An Exploration of Machine Learning & Artificial Neural Networks to Predict the Unemployment Rate in Australia

[Document subtitle]

Jayden Dzierbicki

[Year]

Contents

[Abstract/Executive Summary 2](#_Toc127986772)

[Introduction 2](#_Toc127986773)

[Unemployment Rate 2](#_Toc127986774)

[Australia Unemployment Between 1999 and 2020 2](#_Toc127986775)

[Machine learning to predict Unemployment 1](#_Toc127986776)

[Data 1](#_Toc127986777)

[Overview 1](#_Toc127986778)

[Imputing missing values 1](#_Toc127986779)

[Distribution, Outliers & Summary statistics 1](#_Toc127986780)

[Training and Test data set 1](#_Toc127986781)

[Method 2](#_Toc127986782)

[Analysis and Investigation of Machine Learning Methods 3](#_Toc127986783)

[Model selection 3](#_Toc127986784)

[The MARS Model (NB Marker: This was covered in R-Demo week 3/Topic 3) 3](#_Toc127986785)

[Hyper-parameters 4](#_Toc127986786)

[Report On Performance and Interpretation of obtained MARS Model on Training Data 5](#_Toc127986787)

[Report On Performance and Interpretation of obtained MARS Model on Validation Data 5](#_Toc127986788)

[Analysis and Investigation of Neural Network 6](#_Toc127986789)

[Analysis – NN 8](#_Toc127986790)

[Comparison 8](#_Toc127986791)

[Conclusion 8](#_Toc127986792)

[Appendix 9](#_Toc127986793)

[References 11](#_Toc127986794)

# Abstract/Executive Summary

Over the past 21 years between 1999 and 2020 the Australian economy has experienced various macroeconomic events which have had varies impacts on the unemployment rate, a metric which is compiled by the ABS and used to influence policy; and even reported to be linked to election outcomes. This paper aims to explore how machine learning and neural networks can be used to ingest various macroeconomic indicators such as GDP, estimated resident population and XXX to name a few; which we hypothesized could be used to predict and model unemployment due to the correlation amongst these features with unemployment. A common trend we observe for example is when macroeconomic variables such as GDP decrease tends to result in unemployment figures increasing; and we observed this during the global financial crisis around 2008 in the context of Australia. This type of predicting power can be used to predict unemployment at times when agencies are waiting on the ABS to report on the unemployment figure. In our analysis all models were compared against a basic naïve model; the machine learning model we developed was the multivariate adaptive spline regression model (MARS); this model was selected over random forest and boosting due to having the smallest validation RSME when we compared the three models using out of box setting, and we reported the other two models performing worse than the naïve model. Overall, the final MARS model performed better than the naïve mode, with a lower test RSME and cross validation RSME; though the test RSME was slightly larger than our cross validation which we hypothesised due to modelling unemployment during Covid-19 pandemic. In addition, we explored the most basic neura

# Introduction

## Unemployment Rate

The Australian Bureau of Statistics (ABS) is Australia’s national statistical agency which is responsible for collecting employment statistic such as the unemployment rate, one of many employment statistics. The unemployment rate is a key indicator of the labour market performance, providing a snapshot of the available labour supply at a particular time. The survey is conducted on about 26,000 dwellings, with responses from around 52,000 people every month, ensuring it forms a representative sample of the Australia population, with respondents being asked self-guided questions (How the ABS Measures Unemployment, 2022).

## Australia Unemployment Between 1999 and 2020

The unemployment rate is frequently used by the Australian Government to influence policy decisions, especially during election cycles the term gets thrown around a lot, with values allegedly being cherry picked at times when different ministers report on different values to fit their agenda along the campaign trail, possibly using outdated figures (Cockburn et al., 2022; Visentin et al., 2022). This suggests that the unemployment rate is a well-known economic indicator comprehensible by the average citizen as when it is considered high the repercussions have been reported to be linked to election outcomes (Leigh & McLeish, 2009). Over the past 21 years between 1999 and 2020 the Australian labour market has experienced various macroeconomic events which have had influence on the unemployment rate. In the early 2000s it was reported that Australian labour market was performing relatively well after the recession it ‘had to have’ in the decade prior, entering the 2000s with unemployment at around 6 per cent (RBA, 2011). From this period of time unemployment was on a steady decline in Australia until the occurrence of the global financial crisis (GFC) which occurred between 2007-2008, causing extreme pressure on the global financial markets and banking systems; this seen an increase in the Australian unemployment rate between 2008 and 2009 raising from around 4 per cent to 5 ¾ percent due to the lagging effects of the unemployment rate (RBA, 2010), this was the first time since 2000 that Australia experienced such a sharp increase in the unemployment rate, which caused an apparent uptake in interest in policy-makers aiming for lower unemployment rate (RBA, 2011).

