Predicting the Backing Ratio for the ECCU

Verne Cazaubon

2023-12-03

Introduction

The Organisation of Eastern Caribbean States (OECS) is an International Inter-governmental Organisation dedicated to regional integration in the Eastern Caribbean. The OECS is an (11) eleven-member grouping of islands spread across the Eastern Caribbean. Together, they form a near-continuous archipelago across the eastern reaches of the Caribbean Sea. They comprise the Leeward Islands: Antigua and Barbuda, Saint Christopher (St Kitts) and Nevis, Montserrat, Anguilla and the British Virgin Islands; and the Windward Islands: the Commonwealth of Dominica, Saint Lucia, Saint Vincent and the Grenadines and Grenada, Martinique and Guadeloupe. [link].

The Eastern Caribbean Central Bank (ECCB) was established in October 1983. It is the Monetary Authority for a population of more than 600,000 people spanning six sovereign states: Antigua and Barbuda, the Commonwealth of Dominica, Grenada, Saint Christopher (St Kitts) and Nevis, Saint Lucia, and Saint Vincent and the Grenadines, and two overseas territories of the United Kingdom: Anguilla and Montserrat. [link]. These comprise the Eastern Caribbean Currency Union (ECCU). The Eastern Caribbean Dollar (XCD), less formally the EC$, is pegged to the United States Dollar (USD) at XCD 2.70 = USD 1.00 or EC$ 2.70 = US$ 1.00.

In the Governor's Foreword of ECCB's Report and Statement of Accounts for the Financial Year ended 31 March 2023, p.2, he stated that "the Central Bank reports that the EC dollar remains strong and stable with a backing ratio around 92.0 per cent, well above the legal requirement of 60.0 per cent." [link]. What is this backing ratio? And why is it important?

First, the backing ratio is the ratio of Total External Assets (International Reserves) to Total Demand Liabilities. In other words, Backing Ratio $= \frac{\text{Total External Assets}}{\text{Total Demand Liabilities}}$.

Now to understand its importance, consider the following from the same document. Under the section titled Foreign Reserves Management, p. 62, it was stated that "The Bank will seek to strengthen the management of Foreign Reserves by: 1. Relooking the risk appetite of the Bank while keeping in mind the backing ratio". Here we note that the backing ratio is of extreme importance to foreign reserves management, particularly given the exchange rate agreement of the currency union, that is, being pegged to the USD. Thus, critical decisions taken by the ECCB always consider the impact on the backing ratio.

With the above in mind, we set out to predict the backing ratio of the ECCU in an effort to provide a tool to monitor the indicator, project its trajectory, and eventually assist with decision making at the highest level of the organisation.

Abbreviations and Acronyms

Organisations:

- ECCB - Eastern Caribbean Central Bank
- ECCU - Eastern Caribbean Currency Union
- OECD - Organisation for Economic Co-operation and Development
- OECS - Organisation of Eastern Caribbean States

Countries:

- CAN - Canada
- CHN - People's Republic of China
- GBR - United Kingdom
- JPN - Japan
- USA - United States of America

Other:

- CPI - Consumer Price Index
- csv - comma-separated values
- GFC - Global Financial Crisis

Description of data sets

For the purpose of this analysis, an extremely limited data set will be considered. It will consider data obtained from the websites of the ECCB [link] and the OECD [link]. From the ECCB's website, the total external assets and total demand liabilities were downloaded in csv format. The indicators downloaded from OECD's website were forecasts for: CPI or inflation, long-term interest rates, and unemployment rates.

Forecast data was used as opposed to the actual data because when predicting the backing ratio only forecast data would be available at that moment. We note that this would increase error but it presents a more realistic scenario. For the target variable (backing ratio), however, the actual values were selected as our goal is to predict it.

In the end, thirteen (13) independent variables (will also be referred to as predictors, predictor variables, or features) were used to forecast target variable. It is expected that in the full implementation of the project more than thirteen independent variables will be selected. More on this will be mentioned in the section *Future Work*.

Goal of project

The goal of this project is to predict the backing ratio of the ECCU given forecast economic data from the main international trading partners and tourism source markets of the ECCU countries. These main international trading partners and tourism source markets are: Canada (CAN), People's Republic of China (CHN), Japan (JPN), United Kingdom (GBR), and United States of America (USA). It should be noted that two indicators, long-term interest rates forecasts and unemployment rates forecasts, were not available for the People's Republic of China from the OECD's database. In a full scale implementation of this project, the goal is to produce results at least as accurate as those of any econometric model that is used to predict the backing ratio.

Key steps performed

An outline of the key programming steps performed are as follows:

1. Install packages if necessary, and load the required libraries.
2. Download data from the respective websites/databases.
3. Input the data.
4. Clean the data.
5. Select the relevant records and independent variables (or features).
6. Visualize the data. Gain insight into variables/features.
7. Combine the data into one tibble. Format the data to ready it for analysis.
8. Separate the data into a training set and a test set.
9. Analyse the data using different machine learning algorithms.
10. Compare the performance of these algorithms.
11. Conclude on which model best fits the data.

Analysis

Processes and techniques used in data cleaning

Data were obtained from different websites and as a result the downloaded data were formatted differently. Therefore the process used to clean data sets was dependent on the website the data came from. The format of downloaded data from any one website is similar which makes cleaning data from that website an easier process when multiple files need to be inputted and cleaned.

For this project, one csv file was downloaded from the ECCB's Statistics webpage and stored in a folder named Data. The file was read into the program. It had no header row. There was some metadata in the first few lines of the file followed by column headings in row 8. After inputting the data the column headings were renamed to those in the file and the extra rows at the beginning were deleted. Next, the column the column types were inspected to know how to proceed. The data were cleaned so as to make analysis possible. This included converting dates from character to date format, and converting the Total External Assets and Total Demand Liabilities from character to numeric. We also checked for any records containing NAs so those would have been handled prior to moving on, if any were found. Finally, the Backing Ratio was calculated by dividing the item in column Total External Assets by the item in column Total Demand Liabilities for each record. Intermediate tibbles used in the cleaning process were deleted to preserve computer memory.

