Predicting Shitcoins

Using Machine Learning to Identify High-Risk Cryptocurrencies

By: Neha Awasthi, Sukeerthi Adi, Jared Rodriguez, Sanjeev Hirudayaraj, Ankita Sindhu

Introduction

Cryptocurrencies have been around for over a decade now and have taken the world by storm. The underlying technology, blockchain, has brought a paradigm shift in the way we transact and store value. The decentralized nature of cryptocurrencies and their resistance to censorship and government interference have made them popular among people seeking financial freedom and privacy. Despite their relatively short existence, cryptocurrencies have already gone through several booms and busts, with the market experiencing significant volatility. Some cryptocurrencies have gained massive popularity and have become household names, while others have faded away into obscurity. With the proliferation of new cryptocurrencies and the ever-changing dynamics of the market, the need for reliable tools and techniques to predict their success or failure has become increasingly important. This is where predictive models come in, offering a data-driven approach to identifying potential "shitcoins" and minimizing investment risks.

Shitcoins

Shitcoins, in the context of the cryptocurrency world, refer to any digital currency or token that has little to no value, is not backed by any tangible assets, and is often created as a joke or as a way to scam unsuspecting investors. They are essentially useless and do not serve any practical purpose, making them a risky investment.

In the past, there have been several instances of shitcoins rising in popularity, such as Dogecoin, a cryptocurrency created in 2013 as a joke based on the "Doge" meme. Despite its origins, Dogecoin gained a significant following and even gained the attention of celebrities such as Elon Musk, leading to a surge in its value.

However, the rise of shitcoins has also had a negative impact on the cryptocurrency market, leading to a loss of credibility for the industry as a whole. Many investors have been burned by investing in shitcoins, leading to a lack of trust in the cryptocurrency market.

As such, it is important for investors to be cautious and conduct thorough research before investing in any cryptocurrency, especially those that seem too good to be true or lack any real practical use. The rise of shitcoins has highlighted the need for more regulation and transparency in the cryptocurrency market to protect investors and ensure the long-term viability of the industry.

This is where a shitcoin prediction model can be useful. By analyzing various factors and metrics, such as market trends, historical data, and sentiment analysis, the model can predict whether a cryptocurrency is likely to be a poor investment or not. This can help individuals and businesses in the cryptocurrency industry make more informed investment decisions, potentially saving them from significant losses. The model can also benefit the industry as a whole by promoting more responsible investing and weeding out fraudulent or misleading cryptocurrencies.

Business Problem

The business problem that we are trying to solve is the issue of cryptocurrency investors losing their investments due to the volatility and unpredictability of the market. By building a predictive model that can identify potential "shitcoins," we can help investors make more informed decisions and avoid investing in cryptocurrencies that are likely to fail. This can be useful for the finance and investment industry, as well as individual investors looking to make profitable investments in the cryptocurrency market. With our model, investors can save time and effort in conducting extensive research on various cryptocurrencies, as the model can provide them with valuable insights into the potential success or failure of a cryptocurrency.

How can we accurately predict whether a cryptocurrency is a 'shitcoin' or not to assist investors in making informed decisions and mitigate financial risks?

Part 1- Coingecko Analysis

Analyzing various cryptocurrencies, their features, and market data to predict the possibility of a currency being a shitcoin.

About the Data

Data acquisition:

In the analysis of CoinGecko, we utilized web scraping techniques on the Chrome web-browser through the Selenium package. In the setup portion, we opened an empty web browser for the CoinGecko website. The first iteration was done on the homepage, the second with the meme coin page. Next, we created a set of empty lists, web elements using get_elements, and loops that would scrape all instances of the element on the page.

Then, we combined them all into one larger for loop that ran for the length of the pages. Outside of the loop, it would click on the “Show Fully Diluted Valuation” button. Inside of the loop, it would run all of the web elements, loops to scrape the elements, and then click on the next page with a sleep timer attached. The timer allowed for the page to fully refresh to avoid grabbing unloaded elements, which would result in an error.

After that, each list was put into it’s own data frame for initial cleaning. The “fdv” and “fdv_mkt” had the same web element value, so we had to split it into two separate data frames by every other index value. The data frame containing “price” had three empty rows at the top of every page. In order to delete these rows, I set empty strings to NaN using the NumPy package. After cleaning, I reset all indexes, so that they would properly merge. We then concatenated all of the data frames into one data frame (one for the base page and meme coin) and exported it to a CSV file.

Lastly, some additional cleaning was performed on the newly created CSV files. The datatypes were changed to strings to utilize regular expressions. There were four patterns in total. Two patterns replaced characters (“$” and “)”) into empty strings. One pattern replaced hyphens with a zero to properly represent its value. The other pattern replaced a left parenthesis with a hyphen to denote a negative number. The purpose of these conversions was to alter the columns to float64 datatypes. The meme coin dataset had additional undesired

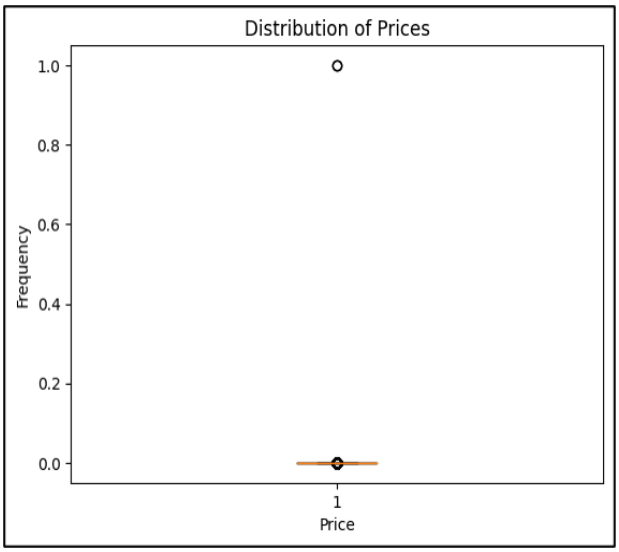
values and those were dropped. These cleaned data frames were once again exported to CSV files.

