Table of Contents

Introduction

A friend of yours comes to you with questions about the stock market. You tell them that now would be a terrible time to buy stocks. After all, in the midst of a global pandemic, surely stock prices would drop dramatically. You quickly google to confirm and see that contrary to your assessment, stock prices have surged in recent days, despite news of rising death totals and record unemployment. You tell your friend to speak with someone with a financial background.

The rises and falls of the stock market are seemingly impossible to predict. At the same time, there is a vast wealth of historical records to explore going back decades. Armed with data science techniques and a little computing power, would it be possible to make sense of stock market prices? Are there any trends that have repeated over the years? What features are most useful in predicting tomorrow's prices?

Datasets

1. https://www.kaggle.com/jacksoncrow/stock-market-dataset
   This dataset contains multiple decades of daily historical stock prices sourced from Yahoo Finance. There are over 5,000 stocks contained in the dataset, which has been updated to the current month.

2. Data will also be sourced from https://iextrading.com/ via the API provided by https://alpaca.markets/, a trading platform with many convenient features such as paper trading and backtesting. This data can be used to trade on the current day's market information.

3. If you have a verified Alpaca API account, which is currently available only to US citizens, you can use the Polygon API. It features more symbols, more data points by default, and more accurate historical data.

What You'll Need

- Python 3 Environment

- Data Science Python Libraries (Pandas, Numpy, Matplotlib, etc.)
- Technical Indicator Library (TA-Lib, ta)
- Alpaca API Paper Account (Verified Account is optional but recommended)
- My GitHub repository which contains the code necessary to run the pipeline.

Data Acquisition

Initially, I used a Kaggle Dataset which sourced its data from Yahoo Finance. It contained over 5000 stock symbols, an overwhelming number for testing purposes. Using Volume (a value representing the number of trades performed within a period) as a metric, I reduced this number to the top 500 stock symbols.

I encountered a number of issues with this dataset. 'Zero' values were a common sight, and for time series data these missing values are difficult to impute without introducing inaccuracies. The dataset was also static, requiring manual downloads to update to a newer version.

I then switched to the Alpaca Data API, which is available to all Alpaca accounts. Using the Alpaca Python API endpoints, you can download up to 1000 previous data points for 100 symbols per call. A timeframe value can be adjusted for daily or intraday results. While day trading is allowed on an Alpaca paper account, please note that there are special rules and limits for day trading when using real money.

The Data API historical data was more accurate, but after encountering some strange results, I noticed that errors were still present in the data. These can be mitigated by removing the offending values after normalization. If you are a US Citizen and don't mind signing up for a verified Alpaca account, there is a third option available to you.

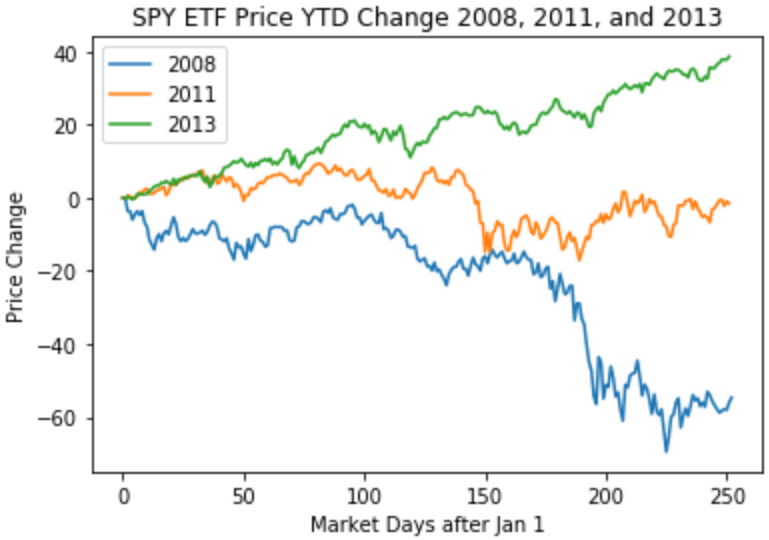
The Polygon API provides real time historical stock data and boasts higher accuracy. You can pull up to 3000 previous data points per call. However, you can only download one stock symbol per call, so be wary of reaching the API rate limit. In the included GitHub repository, I have separate notebooks for both the Data API and the Polygon API data acquisition. When running the pipeline, choose one or the other according to your account.

