r/Bitcoin and r/Ethereum

Hayden Tsutsui

Using NLP on Reddit to Understand Bitcoin and Ethereum

Background

Bitcoin
2020 Performance (YoY)
- 126.3B to 536B (324%)
- Current MC: 721B

Ethereum
2020 Performance (YoY)
- 14B to 85B (507%)
- Current MC: 133B

How? Why?

Macro Drivers

- Fed and central banks around the world continue to print fiat and devalue their currencies
- Product Zeitgeist Fit (PZF)
  - PMF with perfect timing
- Blockchain technology
  - Decentralized, trustless, permissionless, peer-to-peer transactions and information

Institutional Adoption

- PayPal, Square, OCC, investment funds, and more

Wait, what?

- What is blockchain?
- So, Bitcoins come from thin air?
- What can I buy Bitcoin with?
- What is difference between Bitcoin and Ethereum?
- How do I buy it?

FUD

Fear. Uncertainty. Doubt.


I DON'T KNOW
HOW BITCOIN WORKS
AND AT THIS POINT
I'M TOO AFRAID TO ASK
imgflip.com


ME TRYING TO EXPLAIN
ETHEREUM SHARDING

Machine Learning to Understand r/Bitcoin and r/Ethereum

Process

- Scrape most recent posts from each subreddit
- Clean text data. Create Sentiment column.
- Use machine learning to classify whether a post belongs to r/Bitcoin or r/Ethereum
- Observe key terms that are good indicators of r/Bitcoin and r/Ethereum subreddits

Learn a little about what these communities are currently talking about

GridSearch and Tune Hyperparameters of RandomForestClassifier

	0
features__text_features__cvec__max_df	0.15
features__text_features__cvec__max_features	7185.00
features__text_features__cvec__min_df	2.00
features__text_features__cvec__ngram_range	1.00
rf__max_depth	500.00
rf__min_samples_leaf	20.00
rf__min_samples_split	20.00
rf__n_estimators	50.00

