A multi-variate statistical analysis of hedge fund asset suppression and community response to meme stocks with Machine Learning

Rohit Rangaraj
Lumiere Education
rohitrangaraj2005@gmail.com

Special thanks to Samuel Showalter from The University of California, Irvine, for his endless support and guidance throughout the project.

Abstract

The stock market refers to the collection of markets and exchanges where regular buying, selling, and issuing shares of publicly-held companies occur. This ideally is a very efficiently priced system where all commodities are priced fairly, but investors find inefficiencies in the market and capitalize on these inefficiencies. In this paper, we look at a newer kind of stock called "meme stocks". These stocks rely entirely on the community for their survival in the market. This paper will explore the factors and parameters on how two of these stocks, $GME, and $AMC, had their push and how it might affect the market as a whole. This is an analysis of the "short-squeezed" $GME and the parameters which caused it, and a prediction of how an upcoming stock, $AMC, might follow the same path if said parameters are satisfied. We also look upon how each kind of market and social parameter has the most significant impact on the stock and observe how data like short interest, utilization are crucial for the movements of the stock. We also predict that $AMC will have a similar movement as $GME as we are able to look at straightforward correlations between the parameters and the stock.

1. Introduction

1.1. Machine Learning for Financial Analysis

"Past results are no guarantee of future performance."

We go by this very well-known quote in the stock market, but past data is very useful in predicting what might happen. With vast amounts of data appropriate to our needs, almost anything can be predicted to assist human traders in making better decisions and trades. Machine Learning has been growing rapidly in the finance market because of the pure statistical and mathematical advantage of a machine crunching terabytes and petabytes of data in a small amount of time which was not possible for humans to do unassisted.

1.2. Meme Stocks

The term *meme stocks* refer to a particular class of stock where the fundamentals of the underlying company are so bad that huge market-makers and hedge funds start entering into huge short positions. This basically forces vulnerable companies to lose all their market capitalization, thus making the companies bankrupt. This trend of excessively shorting companies became much more prominent in the market crash during the COVID-19 Pandemic as most companies were losing a huge part of their revenue, and companies that didn't have a huge online presence, suffered huge losses due to the extended lockdown periods. This provided an opportunity for the short sellers to double down on their positions as they predicted that these companies which weren't good businesses in the modern era to go even lower in price, and ultimately go into bankruptcy. But, retail investors started noticing patterns in these stocks and started to rally a movement in which they bought shares of companies to oppose against the institutional traders who basically controlled the market's every move. Some retail investors saw some hope for companies which hedge funds considered to ultimately go bankrupt and sizeable amounts of induvidual investors started to delve deeper into analysing each part of their financials and their possible plans for the future. One such Reddit (discussed in 1.3) user, *Keith Gill* started sharing his thoughts on $GME. He theorized how it wasn't as bad as the institutions were thinking and how it had a potential come back in terms of financials like revenue. He began posting extremely detailed analysis on the stock and potential strategies for the company to succeed in the current market. He also entered into huge long positions on the share when it hit the all time low of $2.57 in March of 2020 and started posting monthly updates on his share positions and his options contract plays. He began this by investing nearly $50,000 and by the end of the GameStop Short Squeeze, he walked out of this trade with nearly $40 million in his pockets. This lured in a lot more retail investors to pick up trading as a productive hobby and as a learning oppurtunity while people had been forced to stay inside due to the pandemic. At the same time, a platform named Robinhood, started providing zero-commission trading thus allowing heaps of people to invest in the stock market. This huge wave of retail investors enabled very high buying pressure in meme stocks and resulted in the GameStop Short Squeeze.

1.3. Reddit and r/wallstreetbets

The community behind these meme stocks is mainly from a social media forum called Reddit. This is a place where thousands of topic-specific forums exist, called "subreddits". One such subreddit is r/wallstreetbets; this internet forum was behind the biggest community uprising against the hedge funds on Wall Street, i.e., the GameStop Short Squeeze. r/wallstreetbets is filled with posts putting up the huge gains users have attained by leveraging in very high ratios and "YOLOing" (Using all your money in a single trade) them in risky trades. Spectators of these posts get motivated by these gains and try to invest a little money on their own, growing this community exponentially day by day. Recently, the subreddit has been filled with memes, detailed due diligence on market trends by financially talented users and the positions on said stocks. Stock specific subreddits have also branched off the main r/wallstreetbets to places like r/amcstock, r/GME, r/superstonk, etc.