Three csv files were downloaded from the OECD's Data webpage and stored in a folder named Data. As the files came from the same data source, the process to clean the data were similar for each file. The files were read into the program. Their columns were examined to see what kind of data they contained. The files already had column headings but the names of some of those headings needed to renamed to something more readable. NAs were checked for and dealt with if any were found. Dates were converted to a date format. This conversion was more computationally involved than for the previous data set as the dates in these tables were formatted as year-quarter, for example 1980-Q2. This format could not be handled by any packages that were previously loaded. Once this was done, the countries and dates that were needed were filtered out into a tibble, and intermediate tibbles were deleted.

Processes and techniques used in visualization and data exploration

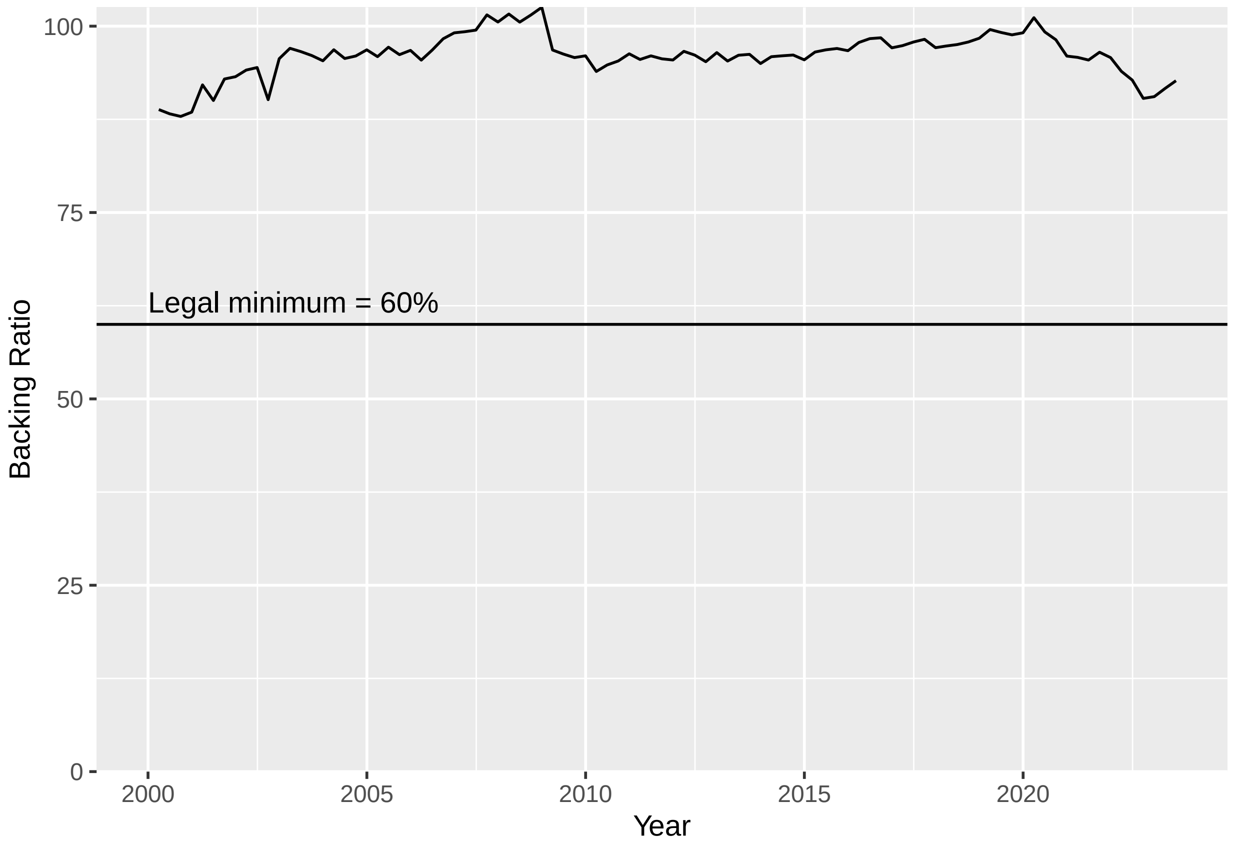
Due to the time series nature of the data, line graphs were primarily used for visualization. We now examine the variables/features to see whether they coincide with events that we know. Sometimes the data may reveal events which we were not aware of, and would prompt an investigation into the cause. Any extreme values (outliers) seen were worth investigating.

Insights gained

We now move to see what insight we can gain from the data. What story is the data telling? How does the story of the independent variables, or features, coincide or differ from that of the target variable? We look at visualizations of each of these in turn, then we consider a correlation matrix showing any relationships among the variables.

Backing ratio

By examining the graph below it can be seen that the backing ratio has remained above 87.5% through the entire period. Two events occurred which noticeably affected the backing ratio for an extended spell. The first occurred around 2008-2009 which coincides with the Global Financial Crisis (GFC). To this point the backing ratio has never returned to the pre-GFC level. The second event commenced sometime in 2020 and lasted approximately two years. This, of course, would be the COVID-19 pandemic. The backing ratio seems to just be recovering from this last economic shock.

