GRAPES: semi-automatic approach for forecasting models to predict GameStop prices using cloud computing and machine learning

Tan Van Vo and
Sukhpal Singh Gill*

School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London, England, UK
Email: t.vo@se13.qmul.ac.uk
Email: s.s.gill@qmul.ac.uk
*Corresponding author

Abstract: Since the Covid-19 pandemic, we have seen a surge of retail investors that now can easily trade anywhere in the world with just a Smartphone. Social media groups like Reddit's WallStreetBets have almost put a few hedge funds close to bankruptcy by driving GameStop share prices to the sky. In this work, we propose a framework called *GRAPES* which uses Cloud Computing and Machine Learning to explore various forecasting techniques in predicting GameStop prices. In addition to this, this work also provides light insight into semi-automating forecasting models using tools such as Google Cloud Platform (GCP), Airflow and Streamlit. Moreover, we monitored the investment funds from Ark Invest to provide additional insight into the market in general. Overall, the paper shows the Autoregressive Moving Average (ARMA) model gives the best accuracy based on the Mean Absolute Percentage Error (MAPE) of 1.12%. This means the predictive model is out with an average of 1.12% from the actual price.

Keywords: apache airflow; Google cloud platform; docker; stock; WallStreetBets; streamlit; forecasting; GEM; GameStop; ETF; exchange-traded fund.

Reference to this paper should be made as follows: Vo, T.V. and Gill, S.S. (2022) 'GRAPES: semi-automatic approach for forecasting models to predict GameStop prices using cloud computing and machine learning', *Int. J. Grid and Utility Computing*, Vol. 13, No. 5, pp.538–550.

Biographical notes: Tan Van Vo is a Data Science MSc student at Queen Mary, University of London, UK. He has a BSc in Mathematics. Currently, he is employed as R&D Data Engineer. Before that, he did Reserving and Pricing for Leading Insurances in the UK. His research interests include stocks, machine learning and cloud technologies.

Sukhpal Singh Gill is a Lecturer (Assistant Professor) in Cloud Computing at School of Electronic Engineering and Computer Science, Queen Mary University of London, UK. Prior to this, He has held positions as a Research Associate at the School of Computing and Communications, Lancaster University, UK and also as a Postdoctoral Research Fellow at CLOUDS Laboratory, The University of Melbourne, Australia. He is serving as an Associate Editor in *IET Networks* Journal. His research interests include cloud computing, fog computing, software engineering, internet of things and healthcare.

1 Introduction

During the coronavirus pandemic, we have seen a surge of amateur investors to fend off the boredom of lockdowns. According to Financial Times research, 'Retail trading now accounts for almost as much volume as mutual funds and hedge funds combined' (Martin and Wigglesworth, 2021). Many unemployed people traded with commission-free trading platforms in the market to replace lost income.

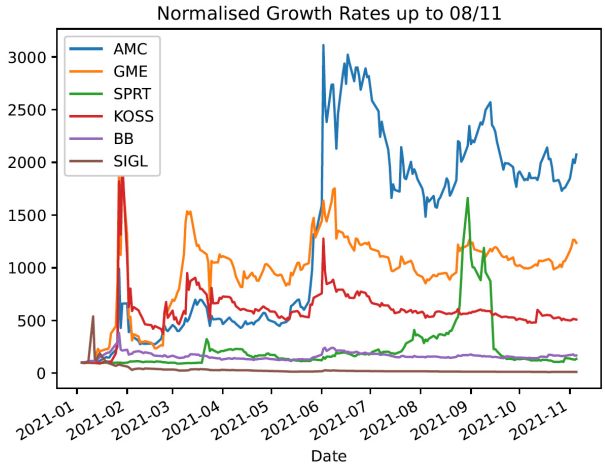
On 27th January 2021, a struggling chain GameStop (NASDAQ:GME) that had hit global headlines reached $347 closing price from $17 at the beginning of the year. WallStreetBets (WSB) community was the main driver of GameStop. During the same week, WSB subreddit had gained over two million new subscribers to six million total members (Ghosh, 2021). Similar to GME, there are other companies we selected that were also impacted by Covid-19 that went parabolic by Reddit traders shown in Figure 1. The graph in Figure 1 shows the performance of several

meme stocks from year-end 2020 normalise closed price series growth rate at 100 as followed:

- A movie theatre chain AMC Entertainment (NASDAQ:AMC) escaped bankruptcy climbed from \$2.12 to \$19.90 per share on 27th January 2021. Its 52-week high as of 8th November 2021 is \$72.62.
- A manufacturer and designer of headphones Koss Corporation (NASDAQ:KOSS) with declining sales climbed from \$3.44 to closed price of \$64 on 29th January 2021. It reached its 52-week high of \$127.45 on 28th January 2021.
- Blackberry Ltd (NASDAQ:BB), their mobile phones were popular in the early 2000s and now develops security softwares. Share price climbed from \$6.63 to closed price of \$25.10 on 27th January 2021. On the same day, it also reached its 52-week high of \$28.77.

In brief, it is estimated the short-selling hedge funds have suffered \$20billion losses in GME in January 2021 according to S3 Partners (Li, 2021).

Figure 1 Normalised growth rates from 01/01/21



The normal market trading hours are 9:30 am to 4 pm (Eastern Time). Trading outside the market normal hours is more volatile and riskier because of fewer investors (Beers, 2021). Scanning the market to find the right stock to trade can be time-consuming. Stock prices are commonly driven by a variety of fundamental catalysts such as earning reports, Food and Drug Administration (FDA) approvals or disapprovals, and major contracts win or lose. Furthermore, a misinterpreted tweet from Elon Musk can send a random stock soaring, i.e., a technology company Signal Advance (NASDAQ:SIGL) in Figure 1, climbed from \$0.60 on 7th January 2021 to \$38.70 on 11th January 2021 (Ho, 2021). A lack of due diligence can make investors lose a lot of money from extreme volatility stocks or be extremely lucky to profit. So, a \$100 investment would give you a return of \$6,450 if you successfully cash out at the peak.

1.1 Motivation and our contributions

For this paper, it was of interest to investigate if we can build forecasting models to predict GameStop prices. In this work, we proposes a framework called *GRAPES*, which is predicting **G**ameStop p**R**ices and introduces a semi-**A**utomatic a**P**proach for **E**casting model**S** using cloud computing and machine learning. We explored three techniques as follows: Prophet, Time Series and Long Short-Term Memory. Prophet was the quickest machine learning algorithm to produce results. However, the Time Series dynamic prediction had the best result. One of the metrics showed the Mean Absolute Percentage Error (MAPE) in the time series forecast to be out of range by 1.12% from the actual price. The methods and results will be discussed further.

In addition to this, it is an interest to follow one of the top 100 Exchange Trade Fund (ETF), Ark Invest (ETFs, 2021) which invest in disruptive innovation. However, they do not invest in GameStop. In the US alone, ETF accounts for approximately \$5.45 trillion and has been growing year on year in the last two decades (Statista, 2020). Ark Invest caught media attention in the last few years, so it would be interesting to monitor their portfolios to see if they can make informed decisions. This will be explored further by tracking the number of shares they hold and stock prices. Moreover, WSB has attracted more attention this year. In brief, we will examine the impact from WSB to GME. Overall, so many factors affect the stock market, and it is time-consuming research in stocks. One way to overcome this is to build a simple web application to centralise all data collected for this paper in a single place and allow users to construct forecasting models.

