Topic Identification and FinBERT: Developing Sentiment-Based Predictors from the Beige Books to Supplement US 10-Year Treasury Yield Prediction

Kathryn Painter
UC Berkeley MIDS
December 2024

Abstract

The US 10-Year Treasury yield serves as an important benchmark for interest rates and economic performance. Given the dynamic nature of capital markets, financial institutions rely on benchmark projections, such as 10-year Treasury yields, to inform their financial strategy. This paper investigates the use of LDA and FinBERT to perform topic identification and sentiment scoring, respectively, to create a set of sentiment-based predictors to supplement the prediction of US 10-year Treasury yields. The final XGBoost model did not show improvement over the baseline autoregressive model emphasizing the complex nature of yield prediction and reinforcing the use of traditional econometric approaches. Nevertheless, there are opportunities to address limitations in further works.

1 Introduction

Yields on US Treasury securities have long been considered one of the most trusted proxies of investor expectations and the US economic outlook. The US 10-year Treasury yield has particular interest to investors as it serves as a benchmark for other interest rates such as mortgages and corporate debt. Movements in the 10-year Treasury yield require investors, financial institutions, and policymakers to adjust their strategies accordingly. Therefore, financial analysts and researchers have worked to refine approaches to increase accuracy in predicting yields. Approaches have evolved from univariate autoregressive models to multivariate econometric models, and more recently, to machine learning (ML) models such as Gaussian Process Regression (GPR), XGBoost, and even Long Short-Term Memory (LSTM) networks. Nevertheless, these approaches primarily focus on the use of quantitative data such as historical yields at different maturities, market indices, and macroeconomic indicators to predict future yields. Only recently, via the advancements in natural language processing, have researchers begun to explore the use of qualitative data to supplement yield predictions.

This paper expands on existing research, leveraging Beige Book data to develop a set of sentiment-based predictors to supplement the prediction of US 10-year Treasury yields using an XGBoost regression model. This work is largely motivated by the notion that the Federal Open Market Committee's (FOMC) decisions on monetary policy have a direct impact on US 10-year Treasury yields and that the committee itself relies heavily on the information provided by the twelve Federal Reserve regions via the Beige Book (Fujiwara et al., 2023).

2 Background

US Treasury yield prediction is a complex task due to the high variability of multiple factors such as macroeconomic conditions and market movements. Until recently, yield projections have been produced via ordinary econometric models. Oosterlaken (2020) attempted to use machine learning methods to demonstrate their ability to outperform traditional autoregressive (AR) and vector autoregressive (VAR) models. The ML models considered included Regression Trees, Random Forests, Gradient Boosted Trees, Regularization, and Gaussian Process Regression (GPR). To assess model performance, 1, 2, 3, 5, and 10-year maturity yields were projected on four different subsamples at three different horizons (1, 3, 12 months). Model inputs included historical yield data and a wide range of macroeconomic indicators. Oosterlaken concluded that while some ML models achieved lower rMSE values, ultimately they did not significantly outperform AR models. The author highlights that one of the primary limitations could be a lack of appropriate explanatory variables. Other possible limitations include the inherent lag present in macroeconomic data and the granularity of the data. National macroeconomic indicators may not be granular enough to explain the variation in treasury yields.

Fillippou et al. (2024) leveraged regional Beige Book data to show that economic sentiment is “extremely heterogeneous” across the twelve Federal Reserve regions, providing evidence that national economic sentiment is not a simple aggregation of regional sentiment. They used FinBERT–a variant of BERT fine-tuned on financial texts–to calculate the sentiment associated with each Beige Book release at the national and regional levels. They also demonstrated that regional sentiments improved the predictive power in forecasting US recessions compared to national sentiments. Fujiwara et al. (2023) showed that sentiment derived from the Beige Books, specifically regarding inflation and unemployment, lead to more accurate Treasury yield spread predictions using eXtreme Gradient Boosting (XGBoost). However, simple topic identification was used to extract topics related to inflation and labor. Sentences containing at least on of the following keywords [’inflation’, ’price’, ’cost’, ’energy’, ’labor’, ’stability’, ’wage’] were considered to be inflation and labor related topics.