The paper's objective is to use the data from financial sources for price movements, market trends and indicators and to scrape data off sites like Reddit to analyse the impact of the community on these highly volatile moves. We will look at how each social and financial parameter influenced the huge moves in these meme stocks. We have used a Recurrent Neural Network with LSTMs to analyse and predict the movements of the stock. We will discuss how we proceed to this in the upcoming sections.

2. Related Work

2.1. People who took part in the Retail Investor Movement

From the paper, "Hasso, Muller, Pelster, Warkulat, Who participated in the GameStop frenzy? Evidence from brokerage accounts", we can have a clear view at how retail investors who took part in the meme stock frenzy, had their portfolios looking prior to 2020 and how a *personality* changes their financial decisions and how they have reacted to previous market trends [1]. This gives us more insight into what kind of people took part in this movement against the market-makers. We also look at how these *ultra-risky* trades worked out for them and how their portfolios have performed at the conclusion.

2.2. Small factors that have played a huge role in the meme stocks

In the Paper, "Hu, Danqi and Jones, Charles M. and Zhang, Valerie and Zhang, Xiaoyan, The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery", the authors talk about how the crazy movement of the meme stocks have affected the market as a whole and they look up on new metrics such as the Robinhood 50 stocks which the platform had *restricted* trading due to the crazy market movements and the settling costs the platform had to endure [2]. This paper gives us a clear idea of how things would have been if small changes are made to the market's environment. The authors also clearly summarizes how the market has responded to each factor which will provide the required weightage for each parameter in the grand equation.

2.3. Community Sentiment and effect on price

In the paper, "Long, Cheng and Lucey, Brian M. and Yarovaya, Larisa, 'I Just Like the Stock' versus 'Fear and Loathing on Main Street': The Role of Reddit Sentiment in the GameStop Short Squeeze", the authors explore how the Reddit sentiment brought about unnatural momentum in an otherwise dying company [3]. This paper shows the impact of daily activity on r/wallstreetbets directly correlating with the price movements and the intraday volume of $GME. They achieve this by using Natural Language Processing models found in the SentimentIntensityAnalyser and Text2Emotions packages in Python to analyse over 10.8 million comments on r/wallstreetbets. This however doesn't create a concrete set of parameters, which can then be used as a benchmark for upcoming meme stocks like $AMC, to predict their trends, movements and to determine whether the communal support provided by Reddit will be as influential in their pricing in the long term as it was with GameStop.

3. Experimental Design

3.1. Dataset Collection and Data Cleaning

The data for the analysis in this paper can be classified into two categories:

- **Financial Data**
- **Reddit Data**

3.1.1 Financial Data

The financial data is all the price movements, market trends, indicators, technicals from the following data sources and websites:

- www.alphavantage.com
- www.finance.yahoo.com
- www.alpacafinance.org

This will help us in analyzing how the market moves from a completely statistical perspective.

We use the following primary indicators, fundamentals and metrics:

- **Price**
- **Volume**
- **Average Volume**
- **Short Interest**
- **Short Volume**
- **Utilization**
- **Shares on Loan**
- **Aroon**
- **Bollinger Bands**
- **Relative Strength Index (RSI)**
- **Exponential Moving Average (EMA)**
- **Volume-weighted Average Price (VWAP)**
- **Moving Average Convergence/Divergence**
- **Chaikin Accumulation-Distribution**
- **On Balance Volume (OBV)**
- **Earnings Per Share (EPS)**
- **Earnings Estimate**
- **Earnings Difference and Surprise**
- **Analyst Ratings**

This particular set of data may be more useful to said *meme stocks* as these are heavily shorted and are highly volatile, therefore we give more weightage to the momentum indicating oscillators and importance to metrics like Short Interest, Utilization and Shares on Loan as they indicate how heavily the stock is being shorted and we could interpret the community's fightback from the performance of price with such selling pressure.

3.1.2 Reddit Data

On the other hand, the Reddit data will help us determine how influential the community has been in a particular period, affecting the community's momentum and morality on bearish days and hyping up the community on bullish days. Majority of the data is from www.subredditstats.com as it provides an accurate measure of the number of people active on each subreddit that we are interested in.

The subreddits we analysed are :

- **r/wallstreetbets**
- **r/amcstock**
- **r/gme**
- **r/superstonk**
- **r/spce**
- **r/pltr**
- **r/investing**
- **r/stocks**
- **r/stockmarket**

We also analysed the peak activity times on these subreddits and infer any impact on the market at that particular period of time.

