Information Cascade Prediction for r/wallstreetbets

Preface

Information cascades are phenomena in network theory in which people make the same decision sequentially. They have attracted significant research in the past few years, often in the context of social media sites such as Reddit. In this project, I will be using machine learning methods to attempt to predict the growth of information cascades on r/wallstreetbets, a subreddit in which users discuss investments. Specifically, I will also attempt to predict the massive surge of interest in the GME stock in January 2021.

In this context, an information cascade begins with a post on a specific stock (or group of stocks), and grows as comments responding to that post are made. I will study two types of cascades here: cascade trees, which are defined by a single post as the root, and the network of responses to the post, and cascade forests, which also begin with a post, but can include other posts on the same topic and their responses in their network if the posts are made within a specific time window (for this project, I will be using a window of 24 hours).

1. Data

The data for this project was acquired from the r/wallstreetbets subreddit, via several methods.

I began with a list of official NASDAQ stock ticker symbols, and used the Pushshift Reddit API to iterate over the list of symbols, and acquire all posts and post metadata containing the symbol, storing them in .csv files, one per symbol. The data acquired included the post ID, the post author, the timestamp, the number of comments in response to the post, and the post content.

I then merged all these files into a single pandas DataFrame object. Pushshift had some notable gaps in the data it could extract, so I had to use the PRAW API to iterate over all post IDs within the missing time ranges. This was a time-consuming process, but I eventually filled in the missing data.

I also needed to use the PRAW API to adjust some additional metadata such as the number of comments for a specific post, since the Pushshift data was not always reliable.

I used PRAW one last time to acquire the comments for each post I had scraped, retaining the information on comment hierarchy (comments which are direct responses to the post, comments which are responses to those comments, etc.)

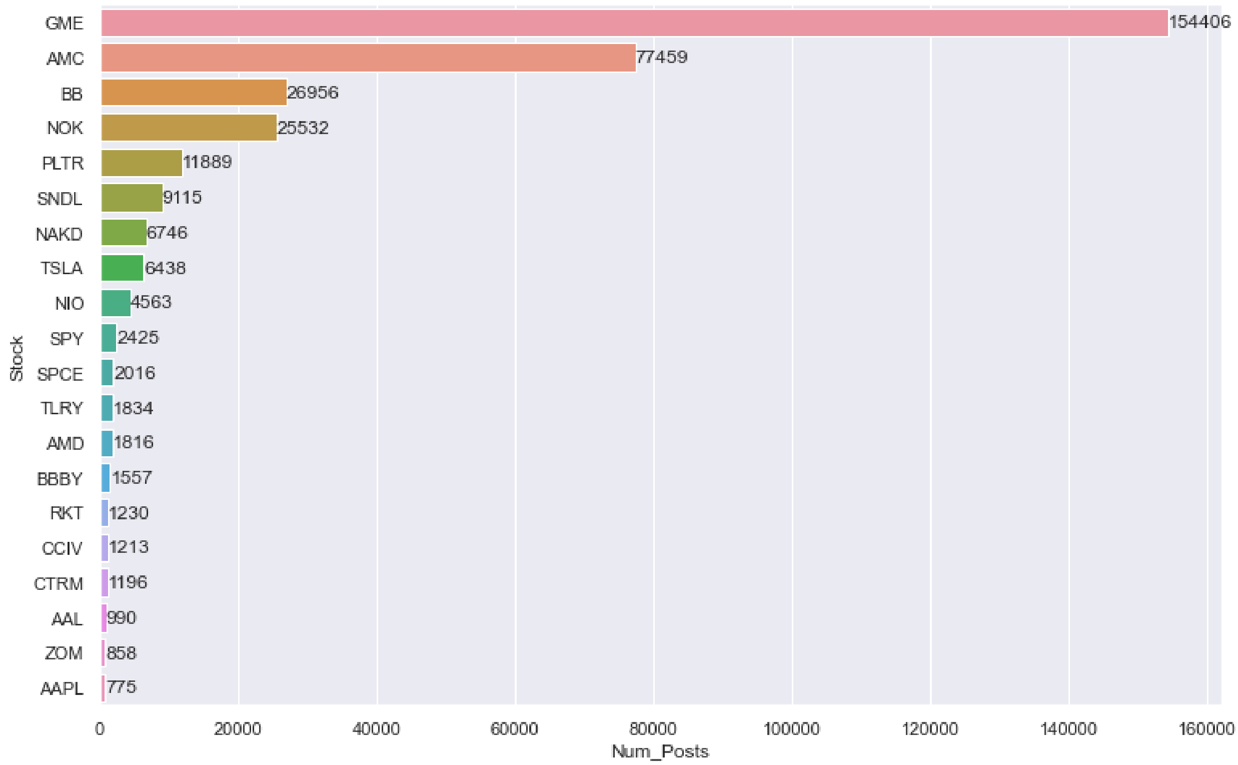
Finally, I performed some minor processing on the data, removing posts by deleted users, bots, and making sure the posts and comments data frames were consistent with one another.