The following methodology builds upon existing research combining regional economic sentiment derived from the Beige Books using FinBERT and Local Latent Dirichlet Allocation (LDA) to identify topics present in the Beige Books to generate a set of sentiment-based predictors to supplement yield prediction via XGBoost. This attempts to overcome the limitations of previous research by decreasing the lag in macroeconomic data and increasing the granularity of topics rather than using a handful of broader predetermined topics.

3 Methodology

3.1 Data

The data underlying this report was collected from two sources. The 10-year Treausry yields and macroeconomic indicators such as consumer price index (CPI), unemployment rates, producer price index (PPI), and personal savings rate are pulled directly from the Federal Reserve Economic Data (FRED) API. The Beige Book data was web scraped from the Federal Reserve Bank of Minneapolis Beige Book Archive which hosts all releases of the Beige Book since 1970 (see Appendix 6.2). The data is of monthly frequency, spanning from 1970-01-01 to

2024–03-01. To align 10-year Treasury yields to monthly macroeconomic data, yields were aggregated as a daily average for a given month.

3.2 Baseline Models

Given the mixed results on whether ML models improve prediction accuracy compared to traditional univariate timeseries models, two baselines were established.

Baseline #1 was chosen to be a first order autogressive model, denoted as AR(1), as described by the equation below.

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

where y_t is the yield at time t, c is a constant, ϕ_1 is the coefficient, and ε_t is the error term. Oosterlaken (2020) showed the highest accuracy was obtained with an AR(1) model, therefore the same model is considered here.

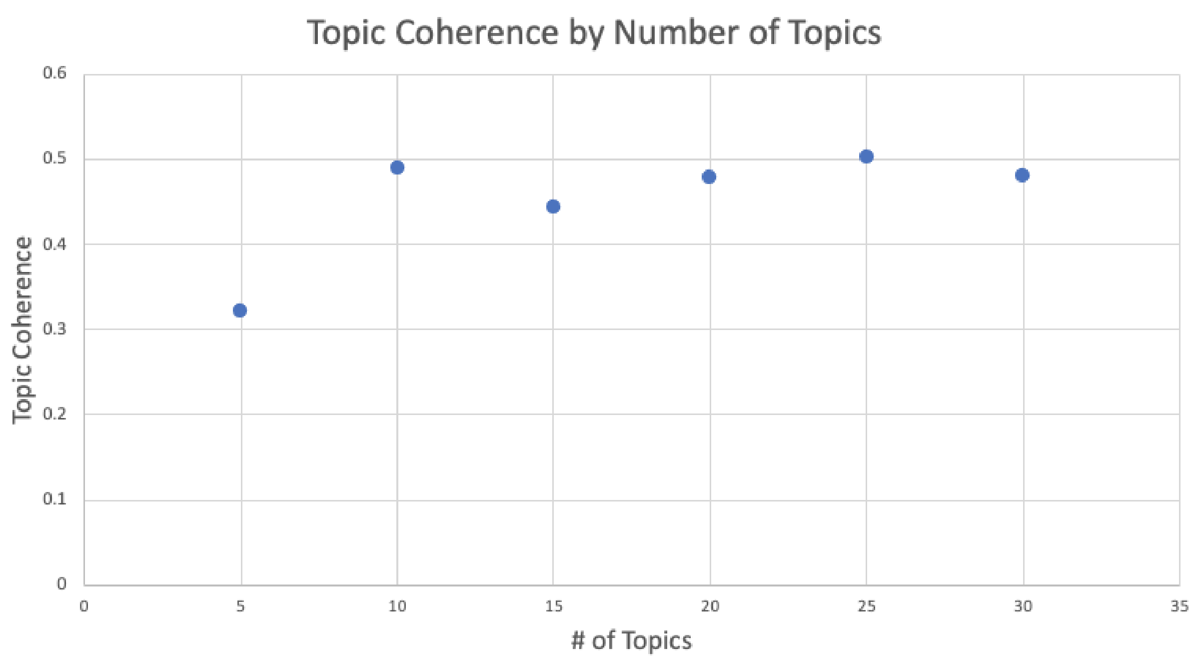
Baseline #2 leverages XGBoost regression to predict future 10-year Treasury yields using past yields and covariates such as CPI, PPI, unemployment, and personal savings rate. A lag value of 12 was used for both past yields and past covariates due to an observed annual seasonal cycle present in 10-year Treasury yields (see Appendix 6.5). Fujiwara et al. (2023) demonstrated improved accuracy in predicting treasury yield spread when supplementing an XGBoost model with sentiments from the Biege Books. Therefore, this model is considered as a basis for comparison.

