**Text

Description automatically generated**

**A picture containing schematic

Description automatically generated**

**Nowcasting Macroeconomic Indicators using Google Trends**

**Icon

Description automatically generated**

**Submitted By:**

**Master of Data Science Capstone**

**June 20, 2022**

**Submitted By:**

**Aishwarya Sharma**

**Harpreet Kaur**

**Jagdeep Brar**

**Project Partners:**

**Nick Newstead**

**Marina Smailes**

**Instructors:**

**Irene Vrbik**

**Firas Moosvi**

Image Sources: [UBC Logo](https://www.abdn.ac.uk/study/undergraduate/canada-university-of-british-columbia-4250.php), [Statistics Canada Logo](https://crippledscholar.com/2020/05/30/statistics-canada-isnt-collecting-information-on-disability-during-the-pandemic/), [Google Trends](•%09https:/techcrunch.com/2014/04/18/google-trends-debuts-email-notifications-for-search-topics-hot-searches-and-more)

**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. So, the need for alternative data sources is of growing importance for both supplementing Statistics Canada’s data holdings and for nowcasting economic activity. The goal of this project is to develop a methodology to predict macroeconomic factors 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 provides daily, weekly, and monthly reports on the volume of Google queries related to different industries which can help to understand the business cycles and provide signals about multiple aspects of the economy that can further be used to estimate the economic factors in real time. The nowcasting of economic indicators will provide more timely information for policymaking.

**Methods and Results:** To nowcast the indicators, initially time series data was made stationary by removing the trends, normalizing the data and by removing the seasonality. After achieving the desired stationary data, different models, both econometric and machine learning were applied. Since, the number of predictors were more than the samples, Dynamic Factor Model (DFM) and Principal Component Analysis (PCA) were used. Comparative analysis was done in order to choose the perfect model for all the three indicators as per the prediction error and model fit.

DFM along with ARIMA best suited for the prediction of GDP and Random Forest was most appropriate for retail trade sales and E- Commerce.

This nowcasting was made user interactive by creating a dashboard where user can select the indicator and year range and could get the predictions in a table along with the time series plot for the predicted value and growth rate for all the three indicators.

**Conclusions drawn:** The nowcasted value for all the three indicators were done successfully and accurately. Nowcasted value for GDP as of April was predicted as ‘2,163,537.05’, retail trade sales value for March and April 2022 was predicted as ‘60,305,994.38’ and ‘60,411,554.47’ respectively and finally for E- Commerce for March to June 2022 were nowcasted as ‘3,315,434.35’, ‘3,333,537.08’, ‘3,388,988.16’ and ‘3,581,841.75’.

1. **Introduction**

Nowcasting is basically predicting the future. 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**. The volume of queries for different keywords and categories from Google Trends API has served 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 variables 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 available monthly, so our objective is to nowcast the monthly retail e-commerce sales at national level.

To predict the indicators, python libraries are used to collect the data as predictors from google trends and the nowcasted value is displayed in a tabular form in a dashboard. The dashboard created is user interactive having all the information of how to operate stated there on dashboard. In addition to this, time series plots are shown which will display the growth rate and predicted value for all the three indicators.

1. **Background and Related Work**

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 hasrecently 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.

1. **Data**

Data set for this project are open ended and the short description about data is provided below:

1. [**Gross domestic product (GDP), expenditure-based, Canada, quarterly**](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610010401)**:** This dataset is comma separated file containing the information about the GDP quarterly.
2. [**Retail trade sales by industry**](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2010000802)**:** This is a comma separated data set containing the information about the retail sales trades as per the industry.
3. [**Retail E-commerce sales**](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2010007201)**:** This is a comma separated dataset containing the information about the retail e-commerce sales.
4. [**Google Trends API**](https://trends.google.com/trends/?geo=CA): We have accessed Google Trends website to get real time data for the macroeconomic indicators. Different keywords, categories and subcategories are used to extract Google Trends predictors such as Economic crisis, loans, GPS, unemployment, affordable housing, economy news, agriculture, and forestry [3, Annexure B].

The timeline and frequency for the data extracted from statistics Canada website is as shown below:



Our focus was on the data starting from 2004 for GDP and retail trade sales data as we have Google trends available from that period. The e-commerce data is available from the year 2016. This range of data has provided us enough information for nowcasting.

