Assessment 3 – MA5832: Capstone Report

by David Northey

Abstract / Executive Summary

This report investigates key factors influencing Australia's unemployment rate using Neural Networks and XGBoost. A major challenge was the small test dataset (7%), which led to erratic results and reduced model reliability. Additionally, extreme economic events, such as the global pandemic introduced significant variations, making accurate predictions difficult.

The dataset was divided into training and testing sets and the models were assessed using RMSE, MAE and R2. The XGBoost model seemed to be overfit to the training dataset, suggesting that hyperparameter tuning was not sufficient to handle the multicollinearity present in the data. Although data imputation was performed, the Global Financial Crises (GFC) data was removed due to reliability concerns. This may have reduced the model's ability to predict across economic downturns. In contrast, the Neural Network Model demonstrated more robust performance with R2 to data variability.

To enhance predictive performance in future work, random sampling should be considered for data selection using a larger test group. Using data from the GFC era would improve model robustness, the missing data could be imputed and evaluated. Using cross validation will help for the model performances and should be incorporated in future work. Applying Principal Component Analysis (PCA) should also be applied to reduce dimensionality and multicollinearity in the dataset. Additionally, testing alternative models such as Random Forests would be interesting to see and provide valuable comparative insights.

By addressing these challenges, future models can capture the complexities of the Australian labour market and improve unemployment rate forecasting.

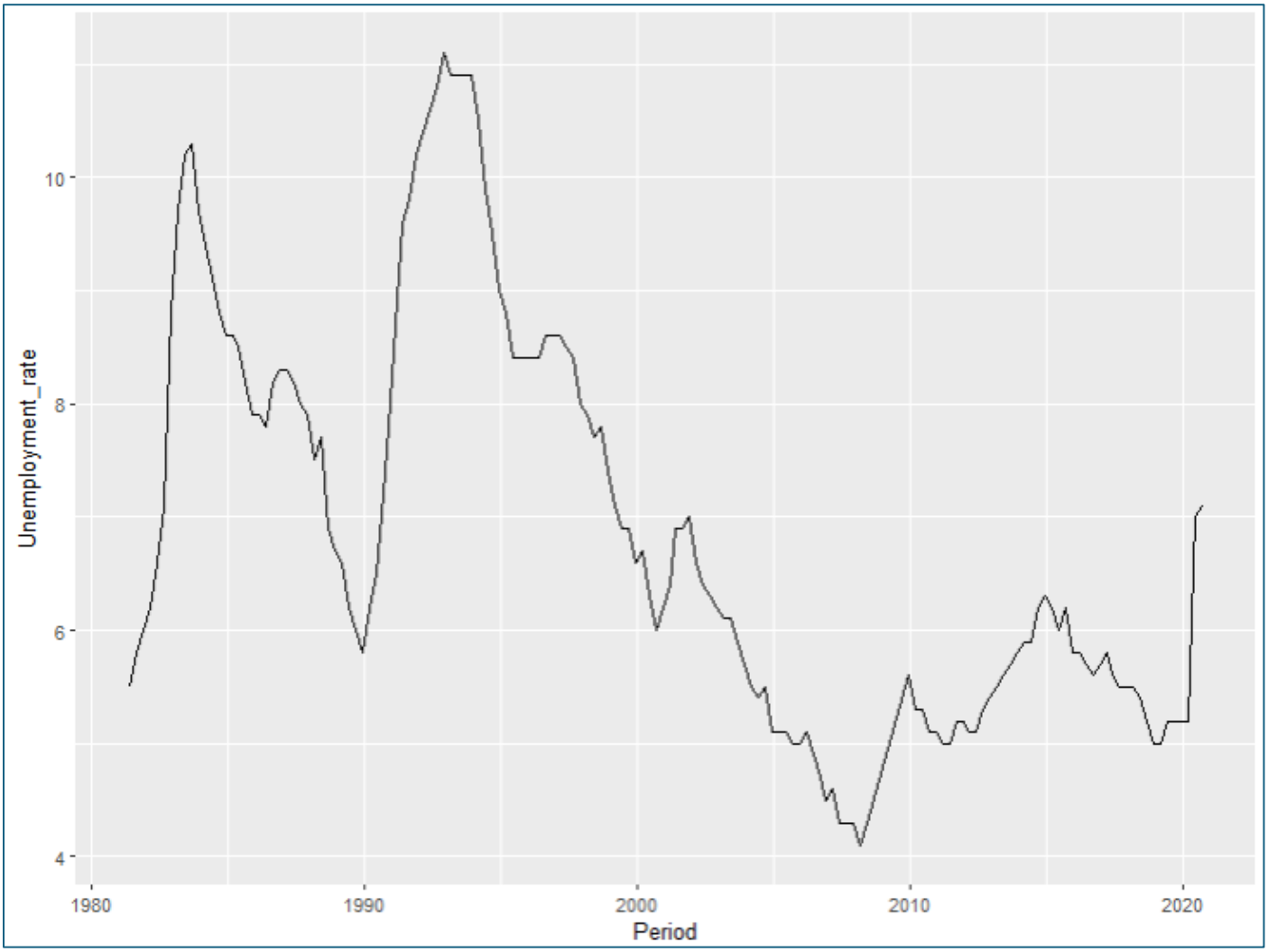
Introduction

The Australian Bureau of Statistics (ABS) defines unemployment as individuals aged 15 and over who are not employed but are actively looking for work (*ABS, 2022*). Employment is a key driver of a country's economic output, enabling people to support themselves and their families. It Is also closely linked to physical and mental health, contributing to overall well-being (*AIHW, 2023*).

Between 1999 and 2007, Australia experienced a period of economic growth, largely driven by a boom in global demand for natural resources (*Kearnes et al, 2011*). During this time, the unemployment rate dropped to nearly 4% by 2007. In July 2000, the Goods Service Tax (GST) was introduced at a 10% rate, providing Australia with a stable source of tax revenue. This revenue supported essential services such as Healthcare, Education and transport, while also creating new jobs (*ATO, 2024*).

The Global Financial Crisis (GFC) occurred in 2008-2009, triggered by the collapse of a housing bubble in U.S. due to Subprime mortgage lending. '*How do you make poor people feel wealthy when wages are stagnant? You give them cheap loans'* (*Lewis, 2015*). Despite the severity of the crisis, the Australian unemployment rate did not experience a dramatic increase, staying below 6% (**Graph 1**). The slight increase was likely due to a decline in exports to countries such as China and the U.S. The Australian economy's resilience can be attributed to government stimulus measures, such as cash handouts and infrastructure projects (*The Treasury, 2012*).

