

CSC110 Project Proposal: Predicting Canadian GDP During Covid-19

Shu Fan Nicholas Au, Jack Sun, Jerry Zhu

Friday, November 5, 2021

1 Introduction

Gross Domestic Product (GDP), whose official definition is the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period, is the most commonly used indicator tracking the health of an economy (Kramer, 2021). It is a measurement of the aggregate economic activity, so not surprisingly, reduced economic activity during the pandemic has negatively impacted Canadian GDP.

In fact, the Canadian GDP growth rate in 2020 was not only negative for the first time since the Financial Crisis of 2008, it decreased by a staggering 5.4% in 2020 in contrast to a yearly decrease of 2.9% in 2009 as a result of the Financial Crisis (Government of Canada, 2021). Given this unprecedented situation, we wish to understand whether the relevant economic indicators today can convincingly predict the GDP as well as it has in the past, or whether they fail to predict the GDP today because the pandemic has created an economic situation significantly different from anything before. Specifically, we consider, for quarterly periods from 2008 to the end of 2019 (just before COVID-19 hits in Canada), the following economic indicators:

1. Balance of Payment
2. USD-CAD Exchange Rate
3. Bank of Canada Overnight Rate
4. Inflation Rate
5. 10-Year Canadian Government Bond Yield
6. Unemployment Rate
7. S&P/TSX Composite Index
8. Population

Of these, a study from Vignan University established that 1 and 2 significantly predict GDP (Divya Devi, 2014). We felt that the rest are plausibly relevant. Further, some of these indicators warrant an explanation. The balance of payments is the record of all international financial transactions made by the residents of a country (Heakal, 2021). The Bank of Canada overnight rate is the interest rate at which other banks borrow from it. The inflation rate measures the rate at which prices are rising. We calculate the inflation rate through the Consumer Price Index, which measures the average prices over time for fixed goods. The 10-year government bond yield is the interest rate that the Canadian government pays to borrow money for 10 years. Finally, we consider the value of the SP/TSX Composite Index, which tracks the performance of the largest companies listed on Canada's primary

stock exchange, the Toronto Stock Exchange (TSX). Concisely, we ask: **How well can the changes to Canadian GDP Growth Rate caused by COVID-19 be predicted?**

2 Dataset Description

We are using datasets for the eight Canadian economic indicators and the Canadian GDP from 2008 - Q2/June 2021. Some of the datasets might also include data that are outside this date range, but we will only extract data that are in this range. The table below shows details of our datasets. Each dataset contains data either per day, per month or per quarter. We will only use data stored in the variable(s) described in “Variable to Use” (table) as well as their corresponding date.

Economic Indicators / GDP	Dataset File	Variable(s) to Use	Source
Canadian GDP (2008-2021)	GDP.csv	“Gross domestic product at market prices” at row 43	Statistics Canada
Balance of Payment (2008-2021)	Balance_of_Payment.csv	“Total investment income, receipts” at row 12	Statistics Canada
USD-CAD Exchange Rate (2008-2021)	USD-CAD_Exchange_Rate.csv	“Price” at column B	Investing.com Canada
Bank of Canada Overnight Rate (2011-2021)	Bank_of_Canada_Overnight_Rate.csv	“V390779” at column B	Bank of Canada
Bank of Canada Overnight Rate (2008-2010)	Bank_of_Canada_Overnight_Rate_2.csv	“Overnight money market financing” at row 12	Statistics Canada
Inflation Rate (2008-2021)	Inflation_Rate.csv	“INDINF_CPLM” at column B	Bank of Canada
10-Year Canadian Government Bond Yield (2008-2021)	10_Year_Canadian_Government_Bond_Yield.csv	“BD.CDN.10YR.DQ.YLD” at column F	Bank of Canada
Unemployment Rate (2008-2020)	Unemployment_Rate.csv	“Canada” at row 14	Statistics Canada
Unemployment Rate (2021)	Unemployment_Rate_2.csv	“Unemployment rate 13” at row 28	Statistics Canada
S&P/TSX Composite Index (2008-2021)	S_PTSX_Composite_Index.csv	“Standard and Poor’s/Toronto Stock Exchange Composite Index, close” at row 14	Statistics Canada
Population (2008-2021)	Population.csv	“Canada” at row 11	Statistics Canada