All the data files required wrangling. The format and data type of dates are adjusted, and column names were renamed. Also, few columns were removed which were not that useful. Google trends data was extracted using the python library ‘Pytrends’. The data fetched from google trends was stored in a csv files and that could be downloaded directly from the dashboard. The links to access the script for accessing the data and the way data was wrangled are :

* To get data from Google Trends: <https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script1_extractGoogleTrendsData.py>
* To wrangle data : <https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script2_fitModels.py>

The code is written in a scripts file so that it is easy to run and understand.

The information for the data selected for all the three indicators is as shown below:



Table \_ : Indicators data summary

1. **Tools, Methodology and Techniques**

The detailed description on the methodology of the whole project along with the tools and techniques used is as mentioned below and the workflow is also depicted in Figure 1.

Diagram

Description automatically generated

**Figure 1**. Workflow of the project

1. **Data Cleaning and Wrangling**

The dataset for GDP, retail sales and e-commerce sales is filtered so that it contains information starting from the year 2004. Moreover, the data was time series and therefore, rolling predictions were done in order to capture the trend which was varying with time. In addition to this, growth rate was also calculated, by taking the difference from the previous day, to get the accurate data. Finally, it was made stationary by normalizing the data, by removing the trend and by removing the seasonality.

Google trends data was retrieved by using the python library ‘Pytrends’ and different keywords were stored which served as a predictor for the nowcasting. Furthermore, Google Trends data was also time series data and hence was made stationary.

1. **Model Fitting**

The data set is a time series data and thus required the nowcasting using the real time data from Google Trends. The volume of real time keyword queries from the Google Trends are used as the predictors to nowcast the macroeconomic indicators. As mentioned in Section 1, the econometric models and machine learning models both are used to estimate the relationship between predictors and macroeconomic factors. Following the similar path, we used methods of both categories to fit our data. The methods implemented are discussed below:

* **Econometric models**: Since the data is time series and DFM have been widely used for such analysis, we have used DFM which has served as a dimension reduction method and was helpful in analysing the time series data. Also, have used ARIMA in order to predict the factors.
* **Machine learning models**: The literature claims that machine learning models have been serving well to estimate abrupt changes in economic activity. So, we have used LASSO, Random Forest and XGBoost in order to nowcast the indicators.

1. **Validation**: After implementing the above-mentioned models 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:
2. Retail Trade Sales: Different models were applied but the best one was random Forest.



**Table \_ :** Retail Trade Sales model fit summary

1. E- Commerce: Different models were applied but the best one was random Forest.



**Table \_ :** E-commerce model fit summary

1. GDP: Different models were applied and the best one was DFM with ARIMA.



**Table \_ :** GDP model fit summary

The code is well organized and is written in three scripts.

Script 1: This is executed to extract the data from the google trends and store that in a csv format.

Script 2: This is executed to clean and wrangle the data. In addition to this, this script is executed to fit the models.

Script 3: This is executed in order to create the dashboard.

# Analysis, and Interpretation

After performing the data wrangling and model fitting on the Google PyTrends time series and the Statistics Canada 18years historical retail and e-commerce sales and GDP values, a comparative study was performed using the results obtained from both the machine learning and the econometric modelling techniques for all three indicators.

The model was selected on the basis of prediction errors obtained from these fitted models. The model selection criteria were the Root Mean Squared Error (RMSE). The model with least predicted error was selected for every individual indicator.

The study uses Google PyTrends which are the searches made by the user over the Google with respect to the relevant categories, sub-categories and keywords used for any of the macroeconomic indicator. The plot below shows the nowcasting extracted by just building model over the historical data of the GDP indicator.

**Chart, line chart

Description automatically generated**

**Figure \_ :** GDP fitted and predicted value without using Google trends

As it is clearly visible that the model is under-fitting the data and one can differentiate a clear difference in the model fit (Figure ). The model is capturing the trend within the data. Every variation has been captured. It is interesting to see how the pandemic and the recession dip have been captured just with the use of Google trends.

