Big Data - Final Project

DataFinance Analytics

Alvin Isaac

Gerald Fattah

Max Savasta

Saif Alzaabi

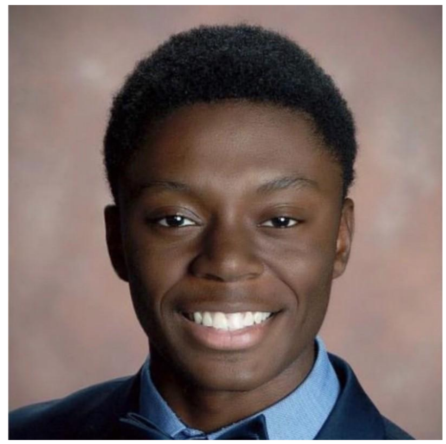
CSCI 4907- Introduction to Big Data and Analytics

Dr. Roozbeh Haghnazar

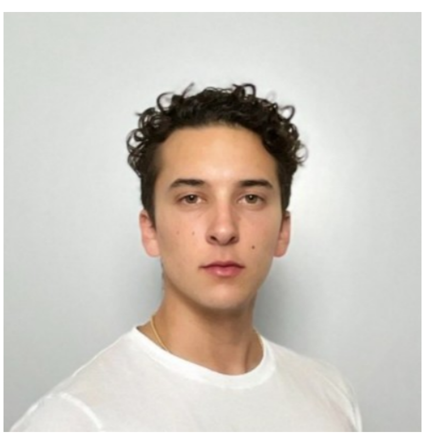
December 15, 2023

Meet the Team!

Gerald Fattah



Alvin Isaac



Max Savasta

Saif Alzaabi

Table of Contents

Abstract

Big Data Project Abstract DataFinance Insights In an era where financial markets are influenced by an ever-expanding volume of data, this project aims to harness the power of big data analytics to gain deeper insights into stock valuation and market sentiment. The primary idea revolves around the collection and storage of both quantitative and qualitative financial data. Quantitative data, readily available through financial institution APIs, forms the foundation of this project. Additionally, qualitative data, including news articles and sentiment analysis from social media platforms like Twitter, will be incorporated as an added complexity and storage challenge. The project's data sources encompass a range of financial information. Quantitative data is obtained from established financial institution APIs, offering comprehensive insights into historical stock prices, trading volumes, and financial metrics. Qualitative data involves web scraping of news articles and social media content, followed by natural language processing (NLP) techniques to gauge market sentiment. These diverse data sources provide a holistic view of the financial landscape. The group plans to use primarily AWS tools to build this pipeline. The challenges in this project encompass efficiently handling and storing large volumes of financial data, ensuring data quality and consistency across diverse sources, developing accurate natural language processing (NLP) models for sentiment analysis from news articles and social media, selecting appropriate machine learning algorithms for stock valuation, and ensuring model interpretability for transparent and trustworthy indicators in the dynamic financial market landscape.

1. Introduction

In the world of Finance, big data, analytics, AI, and ML have become infinitely powerful tools. These tools have revolutionized areas such as modeling of stock prices, portfolio creation and optimization, correlation between assets, market dynamics, high-frequency trading (HFT), algorithmic trading, and more. Proving its worth, hedge funds incorporating AI have outperformed the industry average in returns and lower volatility from 2016 to 2019. The speed of finance is also increasing due to these tools and doesn't show signs of slowing. Automated trading generates much of the volume in the markets, accounting for 50-70% in the Equity markets, 60% in futures markets, and 50% in the treasury markets. These systems work in milliseconds when it comes to receiving new information and making a consequent decision.

The goal of DataFinance Insights is to join this movement and create a near real-time pipeline that will seamlessly update and incorporate new data, as well as efficiently store historical data. The accessibility to large datasets is paramount. With multiple streams of emerging data, this pipeline can uncover new trends and correlations in real time that will provide users with insights into the market and possibly new trading strategies. With a rich set of historical data, this pipeline can also backtest the developed algorithms to ensure their accuracy, as well as uncover historical trends that may apply today.

This project aims to harness the power of big data analytics to gain deeper insights into stock valuation and market sentiment. The primary idea revolves around the collection and storage of both quantitative and qualitative financial data. Quantitative data, readily available through financial institution APIs, forms the foundation of this project. Additionally, qualitative data, including news articles and sentiment analysis from social media platforms, will be incorporated as an added complexity and storage challenge.

The project's data sources encompass a range of financial information. Quantitative data is obtained from established financial institution APIs, offering comprehensive insights into historical stock prices, trading volumes, and financial metrics. Qualitative data involves web scraping of news articles and social media content, followed by natural language processing (NLP) techniques to gauge market sentiment. These diverse data sources provide a holistic view of the financial landscape. The group plans to use primarily AWS tools to build this pipeline.

2. Approaches - Literature Review

Ranco et al in 2015 - Social Media in Financial Markets:

The presented literature provides a detailed exploration of the intersection between social media, particularly Twitter, and financial markets, shedding light on the intricate relationship between online sentiment and stock prices. The initial studies discussed in this article provide a foundational understanding of the relationship between web-derived data and financial markets. These works paved the way for subsequent research efforts, illustrating the potential of online activities in predicting market behavior. The authors highlight various methods employed, including analyzing web news, search engine queries, and social media, indicating the evolving significance of social media platforms like Twitter. The authors uncover a strong correlation between Twitter sentiment and stock returns. The significance of this finding lies not only in the confirmation of expected events, such as earnings announcements, but also in the revelation that unexpected peaks in Twitter activity influence market behavior. This novel insight expands the understanding of how social media sentiment impacts financial markets, pointing towards a need for more nuanced analyses.

