# QBUS6840 Group Assignment (30 marks)

October 2, 2019

## 1 Background and Task

Gross domestic product (GDP) is defined as "an aggregate measure of production equal to the sum of the gross values added of all resident and institutional units engaged in production and services (plus any taxes, and minus any subsidies, on products not included in the value of their outputs)"[1]. Another common definition of GDP is that "GDP measures the monetary value of final goods and services—that are bought by the final user—produced in a country in a given period of time (say a quarter or a year)"[2]. As GDP is one of the most important indicators used to analyse the health of a nation's economy, building statistical models to forecast future GDP growth rate (quarterly or yearly) is crucial for government officials and business managers to determine fiscal and monetary policies and plan future operating activities [3].

In this group project, your task is to develop a predictive model to forecast the future GDP of a country given its historical quarterly GPD information. The GDP data set **GDP_training.csv** (in the unit of billion US dollars) contains the quarterly GDP values measured from 31/3/1959 to 31/12/2015. This data set is based on a real GDP data set with some added noise for the de-identification purposes.

The test data set **GDP_test.csv** (not provided) has the same structure as the training data, and contains the quarterly GDP values from 31/3/2016 to 31/12/2018.

Your task is to develop a predictive model, using **GDP_training.csv**, to forecast the quarterly GDP of the country from 31/3/2016 to 31/12/2018. Note that, this is a multiple–step-ahead forecast problem.

### 1.1 Test error

For the measure of forecast accuracy, please use mean squared error (MSE). The MSE, computed on the test data, is defined as follows. Let $\widehat{y}_{T+h|1:T}$ be the $h$-step-ahead forecast of $y_{T+h}$, based on the training data $y_{1:T}$, where $y_{T+h}$ is the $h$-th GDP value in the test data **GDP_test.csv**. The test error is computed as follows

$$\text{test\_error} = \frac{1}{12}\sum_{h=1}^{12}(\widehat{y}_{T+h|1:T} - y_{T+h})^2,$$

where 12 is the number of observations in the test data.

[1] https://stats.oecd.org/glossary/detail.asp?ID=1163

[2] https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm

[3] https://www.investopedia.com/terms/e/economic-forecasting.asp

## 2 Submission Instructions

1. Each group needs to submit TWO files (or more if necessary) via the link in the Canvas site.
   - A document file, named **Group_xxx_document.pdf**, that reports your data analysis procedure and results. You should replace the xxx in the file name with your group ID.
   - A Python file, named **Group_xxx_implementation.py**, that implements your data analysis procedure and produces the test error. You might submit additional files that are needed for your implementation, the names of these files must follow the same format **Group_xxx_<name>.py**. You should replace the xxx in the python file name with your group ID.
2. About your document file **Group_xxx_document.pdf**
   - Describe your data analysis procedure in detail: how the Exploratory Data Analysis (EDA) step is done, what and why models/methods are used, how the models are trained, etc. with sufficient justifications. The description should be detailed enough so that other data scientists, who are supposed to have background in your field, understand and are able to implement the task. All the numerical results are reported up to four decimal places.
   - Clearly and appropriately present any relevant graphs and tables.
   - The page limit is 25 pages including EVERYTHING: appendix, computer output, graphs, tables, etc.
   - You must use the cover sheet provided on Canvas.
3. The Python file is written using Spyder or Jupyter Notebook as the editors, with the assumption that all the necessary data files (**GDP_training.csv** and **GDP_test.csv**) are in **the same folder** as the Python file. If you use deep learning models, then please assume that **Keras (with Tensorflow backend)** has been installed.
   - If the training of your model involves generating random numbers, the random seed in **Group_xxx_implementation.py** must be fixed, e.g. `np.random.seed(0)`, so that the marker expects to have the same results as you had.
   - The Python file **Group_xxx_implementation.py** must include the following code

     ```
     import pandas as pd

     GDP_test = pd.read_csv('GDP_test.csv')

     # YOUR CODE HERE: code that produces the test error test_error

     print(test_error)
     ```

     The idea is that, when the marker runs **Group_xxx_implementation.py**, with the test data **GDP_test.csv** in **the same folder** as the Python file, he/she expects to see the same test error as you would if you were provided with the test data. The file should contain sufficient explanations so that the marker knows how to run your code.
   - In case you want to test your code to see if a test error is produced, a "fake" test data is provided. This data set has the same format as the real test data **GDP_test.csv**, except that the GDP values in there are not the actual values.

4. Your group is required to submit meetings minutes. You may use the templates provided for preparing agendas and meetings minutes. In case of a dispute within a group, I will use the meeting minutes and/or request for more information to make adjustment to the individual marks. If there is sufficient evidence that a group member has contributed nothing, this student will get a mark of zero.
5. Each member of the group is also required to submit a peer assessment. Please use the peer criteria sheet `Peer Assessment Criteria` and assessment form `Peer Assessment of Team Members`, provided on Canvas, for this purpose.

## 3 Marking Criteria

This assignment weighs 30 marks in total. The content in **Group_xxx_document.pdf** contributes 15 marks, and the Python implementation contributes 15 marks, but the priority is given to the forecast accuracy (see below). The marking is structured as follows.

1. The accuracy of your forecast: Your test error will be compared against a benchmark test error obtained by the teaching team (sure, you can beat us!). The marker first runs **Group_xxx_implementation.py**
   - Given that this file runs smoothly and a test error is produced, the 15 marks will be allocated proportionally to your prediction accuracy. **If your test error is as small as the benchmark error, you're highly likely to be awarded the total 30 marks (the document file is then no longer needed).**
   - If the marker cannot get **Group_xxx_implementation.py** run or a test error isn't produced, some partial marks (maximum 5) will be allocated based on the appropriateness of **Group_xxx_implementation.py**.
2. Your report described in **Group_xxx_document.pdf**: The maximum 15 marks are allocated based on
   - the appropriateness of the chosen forecasting method.
   - the details, discussion and explanation of your data analysis procedure.

## 4 Award

The group with the best forecast accuracy will be announced (unless they do not want to be named) and receive a prize! Their test error and the test error obtained by the teaching team, the model chosen, etc. will also be made available.

## 5 Errors

If you believe there are errors with this assignment please email the coordinator immediately at minh-ngoc.tran@sydney.edu.au.