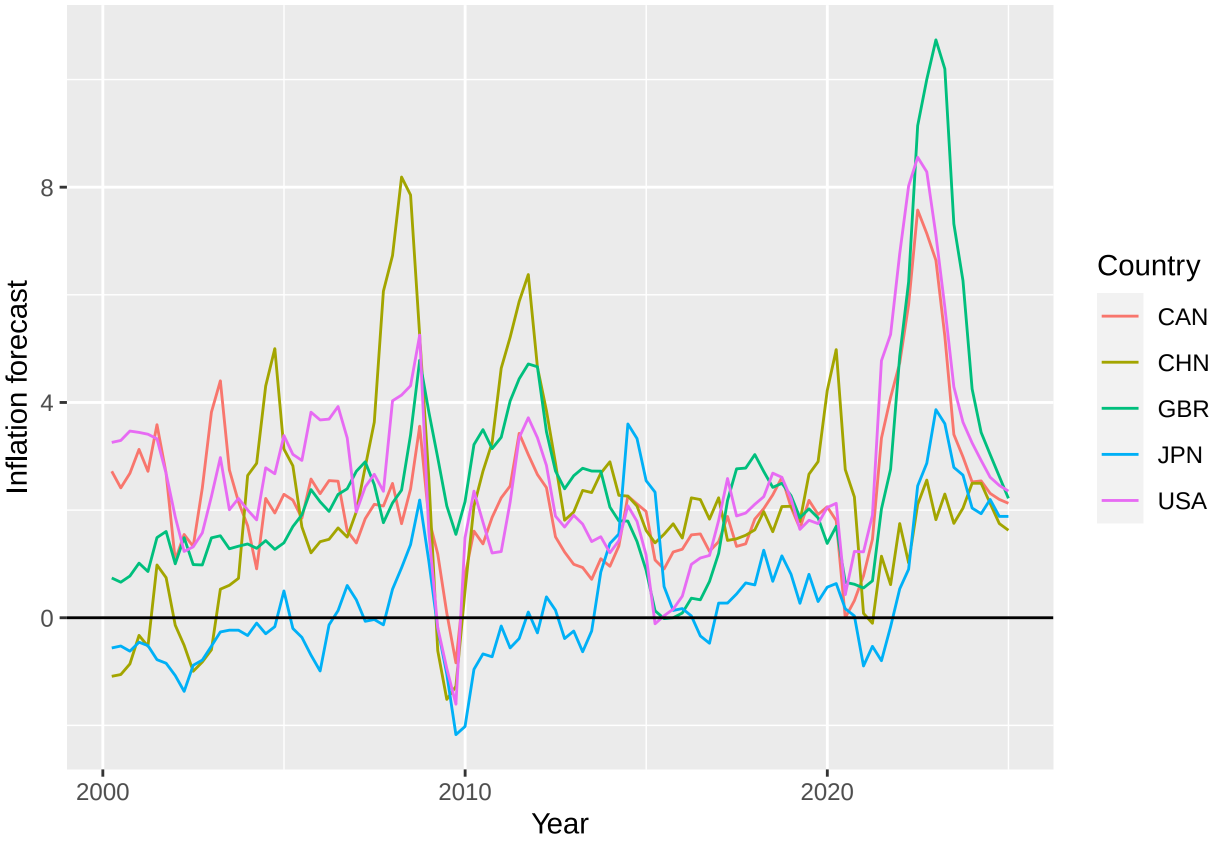


Inflation

Bear in mind that the graph below, and those which follow, show forecast data and not actual data. Also noteworthy is that they contain forecasts through to Q4 2024.

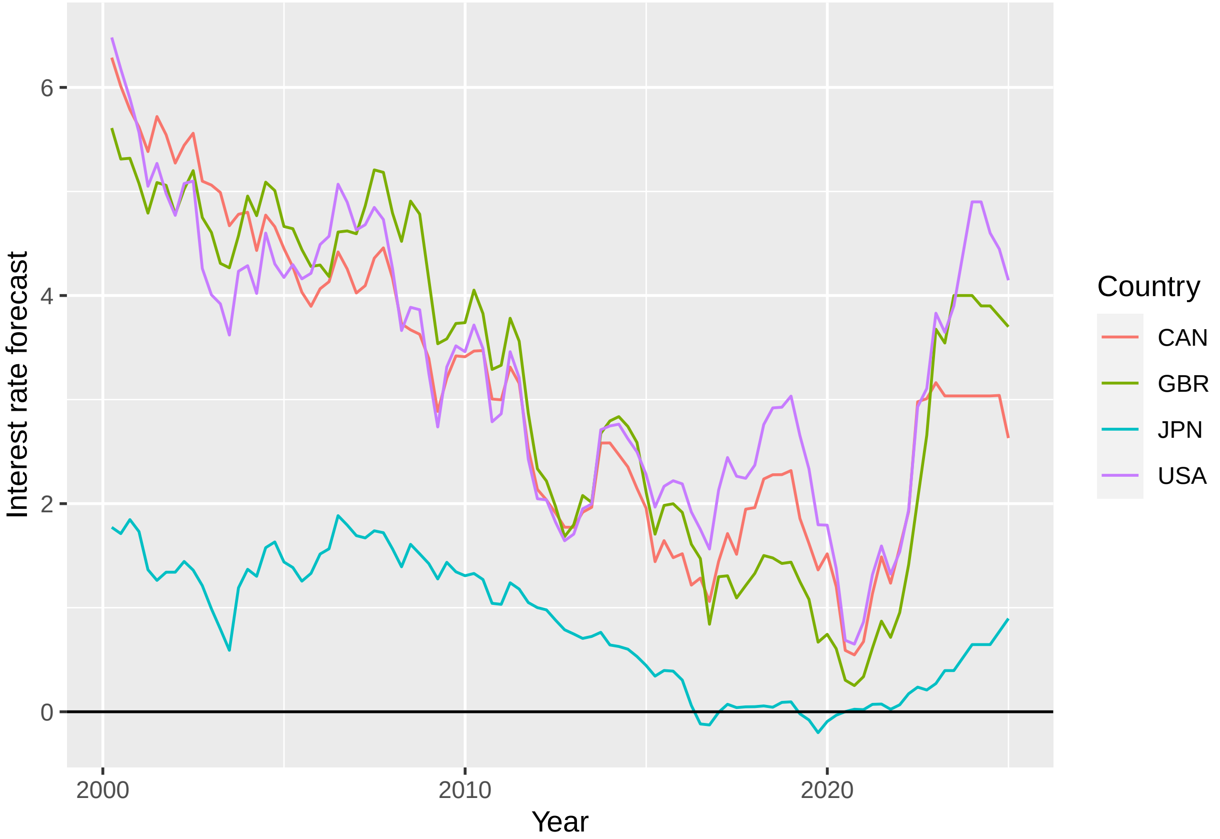
The CPI forecast shows similar trends to the backing ratio. All countries show huge spikes in the CPI around the same economic shocks of the GFC and the pandemic. Over the first two decades there are other spikes seen by individual or multiple countries, however, it is only during the two major economic shocks that these spikes coincide for all five countries.

It is also evident that through most of the series, Japan's inflation remained relatively low and comparably stable. In addition to the spikes in inflation around the GFC and pandemic, Japan's data reveals that another significant economic shock occurred sometime in 2011. That would be when a powerful earthquake struck the northeastern part of Japan causing a huge tsunami which killed thousands and devastated that part of the country.



