UCD PROJECT: CAN SOCIAL MEDIA AFFECT A SHARE PRICE?

Brian Higgins | 1st June 2022 | UCD Project

GITHUB URL

https://github.com/bhiggi01/UCD2

ABSTRACT

This project aims to answer if 'social media can affect the share price of a company?' We will be looking at the Share Price for Game Stop, an American video game, consumer electronics and gaming merchandise retailer. Could users on Social Media decide to buy Game Stop shares in a bid to raise the price?

This project has real-world applications as there are over 5.8 billion users of social media in the world. It has never been easier for people to connect and share views and opinions. If even a small percentage of these users decided to buy a share they could affect the price.

INTRODUCTION

On the 1st of January 2021, the Share Price of Game Stop was worth 17us dollars. By the 29th of January, the price had increased to 325us with some daily highs reaching over 500us. This was an incredible 1800% increase in the share price. All this is from a company that had been increasingly struggling with a changing industry as traditional physical stores have lost ground to online stores such as Amazon. Why caused this spike?

This huge jump in share price was due to a campaign that started on Reddit by users to buy and hold Game Stop shares with two aims, first to raise the share price and second as a way to punish Short Term Wall Street Traders who bet on the price of shares to go up or down. With the 1800% increase in Game Stop Share Price, many short term traders initial bought the stocks as they expected the shares to go down. However, Reddit users held on to their shares and encouraged other users to also buy and hold Game Stop shares which further raised the price. The Short Term traders were left in a situation where they had to offload their shares at huge losses which added up to more than 13 Billion US. We will be looking at this activity in more detail to see if we can create Machine Learning Models to capture this activity.

DATA SET

REDDIT SOCIAL MEDIA DATA

14 different subreddits with a financial label that have been scrapped from the Social Media Site Reddit and combined to make one large dataset of over 1.5 million entries for the year 2021. These subreddits are used by users to discuss various stocks and shares, what to buy and what investments to make. This data has been sourced from Kaggle but originally Reddit (1). The Reddit data is made up of 24 different columns expected from a social media site. Only 6 of these columns are of use; 'author', 'created' (date), 'title' (comment), 'score' (up_vote Vs down_down), 'upvote_ratio' and 'Num_comments' (amount of comments).

GAME STOP SHARE PRICE

The second dataset was the Share Price of the Game Stop from the 'New York Stock Exchange' sourced from Kaggle (2). Information includes the 'Open', 'Close', 'Low', 'High' and 'Adjusted Close' prices, as well as the 'Date' and 'Volume' of sales.

IMPLEMENTATION PROCESS

DATA PROCESSING

With the Reddit data, we combined the datasets, removed unwanted columns and kept the 6 mentioned columns, checked for null and duplicate values, changed some object types to more useful types, changed the comment text to lower case and took a smaller dataset with only game stop mentions which had 150,505 entries.

With the Share Price data, we took a time range for the year 2021 to match the Reddit data, removed the 'Adjusted Close', and changed some object types. We then joined both data sets together on their Date columns to create a combined Dataset.

EXPLORATORY DATA ANALYSIS

Looking at the Number of Game Stop mentions made by users and the Share Price we can see a clear correlation in Fig.2 between the Share Price (red line) and the Number of Game Stop Mentions (blue line). The number of Game Stop Mentions on Reddit Data peaked with over 50,000 mentions at the same time the share price reached over 325us.

