Volatility Prediction
Midterm Report

Wanlin Li wl596, Zhiwei Zhou zz498, Meiyi Li ml2549

October 28, 2017

Abstract

This project aims to develop a deeper understanding of stock return volatility. Traditionally, people use historical volatility as an estimation for future volatility. This project hypothesizes that a better volatility estimation could be attained through other information such as stock fundamentals. This midterm report focuses on the description and exploratory analysis of the raw data, as well as some preliminary analysis on the data that could pave way to further and deeper analysis.

1 Exploring Raw Data

1.1 Description

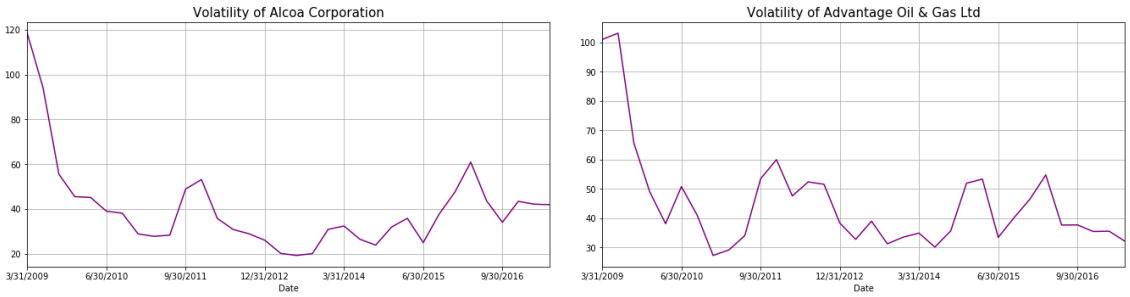

Figure 1: Volatility Data Sample

The volatility that we wish to predict is the volatility of all stocks in NYSE. (Sample in Figure 1) Our raw data come from Bloomberg, with a time span of 10 years. The data consist of one outcome variable, which is the volatility of stock daily returns over the past three months, and seven feature variables which are quarterly data of Asset Turnover, Current Ratio, Debt to Asset, Debt to Equity, Free Cash Flow per share, Invested Capital, as well as PE Ratio. Total number of data point is 107,304. Since we have feature data that scale only from 0 to 1, and we have feature data that are in hundred-level, we normalized our feature data and volatility data for truthful investigation of their relationship. Please refer to the files in the folder for sample data.

1.2 Basic Exploration

Since we are looking for relationship between each feature and volatility, we have graphed the scatter plot of volatility against each feature. At this stage, we are also considering close price as a new feature. Therefore we have added color in the graph to represent close prices for an in-depth understanding of the dynamics. Selected examples are demonstrated in Figure 2.

There are several preliminary understandings from the graph. For the upper left graph of Volatility against PE, we could see that there is a negative relationship. Moreover, the variance of Volatility of a given PE level decreases as PE increases. So our preliminary assumption is that there is a negative correlation between volatility and PE.

For the lower left graph of Volatility again Invested Capital, we can still see a vague negative correlation between the two variables. However, whether the correlation relationship is truly negative, and whether the

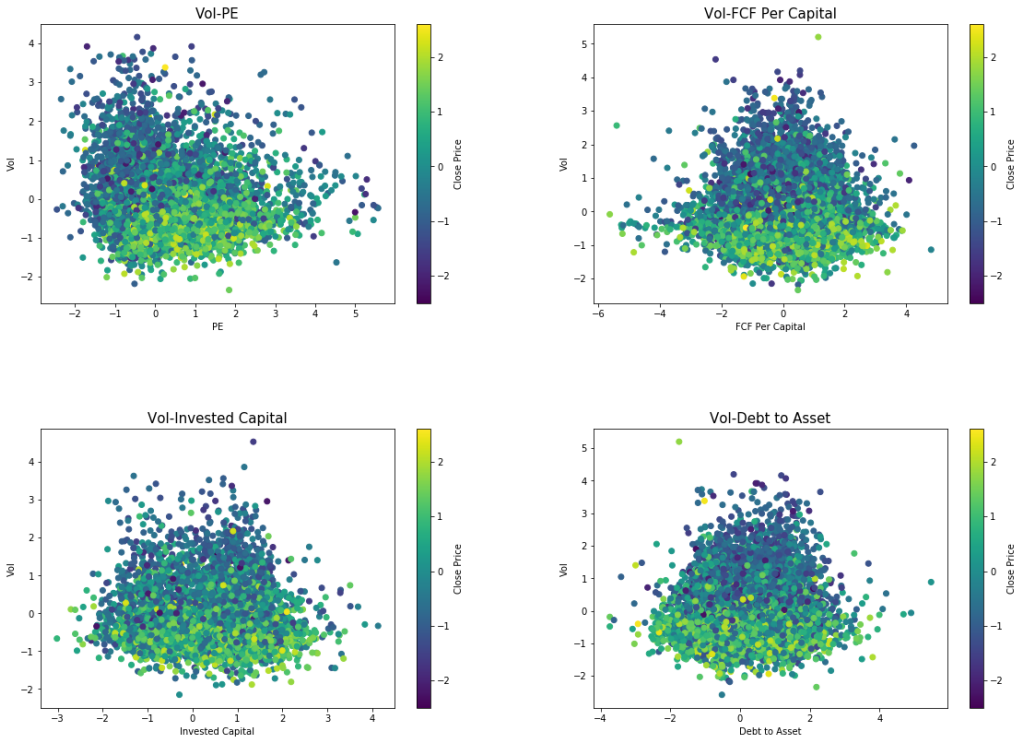

Figure 2: Scatter plot

relationship is linear or non-linear can not be immediately concluded from simply looking at the graph.

