Nowcasting Macroeconomic Indicators using Google Trends

June 21, 2022
Master of Data Science Capstone
University of British Columbia, Okanagan Campus

Submitted By	Project Partners	Instructors
Aishwarya Sharma	Nick Newstead	Irene Vrbik
Harpreet Kaur	Marina Smailes	Firas Moosvi
Jagdeep Brar	(Statistics Canada)	

Image Sources: UBC Logo, Statistics Canada Logo, Google Trends

Table of Contents

Executive Summary

The information on economic indicators is crucial for policymaking and taking decision at right time but this information is usually available with a lag. The nowcasting of economic indicators can help Statistics Canada to fill these lags and provide more timely information for policymaking. The economic indicators are the representatives of economic activity of country which may be captured by Google Trends. Thus, the goal of this project is to develop a methodology to predict macroeconomic indicators such as Gross Domestic Product (GDP), Retail Trade Sales and Retail E-Commerce Sales in real time by using the real time data source, Google Trends. Google Trends provide daily, weekly, and monthly reports on the volume of Google queries related to different industries/keywords which can exhibit the business cycles and provide signals about multiple aspects of the economy that can further be used to estimate the macroeconomic factors in real time. To nowcast these macroeconomic indicators, both econometric and machine learning models are used, and a comparative analysis is provided. The best suited models are chosen for each indicator and the predictions are made to fill up the lagged information. All the obtained results are displayed on an interactive dashboard for easy access and understating of results.

Dashboard: https://nowcasting-indicators-canada.herokuapp.com/

Note: *The GDP values in the entire report need to be multiplied by 1,000,000 to get the actual amount in Canadian dollars. The Retail Trade Sales and E-Commerce sales need to be multiplied by 1,000 to get the actual amount in Canadian dollars.*

1. Introduction

The macroeconomic indicators influence the market movements of a nation and plays a significant role in many analyses. Unfortunately, these indicators exhibit a lag, for instance, the GDP of first quarter is released in second quarter as the evaluation process takes time. This lag causes delay in crucial information which could have been used for policy making or to make other important investment decisions. On the other hand, Google Trends capture interest of people on daily basis so can explain the economic activity of a nation at present as this information is available without any delay. Therefore, the economic activity exhibited by Google Trends can be used to estimate the trends of macro-economic indicators, for instance, the GDP for second quarter may have connection with Google searches in second quarter and these Google searches can be used to predict GDP of second quarter. That also explains why it is “nowcasting” not forecasting as Google Trends of specific month can be used to predict the macro-economic indicators for that month only not for the next month.

Nowcasting is basically predicting the present. This project aims to develop a methodology to predict macroeconomic indicators such as GDP, Retail Trade Sales and retail E-Commerce sales with real-time data source Google Trends for Statistics Canada. The volume of queries for different keywords and categories from Google Trends API serves as the predictors for nowcasting the desired economic factors. The key goals of the project are discussed below:

1) **Nowcasting quarterly GDP**: Our first goal is to nowcast the macroeconomic indicator GDP quarterly by using the real time Google Trends as predictors.
2) **Nowcasting monthly Retail Trade Sales**: The retail sales data are available monthly, so our objective is to nowcast the monthly retail trade sales at national level**.**
3) **Nowcasting retail E-Commerce sales**: The retail sales data are also available monthly, so our objective is to nowcast the monthly retail e-commerce sales at national level.

To fulfil these objectives, two types of models are used (i) Econometric Models which includes Dynamic Factor Model (DFM) and Autoregressive Integrated Moving Average (ARIMA) (ii) Machine learning models which includes Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest and XGBoost in combination with Principal Component Analysis (PCA). Thereafter, the model with least prediction error has been used for three considered indicators and implemented to nowcast the GDP, Retail Trade Sales and E-Commerce Sales and their growth rate as well. The obtained growth rate, values and 95% prediction interval are presented in visual and tabular form on dashboard in a user interactive format.

2. Literature Review

Macroeconomic factors are the key drivers of economy, and their timely information helps in good policymaking. However, this information is available with a lag, for instance, the data for the present month’s GDP is generally published in the coming month/quarter which causes delay in decision-making. To overcome this issue of delayed information gave rise to nowcasting approach. This approach has recently gained the interest of economists and researchers as this approach provides the information on economic indicators in real-time. Traditional macroeconomic indicators have some lag, and to fill this gap of information, Google Trends have been widely used as it may help in predicting the present [1]. The volume of queries on different industries may be correlated with the current level of economic activities in respective industry and may help to predict the subsequent data release [1].

Many researchers have used Google trends for nowcasting the economic activity. Google Trends provide information of business cycles and economic activities in economy and the salient features of these business cycles can be captured with few unknown factors using dynamic factor analysis models [2]. These models are applicable to high-dimensional data and can reduce the dimensionality of economic systems. Dynamic Factor Model (DFM) became the mainstream tool for nowcasting GDP growth over the time. Later on, new techniques emerged, and researchers have started to use machine learning algorithms for nowcasting economic factors. Woloszko [3] proposed a weekly tracker to estimate GDP in 46 Organisation for Economic Co-operation and Development (OECD) countries and G20 countries (excluding European Union). The proposed OECD tracker is based on a machine learning algorithm that estimates the relationship between Google Trends variables and GDP growth.

Dauphin et al. [4] have also used Google Trends data to estimate GDP growth, they provide comparative analysis of different nowcasting approaches such as Auto-Regressive (AR) models, DFM and some machine learning algorithms like Regularized Regression models, Random Forest, Support Vector Machine (SVM) and Neural Networks, and state that there is no one-size fits all model as different models are suitable for different datasets. Richardson et al. [5] used machine learning algorithms to nowcast GDP growth in New Zealand and their results show that machine learning algorithms boosted trees, SVM and neural networks outperformed the traditional AR models for their study. The aforementioned studies indicate that traditional econometrics models and machine learning models both can be used for the nowcasting economic factors, but the success and accuracy of the model may vary for different datasets. Therefore, a comparative study between traditional and modern machine learning algorithms may be more appropriate to fit a model on data in hand.

3. Description of Dataset

As this project aims to nowcast the GDP, Retail Trade Sales and E-Commerce Sales for Canada, we use the data Statistics Canada for these three indicators. The Datasets are publicly available, and a brief description of data is provided below.

3.1. GDP Data

The GDP data is extracted from Statistics Canada website which is comma separated file containing the information about the GDP quarterly. The attribute "**Gross domestic product at market prices**" presents the GDP at national level and is used for our analysis. The time series of GDP value over the years in presented in Figure 1.

Data source: **Gross domestic product (GDP), expenditure-based, Canada, quarterly**

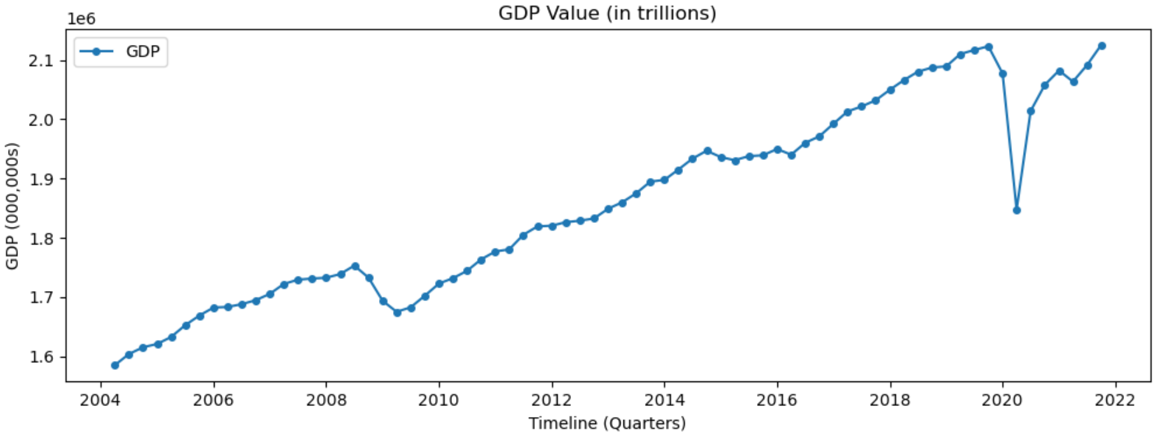


Figure 1: GDP Value over the Years

3.2. Retail Trade Sales Data

The Retail Trade Sales data has been obtained from Statistics Canada and the data can be downloaded as a comma separated file containing the information about the retail sales trades as per the industry. The attributes "**Retail trade [44-45]**" presents the national level Retail Trade Sales and is used for our analysis. The Retail Trade Sales in Canada are depicted in Figure 2.

