to the last model, the simplicity of the last model triumphs over the new one.

Model Risk, Bias, Ethics

Acknowledging the significance of identifying protected groups within machine learning is a crucial step towards ensuring equitable and just outcomes. By systematically identifying attributes such as gender and age that fall under protected categories, machine learning practitioners can proactively evaluate the potential for bias in their models. This approach not only guards against inadvertent discrimination but also fosters a culture of openness and responsibility in the creation and application of AI systems. As a result, recognizing and addressing protected groups plays a pivotal role in upholding the values of fair and impartial machine learning practices.

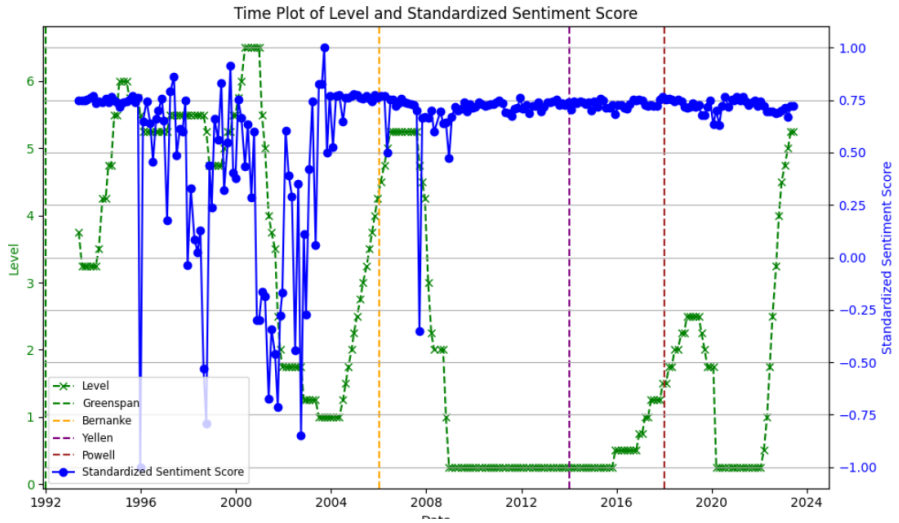
The dataset implicitly incorporates a diverse array of Federal Reserve Chairpersons spanning different historical periods. This diversity encapsulates various attributes including gender, age, and potentially religious affiliations. A notable instance is Alan Greenspan, the longest-serving Chairperson, whose tenure extended from 1987 to 2006, encompassing ages ranging from 61 to 80. An important milestone was marked by Janet Yellen, who became the first woman to head a central bank and the inaugural Democratic Chairperson since 1979, holding office from 2014 to 2018. This dataset comprises four distinct Chairpersons, effectively representing attributes like gender and age, which are considered protected categories.

Though the concept of protected age groups often pertains to individuals aged 40 and above, the unique nature of this role introduces a tenure spectrum spanning from 53 to 80. This variability introduces the potential for introducing bias into the dataset's dynamics. A noteworthy aspect of the data distribution pertains to the division of data into training, validation, and test sets. In this scheme, the validation and test sets predominantly capture the tenure of Jay Powell, the most recent Chairperson. Consequently, the training set faces the challenge of predicting outcomes under the leadership of a Chairperson characterized by distinct speech patterns, monetary policies, and differing levels of influence over the Board of Governors—a key policy-setting body.

The situation, however, presents an interesting dichotomy. While the inherent variability might introduce potential bias, the extended duration of the training set spanning three Chairpersons from the protected categories helps mitigate this concern. The objective is for the model to adeptly navigate biases linked to gender and age, eventually establishing a robust average model suitable for any future Chairperson. However, acknowledging the emergence of unique patterns under a new Chairperson is crucial, as it could potentially introduce unanticipated variability in the model's performance.

One of the variables examined for potential bias is the "Proportion of Negative Words." This bias consideration emerged during the initial phase of data exploration, uncovering a distinct period from 1996 to 2004 characterized by a notable increase in word density within FOMC Minutes, coupled with a corresponding rise in the frequency of both Positive and Negative Words. Exploring its alignment with the tenure of different Federal Reserve Chairs revealed a statistically significant variation in the Proportion of Negative Words across these terms. An essential concern, however, rested in assessing whether the same variable in the training data significantly differed from that in the validation and test sets, as such discrepancies could introduce bias. Delving into these differences based on individual Chair terms demonstrated a statistically significant variance in the Proportion of Negative Words between Fed Chair Powell and the other three Chairpersons. Guided by the notion that a training dataset encompassing three

distinct Chairpersons' tenures could mitigate bias, a comparison of Powell's term with the other three was calculated.



To delve into the protected groups bias, an analysis of mean difference in Proportion of Negative Words between Janet Yellen and the other three Chairmen was also performed, which resulted in a statistically significant difference in average means of -0.175. This represents an implicit bias within the training set, and requires factoring out. Thus, all observations within the timeframe of Yellen's tenure received an increase of 0.175, and all interaction effects were multiplied by the adjusted value.

Consequently, a final data augmentation step involved log-transforming the "Proportion of Negative Words" variable, as well as adjusting for the bias against sex as a protected group, before reapplying the model. The reapplied model showed similar accuracy across all data groups as previous models, and did not perform well enough in any to justify being used as the primary model.