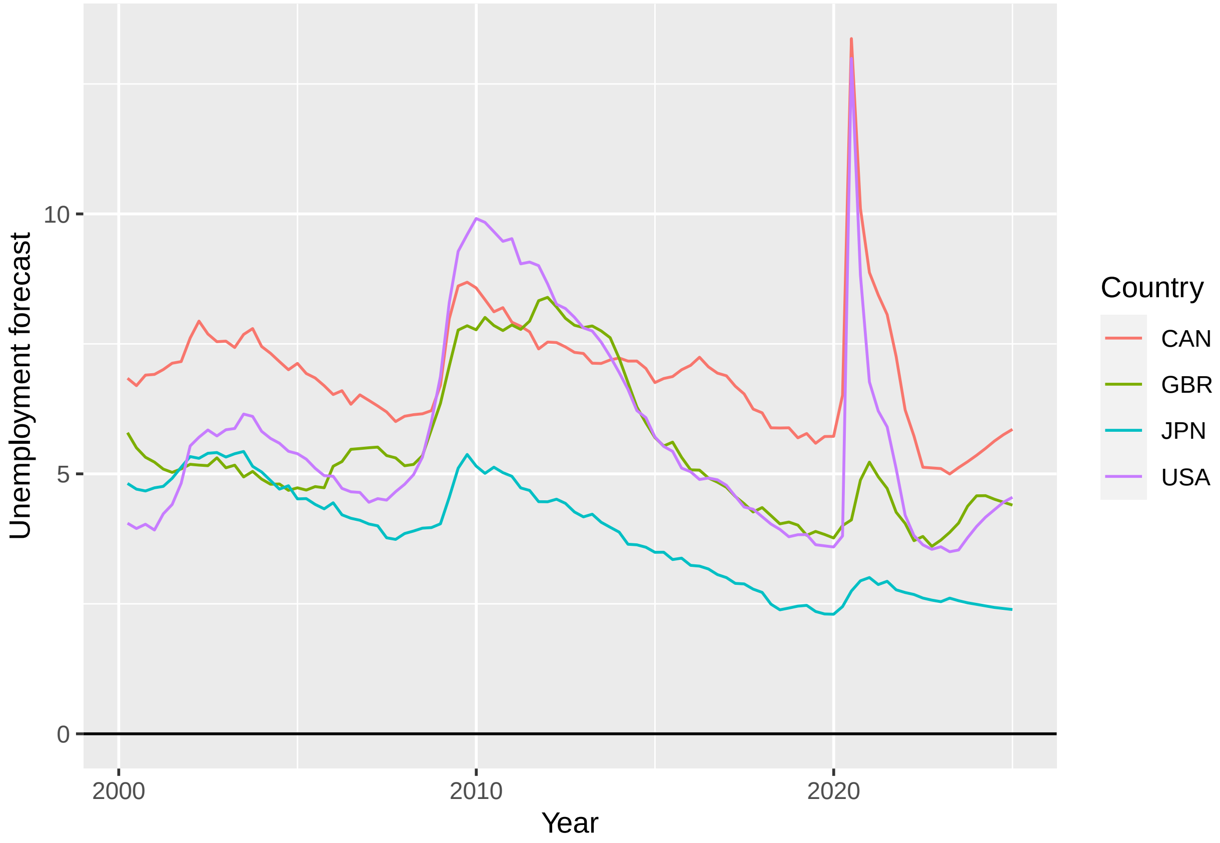
Interest rates

The interest rates forecasts of Canada, the United Kingdom, and the United States of America all seem to move very closely with each other. Japan's interest rate forecasts are always below that of the other three countries, however, it still follows their trend. The graph below shows that interest rate forecasts have trended downward from 2000 to 2020. From the end of 2020, though, we are seeing the sharpest increase in interest rates ever. In just two years, interest rates seem to reach levels seen a decade prior. The forecasts through the end of 2023 and through 2024 do show a leveling off and slight decrease for Canada, the United Kingdom, and the United States of America. Japan's interest rates are forecast to continue rising.



Unemployment

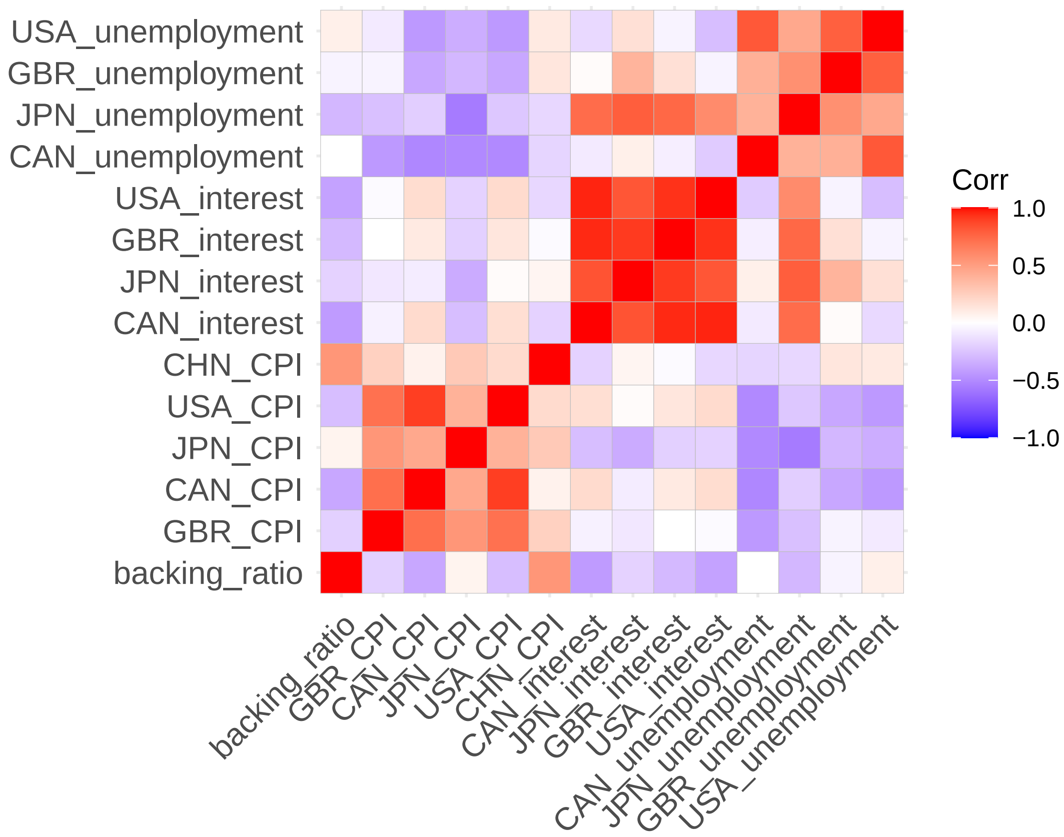
Unemployment, similar to the previous two features/variables, saw an increase around the GFC and pandemic. At the start of 2020, unemployment was at its 20-year low for all countries shown. At the start of the pandemic, unemployment rates forecasts rose sharply for all countries. The forecasts for Canada and the United States of America spiked to levels unseen in the first two decades of the century. These forecasts quickly fell but are on the rise through 2024. Contrarily, Japan's and the United Kingdom's forecasts for unemployment are heading downward through 2024.