Variable	Definition
Rank	The rankings of each cryptocurrency according to CoinGecko
Ticker	The ticker for each cryptocurrency
Price	The trading price for each cryptocurrency
24h_Volume	The twenty-four-hour trading volume for each cryptocurrency
Mtk_Cap	The market cap for each cryptocurrency
FDV	The fully diluted valuation for each cryptocurrency (current price x max supply)
FDV_MKT	The market cap divided by the fully diluted valuation divided for each cryptocurrency
Meme_Coin	Whether the cryptocurrency is a meme coin or not

Exploratory Data Analysis

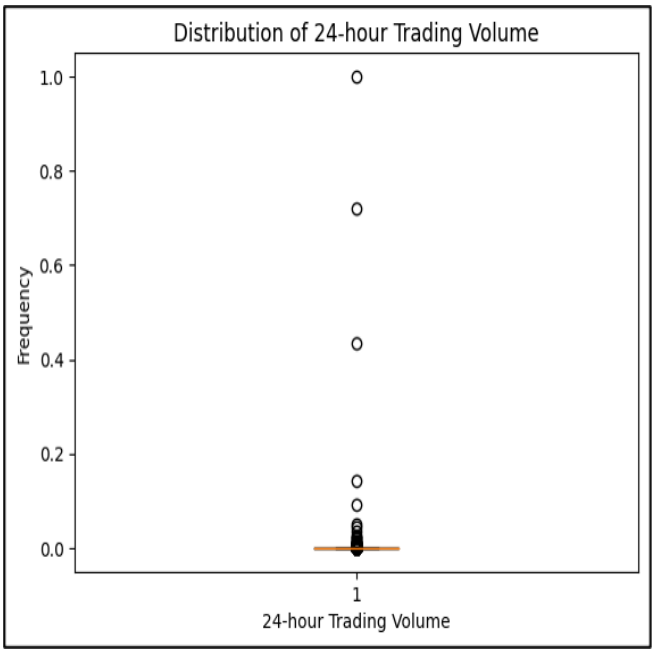
Once we had our final dataset, we performed some basic exploratory data analysis.
We started with the distribution of our variables and then explored the relationships amongst them.

Price:



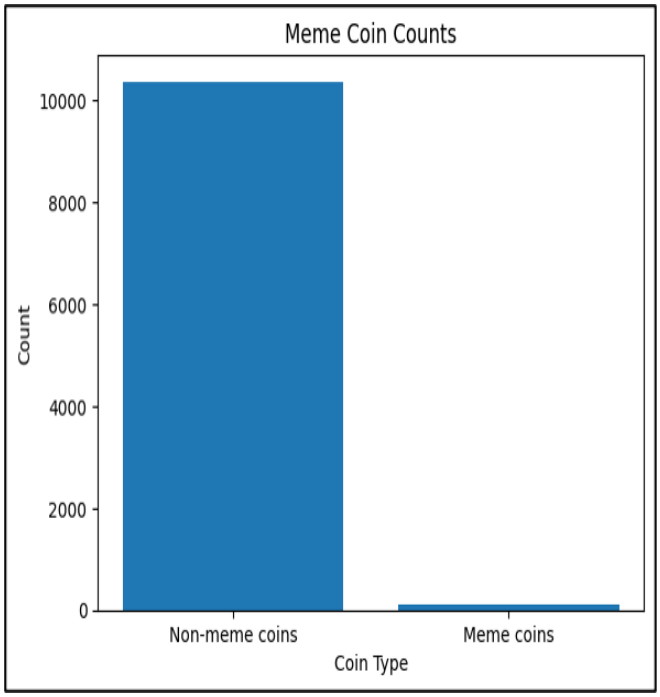
Price distribution suggests that most of our crypto coins are priced around 0 with only a few outliers having prices much higher than that.

24- hour Trading Volume:



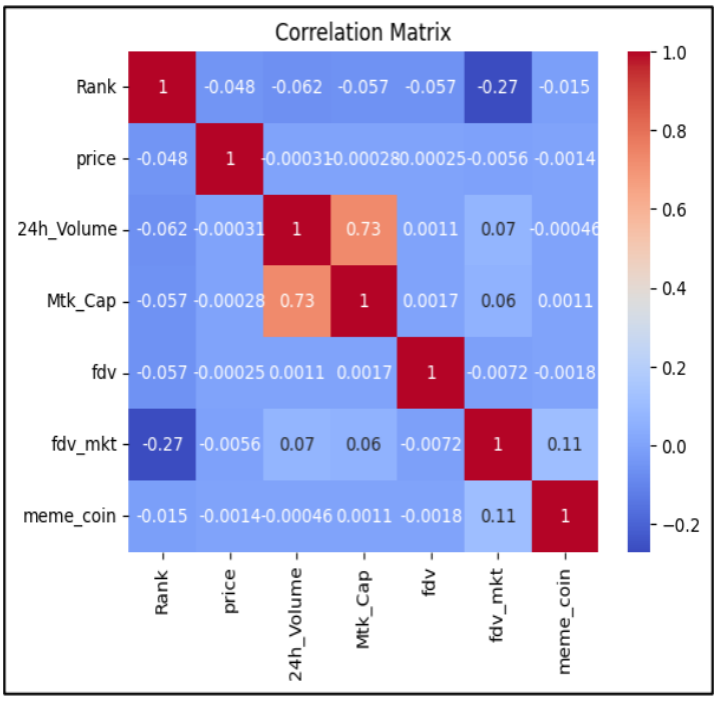
Looking at the distribution of 24-hour trading volume, we observe that again most of the coins are trading at or around 0 with some outliers.

Shitcoins:



We looked at the distribution of our predicted variable to see how many of the coins in our datasets are meme/shit coins and it seems that most of the coins are actually non-meme coins and only a small percentage of these coins are meme coins.

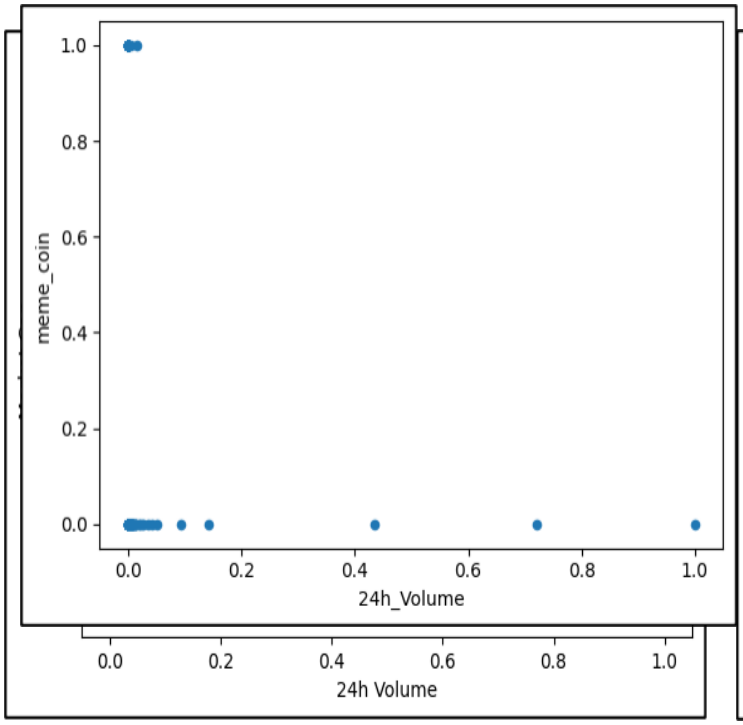
Correlation Analysis:

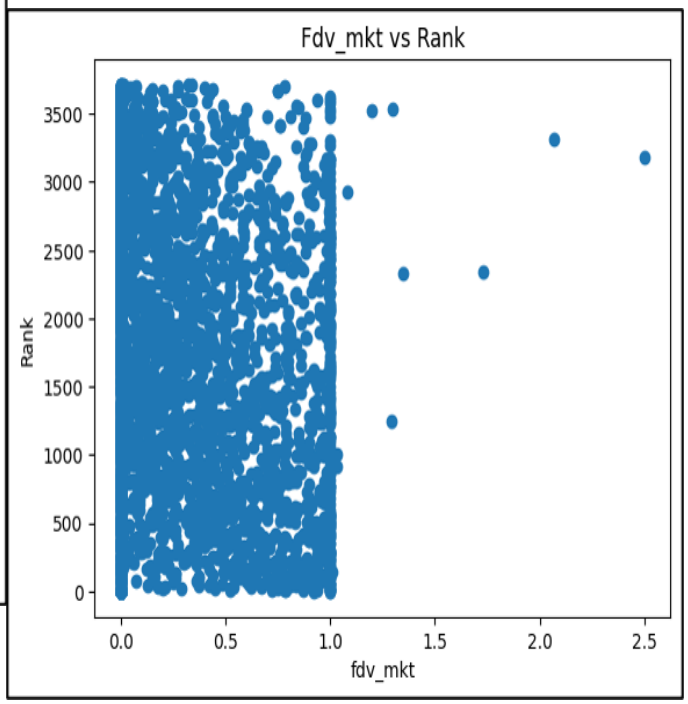


We now look at the correlation of all of our variables to see the strength of linear relationships among them. Market cap is highly correlated with 24-hour trading volume. Other than these two, there doesn't seem to be any strong correlation between any other variables. We do see that rank is negatively correlated with Fdv_Mkt.

Relationship between different variables:

Once, we had examined the distribution and correlation, we performed some bivariate analysis by exploring relationships between the variables.



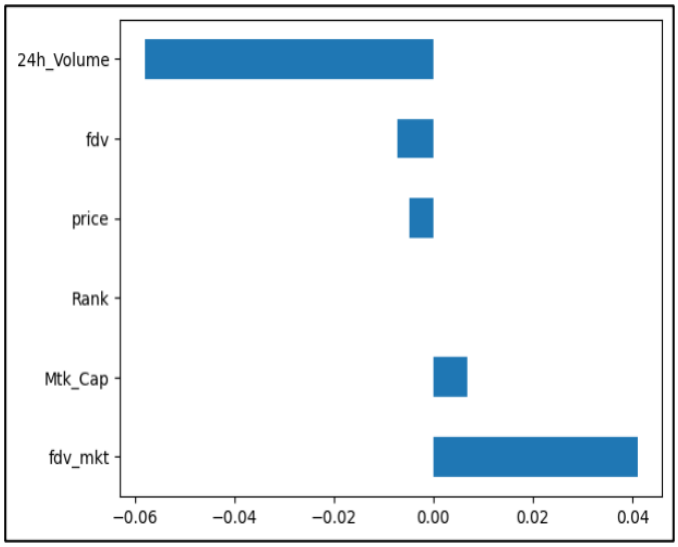


Our bivariate analysis, didn't reveal any strong linear relationships between different variables except for a steady linear relationship between Market capitalization and 24-hour trading volume. This is consistent with our correlation analysis.

Predictive Analysis

Now that we know how our variables are related, we move on to our main analysis where we use these variables that we have created to predict whether a coin is a shitcoin or not. After preparing the data and splitting it into a training and testing set, we tried multiple models on it to find out the best fit for our predictive analysis.

- **Linear Regression**

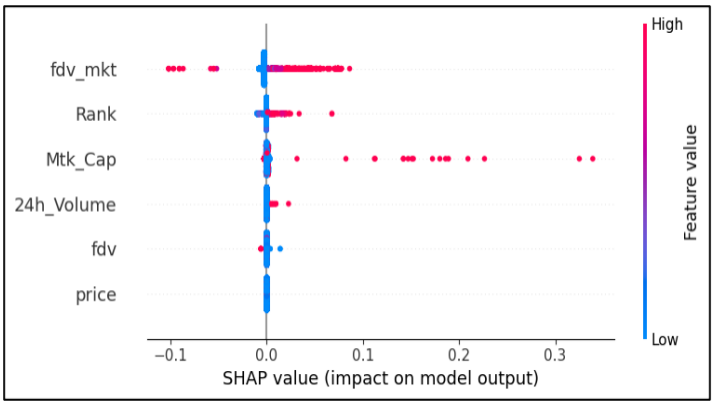


We started with a simple linear regression. The output score for this model was at 1.5 so obviously this is considerably worse than just random guess and we moved on to other models.

We checked the feature importance here and 24-hour trading volume seems to be the most important feature followed by fdv_mkt for this model.