Data Source: **Retail trade sales by industry**

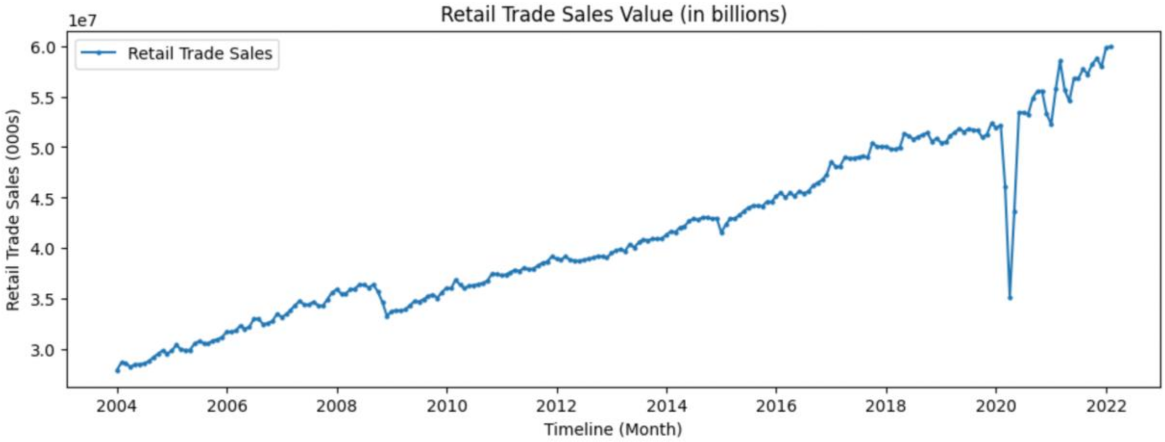


Figure 2: Retail Trade Sales over the years

3.3. E-Commerce Sales Data

This data is also obtained from Statistics Canada as comma separated file containing the information about the retail e-commerce sales. The attribute “**Retail E-commerce sales, seasonally adjusted**” presents the national level E-Commerce Sales and is used for our study. The E-Commerce Sales in Canada are depicted in Figure 3.

Data Source: **Retail E-commerce sales**

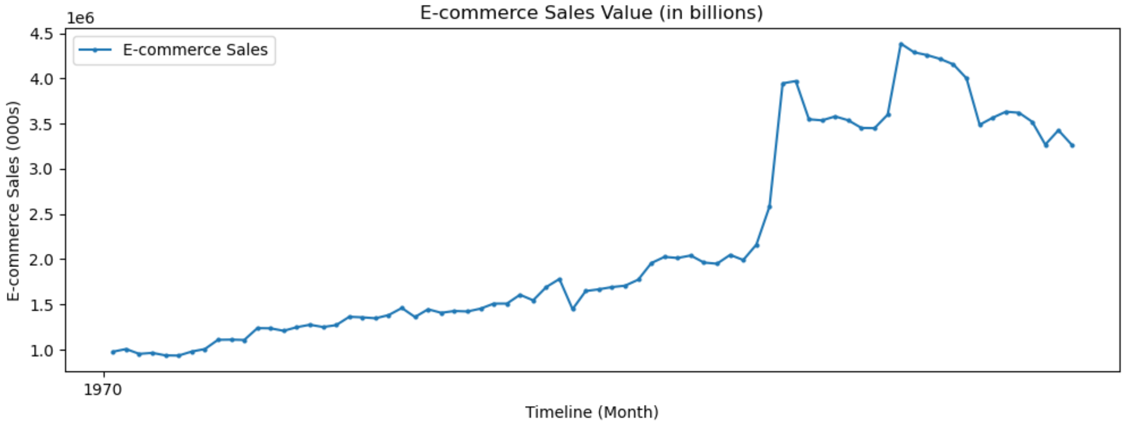
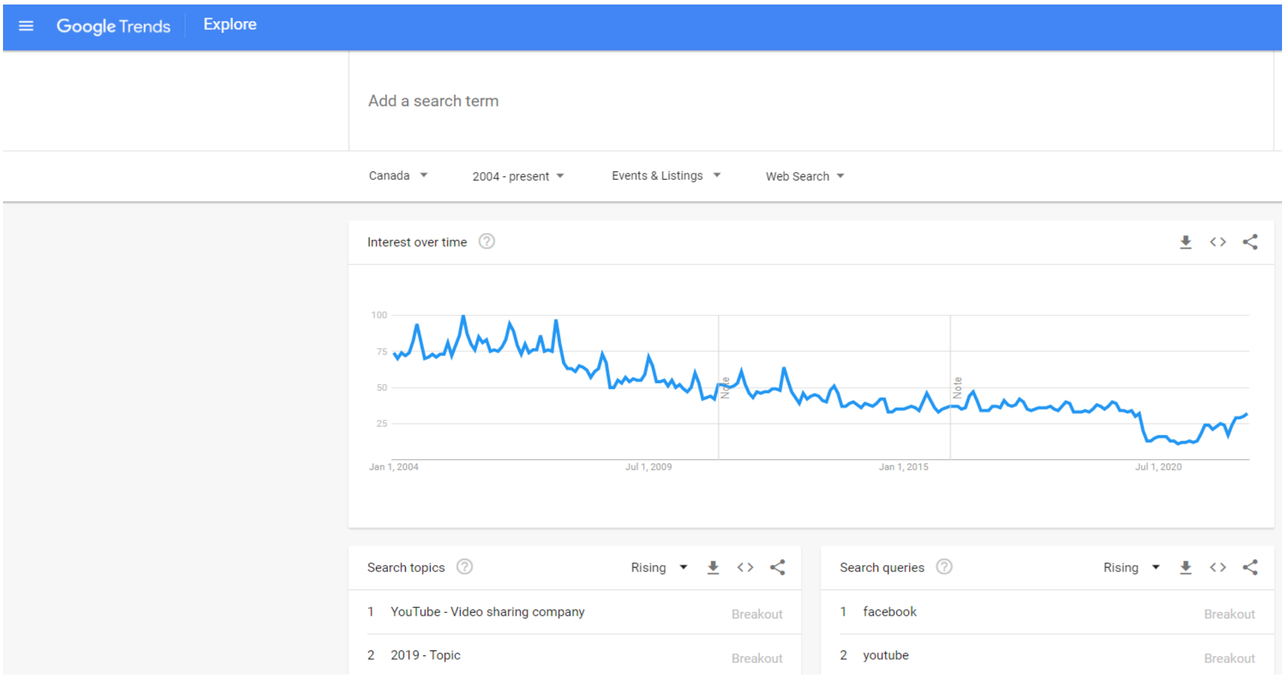


Figure 3: E-Commerce Sales over the years



Canada ▾ 2004 - present ▾ Events & Listings ▾ Web Search ▾

Search topics	
1 YouTube - Video sharing company	Breakout
2 2019 - Topic	Breakout

Search queries	
1 facebook	Breakout
2 youtube	Breakout

Snapshot 1: The demonstration of Google Trends for category “Events & Listings”

3.4. Google Trends Data

Along with Statistics Canada’s data, Google Trends of different categories and queries have been used as predictors in our analysis, for example, some of the queries are “Economic crisis”, “loans”, “GPS”, “unemployment”, “affordable housing”, “economy news”, “agriculture”, “forestry” and many more [Annexure A]. The snapshot 1 exhibits the how the Google Trends look like for a specific category “Events & Listings”.

We use python library “**Pytrends**” that provides access to Google Trends API to extract trends for desired categories and/or queries. The trends for few categories are shown in Figure 4.

