Explain the model, analyze risk, bias and ethical considerations

1. Protected Groups

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.

2. Potential Bias

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. Because the previous iteration of the model relied on the interaction effect between the Proportion of Negative Words and other macroeconomic variables, the value of those proportions was factored out first from the relevant variables and then reapplied with the correct log-transformed proportion. The variable itself was also added for consideration. The reapplied model showed much greater accuracy in the training and validation sets, even near-perfect accuracy in the training set. It did, however produce equal accuracy in the test set, suggesting gross overfitting. It also did so with 74 more features, and thus suffers from model complexity. As such, the previous "best model" will remain so.

One interesting note about the data augmentation in favor of the log-transformed independent variable is the resulting feature importance. Whereas the "best model" included a single variable that had an outsized importance relative to others (as seen in the next section), the new model consisted of roughly four variables that had equal importance and seemed to describe most of the importance in the "Difference" variable. These variables were:

- Unemployment Rate_x_Short-Term_Treasury_Diff: 0.130
- Standardized Sentiment Score x_Short-Term_Treasury_Diff: 0.110
- Proportion Negative Words_ x_Short-Term_Treasury_Diff: 0.106
- Proportion Positive Words_x_Short-Term_Treasury_Diff: 0.097

The interaction effect between the unemployment rate and the difference in Short-Term Treasury bonds has the highest feature importance in the model. In fact, the difference in Short-Term Treasury bonds is chosen as an interaction effect for the top four highest importance features in the model, finally standing on its own as the fifth highest importance following that. This makes some sense, as the difference in Short-Term Treasury bonds has generally been the greatest predictor of the change in the Federal Reserve Target Rate. Fittingly, the remaining variables in the top 3 features in terms of importance also make sense in model predictability, in that they are the unemployment rate, Proportion of Positive and Negative Words, and Standardized Sentiment Score. These variables properly capture the remaining macroeconomic variables of highest importance (Unemployment Rate) and sentiment variables of highest importance (Proportion of Positive and Negative Words, Standardized Sentiment Score).

Given the prevalence of the variable of the difference in Short-Term Treasury bond prices, analysis of average means was again performed between all the Fed Chairs as well as Powell against the other Chairs, and no statistical significance was found, and thus remains.

All of this together does suggest a more balanced and less biased model. The only drawback is the model complexity. Using 164 variables simply isn't easily reproducible, nor is it easily interpretable. As such, the previous model remains as the "best model".

Due to the fact that it is clearly overfit, a subset of the variables with feature importance greater than 0.015 was generated, which was then again fit to the same model. This threshold was chosen as the previous "best model" had 0.015 as the lowest feature importance amongst its models. Running the model with this subset generated a much more generalizable model, with improved validation and test accuracy statistics at minimal cost to training accuracy.

Metric	Training Set	Validation Set	Test Set
MSE	0.00439238	0.0242133	0.0714057
RMSE	0.0662751	0.155606	0.267218
R^2	0.890293	0.627463	0.124901
Adjusted R^2	0.828245	1.18627	1.43755
Accuracy	0.984375	0.958333	0.75

As such, averaging for bias against protected groups and against the test set as well as dimension reduction techniques resulted in a model that is vastly superior to the previous "best model", and only uses 35 features, one fewer than the "best model". While it still seems to be incredibly overfit, it has strong predictive power over the validation set and has improved predictive power over the test set compared to the previous "best model". As such, this model will be used going forward.

3. Feature Selection

Random forest models generate feature importance by constructing a collection of decision trees. Each tree is trained on a bootstrap subset of the data and is split based on differing features at each node. When a tree is built, it makes splits to reduce the impurity of the data within each node, based on Gini impurity. During this process, when a feature is chosen for a split, that Gini impurity reduction is recorded—the larger the impurity reduction, the greater the feature importance. Then, this process is iterated over each feature. So, for the feature with the greatest importance, the interaction effect between the Proportion of Negative Words and Nonfarm Payrolls, the average decrease in the impurity measure caused by splitting on the 'Difference' variables was 0.59. As the feature with the highest importance, it was able to explain away the most randomness in the 'Difference' variable. The top 10 predictors are as follows:

- Unemployment Rate_x_Short-Term_Treasury_Diff: 0.323
- Proportion Negative Words_x_Short-Term_Treasury_Diff: 0.125
- CEI_RollingMean_x_Level: 0.082
- contraction_x_Short-Term_Treasury_Diff: 0.081
- LEI_x_Level: 0.054
- Short-Term Treasury Diff: 0.042

- Standardized Sentiment Score_x_Short-Term_Treasury_Diff: 0.031
- Housing Sales_x_LAG_RollingMean: 0.027
- Nonfarm Payroll: 0.023
- Short-Term Treasury Bond Rate_x_LAG_RollingMean: 0.022

The first feature, “Unemployment Rate_x_Short-Term_Treasury_Diff”, explains most of the reduction in impurity, with a feature importance roughly ten times that of the next variable. That said, the “Short-Term_Treasury_Diff” in some interactive effect occurs four times in the top ten features, and itself selected as sixth most important in the model. This may prove that there is value in interacting sentiment and macroeconomic data.