Most of the data we use is pre-cleaned and processed as it's a huge area of research and as higher implementations of Machine Learning and Artificial Intelligence are used in the financial markets, data collection and cleaning have become exponentially easier. But, nothing is perfect and we did come across unavailable or inaccessible data. Therefore we have taken up usual missing data imputation methods to substitute the values to the closest possible approximations. One such approximation was used for the member count of subreddits that went private for a period of time (making the count unavailable). For this particular case, a stochastic regression imputation was used to impute the values. We also tried various scaling methods to standardize our inputs and the varying results will be published in the Results and Discussion section.

3.2. Overview of predictive model

We have used a Recurrent Neural Network (RNN) to analyse the price movements of $GME and use the factors to develop a set of parameters to predict other stocks future movements. We implemented LSTM (Long Short Term Memory) layers to create a recurrent neural network. The network was then trained on historical data collected in the section 3.1. As we will be analysing various different input values and different features to predict and analyse the stock movements, we used a multi-variate time series forecasting approach. The activation function is the *tanh* function and the optimizer is the Adam optimizer. The following parameters were tuned to achieve the best results: The number of hidden layers, number of neurons in each layer, number of epochs, batch size and learning rate.

3.3. Variations of Experimental Parameters

To predict which features will be most significant in predicting the stock movements, we split the training data into 4 to state the significance of each feature.
The four subsets are :

- **Basic Price Data**
- **Meme Stock Specific Data**
- **Technical Indicators**
- **Company Financials**

3.3.1 Basic Price Data

This subset of data is the most important as it contains the price movements of the stock and the volume of the stock. Training without feeding these important features will result in a model that will not be able to have any idea of the movements of the stock. So we don't remove this from any subset of the training data.
The data points used are :

- **Daily Open**
- **Daily Close**
- **Daily High**
- **Daily Low**
- **Daily Change %**
- **Daily Volume**
- **Average Volume (Past 10 Trading Days)**
- **Volume Weighted Average Price**

3.3.2 Meme Stock Specific Data

This subset of data is only relevant to the stocks we are analysing as they are not of much use to regular stocks but are of great significance to the meme stocks. This is the case because as we established earlier (Section 1.2), the meme stocks are highly shorted and are highly volatile and have a huge community following. Parameters here contribute to the phenomena like *Short Squeeze* or a *Gamma Squeeze*.

The features we used are :

- **Short Volume**
- **Short Percentage**
- **Free Float %**
- **Shares Outstanding**
- **Utilization %**
- **Short Interest**
- **Shares on Loan**
- **Cost to Borrow**
- **Subreddit Active Users**
- **Subreddit Members**
- **Reddit Post Upvotes**
- **Reddit Post Downvotes**
- **Reddit Post Count**

We compared the results **with** and **without** this split of data to signify the impact of the features on the stock movements.

3.3.3 Technical Indicators

We also included highly used technical indicators as trainable features to more accurately predict the market. The selection will also be featuring a lot of oscillators and momentum indicators as these stocks are notoriously volatile.
The indicators we used are :

- **Aroon**
- **Bollinger Bands**
- **Relative Strength Index (RSI)**
- **Simple Moving Average (SMA)**
- **Exponential Moving Average (EMA)**
- **Stochastic Oscillator (STO)**
- **Momentum**
- **Williams %R**
- **Moving Average Convergence / Divergence**
- **Chaikin Accumulation - Distribution**
- **On Balance Volume (OBV)**

These features were also be split from the training data to analyse the impact of the indicators on the predictions.

3.3.4 Company Financials

This is the section where the biggest problems arise as the company financials are not very good and it is hard to predict how the company might improve sales numbers or operating principles. This is also debated by professional analysts, but we included a healthy set of analyst ratings and predictions to make the model more accurate.
The financials we used are :

- **Earnings Per Share (EPS)**
- **Market Capitalization**
- **Earnings Estimate**
- **Earnings Difference**
- **Earnings Surprise %**
- **Analyst Ratings**
- **Short Term Debt**
- **Long Term Debt**
- **Cash Flow**
- **Revenue**
- **Post Tax Earnings**
- **Total Assets**
- **Total Liabilities**
- **Total Stockholders' Equity**
- **Book Value Per Share**

This will determine how a poorly performing company might have a chance of improving in the future.

4. Results and Discussion

In this section, we'll discuss the results of the model and the significance of each subset of data we used to train.

4.1. Results without Meme Stock Specific Data

In this subset of data, we will be using only the basic price data and key parameters from Sections 3.3.3 and 3.3.4 without including the data discussed in Section 3.3.2. As shown in Figure 1, the results are not accurate at all. The model is not able to predict the future price of the stock with just the basic price data. We will also test the changes in predictions when we feed the model with the meme stock specific data in Section 4.2.