- **Random Forest**

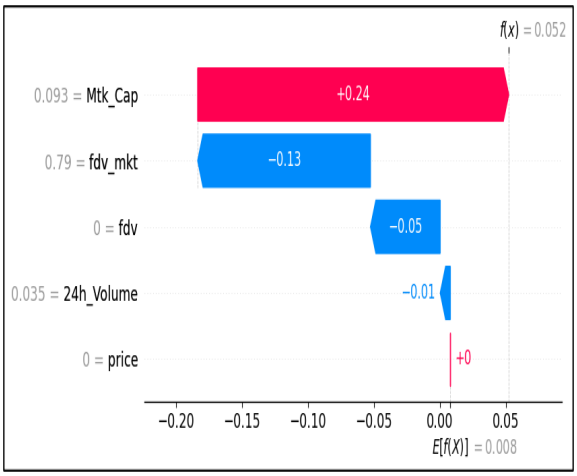
a) **Random Forest Regression**



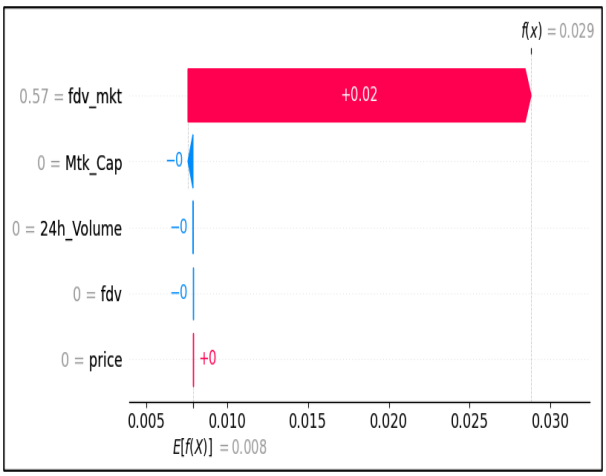
Next, we tried Random forest model. First we tried it as a regressor and while it slightly improved from linear regression, the output score was still at only .8.

Using SHAPLEY, we tried to determine the feature importance and fdv_mkt seems to have the most impact on the output followed by rank and Market cap.

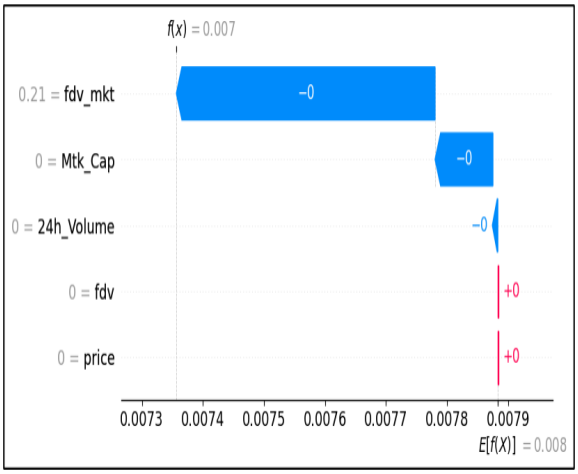
Sample 3



Sample 1229



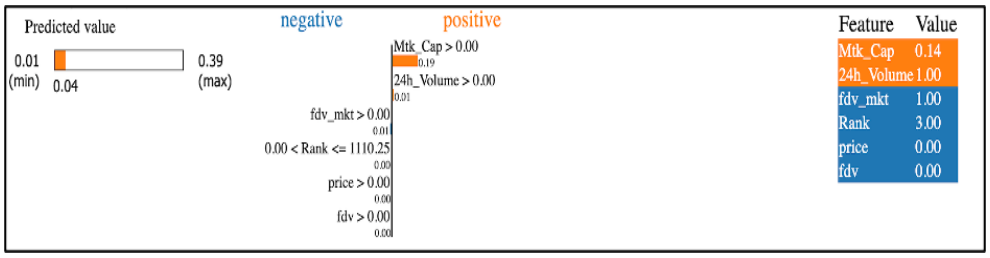
Sample 1449



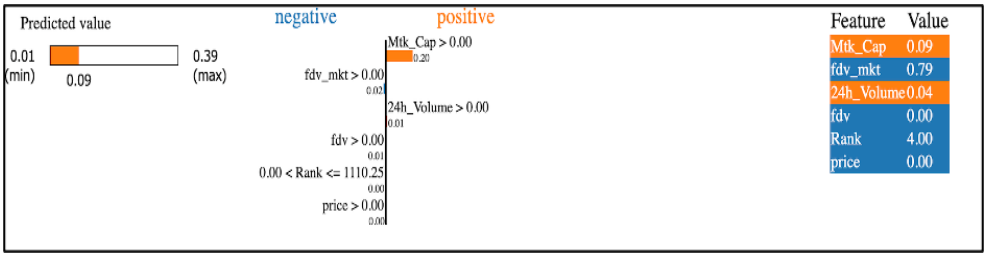
Now, we see a few samples where we used SHAPLEY explainer to explain feature importance and the score.

We also tried LIME to the same end and following are a few instances of using LIME explainer to examine feature importance and the score of that instance