3 Computational Overview

We constructed a machine learning model to predict quarterly GDP growth under the effect of Covid-19 through the 8 indicators. We train the model with data from 2008 to 2017, and have it predict the GDP growth rate by inputting economical indicators for 2018-2021. 2018-2019 is before covid period and 2020-2021 is during covid period. The predicted and actual GDP growth rate will be compared in their corresponding time period. We will compare the accuracy of the model predicting GDP growth rate before covid (2018-2019) and after covid (2020-2021) to make our final conclusion. The reason for selecting 2018-2019 is because this data wasn't used as data training for the model, same as 2020-2021.

There are two major steps in our program, data extraction and computing the model.

In the step of data extraction, the function `loading_data()` in `reading_data.py` helps us to extract all the useful data into a structure that is ready for us to easily implement it into the model.

The `loading_data()` function used a lot of other helper functions. A few major ones will be listed here. We used the functions `load_horizontal_data()` and `load_vertical_data()` to systematically filter out specific values of the data that we need (the value's corresponding variables are described in the above table) and transform it into a standard data structure (`list[tuple[datetime.date, float]]`). As some of the datasets have data per day, while others have data per week and per quarter. We implement functions `change_day_to_month()` and `change_month_to_quarter()` to convert all data into per quarter frequency.

For the part of computing the model. We first convert the data from `loaded_data()` into a `DataFrame` - a data type in the pandas library conducive for machine learning using scikit-learn. We process it as follows: add "dates" as a separate column, strip the dates from the indicator value, and fill in missing null values with official figures on the given date. We also compute "gdp_growth" as follows: set the first row to be 0%. For an arbitrary row $i \geq 1$, the gdp growth is $\frac{gdp[i] - gdp[i-1]}{gdp[i-1]} * 100\%$. Adding it to the `DataFrame` finishes the processing part.

With the processed `DataFrame`, we plot a correlation matrix of all the variables using Plotly and histograms of the columns using matplotlib, which allows us to make useful observations about the data to select the best pre-processing transformations. In particular, we conclude that transforming x with `StandardScaler()` and `PCA()` will benefit the model. We highlight these observations in 6.1.

To train the machine learning model, we use only the data from 2009-2017. We split the data into an array of independent variable values (x) and an array of the dependent variable (GDP) value (y), both of type `np.ndarray`. We then split x and y into training and testing sets using the `train_test_split` method from `sklearn.model.selection`. Since the number of rows is low, the model is prone to overfitting; that is, the state of the model fitting the training data so tightly that it does not generalize in predictions. Therefore, we train our model on only 55% of the data and use 45% of the data to test it. Now, based on the aforementioned observations made on the graphs, we fit `StandardScaler` and `PCA` on the training set of the indicators to then use them to transform the entire indicator set.

Our candidate machine learning models are 'Elastic Net', 'Linear Support Vector Regressor', 'Random Forest Regressor', 'Gradient Boosting Regressor', and 'Extra Trees Regressor'. To select the best candidate, we use 5-fold cross validation on all the candidates with the help of the `cross_val_score()` method from `sklearn.model.selection`. It splits x and y into 5 "folds" and calculates the Mean Squared Error (MSE) of the particular regressor on the "fold." Then, the method returns an `np.ndarray` of the errors. We then compute the Root Mean Squared Error (RMSE) with the average error, selecting the model that gives the lowest RMSE.

