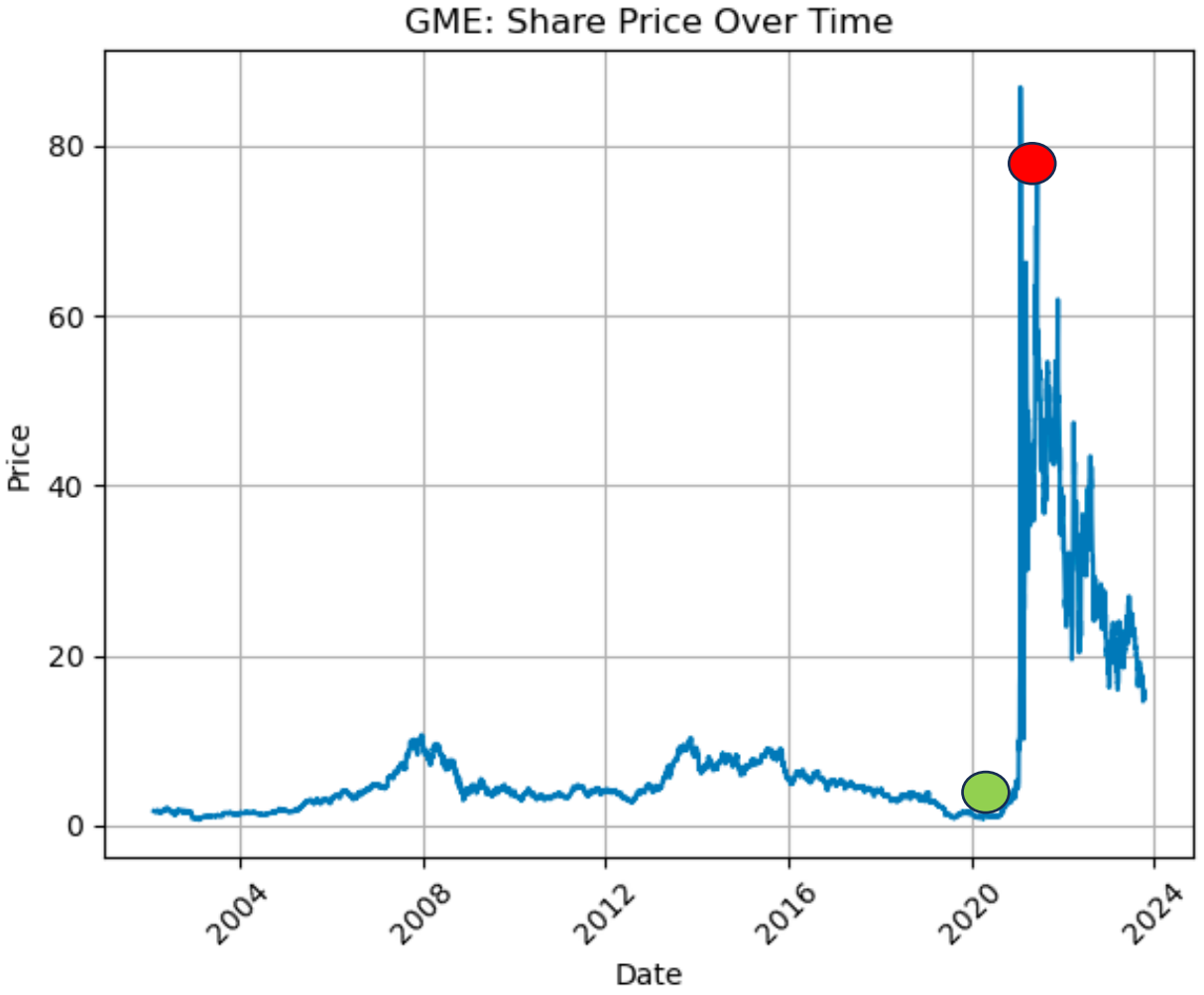
