Using Logistic Regression and Random Forests to Predict US Equity Price Movements

Shankrith Natarajan
Columbia University
sn2955@columbia.edu

Jordan Boucher
Columbia University
jpb2232@columbia.edu

Behshad Jahan Pour
Columbia University
bj2400@columbia.edu

1 Abstract

This project is an effort to estimate the performance of the top 250 most liquid US equity stocks using machine learning techniques. The data is pulled from STOQ, a publicly available database, and technical indicators including momentum and volatility indicators are used as features. To create a baseline case, linear regression models were run and the final models comprised of random forest model using the LightGBM framework. Information coefficient was the main metric used to quantify accuracy. To establish robustness and the extent to which future predictions are possible, multiple models were run across multiple timeseries for both training and testing. The random forest models were more accurate than linear regression in predicting stock prices in the near future as well as farther away datetimes, but the difference was more pronounced for the near future.

2 Introduction

2.1 Decision Trees

The basic motivation for decision trees is wanting to predict some independent variable using a set of dependent variables. At each node in the decision tree, one particular predictor variable is considered, and the data point being tested is given a Boolean value based on whether or not it meets that condition. This process is repeated until the point reaches the end of the tree and is categorized into one of the possible outcomes for the dependent variables.

The machine learning aspect comes from the fact that there are theoretically an infinite number of decision trees given that at least on the independent variables is continuous. The machine, then, will consider many different trees and choose the single best one, where best is defined as the one that gains the most information at each point.

We will be using the Spearman rank-order correlation coefficient to judge our information gain, but a couple of other possible measure are:

$$\text{Entropy} = -\sum_{i=1}^{n} p_i log(p_i)$$

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} p_i^2$$

Notice that information gain can be defined equivalently as a decrease in randomness (i.e., entropy) or an increase in likelihood.

2.2 Random Forests

The procedure behind a random forest follows directly from decision trees. We simply take a random sample with replacement of the data that is the same length as the data. This process–along with selecting a subset of the predictors–is called bootstrapping, and we create a decision tree using the

same method as we did previously. We repeat this process numerous times (with different subsets of the data and the predictors each time) to arrive at a forest of decision trees:

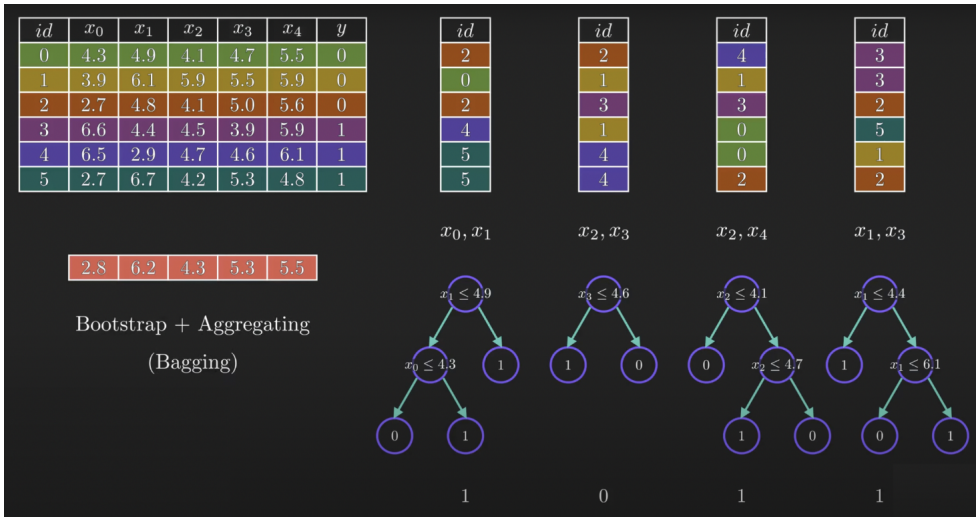

Figure 1: An Example Random Forest

Naturally, when we send a new data point through each tree, we are not going to receive the same categorization each time. In other words, the decision trees may disagree on how to classify the new data point. What we receive instead, then, is a probability distribution of possible outcomes for that point. The process of considering all of the decision trees for a new data point is called aggregating.

