Predicting Stock Market Future Returns Using Past Returns

# Introduction

In 1792 the New York Stock Exchange was created, which marked the creation of the stock market, a revolutionary tool in American economics, where individuals were now able to buy a percentage of a company they like and which gave companies more financial support to expand and create jobs. For individuals the stock market was also revolutionary, as it provided them a platform to invest their extra money in companies they think will do well and make more money passively.

In today’s day and age, the stock market has become the source of a hyper-obsession with trying to predict the future value of stocks in an effort to get rich quick. However, the reality is that the value of stocks are very volatile to what is “trending,” as stocks are only valued at what people think they are worth, not what they are actually worth. Public opinion and hype can over inflate or deflate a stock’s true worth so it is often thought impossible to predict the market. In the movie Wolf of Wall Street, the character Mark Hanna states “Number one rule of Wall Street. Nobody - and I don't care if you're Warren Buffet or if you're Jimmy Buffet - nobody knows if a stock is going to go up, down, sideways or in circles. You know what a fugazi is? It's a whazy. It's a woozie. It's fairy dust. It doesn't exist. It's not real.”

Today we set out to test this theory using Machine Learning on the S&P 500. The S&P 500 is a stock of a collection of the top 500 companies, and it is often a useful measure of how the entire stock market is performing. In comparison with looking at the stock value of only one company, which is more susceptible to dramatic shifts and trends, the S&P 500 is much more robust, as no one company can skew the value on its own, which is why we have selected this particular stock. We will use Linear Regression to use previous data of the S&P 500 to predict the future of its value.

# Abstract

The idea behind this project is to use past returns of an index to predict what will happen in the future. There is a theory in Financial Modelling and Engineering called Momentum. It is the idea that stocks that have strong sustained changes in price in a certain direction, will continue to move in that direction given no external event occurs to break the trend. Price momentum is related to some psychological concepts related to stock investing. If the price of a stock moves up, it encourages others to bet with the trend because they do not want to miss out on the opportunity to grow their money with the trend. Human greed drives investors to buy to ensure that they not missing out on the opportunity of a upwards moving market. This theory also applies for stocks which are falling in value. As stocks falls in value, there is mass panic that causes more people to sell and this causes a correlation between past returns and future returns.

# Prior Studies on Momentum

Momentum is a popular Quantitative Investing Strategy that has been used by Hedge Funds to maximize returns. A popular study which first studied momentum was by Jagadeesh and Titman (1993). The strategy they studied the returns generated from going long on stocks which had positive momentum and short on stocks which had the greatest negative momentum. This strategy had generated a 1% return on average for the time it was tested.

Another study showed that stocks which are close to their 52-week high, are more likely to move upwards than they are to move downwards. And the same was proved to be true for stocks near their 52-week low, as they were more likely to move down than up.

The studies on Momentum highlight the psychological factors that are highly dominated in the market and their effects on stock returns.

# Methods

This project had 3 major steps:

1. Data Procurement
2. Data Processing
3. Visualization
4. Modelling
5. Evaluation

# Data Procurement

The data we used for this project is from Yahoo Finance. We used the Yahoo Finance module. Yahoo Finance allows for calling their API to get time series data of most financial assets in their databases. We pulled data for the SPY symbol which represents the value of the 500 largest companies in the USA. The data pulled from Yahoo Finance is returned as a Data Frame.

# Data Processing

The relevant features required to conduct this regression model was the ‘close’ feature from the Yahoo Finance Dataset. Once we had this, we had to use a finance formula to calculate the price momentum for a single day and a week.

Here, W1 to Wm represents the weights associated with each Price Return Calculation. P0 represents the current price of the financial asset. n1 to nm represents the time period to lookback on to calculate, for example, P12 represents the price 12 days ago.

