Nowcasting Monthly Macroeconomic Variables in India Using Google Trends: A Machine Learning Approach

ABSTRACT

Timely availability of macroeconomic variables is essential for better policies. However, in India, official data are released with significant delays, and not all variables are published at a monthly frequency. This often leads to changes in macroeconomic conditions by the time policy decisions are made, which are based on previously published data. To overcome this challenge, this study examines the use of Google Trends data for nowcasting three key macroeconomic variables of India: unemployment, consumer price index (CPI), and GDP growth rate. Adopting the two-step approach of Woloszko (2020), we combine Google Trends data with machine learning models to evaluate their predictive performance. Results show that the choice of machine learning method is variable-dependent: unemployment is best predicted by Lasso regression, CPI by decision trees, and GDP growth by gradient boosting. We have included SHAP analysis to identify which features contribute most to the prediction of each variable. We have also analysed two major economic downturn periods, the 2008 Global Financial Crisis and the COVID-19 pandemic, to assess the model's ability to capture turning points. To check the robustness of our results, we analyse the correlation and regression of the predicted series with actual data, and also compare the model prediction with benchmark models. Overall, the results suggest that Google Trends can serve as a useful tool for nowcasting macroeconomic variables.

Keywords: Nowcast, Machine Learning, Macroeconomic variables, Google Trends Data, Correlation, Regression, Shapley Analysis

1. INTRODUCTION

Indian policy-making relies heavily on macroeconomic indicators. However, two key challenges limit their effectiveness. First, there is a significant publication lag: CPI data is released around 12 days after the reference month, unemployment data has a delay of about 45 days, and GDP growth rate figures are published roughly two months after the end of the quarter (CMIE, 2022; Ghate et.al, 2013). Second, we do not get these values at monthly frequency. Only CPI and unemployment figures are available monthly, while GDP growth is published quarterly.

In order to bridge this gap of delay in the publication of official data and frequency with which they are released, we need to estimate monthly data based on other available data. One way is to use survey data, but that has been challenged by studies such as Vermenleu (2012). According to this study, the underlying relationship between survey-based indicators and macroeconomic variables becomes unreliable when changes in economic activity are abrupt and massive. They are also associated with substantial cost and resource implications, along with certain conceptual and operational challenges in terms of sample representativeness, estimation techniques (Bai et.al, 2016).

A prior version of this paper was made available as a preprint on SSRN (DOI: 10.2139/ssrn.5620310) and Zenodo (DOI: 10.5281/zenodo.17383219).

An alternative approach is to leverage high-frequency datasets, usually referred as “big data”. Diebold (2003) was the first to use the term “big data”. IBM classify big data into four categories, which include volume, velocity, variety and veracity. According to Kapetanios et al. (2018), types of big data include traditional business system, social networks, and the internet of things. Big data involves dataset such as flight departures, satellite nighttime lights data, mobility reports based on anonymized personal data from Google and Apple, air quality indices, electricity consumption, credit card transactions, restaurant bookings, news-based indicators, etc. These alternative data sources have been increasingly used as a proxy for economic activity (Henderson et.al, 2012; Chen et.al, 2019; Chetty et.al, 2020). In this study, we have used one such high-frequency dataset, which is “Google Trends”, that is generated through the volume of searches people make. Google Trends is provided by Google in the form of relative searches.

Internet search behaviour can serve as an indicator of revealed expectations. People typically search for topics they are curious about or issues that concern them. For example, if someone want to buy bike, they might search for information related to bike, like which bike cost how much? what is the mileage? etc. Similarly, a person worried about rising household expenses and expecting an increase in prices might search for information on inflation. Also, if someone does not want to buy mobile phone, they probably would not search for related information. According to Guzman (2011), a rise in search volumes for the term “inflation” reflects public concern over potential price increases. Therefore, changes in the volume of search queries for inflation can be viewed as a proxy for the public’s expectations about future price movements. Previous studies have demonstrated that Google Trends can help forecast unemployment, private consumption, and food prices across different economies (Choi & Varian, 2012; Vosen & Schmidt, 2011; Seabold & Coppola, 2015).

According to Giannone *et al. (*2008), Nowcasting involves prediction estimates of the present, the near future, and the recent past. Mostbeck (2019), defined nowcasting as the real-time prediction of current economic conditions using high-frequency data before official statistics become available. It has emerged as a crucial tool for economic analysis and policy formulation. Nowcasting monthly macroeconomic variables can help address two challenges faced by the Indian economy:

i. Delay in publication, and
ii. Publication in different frequency

It will also be helpful in tackling other challenges of Indian economy such as complex economic structures and diverse regional variations. It can also support real-time policy decisions. Nowcasted values of macroeconomic variables could make government prepared to economic shocks. Government could respond to these shocks by adjusting policy according to the nowcasted values of macroeconomic variables.

In recent years, the application of machine learning techniques in macroeconomic forecasting has gained significant attention, particularly with the growing availability of alternative and real-time data sources such as Google Trends (Coulombe et.al, 2021; Medeiros et.al, 2021). It provides the timeliness and accuracy of macroeconomic variables forecasts. Traditional models

often rely on lagged economic indicators and assume linear relationships; hence they are unable to capture the dynamic nature and non-linear relationship of macroeconomic variables.

In developed economies, nowcasting is primarily focused on refining already robust statistical systems. But India is a developing economy and unlike developed economies nowcasting in India must address fundamental challenges such as data scarcity, irregular publication schedules, and the need to capture economic activity across formal and informal sectors. According to Agnihotri & Mishra (2021) due to the unavailability of variables and the volatility in the nature of macroeconomic variables, the prediction is harder for developing economies compared to stable and developed nations.

Another challenge in front of policy makers is the structure of Indian economy. It is dominated by informal sector, which is not captured in the official statistics. As per the PLFS data for 2021-2022, around 83% of India's labour force is engaged in the informal sector (NSSO, 2022), representing about 390 million workers. This includes domestic workers, home-based workers, casual labourers and unregistered micro-enterprises, whose activities are typically characterised by low documentation and limited visibility in administrative records. Because a large part of this informal activity is difficult to track through formal reporting systems, its contribution is only partially captured in official statistics. Hence, official GDP measures might not be an actual representation of growth. This creates substantial gaps in understanding current economic conditions.

This paper addresses this gap by combining Google Trends data with machine learning techniques to predict three core indicators: unemployment, GDP growth rate, and Consumer Price Index (CPI). Using a systematic keyword normalization strategy and Woloszko's (2020) two-step modelling framework, we demonstrate that search-based behavioural data can complement official statistics, offering granular, high-frequency insights for real-time policymaking. Rest of the paper is structure as follows: Section 2 contains the literature review, Section 3 deals with data collection, Section 4 covers the research methodology, and Sections 5 and 6 present the results and discussions, and conclusions, respectively.

2. LITERATURE REVIEW

In recent years, Google Trends have been widely utilized in studies aiming to predict macroeconomic variables. Notably, Choi and Varian (2009a) were among the pioneers who explored the potential of Google Trends for forecasting macroeconomic indicators. The rise of digital footprints, especially online searches, enables real-time economic analysis. Google Trends offers timely, high-frequency, and location-specific search data. In contrast to traditional statistics that come with significant delays, Google Trends reflects public sentiment almost instantly. This makes it useful for nowcasting macroeconomic indicators. Researchers have recognized its ability to provide early signals and complement traditional data sources.

Choi and Varian (2009a) explore the application of Google Trends data for nowcasting economic activity, illustrating how search query information can help estimate home sales, retail sales, automotive sales, and travel trends. In a subsequent study, Choi and Varian (2009b)

demonstrate that Google Trends can also be used to forecast initial unemployment insurance claims. Similarly, some scholars have also shown that google search data can be used for predicting unemployment, for instance D'Amuri and Marcucci (2010) have forecasted unemployment for US, Suhoy (2008) for Israel, Askitas and Zimmerman (2009) for Germany, Anvik and Gjelstad (2010) for Norway, and D'Amuri (2009) for Italy. Suhoy (2010) have used Google Trends data to nowcast private consumption for Israel. Vosen and Schmidt (2011) forecasted consumption for the US using Google Trends data. This study also shows that Google search query data outperforms survey-based indicators. Kulkarni et al. (2009) applied Google Trends data for predicting housing price, Constant and Zimmermann (2009) have predicted politics and social trends, Song et al. (2009) have used it in tourism sector and similarly Goel et al. (2010) in entertainment sector.

The application of Google Trends data specifically to the Indian economic context remains limited but promising. Ahani et al. (2021) has demonstrated that Google search data can improve forecasting accuracy for various economic indicators, including unemployment and consumer confidence. The methodology of using search query statistics as economic indicators has shown particular promise during crisis periods when traditional data sources may be less reliable. Pratap and Priyaranjan (2023) have predicted GDP growth rate of India using Google trends. This study used MIDAS model. Jha and Sahu (2020) used hybrid New Keynesian Phillips curve along with google search data for forecasting inflation in India. Similarly, Samanta (2019) use "price" and "inflation" keywords for predicting inflation rate in India. Some studies which focused on particular macroeconomic variable forecasting are listed below:

2.1 Inflation Forecasting

Google Trends is widely used for inflation nowcasting. Choi & Varian (2009) found that automotive-related searches improve regression models for inflation, benefiting central bank monitoring. Guzmán (2011) created the Google Inflation Search Index (GISI), which can lead actual inflation by up to 12 months, surpassing survey and market measures.

India-focused studies include Samanta (2019), who found that searches for "inflation" and "price" predict CPI-based inflation rates, and Jha & Sahu (2020), who used Google Trends in a Hybrid New Keynesian Phillips Curve model. Their results show that internet-derived inflation expectations enhance forecast accuracy over autoregressive and survey methods. These studies suggest that online inflation expectations are forward-looking, timely, and adaptable.

2.2 Unemployment Nowcasting

Unemployment prediction is another well-explored area. D'Amuri & Marcucci (2012) developed the "Google Index" (GI) using job-search queries, showing it predicts U.S. unemployment better than professional forecasts. Pavlicek & Kristoufek (2015) applied this to the Visegrad countries (the Czech Republic, Hungary, Poland and Slovakia), finding strong links between search data and unemployment, especially in the Czech Republic and Hungary.

Search data provides insights into job-seeking trends in real time, often preceding labour market statistics, which gives policymakers and agencies timely opportunities for intervention.

2.3 Tourism Demand Estimation

Siliverstovs & Wochner (2018) used a regional comparison approach with Google Trends to address normalization. Their model, applied to Swiss tourism, yielded accurate demand forecasts, showing that search data can support sectors needing advance planning. Tourism's sensitivity to shocks makes such forecasts valuable for regional management.

2.4 GDP Growth Tracking

Woloszko (2020) created the OECD Weekly Tracker, a neural network model using Google Trends and bias adjustment to estimate GDP growth for 46 countries. This method was especially helpful during the COVID-19 pandemic, when traditional indicators lagged. McLaren and Shanbhogue (2011) also found search data useful for tracking UK labor and housing markets. Their findings support Google Trends as a timely complement to GDP measurement, especially in volatile times. Grath (2025) use machine learning model to nowcast Swedish GDP growth. Heikkinen (2019) predicted GDP growth rate of Finland and Germany

While research confirms Google Trends can improve macroeconomic forecasts, most works cover only one or two variables, such as inflation or unemployment, or focus on specific sectors like tourism. Few try to unify multiple macroeconomic estimates with search data, especially in developing economies.

2.5 Research Gap

Despite broad use, most research focuses on developed economies or single-variable forecasting. There is little study of multi-variable, monthly estimation for emerging markets like India, especially using a consistent approach for inflation, unemployment, and GDP growth rate.

2.6 Research Objective

The objective of our study is to use Google Trends data for nowcasting monthly macroeconomic variables of India such as unemployment, Consumer Price Index (CPI), and GDP growth rate. Our goal is to systematically assess whether this approach provides robust and policy-relevant insights into these key indicators.

3. DATA COLLECTION

3.1 Google Trends Data

According to Guzmán (2011) and Seabold and Coppola (2015), each Google Trends series has two key characteristics. First, the values reported for different time points do not represent absolute search volumes for the selected keywords; rather, they are relative measures. The point of maximum search interest during the selected time frame is assigned a value of 100, and all other values are scaled in proportion to it. Second, the time series representation of a search index for a given keyword and period may vary depending on the date the query is made. This means that even if the reference period remains fixed, the reported data can change over time.

Keywords selection process for Google Trends

From google trends we have collected the monthly data for the period January 2004 to June 2025, for more than 170 keywords. These keywords were extracted based on 10 thematic categories, backed by existing literature and tailored to the Indian context: -

1. Labour Market & Careers: Inspired by studies such as D'Amuri and Marcucci (2012), which demonstrate that Google search data can serve as a powerful and reliable leading indicator for forecasting unemployment, using keywords like *"jobs"*, *"public jobs"*, *"California jobs"*, *"collect unemployment"*, and *"job centre"*. Similarly, Pavlicek and Kristoufek (2015) examined the relationship between unemployment rates and job-related Google searches for the Visegrad countries, using terms associated with *"job"* and *"work."* In line with these approaches, we included India-specific keywords such as *"job near me", "temporary jobs", "private employment agency", "employment news", "sarkari job", "naukri", "lawyer", "government jobs", "career search", "recruitment", "employment news", "sports", "developer jobs",* and general jobs search terms.
2. Housing, Construction & Real Estate: Inspired by McLaren and Shanbhogue (2011), which have explored the potential of using internet search data to nowcast economic activity in the United Kingdom, specifically for the labor and housing markets. The study found a significant correlation between search terms like "JSA" (Jobseeker's Allowance) and "unemployment", "unemployment benefit", "unemployed" and "foreclosure" and terms related to real estate. We have used keywords such as *"affordable housing", "house price index", "home insurance", "residential rentals", "home improvement", "housing bubble", "foreclosure", "real estate", "property", "apartments", "office space", "commercial building", "home loan", "mortgage", "flooring", "construction", "civil engineering", "maintenance", construction "consulting", "swimming pools",* and other related keywords.
3. Inflation & Price Level: Building on studies such as Samanta (2019), which examined the usefulness of Google search data for tracking and predicting inflation in India; Guzman (2011), who investigated the potential of Google Trends for forecasting inflation expectations; Jha and Sahu (2020), who incorporated internet search-based inflation expectations into a Keynesian Phillips Curve framework for India; and Choi and Varian (2012), who explored the application of Google Trends in forecasting inflation expectations and consumer sentiment, we include a range of inflation-related keywords in our analysis. These keywords include *"WPI", "CPI", "consumer price", "inflation", "onion prices", "petrol price", "food inflation", "gold rate today", "USD to INR", "rupee exchange", "tomato price",* among others, reflecting key items that influence public perception of inflation in the Indian context.
4. Finance and Economic Activity: inspired by Woloszko (2020), who employed Google Trends search data in combination with machine learning algorithms to capture the non-linear relationship between search behavior and GDP growth, using keywords such as *"bankruptcies"*, *"investment"*, and *"luggage"*. we incorporate India-specific keywords including *"bankruptcy", "financial crisis", "economic crisis", "recession", "unemployment", "public debt", "Sensex", "stock market", "Nifty50", "returns",*

"mutual funds", "FDI", "SIP", "Fixed deposite", "business news", "auditing", "accounting", "economy news", "economic policy", "exports", "imports", "trade deficit", "BOP", "export promotion", "custom duty", "remittances", "budget", "income tax filing", "income tax slabs", "budget", "direct tax", "EMI", "layoffs", and other related keywords.

5. Tourism, Travel & Hospitality: inspired by Siliverstovs and Wochner (2018), who investigated the predictive power of Google Trends data for economic activity, using keywords related to tourist destinations such as *"Graubünden"*, *"Zürich"*, and *"Wallis"*. We include search queries on *"tourist destinations", "travel", "travel agencies", "travel services", "accommodations", "spas", "vacation deals", "holiday packages", "hotel booking", "hotel", "flight tickets", "luggage", "rental and taxi services",* and other travel-related keywords.
6. Retail, Consumer Goods & Lifestyle: we included keywords that can reflect consumer spending and lifestyle. For these we have included keywords such as *"grocery", "food retailers", "electronics sale", "discount deals", "online shopping", "book retailers", "footwear", "accessories", "fast food", "tobacco products", "brands", "fashion and style", "home appliances", "brands", "beauty services", "wedding", "birthday", "fitness", "health",* and *"pharmacy"*.
7. Technology, Industry & Business Services: we have also collected the keywords related to technology, industrial activity, and professional services. These keywords include *"computer servers", "data management", "development tools", "outsourcing", "consulting", "corporate", "distribution and logistics", "printing and publishing", "food production", "manufacturing", "dyes", "pigments", "animal products", "services",* and *"agriculture"*.
8. Transport & Vehicles: Searches in this theme relate to mobility, logistics, and the automotive sector, including *"car", "commercial vehicles", "vehicle shopping", "vehicle licensing and registration", "vehicles", "freight", "trucking", "tractor", "transportation", "logistics", "package delivery", "house moving",* and *"aviation"*.
9. Arts, Entertainment & Events: Search activity in this group includes *"acting", "theatre", "performing arts", "events",* and other related keywords.
10. Miscellaneous Services: This catch-all includes *"mail", "GPS and navigation"*, "politics", "world news" and other such keywords.

Each Google Trends series is relative rather than an absolute measure of search volume. The time point with the highest search interest within the selected period is assigned a value of 100, and all other observations are scaled in relation to this peak. As a result, the data do not indicate the actual number of searches, and factors such as changes in internet penetration or variations in Google usage over time do not affect the interpretation of the series. Also, the terms themselves are not independently comparable with each other. But google trends provides us facility to compare the search intensity of up to five keywords. In any batch of five keywords which will be searched using Google trends, the keyword which is searched highest among them will get a value of 100 for that particular time period and rest will receive value related to the highest search for that keyword.

Normalization Strategy

In order to keep all the keywords' data relative to only one keyword, so that our overall data could become meaningful, we have used one anchor keyword in every batch of five keywords. we have used *"unemployment"* as anchored keyword. So, each batch of data that we have extracted contain *"unemployment"* and other four different keywords. The resulting data were stored in more than 45 spreadsheets, each containing four target keywords and one anchor. These spreadsheets were then organized into three groups based on the relative search frequency of the keywords compared with "*unemployment"*:

Group 1: Similar Scale to *"Unemployment"*

Keywords with search volumes close to "*unemployment"*. In this case, "*unemployment"* was retained as the anchor keyword, allowing the sheets to be merged directly using it as a common reference.

Group 2: Much Lower Search Volume than *"Unemployment"*

Keywords with significantly lower search intensity. For these, we employed *"tourist destination"*, a medium-volume keyword, as the reference instead of "*unemployment"*.

Group 3: Much Higher Search Volume than *"Unemployment"*

Keywords with substantially higher search volumes. For these, *"loan"* was used as the reference keyword.

In addition, two comparison sheets were created:

1. *"Unemployment" vs. "Tourist Destination"*
2. *"Unemployment" vs. "Loan"*

These comparison sheets enable rescaling across groups, allowing all keywords to be placed on a common scale. Ultimately, "*tourist destination"* serves as the final reference point, making it possible to integrate all keywords into a single, consistent dataset. To the best of our knowledge, this study presents one of the most comprehensive India-specific compilation of Google search activity for macroeconomic forecasting, spanning two decades and incorporating both conventional indicators and uniquely country-specific terms.