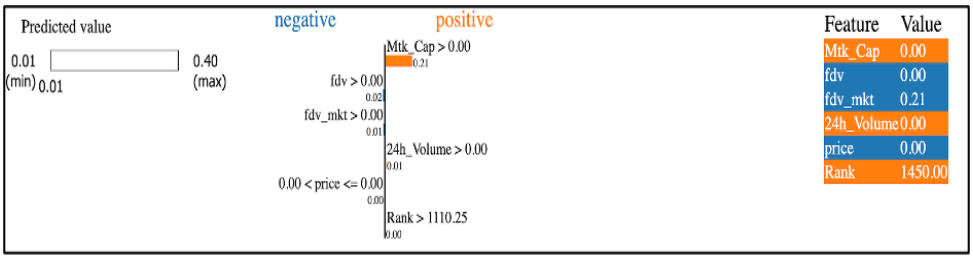
i = 3



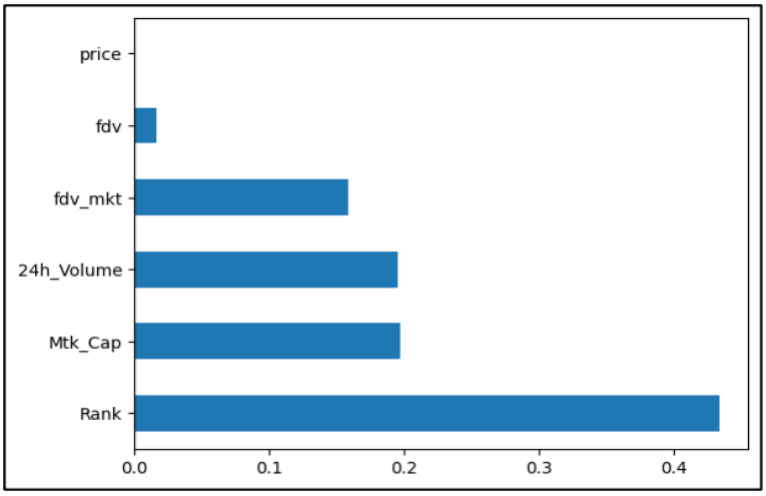
i = 2



i = 1449



b) Random Forest Classification

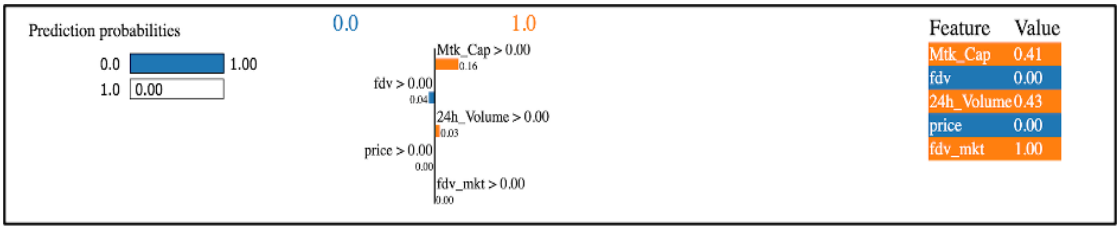


We also tried Random forest model as a classifier. This actually decreased the output score to -0.6.

The most important features in this model was Rank followed by Market cap amd 24-hour volume.

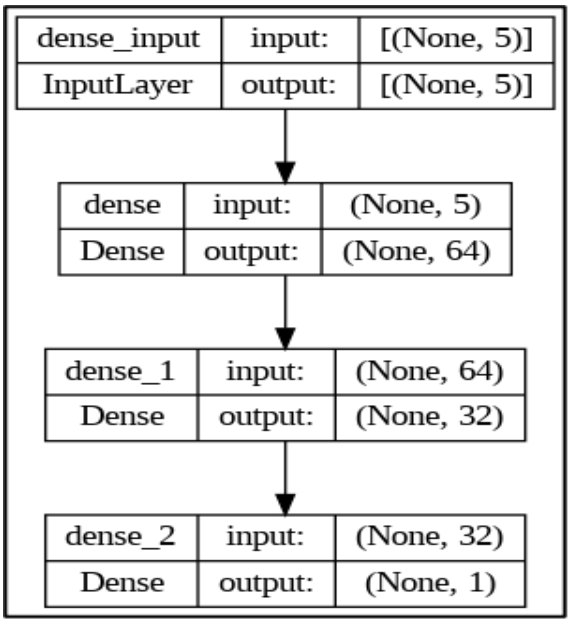
We tried LIME in this model too to explain an instance

i = 1

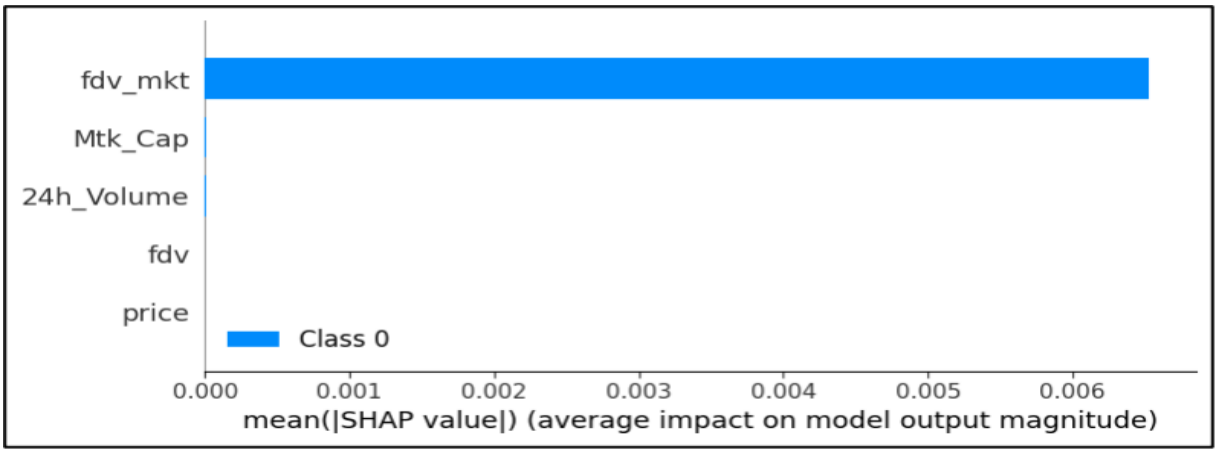


- **Other Models:** After Linear regression and Random Forest models, we tried SVM, Histogram gradient boosting regressor and Histogram gradient boosting classifier with the scores of -86.9 , 25.3, 98.05 respectively. So far, our best model was Histogram gradient boosting classifier.

- **Neural Network :**
  The final model we decided to try was a neural network model using using Keras.



We used a simple neural network with two hidden layers and a final output layer with a sigmoid activation function.
This is our best predictive model which can predict whether a crypto coin is a shitcoin or not with an accuracy of 99.05%
We used SHAPLEY to figure out feature importance of the model and this model had fdv_mkt as the most important feature which is actually consistent with our other models that did well.



Insights from Predictive Analysis:

This analysis provides some noteworthy insights and provides us with a viable model which can be reliably used to predict if the coin is a “shitcoin” or not based on the data we can get from CoinGecko website. We see that the most important factor in our prediction is actually Fully diluted valuation to Market Capitalization ratio which provides us a reasonably accurate model to work with. We can also infer that the relationships between our predicted variable and the data we are using to predict it are probably not linear in nature by the performance of our previous model. The final model that we have decided to use can be used to predict the occurrence of a “shitcoin”.

The correlation analysis revealed that market capitalization and 24-hour trading volume have a strong linear relationship, indicating that highly valued coins are likely to be highly traded. Other variables in the dataset did not show strong correlations, suggesting that they may have less impact on coin valuation. The predictive analysis conducted on the dataset used various models to determine the best way to predict whether a coin is a "shitcoin" or not. The final model, a

neural network, achieved an impressive accuracy rate of 99.05%, indicating that this model can be a valuable tool for investors seeking to evaluate the potential of different coins.

Overall, this analysis provides a comprehensive picture of the crypto market, and can be used by investors to make more informed investment decisions. The insights on the distribution of prices and trading volumes can be used to identify undervalued coins with high growth potential, while the prevalence of meme coins can be used to avoid risky investments. The correlation analysis highlights the importance of market capitalization and trading volume, while the predictive analysis provides a valuable tool for assessing the potential of different coins. By providing a clear picture of the crypto market, this analysis can help investors to make better decisions and achieve greater success in their investments.

