QBUS6840 Group Assignment

Key information

1. **Required submissions:**
    a. **ONE** written report (word or pdf format, through Canvas- Assignments- Report submission (group assignment)).
    b. **ONE** code file (Jupyter Notebook ".ipynb" or Python ".py", through Canvas- Assignments- Code submission (group assignment)).
2. For the submission, each group should pick up a group **representer** who needs to submit both files. **Each group should only submit one report and one code file**.
3. **Due date/time: Thursday 3-Nov-2022, 2:00 pm** (Report and Code submission).
4. The late penalty for the assignment is 5% of the maximum mark per day, starting after 2pm on the due date. The closing date **Sunday 13-Nov-2022, 2:00 pm** is the last date on which an assessment will be accepted for marking.
5. **Weight: 25%** of the total mark of the unit.
6. The full marks of this group assignment are **65 marks**. **In addition**, the **maximum bonus marks** based on the class forecasting competition are **5 marks**.
7. Groups: you should complete this group project in a group of **four students**. You must follow the allocated group on Canvas-People page.
8. Presentation: please refer to the Presentation Instructions section of this file for more detailed instructions, including the length requirement of the report, font size, etc. To facilitate your report writing process, a **Report_Instructions.pdf** file is also provided on Canvas.
9. Numbers with decimals should be reported to the **four-decimal point**.
10. Marking Criteria: please refer to the Marking Criteria section of this file for more detailed instructions.
11. Please include the **name and student ID of all group members and group ID** in the submitted report and code file. The names of your report and code should follow the following formats respectively, by replacing "123" with your group ID. Example: **Group_123_Report, Group_123_Code**.

Key rules

- Carefully read requirements of the assignment.
- Please follow any further instructions announced on Canvas.
- You must use **Python** for the assignment.
- If the training of your model involves generating random numbers, your Python code **random seed must be fixed**, by using **np.random.seed(0)**.
- Reproducibility is fundamental in data analysis, so that you will be required to submit a code file that generates your results. Not submitting your code will lead to **a loss of 50%** of the assignment marks.
- Failure to read information and follow instructions may lead to a loss of marks. Furthermore, note that it is your responsibility to be informed of the University of Sydney and Business School rules and guidelines, and follow them.
- Referencing: Harvard Referencing System. (You may find the details at: http://libguides.library.usyd.edu.au/c.php?g=508212&p=3476130).

Background

The **underemployment** rate is the number of **underemployed** people expressed as a proportion of the labour force. The underemployment refers to the condition in which people in a labor force are employed at less than full-time or regular jobs or at jobs inadequate with respect to their training or economic needs. The underemployment rate is reported by the relevant government department in most countries. The underemployment rate can be used as an important indicator by the central bank of the country to determine the health of the economy when setting monetary policy.

Tasks and Datasets

For this group project, we have obtained the **monthly** historical underemployment rate data in a country from **June 1978 to December 2017**, as in dataset **UnderemploymentRate_InSample.csv**, which can be downloaded from the Canvas. The dataset contains information of Date (1/month/year, so monthly data) and Underemployment Rate.

Your task is to develop a predictive model, trained with **UnderemploymentRate_InSample.csv**, to forecast the monthly underemployment rate from **January 2018 to December 2019**. Note this is a **24-step-ahead forecast** task.

An out-of-sample test dataset which contains the true 2018 and 2019 underemployment rates, named **UnderemploymentRate_OutofSample.csv** in the same format as the in-sample data, is provided on Canvas. They will be used to assess the forecast accuracy of your produced models. Since you should assume the out-of-sample data is completely hidden from your model training/selection process, you **must NOT use the out-of-sample test dataset in your model training/selection process**. Otherwise, your model training process will be treated as having critical issues and you will receive heave penalty on the methodology and forecasting results, no matter how good your forecasting results are.

In other words, the **out-of-sample test dataset should be only used to evaluate your forecast accuracy (details to be shown later).**

Please note the assignment tasks are designed to be open-ended questions. This gives more freedom for you to explore a good solution and is similar to the situations that you might encounter in the real world.