Deployment Strategy

The final model is able to predict the FOMC monetary policy action 70.83% of the time. Regarding potential data risks in the analysis, due to the limited number of available FOMC data, the entire dataset requires more data to make a more adept model. As the validation and test sets are only 24 observations large, a single missed observation corresponds to a 4.2% decrease in accuracy. This goes to show that more data is required to make calculating the strength of each model more trustworthy.

This model will utilize batch reporting, and will ideally be run some time—roughly a week—before each FOMC meeting, which happens eight times a year during normal years (exceptions were made during the Covid-19 pandemic, which required the Federal Reserve to meet to discuss the providing of emergency stimulus to American citizens and how to adjust interest rates accordingly). Batch prediction has the benefit of delayed latency, as downloading and utilizing real-time data would be time-consuming and impractical.

Tracking the performance of this model post-launch will be very importance given the degree of overfitting in the final model. A model that has only a moderate amount of predictive success is not worth publication, or could lead to some adverse effects in public usage. In a worst case scenario, though unlikely, a great deal of money is hedged against Short-Term Treasury Rate price changes using this model, and a

market participant loses a good sum of money. As such, tracking model performance is incredibly important.

To do this, the predicted values will be calculated for ordinal log-loss. This is a modified version of the standard log-loss that is designed for ordinal data, like the monetary policy action predicted. This performance metric seeks to penalize a model's predictions based on the degree of misclassification. Ordinal log-loss also benefits from being able to return the direction of error. So, if the model is routinely predicting too high, then ordinal log-loss will be able to capture that. While simple accuracy sufficed for model creation, when considering the gradual degradation of model accuracy over time, the degree to which the model incorrectly predicts monetary policy action will be necessary. Selecting a baseline value for log-loss error will help automatically flag when the model is no longer meeting required performance standards.

For performance tracking, it is best practices to identify a green light (all is well with the model), yellow light (errors require tracking), and a red light (model should be refit) thresholds. Given that ordinal log loss takes directional error into account, the absolute value of ordinal log loss will be considered. These thresholds will be [0.0 to 0.2], [0.2 to 0.5], and [0.5<] for green, yellow, and red lights respectively. Because of the low frequency of prediction (8 observations every year), unless the model dips egregiously in predictability, it is not likely that the model will immediately reach critical.

That said, there are certain ways to prevent this from happening. During the "yellow light" period, it could be useful to analyze the actual residuals rather than just using the log-loss value as a signal. Should the loss values be negative, this can indicate that the model is over- or underestimating. If multiple iterations and observations in a row are incorrect in a single direction, this may be the result of bias, which would prompt an investigation into further variables in the training set that may cause bias.

Similar things can be said for ordinal Mean Average Error. This is a much more straightforward performance metric that can show the magnitude and direction of prediction error versus real values. It also has a much more straightforward interpretation, as it would represent the average degree and direction to which the model incorrectly predicts monetary policy action. Similarly, the thresholds for yellow and red lights will be the absolute value of 0.5 and 0.75 respectively.

The model will also go through a natural retraining lifecycle. Retraining would feel necessary, as the limited nature of the dataset would call for a more accurate model when given more datapoints. Further, naturally pushing the training/validation set cutoff further into Jay Powell's tenure would strongly reduce the bias in the training set, and could result in a more generalizable model. Eventually including the tumultuous period during the early months of the Covid-19 pandemic into the training set would further allow the model to be able to capture exigent scenarios and make for a more malleable model. Because the original dataset was already so limited, retaining may result in a wildly different model, even with the same model and features. Retraining every four years, in accordance with the general length of tenure for a Fed Chair as well as President, would be practical, but could lead to overfitting and bias in the training set. That said, retraining it every year or every eight meetings would feel redundant, and any inherent seasonality in the model would be reset each time. Thus, retraining the model every two years or 16 meetings would be apt for this model.

Should the model begin to be less and less reliable, retraining will become necessary, requiring updating data inputs and refitting the model. Increasing error could be the result of the natural relation

between variables changing over time, a key aspect of data drift. For example, it could be that the FOMC Minutes change in their intention, and thus naturally alter the Net Sentiment Score without intending to alter their monetary policy decisions. It could also be that in 2024, following an election, a new Fed Chair may take over for Jay Powell, who could have wildly different vernacular in their speeches, and could alter the effect of Net Sentiment Score.

For performance metrics, Kolmogorov-Smirnov Testing will help detect if the distributions of specific features have changed over time. This will help determine if the model requires reformatting. One benefit of using Kolmogorov-Smirnov Testing (KS Testing) is its non-parametric nature, which makes it versatile and applicable to all sorts of data distributions. It also is widely known, and has an easy interpretation, with a significance level that can be flagged should the data drift outside of preferred distributions.

Result

Through multiple ML iterations and models, as well as newer techniques such as Data-Centric AI, this analysis proved that to some degree, the FOMC's forward guidance does in fact have some predictive power over the next monetary policy decision. Starting with simpler models such as OLS and progressing to more advanced techniques like Random Forests and XGBoost, each approach was chosen strategically to leverage its strengths and address distinct challenges. The process involved meticulous steps like data preprocessing, feature selection, and hyperparameter tuning, all with a steadfast focus on accuracy and interpretability. The study thus offers a comprehensive roadmap for refining predictive models in the context of monetary policy actions, yielding insightful perspectives on the intricate dynamics between financial indicators and policy decisions.