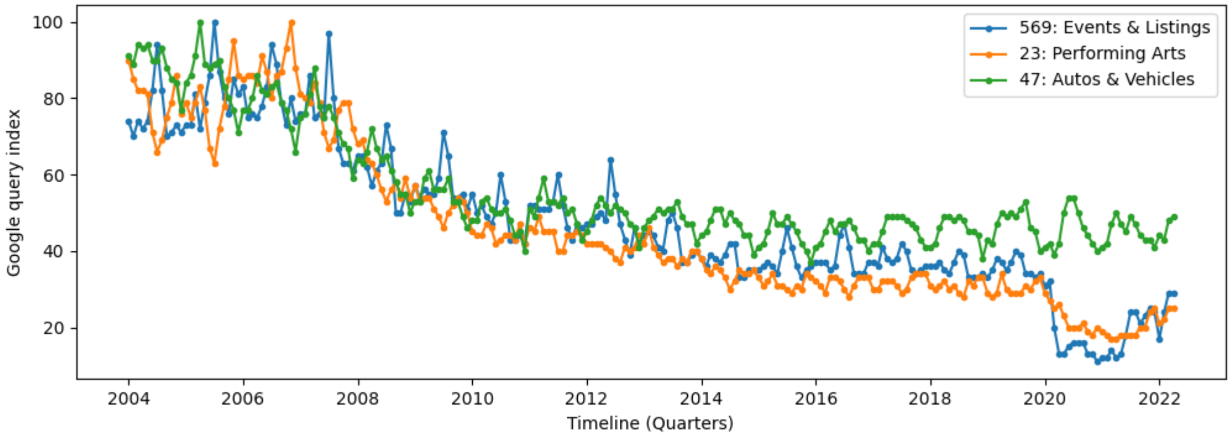


Figure 4: Google Trends of three categories

The y-axis represents Google query index which is calculated as below:

$$Google\ Query\ Index_t = \frac{Number\ of\ searches\ of\ keyword\ \text{"x"}\ at\ time\ t}{Total\ number\ of\ searches\ at\ time\ t}$$

The Google Query Index represents the volume of queries at some time t which is rescaled to take minimum value 0 and maximum 100. It is observant from Figure 4 that Google Trends time series start from 2004 as the Google Trends are available from 2004 onwards. On the other hand, GDP data is available from 1961, Retail Trades Sales available from 1991 and E-Commerce Sales are available from 2016, so to maintain the consistency of data we opted the timelines, mentioned in table 1, to use data for our analysis.

Table 1: Timeline of data selected for analysis

Economic Indicator	Timeline	Frequency
Retail trade sales	2004-2022	Monthly
E-commerce sales	2016-2022	Monthly
GDP	2004-2021	Quarterly

For GDP, we chose 141 categories and top 2 related queries of each category for further analysis and for Retail Trade Sales we consider 36 categories and top five related queries of each category, and 31 keywords has been selected for E-Commerce Sales analysis which comes under "E-Commerce Services". The list of categories and keywords considered is provided in Annexure A and the brief description of data is provided in Table 2.

Table 2: Dimension of the data selected |for analysis

Indicator	Number of observations	Number of predictors (trends)
Retail trade sales	217	396
E-commerce sales	74	31
GDP	72	446

Data Source: **Google Trends API**

We have appropriate amount of data of GDP and Retail Tarde Sales for analysis but the data for E-Commerce is not that much as the timeline starts from 2016. Thus, we implement different methods in order to find the suitable one for all the three indicators. The adopted methods are discussed in the following section.

4. Methodology

In this section, we describe the workflow opted to nowcast the considered macro-economic indicators and the workflow is visually presented in Figure 5. The detailed description on the methodology of the whole project along with the tools and techniques used is as mentioned below.

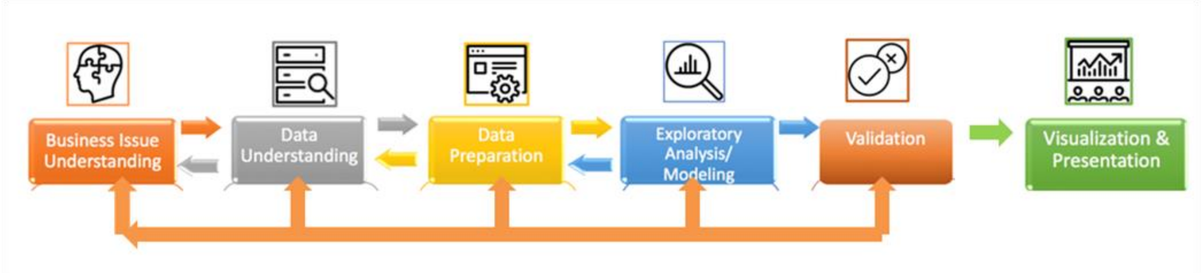

Business Issue Understanding
Data Understanding
Data Preparation
Exploratory Analysis/ Modeling
Validation
Visualization & Presentation

Figure 5: Workflow of the project

4.1. Data Extraction and Cleaning

All the data files required little cleaning and a lot of wrangling. Thus, the first step is to change the format and data type of dates, and the required columns names renamed, and the extra columns are dropped. We need only one column from each file i.e., GDP value, Retail Trade Sales, and E-Commerce Sales at national level. Google trends data was extracted using the python library ‘Pytrends’. The trends data, extracted from Goggle Trends, are in monthly frequency. The fetched trends data was stored in csv files. The link to scripts to access the Google Trends data are provided in Annexure A.

4.2. Data Wrangling and Transformation of Time Series

The GDP, Retail Trade Sales and E-Commerce Sales time series are transformed into growth rate time series. The obtained growth rate time series are stationary which can be used for further analysis. The growth rate is calculated as below:

$$Growth\ rate\ at\ time\ t = \frac{x_t - x_{t-1}}{x_{t-1}},$$

where x_t represents the value of any indicators in month/quarter t. The obtained growth rate time series serve as response variable. The predictors/trends extracted from Google Trends also require the following transformations to make them stationary:

(i) **Normalization**: The time series of different categories and predictors (of monthly frequency) have been normalized to have mean zero and standard deviation one.

(ii) **Detrending**: After normalization, the trends have been removed by using first difference for GDP’s predictors/trends and second order differencing for the predictors of remaining two indicators.

(iii) **De-seasonality**: The predictors exhibit seasonal pattern which we remove by subtracting the average Google Query Index of each month from respective month, for instance, the average value of all the January months in dataset has been subtracted from each January month.

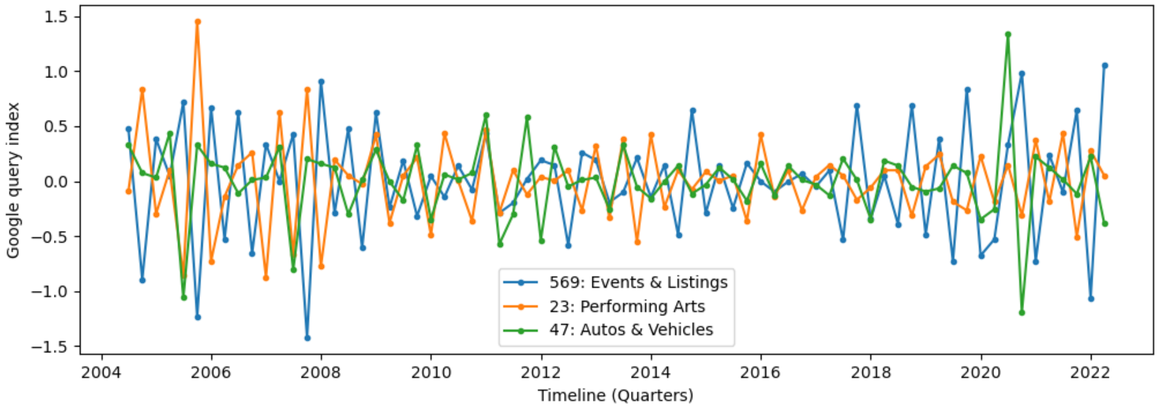


Figure 6: Stationary time series of three categories

These steps provide us the stationary time series and we check the stationarity of time series by using Augmented Dickey Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. The following Figure 6 demonstrates the transformed time series of three categories presented in Figure 4.

The GDP predictors are fetched form Google Trends have monthly frequency like the other predictors, but the GDP data is in quarterly frequency, so the GDP predictors require one more step of wrangling in order to match the frequency of response variable i.e., GDP growth rate.

Transform GDP predictors monthly data to quarterly: Each quarter of GDP data has three months, so we have one observation for three months. On the other hand, we have three Google query indices for respective three months and three query indices are used as three predictors for that one quarter, demonstration of this transformation is provided in the Tables 3 and 4 below for easy understanding.

Later on, we found that only the “Predictor 3” (highlighted in Table 4) of each category is helpful in making predictions, so we dropped the first and second months Google Query Indices of the predictors selected for GDP and only the third months Google Trends have been used for further analysis.