The following performance metrics were considered for each baseline: Mean Average Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (rMSE). These metrics were calculated for three different prediction horizons (1, 6, and 12-months) using a series of historical forecasts starting from 2010-01. For example, for a prediction horizon of 12-months, yields up to 2009-12 were used to train the model and predict yields from 2010-01 to 2010-12. This process was repeated until predictions reached the final time index, 2024-03. MSE was chosen as the final evaluation metric for model comparison as it is widely used for timeseries prediction tasks.

3.3 Topic Modeling

Latent Dirichlect Allocation (LDA) is a generative probabilistic model that attributes words to topics across a collection of documents (Blei et al., 2003). In the context of documents, LDA aims to assign each word to a distinct topic and each topic to the appropriate documents. Since topics can belong to the same document, and words can belong to the same topic, Dirchilect distributions are favored. In this case the ability to model multiplicity is preferred due to the the natural overlap of economic themes present in the Beige Books. Dirchilect distributions are multivariate distributions that are modeled by two parameters α and β. In the LDA algorithm, α is the document-topic density and β is the word-topic density. Document-topic density refers to the number of topics present in a given document, therefore higher values of α imply more topics per document. Whereas word-topic density refers to the number of words present in a given topic, higher values of β imply more words per topic. LDA uses conditional probabilities of topics given documents and words given topics to assign words to topics and topics to documents. In 2010, Brody & Elhadad demonstrated that LDA could be applied to shorter texts such as product reviews by treating each sentence as a separate document–describing the approach as Local LDA as it prevents the inference of global topics.

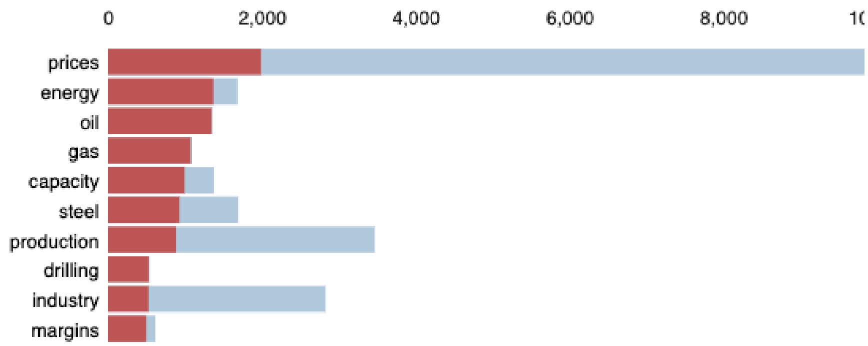
To ensure effective topic discovery and reduce noise, each Beige Book sentence was split into a list of words which were then lemmatized and scrubbed of all stop words and adjectives. In an effort to reduce runtime, 25% of the data was randomly sampled and used to train a Local LDA model initialized with 30 topics. This model achieved a baseline topic coherence score of 0.3658. In this context topic coherence was selected to measure how well the topic relates to the information presented in the Beige Books. Hyperparameter tuning was performed by training a collection of models using different combinations of N (number of topics), α, and β . The model that achieved the highest coherence (0.4911) with less than 25 topics exhibited the following parameters: N = 10, α = 0.01, β = 1.



The resulting LDA-derived topics were investigated and each was assigned a more intuitive name. For example, the LDA representation of topcic_8 is shown below as a weighted sum of the individual words that belong to the topic.

> 0.035*"prices" + 0.024*"energy" + 0.024*"oil" + 0.019*"gas" + 0.018*"capacity" + 0.016*"steel" + 0.016*"production" + 0.009*"drilling" + 0.009*"industry" + 0.009*"margins"

The frequency of top 10 words in this topic relative to the overall text is shown below. As shown in the chart, the word “energy” appears in the overall text approximately 1,800 times. It’s high estimated frequency within topic_8 suggests it is primarily associated with topic_8. Given the top 5 highest frequency words within the topic, this topic was assigned the following name "energy".



The final intuitively labeled topics are: mortgages, percent, consumer goods, wages, credit, agriculture, capital, energy, travel, and services.