2.3 Information Coefficient

The Information Coefficient (IC) is defined as the mutual information between the two variables divided by the square root of the product of their entropies, which are measures of the randomness or uncertainty in a variable. We will be using the IC to judge the possible node conditions for each decision tree.

In general, the IC considers the correlation between two data sets. Naturally, then, its value will fall between -1 and +1, the former indicating that the data has a perfect negative correlation and the latter, that the data has a perfect positive correlation. The IC is often used as an alternative to the Pearson correlation coefficient, which measures the linear dependence between two variables. The IC is generally more robust to nonlinear relationships and can handle both discrete and continuous variables.

In particular, we will be using the Spearman rank-order correlation coefficient which has the feature of not assuming that the two datasets come from a normal distribution. It has been shown that daily returns to stocks do not follow such a 'nice' distribution, so it is advantageous to have the added flexibility.

3 Data Overview

3.1 Stooq

STOQ (Stooq) is a financial website that provides free access to a variety of financial market data. It is based in Poland and was founded in 2002.

STOQ contains a wide range of financial market data, including stock quotes, indices, currency exchange rates, and commodities prices. It covers a wide range of financial instruments, including equities, bonds, currencies, commodities, and more.

One of the key features of STOQ is its historical data, which allows users to access and analyze financial data over a specific time period. This can be useful for investors and traders who want to

track the performance of a particular asset or market, or for analysts and researchers who want to study trends and patterns in financial data.

Overall, STOQ is a useful resource for anyone interested in financial markets. It provides a wide range of data and tools that can help users to analyze financial data and make informed investment decisions.

3.2 Data Preprocessing

In order to obtain a comprehensive and accurate dataset for US equities, the STOQ platform was utilized to extract all available information for the specified time frame of January 2010 to November 2022. In order to ensure the integrity and usability of the data, the following steps were taken:

3.2.1 Data Cleaning

During the data cleaning phase, a number of issues were identified and addressed in order to ensure the integrity and accuracy of the data. These issues included the presence of missing values, missing ticker symbols, outliers, and potential junk values, as well as the possibility of duplicate records. To address these issues, various data cleaning techniques were implemented. Missing values were either imputed or removed, depending on the specific context and the impact on the overall dataset. Ticker symbols were checked against a reference list and any discrepancies were corrected. Outliers were identified using boxplot analysis and either removed or transformed as appropriate.

Junk values were identified and removed, and the dataset was checked for duplicates and any duplicate records were removed. It is possible that these issues arose due to errors in the maintenance and updating of the STOQ database. By addressing these issues through data cleaning, a more accurate and reliable dataset was obtained for further analysis and use.

3.2.2 Identifying Most Liquid Stocks

Liquidity is an important factor to consider when analyzing stocks, as it reflects the ease with which an asset can be bought or sold in the market. In general, stocks with higher liquidity tend to be more popular and more actively traded, which can provide a more accurate representation of market trends and sentiment. To identify the most liquid stocks, a calculation was performed by multiplying the close price and the volume traded for each stock on a given day, and then ranking the stocks in descending order. Only the top 250 most liquid stocks were included in the analysis, as this provided a sufficiently large and representative sample of the market. By focusing on the most liquid stocks, the analysis was able to more accurately reflect the overall market trends and dynamics.

3.2.3 Feature Engineering

The features included both momentum indicators, which measure the strength and direction of a trend, and volatility indicators, which measure the level of fluctuation in the price of an asset. Time indicators, such as the date and time of the data points, were also included in the analysis in order to account for any temporal patterns or trends. Additionally, various stock information features, such as the ticker symbol and sector, were included in order to provide context and additional insight into the data.

To ensure standardization and comparability across stocks, the relative return percentile was used instead of absolute returns. This allows for the returns of different stocks to be compared on a common scale, rather than being affected by the absolute size of the stock. In addition to these features, a number of technical indicators were also considered in the analysis. These included the percentage price oscillator, which measures the difference between two moving averages, the relative strength indicator, which compares the strength of a stock to a benchmark index, the normalized average true range, which measures the degree of price volatility, and the bollinger bands, which provide a measure of the volatility of a stock based on its past price movements. These will be covered in detail in the next section.

