

stock prices

An Analysis of r/Wallstreetbets's effect on stock prices

Capstone Project - Ryan

## **Business Problem Statement**







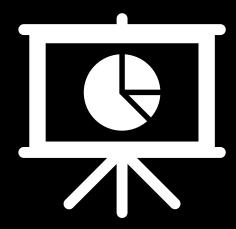
- ► Seen WSB influence on gme
- Wants to use WSB's activity to give the bank a trading edge



## Data Problem Statements

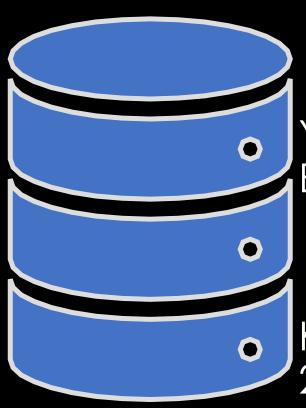


Is there a relationship between the sentiment or number of comments on WSB and stock prices?



Can we use stock data and WSB comments data to predict the next day's price movement?

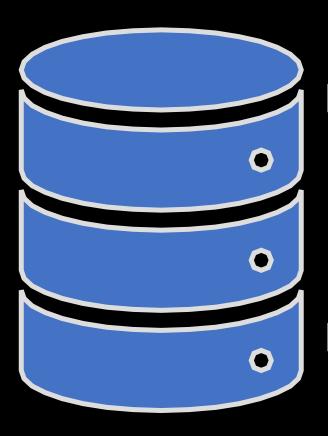
## Data sources



Yahoo Finance: GME, TSLA, PLTR, NOK, BB, SP500, AMC

Kaggle: WSB comments (Feb-2018 to Feb 2021)

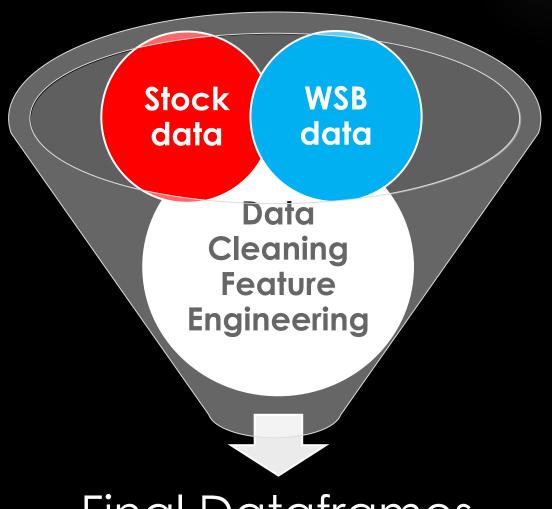
### Data sources



- ▶ Stock data columns:
  - ▶Open, High, Low, Close, Volume

- ▶ WSB comments columns:
  - ▶created utc, text

# Data preparation



Final Dataframes

## Data Cleaning / Feature Engineering

(WSB comments)



Remove NA values



Apply spacy's sentiment and subjectivity models.

#### Available Columns

- created\_utc
- ▶ text
- sentiment
- ▶ Subjectivity
- ▶ month
- day\_of\_week



- Convert unix timestamp (created\_utc) to tz-aware datetime object
- Get day\_of\_week and month label from created\_utc

# Data Cleaning / Feature Engineering

(Stock data)



Remove all data prior to 16<sup>th</sup> Feb 2018



Compared High and Low values to get pct\_volatility column



Compared Open and Close values to get pct\_change column

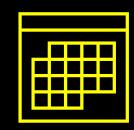


Converted

Date column to
tz-aware
datetime object



Compared Open and Close values to get bool column up\_today and up tomorrow

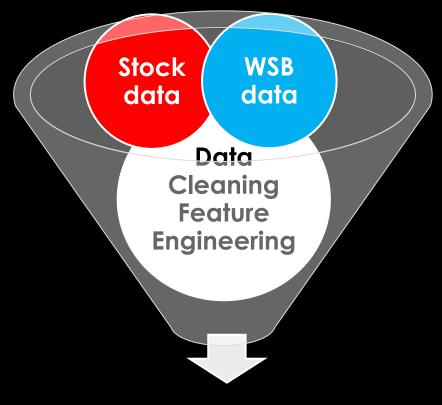


Get
day\_of\_week
and month
label from Date

#### Available Columns

- Date
- Open
- ▶ High
- ▶ Close
- ▶ Low
- ▶ Adj Close
- Volume
- pct change
- up\_today
- up\_tomorrow
- pct\_volatility
- day\_of\_week
- month