The rest of this article is organised as follows: Section 2 discusses the related work. Section 3 presents methodology. Section 4 discusses implementation and experimental results. Section 5 concludes and summarises the paper and highlights the future directions.

2 Related work

At the time of writing this paper, there was no academic research in building forecasting models for GameStop. Google Scholar has only eight papers, and IEEE has only two papers related to GameStop. However, those researches are more focused on WallStreetBets and the sentiment analysis from Reddit posts.

One interesting article from Letelier (2021) used anomaly detection methodology on GameStop with the *k*-nearest neighbours algorithm (*k*-NN). He provided an animation of 5-minute interval data up to two hours data over two months period to detect unusual behaviour. However, he concluded that the algorithm could not predict the short squeeze in January 2021 because there was already abnormal behaviour two weeks prior.

Despite the outcome, another article by Xin (2021) used time series method to forecast GameStop price. The data he collected was up to mid-April 2021, and he forecast the last 30 entry points. However, the forecast data points do not seem to capture the trend of the test data points. The pre-processing step seems to be flaw. For example, he did not use log difference transformation but used the difference method instead. By using the log transformation allows the variance of time series to stabilise. Also, by observing his Autocorrelation (ACF) and Partial Autocorrelation (PACF) graphs, this suggests the time series is an ARMA (0, 0). Effectively, suggesting the model does not have a serial correlation. An improvement of this model will be discussed in the result section.

To update the database on a daily basis, a concept from an article by Benton (2020) demonstrate using Airflow to schedule daily task to refresh Apple Inc. (NASDAQ:AAPL) market data from his desktop. To leverage cloud technologies and to reduce local storage space, we created pipelines to refresh data into Google BigQuery (Gill et al., 2022).

Finally, to create a simple semi-automatic forecasting model quickly. We explored an interactive financial dashboard with Streamlit by Marx (2020). To improve on this to reflect on this paper, we used Streamlit to connect to Google BigQuery to allow users to select a few options to produce forecast results.

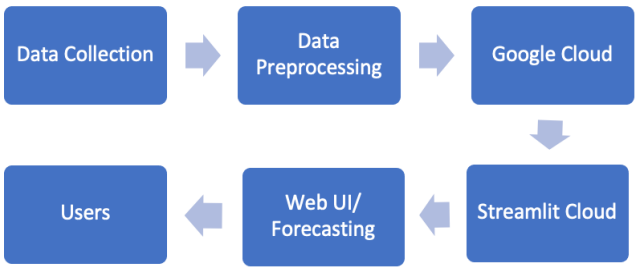
3 Methodology

3.1 Methodological approach

This section discusses the tools we used. The idea behind this is to centralise our data into one place to reduce repetitive tasks and gain a competitive advantage by building a web application. The app will allow users to forecast stock prices with additional information from WSB and Ark-Invest portfolios.

We have used Apache Airflow to automate tasks by writing scripts and scheduling data pipelines. The diagram in Figure 2 gives a high level overview how data are moved across the platform. Airflow was set up on Docker instead of a local machine to reduce the dependency issues. The data collected are stored in Google BigQuery and Google Storage. It is fast, low cost and provides backups while retaining immediate access.

Figure 2 Data flow framework



3.2 Data collection

The data is collected from three external data sources. The first two data sets discussed below are not used in to build forecasting models. Those data sets are collected are experiments to make observation of the potential uses. The data are reviewed and discussed the impact it may have in performance of stock in general. The third data set below is used for forecasting.

1 ARK's Exchange Traded Funds are investment funds managed by Cathie Wood. She is known for being the best stock picker and for making bold predictions (Wikipedia, 2021), i.e., she predicted Tesla stock would hit $4000 in 2018 (Fox, 2021). The data has been collected on a daily basis since October 2021 to monitor the changes in shares and weights of the portfolio (*ARKK, ARKW, ARKG, ARKQ, ARKF, ARKX, PRNT, IZRL*) (ARK Fund, n.d.). However, the study size in this study is not large enough, but it still useful to gain valuable insight. They have a combined market portfolio value of over $40,000 m in at least 100 companies. By monitoring their portfolios may impact the direction of the market movements.

2 WallStreetBets (WSB) (WallStreetBets, 2021) is a subreddit where users discuss stock from Reddit. The social platform hit headlines globally because Reddit day trader was the main driver of GameStop.

   This data aims to extract the number of tickers that contains 'cashtag' mentioned from the subreddit's post using the PSAW library to see some correlations with the market movements. PSAW library searches public Reddit comments/submission via the pushshift API (Marx, 2018). However, extracting more data beyond cashtag($) are challenging, such as sentiment or emotion analysis. This is a difficult task for machines to understand sarcasm and the sense of meaning.

3 Historical daily NASDAQ data is collected from Kaggle and maintained by Mooney (2021). However, it is only updated on a weekly basis. This is however used as a base table. To collect data regularly, we have used Airflow to update the database on a daily basis using Yahoo Finance API. The Airflow is schedule to update the database in Google BigQuery. The data set consist of the date, open, close, volume and adjusted close price.

3.3 Data technology

In this section, the paper discusses the technology used to build the web application and forecasting models. Figure 2 gives high level overview how the data are move across.

1 *Apache-Airflow:* Is an open-source platform that manages workflows as Directed Acyclic Graphs (DAGS), schedule and monitor workflows (Airflow, 2021a). Figure 2 shows an example of a DAG diagram.

   Airflow has four main components as follows as shown in Figure 3:

a) *User interface (web server)*: Is a flask application that allows you to monitor and manage your workflows. You can review the logs to debug failed tasks.

b) *Metadata database*: Keeps a record of all the DAGS tasks and their status.

c) *Scheduler*: Is responsible for triggering workflows and tasks it monitors that reads from the metadata database. It is also responsible for staying in sync with the folder where are DAG stored.

d) *Executor*: Once the tasks are triggered, it will pass on to the executors which will run the task. It figures out how resources are managed and which workers should execute each tasks. Apache Airflow stated 'There are two types of executor – those that run tasks locally (inside the scheduler process), and those that run their tasks remotely (usually via a pool of workers)' (Airflow, 2021c). For this experimental design, this study used the 'Sequential Executor' because the ram resource is low and straightforward to set up. However, it is not scalable, so it can only run one task instance at a time.

2 *Docker:* Is an open source container that allows developers to package up an application it needs to run. From our experience, installing Airflow in Docker provides smoother process in setting up in a Docker container than local machine, i.e., Window or M1 Mac.

3 *Google cloud platform:* Is a cloud-based environment. Cloud technologies are increasingly popular. Resources are flexible and scalable, provide better security and all applications are updated automatically. Here are the list of products that were used for this study:

a) *Cloud storage*: Is a secure and scalable storage that can be accessed globally (Google, 2021a). The data collected above are stored in the cloud. As the data set size increases and/or add new data sources will require more space. Without taking up physical space on your computer, Cloud Storage can increase your storage size fast.