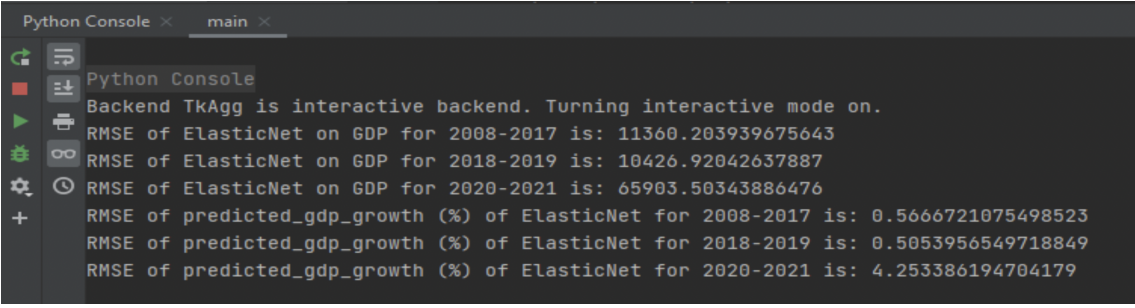
We find that ElasticNet gives the lowest RMSE. To optimize its performance, we perform hyperparameter tuning on it with the help of GridSearchCV method from `sklearn.model_selection`. Notable parameters to tune are "alpha", "l1_ratio", "tuned_model", we then make predictions on the GDP using indicator values from the entire DataFrame. In the function `visualize_predictions`, we make predictions for 2008–2017 (training set), 2018–2019 (pre-covid), and 2020–2021 (during covid). The graphs and printed statistics allow us to draw in-

4 Instructions for obtaining data sets and running your program

As our datasets aren't that large and the Canda.ca has broken down again recently, we don't have a direct link for downloading the csv. All of our datasets are uploaded on Markus.

After running the main.py file, some values and several figures will be seen.

Under the python console, the root mean square error (RMSE) of our model on the GDP and GDP growth rate in their corresponding date range are presented.