Graph 1: Measure of Unemployment rate as a percentage between 1981-2020



After 2009, Australia entered a recovery phase, initially seeing a decline in unemployment. However, this period was also marked by significant structural changes, including technological advancements in automation that led to a reduction in some low skilled jobs (*McKinsey & Company, 2019*). The COVID-19 pandemic in 2020 further impacted the economy with lockdowns, border restrictions and reduced demand causing widespread job losses particularly in tourism, hospitality and retail sectors (*ABS, 2021*). In response, the government introduced significant economic stimulus packages such as JobKeeper and JobSeeker which helped control the unemployment rate (*Cassells et al, N.D.*).

Data

The dataset provided was an excel spreadsheets with 9 variables by 158 observations:

Table 1: The list of variables used in the provided dataset with description and class

Variable Name (renamed)	Description of variable	Variable class
Period	Time period for the dataset measured quarterly from 1981-2020	Date
Y: Unemployment_rate	Proportion of the labour force that is jobless	Continuous (percentage)
X1: GDP	Numeric value representing economic growth or contraction over a set period	Continuous (percentage)
X2: Goverment_expenditure	Change in government spending eg healthcare, education, defense etc)	Continuous (percentage)
X3: All_sectors_expenditure	Change in economic activity across all industry sectors.	Continuous (percentage)
X4: Trade_index	A measure of export relative to import as a percentage (purchasing power)	Continuous (percentage)
X5: CPI	A key indicator of inflation of a variety of services (food, housing, transport, education etc)	Continuous
X6: Job_vacancies	Number of unfilled jobs actively recruited by employers in the thousands	Integer
X7: Resident_population	Number of austrlian citizens, permanent residents and long-term temporary residents.	Integer

The data was loaded into R studio using the *read_xlsx()* function, with the second row used as the variable names through subsetting. The first two rows were then removed as these were no longer needed. When exploring the dataset, the classes have been assigned the 'character' class. These were then reassigned to 'numeric' class and rounded to one decimal place, except for the first column which was converted to 'Date' class.

When viewing the dataset summary, some unusual formatting appeared in the variable names (e.g. `` `Unemployment rate\r\nPercentage\r\n` ``). To address this, the *gsub()* function was used to remove the unwanted characters. The variable names were then renamed to condense them for better data visualisation and ease of model building.

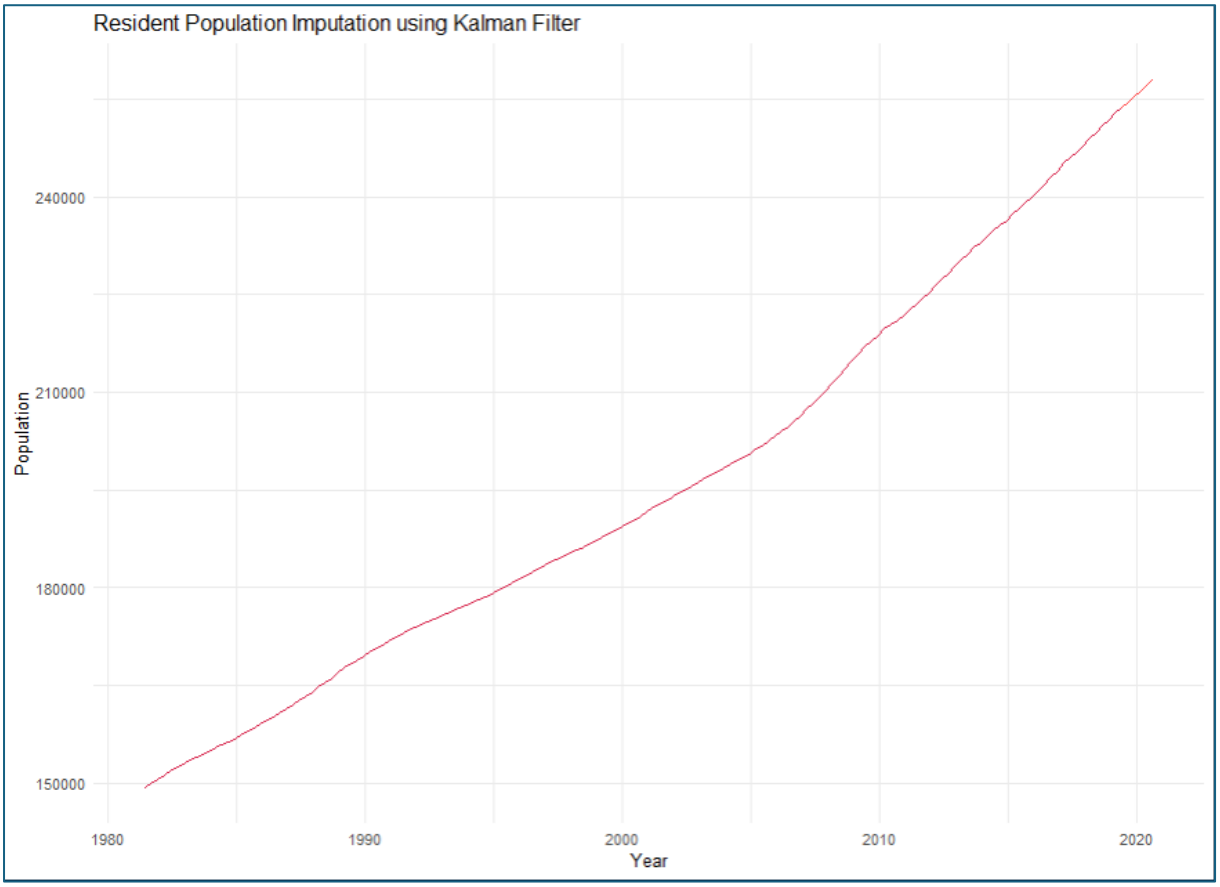
The summary of the dataset also revealed a total of 10 missing values, which will need to be handled since the neural network model cannot work effectively with missing data. There were 5 consecutive missing values for '*Job_vacancies'* between 2008 and 2009. This gap occurred because, due to budget pressures, the Australian Bureau of Statistics (ABS) halted the vacancy survey after May 2008 and did not resume it until November 2009. As a result, the ABS was unable to generate reliable estimates for five quarterly surveys (*The Sydney Morning Herald, 2010*). While it would be ideal to retain this data for model building, these values will be removed to maintain data integrity, as imputing them would be difficult due to the unique nature of the Global Financial Crisis (GFC).