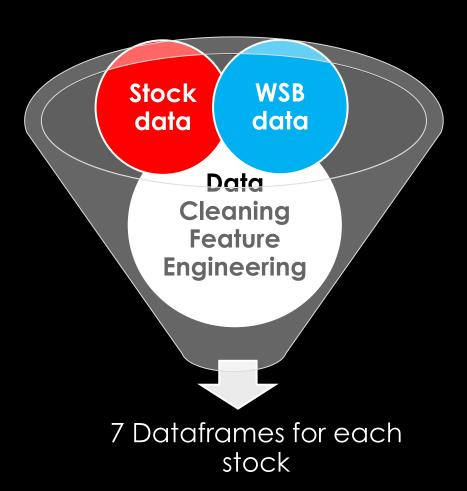
## Data preparation



Final Dataframes

- Group wsb data by date and obtain the following columns:
  - ▶ Mean subjectivity
  - ▶ Mean sentiment
  - ▶ Count of comments
- ► Left join wsb data on stock data to obtain 7 different dataframes
- ► Target variable is up\_tomorrow

## Final Dataframe Columns



#### Features:

▶ Open: float64

► High: float64

▶ Low:float64

► Close: float64

▶ Volume: float64

up\_today:bool

pct\_change:float64

▶ sentiment : float64

subjectivity: float64

► comment\_count:int64

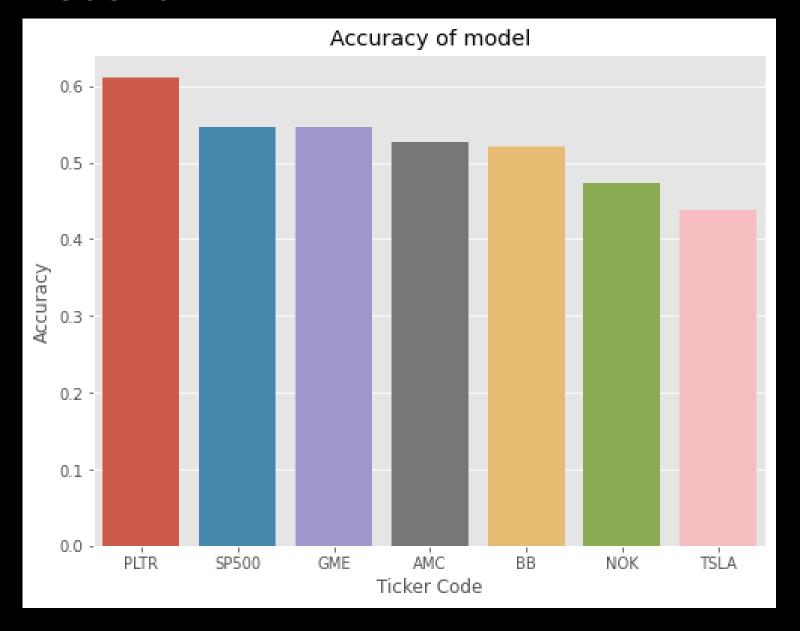
day\_of\_week: category

month: category

Prediction column:

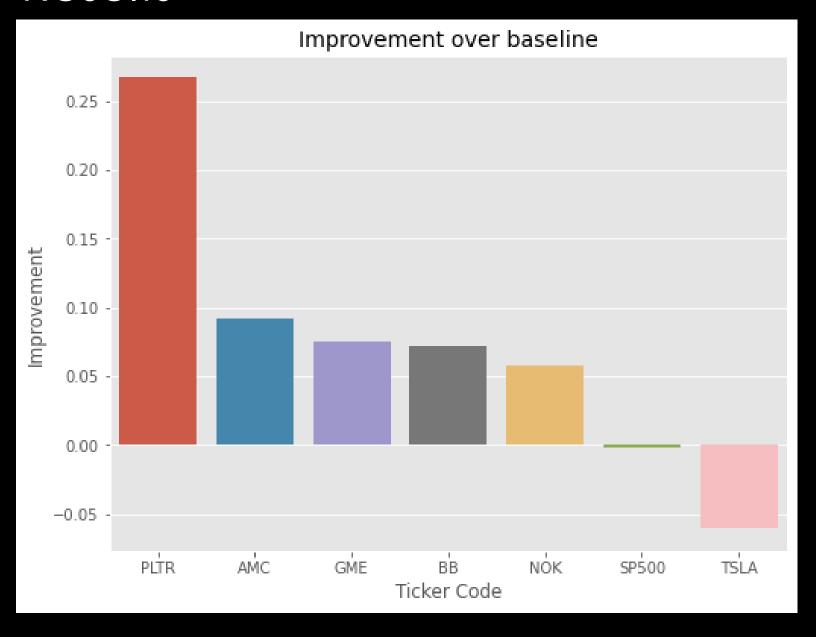
up\_tomorrow:bool

#### Results



- ► Model used:
- ► XGBoost Classifier:
- ▶80%:20% train\_test\_split
- ▶ Best accuracy performer is PLTR, over 60% accuracy
- Model predicted the right movement 60% of the time