In addition to individual stock data, I also acquired information on the market performance as a whole. This can be done either by averaging values from all stocks, or more simply using data from an ETF (Exchange-Traded Fund). ETFs use multiple stocks to derive their value. The more popular ETFs include SPY, which tracks the S&P 500, and DIA, which tracks the Dow Jones Industrial Average. Utilizing ETFs for market data saves time and disk space.

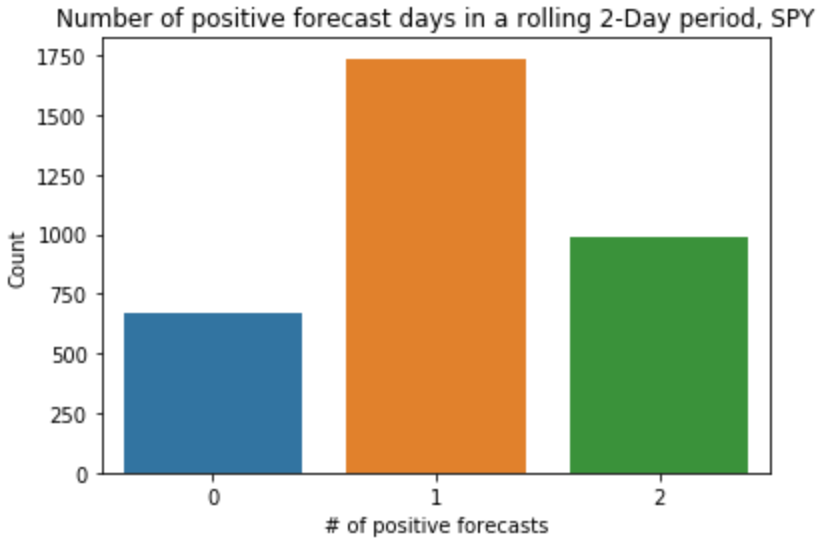
Exploratory Data Analysis:

What trends can be found with over a million data points? To begin, it's best to look at the market as a whole. The SPY ETF approximates the market using a collection of stocks. From 2007 to 2020, the market has doubled its value. However, there have been periods of growth and recession.

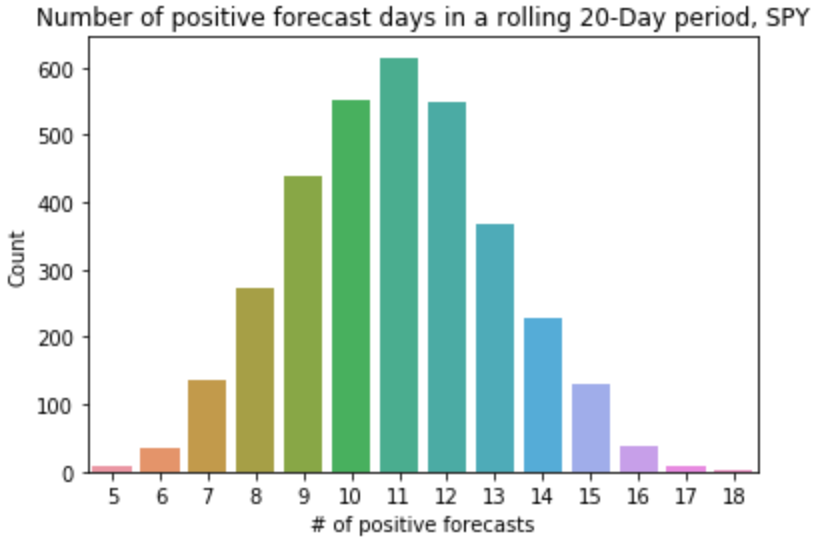
Let's take a closer look. The years 2008, 2011, and 2013 represent the different types of markets. 2008 shows a market in decline. In 2011 the recovery had yet to take full effect, and the price at the beginning and the end of year were nearly the same. 2013 was a year of consistent growth.