Kolasani et al in 2020 - Social Media and Stock Market Predictions:

This research paper delves into the intersection of social media, particularly Twitter, and stock market prediction. The study addresses the significance of external factors, including social media and financial news, in influencing stock price movements. Social media platforms, such as Twitter, are explored as valuable resources for precise market predictions, emphasizing the need for automated analysis systems due to the vast amount of data generated daily. Additionally, the paper highlights the challenges posed by market volatility and the pivotal role of external factors, especially social media. Twitter, with its vast user base, emerges as a critical platform for understanding market sentiments. The study focuses on leveraging machine learning models, specifically Support Vector Machines (SVM) and Neural Networks, to predict stock prices based on Twitter data. The authors reference prior studies on sentiment analysis, emphasizing the relevance of social media data in predicting stock trends. Notable research efforts in sentiment analysis, including the use of Convolutional Neural Networks (CNN), are discussed. The review identifies the need to bridge the gap between social media data and historical data, emphasizing the role of human behavior reflected in social media interactions. The authors pinpoint the limitations of previous models and propose improvements through their research. This paper significantly contributes to the field by integrating sentiment analysis of Twitter data with advanced machine learning models. By establishing the superiority of neural networks over traditional models in predicting stock movements, the study advances the understanding of social media's impact on financial markets.

Hongxing He et al in 2006 - Algorithmic Predictions of Stock Market Trends:

This paper focuses on a data mining process for analyzing and predicting stock trends. The three major components of this process were partitioning, analysis, and prediction. A k-means clustering algorithm is utilized to partition stock price time series data. After this they used linear regression to analyze the trend within each cluster and the results of the regression are used to predict trends in sectioned time series data. Overall the process aims to predict future trends in stock prices accurately and efficiently. The proposed trading strategy is called TTP (Trading based on Trend Prediction). Another big focus in the paper is another trading strategy, “Naive Trading” (NT). Their results are reported for stock trading in select countries during the test period of 1999-2000. The results of the different strategies (NT, TTP, GP (Genetic Programming), and other traditional trading methods were compared. In this comparison there was evidence for the effectiveness of the proposed trading strategy (TTP).

Olaniyi et al in 2011 - Regressions to Predict Stock Prices:

This paper discusses regression analysis for predicting the stock prices. It uses the banking sector in the Nigerian economy to conduct this study. It focuses on the methods of data mining for valuable information and patterns from big datasets. This data was being used to generate stock price predictions. Data was collected from activity summaries put out by the Nigerian Stock exchange. Linear regression was utilized to predict the stock prices. One of the ways they separated these regression equations was by bank. They analyzed a linear relationship between the current market price and P.E ratios. A database of all the stock price data was also generated that was analyzed to identify patterns and trends. The paper highlights the importance and strength in using these techniques.

Robinson et al in 2011 - News and its Consequences:

Companies worldwide have bolstered their environmental commitments to boost their competitiveness and enhance overall performance. Existing research on businesses and news reveals that short-term fluctuations in stock prices are often influenced by news updates. Nevertheless, numerous eco-conscious firms invest in industries where the benefits may materialize over a more extended period. Consequently, it is reasonable to assume that the stock prices of these companies might exhibit lower sensitivity to daily news reports. A study conducted using a database of green firms in emerging markets reveals that news can indeed affect the daily returns of environmentally focused companies. However, the impact of this news appears to be transient and is not consistently observed across the majority of the firms examined.

Birz et al in 2011 - Value of an Undervalued Metric: Macroeconomics:

This article addresses a gap in the research on the effects of macroeconomic data releases and stock market returns. Birz (2011) states that previous research did not find significant correlation between the effects of changes in macroeconomic variables and the stock market due to the methods of measurement of investor sentiment. This article uses newspaper headlines as a way to take into account the economic environment. For example, a 6% unemployment rate would result in different effects based on the state of the market at the time, if its in a recession or a boom. The newspaper headlines can be positive, negative, or neutral when reporting GDP or unemployment and that is measured and compared to the resultant stock market prices. The research has found a strong correlation between the newspaper headlines sentiment and the effects on stock market returns. What we can learn from this is that it is important to find a consistent way of measuring the sentiment of investors/news while taking into account the state of the market when exploring the effects on stock market returns.

Bomfim et al in 2003 - Monetary Impact on Market Prices:

This article examines the effects of monetary policy announcements on stock market prices. The researcher places an emphasis on the prices before and after dates of the meetings as well as the element of surprise and pre-announcements. The study shows that stock market price changes are strongly correlated with meeting dates and monetary policy announcements. Financial economics research was not able to establish a strong correlation due to unaccounted for surprise effect of announcements and monetary theory economics research implicitly assumed that the nature of stock market returns is time invariant. Bonfim (2003) used the strengths from both areas of research and avoided the mistakes that were made to find that more surprising news had heavier consequences on the prices of stocks. What we can learn from this research is the importance of the attitude of the market before announcements are made. The more surprising and unexpected an announcement is, the more the effect is on the stock market.