Chart, line chart

Description automatically generated

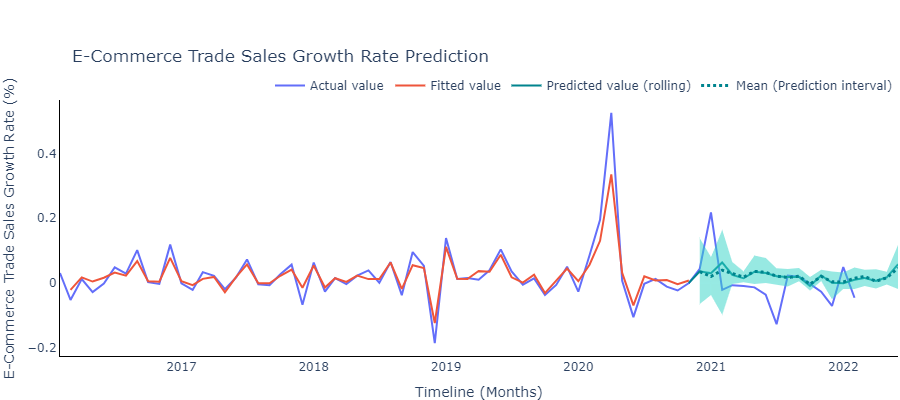
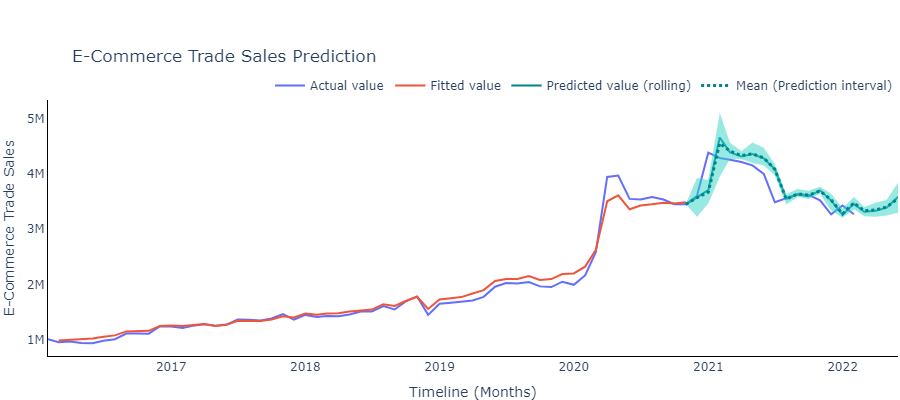
**Figure \_ :** GDP fitted and predicted value with Google trends

## E-commerce Sales

The indicator included Google PyTrends keywords with respects to a single category ‘E-commerce Services’ with id ‘340’. There were less predictor than the number of observations; therefore, we have applied random forest directly to the keyword`s time series. The prediction errors obtained from the models are depicted in the table below. As per the presented information Random Forest model had the least error with \_ number of fitted trees , these were selected after performing cross validation over the model.

|  |  |  |
| --- | --- | --- |
| **Method** | **Prediction Error (RMSE)** | **Parameter Tuning** |
| ARIMA | 390,077 | -- |
| LASSO | 246,766 | Penalty parameter |
| Random Forest | 260,128 | Number of trees |
| XGBoost | 212,289 | *In progress* |

The plot depicts the growth rate`s rolling prediction achieved using the random forest machine learning model. It is interesting to see how the keywords have all together contributed towards the sales prediction. The sudden dip in the market trend due to certain unforeseen events have been captured my model using the change in the search trend over the Google with respect to this category.



**Figure \_ :** E-commerce`s growth rate fitted and predicted

**Figure \_ :** E-commerce sales fitted and predicted values

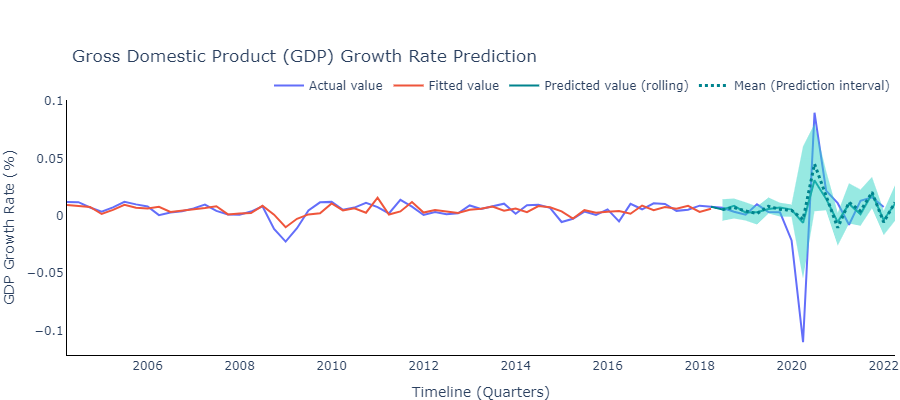
The plot is presenting the actual e-commerce retail sales value calculated using the growth rate`s variation in predictions as provided in the above graph. The green brand depicts the 95% prediction interval which is extracted using \_ bootstrap samples.