To change the output significantly, one could focus on altering the features with the highest importance scores. If the aim is to flip a prediction from one class to another, the features with the highest impact in the current prediction class would be modified. For instance, if the prediction class is influenced by "Unemployment Rate_x_Short-Term_Treasury_Diff” and “Proportion Negative Words_x_Short-Term_Treasury_Diff” in the first prediction, to flip the prediction, these values of these features could decrease, thus reducing their impact on the prediction. By adjusting these important features strategically, it's possible to create a shift in the prediction outcome.

4. Model risk

One issue to consider when thinking about model risk is “poor problem-solution alignment”. This is when ML models are generated in service of creating ML models rather than problem solving or adding value. This can happen when modelers are more interested in utilizing a certain model or type of ML rather than selecting the best process for capturing data. That said, like other models published on Medium before it[1], this model has precedent in leveraging Federal Reserve Minutes data to forecast monetary policy. Because the Federal Reserve consistently utilizes the FOMC and a key signal in forward guidance, it is only natural to attempt to glean deeper meaning from the specific choice of wording that the Federal Reserve insists on using, as they take wording extremely seriously so as not to disturb market forces simply by releasing Minutes.

Another model risk is overfitting, which this model clearly is in violation of. Overfitting is when a model that is trained on a training set is overly biased toward that data, and is limited when it comes to predictability in the validation and test sets. Looking back at the validation and test results for the best model

Metric	Training Set	Validation Set	Test Set
MSE	0.00716146	0.0390625	0.0572917
RMSE	0.0846254	0.197642	0.239357
R^2	0.821131	0.398998	0.297872
Adjusted R^2	0.797846	-12.823	-15.1489
Accuracy	0.885417	0.708333	0.708333

The model clearly outperforms in the training sets versus the validation and test sets, which could indicate overfitting. That said, there are a few caveats. The model actually performs equally well against the validation and test sets, which suggests that it is not overly overfit. Another point is the size of both the validation and test sets. Due to the limitations in data, the number of observations to predict over is only 24 each. As a result, one incorrect predicted observation is equivalent to a .042 reduction in accuracy. Comparing the training set against the validation and test sets, this is equivalent to predicting only 4 fewer observations. As such, given the disparities in relative size of the training and validation sets, this may suggest again that the model is not overly overfit. This does not mean that the lesser accuracy is acceptable, only that the model may not be too overfit.

When looking at the new "best model", it runs against some of the aforementioned issues

Metric	Training Set	Validation Set	Test Set
MSE	0.00439238	0.0242133	0.0714057
RMSE	0.0662751	0.155606	0.267218
R^2	0.890293	0.627463	0.124901
Adjusted R^2	0.828245	1.18627	1.43755
Accuracy	0.984375	0.958333	0.75

It performs incredibly well in the training and validation sets, but drops off dramatically in accuracy in the test set. That said, it performs better in all categories versus the previous model (the Adjusted R^2 being higher in the validation and test sets is an indication that there are more features than there are observations in those sets—not true in the training set). Given that the test set comprises the exigent monetary policy reaction to the Covid-19 pandemic, the success in the validation set and weaker accuracy in the test set is not a dealbreaker.

Another thing to consider around model risk and ethics is data availability. With the increasing sophistication of data scraping methods, there is concern whether models are pulling data from confidential sources that should not be made public. As for this model, all data is pulled from public, open-source APIs such as the Federal Reserve of St. Louis (FRED) and FedTools. Further, all FOMC Minutes are inherently public, and intended to be heard by as many market participants as possible. As such, this model does not infringe on any data confidentiality concerns or privacy issues.

One potential risk is that of a feedback loop. While this model is not adept enough to become standard practice, a better model adopted by many market participants could cause the Federal Reserve to even further scrutinize their choice of words when releasing Minutes. They could even more specifically intentionally avoid certain words used in the model, requiring a reconfiguration due to selection bias. While this model is unlikely to galvanize enough attention to force the Fed to rethink their forward guidance strategy, another iteration of the model could.