Team Safari – Markethub: Combining Market Data and Reddit Sentiment for Equity Event Prediction

I. Introduction

Inspired by the events surrounding the “Meme Stock” phenomenon and the social media fueled short market disruptions of the early COVID pandemic, we sought to create a model to screen for similar situations. We began with a brief inspection of the market mechanics behind what is known as a ‘short squeeze’ and expanded our analysis to look at more broadly disruptive events in the equity market. Our hypothesis was that these events might have telltale signals in existing market data.

When investors anticipate a decline in a stock's value, they engage in 'short selling' by initially selling borrowed shares, aiming to buy them back later at a lower price, thereby profiting from the difference. Short squeezes happen when a stock that is heavily shorted unexpectedly surges in value. This forces short sellers to quickly buy back the stock to close their positions and limit losses. This rush to buy adds to the upward momentum, further inflating the stock price. Notable instances of this dynamic were observed in stocks like GameStop and AMC, which experienced extraordinary surges of 2300% and 3500% respectively in 2021. Charts of this movement can be seen in Figure 1.

These events are both rare and rather difficult to define a priori. Thus we broadened our search for moves that appeared to be both larger than historical price movements and seemingly unanticipated by the options market. The implied moves derived from options pricing can be interpreted as a view into the overall market sentiment towards upcoming volatility. This uncertainty is most often centered around an anticipated event such as earnings, where the outcome is unknown, but semi-predictable. We wanted to find large resulting moves, when some characteristics of short squeezes were present, coupled with a surprise to the market. We then wanted to link the sentiment of social media, given the demonstrated impact of retail traders in recent market events.

Market sentiment is generally described as bullish (anticipating higher prices) or bearish (anticipating lower prices). To gauge that sentiment, we incorporate Reddit text data into our analysis in the form of submissions and comments by users of the popular financial forum (or ‘subreddit’) r/WallStreetBets (WSB) which has grown to over thirteen million users as of 2022. We found that applying NLP techniques such as VADER sentiment analysis let us peek into the stock ticker-specific sentiment found in this online community.

Sample Short Squeezes:

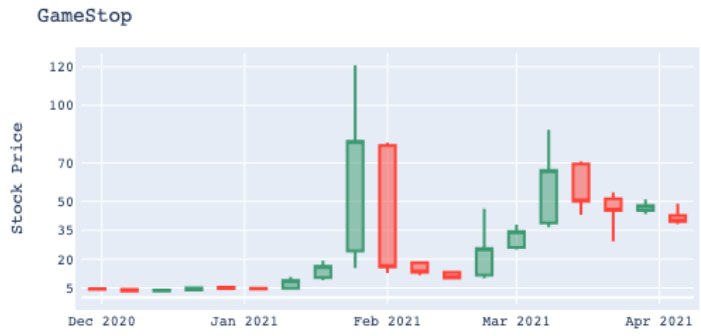


Figure 1-1: In January 2021, **GameStop** observed an uptick in buying volume. The short squeeze saw the stock price ascend from the single digits to an intraday high of $120 per share

Figure 1-2: Jumping the GME short squeeze bandwagon, **AMC** shares rallied about *3500%* in the first half of 2021. It should be noted that at the time, and as of this writing, AMC's fundamental financial metrics were less than stellar.

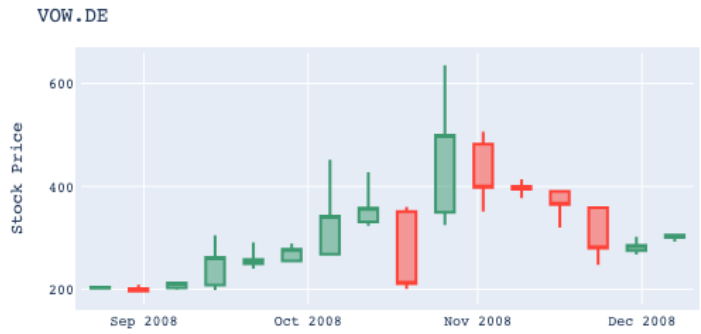


Figure 1-3: In October 2008, Porsche purchased 75% of shares of fellow German automaker **Volkswagen**, leaving short-sellers unable to close their positions. The subsequent short squeeze saw shares of Volkswagen rally approximately *300%* over a two month period.

II. Project Statement and Objective

- **Our objective is to build a predictive model using sentiment analysis of Reddit posts, combined with historical cross-market data to forecast surprise equity market events, like short squeezes, and resulting price action. We explore the impact of Short Interest trading volume and Stock Option Metrics like Implied Volatility (IV) and Gamma on future stock prices.**
- **We also present a real-time dashboard displaying these predictions, for proposed use by traders and investors.**

The problem of combining sentiment, historical data, and reaching trustworthy conclusions is a complex task due to the myriad of inputs that affect market and individual stock price fluctuations. For instance,

macroeconomic catalysts such as unemployment and interest rates, and large-cap company earnings reports can set the tone for overall market performance. Stock specific catalysts are commingled with the broader market impact, and the effects are sometimes hard to distinguish.

Participant sentiment also has the potential to sway the direction of the market or a single stock's price. In the case of the GME and AMC short squeezes, both were driven by the online forum WSB where users informally banded together to collectively drive the price of these stocks higher, producing exponential gains in an unusually brief timespan. A challenge we faced in this facet of our work was working with user comments with a moderate likelihood of containing satire or inappropriate language. This works to confuse a model and create unintended interpretations. We also faced limitations on the real time access of market data, but fortunately our analysis of historical options data was made possible by our team member's vast archive of end-of-day (EOD) options chain for every stock ticker and subsequent risk calculations.

III. Data Collection and Methods

Our data was diverse and collected according to the requirements of each. We broadly stored data in the Amazon Web Services (AWS) cloud computing platforms, and conducted our analysis there using AWS SageMaker Jupyter notebooks and AWS Postgres databases.

Table 1: Data Definitions

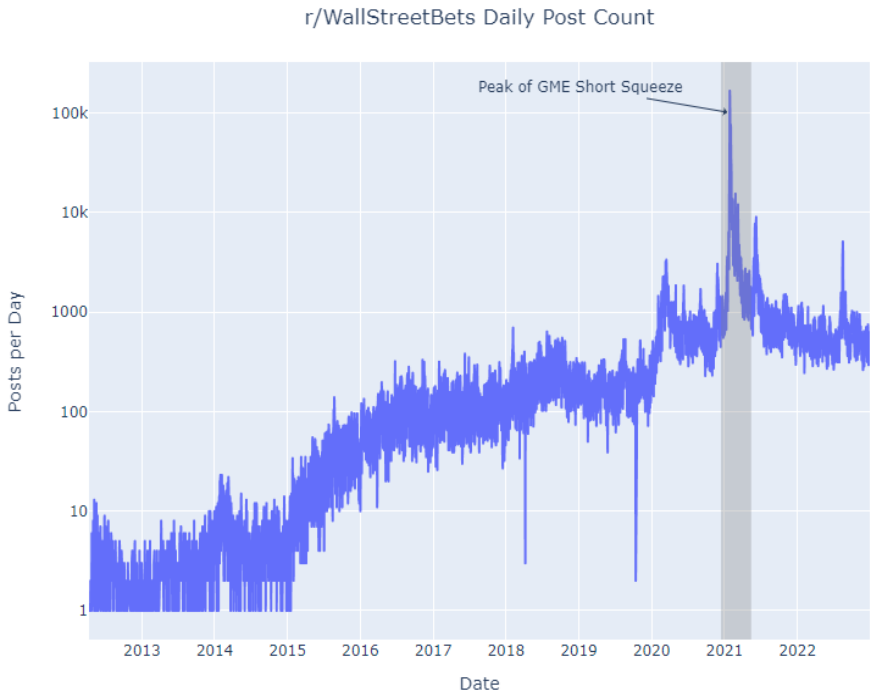
Data Set	Source	Method
Stock Tickers	Nasdaq Stock Screener	Created a universe of 7335 symbols listed on the NYSE, AMEX, and NASDAQ. Downloaded as a csv file
Daily Open, High, and Closing stock prices	Yahoo Finance	Collected using the YahooFin API
ETF constituent data:	Zacks.com	Web scraping
Short Sale data:	FINRA	Archived into a database over many years
Historical option contract metrics (Greeks)	Chicago Mercantile Exchange (CME)	EOD options chain data accumulated into a database. Black Scholes calculation, including IV and Gamma Accumulations done for relevant periods (30, 90 days) Uploaded to AWS Postgres database.
Historical text data from r/WallStreetBets (2012-2022)	Reddit/Pushshift Archives https://the-eye.eu/redarcs/	Downloaded archive in Z standard compression format, extracted posts and comments in JSON. Uploaded the extracted text (5.6 GB) into an AWS database

API text data from r/WallStreetBets (Realtime)	Reddit Python Reddit API Wrapper(praw-dev, 2023)	PRAW is a Python library that enables developers to acquire data with the Reddit API.

IV. Sentiment Analysis

As mentioned, sentiment analysis on r/WallStreetBets can provide trading ideas by gauging the collective sentiment of the community towards specific stocks, helping identify potentially popular or volatile assets. By uncovering trending stock discussions, traders can make informed decisions based on the sentiment dynamics within the subreddit.

Figure 2: WallStreetBets Daily Post Count



Reddit archival data is a collection of past Reddit content, including posts, comments, and associated details. Data is gathered by third parties, like Pushshift.io, using Reddit's API or web scraping. It's valuable for understanding trends, content, and user behaviors on the platform. The volume of posts in the subreddit notably experienced exponential growth, particularly during the meme stock era (see Figure 2), reflecting the community's surging interest and engagement with stock-related discussions during the COVID pandemic timeframe.

