

P-STOCKS

Stock Price Prediction

Project Manager: Arjun Govind, Pomona '25, Majoring in Computer Science

Preferred Contact: avga2021@mymail.pomona.edu



Project Background

The stock market is an exciting place for investors to use their knowledge to buy and sell stocks. Both investors from large financial institutions and small individual investors are actively challenged to “beat the market”. There are many technical metrics to gauge the health and potential growth of a company but at the end of the day a stock's price is simply determined by the supply and demand for its shares in the market.

This project will aim to predict the demand for a stock using Google Trends. Google Trends is a website which collects data on the amount of times a term is searched. By looking at this data we can see the volume of searches related to a stock at any given time. We will hopefully be able to draw on correlations between search volume and price fluctuations. This would enable us to determine when it might be best to buy or sell a stock given the volume of searches related to it.

Dataset

There is an abundance of data available on the stock market on the web. The challenge is pulling that data into a usable data set. There are a variety of tools and scrapers on the web to help with this. Using **yfinance**, a python module, we will be able to pull basic data on stocks such as opening and closing prices and begin creating our csv file. We may also choose to utilize existing tools on the web to create our data set. We will then use **pytrends**, a python module to pull data from Google Trends and add that to our existing dataset. Our completed data set will have the opening and closing prices of a number of stocks as well as the weekly, monthly and yearly changes in that stocks price. It will also include the search volume of each

stock name over each time interval. And a list of other technical indicators that we will use to compare price correlations with.

Analysis

With this data set we will use some of the supervised classification and regression algorithms found in the **sklearn** library to create our predictive models. These may include Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbor and Neural Network. The goal is that these models will be able to determine how much a stock price may change based on the volume of searches.

After running our algorithms we will work to optimize the models using gridsearch on the model's parameters. We may also explore pre-processing techniques such as standardization or normalization to improve the accuracy of our models.

We will run these models with search volume as our input as well as the variety of other technical indicators we compiled into our database and use feature importance to determine which have the strongest influence over stock price.

Goal

The goal of the project is to produce a successful predictive model that is able to determine the amount a stock might change given the stock's volume of searches. Given how hard the stock market is to predict I am fully expecting our model to be far from 100 percent accuracy, but I am hoping that we will find some correlation between the number of times a stock is searched and the price change.

The outcome of this project can provide important implications to any individual interested in investing money in the stock market. If our project is successful we will have determined a valuable indicator to look at before buying or selling a stock.

An important aspect of this project is that we will be measuring the *absolute value* of change in a stock's price as opposed to how much it increases/decreases. This is because it's likely that a high volume of searches could mean a lot of people are about to buy a stock or bad news came out about that stock and a lot of people are getting ready to dump it. Using the absolute value will prevent our model from being very unsuccessful. If we do find a correlation in our model we may then try to also factor in the sentiment surrounding the stock so we can differentiate when the stock may increase or decrease based on the number of searches,

Packages: scikit-learn, pandas, requests, yfinance, pytrends

Planned Timeline

Introductions to: - Stock market and technical indicators	
---	--

- Google Trends and working with pytrends - Web scraping (beautiful soup, and requests) - Introduction to Github collaboration workspace, communication methods	September
- Choose the list of stocks we will use - Build initial csv dataset using the module yfinance - Use pytrends to pull data about search volume for all of our stocks and add them to csv file - Clean and finalize the data set.	September-October
- Begin running machine learning algorithms on data set - Optimize our models with grid search - Try pre-processing techniques on data to improve accuracy - Make any necessary changes to data set, change keywords, change time periods we analyze, change stocks we analyze.	October-November
- Evaluate each models accuracy - Possibly research and pull data that about stock sentiment and add that to our data set - Try running our model with other technical indicators and use feature importance to determine most valuable metrics - Try using our model to make predictions on current stocks - Summarize results and design presentation slides	November - December

Project Member requirements:

This project is open to anyone! Having a basic knowledge of Python will certainly help but we will go over all of the methods we will use. Existing knowledge of the stock market is a bonus but not a requirement. In general stocks are very complex, so having a strong desire to learn and problem solve will certainly be beneficial for this project.

This is a great introduction to real word applications of machine learning. The project will also teach its members how to scrape and parse through data from the web, clean and organize data, build a CSV file in python, visualize data and learn about the stock market!