3.2.4 Defining Training Periods

To ensure a consistency in our models in terms of dates, a mechanism called lookaheads were incorporated. Multiple models were ran across different timeseries to ensure robustness of the

model. To do this, there had to be no overlap between train and test data, all the while increasing or decreasing testing and training periods. This would help in identifying robustness of the model as well as checking if the model is good at predicting timestamps in the near future or farther dates. To incorporate this, a variable called lookahead is defined. This defines the number of days post which testing period begins. The training, testing, and lookahead sets are given below:

- training = [63 days, 126 days, 1 year, 2 years, 3 years, 5 years]
- testing = [5 days, 21 days]
- lookaheads = [1 day, 5 days, 10 days, 21 days]

4 Technical Features

4.1 Percentage Price Oscillator

The percentage price oscillator (PPO) is a technical indicator used to measure the difference between two moving averages of an asset's price Robert (2004). It is typically used to identify trends and potential buy or sell signals in the market. The PPO is calculated as follows:

$$\text{PPO} = (\text{12-day EMA} - \text{26-day EMA}) / \text{26-day EMA} * 100$$

where EMA stands for exponential moving average. The 12-day and 26-day moving averages are typically used, but other time periods can be used as well. The PPO is then plotted as a line on a chart, with the centerline at zero and values above and below the centerline indicating upward and downward trends, respectively.

The PPO can be used in conjunction with other technical indicators and analysis techniques to identify trends and make informed investment decisions. It is important to note, however, that like all technical indicators, the PPO should not be used in isolation and should be considered as part of a broader analysis of the market.

Figure 2: Percentage Price Oscillator Plot

4.2 Normalized Average True Range

The normalized average true range (NATR) is a technical indicator used to measure the degree of price volatility in a financial asset Robert (2004). It is calculated as follows:

$$\text{NATR} = \text{ATR(n)} / \text{Close(n)}$$

where ATR(n) is the average true range over a specified number of periods (n) and Close(n) is the closing price over the same number of periods. The average true range (ATR) is a measure of volatility that takes into account both the magnitude of price movements and the gap between the high and low prices during a given period. It is calculated as follows:

$$ATR(n) = ((n-1) * ATR(n-1) + TR) / n$$

where TR is the true range for a given period, which is the maximum of the following three values:

- High(n) - Low(n)
- Absolute value of High(n) - Close(n-1)
- Absolute value of Low(n) - Close(n-1)

The NATR is typically plotted as a line on a chart, with higher values indicating higher volatility and lower values indicating lower volatility. It can be used in conjunction with other technical indicators and analysis techniques to identify trends and make informed investment decisions.

Figure 3: Naturalized Average True Range Plot

4.3 Bollinger Bands

Bollinger bands are a technical analysis tool used to measure the volatility of a financial asset Robert (2004). They consist of three lines plotted on a chart: a simple moving average (SMA) in the middle, and an upper and lower band that are typically two standard deviations above and below the SMA, respectively. The bands are calculated as follows:

$$\text{Upper Band} = \text{SMA} + 2 * \text{Standard Deviation}$$
$$\text{Lower Band} = \text{SMA} - 2 * \text{Standard Deviation}$$

where SMA is the simple moving average over a specified number of periods and Standard Deviation is the standard deviation over the same number of periods. The SMA and Standard Deviation are calculated as follows:

$$\text{SMA} = \text{Sum of Prices over n periods} / n$$
$$\text{Standard Deviation} = \text{Square root of Sum of Squared Differences from Mean over n periods} / (n-1)$$

The Bollinger bands are plotted on a chart with the SMA in the middle and the upper and lower bands above and below it. They are typically used to identify trends and potential buy or sell signals in the market. For example, if the price of an asset breaks through the upper band, it may indicate an overbought condition and a potential sell signal. Similarly, if the price breaks through the lower band, it may indicate an oversold condition and a potential buy signal.

It is important to note, however, that Bollinger bands should not be used in isolation and should be considered as part of a broader analysis of the market. Additionally, the choice of the time period

used to calculate the bands can affect the resulting values and should be chosen carefully based on the specific context and the investment horizon.