## Machine learning to predict Unemployment

The aim of this paper will be to explore the ability of machine learning to predict unemployment based on various macroeconomic indicators, allowing us to explore the possibility of moving away from costly quarterly surveys which are conducted at the expense of the taxpayer, and the possibility of utilising technology to provide metrics which can be updated based on the most recent macroeconomics indicators. We will exore

# Data

## Overview

Data was obtained from the JCU MA5832 content page and downloaded from the assessment 3 folder on 14th February 2023, labled “AUS\_Data.xlsx”; additional data labelled “Dec2020.xlsx” was downloaded on 23rd February 2023 and the row associated with December 2020 was manually pasted to “AUS\_Data.xlsx” to include in the analysis. The data was saved locally and imported into R (version 4.2.2 (2022-10-31 ucrt) for data pre-processing. The dataset provided is aggregated and collected from the ABS which contains data provided quarterly from June 1981 to September 2020, which includes the Australian unemployment rate as well as various macroeconomic indicators which are described in table 1; with a total of 9 inputs and 158. We observed some extreme values and shifts coinciding with the Covid-19 pandemic for X1, X2 & X3 though these appeared to be correct (ABS, Job vacancies, Australia, November 2022; Massimo et al., 2022)

Once loaded into R we called on the str() function to ensure data was imported in the correct data format, we expected all our inputs to be numeric data type except ‘period’ and observed that X1 was imported as character data type, utilising the mutate function we converted X1 to the numeric data type via as.numeric() and summarise our data in table 1 with descriptions. In addition we utilised the lubridate package to split our period into month and year via the month() and year() function respectively, allowing us to capture any time-related pattern or trends based as we anticipate seasonal influence on the unemployment rate (Pettinger, ND).

## Missing values

Utilizing the sapply() function we produced a column wise summary to detect any missing observations in our data set. We observed that both X6 and X7 contained a total of 5 missing observations each, we then called on the ts() function and plot() function to visually inspect the missing observations. We observed missing observations for X6 around the period of the GFC 2008-2009; whilst missing observations for X7 around the period of the covid-19 pandemic 2019-2020. We attempted to locate both original data before attempting to impute any missing observations and were able to locate a data set from the ABS for X7.

**Estimated Resident Population X7:** Data was obtained from the ABS directly (ABS, *National, state and territory population, June 2022*) and saved locally as “310101.xlsx”. Upon further inspection we observed that X7 in the original data frame was reported in the thousands when it should have been in the hundreds in line with ABS reporting. We elected to remove X7 and appended the correct figures from the “Data1” sheet and renamed it X7\_2 linking the data by period onto our data frame.

**Job Vacancies Imputation X6:** We were unable to locate any data directly from the ABS, utilising domain knowledge we assumed that we would expect a decrease in X6 due to the GFC which occurred during this period. Utilizing the imputeTS package we deemed the na\_interpolation() function appropriate which produced values we considered appropriate based on the conditions at the time, which we confirmed visually.

## Distribution, Outliers & Summary statistics

We observe that many of our variables have various distribution types with the plots provided in the appendix, in addition we observe differences in spread and potential outliers evident from the boxplots. Given that many variables are not normally distributed we decided to explore for outliers utilising the IQR approach as opposed to the z-score approach (site) which assumes gaussian distribution which we did not observe with figure 1; we observed many possible outliers based on this approach but assumed these to be true observations and as such did not omit them as they represent extreme economic events such as recessions, and could be valuable to keep them in the dataset to train the model; we confirmed some of the extreme data points by consulting source data and external sources as a sanity test (ABS, Job vacancies, Australia, November 2022; Massimo et al., 2022).

Including such extreme events can help the model to capture the full range of variation in the data and produce more robust predictions. When we develop the machine learning and neural network model, we will discuss the potential impact of these outliers on the chosen model as this can vary between models. Overall table 1 suggests that between the whole data set (1981 and 2020) the unemployment rate in Australia has had a mean of 6.85% and a standard deviation of 1.78%, with a minimum value of 4.10% and a maximum value of 11.13%; this highlights the dynamic nature of unemployment, and other macroeconomic indicators which experience variation in the spread suggesting the importance of retaining the whole data set for the model to train on.