## GDP

The indicator included Google PyTrends keywords with respects to a single category ‘E-commerce Services’ with id ‘340’ [Appendix ]. The prediction errors obtained from the models are depicted in the table below. As per the presented information Random Forest model had the least error with \_ number of fitted trees , these were selected after performing cross validation over the model. The predictors for the data were comparatively more than the observation count; therefore, applying DFM with ARIMA modelling technique was most optimal in capturing the trend over time. The number of trees fitted for the model was \_, these were selected after performing cross validation over the model.

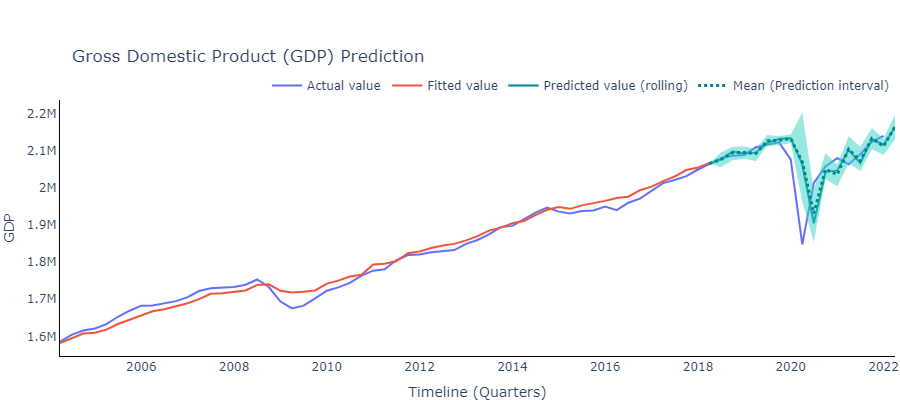
|  |  |  |
| --- | --- | --- |
| **Method** | **Prediction Error (RMSE)** | **Parameter Tuning** |
| DFM + ARIMA | 65,511 | Number of factors |
| LASSO | 84,146 | Penalty parameter |
| PCA + Random Forest | 78,651 | Number of trees |
| PCA + XGBoost | 83,641 | *In progress* |

The plot depicts the growth rate rolling prediction achieved using the ARIMA model. DFM helped in dimension reduction which contributed towards achieving an apt ARIMA model capturing the recessional (2008) and the pandemic (2020) impact over the GDP of the country.



**Figure \_ :** GDP growth rate`s fitted and predicted values

The plot is presenting the actual GDP values calculated using the growth rate variations predictions as provided in the above graph. It is interesting to see how the model has captured the unforeseen/sudden variations using the Google trends of the related categories. The green brand depicts the 95% prediction interval which is extracted using \_ bootstrap samples.



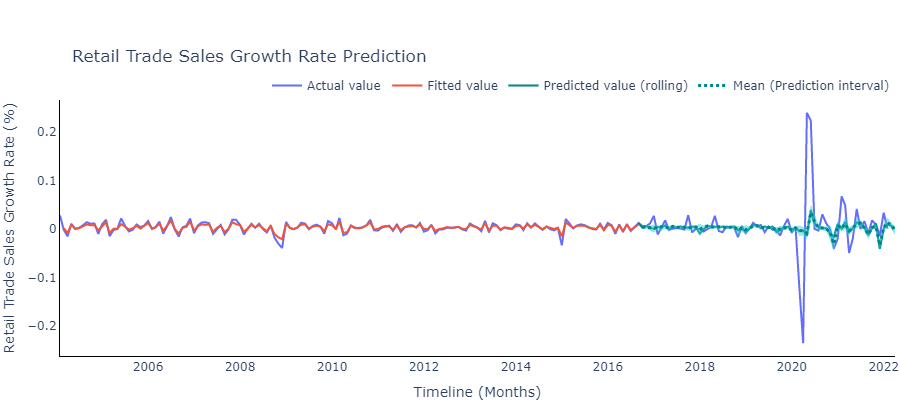
**Figure \_ :** GDP fitted and predicted values

