**Title: Statistical Analysis of Stock Volatility**

**Background and Motivation**

The primary motivation of the project is to highlight the "risk" associated with Stocks by applying various statistical analysis to quantify the risk factors such as “Market Impact” ("Beta", "Standard Deviation") and “Social Sentiment” (“Twitter feeds”). The fundamental reason to choose a topic in finance is partly current experience working in the industry and partly to understand the significance of "risk" being an end investor. Many of us manage different types of brokerage accounts (stock trading, 401k and IRA accounts) without clearly understanding the differing risk associated with the performance of these accounts. The risk appetite itself differs widely with different classes of investors depending on their personal situation and other factors. Hence it is prudent for such investors to understand the risk factors associated with their individual stocks or funds. Further, the recent advances in the social media has provided a wealth of information in the form of Twitter feeds or Google trends that can further provide good insight on the sentiment of investors influenced by current market events. This research is an attempt to highlight the concepts around "risk" factors including investor "sentiment" by performing a thorough statistical analysis backed by appropriate models.

**Project Objectives**

Volatility is a statistical measure of the dispersion of returns for a given security or market-index. Commonly, the higher the volatility, the riskier the security. The volatility itself is affected by multiple factors such as market (“Beta”, “Standard Deviation”), performance (“Earnings”), research (“Analyst Recommendations”) and social factors (“Social Sentiment”). We will start with the concept of Beta as a measure of volatility or risk and analyze how it affects the returns of a stock relative to the benchmark index (like S&P 500). In particular, we will run a simple case study by choosing three stocks with widely differing returns relative to the benchmark (S&P 500) to emphasize the effect of beta on different classes of stocks. We will then generalize this a bit more, by comparing Beta (risk) for a universe of stocks against their respective returns for a chosen period and more important, we want to understand **if a portfolio of low-volatility stocks earn returns different from a portfolio of high-volatility stocks.**This is one of the fundamental questions that this paper ponders upon by providing a statistical insight into the results. We will then repeat the analysis using Standard Deviation as the volatility measurement. The way we present this part of the material is by running a multiple regression model on stock’s return using beta and standard deviation as predictors.   
While the main focus of this research paper is on stock’s beta and standard deviation, we will also introduce other significant measures of volatility such as Twitter feeds that is gaining momentum since the advent of Social Media Revolution. The is primarily done to make this investigation more complete. .

**Summarizing high level questions and underlying analysis:**

**1. What impacts stock volatility? beta, SD, earnings, analyst recommendations, market impact, social sentiment etc.  
 Analysis of Beta and SD on stock performance**

- Beta estimation for chosen 3 stocks   
- Distribution of Beta for Russell 5000 basket  
- Distribution of SD for Russell 5000 basket  
- Sentiment analysis using twitter feed. Analyze the volatility of chatter for chosen stocks against its returns

**2. Does a portfolio of low-volatility stocks earn returns different from a portfolio of high-volatility stocks?** - Build regression model using one predictor – Beta vs Stock returns

    - Monte Carlo simulation by dividing stocks into 5 buckets (low to high beta) to answer above question.

- Build multiple regression using 2 predictors (Beta and SD)

- repeat above experiments using 1-yr/5-yr/10-yr beta annualized values.

**3. Does Social sentiment has any impact on stock volatility short or longer term?**

- Build regression model using raw #tweets as predictor - #tweets vs stock returns (4 week period)  
 - Build regression model using sentiment score as predictor – Sentiment Score vs stock returns  
 - Using a window of T-n days (T is current day) to see if the sentiment score has a better fit

What Data?

[http://finance.yahoo.com](http://finance.yahoo.com/)

**Data Set-1 (historical prices for chosen stocks)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Open**    **(Numerical)** | **High**    **(Numerical)** | **Low**    **(Numerical)** | **Close**    **(Numerical)** | **Volume**    **(Numerical)** | **Weekly Returns (%) (Numerical)** |
| 11/30/2015 | 2090.95 | 2103.37 | 2080.41 | 2102.63 | 3.98E+09 | 0.599000787 |
| 11/23/2015 | 2089.41 | 2095.61 | 2070.29 | 2090.11 | 2.95E+09 | 0.045002802 |
| … | … | … | … | … | … |  |
| 1/10/2005 | 1186.19 | 1194.78 | 1175.64 | 1184.52 | 1.48E+09 | -0.140780236 |
| 1/3/2005 | 1211.92 | 1217.8 | 1182.16 | 1186.19 | 1.6E+09 | 0 |
|  |  |  |  |  |  |  |

**Data Set-2 (Russell 5000 stock performance attributes)**

**Data Set-2**

[**https://research2.fidelity.com/Screener**](https://research2.fidelity.com/Screener)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol**      **(Categorical)** | **Market Capitalization**    **(Numerical)** | **Sector/Industry**    **(Categorical)** |  | **Volume (90 Day Avg)**  **(Numerical)** | **Total Return (5 Year Annualized) (Numerical)** |
| TALN | $13.8M | Consumer Discretionary |  | 0.0285 | 10.75663 |
| ED | $18.2B | Utilities |  | 2.09986 | 9.82558 |
| … | … | … |  | … | … |
| NEM | $9.2B | Materials |  | 8.75166 | -19.88953 |
| APPS | $98.4M | Information Technology |  | 0.31961 | -2.60258 |

**Data Set-3 (Sentiment analysis using twitter/stocktwits feeds)**

|  |  |  |
| --- | --- | --- |
| **Symbol**      **(Categorical)** | **Timestamp** | **Tweet** |
| GILD | 2016-04-06T16:20:07Z | $GILD here we go 109+ this month... |
| GILD | 2016-04-06T16:12:09Z | $GILD Nobody should be surprised. Faithful GILD nation knows 90&#39;s isnt where we belong. |
| GILD | 2016-04-06T16:06:53Z | @Stevemed Between $GILD, $AMGN and $CELG, you let these ride for a decade or two and you&#39;ll be happy. Simple. |
| GILD | 2016-04-06T16:02:10Z | $GILD 100 by end of week! |
| .. | 2016-04-06T16:01:50Z | $GILD why isn&#39;t this running on part with the sector? |

**Design Overview**

**Statistical and computational methods plan to use:**

1. Exploratory data analysis of the

2. Beta estimation using linear regression model for 3 chosen stocks. Use the first dataset here.

3. Impact of "beta" and "sd" on stock returns by building linear and multivariate regression model for Russell 5000 stocks. Use the second dataset here.

4. Monte Carlo simulation by dividing stocks into 5 buckets of differing beta to understand low vs high volatility stock performance.

5. Sentiment analysis using twitter feed. Analyze the volatility of chatter for chosen stocks and measure its correlation against stock returns.

6. Perform a Naive-Bayes analysis to predict the returns of the stocks based on the posts in the forums.

Or 6. Classify tweets into Bearish, Bullish or Neutral category using Naive-Bayes classifier to predict the returns of the stocks based on the tweets using Cross-Validation.