At the end of this step in the process, I had two data frames: one for posts and one for comments. The focus of the remainder of the project is to use this data to evaluate the development of information cascades on the subreddit.

2. EDA

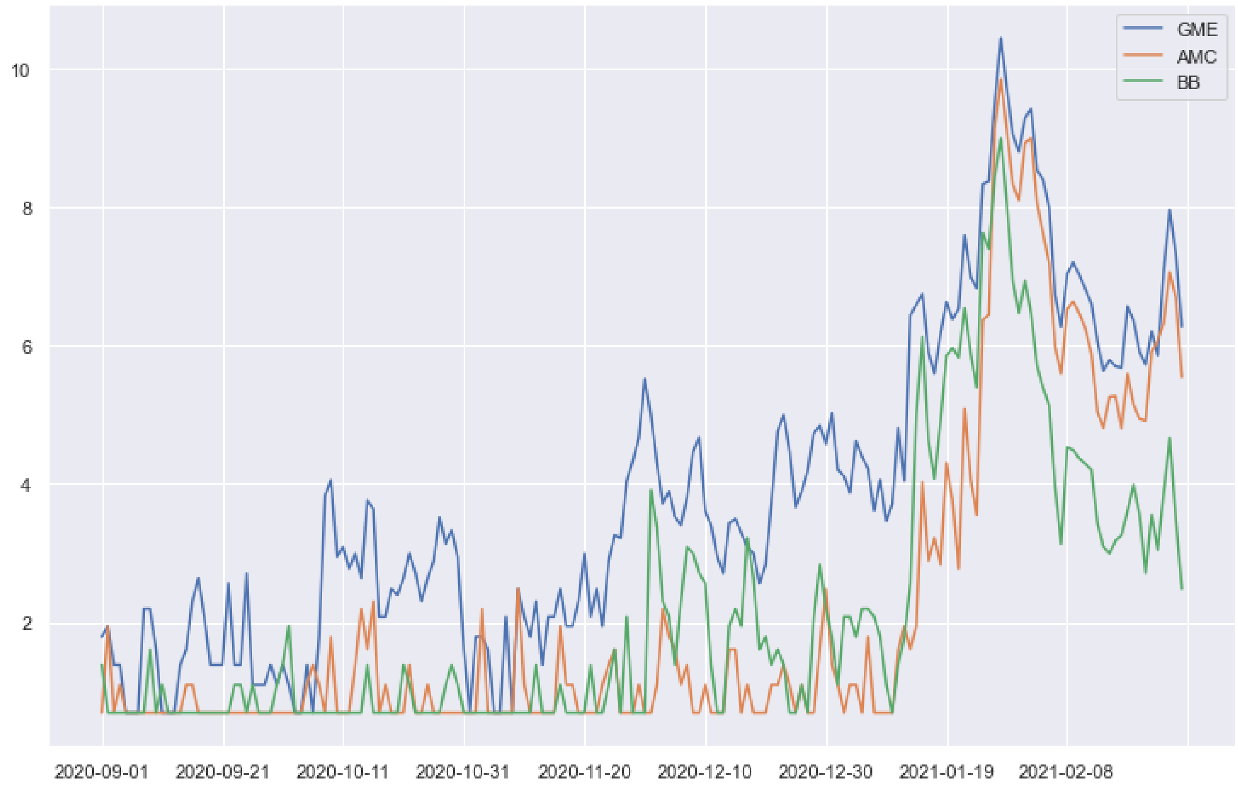
For this part of the project, I produced a number of visualizations to allow myself and the reader to better understand the data. I used the libraries matplotlib and seaborn for the visualizations.