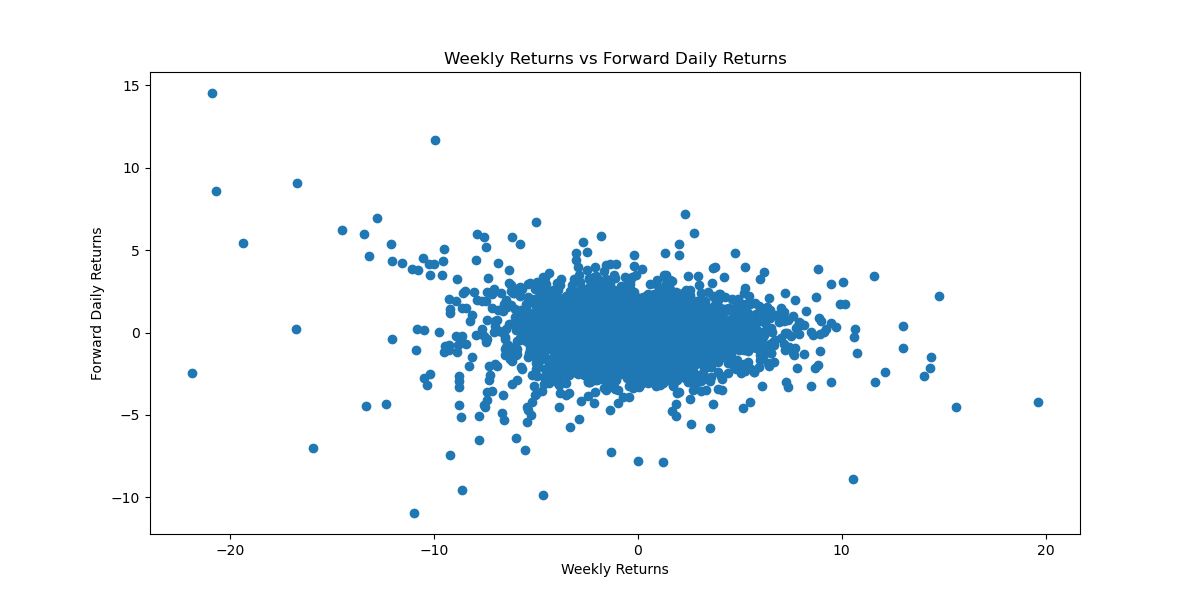
We utilized this formula for 2 separate time periods of 1 day and 6 days (1 week) to calculate momentum values for each time datapoints. The new features created in our dataset was the 'Weekly Returns’ and 'Daily Returns'. We weighted the returns for each time frame with a weight of 1.

The next column we had to calculate as the forward price return. This functioned as the target variable for our data set.

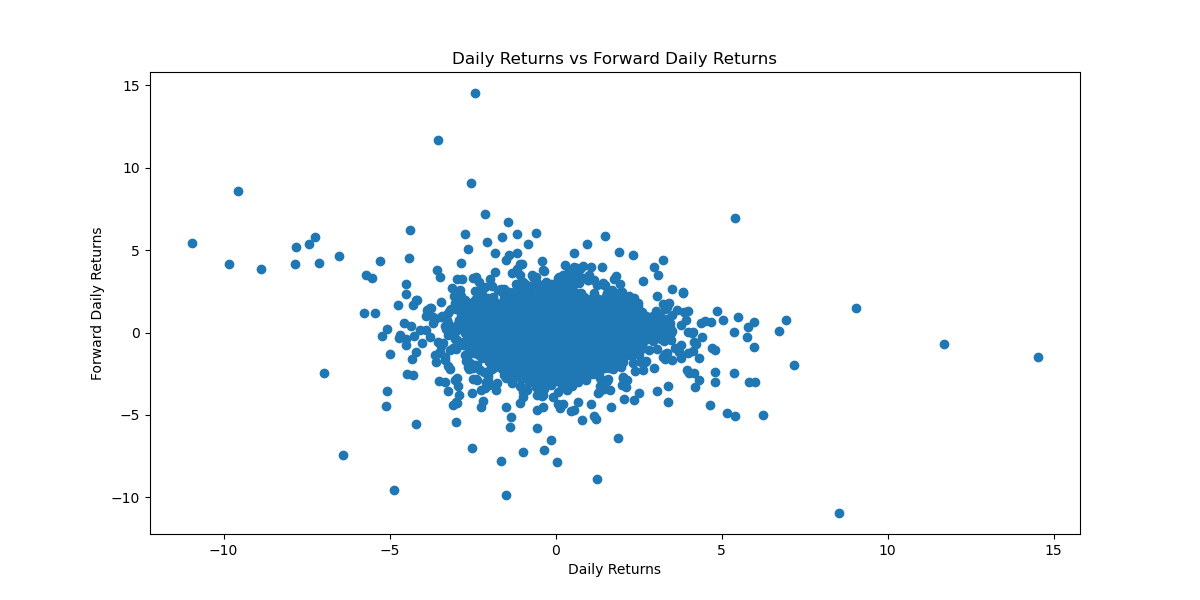
Here Pn represents the forward-looking price. So P6 would be the price of the asset 6days in the future. This will allow us to see if the past daily and weekly returns allow us to predict the future returns.