Table 3. Presentation of monthly predictor for quarterly indicator

Response Variable	Google Query Index for predictor category “x”
GDP growth rate quarter 1	January’s query index
	February’s query index
	March’s query index

Table 4. Transformation of monthly predictor to quarterly predictor

Response Variable	Google Query Index for predictor category “x”		
	Predictor 1 of category “x”	Predictor 2 of category “x”	Predictor 3 of category “x”
GDP growth rate quarter 1	January’s query index	February’s query index	March’s query index

4.3. Tools and Technologies

The data set is a time series data and thus required the nowcasting of macro-economic indicators using the real time Google Trends data which are also time series data. As discussed in Section 2, econometric models have been widely used for nowcasting of economic indicators but machine learning models have been outperforming the econometric models for some studies. Thus, we aim to use both the econometric models and machine learning models to estimate the relationship between predictors and macroeconomic factors. The considered models in both categories are depicted in the Figure 7.

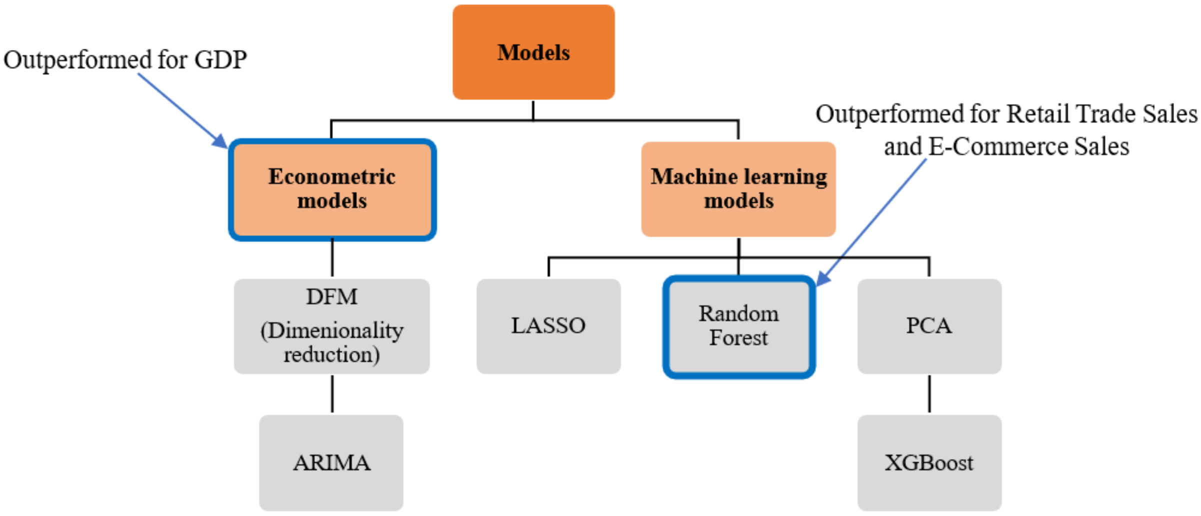


Figure 7: Outline of models used for analysis

Why to choose these models ?

As described in Table 2, the dataset has more predictors than observation for two macro-economic indicators and for the third predictor E-Commerce Sales the timeline starts from 2016. Therefore, we choose models that can work well with more predictors and less samples. The econometric model ARIMA has been an established model for time series data, but it can not be used directly when dataset has more predictors than samples. So, to reduce the dimension of data first DFM model is applied and then ARIMA is implemented on the factors obtained by DFM. The machine learning model LASSO can handle any number of predictors due to the existence of regularization in objective function and it can remove less informative predictors by making their coefficients zero. This feature of LASSO makes it suitable model to try for our analysis. Next, we consider Random Forest as it can also be used when dataset has more predictors than samples as only a subset of predictors is applied on each split of decision trees, along with that it can capture the non-linearity or interaction of predictors well which make it suitable to consider. Lastly, we consider XGBoost as it is well suited model to make predictions and provides the regularization feature as well. However, it can not be applied directly to dataset with more predictors and less observations, so we use PCA first to reduce the dimension of data and XGBoost is implemented on the obtained components.

We provide the comparative study of all the selected models in the coming section and based on their accuracy of predictions we have selected the best suitable model for the three macro-economic indicators. Here, we provide the description of models that outperformed for our considered indicators and used for making predictions.

Model Selection to nowcast GDP growth rate

After making comparison of all the chosen models, we found that the econometric models performed well than other models for GDP growth rate. The model is implemented by using the following steps:

(i) **Prepare predictors data frame**: Prepare one data frame from all the selected categories and queries/keywords trends. We have 141 categories and 282 related top queries (top two for each category) and 23 manually selected keywords, so in total we have 446 predictors for GDP indicator. After our analysis, we realised most of the queries/keywords are just adding noise to data and making the predictions worse and using only categories produce better results. Therefore, we move forward by choosing 141 categories data only to make predictions for GDP growth rate. Then we checked the correlation of the time series of 141 categories with the GDP growth rate time series and selected categories which has positive or negative correlation 0.6 or more. As a result, we are left with 92 categories to consider for model fitting and making predictions. We did made comparison of results if we use all the 446 predictors or all 141 categories or 92 correlated categories then later two outperform than the former (all 446 predictors) and we get almost same results by using 141 categories and 92 categories. We chose 92 categories as we need to apply DFM on the predictors and selecting less predictors makes this model more efficient.

(ii) **Apply DFM**: After choosing the appropriate predictors i.e., trends of 92 categories, DFM is implemented on the 92 predictors and 20 factors are calculated which captures the variation of 92 predictors and reduces the dimension of the predictor space. Then we need to choose appropriate number of factors out of 20 obtained factors as the 20 factors capture almost all the information and we do not require that much information as some of predictors may have nothing to do with GDP growth rate so want to throw some information away. Selection of number of factors is a parameter that need to be tuned and after tuning we get 13 factors (parameter tuning is described in the coming section) exhibit the appropriate information and should be used for further model fit. While calculating the factors the order of factors is chosen to be one so that the resultant residuals has no correlation and the 20 resultant factors will remain time series as well.

(iii) **Testing and training set split**: We use 80% of data as training set and latest 20% of data as testing set.

(iv) **ARIMA model**: After getting the training set, ARIMA model is implemented on the training set and the parameters for lags are selected after observing the lag plots presented in Figure 8. The autocorrelation and partial autocorrelation plots show spikes at order one

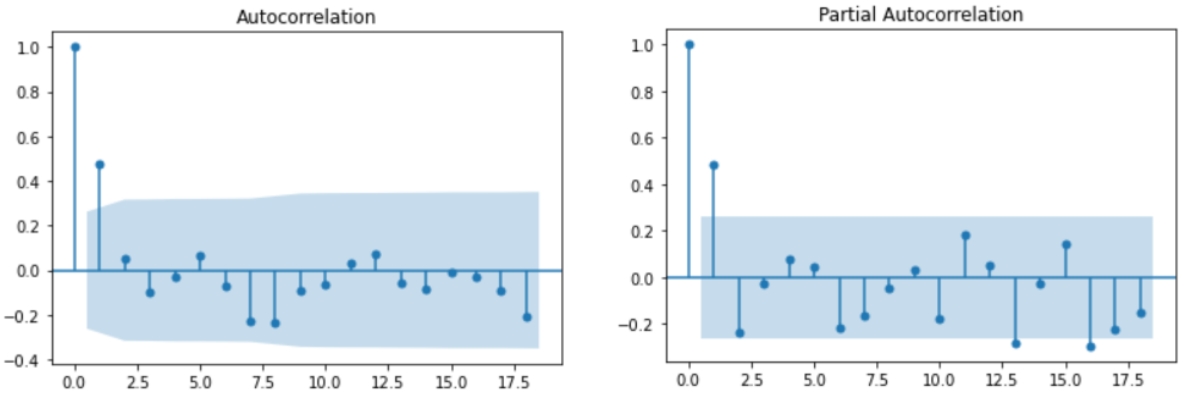


Figure 8: Autocorrelation and Partial Autocorrelation plots

which suggest the moving average (ma) order and autoregression (ar) parameters of order 1 respectively. The "Integrated" order is chosen zero as the seasonality of data has already been removed.

Thus, we implement ARIMA model with parameters ar = ma = 1 and seasonality parameter zero on the 13 factors. The parameter tuning and predictions of this model are discussed in the coming sections.

