FE-582: Foundations of Financial Data Science

Project Description
October 19, 2017

Project Overview:

The objective of the project is to analyze financial market data by applying data collection, data preparation, feature extraction, data cleaning, analytical processing and algorithms. Data clustering, data classification, outlier analysis, and data mining techniques may be implemented. The data can be obtained from the financial markets (stocks, financial statements, etc.) or from text data sources (twitter, news, etc.). You may implement new methods and argue the advantage of them over traditional methods.

Problem Statement:

Can we predict when the stock market has overreacted to bad news for individual stocks? If so, can we create a set of rules, based on analysis of historical data, to isolate individual stocks for selection in a long portfolio that will generate abnormal positive returns when compared to benchmarks?

Background:

Every day in markets around the world, a small number of assets fluctuate in their value significantly. A recent glance at the trading activity on the NYSE shows that on September 29, 2017, two stocks - Revlon (Ticker: REV) and W&T Offshore (Ticker: WTI) - both lost more than 10% of their price and market capitalization within the span of a few hours.

Over longer periods of time, according to the Efficient Market Hypothesis, all publicly available information is priced into the price of an security. However, over short periods of time, it has been shown that markets can act irrationally and over / underprice information into the price of a security. Our goal is to study whether or not it is possible to identify a set of criteria that can be used to understand whether or not investors have overreacted to negative information about an individual security.

Our Approach:

We will isolate our analysis to stocks trading on the NYSE over the past 10 years. For each trading day, we will isolate the stocks that declined by more than 10%. This will allow us to start with a manageable dataset, which can be expanded in the future using other thresholds. We will aggregate quantitative (e.g. P/E ratio, market capitalization, trading volume, float, % of institutional investors, Piotroski score) and qualitative (e.g. industry, month of event) data for each security from various data sources.

Once isolated, we will use supervised learning techniques to identify which variables are correlated with positive abnormal returns over multiple time periods (e.g. 1-day, 3-day, 5-day, 14-day, 30-day, 60-day, 180-day, 360-day), which we can use to isolate investment opportunities.

Solution Vision:
If we are able to identify a set of rules that consistently predict positive abnormal returns, we envision using them on a daily basis to evaluate losers (and eventually winners) in the stock market, making investment decisions at the end of intraday trading activities.

Challenges:
Our model seeks to distill a binary target variable (Yes - security price decrease is an overreaction, No - security price decrease is a valid valuation) into homogenous groupings. In creating these categorical variables, we will need to answer the following questions:

1. Threshold for one-day loss (current bench mark is 10%)
2. Gain threshold to determine over negative reaction:
    a. Time horizon
    b. Absolute gain
    c. Gain relative to industry or sector index

Data:
The study described here within will utilize three types of financial market data: i) Stock Price, ii) Company Financials, and iii) Composite Measures. The highest frequency data will be daily stock price, which will be considered a uniformly sampled time-domain signal. Statistical measures, including the autocorrelation of the adjusted daily close price and that signal's first derivative, are also candidates for inclusion in this study. Company financials, with quarterly and yearly frequency, will be included in the training and test data sets. Lastly, composite measure, such as Piotrowski's score, that are derived using both asset price and company financials, will be utilized.

The scope of the training data set will be limited to a to-be-determined set of securities selected from the index-under-study's (NYSE) biggest daily losers on days during a multi-year period (3-5 years) thru the end of 2015. Our goal in sizing the model training data set will be to use a computationally tractable number of securities that still provides strong enough basis to span the prediction space. The test data set will be an entire order of magnitude smaller than the training set and selected from the pool of daily losers during the year 2016.

The following **Data Sources** are considered candidates for this study:

1. Bloomberg Data
2. WRDS Asset Price and SEC databases
3. QuantMod
4. Quantl

Analytical Processing and Algorithms:

Candidate techniques for **Feature Extraction** from the time-domain signals are itemized below:

1. Statistical measures, including autocorrelation of the daily adjusted close price
2. Model-based parametric features, e.g.using Linear Predictive Coding, Vector Quantization (VQ)
3. Multidimensional scaling and Principal Component Analysis will be explored in search of low-dimensional representation of the mixed-domain signal (consisting of varied data types).

Clustering techniques presently under consideration for this project are K-Means and Support Vector Machines. **Classification** may be performed using Classification and Regression Trees (CARTs) or directly in a lower-dimensional component space obtained using Principal Component Analysis. A neural-network based classification approach is considered out of scope for this project. Where possible, availability of a prior implementation in R may direct the final selection of processing techniques.

Literature:

Thaler, R., & De Bondt, W. F. M. (1985). Does the Stock Market Overreact? *Journal of Finance*. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

Reviewed, P., & Paik, R. (2012). University of California Statistical Approach to Stock Market.

Naderi, M., & Mekanik, S. (2012). An Analysis of Short- Term Overreaction to Stock Market News : Iranian Evidence. *International Journal of Academic Research in Business and Social Sciences*, *2*(9), 230–242.

Ali, R., Ahmad, Z., & Anusakumar, S. V. (2012). Stock market overreaction and trading volume: Evidence from Malaysia. *Asian Academy of Management Journal of Accounting and Finance*, *7*(2), 103–119.

Veronesi, P. (1999). Stock Market Overreactions to Bad News in Good Times: A Rational Expectations Equilibrium Model. *Review of Financial Studies*, *12*(5), 975–1007. https://doi.org/10.1093/rfs/12.5.975

Salighehdar, A., Liu, Y., Bozdog, D., & Florescu, I. (2017). Cluster Analysis of Liquidity Measures in a Stock Market Using High Frequency Data. *Journal of Management Science and Business Intelligence*, *9264*(1985), 2–2. https://doi.org/10.5281/zenodo.808452

Mendel, J. (1991). Tutorial on higher-order statistics(spectra) in signal processing and system theory- Theoretical results and some applications. *IEEE, Proceedings*, *79*, 278–305.

Lathauwer, L. De, Moor, B. De, & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis …*. Retrieved from http://epubs.siam.org/doi/abs/10.1137/S0895479896305696