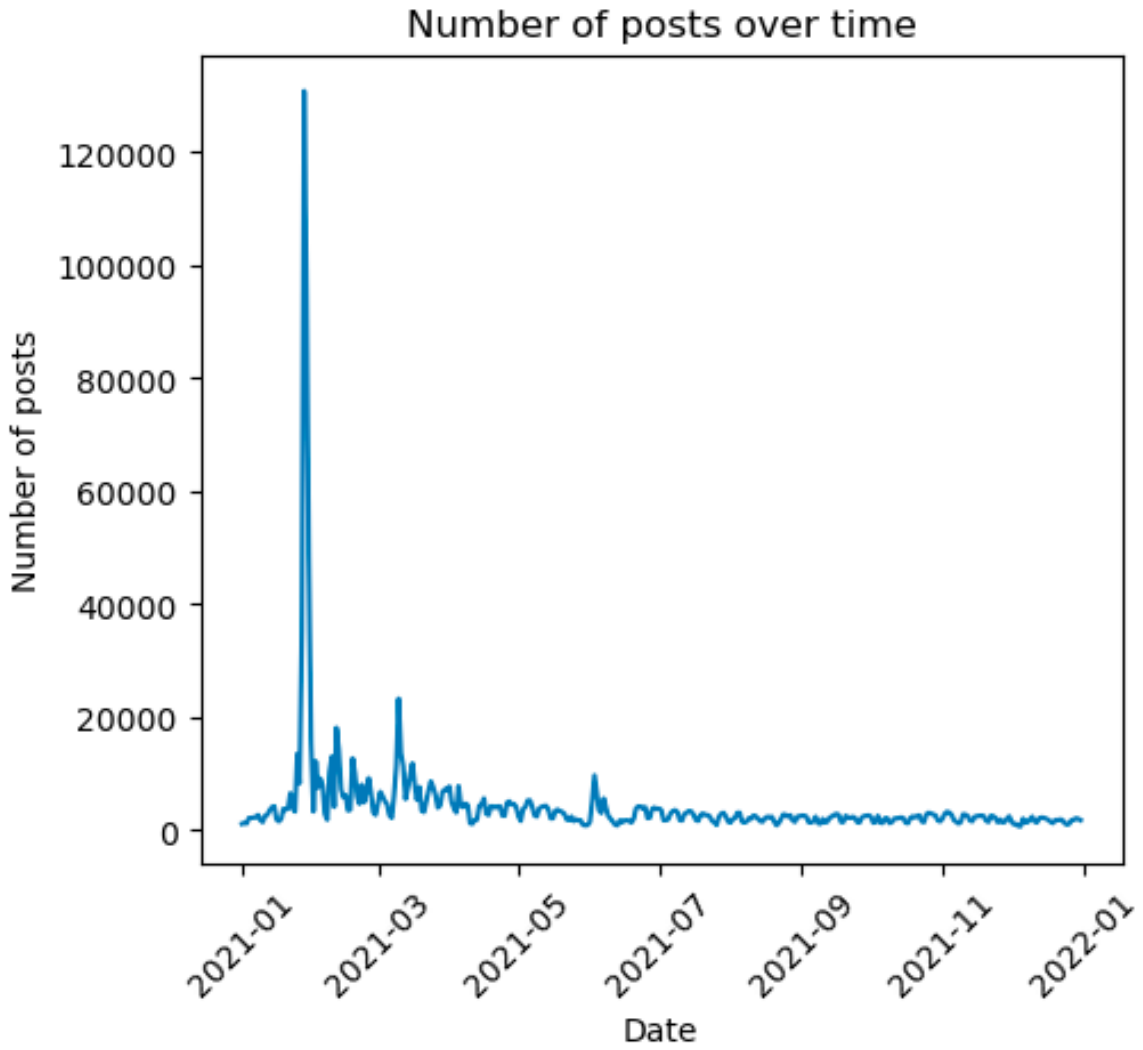
Identifying precise forecasters