Part 2- Reddit Analysis

Analyzing the general sentiment around certain cryptocurrencies on subreddits to predict the possible sentiment around coins

About the Data

Variable	Definition
comment	The textual content of the comment made on a Reddit post.
id	A unique identifier is assigned to each comment.
sentiment_score	A numerical value representing the sentiment of the comment on a scale from -1 to 1,

	where -1 indicates a negative sentiment and 1 indicates a positive sentiment.
word_count	The total number of words present in the comment.
char_count	The total number of characters present in the comment.
avg_word_len	The average length of the words in the comment is calculated by dividing the total number of characters by the total number of words.
sentiment	A categorical variable indicating the sentiment of the comment as either "positive", "negative" or "neutral".

Analyzing shitcoin-related data from Reddit can be a valuable tool for predicting the success or failure of a cryptocurrency. The cryptocurrency market is highly sentiment-driven, and Reddit provides a platform for crypto enthusiasts to express their opinions and share news and updates on various coins. The sentiment expressed on Reddit can often be an early indicator of market trends and can help investors make informed decisions.

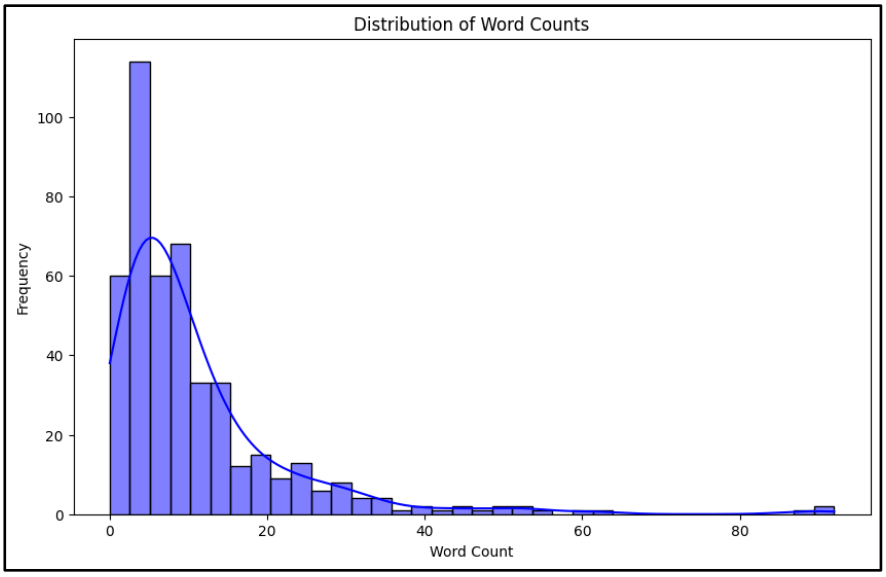
In this analysis, we obtained data from Reddit by scraping several cryptocurrency-related subreddits. The data includes variables such as sentiment score, word count, character count, and average word length, which provide valuable insights into the opinions and attitudes expressed by Reddit users.

The value of this data lies in its ability to provide a glimpse into the collective sentiment of a community of crypto enthusiasts. By analyzing this sentiment, we can gain a better understanding of the potential success or failure of a particular cryptocurrency. This information can be invaluable to investors, who can use it to make informed decisions and maximize their returns.

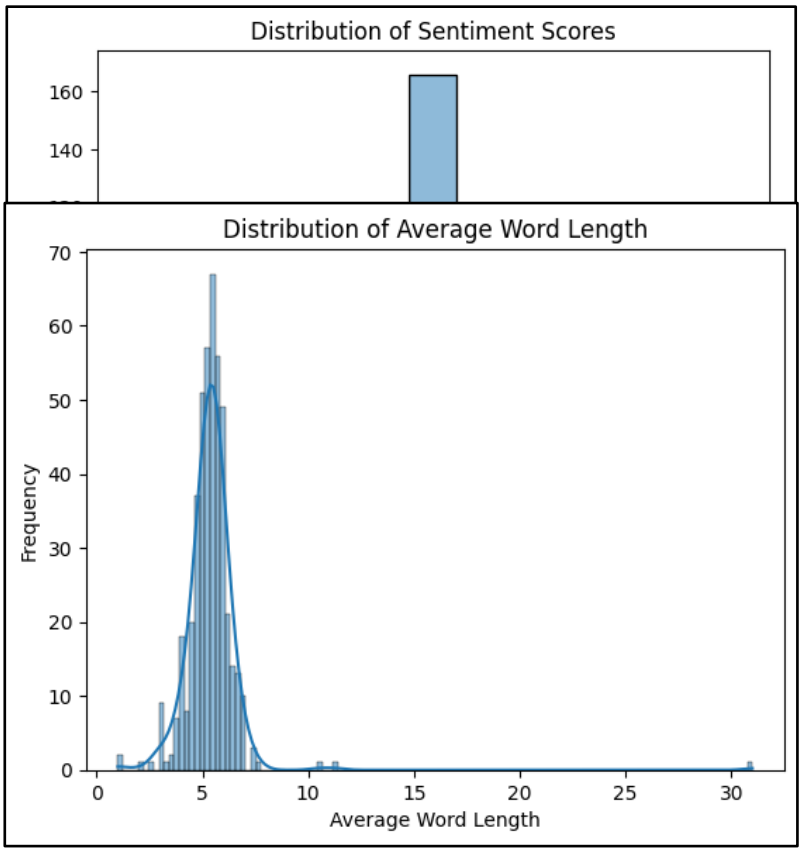
Exploratory Data Analysis

The distribution of word count in the analyzed Reddit data reveals that a large number of posts related to shitcoins have a relatively low word count, with the highest bar being observed

between 0-5 words. This suggests that many users are expressing their opinions or providing quick updates on the latest developments in the crypto market in a concise manner. The next most common range of word counts is 5-10 words, indicating that while users are providing slightly more detail, they are still keeping their posts relatively short. This pattern may be a reflection of the fast-paced and rapidly changing nature of the crypto market, where users need to be able to quickly disseminate and absorb information. On the other hand, there are few posts with word counts over 80, which suggests that in-depth analysis or longer-form content may not be as common in the context of shitcoin discussions on Reddit. The relatively small number of posts in the 20-40 word count range may indicate that users are either posting short comments or longer, more detailed analyses, but not many posts that fall in between. Understanding these patterns in the distribution of word counts can provide valuable insights into the nature of discussions surrounding shitcoins on Reddit.