First, I graphed the top 20 stocks mentioned on r/wallstreetbets during the period of interest. Here is the result:

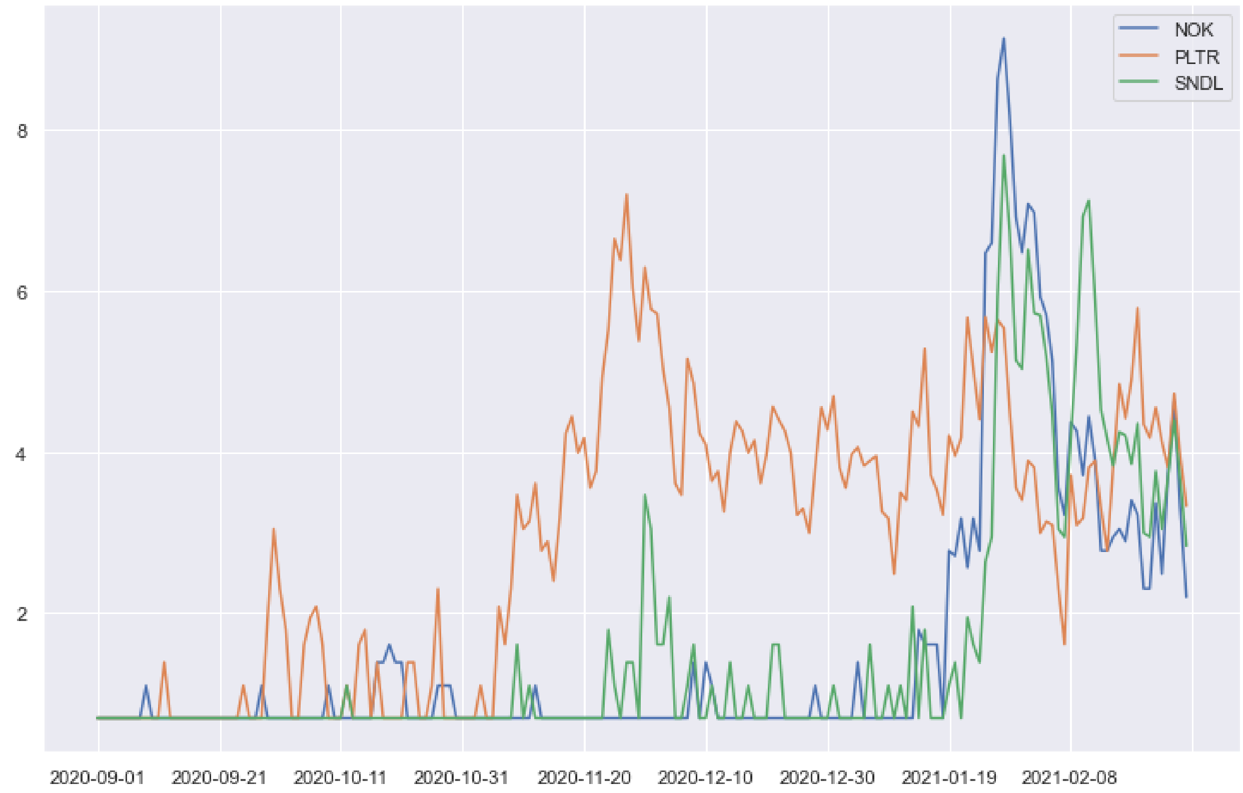


As one might expect, GME had significantly more mentions than any other stock, followed by other 'bandwagon' stocks such as AMC, BB, NOK, and PLTR.