Lucie Gavrilova

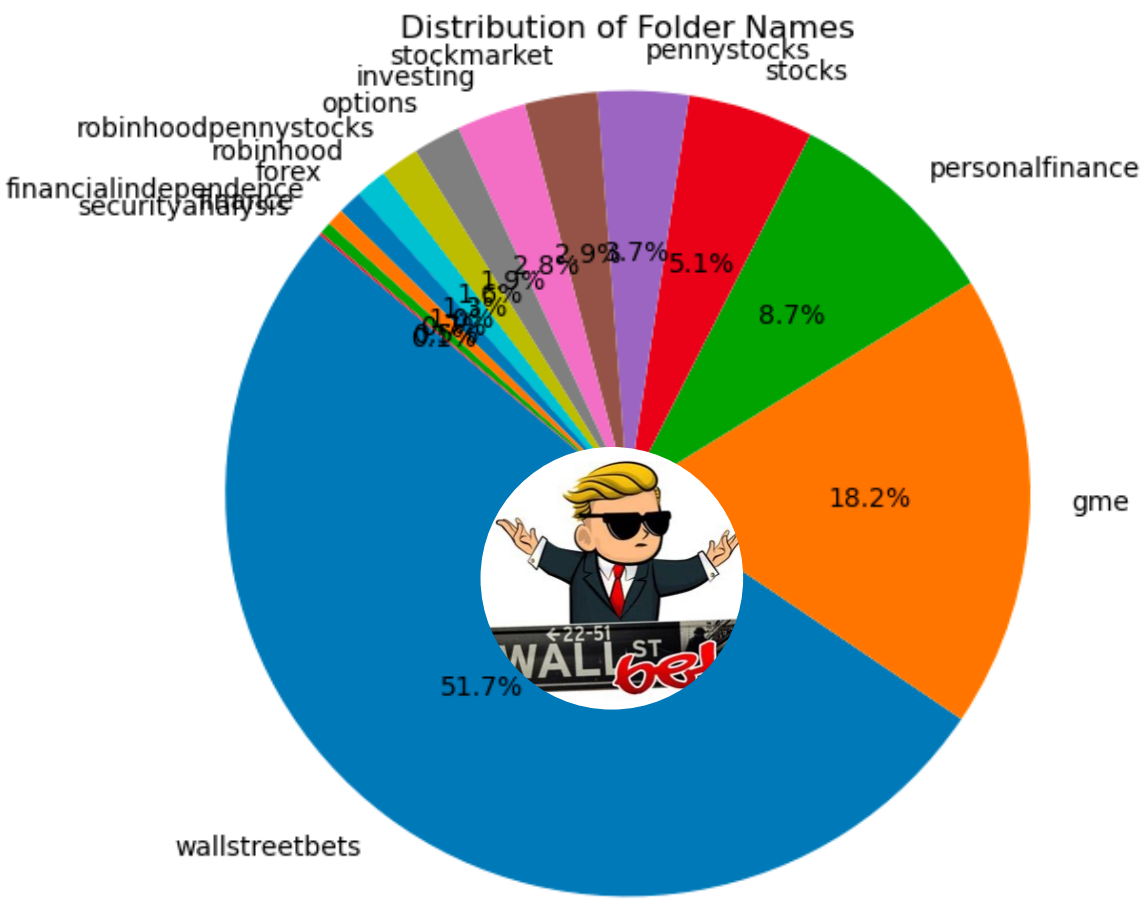
BrainStation Capstone Project – Sprint 2