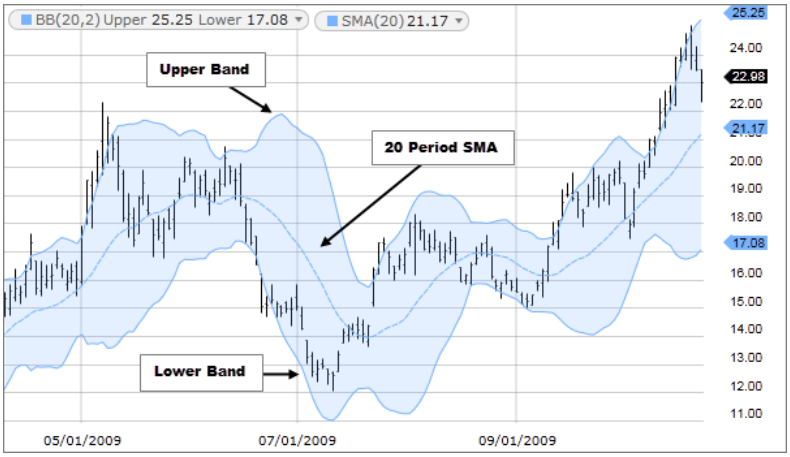

Figure 4: Bollinger Bands

4.4 Relative Strength Index

The relative strength index (RSI) is a technical analysis tool used to measure the strength of a financial asset relative to a benchmark index. It is calculated as follows:

$$\text{RSI} = 100 - 100 / (1 + \text{RS})$$

where RS is the relative strength, which is calculated as follows:

$$\text{RS} = \text{Average Gain over n periods} / \text{Average Loss over n periods}$$

The RSI is typically plotted as a line on a chart, with values ranging from 0 to 100. It is typically used to identify overbought or oversold conditions in the market, with values above 70 indicating an overbought condition and values below 30 indicating an oversold condition. However, it is important to note that the RSI should not be used in isolation and should be considered as part of a broader analysis of the market. Additionally, the choice of the time period used to calculate the RSI can affect the resulting values and should be chosen carefully based on the specific context and the investment horizon.

The RSI can be used in conjunction with other technical indicators and analysis techniques to identify trends and make informed investment decisions. It is important to note, however, that like all technical indicators, the RSI is a statistical measure and should not be relied upon as the sole basis for making investment decisions.

Figure 5: Relative Strength Index

4.5 Visualizing Technical Features against a single stock

Consider the stock behavior of SLGN.US, Silgan Holdings, a manufacturer of consumer goods packaging products and metal containers for the food and beverage industries. The company was founded in 1987 and is headquartered in Stamford, Connecticut, United States. Technical indicators are plotted alongside stock performance from January 2010 to November 2022.

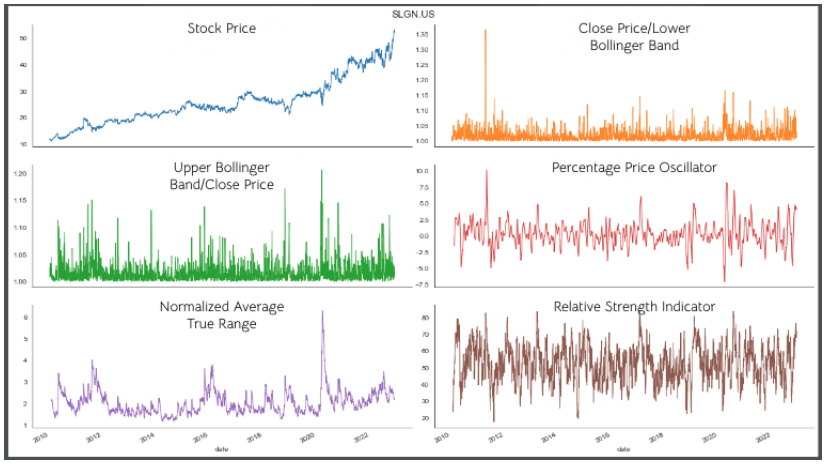

Figure 6: SLGN.US

As seen, the stock price behavior is effectively captured in the technical indicators. Post 2020, markets experienced extreme volatility due to the COVID crisis which has been captured by the NATR plot. The spikes in stock price are also captured by both PPO and Bollinger bands. In this way, the stock prices can be indicated using technical indicators.

