Correlation One: Machine Learning Challenge

Business Context

You are hired as a machine learning engineer at a quantitative hedge fund. The fund trades stocks in the S&P 500 Index.

Historically, the fund's research has relied market knowledge and intuition to discover *signals* that predict the returns of stocks. Now, the fund wants to use machine learning techniques to cull predictive signals from data.

The Prediction Problem

You are given a dataset that consists of the (simulated) daily open-to-close changes of a set of 10 stocks: S1, S2, ..., S10. Stock S1 trades in the U.S. as part of the S&P 500 Index, while stocks S2, S3, ..., S10 trade in Japan, as part of the Nikkei Index.

Your task is to build a model to forecast S1, as a function of S1, S2, ..., S10. You should build your model using the first 50 rows of the dataset. The remaining 50 rows of the dataset have values for S1 missing: you will use your model to fill these in.

The fund's researchers believe that some but not all of the lagged values of the other variables are important in predicting S1.

Answer Submission Guidelines

Your response should be emailed to submission@correlation-one.com. Please treat your responses as evidence of your work product for prospective employers

Your email should consist of four items, namely your

- Predictions for S1 on the test dataset. Predictions should be submitted in a .csv file with two columns. The first column is a date and the second column consists of predictions for S1 on that date. The data should be comma delimited and the file titled "predictions.csv", and should have a header row with column names "Date" and "Value".
- Code used to answer the question. We should be able to run the code; you can assume the dataset is placed in the directory "./data/". Please include the code that outputs your final predictions as well as any other code you used in the exploration and model building process.
- Resume in .pdf format
- Answers to the 4 questions below. **Please explain your analytical approach and how you arrived at an optimized result given the available data, including any relevant experiments or model tuning that didn't make it into your final chosen model. You answer write-up and**

code submission should read like work product submitted to a business unit head. Please submit your answers in .pdf format

(1) Which variables matter for predicting S1?
(2) Does S1 go up or down cumulatively (on an open-to-close basis) over this period?
(3) How much confidence do you have in your model? Why and when would it fail?
(4) What techniques did you use? Why?

Additional Technical Details

Your predictions will be compared to the actual values of S1, and the error will be evaluated using the sum of absolute deviations.

The daily returns for stocks are daily percentage changes of opening price to closing price.

For a given date (e.g. 2011-10-06), the return for S1 is measured from 9:30am EST to 4pm EST on 2011-10-06. The return for S2, S3, ..., S10, which trade in Asia, are measured from 8pm EST on 2011-10-05 to 2am EST on 2011-10-06.