GameStop share price and Wallstreetbets subreddit

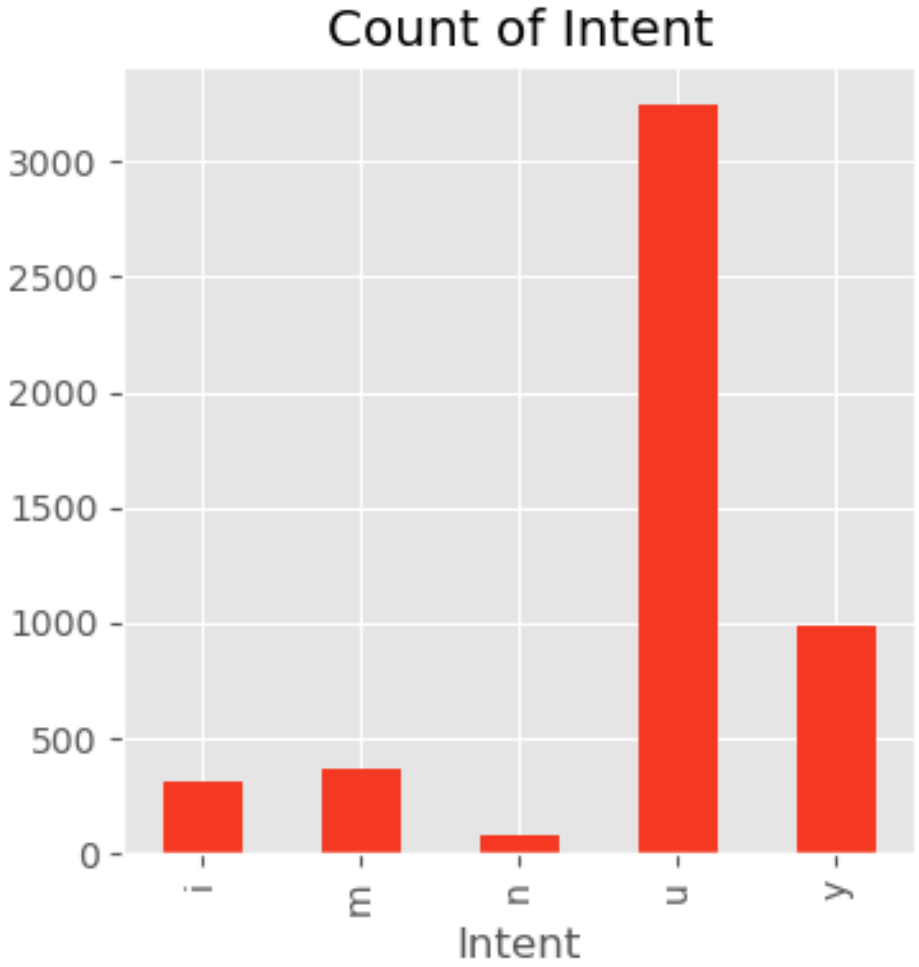




The Data Sources



1.5m rows



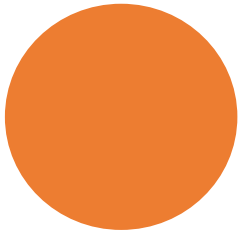
5000 rows

The Datasets: sneak a peek

	link_id	parent_id	User	Text	Intent	Support
737	t3_mk3mcj	t3_mk3mcj	Caeden27	This is why I put all of my money into GME	y	y
4206	t3_ladzdt	t3_ladzdt	MetalliTooL	They should really ban all GME discussion, out...	u	y
4659	t3_lbgjic	t3_lbgjic	EnergyMatrix	Buy more GME🚀	u	y

	link_id	parent_id	User	Text	Intent	Support
2593	t3_lxva6s	t3_lxva6s	violauh	I hope gme doesn't moon till the 12th so I can...	y	u
284	t3_l7kfrj	t3_l7kfrj	chrisred244	Okay im going 5g deep into GME. I'm in UK what...	y	u
1526	t3_l6er79	t3_l6er79	KeybordKat	Just emptied all my other investments to get m...	y	y
4802	t3_kxozk1	t1_gjbjosx	willard507	Lies! Don't listen to the 🍭🐻\n\nGME🚀🚀🚀🚀🚀🚀🌙🌙🌙	u	y
2782	t3_lbqs4l	t1_glvhmz6	InvincibearREAL	Someone bought min $21.5M worth of $GME	u	u