Model Selection to nowcast Retail Trade Sales and E-Commerce Sales growth rate

Likewise, we implemented all the considered model to predict Retail Trade Sales growth rate and E-Commerce sales and on the basis of prediction accuracy, we observe Random Forest outperformed other models for both indicators and so the steps used to implement Random Forest are described as follows:

(i) **Prepare predictors data frame**: Following the similar path, we prepare one data frame of all the considered categories and related top five queries and got predictors dataset with 396 predictors for Retail Trade Sales. Similarly, we prepare a data frame of all 31 time series for E-Commerce sales to use further for model fitting.

(ii) **Testing and training set split**: As we have more observations (217 observations) for Retail Trade Sales data set, so we can keep a few more data for test set. We use 70% of data as training set and latest 20% of data as testing set for Retail Trade Sales growth rate prediction. On the other side, the data for E-Commerce Sales starts from 2016 that depicts we do not have much observation for this indicator, so we keep 80% data in our training set and use 20% latest data as testing set.

(iii) **Apply random forest**: It is well known that machine learning cannot be applied to time series data directly as they can not capture the correlation between consecutive terms. So, to use machine learning models for time series, we need to add the lagged series of response variable to the predictors in order to capture the timely pattern. For, Retail Trade Sales and E-Commerce sales we added lagged one series of the response variable 'growth rate' to the predictors as the autocorrelation plots suggest. Random Forest model can work with the dataset with more predictors than observations as it tries a subset of predictors for each split in decision trees. So, we do not need to reduce the dimension of data and model can be directly applied to all the predictors. Whereas we need to decide the number of trees to use for model which is parameter that needs to be tuned. The parameter tuning is performed in the coming section, and we select 100 number of trees as it provides a model fit with minimum prediction error for Retail Trade Sales growth rate. On the other hand, we select 600 number of trees for E-Commerce Sales growth rate model as it produces best results comparatively with least prediction error.

4.4. Rolling Prediction

In this section, we made prediction in two ways. First is if prediction is required for next month or quarter (next does not mean future here it means next from the published month or quarter) only and other is if prediction is required for next two months or quarters. These two ways are described below:

One-step ahead prediction

For our analysis, one step ahead rolling prediction makes sense as we just require the information of the next step. Over the time values of the considered macro-economic indicators will be released and can be used to make further prediction. Therefore, we follow one step ahead rolling prediction technique where first we use the training set data to train the respective model and then a prediction is made for the next one step (month for Retail Trades Sales and E-Commerce Sales, and quarter for GDP) using the predictors of next step. Once the prediction is made, the real data of that next step is appended to the training set and model (with same parameters) is fitted again on the appended train set and so on.

Prediction for two or more steps

Sometimes, there is more lag in macro-economic indicators than just one step, for instance, the Retail Trade Sales and E-Commerce Sales are published till March, and we do not any information on April, May sales of the same. So, here two steps are missing, and we need to make prediction for two steps. First step can be predicted as mentioned in above paragraph and for second step, predicted value of first step is appended to the train set (as the actual value of first step is not released yet) and model is fitted again on the appended train set and prediction is made for the second step. This process is followed to make prediction for the lagged values of considered three indicators. Although, the first step can be considered more reliable than the following steps as the following steps are based on the prediction of first step so may have more variability.

For the testing set, one step ahead rolling prediction is used as we already know the actual value of the response variable. Then the model is implemented to nowcast the indicators values for months/quarters we do not have the actual values where we use one step ahead rolling prediction if one value is required and two or more steps prediction of more than one values are required.

4.5. Prediction Error

Prediction error gives us the idea of how much our predictions can vary from the actual values. For prediction error, we have calculated Root Mean Square Error (RMSE) on the test set and the error is computed as below:

$$Prediction\ error\ (RMSE) = \sqrt{\frac{\sum_t (x_t - \widehat{x_t})^2}{\sum_t 1}}$$

Where t is the timeline of the series and the x_t and the actual value of indicator at time t and $\widehat{x_t}$ is the predicted value of the indicator at time t.

4.6. Tuning Parameter

The selected models have parameters that need to be tuned. To tune these parameters the following approaches have been used:

Parameter tuning for DFM + ARIMA

The tuning parameter for ARIMA model is to decide how many '**number of factors**' (obtained by DFM) should be used to implement ARIMA model. To select appropriate number of factors, we looped over all the number of factors starting from 1 to 20 and implemented ARIMA model and compared their prediction error. The prediction error with 13 number of factors is least so 13 factors are used to apply ARIMA and made predictions for GDP growth rate.

Parameter tuning for Random Forest

The tuning parameter for Random Forest model is to decide '**number of trees**' to use for modeling. We have cross-validation technique to tune this parameter. The steps used to cross validation are discussed below:

(i) Split the train data into five folds (five folds are chosen keeping in mind the samples in our data) without any shuffling.

(ii) Apply model on first fold and then make one step ahead rolling prediction for the second folds and store the prediction error for second fold.

(iii) Combine first two folds and apply model in them and make one step ahead rolling prediction for the third fold and store the error for third fold.

(iv) Following the same manner, apply model on $(n-1)$ and make prediction the n^{th} fold and store the prediction error. Here we consider 5 folds so $n = 5$.

(v) The cross-validation error is calculated by taking average of all the $(n-1)$ stored prediction error.

We compared the cross-validation error for the number of trees starting from 100 to 1000 and selected the number of trees that produce least cross validation error. For Retail Trade Sales 100 trees and for E-Commerce Sales 600 trees provide least cross-validation error.

we have performed the comparative analysis. This has provided us the accurate predictions and has let us choose the best model for nowcasting economic indicators. Results for all the models applied along with the one chosen is as shown below:

4.7. Block Bootstrap Resampling

The aforementioned process provides us the one path of prediction for each indicator. To make it robust, we make prediction by using bootstrapped samples as well that further enables us to calculate the 95% prediction interval for each indicator's growth rate. The following steps has been used to get prediction band for growth rate and value of each indicator as well:

(i) We choose block size 9 to get bootstrapped samples from training set of each indicator. The block size seems appropriate to capture the pattern of time series and provide variation in samples The block size chosen by using inbuilt function '**optimal_block_length()**' under package '**arch.bootstrap**'.

(ii) After deciding on block size, we get 100 bootstrap samples for each indicator. (100 and 500 bootstrap sample provide almost same prediction band so we chose to use 100 samples instead of 500 to increase the efficiency of selected models.

(iii) Apply the selected model on each bootstrapped sample make prediction for the test set and we get 100 prediction paths for each indicator as a result.

(iv) Then, 95% prediction interval is obtained by calculating the respective quantiles of the 100 prediction paths. The 95% prediction band around the prediction made with original training set provides us information about the predictions.

5. Analysis and Interpretation of Results

After fitting model on the chosen predictors, in this section a comparative study is performed using the results obtained from both the machine learning and the econometric modelling techniques for all three indicators. The final model for each indicator (discussed in Section 4.3) is selected on the basis of prediction errors obtained from these fitted models. The model selection criteria is the comparison of predicted RMSE. The model with least predicted error was selected for every individual indicator.

5.1. Results Obtained for GDP

The prediction error of the econometric and machine learning models for GDP prediction is depicted in Table 5. As per the presented information, combination of DFM and ARIMA has the least error by using least number of predictors.

Table 5. Prediction error with different model for indicator GDP

Method	Predicted RMSE	Parameter Tuning	Used predictors
DFM + ARIMA	65,511	Number of factors (13)	92 predictors
LASSO	84,146	Penalty parameter	446 predictors
PCA + Random Forest	78,651	Number of trees	446 predictors
PCA + XGBoost	83,641	Number of trees	446 predictors

The predicted growth rate for GDP is presented in the Figure 9. The plot depicts the growth rate rolling prediction obtained using the ARIMA model. The 'red' line represents the fitted values to GDP growth rate which captures the 2008 recession period well.