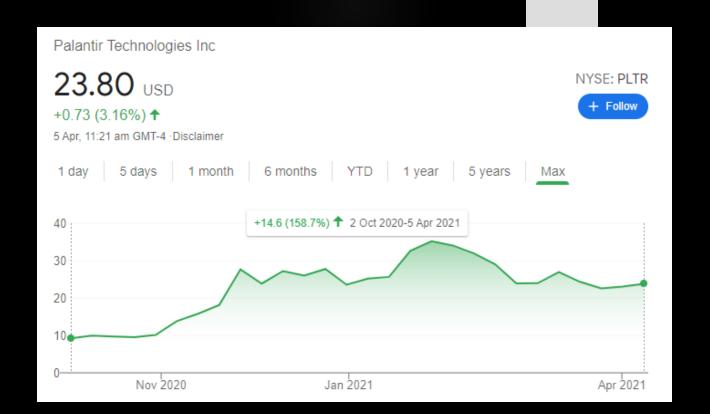
## Results

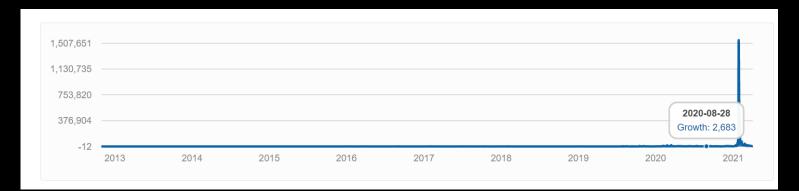


- Best improved model over baseline is PLTR
- ▶ Baseline: predict stock is up every day

## Analysis of results

- ▶ PLTR IPO'd recently and stock data is only from 2<sup>nd</sup> Oct 2020 onwards
- ► For all other stocks, the analysis runs from Feb-2018 to Feb-2021
- Aug 2020 was the start of WSB's huge membership growth
- WSB's influence initially is low due to low membership count





# Feature Importances

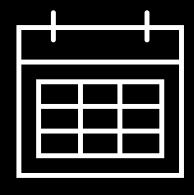
- ▶ PLTR affected by low, and whether or not it is wednesday
- ► TSLA affected by whether or not it is monday and may
- Most affected by WSB sentiment & subjectivity—PLTR
- Most affected by WSBcomment\_count AMC
- Most affected by the month of January – GME
- Most affected by the month of February

Results Heatmap									
Open -	0.032	0.04	0.031	0.042	0.029	0.028	0.062		- 0.200
High -	0.037	0.038	0.035	0.064	0.038	0.026	0.072		0.200
Low -	0.033	0.038	0.044	0.049	0.04	0.018	0.17		
Close -	0.03	0.035	0.025	0.028	0.033	0.028	0.082		- 0.175
Volume -	0.025	0.034	0.039	0.044	0.03	0.046	0.082		-0.175
up_today -	0.034	0.021	0.048	0.024	0.035	0.0059	0.053		
pct_change -	0.039	0.032	0.031	0.039	0.032	0.04	0.066		
pct_volatility -	0.03	0.036	0.032	0.041	0.032	0.028	0.072		- 0.150
sentiment -	0.03	0.031	0.036	0.033	0.034	0.034	0.05		
subjectivity -	0.035	0.032	0.036	0.044	0.034	0.025	0.049		
comment_count -	0.031	0.031	0.031	0.046	0.033	0.028	0.04		- 0.125
day_of_week_Monday -	0.058	0.012	0.044	0.031	0.052	0.13	0		
day_of_week_Thursday -	0.04	0.034	0.039	0.055	0.024	0.045	0		
day_of_week_Tuesday -	0.036	0.037	0.04	0.025	0.035	0.037	0		- 0.100
day_of_week_Wednesday -	0.032	0.054	0.024	0.049	0.052	0.028	0.2		
month_August -	0.034	0.062	0.014	0.087	0.032	0.088			
month_December -	0.047	0.02	0.034	0.043	0.032	0.093			- 0.075
month_February -	0.044	0.11	0.014	0.067	0.061	0	0		
month_January -	0.08	0.048	0.051	0	0.14	0	0		
month_July -	0.057	0.037	0.036	0.036	0.063	0.021			- 0.050
month_June -	0	0.015	0.032	0.03	0.017	0			
month_March -	0.029	0.045	0.029	0.05	0.032	0			
month_May -	0.039	0.044	0.094	0.0088	0.011	0.21			- 0.025
month_November -	0.029	0.054	0.053	0.023	0.029	0	0		0.025
month_October -	0.051	0.023	0.059	0	0.026	0.05	0		
month_September -	0.067	0.039	0.051	0.043	0.024	0	0		0.000
	SP500	GME	вв	AMC	NÓK	TSLA	PLTR		- 0.000

#### Future work



- ▶ Train a BERT Model on labelled data for sentiment analysis
- ► (Bidirectional Encoder Representations from Transformers)



- Narrow the timeframe down to Aug-2020 onwards as this is the period of WSB's explosive growth
  - Downsides: lack of data

### Conclusion

- ► The results of PLTR show that WSB is starting to have influence on the stock market
  - ► The cult-like behavior in WSB is contagious
  - WSB encourages people to hold stocks through memes something never done before
  - ▶ There is also the desire to get back at hedge funds for the 08 crisis
- The investing landscape has changed significantly since the Warren Buffet days
  - People were expecting a crash but no crash came
  - ► S&P broke 4000 during the middle of a pandemic
  - ▶ The economy and the stock market has decoupled