I also graphed some of the most popular stocks' mentions in relation to one another temporally, to observe how correlated they were.

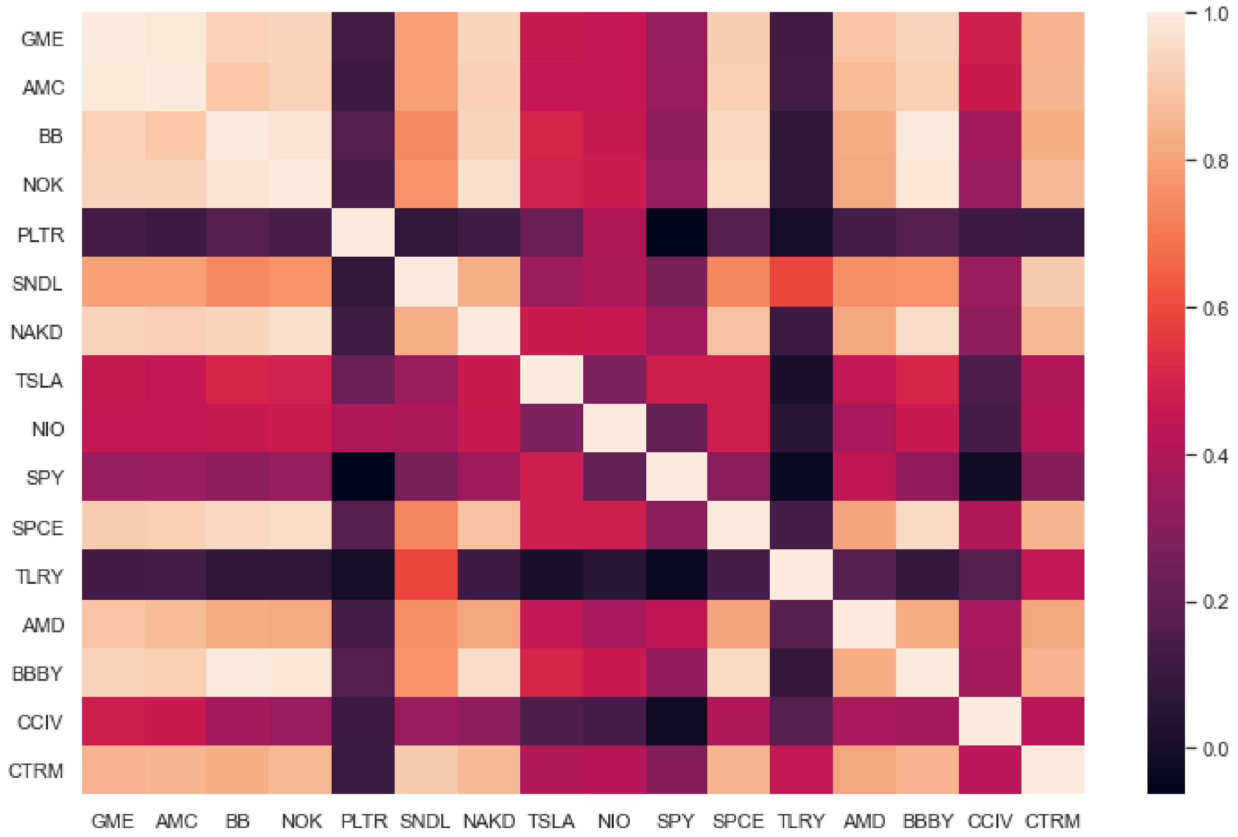


For GME, AMC, and BB, it is evident that stock mentions are highly correlated. This makes it apparent that people following one stock are also likely to be following the others, and that if sufficient action is taken to influence the market regarding one, others may follow suit.



