Assignment

1. Create a time plot of the target variable and of Variable74OPEN using temporal aggregation. Explore the data for patterns, extreme and missing values.

   (a) One participant reported that differencing the predictor variables at lag 12 was useful. Compare boxplots and histograms of Variable74OPEN by target variable to the same plots of the differenced Variable74OPEN by target variable.

   (b) Find the three dates when there were days with no data. What are solutions for dealing with these missing values?

   (c) Examine carefully the data at 3:55pm daily. Some competitors noticed that this period always had larger gains/losses, suspecting that it represents the start/end of a trading day, and therefore more than 5 minutes. This is an example where a competition differs from real life forecasting: in real life we would know exactly when the trading day starts and ends. How can this information help improve forecasts for these periods?

2. Partition the data into training and validation, so that the last 2539 periods are in the validation period. How many minutes does the validation period contain?

3. What is the percent of periods in the training period that have a value of 1?

4. Report the classification matrix for using majority-class forecasts on the validation period.

5. Generate naive forecasts for the validation period, using the most recent value. Report the classification matrix for the naive forecasts.

6. One of the top performing competitors used logistic regression, initially with Variable74 variables (high, low, open, and close) as predictors. In particular, he used lagging and differencing operations. To follow his steps, create 12-differenced predictors based on the Variable74 variables, and lag the target variable by 13 periods. The model should include the original Variable74 predictors and the differenced versions (eight predictors in total). Report the estimated regression model.

7. Use the logistic regression model to forecast the validation period. Report the classification matrix. How does the logistic model perform compared to the two benchmark forecast approaches?

8. The winning criterion in this contest was the highest *Area Under the Curve*[3] (AUC) averaged across the results database. Recall that most forecasting methods for binary outcomes generate an event *probability*. The probability is converted to a binary forecast by applying a threshold. The AUC measure is computed from the classification matrix by considering all possible probability thresholds between 0 and 1. Consider the following classification matrix, where a, b, c, and d denote the counts in each cell:

	predicted events	**predicted non-events**
actual events	a	b
actual non-events	c	d

The AUC is computed as follows:

- Obtain the classification matrix for a particular probability threshold (recall that the default is a threshold of 0.5).
- Compute the two measures $sensitivity = \frac{a}{a+b}$ and $specificity = \frac{d}{c+d}$.

[3] See also the evaluation page on the contest website `www.kaggle.com/c/informs2010/Details/Evaluation`

- Repeat the last two steps for probability thresholds ranging from 0 to 1 in small steps (such as 0.01).
- Plot the pairs of *sensitivity* (on the y-axis) and *1-specificity* (x-axis) on a scatterplot, and connect the points. The result is a curve called an *ROC Curve*.
- The area under the ROC curve is called the *AUC*. Computing this area is typically done using an algorithm.

High AUC values indicate better performance, with 0.50 indicating random performance and 1 denoting perfect performance.

(a) Using the logistic regression model that you fitted in the last section, compute *sensitivity* and *1-specificity* on the validation period for the following thresholds: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. This can be easily done by modifying the probability threshold on the Excel LR_Output worksheet.

(b) Create a scatter plot of the 11 pairs and connect them. This is the ROC curve for your model.

9. While AUC is a popular performance measure in competitions, it has been criticized for not being practically useful and even being flawed. In particular, Rice (2010) points out that in practice, a single probability threshold is typically used, rather than a range of thresholds. Other issues relate to lack of external validity and low precision. He suggests[4]:

> "[I]nstead of picking a model winner in what could be a random AUC lottery, apparently more accurate measures - straight classification error rate and average squared error - with much better statistical and external validity should probably now be considered."

Compute the classification error rate $\left(\frac{b+c}{a+b+c+d}\right)$ for the logistic regression model, using the validation period.

[4] D. M. Rice. Is the AUC the best measure?, September 2010. available at www.riceanalytics.com/_wsn/page15.html

10. The same competitor, Christopher Hefele, then added more predictors to his model: Variable167 and Variable55 (each consisting of four series). His AUC was increased by 0.0005. Is this additional complexity warranted in practice? Fit a logistic regression with the additional predictors (taking appropriate differences), generate forecasts for the validation period, and compare the classification matrix and classification error rate to that of the simpler model (with Variable74 predictors only).

11. Use a neural network with the three sets of differenced and original variables (74, 167, and 55) as predictors. Generate forecasts and report the classification matrix and classification error rate. How does the neural network's performance compare to the two benchmark methods and the logistic regression model?

12. Which of the different models that you fitted would you recommend a stock trader use for forecasting stock movements on an ongoing basis? Explain.