**SDSC2005-22B**

**Exercise 4. Time Series Analysis for Predictive Research**

**(V3, Feb 28, 2023)**

Your EID (required): fzehra3

**Data1**: Ex4\_data1.xlsx, which contains the following sheets:

* Intraday\_data: the price of HSI per minute from Jan 16, 2023 to Feb 13, 2023 (in total of 18 trading days)
* Interday\_data: the price of HSI per day from Jan 2, 1987 to Feb 10, 2023 (in total of 9,167 trading days of 36 years)
* Use linear interpolation to fill any empty cell.

**Task**:

Identify the following intraday cycles of HSI Price to answer the question that when is the *best* time of the day to buy (i.e., at the lowest price) and sell (the highest price) stocks on average in HK stock market:

1. **Intraday time**:
   1. Detrend (removing the overall trend of) Price throughout the entire period of the Intraday Data;
   2. Use each 30 minutes of trading time as a half-hourly unit within each trading day;[[1]](#footnote-1)
   3. Use 1-hot encoding to create the half-hourly variables;
   4. Use an OLS regression with Price as the DV and hourly variables as the IVs to measure the effect of “half-hour of the day” on Price;
   5. Optional: use alternative way(s) to measure the best time to buy/sell based on OLS regression.[[2]](#footnote-2)
2. **Interday time**:
   1. Detrend Price throughout the entire period of the Interday Data;
   2. Use each trading day as a daily unit within each trading week and each month as a monthly unit within each month, respectively;
   3. Use 1-hot encoding to create the daily and monthly variable(s), respectively;
   4. Use an OLS regression with Price as the DV and daily and monthly variables as the IVs to measure the effect of “day of the week” and “month of the year” on Price, respectively;
   5. Optional: use alternative way(s) to create best day/month to buy/sell on the OLS regression report the resulting effect if significantly greater than the 1-hot encoding approach (grading policy: extra point(s) for significantly improved results, depending on the size of the improvement; no penalty for wrong answers).

**Report**:

1. Quantitative findings in Table 1.
2. A summary paragraph to interpret what investors may learn from the results, if any, for their trading strategies.

Table 1. Results of OLS Regressions

|  |  |  |  |
| --- | --- | --- | --- |
|  | Intraday Effect (Half-hour of the day) | Interday Effect  (Day of the week) | Interday Effect (Month of the week) |
| *Required*: | | | |
| Best time to buy | Which half-hour? | Which day? | Which month? |
| Best time to sell | Which half-hour? | Which day? | Which month? |
| Ratio of sell-to-buy price[[3]](#footnote-3) | the s2b ratio | the s2b ratio | the s2b ratio |
| Model R-squared |  |  | |
| *Optional*: | | | |
| Best time to buy | Which half-hour? | Which day? | Which month? |
| Best time to sell | Which half-hour? | Which day? | Which month? |
| Ratio of sell-to-buy price | the s2b ratio | the s2b ratio | the s2b ratio |
| Model R-squared |  |  | |

**Data2**: Ex4\_data2.xlsx, containing the following sheets (using “Adj Close” in column F as Price for all questions below):

* 0005.hk: the price and volume of HSBC (bank)
* 0027.hk: the price and volume of Galaxy Entertainment (casino)
* 0101.hk: the price and volume of Hang Lung Properties
* HSI: the price and volume of Hang Seng Index (Hong Kong)
* DJI: the price and volume of Dow Jones Index (U.S.)
* SSEC: the price and volume of Shanghai Stock Exchange Composite (China)
* Use linear interpolation to fill any empty cell.

**Task**:

1. **ARIMA parameters**:
   1. Data: use all dates up to Dec 31, 2022 for the three stocks (HSBC, Galaxy, and Hang Lung), respectively;
   2. Use ACF (autocorrelation function) and PACF (partial autocorrelation function) toidentify the autoregression (AU), integration (I), and moving average (MA) parameters for each stock price;
   3. Fit a univariate ARIMA model (i.e., only Price plus AR, I, and MA, without any IV) for each stock
   4. Report the results in Table 2.
2. **Predictive models**:
3. Data: split the data to a training set (up to Dec 31, 2022) and a test set (from Jan 1 to Feb 21, 2023) for each stock;
4. Model: build a predictive model for each stocks, respectively, using Price as the DV and any of the following as the IVs:
   1. Time-effects: day of the week, month of the year (“seasonality”), and any other features that represent repeated cycles of time (see questions for Data 1);
   2. Internal factors: the previous price and volume of the stock (no need for previous price if you use ARIMA/SARIMA because it will be automatically included);
   3. Market influences: the previous price of the stock market in Hong Kong (HSI), the U.S. (DJI), and mainland China (SSEC);
   4. Optional IVs: any other time series data measured on a daily unit to be collected by you and add to the model as the IVs (same grading policy as in Data 1 applies here);
5. Estimation (based on the training set) and test (based on the test set) method: use any method of your choice, including an ensemble of several methods, e.g.,
   1. OLS;
   2. Exponential smoothing;
   3. ARIMA/SARIMA;
   4. Machine learning/deep learning;
   5. Anything else;
6. Reportyour model specification and resultsin Table 3**.**

1. **Forecast future values**: use your predictive model to forecast the price of each stock on March 13, 15, and 17. Report the results in Table 4 and Figure 1.

