FE520 Assignment 5

Dan Wang, Zhiyuan Yao

Fall 2020

Submission Requirement:

For all the problems in this assignment you need to design and use Python 3, output and present the results in nicely format.

Please submit a written report (pdf), where you detail your results and copy your code into an Appendix. You are required to submit a single python file and a brief report. Your grade will be evaluated by combination of report and code.

You are strongly encouraged to write comment for your code, because it is a convention to have your code documented all the time.

Python script must be a '.py' script, Jupyter notebook '.ipynb is not allowed.

Do NOT copy and paste from others, all homework will be firstly checked by plagiarism detection tool.

1 Time Series Data Practice(60 pts)

Recall what we mentioned in the class, we have two types of data splitting for training and testing data: out of sample and out of time. It is proper to use Out of Time splitting method for time series dataset. Writing a function to spilt "Energy" dataset into training and testing data.

Parameter input:

StartYear: int (default value = 2012), **EndYear**: int (default value = None).

Ontput:

Train, **Test** (Data type: Array(Numpy))

If EndYear is None, we will only choose all data with "Data Date" == **StartYear** as **Test** data, all other data as **Train** data. By default, all company Data Date within 2012 will be selected as Testing data. If EndYear is **NOT** None, we will choose all data with "Data Date" == **StartYear** to **EndYear** as **Test** data, all other data as **Train** data, For example, **StartYear** = 2010, **EndYear** = 2013, all data in 2010, 2011, 2012, 2013 will be selected as Testing data .

All return should be array from column "**Accumulated Other Comprehensive Income (Loss)**" to column "**Selling, General and Administrative Expenses**".

2 Momentum and Mean Reversion(40 pts)

Momentum and mean reversion are common trading strategy. For the purpose of this homework, a stock exhibits momentum is defined as an asset whose price returns are more likely to go up(down) on day t if the return went up(down) on day t-1. In other words, the stock exhibits a positive auto-correlation. Mean reversion is the opposite of momentum. Stocks are more likely to go up(down) on day t if that stock went down(up) on day t-1.

You are provided below with a simulated dataset of series of stock returns. These returns have been generated with a predetermined average momentum during one period and a predetermined average mean reversion in another period.

Please use dataset provided to answer the questions below. In order to do so, you will need to clean the dataset. It comes with a number of flaws commonly seen in dataset we receive.

Question:

1. In what month did the returns shift from exhibiting mean reversion to exhibiting momentum, or from momentum to mean reversion. Please output the last month that momentum(mean reversion) shift to mean reversion(momentum).

2. During the time period when these stock returns had momentum property, what was the average momentum? Please note this is a single number, average cross over all stock returns.

3. During the time period when these stock returns had mean reversion property, what was the average mean reversion?Please note this is a single number, average cross over all stock returns.

This is an interview question in industry, I have provided all information and hints in the question and in the zoom video. This question aims to practice your ability to clean the data using the technique we covered in the pandas lecture (especially in time series). Please try to think about how to solve this question before looking into the hints.

Hints: The dataset provided is multiple stock returns, but questions are asked for a single number of month that returns from one state to another state. Thus, what you need to do is find a way to aggregate multiple stock into on market index, and observe the index to find the answer.

If this is not enough, I have provided more detail Q & A in ***Discussion*** - **Question about Assignment 5**.

3 Regression (Bouns: 30pt)

1. In this question, we are going to use the diabetes data set. Use `sklearn.datasets.load_diabetes()` to load the data and labels.

2. Randomly split the data into training set (80%) and testing set (20%).

3. Create a linear regression model using sklearn, and fit training data. Evaluate your model using test data. Give all the coefficient and R-squared score.

4. Use 10-fold cross validation to fit and validate your linear regression models on the whole data set. Print the scores for each validation.

5. (Bonus 3pt) Use sklearn to create RandomForestRegressor model, and fit the training data into it.

6. (Bonus 7pt) Use Grid Search to find the optimal hyper-parameters (max_depth:{None, 7, 4} and min_samples_split: {2, 10, 20}) for RandomForestRegressor.