Iyinoluwa et al. in 2019 - Data Mining Techniques to Predict the Market;

This research focuses on the challenging task of stock market prediction, a topic of immense interest due to its potential financial benefits. The unpredictable nature of the stock market necessitates advanced analytical methods for reliable predictions. This study introduces a novel approach combining Frequent Pattern growth, Fuzzy C-means clustering, and K-Nearest Neighbor algorithms to predict stock market trends accurately. The research addresses the need for deeper analysis beyond traditional methods, aiming to provide investors with valuable insights for making profitable decisions. The research distinguishes itself by proposing a method that not only identifies patterns but also evaluates them into actionable insights, enabling investors to strategize effectively. Their methodology involves transforming raw stock market data into interpretable historical data. From this, the Frequent Pattern Growth algorithm is used to identify frequent patterns, providing the foundation for a Fuzzy C-means clustering to match facts with frequent patterns. Finally, the K-Nearest Neighbor classifier, with k=1, is used for classification, with three distinct trend categories: static, uptrend, and downtrend. The researchers used financial data from multiple banks to test this model. A key highlight of the study is the comparison with a neural network model, a benchmark in the field. The evaluation results demonstrate the superiority of the proposed model over the neural network, establishing its effectiveness in predicting stock market trends.

Pricope in 2021 - Deep Reinforcement Learning in Quantitative Algorithmic Trading: Overview:

This paper gives an overview of the applications of deep reinforcement learning (DRL) methods to Finance, specifically in low-frequency quantitative algorithmic trading. The value of low-frequency (hourly to a few days) trading systems, as compared to high frequency trading (HFT), trading in pico- to milliseconds, is the accessibility to these systems. HFT research is generally safeguarded by institutions with the capital to fund it as well as the power of a machine working in hard-real time.

Proposed DRL systems can work in three ways, critic only RL, actor only RL or actor-critic RL. Any DRL system is given a state representation of its environment and a pool of actions that can be made, the next state is set based on the previous actions and the numerical rewards applied to it. The *critic* in these systems is the Q or action value function. This takes as input a state and possible actions and outputs the expected Q value – a probability distribution of events. Lastly, there is an *actor* who uses the probability distribution of Q and maximizes the policy for the outcomes. Since Quant Finance is not an uncharted field, the metrics to measure the goodness of a system are based on Sharpe Ratio, Sortino Ratio or annualized returns.

These systems run on quantitative financial data and usually have options such as buy, sell or hold, but can learn or be given more complex trading strategies as they expand. Many of the algorithmic trading algorithms correlate to other ones, minimizing the opportunity to profit. These are based on MACD, RSI, ADX, CCI, OBV, moving average and exponential moving average. It is also important to note that research supports that simple algorithms based on data mining and one indicator can achieve annual returns of up to 32%.

The paper then introduces many researched DRL methods and analyzes their outcomes. Overall conclusions are that low-frequency trading with realistic setups can obtain over 20% annual returns. Many programs are run on a daily scale which cannot predict social, political, IB moves or public sentiment (to avoid that interference use a smaller timescale). Real world edge cases (API issues, request throttling, market crashes) are a significant risk to these systems.

Zeng, Z, et. al 2021 – Deep Video Prediction for Time Series Forecasting:

This paper, by J.P. Morgan AI researchers, highlights a novel form of ML utilizing convolutional neural networks (CNN) – computer vision – to predict stock prices. The theory behind this is that humans rely a lot on graphical representations to predict stock movement, rather than just quantitative metrics. By turning time series data into a sequential set of images, CNN can use past images to predict how the next image will look, correlating to change in price. Many tests were provided, but the best performing was a 3x3 heatmap of 9 stocks/indices, arranged by asset correlation. Using data from June 29, 2010 through December 31, 2019 with a 95% training/validation split, they achieved high levels of accuracy. The results were astonishing as they achieved 65%-75% accuracy dependent on lambda. This outperformed ARIMA modeling which had an accuracy of 59%-65%, dependent on lambda. We hope to be able to implement this novel form of analysis within our system.

Cao, L 2022 - AI in Finance: Challenges, Techniques and Opportunities:

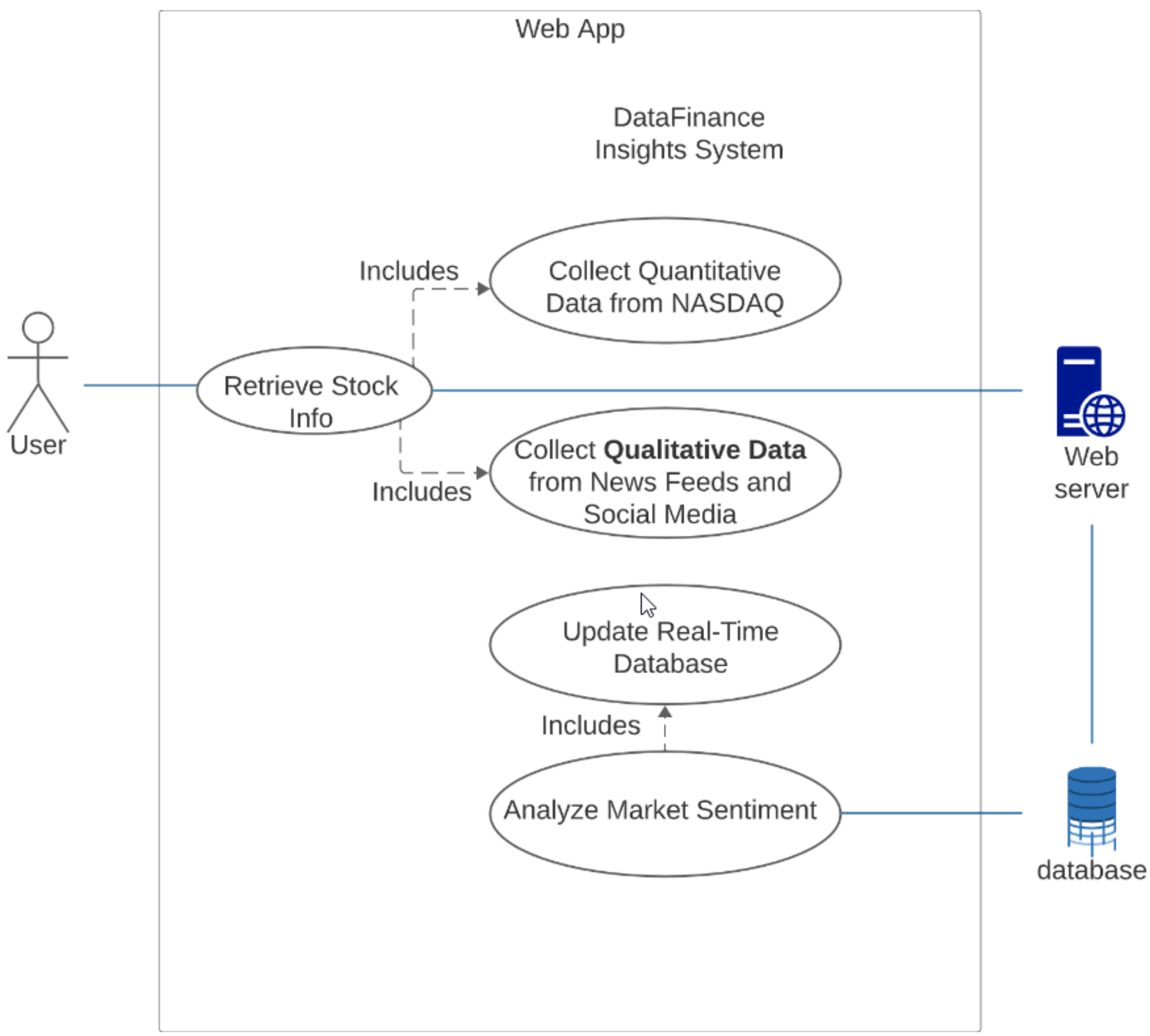
This paper takes an overview of AI in Finance, summarizing multiple ideas and forms of analysis. For our simplicity I will summarize main topics that can be useful to us in conducting research and analysis. An overview of challenges in this field are mechanism design and optimization, forecasting and prediction, portfolio management, sales and marketing analysis, business profiling, sentiment analysis, anomaly detection, compliance, risk management, and objective optimization and operations optimization.

Alquina M, et al. 2021 - Overview of High Frequency Trading (HFT)

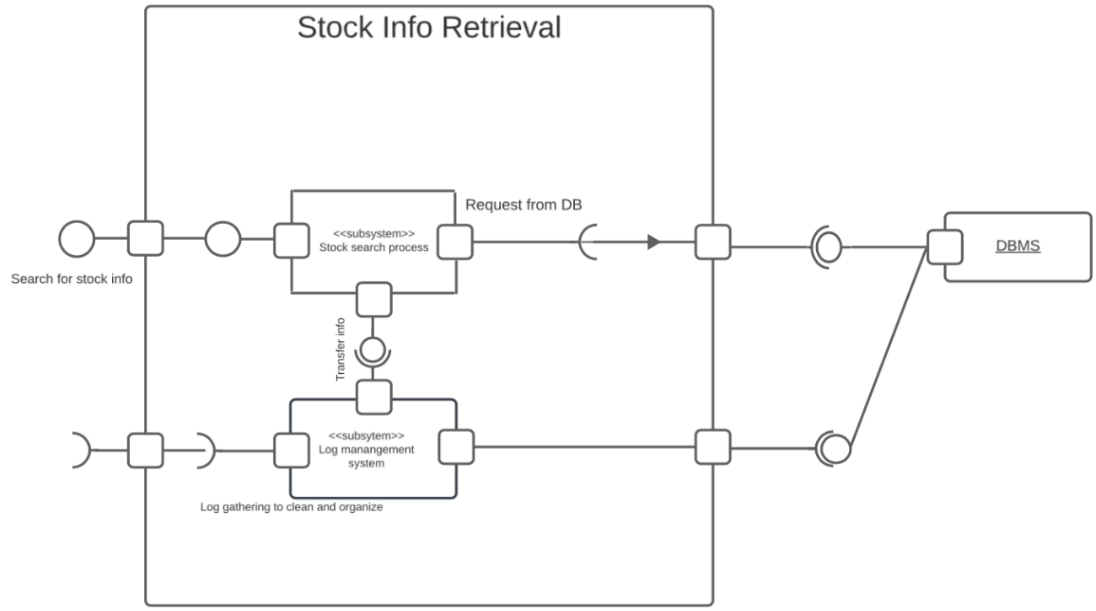
This paper gives an overview of the speed at which markets are made up. HFT or latency arbitrage races (5-10 millionths of a second) account for about 20% of all volume. This is concentrated within the top 6 firms. The top 6 firms due to this take about 80% of the liquidity of the markets while only providing 42%. This means they will have to pay higher fees, but are also removing liquidity of the markets. Conversely, outside of the top 6 firms the rest only take 20% of liquidity and provide 58%. Basically how this works is that when prices move there are still bid-ask quotes that are 'stale' and profit can be made on these. This can be done in many markets but an example is if the price of S&P futures contract changes by a large enough amount, the race would pick off stale quotes in every asset highly correlated to the S & P 500 index. So these minor changes in highly correlated assets result in arbitrage opportunities. Other places where this can be done is T-bonds, cash vs future markets, options and underlying stocks, ETFs, currency triangles, commodities at different delivery dates and lastly, this can be done for the same assets at different platforms – there are 16 exchanges and 50+ alternative trading platforms where arbitrage opportunities can occur. While we will not be implementing HFT due to high costs and heavy engineering, it is important to understand how the market operates.