*Table 1: Summary statistics of numeric variables, demonstrating a wide spread of many macroeconomic indicator. Refer to appendix corresponding boxplot summary*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable full name | R-name | Mean | SD | Median | Min | Max | Missing Values Imputed |
| Unemployment rate | Y | 6.85 | 1.78 | 6.26 | 4.10 | 11.13 | NA |
| GDP % | X1 | 0.74 | 0.99 | 0.70 | -7.00 | 3.40 | NA |
| General government ; Final consumption expenditure: Percentage | X2 | 0.85 | 1.66 | 1.00 | -4.60 | 7.5 | NA |
| All sectors ; Final consumption expenditure: Percentage | X3 | 0.77 | 1.11 | 0.80 | -8.30 | 5.90 | NA |
| Terms of trade: Index - Percentage | X4 | 0.35 | 2.95 | 0.30 | -8.10 | 13.20 | NA |
| CPI (all group) | X5 | 75.34 | 25.16 | 73.90 | 28.40 | 117.20 | NA |
| Job Vacancies (000) | X6 | 114 | 57.31 | 103.0 | 26.80 | 259.2 | Yes, imputed utilising imputeTS package |
| Estimated Resident Population (000) | X7\_2 | 19715 | 3103.70 | 19225 | 14923 | 25655 | Yes, all values replaced with updated ABS data |

## Training and Test data set

At this stage the data was separated into two data frames, one titled ‘train’ and the other titled ‘test’. The split for train was applied via the filter() function on data points occurring prior to March 2018 whilst test included March 2018 and onwards, the train and test data frames contained 147 and 12 observations respectively and each with the 10 variables summarised in table 2. In addition, we elected to remove “period” from both test and train via the select(-period) function as we would capture the cyclical nature of unemployment from the engineered feature of month and year. Additional data processing steps will be discussed separately depending on the model chosen. In addition, we decided to train our models on all available data which includes from 1980 onwards to capture the various economic events which have occurred throughout history and not limit us from 1999; whilst not shown here or discussed any further we observed an increase in RSME/MSE when we excluded data points prior to 1999.

*Table 2: Summary description of variables both, with the inclusion of engineered features such as month and year. Additional data processing steps might occur at a later stage depending on underlying model used.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable full name | R-name | Datatype/R Coercion | Description | Comment |
| Unemployment rate | Y | numeric | Unemployment rate in Australia |  |
| GDP % | X1 | Numeric/as.numeric() | Gross | Imported as character, converted to numeric manually. |
| General government ; Final consumption expenditure: Percentage | X2 | numeric |  |  |
| All sectors ; Final consumption expenditure: Percentage | X3 | numeric |  |  |
| Terms of trade: Index - Percentage | X4 | numeric |  |  |
| CPI (all group) | X5 | numeric |  |  |
| Job Vacancies (000) | X6 | numeric |  | Imputed variables around time of GFC |
| Estimated Resident Population (000) | X7\_2 | numeric |  | Replaced X7, obtained from ABS due to errors in original supplied data. |
| Month | Month\_lag | Ordinal categorical/as.factor() | Month derived from period, converted to factor with 3 levels | Manually derived from period, allowing us to capture potential cyclical nature. |
| Year | Year | Ordinal categorical/as.factor | Year derived from period, converted to factor with multiple lelevels |  |

# Method

We undertook two forms of supervised learning in the form of (1) multivariate adaptive regression spline model and (2) neural network model; with both approached utilising the training data set described above; with some minor amendments discussed in the preceding section depending on the model. All algorithms were conducted with R and RStudio(Version 4.2.2) on **a x86\_64-w64-mingw32** platform. We utilised set.seed() function for reproducibility via selecting the seed as 123. An important assumption made at this stage is that all the data is correct despite some minor imputations occurring for X6; and X7 being replaced as discussed previously. We assume that any outlier represents extreme economic conditions which have been document in the literature and as such should be retained (ABS, Job vacancies, Australia, November 2022; Massimo et al., 2022). The goal of this task is to test our ability to predict unemployment based on various current metrics which we summarise with the following equation, we did not explore the possibility of forecasting unemployment, though lagging the variables and removal of the unlagged variables could be used instead and explored, though the discussion and comparison around machine learning and neural networks would still be valid:

Since our primary objective was to develop a predictive model, we elected to utilise the root mean squared error (RMSE) as our metric of choice when evaluating the performance of our model, which we can define formally with the following equation:

In addition, we compared all our model’s performance against a naïve persistence model. The naïve model simply assumes that there is no change in unemployment rate from one period to the next. This is a very basic and naïve model; but serves an important purpose as it allows us to compare a reference point on our more advanced models (Nair, 2022) which we describe below. We defined our naive model with the following simple equation, which resulted in a validation MSE of 0.6122222 (RSME 0.7824463) as a baseline for our reference, which can be explained utilising the following equation:

# Analysis and Investigation of Machine Learning Methods

## Model selection

We compared various supervised machine learning models such as boosting, random forest and multivariate adaptive regression splines (MARS); we briefly reviewed the strengths, limitations and assumptions of each model as summarised in the appendix to assist with selecting a model based on the characteristics of the particular data set which we described in the data section; allowing us to take into account elements of the underlying data such as (1) size, (2) complexity and (3) variable type. Though we observed these models shared many similarities making selecting the model more complex; we decided to run three “out of box” models with default setting in R, evaluating the residuals and MSE before selecting the desired model we would then attempt to optimise.

