Analyzing S&P500 Companies and Predicting Annualized Returns

ehan31, elu14, ewang77, zchen186

Overview

In this project, we delve into the intriguing world of financial arbitrage, specifically investigating the stock market through the lens of data science. Our core objective is to uncover hidden arbitrage opportunities that emerge from significant events and their subsequent impact on various stocks. These events can range from corporate actions like mergers and acquisitions, and earnings announcements, to broader macroeconomic events such as policy changes, geopolitical developments, or even unexpected occurrences like natural disasters.

We hypothesize that significant events, ranging from corporate actions to macroeconomic developments, lead to predictable patterns in stock price movements. We aim to use these patterns to identify and exploit arbitrage opportunities. Consequently, we are mostly interested in the following task setting: given beta values and the mean annualized risk values of S&P500 companies, predict their mean annualized returns.

Data

The data attained was collected using webscraping techniques from Wikipedia and the IEX Fianance API, which is one of the top global sources of financial data and news for stocks. We navigated the Wikipedia webste and collected data about company names and stock ticker symbols that was nested within different HTML and react tags, which made the collection process difficult. We specifically scraped the list of S&P500 Stocks on Wikipedia, and we removed those stocks that did not have active data. This sample is comparably small when taking into account the number of total US stocks, but it is more likely to be representative of the stocks that most investors would choose to invest in. After obtaining the list of stock names and symbols, we used the IEX Finance API to obtain more detailed financial information, such as price history and detailed returns.

Findings

Below are the analyses for each of our individual hypotheses.

Claim #1: Higher beta is correlated with a higher annualized mean return.
Support for Claim #1: The scatterplot visualization, with linear regression, shows the relationship between the beta and the annualized mean return for the stocks. A two-sample t-test assessing the significance of the difference in mean annualized returns between high-beta and low-beta stocks supports this claim.

Hypothesis	Statistical Test	Test Statistic (t/Z-Value)	Degrees of Freedom	P-Value	Alpha Level	Null Hypothesis Rejected?
1	Two-Sample t-Test	9.027627434487139	N/A (assumed equal variances)	1.4426540224673013e-14	0.05	Yes

Claim #2: The US presidential election on November 3, 2020, did not have a significant impact on stock market returns.
Support for Claim #2: We used a one-sample t-test to examine the average abnormal return of the stock market on the day of the election. The results showed that the average abnormal return was not statistically different from zero, indicating no significant impact.

Hypothesis	Null Hypothesis	Statistical Test	Test Statistic (t/Z-Value)	Degrees of Freedom	P-Value	Alpha Level	Null Hypothesis Rejected?
2	mean=0	One-Sample t-Test	0.21373513864376223	N/A (single sample)	0.8307536077929717	0.05	No
2	mean=sample_mean	One-Sample t-Test	0.1048038305434851	N/A (single sample)	0.9165314704574674	0.05	No

Claim #3: Companies with higher annualized risk have higher returns.
Support for Claim #3: The results of a two-sample t-test revealed a statistically significant difference in mean annualized returns between companies with higher and lower annualized risks.

Hypothesis	Statistical Test	Test Statistic (t/Z-Value)	Degrees of Freedom	P-Value	Alpha Level	Null Hypothesis Rejected?
3	Two-Sample t-Test	7.39673826859103	N/A (assumed equal variances)	4.285239077218654e-11	0.05	Yes

ML Results
In the machine learning component, we aimed to predict the annualized returns of S&P 500 companies using beta and annualized risk values. While both models showed some predictive power, neither achieved high accuracy, suggesting a need for further refinement or alternative approaches. It is furthermore likely that using a linear regression model may not have been applicable since the data is not easily modeled with only the variables at hand.

Machine Learning Component	Mean Cross-Validation Score	Test RMSE
Component 1: Beta Values	-0.8040566023966509	0.4480831366870048
Component 2: Annualized Risk Values	-0.7623398741108754	0.4150921783592476

Discussion & Significance
The statistical analysis supports Claims #1 and #3, indicating significant differences in the means between the respective groups. However, Claim #2 was not supported. Future research could benefit from an increased sample size, the inclusion of control variables, a longitudinal study design, and the use of additional statistical tests where appropriate.