The remaining 5 missing values were for the resident population variable between 2019 and 2020. To ensure valid comparisons in the model, these missing values will be imputed. However, it is important to consider that this period coincided with the COVID-19 pandemic, which had a significant impact on migration to Australia. To impute these values, we will use Kalman filtering which is well suited for time series data. Kalman filtering leverages past trends to estimate missing values and is particularly effective when handling steady, predictable patterns (**Graph 2**). It can adapt to fluctuations and account for uncertainty. Additionally, Kalman filtering maintains the temporal structure of the dataset, unlike other imputation methods (e.g. mean imputation), which do not preserve the structure (*Weeks. M, 1999*).

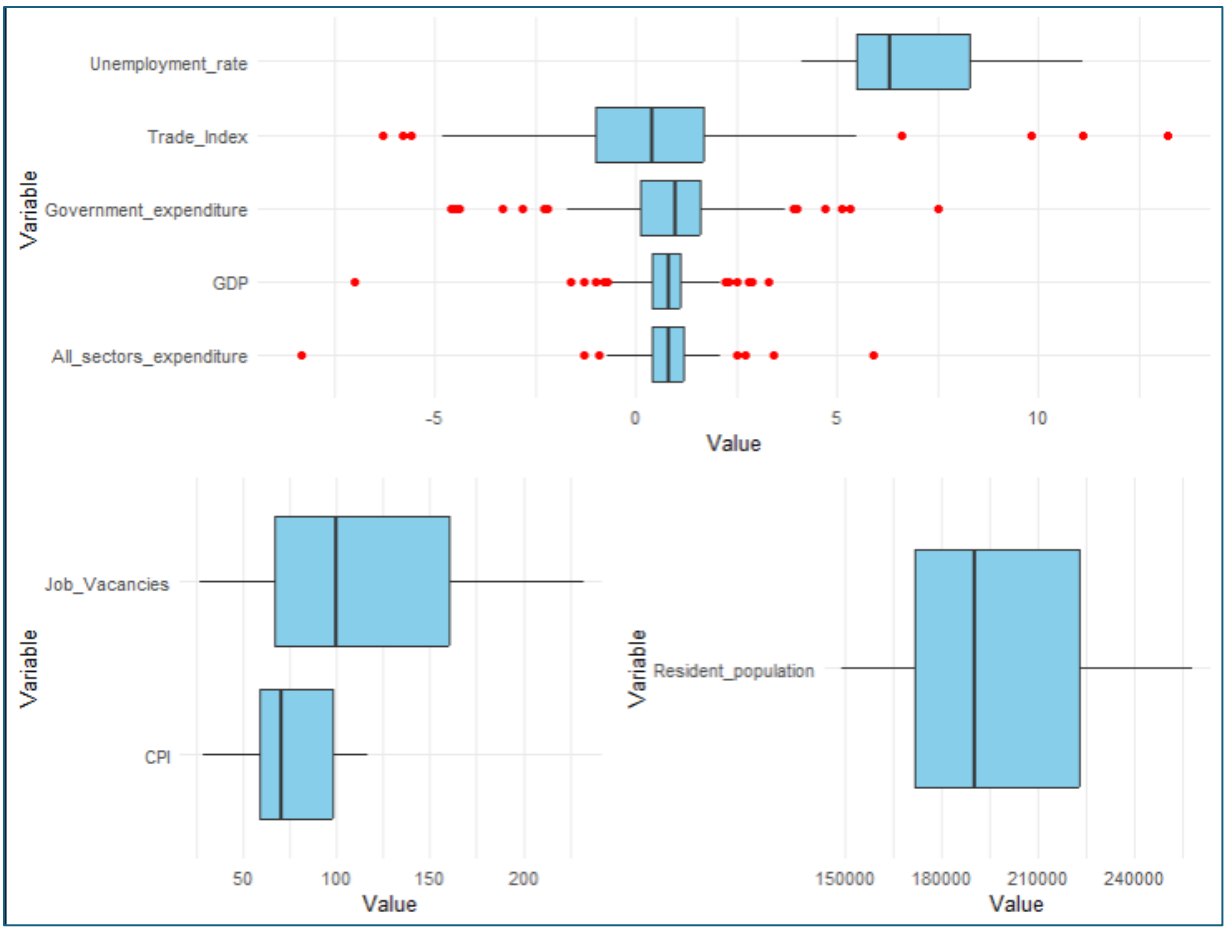
When investigating the dataset, no outliers were detected that could be considered data entry errors. Although large negative outliers were observed for '*GDP'* and '*All_sectors_Expenditure'* these can be explained by the impact of COVID-19 (**Graph 3**).

When checking the dataset for multicollinearity, a strong correlation was found between '*CPI*' and '*Resident population*'. As new migrants enter the country, they often need to purchase items like furniture, which can contribute to inflation. Additionally, the '*Job_vacancies'* variable is highly negatively correlated with the unemployment rate, which is expected. '*Job_vacancies'* is also strongly correlated with '*Resident_population*', as the availability of more jobs tends to attract more people seeking new opportunities. The other variables in the dataset have lower levels of correlation present (**Graph 4**). It will be important to take this into consideration when building the models to fit the regression.