	score
cross_val_score	0.816430
train_set_score	0.835282
test_set_score	0.820825

Classification Metrics

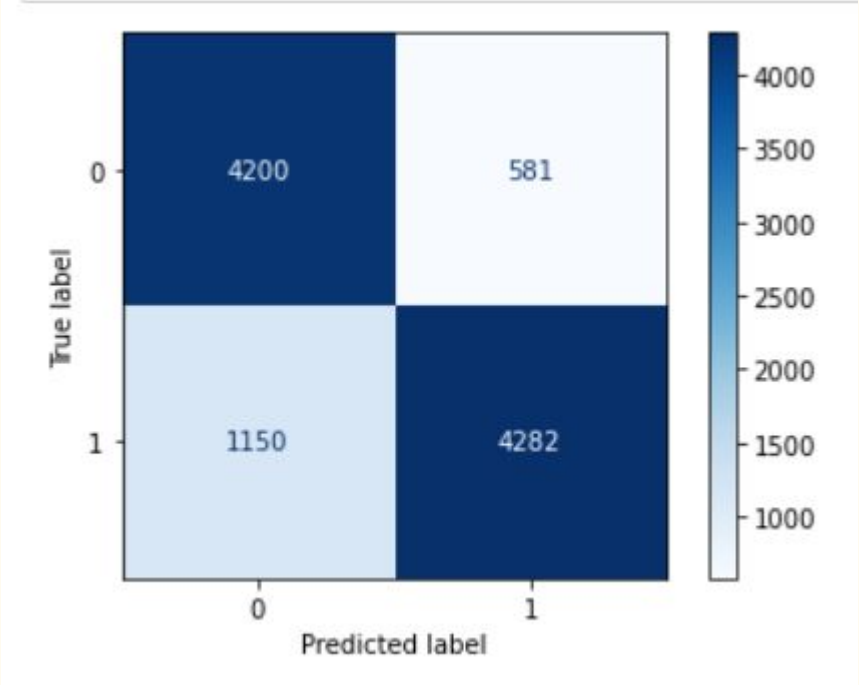
Accuracy: 82%

Sensitivity: 78%

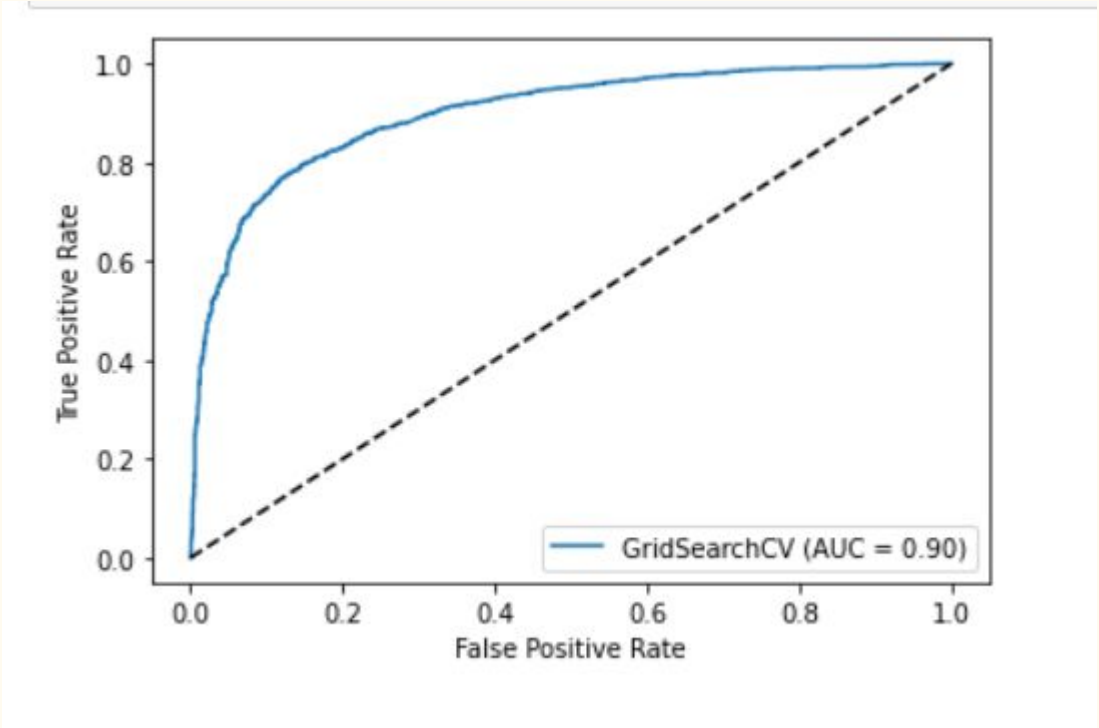