5 Baseline Linear Regression Results

Initial results were produced with a linear regression model to establish a baseline. As expected, the linear regression model was not able to effectively capture the behavior and the information coefficient was negative. This indicates a negative correlation between the results predicted by the model and the actual stock behavior. There were some promising results in the 5 day lookaheads but nothing which can be used in an actual trading scenario.

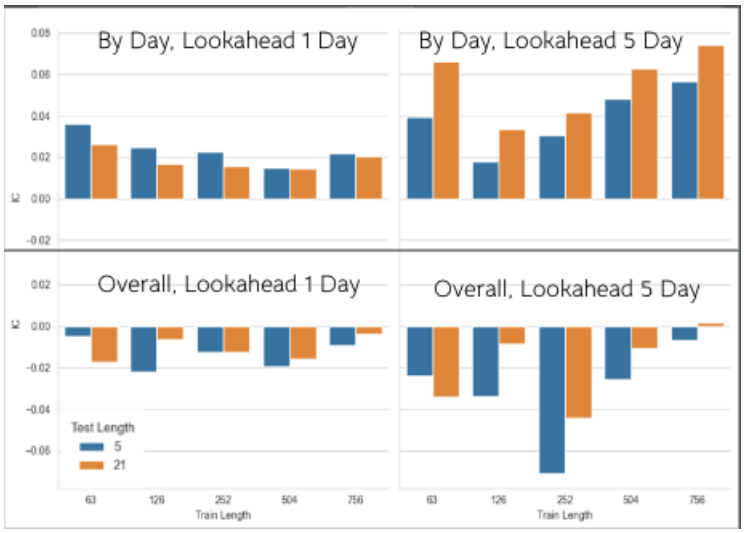

Figure 7: Baseline Linear Regression Results

As seen in the above graphs, the overall information coefficients, which is an average of each day's information coefficient is negative which indicates no relation between predicted and actual results. But when predicting for a single stock after a period of 5 days, the model seems to be performing reasonably well.

6 Random Forest Results

Next the results for the Random Forest model were produced, utilizing the open source LightGBM framework. Generally, the information coefficient was positive on average for all lookaheads, however, as seen in the figure below, higher lookaheads had higher variability in information coefficient.

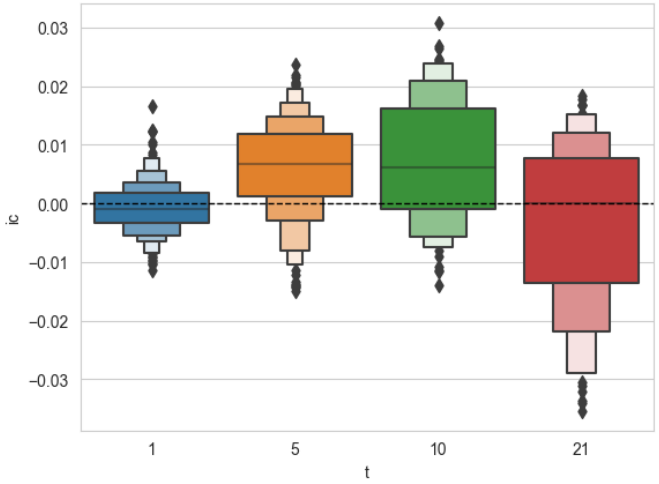

Figure 8: Random Forest Results

The graphs below display more specific data for a test length of 21. Overall, the model with a training length of 126 and test length of 21 had the highest average information coefficient across all four lookaheads.