Conversely, PLTR’s following appears to be independent of NOK and SNDL (which in turn are also correlated with GME), so not all major stocks have the same following on Reddit.

To reinforce this, here is a heatmap representing the correlation in mentions between all the major stocks on Reddit:



As can be seen, there is considerable variability in which stocks are correlated and which aren't.

3. Preprocessing

During this step, my primary goal was to convert the data I had obtained from Reddit into a Networkx graph. Networkx is a Python library which provides functionality to perform a number of graph-based algorithms on data in graph format.

The format of the graphs is that each post or comment is represented by a node, and each 'relation' between posts is represented by an edge.

I constructed two graphs from the data. The first was a 'disconnected' graph, by which I mean that the only edges between nodes are comments on a specific post. This means that there is a single root node for each component of the graph.

The second graph, the 'connected' graph, includes edges between posts if they satisfy the following requirements:

1. The posts are less than 24 hours apart.
2. The author of the second posted commented on the first post.
3. Both posts address the same stock.

The assumption behind this graph is that, having commented on the first post, and having made their post on the same topic in a timely fashion, the author of the second post is perpetuating the ‘message’ of the author of the first post, and thus their post can be seen as a continuation of the cascade associated with the first post.

For the connected graph, connected components are more complex, as one would expect.

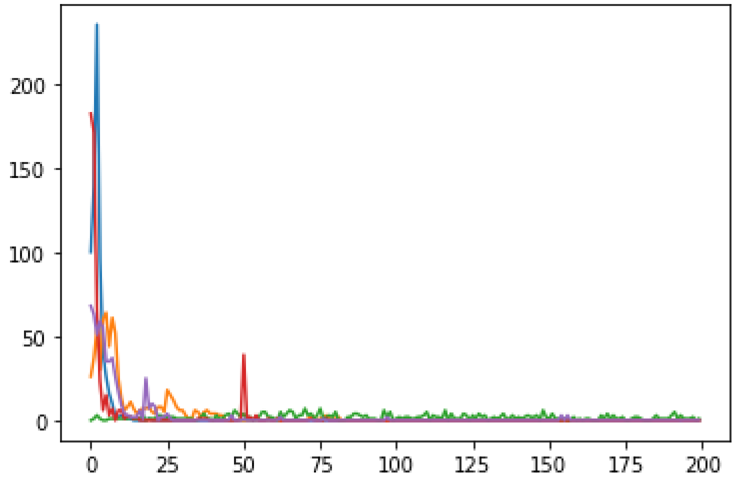
Both graphs contain both temporal information (each node is timestamped), and contain information on which stock (or stocks) are discussed in the posts.

4. Modeling

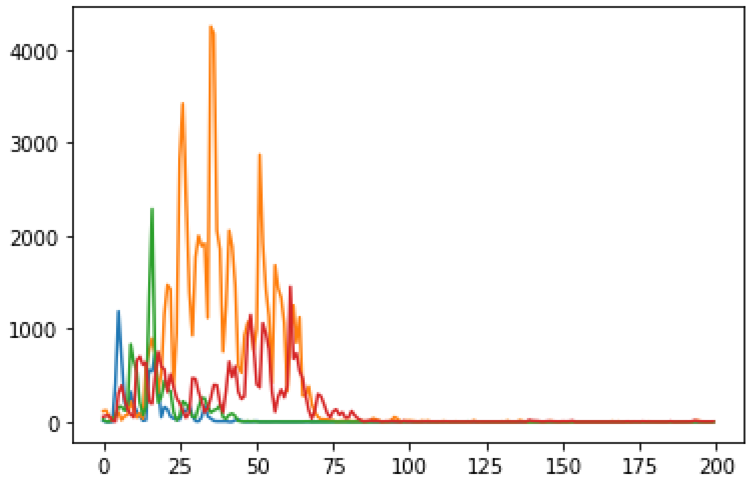
This is the final stage of the project. Here, my task was to extract features from the graph data I developed in the previous step, and then to use various supervised learning techniques to make the relevant prediction, that is, whether the cascade in question will at least double in size.