3. Analysis of Pipeline

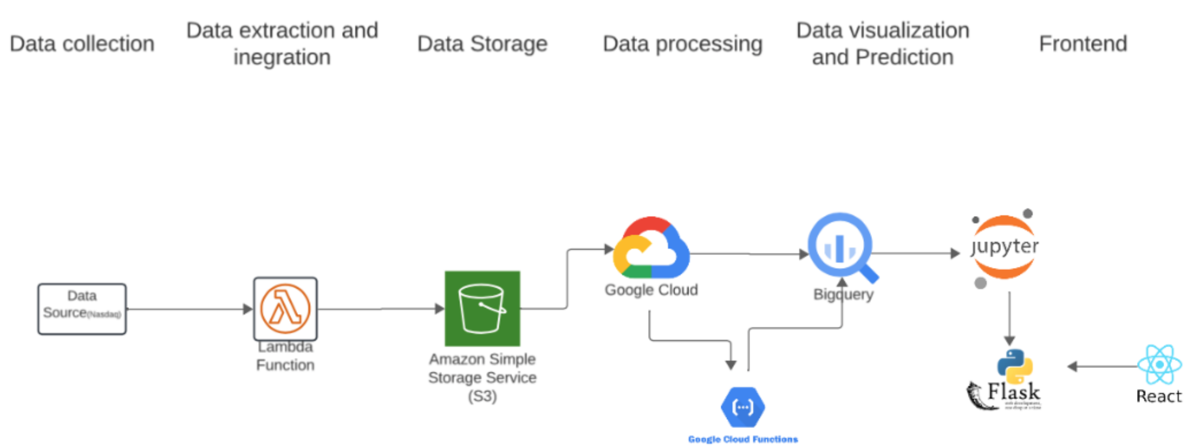
3.1 System Architecture and Diagrams



Use Case Style Diagram



Component Level Diagram



Architectural Diagram

3.2 Description of Dataset

Our dataset is sourced from the AlphaVantage API. This API offers a wide variety of financial data related to the stock market. We collected daily price values, including open, high, low, close, and volume (OHLCV), from the past four months for each stock. With the OHLC values, we would be able to run our predictive analytics and give the user a good representation of stock movement over time.

3.3 Data Ingestion

In the data integration phase, Lambda served as a pivotal component in seamlessly pulling stock information from the AlphaVantage API and efficiently inserting it into an S3 bucket. Leveraging the serverless computing capabilities of Lambda, the integration process was streamlined, allowing for automated and scalable data transfer. Lambda's event-driven architecture enabled the immediate response to triggers, ensuring that stock information was regularly updated in the S3 bucket. This dynamic integration not only facilitated real-time data retrieval, but also optimized storage and accessibility within the AWS ecosystem. The serverless approach offered by Lambda enhanced the efficiency of data integration, contributing to the overall effectiveness of the project's architecture. Next in our data integration process, we transferred the information stored in the S3 bucket to Google Cloud. This cross-cloud data transfer was executed to leverage the unique benefits offered by Google Cloud Platform (GCP). The integration between S3 and GCP facilitated a more comprehensive and flexible data ecosystem. By harnessing the strengths of both AWS and GCP, we ensured a robust data flow, allowing for improved analytics and visualization capabilities. This dual-cloud approach not only maximized the advantages of each platform but also enhanced data redundancy and accessibility.

3.4 Data Processing

In the data processing phase, Google Cloud Functions played a pivotal role as we created functions tailored to the requirements of our project. These functions were specifically designed to refine and enhance the incoming data within Google Cloud, eliminating redundancy and ensuring proper formatting. Leveraging the serverless capabilities of Google Cloud Functions, we achieved an agile and scalable data cleaning process. The software processed the data before sending it to be stored in tables within BigQuery. This approach not only streamlined the data processing pipeline but also allowed us to use the analytical power of BigQuery for insightful queries and data analysis.

When the data arrived in our database we made sure to run simple analytics on the data to make sure all of the data is there, in the format we expect, and clean enough to use for analysis. Our expected output is to have 8 stocks, each with 100 days of historical data, so 800 rows in total. The list of commands we used for cleaning analysis are below.

I. We used describe() to make sure the count for each of our variables was the same.

```
df.describe()
```

	open	high	low	close	volume
count	800.000000	800.000000	800.000000	800.000000	8.000000e+02
mean	182.485083	184.400177	180.620709	182.600612	4.004052e+07
std	86.501113	87.330252	85.589010	86.491484	3.674617e+07
min	32.230000	32.910000	32.020000	32.520000	1.640931e+06
25%	134.766800	136.396250	133.242500	135.030000	1.306773e+07
50%	165.640000	167.650000	164.295000	165.800000	2.860251e+07
75%	241.422500	244.405000	239.178175	241.695000	5.282563e+07
max	383.760000	384.300000	378.160000	382.700000	1.746679e+08

