Project 3: Classifying subreddits between CryptoCurrency and stocks

Khalis Kassim

Agenda

- Problem Statement
- CryptoCurrency vs stocks
- Data Cleaning
- Feature Analysis and Importance
- Naive Bayes vs Random Forest
- Evaluation & Recommendation
- Conclusion & Learnings

Problem Statement

Our client, Diamond Hands Pte Ltd, is looking to collate a list of posts on reddit for investment-related trends, which they can recommend the type of investment to a client or user based on their comments. Noting that the comments are mostly from retail investors discussing between stocks and cryptocurrency, they want to classify these comments accordingly.

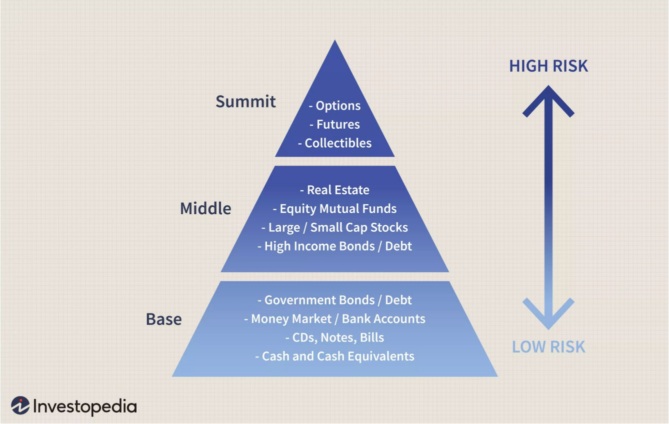
The company has commissioned you to build a classification model based off subreddits specifically from r/CryptoCurrency and r/stocks, to help train and build the model to classify these comments to the correct topic. This would allow the the company to suggest the correct type of investments a client should make based on their comments. Additionally, this would allow the staff have up-to-date info on changes within the market, from the general public, to better plan and forecast of any upcoming changes of their clients' portfolio.

- Scrape comments off subreddits r/CryptoCurrency and r/stocks
- Provide possible trend analysis based on scraped data
- To build a classification model that can classify the subreddits with minimal misclassifications

CryptoCurrency vs stocks

- Decentralised digital currency based off the blockchain technology
- Classified as extremely high-risk investments due to volatility
    - Speculative
    - Unregulated
    - Susceptible to errors and hacking

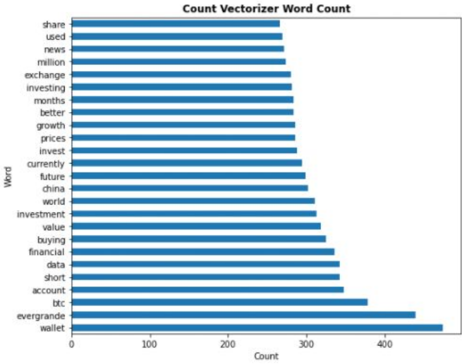
- Share, or equity, representing the fractional ownership of a company or corporation
- Different risk profiles for different types of investments

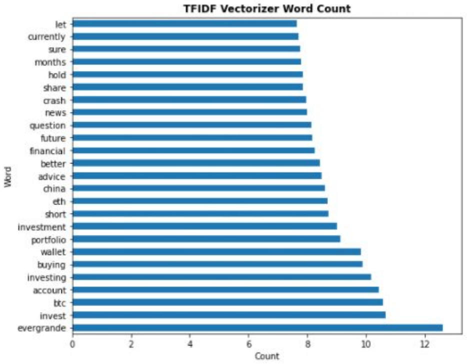

HIGH RISK
LOW RISK
Summit
- Options
- Futures
- Collectibles
Middle
- Real Estate
- Equity Mutual Funds
- Large / Small Cap Stocks
- High Income Bonds / Debt
Base
- Government Bonds / Debt
- Money Market / Bank Accounts
- CDs, Notes, Bills
- Cash and Cash Equivalents
Investopedia