Figure 3: Sample of Initial Reddit Dataset

	id	subreddit_id	subreddit	author	created_utc	permalink	title	selftext	num_comments	score
0	18eiq1v	t5_2th52	wallstreetbets	thecuzzin	1.702144e+09	/r/wallstreetbets/comments/18eiq1v/crypto_bros...	Crypto bro's growing bolder by the day		4	80
1	18eilgq	t5_2th52	wallstreetbets	grip_n_Ripper	1.702143e+09	/r/wallstreetbets/comments/18eilgq/upst_are_we...	UPST - are we looking at the making of a short...	It's had a bit of a run lately, and the short ...	5	0
2	18eig6u	t5_2th52	wallstreetbets	FutureOwn5191	1.702143e+09	/r/wallstreetbets/comments/18eig6u/yolo_and_di...	YOLO and Diamond Hands: Why BYND is a Weakend ...	Hey fellow degenerates, ...	2	3
3	18eiaj5	t5_2th52	wallstreetbets	go_far_go_together	1.702142e+09	/r/wallstreetbets/comments/18eiaj5/arm_put_the...	ARM put thesis	I don't ever post here. Don't listen to anyth...	9	14
4	18ei2y2	t5_2th52	wallstreetbets	ChewThirty	1.702142e+09	/r/wallstreetbets/comments/18ei2y2/augmedix_me...	Augmedix - Medical AI	\nWith Augmedix and Google Cloud entering a pa...	2	1
...	...	...	...	...	...	...	...	...	...	...
891	180lpiu	t5_2th52	wallstreetbets	audiomuse1	1.700586e+09	/r/wallstreetbets/comments/180lpiu/lucid_says_...	Lucid says its new all-electric SUV beats Tesl...		187	1053
892	180l96m	t5_2th52	wallstreetbets	StretchyJeff	1.700585e+09	/r/wallstreetbets/comments/180l96m/get_in_on_r...	Get in on RBT making garbage smart	\n\nhttps://preview.redd.it/v30g304vbq...	3	7
893	180kj0k	t5_2th52	wallstreetbets	dwittherford69	1.700583e+09	/r/wallstreetbets/comments/180kj0k/cramer_exam...	Cramer examines OpenAI shakeup, says Microsoft...		4	7

The libraries we used to process the raw text data require limited preprocessing. We removed non-standard characters via regex, stripped and trimmed whitespace, and removed duplicate submissions. Given the colorful language used on Reddit.com, we also replaced profanity (via LDNOOBW, 2023). A sample of raw Reddit data can be seen in Figure 3, and field descriptions are available in Appendix D.

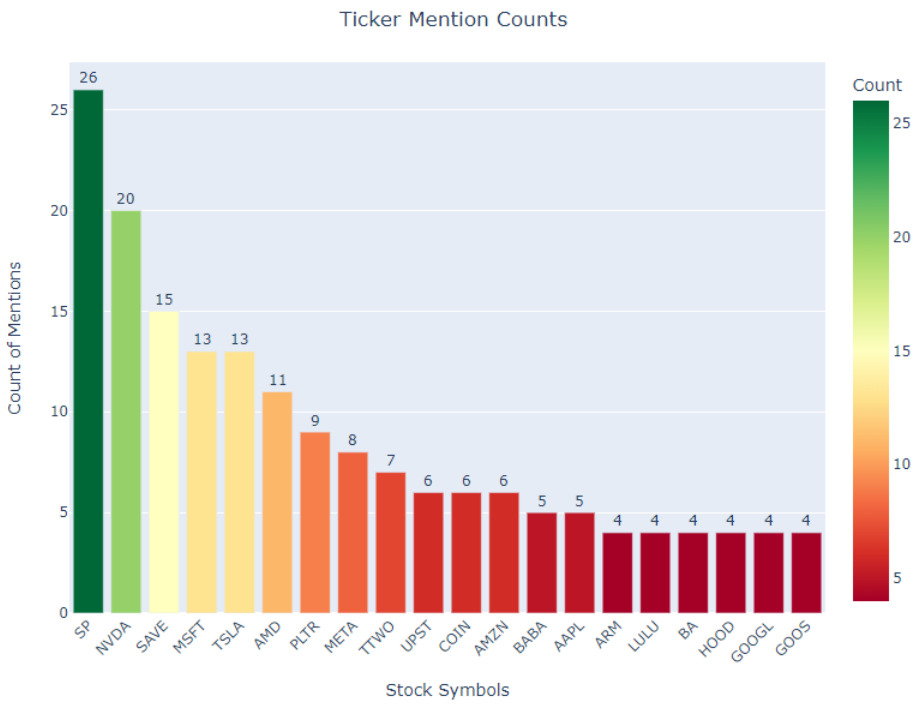
Reddit Submissions vs Comments

We initially considered using both the posts and the user comments in our analysis. Our decision was to focus on Reddit submissions (posts), while also considering the post's content and score, allowing us to better control and analyze the core themes and sentiments that users want to convey. Additionally, the post's score, representing the number of upvotes and downvotes, serves as an indicator of content significance. Posts with higher scores tend to reflect information that has captured the community's attention, making them vital in understanding trends and discussions. Ultimately, our decision to focus on submissions and their scores allows us to extract information while keeping the analysis targeted to our research objectives.

Stock Ticker Extraction

To identify stock symbols within Reddit posts, our process begins by sourcing known stock symbols, with provisions for adding additional stocks as needed. A blacklist is implemented to filter out non-relevant terms (Dominguez, 2021). Regular expressions are applied to the text data, ensuring the identification of valid symbols while excluding those found in the blacklist. This process uncovers tickers within posts and allows tracking of stock-related discussions.

Figure 4: Ticker Mention Counts



One challenge is the inclusion of unintended false positives, where common words or abbreviations inadvertently match ticker symbols. Stock tickers often consist of short letter sequences, which can overlap with everyday language or internet slang. We controlled for some of these, but we also left some to be included since they are indications of sentiment. For example, in Figure 4, which represents the frequency of stock symbol mentions in our data, we observe 'SP' and 'HOOD' appearing frequently. In this context, 'SP' usually refers to the S&P 500 stock index, and 'HOOD' often pertains to discussions about the Robinhood brokerage service rather than its underlying stock.

Valence Aware Dictionary and sEntiment Reasoner (VADER)

VADER Sentiment Analysis is a natural language processing tool designed to assess the sentiment expressed in a piece of text (Hutto, C.J., & Gilbert, E. ,2014). VADER is specifically tailored to handle social media text, making it well-suited for platforms like Twitter, Facebook, and Reddit. This tool is valuable for quickly gauging sentiment in social media data, customer reviews, or any text data where traditional sentiment analysis may struggle due to slang, context, and informal language. However, as others have noted, VADER is not specifically designed for use in the finance domain (Shapiro, et al, 2020). VADER works by assigning polarity scores to individual words and emoticons within a text. These polarity scores indicate whether a word or phrase conveys a positive, negative, or neutral sentiment, as well as the intensity of that sentiment.

The community of r/WallStreetBets is known for its unconventional trading strategies, humor, and enthusiastic discussions about stock market investments. Its members often use a specific set of terms,

abbreviations, and slang to communicate, making it a challenge to analyze. Therefore, to account for their vocabulary and communication style we used two enhancements:

> **Custom Lexicon**: Expanding the VADER lexicon with community-specific terms and slang used in r/WallStreetBets. For instance, including terms like "YOLO" (You Only Live Once) or "diamond hands" (holding onto a stock regardless of its performance) to accurately gauge sentiment within the subreddit.
>
> **Weighting**: Assigning different weights to sentiment scores based on the significance of certain terms within the community. For example, giving higher weight to terms that are particularly relevant or impactful in r/WallStreetBets discussions to better reflect the community's sentiment trends.

By customizing the model with community-specific terms and weightings, we can enhance the accuracy and better capture the nuances of sentiment expressed (Dominguez, 2021). This helps in understanding the sentiment dynamics and market sentiment. For more on VADER and its use in financial news, see Appendix E.

Cumulative Sentiment by Ticker

For each ticker, the sentiment scores of all posts mentioning that stock are summed. This cumulative sentiment score provides an overall relative assessment for that particular stock based on the posts within r/WallStreetBets for a given timeframe. As Figure 5 shows, Spirit Airlines (SAVE) has strong positive sentiment surrounding a pending merger, while the AI and technology related names also are popular as of December 2023.