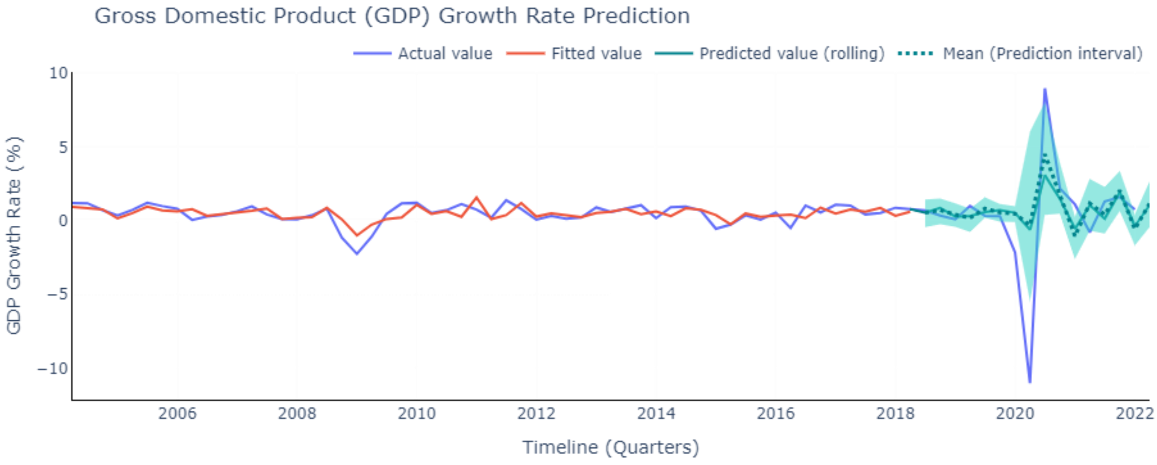


Figure 9: Fitted and predicted GDP growth rate along with prediction interval

The solid 'green' line shows the rolling prediction made by using the training set. The 'green brand depicts the 95% prediction interval which is extracted using 100 bootstrap samples and the dashed 'green' line represents the mean of prediction interval. The model is able to capture the dip in GDP growth rate around 2021 due to pandemic, although this is not exactly same in magnitude but gives us the indication of drop in growth rate due to unseen reasons.

The predicted growth rate used to calculate the GDP value. The fitted and predicted GDP values are depicted in Figure 10 and the 95% confidence interval is also presented.

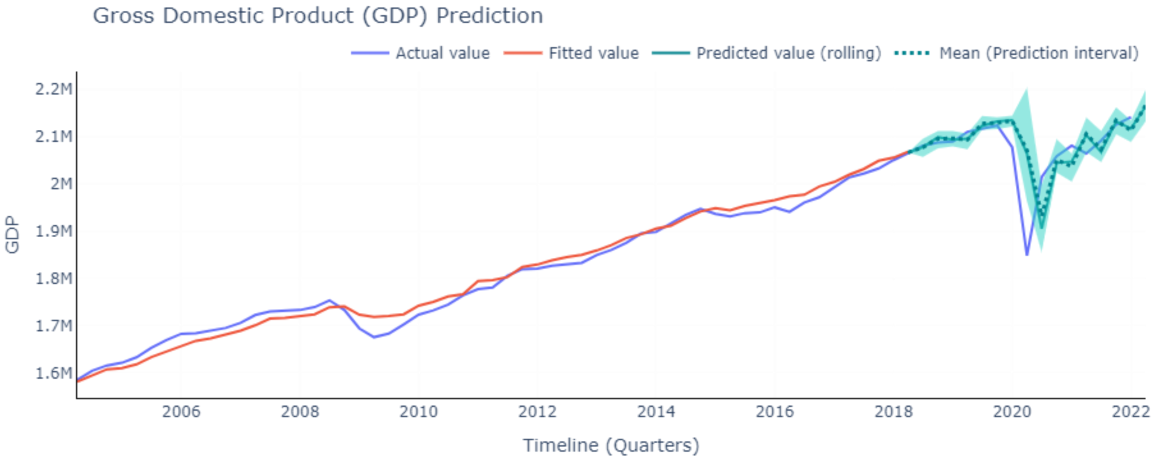


Figure 10: Fitted and predicted GDP value along with prediction band

5.2. Results Obtained for Retail Trade Sales

Following the same path, we select model for Retail Tarde sales by making comparison of prediction error of all the considered models, presented in Table 6.

Table 6. Prediction error with different model for indicator Retail Trade Sales

Method	Predicted RMSE	Parameter Tuning	Used predictors
DFM + ARIMA	2,828,358	Number of factors	396 predictors
LASSO	2,379,342	Penalty parameter	396 predictors
Random Forest	2,281,435	Number of trees (100)	396 predictors
PCA + XGBoost	3,410,734	Number of trees	396 predictors

The prediction error of the model Random Forest model is least so this model is selected for making predictions.

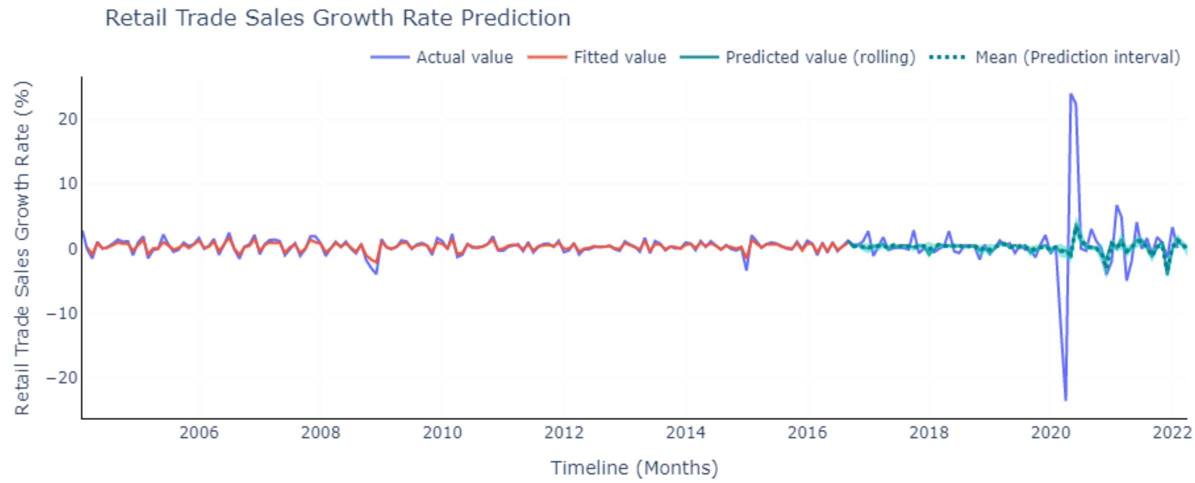


Figure 11: Fitted and predicted Retail Trade Sales growth rate along with prediction band

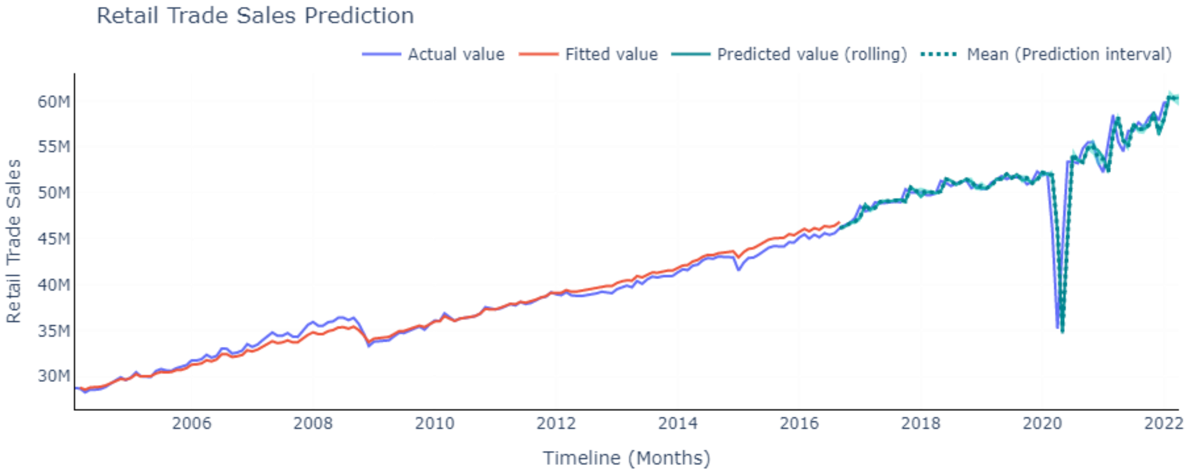


Figure 12: Fitted and predicted Retail Trade Sales along with prediction interval

The obtained results for Retail Trade Sales growth rate are depicted in Figure 11 and the calculated fitted and predicted value for Retail Trade Sales are presented in Figure 12, the prediction band for Retail Tarde Sales is also shown along with its mean. Both the plots exhibit a good model fit and predictions. We can observe the prediction band is not much visible here as the prediction band is really narrow for this indicator, the reason could be the variations in growth rate of Retail Trade Sales are very small which provides us almost similar bootstrap samples and makes prediction band somewhere similar to the predictions made by using training set. The zoomed-in version of plot can be seen at dashboard which provides more detailed information on prediction band.