Exploratory Data Analysis

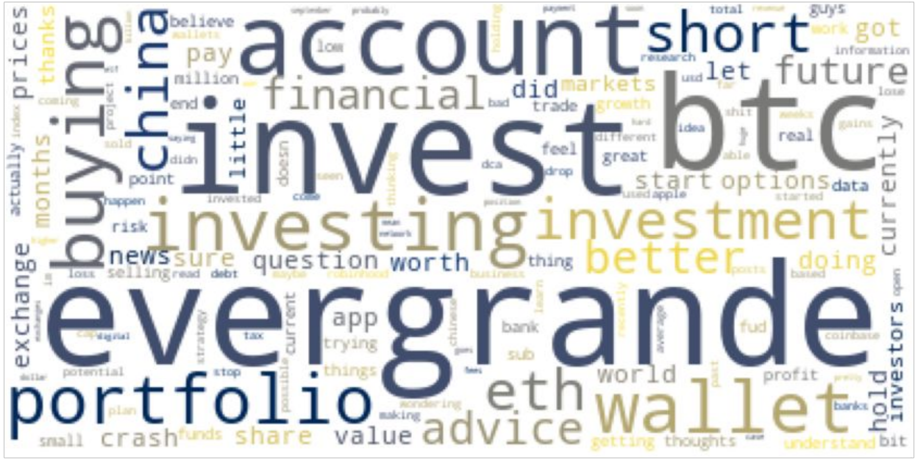
- Repeated posts by AutoModerator bots
- Lemmatizer, Stemmer or None?
  - Not complete grouping with Lemmatizer
  - Different variations of words with Stemmer
  - Complete, but no grouping of common words

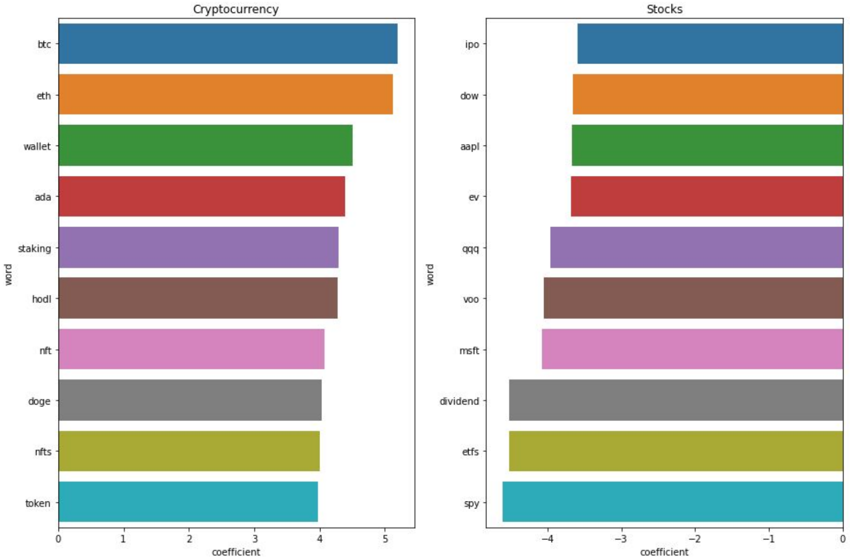
- CountVectorizer or TfidfVectorizer
  - Higher occurrence of words vs scoring words with lesser occurrence in many documents





Feature Analysis & Importance





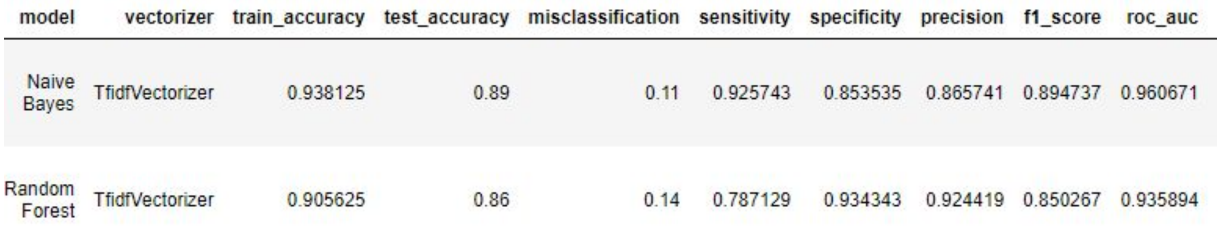
Naive Bayes vs Random Forest

Naive Bayes

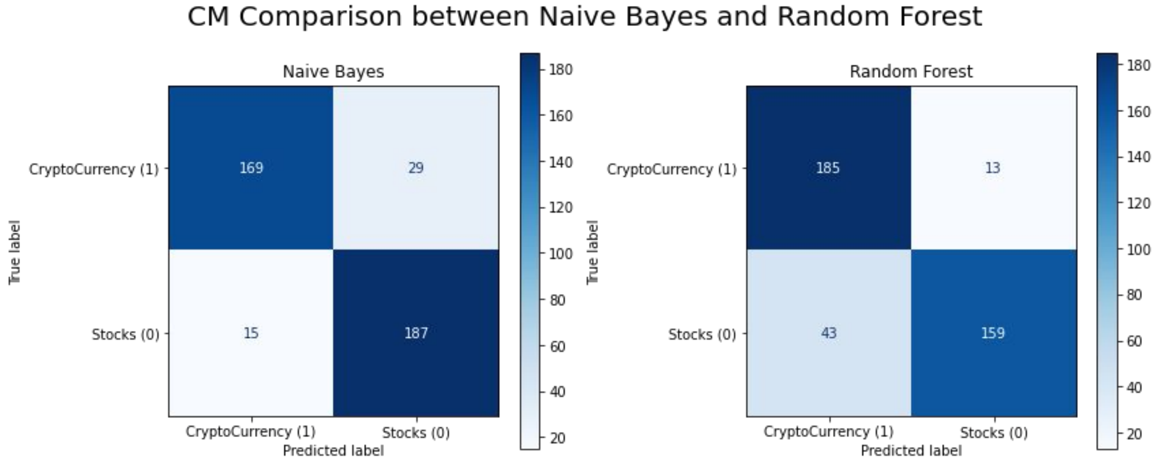
- Low Misclassification Rate (0.11)
- Slightly higher accuracy