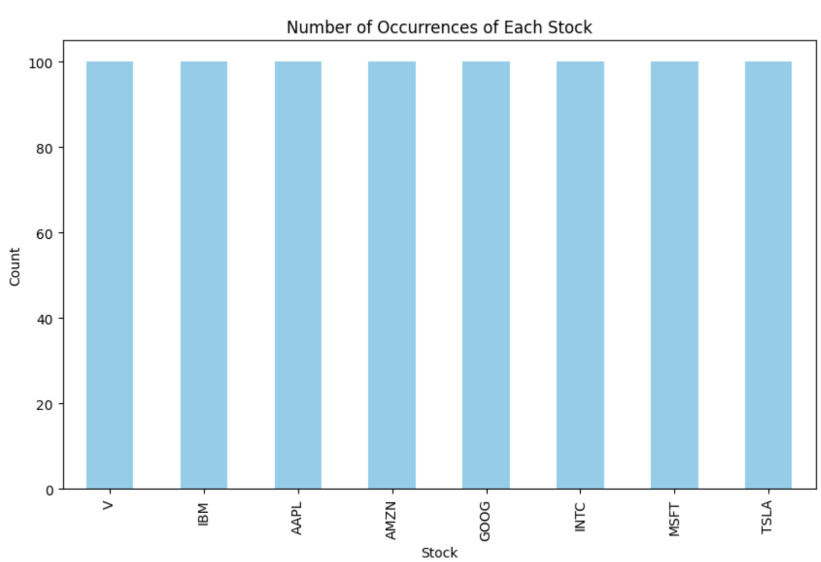
II. To check if the correct stocks and count for each stock were loaded, we used these three functions below, and we did in fact have the right counts for all.

```
df['stock'].describe()
```

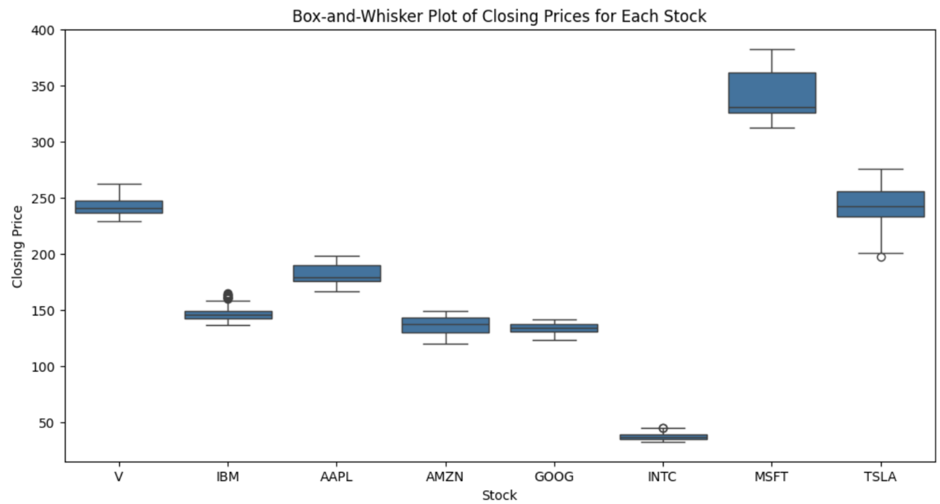
```
count      800
unique       8
top          V
freq       100
Name: stock, dtype: object
```

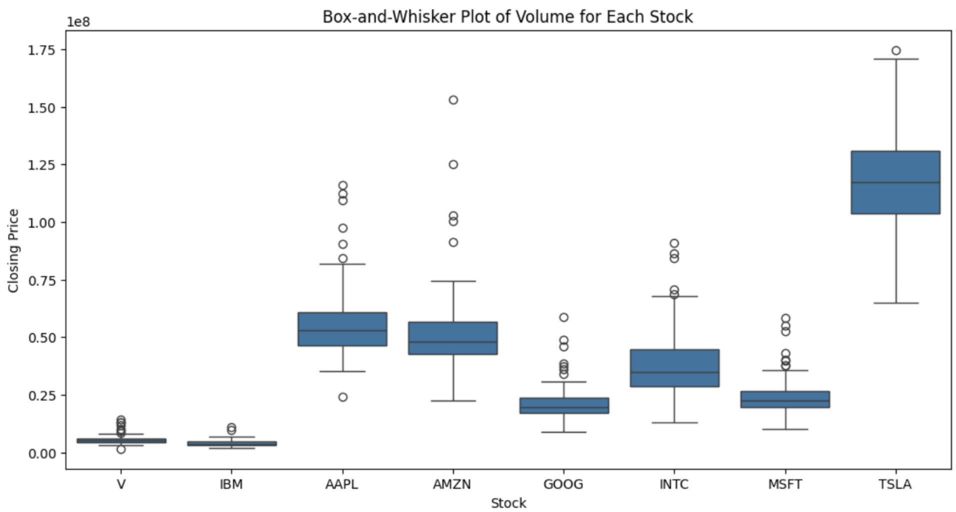
```
list(df['stock'].unique())
```

```
['V', 'IBM', 'AAPL', 'AMZN', 'GOOG', 'INTC', 'MSFT', 'TSLA']
```



III. Lastly, we had to make sure the values we got for closing price, volume and other metrics were correct. We created box and whisker plots for each of the applicable variables and plotted. The examples shown below incorporate closing prices and volume, do not show any outliers that cause concern and with that, we can continue to model the data.