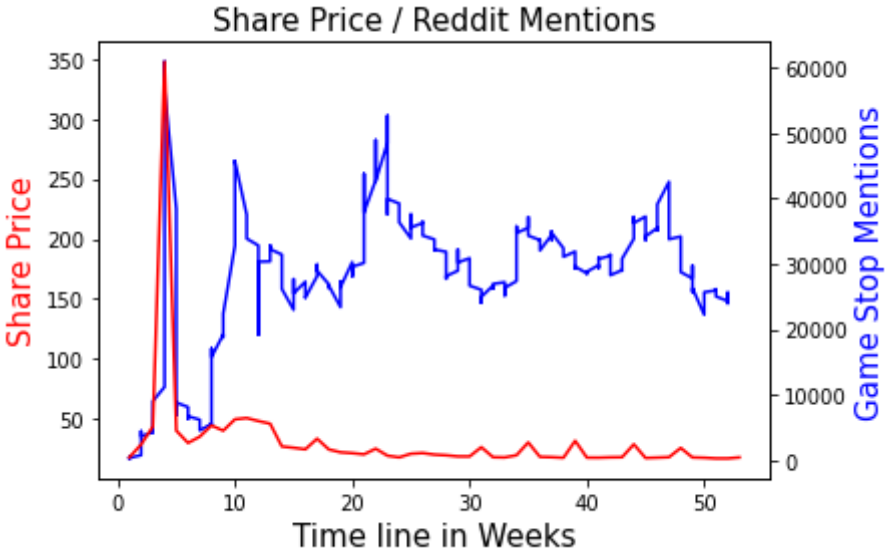

FIGURE 2: SHARE PRICE / REDDIT MENTIONS

Interestingly the share price continues to fluctuate but the number of Reddit comments made about Game Stop decreases over the year.

'CORRELATION DOES NOT MEAN CAUSATION

There are many examples of links between separate things so just because two separate things happen at the same time does not mean they are linked. To prove there is a connection we will need to be able to model this link with Machine Learning Models.

NATURAL LANGUAGE PROCESSING (NLP)

We used the 'Bag of Words' operation to look for the most common words in the title comments in the Reddit data. In Fig 3 we can see that the GME and Game Stop are the most common words found in the Reddit comments, this is not surprising as these were how we took a smaller sample of the larger 1.5 million entries.

Some other words are quite interesting; 'buy', 'hold', 'stock', 'moon', 'short', 'market' are all words that would be associated with shares.

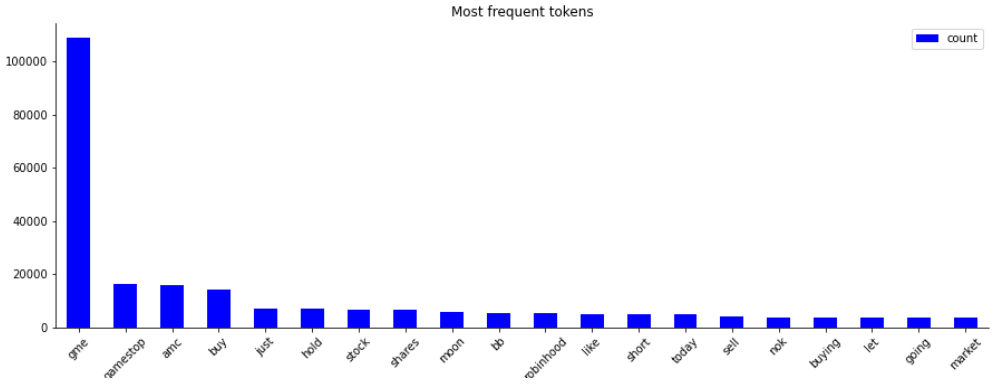

FIGURE 3: MOST FREQUENT TOKENS (WORDS)

MACHINE LEARNING MODELLING

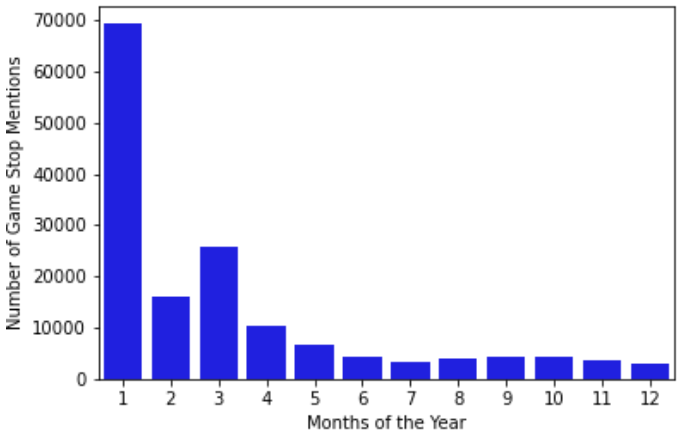

FIG 4: GAME STOP MENTIONS BY MONTH NUMBER

As you can see from fig 4, the majority of our Game Stop mentions in the Reddit data are in the first 2 months. Since we are looking to predict the price of a Share we are also using time as a factor to arrange our data into Train and Test datasets. If we did a normal 0.20 split of the test data then that spilt would need to be

6 months for the Train and 6 months for the test, which is not ideal. So we tested on three different time ranges.

