



DATA SCIENCE SENIOR CAPSTONE, BY JALIN ROBERTS

Predictive modeling for stock analysis and advanced financial portfolio
forecasting



JANUARY 15, 2024
BELLARMINE UNIVERSITY
DS-450-01

Executive Summary

My proposal for my Data Science Senior Capstone is that which encompasses an important aspect of financial analysis and decision making. That aspect is the ability to compare previous data utilized by professionals with many years of relevant knowledge and skillsets and use that information to make better informed stock investments. Through predictive analysis, I will use exploratory data analysis to find stock trades made by Citadel Advisors LLC for 2022 and run a regression model to get a prediction for which to invest in 2023. I will then compare between the two to find if building a predictive analytics model using Citadel's 2022 trades can yield an investment portfolio for 2023 that yields higher than 10% positive return by year-end. 10% will be the baseline for which gauges success, as most stock portfolios are based off the S&P 500, which yearly shows returns between 7%-10%.

The predictive analytics model by which we will train and test will encompass 13F-HR form filings found through the SEC. I will also utilize data from other prolific investment companies such as Black Rock, State Street, and Vanguard in my 2023 portfolio to find the portfolio that yields the highest potential returns. Citadel's 2022 trades will be the initial regression used for evaluating what other relevant trade information will be added or left out in my model. The project will begin with using Citadel's trade information as a base, and other investing company's 2022 trades will be added to increase performance of my portfolio prediction for 2023. The output will include percentage of portfolio investment recommendation, and how many stocks can be purchased relative to the amount desired to be invested.

Project Idea

The idea for this project stems from a myriad of similar ideas and opinions from which I have encountered throughout my twenties. A consistent criticism that I have encountered of banks, hedge funds, and politicians is that they somehow always come out on top financially. It is often concluded that it is possible that hedge funds, prolific investors, and/or politicians could have access to personnel and information for which gives them an unfair, and

potentially illegal advantage. Knowing these complaints lead to this very project idea – Is it possible to outperform the S&P 500 by building an investment portfolio using trading information from one of the most successful hedge funds in the United States?

This question actually leads into a follow-up question that fits perfectly for this project: If I'm going to use a single hedge fund's 13F filings, why not use several hedge fund's investments for 2022 and build a portfolio from all of the information gathered? This follow-up question is actually where the overall idea for this project is centered. I do not have hedge fund or investor levels of money, but what if I can invest like a professional using information I gather and models I build? Not only does this method teach how to trade like a professional, but I can build a portfolio projection for myself and mold it to a reasonable amount of money for a university student to invest. This technical analysis will be the basis for which I can forecast the price movement of trade data to identify potential short-term and long-term investing opportunities. Building a model that only looks at opening and closing prices is not sufficient enough for what I am trying to build, and so I will analyze percent changes of stocks quarterly to have a better understanding of forecasting and predicting.

Background

The idea of using existing information to predict portfolio suggestions is nothing new in the investing sector, or even the greater financial sector. The idea of this project is the old saying "history tends to repeat itself." However, this project at face value is nothing new. At my previous job, I worked for the United States Postal Service as a Rural Mail Carrier where I spoke to members of my local community almost every single day of the week. I heard their criticisms of congress, the financial state of the country, economic performance, and how it just seems as though some people always come out on top. I had a lot of free time on my mail routes in order to ponder their complaints, and that gave the idea of another old saying: "If you can't beat them, join them." That saying gave the idea to this very project and the problem I would like to explore and even potentially create an interesting solution.

Starting off, I had to ask the obvious question of "where exactly do I even find this data?" I was not even sure it was public information. However, after some initial searches, I found the key that would unlock this project's potential. Every company is required to publicly file their trades via the U.S. Securities and Exchange Commission in form known as a 13F. This form filing gives a description of the class (such as common stock and put/call options), and the relevant information such as price, quantity, option type, and names. Form 13F is required to be

filed within 45 days of the end of a calendar quarter. Even though there is a delay of getting this information, the form 13F filings contain all the information I will be utilizing for my project. These filings are the closest information I can obtain outside of working for the respective companies and making the investments for them.

That being said, solutions already exist in many respective forms, but they seem to share a major roadblock which is being hidden behind a subscription, or a payment in general. A specific example is an iPhone app called “Autopilot” which encompasses the idea of hands-off trading for \$10 per month. The substance and allure of this app is that it is an auto-trading application that allows users to “auto-pilot” their trades by copying politician’s trades, and even companies such as citadel, and even CEOs. However, even though the idea for this app is rooted sound logic, it still has room for improvement, and that is where I want this Senior Capstone to bridge the gap. Utilizing Citadel as an initial regression, I will use other existing 13F filings and add their trades to predict model compositions and performance outcomes in order to minimize investment risk. By utilizing this strategy, I can have stocks that are risky investments yielding potentially higher returns, but also have the option of weighing a portfolio heavier in favor of more conservative long-term options to balance out of the risk of volatile stock price fluctuation.

The existing data that will be used in my Capstone Project will come from a website titled 13f.info, a website in which 13F filings are readily available and downloadable in CSV format. The data for this project is already fairly clean, and will not need extensive amounts of data analysis and cleaning. However, though this data is relatively clean, it can be improved and better integrated for modeling purposes. This will be further explored once a thorough analysis of all the data has been performed.

Modeling

Modeling this project is extremely important considering the vast amount of trade data available and needed in order to make an appropriate model. I am going to use linear regression as a baseline for evaluation, however I will utilize a mix of statistical, rule based, and machine learning models for prediction and forecasting. For my statistical rule-based modeling, I will use random forests for my predictions. I will then utilize Q-learning in order to determine a better course of action. I will then utilize a machine learning model known as a Long short-term memory (LSTM) network to forecast predictions.

Tools

The tools I plan to use are VS Code and Microsoft Excel. Within VS Code, I will utilize Jupyter Notebook and Python for programming purposes. I will then utilize Python libraries such as Pandas for data analysis, Scikit-Learn for machine learning, and NumPy for mathematical operations. I will also utilize the Keras and PyTorch libraries in my forecasting data with LSTM. I am going to use VS Code because of the UI simplicity and quick availability and integration of Python libraries. I am also going to use the relevant machine learning libraries under the suggestion of my professor Dr. Robert Kelley. At the end of my project, I would like to visualize my data and findings through Tableau. Any additional programs, libraries, or relevant tool(s) may come along the building pipeline as problems occur and remedies are needed. However, what tool(s) are currently listed should suffice my needs in my project's current state.

Conclusion

The end-state of this project is to build a model in which gives the end-user an ability to both learn about stock analysis, and make better informed investment decisions. Through the use of machine learning, I will be able to build a model that changes its suggestions and forecasting based on quarterly performance, thus allowing for robust decision making and updated accurate predictions. This methodology also allows for the possibility of short term and long term investment strategies. This proposal is to encapsulate resources and materials that are free to use, and encompassing publicly available SEC filings. If the relevant companies included in this project have some kind of advantage over an individual investor, I aim to personally test and evaluate if I can teach myself to invest like a true professional utilizing advanced artificial intelligence tools and techniques. I will then visualize my findings and thoroughly explain the process, complications, any necessary and/or relevant changes to inquiring parties.

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](#)