Based on the table below we observe the following observations about the bias-variance trade off associated with the three “out of box” models, we proceeded with the MARS model as it had the lowest test MSE and second lowest train MSE. Whilst the random forest had the lowest train MSE the naïve model outperformed the boosting model, and the random forest was just better with a 0.01 reduction in test MSE comparatively. Reviewing the training residual plots the random forest displays evidence of systematic bias for when unemployment rate was below 5%, then proceeds to show an upward trend as unemployment increases; while the MARS training residual plot showed improved randomness with little obvious pattern or structure which is supplied in the appendix. Overall, all models appear to be eluding to overfitting due to differences in high test MSE vs low train MSE which could be associated with low bias and high variance. All models except boosting outperformed the naïve model. Whilst not shown here, we undertook grid searching to optimise hyperparameters for both random forest and boosting which yielded results worse than MARS.

*Table 3: Performance of out of bag models comparing test and train RSME, both Random Forest and Boosting performed poorly with a test RMSE larger than the naïve model. We selected the MARS model due to performance at this stage.*

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | Random Forest | MARS | Boosting |
| Train MSE | RMSE | 0.02870791 | 0.1694341 | 0.04736039 | 0.2176244 | 0.08096076 | 0.284536 |
| Test MSE | RMSE | 0.6030101 | 0.7765373 | 0.4132105| 0.6428145 | 1.723489 | 1.312817 |
| Comment on train residual plot | Residual plot is not random, upward trend from left to right. All residuals when unemployment less than 5% are negative. | Residual plot is random, some possible outliers. All residuals when unemployment less then 5% are negative | Many residuals are below 0 indicating possible bias in the model |
| Comment on test/train MSE | Model could be overfitting to training data. Low train and high test MSE suggest model not generalizing well to unseen data | Model could be overfitting to training data. Low train and high test MSE suggest model might not be generalising well to unseen data | Smallest difference, though test MSE almost identical to naïve model. |
| Run time | 0.03983212 secs | 0.00947094 secs | 0.006984949 secs |
| R-Package used | randomForest | earth | gbm |

## The MARS Model (NB Marker: This was covered in R-Demo week 3/Topic 3)

The MARS model is commonly used for complex non-linear regression problems which we observed based on our underlying data, in addition the MARS model has been reported in the literature to predict and forecast models such as unemployment, energy demand and sales (Lu et al., 2012; Katris, 2019) ; all which have elements of time and non-complex linear relationships. The MARS model was first proposed by Fridman (1991) as a nonparametric statistical method based on divide and conquer strategy in which the training data set is partitioned into separate piecewise linear segments (splines) with various gradients, it can be thought of as a modification of the CART which allows it to improve the performance in the regression setting; with no underlying assumptions made about the underlying functional relationships. Like CART, the MARS model employs recursive portioning strategy to divide data into subsets on their predictor variables, allowing it to fit models on these subsets. Though, MARS utilises basis functions (BF) such as hinge functions and constant function to model non-linear relationships between the predictor and the response variables within each subset, allowing MARS to capture more complex nonlinear relationships then CARTS. The connection points between the pieces are referred to as knots, marking the end of one region of the data and the beginning of another which is achieved though minimising some loss function such as the RMSE.

Intuitively we can describe MARS by a simple model in 2-dimension using our data, that is modelling unemployment on estimated population. The MARS procedure will first look for the single point across the range of x values (estimated population) were the two different linear relationships between Y (unemployment) and X (estimated population) achieves the smallest error through the loss function. This result is referred to as a hinge function h(x-a), where a is the cut point value. For a single knot we have the following hinge function as h(x-21,000) approximately as seen in figure 1 summarising our simple two linear model MARS model:

Chart

Description automatically generated

Figure : Simple 2-dimensional demonstration of MARS model and an equation with a single knot corresponding to the blue line.