First, I visualized some of the cascades I had collected, graphing the addition of new nodes over (normalized) time.

The shape of tree cascades consisted of a sharp descent from a maximum value near the start. This is to be expected since the activity has a single root, and the way Reddit is structured means that activity will blow up close to when the post is made, and then dwindle.



Conversely, the shape of cascade forests had more variation. Again, this makes intuitive sense, they have multiple source posts, some of which will be more effective at driving activity than others, so the overall activity will be more sporadic.



For performing the split between training and test data, I divided both the cascade tree data and cascade forest data into five separate folds, four of which would be used as training data and one of which would be used as test data for each round of validation.

To expand my data beyond the full cascades mined from Reddit, I produced multiple data entries per cascade, representing various stages of the cascade's growth, eg. At 20,50,100 nodes. These were produced during the feature extraction stage, documented next, so it was necessary to split train and test data prior to this, so that partial cascades derived from the same full cascade weren't in both the training and the test sets.

4.1 Feature Extraction

Due to the massive variation in the sizes of information cascades, and additionally due to the findings of Leskovec et al, I decided to focus on the temporal features of the cascades and ratios between them in order to make predictions.

I wrote Python methods to extract features for both the cascade trees (single root) and cascade forests (multiple roots).

Some of the features extracted for the cascade trees include: the average time for a new node to be added for the first half and second half of the cascade, the ratio between them, the average time for a new node to be added in each quarter, and the corresponding ratios, and the average time for a new node to be added in each tenth, and the corresponding ratios. Using multiple criteria like this allows for both large and small cascades to be taken into account, the more granular measures (tenths) providing more information for the larger cascades.

In addition to the features listed above, for the cascade forests I extracted some additional features differentiating roots (posts) and other nodes (comments).

