Forecasting Stock Market Dynamics in the Tech Sector: A Multi-Stock Gradient Boosting Approach*

Jamie Lee

November 21, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section **??**….

2 Data

2.1 Overview

The data for this paper, obtained from the Yahoo Finance API via the `tidyquant` by library in R on November 19, 2024, encompasses historical stock price information for four leading technology companies: Google (GOOG), Apple (AAPL), Amazon (AMZN), and Microsoft (MSFT). This dataset spans from January 1, 2018, to December 31, 2024, providing daily stock market data for each company. Key variables include **`symbol`**, **`date`**, **`open`**, **`high`**, **`low`**, **`close`**, **`volume`**, and **`adjusted`** prices. Our analysis focuses on predicting daily price differences

*Code and data are available at: https://github.com/jamiejiminlee/Tech-Stock-Forecast.git

and percentage changes across this basket of stocks, leveraging additional features constructed from the raw dataset. These features include `Price_Lag1` (lagged closing prices), `Price_Diff` (daily price differences), and `Price_Change_Percent` (percentage changes). By incorporating these derived variables, we aim to capture sequential trends and normalize differences across stocks with varying price ranges.

To ensure reliability, the data was pre-processed to remove missing values, align time-frames across stocks, and compute the constructed variables. Observations with incomplete data were excluded to maintain consistency in analysis. Data cleaning and analysis were conducted using the following packages: the `tidyverse` package of (**Tidyverse?**), `lubridate` package of (**Lubridate?**), and `arrow` package of (**arrow?**). Details on the data cleaning and manipulation process are provided in **?@sec-data-cleaning**.

2.2 Measurement

Stock prices are influenced by numerous factors, including macroeconomic conditions, industry trends, corporate actions, and investor sentiment. These dynamics fluctuate frequently, often in response to earnings reports, geopolitical events, or shifts in market expectations. Financial data simplifies these complex market behaviors into numerical indicators such as `open`, `high`, `low`, and `close` prices, which, while precise, may not fully encapsulate underlying investor motivations or market sentiment. For instance, the closing price represents only the last traded price of the day, smoothing out intraday volatility and obscuring more nuanced trading patterns.

Constructed variables like `Price_Diff` and `Price_Change_Percent` aim to quantify daily stock price movements by measuring differences or proportional changes relative to the previous day. However, these metrics assume that price movements reflect consistent investor sentiment across all stocks, potentially overlooking idiosyncratic factors specific to individual companies, such as leadership changes or product launches. Additionally, these metrics equally weight all days, despite certain trading days (e.g., earnings announcements or market holidays) being more influential.

The `volume` variable, representing the total shares traded, is used as a proxy for market activity and investor interest. This assumes that higher trading volumes uniformly indicate significant market events, despite possible differences in trading motives, such as speculative activity or algorithmic trading. Similarly, the temporal predictors, including `date`, `day of the week`, and `month`, aim to capture seasonal trends or trading patterns. However, these features simplify complex market behaviors into discrete time intervals, potentially overlooking shorter-term dynamics like market reactions to breaking news or macroeconomic data releases.

By aggregating these variables across a basket of stocks, the model assumes comparability among them, despite differences in market capitalization, industry focus, and investor bases. This simplification may obscure unique drivers of price changes for each stock, potentially limiting the model's ability to capture more granular patterns. Nonetheless, these variables