Once we find the first knot the model will then continue and search for a second knot, if successful it will result in three linear models for Y (unemployment) like the above with the following hinge functions h(x-17,000) & h(x-20,500) approximately. This highlights that the number of knots is an important hyperparameter which we must tuned to avoid over fitting or underfitting of our model which we will be discussed below. The concept of using knots in MARS is like the use of splits in the CART model. Both models employ a recursive partitioning strategy to split the data into subsets based on the predictor variables. The difference is that CART uses binary splits to create branches, while MARS uses a continuous set of knots to create piecewise linear segments or splines. In MARS, the number and position of the knots are selected to minimize the loss function and create the best model, this can be achieved by toggling nprune which are the maximum number of terms (including intercept) in the final pruned model, keeping in mind the algorithm will determine the optimal. In CART, the number and position of the splits are selected to optimize the impurity reduction or the information gain criterion.

## Hyper-parameters

In the MARS model there are two important tunning parameters which must be considered, the maximum degree of interactions and the number of terms retained in the final tuned model. This can be achieved by performing a grid search to find the optimal number of combinations which minimises a loss function such as the RSME. A good starting point is to set up a grid which assess 30 different combinations of interaction complexity (degree) and the number of retained terms in the final model (nprune) it is noted that there is little benefit in assessing greater than 3-rd degree interaction and starting out with 10 evenly spaced values for nprune; which we can later tweak if needed (Greenwell & Boehmke , 2020). This was achieved using the caret package with the earth method applied to the train() function, using the train data set we previously described. The model took 38 seconds to run and provided the following optimal combinations include first degree interactions with 23 retained terms. Furthermore, we attempted to improve our model by performing a second grid search with a range from 10:23 to see if we could further improve our model based on the upper bound discovered in the first grid search, which resulted in nprune 21 being selected with first degree interaction. The cross validation RMSE for the final optimal model was 0.3548613 with a total run time of 61 seconds including both iterations of the grid search, with the plot provided in the appendix

## Report on Performance and Interpretation of obtained MARS Model on Training Data

We have compiled several metrics in table 4 in relation to the performance of our MARS model on the training data set. What we observe is that the training RMSE of 0.2176244 is less than the cross-validation RMSE of 0.3548613 which could suggest that the model is overfitting to the training data; though the difference is relatively small which could suggest that the model may not be significantly overfitting, though we did not explore this further. Given the cross validation RMSE is less than the RMSE on our naïve model suggest the MARS model is performing relatively well on unseen data. In addition the MARS model deetmined the following top five varibles as being the most important when predicting unemployment, (1) job vacanices, (2) CPI, (3) estimated population, (4) year 1983 and (5) year 1999. We were not shocked by the inclusion of job vacanices and CPI as being the most important varibles; though the inclusion of 1983 and 1999 as cut-off periods for the hinge function as important varibles hihglights an area for us to invesitgate, as such we discovered these periods to be associated with an increase in the unemployment. Whilst they are indeed important for predicitng unemployment based on the training data, they are not representative of the current state of the economy and as such the model may overfit to these vribles and perform poorly when presented with new data from different periods; this could be a reason why we hyptohsis an increase in the repeated 10-fold-cross validation RMSE and highlgiths an area for us to further investigate to consider altenrtive approaches.

*Table 4: Performance of tuned MARS model on training data with cross valdation accuracy included. We observe that the model perfoms relativly well on the training data. In addition, cross validation accuracy was less than the naïve model RMSE suggesting the model is able to adapt to unseen data. Possible evidence of overfitting.*

|  |  |
| --- | --- |
| Metric | MARS Model (nprune = 20, degree = 1) |
| RMSE | 0.2176244 |
| Repeated 10-fold- Cross-Validation RMSE | 0.3256355 |
| Repeated 10-fold-Cross-Validation R Squared | 0.9614283 |
|  |  |

Whilst the residual plot shows randomness, we do observe some possible outliers which could warrant further investigation, though we did not explore this further and mark it as a limitation of the model. We also observe that as unemployment increases the residual plot appears to become narrower which could suggest improved accuracy in predicting unemployment when unemployment is higher.

