INCORPORATING GOOGLE TRENDS AS BIG DATA FOR ENHANCED INFLATION FORECASTING: EVIDENCE FROM KAZAKHSTAN

BY

RAKHAT BEISENBEK

THESIS

Submitted in partial fulfillment of the requirements

for the degree of Master of Science in Finance

in the Graduate School of Business

Nazarbayev University, 2024

Astana, Kazakhstan

Advisor: Dr. Aigerim Yergabulova

Abstract

Accurate inflation forecasting is essential for policymakers and businesses, allowing for informed decision-making and economic stability. Most historical forecasting methods are based on a few key macroeconomic factors while leaving a wide variety of more detailed data from the big data sources unused. This paper aims to analyze the performance of using Google Trends data on improving the inflation forecast of Kazakhstan and compare the result with the traditional macroeconomic models.

The research develops two primary models: a baseline model that includes conventional macroeconomic indicators such as GDP growth, oil prices, NEER, and inflation expectations, and an enhanced model that integrates Google Trends data for key terms like "inflation," "GDP," and "exchange rate." These data are preprocessed with standardization, lagging, and percentage change transformations. Machine learning techniques, specifically random forest and gradient boosting regressors, are applied to evaluate model performance. Statistical validation includes Likelihood Ratio tests for out-of-sample density forecast accuracy, as well as the Mean Squared Error (MSE), Mean Absolute Error (MAE) evaluation metrics calculation, Mincer-Zarnowitz regression for bias, and the Diebold-Mariano test for forecasting accuracy.

Findings show that incorporating the set of variables of Google trends improves the forecasting accuracy of the enhanced model by making relatively small MSE and MAE compared to the baseline. The Likelihood Ratio test supports the improvement of the models for density forecasting, and in terms of feature importance, Google Trends data turns out to be critical for the enhanced model. While the result of the Diebold-Mariano test turned out to show marginal significance, extending the dataset period and applying advanced techniques further maintained the robustness of the enhanced model.

This research proves that Google Trends as a big data contributes to enhancing the accuracy of the inflation forecasts for developing economies. Although sentiment analysis was initially considered, it was excluded from the study due to data limitations in Google news and the absence of access tokens for media sources like Facebook.

Acknowledgements

I am deeply grateful to the Graduate School of Business at Nazarbayev University for providing the academic foundation and resources necessary for the successful completion of this research.

First and foremost, I express my sincere gratitude to my thesis advisor, Dr. Aigerim Yergabulova, for her unwavering guidance, insightful feedback, and constant inspiration throughout the research. Their expertise and patience have been priceless during this thesis work.

I would also like to express my heartfelt thanks to all the professors at the Graduate School of Business who have contributed to my academic journey. Their knowledge, mentorship, and dedication have not only broadened my intellectual horizons but also inspired me to push the boundaries of my understanding.

This thesis is a culmination of collective efforts, and I am truly indebted to everyone who has been part of this journey. Many thanks for making this possible.

Table of Contents

1. Introduction

Inflation forecast is one of the most important inputs to policymaking, especially for emerging markets like Kazakhstan. Inflation forecasts help in the decision making of monetary policy for its stability for the financial market and the overall perception of the public. In recent years, central banks worldwide have increasingly explored unconventional data sources, such as internet search trends and social media sentiment, to enhance traditional forecasting models. This thesis investigates whether incorporating Google Trends data into inflation forecasting models can improve their predictive accuracy within the Kazakhstani context.

Google Trends data serves entirely as the real-time proxy of consumers' concerns and interests that are not reflected through the conventional economic indicators. The application of Google Trends in economics was initiated by Choi and Varian (2012) to prove that Google Trends can facilitate the improvement of nowcasting for unemployment claims, automobile sales, and consumer confidence in the United States. Their work focused on time factors of search data that provide insights into public sentiment and behavior almost instantaneously, compared to the delayed nature of official macroeconomic statistics. This discovery provided some solutions for using trends in internet searches for other purposes and opportunities in many different fields involving economic forecasting such as inflation.

Choi and Varian's also showed that such data can complement standard economic measures for short-term forecasting particularly in increased volatility. For instance, during the global financial crisis of 2008, changes in the search terms offered signals of the forthcoming changes in consumer behavior before they appeared in official statistics. Such capabilities are particularly important for emerging economies such as Kazakhstan because, for example, local data collection systems might be characterized by time gaps or incomplete coverage.

Kazakhstan's case study needs to be considered properly because of the commodity export dependence, young financial markets, and the official dual-language information release channels. ARIMA and VAR have been commonly used for inflation forecasting in Kazakhstan but have weaknesses and fail to incorporate nonlinear movements and external shocks in many cases. By incorporating Google Trends data for key search terms like "inflation," "GDP," and "exchange rate," this thesis explores whether these additional data sources can enhance predictive accuracy. Also, this research extends Choi and Varian's methodology application literature to a relatively underexplored emerging market context.

The performance of the proposed models is assessed statistically rigorously by using several statistical tests and machine learning techniques. To compare different forecasting models, Mean Squared Error (MSE) and Mean Absolute Error (MAE) performance criteria are used and the performances of each are compared through

the Diebold-Mariano test. Additionally to these traditional metrics, more advanced machine learning techniques are used to compare the performance of models. One of them is Lasso Regression, and this helps in identifying the most influential predictors and serves as a benchmark for evaluating model interpretability and performance. To one level, Random Forest and Gradient Boosting methods are used as an expansion to Hastie et al. (2009) that would capture the intricate nonlinearity and that would enhance the robust sample predictions Withers (2014) united with the real-time search data of Choi and Varian.

The findings of this thesis are useful for both academic literature and policy discussions. Hence from a methodological perspective they give an idea about the efficiency of alternative data sources to improve the inflation forecasting. Similarly, for policymakers, the results offer tools to develop more responsive and informed monetary policies, particularly during times of economic uncertainty. Finally, this thesis demonstrates the efficiency of combining the big data sources, particularly Google trends, with traditional economic indicators to attain more accurate and timely forecasts.

Here is the structure of the thesis: Chapter 1 overall introduces the research question of the thesis, and Chapter 2 makes the literature review, focusing on the integration of unconventional data sources into economic forecasting. Chapter 3 outlines the methodology and data collection process, including the preprocessing of

Google Trends data. Chapter 4 gives results of tests and regressions, evaluating the performance of various models and their policy implications. Finally, Chapter 5 is the conclusion part.