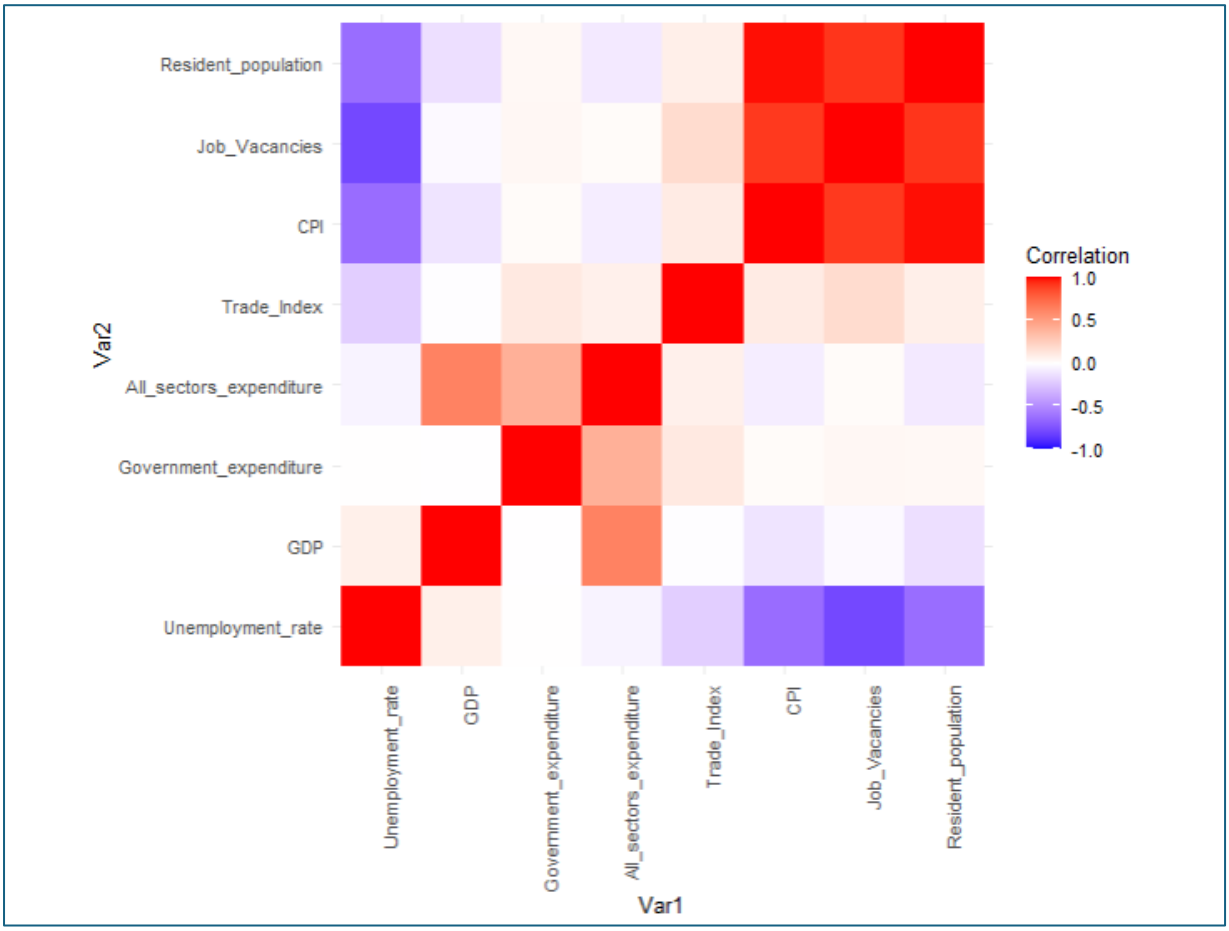
Graph 2: Graph showing the steady predictable pattern of resident population against Period



Graph 3: Boxplots showing the spread of the data for each variable



Graph 4: Heat map showing the correlation between the variables in the dataset



Analysis and Investigation of Machine Learning (ML) method

When selecting the appropriate Machine Learning model from weeks 3 and 4 from this course, we need to take into consideration the information we have learned during our exploratory data analysis. Although the dataset is relatively small, which typically makes Support Vector Machines (SVM) a suitable choice, the presence of multicollinearity (**Graph 4**) suggests that SVM may not perform well in this case. When viewing the dataset, a lot of the variables are non-linear which will lead to very long processing times. Another important consideration is the presence of outliers in the dataset. Given these issues, Random Forest and Boosted trees are good algorithm choices. To develop a model with higher accuracy, we will use the Boosted Trees model for this assessment. Boosted Trees are effective in handling complex, non-linear relationships in the data, and can also be tuned to account for correlation among the predictor variables (**Appendix 1**).

Hyperparameter Selection

The hyperparameters used to build the model were selected according to the instructions from the caret package for the 'available models' section (*Kuhn, 2019*):

nrounds: This controls the number of trees used in the model. More trees can lead to better performance but can increase the chances of overfitting.

Max_depth: The depth of the individual tree is controlled using this hyperparameter. Deeper trees capture complex patterns but increase the chances of overfitting.

Learning rate (eta): This hyperparameter can control how much the model is overfitting, lower values can reduce overfitting but will be more computationally expensive.

Gamma: This is a regularisation parameter, this value controls how many splits occur. A higher value will reduce the number of splits.

colsample_bytree: Controls the fraction of features used for each tree and can help with issues with overfitting.

Min_child_weight: Defines the number of samples required in a node leaf. Higher values makes the model focus on stronger patterns. Controls how the splits can occur.

Subsample: This is the fraction of the training data used for each boosting round. This can help prevent overfitting when used.

To optimise the hyperparameters for the model, we will use the *trainControl()* function, which cycles through different hyperparameter setting and picks the best model based on various different error measures such as RMSE. Cross-validation was also used to ensure the best results by splitting the dataset into sections. The best hyperparameters obtained were:

nrounds = 200, max_depth = 6, eta = 0.1, gamma = 0, colsample_bytree = 0.7, min_child_weight = 1, subsample = 0.7.

These hyperparameters will be used for the XGBoost package, which allows for adjustments of the lambda value. This is important because it ensures the model accounts for the correlation among the variables in the dataset.

Training data performance and Interpretations

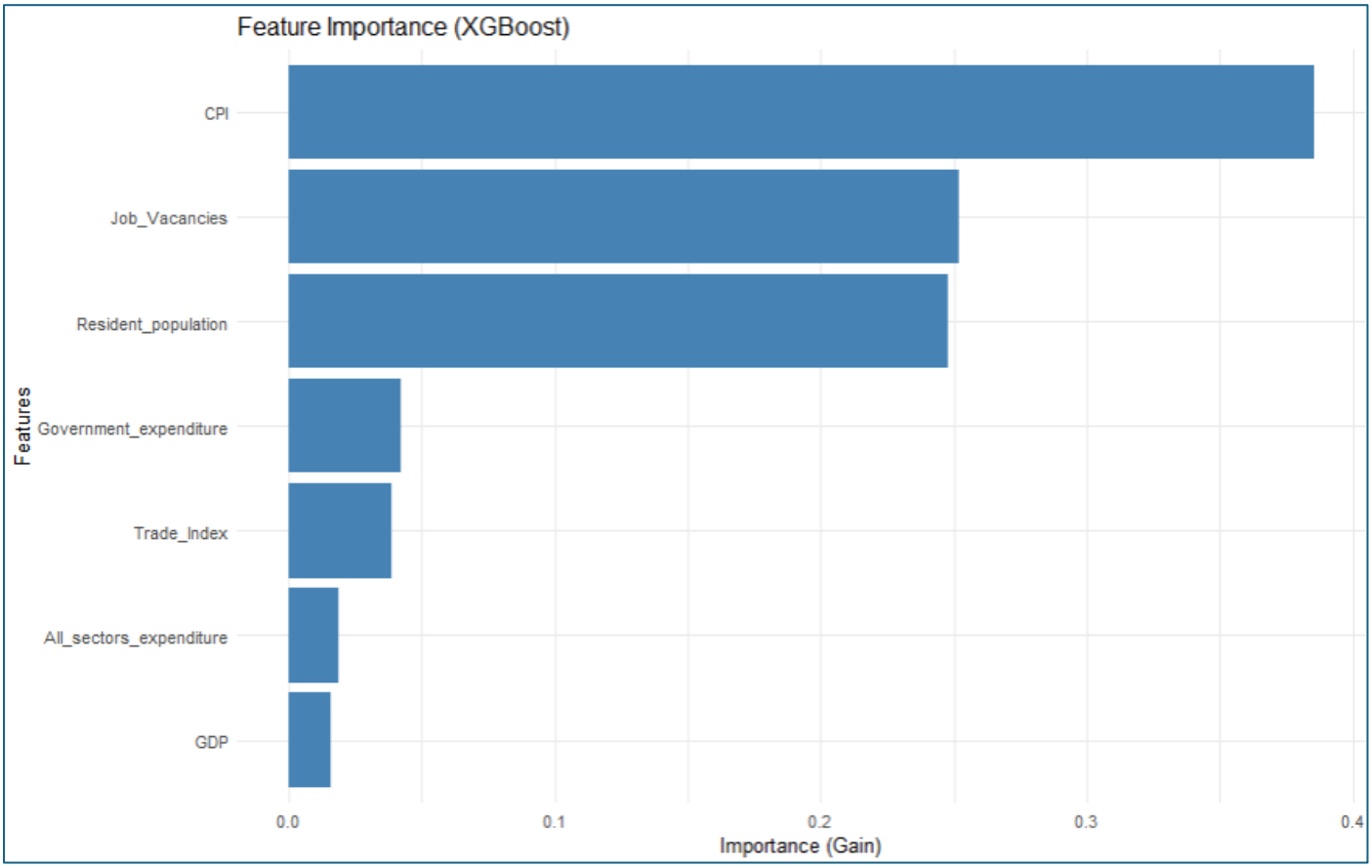
Table 2: Comparison for the XGBoost model with Ridge regression analysis, Training data.