3.2 World Bank Data

In addition to Google Trends, we compile annual macroeconomic indicators from the World Bank's World Development Indicators (WDI, 2025) for the period 2004–2024. The variables include:

1. **GDP Growth Rate**: Annual percentage change in GDP at constant 2015 USD.
2. **Unemployment Rate:** Share of the labor force actively seeking but not engaged in work.
3. **Consumer Price Index (CPI):** Inflation measure tracking changes in a representative consumption basket.

These indicators provide the macroeconomic benchmarks against which Google search-based measures are evaluated, enabling a systematic comparison between conventional data sources and high-frequency digital indicators.

3.3 Rationale for Combined Dataset

The integration of Google Trends with World Bank data offers both timeliness and credibility. While official macroeconomic indicators such as GDP growth rate, CPI, and unemployment are authoritative, they are published with considerable delays and limited frequency. In contrast, Google Trends provides high-frequency, real-time information on public search behaviour that can serve as an early signal of economic conditions. Anchoring Google-based measures to established World Bank indicators allows us to test whether search intensity enhances forecasting performance beyond traditional data (Kohns & Bhattacharjee, 2023). This integrated approach enhances the robustness of the analysis and contributes to the growing literature on digital indicators in economic forecasting.

4. Research Methodology

Following Woloszko (2020), this study adopts a two-step forecasting framework which combine machine learning techniques with Google Trends data to predict monthly macroeconomic variables for India.

Step 1: Model Selection Using Annual Data

In the first stage, the monthly Google Trends data were aggregated to the annual frequency and merged with official macroeconomic variables obtained from the World Bank. We conducted exploratory analysis to examine the statistical properties of the dataset and performed feature engineering to construct suitable predictors. The dataset was then divided into training and testing subsets.

Several machine learning models, such as Random Forest, Gradient Boosting, K-Nearest Neighbors, AdaBoost, Lasso Regression, Decision Tree, and Linear Regression, were trained and evaluated. These algorithms are commonly applied in macroeconomic forecasting due to their ability to capture nonlinear relationships and high-dimensional patterns (Mullainathan & Spiess, 2017; Medeiros et al., 2021). Model performance was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), which are widely recommended in predictive modelling (Hyndman & Koehler, 2006). The model with the lowest prediction error was selected as the best-performing machine learning model for that macroeconomic variable.

Step 2: Monthly Forecasting Using Google Trends

In the second stage, the selected best- performance model was applied to the high-frequency monthly Google Trends, assuming that this relationship is frequency-neutral, to generate monthly predictions of macroeconomic variables.

To perform feature selection and mitigate overfitting in the high-dimensional Google Trends dataset, this study employs the Least Absolute Shrinkage and Selection Operator (LASSO). LASSO is a regularized regression technique that imposes an ℓ_1 penalty on the magnitude of regression coefficients, encouraging sparsity by shrinking less informative coefficients exactly to zero (Tibshirani, 1996). Formally, the LASSO estimator is obtained by solving the following optimization problem:

$$\hat{\beta}^{LASSO} = arg\ \min_{\beta}\left\{1/_{2n}\sum_{t=1}^{n}(y_t - x_t^T\beta)^2 + \lambda\sum_{j=1}^{P}|\beta_j|\right\} \quad (1)$$

where y_t denotes the dependent variable, X_t represents the vector of predictors, β is the parameter vector, and $\lambda \geq 0$ is the regularization parameter that controls the degree of shrinkage. By driving irrelevant coefficients to zero, LASSO performs simultaneous variable selection and estimation, making it particularly suitable for handling noisy, correlated, and high-dimensional predictors such as search query data (Hastie, Tibshirani, & Friedman, 2009; Molnar, 2022). This property enhances both the predictive accuracy and interpretability of the model.

After applying feature selection, we have applied different machine learning models on the selected features. Each model is evaluated using three standard error metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). We have selected that model, which is best suited for forecasting each indicator and using it we have predicted monthly values for those macroeconomic variables.

We subsequently analysed the SHAP (SHapley Additive Explanations) values to interpret the machine learning models. SHAP values are grounded in cooperative game theory and provide a consistent way of attributing the contribution of each feature to a model's prediction (Lundberg & Lee, 2017; Shapley, 1953).

Formally, given a machine learning model f and an instance $x \in R^M$ with M features, the Shapley value for feature i is defined as:

$$\phi_i(f,x) = \sum_{s\subseteq N\setminus\{i\}} \frac{|s|!(M-|s|-1)!}{M!}\left[f_{s\cup\{i\}}\left(x_{s\cup\{i\}}\right) - f_s(x_s)\right] \quad (2)$$

Where:

- N = {1, 2, …, M} is the set of all features,
- $s \subseteq N\setminus\{i\}$ represents a coalition (subset) of features not including i,
- $f_s(x_s)$ *is the model prediction using only features in subset S,*

The combinatorial coefficient ensures fair weighting across all possible coalitions (Molnar, 2022). Intuitively, $\phi_i(f,x)$ measures the marginal contribution of feature i to the prediction, averaged over all possible subsets of the p features.

The overall model prediction for an observation x can then be decomposed as:

$$f(x) = \phi_0 + \sum_{i=1}^{M}\phi_i(f,x) \quad (3)$$

Where ϕ_0 is the base value, i.e., the expected prediction across the dataset:

$$\phi_0 = E[f(x)] \tag{4}$$

Thus, the SHAP framework ensures that the prediction is additively explained by the contributions of all features.

In our study, we visualize SHAP values through three complementary plots:

a. SHAP summary plot, which provides a high-level view of how each feature influences the model's predictions across the entire dataset. The features are listed, ordered by their overall importance on the y-axis. The x-axis represents the SHAP value, which measures the impact (ϕ_i) of a feature on the model's output. A positive SHAP value increases the predicted macroeconomic variable, while a negative value decreases it. The colour of each dot in the plot indicates the feature's value for a specific data point. High feature values are represented by red dots, and low feature values are represented by blue dots. This plot reveals the relationship between feature values and the prediction of macroeconomic variables.

b. Bar chart showing the mean absolute SHAP value for each feature. This provides a more straightforward measure of global feature importance. The x-axis represents the average impact (magnitude) of the feature on the model output, regardless of the direction (positive or negative). This is calculated by taking the average of the absolute SHAP values for each feature across all data points:

$$MeanAbeSHAP(i) = \frac{1}{n}\sum_{J=1}^{n}\left|\phi_i\left(f, x^{(j)}\right)\right| \tag{5}$$

Where n is the number of observations. The features are listed ordered by their mean absolute SHAP value on the y-axis. This plot identifies the top predictors in the model.

c. SHAP force plot that explains a single prediction. This plot breaks down a specific prediction, showing how each feature contributed to moving the model output from the base value (ϕ_0) to the final predicted value $f(x)$. The base value is the average model output over the entire dataset. This is the starting point for the explanation. The plot represents two coloured bars- red bars, for features with positive SHAP values (which pushes the prediction higher). The length of the bar indicates the magnitude of the impact; and blue bars, for Features with negative SHAP values (which pushes the prediction lower). This plot explains why the model predicted the final prediction value.

These plots collectively enhance the transparency of our models by providing both global interpretability and local interpretability, which is particularly valuable for macroeconomic forecasting in the Indian context where policymakers require clear explanations of model predictions. we have also analysed the two economic downturn periods, 2008 Global Financial Crisis and Covid-19 pandemic.

We have also compared our predictions with the prediction of traditional benchmark models such as Random Walk with drift (RW-drift), AR (1), and OLS with lag and trend.

i. Random Walk with drift

The Random Walk with Drift model assumes that shocks to the series have permanent effects and that the series evolves according to a stochastic trend:

$$y_t = y_{t-1} + \mu + \varepsilon_t \quad (6)$$

Where ε_t is an independently and identically distributed error term with zero mean and constant variance, $\varepsilon_t \sim$ i.i.d. $(0, \sigma^2)$, y_t denotes the macroeconomic variable of interest at time t, μ captures the deterministic drift component, and ε_t is a serially uncorrelated innovation. This specification is widely regarded as a robust benchmark in macroeconomic forecasting, particularly for highly persistent series such as GDP, inflation, and exchange rates (Nelson & Plosser, 1982; Stock & Watson, 2007).

ii. First-Order Autoregressive Model (AR (1))

The AR (1) model captures short-run persistence by allowing current outcomes to depend linearly on their immediate past values:

$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t \quad (7)$$

$where \quad |\phi| < 1$, α is a constant term and ϕ measures the degree of persistence in the series. Autoregressive models form the backbone of empirical time-series analysis and have been shown to perform competitively in forecasting macroeconomic aggregates (Box & Jenkins, 1976; Hamilton, 1994).

iii. Lag–Trend Ordinary Least Squares (OLS) Model

To account for both stochastic persistence and deterministic long-run dynamics, we estimate the following OLS regression with a lagged dependent variable and a time trend:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 t + u_t \quad (8)$$

Where u_t is an independently and identically distributed error term with zero mean and constant variance, $u_t \sim i.i.d.(0, \sigma^2)$ and t is a deterministic linear time trend. This specification is particularly useful for macroeconomic series exhibiting secular trends and has been widely employed in empirical growth and inflation forecasting studies (Harvey, 1989; Diebold, 2015).

Together, these models represent widely accepted low-dimensional forecasting benchmarks. Their inclusion enables a transparent evaluation of whether machine learning methods deliver statistically and economically meaningful improvements in predictive accuracy beyond traditional linear time-series approaches.

To address potential sensitivity to specific error metrics, we evaluate forecast performance using four complementary measures which includes RMSE, MAE, MASE, and Theil's U. RMSE measures the effect of large errors, while MAE indicates overall average accuracy. MASE offers a scale-free assessment in relation to the naïve random walk benchmark, and Theil's U allows performance comparison with historical forecasting studies. This

comprehensive set ensures robustness when individual metrics yield conflicting rankings. We have also estimated Pearson correlation analysis and simple regressions as preliminary checks of consistency, supplemented by robustness tests to examine whether monthly forecasts could meaningfully explain annual variation.

For comparison, we have taken the annual data till 2020 and predicted the macroeconomic variables values for the year 2021, 2022, 2023 and 2024. We also have the official annual values for these years. Then we have compared the prediction from our model (aggregated to annual values) and three traditional benchmark models with the official data. On the bases of performance indicator such as RMSE, MAE, MASE, and Theil's U, we have evaluated the prediction of our model.

5. RESULT AND DISCUSSION

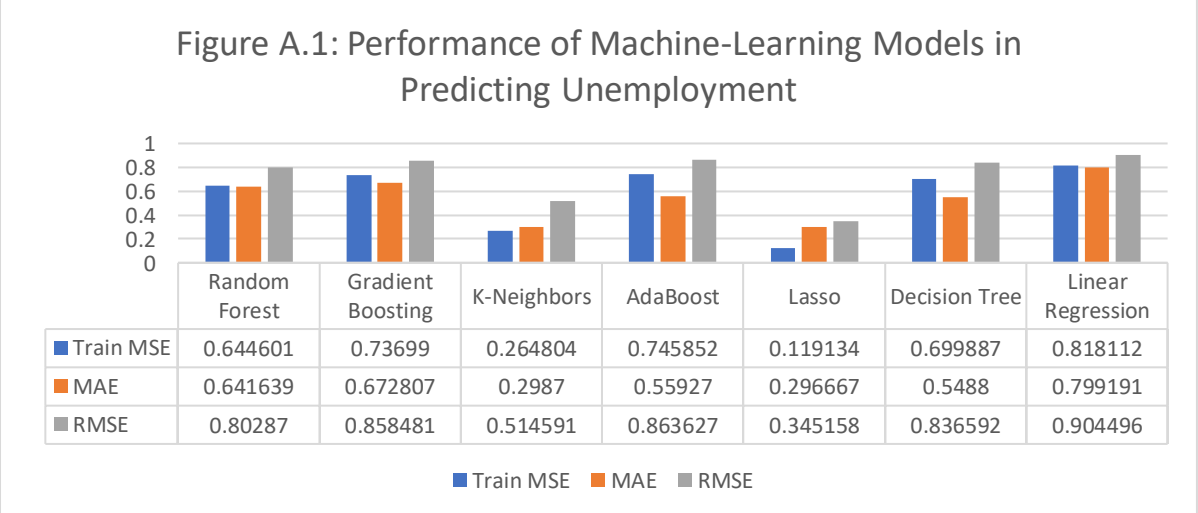
This section explores the predictive performance of various machine learning models such as Random Forest, Gradient Boosting, K-Nearest Neighbors, AdaBoost, Lasso Regression, Decision Tree, and Linear Regression, across three key macroeconomic variables Unemployment Rate, GDP Growth Rate and Consumer Price Index (CPI). Prediction and assessment of each series is provided below:

5.1 Unemployment

5.1.1 Model Selection and Feature Identification

We have applied Lasso with alpha value 0.005 for feature selection and results shows that 84 features are suitable for predicting unemployment using machine learning.

Figure A.1: Performance of Machine-Learning Models in Predicting Unemployment



	Random Forest	Gradient Boosting	K-Neighbors	AdaBoost	Lasso	Decision Tree	Linear Regression
Train MSE	0.644601	0.73699	0.264804	0.745852	0.119134	0.699887	0.818112
MAE	0.641639	0.672807	0.2987	0.55927	0.296667	0.5488	0.799191
RMSE	0.80287	0.858481	0.514591	0.863627	0.345158	0.836592	0.904496

Source: Computed by authors using Google Trends data

Figure A.1 presents the performance of different machine learning models in predicting unemployment, evaluated using three commonly employed error metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Among the models, Lasso regression stands out as the best performer, achieving the lowest MSE (0.119),

MAE (0.297), and RMSE (0.345). This indicates its ability to provide more accurate and consistent predictions compared to other approaches.

The K-Neighbors model also performs relatively well, with a notably lower MSE (0.265) and RMSE (0.515), suggesting it can capture non-linear relationships in the data effectively. In contrast, models such as Linear Regression and Gradient Boosting show higher error values, reflecting weaker predictive accuracy in this context. Interestingly, while Random Forest and Decision Tree perform moderately, they do not outperform simpler regularized regression approaches like Lasso.

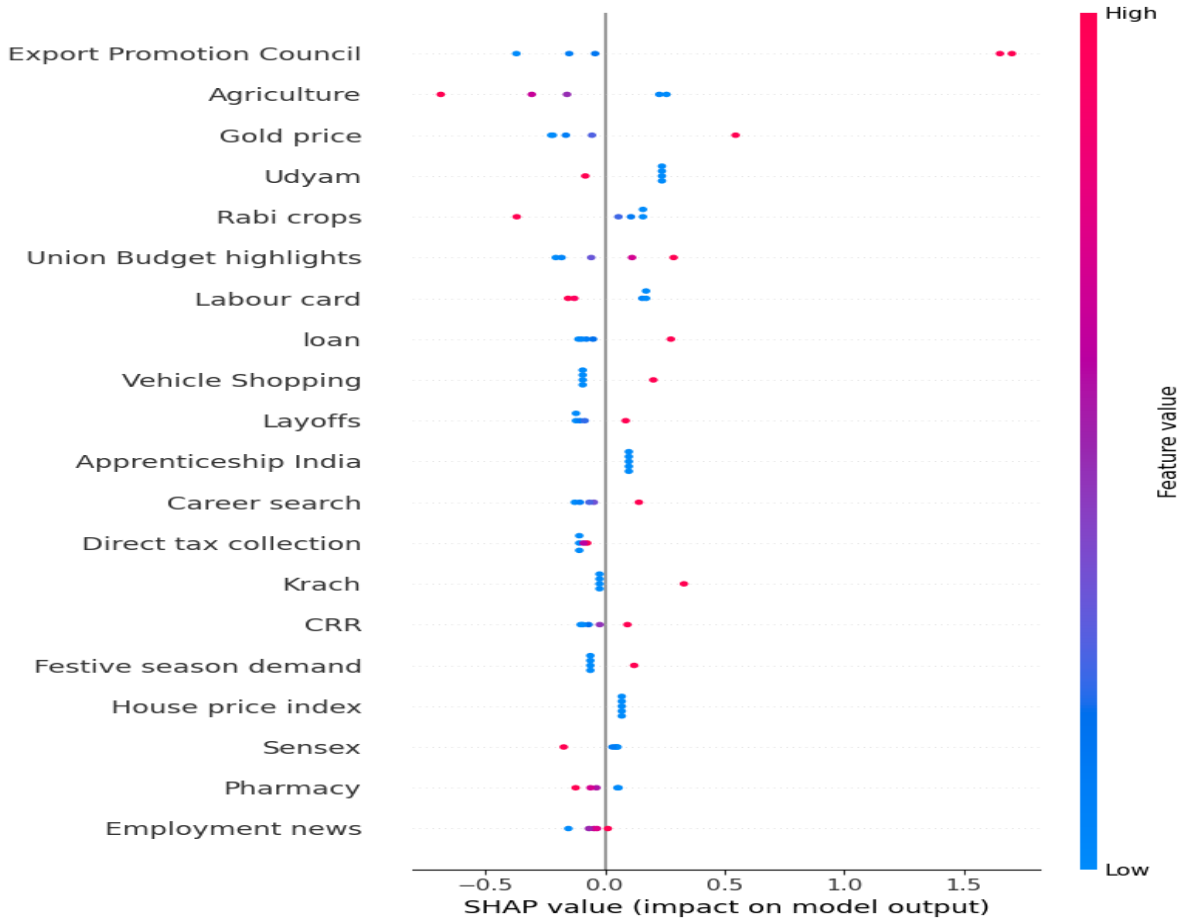
Overall, the results highlight that Lasso is better suited for unemployment prediction in the given dataset, outperforming other machine learning models.

5.1.2 Prediction Performance of Machine-Learning Models

Using the selected feature set, the model was trained on yearly data and subsequently used to predict India's *monthly* unemployment rate. The full series of monthly predictions is presented in Appendix A. By exploiting the high-frequency nature of Google Trends, the model is able to capture short-term fluctuations in labour market conditions that are not visible in traditional annual indicators.