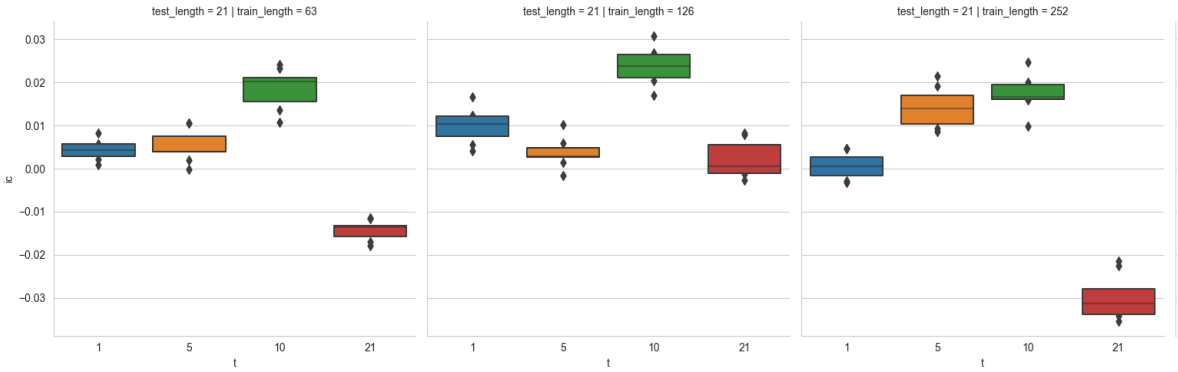

Figure 9: Random Forest Results for Test Length of 21

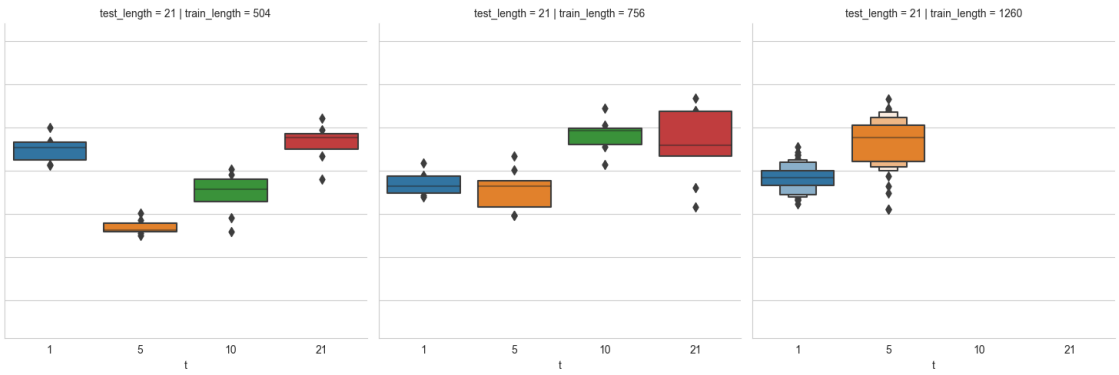

Figure 10: Random Forest Results for Test Length of 21 [2]

7 Analysis

7.1 Comparison of Random Forest with Linear Regression

Overall, the results for the Random Forest model significantly outperform the baseline Linear Regression results, as seen in the graph below; however, linear regression is competitive for the lookahead of 21 days. One reason for this difference in performance could be that decision tree models are non-linear, and thus they can factor in non-linear relationships between signals and the price.

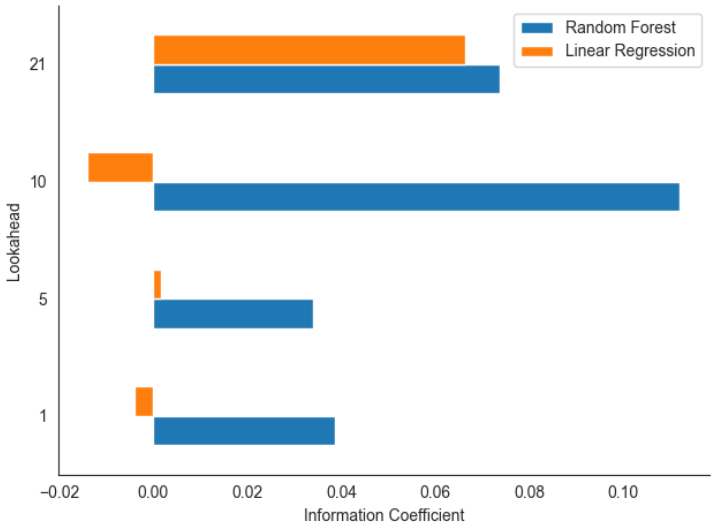

Figure 11: Comparison of Random Forest with Linear Regression

7.2 Factor Quantile Analysis

Finally we conducted a factor quantile analysis based on the results. The two graphs below plot mean period wise return by factor quantiles. We split our collection of stocks into five quantiles based on the predictive factor generated by the Random Forest model, and collected the mean returns of each quantile for the four lookahead periods, as seen in the graphs below.