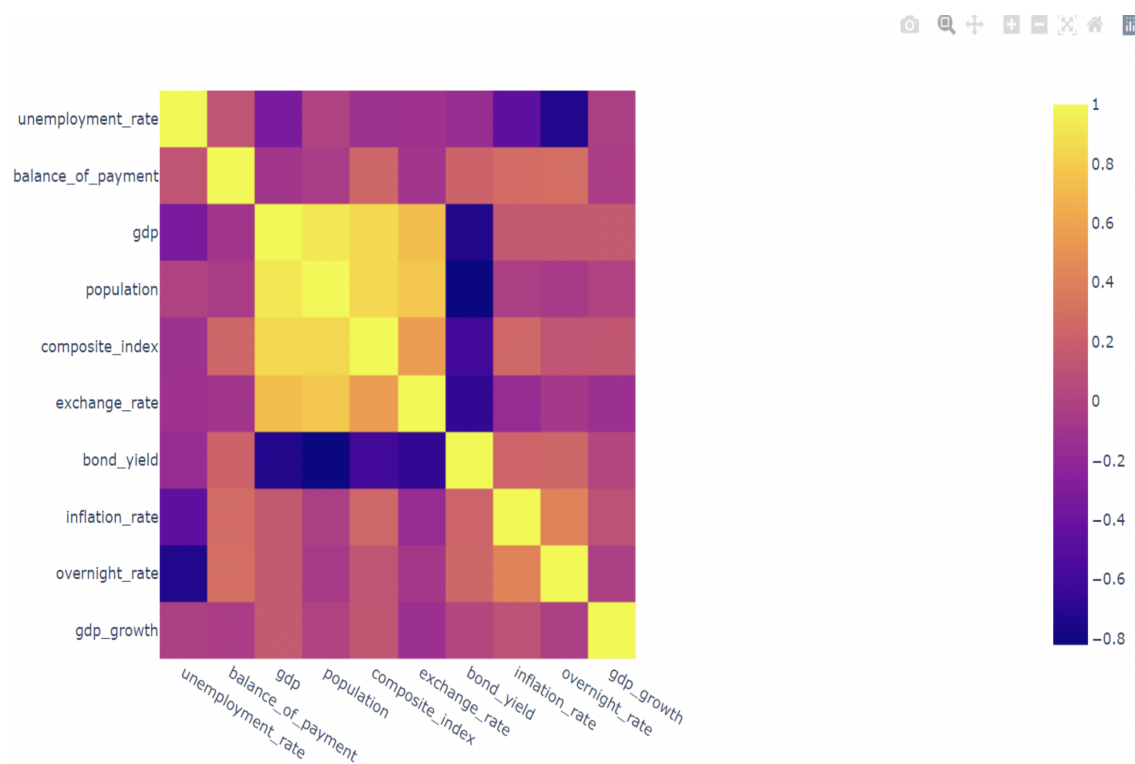


```

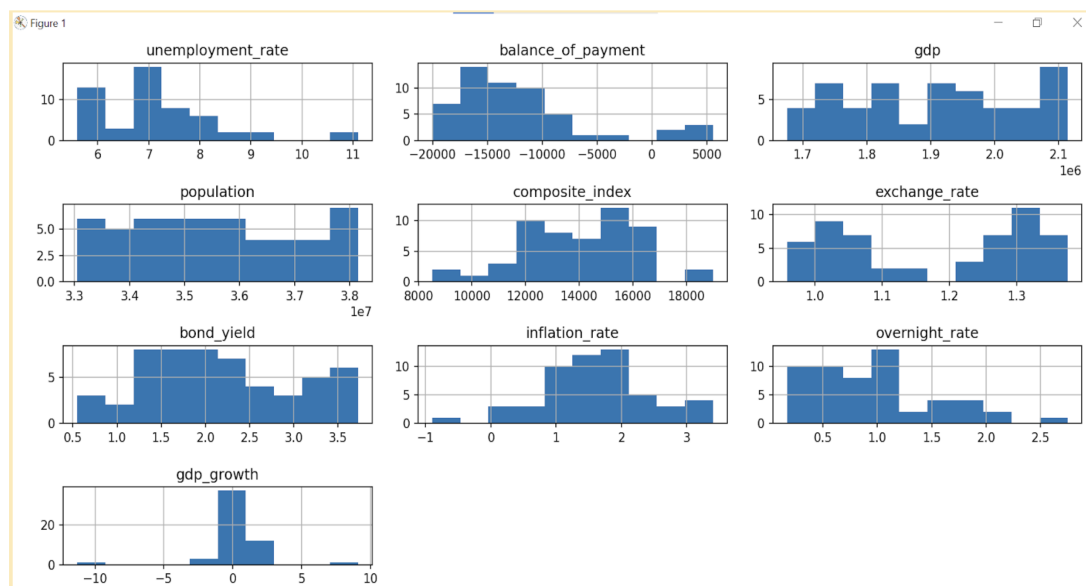
Python Console × main ×
Python Console
Backend TkAgg is interactive backend. Turning interactive mode on.
RMSE of ElasticNet on GDP for 2008-2017 is: 11360.203939675643
RMSE of ElasticNet on GDP for 2018-2019 is: 10426.92042637887
RMSE of ElasticNet on GDP for 2020-2021 is: 65903.50343886476
RMSE of predicted_gdp_growth (%) of ElasticNet for 2008-2017 is: 0.5666721075498523
RMSE of predicted_gdp_growth (%) of ElasticNet for 2018-2019 is: 0.5053956549718849
RMSE of predicted_gdp_growth (%) of ElasticNet for 2020-2021 is: 4.253386194704179

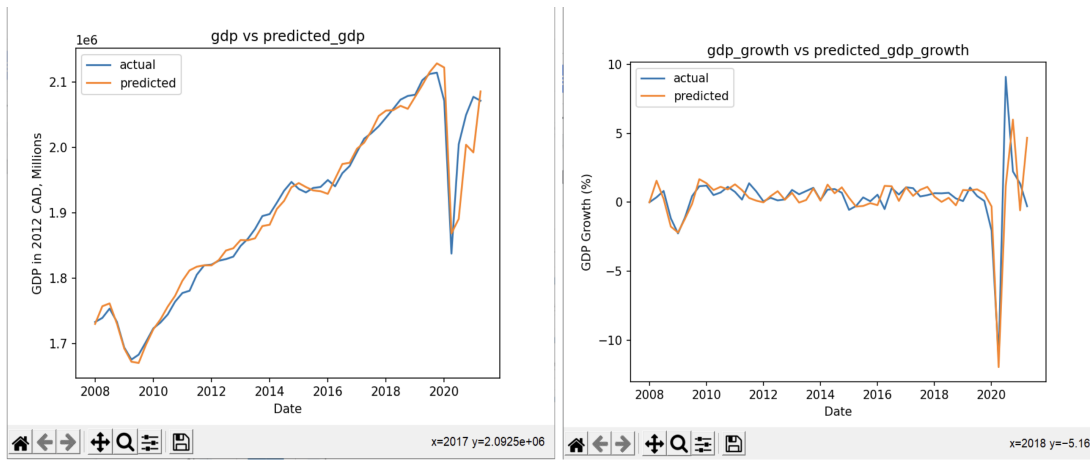
```