Figure 5: Cumulative Overall Sentiment by Symbol

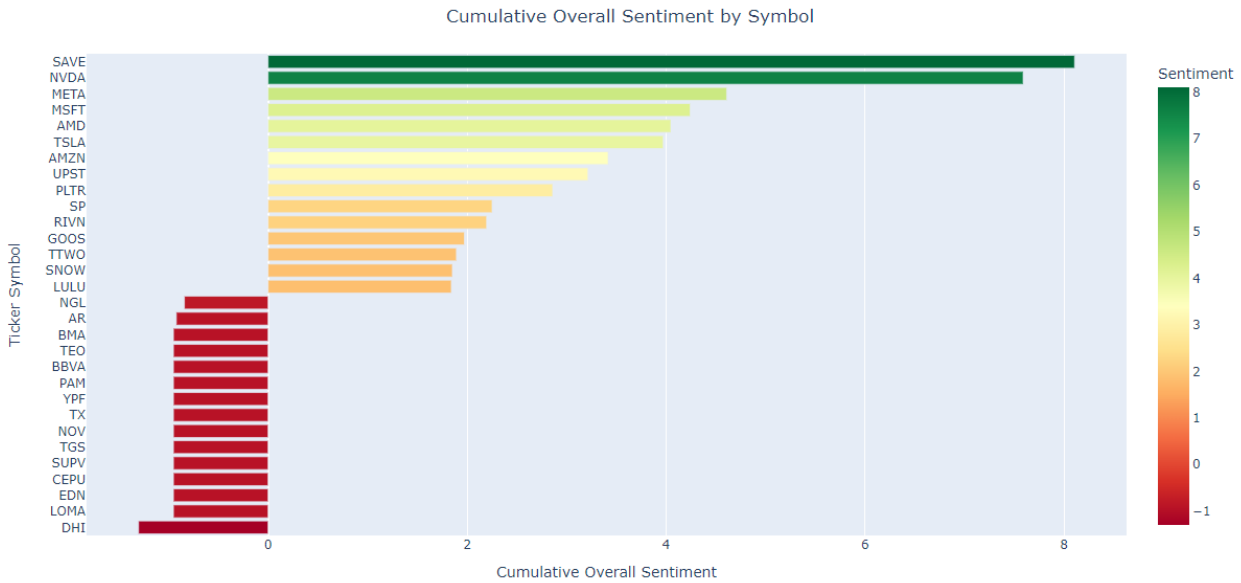
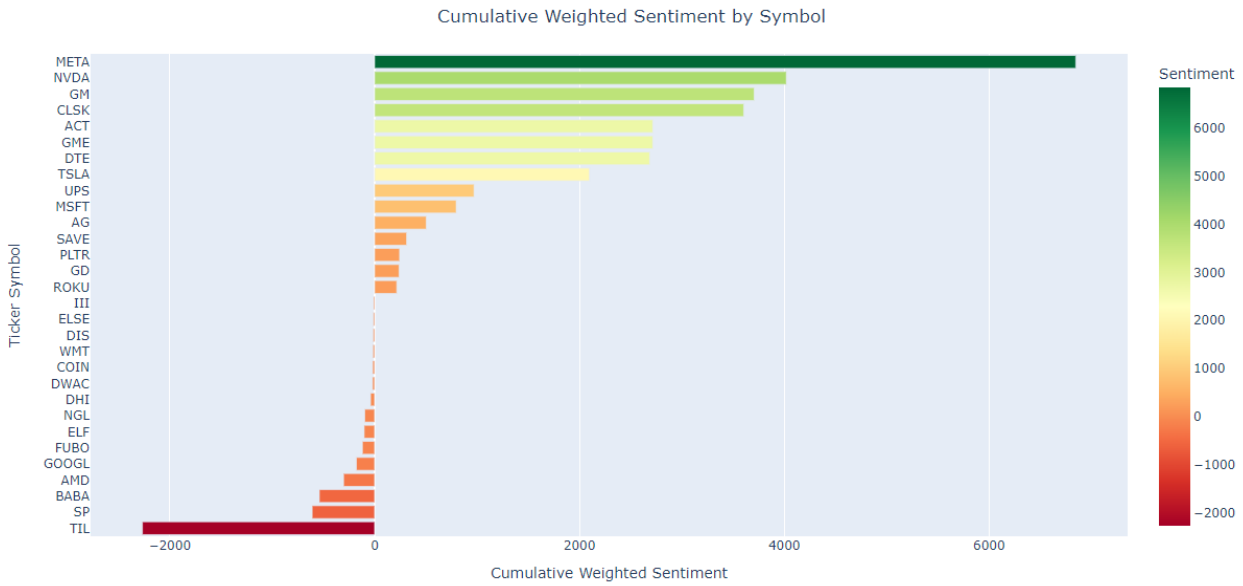


Figure 6: Cumulative Weighted Sentiment by Symbol



Reddit allows users to give feedback on a particular post by up and down voting, and that is reflected by a "score" of the post. This allows for a more nuanced perspective on sentiment, giving weight to posts that have garnered attention and engagement within r/WallStreetBets. This can help identify not only sentiment trends but also which discussions or posts are driving those sentiments. This is shown in Figure 6, where META and NVDA show strong community affinity and sentiment. However, this can be skewed in the r/WallStreetBets universe where humor and memes are often used and receive strong feedback. This is possibly the case in "TIL" being an abbreviation for "Today I Learned". As such, it could be added to the blacklist and excluded from analysis, but this is a judgment call since "TIL" also represents Instil Bio, Inc. on the NASDAQ exchange.

N-Grams

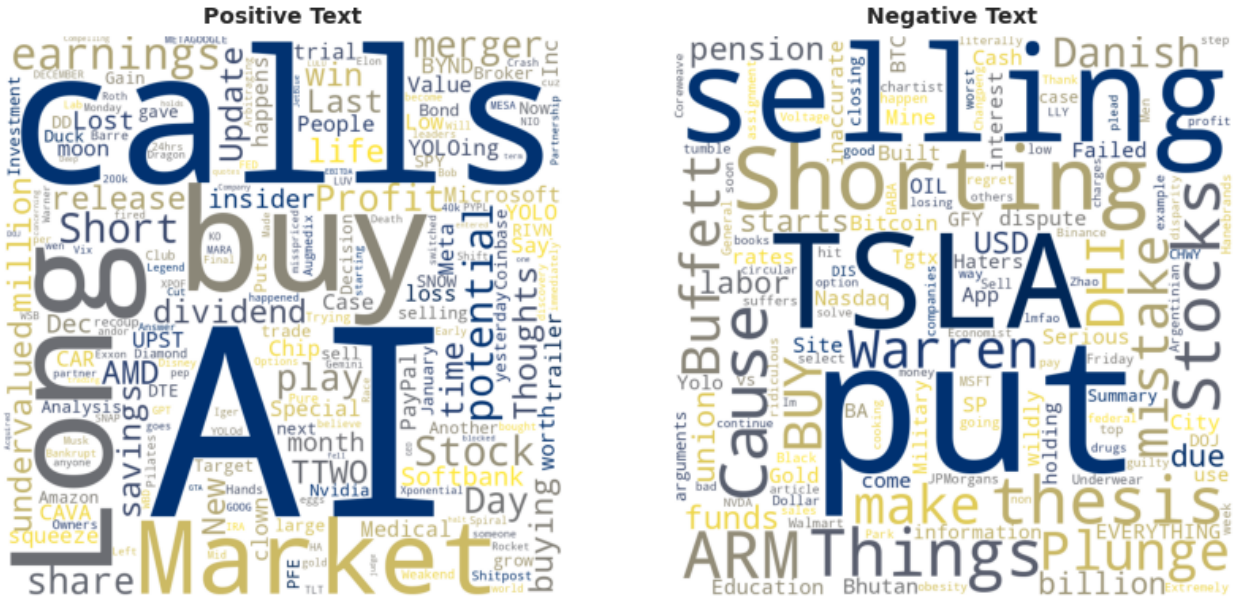
Bigram and Trigram counts play a role in deciphering the co-occurrence patterns of words within our textual data. These counts extract valuable insights about specific language and stock-related topics discussed. By identifying common phrases, trading terms, and frequently occurring word combinations, we gain an understanding of the language used within the subreddit. Analyzing these counts allows us to uncover contextual nuances, track sentiment trends, and identify recurring themes in the discussions. This enhances our ability to perform tasks like sentiment analysis, topic modeling, and information retrieval. Here again, we see references to the Spirit Airlines mergers with "SAVE MERGER", and the common refrain "TO THE MOON" referring to a stock with anticipated extreme upside.

Figure 7: Sample Bigrams and Trigrams Present in Post Titles

Word Cloud

A word cloud can provide insights into the subreddit's current hot topics, stock picks, or prevailing sentiments, making it useful for staying informed about the latest trends within the subreddit. The resulting visualization shows the most commonly mentioned vernacular and key themes (Konstantin, 2021). The larger words in the cloud would typically represent frequently discussed stocks, trading terms, and popular expressions. As Figure 8 highlights, recent sentiment towards AI is positive, while Tesla(TSLA) faces negativity.

Figure 8: Sample Text Word Clouds by Sentiment



V. Model Development and Validation

Model Development

Our model was specifically designed to identify large, unexpected upward moves in stock prices. We adopted a classification approach, focusing on creating a balanced dataset with positive and negative classes. This approach stemmed from our goal to not predict specific stock forecasts or movement degrees, but rather to identify stocks within a broad universe that have a higher likelihood of experiencing significant events. Logistic regression was chosen as our primary model due to its probabilistic nature, allowing for the ranking of stocks based on the model's confidence in predicting events. Additionally, we utilized LightGBM, a tree-based learning algorithm known for its efficiency with large datasets, further aiding in stock ranking based on event likelihood.

Data Sources

In our model, designed to pinpoint unexpected upward movements in stock prices, the integration of diverse data sources, especially options data, is pivotal in enhancing its generalization capabilities. We leverage the contrast between implied volatility (IV) and historical volatility (HV) to tap into market sentiment and expectations. IV's forward-looking nature offers insights into future stock price fluctuations, while HV anchors our understanding in historical price trends. This duality enables more accurate assessment of investor sentiment by tracking the allowing our model to measure when these values deviate significantly from their long term trend, possibly implying an imminent large price movement in the underlying stock.

Further refinement of our model's predictive power comes from analyzing IV over varying time frames, such as 30-day and 90-day options. Short-term IV sheds light on immediate market responses, while long-term IV captures broader trends. This time-based comparison helps our model identify sentiment shifts. Additionally, examining IV in call and put options potentially enriches our model’s understanding of market biases, offering a detailed view of investor psychology.

Cumulative Gamma, reflecting the sensitivity of option prices to changes in the underlying stock, also plays a key role. Its inclusion in our model accounts for the impact of market maker hedging, particularly important near option expirations, and serves as an indicator of potential price stability or volatility.

The inclusion of sentiment scores from platforms like Reddit parallels the incorporation of consumer confidence indices in econometric models, acknowledging the sentiment's influence on market liquidity and volatility.

To bolster our model's ability to forecast significant stock movements, we conducted a comparative analysis of features like IV and short sale volumes across one to eight-week periods. This strategy is crucial to spot rapid feature fluctuations, potentially signaling upcoming major market events and ensuring our model captures early signs of impactful market dynamics.

Event Identification

In our study, we developed a systematic approach to identify significant stock price events such as short squeezes or major market news impacts. This approach begins with the calculation of daily percentage

changes in stock closing prices. We then employ rolling statistics specifically rolling mean and standard deviation over a typical lookback period of 90 days to dynamically assess these price changes against recent market trends.

Data Preprocessing

In the data preprocessing and processing phase of our study, we focused on meticulously handling missing or infinite values and standardizing the dataset. This step was essential to ensure the integrity of our data and the accuracy of the model's coefficients reflecting the true influence of the features. Each feature in our dataset including variables like Greeks, short sale volume, and others was subjected to a comprehensive rolling time frame analysis.

Validation Strategy

We employed the TimeSeriesSplit method for cross-validation from the sklearn's model selection library. This approach ensured that our validation process followed the chronological sequence of financial data, preventing any lookahead bias during model training. Lookahead bias refers to a situation where a model learns from events that haven't occurred yet, and using this strategy helped us avoid such issues.