4.2. Results with Meme Stock Specific Data

Once we have included the meme stock specific data, the model is able to predict the future price of the stock with a decent degree of accuracy. The results of this model is shown in Figure 2. This shows that this subset of data is very crucial to our predictions.

4.3. Results without Technical Indicators

In this subset of data, we will be using only the basic price data and key parameters from Sections 3.3.2 and 3.3.4 without using any technical indicators. As shown in Figure 3, the results are not perfectly accurate but the model is predicting with decent accuracy. We will now check the change in accuracy once we include the technical indicators in our models.

4.4. Results with Technical Indicators

After including the technical indicators, the model is able to predict the future price of the stock with a better accuracy. As we can see in Figure 4, eventhough the subset has improved the accuracy, the significance of adding indicators weren't as huge as results in Section 4.1.1 and 4.1.2.

4.5. Results without Company Financials

In this variant of the model, we will be using only the basic price data and key parameters from Sections 3.3.2 and 3.3.3 without using any financial data. As shown in Figure 5, the model predicts the price to be very varied from the actual price. This is due to the poorly performing financials of the company and how the improvement in financials might prove very significant to the market capitalization of $GME.

4.6. Results with Company Financials

Once we include the financials, the model is able to predict the future price of the stock with a far better accuracy. The results of this model are shown in Figure 6. It also indicates the price predictions are far lower than the one without the financials and thus adding to our point of poor performance of the company.

4.7. Results with Complete Superset of data

The results of the model with the complete superset (as shown in Figure 7) show us that the overall dataset collection and model training were very accurate. The mean squared error is just *0.0006129711337196139* , so the model is able to predict the stock movements with a high degree of accuracy.

4.8. Comparison and Relatability to upcoming stocks

The data analysis is on the GameStop data as it is more complete and there has been a clear short squeeze in $GME. This analysis will give us insight into the parameters used to determine whether a stock has a strong communal impact. These parameters can be used as a benchmark for upcoming stocks to determine whether they have a similar movement to $GME.

5. Conclusion and Further Research

The experiments we have done on the GameStop data show that the model is able to predict the stock movements with a high degree of accuracy. It is also able to state the significance of the various parameters used in the model.

As we can see from the results, the model's accuracy is very volatile and very dependent on some subsets of data. This is because the model is not able to predict the future price of the stock with just the basic price data or just the technical indicators as any normal stock would indicate. The predictions are very dependent on the meme stock specific data and financials. This shows us that the parameters like short interest, utilization, etc. are very important to the stock price. Without including key stats in the meme stock specific data, the model simply cannot understand or predict the stock movements. This further proves our point that asset suppresion from the hedge funds plays a huge role in the stock price of meme stocks.

We also realize that this model is not perfect and significance of parameters might be different for different meme stocks. So, we propose many more research ideas to continue to improve our understanding of the stock market and specifically the meme stock market.

One such idea could be to analyze the movements of $AMC as we go ahead and predict if we could see a simliar movement of price as seen in $GME. We also want to see if we could develop a concrete model to predict if a stock has similarities and if it could be classified as meme stocks.

Our research adds up to a lot of work and we hope to continue to improve our understanding of the meme stock market as it is a very complex and dynamic market. We believe that the model has provided us with a good starting point to understand the meme stock market and we hope to see many more research ideas to continue to improve our understanding of the meme stock market and provide the induvidual retail investor a competitive advantage over the huge market makers and hedge funds to make it a fair playing field.