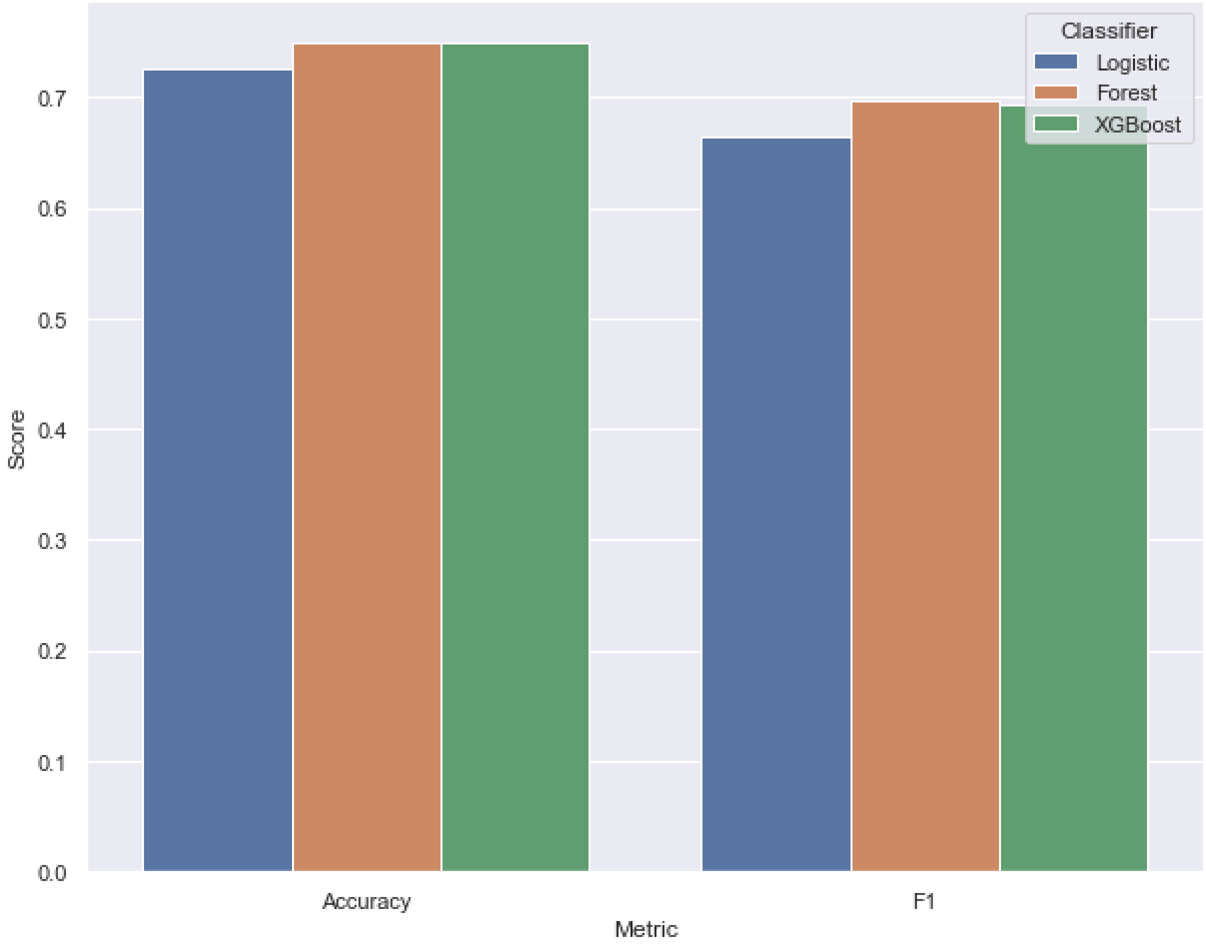
While saving the extracted features, I also saved the class of each cascade, which is a binary value, zero (0) if the cascade does not at least double, and one (1) if it does.

4.2 Evaluation

I employed three different machine learning methods to the dataset I generated: logistic regression, random forest, and XGBoost (a popular boosting algorithm).