After visualizing each data set individually, the data were combined into one tibble. We now look at the relationships between the different variables/features. The correlation matrix (or correlogram) below was used to examine the correlations/relationships between the features/independent variables and the target variable. It showed that there was a strong positive correlation between the interest rate forecasts of Canada, Japan, the United Kingdom, and the United States of America. This is not unusual as it was seen in the line graph earlier. There were also positive correlations among the countries with regard to unemployment and CPI forecasts, although the relationship were not as strong as that of interest rates.

Also noteworthy from the correlation matrix is the moderate negative correlation between unemployment and inflation. Again, this relationship is not unusual.

The backing ratio showed a moderate positive correlation with the People's Republic of China's inflation. Its relationship with other features was weak to moderately negative. There was very weak to no correlation with Japan's inflation, Canada's unemployment, and the United Kingdom's unemployment.



Modeling approach

The data from this new single tibble was used for further analysis. The data were divided into two new data subsets - a training set and a test set. Regression algorithms were used as opposed to classification as our target variable, the backing ratio, is a continuous variable. The algorithms were supervised learning algorithms and were fed the training set to learn on. After learning, their performance was evaluated using the test set.

The algorithms utilized were: generalized linear model (baseline), k-nearest neighbors, random forest, support vector machine with a polynomial kernel, principal component analysis, and extreme gradient boosting. We also created an ensemble model which dropped the highest and lowest prediction of the five non-baseline algorithms, and computed the mean of the remaining three as its prediction.

The root mean square error (RMSE) was used as the loss function to analyse the performance of the models. The formula is $\mathrm{RMSE}(\hat{y}, y) = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$. The units of RMSE would be the same units as the predicted variable - backing ratio, which is percent (%). Therefore, a RMSE < 1 would be a good target particularly since we are using forecast data and not actual data as predictors. This would be interpreted as, on average, the model predicted values within +1% or -1% from the actual value.

The required indicators from the ECCB's website were available only back to the year 2000. The data were available with quarterly periodicity up to the second quarter (Q2) of 2023. As this data forms the target variable, the independent variables had to be limited to this same time frame (Q1 2000 to Q2 2023). This meant that only 94 records of data were available for the process. Due to the small number of records, 90% of the data were used to train the algorithms and 10% to test them.

To tune the algorithms' parameters, the methods used included 10-fold cross-validation, 5-fold cross-validation, and the default bootstrap method.

Generalized Linear Model (Baseline model)

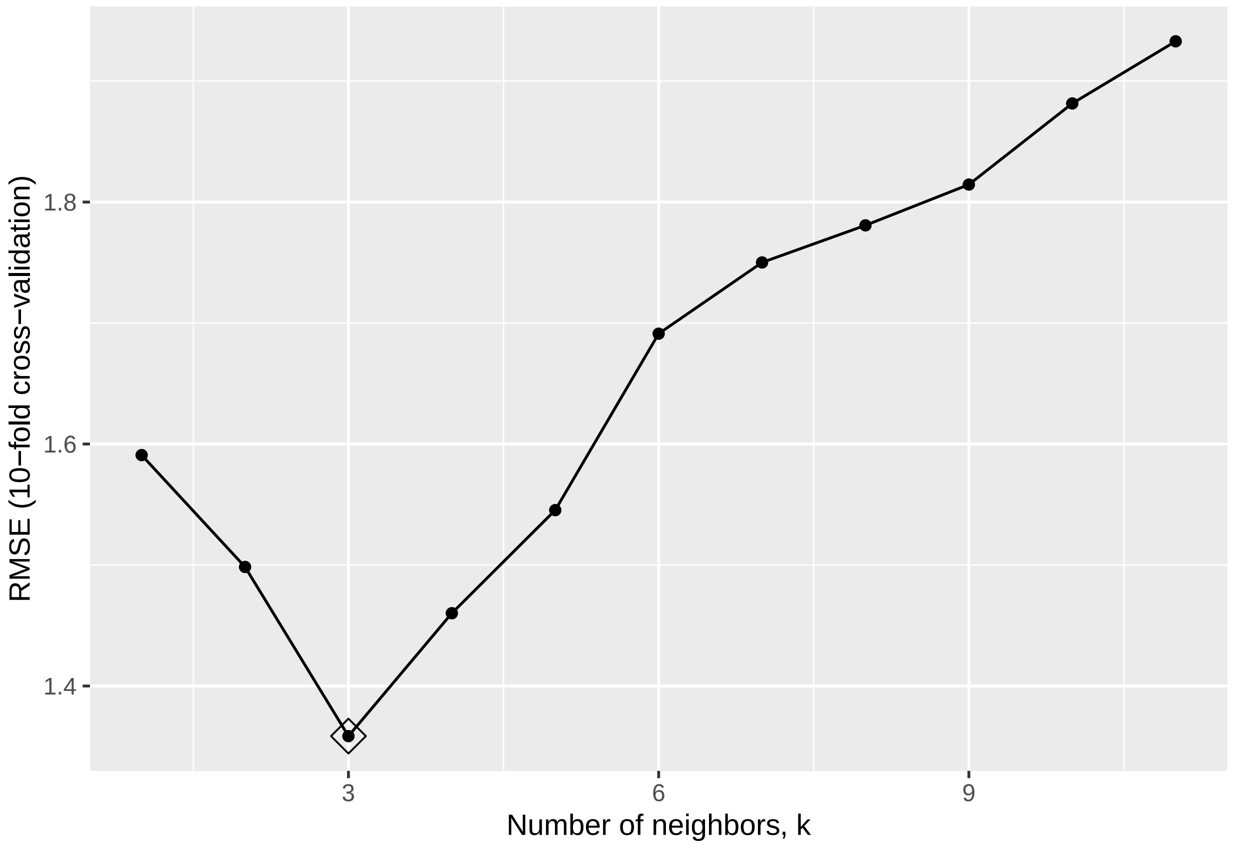
As a baseline, a generalized linear model was used. This is at its core linear regression, however, it can be used to fit non-linear data - and the backing ratio is non-linear. There are no parameters to be tuned for this mode.