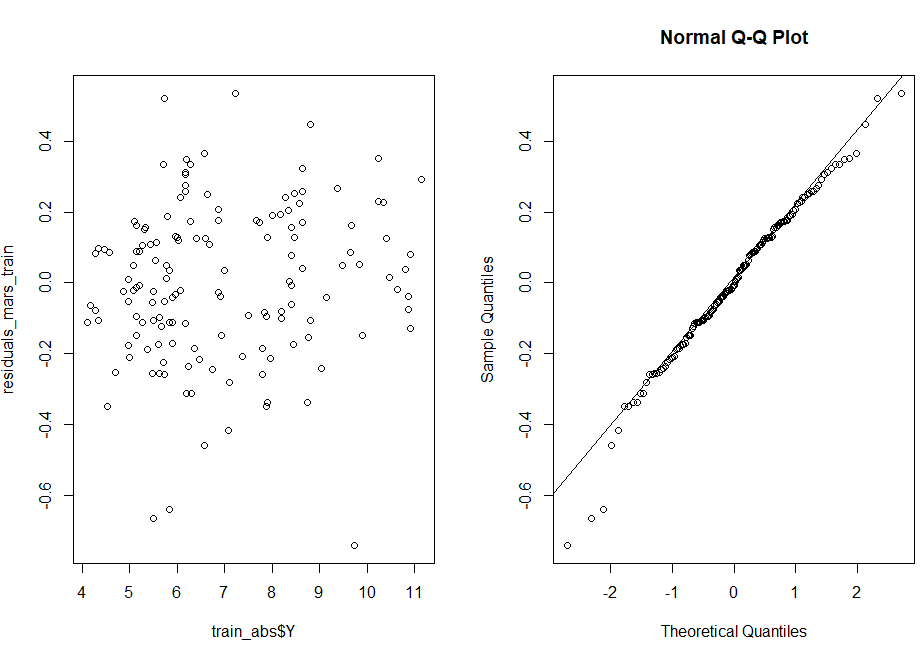


Figure : Residual plot predicted against actual on training data set. We observe evidence of randomness with possibility of outliers. As unemployment increase the residual plot becomes narrower suggesting possible improved accuracy as distribution around 0 still appears random.

## Report On Performance and Interpretation of obtained MARS Model on Validation Data

We have compiled several metrics in table 5 in relation to the performance of our MARS model on the validation data set. What we observe is that the validation RMSE of 0. 6428145 is higher than the cross-validation RMSE of 0.3548613. This suggests that our model did not perform as well on the validation data as it did on other unseen data. However, the model's performance on the other unseen data was comparatively better. Going forward we would suggest acquiring new data for the period 2021 to 2023 to see how it performs on new unseen data as the validation data set only contained 12 observations which may not be sufficient to accurately assess the model’s performance; a common approach is to use a 80:20 ratio for training and validation respectively, which would result in 31 observations instead of the 11 we used based on the current sample size. Ideally going forward we would recommend increasing the overall sample size to include observations from 2020 to 2023 and applying the commonly used ratio. In addition, we highlighted previously that the importance of 1983 and 1999 may lead to overfitting and could be a reason why our model performed poorly on the validation data set as these periods are not associated with the current state of the economy.

*Table 5: Performance of tuned MARS model on validation data with cross valdation accuracy included. We observe that the model perfomrses worse on the validation data then it did on cross validation data, both metrics provided on utilising unseen data. We hyptohisis this could be due to extreme condions caused by covid-19 pandaic as the model was not trained on pandamic like conditions.*

|  |  |
| --- | --- |
| Metric | MARS Model (nprune = 20, degree = 1) |
| RMSE | 0.6428145 |
| Repeated 10-fold- Cross-Validation RMSE | 0.3256355 |
| Repeated 10-fold-Cross-Validation R Squared | 0.9614283 |
|  |  |

In addition to the high validation RMSE we reviewed the residuals of the validation data set and observed that two of the observations in the dataset was underestimated. This occurred when unemployment was predicted lower than it should have been, at around 7% and 6.8% for quarter three and four of 2020 respectively. We have associated this deviation to the impact of the Covid-19 pandemic, furthermore this highlights the need for additional data post Covid-19 to accurately assess the model's ability and increase the validation data set as Covid-19 represents a deviation from the normal, and the model's performance may be affected by such unforeseen events such as Covid-19; again highlighting the vastly different economic conditions from 1993 & 1999 as previously discussed. Whilst unemployment was high at this time, Covid-19 still experienced positive GDP growth at times, as well as high job vacancies due to various government policies and interventions. Whilst not shown here, we were able to improve RMSE for the validation data when we developed a lag MARS model, lagging all variables by one; this resulted in a revised RMSE of 0.6076795 on validation data. We propose to further refine the model the inclusion of lagged variables as well as feature selection to reduce possible overfitting such as removal of year.

# Analysis and Investigation of Neural Network

For this question we elected to use the neuralnet package in R over the keras package, as it allows us to discuss the basic principles of the neural network (NN) and get an understanding of the NN before utilising more sophisticated packages such as keras which are better suited for advanced users with experience in NN models. It will allow us to explore the concept of backpropagation, activation functions and hidden layers; whilst the keras package is better suited once you have more advanced knowledge of NN it also contains a steeper learning curve and as such one should be familiar with the basics and have a strong foundation before exploring more advanced methods, for this reason we believe the neuralnet package is better suited for our learning journey to more advanced packages and programs such as python given this is our first NN model we have developed. We will explore two NN models, a basic vanilla neural network with one hidden layer and another neural network with two hidden layers, utilising a grid search approach to optimise both models with the goal of selecting the model with the smallest test MSE.

