Project
OzCorp, LLC
Due Date: 3/28/2015

Overview

Day-Trading Algorithms are big in the world of Finance. Companies without tradings algorithms will quickly fall behind. It is your task to show-case preliminary results to predict forward stock returns based on twitter sentiment and volume.

Goal

To be able to predict forward returns based on twitter sentiment, and volume. In the world of day-trading, accuracy is important. Using sentiment to predict forward returns is a fairly new art. Any preliminary result with accuracy over 50% is interesting however your goal is to maximize the accuracy of your model. We at OzCorp are still learning the ins and outs of algorithm trading so we are relying on you, our core team of data scientists to produce results and to teach us about the data that we have. Please see the following page for the description of the data.

The Data

We have been collecting raw data for several years now. Mainly we have focused on collecting both stock prices and tweets regarding the stock in question. Of over 100 tickers in the database, we have included the data for stock ZYX here. The columns in the database are:

- **time** datetime of the recorded event
- **ZYXprice** Price of the stock at that moment
- **ZYX1MinSentiment** raw sum of sentiment of tweets in past minute.
- **ZYX5MinSentiment** raw sum of sentiment of tweets in past five minutes.
- **ZYX10MinSentiment** raw sum of sentiment of tweets in past ten minutes.
- **ZYX20MinSentiment** raw sum of sentiment of tweets in past twenty minutes.
- **ZYX30MinSentiment** raw sum of sentiment of tweets in past thirty minutes.
- **ZYX60MinSentiment** raw sum of sentiment of tweets in past sixty minutes.
- **ZYX1MinTweets** number of tweets about the stock in past minute.
- **ZYX5MinTweets** number of tweets about the stock in past five minutes.
- **ZYX10MinTweets** number of tweets about the stock in past ten minutes.
- **ZYX20MinTweets** number of tweets about the stock in past twenty minutes.
- **ZYX30MinTweets** number of tweets about the stock in past thirty minutes.
- **ZYX60MinTweets** number of tweets about the stock in past sixty minutes.
- **ZYX1minPriceChange** percent change in price of ZYX in past minute.
- **ZYX5minPriceChange** percent change in price of ZYX in past five minutes.
- **ZYX10minPriceChange** percent change in price of ZYX in past ten minutes.
- **ZYX20minPriceChange** percent change in price of ZYX in past twenty minutes.
- **ZYX30minPriceChange** percent change in price of ZYX in past thirty minutes.
- **ZYX60minPriceChange** percent change in price of ZYX in past sixty minutes.
- **5fret** percent change of ZYX five minutes into the future
- **10fret** percent change of ZYX ten minutes into the future
- **20fret** percent change of ZYX twenty minutes into the future
- **30fret** percent change of ZYX thirty minutes into the future
- **60fret** percent change of ZYX sixty minutes into the future

As we are not sure which of the frets (forward returns) will be your best response, we have included multiple options at varying time intervals.

Guiding Questions

1. This problem be viewed as regression but also as a classification. How?
2. There are 20 possible features and 5 possible responses. You will have to pick just one response but you don't need to use all 20 possible features. Try picking and choosing features
3. OzCorp mentions maximizing accuracy, however they do not mention sensitivity and specificity. What do those words mean relative to this project? Are one of these three metrics more important than the other?
4. In trading, you can either long or short a stock (ie. you can either bet the stock will go up or down). Will training a model to do both be accurate enough or should you only focus on one?
5. Always think about future work. What can be shown for preliminary results, and what other predictors might exist out there that you could ask for?

Plan of Action

- Brainstorm a few possible solutions (ie. maybe a regression, maybe a classification on these variables, maybe one plan uses 60fret as a response and the other uses 5fret). The idea is to rapid prototype
- Build out these 3-5 models using examples from class
- Pick the best 2-3. Remember your metrics determine if your models are performing well.
- enhance those models if possible
- finally choose the "best" model as determined by metrics.