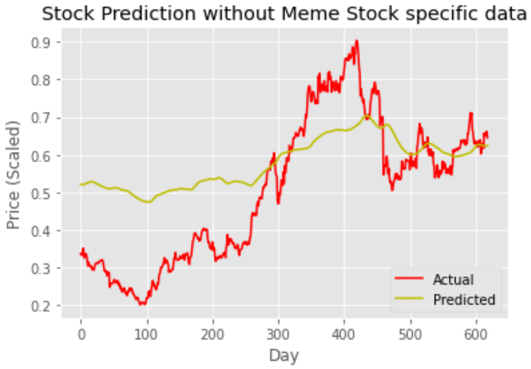

Figure 1: $GME Prediction without data from 3.3.2

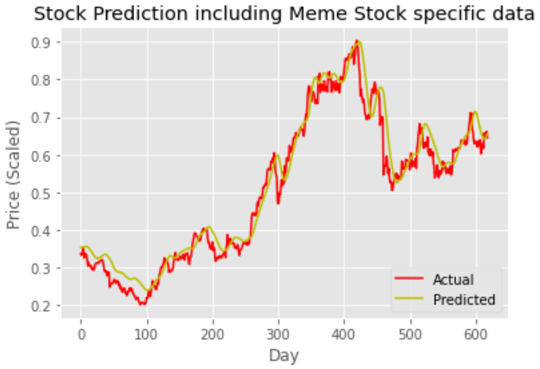

Figure 2: $GME Prediction including data from 3.3.2

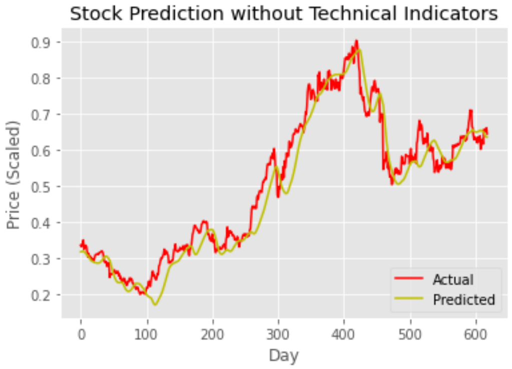

Figure 3: $GME Prediction without data from 3.3.3

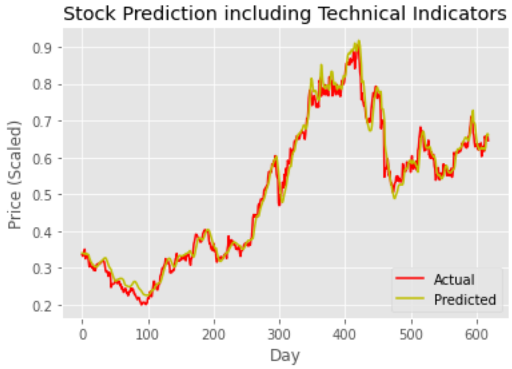

Figure 4: $GME Prediction including data from 3.3.3

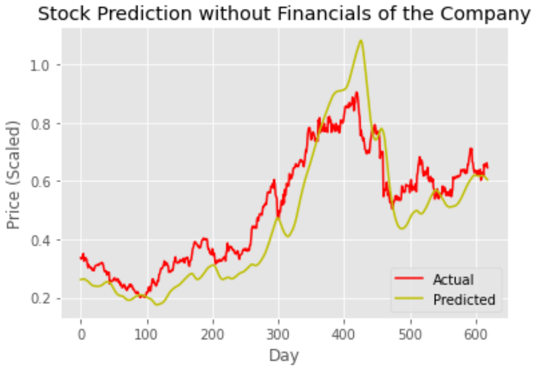

Figure 5: $GME Prediction without data from 3.3.4

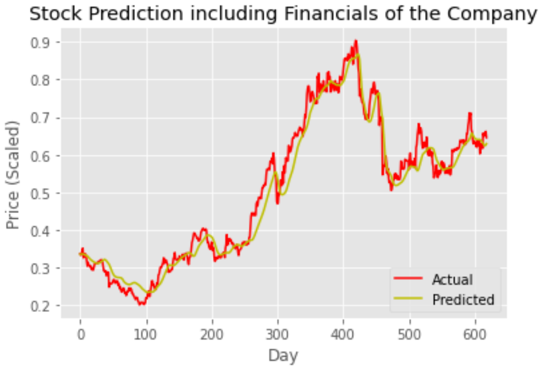

Figure 6: $GME Prediction including data from 3.3.4

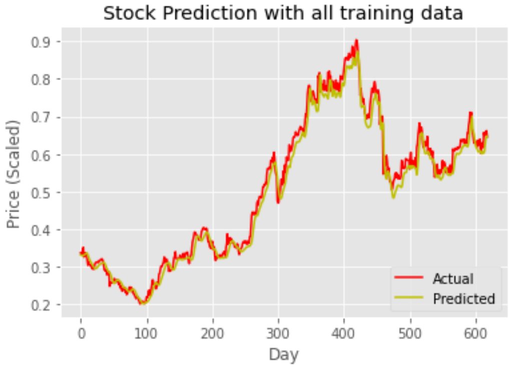

Figure 7: $GME Prediction with the complete superset of training data

References

[1] Tim Hasso, Daniel Müller, Matthias Pelster, and Sonja Warkulat. Who participated in the gamestop frenzy? evidence from brokerage accounts. *Finance Research Letters*, page 102140, 2021. 2

[2] Danqi Hu, Charles M. Jones, Valerie Zhang, and Xiaoyan Zhang. The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. 2021. 2

[3] Cheng Long, Brian M. Lucey, and Larisa Yarovaya. I just like the stock versus fear and loathing on main street : The role of reddit sentiment in the gamestop short squeeze. 2021. 2