Test	Square Root Mean Error (RMSE)	Mean Absolute Error (MAE)	R2
Ridge_regression	0.9287	0.7566	0.7182
XGBoost	0.0413	0.0243	0.9995

The ridge regression model will be used as a benchmark to compare the boosted model. The MAE and RMSE of the ridge regression model were 0.7566 and 0.9287 respectively. In contrast, the RMSE (0.0413) and MAE (0.0243) of the Boosted model are very low, indicating that it fits the dataset very well with a very low error rate. However, this could also suggest that the model is overfitting the data. We can assess this more clearly by testing the model on the unseen data. Another import note was that the R2 value was measuring almost perfect correlation, this could also suggest the model has been overfit to the training data.

When evaluating the importance, the top three most influential predictors are '*CPI*', '*Job_vacancies*' and '*Resident_population*' (**Graph 5**). This is concerning, as these variables show a strong correlation, which may be contributing to issues with model performance, particularly multicollinearity.

Graph 5: Bar plot showing the variable importance in the data for predicting Unemployment rate



It seems unusual that '*CPI*' is considered a more important predictor of unemployment rate than other variables in the data. Without examining the graph, it could be assumed that variables such as '*GDP*' and '*government_expenditure'* would have a significant effect on the unemployment rate.

Predictive performance of the XGBoost model on the testing data.

Table 2: Comparison for the XGBoost model with Ridge regression analysis for the Testing data.

Test	Square Root Mean Error (RSME)	Mean Absolute Error (MAE)	R2
Ridge_regression	3.1494	1.5731	0.2564
XGBoost	0.7687	0.5139	0.106

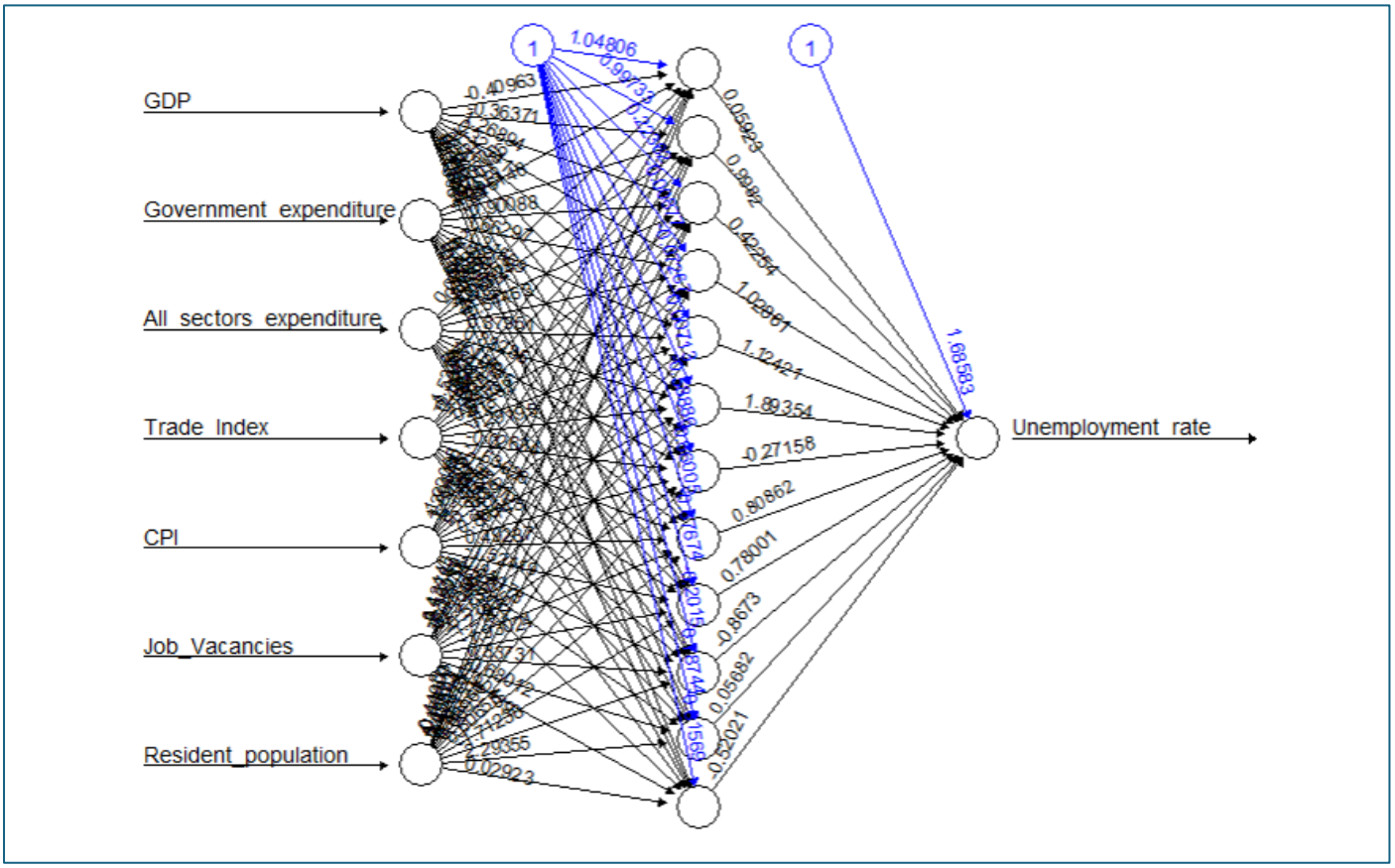
The performance of the model on the test data was decent when compared to the ridge regression analysis. The RMSE (0.7687%) and MAE (0.5139%) are lower than those of the ridge regression model suggesting that the model is predicting the unemployment rate with a reasonable level of accuracy. However, when comparing the errors on the training data to the test data, the test results are significantly less accurate than the training results. Another issue observed is that the R2 value dropped to 0.106, after showing near perfect correlation on the training data. This suggests that the model may have been overfit to the training data, although this will be explored later in the assessment.