2. Literature Review

Many traditional models of macroeconomic forecasting have been improved through big data sources. Indeed, a few years ago such forecasters used simple econometric models primarily based only on a few aggregate data, whereas at present forecasters utilize complex sets of indicators that contain large volumes of information. With the development of the different economic structures there are more concerns about the real-time and higher frequencies data. More conventional econometric models, which usually require quarterly or annual data only, are now complemented by non-traditional indicators, including the volume of internet searches and social networks activity. Of these tools, Google Trends has risen to the level of a macroeconomic forecasting tool that provides real-time analysis of trends in the sentiment, behavior and expectations of the public (Choi and Varian, 2012).

Big data is especially helpful in developing economic models when there is not enough readily available reliable information on the emerging economies like Kazakhstan. Scholars and policy makers are trying to figure out how they can make more accurate and timely forecasts of critical measures of economic activity. This

chapter reviews the body of literature on the use of Google Trends for economic forecasting, with a particular focus on inflation forecasting, and identifies the gaps that this study aims to fill.

As a trend analysis, Google Trends, which monitors the relative popularity of search terms over time, has proven useful for predicting a range of macroeconomic factors. Choi and Varian (2012) were the first who used Google Trends data for economics predictions. It turned out that the use of keywords such as unemployment, consumer mood, and retail sales could serve as valuable leading indicators, often predicting economic shifts before they were reflected in traditional statistics. Their work can be credited for the expansion of the use of internet search data to forecast models especially in fields like labor markets and consumer behavior.

Varian (2014) built on this by explaining how Google search data could reveal patterns and trends that are not apparent via traditional approaches. This was especially important during periods of economic uncertainty, when traditional data might lag behind the actual economic changes. When search data was incorporated into the forecasting models, researchers were able to capture near real-time shifts in consumer expectations and behaviors, offering a timelier alternative to traditional economic indicators.

Similar findings have been established by other researchers to support the prediction capability of Google Trends. For example, Preis et al. (2013) showed that Google trends data significantly enhanced forecasting of stock market dynamics especially during periods of volatility. Further to the research, Goel et al. (2010) supported the use of search data in financial markets by proving the existence of relationships between stock prices and volumes of search and trading. In a similar vein, Askitas and Zimmermann (2009) have used Google Trends to monitor job search patterns during adverse economic conditions, thus underlining the importance of search data in the context of changes on the labor market. All these studies highlight the power of Google Trends data in predicting various economic indicators across different sectors.

Additionally, forecasts of inflation are important for central banks and other policy makers as they will direct monetary policy and financial stability and shape the public's understanding of inflation. Nonetheless, inflation forecasting is a difficult task since data are often scarce, delayed, or even sometimes unreliable, particularly for emerging economies and because economic systems are complex. Such conventional indicators include GDP growth, exchange rates and the prices of commodities, most of which are provided only on a quarterly or annual basis. At the same time, Google Trends provides more frequent data that can reveal more specific short-term changes in the focus of people's concerns, behaviors, and expectations.

On the same premise, Hassani et al. (2019) incorporated Google Trends into machine learning models for inflation forecasting. Surprisingly, the authors discovered that information from search queries drastically lowered the forecasting errors especially during the economic uncertainty. Their study highlighted that real-time search data could enhance the forecasting power of traditional econometric models, especially in periods of high inflation volatility. In the same line of thought, Ayivodji (2024) showed the effectiveness of Google Trends data in estimating inflation in developing economies, especially where the data reporting is less timely and reliable. Ayivodji then demonstrated how by using key words connected to inflation and economic conditions, Google Trends could help mitigate the lag in traditional datasets, providing policymakers with more up-to-date insights.

However, there are still some issues left for future research in the following context. Accordingly, many of the prior works are concentrated on the advanced economies while there is very limited research on the emerging markets such as Kazakhstan. This research will therefore try to cover this gap by using Google trends data in inflation forecasting for a country such as Kazakhstan where traditional techniques suffer from delays and missing information. Furthermore, as Google Trends has been utilized in different studies to make predictions on inflation, limited research has systematically integrated numerous search terms linked with economic factors including “inflation”, “GDP” and “exchange rate”. More notably, this study

aims to overcome that limitation by investigating the net impact of the aforementioned search terms on inflation forecasting precision. Lastly, many papers have addressed the issue of how many indicators the model can predict independently, rather than how to incorporate these indicators simultaneously into the model. This research helps to address this issue by developing an improved model of Google Trends coupled with several economic variables.

Therefore, the use of Google Trends data when making economists' inflation forecasts can be considered promising. Real time, high frequency data allows policies to come up with accurate and timely decision making because it has more accurate data. Following the literature review, the current research adds to the body of knowledge by employing Google Trends data to forecast inflation in Kazakhstan and by incorporating various economic variables in a single model. This research hopes to contribute to the existing literature by identifying four significant areas that deserve further attention and by proposing a set of guidelines for utilizing unconventional data sources in economic forecasting.

3. Data and Methodology

The main methodology includes the construction of the baseline model using the standard macroeconomic variables and enhanced model using Google trends of terms such as inflation, GDP and exchange rate. Google trends provide real-time search interest data as it is aforementioned, which adds predictive power to the baseline model using conventional data. The methodology used in this study includes a collection of Google trends using the pytrends in Python, and it is followed by preprocessing, adjustment and smoothing in order to make it consistent and suitable for further analysis. Google trends relative interest search score has been downloaded for both languages (English and Russian) in order to have broad language coverage of the data, and these terms are also smoothed using the monthly average in order to reduce the variability across data.

First, let's describe the data sources from each model: the baseline and enhanced model. The baseline model relies on monthly data spanning from 2013 to 2024. The key macroeconomic variables included in this model are:

1. **Inflation**: Monthly inflation data derived from national statistics, serving as the target variable for forecasting.
2. **GDP**: Year-over-year growth rates for Kazakhstan, available on a monthly basis.

3. **Exchange Rates**: The Nominal Effective Exchange Rate (NEER) index, along with its percentage changes.

4. **Brent Oil Prices**: Monthly average prices of Brent crude oil.

5. **Inflation Expectations**: Survey-based data collected by the National Bank of Kazakhstan, available starting from 2016.

$$\boldsymbol{Inf_t = \beta_0 + \beta_1 \times GDP_t + \beta_2 \times Inf\_exp_t + \beta_3 \times Exchangerate_t + \beta_4 \times Brent_t + \varepsilon_t} \quad \textbf{(1)}$$

These traditional variables form the core structure of the baseline forecasting model. Now the turn of the enhanced model, which extends information from Google Trends that will offer higher frequency and real time data than the standard economic indicators. Due to Kazakhstan's linguistic diversity, the bilingual approach is used for the data collection of Google search results. Both English and Russian key terms regarded economic activity to combine broad results and cover the public interest and behavior.