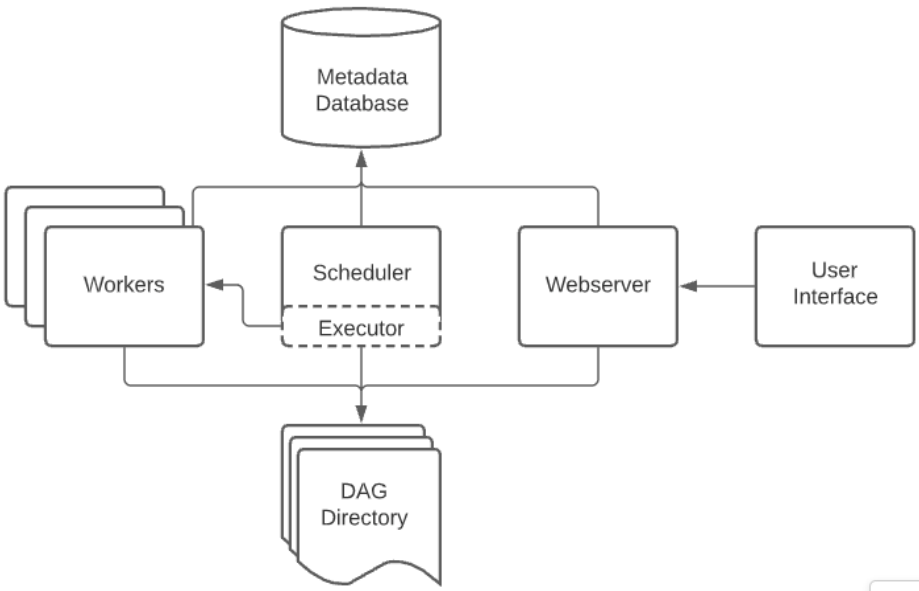
b) *Big query*: Is a serverless data warehousing that uses standard SQL to query the data (Google, 2021b). The data from cloud storage are added to the BigQuery warehouse seamlessly by scheduling tasks in Airflow.

c) *Dataproc*: Is a Spark and Hadoop service that can set up clusters quickly. It is also integrated with BigQuery and Cloud Storage that can orchestrate ETL pipelines (Google, 2021c). The code below is a snapshot in creating clusters with Airflow. In brief, you can adjust the number of clusters you need. By increasing the clusters (worker) will improve the query performance, but it cost more to run (Google, 2021d). To save money, you can schedule Airflow to delete the clusters after finishing the query.

```
create_cluster =
DataprocClusterCreateOperator(
task_id='cluster_id',
project_id= 'google_project_id',
cluster_name = 'cluster_name',
num_workers = 2,
storage_bucket  =  'google-storage-
bucket',
zone='location' )
```

4 *Web application:* The data hosted in BigQuery are accessed through Streamlit Cloud. It is the fastest way to build a web application than writing the code from ground up like Flask (Schmitt, 2021). We have created dynamics dashboards that can be updated in real time. The user can select forecasting methods, that will be discussed in the next section, to predict future prices and review summary statistics collected from BigQuery.

Figure 3 Airflow architecture overview (Airflow, 2021b)



3.4 Forecasting models

In the app, the user can select one of the following methods below to predict future prices. In addition to this, the user can choose the date range they wish to model on shown in Figure 4.

1 *Linear regression*: This is used to model the relationship by fitting a straight line between the series data and the price that minimises the mean squared error.

2 *Polynomial regression*: It fit a nonlinear relationship between the time series dates and the model you want to predict, i.e., Closed Price. Increasing the nth degree of the polynomial allows the flexibility to fit a wide range of curvatures. However, this will overfit the data and will lead to poor performance in predicting the daily prices.

3 *ARMA(p, q)*: An ARMA model is a combination of Autoregression (AR) and Moving Averages (MA). The p is order of AR part and the q is order of MA part. Before forecasting the time series, the data needs to be stationary. The trend, variance and AR needs to be constant to be stationary (Nau, 2014). This can be done using the difference and or transformation methods with Dickey-Fuller test statistics shown in Figure 5. The statistical test is where the null hypothesis is that your time series is non-stationary due to trend. For this paper, a log-transformation is applied first then a difference method. To select the order of p & q for the model can be done by visualising the Autocorrelation Function (ACF) and partial autocorrelation (PACF) graphs. However, a quicker approach is done by using the grid search process to find the best model that minimise the Akaike Information Criterion value (Zajic, 2019). Once all the steps processed are completed, we will experiment with two predictive methods, static and dynamic. The static uses the previous series value to estimate the next value, whereas the dynamic method uses the predictive value to forecast the next value.

4 *ARMAX(p, q) model*: Similar to the ARMA model, we can model the time series with additional independent variables known as an exogenous variable. For instance, the amount of volume of shares traded between buyer and seller on a given period is shown in Figure 6. If there is a significant movement in the volume compared to the average last seven days, it can make the share price more likely to move in any direction.

5 *Prophet*: This is a fully automatic forecasting method and fasts which don't require advanced knowledge in time series forecasting. Prophet was developed by Meta (Facebook). You can get results in just a few seconds. The time series model contains four components: trend, seasonality, holidays and error term. They are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + e(t) \quad (1)$$

Stock markets in the USA have similar holidays to the federal government's holiday schedule (Phung, 2021). Prophet has a built-in function of country-specific holidays (Facebook.github.io, n.d.).

6 *Long Short Term Memory (LSTM)*: Is an artificial recurrent neural network that reuses the output step as an input for the next step and so on. So, it can learn a sequence of time series to predict the subsequent output. For this experiment, the mean square error was used as a loss function and ADAM as the optimisation algorithm (Agrawal, 2021). The data set values were scaled down in the range of 0 to 1. The number of epochs was set at 20 to allow the algorithm to minimise the error in the training set.