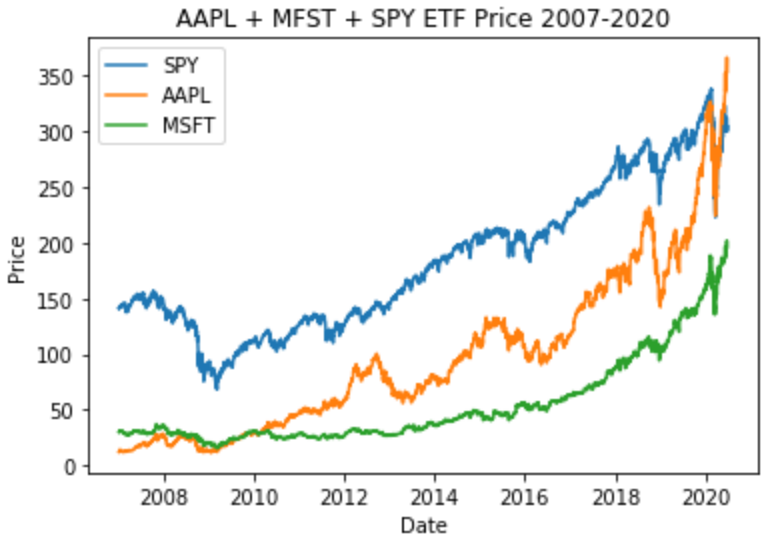
A naive trader might buy when the previous day's price change was positive, and sell when the previous day's price change was negative. However, when a rolling period of two days is sampled over the entire dataset, days where the price change trend reverses are more common than days that continue a trend. In the short term, the market is volatile.



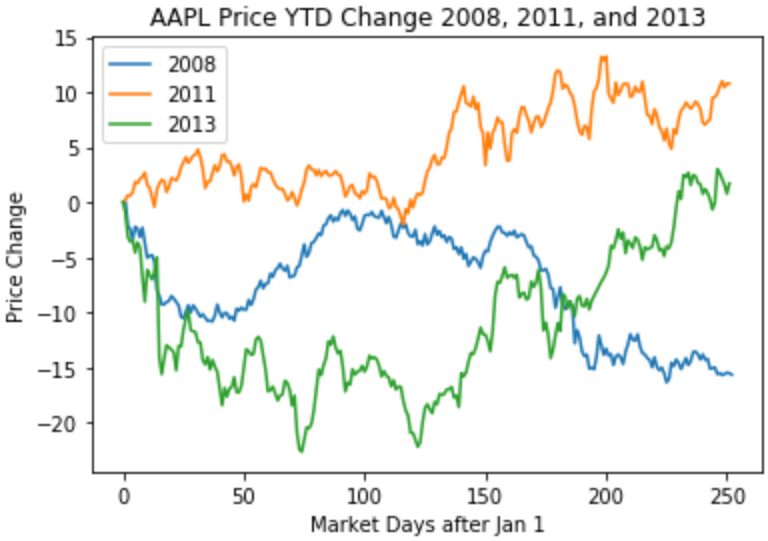
Over a rolling period of 20 days, the market shows an overall positive trend. The expected number of positive days is 11. The expected monetary return during a 20 day period is also positive. Going by these results, long term strategies have been safer for the average trader.



How closely do individual stocks match the market? To visualize this, I chose two well known stocks to plot alongside SPY, AAPL (Apple) and MSFT(Microsoft). Just by looking at the prices from 2007 to 2020, both stocks follow a similar pattern of growth.



Microsoft's prices in 2008, 2011, and 2013 matched the SPY prices. However, Apple actually performed worse in 2013 than it did in 2011. Slowing sales and profit margins triggered a decline in the stock price, despite the rest of the market performing well.



Individual stock price forecasts match the market forecasts only 63% of the time. The market is a driving force for stock prices, but individual events can override this effect. Certain industries do match the market more closely, however. For example, financial institutions make up five of the top ten stocks by market influence and match the market forecast over 75% of days.

Technical Indicators

Just the market data alone is not enough to predict future prices. To create a robust model, feature generation is required. To accomplish this, I will use technical indicators. A technical indicator is a fancy term for a function that modifies historical stock prices. This modification usually reveals a trend or a signal that can be used to make decisions when buying or selling stocks. For example, a simple Moving Average may be used to determine if a stock is trading above or below its average price for a specific time window.

Technical Indicators are not perfect. While they do provide additional insight into a stocks current strength, they have a few drawbacks. Firstly, they are reactionary. A moving average can only go up once the stock has already climbed a little. Most indicators have various levels of “lag” before they reveal any sort of trend. At best, you can catch onto a trend quickly, or be more sure of the strength of a trend.

Additionally, technical indicators cannot predict world events. Stocks are (usually) a reflection of the strength of the company they represent. If something happens that boosts the value of the company, it will (usually) be reflected in the stock’s value as well. Technical indicators cannot account for sudden changes in a stock’s future value.