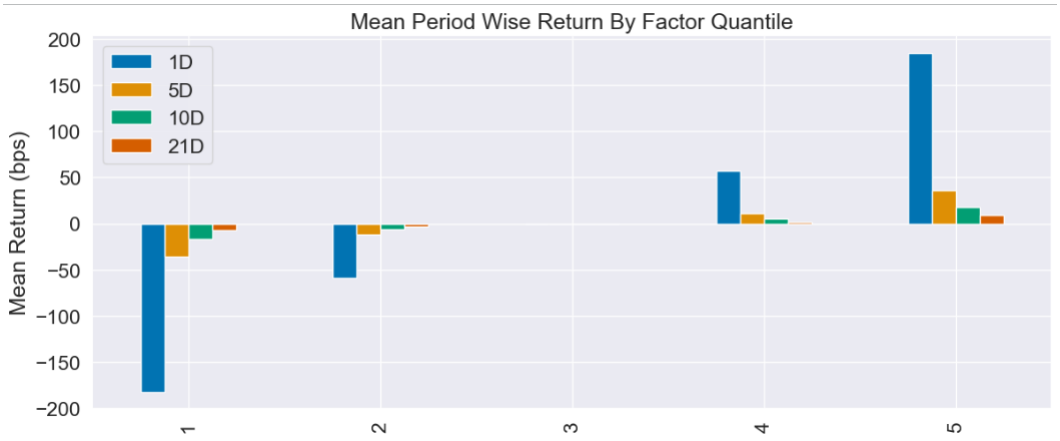

Figure 12: Factor Quantiles (In Sample)

Looking at Figure 12, we see an symmetric split between the quantiles, which could potentially be used as part of a trading strategy: we can short the stocks in the bottom two quantiles and go long on stocks in the top two quantiles. This strategy leverages the symmetry of the quantile split to potentially generate profits.

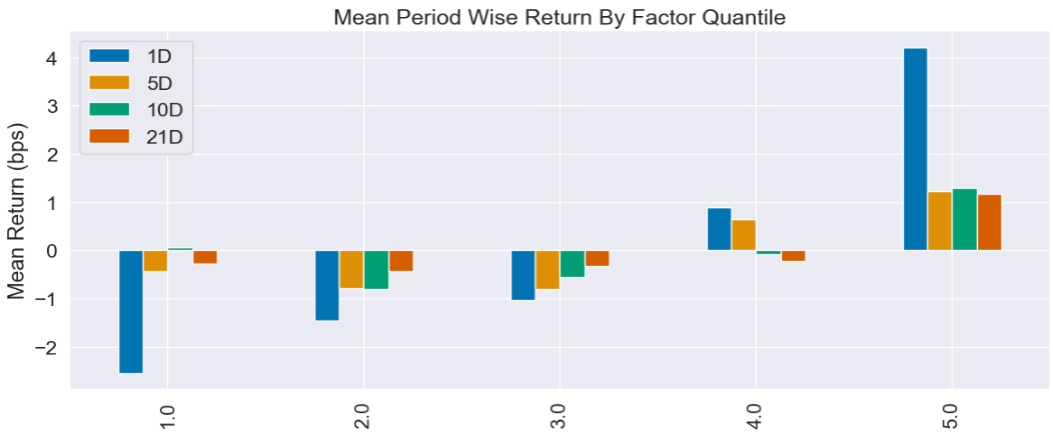

Figure 13: Factor Quantiles (Out of Sample)

The factor quantiles for the out of sample data do not have the same split desired, and in general the random forest model did not perform very well on out of sample data. One potential reason for the model's underperformance compared to our expectations is that out of sample data ranged from January 2020 to December 2021, coinciding with the COVID-19 pandemic. Therefore, the effectiveness of the model could be affected by the impact of the pandemic on the market. The global economic downturn caused by the pandemic's impact on supply chains and consumer spending may have made it difficult for the model to accurately predict market movements using internal signals.

References

[Robert 2004] ROBERT, D.: Technical Analysis of Stock Trends. (2004)