Figure 4 App – Date range selection

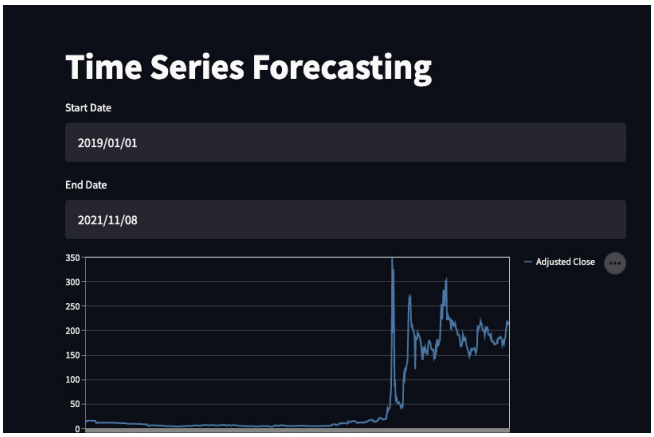


Figure 5 App – LogDiff and Dickey-Fuller

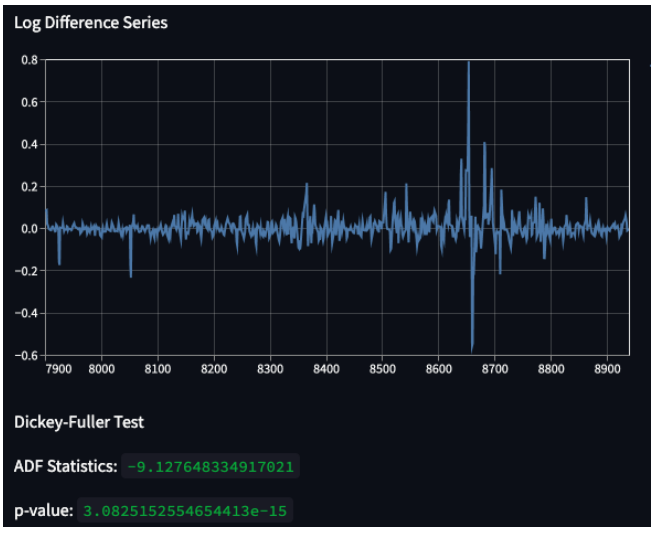
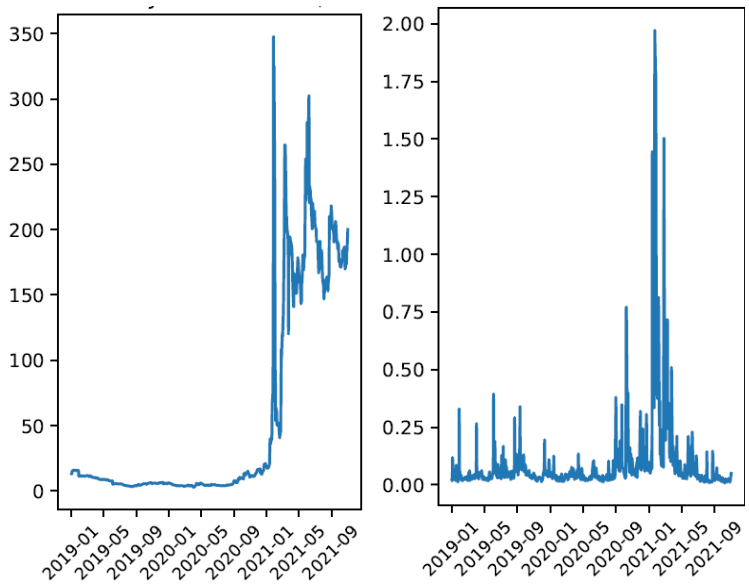


Figure 6 (a) GME daily close price (b) GME daily volume



(a) (b)

3.5 Evaluation metrics of forecasting models

To assess the quality of the forecasts, we measure the performance of the models as follows: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Further details of the metrics can be found (Rink, 2021).

4 Experimental results

In this section, we discussed briefly about ARK Invest, and comparing the query performance for BigQuery and DataProc. After that, the main focus present out findings on predicting GME and evaluating the model performances.

4.1 ARK invest

Two companies were selected from the ARKK fund. As shown in Figures 7 and 8 show continuous change of the closed price (LHS) and the total number of shares (RHS) between 5th October 2021 to 11th November 2021. Some data points are missing because the ETF funds were not downloaded on a daily basis. Despite the lack of data, the result demonstrates two things. First, Figure 7 shows a positive relationship. By raising their stake of ownership in SKLZ to suggest growth opportunities and expecting good financial reports. Secondary, Figure 8 shows an inverse relationship. They have double the number of shares despite a decline in price per share after HOOD announced data security breach (Vishnoi, A. and Bloomberg, 2021). This could be a strategy to push to average cost down and hope for the market to recover above the new average price or just a long term investment because of their products.

Figure 7 Skillz Inc. – an online mobile gaming

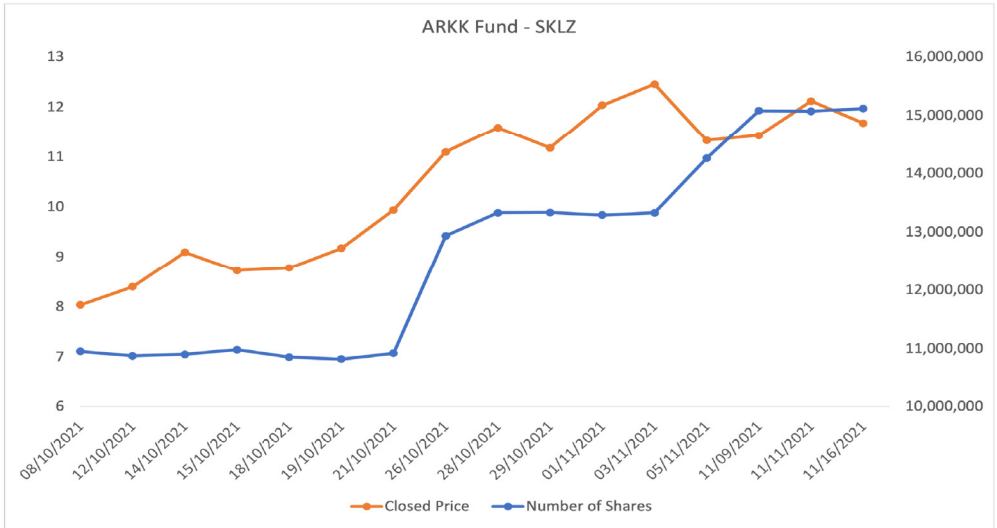
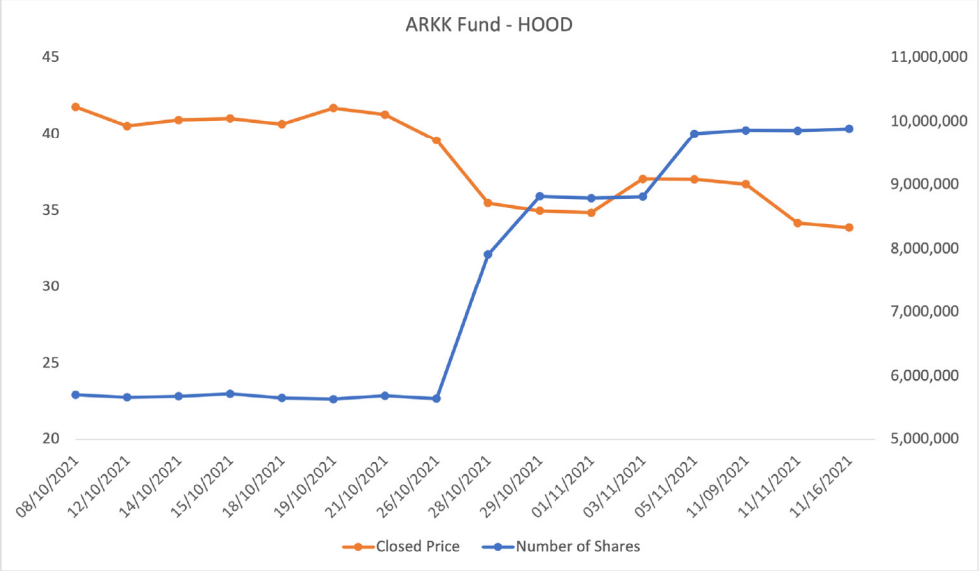


Figure 8 Robinhood – a financial service



4.2 BigQuery vs. Dataproc

An experiment has been tested on query response times with Airflow on a small data set. The result shows BigQuery response time was quicker to retrieve data within a few seconds. Whereas Dataproc can take a few minutes as the Airflow is scheduled to set up the clusters and then delete it when it finishes the query. A similar experience had been performed on a large scale without airflow can be found in Srivastava (2021).