5.1.3 Interpretability of the Model (SHAP Analysis)

Figure A.2: Global Feature Importance (SHAP Summary Plot) for unemployment



Source: Computed by authors using Google Trends data

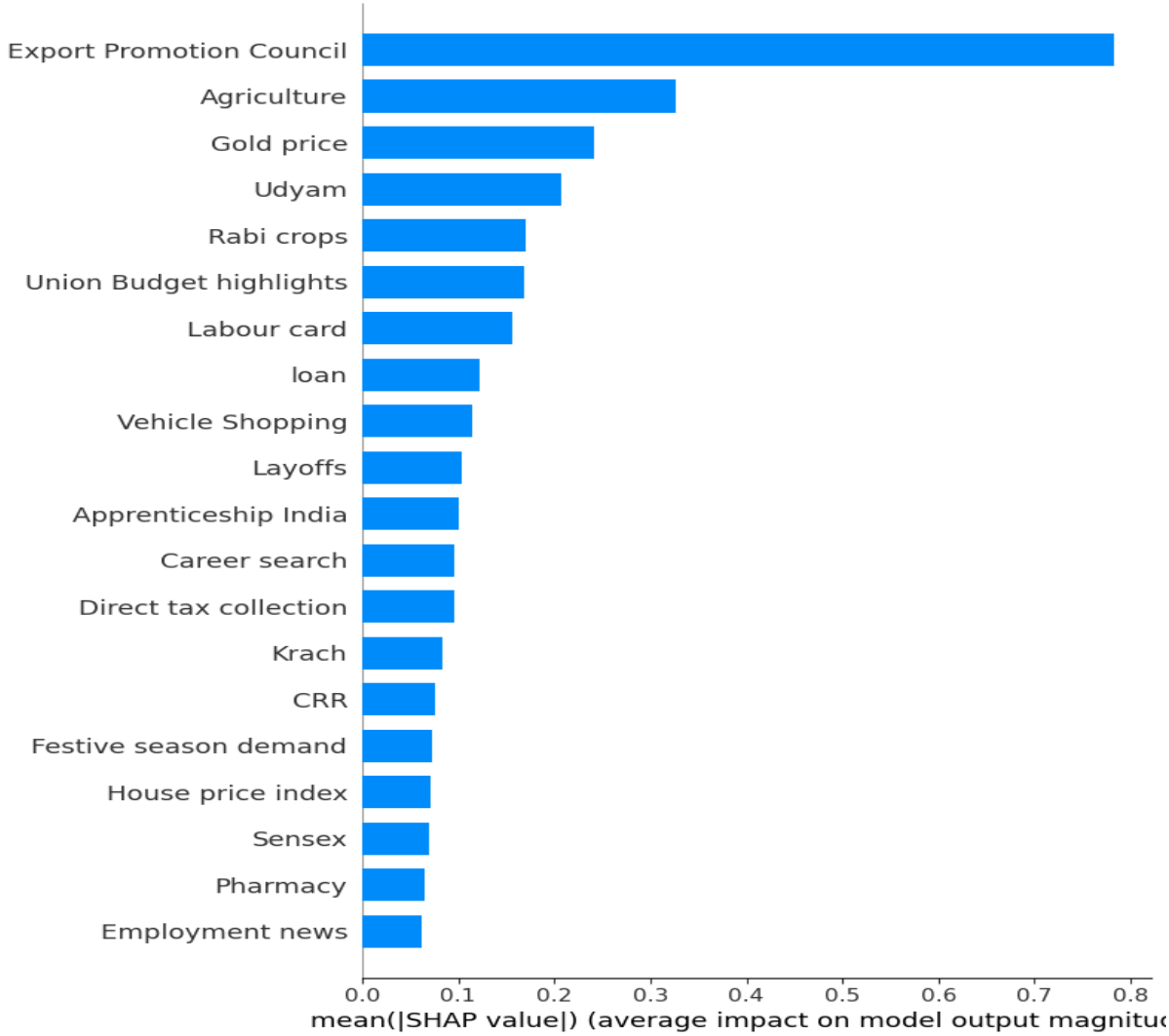
To know which google search keyword is contributing in predicting the monthly unemployment values, we have applied SHAP analysis, which provides both global and local interpretability of the model.

The SHAP summary plot (Figure A.2) reveals that the *Export Promotion Council* emerges as the most influential and volatile predictor, with higher search interest consistently contributing to upward movements in predicted unemployment. This suggests that disruptions or uncertainty in export-related activities, often associated with demand shocks or global trade fluctuations, tend to coincide with elevated job-market stress.

By contrast, *Agriculture* shows a pronounced negative association, with higher search volumes reducing unemployment predictions. This aligns with the counter-cyclical nature of agricultural activity in India, where rural employment buffers short-term labour market shocks. Similarly, *Gold price* exerts downward pressure on unemployment predictions, indicating the stabilizing influence of commodity-driven liquidity and household wealth effects.

Policy-related keywords such as *Union Budget highlights* and employment-linked terms like *Labour card*, *Udyam*, and *Layoffs* contribute meaningfully but less strongly, capturing policy announcements, compliance behaviour, and job-search intensity.

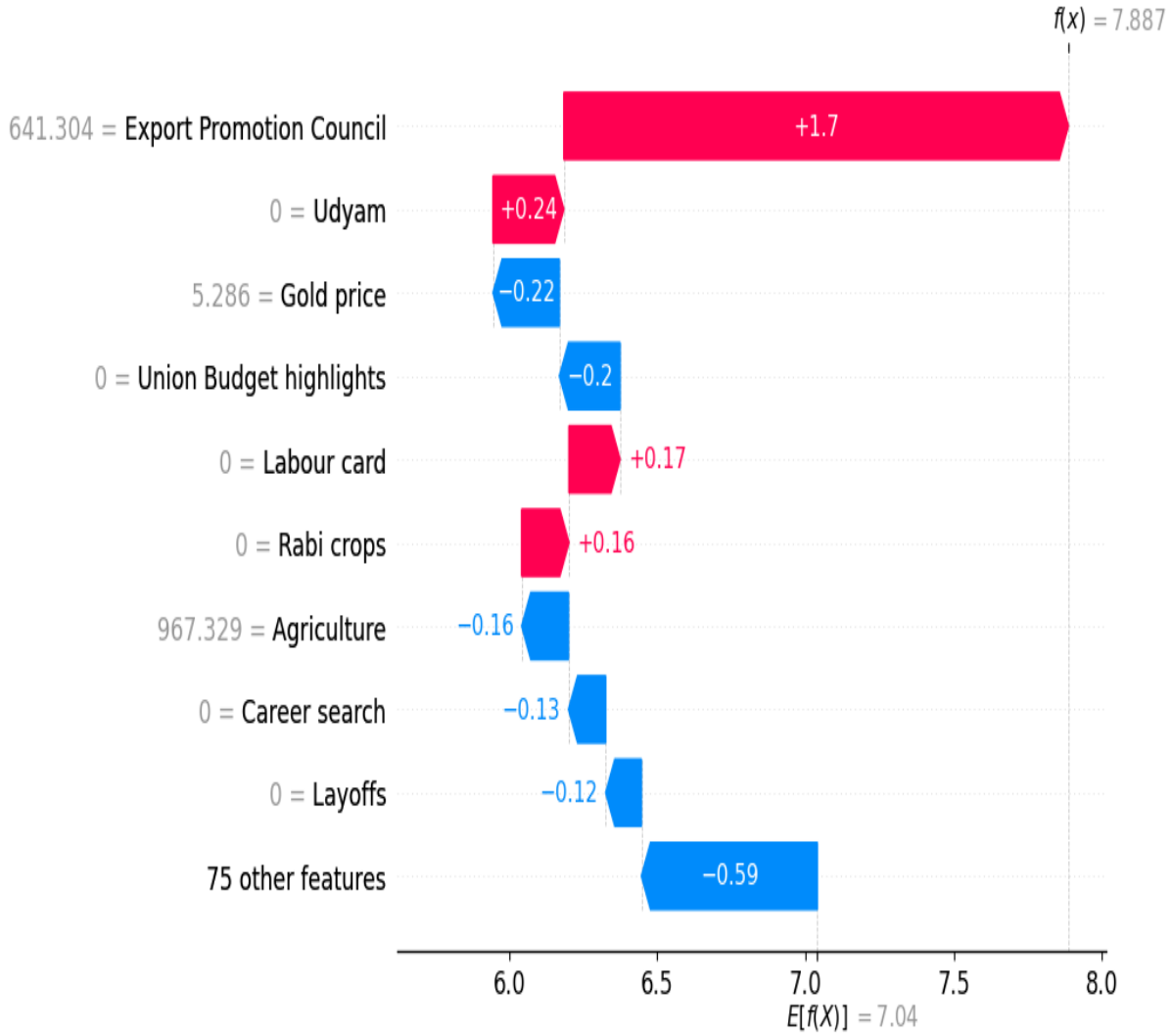
Figure A.3: Mean Absolute SHAP Value for unemployment prediction



Source: Computed by authors using Google Trends data

From the figure A.3, we observe that the '*Export Promotion Council'* stands out as the single most influential predictor, followed by '*Agriculture'* and '*Gold price'*. These three dominate the feature space, accounting for the largest share of model influence. Secondary but relevant predictors include '*Udyam', 'Rabi crops', 'Union Budget highlights', 'Labour card',* and '*Layoffs'*, which play supportive roles in refining forecasts. By contrast, macroeconomic and financial indicators such as '*CRR', 'Sensex', 'House price index',* and '*Direct tax collection'* contribute less consistently, suggesting that while they are informative, their predictive strength is relatively weaker. This ranking reinforces the primacy of trade, agriculture, and commodity dynamics as the backbone of unemployment forecasting in the model.

Figure A.4: Local SHAP Explanation for a Single Prediction



Source: Computed by authors using Google Trends data

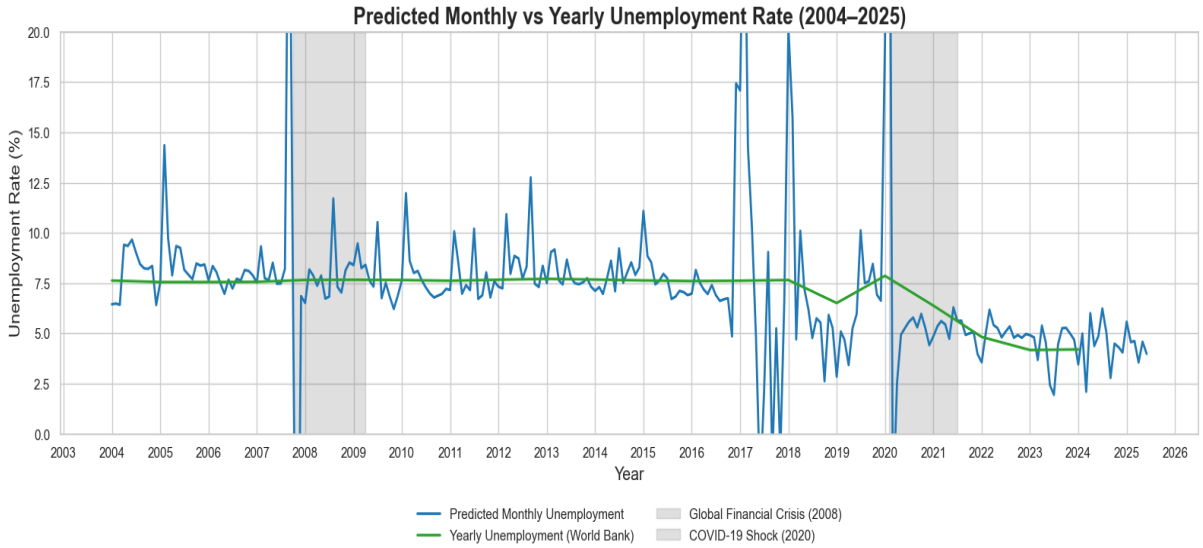
From figure A.4, we can infer that the '*Export Promotion Council'* shows the strongest upward influence, pushing the predicted unemployment significantly above the baseline expectation. This suggests that trade-related fluctuations captured by export promotion activities are closely tied to localized spikes in unemployment. On the other hand, features such as '*Gold price', 'Union Budget highlights',* and '*Agriculture'* exert downward pressure, counteracting the prediction to some extent. Additional moderate positive contributions come from '*Udyam',*

'Labour card', and '*Rabi crops'*, which reinforce the forecasted increase. This localized explanation illustrates the model's reasoning for an above-average unemployment prediction, emphasizing the dominant role of export-related factors while also showing how commodity markets and agricultural indicators mitigate the effect.

Across both local and global perspectives, the analysis consistently identifies '*Export Promotion Council'*, *'Agriculture'*, and '*Gold price'* as the dominant drivers of unemployment predictions. Export-related activities tend to amplify unemployment forecasts, while agricultural performance and commodity prices provide stabilizing effects. Policy indicators and labour-related terms contribute meaningfully but with moderate influence, reflecting their supplementary role in shaping labour market outcomes. The convergence of evidence from individual predictions, global distributions, and average feature importance strengthens the reliability of these findings. Together, the results demonstrate that machine learning models not only capture predictive accuracy but also offer interpretable insights into the structural and sectoral determinants of unemployment, thereby providing valuable guidance for policymakers and economic analysts. These insights suggest that policymakers should closely track trade shocks and agricultural cycles, as they play a pivotal role in shaping short-term employment conditions. Early identification of such drivers can enable timely interventions, particularly in export-dependent and rural labour markets.

5.1.4 Monthly Unemployment Dynamics

Figure A.5: Predicted Monthly Unemployment Rate and Official Annual Unemployment Data



Source: Computed by authors using Google Trends data

Figure A.5 highlights key dynamics in unemployment trends. The predicted monthly unemployment series shows greater volatility than the yearly data, capturing short-term fluctuations that are averaged out in annual estimates. This volatility is particularly evident during major economic disruptions, such as the Global Financial Crisis (2008) and the COVID-19 shock (2020), where sharp spikes and irregular movements are observed in the monthly predictions.

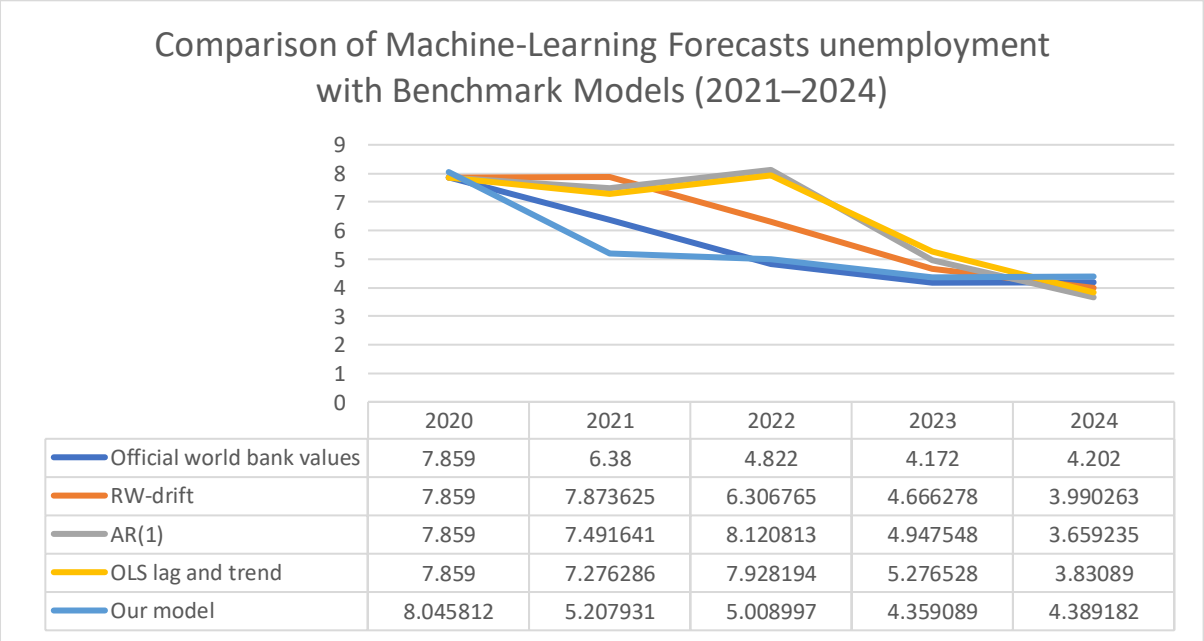
Despite these fluctuations, the general trajectory of the monthly unemployment series aligns with the yearly World Bank estimates, especially in capturing long-term trends. For instance, the relative stability in unemployment from 2004 to 2015, followed by a gradual decline after 2018, is visible in both series. The impact of COVID-19 is also clearly reflected, with a sharp rise around 2020 before a declining trend in subsequent years.

The figure A.5 illustrates that while annual data provide a smooth representation of unemployment dynamics, monthly estimates offer a richer picture of short-term labour market stress and recovery. This makes the predicted series particularly useful for real-time monitoring and policy assessment, especially during periods of economic crisis where rapid shifts may not be immediately visible in annual data.

5.1.5 Benchmark Comparison (Annual Forecast Evaluation)

For a formal comparison with conventional forecasting models, we aggregated our monthly unemployment predictions to annual frequency and contrasted them with out-of-sample forecasts generated using three benchmark methods: Random Walk with drift, AR(1), and OLS with lag and trend. The benchmark models were estimated on official World Bank data for 2004–2021 and used to predict unemployment for 2021–2024.

Figure A.6: Comparison of Machine-Learning Forecasts for unemployment with traditional benchmark Models (2021–2024)



	2020	2021	2022	2023	2024
Official world bank values	7.859	6.38	4.822	4.172	4.202
RW-drift	7.859	7.873625	6.306765	4.666278	3.990263
AR(1)	7.859	7.491641	8.120813	4.947548	3.659235
OLS lag and trend	7.859	7.276286	7.928194	5.276528	3.83089
Our model	8.045812	5.207931	5.008997	4.359089	4.389182

Source: Computed by authors

Figure A.6 compares the unemployment forecasts generated by the benchmark models, Random Walk with drift, AR (1), and the OLS with lag and trend specification, with those from our machine learning model, using the official World Bank series as the reference.

The official data show a sharp decline in unemployment after 2020, followed by a gradual adjustment. Across the benchmark models, forecasts tend to either overstate the unemployment rate or fail to fully capture the magnitude of the post-pandemic decline. In contrast, our model

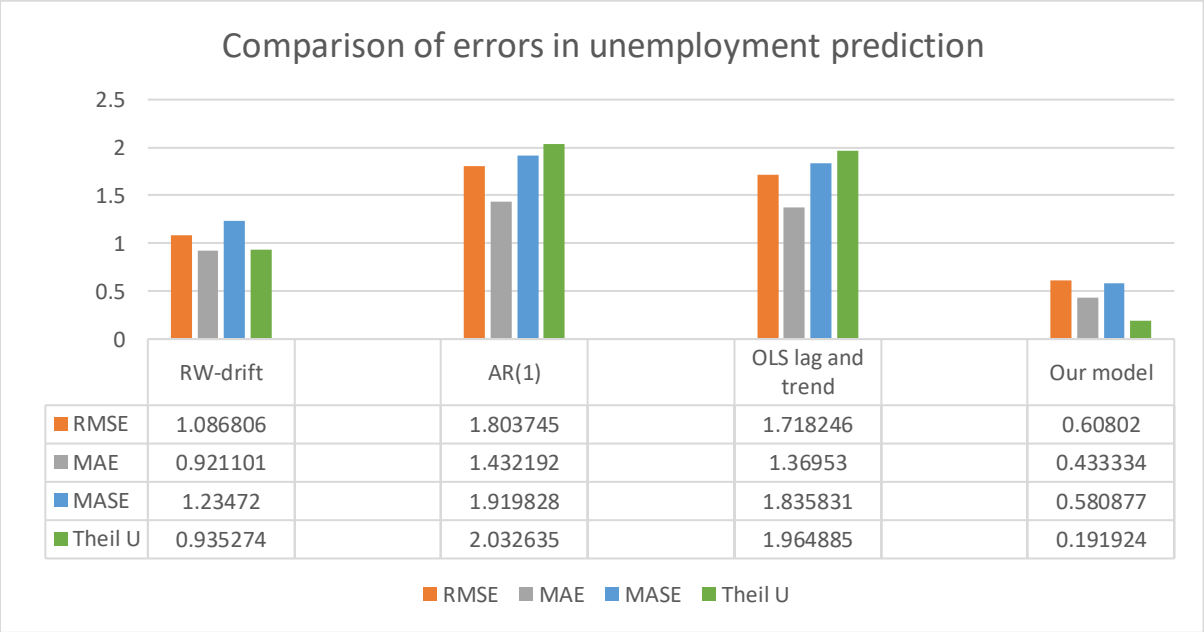
more closely tracks the turning points and levels observed in the official series during these four years, reflecting improved responsiveness to the structural shifts in the labour market.