**Report**: Present your results in the following tables:

Table 2. ARIMA Parameters of Individual Stock Price

|  |  |  |  |
| --- | --- | --- | --- |
|  | HSBC (005) | Galaxy (027) | Hang Lung (101) |
| Autoregression (AR) | | | |
| * Order (e.g., 0, 1, etc.) | 0 | 2 | 1 |
| * Coefficient | NA | 1.5774, -0.9525 | 0.9087 |
| Integration (I) | | | |
| * Order | 1 | 1 | 1 |
| Moving Average (MA) | | | |
| * Order | 0 | 3 | 1 |
| * Coefficient | NA | -1.5491, 0.8904, 0.0291 | -0.9272 |
| Model fit (AIC) | 12842.029 | 13515.114 | 1855.550 |

Table 3. Predictive Models of Individual Stock Price

|  |  |  |  |
| --- | --- | --- | --- |
|  | HSBC (005) | Galaxy (027) | Hang Lung (101) |
| *a. Training Set:* | | | |
| * Model type | ARIMA | ARIMA | ARIMA |
| * Equation | Refer to image attached | Refer to image attached | Refer to image attached |
| * Accuracy (MAPE)[[4]](#footnote-4) | 1.0216792933530783 | 2.165622355765563 | 1.4986463615978578 |
| * Justification for using the model | ARIMA models are a type of statistical model used to analyze and forecast time series data. It explicitly caters to a set of standard time series data structures, and as such provides a simple yet powerful method for making reasonable time series forecasts. | ARIMA models are a type of statistical model used to analyze and forecast time series data. It explicitly caters to a set of standard time series data structures, and as such provides a simple yet powerful method for making reasonable time series forecasts. | ARIMA models are a type of statistical model used to analyze and forecast time series data. It explicitly caters to a set of standard time series data structures, and as such provides a simple yet powerful method for making reasonable time series forecasts. |
| *b. Test Set:* | | | |
| Accuracy (MAPE)3 | \*13.256046881707736 | 4.201882733827262 | 3.308485917402954 |

\*Note: Different orders of p and q were used for this dataset, yet MAPE only improved insignificantly. Thus, I chose the model which had the best parameters with the lowest AIC.

Text

Description automatically generated

Table 4. Forecasted Stock Price on March 13, 15, and 17

|  |  |  |  |
| --- | --- | --- | --- |
|  | HSBC (005) | Galaxy (027) | Hang Lung (101) |
| March 13 | 48.549999 | 51.37246554166069 | 15.065717896672437 |
| March 15 | 48.549999 | 51.312410237693996 | 15.065099646379004 |
| March 17 | 48.549999 | 51.359932541292835 | 15.064589134915384 |

Figure 1.

1. Use a scatterplot with Price in the *y*-axis, and date in the *x*-axis, including include an “observed period” (up to Feb 21, 2023) and a “forecast period” (March 13, 15, and 17);
2. Show two lines (the observed and estimated prices) in the observed period and three lines (forecasted price, and the confidence interval at the 95% confidence level) in the forecast period;
3. See slide 34 of Week 5 as two examples.

Figures in ipynb file.

Note: I plotted forecast steps for training set up till 17 March in separate graph because the ratio of the observed period and forecast period was distorting the visualized outcome, as a result of which we could not see the trend of forecast period properly.

**Optional Question** for both Data1 and Data2 (the above grading policy applies)

Note that you are required to detrend for Data1 but not required to do so for Data2. Discuss what the detrend (for Data1) and non-trend (for Data2) will do to the respective results? If you think either Data1 or Data2 should be done differently, why and what will happen to the results?

Detrending time series data means removing an underlying trend from the data. The main reason for doing so is to make seasonal or cyclical subtrends in the data more visible. Detrending was done in Data2 by the process of differentiation. All the datasets were non-stationary as per the AD fuller test. If it were not done, then data would not have been stationary, which otherwise helped stabilize the data and reduce trend and seasonality.

**Submission**:

1. Write your answer the above questions in this Word document and save it in Word format (doc or docx);
2. Attach your programming codes for both Data1 and Data2 in the original format (e.g., \*.py, \*.ipynb, etc.);
3. Put the Word file and the programming codes in a zip/rar file package (i.e., \*.zip or \*.rar) and upload it to the Assignment box.

1. For example, create a set of half-hourly variables to represent records, each representing the price during 9:30-9:59, 10:00-10:29, …, 11:30-12:05, 13:00-13:29, etc. respectively. Note that the interval during 11:30-12:05 is longer than 30 minutes. Theoretically, there should be no trading between 12:00 and 12:05, but these minutes are included in the data for unknown reason. Let’s consider these are make-up records for 11:30-11:59. [↑](#footnote-ref-1)
2. Grading policy for optional solutions: extra point(s) for significantly improved results, based on the size of model fitness (R-squared); no penalty for wrong answers. [↑](#footnote-ref-2)
3. Ratio of sell-to-buy price = price at best time to sell / price at best time to buy. [↑](#footnote-ref-3)
4. Mean Absolute Percentage Error (MAPE) = , where are actual and estimated Price at date *t* (= 1 to *n*). [↑](#footnote-ref-4)