4.3 Models

This section explores modelling techniques were discussed previously to forecast the closed price. The models are trained from January 2019 to compare and predict the last 25 entries of the training data. The natural assumption here is that the historical trend will be repeated in the future. However, predicting the future far ahead increases uncertainty, especially with volatile stocks uncertainty can proliferate.

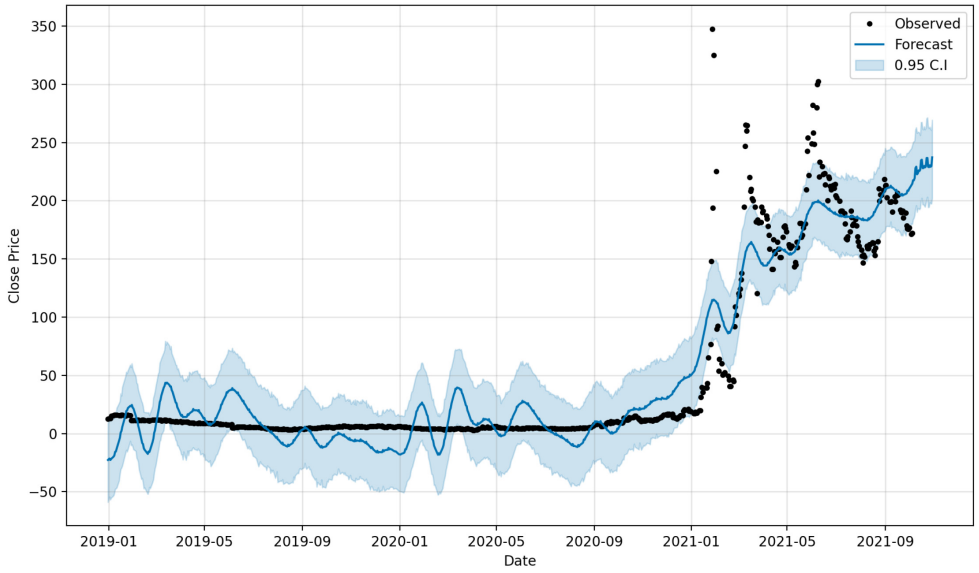
Both linear regression and polynomial would be the least likely models and perform poorly in predicting. They are highly affected by outliers, therefore scored extremely high on the evaluation metrics. Increasing the degree of the polynomial to capture the complexity of closed price data points will lead to a larger variance in predicting.

On the other hand, there is a powerful, fast and automated forecasting tool, Prophet, shown in Figure 9. We obtained good results with default parameters in Figure 9(a). However, even better results are achieved when adjusting the parameters shown in Figure 9(b). The prophet method could not able to capture the uptrend of the short squeeze in January 2021 because the default parameters of the 'change-points' is the first 80% of the data. By increasing the change-points will capture more of the time series data. Ideally, you don't want to increase the change-points as this will overfit the data. Towing the uncertainty, we should adjust the parameters to reflect current trend. To improve the accuracy of the 2nd model further, we have removed the daily and weekly seasonality because the future trend is uncertain. As shown, the forecast line fits better.

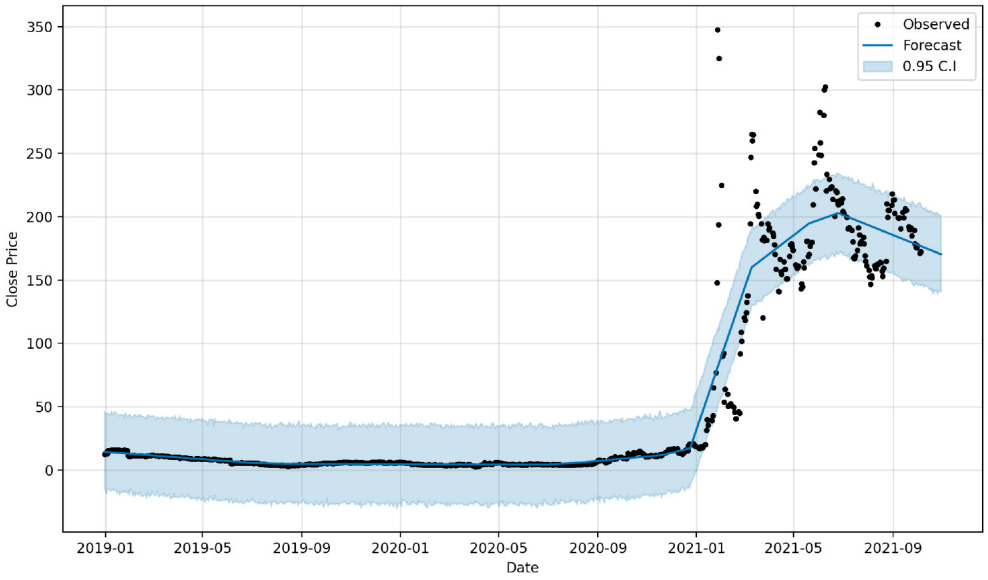
Figure 10 shows the ARMA and ARMAX static models prediction for two different timeframes using the one-step-ahead method for GME. This is an experiment to see how well the model would perform during the rally in January 2021 and the latest data after the market stabilised. Also, the ARMAX has an additional variable, Volume. This is particularly interesting to see if the Volume has an effect on the prices. As shown, both models have similar predictions for both timeframes. Despite the outcome, Volume is still an essential factor for a day trader, but perhaps not for predictive modelling. It is evident from the data that the top two models performed poorly and were misleading for January 2021 due to unprecedented volatility. As shown, an extreme change in the future increases the uncertainty to forecast. Whereas the bottom two models performed better when price action are calmer.

Figure 11 is similar to the previous result, but this time uses the dynamic forecasting method where the predictive value is used to forecast the next value instead of the actual value. However, the uncertainty can overgrow if we do not know the error term at each step. Considering the results, the forecast line in Figure 11 fits better than in Figure 10.

Figure 9 Graphs of GME Prophet models. (a) [GME – Prophet default] (b) [GME – Prophet adjusted]



(a)



(b)

Our last approach is a gentle introduction to the deep learning Long Short-Term Memory (LSTM) method. To forecast next outcome, the data comprises 60 sequences of data points. The whole process was then repeated iteratively. As shown in Figure 12, the data points are split into training set [blue line] and the test set [orange line]. The forecast line [green line] captures the trend very well over a more extended period than the usual 25 entries we tested previously.