3.4 Sentiment Based Predictors

FinBERT was used to estimate the sentiment of each topic present for a given release of the Beige Book. FinBERT is a language model based on BERT created to examine financial texts (Araci, 2019). The model achieved an 86% accuracy when applied to sentiment classification tasks on the TRC2-financial data set. The pretrained model was pulled from HuggingFace (see Appendix 6.3) and ran on each sentence to come up with an overal sentiment label ("positive", "neutral", "negative") and score (0-1). The scores associated with a "negative" label were multiplied by -1 and scores associated with the "neutral" label were set to 0. Individual scores were then grouped by release date, region, and topic and averaged to derive topic-level sentiments.

3.5 Final Model

The final model expands on *Baseline #2* by leveraging the same XGBoost Regression architecture with the dervied regional topic-level sentiments as additional inputs. Since the Beige Books are released eight times per year, missing months are forward filled using prior month's sentiments.

4 Results

The table below reports the final MSE values at various prediction horizons for both the AR(1) and XGBoost baselines, as well as the final XGBoost model supplemented with regional topic-level sentiments.

Horizon	1-month	12-month	6-month
Model			
AR(1)	0.04	0.83	0.41
XGBoost + Macro Ind	0.08	1.13	0.51
XGBoost + Macro Ind + Sentiments	0.10	1.41	0.78

As seen above the model that achieved the best MSE score across all three prediction horizons was the AR(1) model. While this result aligns with Oosterlaken's (2020) findings, it contradicts the findings of Fujiwara et al. (2023) that showed an XGBoost model supplemented with both macroeconomic indicators and sentiments from the beige books improved the predictive power. However, this result is not directly comparable to the result of Fujiwara et al. (2023) due to the use of different macroeconomic indicators and prediction of monthly daily average yields instead of daily yield spreads.

5 Conclusions

This paper aimed to build upon existing research combining regional economic sentiment derived from the Beige Books using FinBERT and Local Latent Dirichlet Allocation (LDA) to identify topics present in the Beige Books to generate a set of sentiment-based predictors to supplement US 10-year Treasury yield prediction via XGBoost. While existing research showed promise, the final model did not outperform the autoregressive baseline. This may be indicative of the nuanced complexities present in predicting economic data, such as the 10-year Treasury yield. The inherent timeseries nature of the data further complicates this task due to the presence of intricate temporal dependencies between lagged values and covariates.

A few limitations to consider in further works include the use of FinBERT to score sentiments of economic textual data and more thorough feature selection prior to re-training with sentiments. FinBERT was trained on the Financial PhraseBank data set which is a collection of financial news articles. The author's goal was to fine-tune a language model so that it could interact with trading data. While economics and finance are closely related, it may be worth fine-tuning BERT on the Beige Books themselves or federal economic reports. Another opportunity for improvement would be conducting a thorough feature selection to reduce the feature set to only include topic-level sentiments that demonstrate a relationship with treasury yields.

6 Appendix

6.1 References

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models* (No. arXiv:1908.10063). arXiv. http://arxiv.org/abs/1908.10063

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003). Latent dirichlet allocation | The Journal of Machine Learning Research. (n.d.). Retrieved December 8, 2024, from https://dl.acm.org/doi/10.5555/944919.944937

Brody, S., & Elhadad, N. (2010). An Unsupervised Aspect-Sentiment Model for Online Reviews. In R. Kaplan, J. Burstein, M. Harper, & G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 804–812). Association for Computational Linguistics. https://aclanthology.org/N10-1122

Filippou, I., Garciga, C., Mitchell, J., & Nguyen, M. T. (2024). Regional Economic Sentiment: Constructing Quantitative Estimates from the Beige Book and Testing Their Ability to Forecast Recessions. *Economic Commentary*, *2024–08*. https://doi.org/10.26509/frbc-ec-202408

Fujiwara, M., Suimon, Y., & Nakagawa, K. (2023). Treasury yield spread prediction with sentiments of Beige Book and macroeconomic data. *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 337–342. https://doi.org/10.1109/IIAI-AAI59060.2023.00073

Li, Z., Wang, B., & Chen, Y. (2024). Incorporating economic indicators and market sentiment effect into US Treasury bond yield prediction with machine learning. *Journal of Infrastructure, Policy and Development*, *8*(9), Article 9. https://doi.org/10.24294/jipd.v8i9.7671