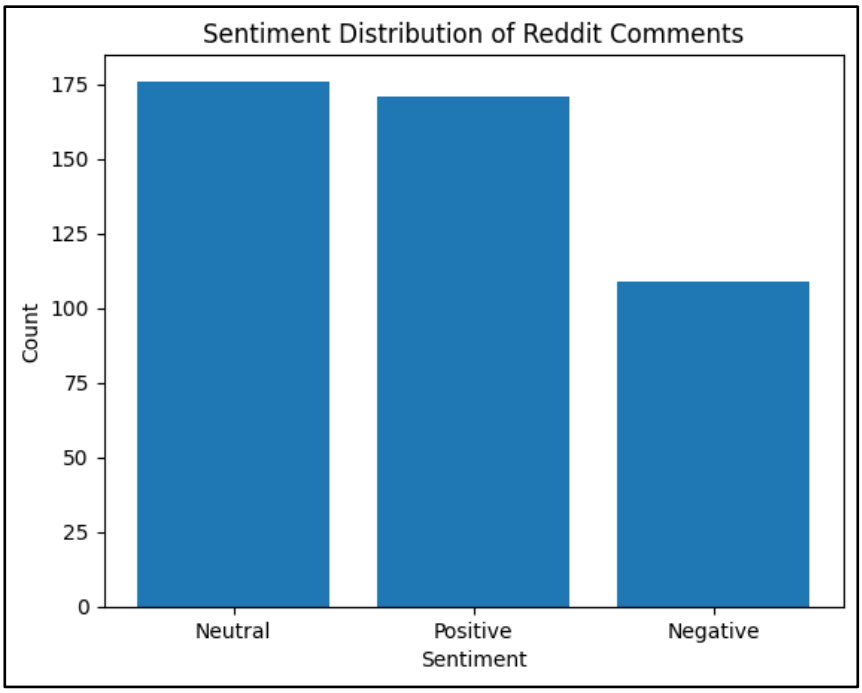


In the context of shitcoin Reddit analysis, the distribution of sentiment scores with the highest bar around 0 indicates that the sentiments expressed about shitcoins on Reddit are largely neutral. The fact that the -0.5 space is higher than the 0.5 space suggests that there may be slightly more negative sentiment expressed about shitcoins compared to positive sentiment. This insight can be useful in understanding the overall sentiment of the community towards

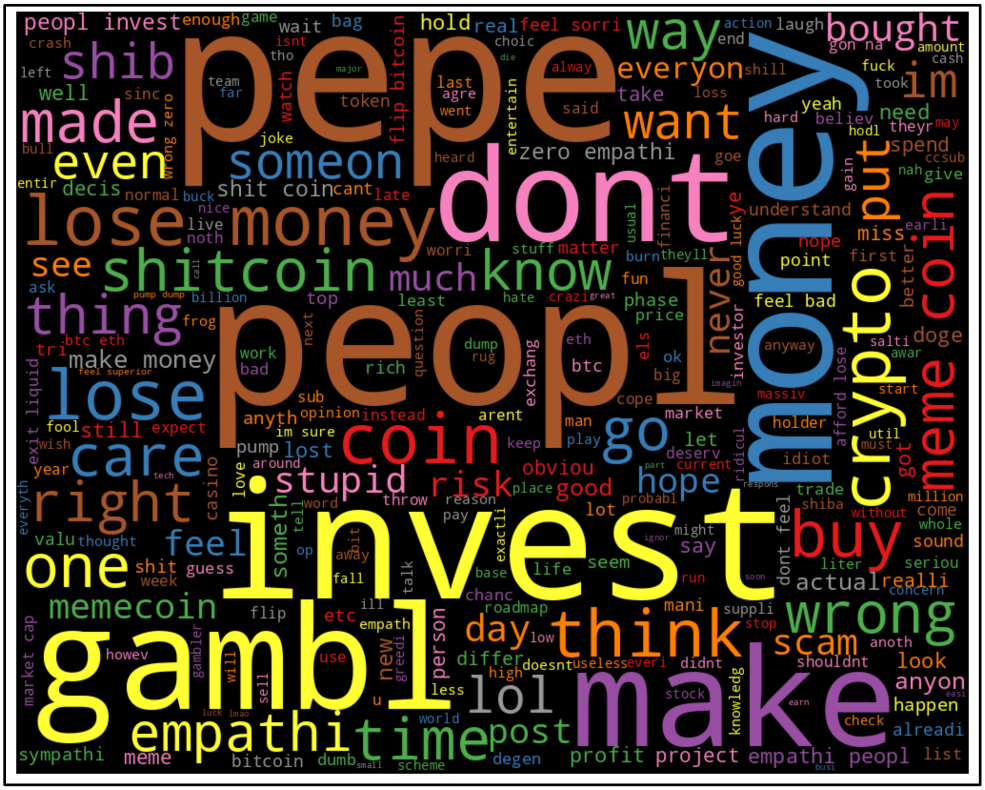


shitcoins and can inform our analysis and predictions about whether a particular cryptocurrency will be a shitcoin or not.

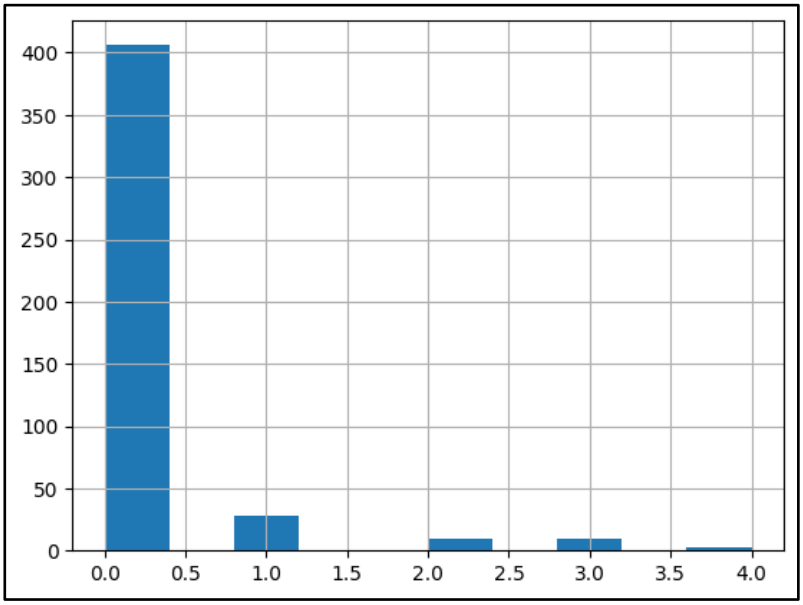
The distribution of the average word length in the Reddit data for shitcoin analysis shows that most of the posts have an average word length between 2-7. This could indicate that the community discussing shitcoins on Reddit is trying to communicate their thoughts and ideas using simple language, which is easy to understand for everyone. The fact that very few posts cross the 30-word length mark further supports this hypothesis. It's possible that the simplicity in the language used by the Reddit community discussing shitcoins may be because of the nature of the subject matter. As cryptocurrencies and the associated technology can be complex and technical, the community may feel it's necessary to keep things simple and accessible to everyone. However, further analysis is needed to confirm this hypothesis.