Overall, the figure demonstrates that although traditional time-series benchmarks provide reasonable directional forecasts, the machine-learning model achieves a noticeably closer alignment with the official unemployment values during the evaluation period (2021–2024), highlighting its superior nowcasting performance.

5.1.6 Relative Performance and Literature Comparison

Now in order to confirm our results, we have analysed the performance indicator such as RMSE, MAE, MASE, and Theil's U.

Figure A.7: Comparison of errors in unemployment prediction



	RW-drift	AR(1)	OLS lag and trend	Our model
RMSE	1.086806	1.803745	1.718246	0.60802
MAE	0.921101	1.432192	1.36953	0.433334
MASE	1.23472	1.919828	1.835831	0.580877
Theil U	0.935274	2.032635	1.964885	0.191924

Source: Computed by authors using Google Trends data

The results from figure A.7, clearly indicate that our machine learning based approach outperforms all benchmark models across every evaluation metric. For unemployment prediction, the Random Walk with drift achieves an RMSE of 1.0868 and MAE of 0.9211, while the AR (1) and OLS lag-trend models perform even worse, with RMSE values of 1.8037 and 1.7182, respectively. In contrast, our model attains a substantially lower RMSE of 0.6080 and MAE of 0.4333, representing a reduction in forecast error of more than 40% compared to the Random Walk benchmark. Similarly, the MASE values for our model of 0.58 is far below those of the benchmark models, demonstrating its superior accuracy and robustness. The markedly lower Theil's U statistic (0.19) further confirms that our approach not only improves predictive precision but also contains genuine informational content beyond naïve or autoregressive methods. These results highlight the significant value added by incorporating Google Trends data and machine-learning techniques for timely and reliable unemployment nowcasting in the Indian context.

Compared with earlier studies, the improvement is particularly noteworthy. For example, D'Amuri and Marcucci (2012) reported only a marginal enhancement when augmenting an

AR(1) model with Google Trends data for U.S. unemployment (RMSE improving from 0.136 to 0.133). In contrast, our model achieves a much larger performance gain in the Indian context, reducing the AR(1) RMSE from 1.804 to 0.608. This reflects the higher informational content of Google Trends signals in emerging economies, where official statistics are infrequent and subject to delays.

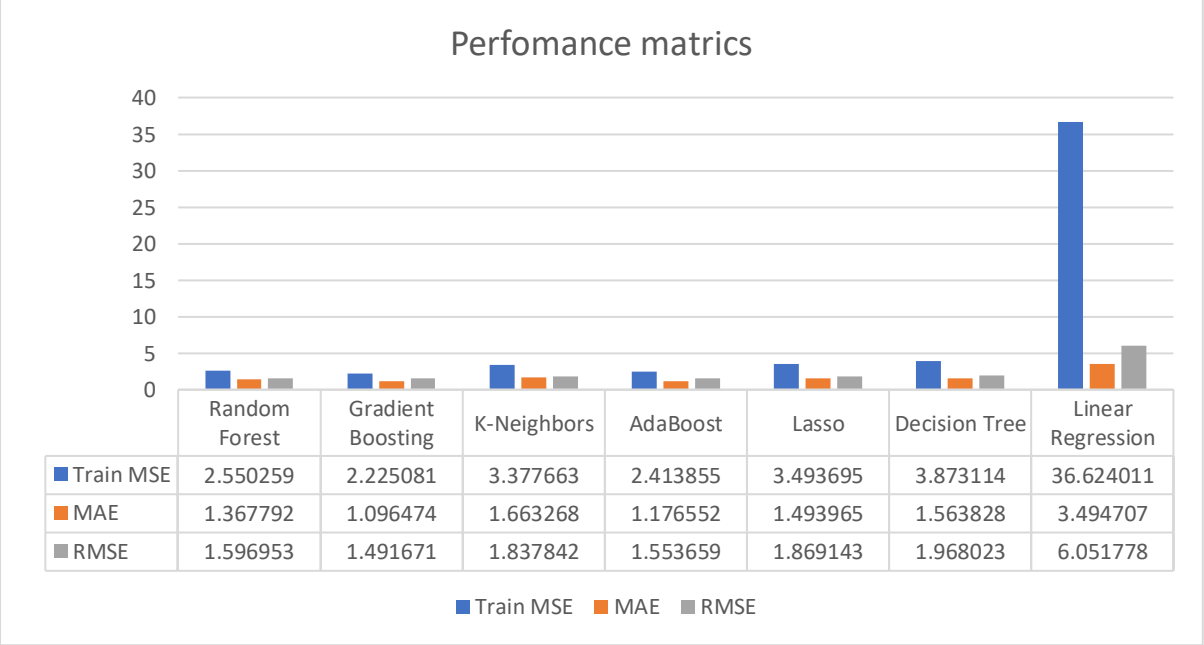
5.2 GDP Growth rate

5.2.1 Model Selection and Feature Identification

To identify relevant predictors of GDP growth, we aapplied Lasso with a penalty parameter (alpha value) of 0.005. Results indicates that 106 features are suitable for predicting GDP Growth Rate using machine learning.

The predictive performance of candidate models is shown in Figure B.1. From the figure we observe that among the models, Gradient Boosting delivers the best results, with the lowest MSE (2.225), MAE (1.096), and RMSE (1.492). This indicates that it is more effective at capturing the complex, non-linear dynamics of GDP growth compared to other approaches. AdaBoost and Random Forest also perform relatively well, showing moderate error values and reinforcing the strength of ensemble-based methods in this forecasting task.

Figure B.1: Performance metrics of various machine learning models for predicting GDP Growth rate



	Random Forest	Gradient Boosting	K-Neighbors	AdaBoost	Lasso	Decision Tree	Linear Regression
Train MSE	2.550259	2.225081	3.377663	2.413855	3.493695	3.873114	36.624011
MAE	1.367792	1.096474	1.663268	1.176552	1.493965	1.563828	3.494707
RMSE	1.596953	1.491671	1.837842	1.553659	1.869143	1.968023	6.051778

Source: Computed by authors using Google Trends data

In contrast, models such as K-Neighbors, Lasso regression, and Decision Tree exhibit higher error levels, suggesting limited ability to handle the variability and fluctuations inherent in GDP growth data. The weakest performance is observed for Linear Regression, with a substantially higher MSE (36.624) and RMSE (6.052), highlighting its inability to capture the underlying patterns in the data.

Overall, the findings suggest that ensemble learning approaches, particularly Gradient Boosting, are the most reliable for predicting GDP growth rates, while traditional linear models perform poorly in this context.

5.2.2 Predictive Performance of the Machine-Learning Model

Using the 106 selected features, the Gradient Boosting model was trained to generate **monthly** GDP growth estimates for India. These high-frequency predictions leverage granular behavioural signals from Google Trends, offering insights into economic turning points that traditional annual aggregates may obscure.

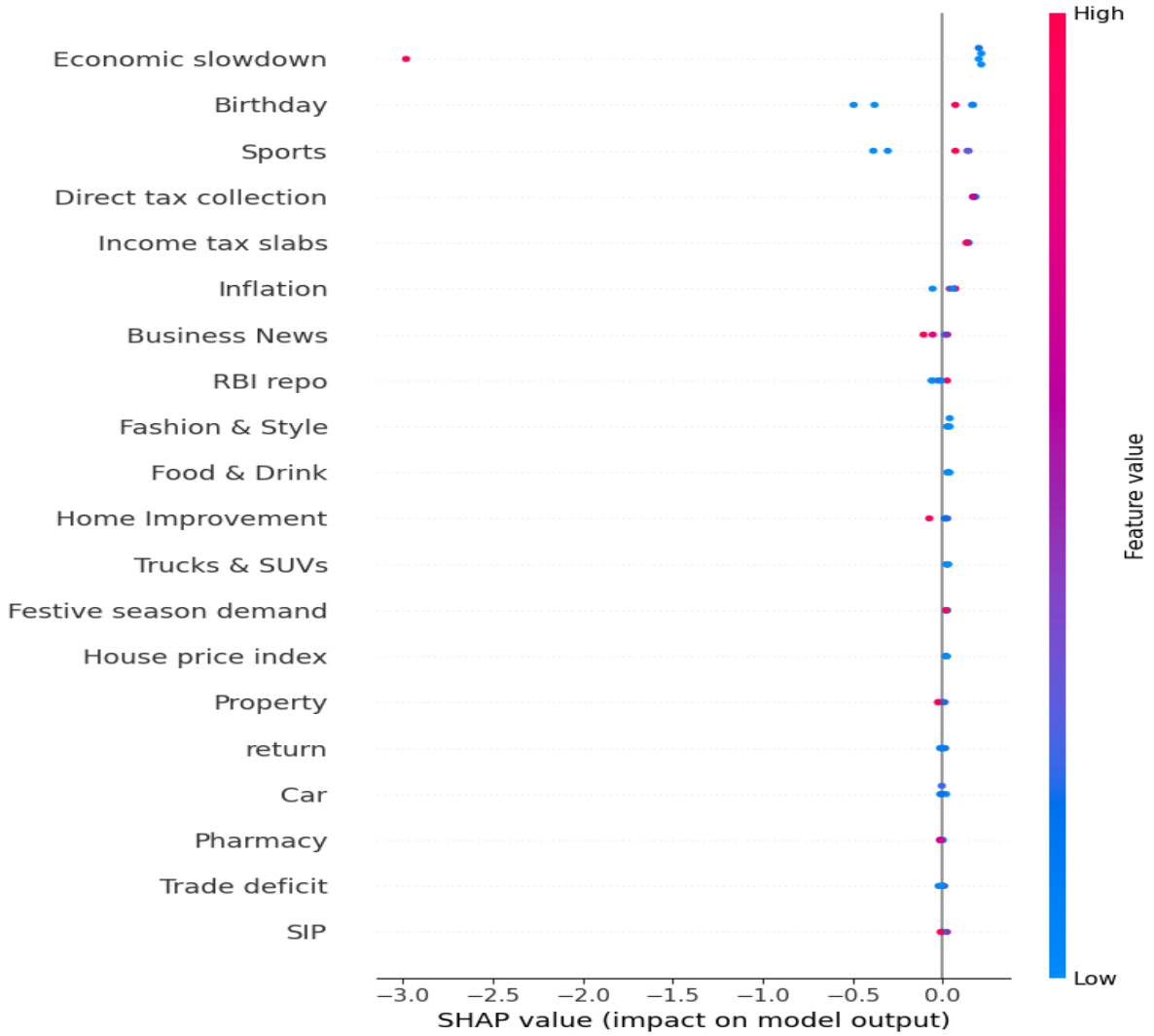
The complete series of monthly predictions is provided in Appendix A.

5.2.3 Model Interpretability: SHAP Analysis

To evaluate the economic drivers influencing the model's behaviour, we conducted a SHAP analysis providing global, average, and local explanations.

Global Feature Contributions

Figure B.2: Global Feature Importance for GDP growth rate prediction

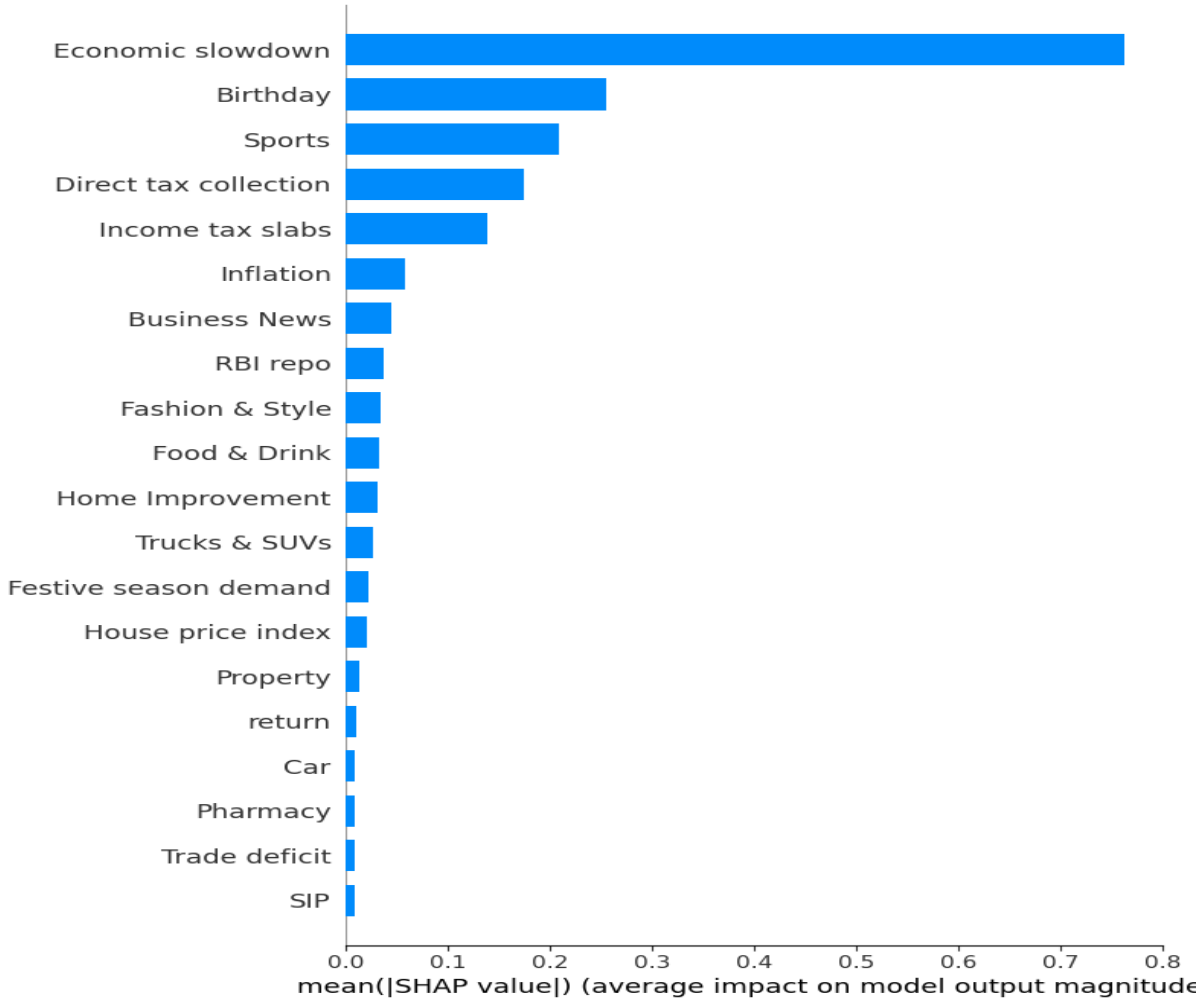


Source: Computed by authors using Google Trends data

According to figure B.2, *Economic slowdown* emerges as the most consistent positive driver of predicted GDP growth, while features like *RBI repo* and *Inflation* frequently push predictions downward. Interestingly, consumer- and sentiment-linked terms such as *Birthday* and *Sports* consistently show up on the positive side, pointing to their association with periods of stronger demand conditions. The heterogeneity in SHAP distributions indicates that the impact of certain features varies across time, with some drivers having episodic but large effects.

Mean Absolute SHAP Importance

Figure B.3: Mean Absolute SHAP Value for GDP growth rate prediction



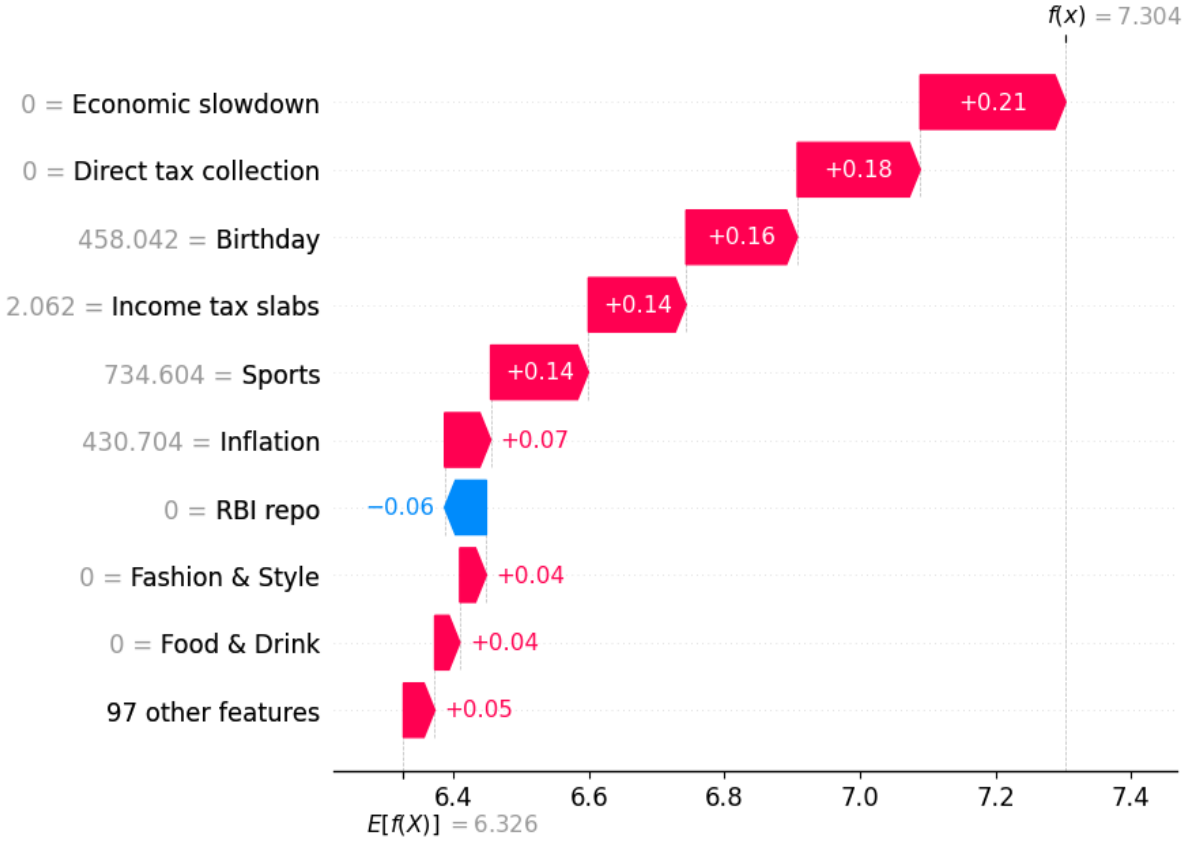
Source: Computed by authors using Google Trends data

From Figure B.3, we observe that when features are ranked by their average importance, *Economic slowdown* stands out as the dominant factor, overshadowing other predictors. This highlights that the model interprets search intensity around slowdown-related terms as a highly reliable signal of shifts in growth dynamics. Other top features include *Birthday*, *Sports*, and *Direct tax collection*, reinforcing the role of consumer behaviour and fiscal policy as central to the model's explanatory power. Traditional variables like *Inflation* and *RBI repo* matter, but their contributions are more modest compared to the unconventional Indicators.