We used the shift function for pandas in python to calculate these momentum scores. We cleaned the data set of Nan values using dropna function in pandas.

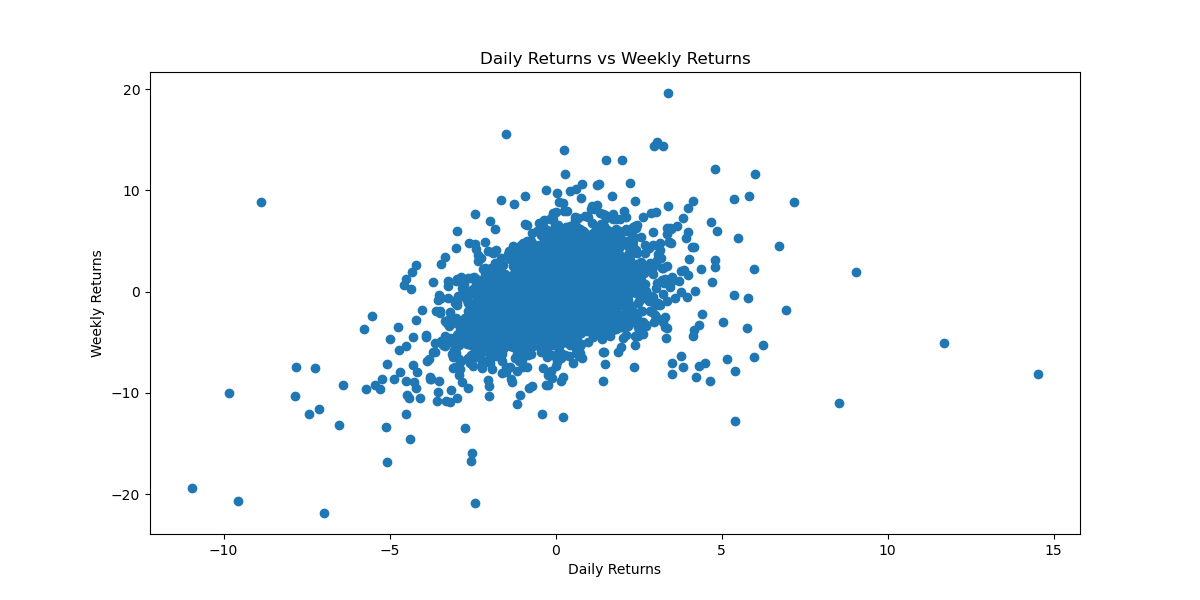
# Visualization

Since all the data in our project is numerical. We used Scatter plots and 3D graphs to visualize the data. We generated 3 2D scatter plots to map out our features to the target variable.

As we can see here there is not a strong relationship between weekly returns and forward daily returns, with a correlation coefficient coming at -0.02. This negative value could be due to the outliers in the data set. The stock market’s volatility seems to spike in both directions during market turmoil and this might have caused this.



Again, the correlation is extremely low. There values of these data points cluster around 0 and do not stray away from 0. There are some outlier values again due to the volatility during market turmoil. The correlation coefficient is not very high during this period again. This might be due to the short-term time period used and possibility of random prices during this time period.



We also mapped out the Daily Returns vs. Weekly Returns, this graph shows a relatively strong positive correlation of 0.56. This signifies that there is some extent of positive correlation between weekly and daily returns. But the caveat is that there is an overlapping data in this set, which might lead to this.

If we were to visualize this data set in 3 dimensions, we should expect a plane with which returns moves upwards from negative daily and weekly returns towards positive daily and weekly returns.

# Modelling

We decided to use Linear Regression Model for our Machine Learning Algorithm. But we had also considered using other ones like KNN with numerical data. The reason for using Linear Regression was that we feel that we should be able to get a well fitted plane for our data. As the past returns are negative we should expect negative future returns and vice versa.