Random Forest

- High specificity
- A lot of Type I errors

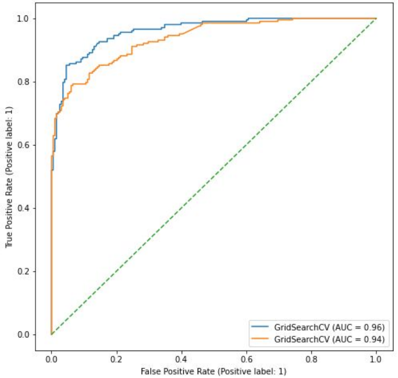


model	vectorizer	train_accuracy	test_accuracy	misclassification	sensitivity	specificity	precision	f1_score	roc_auc
Naive Bayes	TfidfVectorizer	0.938125	0.89	0.11	0.925743	0.853535	0.865741	0.894737	0.960671
Random Forest	TfidfVectorizer	0.905625	0.86	0.14	0.787129	0.934343	0.924419	0.850267	0.935894

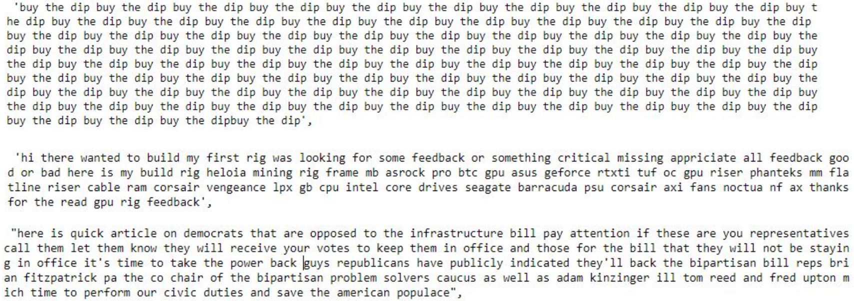


Evaluation & Recommendation

- Naive Bayes as the recommended model
  - Higher sensitivity and ROC AUC score
  - Smaller amount of Type I and II errors



- Misclassification Analysis
  - Posts with repeated text to hit word count
  - Posts that has content on both stocks and cryptocurrency
  - Words with different meanings; ‘mining’
  - Posts that don’t have key features, but are indirectly related to subreddit topic



```
 'buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy t
he dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip
buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the
dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy
the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip
buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the
dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy
the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip buy the dip
buy the dip buy the dip buy the dipbuy the dip',

 'hi there wanted to build my first rig was looking for some feedback or something critical missing appriciate all feedback goo
d or bad here is my build rig heloia mining rig frame mb asrock pro btc gpu asus geforce rtxti tuf oc gpu riser phanteks mm fla
tline riser cable ram corsair vengeance lpx gb cpu intel core drives seagate barracuda psu corsair axi fans noctua nf ax thanks
for the read gpu rig feedback',

 "here is quick article on democrats that are opposed to the infrastructure bill pay attention if these are you representatives
call them let them know they will receive your votes to keep them in office and those for the bill that they will not be stayin
g in office it's time to take the power back guys republicans have publicly indicated they'll back the bipartisan bill reps bri
an fitzpatrick pa the co chair of the bipartisan problem solvers caucus as well as adam kinzinger ill tom reed and fred upton m
ich time to perform our civic duties and save the american populace",
```

Conclusion & Learnings

- Random State
  - Shifts in accuracy after applying suggests lack of robustness
- Majority of top features are cryptocurrency related
  - Textual data not well scored for other subreddit; tickers, different methods of investment for stocks (bonds, etf, etc)
  - Data with better textual quality in relation to subreddit
  - More data, more documents, more noise -> Might affect vectorizer scores
- Scikit-learn recommends Naive Bayes for text data classification < 100,000 samples