Besides, there will be a web page pop up in the browser. There is a correlation matrix representing the coefficient correlation between economic indicators and GDP. You may move your mouse to the box to see each of the specific values in the table. We can focus on reading the 3rd row, which focuses on describing the GDP.



There will also be 3 figures popping up. Figure 1 is a histogram of the economic indicators and the GDP. Figure 2 shows the comparison between the predicted gdp and the actual gdp. Figure 3 shows the comparison between the predicted gdp growth and the actual gdp. For figure 2 and 3, you may move your mouse to a specific point to see the specific x and y coordinate of that point.





5 Changes in Project Plan

We didn't make notable changes. We only changed some datasets that we are using.

6 Discussion section

6.1 Results Interpretations

From the correlation matrix, we observe that while indicators like "bond_yield" and "population" highly correlate with the dependent variable (learn developers).

From the values printed in the python console, we could see that the RMSE of our model on gdp growth rate before covid times (2018-2019) are pretty decent with 0.505%. However, with a RMSE of 4.25% during covid times (2020-2021) is not that decent. By looking at the figure 3, we indeed see that despite the model accurately predicting the first two quarters of 2020, it doesn't do a great job afterwards and makes a lot of wrong predictions, such as predicting a drastic increase while there is actually a drastic drop in GDP growth rate. This could be seen in figure 2, in which most of the predicted GDP are a lot lower than the actual GDP after 2020.

The fact that the model is doing a good job predicting GDP before covid and did a bad job in predicting after covid shows that the economical indicators might not be that effective to predict changes to Canadian GDP Growth Rate caused by COVID-19.

6.2 limitations

The different precision level and time interval of data sets are limitations. The overnight rate, inflation rate and bond yield indicators are recorded in days, which are very frequent. However, because data sets like GDP and population are updated 1 time per quarter, we have to unify the frequency of all 9 data sets to be 1 time per quarter by taking the average. This indeed keeps time consistency of data, but the number of available data is decreased to 54. Consequently, there are less available data sources for models to learn, which potentially

undermines performance of models. Besides, the precision level of two data sets of overnight rate is inconsistent. *Bank_of_Canada_Overnight_Rate_2.csv* has a precision level 0.0001, while *Bank_of_Canada_Overnight_Rate.csv* only has 0.25. Such precision

6.3 further exploration

In this project, we demonstrated that the Covid-19 does undermine the accuracy of the model to predict GDP growth. The reason for inaccuracy is that we intentionally neglected Covid-19 indicators, such as number of active cases and days of quarantine. These indicators can negatively impact GDP growth, because Covid-19 kills labour forces and impedes staff-intensive industries. Next step, we will build an upgraded model to predict GDP growth, which inherits the structure of the original model and adds Covid-related variables. We will explore to what extent the Canadian GDP Growth Rate caused by COVID-19 be predicted in optimum.

References

Bank of Canada. (n.d.). Canadian interest rates and monetary policy variables: 10-year lookup. Retrieved November 5, 2021, from <https://www.bankofcanada.ca/rates/interest-rates/canadian-interest-rates/>.

Bank of Canada. (n.d.). Inflation: Definitions, graphs and data. Retrieved November 5, 2021, from <https://www.bankofcanada.ca/rates/indicators/capacity-and-inflation-pressures/inflation/>.