The NN model has evolved to include a large class of models and learning methods which is a method of AI which is inspired by that of the human brain; for us to have an understanding of the NN we can use the analogy of the human brain cells, referred to as neurons which form a complex interconnected network which sends electrical signals to each other to help process information. The NN is made up of artificial neurons which aim to solve a problem such as image recognition, text recognition just to name a few applications; these neurons are also referred to as nodes with the most basic vanilla neural network having interconnected artificial neurons in three layers. These three layers comprise of an (1) input layer where information from the outside world enters the NN and is passed onto the next layer, (2) the hidden layer takes input from the input layer (or other hidden layers if more complex) and analyses the output from the previous layer, processes it further and passes it onto the next layer, (3) the output layer gives the final result and prediction by the NN. In the case of regression problem we have one output layer; though classification problems will be set up differently which we do not explore.

# Analysis – NN

# Comparison

We observed in our MARS model, and every out of box model such as boosting and random forest appeared to show systemic bias when unemployment was less then 5%; constantly underestimating unemployment based on the relationship between residuals and observed values of unemployment.

# Conclusion

# Appendix

Diagram

Description automatically generated with medium confidence

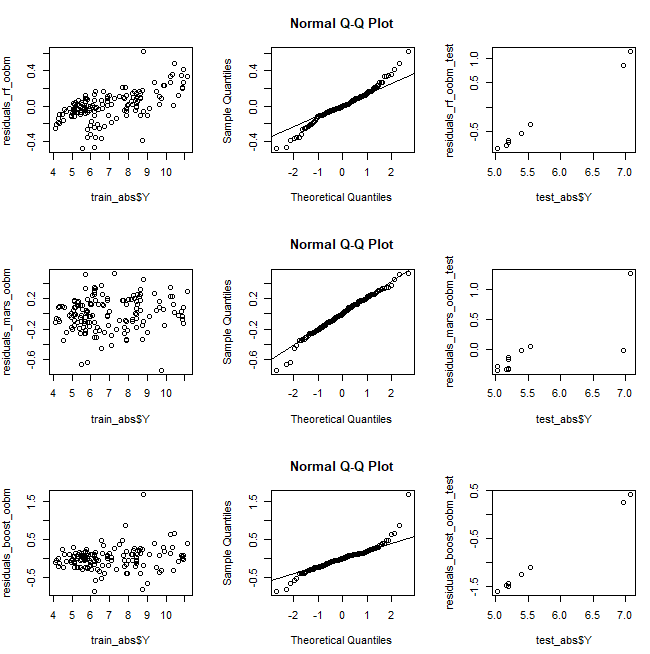
<https://medium.com/swlh/its-all-about-assumptions-pros-cons-497783cfed2d> RF

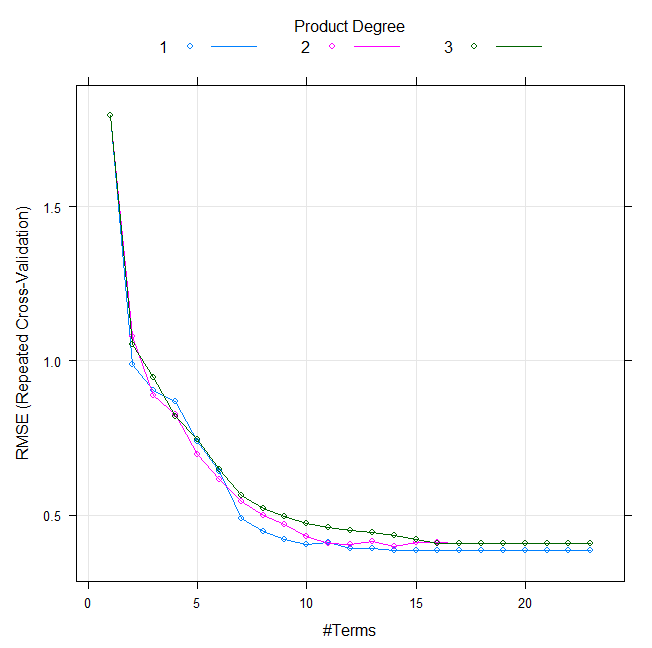
<https://support.bccvl.org.au/support/solutions/articles/6000083217-random-forest> RF

