Building the stock market prediction Engine

Saankhya, one of the Financial Analytics company wants to build a statistical algorithm for algorithmic trading and they hired you as the data scientist to help them.

Data Description

The data contains anonymized features relating to a single financial asset. The data contains 111 columns where the targets are y1 and y2.

First column is the timestamp which indicates the day at which the event occurred

"y1" (Numeric) is the percentage change (http://www.investopedia.com/ask/answers/03/100303.asp) in asset price of w.r.t. the previous day's asset price.

"y2" (Binary) is the volatility (http://www.investopedia.com/terms/v/volatility.asp) of the asset w.r.t. previous 2 weeks – 0 implies the asset is not volatile and 1 means it is.

You are expected to build at one classification model on "y2" and one regression model on "y1" using either of regression or time-series techniques.

Where do I see the data?

The datasets are attached with this note. There are two datasets. "train_data" which you would use for training your models.

"test_data" - the unseen dataset without target variables. You will predict on this test_data and submit your predictions in the format stated (Please refer Submission folder).

Submission Instructions:

1. ### Where do I submit?

   Submit as a single zip file to piazza under module CSE 7302c.

   Naming convention of zip file and subject line - TeamNo_CSE7302c_CUTe

2. ### What should I submit?

   You need to submit a zip file (with R code, prediction results and team list) and a presentation in the PDF format.

Deadline to submit: 17th December, Sunday 1:00 pm

3. What is the submission format?

Use the train_data to make prediction models for the two target variables y1 (continuous) and y2 (categorical). Submit your predictions for the test_data.
Submit your zip file that contains your prediction, team description and the R/HTML code.

4. What happens after submission?

Prepare for the 10 minute presentation and 5-10 minute Q&A.

5. What is the evaluation criteria?

We will evaluate on the following metrics.

S.No	Metric
1	Translation of business or technical case into Analytics or ML problem
2	Exploratory Data Analysis
3	Variety of models
4	Model performance and justification for the choice of the error metrics
5	Data Handling
6	Code correctness and Readability: Ability to code the model effectively
7	Code Readability: Functions, comments, etc.
8	Smooth flow of presentation and thought
9	Presentation Appearance: Font consistency, spell check, clean and informative slides
10	Presentation related Questions
11	Questions related to the CSE7302c module

6. What if you are absent to the Viva?

We will score your performance accordingly

7. What if you are not participating in the CUTe?

It is difficult to create such an exciting environment again. So, please ensure you have 100% participation. We will not re-conduct the exam.

Other miscellaneous questions:

1. There are missing values in the data-frame, what do I do?
   It is up to you how to handle them.

2. What is the viva format?
   Questions based on concepts from CSE 7302c module and your code

3. What should be the contents of the presentation?
   The allotted presentation time is 10 minutes. Ensure the presentation reflects all your work.

Basic guidelines for slides (but not limited to):

1. Problem statement(s)
2. Evaluation metrics and why,
3. *Exploratory Data Analysis (EDA) with basic plots*

(guidelines - univariate (single variable analysis),bivariate (analysis of two variables),multi-variate analysis with correlations, chi-sq tests of independence wherever relevant) - needn't cover all the variables/attributes given in dataset. Choose which you find interesting and report your findings.)

4. Model building (restrict yourself to only 7302 module concepts/algos)
5. Model evaluations for train data (with plots)
6. Predictions and error evaluations on test data you created from given data (with plots)
7. Your predictions for actual test data (plots, summary tables, anyway you want to show it).
8. Final model selection for each problem statement and why
9. Conclusions and Improvements

4. Will the deadline be extended?
   No. 17th December 2017 1:00pm is the submission deadline for both code/HTML as well as presentation. We will start the presentation/viva sharp at 2:00pm. The team-wise schedule for this would be published soon and would be randomly allocated.