Analysis and Investigation neural network (NN) method

The selected Neural network model for this assessment is a feed forward artificial neural network. The input layer consists of neurons corresponding to the number of variables in the dataset. Connected to

the input layer is a hidden layer with 24 neurons using the ReLu activation function. To avoid overfitting, a 10% dropout rate has been applied. The output layer consists of a single neuron, as this is a regression task. The model is trained using RMSprop optimisation with a Mean Squared Error (MSE) loss function, using the MAE to evaluate the performance. The training is performed over 39 epochs with a batch size of 16, using both training and validation datasets.

Graph 6: Neural net model for the dataset using numeric predictors for the unemployment rate



Predictive performance of the Neural model on the training data.

To compare the results, we will again be bench marking against the ridge regression algorithm.

Table 3: Comparison for the Neural net model with Ridge regression analysis -Training data

Test	Square Root Mean Error (RSME)	Mean Absolute Error (MAE)	R2
Ridge_regression	0.9287	0.7566	0.7182
Neural_net 1	0.739	0.5602	0.834

After 39 epochs, the model settled on around 0.9% for the loss (roughly 0.9% RMSE) and 0.8% for MAE after running the model several times. It was noticed that there were some fluctuations in the errors each time the model cycled through all the epochs. When fitting the model to the training data, the neural network model outperformed the Ridge regression model (RMSE: 0.7390, MAE: 0.5062), although the difference in performance is not substantial. The R2 value (0.8340) suggests that the model is making good predictions. The steady decline in both loss and MAE with each epoch indicates that the model is learning correctly and is not overfitting to the training data (**Graph 7**).

Predictive performance of the Neural model on the test data (March 2018 to September 2020).

Table 4: Comparison for the Neural net model with Ridge regression analysis for the Testing data.

Test	Square Root Mean Error (RSME)	Mean Absolute Error (MAE)	R2
Ridge_regression	3.1494	1.5731	0.2564
Neural_net 1	0.8204	0.432	0.5955

The results for the test data were inconsistent each time the model was run. The MAE and loss values varied significantly across different runs. For example, one run produced an MAE of 1.0% and a loss of 2.8%, while another run showed an MAE of 0.6% and a loss of 1.9%, and another yielded an MAE of 1.2% and a loss of 3.9%. These fluctuations suggest that the error rate differs from one run to the next, with the model sometimes predicting the test set accurately and other times not.

When fitting the model to the training data, the RMSE (0.8204) and MAE (0.4320) outperformed the ridge regression model significantly. However, the R2 values are much lower for the test data compared to the training data, indicating that the model is not performing well on the unseen data as it did during training.

Vary the number of the hidden layers in the model. Explore the impacts of the change on the prediction performance of the model

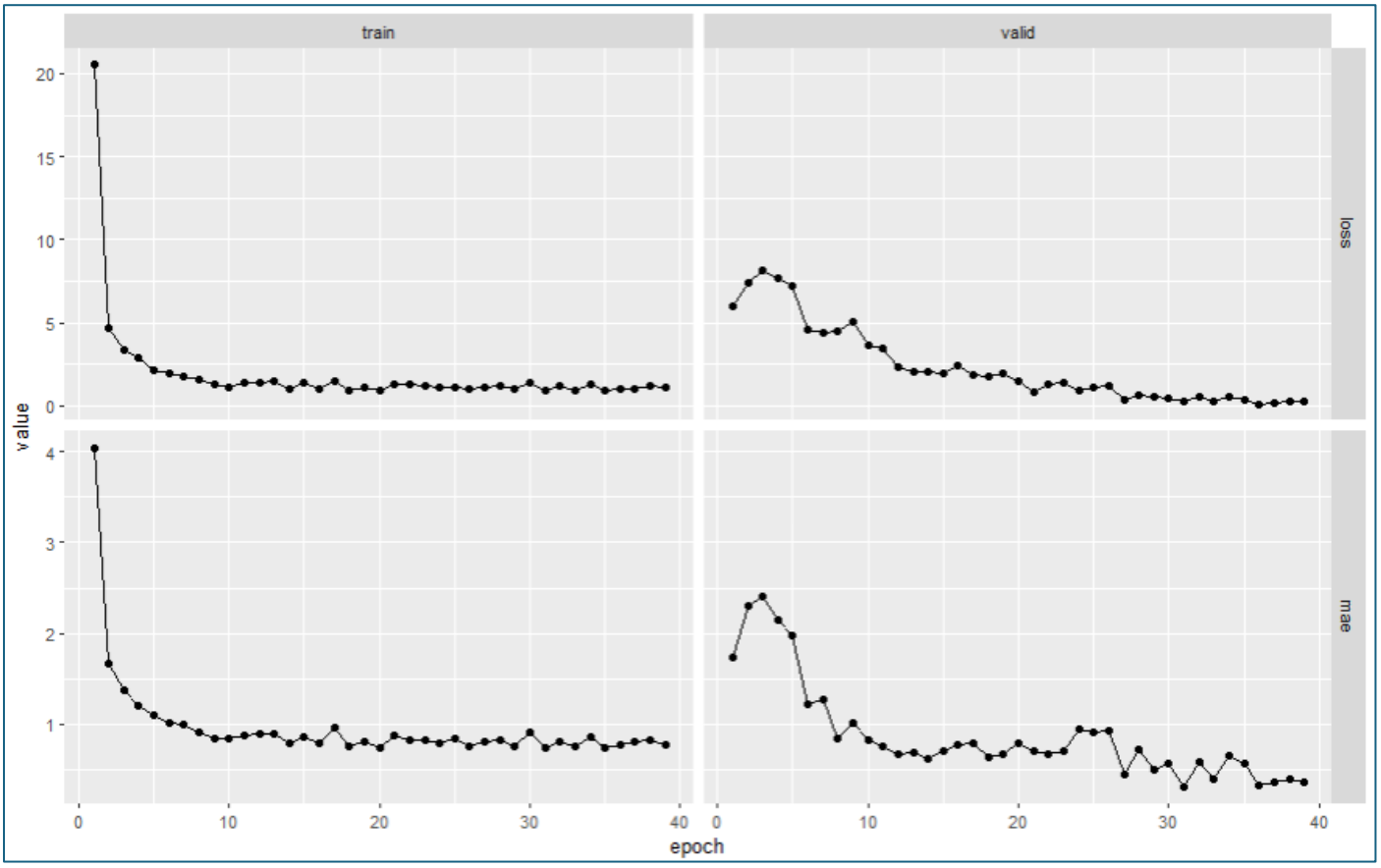
Table 5: Comparison for the Neural net models with Ridge regression analysis for the Testing data.