Local Prediction Explanations

From figure B.4, we observe that the largest positive drivers are *Economic slowdown*, *Direct tax collection*, and *Birthday*, all of which substantially increase the prediction. These results suggest that both macroeconomic concerns (slowdown, tax policy) and behavioural patterns (search intensity around birthdays) carry significant weight in shaping growth expectations. On the other hand, the *RBI repo* rate acts as a mild drag on the prediction, consistent with the contractionary role of monetary tightening.

Figure B.4: Single Prediction Explanation for GDP growth rate prediction



Source: Computed by authors using Google Trends data

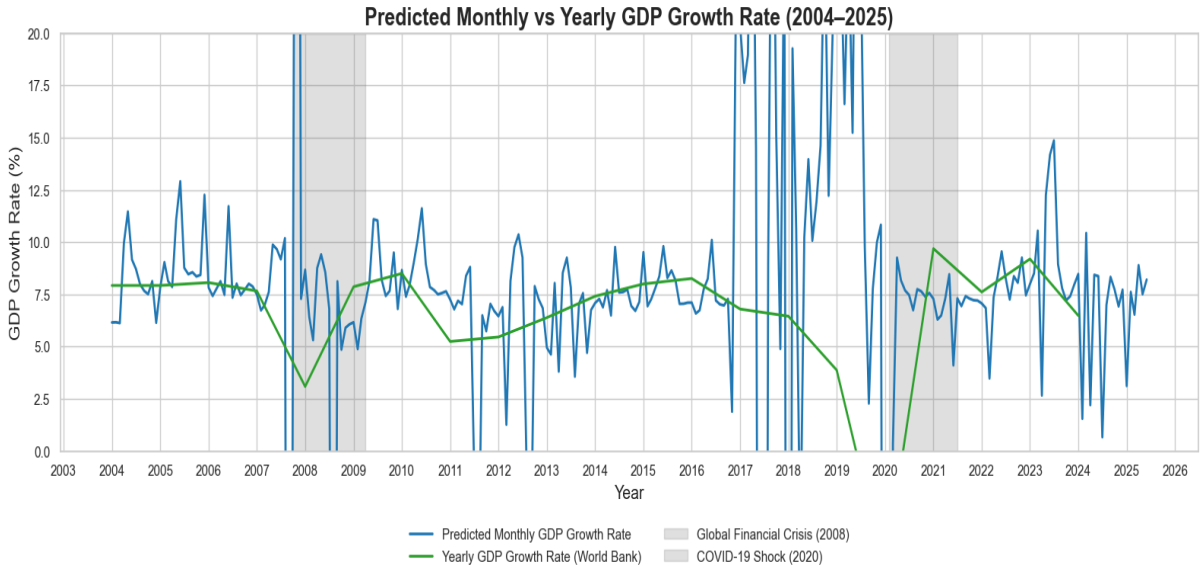
Taken together, the three figures reveal that the model assigns high importance not just to conventional macroeconomic indicators but also to behavioural and sentiment-driven search patterns. *Economic slowdown* dominates as the single most influential driver, while terms like *Birthday* and *Sports* highlight the predictive power of nontraditional data in capturing shifts in consumption and sentiment. Fiscal policy indicators, particularly *Direct tax collection* and *Income tax slabs*, further complement the model by capturing institutional and policy-related dynamics. Overall, the results suggest that integrating Google Trends with macroeconomic variables enriches the model's ability to anticipate GDP growth by combining real-time consumer signals with structural policy drivers.

The prominence of search-based sentiment indicators, such as economic slowdown and tax related terms, underscores the role of behavioural signals in growth monitoring. Incorporating these real-time insights could help policymakers respond faster to shifts in consumption, confidence, and fiscal dynamics.

5.2.4 Monthly GDP Growth Dynamics

Figure B.5 shows that the monthly predictions exhibit substantially greater volatility than the annual series, underscoring the responsiveness of high-frequency estimates to short-term shocks and cyclical fluctuations that are otherwise smoothed out in yearly data.

Figure B.5: Predicted monthly GDP growth rate compared with officially reported annual GDP growth rate data from the World Bank



Source: Computed by authors using Google Trends data

The Global Financial Crisis (2008) is clearly reflected, with the annual series showing a sharp slowdown to nearly 3%, while the monthly predictions capture more severe contractions and rapid rebounds around the same period. Similarly, the COVID-19 shock (2020) produces a dramatic collapse in the monthly growth series, followed by a sharp rebound, whereas the annual estimates smooth this effect into a deeper but less abrupt trough.

Over the longer horizon, both series broadly move in tandem, reflecting the underlying trajectory of the economy: strong growth in the mid-2000s, deceleration during global and domestic shocks, recovery in the early 2010s, and renewed volatility around 2018–2020. Importantly, the predicted monthly series provides finer granularity, showing turning points earlier than annual estimates, for example, the post-2008 recovery and the rebound after COVID-19.

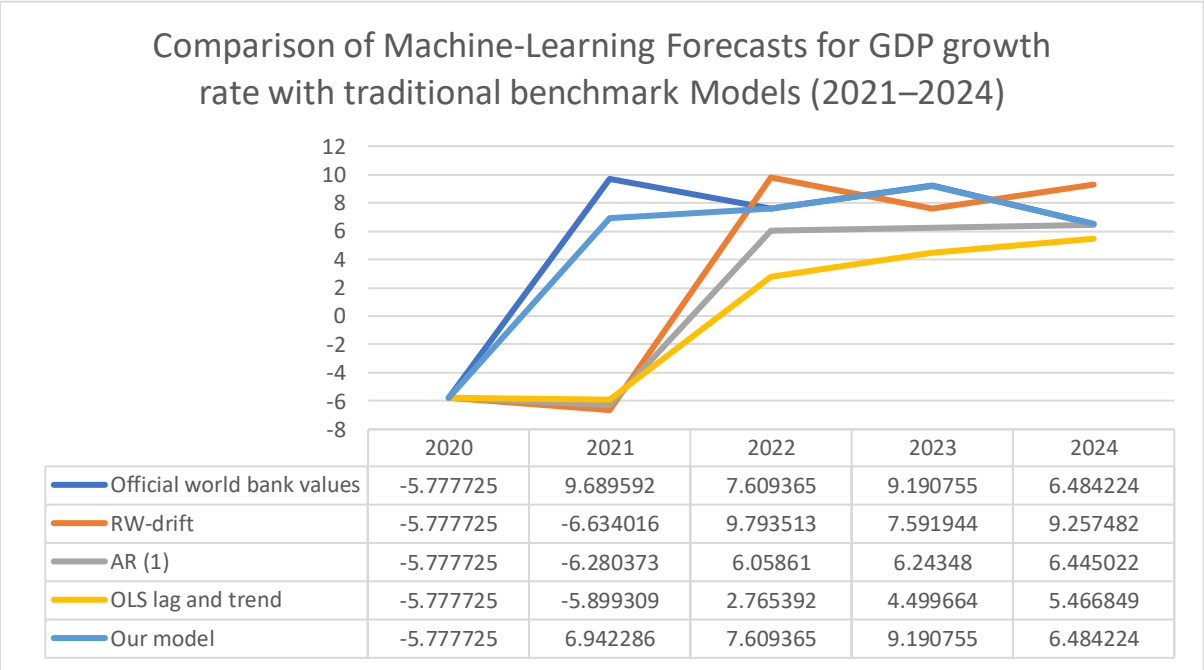
Overall, the figure demonstrates that while annual GDP growth rates provide a stable long-term perspective, monthly predictions offer valuable insights into the timing, depth, and volatility of growth cycles, making them especially useful for real-time policy monitoring during crises.

5.2.5 Benchmark Comparison (Annual Forecast Evaluation)

To compare the machine-learning model with traditional forecasting approaches, the monthly predictions were aggregated to annual frequency and contrasted with out-of-sample forecasts

from RW-drift, AR(1), and OLS lag-trend models. Each benchmark model was estimated on World Bank data from 2004–2020 and used to forecast GDP growth for 2021–2024.

Figure B.6: Comparison of Machine-Learning Forecasts for GDP growth rate with traditional Benchmark Time-Series Models (2021–2024)



	2020	2021	2022	2023	2024
Official world bank values	-5.777725	9.689592	7.609365	9.190755	6.484224
RW-drift	-5.777725	-6.634016	9.793513	7.591944	9.257482
AR (1)	-5.777725	-6.280373	6.05861	6.24348	6.445022
OLS lag and trend	-5.777725	-5.899309	2.765392	4.499664	5.466849
Our model	-5.777725	6.942286	7.609365	9.190755	6.484224

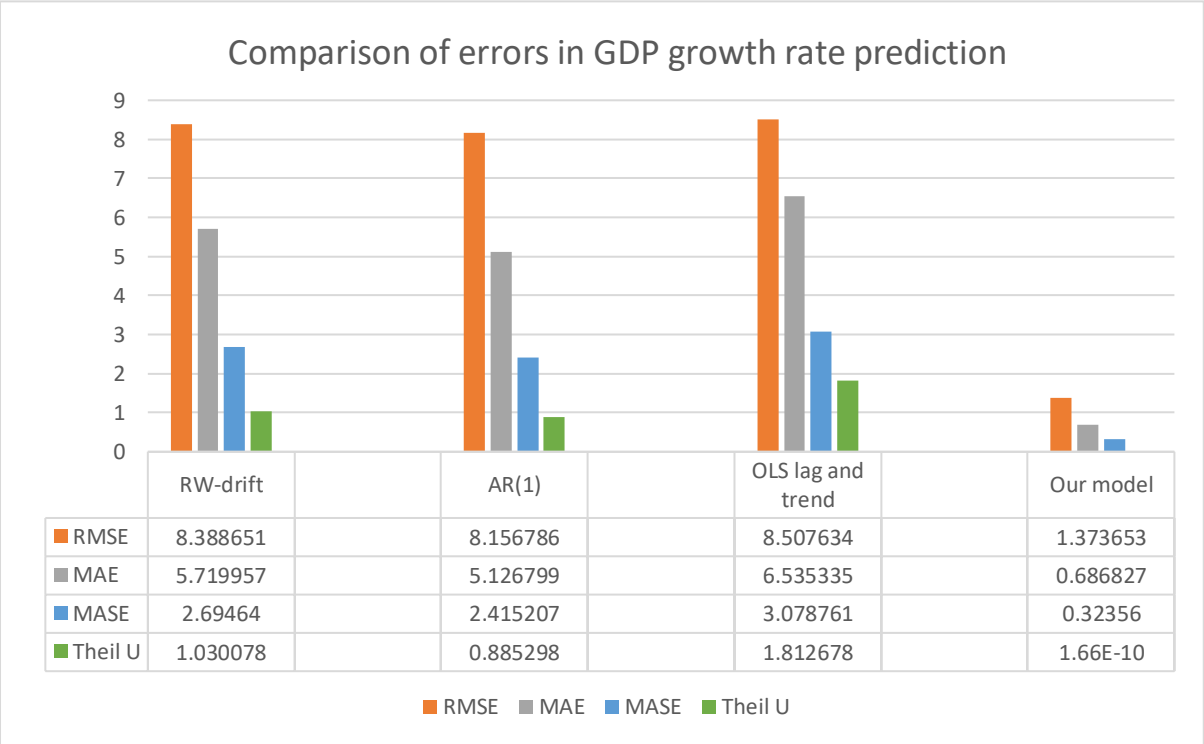
Source: Computed by authors using Google Trends data

The results indicate that traditional benchmark models exhibit substantial lag in capturing the economic rebound in 2021, with both RW-drift and AR(1) continuing to predict negative or weak growth. In contrast, the machine-learning model generates a positive growth estimate for 2021 that is closer to the official data, suggesting improved responsiveness to rapidly changing economic conditions. From 2022 to 2024, the machine-learning model closely tracks the official growth trajectory, with particularly strong alignment in 2022 and 2023, while the OLS lag-and-trend model consistently underestimates the level of economic recovery.

5.2.6 Relative Performance and Literature Comparison

From the figure B.7, we observe that the benchmark models exhibit relatively high forecast errors, with RMSE values ranging from 8.15 (AR(1)) to 8.50 (OLS), and MAE values between 5.12 and 6.53. Their MASE statistics further indicate weak predictive accuracy, reflecting their limited ability to capture the high volatility and non-linear dynamics inherent in GDP growth. In contrast, our machine learning based model delivers a dramatic improvement, achieving an RMSE of 1.37 and MAE of 0.69, an almost 85% reduction in forecast error relative to traditional approaches. The MASE (0.32) values is equally superior, highlighting the model's consistent accuracy across different scales of growth variation. Notably, the Theil U statistic for our model approaches zero, indicating near-perfect predictive performance and demonstrating that the information captured by Google Trends based behavioural indicators, combined with machine learning, significantly outperforms both naïve and autoregressive benchmarks. These results underscore the substantial gains in precision and reliability achieved through our high-frequency, search-based nowcasting methodology.

Figure B.7: Comparison of errors in GDP growth rate prediction



	RW-drift	AR(1)	OLS lag and trend	Our model
RMSE	8.388651	8.156786	8.507634	1.373653
MAE	5.719957	5.126799	6.535335	0.686827
MASE	2.69464	2.415207	3.078761	0.32356
Theil U	1.030078	0.885298	1.812678	1.66E-10

Source: Computed by authors using Google Trends data

Grath (2025) nowcasts Swedish GDP growth and reports that the RMSE of an AR (1) model for a three-step-ahead forecast is 1.94, which improves marginally to 1.91 (1.55% improvement) when using a Bayesian neural network with two hidden layers. In terms of MAE, the AR (1) model yields 0.98, and this is reduced to 0.95 (3.06% improvement) using a random forest model. Similarly, Pratap and Priyaranjan (2023) find that for India's GDP growth (pre-COVID), the RMSE of an AR(1) specification is 2.11, which improves to 1.95 (7.58% improvement) when using a MIDAS approach; in the post-COVID period, the AR(1) RMSE rises sharply to 15.16, but is reduced to 13.54 (9.36% improvement) using a VAR model. Woloszko (2024) makes an improvement of 51% over AR(4) for predicting weekly GDP growth rate of India for the period 2020-21.

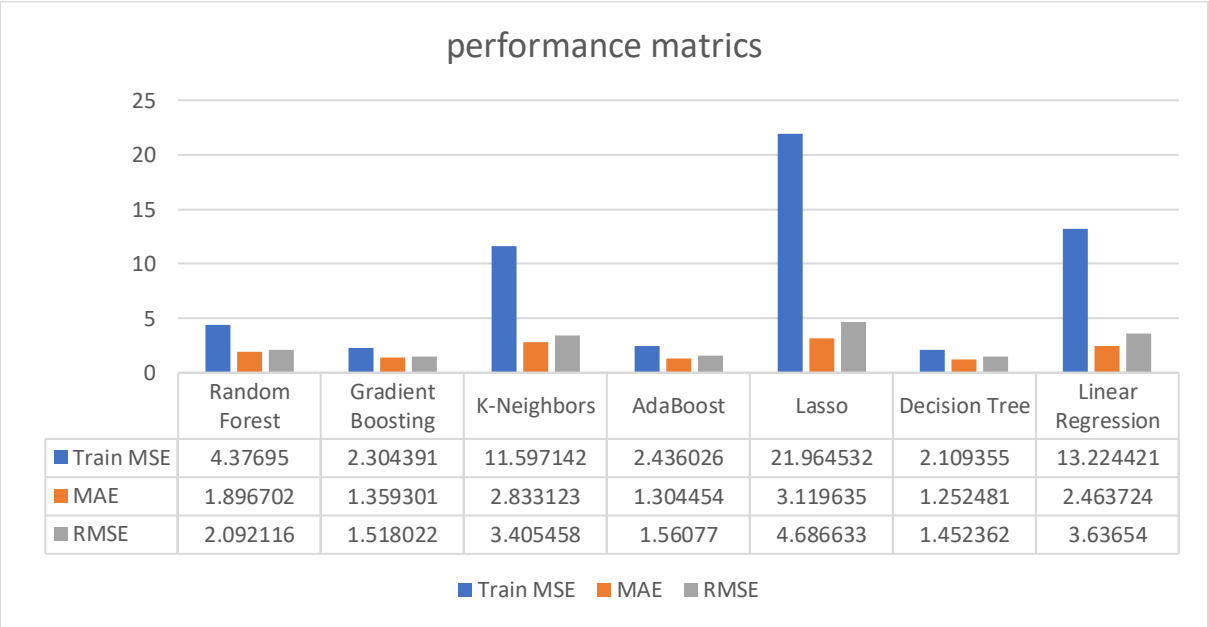
In comparison, our results show a substantially larger improvement: the RMSE of the AR (1) benchmark in our setting is 8.157, which our machine-learning model reduces to 1.374 (83.17% improvement), while the MAE drops from 5.127 to 0.687 (86.66 improvement). These gains highlight the significantly higher predictive power of our search-based machine-learning framework relative to traditional time-series methods.

5.3 Consumer Price Index (CPI)

5.3.1 Model Selection and Feature Identification

To identify relevant Google Trends indicators for inflation forecasting, we applied Lasso regression with a penalty parameter of $\alpha = 0.005$. The model selected 102 search-based features suitable for predicting CPI of India

Figure C.1: measurement matrix of different machine learning methods for predicting CPI



	Random Forest	Gradient Boosting	K-Neighbors	AdaBoost	Lasso	Decision Tree	Linear Regression
Train MSE	4.37695	2.304391	11.597142	2.436026	21.964532	2.109355	13.224421
MAE	1.896702	1.359301	2.833123	1.304454	3.119635	1.252481	2.463724
RMSE	2.092116	1.518022	3.405458	1.56077	4.686633	1.452362	3.63654

Source: Computed by authors using Google Trends data

Figure C.1 reports that among the models, the Decision Tree demonstrates the best predictive accuracy, with the lowest MSE (2.109), MAE (1.252), and RMSE (1.452). Close behind, Gradient Boosting and AdaBoost also achieve strong results, with relatively low error values across all three measures, suggesting that these ensemble and tree-based methods are particularly effective in capturing the dynamics of CPI.

Random Forest performs moderately well but with slightly higher errors compared to Gradient Boosting and Decision Tree. Models such as K-Neighbors and Linear Regression exhibit weaker performance, reflected in higher error values, while Lasso regression performs the poorest, with the largest MSE (21.964) and RMSE (4.687). This indicates that simple regularized linear methods may not be well-suited for predicting CPI, likely due to the complex and non-linear structure of the data.

Overall, the findings suggest that tree-based methods, particularly Decision Tree and Gradient Boosting, provide the most reliable forecasts of CPI.

5.3.2 Predictive Performance of the Machine-Learning Model

Using the selected feature set, the Decision Tree model was trained to produce **monthly estimates** of India's CPI. These high-frequency predictions exploit the real-time responsiveness embedded in Google search activity, enabling more granular assessment of inflationary pressures compared to annual data.