Figure 10 GME static time series: Top – 08/02/21, Bottom – 08/11/21

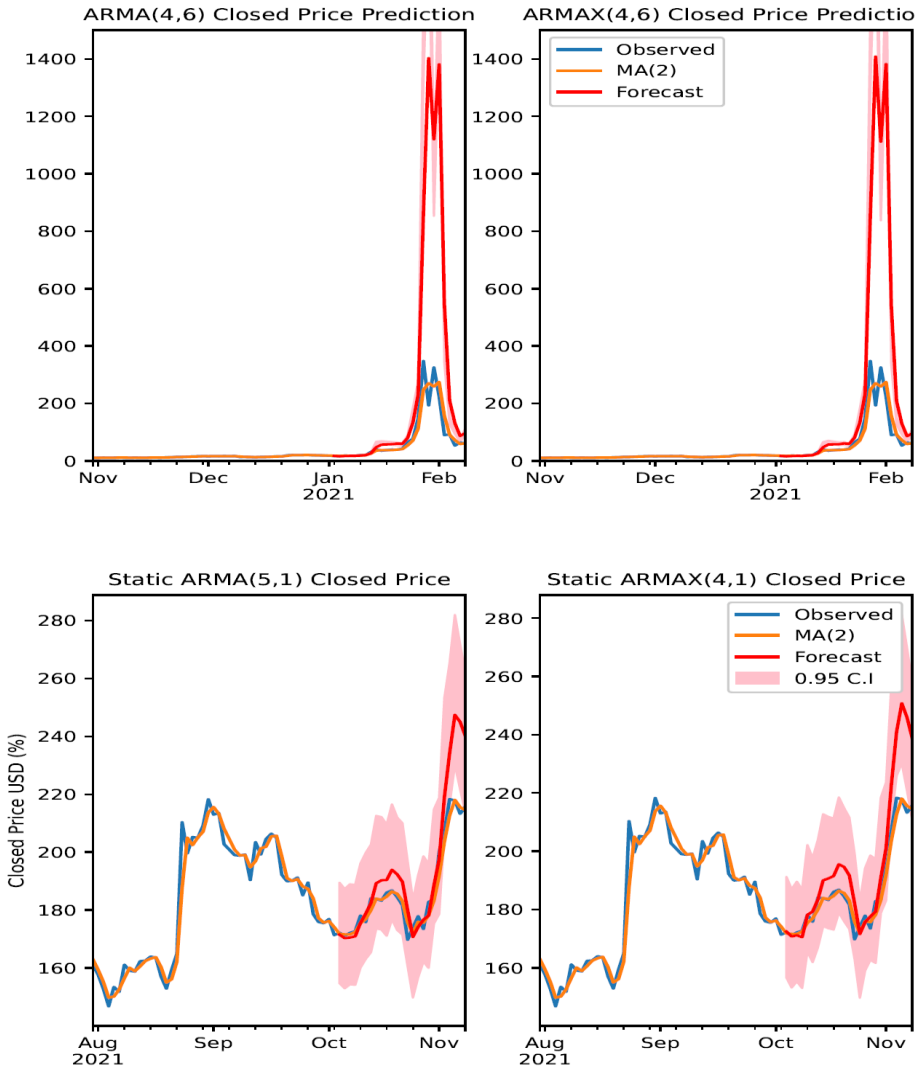


Figure 11 GME dynamic time series

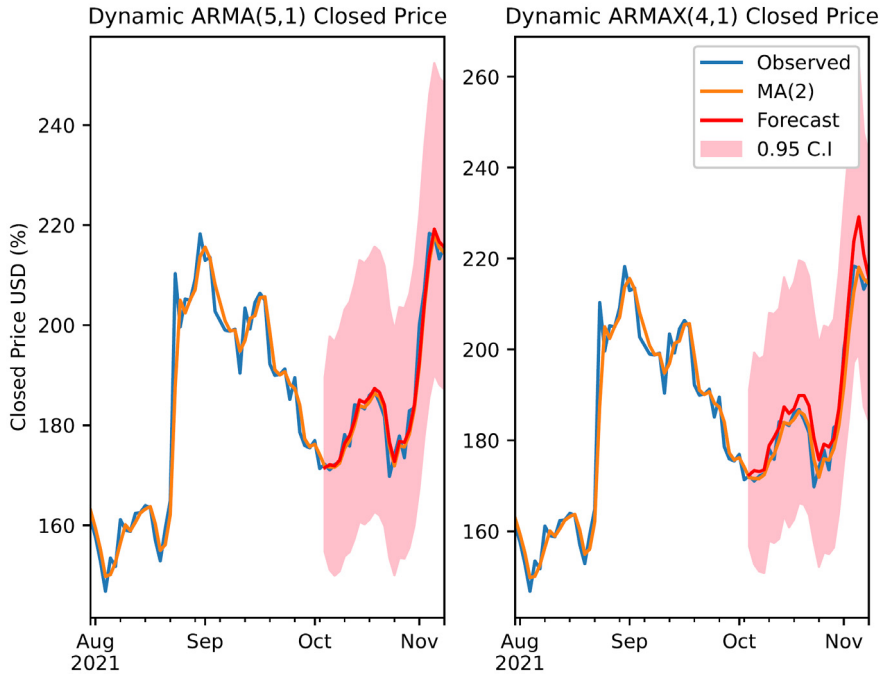
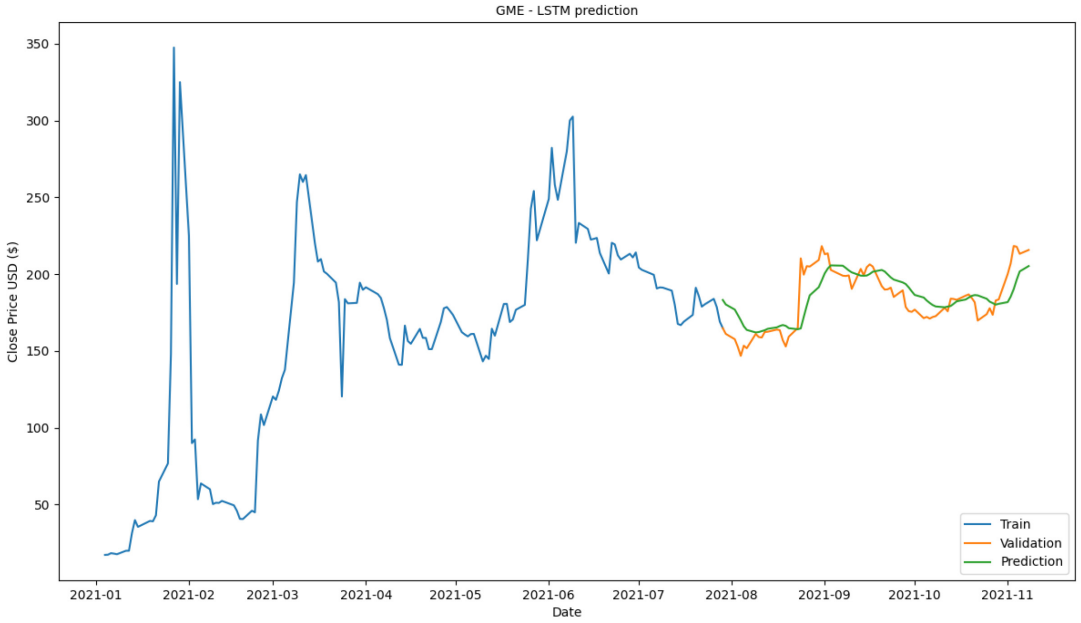


Figure 12 LSTM prediction



4.4 Evaluating model performance

Table 1 shows the breakdown of the model's performance across the four metrics. From this table, we can see the dynamic ARMA (5, 1) resulted the lowest value across all the metrics. The calculation is based on the last 25 entries of the actual and predictive values. Interestingly, the MAPE indicates that on average, the predictive values are out of 1.12% from the actual value. Also the MAE tell us the forecast error on average is $2–$3 per day. These results therefore need to be interpreted with caution.

