

MA5832-Capstone Report

Weighting: 45% Total marks: 100

Overview

During this assessment you will produce a written report on analyses of a real-world problem using a neural network and an alternative machine learning method that have been presented in this course. For both techniques, you will investigate the properties of the methodologies by altering the machine learning parameters. Additionally, you will compare and contrast the two methodologies using the same data source.

The purpose of the assignment is to enable you to:

- apply machine learning methodologies;
- undertake independent research investigating machine learning parameters;
- compare and contrast machine learning methodologies;
- construct a written communication and interpretation of findings resulting from machine learning methodologies.

Submission

Report Structure

The assignment **MUST** follow the below structure. Otherwise, the assessment will not be marked.

- Title and author
- Abstract/Executive Summary (one page)
- Brief overview on the Australian unemployment rate
- Data. Including any data preparations and sub-setting
- Analysis and Investigation of Machine learning (ML) method
- Analysis and Investigation neural network (NN) method

- Comparison and Contrasting ML and NN models
- Conclusions/findings, together with any recommendations and lessons learned
- Bibliography/References - not included in page count
- Appendices, (e.g. Rcode, additional graphs, keys and definitions) - not included in page count

Submission

You will need to submit the following:

- A PDF file. The assignment can be a **maximum of 12-A4 pages**. References and Appendices are not included in page count.
- **Rmarkdown/R** script file to reproduce your work.
- The task cover sheet.

You have up to three attempts to submit your assessment, and only the last submission will be graded.

A word on plagiarism:

Plagiarism is the act of using another's words, works or ideas from any source as one's own. Plagiarism has no place in a University. Student work containing plagiarised material will be subject to formal university processes.

1 Capstone scenario

Dataset

The data, “AUS_Data.xlsx”, used in the capstone is aggregated and collected from the Australian Bureau of Statistics (ABS).¹ The data is available quarterly from June 1981 to September 2020. The data includes the response variable (unemployment rate) and 7 predictors:²

- Y : unemployment rate measured in percentage
- X_1 : Percentage change in Gross domestic product;
- X_2 : Percentage change in the Government final consumption expenditure;
- X_3 : Percentage change in final consumption expenditure of all industry sectors;
- X_4 : Term of trade index (percentage)
- X_5 : Consumer Price Index of all groups (CPI) ;
- X_6 : Number of job vacancies measured in thousands;
- X_7 : Estimated Resident Population measured in thousands.

Further explanations about the variables can be found in the ABS website.

2 Assessment Tasks

1. Provide an overview of the Australian unemployment rate over the last 21 years (1999-2020) and some insights on factors driving the unemployment rate (provide relevant references when needed; maximum one A4 page). **(10 marks)**

Data

2. Prepare data appropriate for the proposed supervised machine learning methodologies such as: **(10 marks)**
 - (a) implementing appropriate data wrangling procedures, e.g. missing values treatment /transformation of variables.
 - (b) provide and comment on descriptive statistics of the variables.

¹The Australian Bureau of Statistics (ABS) is a national statistical agency, which provides trusted official statistics on a wide range of economic, population, social and environment matters of importance to Australia (<https://www.abs.gov.au/about?OpenDocument&ref=topBar>).

²Some variables are aggregated from monthly to quarterly for the assignment purpose.

Machine Learning

3. Apply one of the supervised machine learning (ML) algorithms from either Week 3 or Week 4 to the data prepared in Question 2 to predict the Australian unemployment rate from March 2018 to September 2020. **(25 marks)**
 - (a) Justify your choice over the other supervised machine learning algorithms.
 - (b) Justify the choice of the hyper-parameter(s) which is required to be specified in R to estimate the selected model.
 - (c) Report the performance(s) and interpretation(s) of the obtained ML model(s) on the training dataset.
 - (d) Discuss the predictive performance of the model on the test dataset (March 2018 to December 2020).

Neural Network

4. Apply a neural network (NN) to the data prepared in Question 2 to predict the Australian unemployment rate from March 2018 to September 2020. **(35 marks)**
 - (a) Describe the structure of the selected neural network model.
 - (b) Report the performance(s) and interpretation(s) of the produced NN models on the training dataset.
 - (c) Discuss the predictive performance of the model on the test dataset (March 2018 to December 2020).
 - (d) Vary the number of hidden layers in the model 4(a). Explore the impacts of the change on the prediction performance of the model.
 - (e) Vary the number of neurons in each layer in the model 4(a). Explore the impacts of the change on the prediction performance of the model.

Comparison and Suggestion

5. Compare the chosen ML model in Question 3 with the NN model in Question 4, and then provide a recommended model. At a minimum, include **(10 marks)**
 - (a) Cross-validated accuracy
 - (b) Computational time to train models
 - (c) Interpretability

6. Provide some suggestions regarding the methodologies/data to further improve the prediction of the unemployment rate of Australia. **(10 marks)**