You need to prepare a **report** for this assignment. The purpose of the report is to describe, explain, and justify your solutions with polished presentation. Be concise and objective. Find ways to say more with less.

You **MUST** submit your **Python code** which can be used to **replicate** the results in your report. Please note even if you fix your Python code random seed by using np.random.seed(0), changing the computer/CPU could have impact on random number generation and produce slightly different results. Please note the key target of having replicable results is to make sure that **every group has genuine results reported**. Therefore, if you have slightly different results for different runs/computers, it is ok. As long as the marker can re-run your code and have results that are very close to yours, then it is fine.

Suggested Report Outline:

1. **(2 marks)** At the **beginning (the first line)** of your report, you should report your best out-of-sample forecasting result, by stating: **"The best out-of-sample forecasting Root Mean Squared Error of our group is: ……".** Please note the markers will **run your code** and check whether your reported results can be produced/replicated. **Reporting false results deliberately can result in an up to 30% penalty of the assignment marks**.
2. **(5 marks)** Introduction. Write a few paragraphs stating the business problem and summarising your works, etc. Use plain English and avoid technical language as much as possible in this section (it should be for the general audience).
3. **(10 marks)** Data pre-processing and exploratory data analysis (EDA). Write python program to clean the data, e.g., checking/deleting incomplete information if any, making sure data is complete, or transforming the data if needed, etc. It is up to you on whether/how to transform the data so that the resulting dataset can be well incorporated in training your chosen models.

   Conduct initial analysis of the time series by plotting them or do what you can to reveal any patterns. Summarise what you have revealed or observed. In your report, carefully present your EDA procedure and findings, and discuss how the EDA results inform you on the methodology section.
4. **(40 marks)** Methodology and forecasting results. In your report, you should present the details of **three different models**. The three models should be different **types** of models. For example, ARIMA(1,1,0) and ARIMA(2,0,1) are the same type of models. ARIMA and Seasonal ARIMA models will be counted as different types of models. Simple Exponential Smoothing, Trend Corrected Exponential Smoothing, and Seasonal Holt-Winters Smoothing models will be counted as different types of models. Additive and Multiplicative Seasonal Holt-Winters Smoothing models will be counted as different types of models. Neural Networks Autoregression and Recurrent Neural Networks models will be counted as different types of models.

   The details of the methodology/model should include: your rationale, how you train your models, model selection process, some interpretations, your findings and justifications of your choices. You can try models that are not covered in our unit. However, for the three models presented, **at least two models should be the models that we have covered in the lecture**. The types of models could be the Moving Average, Decomposition method, Exponential Smoothing, ARIMA, Neural Networks Autoregression Model, Recurrent Neural Networks (RNN), Forecasting Simple Average, Forecasting Combination, etc. This is your choice. Below list contains some further clarifications.

   - As mentioned above "In your report, you should present the details of three different models.", while in the assignment working process in general you **should try more than three models**. This is because you need to provide **rational and justifications** on your choice, i.e., why do you initially choose to test these 5 or 10 models (rational)? Why do you finally decide to presentation these 3 models (justifications of your choices)? If I were to work on the assignment, I would try 5 or 10 or even more models and use the model

selection technique (train/validation split) and/or out-of-sample forecasting results to decide my final three models.

