# 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**

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

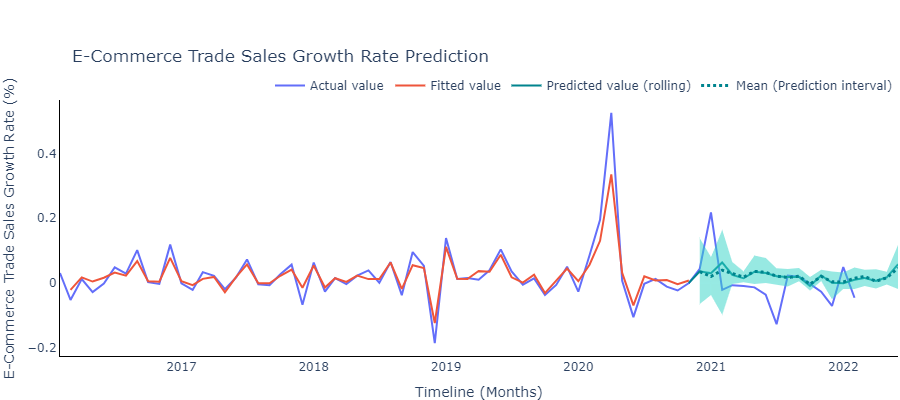
Description automatically generated

## E-commerce Sales

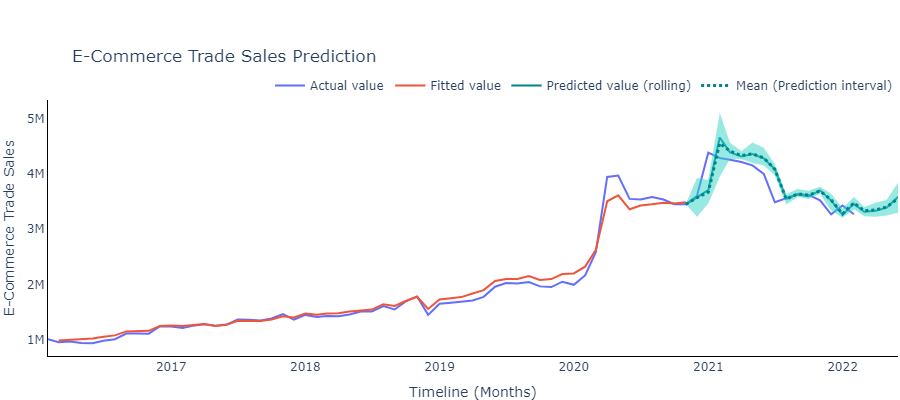
The indicator included Google PyTrends keywords with respects to a single category ‘E-commerce Services’ with id ‘340’. There were less predictor than the number of observations; therefore, we have applied random forest directly to the keyword`s time series. The prediction errors obtained from the models 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.



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.