Table 1 Evaluation metrics

Models	*MSE*	*RMSE*	*MAE*	*MAPE*
Prophet Default	1931.97	43.95	41.64	22.99%
Prophet Adjusted	269.92	16.43	11.77	6.01%
Static ARMA(5, 1)	134.46	11.60	7.91	3.98%
Static ARMAX(4, 1)	163.78	12.80	8.61	4.32%
Dynamic ARMA(5, 1)	7.94	2.82	2.11	1.12%
Dynamic ARMAX(4, 1)	21.23	4.61	3.53	1.87%
LSTM	173.85	13.19	10.19	5.52%

Both MSE and RMSE are not good indicators in evaluating the model performance if the data set contains outliers. These metrics are effected large outliers and puts more weight on larger errors resulting larger errors. In contrast, both MAE and MAPE are better indicator if the data set contains some outliers. Overall, it is important to correctly interpret the results

Limitations: One of the limitations requires users to have prior knowledge in forecasting. Otherwise it will lead to misleading results.

Another limitation is making a decision solely on prices. There are a number of factors affecting the prices, and the user should not always depend on forecasting models. Also, the most significant risk in forecasting is the change in trend.

One concern about the findings of evaluation metrics was that we did not use any optimisation methodology to find the best parameters for both Prophet and LSTM because it is computationally expensive. The parameters were adjusted accordingly by experimenting with combinations of different parameters and selecting the best results. Therefore, the ARMA was a favourable model because the parameters were optimised.

One primary source of limitation is the lack of testing data which represent less than 10% of the overall data. Therefore, the training data is overfitted. This is because the data needs to capture the latest trend.

Google Cloud provides $300 free credits for three months, after that will require future payments to continue with their products to connect to BigQuery. So, the web application will no longer connect to the cloud.

5 Conclusions and future work

This work presents a framework called GRAPES, which is predicting GameStop prices and introduces a semi-automatic approach for forecasting models using cloud computing and machine learning. From the result, the web application will only have two models, Prophet and ARMA.

We show that the models above can forecast GME prices. However, it has not been appropriately tested on other volatile stocks such as AMC. So, the reader should not be reliant on the model because it doesn't always predict success. Models like ARMA/ARMAX parameters do require to be updated regularly because psychological factors and consumer behaviour changes over time.

Although the Dynamic ARMA (5, 1) was statistically the best model, our preferred models would be Static

ARMA or Prophet Adjusted method because the MAE results. An average price change of between $8 and $9 seems more plausible than average price change of $2–$3.

However, for a quick turnaround results, Prophet would be the best model. Time Series method requires the data to be stationary before forecasting. In addition to this, if the series of data contain seasonality then there is more requirement to fit the model. Also LSTM is difficult to learn if you have no knowledge of neural network architecture. Furthermore, Prophet can handle missing data quite well, and produce powerful forecasting models by default.

Scrapping large amounts of WallStreetBets data exhausted our memory and it was time consuming. We were not able to extract the whole data set from January 2021 under 5 hours. However, by taking a snapshot of the data set, it requires special attention to be cleaned and normalised before effective analysis. Therefore, this will be discussed further in future work.

5.1 Future work

This section discusses the desirable future works.

5.1.1 WallStreetBets

This is still an interesting piece to analyse. However, unstructured data such as users' comments and images are challenging. It would be interesting to explore 'Name Entity Recognition' techniques to classify texts into categories such as organisation and others. This way, it can increase the accuracy of identifying organisation names instead of counting company names with cashtag, i.e., '$GME'. The organisation name can be in varying formatting such as game, GAMESTOP and misspelt words.

5.1.2 File storage

The amount of data created and replicated is increasing rapidly globally. It is estimated 175 zettabytes will be made by 2025 compared to 44 zettabytes in 2020 (Vuleta, 2021). The data collected for this project was in CSVs file format; however, they are not efficient to store data and are slow to query (Radecic, 2018). An alternative solution would be to keep the file in a parquet file format. It has been shown for an identical data set (1TB CSV), it is 16 times cheaper and faster to scan data in Parquet format (Radecic, 2018). Therefore, as we collect more data from different sources, it would be cheaper to store parquet files in Cloud.

5.1.3 Additional data

It would be interesting to see how models would perform on stocks from different free-float (public float), market capitalisation and sectors. For example, the free-float effectively is the number of shares available to trade in the public. So, a limited number of shares available with high short interest can lead to extreme volatility. Whereas, large-cap market (over $10 bn) such as AAPL, MSFT or GOOG are more stable and have better performance in predicting than the non-large cap. In addition to this, incorporating additional independent factors such as inflation, unemployment rate earning per share.

5.1.4 Social media

It has been seen that Elon Musk's tweets have so much influence in the stock market. Recently, he announced on Twitter to sell 10% of his shares which led to Tesla stock down by 12% (Kolodny, 2021). It could be interesting to follow CEOs, analysts and politicians. In addition to this, we could also analyse the Stocktwits platform too. It is one of the large communities of investors and traders to discuss stocks (Businesstimes.com, 2021).

5.1.5 Multivariate models

It would also be interesting if we could perform a multivariate analysis with other factors besides Volume shown in Figure 10. We could explore the open price, relative strength index and exponential moving average. Adding too many independent factors can make the modelling process more complex. However, a univariate forecast model may appear approach as it is straightforward to produce.

5.1.6 Airflow executors

When we start extracting more data from various sources, the data pipelines will be more complex. It would be interesting to see which executor types will perform well when executing multiple tasks in parallel. However, during this study, there are challenges when experimenting with different Airflow executors as it requires at least 8 GB to run.

5.1.7 Web application

The web application needs to optimise with cache (Streamlit, 2021). This way, the app can store information without rerunning the app. This will increase the application performance.

5.1.8 Infrastructure efficiency

The infrastructure efficiency result can be evaluated in the future. For example BigQuery vs. DataProc, or the execution time of each components can be analysed. In future, design options to design the data pipeline can be explored to make intelligent design decisions to lay the components.

5.1.9 Performance comparison

As GameStop prices prediction is emerging area and the COVID-19 situation makes it more unique as reported in the literature, though future researchers can compare the performance of proposed work with traditional price prediction models in terms of experiment results. For examples, the performance of proposed work can be compared with Generalised AutoRegressive Conditional Heteroskedasticity (GARCH), Geometric Brownian Motion and emerging AI/ML models to assess volatility in financial markets.

Acknowledgement

We gratefully acknowledge Google for their support of this research through their GCP research credit program.

References

Agrawal, A. (2021) *Loss Functions and Optimization Algorithms*, Demystified, Medium. Available online at: https://medium.com/data-science-group-iitr/loss-functions-and-optimization-algorithms-demystified-bb92daff331c (accessed on 19 November 2021).

Airflow (2021a) *Apache Airflow Documentation*, Airflow. Available online at: https://airflow.apache.org/docs/apache-airflow/stable/index.html

Airflow (2021b) *Architecture Overview*, Airflow. Available online at: https://airflow.apache.org/docs/apache-airflow/stable/concepts/overview.html

Airflow (2021c) *Executor*, Airflow. Available online at: https://airflow.apache.org/docs/apache-airflow/stable/executor/index.html

ARK Fund (n.d.) *ARK invest fund holding*. Available online at: https://ark-funds.com/download-fund-materials/

Beers, B. (2021) *How after-Hours Trading Affects Stock Prices*, Investopedia. Available online at: https://www.investopedia.com/ask/answers/05/saleafterhours.asp (accessed on 7 November 2021).