## Retail Sales

The indicator included Google PyTrends keywords with respects to a single category ‘E-commerce Services’ with id ‘340’[Appendix ]. The predictors for the data were comparatively more than the observation counts; therefore, applying PCA with \_% of explained variation with Random Forest modelling technique was most optimal in capturing the trend over time. The number of trees fitted for the model was \_, these were selected after performing cross validation over the model.

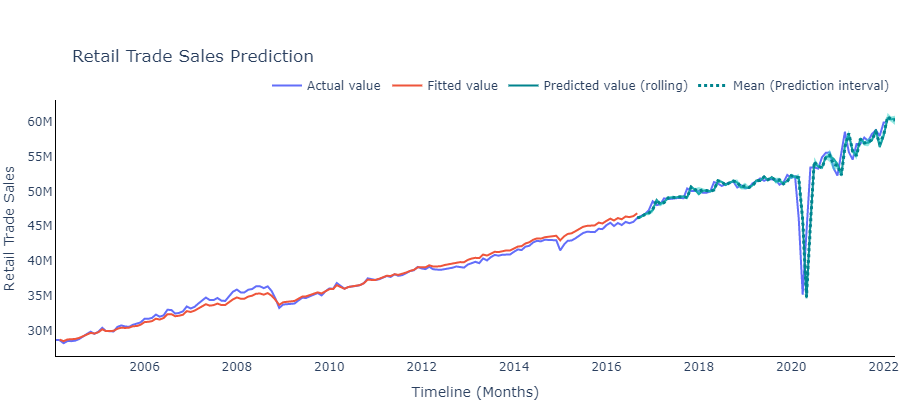
|  |  |  |
| --- | --- | --- |
| **Method** | **Prediction Error (RMSE)** | **Parameter Tuning** |
| DFM + ARIMA | 2,828,358 | Number of factors |
| LASSO | 2,379,342 | Penalty parameter |
| PCA + Random Forest | 2,281,435 | Number of trees |
| PCA + XGBoost | 3,410,734 | *\_* |

The below plot depicts the growth rate`s rolling prediction achieved using the random forest machine learning model. PCA helped in dimension reduction which contributed towards avoiding over-fitting of the random forest model and further capturing the unforessen recessional (2008) and the pandemic (2020) impact over the retail sales of the country.



**Figure \_ :** Retail sales growth rate`s fitted and predicted

The plot presents the actual retail trade sales value calculated using the growth rate variations predictions as provided in the above graph. It is interesting to see how the model has captured the unforeseen/sudden variations using the Google trends of the related sub-categories. The green brand depicts the 95% prediction interval which is extracted using \_ bootstrap samples.



**Figure \_ :** Retail sales fitted and predicted values

# Conclusion

It seems interesting throughout the entire research that how the Google PyTrends help in capturing the economic trends in a country. We could compare how the prediction trend changes in case a model was built over the historical GDP values over the past 18years and how it differs while building a model over the time series of keywords or the list of categories, sub-categories searches over the Google.

To recapitulate, the macroeconomic indicators used throughout the study had different list of predictors (keywords, categories, subcategories) in order to nowcast the values. The comparative study helped her select the respective models for each of the indicator. For E-commerce we have selected random forest machine learning model, for retail trade sales PCA with random forest and for GDP DFM with ARIMA.

In future work, as 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. There is lag coming…..

# 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.
6. Logo source: [UBC Logo](https://www.abdn.ac.uk/study/undergraduate/canada-university-of-british-columbia-4250.php), [Statistics Canada](https://crippledscholar.com/2020/05/30/statistics-canada-isnt-collecting-information-on-disability-during-the-pandemic/), [Google Trends](https://towardsdatascience.com/google-trends-api-for-python-a84bc25db88f)

# Appendix

Appendix A

[keywords, category list]