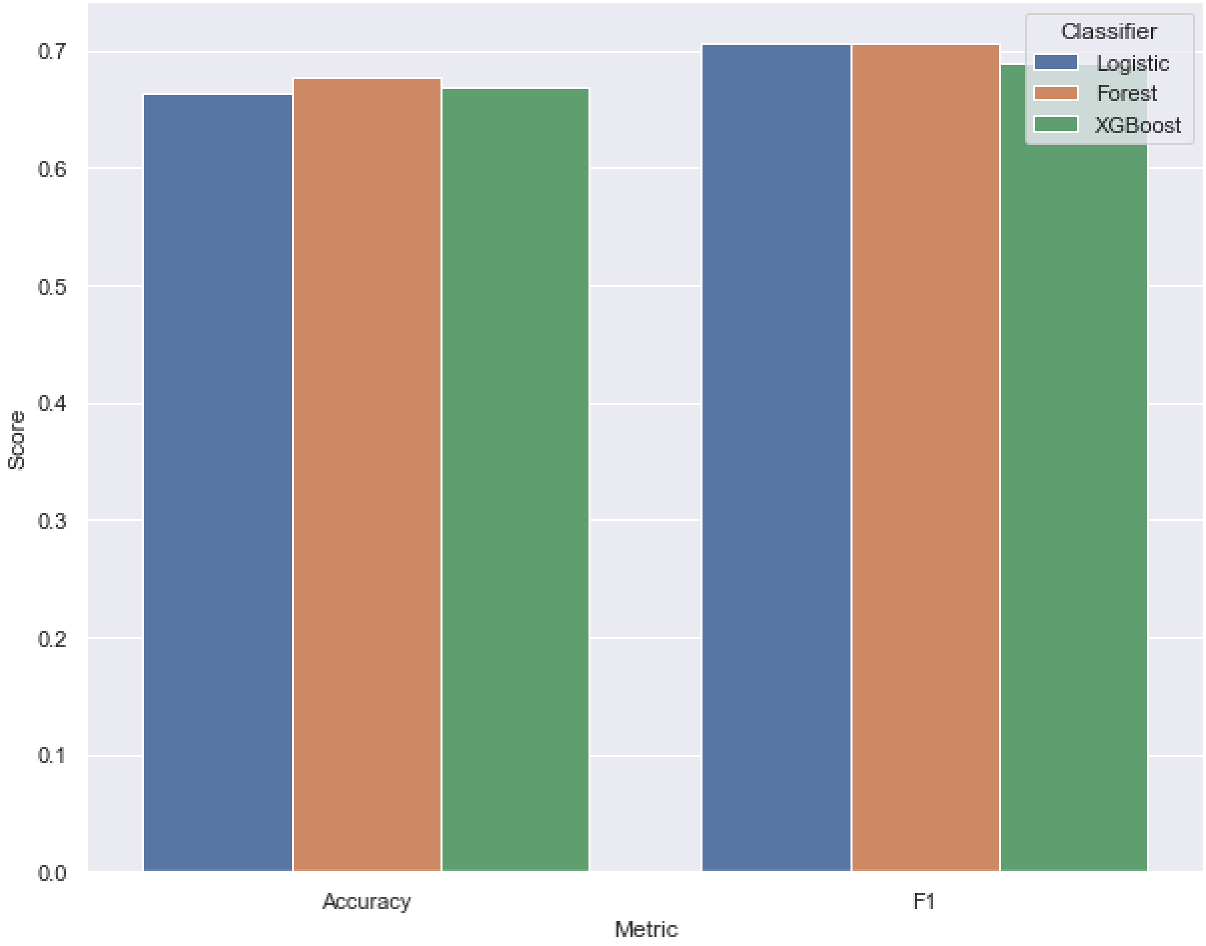
For the random forest and XGBoost models, I iterated through various parameters for depth and the number of estimators in order to find the optimal ones.

For the cascade trees, the results are as follows:



The accuracy was above 70% for all estimators, though the random forest and XGboost models performed better than the logistic regression model. The F1 score, an alternative metric which is useful for unbalanced data, was slightly below 70% across the board.

For the cascade forests, the results are as follows:



The accuracy in the case of the cascade forests was lower than for the trees, but the F1 scores were higher, with the logistic regression and random forest models both achieving over 70%.

Overall, the random forest model was the most performant.

Unfortunately, the model was still not able to make accurate predictions for the huge GME cascade, most likely because its size and duration caused it to have a more complex structure than other cascades used to train the model.

5. Conclusion

My initial inspiration for this project was to determine whether I could predict the GME cascade based on other cascade structures seen in the WSB subreddit. While the outcome of this specific task was disappointing, the model still performed fairly well in predicting the growth of cascades on average.

Collective action phenomena are only going to become more commonplace with social media's ever-increasing prominence and capacity to connect people, so additional research into this topic could prove very valuable.

Possible improvements to this specific project could include having a more diverse and complex set of features to serve as the basis for predictions. Also, since I was focusing on purely structural and temporal aspects as opposed to subject matter, I could have extracted cascades from other subreddits, since the network structure would arguably be similar. More data would certainly have been valuable to the project.

This has been a valuable learning experience for me, however, and has been an excellent opportunity to dive into a topic with applications very relevant to the world we now live in.

References

[1] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. ICDM, 2007.

[2] J. Cheng, L. Adamic, P. Dow, J. Kleinberg, J. Leskovec. Can Cascades be Predicted? https://arxiv.org/pdf/1403.4608.pdf , 2014.