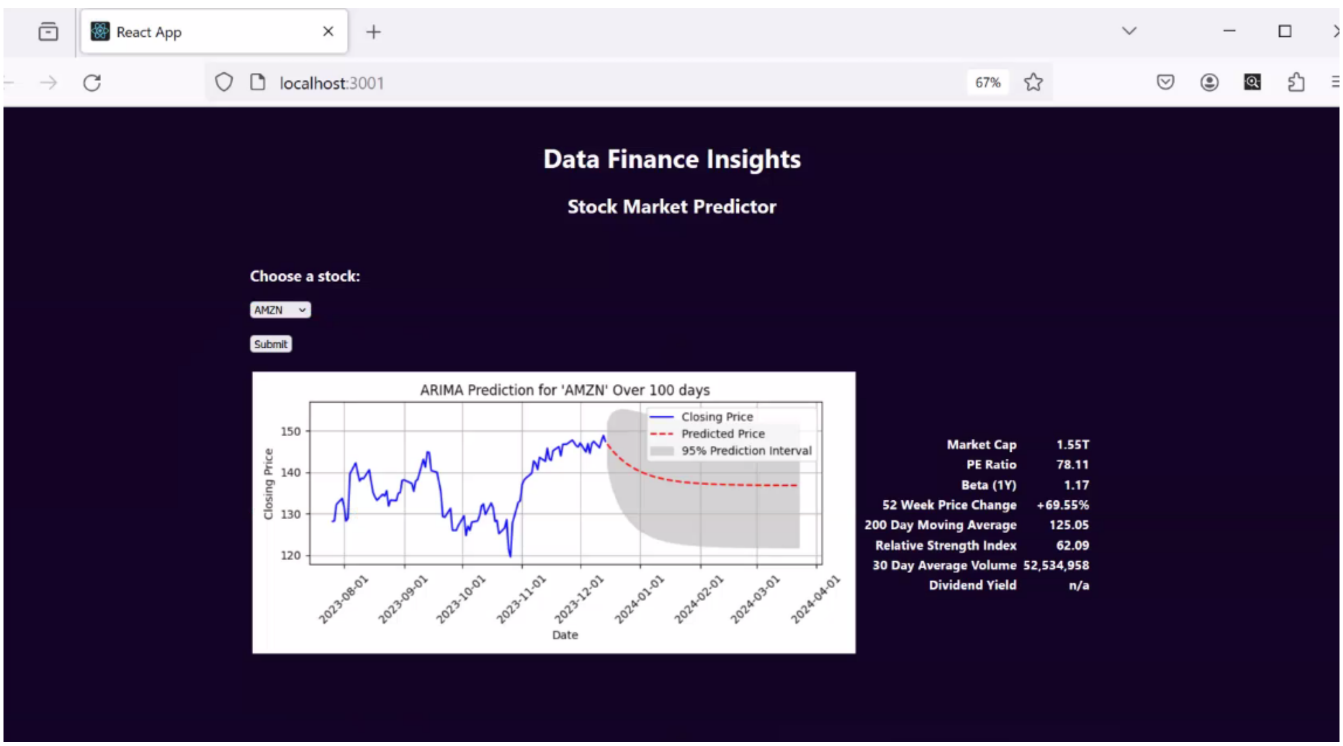


3.5 Predictive Analytics

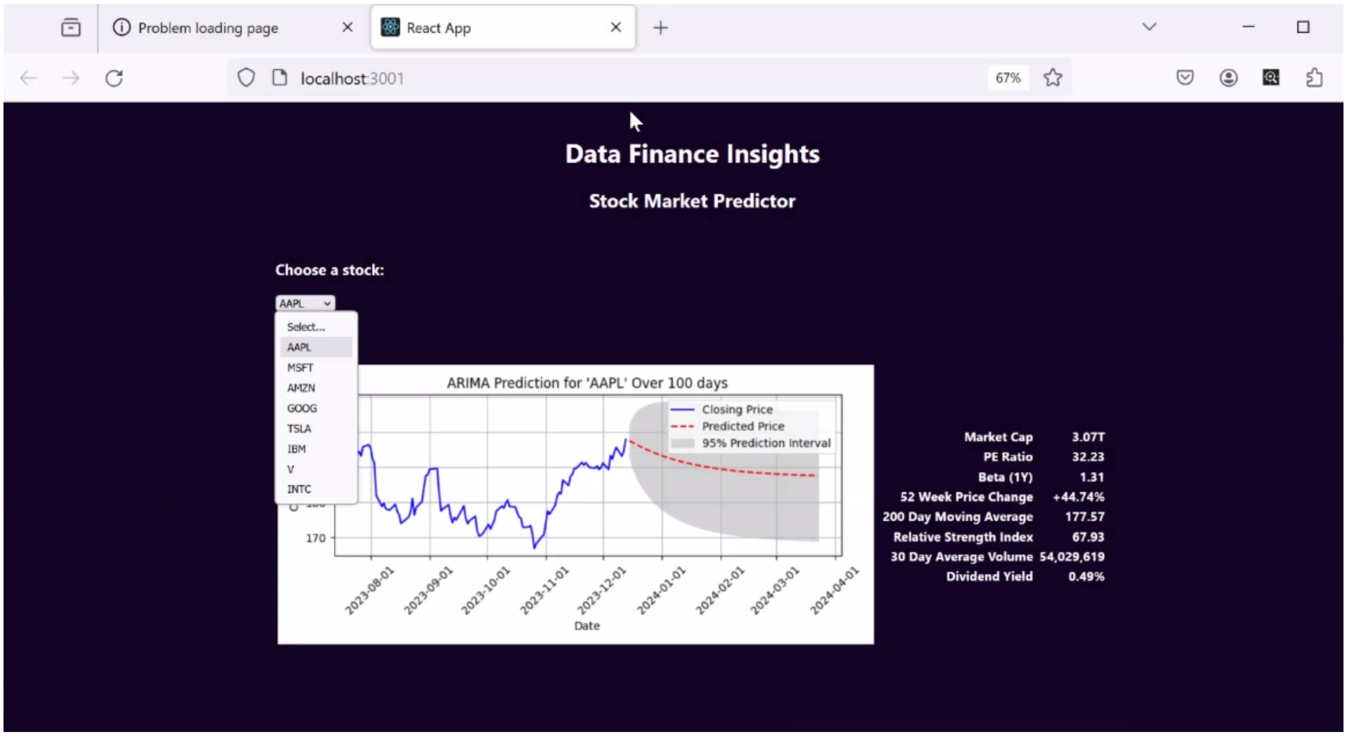
Predictive analytics was the next step in our process. We used an autoregressive integrated moving average (ARIMA) model and the Facebook Prophet (Prophet) model to offer two different forms of time-series analysis. ARIMA can be broken down into three pieces, AR, I and MA. AR indicates that the variable of interest is regressed against its lagged (prior) values, I indicates how many times the data has been different, and MA incorporates the dependency between an observation and a residual error from a moving average model applied to the lagged observations. Prophet is a model for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality.

