Data Centric AI

The proposed methodology entails a recalibration of the prediction generation framework. In contrast to the preceding three reports, which emphasized the identification of optimal models for pre-processed training, validation, and test datasets, the present approach adopts the best-model paradigm while introducing modifications to the training dataset for enhanced model refinement. Consequently, the analysis commences by importing the complete dataset without undergoing dimensionality reduction, as previously undertaken in Week 5. Subsequent phases will involve the construction of additional data constructs at a later juncture.

One final point is that due to data errors in uploading and maintenance, updated data was imported at the top of the funnel, including new FOMC monetary policy data and updated macroeconomic data. All previous models were re-run due to this fact, and comparisons are made with updated metrics.

1. Error Analysis

This report required a thorough reconsideration of all of the variables included and how they may have led to errors in the model. One identified error was that sentiment for negative words was incorrectly applied to the Loughran-McDonald Master Dictionary, and the data was actually considering words that had an 'uncertain' sentiment rather than 'negative' sentiment. This may have increased randomness in the data, as it was a weak predictor for monetary policy action. This led to a reconsideration of the univariate variable selection process, as different words representing properly negative sentiment were included while words with an uncertain sentiment were dropped. Another was the fact that some errors were clearly autocorrelated—that when monetary policy increased over a few meetings, the model would not learn from its mistake after the first increase. Thus, autocorrelated errors were a concern, and 'Date' was reincluded as a necessary variable.

2. All Variables

The first approach leverages all of the variables generated in data ingestion in a random forest model that optimizes over estimators, the value of cross-validation, and features. It also tests a few random seeds to return the random seed associated with the best model. The previous iteration of this model calculated over thousands of random seeds, which helped make it the most accurate model of all three reports. That said, given that this model has to wrangle a dataset of 833 variables and 192 observations each, iterating so many times is structurally impossible. As such, this model has to adjust and only use 10 random seeds. This still results in long run times, but a model that performs admirably.

Metric	Training Set	Validation Set
MSE	0.00455729	0.03125
RMSE	0.0675077	0.176777
R^2	0.886174	0.519199
Adjusted R^2	1.00323	1.0016
Accuracy	0.927083	0.625

The model utilizes 50 variables, which is not only close to as small in dimensionality as the best model (44 variables), but performs relatively well against the validation set. That said, it does not reach a level of validation accuracy that would be acceptable in practical application.

3. Relevance Score Subset

Given that the data was updated with newer words with proper negative sentiment, running univariate variable selection to assess whether these updated data would provide a more accurate model than the previous iterations was tantamount. Yet the size of the dataset of vectorized words continued to prove a hassle in model production, so a subset of the vectorized words with an average of 3 instances in each FOMC was generated. This would reduce dimensionality greatly while also highlighting words that regularly occur in FOMC meetings. After creating a subset of vectorized words, the full merged dataset with macroeconomic variables has feature importance extracted. With model complexity in mind, all variables with feature importance greater than an absolute value of 0.2 were used for a further subset, resulting in a dataset with 23 variables. Running this dataset through the aforementioned model returns the following accuracy measures:

Metric	Training Set	Validation Set
MSE	0.00260417	0.0260417
RMSE	0.051031	0.161374
R^2	0.934957	0.599332
Adjusted R^2	0.926489	-8.21536
Accuracy	0.958333	0.708333

This is quite an admirable model, especially considering that relative to the best model's 56 features, this one only uses 20. This shows great accuracy in the training set, but does seem to fall off in the validation set, suggesting overfitting. That said, given that it has the highest accuracy in the validation set, this will be the best model so far.

4. Interaction Effect Model

One final consideration for variable augmentation is to consider the interaction effect between the vectorized words and the macroeconomic data. Intuitively, these words are always used to describe macroeconomic data and do not exist without reference. As such, within the subset of data from the previous model, interaction effects for each of the vectorized words and each of the macroeconomic variables were created by multiplying the two. Not only would this instill context into each word, but also hopefully extract interaction terms that have no meaning. The resulting model utilizing this data has these accuracy measures:

Metric	Training Set	Validation Set
MSE	0.00846354	0.0390625
RMSE	0.0919975	0.197642
R^2	0.788609	0.398998
Adjusted R^2	0.629581	1.23429
Accuracy	0.880208	0.708333

There is a similar validation set accuracy (approximate to 17/24 observations predicted) as the previous model, though it is fairly weaker in the training set. One note should be that given the limited dataset, getting a single observation more predicted results in a dramatically greater validation set accuracy. When the validation set is so small, predicting a single observation more is equivalent to 4.2% greater accuracy. This is a huge limitation of this dataset, and could result in some models being excluded from the selection process due to random errors against the validation set with huge accuracy implications. Given that fact, the greater accuracy measures in the previous model such as MSE, and RMSE as well as the greater accuracy in the training set, despite the implications of overfitting, make it a better model.

5. Test Set Accuracy

Running the model against the test set, the following accuracy measures are returned:

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

These accuracy checks are heartening, specifically because they are so much better than the last “best model” in terms of generalization in the validation and test sets, where the last model scored 0.542 and 0.667 respectively. This alone would be qualifying enough, were it not for the fact that even the MSE and RMSE are better than the previous “best model” in both the validation and test sets as well. As such, this model will be used as the “best model” going forward.

6. Insights

Through this process, some insights were gleaned from model creation to data centric AI. Firstly, model creation certainly has its limits, and finding the best predictive system is not just about using the right code or choosing the best model, but also hyperparameter selection. This fine-tuning is what can make a good

model an even better one. Secondly, the error identification is key in data centric AI. This can be seen in the fact that the “best model” in fact uses the ‘Date’ variable whereas the previous iteration did not. This could be why the accuracy through the validation and test sets was consistent, even when faced with varying monetary policy periods.