Stock Price Prediction

Josiah Kek (jrk322) Sukriti Poddar (sp873) Yichen Shao (ys2233)

October 3, 2021

Question

How can we use a company's financial indicators to determine if its stock is worth buying?

Dataset

We will be using a Kaggle dataset (here) that tracks 200+ financial indicators (obtained from SEC 10-K filings) of 4000+ companies from 2014 to 2018. These indicators are multifaceted, and they span the range of company earnings, expenses and inventory. To obtain month-on-month stock performance from 2014 to 2018, we will scrape such data from the Nasdaq website (here) and join it to the Kaggle dataset.

Significance of Problem

Predicting stock prices is a notoriously difficult task, due to high levels of randomness and noise in data. However, even small advances in understanding the drivers of stock price can help our hedge fund to boost its revenue.

The financial indicators on a company's SEC 10-K filings provide a good snapshot of its financial health, which is a key driver of mid- to long-term changes in stock price. It would be useful to identify which financial indicators are most correlated to stock price increases (or decreases), with the aim of using the indicators to predict stock performance over a 5-year period. One advantage of relying on SEC 10-K filings is that this is publicly available data that our hedge fund does not need to pay to obtain.

Relevance of Dataset

Each entry of the Dataset has two possible labels: class and percent price variation. Class indicates whether the trader should buy the stock or not. And percent price variation indicates positive or negative price variation for each stock during the whole year. Thus, we could use the dataset for both classification and regression predictions.

The dataset also has 223 features excluding two possible labels. Those features cover a wide range of the information about each company (liek revenue, cash flow, tax liability, etc). Thus it allows us to narrow down to the specific features that are most relevant to stock price. Moreover, since we have datasets from 2014 to 2018, it would be sufficient enough for us to validate and backtest our models.