Hyperparameter Tuning

In our hyperparameter tuning process, we focused on optimizing our logistic regression model using specific parameter settings. We defined a parameter grid that included values for 'C' (regularization parameter), 'penalty' (norms for penalization), 'solver' (solver type), and 'max_iter' (maximum number of iterations). This approach allowed us to systematically explore various combinations of these parameters to find the configuration that yielded the best predictive performance for our model while minimizing the risk of overfitting.

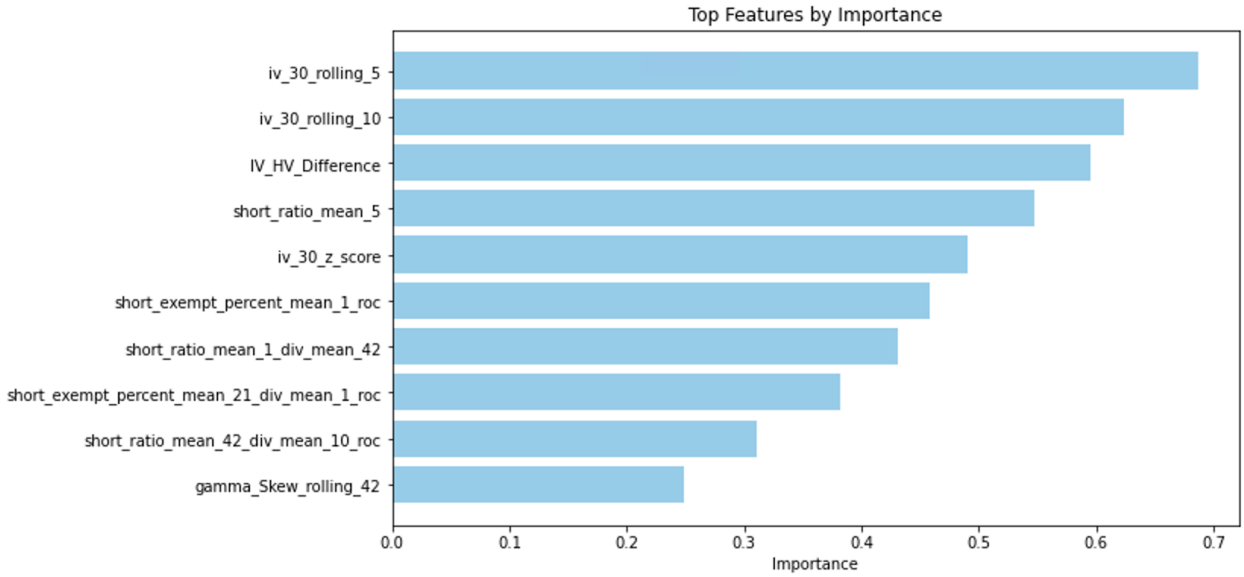
VI. Model Performance and Results

Initial Results & Feature Importances

The initial performance of our logistic regression model yields nuanced insights into its capability to identify stock market events, such as short squeezes. An accuracy of 59% with a margin of plus or minus 3% across different folds of the TimeSeriesSplit indicates a moderate level of prediction reliability. This suggests that while the model captures some underlying patterns in the data conducive to stock events, there is room for improvement in prediction consistency and possibly in reducing the variance across time series folds through feature engineering, model tuning or alternative modeling techniques.

Feature importance, depicted in the below visualization, provides a deeper understanding of which market indicators are most influential in driving the model's predictions. The 5-day rolling average of 30-day implied volatility ('iv_30_rolling_5') emerges as the most significant feature, emphasizing the market's forward-looking sentiment. Similarly, the 'IV_HV_Difference' feature underscores the divergence between market expectations and historical price movements as a critical predictive factor. The importance of short interest levels is captured through 'short_ratio_mean_5', highlighting short-term market pressures.

Figure 9: Feature Importance



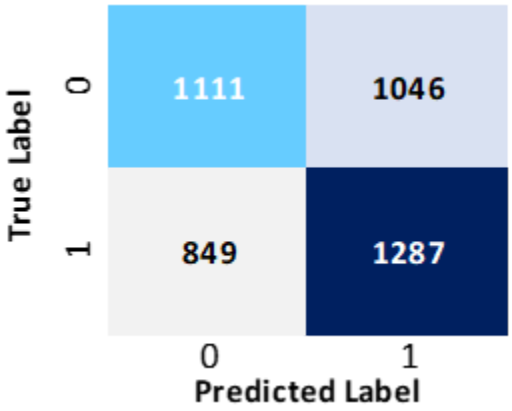
The model's feature importance profile underscores the complexity of market dynamics where both recent trends in volatility and more gradual shifts in investor sentiment and behavior play crucial roles. Understanding the model's reliance on these features is vital for interpreting its predictions and for guiding subsequent model iterations.

Confusion Matrix Analysis

The confusion matrix offers a visual and quantitative understanding of the model's performance.

The True Positives (TP) at 1,287 indicate the instances where the model correctly predicted a significant event, while the True Negatives (TN) at 1,111 suggest accurate predictions of non-events. Conversely, False Positives (FP) and False Negatives (FN) at 1,046 and 849, respectively, represent the model's erroneous predictions. The relatively balanced nature of TP and TN indicates a level of competence in the model, yet the high FP and FN highlight areas for improvement, particularly in minimizing incorrect predictions which could be critical in financial decision-making.

Figure 10: Confusion Matrix



F1 Score

The F1 score, at approximately 0.576, reflects the balance between precision and recall in the model's predictions (Czakon, 2023). In the context of stock market events, where accuracy is paramount, this score suggests that while the model can identify events with moderate reliability, there is still a considerable risk of misclassification that could lead to potential financial losses.

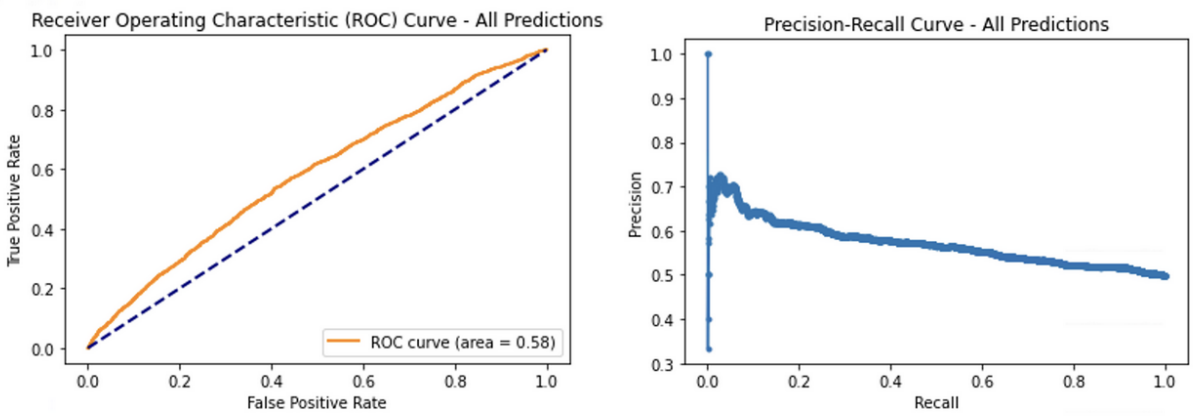
Precision and Recall Interpretation

The calculated precision of about 0.552 indicates that when the model predicts an event, it is correct around 55% of the time. The recall of approximately 0.603 means that the model identifies about 60% of actual events. In the high-stakes environment of the stock market, these figures suggest that while the initial model has some predictive power, it may not yet be reliable enough for high-confidence trading decisions.

ROC and Precision-Recall Curve Insights

The ROC curve, with an AUC of 0.58, and the precision-recall curve illustrate the model's trade-offs between true positive rates and false positive rates. These curves indicate that the model's ability to discriminate between events and non-events is only slightly better than random chance (Czakon 2023). This is especially critical when considering the financial implications of false predictions in stock market events, signifying a need for a more nuanced threshold to optimize the balance for the specific financial context.

Figure 11: ROC/AUC

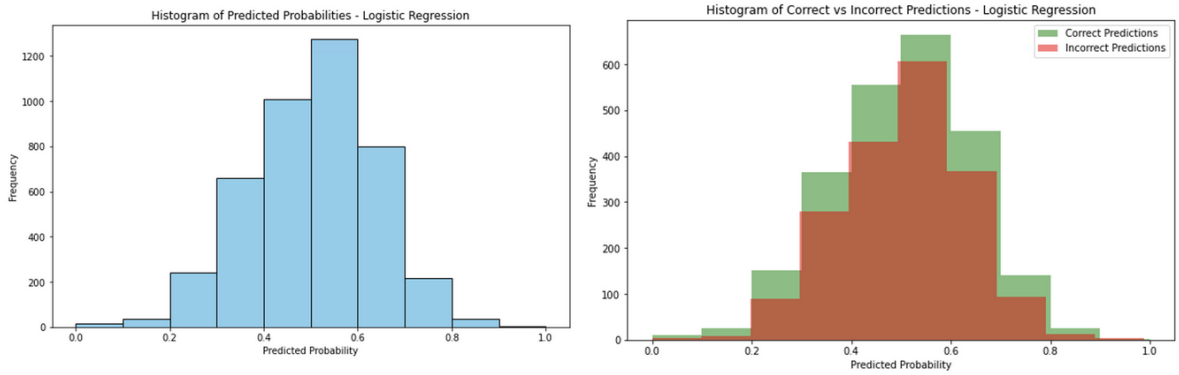


Histogram of Predictions

The histograms of predicted probabilities reveal that correct predictions predominantly cluster at higher confidence levels, suggesting the model's strength in certain areas. Incorrect predictions, however, are dispersed across the confidence spectrum, indicating a consistent level of uncertainty when the model is incorrect. This distribution underscores the model's reliability in high-confidence scenarios while

highlighting the need for caution at lower confidence levels, especially in the context of identifying market events like short squeezes.

Figure 12: Histogram of Predictions



High Confidence Predictions Analysis

Transitioning from the previous histogram insights, we observed a concentration of accurate predictions at higher confidence levels. This prompted a deeper investigation into high-confidence predictions, under the hypothesis that focusing on such predictions could enhance the model's practical utility in identifying market events like short squeezes.