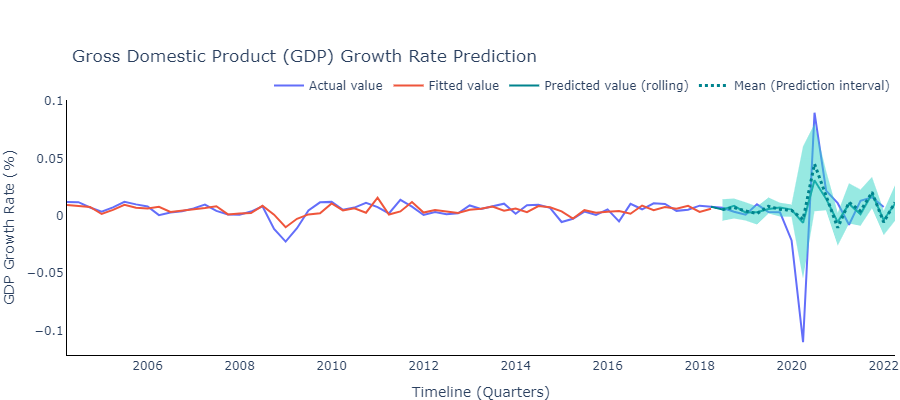


## GDP

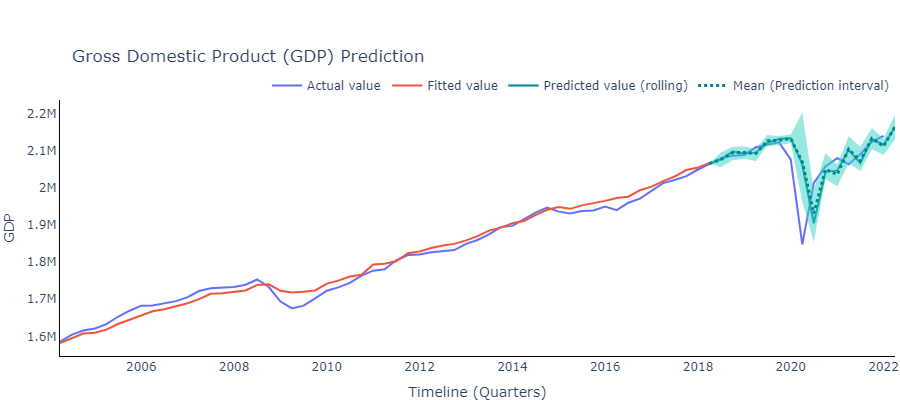
The indicator included Google PyTrends keywords with respects to a single category ‘E-commerce Services’ with id ‘340’ [Appendix ]. The prediction errors obtained from the models are depicted in the table below. As per the presented information Random Forest model had the least error with \_ number of 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.



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.

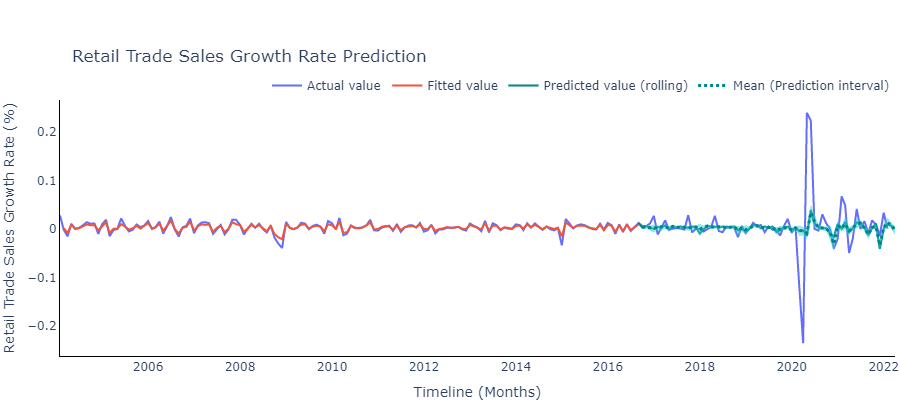


## Retail Sales

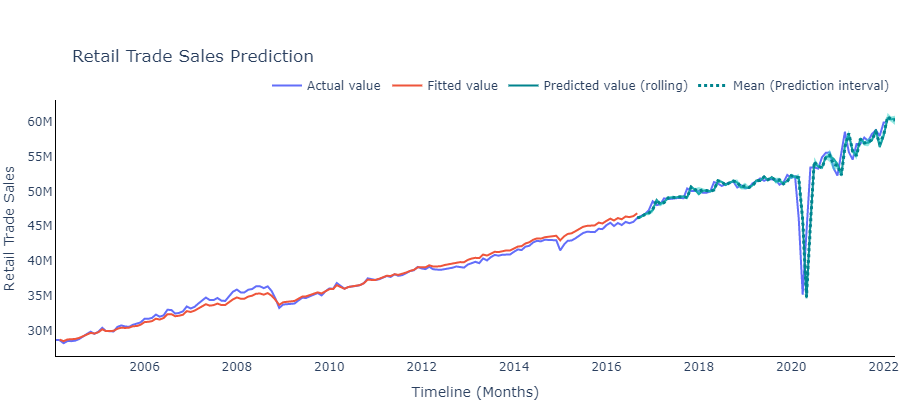
The indicator included Google PyTrends keywords with respects to a single category ‘E-commerce Services’ with id ‘340’[Appendix ]. The predictors for the data were comparatively more than the observation counts; therefore, applying PCA with \_% of explained variation with Random Forest modelling technique 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.



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.



# 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)
7. **Appendix** (if needed):  If you feel the need to include a technical discussion of methods, how to use the tools developed to perform future analysis, deployment of dashboards, delivery and deployment package for the client, code listings, etc, please place them here.

### Appendix A

[keywords, category list]

### Appendix B

It is convenient to operate over the work in the future. The changes with respect to the keywords, categories and subcategories can be added to the respective files of the indicators. The [script1\_extractGoogleTrendsData.py](https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script1_extractGoogleTrendsData.py) which would extract all the time series with regards to the relevant keywords added in the list of the indicator. The [script2\_fitModels.py](https://github.com/ubco-mds-2021-labs/capstone-project-googletrends_capstone/blob/main/src/code/script2_fitModels.py) contains the model fitting relevant code