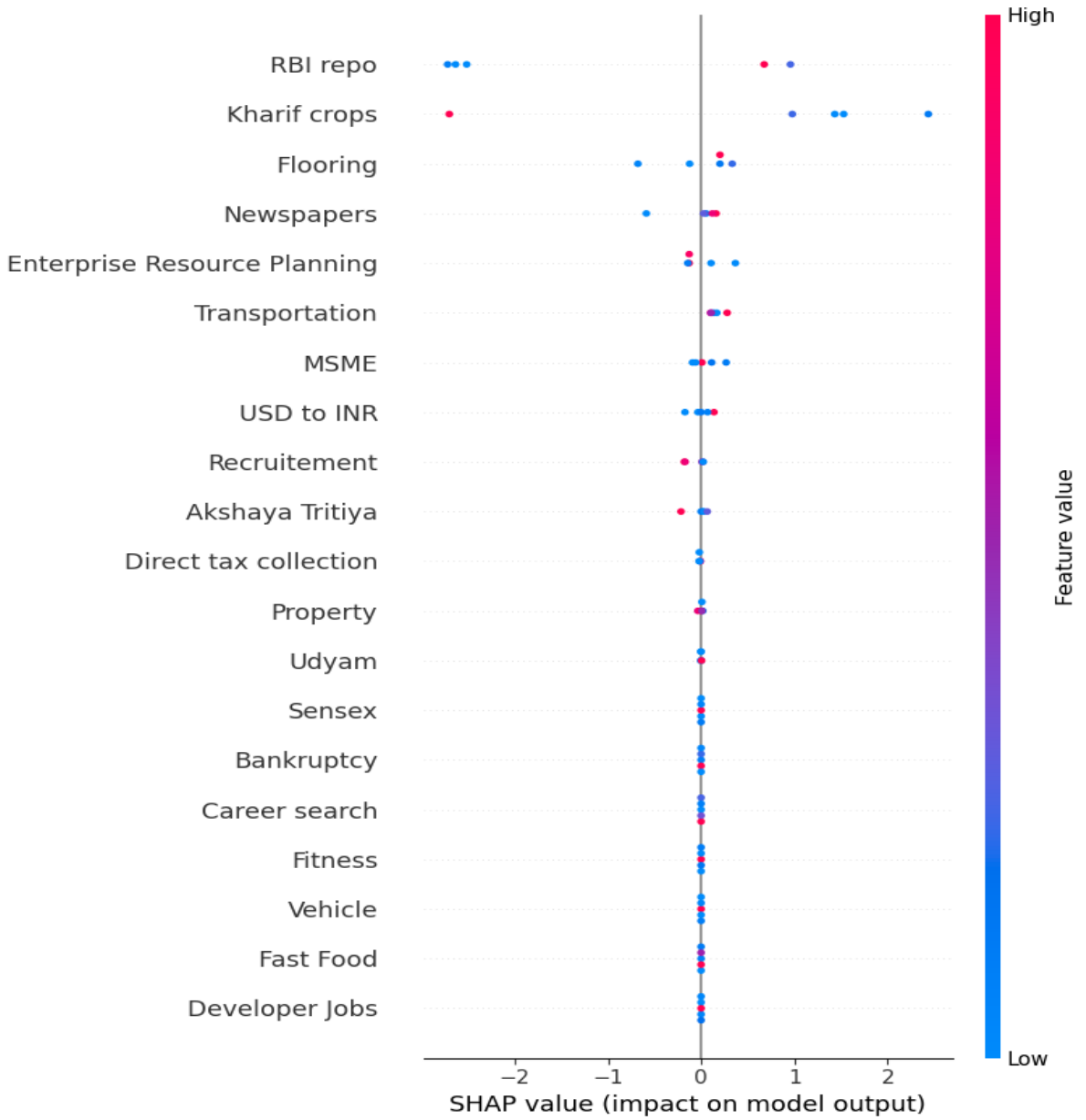
The full monthly CPI series is provided in Appendix A.

5.3.3 Model Interpretability: SHAP Analysis

To better understand the determinants of CPI predictions, we conducted a SHAP analysis offering global, average, and single-instance explanations.

Global Feature Contributions

Figure C.2: Global Feature Importance for CPI prediction



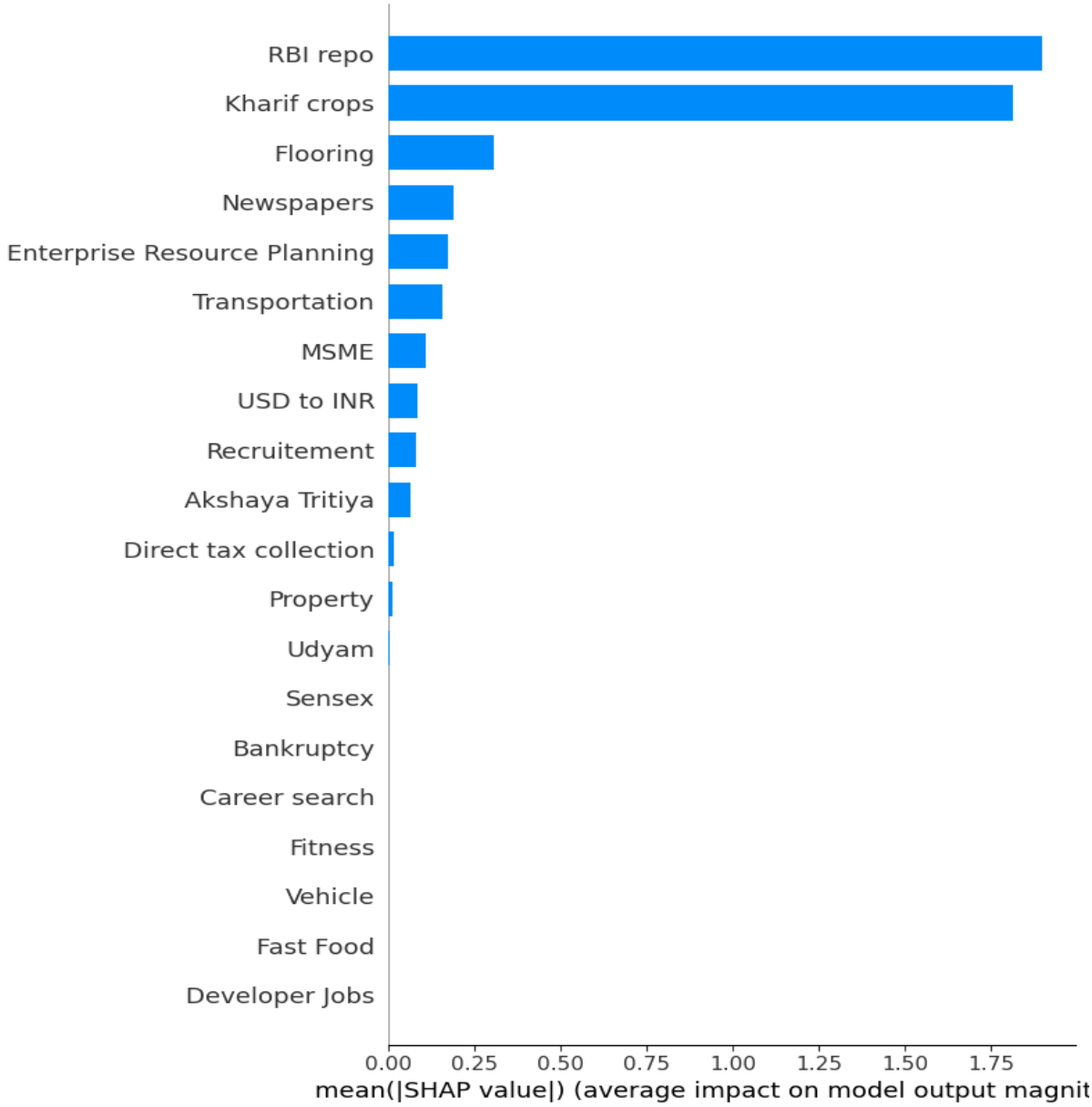
Source: Computed by authors using Google Trends data

Figure C.2 shows that *RBI repo rate* emerges as a dominant feature, consistently exerting a **negative** influence on predicted CPI, reflecting the contractionary effects of monetary tightening on inflation. In contrast, *Kharif crops* exhibit a **positive** contribution, consistent with the role of seasonal agricultural output in shaping food inflation, a major component of India's CPI basket.

Other features, such as *Flooring*, *Newspapers*, *Transportation*, and *MSME*, display both positive and negative SHAP values across observations. This heterogeneity indicates that their effects on inflation vary depending on broader economic conditions, capturing shifts in consumption, supply chains, and business activity.

Mean Absolute SHAP Importance

Figure C.3 Mean Absolute SHAP Value for CPI prediction



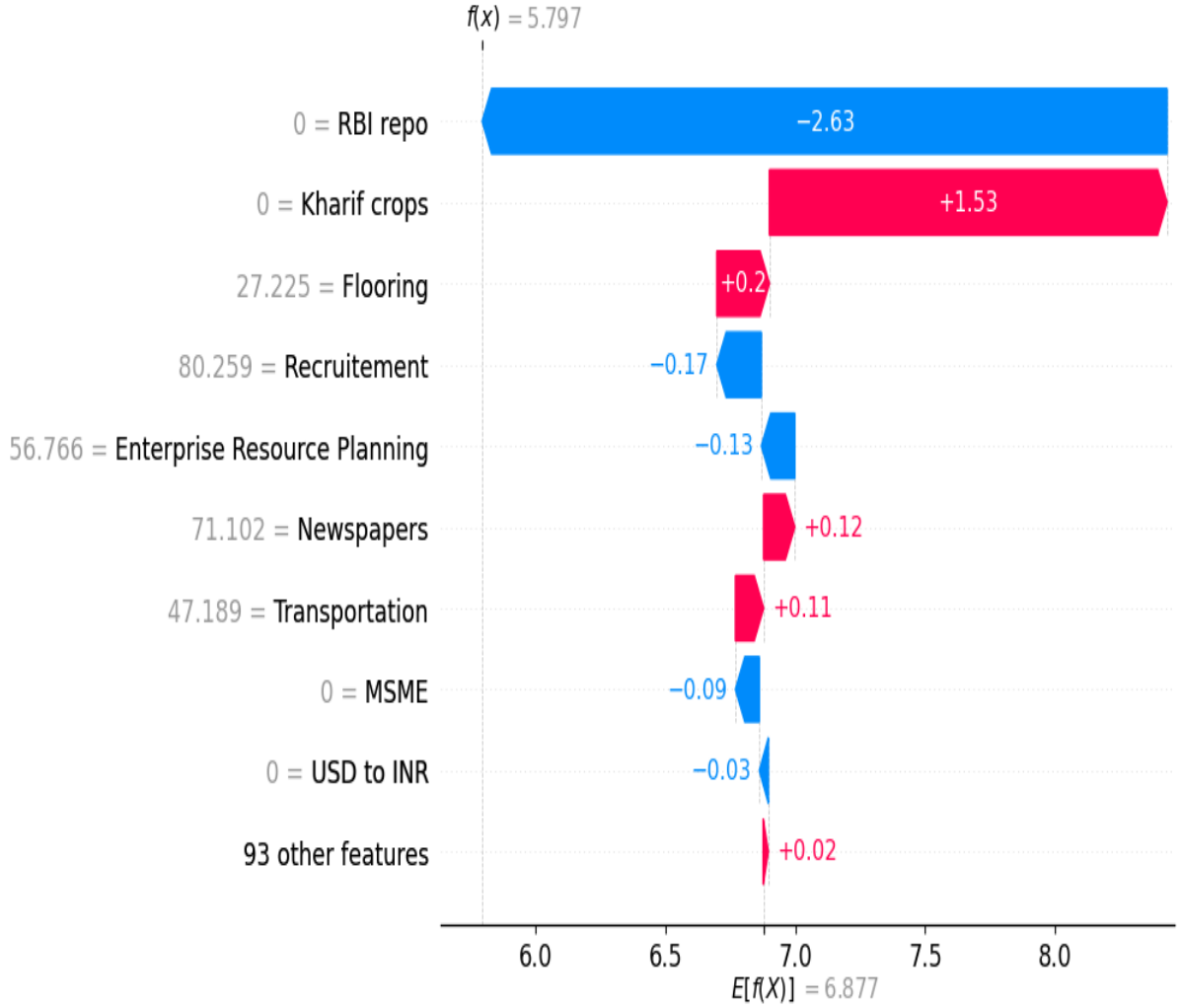
Source: Computed by authors using Google Trends data

Figure C.3 ranks the predictors by their average absolute contribution. *RBI repo rate* and *Kharif crops* are the two most influential variables, indicating that monetary policy and agricultural conditions play central roles in determining inflation outcomes. Secondary contributors include *Flooring*, *Newspapers*, *Enterprise Resource Planning*, and *Transportation*—signals that capture consumption preferences, information flows, corporate activity, and mobility, respectively.

The ranking underscores the model's ability to integrate structural (monetary policy, agriculture) and behavioural (consumer and enterprise searches) factors when forming inflation expectations.

Local Prediction Explanations

Figure C.4: Single Prediction Explanation for CPI prediction



Source: Computed by authors using Google Trends data

From figure C.4 we observe that the *'RBI repo'* exerted the strongest negative influence on the prediction, lowering the model's output significantly. *'Kharif crops'* added positively to the forecast, pushing the predicted CPI upward. Other variables, such as '*Flooring', 'Recruitment',* and '*Enterprise Resource Planning',* had smaller but still noticeable effects.

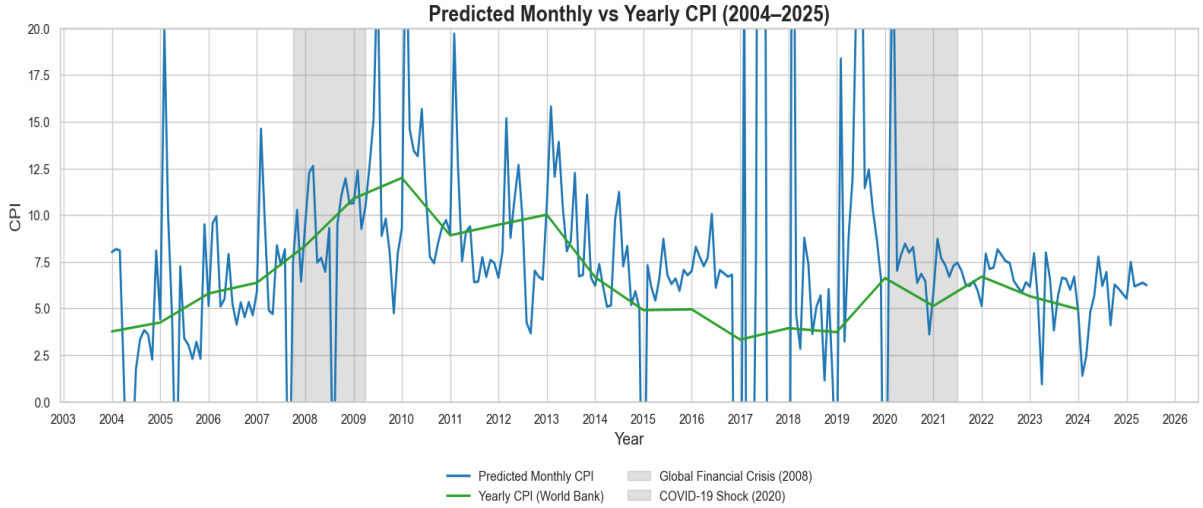
Taken together, the analysis reveals that the keyword *'RBI repo'* plays a central role in shaping CPI outcomes, with consistent downward pressure on inflation predictions. Whereas keyword *'Kharif crops'* tend to push inflation upward. Other indicators, while less dominant, contribute to refining the forecasts in different contexts. These SHAP analysis enhances trust in our machine learning models.

The results emphasize that monetary policy measures and agricultural production patterns remain critical drivers of inflation expectations. Close tracking of these signals, especially during supply-side disruptions, can help fine-tune inflation management strategies

5.3.4 Monthly CPI Dynamics

Figure C.5 compares the predicted monthly CPI series with officially reported annual CPI data from the World Bank. The monthly series exhibits sharper fluctuations, reflecting short-term price volatility that annual data smooth out.

Figure C.5: Predicted monthly CPI values compared with officially reported annual CPI data from the World Bank



Source: Computed by authors using Google Trends data

Notable episodes are captured with greater precision:

Global Financial Crisis (2008–2010):

While the annual CPI rises steadily to nearly 12%, the monthly series reveals sharper spikes and temporary dips, reflecting rapid movements in food and commodity prices.

COVID-19 shock (2020):

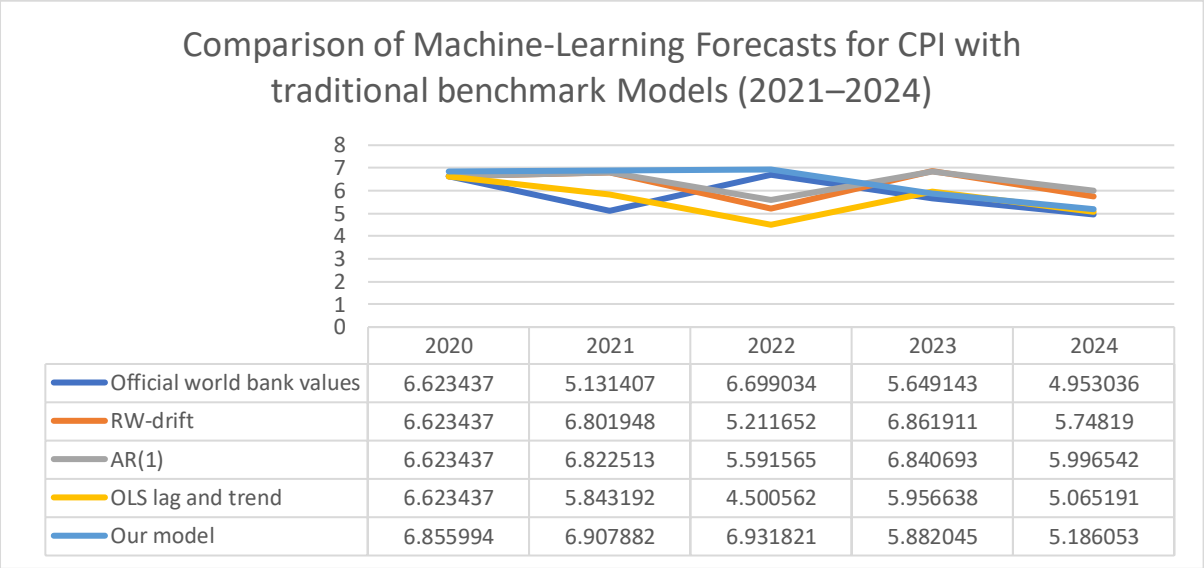
The monthly predictions show erratic and abrupt price changes driven by supply-chain disruptions and fluctuating demand. In contrast, the annual data smooth these variations, showing only the overall inflationary rise.

Over the entire sample, both series exhibit similar long-run movements, rising inflation in the late 2000s, moderation in the early 2010s, and renewed volatility in the late 2010s and early 2020s. However, the high-frequency series detects inflation turning points earlier, underscoring its usefulness for real-time policy analysis.