- A – Train Data: 10 months / Test Data: 2 months
  Train 95% Data / Test 5% Data
- B – Train Data: 8 months and Test Data: 4 months
  Train 91% Data / Test 9% Data
- C - Train Data: 6 months and Test Data: 6 months
  Train 80% Data / Test 20% Data

MACHINE LEARNING MODELS

During the Project we looked at the following Machine Learning Models:

- Linear Regression
  - Ridge Regression
  - Lasso Regression
- Random Forest
- XG Boost

RESULTS

We looked at two metrics to compare our models. Accuracy and how well it predicted the Share Price. Unfortunately, it was very difficult to get good results for either metric. This is not surprising given the subject of the project and the different factors that can affect a share price. Large Hedge Funds have been working on this for years with little results.

ACCURACY OF MODELS

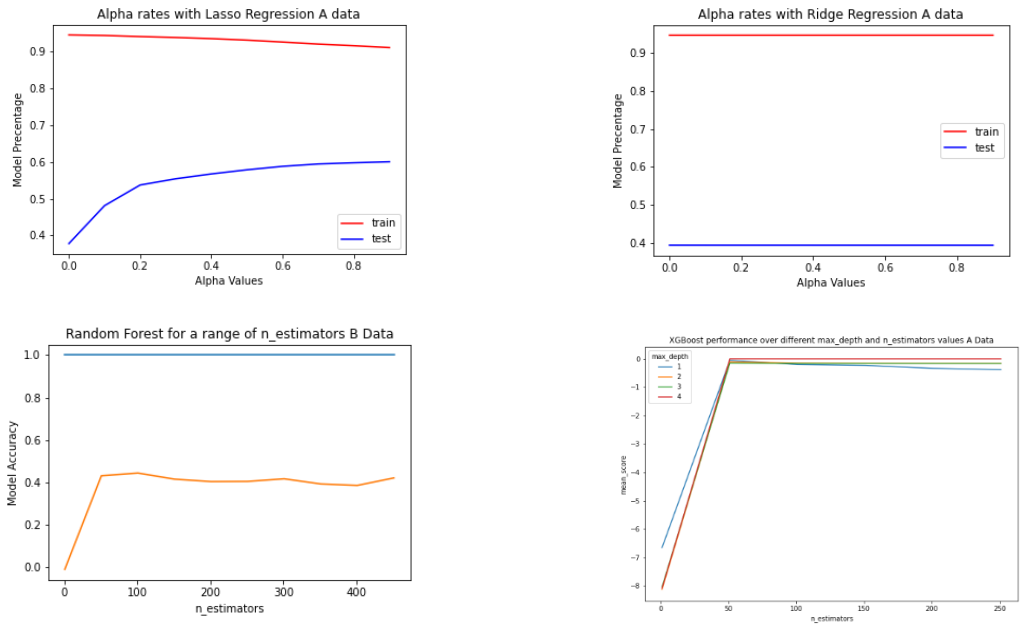

FIG 5: DIFFERENT ACCURACY MODELS FOR THE TRAIN AND TEST DATA

In Fig 5 we can see different models that were used to get the Train and Test Accuracy. All the best models ran on the A Train/Test data sets. The Train Accuracy was always in the high 90's but the Test Accuracy was much lower. The accuracy results were much lower for the B and C Data sets.

TABLE WITH ACCURACY RESULTS FOR TRAIN AND TEST DATA

ML Models	Accuracy
A_Linear_Regresion	0.395258
B_Linear_Regresion	-0.218658
C_Linear_Regresion	-10.303093
A_Ridge_Regression	0.395088
B_Ridge_Regression	-0.217177
C_Ridge_Regression	-10.345777
A_Lasso_Regression	0.597171
B_Lasso_Regression	0.436013
C_Lasso_Regression	-2.170383
A_Lasso_Regression_Alpha_0_9	0.600184

ML Models	Accuracy
A_Random_Forest	0.50109
B_Random_Forest	0.403605
C_Random_Forest	-0.059489
A_XG_Boost	0.396404
B_XG_Boost	0.340001
C_XG_Boost	-0.077485

FIG5: TABLE OF TRAIN/TEST ACCURACIES

In fig 5 you can see the Train and Test results for all our Models. The best Test Model Score is from a Lasso Regression model with an alpha value of 0.9 ran on the A Train/Test data. However, this is only 60% accurate and the Test Data was only 5% of the Data which is a very small Test size. The larger B and C Train/Test models had worse results and lead to large overfitting.