Oosterlaken, J. E. (2020). *Predicting the US Treasury Yields using Machine Learning Techniques*. https://thesis.eur.nl/pub/53492

Zhu, X., Yang, S. Y., & Moazeni, S. (2016). Firm risk identification through topic analysis of textual financial disclosures. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. https://doi.org/10.1109/SSCI.2016.7850005

6.2 Data Sources

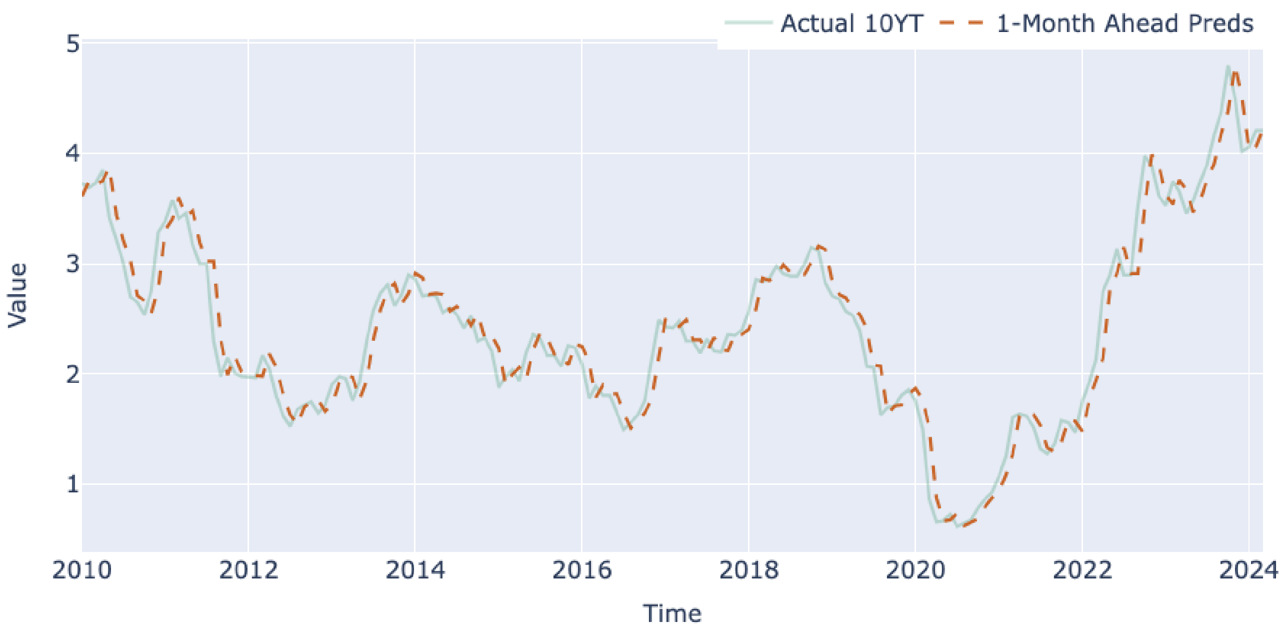
Fred API
https://fred.stlouisfed.org/docs/api/fred/
Beige Book Archives
https://www.minneapolisfed.org/region-and-community/regional-economic-indicators/beige-book-archive