5.3.5 Benchmark Comparison (Annual Forecast Evaluation)

To compare machine-learning performance with traditional forecasting methods, monthly CPI predictions were aggregated to annual frequency and contrasted with forecasts from RW-drift, AR(1), and OLS lag-trend models. These benchmarks were estimated on official CPI data from 2004–2021 and used to forecast inflation for 2021–2024.

Figure C.6: Comparison of Machine-Learning Forecasts for CPI with Benchmark Time-Series Models (2021–2024)

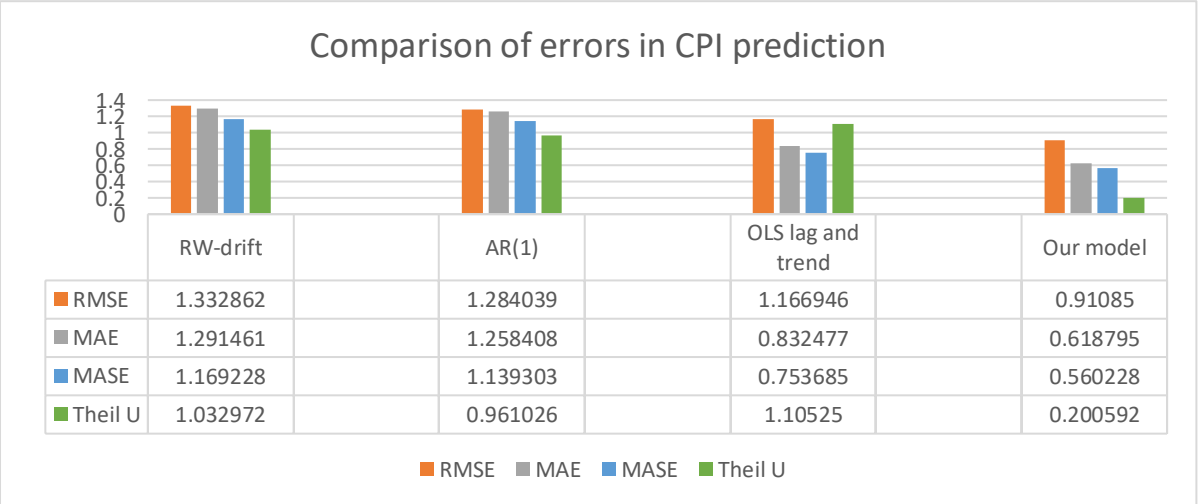


	2020	2021	2022	2023	2024
Official world bank values	6.623437	5.131407	6.699034	5.649143	4.953036
RW-drift	6.623437	6.801948	5.211652	6.861911	5.74819
AR(1)	6.623437	6.822513	5.591565	6.840693	5.996542
OLS lag and trend	6.623437	5.843192	4.500562	5.956638	5.065191
Our model	6.855994	6.907882	6.931821	5.882045	5.186053

Source: Computed by authors

From figure C.6, we can observe that across the four forecasted years, the machine-learning model aligns more closely with official values than the benchmark models, particularly in detecting the inflation spike following the reopening phase and the subsequent decline. This suggests that search-based behavioural data provide additional predictive content not fully captured by traditional time-series approaches, enabling more responsive and accurate short-term inflation nowcasts.

Figure C.7: Comparison of errors in CPI prediction



	RW-drift		AR(1)		OLS lag and trend		Our model
RMSE	1.332862		1.284039		1.166946		0.91085
MAE	1.291461		1.258408		0.832477		0.618795
MASE	1.169228		1.139303		0.753685		0.560228
Theil U	1.032972		0.961026		1.10525		0.200592

Source: Computed by authors

From the figure C.7, we observe that the benchmark models show moderate predictive accuracy, with RMSE values of 1.33 (RW-drift), 1.28 (AR (1)), and 1.17 (OLS), and MAE values between 0.83 and 1.29. Their MASE statistics further reflect the limitations of traditional linear models in capturing the complex, non-linear dynamics of inflation, which is strongly influenced by supply-side shocks and seasonal patterns in the Indian context. In

contrast, our model demonstrates substantially higher accuracy, achieving an RMSE of 0.91 and MAE of 0.62, representing a reduction in forecast error of approximately relative to the best-performing benchmark (OLS). The improvement is also evident in the MASE (0.56), which is noticeably lower than those of all baseline models. Importantly, the Theil U statistic for our model (0.20) is significantly below one, indicating that it outperforms naïve methods and contains meaningful predictive information. Overall, these results confirm that combining Google Trends data with machine-learning techniques yields more reliable and responsive CPI nowcasts than traditional time-series models.

5.3.6 Comparison with the Literature

Our findings show significantly stronger performance improvements relative to earlier studies. Jha and Sahu (2020), using a Hybrid New Keynesian Phillips Curve to forecast inflation in India, report that an AR (1) model yields an RMSE of 3.49 and an MAE of 2.53, which they improve to 2.51 and 2.11 respectively using their enhanced specification. In contrast, our results show substantially better performance: the AR (1) benchmark produces an RMSE of 1.284 and an MAE of 1.258, which our machine-learning model reduces further to 0.911 and 0.619, respectively, these reductions reflect substantial predictive gains and highlight the value of high-frequency behavioural indicators (Google Trends) in capturing inflation dynamics in an emerging economy with strong seasonal and supply-driven characteristics.

6. Model Validation and Scope

To further evaluate the reliability of the our nowcasting framework, we compared the yearly averages of predicted monthly macroeconomic variable with official World Bank data for 2004–2025. For this, two complementary validation approaches were employed:

- Correlation analysis to assess the strength of association between predicted and actual series.
- Simple OLS regression to evaluate explanatory power, using official statistics as the dependent variable and predicted values as regressors.

Table 1 summarizes the pairwise correlations between the predicted yearly averages and official data. Predictions for unemployment (0.94) show exceptionally high correlations, suggesting the models capture the underlying trends effectively. CPI also demonstrates a strong relationship (0.74), while GDP growth rate (0.67) shows moderate alignment, reflecting the difficulty of capturing short-term price and growth dynamics.

Table 1: Correlation between the yearly average of predicted macroeconomic variable and official yearly world bank data

Indicator	Correlation
GDP Growth Rate	0.668
CPI	0.742
Unemployment	0.945

Source: Computed by authors using Google Trends data

We further tested the alignment between predicted and official values using simple linear regressions. Table 2 presents the R^2 values and coefficients. The explanatory power is highest for unemployment ($R^2 = 0.893$), confirming the robustness of predictions for these variables. CPI and GDP growth rate achieve moderate R^2 values (0.55 and 0.45, respectively). In most cases, the estimated slope coefficients are statistically significant and close to one, further validating the predictive framework.

Table 2: Regression between the yearly average of predicted macroeconomic variable and official yearly world bank data

Indicator	Regression R^2	constant	Predicted indicator
GDP Growth Rate	0.446	2.530	0.553
CPI	0.550	1.215*	0.742
Unemployment	0.893	1.067	0.826

Note: * indicates insignificance at 5% level.
Source: Computed by authors using Google Trends data

These results confirm that the framework performs exceptionally well for stock variables such as unemployment, where predictions closely track official trends and capture long-run dynamics. In contrast, GDP growth rate poses greater modelling challenges due to its higher volatility and sensitivity to short-term shocks. CPI forecasts lie between these extremes, reflecting both structural and cyclical influences.

Despite these challenges, the predicted monthly series offer valuable additional insights, especially in times of crisis. They not only mirror official yearly data but also reveal turning points earlier, as demonstrated during the 2008 global financial crisis and COVID-19 pandemic, when the models captured sharp contractions and rapid recoveries that were smoothed in annual estimates. This highlights the potential of search-based behavioural data for early warning systems and real-time policy assessment.

From a scope perspective, this work provides a proof of concept for integrating high-frequency internet search data into macroeconomic forecasting frameworks in emerging economies. The methodology is scalable: it can be extended to state-level indicators, additional high-frequency datasets, and refined ensemble approaches to improve accuracy for inflation and growth measures. By leveraging behavioural signals from search patterns, policymakers can complement traditional statistics with a cost-effective, near real-time decision-support tool that enhances responsiveness to economic shocks.

7. CONCLUSION

Overall, the results highlight the effectiveness of combining Google Trends with machine-learning techniques to generate timely, high-frequency estimates of key macroeconomic indicators for India. The findings show that no single algorithm dominates across all variables; rather, model selection must be indicator-specific, with Gradient Boosting, Decision Trees, and Lasso excelling in different contexts. SHAP analysis adds transparency by revealing that search intensity captures meaningful economic signals, ranging from trade activity and agricultural

output to monetary policy and consumer sentiment, offering a richer understanding of underlying drivers. Comparisons with official World Bank data confirm strong alignment for stock variables such as unemployment, while the high-frequency predictions also capture short-term volatility and turning points that annual statistics tend to obscure. Importantly, the framework proves particularly valuable during periods of economic stress, accurately reflecting the dynamics of the 2008 global financial crisis and the COVID-19 pandemic. Together, these findings provide compelling evidence that search-based behavioural data can complement traditional indicators, enabling more responsive economic monitoring and early policy intervention in data-scarce, fast-changing environments.

When benchmarked against traditional time-series models such as Random Walk with drift, AR(1), and OLS with lag and trend, the machine-learning framework consistently delivers superior accuracy. The gains are most pronounced for GDP growth, where forecast errors decline by more than 80 percent relative to benchmarks, but are also substantial for CPI and unemployment. The ability of high-frequency predictions to identify turning points earlier than annual data further underscores the value of search-based indicators, particularly in emerging economies where data lags and measurement constraints remain significant.

This study also contributes to the growing literature on digital trace-based nowcasting. Consistent with earlier research, the results affirm the strong predictive value of Google Trends for labour-market indicators, while also highlighting the challenges of forecasting highly volatile flow variables such as GDP growth, where search data alone may not fully capture short-lived supply, financial, or commodity-price shocks. Nonetheless, the improvement over conventional benchmarks demonstrate the promise of search-based methods as a scalable and cost-effective component of modern forecasting systems.

Despite the promising results of this study, several methodological limitations must be acknowledged. First, Google Trends data are inherently volatile and subject to sampling variability due to Google's undisclosed data sampling and normalization procedures. This results in fluctuations in search indices for identical queries conducted at different times, posing challenges to reproducibility and reliability. Although the use of anchor keywords helps mitigate batch normalization issues, more robust statistical adjustments and multiple data samplings would enhance the stability of the input features. Second, zero values in Google Trends data may not always indicate an absence of searches but could reflect low search volumes below Google's reporting threshold, which introduces potential missing data bias and may obscure meaningful signals. Third, the relative nature of Google Trends measures means that absolute search volumes and changes in internet penetration or access across different demographic groups are not fully captured, which may limit the generalizability and representativeness of the findings, particularly in emerging markets with uneven digital access such as India. Fourth, while machine learning models provide flexible and powerful tools for nowcasting, they carry risks of overfitting and limited interpretability, which require rigorous validation techniques such as cross-validation and careful interpretation to avoid overstatement of causal relationships. Lastly, the lack of integration with additional high-frequency data sources or survey-based controls constrains the explanatory power of the models, suggesting

avenues for future research to build more comprehensive and representative forecasting frameworks.

Future research could integrate search data with other high-frequency indicators such as mobility, payments, or satellite data to build more comprehensive real-time forecasting frameworks. Such efforts would further strengthen the role of digital trace data in economic surveillance and policy design.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to check the grammar of the sentences. After using this ChatGPT, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

Agnihotri, J., & Mishra, S. (2021). *Indian Economy and Nighttime Lights*. http://arxiv.org/abs/2103.03179

Alberto Cavallo & Brent Neiman & Roberto Rigobon, 2015. "The Price Impact of Joining a Currency Union: Evidence from Latvia," IMF Economic Review, Palgrave Macmillan; International Monetary Fund, vol. 63(2), pages 281-297, September.

Alberto Cavallo & Roberto Rigobon, 2016. "The Billion Prices Project: Using Online Prices for Measurement and Research," Journal of Economic Perspectives, American Economic Association, vol. 30(2), pages 151-178, Spring. https://doi.org/10.1257/jep.30.2.151

Amelie Constant & Klaus F. Zimmermann, 2008. "Face to Face with the Financial Crisis: The U.S. Presidential Election from a Transnational Perspective," Weekly Report, DIW Berlin, German Institute for Economic Research, vol. 4(16), pages 102-112.

Anvik and K. Gjelstad, (2010). "Just Google it." Forecasting Norwegian unemployment figures with web queries, Center for Research in Economics and Management CREAM Publication 11.

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, *55*(2), 107–120. https://doi.org/10.3790/aeq.55.2.107

Bai, J., Ng, S., & Qian, J. (2016). A panel data approach to the estimation of economic indices of regional housing markets. *Journal of Econometrics, 191*(1), 1–14. https://doi.org/10.1016/j.jeconom.2015.10.003

Banbura, M., Giannone, D., & Reichlin, L. (2013). Nowcasting. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 195–237). Elsevier. https://doi.org/10.1016/B978-0-444-53683-9.00004-8

Bantis, E., Clements, M. P., & Urquhart, A. (2023). Forecasting GDP growth rates in the United States and Brazil using Google Trends. *International Journal of Forecasting*, *39*(4), 1909–1924. https://doi.org/10.1016/j.ijforecast.2022.10.003

Bouayad, I., Zahir, J., & Ez-Zetouni, A. (2022). Nowcasting and Forecasting Morocco GDP growth using Google Trends data. *IFAC-PapersOnLine*, *55*(10), 3280–3285. https://doi.org/10.1016/j.ifacol.2022.10.129

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Chen, X. & Nordhaus, W.D. (2011). Using luminosity data as a proxy for economic statistics, Proc. Natl. Acad. Sci. U.S.A. 108 (21) 8589-8594, https://doi.org/10.1073/pnas.1017031108.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., & The Opportunity Insights Team. (2020). How did COVID-19 and stabilization policies affect spending and employment? A new

real-time economic tracker based on private sector data. *NBER Working Paper No. 27431.* https://doi.org/10.3386/w27431

CMIE. (2022). *Consumer Pyramids Household Survey: Methodology and updates.* Centre for Monitoring Indian Economy. https://consumerpyramidsdx.cmie.com

Costa, E. A., & Silva, M. E. (2024). Predicting macroeconomic indicators from online activity data: A review. *Statistical Journal of the IAOS*, *40*(2), 403–419. https://doi.org/10.3233/SJI-230102

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2021). How is machine learning useful for macroeconomic forecasting? *Review of Economic Dynamics, 41,* 225–245. https://doi.org/10.1016/j.reD.3020.09.001

D'Amuri, Francesco and Marcucci, Juri (2012). *The predictive power of Google searches in forecasting unemployment*.

Das, P. K., & Das, P. K. (2025). Artificial Intelligence and Inflation Forecasting: A Contemporary Perspective. *South Asian Journal of Macroeconomics and Public Finance*, *14*(1 Special Issue: Macroeconomic Paradigms since COVID-19 and Emerging Policy), 133–164. https://doi.org/10.1177/22779787251318831

De Bondt, G. J., & Sun, Y. (2025). *Enhancing GDP nowcasts with ChatGPT: a novel application of PMI news releases*. https://doi.org/10.2866/2788332

Diebold, F. X. (2015). *Forecasting in Economics, Business, Finance and Beyond*. Journal of Business & Economic Statistics, 33(4), 465–470.

Erokhin, D. (2025). Applying Google trends to analyze electoral Outcomes: A 2024 cross-national perspective. *Social Sciences and Humanities Open*, *12*. https://doi.org/10.1016/j.ssaho.2025.101846

F. D'Amuri and J. Marcucci, "Google it!" Forecasting the US unemployment rate with a Google job search index, Fondazione Eni Enrico Mattei Working Papers 421 (2010).

Gahlot, S., Mangal, R., Arya, A., Sethy, B. P., & Sethi, K. C. (2025). Prediction of swelling pressure of expansive soil using machine learning methods. *Asian Journal of Civil Engineering*, *26*(2), 549–564. https://doi.org/10.1007/s42107-024-01205-7

Ghani, E., Kerr, W. R., & O'Connell, S. (2016). Local industrial structures and female entrepreneurship in India. *Journal of Economic Geography, 13*(6), 929–964. https://doi.org/10.1093/jeg/lbs032

Ghate, C., Pandey, R., & Patnaik, I. (2013). Has India emerged? Business cycle stylized facts from a transitioning economy. *Structural Change and Economic Dynamics, 24,* 157–172. https://doi.org/10.1016/j.strueco.2012.07.002

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics, 55*(4), 665–676. https://doi.org/10.1016/j.jmoneco.2008.05.010

Götz, T. B., & Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting*, *35*(1), 45–66. https://doi.org/10.1016/j.ijforecast.2018.08.001

Grath, B. (2025). *Machine Learning Based Nowcasting of Swedish Real GDP Growth A Methodological Study on Forecast Accuracy and Model Optimization*.

Guzmán, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. In *Journal of Economic and Social Measurement* (Vol. 36, Issue 3, pp. 119–167). https://doi.org/10.3233/JEM-2011-0342

H. Song, B. Pan and D. Ng, Forecasting demand for hotel rooms with search engine query volume data, College of Charleston Working Paper (2009).

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

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

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review, 102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy.* International Journal of Forecasting.

Hyunyoung Choi & Hal Varian, (2009). Predicting Initial Claims for Unemployment Benefits.

Hyunyoung Choi & Hal Varian, (2011). *Predicting the Present with Google Trends*. http://www.census.gov/retail/marts/www/timeseries.html

Hyunyoung Choi & Hal Varian, 2012. "Predicting the Present with Google Trends," The Economic Record, The Economic Society of Australia, vol. 88(s1), pages 2-9, June. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Kapoor, R. (2020). COVID-19 and the state of India's labour market. *Brookings India Policy Paper, March 2020.* https://www.brookings.edu/research/covid-19-and-the-state-of-indias-labour-market/

Kohns, D., & Bhattacharjee, A. (2023). Nowcasting growth using Google Trends data: A Bayesian Structural Time Series model. *International Journal of Forecasting*, *39*(3), 1384–1412. https://doi.org/10.1016/j.ijforecast.2022.05.002

Li, Y., Ahani, A., Zhan, H., Foley, K., Alshaabi, T., Linnell, K., Dodds, P. S., Danforth, C. M., & Fox, A. (2021). *Blending search queries with social media data to improve forecasts of economic indicators*. http://arxiv.org/abs/2107.06096

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30,* 4765–4774. https://arxiv.org/abs/1705.07874

McLaren, Nick and Shanbhogue, Rachana, Using Internet Search Data as Economic Indicators (June 13, 2011). Bank of England Quarterly Bulletin No. 2011 Q2, Available at SSRN: https://ssrn.com/abstract=1865276 or http://dx.doi.org/10.2139/ssrn.1865276

Medeiros, M. C. ;, & Pires, H. (n.d.). *The proper use of Google Trends in forecasting models Standard-Nutzungsbedingungen*. www.econ.puc-rio.br

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics,* *39*(1), 98–119. https://doi.org/10.1080/07350015.2019.1637745

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Leanpub. https://christophm.github.io/interpretable-ml-book/

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Monge, M., Claudio-Quiroga, G., & Poza, C. (2024). Chinese economic behavior in times of covid-19. A new leading economic indicator based on Google trends. *International Economics*, *177*. https://doi.org/10.1016/j.inteco.2023.100462

Mostbeck, J. (2019). Nowcasting and the use of big data in economic forecasting. *European Central Bank Occasional Paper.* https://www.ecb.europa.eu/pub/pdf/scpops/ecb.op232~123456789.en.pdf

Mr. Paul A Austin & Mr. Marco Marini & Alberto Sanchez & Chima Simpson-Bell & James Tebrake, 2021. "Using the Google Places API and Google Trends Data to Develop High Frequency Indicators of Economic Activity," IMF Working Papers 2021/295, International Monetary Fund.