Test	Square Root Mean Error (RSME)	Mean Absolute Error (MAE)	R2
Ridge_regression	3.1494	1.5731	0.2564
Neural_net 1	0.8204	0.432	0.5955
Neural_net 2	0.8906	0.575	0.496

When increasing the number of hidden layers to 3, each with 24 neurons, the loss and the MAE increased significantly (around 1.7 % and 1.0% respectively after multiple runs) for the training and test data. Additionally, the error fluctuated more than when using a single hidden layer. The error was also slightly higher than when only one hidden layer was used.

When fitting the model to the training data and predicting the test data, the performance was slightly worse compared to the original model with just 1 hidden layer, with higher RMSE (0.8906), MAE (0.5750) and a lower R2 (0.496). As the dataset is not very large, increasing the number of hidden layers added more complexity to the model, making it more prone to overfitting. This resulted in higher and more variable errors across runs.

Graph 7: Plots showing the training and validation data across each epoch for loss and MAE



Vary the number of neurons in each layer in the model. Explore the impacts of the change on the prediction performance of the model

Table 6: Comparison for the Neural net model with Ridge regression analysis for the Testing data.

Test	Square Root Mean Error (RSME)	Mean Absolute Error (MAE)	R2
Ridge_regression	3.1494	1.5731	0.2564
Neural_net 1	0.8204	0.432	0.5955
Neural_net 3	1.113	0.597	0.7137

The model was constructed using one hidden layer, and the number of neurons was increased to 48. The MAE and loss measured about the same as when using 24 neurons and initiating the model, although the error did measure slightly lower (MAE: 0.7%, loss: 0.8%). The test validation data was still quite erratic, although the errors were generally lower and more consistent, with occasional large deviations. When fitting the model, the errors performed worse (RMSE: 1.113, MAE: 0.597), but the R2 (0.7137) slightly improved.

When compared to the ridge regression model, this neural network model outperforms it by a significant margin. Increasing the number of neurons may have allowed the model to capture more complex relationships in the data, contributing to the improved R2 reading.

Comparison and Contrasting ML and NN models

Table 7: Comparison for XGBoost and Neural Net models against ridge regression

Test	Square Root Mean Error (RSME)	Mean Absolute Error (MAE)	R2
Ridge_regression	3.1494	1.5731	0.2564
XGBoost	0.7687	0.5139	0.106
Neural_net 1	0.8204	0.432	0.5955
Neural_net 2	0.8906	0.575	0.496
Neural_net 3	1.113	0.597	0.7137

For each of the models used in this assessment, the RMSE is constantly measuring higher than the MAE. This could suggest that there is a lot of variation in the test dataset, as RMSE is more sensitive to outliers than MAE. Despite this, all models performed significantly better than the benchmark ridge regression model.

Comparing the XGBoost and Neural Network models, I believe the best performing model was the first neural network model, as it had the lowest error rates for MAE. The R2 value for this model was also much higher than that of the XGBoost model, suggesting that the XGBoost model may have been overfit to the training data. The correlation within the dataset likely played a key role here, as high multicollinearity can impact the performance of tree based models like XGBoost. The Neural network models worked more consistently, although if we could eliminate the correlation, XGBoost may perform better.

In terms of interpretability, the XGBoost model could be advantageous, as it allows for the identification of the most significant variables influencing the predictions. As well as this, the hyperparameter tuning process was a lot easier for the XGBoost model. For the Neural Network model, a lot of manual tweaking was required, and it was not clear when the model was performing better or worse due to the large amount of variation in the data.

Computationally, both models performed similarly on this dataset. However, hyperparameter tuning for the XGBoost model was the most time consuming process, making the neural network a potentially better choice in this regard. While the actual training time for both models was comparable, cross validation was only used for hyperparameter selection, not for evaluating the final XGBoost model. Investigating the impact of cross validation on the XGBoost performance in future work would help address overfitting.

Similarly, cross validation fitting was not performed on the neural network model and incorporating it could further improve the performance of the model. Overall, the neural network appeared to be the best performing model, as it seemed more robust to multicollinearity than XGBoost.