In predictive modeling, especially within the volatile domain of stock markets, high confidence predictions are often synonymous with greater reliability. These predictions can be instrumental in decision-making processes where the cost of errors is substantial. The rationale for focusing on these predictions is grounded in the pursuit of a model that offers not only a statistical probability of an event but also a practical certainty that can be acted upon with confidence.

We employed a methodology where predictions were filtered based on a confidence threshold. Only predictions with a probability higher than 70% for the event class, or lower than 30% against, were considered 'high confidence.' This method capitalizes on the model's strongest assertions to potentially increase the accuracy and reliability of its predictions.

Results of High Confidence Prediction Analysis

When the model was constrained to high confidence predictions, the accuracy improved markedly to 71%. The precision of the model, indicating the quality of positive predictions, stood at 0.77, while the recall, reflecting the model's ability to find all relevant instances, was 0.70. The F1 score, balancing precision and recall, reached 0.73 for positive events and 0.69 for negative, indicating a robust model performance in this selective context.

Figure 13: Model Performance scores

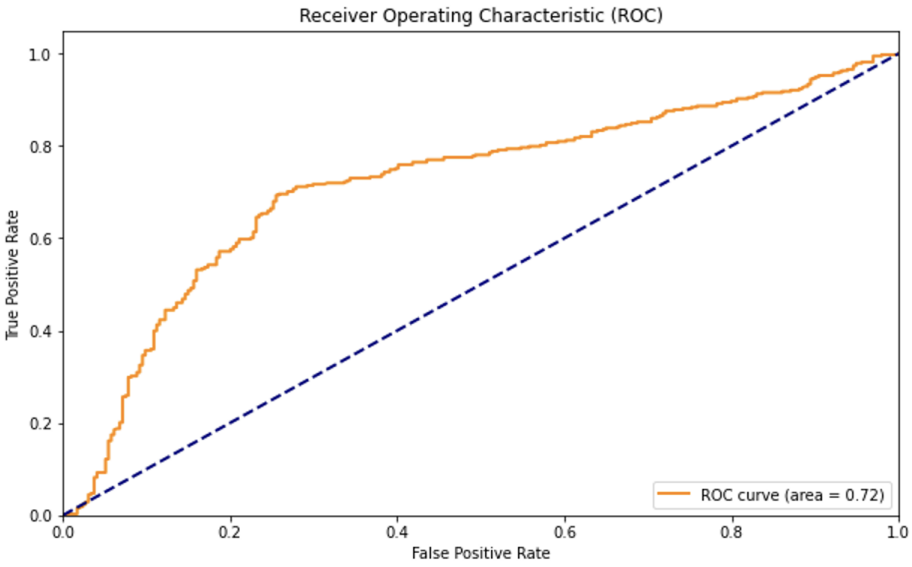
	precision	recall	f1-score	support
False	0.65	0.74	0.69	294
True	0.77	0.70	0.73	378
accuracy			0.71	672
macro avg	0.71	0.72	0.71	672
weighted avg	0.72	0.71	0.72	672

The confusion matrix for high confidence predictions revealed 217 true negatives and 263 true positives, demonstrating the model's effectiveness in correctly classifying events and non-events when sure. The 77 false positives and 115 false negatives suggest that while the model is more accurate, there is still a risk of error, albeit reduced.

A disparity between precision and recall often indicates a trade-off between the two metrics. In the context of our project, a higher precision means that when a short squeeze is predicted, it is more likely to be accurate. However, the lower recall indicates that the model might miss some potential short squeezes, an acceptable compromise to avoid costly false alarms.

The Receiver Operating Characteristic (ROC) curve for high confidence predictions, with an area under the curve (AUC) of 0.72, shows a substantial improvement from the overall model. This uplift suggests that the model is far better at distinguishing between events and non-events when filtering by confidence, a desirable trait for investors who prefer fewer, but more reliable signals.

Figure 14: ROC Curve



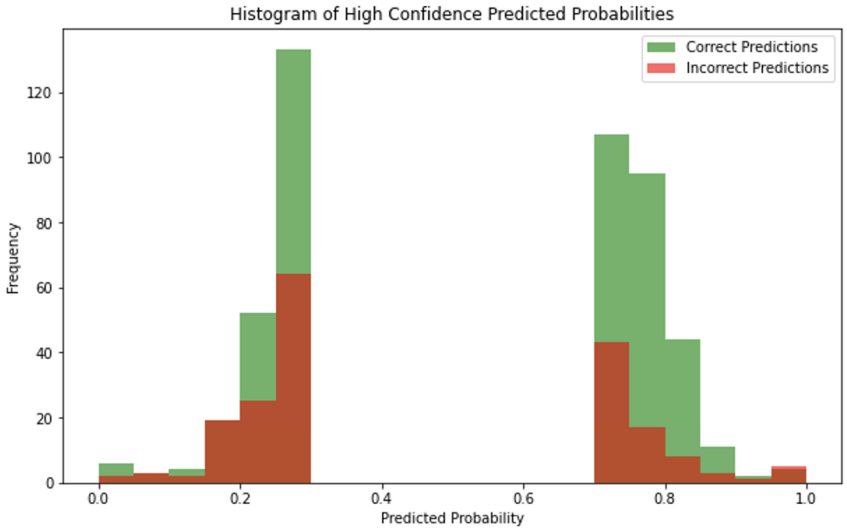
Interpretation and Implications

The analysis implies that by prioritizing high confidence predictions, we can refine the model's applicability in the real world. Investors may use this model to identify potential market events, such as

short squeezes, with a higher degree of certainty, reducing the noise and focusing on the signals with the highest expected payoff. This targeted approach could serve as a valuable tool in a strategy aimed at capitalizing on market inefficiencies presented by short squeezes.

The histogram of high confidence predictions paints a compelling picture. We see a significant clustering of correct predictions at the extreme ends of the probability spectrum, with a stark contrast in the frequency of correct versus incorrect predictions as we approach the 1.0 probability mark. This indicates that when the model is very confident about an event occurring (probabilities close to 1), it is generally correct, reinforcing the model's reliability for high-stakes decisions.

Figure 15: Predicted Probabilities



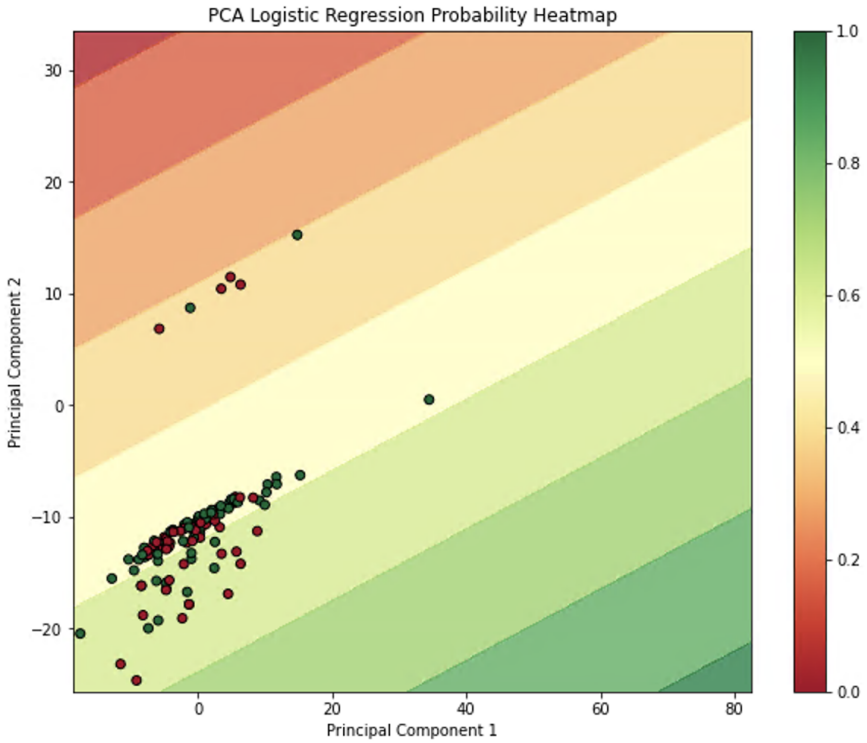
Compared to the initial model's histogram, which displayed a more uniform distribution of correct and incorrect predictions across probability levels, the high confidence histogram shows a clear divergence. Correct predictions dominate at the tails, particularly near the 1.0 probability. This contrast suggests that filtering for high confidence not only refines the accuracy of predictions but also provides a more distinct separation between the model's successes and failures, enabling a clearer understanding of when the model's predictions are most trustworthy. By leveraging high confidence thresholds, the model's predictive reliability is enhanced, suggesting that its use could be particularly effective for risk-averse strategies focused on identifying high-probability events.

Reduced Feature Space

The PCA Logistic Regression Probability Heatmap (Figure 15) condenses the complexity of our logistic regression model into a two-dimensional representation (Prabhakaran, 2022), offering a visual insight into the distribution of predictive probabilities. Within this reduced feature space, we observe gradients of color that correlate with the model's confidence levels: darker shades of green denote higher confidence in predicting significant market events, such as short squeezes, while darker red areas suggest a stronger conviction in the absence of such events. Areas of lighter shades indicate lower confidence, representing regions of the feature space where the model is less certain about its predictions. This visualization not only simplifies the interpretation of the model's internal mechanisms

but also highlights regions where the confidence of correct classification is maximized, which is critical for making informed decisions in the fast-paced environment of stock trading.