The results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.704877

k-Nearest Neighbors

This algorithm outputs the value that is the mean of the k-nearest neighbors of the input value as a prediction. 10-fold validation used to tune k, the number of neighbors. As seen below, the result of cross-validation was choosing k = 3. Therefore, the predicted value of an inputted data point is, in the simplest sense, the mean of its three closest neighbors. In practice, of course, it is more complicated.

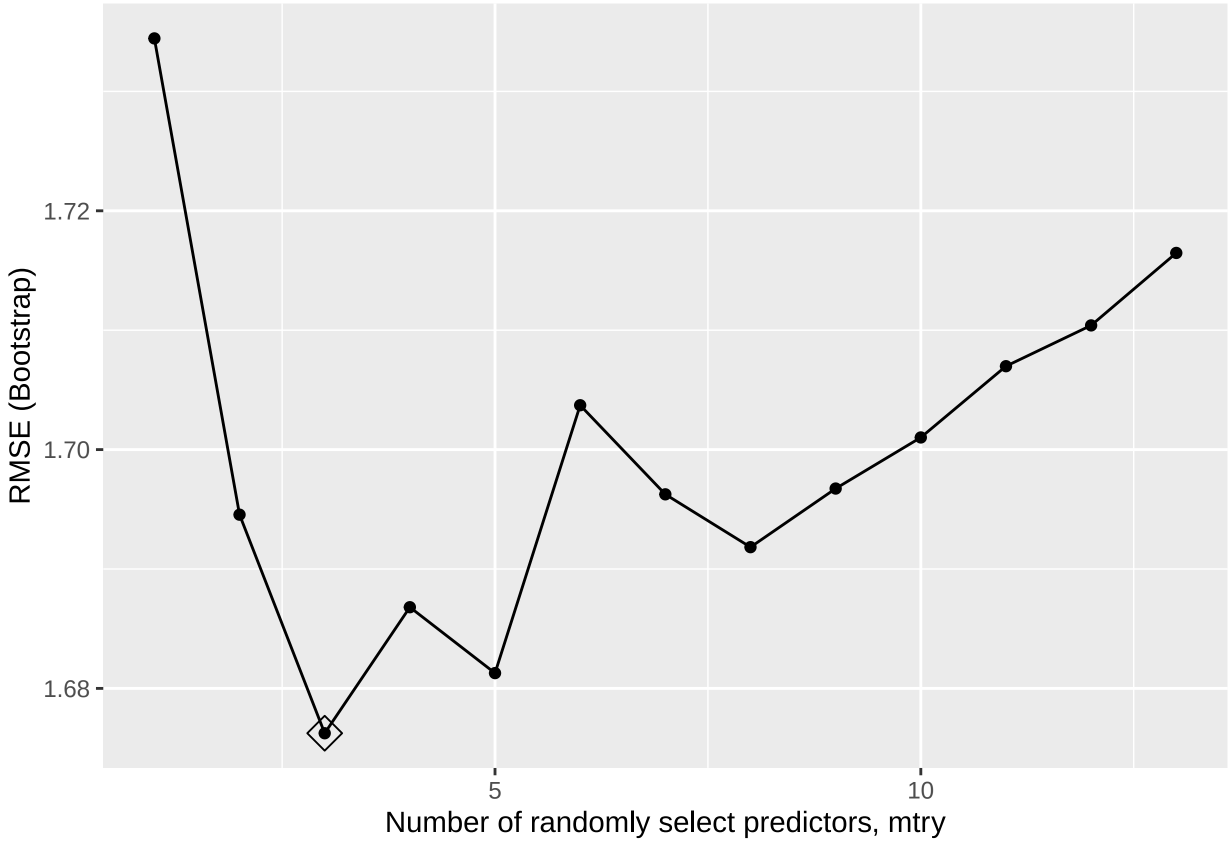


After running the algorithm using the optimal parameter, the results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.7048770
k-Nearest Neighbours	0.8378303

Random Forest

This algorithm is an ensemble algorithm that outputs the mean prediction of the individual decision trees that are formed during the training process. The default bootstrap method used to tune the mtry parameter. The method selected mtry = 3. That is, using 3 randomly selected predictors returned the smallest RMSE.