Bento, A. (2020) *Airflow: How To Refresh Stock Market Data While You Sleep Part 1*, Towards Data Science. Available online at: https://towardsdatascience.com/airflow-how-to-refresh-stocks-data-while-you-sleep-part-1-a464514e45b7 (accessed on 19 November 2021).

Businesstimes.com (2021) *Pump and Dump Finds a New Home – on Social Media and Chat Forums*, The Business Times. Available online at: https://www.businesstimes.com.sg/stocks/pump-and-dump-finds-a-new-home-on-social-media-and-chat-forums (accessed on 15 November 2021).

ETFs (2021) *Largest ETFs: Top 100 ETFs By Assets*, ETFDB. Available online at: /urlhttps://etfdb.com/compare/market-cap/ (accessed on 19 November 2021).

Facebook.github.io (n.d.) *Seasonality, Holiday Effects, and Regressors*, Prophet. Available online at: https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html#built-in-country-holidays

Fox, M. (2021) *ARK's Cathie Wood made a monster call in 2018 that Tesla stock would hit $4,000*, Her prediction just came true 2 years early, Markets Insider. Available online at: https://markets.businessinsider.com/news/stocks/tesla-stock-analysis-cathie-wood-ark-prediction-just-came-true-2021-1 (accessed on 6 October 2021).

Ghosh, S. (2021) *Reddit Group WallStreetBets Hit 6 Million Users Overnight after a Wild Week of Trading Antics*, Business Insider. Available online at: https://www.businessinsider.com/wallstreetbets-fastest-growing-subreddit-hits-58-million-users-2021-1?r=US IR=T (accessed on 6 November 2021).

Gill, S.S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., Golec, M., Stankovski, V., Wu, H., Abraham, A. and Singh, M. (2022) 'AI for next generation computing: emerging trends and future directions', *Internet of Things*, Vol. 19, pp.1–36.

Google (2021a) *Storage*, Google. Available online at: https://cloud.google.com/products/storage/

Google (2021b) *BigQuery*, Google. Available online at: https://cloud.google.com/bigquery

Google (2021c) *DataProc*, Google. Available online at: https://cloud.google.com/dataproc/docs/concepts/overview

Google (2021d) *DataProc Scaling*, Google. Available online at: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters

Ho, K. (2021) *A Misinterpreted Elon Musk Tweet Sent an Obscure Stock Soaring*, Quartz. Available online at: https://qz.com/1956105/elon-musks-tweet-about-signal-boosts-shares-of-the-wrong-company/ (accessed on 7 November 2021).

Kolodny, L. (2021) *Tesla Drops 12% for Biggest Fall this Year after Musk Stock Sale Proposal*, CNBC. Available online at: https://www.cnbc.com/2021/11/09/tesla-shares-drop-as-sell-off-accelerates.html (accessed on 15 November 2021).

Letelier, M. (2021) *Could Machine Learning Have Predicted GameStop Madness*?, Towards Data Science. Available online at: https://towardsdatascience.com/could-machine-learning-have-predicted-gamestop-madness-aae8d9f7f77e (accessed on 18 November 2021).

Li, Y. (2021) *GameStop Short Sellers are still not Surrendering Despite Nearly 20 Billion in Losses This Month*, CNBC. Available online at: https://www.cnbc.com/2021/01/29/gamestop-short-sellers-are-still-not-surrendering-despite-nearly-20-billion-in-losses-this-year.html (accessed on 6 November 2021).

Martin, K. and Wigglesworth, R. (2021) *Rise of the Retail Army: the Amateur Traders Transforming Markets*, Financial Time. Available online at: https://www.ft.com/content/7a91e3ea-b9ec-4611-9a03-a8dd3b8bddb5 (accessed on 1 October 2021).

Marx, D. (2018) *PSAW*. Available online at: https://pypi.org/project/psaw (accessed on 8 October 2021).

Marx, J. (2020) *Creating a Financial Dashboard Using Python and Streamlit*, Towards Data Science. Available online at: https://towardsdatascience.com/creating-a-financial-dashboard-using-python-and-streamlit-cccf6c026676 (accessed on 19 November 2021).

Mooney, P. (2021) *Stock Market Data (NASDAQ, NYSE, S&P500)*, Kaggle. Available online at: https://www.kaggle.com/paultimothymooney/stock-market-data

Nau, R. (2014) *Principles and Risks of Forecasting*, Duke University. Available online at: https://people.duke.edu/ rnau/411diff.htm (accessed on 19 November 2021).

Phung, A. (2021) *What Days Are the U.S. Stock Exchanges Closed*?, Investopedia. https://www.investopedia.com/ask/answers/06/stockexchangeclosed.asp (accessed 7 November 2021).

Radecic, D. (2018) *CSV Files for Storage? No Thanks. There's a Better Option*, Towards Data. Available online at: https://towardsdatascience.com/csv-files-for-storage-no-thanks-theres-a-better-option-72c78a414d1d (accessed on 6 November 2021).

Rink, K. (2021) *Time Series Forecast Error Metrics You Should Know*, Towards Data Science. Available online at: https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27 (accessed on 6 November 2021).

Schmitt, M. (2021) *Streamlit vs. Dash vs. Shiny vs. Voila vs. Flask vs. Jupyter*, Data Revenue. Available online at: https://www.datarevenue.com/en-blog/data-dashboarding-streamlit-vs-dash-vs-shiny-vs-voila (accessed on 20 October 2021).

Srivastava, P. (2021) *Apache Spark on DataProc vs. Google BigQuery*, Sigmoid. Available online at: https://www.sigmoid.com/blogs/apache-spark-on-dataproc-vs-google-bigquery/ (accessed on 19 November 2021).

Statista (2020) *Total net asset under management (AUM) of Exchange Traded Funds (ETFs) in the United States from 2002 to 2020*, Statista. Available online at: https://www.statista.com/statistics/295632/etf-us-net-assets/ (accessed on 19 November 2021).

Streamlit (2021) *Optimize Performance with st.cache*, Streamlit. Available online at: https://docs.streamlit.io/library/advanced-features/caching (accessed on 20 November 2021).

Vishnoi, A. and Bloomberg (2021) *Cathie Wood's Ark ETFs Doubled Down on Robinhood after it Announced its Massive Data Breach*, Fortune. Available online at: https://fortune.com/2021/11/10/cathie-wood-ark-etf-buying-robinhood-shares-data-breach/ (accessed on 19 November 2021).

Vuleta, B. (2021) *How much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*, SeedScientific. Available online at: https://seedscientific.com/how-much-data-is-created-every-day/ (accessed on 6 November 2021).

WallStreetBets (2021) *WallStreetBets*, Reddit. Available online at: https://www.reddit.com/r/wallstreetbets (accessed on 6 October 2021).

Wikipedia (2021) *Cathie Wood, Wikipedia*. Available online at: https://en.wikipedia.org/wiki/Cathie_Wood (accessed on 5 October 2021).

Xin, A. (2021) *Forecasting GameStop Stock Price using Time Series Analysis*, Medium. Available online at: https://medium.com/codex/forecasting-gamestop-stock-price-using-time-series-analysis-794e20b1236d (accessed on 18 November 2021).

Zajic, A. (2019) *Introduction to AIC – Akaike Information Criterion*, Towards Data Science. Available online at: https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced (accessed on 19 November 2021).