6.3 FinBERT

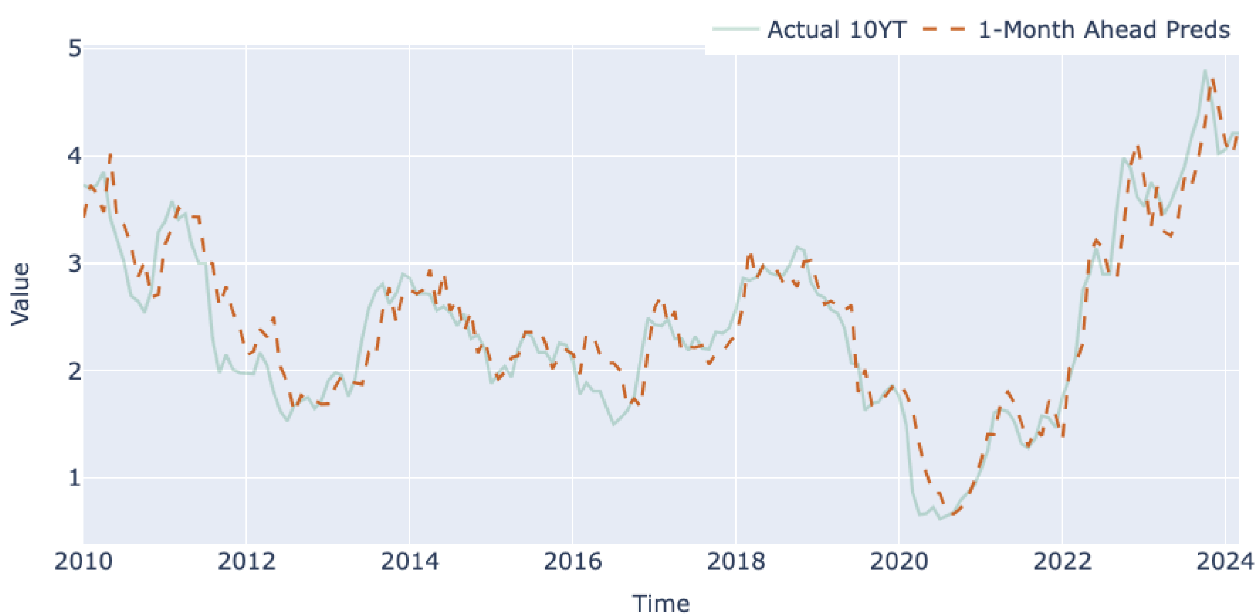
HuggingFace Model
https://huggingface.co/ProsusAI/finbert

6.4 1-Month Ahead Predictions (2010 - 2024)

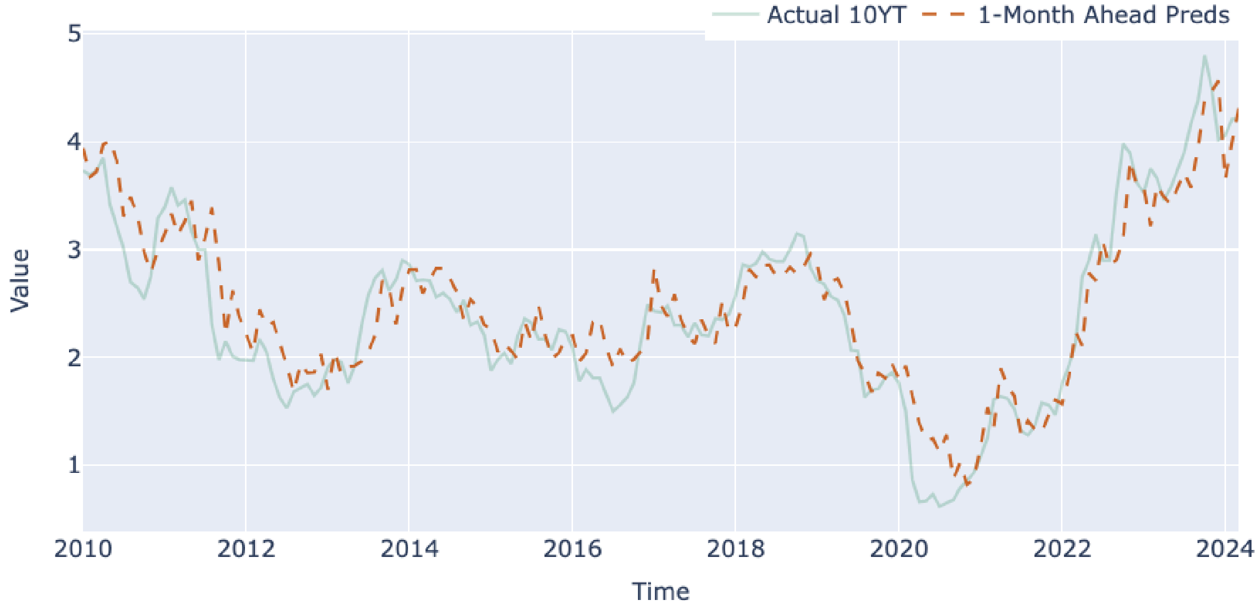
AR(1)



XGBoost + Macroeconomic Indicators



XGBoost + Macroeconomic Indicators + Beige Book Sentiments



6.5 Seasonal Decomposition