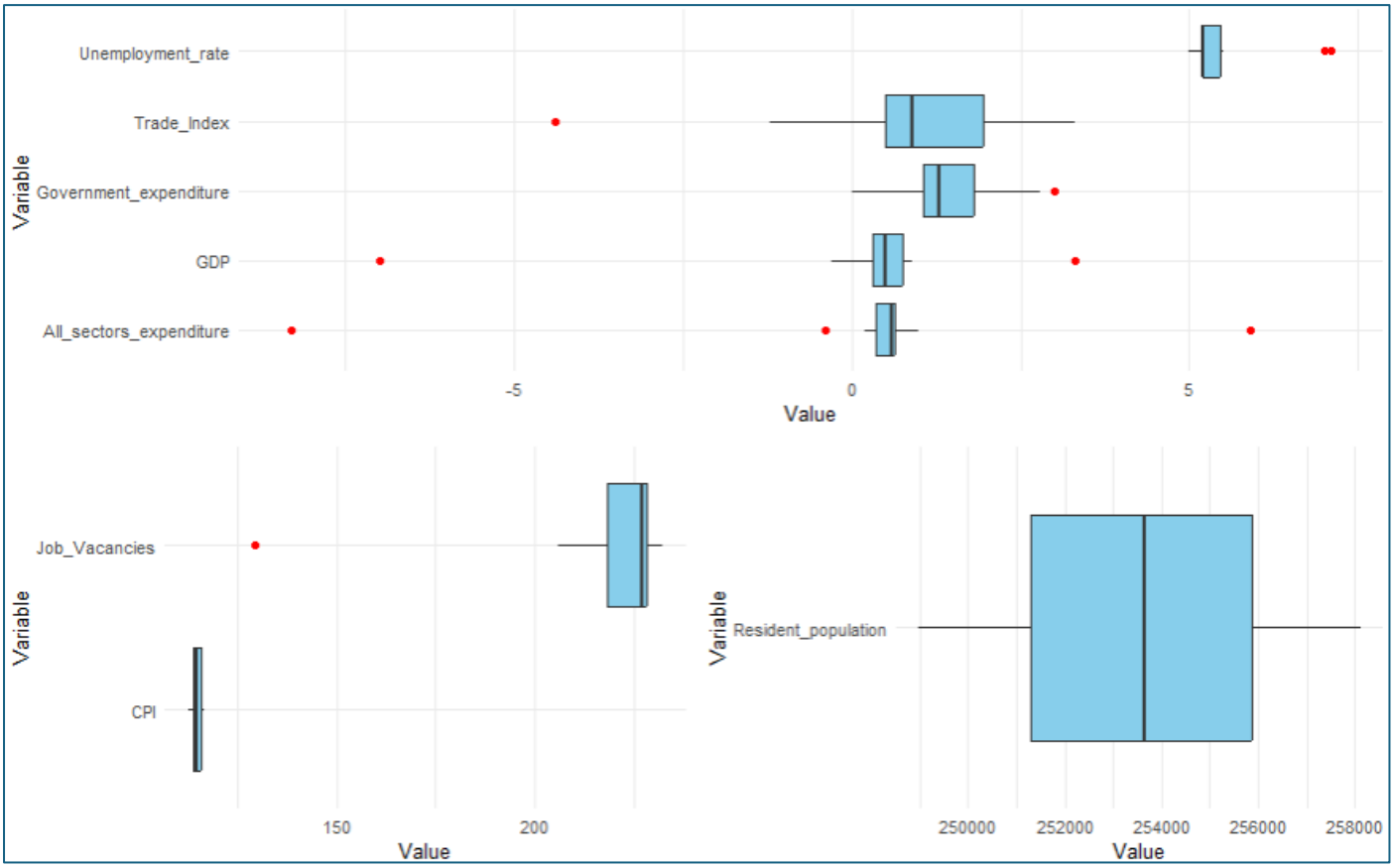
Conclusions

A key issue in this study is the small test dataset, which only represents 7% of the total data and is significantly lower than the conventional 80:20 training-test split. This likely contributed to erratic test errors that were observed throughout this assessment. Additionally, the model was never trained on extreme economic shocks like the global pandemic, making accurate predictions challenging. The test data contains large outliers (**Graph 8**), reflecting significant fluctuations in GDP, expenditure and Job vacancies. Capturing these rapid shifts in a predictive model is challenging.

During data imputation, the decision was made to remove the data during the GFC. However, including this data might have made the training set more robust by introducing greater variability. Future testing could explore whether retaining this data improved the performance of the models. Additionally, a random sampling approach for test data selection could better distribute the noise and improve the performance of the models. It would also be of benefit to explore the data post COVID-19 to assess changes in employment rate and captures the event in full. Future events would take this variability into account when training the model.

The XGBoost model appears overfitted, suggesting that hyperparameter tuning alone, including adjusting the lambda regularisation, was insufficient to handle the dataset's multicollinearity. In future work, Principal Component Analysis (PCA) could be applied to reduce dimensionality, which would improve the performance of the model. Additionally, testing alternative models such as Random Forest, which is more resilient to multicollinearity would be interesting to see how this would perform and provide valuable comparative insights.

Graph 8: Box plots showing the large deviation present in the test data



References

AIHW. (2023). *Employment and Unemployment*. **https://www.aihw.gov.au/reports/australias-welfare/employment-unemployment**

ABS. (2021). *One year of COVID-19: Aussie jobs, business and the economy*. https://www.abs.gov.au/articles/one-year-covid-19-aussie-jobs-business-and-economy

ABS. (2024). Unemployment. https://www.abs.gov.au/statistics/detailed-methodology-information/concepts-sources-methods/labour-statistics-concepts-sources-and-methods/2021/concepts-and-sources/unemployment

Australian Government Australian Tax Office (ATO). (2024). *What GST is and when it is paid*. https://www.ato.gov.au/about-ato/research-and-statistics/in-detail/tax-gap/a-h-tax-gaps/goods-and-services-tax-gap/overview

Cassells. R, Duncan. A. (N.D.). *JobKeepers and JobSeekers: How many workers will lose and how many will gain?* https://bcec.edu.au/assets/2020/03/BCEC-COVID19-Brief-3-Job-Seekers-and-Keepers_FINAL.pdf

Generation. (2024). *Understanding Unemployment: Causes, Consequences and Solutions*. https://www.generation.org/news/understanding-unemployment-causes-consequences-and-solutions/

Kearnes. J, Lowe. P. (2011). *Reserve Bank of Australia. Australia's Prosperous 2000's: Housing and the Mining Boom*. https://www.rba.gov.au/publications/confs/2011/kearns-lowe.html

Kuhn. M. (2019). *The Caret Package*. https://topepo.github.io/caret/index.html

Lewis. M. (2015). The Big Short: Inside the Doomsday Machine, p31. *W. W. Norton & Company*. https://www.imdb.com/title/tt1596363/quotes/

McKinsey & Company. (2019). *Australia's automation opportunity. Reigniting productivity and inclusive income growth*. https://www.mckinsey.com/au/~/media/mckinsey/featured%20insights/future%20of%20organizations/australias%20automation%20opportunity%20reigniting%20productivity%20and%20inclusive%20income%20growth/australia-automation-opportunity-vf.pdf

OECD. (2024). *Unemployment Rate*. https://www.oecd.org/en/data/indicators/unemployment-rate.html?oecdcontrol-4c072e451c-var3=1981&oecdcontrol-4c072e451c-var5=A

The Balance. (2024). *7 Causes of Unemployment*. https://www.thebalancemoney.com/causes-of-unemployment-7-main-reasons-3305596

The Treasury. (2012). *The Australian Economy and the global downturn Part 1: Reasons for resilience*. https://treasury.gov.au/publication/economic-roundup-issue-2-2011/economic-roundup-issue-2-2011/the-australian-economy-and-the-global-downturn-part-1-reasons-for-resilience

The Sydney Morning Herald. (2010). *Job Vacancies rise almost 10%.* https://www.smh.com.au/business/jobs-vacancies-rise-almost-10-20100930-15y9h.html

Weeks. M. (1999). *Methods of Imputation for Missing Data (Fifth Draft),* pp 22-24, https://www.econ.cam.ac.uk/people-files/faculty/mw217/pdf/mispapnw.pdf

Appendix 1: