Predictive and Prescriptive Analytics of Sock Market Volatility
CSC492: Final Project Report

Abdullah Khokhar
University of Toronto Mississauga
abdullah.khokhar@mail.utoronto.ca

Ishaan Sharma
University of Toronto Mississauga
ishaan.sharma@mail.utoronto.ca

Muhammad Zaidi
University of Toronto Mississauga
muhammad.zaidi@mail.utoronto.ca

Michael Liut*
University of Toronto Mississauga
michael.liut@utoronto.ca

Abstract

This work completed herein was part of a 4-month research and development project at the University of Toronto Mississauga. The intention was to create a stock trading bot that could accurately predict trends in the market and complete trades to yield a net profit over a fixed period of time. We describe our contribution and improvements made in this space.

Keywords: Data Analytics, Data Modelling, Prescriptive Analytics, Computational Statistics, Software Development, Machine Learning. Stock Trading.

1 Introduction

Advances and evolution in computing are inevitable, in 2016 a study demonstrated that 80% of stock market trading was completed through the use of algorithms, without a trace of human involvement *(Morton Glantz)*, and it appears that this trend is continuously on the rise. Thus, the development of new, faster, and more efficient computational models, algorithms, and data structures is a necessity. It is also an area of interest and enhancement for those involved in Computational Statistics and Machine Learning. This research project aims to tackle the gathering, organizing, and manipulation of large-scale stock market data to build a model which can successful predictive when to trade stocks. Specifically, the models investigated, theorized, and developed herein will work towards forecasting the securities price value on live markets, specifically the NYSE and NASDAQ.

2 Methods

2.1 *The Machine Learning Model*

There are many common indicators used when performing technical analysis on an asset. These indicators are useful as an individual measurable property to analyze the predictive nature of the asset and allows for the prescriptive analytics we perform to get an accurate representation of the financial market.

Using a radial basis function as the kernel for support vector regression (SVR), allows for the metrics to be linearly independent and the interpolation matrix to be non-singular, which can be beneficial for more dynamic and effective stock-market predictions.

The features set for the model include looking at wider timeframes and involving various periods on the features. Moving averages, standard deviation, exponential weighted moving average, moving average convergence and divergence, high/low rolling average, mean-reversion z-score prediction, difference in momentum, and relative strength index are the features that were considered. Time periods of 5, 10, 20, and 30 days were applied on the features to get a short-term prediction on the value of the asset and having the target price be the rolling mean of the next 14 days. This mean is the fluctuation of the asset and used to measure the volatility of the stock. Given the features set, the predictions are set to be measuring the difference of the close price of the current day and contrasted with the predicted rolling mean.

Starting with a linear kernel that has a standard scaler results in a higher mean variance error with more outliers (see *Figure 1*). Converting it to a minmax scaler with a polynomial kernel results in an anomaly, visually identifiable, with the predicted target being 0 (see *Figure 2*). Once the radial basis function is used as the kernel, and the scaling is improved, the trend is better behaved and outliers more apparent (see *Figure 3*). After refining the features with more periods and adding additional features to measure the difference in momentum and relative strength index, the variance is lowered (see *Figure 4*).

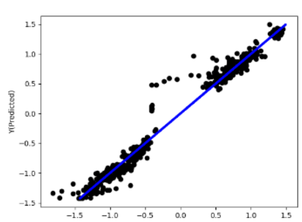

Figure 1: basic

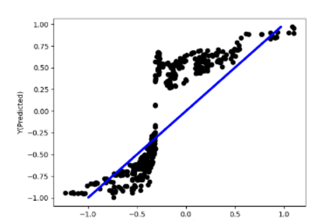

Figure 2: anomaly

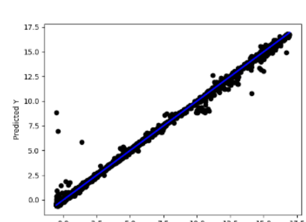

Figure 3: radial basis

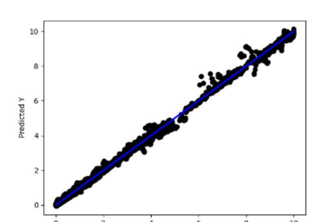

Figure 4: refined final

***course supervisor**

2.2 *Mean Reversion Model as a Tool to Train the Machine Learning Model on Market Volatility*

The return of an investment and the decisions which are made prior to it are directly impacted by the volatility of the current market value. For an investor, the fluctuation seen in the price and its indicators is of utmost importance. The mean reversion model follows the idea that the compilation of these indicators, including stock market volatility, having a strong correlation to an eventual revert in value towards the mean. This model can scope the current price of a security to be either overvalued or undervalued, predicting that these prices or any economic indicator will eventually revert back to the mean.

2.3 *How to Forecast Market Volatility and Price Reversion to Mean?*

The idea of predictive analysis in this model stems from the notion that a securities price and its market behavior are directly correlated to the underlying historical trend of price movement, and merely fluctuates around this barrier.

The fundamental concept of this model then becomes the characterization of what it means to be above and below the mean, as well as how this can be identified. For this, we will use the z-score: $z = \frac{x-\mu}{\sigma}$, where we provide the current mean and standard deviation to get a quantification as to how many standard deviations x is away from the mean. The model then developed on this method is a result of the hypothesis trial and test made identified in this section.

2.3.1 *Hypothesis*

If we take the value of a stock at the begging of each month and calculate there to be a z-score greater or equal than ± 1, then the stock value is either over-valued or under-valued and will revert back to mean in one month's time.

The testing of this hypothesis took place over a series of stocks using 3-years of historical data, with each month being a sample unit of the hypothesis test. A scatter plot (see *Figure 6*) was created to mark the returns of the first week in the X-axis, with the returns of the first week in the Y-axis. However, this did not give the desired results, as the average success plotting in either starting at a low and rising to high or visa-versa was below 50% amongst the 3-year dataset.

After further investigation, we realized that if we base our hypothesis around one singular stock, it may be subject to factors which influence that stock's price being above or below the mean beyond the scope of historical data, trends, and indicators. As a result, the current model's method is based around calculating the current days z-score amongst serval stocks, selecting those with largest deviations. These stocks then are analyzed using their historical data to find instances of when the same z-score was present, and the time it took to revert back to the mean is then averaged. The results can be found in *Figure 6*. However, this stand-alone model can be inaccurate due to the mean being impacted by unrelated stock details, thus, it is used as an additional feature to train the machine learning model.

3 Data and Results

3.1 Backtesting Results

The machine learning model was backtested using the auquan toolbox framework, in Python 3, which allows for backtesting on several timeframes with different instruments. During the testing phase, it became apparent that the model yielded better returns over a larger timeframe; more than 2 months. Further, the cumulative results show a correlation with the historical trend of the price movement. Running backtesting with the timeframe of March 2020 to June 2020, we see an overall profit (see *Figure 5*) which is beneficial as it was able to overcome the market crash from COVID-19 in March 2020.

```
maxDrawdown:84581.98245000001
maxPortfolioValue:1220140.094650
capital: -196788.8369495
portfolio_value: 1195133.6738505
total_profit: 599822.2293065002
capitalUsage: 1221795.2577495
score: 195133.67385049997
total_loss: 404688.5554560002
variance: 83592599.1927191
```

Figure 5: sample output of model execution with a start of $1M in the portfolio over a 3-month period, resulting in a positive return ("score") of $195,133.67.

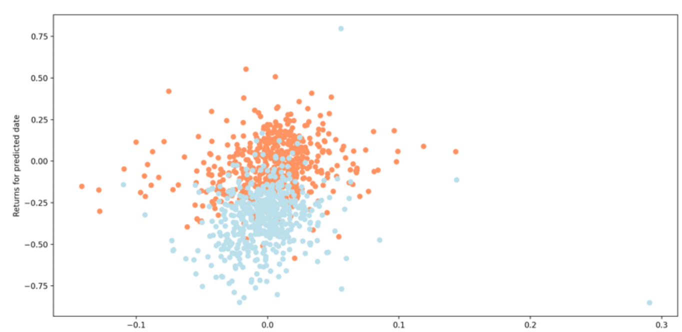

Figure 6: a scatter plot mapping the start date returns by the returns for the predicted revert date, across multiple stocks in the technology sector (blue) and industrial sector (orange).

References

Morton Glantz and Robert Kissell. Multi-Asset Risk Modeling: Techniques for a Global Economy in an Electronic and Algorithmic Trading Era, Academic Press, 2014, "8." pp. 250–260.