DATA SCIENCE SENIOR CAPSTONE, BY JALIN ROBERTS

Hedge Fund Stock Market Analysis, Modeling, and Portfolio Forecasting

JANUARY 27, 2024

BELLARMINE UNIVERSITY
DS-450-01

Executive Summary

This personal project seeks to provide an alternative solution to problem presented from utilizing an automatic trade-copying app known as Autopilot. This app copies historical trading data that is made public as the SEC filings are released. However, because of the 45-60 day lag that occurs when this information becomes public, that leaves potential money on the table, and less accurate investments. I will build a model utilizing 10 stocks from Citadel LLC, and build a training model upon 4 years worth of daily trading data from Yahoo Finance. These 4 years of trading data will be used to forecast 2023 daily price predications, and will be compared to actual 2023 trades to test accuracy. If the ARIMA model predictions for 2023 are within 5% of the actual value for 2023, I will see if it is possible to tighten the gap to shrink it to be as close as possible.

Project Idea

The idea of this project is to slim down an existing idea into a more specific and time-friendly model prediction. Overall, I would like to attempt to accurately (within 5% +/-) predict daily stock prices for companies that are held by the hedge fund Citadel LLC. The overarching idea is that United States' hedge funds seems to consistently and accurately show high return on their portfolio and investments. Because of the sheer amount of daily and quarterly trades taking place via Citadel's investors, I want to analyze stocks held by Citadel, rather than copying their trades as they become public record.

Analyzing Citadel's investment history over the last 5 years provides a reasonable length of time to make specific and random decisions to track stock performance. For balance, I would like to take the top 5 holdings within Citadel's portfolio, and 5 random holdings within Citadel's portfolio and build a model predicting daily stock price. I want to take these 10 holdings, and build a model that aids in creating a visualization to build a foundation of personal investments. I want to take the trade data for these 10 stocks, and train a model using 4 years worth of trades, and compare that to 2023, and analyze how accurately my model predicted daily price change. If there is extra time, I would like to further compare this 2023 testing data using an LSTM and see which methodology is more accurate. The final question this project is seeking to answer is "Is using an LSTM always necessary, and can prediction be done more accurately with simpler methodology?"

Background

A consistently popular opinion within investment enthusiasts is the idea of trying to "beat the system." More specifically, investors point to top companies like Citadel LLC and wonder how hedge funds are able to consistently turn profits year after year; And is there a method that the average citizen can adopt to also feel like they are making smarter investments? During the pandemic, an app called Autopilot was created that sought to beat the system using automatic trading. What this app does is copy prolific investor's trading history such as Citadel, Nancy Pelosi, and other politicians as the publicly available 13F filings are published. The idea is that these portfolios that Autopilot creates to mirror the company or individual you can choose from gives you the leverage to have a portfolio in which you have an advantage over the average investor.

However, there exists a large flaw in this app and idea that does not have a perfect solution for. Autopilot can only copy trades after the SEC publishes the 13F filings, and this can happen as late as 45-60 days after trades have taken place. By the time this app copies this trade, whatever inside information you suspect has taken place for the sake of profit may have already passed the window of opportunity. There is no real way to bridge that gap besides going into the trade history itself and trying to predict daily price changes using individual stock picks. What I aim to do is instead of copying Citadel's trades, I want to pick their trades (5 random, and 5 top holdings) and build a model to predict daily price changes for 2023. I will use daily trade history utilizing Yahoo Finance to build and visualize my findings.

The original dataset that will be used will be Excel (.csv) files that are the publicly available 13F filings for Citadel LLC's stock trades for 2022. I will have four (4) files containing each quarter's investments. I will combine all 4 into one large file. I will then analyze the data to find the top 5 holdings, and randomly select 5 additional holdings. These 10 stocks will be what I build my model upon. I will use the yfinance library within Python to grab the relevant data and begin predicting daily close price using tools in Sci-Kit Learn. I will then utilize K-Nearest Neighbor for a preliminary model, which will be my presentation, I also use decision trees and support vector machines in order more accurately analyze and predict my models. If there is additional time in the semester, I plan to take these 10 stocks and run them through an LSTM using code from a relevant Kaggle project and review the similarities and differences. The goal there would be to analyze the differences in accuracy between the LSTM they are using, and the models I am utilizing to see which is more accurate. This is dependent on time available after the original model is built and is subject to change and possibly be left out.

Modeling

For my predictive modeling, I will utilize three predictive analytics tools. I will use K-Nearest Neighbor to make predictions about the grouping of my stocks. I will also be utilizing a Decision Tree to assess the possible outcomes of my stock choices. A decision tree will be helpful as 5 of my 10 stocks will be chosen at random. Lastly in my arsenal of modeling tools, I am going to also include an ARIMA to forecast stock prices for 2023. ARIMA is an important model here because I will be dealing primarily with Time Series data. All these predictive modeling techniques and tools are to see if it possible to perform more accurately than an LSTM. It has been argued that utilizing an LSTM is the most effective predictive analytics modeling tool to forecast stock prices. The influence for comparing my work to an LSTM came from a similar project on Kaggle which utilized an LSTM instead of what I am aiming to use.

Tools

To begin this project, I will utilize Microsoft Excel to organize my quarterly data. I will then go to Yahoo and utilize Yahoo Finance to download the relevant stock data for my testing. In my testing process, I will use Python in VS Code, utilizing libraries within Python such as Pandas, NumPy, Sci-Kit Learn, and yfinance. In my model building, I will utilize three machine learning modeling tools: K-Nearest Neighbor, Decision Tree, and ARIMA. After I have built my model and sufficiently tested it; depending on the remaining time I have available, I will utilize a Long Short Term Memory (LSTM) model to test my model against. For the visualization aspect of this project, I will also need the matplotlib.pyplot and seaborn Python libraries. Tableau is also going to be a helpful visualization tool for presentation purposes. Along the way there could be the additional of more tools and libraries for which I will utilize, but as of now everything has been accounted for.

Conclusion

A main motivation for this project is to compare machine learning models and learn more about the decision-making process as it relates to financial investing. Using a precursor like a hedge fund gives an interesting context for which I can analyze, interpret, and build functional and realistic predictions. This is an exciting opportunity with a unique twist to existing tools and suffices as an attempt to bridge the gap for a niche portion of investing. These tools, the relevant data, and the processes required to better understand both allows me the creative opportunity to learn and teach at the same time. The primary focus of this work is to have real-world data and exercise the critical steps in the Machine Learning process to analyze, process, forecast, and deploy my model(s).

References

"13F Info – SEC 13F Filings." *13F Info – SEC 13F Filings*, 13f.info/.

Mohan, Darshini. "Unleashing Data Science for Stock Market Trading: Marketing Analytics Companies: Digital Analytics." LatentView Analytics, 23 Nov. 2023, www.latentview.com/blog/unleashing-data-science-for-stock-market-trading/#:~:text=It%20is%20among%20

https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6