Specificity: 87%

Precision: 87%

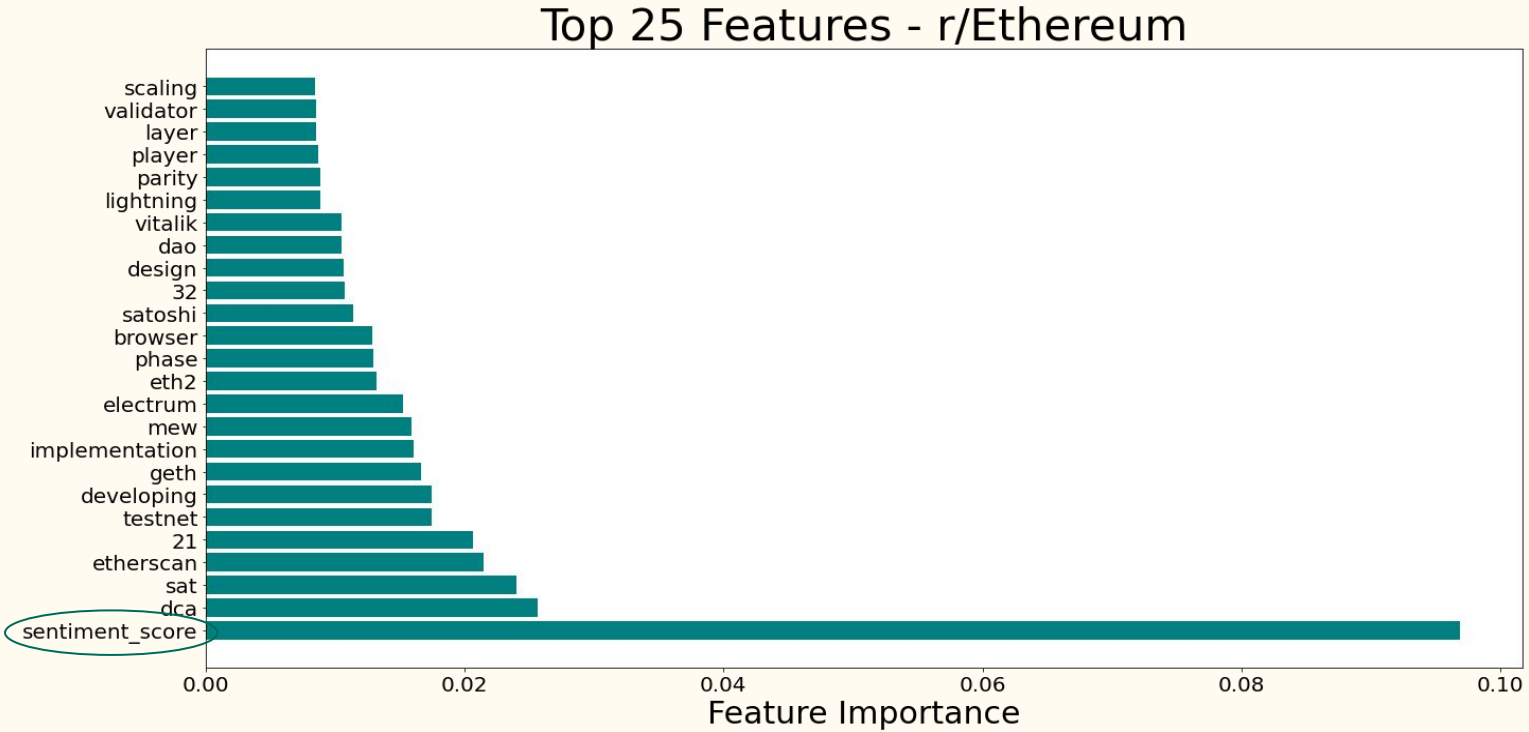
F1 Score: 82%



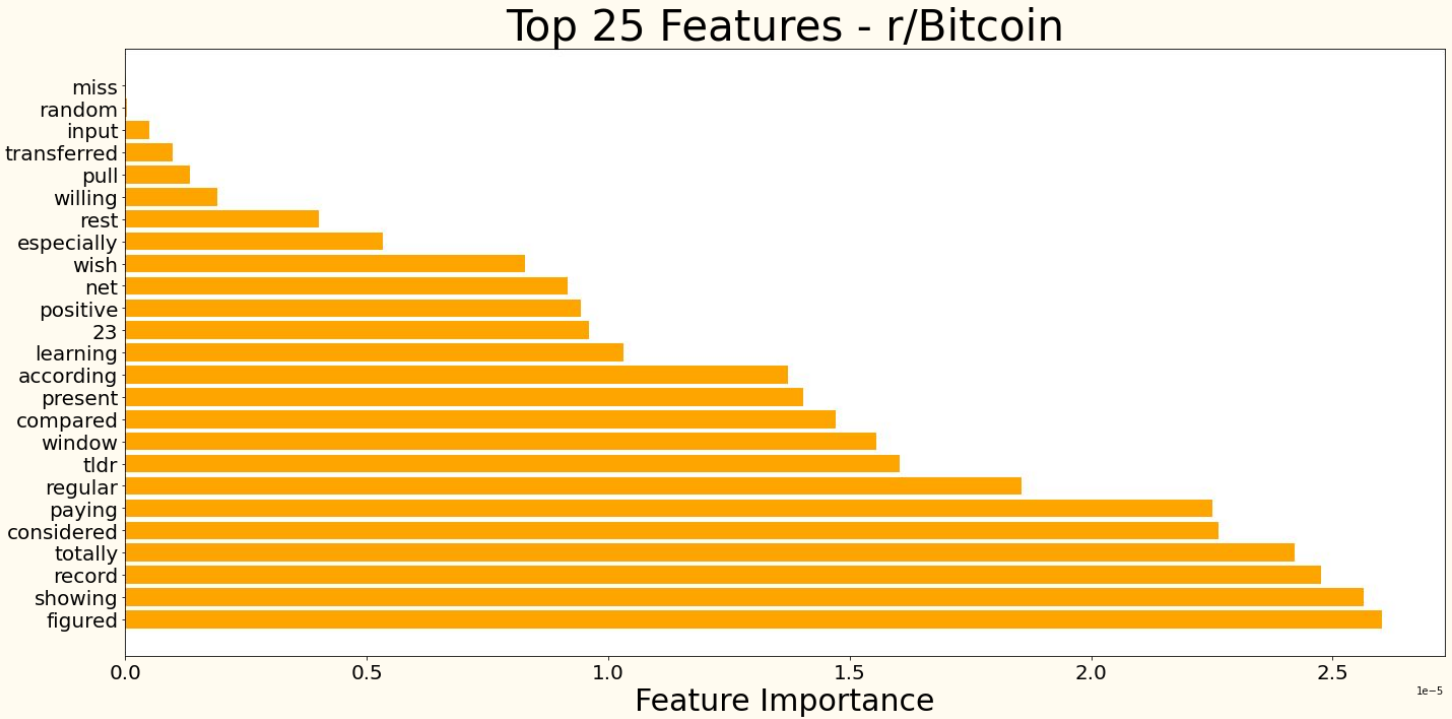
ROC Curve


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
GridSearchCV (AUC = 0.90)