- Then in your report, you can present the details of the final three models and explain your whole assignment working process. Maybe you could also briefly include the working process and test RMSE values of other models that you have tried. By following this strategy, you provide **strong rational and justifications** on your final choice.
- Rational here means why this model is initially used/chosen. For example, suppose you have discovered some seasonality in the data with the EDA, then the rational here means you wanted to try some models that can consider seasonality, i.e., rational means you have a decision in accordance with reason or logic. Then in the rational part, you could mention some theoretical definition with formula of this seasonal model, i.e., how seasonality is modelled in the framework. You could also provide reason/logic on why you think this model could be a good candidate. Later, with the model training/selection and evaluation process, you can have further justifications on your choice.
- If your selected model does not require a model selection process, **clear justifications** on why this model is selected should be well documented. For example, based on the EDA, you can argue that additive HW exponential smoothing model is suitable for modelling the time series data. Since additive HW exponential smoothing model has fixed model complexity, then you do not need to have the model selection process with train and validation split. However, if you choose additive HW exponential smoothing model and decided to do a train and validation split to evaluate its forecasting performance before the final out-of-sample test, this is also fine.
- If the selected model requires a model selection process, such as ARIMA or NN models or your other selected models, a formal model selection process **must** be implemented and well documented.
    - For example, if your selected model is ARIMA or NN models, then you must have a model selection process with train and validation split.
    - This means you need to select one ARIMA model from many potential ARIMA models with different lags and seasonal orders, via using the **optimal validation data performance**. Do the same to select one NN model with different number of hidden layers and hidden neurons, and so on so forth.
    - With the selected 1 ARIMA model specification/complexity etc, **re-train** the selected model with the whole in-sample data and report its final out-of-sample forecasting RMSE.
    - Always remember: "Since you should assume the out-of-sample data is completely hidden from your model training/selection process, you **must NOT use the out-of-sample test dataset in your model training/selection process**".

Then you report the **out-of-sample forecasting Root Mean Squared Error (RMSE) results of your three presented models. In particular, your best model's out-of-sample RMSE forecasting result should be presented at the beginning of the report**, as mentioned in above point (1). This best model's result will decide the

forecast competition bonus marks for your group, refer to the **Marking Criteria** section later for more details.

Calculation of the out-of-sample forecasting results. You need to use your trained models to predict the 24 underemployment rates of 2018 and 2019. Please note that this is a 24–step-ahead forecast, since we assume you are in December 2017 (time stamp T) and have no knowledge about 2018 and 2019. Therefore, as mentioned you should assume the out-of-sample data is completely hidden from your model training/selection process, and you **must NOT use the out-of-sample test dataset in your model training/selection process**.

The 24 predicted values of the underemployment rate should be used to calculate the out-of-sample forecasting error. More specially, you need to use the Root Mean Squared Error (RMSE) to evaluate the forecast accuracy. The RMSE, computed on the out-of-sample data, is defined as follows. Let $\hat{Y}_{T+h|1:T}$ be the h-step-ahead point forecast, based on the in-sample data $Y_{1:T} = \{Y_1, Y_2, Y_3, \dots, Y_{T-1}, Y_T\}$. The true h-th underemployment rate value Y_{T+h} is included in the out-of-sample data **UnderemploymentRate_OutofSample.csv**. The out-of-sample RMSE is computed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{h=1}^{24}\left(Y_{T+h} - \hat{Y}_{T+h|1:T}\right)^2}{24}},$$

here 24 is the number of observations in the out-of-sample data.

5. **(3 marks)** Final analysis, conclusion, limitations and future steps (non-technical).
6. Appendix. In the appendix section, you **MUST** include **three meeting minutes** using the provided **Minutes Template** on Canvas. More detailed instructions are also given below. You can also put any other materials that you see appropriate into the Appendix section. The Appendix will NOT be counted into the length of the main report and there is no page limit for the Appendix.

Meeting Minutes

- Your group is required to submit three meeting minutes which are to be attached to the report as the Appendix. Your group should use the **Minutes Template** on Canvas for preparing agendas and meetings minutes.
- You should document at least three meeting minutes for this group assignment, using the template provided. Each minute should at least record the following information:
    - Meeting dates/time/duration;
    - Key points of the process of discussion, such as who said/did what;
    - Action list, **responsible member(s)**, task due time, etc. It is crucial that you clearly document the actions and works for each member during each meeting;
    - Review/group judgement on the quality of individually;completed/responsible tasks. The purpose of this is to infer

whether a member is doing his/her share of jobs;
    - The minute template contains some example input.

