CARNEGIE MELLON UNIVERSITY
18- BIG DATA SCIENCE
ASSIGNMENT 02

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb].** In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained
  - Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This new process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID-BDS-AssignmentNo. For example, mcsharry-BDS-Assignment1, mcsharry-BDS-Assignment2 and mcsharry-BDS-Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for usual late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is **Rwandan Time (CAT) on Monday 18, April, 2022 / 23:59**.

Questions

This assignment will assess the ability of quantitative models to forecast excess returns on a monthly basis.

Welch and Goyal (2008) find that numerous economic variables with in-sample predictive ability for the equity premium fail to deliver consistent out-of-sample forecasting gains relative to the historical average. They argue that model uncertainty and instability seriously impair the forecasting ability of individual predictive regression models. A copy of their paper ["A Comprehensive Look at The Empirical Performance of Equity Premium Prediction" July 2008, Review of Financial Studies 21(4) 1455-1508] is available here: http://www.hec.unil.ch/agoyal/docs/Predictability_RFS.pdf

Your challenge is to:

1. Load in the Welch and Goyal (2008) dataset (http://www.hec.unil.ch/agoyal/) into Matlab, R or Python

2. Recreate explanatory variables (using the Original data (up to 2005)) and plot to confirm that these have been correctly interpreted. Note that Table 1 gives a list of the explanatory variables. Provide a time series plot for each of the variables with labels. (A single figure with subplots)

3. For the time periods listed below, using the rolling multiple linear regression model as described in Welch and Goyal (2008) and using the updated data (up to 2014), provide a table showing the out-of-sample R^2, RMSE and MAE values.
    - January 1965 to December 2008;
    - January 1976 to December 2008; and
    - January 2000 to December 2008

4. Using the same time frames and providing R^2, RMSE and MAE values (to be compared with results above) for your models, improve on their results using any two of the following ideas:

    (a) Economically motivated model restrictions;
    (b) Forecast combinations;
    (c) Regime switching;
    (d) Machine learning (e.g. Lasso, CART, KNN, RNN or SVM)
    (e) Bagging and boosting.

In your findings, it is important to show that you understand the process and steps required to address these challenges and that you are capable of implementing this in a language of your choice.

Extra task: Installing ArcGIS software.

In Assignment 3, you will work with ArcGIS and it is good to have it installed on your computer. In order to make sure that everything is well set before time, as it may take time to get the access, you are encouraged to request the right to download and install the ArcGIS software that will be used.

Kindly go to this link and follow all the instructions as provided about getting this software: https://www.heinz.cmu.edu/current-students/computing-services/software.
Once you are given the right you will be able to download the software. You can download the ArcGIS 2.6 version.