Figure 16: PCA Probability Heatmap

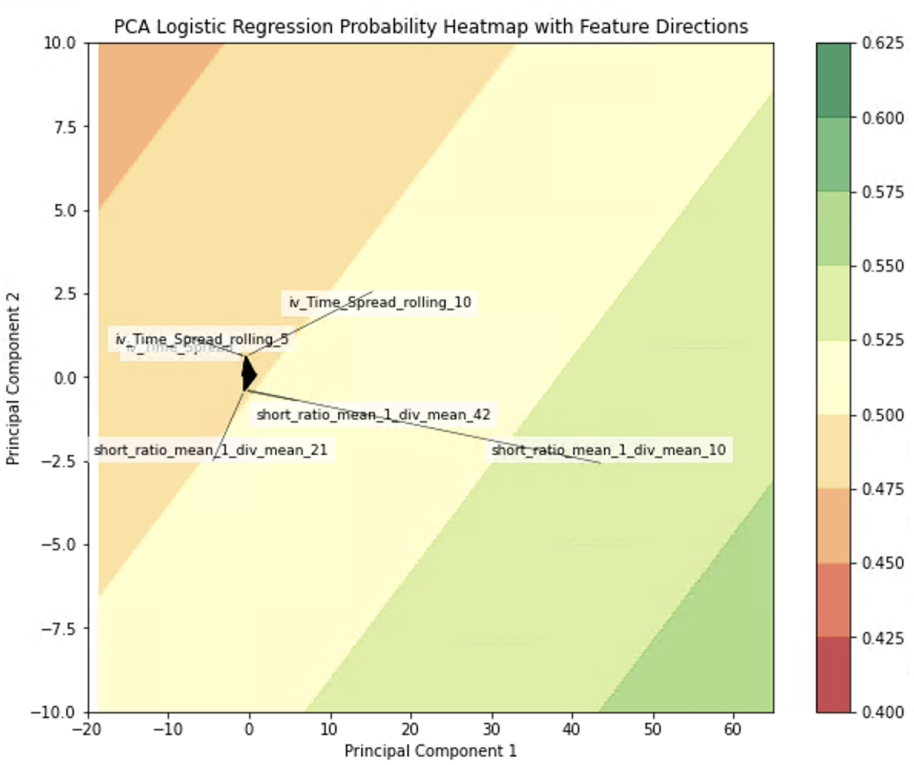


The prominence of short ratio features within the PCA space suggests that our model is highly responsive to the relative change in short sales volume. Specifically, the model is attuned to periods when the short ratio experiences atypical spikes or drops, which is characteristic of market conditions preceding or during a short squeeze. These features—short_ratio_mean_1_div_mean_10, short_ratio_mean_1_div_mean_21, and short_ratio_mean_1_div_mean_42 (Figure 16)—serve as harbingers, capturing the essence of sudden shifts in market sentiment that could lead to significant stock events.

Additionally, the influence of the iv_time_spread feature underscores the model's sensitivity to disparities in option pricing over different time frames. A noticeable increase in the price of short-term options relative to long-term options, as measured by the IV30 to IV90 spread, may signal heightened market volatility or speculation, which are often precursors to events like short squeezes.

In essence, these influential features encapsulate the dynamism of the stock market, where rapid changes in trading behaviors and option pricing can foreshadow significant events. By capturing these fluctuations, our model gains the foresight to potentially identify and anticipate short squeezes, offering valuable insights within the scope of our project.

Figure 17: PCA Probability Heatmap



Additional Model Types

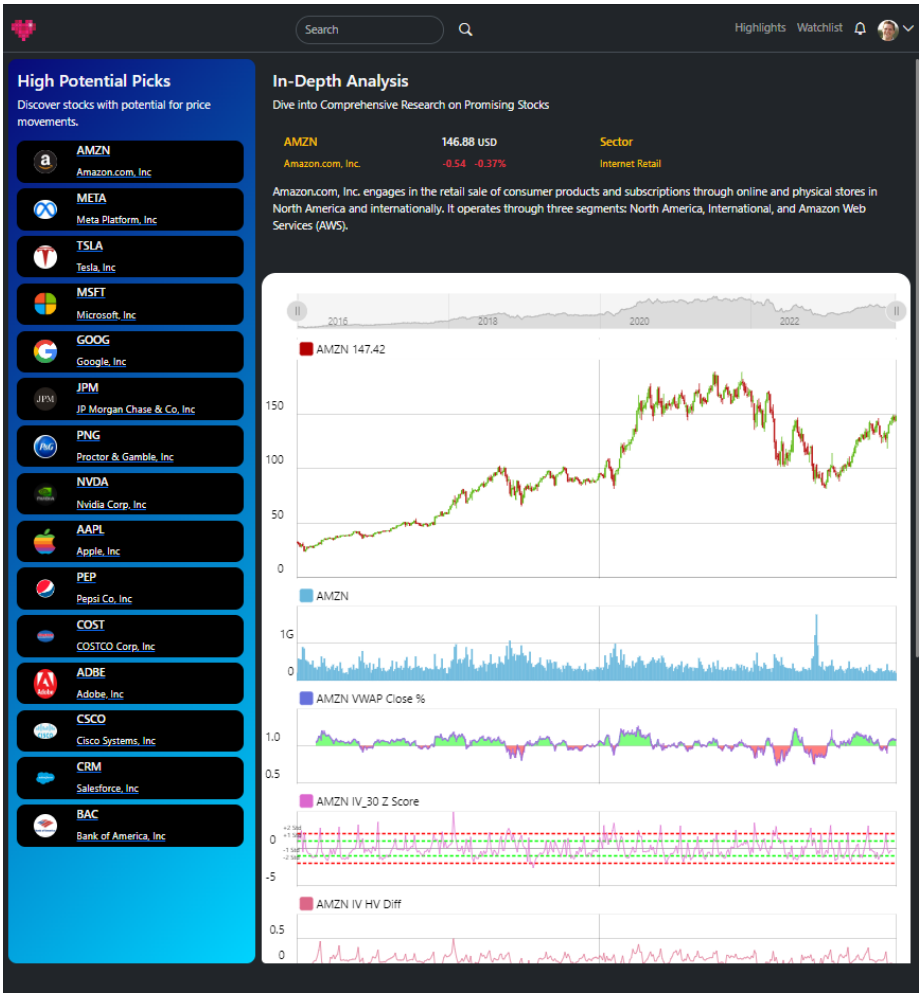
In an effort to enhance predictive accuracy, we applied LightGBM's gradient boosting classifier to our dataset. This advanced machine learning technique resulted in a confusion matrix with 999 true negatives, 443 false positives, 311 false negatives, and 1023 true positives. It achieved a precision of 0.76 for negative classes and 0.70 for positive classes, with recall rates of 0.69 and 0.77, respectively. The overall accuracy was calculated at 73%, with an F1 score of 0.73. These figures, while robust, did not significantly surpass the logistic regression model which had similar accuracy and F1 score. Thus, for its interpretability and parameter tuning capabilities, we opted to continue with the logistic regression model.

Data Sources and Composition

Our model's credibility is underpinned by the reliability of data sourced from established financial APIs and databases. Drawing historical price data from Yahoo Finance—a platform with a proven track record in financial data accuracy—ensured a solid foundation. Moreover, short sale volume data from brokerage reports provided an undercurrent of market sentiment, indicative of hedging activities and speculative positions akin to the Commitment of Traders (COT) reports in commodities trading. The sentiment analysis from Reddit was particularly innovative; by translating qualitative investor discourse into quantitative sentiment scores, we harnessed the burgeoning influence of retail investors—akin to the way market sentiment indices signal investor optimism or pessimism.

VIII. Dashboard

Figure 18: Dashboard



Our real-time dashboard is a streamlined yet comprehensive tool, intended for investors and traders. Central to this is the Sentiment Gauge, which specifically analyzes Reddit sentiment data, offering a unique perspective on retail investor sentiment. Given a stock's mention on Reddit, our dashboard then uses the model to predict and rank the stocks by probability of event. The dashboard also features a price chart, volume, volume weighted price, implied volatility z score, and implied volatility depictions, providing real-time insights into market fluctuations and stock performance. Additionally, it includes sections for recent Reddit Posts, highlighting current discussions among retail investors, and short interest metrics, which can signal potential market shifts.

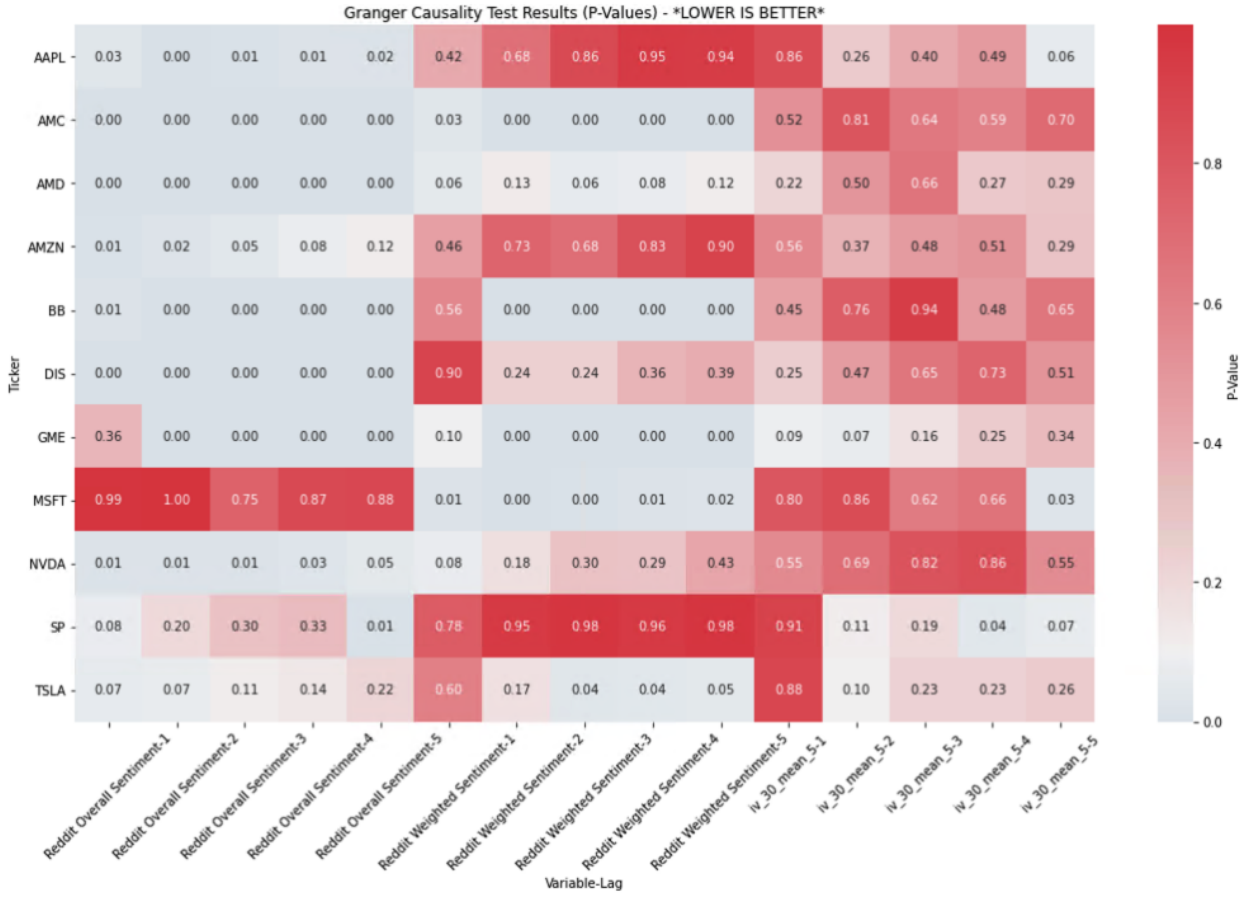
Future enhancements will incorporate news and Large Language Model (LLM) summarization, offering succinct summaries of complex financial information. Textual analysis can be enhanced with Bigrams/Trigrams and Word Clouds, revealing prevalent themes in financial discourse. This dashboard aims to blend quantitative data with qualitative Reddit sentiment analysis, for a holistic visualization of the financial market.