Each ARIMA and Prophet incorporate tuning parameters to increase model accuracy. To have our system be fully autonomous, we created grid search functions that iterated through different combinations of the tuning parameters and returned the parameters that resulted in the model with the smallest amount of error. Once the models were tuned we plotted the graphs as seen in the visualization phase below.

In the data visualization phase, we employed Python scripts to create dynamic and insightful visual representations of the predicted data. To enhance user accessibility, the visualizations were seamlessly integrated into a front-end user interface. This interactive interface provides users with a user-friendly platform to explore and interpret the data comprehensively. The Python scripts, acting as the backbone of this visualization process, empower users to extract valuable insights from the data collected.



Front-end Visualization



Front-end Visualization

4. Conclusion

DataFinance Analytics is a useful representation of all components necessary to create a stock data pipeline with predictive analytics. By utilizing serverless tools like LAMBDA, S3 Buckets, and Google Cloud, we seamlessly transported data from an API, to a database. We ran predictive analytics utilizing ARIMA and Prophet, on the now cleaned and formatted data in our database. Lastly, integrated all necessary data and predictions to a user-friendly front end to deliver the full value of our pipeline.

5. References

[1] Brantmeyer, Collin. “Stock Split Watch: Is Tesla Next?” *The Motley Fool*, The Motley Fool, 7 July 2023, www.fool.com/investing/2023/07/07/stock-split-watch-is-tesla-next/.

[2] Cillizza, Chris. “Americans Read Headlines. and Not Much Else.” *The Washington Post*, WP Company, 26 Nov. 2021, www.washingtonpost.com/news/the-fix/wp/2014/03/19/americans-read-headlines-and-not-much-else/.

[3] Birz, Gene, and John R. Lott. “The Effect of Macroeconomic News on Stock Returns: New Evidence from Newspaper Coverage.” *Journal of Banking & Finance* 35, no. 11 (2011): 2791–2800. https://doi.org/10.1016/j.jbankfin.2011.03.006.

[4] Bomfim, Antulio N. “Pre-Announcement Effects, News Effects, and Volatility: Monetary Policy and the Stock Market.” *Journal of Banking & Finance* 27, no. 1 (2003): 133–51. https://doi.org/10.1016/s0378-4266(01)00211-4.

[5] He, Hongxing, Jie Chen, Jin Huidong, and Chen Shu-Heng. “Stock Trend Analysis and Trading Strategy.” *Proceedings of the 9th Joint International Conference on Information Sciences (JCIS-06)*, 2006. https://doi.org/10.2991/jcis.2006.135.

[6] Iyinoluwa, Oyelade. “Stock Market Trend Prediction Model Using Data Mining Techniques.” *Current Trends in Computer Sciences & Applications* 1, no. 5 (2019). https://doi.org/10.32474/ctcsa.2019.01.000122.

[7] Kolasani, Sai Vikram, and Rida Assaf. “Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks.” *Journal of Data Analysis and Information Processing* 08, no. 04 (2020): 309–19. https://doi.org/10.4236/jdaip.2020.84018.

[8] Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. “The Effects of Twitter Sentiment on Stock Price Returns.” *PLOS ONE* 10, no. 9 (2015). https://doi.org/10.1371/journal.pone.0138441.

[9] Robinson, Justin, Adrian Glean, and Winston Moore. “How Does News Impact on the Stock Prices of Green Firms in Emerging Markets?” *Research in International Business and Finance* 45 (2018): 446–53. https://doi.org/10.1016/j.ribaf.2017.07.176.

[10] Olaniyi, S.A.S. & S., Adewole & Jimoh, Rasheed. (2011). Stock trend prediction using regression analysis-A data mining approach. ARPN Journal of Systems and Software. 1. 154-157.

[11] Pricope, T. V. (n.d.). (publication). *Deep Reinforcement Learning in Quantitative Algorithmic Trading: A Review*. Edinburgh, UK: Informatic Forum (2021).

[12] Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys*, *55*(3), 1–38. https://doi.org/10.1145/3502289

[13] Zeng, Z., Balch, T., & Veloso, M. (2021). Deep Video Prediction for time series forecasting. *Proceedings of the Second ACM International Conference on AI in Finance*. https://doi.org/10.1145/3490354.3494404

[14] Aquilina, M., Budish, E., & O'Neill, P. (2021). *Quantifying the High-Frequency Trading "Arms Race."*https://doi.org/10.3386/w29011