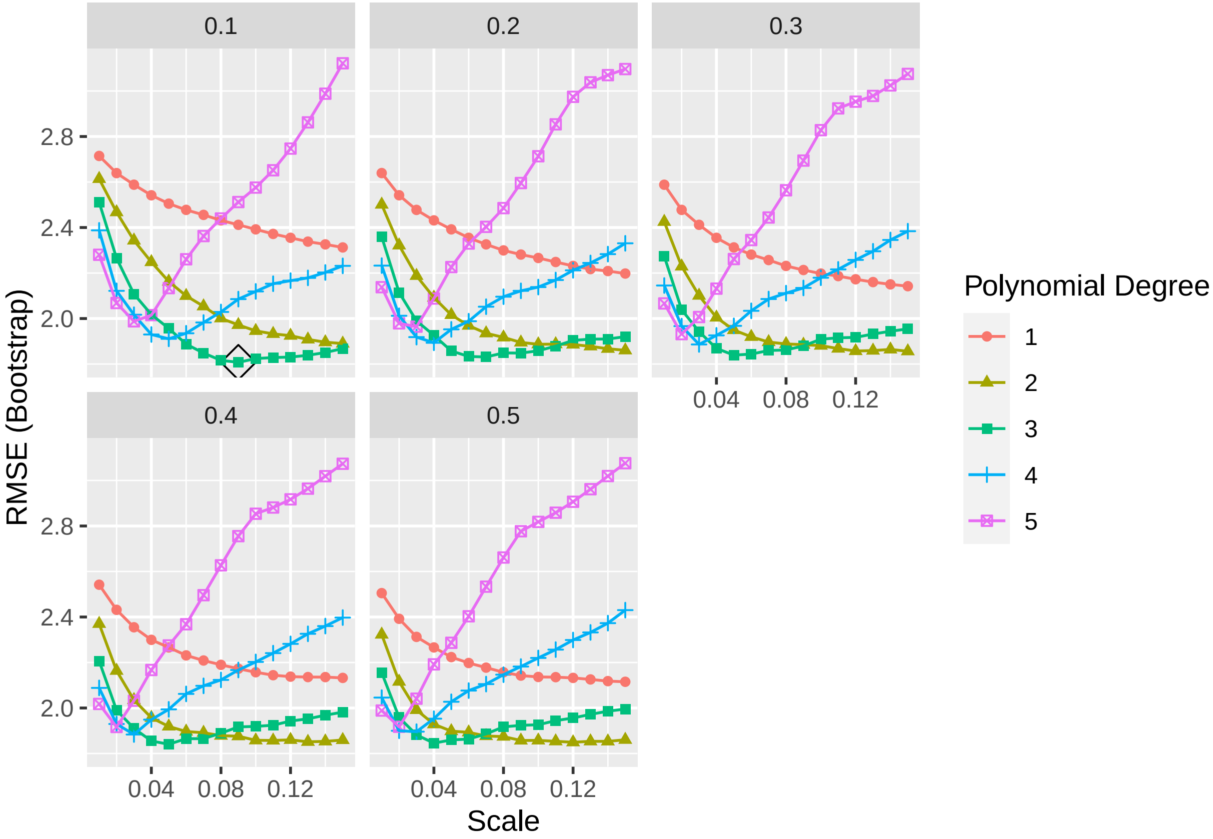


After running the algorithm using the optimal parameter, the results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.7048770
k-Nearest Neighbours	0.8378303
Random Forest	1.1747031

Support Vector Machine with Polynomial Kernel

This algorithm uses hyper-planes (essentially multidimensional lines or planes) to make decisions. The hyperplane which yields the greatest distance between input points is used in the optimal model. A polynomial kernel was used as the predicted variable is not expected to follow a linear model. The default bootstrap method used to tune the polynomial degree, scale, and C parameters of the algorithm. That method resulted in degree = 3, scale = 0.09, and C = 0.1 as the optimal parameters. This is seen in the diagram below.

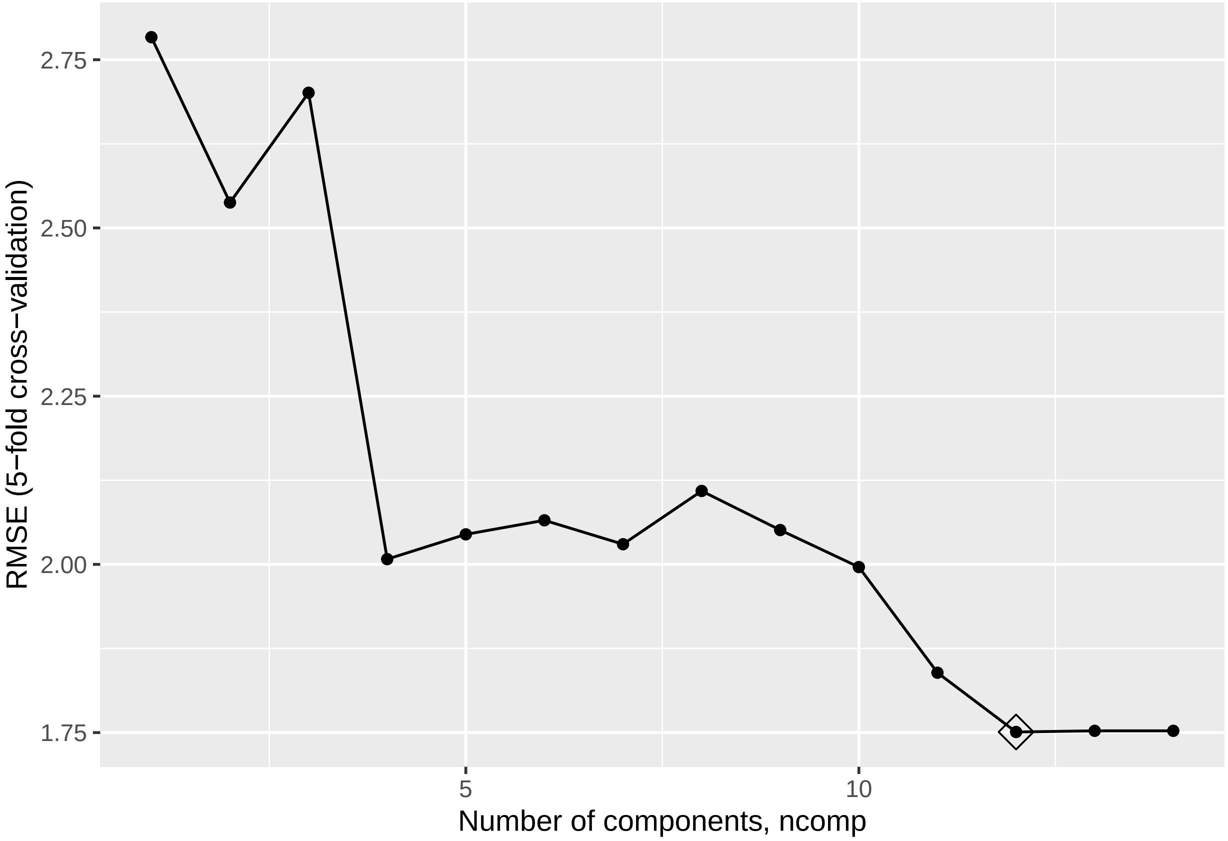


After running the algorithm using the optimal parameters, the results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.7048770
k-Nearest Neighbours	0.8378303
Random Forest	1.1747031
Support Vector Machine with Polynomial Kernel	1.2912638

Principal Component Analysis

This algorithm attempts to reduce the number of predictor variables (or components) by selecting those which explain the most variance in the target variable and are independent among each other. These reduced number of predictor variables are then used to predict the output. 5-fold cross validation was used to tune the parameter, ncomp, number of components. It determined that 12 components were optimal as seen in the diagram below.



After running the algorithm using the optimal parameter, the results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.7048770
k-Nearest Neighbours	0.8378303
Random Forest	1.1747031
Support Vector Machine with Polynomial Kernel	1.2912638
Principal Component Analysis	1.6741932

eXtreme Gradient Boosting

This algorithm is similar to a random forest in that decision trees are a key element in the decision making process. They differ in that random forest create predictions in parallel, whereas this algorithm links the decision trees sequentially. The algorithm was run with default parameters which would have been self tuned. There are four (4) parameters which would have had to be tuned otherwise, but we will not get into these in this project.