MARS: <https://support.bccvl.org.au/support/solutions/articles/6000118097-multivariate-adaptive-regression-splines>

MARS: https://uc-r.github.io/mars

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Strengths | Limitations | Assumptions |
| Random forest (Bagging algorithm) | Robust to outliers, works well for non-linear data and runs well on large data sets | Can overfit datasets which are noisy | No assumptions made about the distribution, can handle skewed and multi-model data |
| Boosting |  | Hard to tune, can easiy overfit if parameters are not tuned correctly |  |
| Multivariate adaptive regression splines | Works well with a large number of predictor variables, robust to outliers | Difficult to understand then other methods | No assumptions made about the distrubtion, however variables should not be highly correlated as this could cause issues with estimations |





# References

ABS (2022) HOW THE ABS MEASURES UNEMPLOYMENT, Australian Bureau of Statistics. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/6102.0.55.001~Feb%202018~Main%20Features~4.%20How%20the%20ABS%20Measures%20Unemployment~50#:~:text=ABS%20measures%20unemployment%20by%20collecting,and%20indigenous%20communities%20throughout%20Australia (Accessed: February 14, 2023).

LEIGH, A.N.D.R.E.W. and MCLEISH, M.A.R.K. (2009) “Are state elections affected by the National Economy? Evidence from Australia,” Economic Record, 85(269), pp. 210–222. Available at: <https://doi.org/10.1111/j.1475-4932.2009.00549.x>.

Visentin, L., Wright, S. and Curtis, K. (2022) 'I made a mistake': Albanese stumbles on unemployment rate and cash rate, The Sydney Morning Herald. The Sydney Morning Herald. Available at: https://www.smh.com.au/politics/federal/i-m-not-sure-what-it-is-anthony-albanese-stumbles-on-unemployment-rate-and-cash-rate-20220411-p5aci2.html (Accessed: FCockburn, G., Jervis-Bardy, D. and Dennett, H. (2022) 'not just luck': Frydenberg Cheers Steady 4 per cent unemployment rate, The Canberra Times. The Canberra Times. Available at: https://www.canberratimes.com.au/story/7699832/not-just-luck-frydenberg-cheers-steady-4-per-cent-unemployment-rate/ (Accessed: February 17, 2023). ebruary 17, 2023).

RBA (2010) The labour market during the 2008–2009 downturn: Bulletin – March 2010, Reserve Bank of Australia. Reserve Bank of Australia. Available at: https://www.rba.gov.au/publications/bulletin/2010/mar/1.html (Accessed: February 17, 2023).

RBA (2011) The Australian labour market in the 2000s: The quiet decade: Conference – 2011, Reserve Bank of Australia. Reserve Bank of Australia. Available at: https://www.rba.gov.au/publications/confs/2011/borland.html (Accessed: February 17, 2023).

Pettinger, T. (no date) Seasonal unemployment, Economics Help. Available at: https://www.economicshelp.org/blog/glossary/seasonal-unemployment/ (Accessed: February 17, 2023).

Kumar, S. (2022) 4 techniques to handle missing values in time series data, Medium. Towards Data Science. Available at: https://towardsdatascience.com/4-techniques-to-handle-missing-values-in-time-series-data-c3568589b5a8 (Accessed: February 17, 2023).

Nair, A. (2022) Baseline models: Your guide for model building, Medium. Towards Data Science. Available at: https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d (Accessed: February 20, 2023).

Job vacancies, Australia, November 2022 (no date) Australian Bureau of Statistics. Available at: https://www.abs.gov.au/statistics/labour/jobs/job-vacancies-australia/latest-release (Accessed: February 22, 2023).

Massimo, Massimo and FocusEconomics (2022) Lifting of restrictions amid massive fiscal and monetary stimulus prompt rebound in Q3, FocusEconomics. Available at: https://www.focus-economics.com/countries/australia/news/gdp/australia-lifting-of-restrictions-amid-massive-fiscal-and-monetary-stimulus (Accessed: February 22, 2023).

National, state and territory population, June 2022 (no date) Australian Bureau of Statistics. Available at: https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release (Accessed: February 22, 2023).

# REVIEW FOLLOWING ARTICLES FROM JCU

Lu, C.-J., Lee, T.-S. and Lian, C.-M. (2012) “Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks,” Decision Support Systems, 54(1), pp. 584–596. Available at: <https://doi.org/10.1016/j.dss.2012.08.006>.

Katris, C. (2019) “Prediction of unemployment rates with time series and Machine Learning Techniques,” Computational Economics, 55(2), pp. 673–706. Available at: https://doi.org/10.1007/s10614-019-09908-9.