Initially, Google trends has been downloaded only for English terms. Excluding the Russian language data would be a big mistake because most of the search volume is happening there. Initial analysis of English search terms revealed several periods where Google Trends reported no search interest, highlighting the need for the inclusion of Russian search data to ensure full coverage.

First, the key search terms related to economic activity, such as “inflation” (“инфляция” in Russian), “GDP” (“ВВП”), and “exchange rate” (“обменный курс, курс доллара”), were identified. These terms were selected based on their relevance to the study and alignment with common economic discussions. As it is mentioned above, separate Google Trends datasets were downloaded for both English and Russian search terms from 2013 to 2024 using pytrends library. The data collection process was standardized to ensure comparability between the two datasets. And the main point here is to calculate the monthly averages to reduce high-frequency fluctuations and ensure consistency across both datasets. These averages (3-month MA) were further smoothed to avoid outliers or missing data that could distort the analysis. Following this, for each economic term, the search interest scores for both English and Russian terms were averaged. This combined data ensures that search activity from both linguistic groups is captured, while minimizing potential noise or gaps in data for any single language. Finally, there is a monthly combined Google trends dataset for the period 2013-2024 ready for incorporation into the baseline model.

$$Inf_t = \beta_0 + \beta_1 \times GDP_t + \beta_2 \times Inf\_exp_t + \beta_3 \times Exchangerate_t + \beta_4 \times Brent_t + \beta_5 \times Inf\_trends_t + \beta_6 \times GDP\_trends_t + \beta_7 \times Exchangerate\_trends_t + \varepsilon_t \quad (2)$$

In order to optimize the predictive models and capture meaningful variations in the data, the following transformations were applied: standardization, lagging, percentage changes. All variables, including both traditional and big data, were normalized to ensure comparability. Additionally, one-month and three-month lags were introduced for certain variables to account for delayed effects on inflation.

In terms of the flow of analysis, there were several models evaluated using key metrics to assess their prediction accuracy and robustness. After getting all the data stored, and models ready to analyze and compare, the following techniques and models were applied in order to investigate any improvement:

- **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**: Following metrics were calculated to evaluate the accuracy of the predictions and compare errors between different models.
- **Likelihood Ratio Test**: This test was used to compare the goodness-of-fit between the baseline and enhanced models, assessing whether the inclusion of Google Trends data improves the model's predictive capability.
- **Mincer-Zarnowitz Regression**: This regression was performed to test forecast bias, ensuring that the predictions made by the models were unbiased and reliable.

- **Diebold-Mariano Test**: Following test was used to compare the forecast accuracy between the baseline and enhanced models, determining whether the inclusion of Google Trends data provided statistically significant improvements.

In order to capture nonlinear relationships and complex interactions in the data, advanced machine learning methods were employed:

- **Random Forest Regressor**: This model was used to assess feature importance and improve predictions by leveraging a non-linear approach.
- **Gradient Boosting Regressor**: Applied for the enhanced model, this method was optimized through hyperparameter tuning to achieve better performance.
- **LASSO Regression**: LASSO was used for feature selection, helping to identify the most predictive variables and improve model generalization.

While this methodology integrates multiple data sources and advanced techniques, there are several limitations that appeared during the analysis. For instance, Google Trends data is only available from 2013 onward, restricting the analysis period. Survey-based inflation expectations data, collected by the National Bank of Kazakhstan, is only available starting from 2016, limiting the time span of this variable. And also, there were occasional gaps in Google Trends data for both English and Russian search terms. In cases where data was missing for either

language, the available data from the other language was used to minimize information loss.

4. Results and Discussion

4.1 Likelihood-Ratio Test and Evaluation Metrics

After data preprocessing and cleaning, the needed evaluations and tests are done in order to prove the target of the thesis. The first test conducted to compare the performance of the baseline and enhanced models was the Likelihood Ratio (LR) Test, which assesses the goodness-of-fit between the two models. The test statistics and corresponding p-value are presented in Table 1.

Table 1: Likelihood-Ratio Test for Model Comparison

Test Statistic	Value	Degrees of Freedom	p-value
Likelihood-ratio (LR) Chi^2	70.18	3	0.0000

$$\mathbf{LR} = -2 \times (ln(L_{baseline}) - ln(L_{enhanced})) \qquad (3)$$

The LR Chi^2 statistic of 70.18 with 3 degrees of freedom results in a p-value of 0.0000, which is highly significant. This indicates that the enhanced model, which incorporates Google Trends data, provides a significantly better fit to the

data compared to the baseline model. The low p-value suggests that the inclusion of Google Trends search data for terms such as "inflation," "GDP," and "exchange rate" notably improves the model's ability to predict inflation in Kazakhstan.

The next step is to compare the evaluation metrics in order to witness the increase in predictive power of the enhanced model. To assess the predictive accuracy of both the baseline and enhanced models, the Mean Squared Error (MSE) and Mean Absolute Error (MAE) were calculated. These metrics provide a clear understanding of the models' performance in terms of average prediction error in Table 2.

Table 2. Model Comparison: Performance Metrics (MSE & MAE)

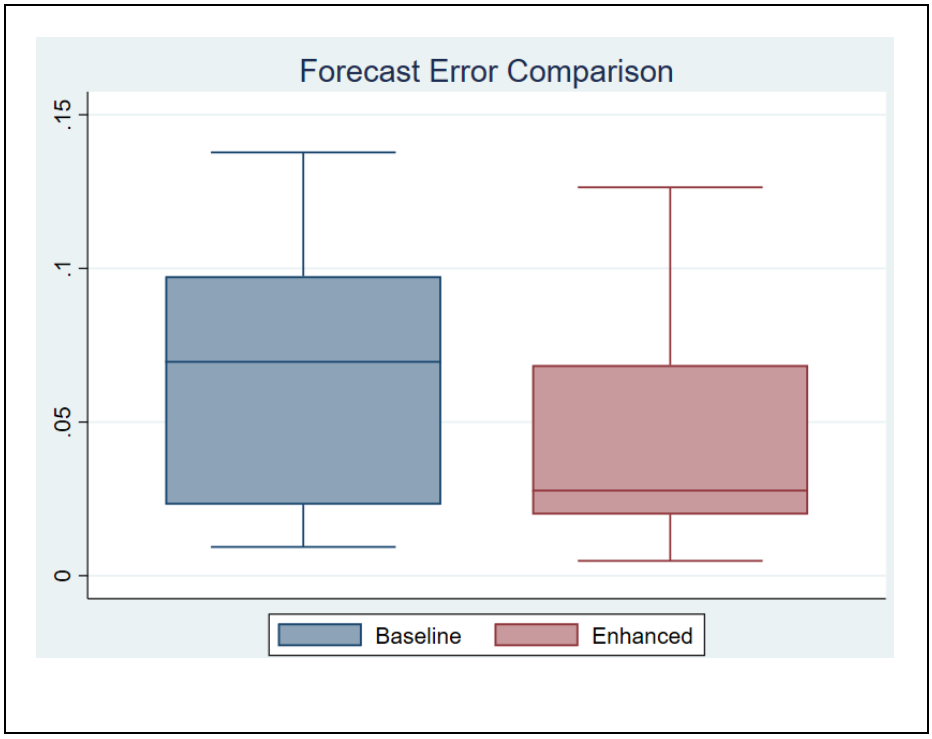
	MSE		MAE	
Metric	**Baseline Model**	**Enhanced Model**	**Baseline Model**	**Enhanced Model**
Mean	0.0058	0.0031	0.0648	0.0448
Standard Deviation	0.0056	0.0042	0.0405	0.0341
Minimum	0.0001	0.0000	0.0094	0.0048
Maximum	0.01897	0.01597	0.1377	0.1264

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(inf_i - inf_{pred})^2 \qquad \mathbf{(4)}$$

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|inf_i - inf_{pred}\right| \qquad \mathbf{(5)}$$

From Table 2, it can be stated that the enhanced model shows better results than the baseline model in both evaluation metrics. Both values of Mean Squared Error (0.0058<0.0031) and Mean Absolute Error (0.0648<0.0448) are lower for the enhanced model. Additionally, standard errors for both evaluation metrics are reduced for enhanced models and have lower variability in inflation prediction, which is another reason to claim that enhanced models provide relatively more accurate and stable predictions. It can visually be seen from Figure 1, there is a clear comparison of Absolute Errors of each model.

Figure 1. Forecast Error Comparison



4.2 Mincer-Zarnowitz Regression Results

The Mincer-Zarnowitz Regression was conducted to test for forecast bias between the baseline and enhanced models. This regression is essential for assessing whether the predicted values from each model systematically deviate from the actual values. The results are presented in Table 3 below.

Table 3: Mincer-Zarnowitz Regression Results

Model	**Coef.**	**Std. Err.**	Intercept	**t-value**	**P-value**	**R-squared**
Baseline Model	3.257	1.739	0.102	1.87	0.072	0.1189
Enhanced Model	0.987	0.326	0.044	3.03	0.005	0.2612

$$\boldsymbol{inf_t = \alpha + \beta \times inf_{pred} + \varepsilon_t} \qquad \textbf{(6)}$$

For the baseline model, the coefficient of the out sample predicted values is 3.257, with a standard error of 1.739. The T-value is 1.87, and the p-value is 0.072, which is significant at the point of 10%. The R squared of the model suggests that only 11.89% of the inflation dynamics can be explained by the baseline model.

In terms of the enhanced model, which incorporates Google Trends data, performs significantly better in the Mincer-Zarnowitz regression. The coefficient of the out of sample predicted values is 0.987, with a standard error of 0.326, and a t-

value of 3.03, which is highly significant with a p-value of 0.005. The p value illustrates that it is marginally significant at 1% significance level. This indicates that the enhanced model's forecasts are much closer to the actual inflation values, with minimal bias. And as can be seen from the R-squared value, the ability to explain the inflation dynamics increased to 26.12%.

When it comes to the interpretation of the coefficients, the predicted value coefficient equaling to 1.0 indicates a perfect prediction model (i.e., the model perfectly predicting the actual values) and intercept having 0 means that there is minimal bias in predictions. As can be seen from regression results of the enhanced model, it moved toward the ideal values, which is way much better than the baseline model's one. For the predicted values coefficient, it has 0.987, and intercept is 0.044.

Right after this regression model, a Diebold-Mariano (DM) test was conducted in order to assess whether the enhanced model significantly improves the predictive accuracy compared to the baseline model and to check the statistical robustness of the model. This test evaluates the differences in forecast accuracy by comparing the squared prediction errors of the baseline and enhanced models. With a p-value of 0.0000, the results are highly significant, allowing us to reject the null hypothesis (H_0: mean = 0). This means that the difference in prediction errors between the two

models is statistically significant, and the enhanced model demonstrates superior performance.

4.3 Lasso Regression Results

In this section, the results of the Lasso regression model applied to predict inflation are presented. Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization technique that not only helps to mitigate multicollinearity issues but also performs automatic variable selection by shrinking the coefficients of less important predictors to zero (Tibshirani, 1996). This characteristic is particularly useful in macroeconomic forecasting models where many variables may interact in complex, nonlinear ways. The model's optimal regularization parameter (α\alphaα) was selected through 10-fold cross-validation.

The Lasso regression model was trained with the following independent variables: GDP, inflation expectations (inf_exp), Brent crude oil price (brent), nominal effective exchange rate (neer), standardized inflation (inflation_standardized), standardized GDP (gdp_standardized), and standardized exchange rate (exchange_rate_standardized). Mainly, all the variables are used including Google trends. The coefficients of the model, along with their respective variables, are shown in Table 4 below:

Table 4: Lasso Regression Coefficients

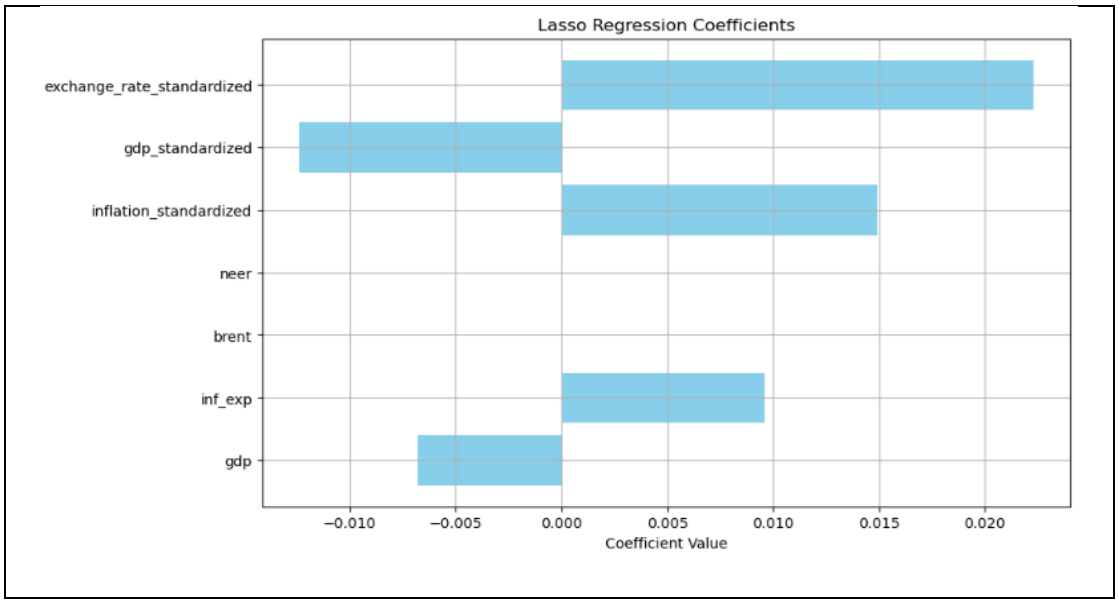
Variable	Coefficient
GDP	-0.006783
Inflation expectations	0.009587
Brent Crude Oil Price	0.00
Nominal Effective Exchange Rate	0.00
Inflation Google trends	0.014885
GDP Google trends	-0.012384
Exchange rate Google trends	0.022276

$$\boldsymbol{\beta = argmin_{\beta_0,\beta_n}(\sum_{i=1}^{n}((inf_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}))^2 + \alpha\sum_{j=1}^{p}|B_j|)} \quad \textbf{(7)}$$

Where, x_{ij} is the value of the predictor variable j for observation i, B_j are the regression coefficients, α is the regularization parameter controlling the penalty strength.

Overall, as can be seen from Table 4, the coefficients of GDP, Inflation expectations, and all three incorporated Google trends terms play a significant role in explaining the variation in inflation rates. And the significance of each component can be seen in Figure 2, where Google trends have relatively higher importance features.

Figure 2. Lasso Regression Coefficients



One of the main aims of showing these coefficients under the Lasso Regression is to show the importance of Google Trends variables along with traditional ones after the shrinkage. Figure 3 and 4 below shows the effect of the Lasso model, the predicted values of inflation become relatively close to the ideal fit, the dispersion of points is reduced. In terms of residuals, it shows normal distribution, which indicates that the model has captured the underlying trend without systematic bias.

Figure 3. Comparison of Lasso Model prediction with Actual values Figure

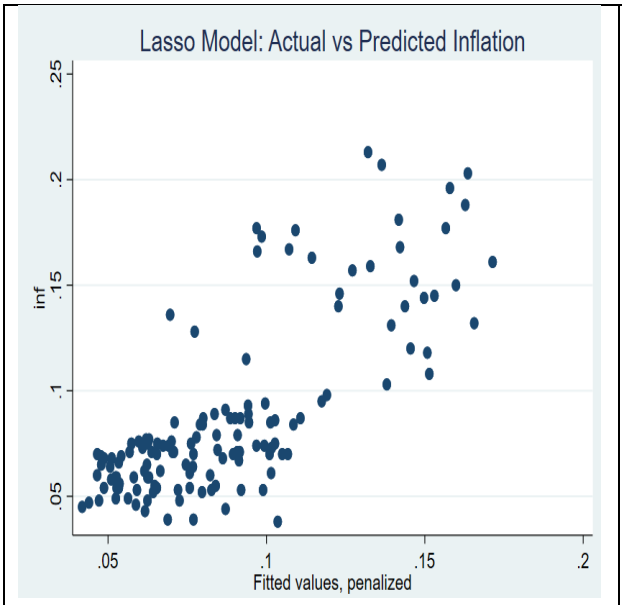
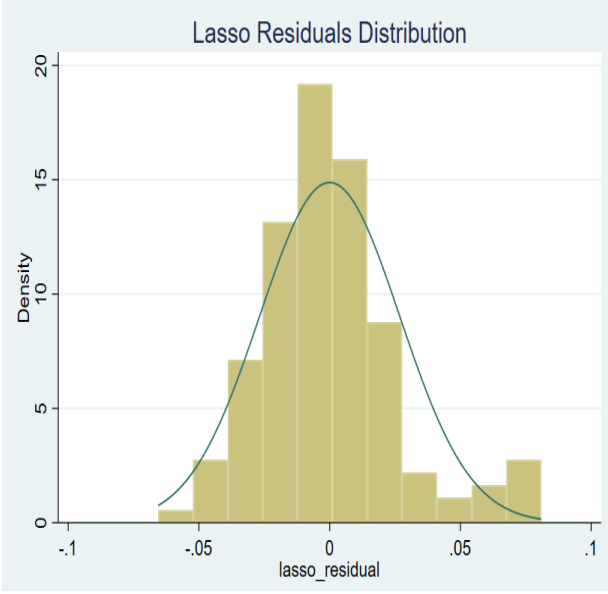


Figure 4. Lasso Model's Residuals Distribution



4.4 Random Forest and Gradient Boosting

Following the steps of Lasso regression, this study also uses other machine learning techniques like Random Forest and Gradient Boosting in order to capture the nuances of inflation forecasting and evaluate model robustness. The inclusion of these two methods is not arbitrary; each algorithm offers unique strengths, and together, they provide a comprehensive evaluation of the predictive potential of the models. The table below summarizes the MSE and MAE for both the Random Forest and Gradient Boosting models in their baseline and enhanced forms.

Table 5: Model Performance Evaluation

Model	MSE	MAE
Random Forest (Baseline)	0.000621	0.016043
Random Forest (Enhanced)	0.000262	0.012155
Gradient Boosting (Baseline)	0.000780	0.019446
Gradient Boosting (Enhanced)	0.000181	0.008846

Using these two machine learning methods, the evaluation metrics of MSE and MAE can also be compared in order to witness the improvement under the enhanced model. From Table 5, it can be noted that for both techniques there is a significant decrease in both MSE and MAE. For Random Forest, the MSE decreased from 0.000621 to 0.000262, while the MAE decreased from 0.016043 to 0.012155. In terms of Gradient Boosting, the values of MSE and MAE declined from 0.000780 and 0.019446 to 0.000181 and 0.008846 respectively after adding Google trends into the baseline model. From both Figure 5 and 6, it can clearly be seen that the enhanced model outperforms the baseline model in terms of inflation predictions

Figure 5. Random Forest

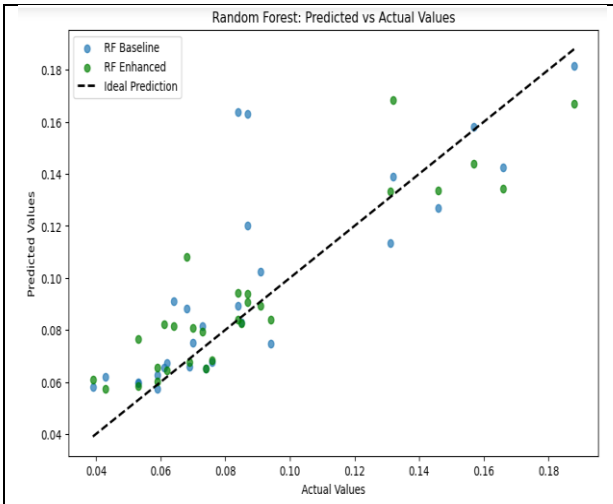
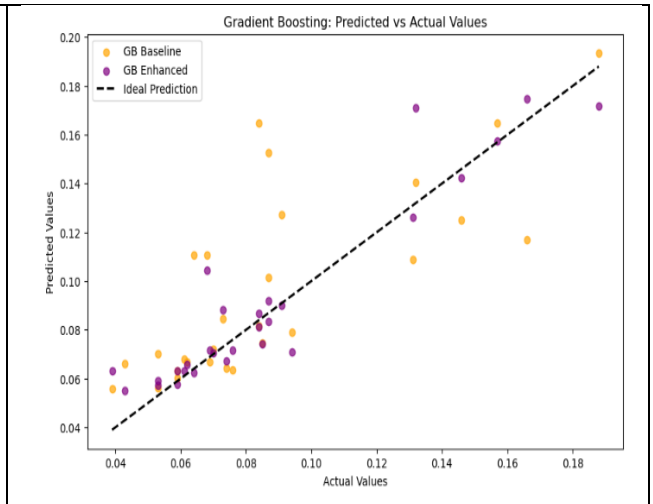


Figure 6. Gradient Boosting

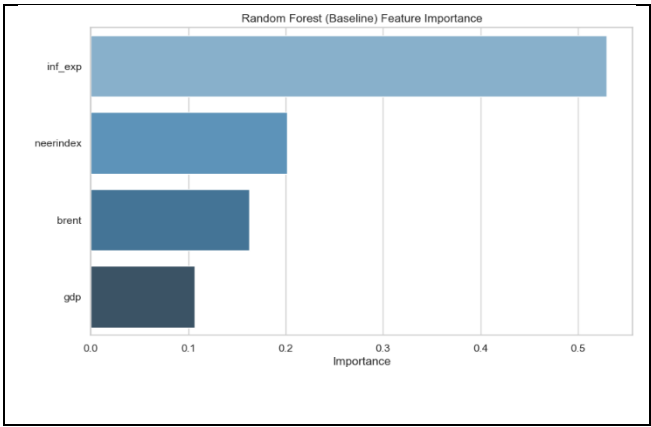


4.5 Feature Importance

Alongside Lasso Regression, there are feature importance reports of each technique for both models below in order to see to what extent Google trends are the important variables in explaining the inflation dynamics.

Figure 7 and 8: Feature Importance Graphs of Baseline Model

Random Forest



Gradient Boosting

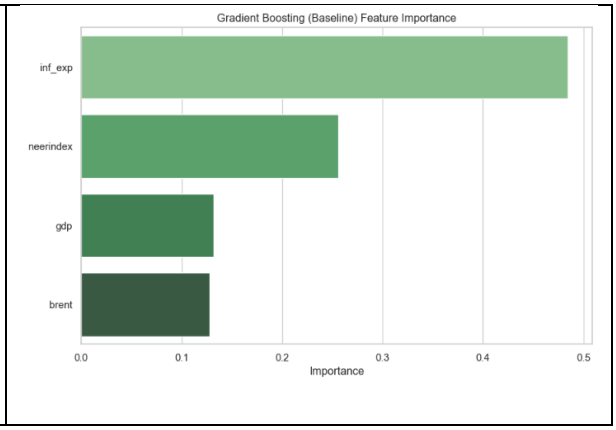
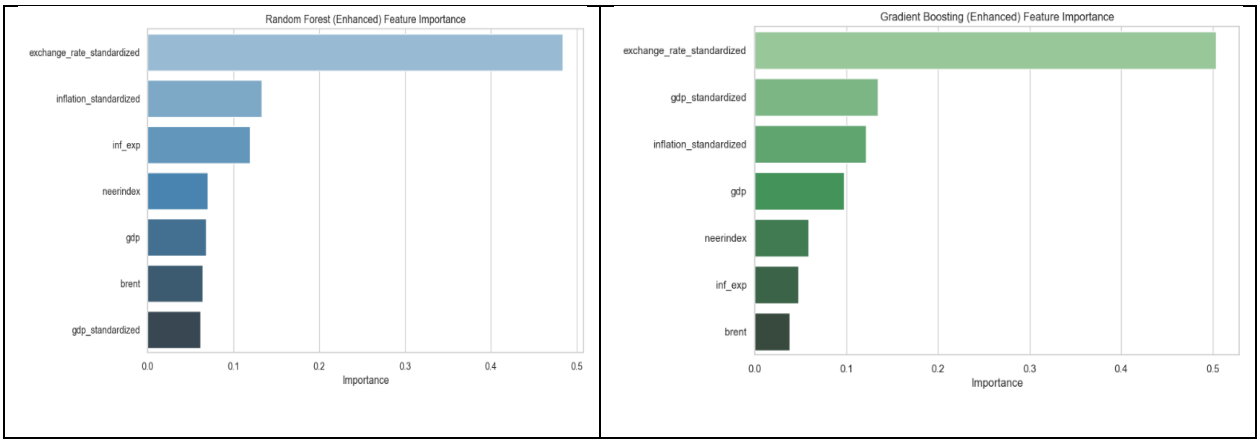


Figure 9 and 10: Feature Importance Graphs of Enhance Model

Random Forest **Gradient Boosting**



As can be clearly observed from figures 9 and 10 that the incorporation of Google trends into traditional data increases the predictive power of inflation and they are the main important features of this forecasting model. From both machine learning techniques, google trends are on the top of the rank in terms of importance. In both models the Google trend for term exchange rate is the most important feature with an importance score of 0.4839 for Random Forest, and 0.5036 for Gradient Boosting.

5. Conclusion

Overall, the thesis demonstrates the power of the Google trends and explores the integration of them with traditional macroeconomic variables to develop the inflation forecasting model in Kazakhstan. The findings aforementioned illustrate that the real-time and high-frequency data downloaded from Google trends significantly improves the accuracy of inflation predictions by reducing the forecast errors, especially in emerging markets where traditional data suffers from delays and gaps. By incorporating Google Trends data to the baseline model, where the main data consists of traditional macroeconomic indicators, it consistently shows that the enhanced model outperforms the first one.

The Likelihood Ratio Test, MSE and MAE evaluation metrics comparison, and Mincer-Zarnowitz regression results all indicate that the inclusion of Google Trends data leads to more reliable and precise forecasts. Furthermore, the application of machine learning techniques such as Random Forest, Gradient Boosting, and Lasso regression reveals that Google Trends significantly influences forecasting performance, and it can be seen from important features that Google trends data takes the significant portion of explanation in variation of inflation.

Despite the derived results from statistical tests, it should be admitted that there are certain limitations, including the restricted availability of data prior to 2013 in

terms of downloading Google trends in the region of Kazakhstan. Also, the sentiment analysis of social media and google news were excluded due to the reason for the absence of access tokens to social media like Facebook, and many occasional gaps in Google news scraping, which as a result did not give a strong improvement in the enhanced model alongside with Google trends. Due to these limitations the social media and news sentiment analysis were excluded from the thesis research.

The initial aim was to bridge the gap between traditional economic indicators and emerging digital data sources. So, this study makes a huge contribution to the role of big data in economic prediction of inflation primarily. In conclusion, the incorporation of Google trends data into macroeconomic traditional data enriches the opportunities of policymakers by offering timely and relatively accurate points to inflation movements in Kazakhstan.

6. Reference list

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *German Council for Social and Economic Data (RatSWD) Research Notes*, No. 41. https://doi.org/10.2139/ssrn.1480251

Choi, H., & Varian, H. R. (2009). Predicting the present with Google Trends. *SSRN*. https://doi.org/10.2139/ssrn.1659302

Choi, H., & Varian, H. R. (2012). Predicting the present with Google Trends. *The Economic Record*, *88*(s1), 2-9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-27. https://doi.org/10.1257/jep.28.2.3

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(41), 17486–17490. https://doi.org/10.1073/pnas.1005962107

Hassani, H., & Silva, E. S. (2015). Forecasting with big data: a review. *Annals of Data Science*, *2*(1), 5-19. https://doi.org/10.1007/s40745-015-0029-9

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, *3*, 1684. https://doi.org/10.1038/srep01684

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

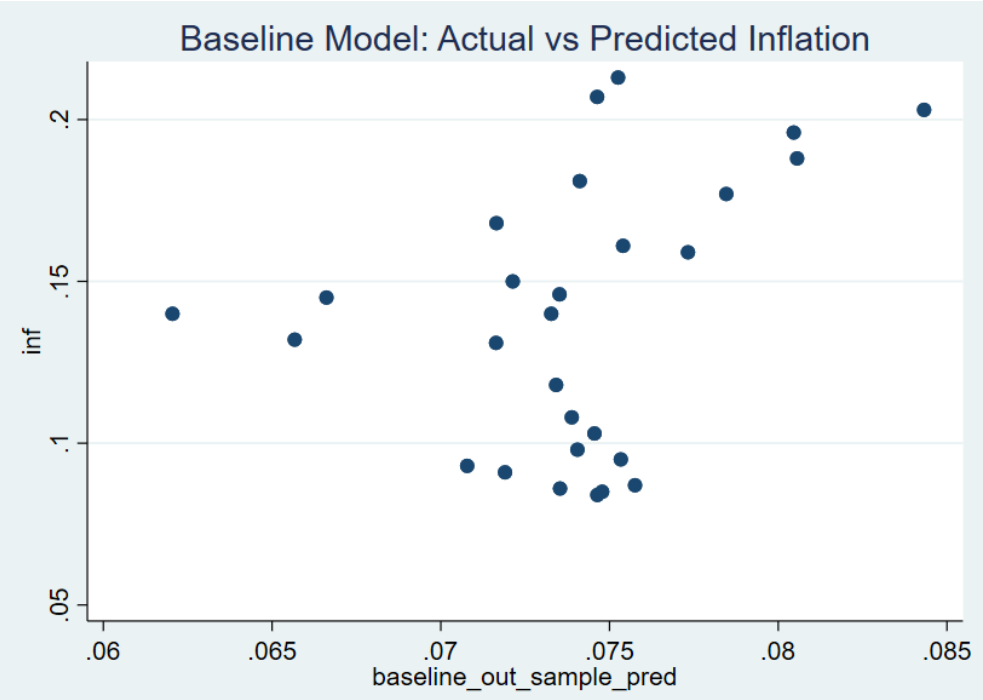
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

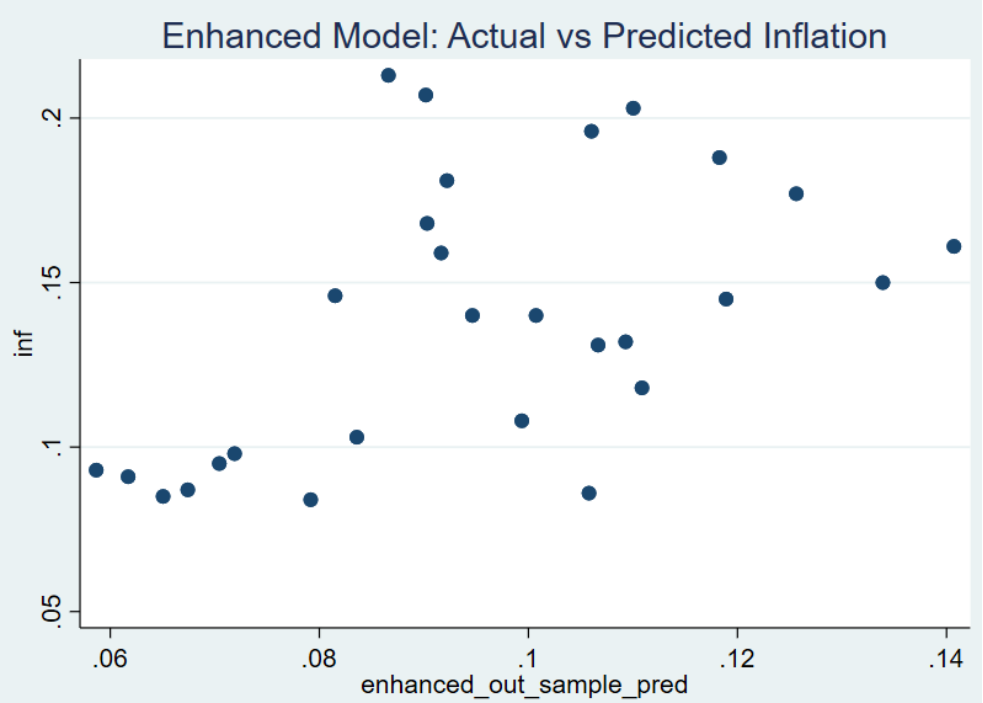
Ayivodji, F. (2024). High-frequency inflation expectations from big data: A natural language approach [Unpublished manuscript]. Université de Montréal and CIREQ.