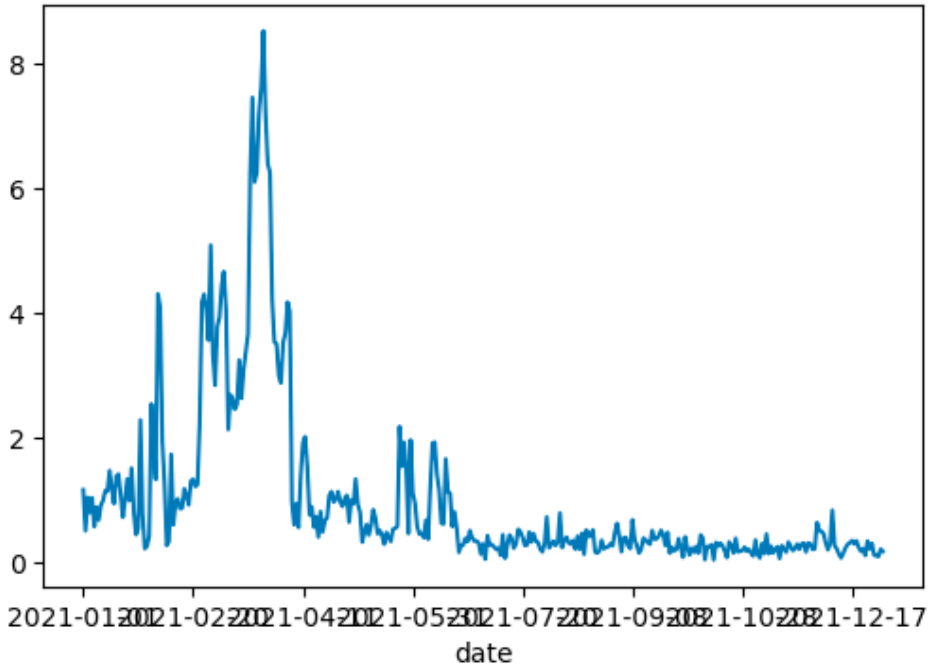
	link_id	parent_id	User	Text	Intent	Support
2481	t3_lzyi84	t1_gq60npy	BIG_MONEY_HUNTER	I haven't wasted any of my retirement on joke ...	y	u
1031	t3_l0hhqg	t1_gjvanag	shallow-pedantic	100% GME or close your account. Period.	u	y
1299	t3_l2x7k0	t1_gk9zsdm	wallawalla_	> potenntial regulations on retail investor...	u	y
2159	t3_kqzpqu	t3_kqzpqu	FentonBlustery	GME 🚀💎🙌	u	y
593	t3_m3hkjf	t3_m3hkjf	Superman0283	gme is the only great stock that is green toda...	u	y



Wallstreetbets dictionary (examples)

to the moon	extremely optimistic
retards	fellow traders
ape	fellow trader
degenerates	folks
diamond hands	patient investors
paper hands	impatient investors
rocket ships	very optimistic
shitpost	not very deep discussion

Rocket emoji usage 🚀🚀🚀🚀🚀🚀
Frequency of usage as % of posts



Model Training

- Pro-processing
  - Custom dictionary applied
  - Emojis removed and counted
- Defining training/test sets
  - 5000 rows only
  - Class 0: 73%, Class 1: 27%
  - Training 80%, Test 20%
- Vectorization into a sparce matrix
  - TF-IDF
- SMOTE applied to address class imbalance
- Machine Learning Models
  - Baseline Logistic Regression
  - Gridsearch for optimised hyperparameters
  - Decision Tree Classifier +PCA+ gridsearch
- Sentiment Analysis
  - TextBlob from NLTK
  - VADER sentiment analysis
- Challenges
  - Possible bias in labelling
  - Small dataset
  - Intentional misspellings