In case of a problem raised within a group, we will request minutes of all group meetings. We will make an individual adjustment to the group mark, if there is sufficient evidence shows that a student has done significantly less works than other members. **If a student has truly done very little works, a mark of 0 will be awarded for the student.**

Marking Criteria

The full marks of this group project are 65 marks, including 60 marks for the report and 5 marks for the presentation. In addition, the **maximum bonus marks** based on the class forecast competition are **5 marks.** The details are shown below:

- The content in your report **Group_123_Report** is worth **60 marks** (with suggested report structure and mark break down as above in the **Suggest Report Outline** section):
    - Focus on the appropriateness of the chosen forecasting methods and provide **full explanation and interpretation** of any results you obtain in your report. Output without explanation will receive 0 marks.
    - Describe your data analysis procedure in detail: how the data pre-processing is completed, how the EDA is done, what and why these models are used, how the models are trained, the model selection process, some interpretations, your findings and justifications of your choices. The descriptions should be detailed enough so that other data scientists, who are supposed to have background in your field, understand and are able to implement your works.
    - Clearly and appropriately present any relevant graphs and tables.
    - You may insert **small section** of your code into the report for better interpretation when necessary.

- The Python implementation. The main program file should be named as Group123_code.ipynb (or Group123_code.py). Your program must be runnable and your out-of-sample forecasting RMSE results must be replicable. **Reporting false results deliberately can result in an up to 30% penalty of the assignment marks.**

  The idea is that, when the marker runs your Group_123_Code.ipynb (or Group_123_Code.py), with the in-sample train data **UnderemploymentRate_InSample.csv** and out-of-sample test data **UnderemploymentRate_OutofSample.csv** in the same folder as the Python file, the marker expects to see the same (or at least very close) out-of-sample RMSE value as you reported. The code file should contain sufficient explanations so that the marker knows how to run your code.

- Presentation is part of the assessment. The marker will assign **5 marks** for presentation. The detailed instructions are shown in the following Presentation Instructions section.

- We will allocate **a maximum of 5 bonus marks** for the forecast competition among the groups. Groups will receive marks according to the rank of your **best out-of-**

sample forecast RMSE value (the value that you reported at the beginning (the first line) of your report; the smaller the better), according to the following rules:

- If the out-of-sample forecast RMSE of your forecast is within top 5 percent in the class, then the full **5 bonus marks** (for each student in the group) will be awarded;
- If the out-of-sample forecast RMSE of your forecast is between 5.1 percent (using one decimal rounding) and 20 percent in the class, then **3 bonus marks** (for each student in the group) will be awarded;
- If the out-of-sample forecast RMSE of your forecast is between 20.1 percent (using one decimal rounding) and 40 percent in the class, then **1 bonus mark** (for each student in the group) will be awarded;
- Otherwise, 0 bonus marks will be awarded.

Presentation Instructions

- Your report should be provided as a word or pdf document.
- Each group should submit one report and one code file.
- To facilitate your report writing process, a **Report_Instructions.pdf** file is provided.
- The report should be **NOT more than 15 pages (excluding Appendix and Reference list)**, with font size **not smaller than 11pt**. The page limit applies to all the content in your report, such as text, figures, tables, small sections of inserted codes, etc, but excluding the Appendix and Reference list. A violation of this rule will incur penalty on the presentation marks.
- You do NOT need to include the cover page and table of content.
- Numbers with decimals should be reported to the **four-decimal point**.
- You report should:
  - Include sections as suggested in **Suggested Report Outline** section.
  - Include all the methodology details and steps as mentioned above.
  - Demonstrate an understanding of the relevant principles of predictive analytics approaches used.
  - Clearly and appropriately present any relevant figures and tables.
- Your group is required to submit **three meetings minutes**. Your group should use the **Minutes Template** provided on Canvas to prepare agendas and meetings minutes. Not providing the meeting minutes will incur penalty on the presentation marks.
- Later, the unit coordinator will collect **peer feedback** on the performance of each group member. Therefore, it is crucial that each group member is contributing genuinely to the group assignment.