5.3. Results Obtained for E-Commerce Sales

On the same direction, we select model for E-Commerce sales by making comparison of prediction error of all the considered models, presented in Table 7.

Table 7. Prediction error with different model for indicator E-Commerce Sales

Method	Predicted RMSE	Parameter Tuning	Used predictors
ARIMA	390,077	Number of factors	31 predictors
LASSO	246,766	Penalty parameter	31 predictors
Random Forest	260,128	Number of trees (600)	31 predictors
XGBoost	212,289	Number of trees	31 predictors

The obtained results for E-Commerce sales growth rate and value are demonstrated in Figures 13 and 14 respectively.

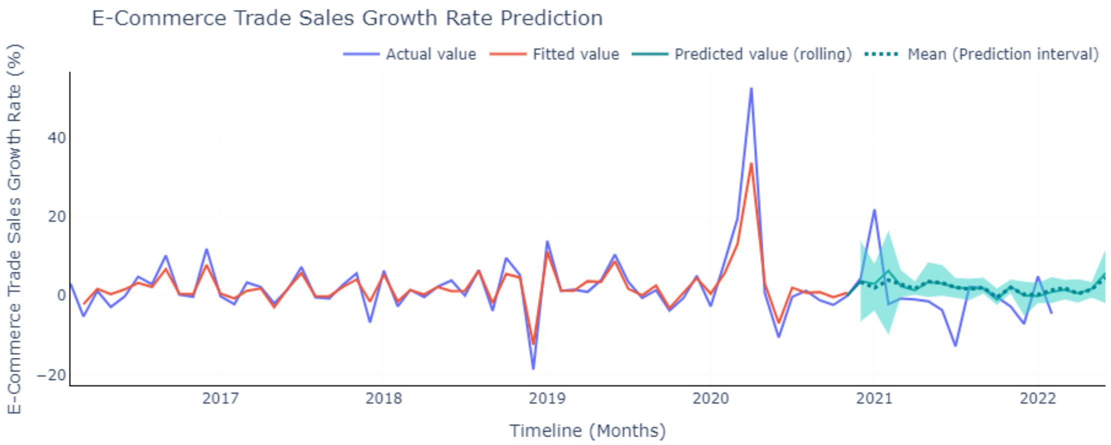


Figure 13: Fitted and predicted E-Commerce Sales growth rate along with prediction band

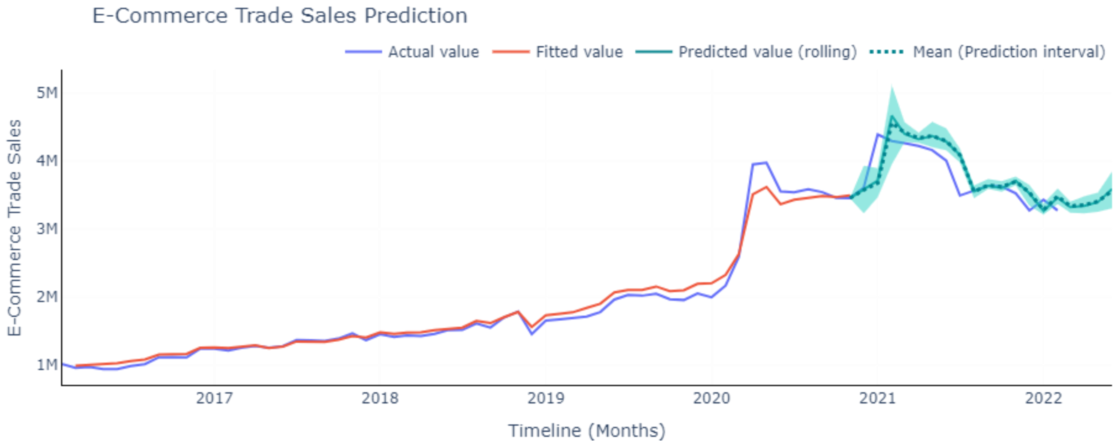


Figure 14: Fitted and predicted E-Commerce Sales along with prediction band

The growth rate prediction follows the trends slightly but not completely as it is really hard to figure out the categories and keywords that may indicate the E-Commerce sales as the E-Commerce sector is relatively smaller than GDP and Retail Trade Sales and less intuitive. We manually extracted 31 keywords which help us to make prediction to some extent but there is room for improvement.

6. Impact of Google Trends

As we use Google Trends data to make prediction of macro-economic indicators, so we wanted to investigate if Google Trends really make any improvement in predictions. To investigate so, we performed check on GDP and used to same model DFM+ARIMA with exactly same parameters firstly considering both lagged series and Google Trends as predictors and then by using only lagged series with no Google Trends. The model fit with Google Trends and without using Google Trends are depicted in Figures 15 and 16 respectively.

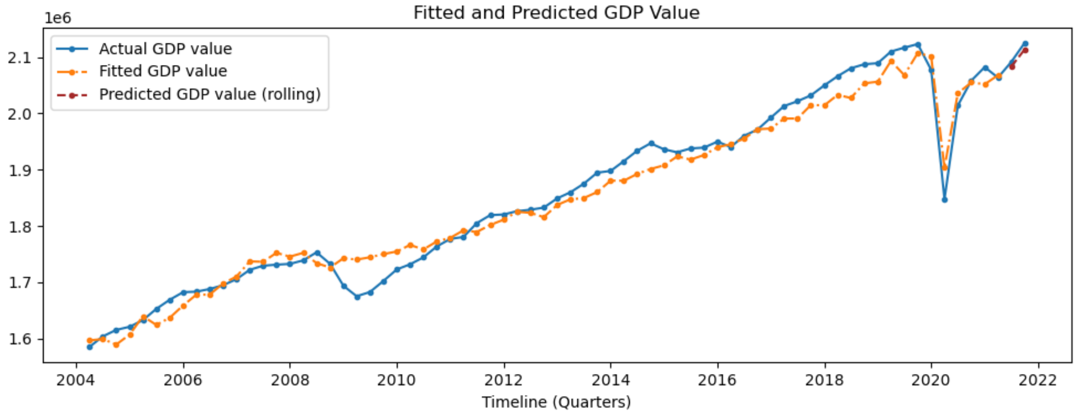


Figure 15: Fitted and predicted GDP by DFM+ARIMA using Google Trends

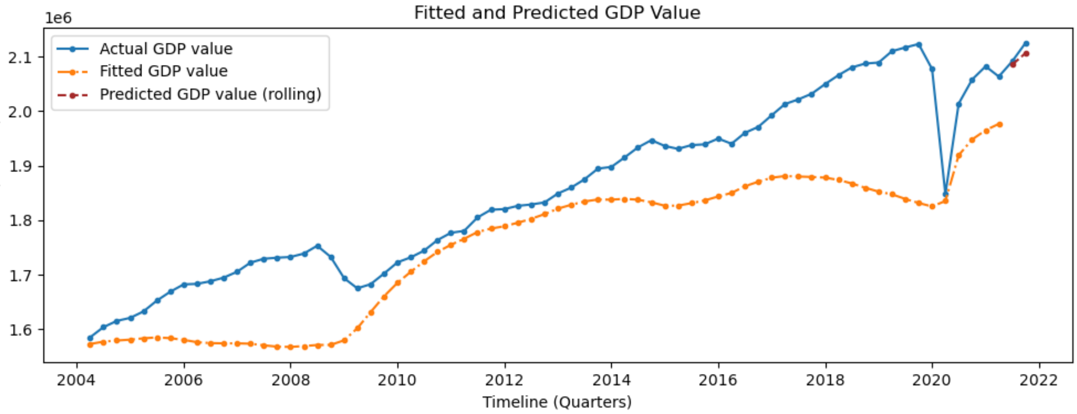


Figure 16: Fitted and predicted GDP by DFM+ARIMA without using Google Trends

We can see the model fit changes a lot if do not change Google Trends and we would not be comfortable to use model fit, shown in Figure 15, for making predictions. Moreover, we have increased train set here in order to include the pandemic period in the model fit which can give us the estimate how the model fit will change if the data sees sudden dip or surge in macro-economic indicators. The figures explains that the model by using Google Trends is comparatively better then the other one.

