

The Proposal

A problem that I am interested in solving is being able to predict stock market prices based on historical data. Often, novice investors do not make well-informed financial decisions, and it is important to be able to understand market trends over time to have a positive return on any money invested. Thus, this project attempts to analyze stock market data over many months and years and predict future prices. A proposed method of solution is to use a combination of Linear Regression, Random Forest, Support Vector, and Neural Networks. Then, after generating each of these models, I will identify which one is the most accurate and then fine-tune it. Some potential stakeholders include students or beginner investors who want to learn more about how the stock market works and make low-risk trades. On a larger scale, a company could potentially use a similar model to understand how their stock price is changing over time and they could compare it to other internal revenue and expenditure data that they may have. One major obstacle is simply the sheer amount of data that I will be working with since I am not sure if my machine will be able to handle it. Furthermore, since I am working with dates, I am not sure how the model will compute “x-values” when I split the data into training and testing sets. One thing that is novel about this problem approach is that I plan to do something called feature engineering which is a concept in machine learning where the developer transforms existing data into additional relevant information to help the model’s accuracy. Rather than just using the data that I have, I plan on creating some additional features like moving averages, volume averages, and more (metrics which are often used in the financial world for understand the position of a security). Furthermore, I am not making this model specific to a single stock; the algorithm will be dynamic. Based on the stock ticker that the user enters, the model and algorithm will be different to generate an accurate prediction.

The Plan

I currently plan to pull stock data using the open-source tool yfinance which uses Yahoo Finance’s publicly available APIs. I will pull a set amount of data and convert it into a local CSV sheet so that I can create a DataFrame and work with it in the code instead of having to constantly pull data. If I supply my own dataset, it will be stock investments that I myself have made and I will graphically compare my purchase dates and prices with the price prediction for that day to see if the model is working correctly. Since this is data that I own (I have a personal Excel spreadsheet with my own purchase history), I give myself consent to use this data for this project. Since my data will already be organized by date, I don’t necessarily need to organize it; I just need to reassign the index (dates) to its own column. I plan to analyze the data by plotting various factors like price, volume, and percent change per day over time to understand basic trends in the stock that I am analyzing. Based on my problem statement and some research that I did, the best model for this data is a Tensorflow Neural Network called LSTM (Long Short-Term Memory) which is very useful to process data over multiple time periods/time stamps as well as calculate short-term and long-term fluctuations. Like I mentioned above in the proposal section, after an extensive data preprocessing/cleaning process as well as some exploratory data analysis, I will attempt to build multiple models and determine which one fits best for this data. I will determine if my model is accurate by first graphing the predicted month of September with the

actual month of September stock prices. The overall accuracy, f1 score, and other factors will also be helpful in identify false positives and wrong predictions. Then, I will have the model predict a certain date or date range and use my own price purchases to compare against the predicted price on that date or date range. There are certainly a lot of peer-reviewed journal articles which use similar approaches for stock price prediction. There are also some websites that attempt to do the same, however, their credibility is questionable. Therefore, to compare to a baseline, I will compare my results with any peer-reviewed, “professional” models that I can find online. Here is one example:

<https://www.sciencedirect.com/science/article/pii/S1877050920304865>