After running the algorithm the results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.7048770
k-Nearest Neighbours	0.8378303
Random Forest	1.1747031
Support Vector Machine with Polynomial Kernel	1.2912638
Principal Component Analysis	1.6741932
eXtreme Gradient Boosting	1.4577395

Ensemble

Our very own ensemble model was also created using the 20% trimmed mean. The highest and lowest predictions of the five non-baseline algorithms were cut off and the mean of the other three algorithms' predictions was returned as its prediction.

After running the algorithm the results to this point are:

Method	RMSE_value
Generalized Linear Model (Baseline)	1.7048770
k-Nearest Neighbours	0.8378303
Random Forest	1.1747031
Support Vector Machine with Polynomial Kernel	1.2912638
Principal Component Analysis	1.6741932
eXtreme Gradient Boosting	1.4577395
Ensemble	1.0302645

Results

Modeling results

As seen in the tables below all RMSE results on the test set were better than the RMSE results on the training set. All models improved the performance on the test set. Also observe that all of the RMSEs were better than that of the baseline model, Generalized Linear Model, which is a generalized linear regression. k-nearest neighbors outperformed all other models by far. It was the only algorithm to reach the target of having a RMSE < 1.

TrainRMSE	TrainRsquared	TrainMAE	method
2.036564	0.5251242	1.525957	glm
1.358578	0.7886549	1.030931	knn
1.676246	0.7079121	1.196898	rf
1.807574	0.6486507	1.279759	svmPoly
1.750912	0.6324201	1.327735	pcr
1.627182	0.6824969	1.186664	xgbLinear

Method	RMSE_value
k-Nearest Neighbours	0.8378303
Ensemble	1.0302645
Random Forest	1.1747031
Support Vector Machine with Polynomial Kernel	1.2912638
eXtreme Gradient Boosting	1.4577395
Principal Component Analysis	1.6741932
Generalized Linear Model (Baseline)	1.7048770

Discussion of model performance

When looking at another element of performance, that is time to execute the algorithm, it was observed that Random Forest's bootstrap method was lengthy but it's run on the test set was efficient. The Support Vector Machine with Polynomial Kernel and the eXtreme Gradient Boosting algorithms took the longest to run with the latter taking up the most time. However, these did not result in the best RMSE.

It should also be noted that if the seed is changed, the above results and the performance of the algorithms may change. If this is the case, then the best algorithm may be the Ensemble model rather than k-nearest neighbors.

Conclusion

Brief summary of report

The report shows that all algorithms are able to predict the backing ratio with root mean square errors ranging from 0.8378303 to 1.704877 including the ensemble model which had a root mean square error of 1.0302645.

The best performing algorithm and only algorithm to reach the performance target was k-nearest neighbors with the RMSE of 0.8378303 which may be because of the time series nature of the data, or perhaps randomly because of the seed that was chosen. The next best algorithm was the Ensemble with a RMSE of 1.0302645.

Potenital Impact

The project will continue to be refined as if it is able to predict the backing ratio with an acceptable level of accuracy, then it can be used to aid in policy creation. In addition to predicting the trajectory of the backing ratio, it could be used to determine the effect of risk appetite on backing ratio. This could eventually become a great asset to the ECCB.

Limitations

There are several limitations to this project. The first is that only a small number of records are available on the ECCB's website. More data are available in-house, that is in the database of the ECCB, and so that would increase the number of records available for analysis.

Another limitation is the small number of predictor variables considered here to limit size of project. Additionally, they were utilized without any transformations being done. Also, only variables which had forecast data were used. This would likely lead to a decrease in the ability of the algorithms to predict accurately, and therefore an increase in RMSE. The RMSE would likely be reduced if actual values of the predictor variables were used. However, when predicting something that occurs in the future, only forecast data would be available. If the backing ratio had to be "now-cast", then the actual values of the independent variables could have been considered.

The training and test sets were chosen randomly. However, this method would not be ideal for time series data as the information that the previous point would provide is lost. For future implementations the sampling process would have to be reconsidered.

Future work

In a full scale implementation of the project, data from the International Monetary Funds' (IMF) World Economic Outlook (WEO) database, the Federal Reserve Bank of St Louis' Federal Reserve Economic Data (FRED) database, Google Finance, and Yahoo Finance would be considered.

These websites contain indicators of interest which include but are not limited to: benchmark rates, bond and stock funds, credit spreads, equity prices, gross domestic product (better known as GDP, for ECCU countries as well as international countries of interest), tourism indicators, volatility indices, and yields. Some of these indicators won't have forecasts and so forecasts would have to be made for them before being used in the updated model.

The available periodicity (how frequently the data are compiled) of the above indicators vary widely (from daily to annually) and so thought has to be put into how the NAs would be dealt with for those available with lower periodicity such as semi-annually and annually.

Consider using method = timeslice with appropriate options for initialWindow, horizon, and fixedWindow in trainControl when cross validating. That is, treat the data as time series data instead of independent, identically distributed data.

With an increase in the number of predictor variables, consideration would have to be given to variables showing higher correlations to the target variable. Although algorithms such as Principal Component Analysis would take that into consideration, other algorithms may need assistance.

Due to the time series nature of the data, another variable to consider would be inclusion of a lag variable, specifically the lag of the backing ratio. As both components of the backing ratio (reserve assets and demand liabilities) are stock variables, as opposed to flow variables, the closing position of the backing ratio in one period is related to its closing position at the end of the subsequent period. As it stands in this project, the target variable was predicted independent of this lag. However, being time series data, there should be some link to the lag of the target variable. The number of periods of lag would depend on the periodicity of the target variable.

The predictor variables need to be examined for stationarity, that is whether the mean (trend), variance and autocorrelation are constant. If these do not hold, then transformations of the data would improve performance. These transformations could include but are not limited to differencing the data and performing log transformations on variables. Analysis of the residuals of the models to determine whether correlations exist can also be performed. If there are correlations it could imply that the model is not considering a key variable.

The above, as well as other considerations would definitely build on this project and provide a stepping stone to improve performance with the next implementation.

References

All references are linked within the document.