7. Challenges

We extract Google Trends sample for different categories or keywords or queries to use as predictors for out analysis. It has been in earlier studies that these samples are inconsistent that means there is possibility to change the data if we call Google Trends API on different days, as shown in Figure 17. The figure 17 shows how the trends for same category 'Auto & Vehicles' has changes slightly when we call API on May 11, 2022, and May 12, 2022.

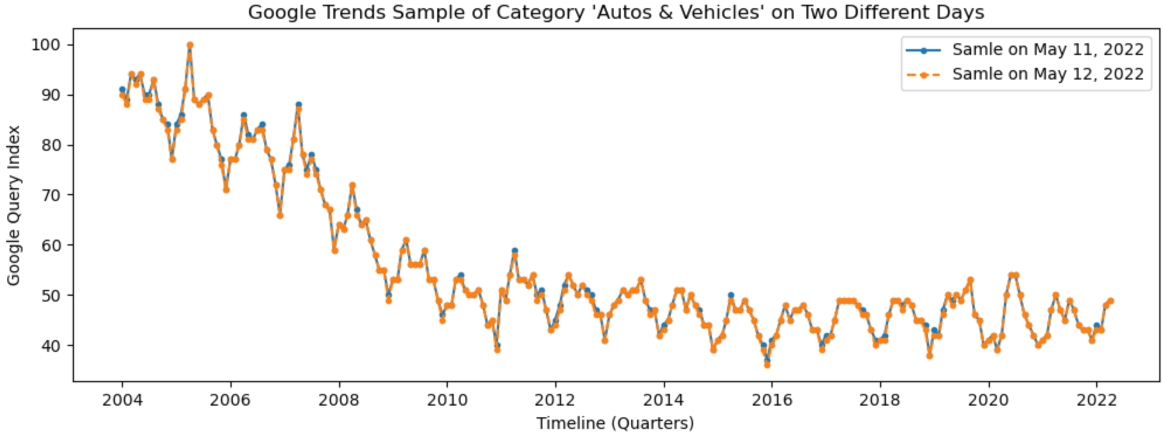


Figure 17: Google Trends for same category on two different days

The ideal way is to extract data of same categories/keywords on different days and take the average of those obtained samples and then use the averaged sample of categories/keywords for further analysis as the averaged sample is consistent. We are extracting monthly data and observe no drastic change in samples on consecutive days, so we have used one sample for our analysis. In addition to that, we have one sample due to time constraint also as we had to wait for 10 days to get 10 samples of the data.

8. Conclusions and Future Directions

This study shows that the use of Google Trends can help in capturing the trends in macro-economic indicators in a nation. This analysis is specifically devoted to nowcast the GDP, Retail Trade Sales and E-Commerce Sales in Canada that can assist Statistics Canada in decisions making and policy making. For making predictions of considered macro-economic indicators both the econometric and machine learning models have been explored and the analysis suggest DFM+ARIMA is best suitable for GDP and Random Forest for remaining two indictors. To get more robust results, prediction band is calculated which ensures the reliability of the predicted values and gives the indication of variability in predicted values.

To recapitulate, macroeconomic indicators are analysed by using a specific set of predictors (keywords, categories, subcategories, queries) in order to nowcast the values. The comparative study of different models helped us to select the respective models for each of the indicator. The obtained results are described by using visualizations and tables in detailed manner on an interactive dashboard that provides easy access to the results.

Through this analysis, we explored econometric and machine learning models and the predicted trends provides the information of economic state of the nation. Although, the predictions are well but there is room for improvement and some more steps may prove fruitful that we were not able to explore due to time constraint. Some of the steps that can be followed in future are mentioned below:

(i) We have many predictors for GDP, 141 categories, 282 top two queries and 23 keywords and we found only 92 predictors useful to make predictions. Although, we tried different combinations of predictors but adding more predictors did not improve our results, but still there is scope to consider some keywords that we might have missed out for machine learning models. Perhaps, considering top 5 related queries and related topics may be helpful as we were not able to collect top 5 queries and topics due to the restrictions on API calls.

(ii) The bootstrap samples for Retail Trade Sales do not vary much, so the obtained prediction band in narrow. There is scope to apply some technique to select the more appropriate block size for bootstrap samples. We explored some block sizes like 5, 7 and 9 but all produced almost similar prediction band.

(iii) The E-Commerce industry is gaining popularity not only in terms of providing convenience to the customers but also the world is moving towards becoming digitalized, there would be other keywords or categories coming up with the expansion in the sector. The current utilises the hot list of such keywords, there are time when people just search for these words to have a peek over creating just a wish list without making any actual purchase such cases need to be handled differently while building a model.

(iv) A dashboard is created to show-case all the obtained results and provide information of the prediction band as well. As more granular data is available for the macro-economic indicators, one new tab can be added to visualize the contribution of different sectors in these indicators and to investigate how this contribution of different sectors has been changed over the time. One more tab can be created which can provide user the ability to train and test the model by using different predictors, that will take 2-3 hours to generate the results as the DFM model takes 10-15 minutes to run and bootstrap for three indicators takes roughly 2.5 hours to generate prediction band. However, if we drop the idea of bootstrap then it can be faster, and user will be able to select the set of predictors and can check how the model fit and predictions change by adding more predictors. Moreover, a dropdown can be provided to select any model among the DFM+ARIMA, LASSO, Random Forest and XGBoost to make predictions.

(v) Moreover, monthly GDP can also be predicted as the predictors are available with monthly frequency and interpolation may be used to get the monthly data for quarterly response variable GDP value or growth rate.

(vi) All the indicators are nowcasted at national level and an attempt can be made to predict the indicators at province level or at industry level. However, it might be hard to figure out the keywords for that, but efforts can be made.

References

[1] H. Choi, H. Varian, Predicting the present with Google Trends, *Economic record, 88 (2012)*, 2-9.

[2] Stock, J.H. and Watson, M.W., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics* (Vol. 2, pp. 415-525). Elsevier.

[3] Woloszko, N. (2020). Tracking activity in real time with Google Trends, OECD Economics Department Working Papers, No. 1634, OECD Publishing, Paris.

[4] Dauphin, M.J.F., Dybczak, M.K., Maneely, M., Sanjani, M.T., Suphaphiphat, M.N., Wang, Y. and Zhang, H., 2022. *Nowcasting GDP-A Scalable Approach Using DFM, Machine Learning and Novel Data, Applied to European Economies*. International Monetary Fund.

[5] Richardson, A., van Florenstein Mulder, T. and Vehbi, T., 2021. Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), pp.941-948.

Annexure A. Important Links

1. Links to the containing list of **Categories and keywords** used for nowcasting three macroeconomic indicators are given below:
   - **GDP**: https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/data/keywords_data/GDP.csv
   - **Retail Trades Sales**: https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/data/keywords_data/RETAIL_SALES.csv
   - **E-Commerce Sales**: https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/data/keywords_data/EECOMMERCE.csv

2. Link to access **Google Trends**: https://trends.google.com/trends/?geo=CA
   Python library 'Pytrends' is used to access the Google Trends data.

3. Link to Github repository to get the North American Industry Classification System (**NAICS**) code of all the categories:
   https://github.com/pat310/google-trends-api/wiki/Google-Trends-Categories

4. Link to our Github repository:
   https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone

5. Link to important scripts in our Github repository:
   - **script1_extractGoogleTrendsData.py**
     This script extracts all the required Google Trends data for all the three macroeconomic indicators and stores it in '/data/storeddata' folder.
     Link: https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script1_extractGoogleTrendsData.py
   - **script2_fitModels.py**
     This script fit model, makes prediction, implements bootstrap and calculates prediction interval for all the three indicators by using the predictors data stored by running script1_extractGoogleTrendsData.py. After fitting models it stores the predicted and fitted data in '/data/storeddata' folder.

Link: https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script2_fitModels.py

- **script3_dashboard.py**

  This script uses the data stored by script2_fitModels.py and created dashboard with the information provided by the predicted and fitted data. To update the dashboard with new predicted data we need to deploy dashboard again on Heroku and this deployment can be automated by choosing option in Heroku account.

  Link: https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script3_dashboard.py