IX. Discussion

Reddit Causality - What drives market movements?

In the investigation of stock events and their predictors, causality refers to the relationship where one variable is a direct influence on the occurrence of another (Shojaie and Fox, 2022). This study probes the causative impact of online discourse, specifically within Reddit communities, on the fluctuation of stock prices. We delve into whether sentiments expressed on Reddit forums precede and potentially incite notable movements in the market, thereby establishing a cause-and-effect link. This exploration is crucial as it offers insights into the predictive power of social media sentiment on financial markets, aligning the complex interplay between trader behavior, information dissemination, and market outcomes with the predictive modeling strategies adopted in our research.

Figure 19: Causality



The logistic regression model's analysis positioned Reddit sentiment as a less significant feature (Figure 18), indicating a relatively limited direct impact on stock prices. However, a closer examination through causality tests reveals a more complex scenario: for certain "meme" stocks like BB and GME, Reddit discussions show a notable causative effect on stock movements, but primarily in lags 2 to 5 rather than at the initial lag. This is in stark contrast to the IV30 feature, which, despite its significant role in the

logistic regression model, did not exhibit a similar causal relationship with stock prices. The delayed causality observed in BB and GME underscores a distinct dynamic, suggesting that shifts in Reddit sentiment may indeed have a delayed but tangible impact on the price movements of these specific stocks.

The causality analysis of overall sentiment indicates a broad influence across various market caps, with Microsoft (MSFT) as a singular outlier, raising questions about potential feedback mechanisms or third-variable influences such as market news. In contrast, weighted Reddit sentiment, which emphasizes significant shifts in community opinion, shows a more selective impact, predominantly affecting smaller "meme" stocks. This nuanced distinction suggests that while broad sentiment trends may reflect or react to market movements, intense sentiment consolidations are more likely to actively drive price fluctuations in stocks with smaller market caps, where collective actions of individual investors can have a more pronounced effect.

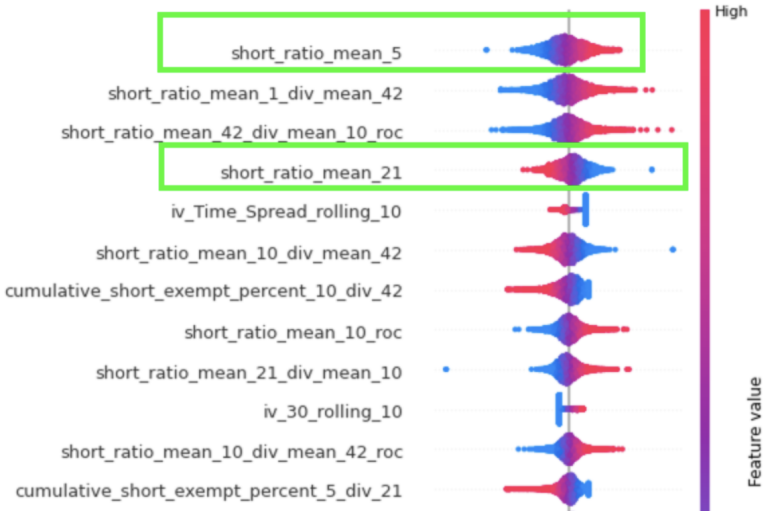
The model's subdued emphasis on Reddit sentiment features, despite their demonstrated causative effect on select stocks, likely stems from their targeted impact. Given that only a subset of "meme" stocks exhibit sensitivity to Reddit discussions, the model pragmatically dilutes sentiment indicators in favor of more universally applicable predictors. This strategic de-emphasis ensures broader applicability across diverse stock profiles, reflecting a calculated trade-off within the model's feature prioritization.

We must also note that financial markets are rife with feedback loops where investor sentiment and stock prices often create a cyclical influence on each other. Our analysis suggests the presence of such loops, particularly within Reddit's community and its impact on stock prices. However, the lack of corresponding patterns in options market causality tests raises questions. To understand the true nature of these dynamics, a deeper dive into the causality between Reddit sentiment and market behavior is warranted, potentially revealing a more complex interplay than our current model captures.

SHAP Analysis

The SHAP analysis largely corroborates the model's feature importance rankings, affirming the key drivers of our predictions. However, discrepancies do emerge, illuminating the nuanced influence of certain variables. For instance, where the model's feature importance may not fully capture the complexity, SHAP analysis provides a more granular understanding of how specific features sway the model's predictions (Awan, 2023), necessitating a closer examination of these differences to refine our predictive framework.

Figure 20: SHAP Analysis



The SHAP analysis reveals intriguing divergences from our model's feature importances, particularly around short sale volumes. For instance, SHAP indicates that a short-term increase in short sale volume (short_ratio_mean_5) predicts an uptick in stock prices, defying conventional expectations. This divergence prompts us to consider regulatory frameworks like REG-SHO, which govern short selling and provide exceptions for market makers. These market participants, vital for liquidity, are allowed to short sell without restriction under certain conditions, potentially explaining the observed volatility around significant trading events and influencing our model's predictions.

The observed short sale volumes in our data might be influenced by hedging activities around earnings announcements, a pivotal period for market volatility. Commonly, holders of significant stock positions, such as investment funds and endowments, seek downside protection by purchasing put options, which often leads to higher prices for short-dated put options. Market makers, who facilitate these trades, often short sell equities to maintain delta neutrality. This activity could explain the short-term surge in short sales captured by our model. However, this remains a hypothesis requiring further investigation. Our findings suggest a complex interplay between market events and stock movements, meriting additional research to fully comprehend and validate these dynamics in the context of stock events prediction.

In future analysis, a refined approach could involve targeting stocks where a causal relationship between Reddit sentiment and price changes is initially evident, especially focusing on "meme" stocks. Then, integrating short sale and options market data into this model could offer a more granular and nuanced understanding of market dynamics. This approach could unveil subtler correlations and causalities, particularly in stocks susceptible to social media influence. By expanding the model's data sources and focusing on specific stock categories, future studies can potentially yield more accurate and insightful predictions of market events influenced by collective sentiment and trading behaviors.

In conclusion, we were able to overcome many challenges associated with disparate parts of the Data Science discipline and produce a useful outcome. Using skills acquired across the MADS program and some newly learned, we worked together to produce our modeling outputs, analysis and dashboard.

X. Limitations and Challenges

Due to the nature of stock market fluctuations, our dashboard, model predictions, and related findings may not be applicable to all future short squeeze occurrences. One natural limitation is the small number of historical short squeeze occurrences which results in a low number of datapoints to reference in our analysis. In a professional setting, models are usually designed to be dynamic and able to pivot to ingest new inputs or be replaced with newer versions apt to face novel market conditions. Without regular maintenance and/or enhancements, our NLP model and dashboard face the risk of becoming obsolete in the future.

We are also subject to survivorship bias in our analysis since some tickers in our Reddit data represent companies that no longer trade. This skews model results towards currently active companies, potentially overlooking historical trends and insights from now-defunct entities. Thus we strongly caution against taking the outputs of the model as perfect. Without considerable more development and backtesting, we do not hold out our conclusions as investment advice.

XI. Broader Impact and Ethical Considerations

In developing a predictive trading tool, it is necessary to address accuracy, transparency, and legal compliance. Clear communication about the tool's limitations, data sources, and capabilities is essential to prevent misuse. Adherence to privacy laws and securing user consent is crucial, especially when using data from sources like Reddit. The tool must comply with securities laws to avoid issues like insider trading or market manipulation.

Effective testing and validation are required to ensure the tool's reliability. Users should be informed about how to interpret outputs and understand potential risks. The tool's reliance on historical data may limit its effectiveness in predicting future market trends, particularly in volatile conditions. There is also a risk of legal liability if users suffer losses based on the tool's recommendations. Furthermore, if widely adopted the tool could create feedback loops, influencing the variables it analyzes and potentially causing market distortions.

Ethical considerations and potential market impacts should be central to the tool's development and deployment. Ongoing updates, straightforward communication, and adherence to legal and ethical standards are necessary to address these issues.