7. Appendices

Appendix A



Appendix B



Appendix C

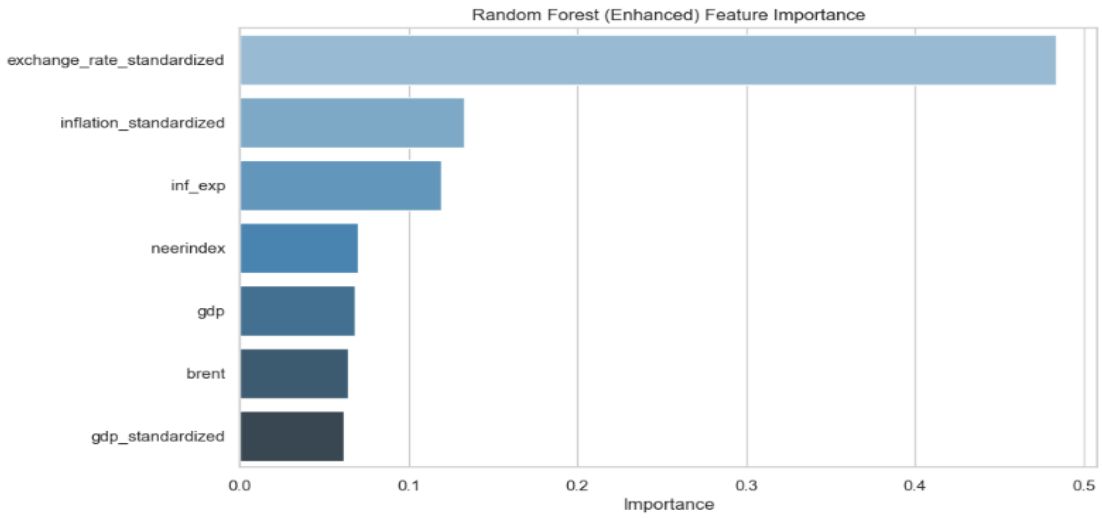
```
Grid value 68:    lambda = .0000545   no. of nonzero coef. =       7
Folds: 1...5....10   CVF = .0007877
Grid value 69:    lambda = .0000497   no. of nonzero coef. =       7
Folds: 1...5....10   CVF = .0007877
Grid value 70:    lambda = .0000453   no. of nonzero coef. =       7
Folds: 1...5....10   CVF = .0007877
Grid value 71:    lambda = .0000412   no. of nonzero coef. =       7
Folds: 1...5....10   CVF = .0007877
Grid value 72:    lambda = .0000376   no. of nonzero coef. =       7
Folds: 1...5....10   CVF = .0007877
... change in deviance stopping tolerance reached ... last lambda selected
Minimum of CV function not found; lambda selected based on stop() stopping criterion.

Lasso linear model                          No. of obs        =        137
                                            No. of covariates =          7
Selection: Cross-validation                 No. of CV folds   =         10
```

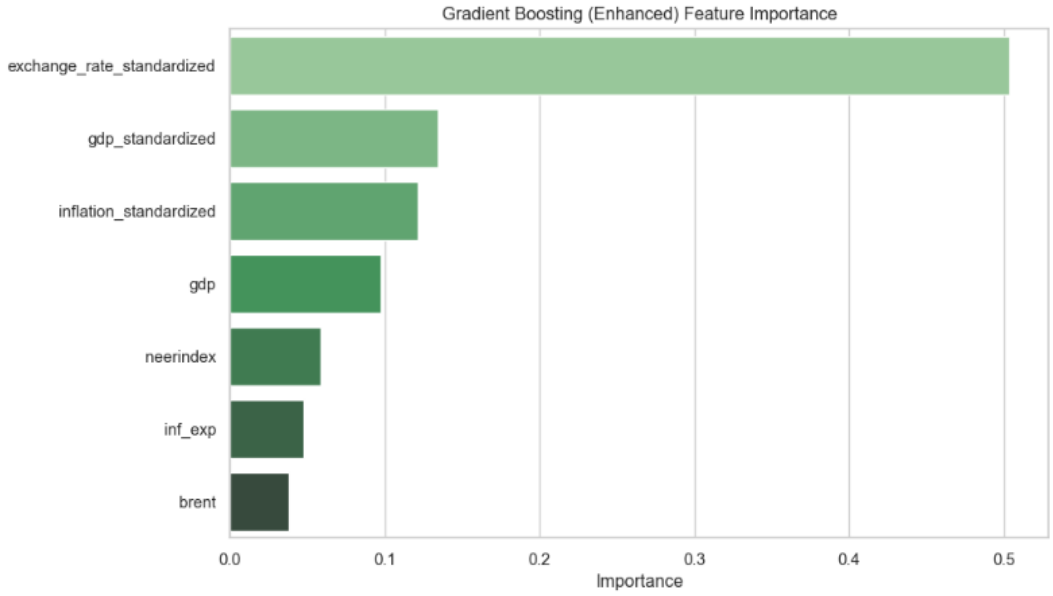
ID	Description	lambda	No. of nonzero coef.	Out-of-sample R-squared	CV mean prediction error
1	first lambda	.0277739	0	-0.0017	.0017797
71	lambda before	.0000412	7	0.5566	.0007877
* 72	selected lambda	.0000376	7	0.5566	.0007877

* lambda selected by cross-validation.

Appendix D



Appendix E


Gradient Boosting (Enhanced) Feature Importance
exchange_rate_standardized
gdp_standardized
inflation_standardized
gdp
neerindex
inf_exp
brent
0.0
0.1
0.2
0.3
0.4
0.5
Importance