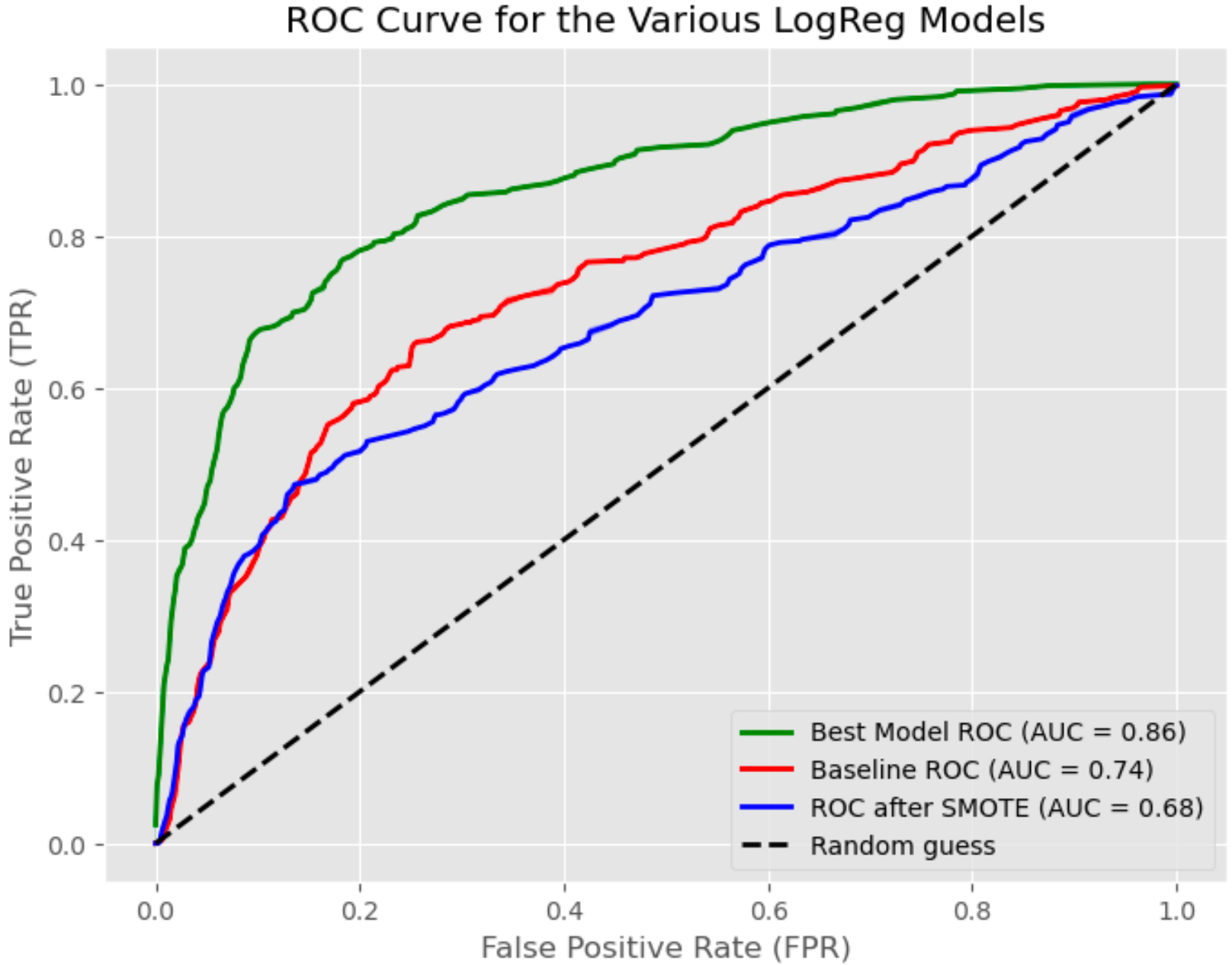
Best model results (optimized logreg):

Classification Report:

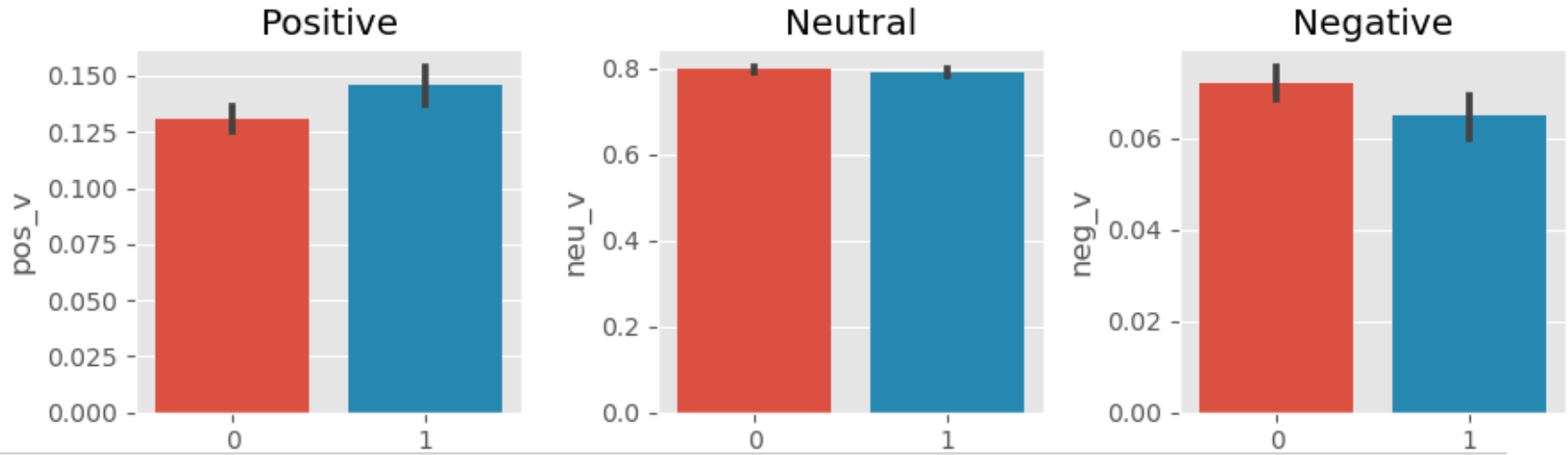
	precision	recall	f1-score	support
0	0.87	0.92	0.89	729
1	0.74	0.62	0.68	271
accuracy			0.84	1000
macro avg	0.80	0.77	0.78	1000
weighted avg	0.83	0.84	0.83	1000

Confusion Matrix:

```
[[669  60]
 [102 169]]
```


ROC Curve for the Various LogReg Models
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Best Model ROC (AUC = 0.86)
Baseline ROC (AUC = 0.74)
ROC after SMOTE (AUC = 0.68)
Random guess

Sentiment Analysis with TextBlob and VADER




```
DADDY_BOPPER
HOLD THE LINE. THE DOWNS ARE THE SHORTERS. WE WIN THIS GAME EASY. Not financial advice
0
1
{'neg': 0.0, 'neu': 0.632, 'pos': 0.368, 'compound': 0.8468}
```

```
Tale_Greedy
STILL HOLDINHG!!!!!!!! GME TO THE FUCKINGH MOON
1
1
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
Pinkpladedlumberjack
Holy shit Boys gme at 405$ gonna be 500 at open
0
1
{'neg': 0.265, 'neu': 0.735, 'pos': 0.0, 'compound': -0.5574}
```

Capstone Project - Wallstreetbets Oct02-Nov29

- Sprint 0 - Choose a topic Oct02-Oct06
- Find a dataset Oct06-Oct09
- Create GitHub, README, environment Oct16-Oct19
- Sprint 1 - EDA Oct13-Oct19
- Clean the text Oct20-Oct23
- Sentiment Extraction (VADER) Oct24-Oct26
- Vectorisation. Pre-processing Oct27-Oct30
- Logistic Regression Nov01-Nov03
- Portfolio of buys and sells Nov03-Nov07
- Sprint 2 - Baseline Modelling Nov07-Nov10
- Backtesting financial preformance Nov10-Nov13
- Refine Entity Resolution, Dependency, Part Of Speech etc. for each post Nov15-Nov20
- Instead of vectorise, try embeddings - word2vec Nov21-Nov23
- Sprint 3 - model optimization, evaluation and interpretation Nov24-Nov27
- Report, video and presentation Nov23-Nov29

Words predictive of intent to buy stock