XII. Statement of Work

William Gorfein	Data Gathering, Data Storage, Options Analysis, AWS Admin, Dashboard, Feature Selection, Modeling

Jake Mason	Data Gathering, Vader Sentiment Analysis using Reddit, Feature Selection, Modeling
Erick Telenchana	GitHub Repository Maintenance, Code Review, Visualizations, Project Management
All	AWS, Mini Deliverables, Visualizations/Design, Final Paper, Video and Slides

XIII. References

Ali Shojaie and Emily B. Fox University of Washington. (2022, December). *Granger causality: A review and recent advances*. Granger Causality: A Review and Recent Advances. https://ar5iv.labs.arxiv.org/html/2105.02675

Awan, A. A. (2023, June 28). *An introduction to shap values and machine learning interpretability*. DataCamp.https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Czakon, J. (2023, September 5). *F1 score vs ROC AUC vs Accuracy Vs PR AUC: Which evaluation metric should you choose?*. neptune.ai. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Dominguez, M. (2021, June 9). *Wallstreetbets sentiment analysis on stock prices using natural language processing*. Nerd For Tech. Retrieved from: https://medium.com/nerd-for-tech/wallstreetbets-sentiment-analysis-on-stock-prices-using-natural-language-processing-ed1e9e109a37

Ganti, A. (2023, September 29). *How implied volatility (IV) works with options and examples*. Investopedia. Retrieved from https://www.investopedia.com/terms/i/iv.asp

Hutto, C., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. In Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Retrieved from https://doi.org/10.1609/icwsm.v8i1.14550

Hutto, C.J., & Gilbert, E. (2014). *VADER Sentiment Analysis* (Version 0.5) [Software]. GitHub. Retrieved from https://github.com/cjhutto/vaderSentiment

Konstantin, T. (2021). *Reddit WallStreetBets Posts Sentiment Analysis*. Retrieved from https://www.kaggle.com/code/thomaskonstantin/reddit-wallstreetbets-posts-sentiment-analysis

LDNOOBW. (2023). *List of Dirty, Naughty, Obscene, and Otherwise Bad Words*. GitHub. Retrieved from https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Mitchell, C. (2023, June 29). *Short squeeze: Meaning, overview, and faqs*. Investopedia. Retrieved from https://www.investopedia.com/terms/s/shortsqueeze.asp

Mitchell, C. (2022, July 14). *What is short interest, and why does it matter to traders?*. Investopedia. Retrieved from https://www.investopedia.com/terms/s/shortinterest.asp

NASDAQ. (2023). *Stock Screener*. Retrieved 11/27/2023, from https://www.nasdaq.com/market-activity/stocks/screener

Prabhakaran, S. (2022, October 10). *Principal component analysis – how PCA algorithms works, the concept, math and implementation: ML+*. Machine Learning Plus. https://www.machinelearningplus.com/machine-learning/principal-components-analysis-pca-better-explained/

Praw-dev. (2023). *PRAW (Python Reddit API Wrapper)* (Version 7.7.1) [Software]. GitHub. Retrieved from https://github.com/praw-dev/praw

Shapiro, Adam Hale, Moritz Sudhof, Daniel Wilson. (2020). *Measuring News Sentiment*, Federal Reserve Bank of San Francisco Working Paper 2017-01. Retrieved from https://www.frbsf.org/wp-content/uploads/sites/4/wp2017-01.pdf

Staff Writer. (2023, March 22). *5 biggest short squeezes in the last 25 years*. RoboMarkets Blog. Retrieved from https://www.robomarkets.com/blog/education/5-large-short-squeezes-of-the-last-25-years/

Appendix A

GitHub Repository: https://github.com/erickts643/SIADS_Capstone_Group17

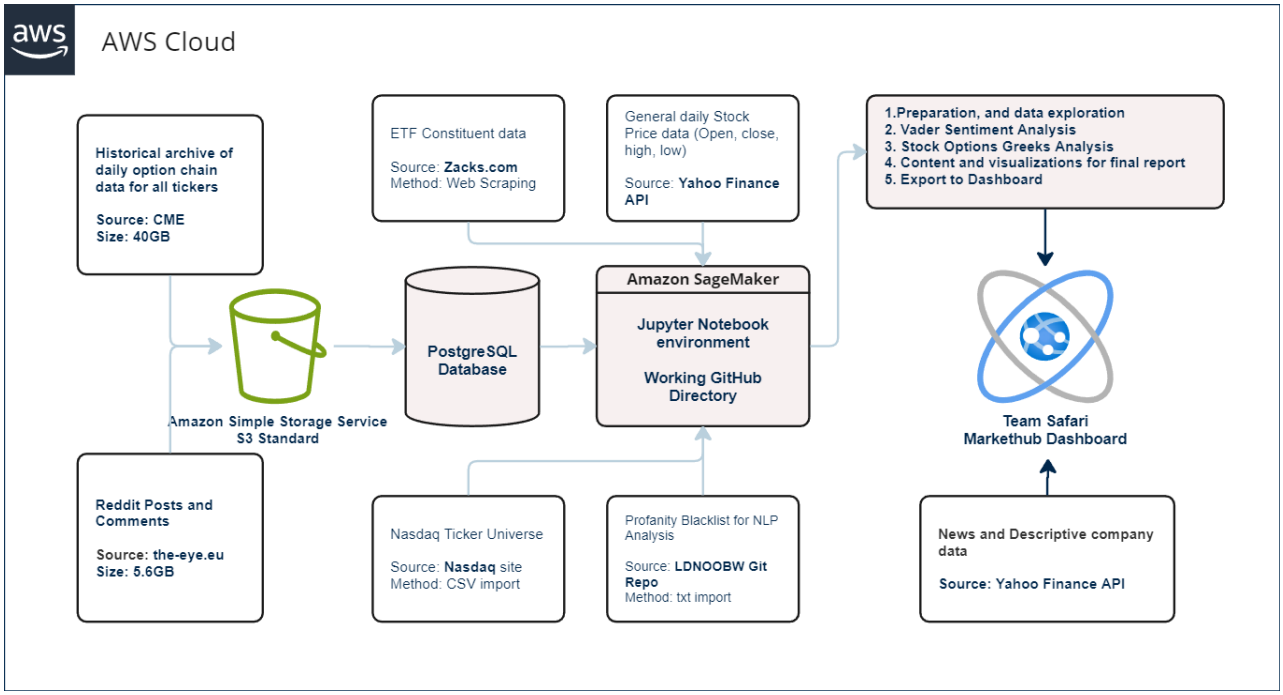
Appendix B

Financial Definitions:

1	Delta	Stock option metric which measures the rate of change in an option's price in relation to the underlying stock. *Example: a delta of 0.5 indicates the option's price changes by $50 for every $1 change in the underlying stock's price.*
2	Gamma	The rate of change in the option's delta per $1 change in the underlying stock's price
3	Implied Volatility (IV)	A forward-looking estimate of the percent move the market expects in the movement of a stock. IV is calculated as part of the Black-Scholes options pricing model
4	Short Interest	The number of outstanding shares that have been sold short and remain outstanding (i.e. waiting for the short position to be closed.)
5	Short Squeeze	A situation in which the price of a stock rises to such an extent that investors who have sold short purchase the stock in order to limit their losses, causing the price to rise further.
6	Stock Option	A contract (either a 'call option' or a 'put option' giving the user the right to buy or sell 100 shares of the underlying stock at a predefined price

Appendix C

AWS Cloud Workflow Diagram



Appendix D

Submissions Table Schema

Field	Description
id	A unique identifier for each submission.
subreddit_id	The name of the subreddit where the submission was posted.
subreddit	The name of the subreddit where the submission was posted.
author	The username of the individual who posted the submission.

created_utc	The timestamp when the submission was created, in UTC format.
permalink	A permanent link to the submission.
title	The title of the submission.
selftext	The body text of the submission, if any.
num_comments	The number of comments on the submission.
score	The net upvotes of the submission.

Appendix E

Federal Reserve Bank of San Francisco: Daily News Sentiment Index

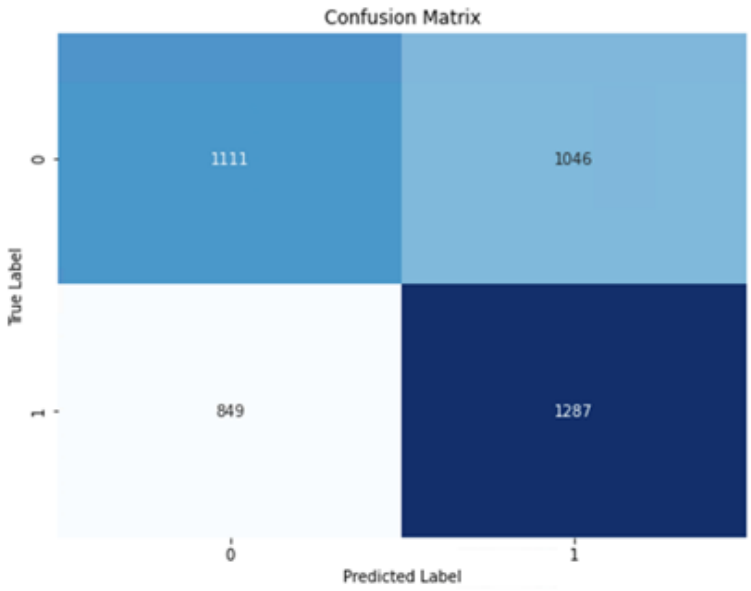
The Daily News Sentiment Index is a high-frequency indicator of economic sentiment derived from lexical analysis of economics-related news articles. It analyzes sentiment scores from news articles in 24 major U.S. newspapers. These articles are selected based on a minimum word count and topics surrounding US economics.

The sentiment-scoring model, which combines public lexicons with a news-specific lexicon created by the authors, is tailored for newspaper articles. The scores from individual articles are aggregated daily. The Daily News Sentiment Index is a trailing weighted-average of these time series. It is then updated weekly, providing a regular measure of economic sentiment from news media.

It can be found here:
https://www.frbsf.org/economic-research/indicators-data/daily-news-sentiment-index/

Appendix F

Original Confusion Matrix and Calculations from Figure 10:



Note: a differently formatted version of the Confusion Matrix (with the same content) was included in the paper for better readability

```
# Confusion matrix values
true_positives = 1287
true_negatives = 1111
false_positives = 1046
false_negatives = 849

# Calculate precision, recall, and F1 score
precision = true_positives / (true_positives + false_positives)
recall = true_positives / (true_positives + false_negatives)
f1_score = 2 * (precision * recall) / (precision + recall)
```

Short Ratio vs. S&P 500 (SPY)

The provided chart offers a comparative analysis of GameStop (GME) and the S&P 500 (SPY), focusing on the rolling short ratio volume rather than price. The visualization includes 1.5 and 2 standard deviation bands for SPY's short ratio volume, establishing a benchmark for typical market behavior. Significantly, the chart highlights instances, shaded in green, where GME's short ratio volume exceeded the 1.5 and 2 standard deviation thresholds of SPY. These peaks coincide with the meme stock phenomenon during 2020 and 2021, emphasizing periods when GME experienced exceptionally heavy shorting activity that far surpassed the market's normative range, reflecting its central role in the extraordinary market events of that time.