There are many types of technical indicators. I used the well organized documentation of TA-Lib, a technical indicator library, to learn about many of the features used in my model.

Types of Technical Indicators

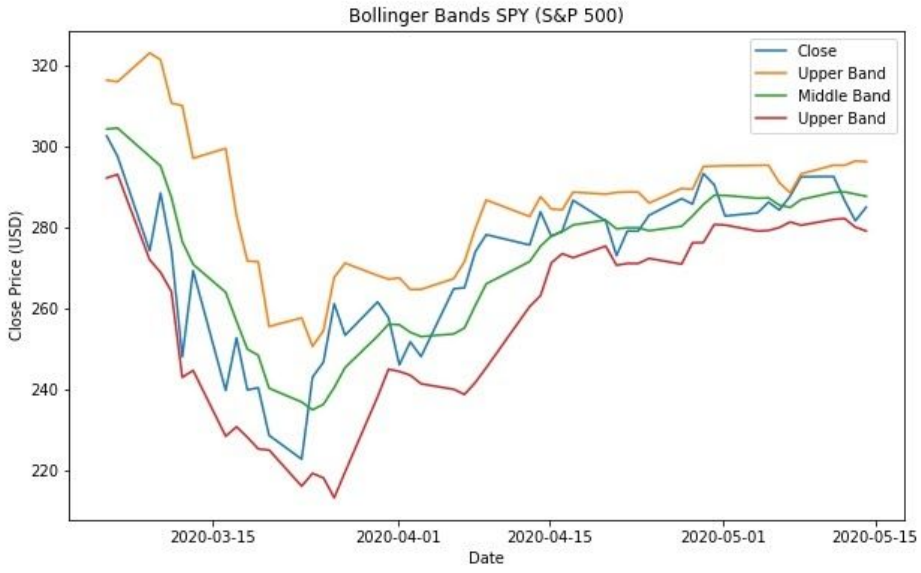
- Overlap Studies
- Momentum Indicators
- Volume Indicators
- Statistic Functions

Overlap Studies

Overlap studies use moving averages to measure stock performance. The term overlap refers to periods in time when the stock price "overlaps" with a particular moving average. This may indicate a time to buy or sell a stock depending on the direction of movement.

Bollinger Bands

The Bollinger Bands are likely the most well known set of technical indicators. They consist of three features. The upper and lower bands represent (by default) two standard deviations above and below the middle band. The middle band is a simple moving average of the stock price, but can be modified to use other types of moving averages. The width between the upper and lower bands may also be used as an indicator, as the distance between them expands when the price changes, and contracts when the price stays the same.



As you can see from the plot above, The closing price crosses over the middle Bollinger Band at multiple points. The width of the bands contracts as the price stabilizes. From just three lines, a lot of information is revealed about the stock's directional trend, as well as the strength of the trend.

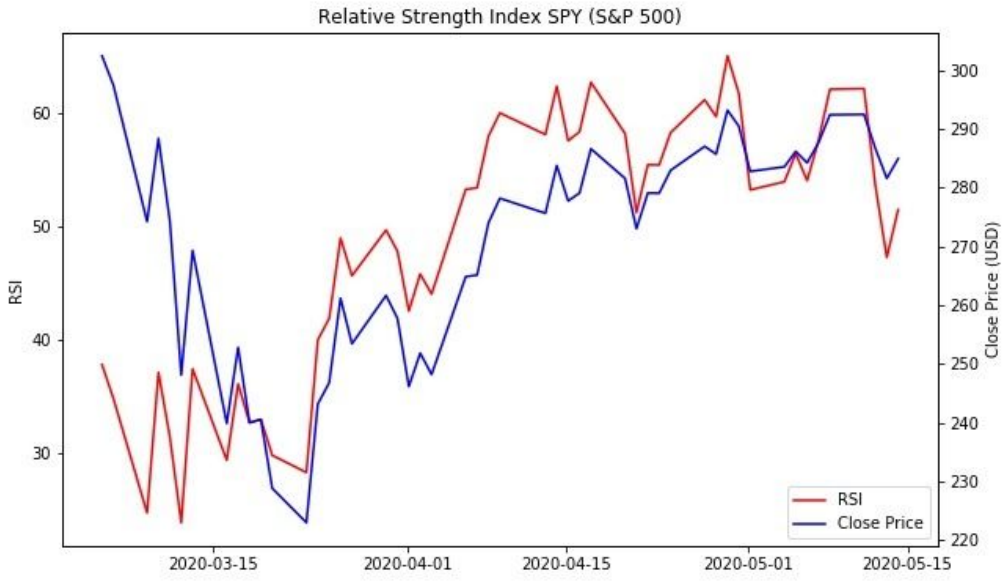
Momentum Indicators

Momentum Indicators look at changes in a stock's price to measure how fast a stock is rising or falling. Unlike overlap studies, many momentum indicators return a score within a specific range.

