**Springboard Foundations of Data Science Capstone Project Milestone Report**

**News Headlines as Predictor of Stock Market Movement**

**author: Raghu Kowshik**

**date: January 21, 2011**

**Introduction**

This is the Unit 6.1 Milestone Report for the Springboard Foundations of Data Science Capstone project. The objective of this project was to build a working predictive model. The data used in the model came from two sources. The first source is the News headline data from Newsmax archive at <http://www.newsmax.com/Archives/Newsfront/16/2008/1/> These headlines are NOT ranked by readers. Each date has more than one headline. The second data source is the Stock data from NASDAQ at <http://www.nasdaq.com/symbol/flex/historical> Only the Flex company stock was extracted. The date range is 09-01-2016 to 12-31-2016 (four months) for both the data. This milestone report describes the major features of the training data with an exploratory data analysis and summarizes the author’s plans for creating the predictive model.

**Obtaining the Data**

Data were down loaded from two sources, newsmax.com and Nasdaq.com, to an MS Excel file. The two data chunks were merged to form a single data file. The initial Nasdaq dataset contained the 84 observations. Each observation consisted of the following variables: date, close, volume, open, high, low. The newsmax.com headlines consisted of 20 headlines per day. The author merged the two Excel sheets into one, with the following variables: Date, close, volume, open, high, HL1, HL2, HL3, HL4, HL5, HL6, HL7, HL8, HL9, HL10, HL11, HL12, HL13, HL14, HL15, HL16, HL17, HL18, HL19, HL20.

**Cleaning the Data**

Before performing exploratory analysis, the author cleaned the data first. This involved converting the text headlines into a document-term matrix via a Corpus object using the “tm” function. Each row of the document-term matrix was combined headlines for each day. Columns were frequency counts of unigrams. The control object will tell the DocumentTermMatrix() function what to do with the text before converting to a term matrix. Punctuation and numbers were removed, everything was converted to lowercase, and common words were removed since they will likely have little predictive power. It was decided that the variables “volume”, “high”, “low” would not be used. The date column was converted into a date object using the “as.date” function.

Feature engineering was performed on the data as follows. A new variable “label” was added. The “label” variable was assigned binary values 0 or 1. If the close value was less than the open value, then the variable “label” was assigned 0, if the close value was greater than the open value, then the variable “label” was assigned a 1.

The final dataset consisted of 84 observations and three variables, namely “Date”, “Label”, and “All”.

The dataset was split into “train” and “test” sets. The “train” dataset was used for developing the model, and the “test” dataset was used for validating the model.

**Model Development – Exploratory analysis**

**A logistic regression model was used. Predictions were generated and the results were put into a data frame for plotting.** Densities of the predicted probabilities were plotted for each of the true values 0 (stock is down) or 1 (stock is up or unchanged).

**Model Assessment**

**ROCR package was used to assess the model performance. The AUC score was 0.6064. This implied that prediction was a little bit better than random guess.**

**Next Steps**

1. The sample headlines and stock price data is for 4 months. There are 84 observations. 42 observations show a upward (1) trend, and 42 observations show a downward (0) trend, are equally divided.
2. Larger sample size may improve the prediction performance. I propose to use 1 year data.
3. Naive Bayes and Support Vector Machines may give better results.
4. Use of bigrams is another possibility to consider for improving the model performance.
5. In this trial, I used Logistic regression. I will also check with the discrete CART model and the Random Forest approaches.