ACTUAL VALUES VS PREDICTED VALUES FOR A SHARE PRICE

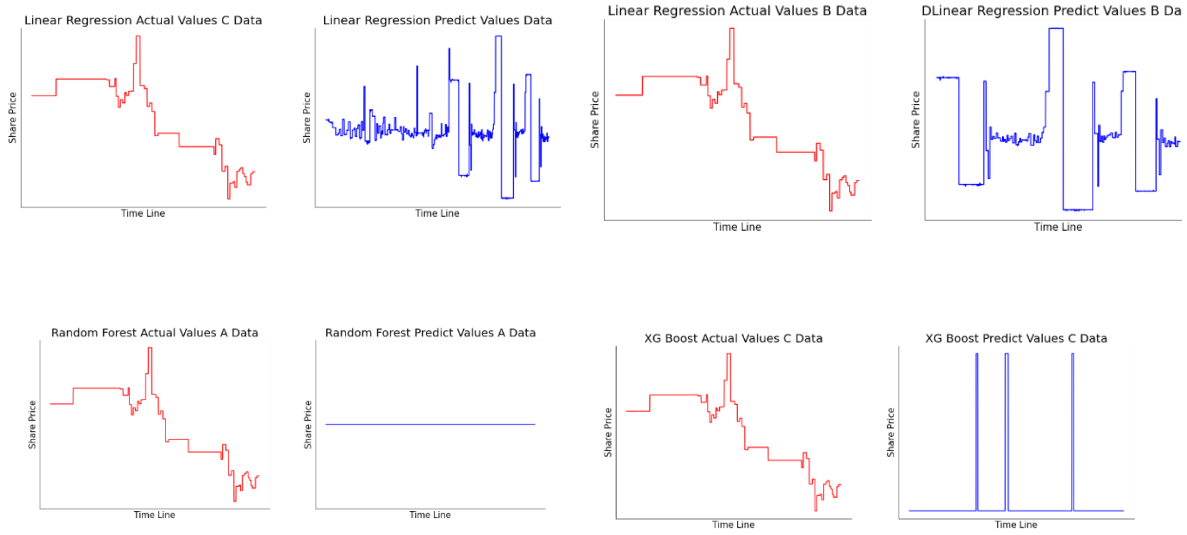

FIG 6: ACTUAL VALUES VS PREDICTED VALUES FOR A GAME STOP SHARE PRICE

In fig 6 we can see the actual values in red (notice they are all the same) and the predicted values on the Test data in blue. As you can see the predicted values were not very good. It was very difficult to get reliable predictions and is a reason why large Hedge Funds with lots of resources have also found this difficult.

PROJECT: NEXT STEPS

Next steps would be:

- NLP: TF-IDF (term frequency-inverse document frequency) to weight the words.
- NLP: Sentiment Analysis to get a Positive, Neutral, Negative value for comments.
- Use Yahoo's Finance API to look for other financial factors that would have affected the data.
- Run larger models with Amazon's AWS machines to get better model results.
- Get more Social Media data from Twitter, Facebook, etc

INSIGHTS

We have shown that there was a correlation between the number of Mentions of Game Stop and the increase in the Share Price. So we have answered the initial question. 'Yes social media can affect the share price of a stock'. Various news reports have linked Reddit to having caused the Game Stop Share Price jump(3, 4, 5). We need to be able to model this to prove without a doubt. There lies the problem.

We have learned that this is very hard to model. Both Accuracy and Predict Share Price values have shown unclear results for Linear Regression, Random Forest and XGBoost Machine learning Models. As there can be many other factors that affect the share price of a stock not just the number of mentions on one Social Media site.

REFERENCES/RESOURCES

1. https://www.kaggle.com/leukipp/reddit-finance-data

2. https://www.kaggle.com/datasets/paultimothymooney/stock-market-data

3. https://markets.businessinsider.com/news/stocks/gamestop-short-sellers-billions-losses-reddit-traders-wallstreetbets-rally-gme-2021-2-1030125873

4. https://www.theguardian.com/business/2021/jan/28/gamestop-how-reddits-amateurs-tripped-wall-streets-short-sellers

5. https://www.thetimes.co.uk/article/gamestop-battle-is-revived-after-week-of-share-price-losses-khg08s9lx