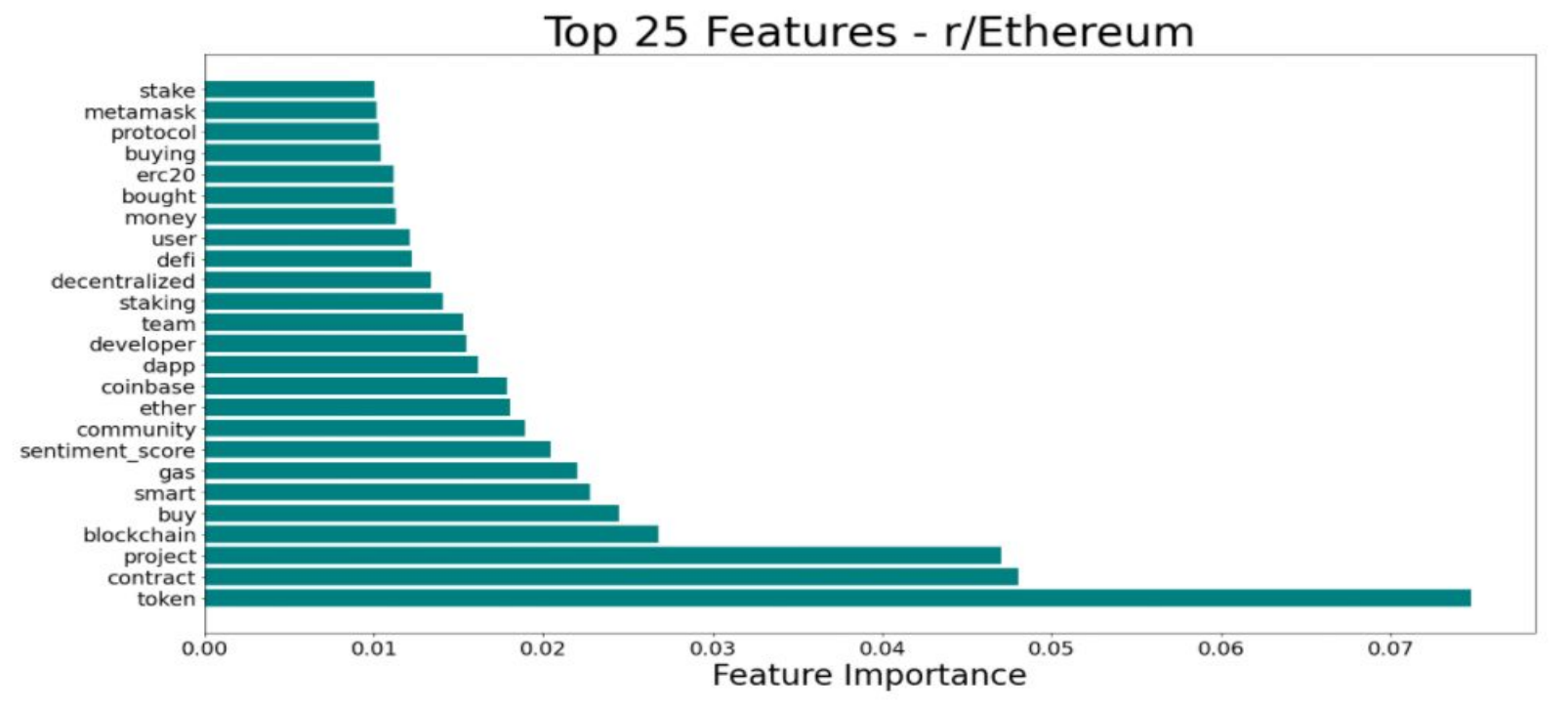
Feature Importance of Random Forest Classification