The features of our modelled were:

1. Daily Price Returns
2. Weekly Price Returns

The Target Variable:

1. Forward Weekly Returns

We used standard procedures to generate the linear regression plane for this model. After implementing the Model, we got the following results for the parameters:

The final equation that fit the data was:

(-0.06276744)\*X1 + (-0.02636234)\*X2 + 0.0530774920171444

We also computed the MSE and RMSE values:

**MSE:** 1.407587118932942

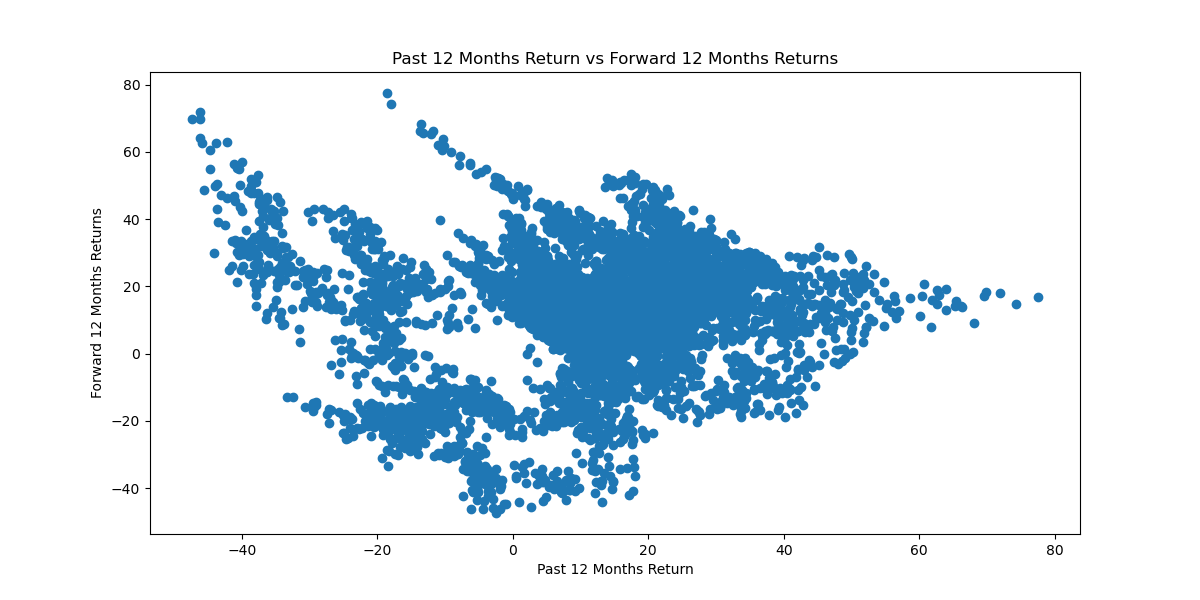
**RMSE:** 1.186417767455015

# Evaluation

Based on the metrics like RMSE calculated for our Model, its value is too high to trust this model in predicting future returns of a stock. The mean returns of the SPY is around 0.2% per day, and with an RMSE value of 1.18%, it represents a very large possibilities of value for the future days price’s return. With this level of error, we can have negative or positive values which would be equally likely in the derived error bound.

One downside of using price momentum to predict the future prices of index funds is that the results are highly dependent on the weights that are used. It is possible the weights used have a high error in their predictive power because their values were not optimized to maximum correlation with future returns. Trying to achieve high levels of optimization might result in overfitting of this formula and thereby leading to chances where the model does not perform well on test data. However, there are ways that the weights can be optimized depending on the formula’s used for price momentum. Studies by Keller, show that there are price momentum formulas that are optimized for predicting future returns. These formulas weight higher time periods more than lower time periods but do not include not Machine Learning models for prediction.

Another improvement would be to increase the time period of the time frame. As we drill down in time frames, the charts become more polluted with false moves and noise. Increasing the time frame allows the model to recognize the structure or the trend it follows disregarding any other factors affecting the price. Here is the chart of the 12 Rolling Month Price Return vs. the Future 12 Month Price Return Graph:



As we can see here the trend has a more positive relationship. But there is still too much variability to come to a valid conclusion weather price momentum can be used as a determining factor. The correlation coefficient in this case was around 0.1.

While a regression model achieves the basic goal of very loosely predicting the price of the stock, a more efficient and successful model would be a classification model. A classification model used in this approach would focus on whether the price increases or decreases or remains stagnant instead of predicting the actual value. This method would be more suitable for short term as well. A SVM algorithm within the classification model would be suitable as it would be able to handle the non-linear and dynamic nature of markets. The value range of each attribute has their own pattern, so it seems that linear classification is very hard to build a correct relation from all the attributes to the class label in this case. However, the SVM is good at non-linear classification and can be tried to develop a strong connection from the input attributes to the class label by supervised learning. SVM has also known to be a better indicator as the feature selection and classification modules built on it would boost up classification accuracy.