QBUS6840
Predictive Analytics

Semester 2, 2019

Homework (15%)

1 Rationale

This assignment is designed to help students to develop basic predictive analytics skills on synthetic and possible real applied problems. The skills include data visualization, model building and analysis in terms of understanding the theory, practicing with raw data and programming in Python.

If you spot any typos or mistakes in this assignment, please report immediately to the coordinator *minh-ngoc.tran@sydney.edu.au*

2 Questions

The data set `SP500_weekly_Jan1988_Nov2018.csv` (downloaded from https://au.finance.yahoo.com), available on Canvas, contains the Standard & Poor weekly adjusted closing stock indexes from January 1988 to November 2018. Denote this time series as $\{y_t, t = 1, ..., T = 1613\}$.

(a) Write Python script to load the data and produce their time series plot. Include the plot and your comments (if any) together with the Python script in your submission.

(b) Write Python script to produce *one-step-ahead* forecasts for the last 100 observations using the naive forecasting method with drift, i.e. compute

$$\widehat{y}_{t|t-1}, \quad t = 1514, ...1613.$$

- Plot these forecasts together with the actual values. Include the plot and the Python script in your submission.
- Report the scale-dependent measure Root Mean Squared Error (RMSE) and scale-independent measure Mean Absolute Percentage Error (MAPE) (the errors between forecasts and the ground true indexes].

(c) Smooth the time series using centered MA-4. Plot the smoothed time series together with the original time series. Include the plot and the Python script in your submission.

(d) Given the stock indexes $\{y_t, t = 1, ...T\}$, the stock returns are defined as

$$r_t = \log \frac{y_{t+1}}{y_t}, t = 1, ...T - 1.$$

Write Python script to compute the stock returns and produce their time series plot. Comment on this plot in conjunction with the plot of the indexes $\{y_t, t = 1, ..., T\}$. Which dataset do you think is more predictable, and why? Include the plot and the Python script in your submission.

(e) For the index dataset $\{y_t, t = 1, ...T\}$

- Use the last 100 observations as testing data, and the previous observations for the training data. Use the training dataset to estimate the parameters (weight α and initial level l_0. You may set $l_0 = y_1$ or l_0 be the average of a few first observations) of the Simple Exponential Smoothing (SES) method.
- Based on these estimates of α and l_0, compute one-ahead-forecasts on the test data, i.e.,

$$\widehat{y}_{t|t-1}, \quad t = 1514, ...1613.$$

Compute the Mean Absolute Percentage Error (MAPE) and plot the forecasts. Please also include your Python code in submission.

(f) For the squared returns dataset $\{x_t = r_t^2, t = 1, ...T - 1\}$

- Use the last 100 observations as testing data, and the previous observations for the training data. Use the training dataset to estimate the parameters (weight α and initial level l_0. You may set $l_0 = x_1$ or the average of a few first observations) of the SES method.
- Based on these estimates of α and l_0, compute one-ahead-forecasts on the test data, i.e.,

$$\widehat{x}_{t|t-1}, \quad t = 1513, ...1612.$$

Compute the Mean Absolute Percentage Error (MAPE) and plot the forecasts. Please also include your Python code in submission.

(g) Compare the MAPE obtained for the index dataset $\{y_t, t = 1, ...\}$ and the MAPE for the squared returns dataset $\{x_t = r_t^2, t = 1, ...\}$. Give your comments

(h) (Optional) Repeat parts (e) and (f) with the Trend Corrected Exponential Smoothing (TCES) method. Do you obtain better forecasts (compared to the SES method)?

3 Instructions and Marking criteria

For parts (e) and (f), the more accurate forecasts (i.e., smaller MAPE) you have, the more marks you're given. Part (h) is optional, bonus marks might be awarded if you attempt this part.

Assignment Report: The assignment report should be presented as a technical report that:

- details ALL required steps.
- provides sufficient explanation and interpretation of any results you obtain. Output without reasonable justifications will not receive full marks.
- Clearly and appropriately presents any relevant tables, graphs and screen dumps from programs if any. You may insert small sections of your code into the report for better interpretation when necessary. Find the best and most structured way to present your work, summarise the implementation procedures, support your results/findings and prove the originality of your work.
- reports numbers with decimals to the **three-decimal point**.
- properly cites all the references if any.

Assessment of your written presentation skill is part of this assignment. Markers will allocate **up to 10%** of the mark for presentation.

Important notes:

- **Required submissions**: **ONE** written report (MS word or pdf format) and **ONE** Python source code file (Jupyter Notebook .ipynb or .py). Please follow instructions for submissions announced on Canvas.
- The late penalty for the assignment is 5% of the assigned mark per day, starting after 4:00pm on the due date. The closing date is the last date on which an assessment will be accepted for marking. See Canvas for the due date and closing date of this assignment.
- As anonymous marking policy, only include your Students ID in the report and do **NOT include your name**.
- The name of the report and code file must follow the format

  <your SID>_QBUS6840_Assignment1_2019S2.
- The report should be **NOT more than 15 pages** including everything like text, figure, tables and small sections of inserted codes, etc, but excluding the appendix.

- The University of Sydney takes plagiarism very seriously. Please be warned that plagiarism between individuals is always obvious to the markers and can be easily detected by Turnitin.

Key rules:

- Carefully read the requirements for each part of the assignment.
- Please follow any further instructions announced on Canvas, particularly for submissions.
- You must use **Python** for the assignment. To avoid any potential issues with your codes, please use the **latest Anaconda version** for your Python programs.
- Reproducibility is fundamental in data analysis, so that you are required to submit a code file that generates your results. Not submitting your code or submitting code that are not runnable will lead to a loss of 50% of the assignment marks.
- Referencing: Harvard Referencing System.
  (Please find the details at: http://libguides.library.usyd.edu.au/c.php?g=508212&p=3476130)