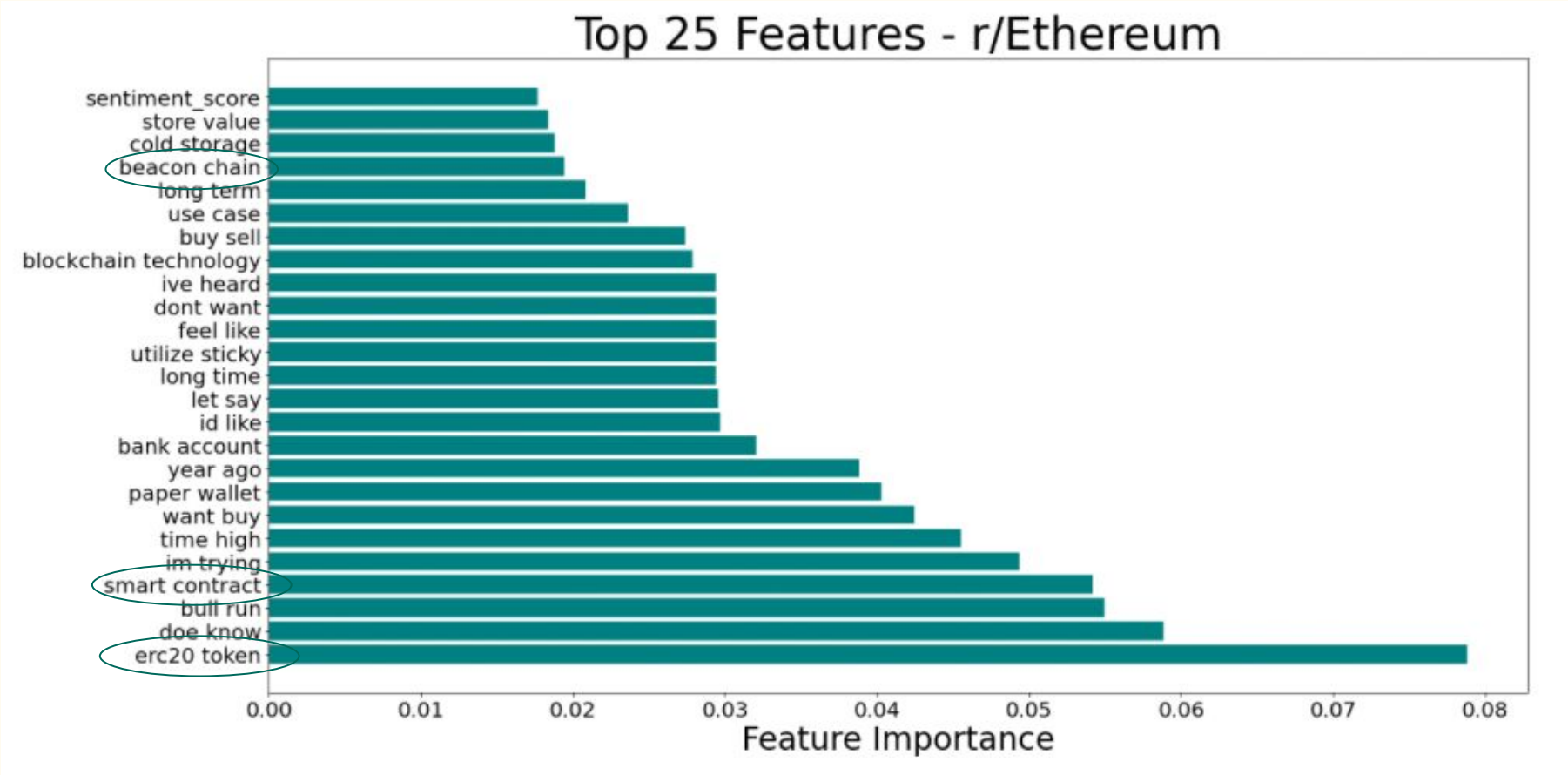
Notable terms: etherscan, testnet, geth, mew, eth2, dao, vitalik validator



Unlike, r/Ethereum most of these words aren't closely related to Bitcoin. Some of these words can be stopwords.



Notable terms: Token, contract, gas, ether, dapp, staking, defi, erc20, metamask


Top 25 Features - r/Ethereum
sentiment_score
store value
cold storage
beacon chain
long term
use case
buy sell
blockchain technology
ive heard
dont want
feel like
utilize sticky
long time
let say
id like
bank account
year ago
paper wallet
want buy
time high
im trying
smart contract
bull run
doe know
erc20 token
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Feature Importance

Limitations

- Sacrificed performance for interpretability of language
  - Tuned hyperparameters to train on less words
  - Extensive stopwords list
  - Multinomial Naive Bayes predicts much better
  - Higher ngrams made predictive power much worse

- Good start for r/Ethereum language, bad for r/Bitcoin
  - Model only found useful terms for predicting Ethereum and assumes other words are predictive of Bitcoin even though they're unrelated to Bitcoin (possibly stopwords)

Further Considerations

- Observe feature importances for other models
- RandomForest did not use all of my dataset's features
- Use more robust stopwords list
- Deeper dive into Sentiment Analysis of both communities

Questions?