Bank of Canada. (n.d.). Selected bond yields. Retrieved November 5, 2021, from <https://www.bankofcanada.ca/rates/interest-rates/canadian-bonds/>.

Brownlee, J. (2020, June 11). How to develop elastic net regression models in Python. Machine Learning Mastery. Retrieved December 13, 2021, from <https://machinelearningmastery.com/elastic-net-regression-in-python/>.

Chen, J. (2021, August 11). What is the overnight rate? Investopedia. Retrieved November 4, 2021, from <https://www.investopedia.com/terms/o/overnightrate.asp>.

Chen, J. (2021, September 21). Treasury yield. Investopedia. Retrieved November 4, 2021, from <https://www.investopedia.com/terms/t/treasury-yield.asp>.

D'monte, D. D. (2020, September 8). Predicting the nominal GDP using economic indicators: A data science approach. Medium. Retrieved November 4, 2021, from <https://towardsdatascience.com/predicting-the-nominal-gdp-using-economic-indicators-a-data-science-approach-7c56cded782>.

Divya, K. H., Devi, V. R. (2014, July 14). A study on predictors of GDP: Early signals. Procedia Economics and Finance. Retrieved November 4, 2021, from <https://www.sciencedirect.com/science/article/pii/S2212567114002056>.

Fernando, J. (2021, September 8). What is the SP/TSX Composite Index? Investopedia. Retrieved November 4, 2021, from <https://www.investopedia.com/terms/s/sp-tsx-composite-index.asp>.

Frost, J. (2021, September 24). Multicollinearity in regression analysis: Problems, detection, and solutions. Statistics By Jim. Retrieved December 13, 2021, from <https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>.

Government of Canada, Statistics Canada. (2021, December 3). Table 14-10-0287-01. Labour force characteristics, monthly, seasonally adjusted and trend-cycle, last 5 months. Retrieved December 10, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410028701>.

Government of Canada, Statistics Canada. (2021, December 9). Table 10-10-0139-01 Bank of Canada, money market and other interest rates. Retrieved December 9, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010013901>.

Government of Canada, Statistics Canada. (2021, January 8). Table 14-10-0294-01 Labour force characteristics by census metropolitan area, three-month moving average, seasonally adjusted and unadjusted, last 5 months, inactive. Retrieved November 5, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410029401>.

Government of Canada, Statistics Canada. (2021, November 1). Table 10-10-0125-01 Toronto Stock Exchange Statistics. Retrieved November 5, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010012501>.

Government of Canada, Statistics Canada. (2021, November 29). Table 36-10-0018-01 Balance of international payments, current account, seasonally adjusted, quarterly (x 1,000,000) Retrieved December 9, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610001801>

Government of Canada, Statistics Canada. (2021, November 30). Table 36-10-0104-01 Gross domestic product, expenditure-based, Canada, quarterly (x 1,000,000) Retrieved December 9, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610010401>

Government of Canada, Statistics Canada. (2021, September 29). Table 17-10-0009-01 Population Estimates, quarterly. Population estimates, quarterly. Retrieved November 5, 2021, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901>.

Heakal, R. (2021, September 8). What is the balance of payments? Investopedia. Retrieved November 4, 2021, from <https://www.investopedia.com/insights/what-is-the-balance-of-payments/>.

Kramer, L. (2021, July 27). What is GDP and why is it so important to economists and investors? Investopedia. Retrieved November 4, 2021, from <https://www.investopedia.com/ask/answers/what-is-gdp-why-its-important-to-economists-investors/>.

Learn. scikit. (n.d.). Retrieved November 4, 2021, from <https://scikit-learn.org/stable/index.html>.

scikit-learn developers. (n.d.). Compare the effect of different scalers on data with outliers. scikit. Retrieved December 13, 2021, from https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html.

USD CAD Historical Data. Investing.com Canada. (n.d.). Retrieved November 5, 2021, from <https://ca.investing.com/currencies/usd-cad-historical-data>