Mullainathan, S., & Spiess, J. (2017). *Machine Learning: An Applied Econometric Approach.* Journal of Economic Perspectives.

N. Askitas and K. Zimmerman, Google econometrics and unemployment forecasting, Applied Economics Quarterly 55(2) (2009), 107–120.

Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, *130*, 114–122. https://doi.org/10.1016/j.techfore.2017.11.022

Nelson, C. R., & Plosser, C. I. (1982). *Trends and random walks in macroeconomic time series.* Journal of Monetary Economics, 10(2), 139–162.

Niesert, R. F., Oorschot, J. A., Veldhuisen, C. P., Brons, K., & Lange, R. J. (2020). Can Google search data help predict macroeconomic series? *International Journal of Forecasting*, *36*(3), 1163–1172. https://doi.org/10.1016/j.ijforecast.2018.12.006

Niesert, R. F., Oorschot, J. A., Veldhuisen, C. P., Brons, K., & Lange, R. J. (2020). Can Google search data help predict macroeconomic series? *International Journal of Forecasting*, *36*(3), 1163–1172. https://doi.org/10.1016/j.ijforecast.2018.12.006

OECD Economic Outlook, Volume 2020 Issue 2. (2020). OECD. https://doi.org/10.1787/39a88ab1-en

Pavlicek, J., & Kristoufek, L. (2015). Nowcasting unemployment rates with google searches: Evidence from the Visegrad Group countries. *PLoS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0127084

Pratap, B., & Priyaranjan, N. (2023). Macroeconomic effects of uncertainty: a Google trends-based analysis for India. *Empirical Economics*, *65*(4), 1599–1625. https://doi.org/10.1007/s00181-023-02392-z

R. Kulkarni, K. Haynes, R. Stough and J. Paelinck, Forecasting housing prices with Google econometrics, George Mason University School of Public Policy Research Paper 10 (2009).

Real-time nowcasting the monthly unemployment rates with daily Google Trends data. (2024). *Socio-Economic Planning Sciences*, 95.

S. Goel, J. Hofman, S. Lahaie, D. Pennock and D. Watts, Predicting consumer behavior with web search, PNAS Early Edition (2010).

S. Vosen and T. Schmidt, (2011) Forecasting private consumption: survey-based indicators vs. Google Trends, Journal of Forecasting https://doi.org/10.1002/for.1213 .

Saakshi Jha and Sohini Sahu, (2020) "Forecasting inflation for India with the Phillips Curve: Evidence from internet search data", Economics Bulletin, Volume 40, Issue 3, pages 2372-2379

Samanta, G.P. (2019). Does Google Search Index Help Track and Predict Inflation Rate? An Exploratory Analysis for India. *Economics Bulletin*, 39. isi_wsc_62/ips188_paper2

Schmidt, T., & Vosen, S. (2022). *Forecasting Economic Activity Using Google Search Data*. International Journal of Forecasting.

Seabold, S., & Coppola, A. (2015). Nowcasting food prices in developing countries using Google Trends data. *World Bank Policy Research Working Paper No. 7398*. https://doi.org/10.1596/1813-9450-7398

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II* (Annals of Mathematics Studies, No. 28, pp. 307–317). Princeton University Press.

Siliverstovs, B., & Wochner, D. S. (2018). Google Trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior and Organization*, *145*, 1–23. https://doi.org/10.1016/j.jebo.2017.10.011

Stock, J. H., & Watson, M. W. (2007). *Why has U.S. inflation become harder to forecast?* Journal of Money, Credit and Banking, 39(S1), 3–33.

T. Suhoy, Monthly assessments of private consumption, Bank of Israel Discussion Paper (2010).

T. Suhoy, Query indices and a 2008 downturn: Israeli data, Bank of Israel Discussion Paper (2009).

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

Tracking activity in real time with Google Trends (OECD Economics Department Working Papers, Vol. 1634). (2020). https://doi.org/10.1787/6b9c7518-en

Tracking GDP using Google Trends and machine learning: A new OECD model. (2024). *cepr.org*. https://cepr.org/article/tracking-gdp-using-google-trends-and-machine-learning

Vermeulen, P. (2012). Survey expectations, uncertainty and disagreement. *European Central Bank Working Paper No. 1467*. https://doi.org/10.2139/ssrn.2026620

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google Trends. *Journal of Forecasting, 30*(6), 565–578. https://doi.org/10.1002/for.1213

Woloszko, N. (2020). *Tracking activity in real time with Google Trends* (1634; OECD Economics Department Working Papers, Vol. 1634). https://doi.org/10.1787/6b9c7518-en

Woloszko, N. (2024). Nowcasting with panels and alternative data: The OECD weekly tracker. *International Journal of Forecasting*, *40*(4), 1302–1335. https://doi.org/10.1016/j.ijforecast.2023.11.005

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

APPENDIX

Predicted monthly macroeconomic variables using machine learning

Date	Monthly Unemployment	Monthly GDP growth rate	Monthly CPI
01-01-2004	6.447943	6.154936	8.013711
01-02-2004	6.48156	6.163442	8.173734
01-03-2004	6.423637	6.125502	8.100885
01-04-2004	9.403047	9.96037	1.133764
01-05-2004	9.350895	11.45949	-6.3914
01-06-2004	9.661152	9.158618	-3.41786
01-07-2004	9.025867	8.711184	1.749352
01-08-2004	8.443567	7.982067	3.342178
01-09-2004	8.232099	7.666353	3.831685
01-10-2004	8.208889	7.50745	3.600696
01-11-2004	8.351769	8.125812	2.262707
01-12-2004	6.410671	6.139792	8.089196
01-01-2005	7.551986	7.844447	4.418067
01-02-2005	14.35245	9.036709	19.94412
01-03-2005	9.724675	8.129407	9.983638
01-04-2005	7.887795	7.854827	4.315247
01-05-2005	9.340331	11.07821	-5.95907
01-06-2005	9.252169	12.9054	7.249617
01-07-2005	8.160842	8.765381	3.387722
01-08-2005	7.922033	8.462	3.051907
01-09-2005	7.696544	8.555778	2.298872
01-10-2005	8.476869	8.355028	3.205685
01-11-2005	8.371398	8.427967	2.313857
01-12-2005	8.427993	12.25782	9.498322
01-01-2006	7.628573	7.825749	5.136336
01-02-2006	8.350035	7.42468	9.567673
01-03-2006	8.053815	7.753289	9.936483
01-04-2006	7.473703	8.124546	5.100602
01-05-2006	6.965234	7.47803	5.505719
01-06-2006	7.667359	11.71321	7.909238
01-07-2006	7.229018	7.342973	5.218769
01-08-2006	7.713941	8.009647	4.131032
01-09-2006	7.631026	7.465236	5.317934
01-10-2006	8.149796	7.719335	4.534907
01-11-2006	8.101902	8.002559	5.333016
01-12-2006	7.885802	7.86954	4.637912
01-01-2007	7.524105	7.491966	5.897733
01-02-2007	9.324699	6.73325	14.61588
01-03-2007	7.739318	6.983218	9.780937
01-04-2007	7.651331	7.610573	4.891158
01-05-2007	8.510002	9.876271	4.700391
01-06-2007	7.469181	9.65681	8.377543

01-07-2007	7.480787	9.174836	7.31265
01-08-2007	8.207093	10.18229	8.162345
01-09-2007	32.53233	-58.471	-8.68246
01-10-2007	6.99051	7.158071	7.486112
01-11-2007	-17.3224	68.24659	10.26607
01-12-2007	6.852408	7.286916	6.438936
01-01-2008	6.514286	8.671404	9.479987
01-02-2008	8.176683	6.414386	12.27876
01-03-2008	7.890717	5.308377	12.62751
01-04-2008	7.36997	8.757092	7.460908
01-05-2008	7.874747	9.40697	7.700317
01-06-2008	6.729696	8.559368	6.965161
01-07-2008	6.829251	6.809863	9.306188
01-08-2008	11.70819	-41.823	-6.01902
01-09-2008	7.31417	8.118534	9.526007
01-10-2008	7.033593	4.852755	11.06289
01-11-2008	8.135327	5.899382	11.95758
01-12-2008	8.523796	6.065265	10.61902
01-01-2009	8.395553	6.163163	10.63735
01-02-2009	9.472536	4.87938	12.38309
01-03-2009	8.248997	6.346233	9.26692
01-04-2009	8.405726	7.11458	10.47113
01-05-2009	7.626993	8.117773	12.58337
01-06-2009	7.324018	11.09191	15.10866
01-07-2009	10.53319	11.03513	23.54192
01-08-2009	6.750174	8.21091	8.879218
01-09-2009	7.517488	7.423623	9.801572
01-10-2009	6.834845	7.663132	7.992046
01-11-2009	6.215875	9.495207	4.745789
01-12-2009	6.872136	6.801627	7.952654
01-01-2010	7.629836	8.658867	9.272078
01-02-2010	11.97044	7.389067	26.62401
01-03-2010	8.610445	7.949096	14.61225
01-04-2010	8.003144	9.011306	13.43835
01-05-2010	8.104146	10.14221	13.15297
01-06-2010	7.633495	11.61155	15.67129
01-07-2010	7.284762	8.949159	11.15425
01-08-2010	6.982779	7.848463	7.760483
01-09-2010	6.782145	7.704434	7.418778
01-10-2010	6.87395	7.498414	8.478824
01-11-2010	6.968082	7.564321	9.339343
01-12-2010	7.211425	7.644125	9.726915
01-01-2011	7.157517	7.251253	9.034266
01-02-2011	10.07384	6.78384	19.71186
01-03-2011	8.653234	7.19259	12.64648
01-04-2011	6.970208	7.03081	7.524883

01-05-2011	7.397012	8.382462	9.117474
01-06-2011	7.14902	8.804612	9.393261
01-07-2011	10.20688	-0.74206	6.407645
01-08-2011	6.714876	-7.77407	6.436512
01-09-2011	6.890938	6.492532	7.739727
01-10-2011	8.031739	5.737278	6.689553
01-11-2011	6.784831	7.040896	7.58598
01-12-2011	7.593668	6.695653	7.432128
01-01-2012	7.322785	6.453478	6.642727
01-02-2012	7.2369	6.881589	8.048176
01-03-2012	10.92583	1.261028	15.17256
01-04-2012	7.959377	8.146387	8.778396
01-05-2012	8.854584	9.759821	10.80445
01-06-2012	8.725216	10.36171	12.67828
01-07-2012	7.718429	9.260406	9.786265
01-08-2012	8.313607	-0.80492	4.234784
01-09-2012	12.75566	-5.90702	3.669128
01-10-2012	7.473393	7.888186	7.018869
01-11-2012	7.305693	7.253397	6.690505
01-12-2012	8.358987	6.841009	6.543745
01-01-2013	7.503802	4.945365	10.60067
01-02-2013	9.044499	4.62297	15.80865
01-03-2013	9.170205	8.034973	12.05273
01-04-2013	7.63994	3.805359	13.91222
01-05-2013	7.431889	8.527283	10.45645
01-06-2013	8.667031	9.250501	8.064825
01-07-2013	7.783061	7.831064	8.610987
01-08-2013	7.490219	3.562698	12.24745
01-09-2013	7.443641	7.047233	6.721805
01-10-2013	7.532747	7.553795	6.79064
01-11-2013	7.739849	4.702109	11.08491
01-12-2013	7.319091	6.749927	6.644206
01-01-2014	7.116413	7.086409	6.215832
01-02-2014	7.302328	7.279551	7.372326
01-03-2014	6.961878	6.880944	6.209909
01-04-2014	7.744926	7.695371	5.094812
01-05-2014	8.613742	6.493096	5.163011
01-06-2014	7.093061	9.761185	9.77636
01-07-2014	9.22325	7.588689	11.23124
01-08-2014	7.518119	7.61185	7.25605
01-09-2014	8.033147	7.734872	8.343532
01-10-2014	8.530343	6.945237	5.179162
01-11-2014	7.906604	6.707926	5.911913
01-12-2014	8.279272	7.1376	5.016124
01-01-2015	11.08478	9.513717	-12.0152
01-02-2015	8.840455	6.931476	7.316336

01-03-2015	8.535374	7.267571	6.151861
01-04-2015	7.435709	7.802087	5.422573
01-05-2015	7.618395	8.362205	6.602973
01-06-2015	7.957663	9.800316	8.724458
01-07-2015	7.73279	8.281588	6.784293
01-08-2015	6.709762	8.640435	6.3077
01-09-2015	6.82738	8.159299	6.595378
01-10-2015	7.113218	7.044774	5.942649
01-11-2015	7.051473	7.051476	7.059393
01-12-2015	6.901191	7.100102	6.775035
01-01-2016	6.973255	7.105196	6.998806
01-02-2016	8.15143	6.583918	8.298473
01-03-2016	7.538264	6.73648	7.768425
01-04-2016	7.180045	7.75675	7.261066
01-05-2016	6.957575	8.23561	7.715337
01-06-2016	7.403268	10.09946	10.05521
01-07-2016	6.90712	7.207315	6.113762
01-08-2016	6.614564	7.022521	7.046843
01-09-2016	6.696551	6.968633	6.874745
01-10-2016	6.746839	7.276366	6.699815
01-11-2016	4.848252	1.882435	6.799172
01-12-2016	17.43214	22.20098	-19.4679
01-01-2017	17.08488	20.38186	-17.8911
01-02-2017	29.44641	17.62224	23.08333
01-03-2017	14.26395	18.95246	-27.743
01-04-2017	10.16047	33.26035	-11.8517
01-05-2017	4.783722	14.50485	7.895665
01-06-2017	-2.74949	-49.044	72.58013
01-07-2017	2.071904	-62.7454	44.19459
01-08-2017	9.045124	8.962333	-20.1355
01-09-2017	-1.3746	34.27887	-14.1018
01-10-2017	5.246301	15.12728	-10.1413
01-11-2017	-0.95339	4.887995	-0.93927
01-12-2017	6.629136	25.35577	-2.22544
01-01-2018	20.19448	-65.8111	-37.9206
01-02-2018	15.7744	19.267	41.57483
01-03-2018	4.703965	10.57127	4.655142
01-04-2018	10.10082	-5.05532	2.839562
01-05-2018	7.229099	10.22764	8.775379
01-06-2018	6.203686	13.96152	7.367511
01-07-2018	4.768328	10.06827	3.629042
01-08-2018	5.748865	11.91024	5.095564
01-09-2018	5.534415	14.61511	5.690708
01-10-2018	2.61932	25.51667	1.146799
01-11-2018	5.915664	12.2219	6.027337
01-12-2018	5.270473	19.953	1.172549

01-01-2019	2.846633	23.60096	-5.35353
01-02-2019	5.088247	23.77935	18.38026
01-03-2019	4.681228	16.61044	3.231003
01-04-2019	3.427025	24.59127	8.850927
01-05-2019	5.264908	15.23745	11.96101
01-06-2019	5.957943	35.88109	22.57095
01-07-2019	10.12206	24.19583	31.82863
01-08-2019	7.499815	9.824696	11.44504
01-09-2019	7.5849	2.277333	12.42954
01-10-2019	8.452328	7.766244	10.28667
01-11-2019	6.923339	9.990801	8.653513
01-12-2019	6.627148	10.82656	6.628622
01-01-2020	19.58084	-59.4934	-16.9338
01-02-2020	35.75193	-79.8369	10.29217
01-03-2020	-3.83882	0.351811	26.02163
01-04-2020	2.5385	9.249744	7.021422
01-05-2020	4.93461	8.163496	7.863837
01-06-2020	5.275223	7.684122	8.460786
01-07-2020	5.582103	7.469922	7.987399
01-08-2020	5.78336	6.741996	8.280251
01-09-2020	5.299756	7.743032	6.366049
01-10-2020	5.967373	7.642793	6.847397
01-11-2020	5.258356	7.386613	6.452152
01-12-2020	4.416508	7.564031	3.612586
01-01-2021	4.855699	7.285381	5.839281
01-02-2021	5.361915	6.29494	8.719622
01-03-2021	5.607775	6.485948	7.685849
01-04-2021	5.44337	7.308098	7.309778
01-05-2021	4.732413	8.45688	6.697145
01-06-2021	6.29068	4.102021	7.2767
01-07-2021	5.620575	7.307745	7.45107
01-08-2021	5.628931	6.937484	7.020835
01-09-2021	4.922429	7.411078	6.288751
01-10-2021	5.000858	7.297667	6.197277
01-11-2021	5.067535	7.218821	6.445917
01-12-2021	3.962989	7.201375	5.962357
01-01-2022	3.558463	7.066894	5.126751
01-02-2022	5.018921	6.844562	7.929353
01-03-2022	6.177018	3.47705	7.116843
01-04-2022	5.410115	7.35883	7.18123
01-05-2022	5.261316	8.344343	8.162393
01-06-2022	4.80091	9.548421	7.840653
01-07-2022	5.087574	8.285585	7.532548
01-08-2022	5.346968	7.25369	7.425704
01-09-2022	4.783936	8.367359	6.464623
01-10-2022	4.91637	8.056335	6.147307

01-11-2022	4.777627	9.25391	5.857141
01-12-2022	4.968744	7.455401	6.397308
01-01-2023	4.909602	7.991229	6.162097
01-02-2023	4.795311	8.506106	7.948499
01-03-2023	3.679524	10.54286	5.516707
01-04-2023	5.381078	2.659413	0.945629
01-05-2023	4.529687	12.24751	7.989149
01-06-2023	2.421397	14.16989	6.587717
01-07-2023	1.944106	14.86121	3.825017
01-08-2023	4.447164	8.943693	5.696408
01-09-2023	5.254815	7.792132	6.641679
01-10-2023	5.273636	7.214866	6.575783
01-11-2023	4.988459	7.405636	5.999358
01-12-2023	4.684293	7.954518	6.696503
01-01-2024	3.457417	8.469089	4.790571
01-02-2024	4.988808	1.543836	1.396608
01-03-2024	2.098806	10.43262	2.432801
01-04-2024	6.00355	2.199254	4.792859
01-05-2024	4.372551	8.42939	5.682748
01-06-2024	4.862742	8.365921	7.76923
01-07-2024	6.232927	0.662647	6.203765
01-08-2024	5.015404	6.996173	6.942224
01-09-2024	2.784826	8.330198	4.101838
01-10-2024	4.488852	7.739941	6.273168
01-11-2024	4.31411	6.924222	6.058113
01-12-2024	4.050187	7.717398	5.788709
01-01-2025	5.579156	3.115239	5.526466
01-02-2025	4.56552	7.612603	7.486073
01-03-2025	4.614919	6.533411	6.180261
01-04-2025	3.553367	8.889993	6.271065
01-05-2025	4.582292	7.506542	6.370589
01-06-2025	3.985091	8.199782	6.241721