RSI

The Relative Strength Index, or RSI, is a value between 0 and 100 that determines the strength of a stock's positive trend. A value of 0 indicates that the stock has had no positive days in the chosen time period. A value of 100 indicates that the stock has only had positive days in the chosen time period.

On its own, the RSI has two weak points. First, the indicator has inherent lag. An RSI can only go up after a positive day, so you might miss the entirety of a quick upward trend. Second, the direction of the RSI matters. A value of 50 will not tell you if the previous day was 40 or 60. For a deep learning model which processes days one at a time, this information will be lost.



The RSI (in red) follows the close price closely, but there are some subtle differences. A value below 50 indicates a negative price trend, which holds steady at the beginning despite the

price dropping. The value only goes above 50 at the tail end of the stock's increase. While the RSI does miss trends on its own, it has value in understanding the current strength and momentum of a stock, which works well in combination with other technical indicators.

Volume Indicators

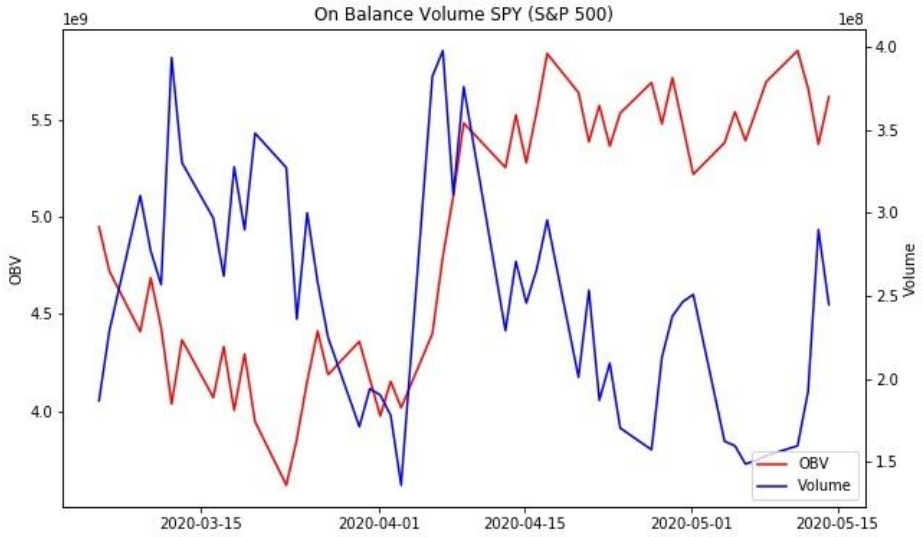
Volume is a measurement that tracks how often a stock is traded. The volume of a stock is the absolute number of trades that occurs on a given trading day for a particular stock. Both buys and sells count as trades, so a stock's volume may be considered "buyer-controlled" or "seller-controlled" depending on the price movement. Some indicators incorporate volume to predict price changes.

On Balance Volume

On Balance Volume (OBV) is a unique indicator. Unlike other indicators, it does not use a time period to calculate its value. Because the calculation is cumulative, the starting date determines the end result. Each value depends on the previous value. If the price goes up, the day's volume is added to the OBV. If the price goes down, the volume is instead subtracted from the OBV.

In theory, the On Balance Volume can show disconnects between a stock's price and volume. A stock which is increasing in price but has a sluggish OBV may indicate a drop in price in the near future. This makes the OBV a leading indicator, or an indicator that may show a change in price before it happens.

However, OBV on its own is not enough, and requires additional information to show a definitive trend. Because the indicator's value is cumulative, it is difficult to normalize to a value that makes sense across multiple stocks, limiting its use in machine learning. Finally, OBV is very sensitive to large spikes in volume, which affects the value long after the spike is over.


On Balance Volume SPY (S&P 500)
1e9
1e8
OBV
Volume
5.5
5.0
4.5
4.0
4.0
3.5
3.0
2.5
2.0
1.5
2020-03-15
2020-04-01
2020-04-15
2020-05-01
2020-05-15
OBV
Volume