For the upper right and lower right graphs, the direction of correlation could not be observed. The only information we could get from these two graphs is that Volatility data have a wider span when the feature data is closer to average value. This makes sense, because when feature data, which is fundamental data take extreme values, market is more certain about the company's future performance, therefore the variance of volatility is low when feature data take extreme values. However, the prediction power of these two features on volatility does not appear to be strong.

Finally, there is a clear relationship between closing price (represented by color) and Volatility. When the closing price is low, Volatility tends to be high. Therefore, closing price is a promising feature to be used to predict Volatility.

1.3 Data Cleaning

There are several challenges during data cleaning. First is that our data were recorded with unaligned date time and indexing. Second is that a large part of the fundamental data is nan.

To tackle the first problem, we replaced the old indexes, which came in mixed forms of string and integer, with new self-created indexing in the form of pandas.datetime(). With the new index, we re-aligned the data with higher precision and generated a read-friendly database.

For the second problem, we investigated all the features for each stock. If the number of missing data is less than 10% of the total data under a feature of a specific stock, then we linearly interpolate the missing numbers. However, if the number of missing data is more than 10%, we drop the entire feature data for that specific stock.

2 Preliminary Regression and Analysis

In this section, we made the temporary assumption that our outcome variable, which is Volatility, is in linear relationship feature variables. We divided our data into training and test set with ratio 3:2.

2.1 Linear Regression with One Feature

We first regressed Volatility over each single feature separately. The motivation is to find out whether any feature has strong enough explanatory only by itself. An example of Volatility regressing against Price Earning Ratio is shown.

$$Vol = a * Feature + b \tag{1}$$

Vol : Normalized Volatility
Feature: We regress volatility against each of the feature variable in normalized form.

Figure 3 (left) shows the predicted Volatility against real Volatility of test set. As can be seen from the graph, data scattered focusing on the red line, which is prediction equals to real value. Therefore, we concluded that PE has explanatory power over Volatility, and we decided to keep this feature.

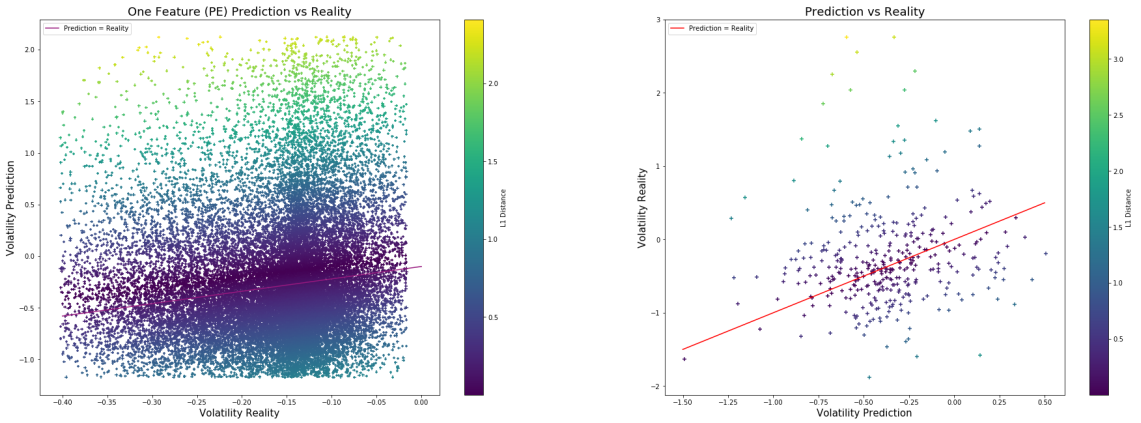

Figure 3: Left: Prediction vs. Reality for Regression Using Price Earning Ratio —— Right: Prediction vs. Reality for Regression Using All Features

2.2 Linear Regression with All Features

$$Vol = w_1 * AT + w_2 * CR + w_3 * DA + w_4 * DE + w_5 * FCF + w_6 * IC + w_7 * PE + constant \tag{2}$$

Vol	Normalized Volatility
AT	Normalized Asset Turnover
CR	Normalized Current Ratio
DA	Normalized Debt to Asset
DE	Normalized Debt to Equity
FCF	Normalized Free Cash Flow
IC	Normalized Invested Capital
PE	Normalized Price to Earning Ratio
Constant	constant (intercept) term

Figure 3 (right) shows that data scattered focusing on the line of prediction real value being equal. Therefore, we concluded that the linear combination of all features together has explanatory power over Volatility.

3 Future Plan

1. Since some of the pre-selected features did not show strong correlations, such as Free Cash Flow, we need to find more features that have higher explanatory power over the movement of Volatility. In our preliminary analysis, closing price could be a promising feature that worth further exploration.

2. At this stage, we only invested the linear relationship between Volatility and our features. There is possibility that the relationship exists in other form, such as non-linear. We will also consider feature engineering possibilities such as adjacent and lags.