In the context of the shitcoin Reddit analysis, the neutral sentiment being the most frequent is not surprising as the sentiment towards cryptocurrencies can often be mixed or uncertain. However, the fact that positive and negative sentiments are almost equal in frequency suggests a significant polarization in the community's opinions. It could be indicative of a volatile market, where the sentiment towards a particular shitcoin can swing quickly, leading to significant price fluctuations. This information can be valuable for investors looking to make informed decisions on whether to invest in a particular shitcoin or not.



Topic Models



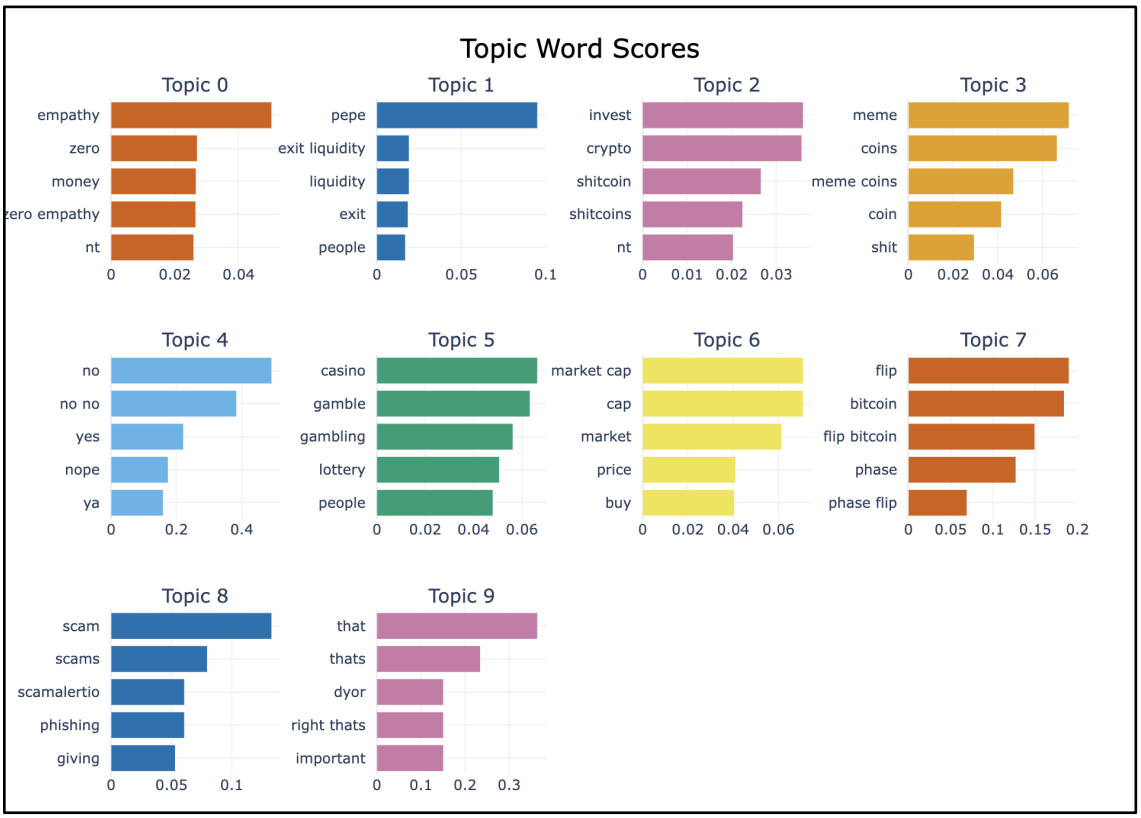
Based on the topic modeling analysis of the Reddit data related to shitcoins, it can be seen that the most commonly discussed topic is related to supply, exit, liquidity, and meme. This could indicate that people are interested in discussing the availability and accessibility of these coins in the market. The second most discussed topic is related to people's opinions on the cryptocurrency industry, which can be seen through keywords such as "dumb," "greedy," "stupid," and "late." This suggests that there may be a lot of negative sentiment surrounding the industry and its players.

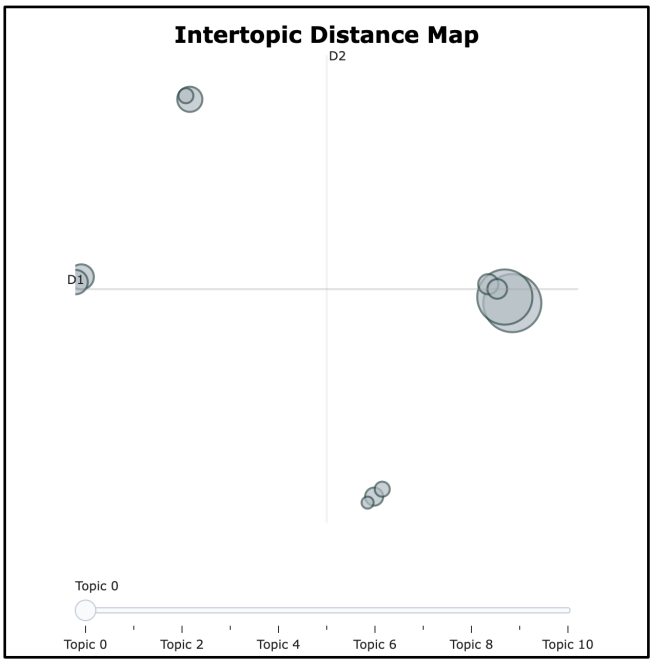
The third and fourth topics are related to scams and alerts about them, which highlights a concern for the potential fraudulent activities surrounding the cryptocurrency market. This can be seen through the use of words like "scam," "alert," and "end." The fifth topic is related to trading, holding, and making profits, but with negative words like "retarded" and "brained stung." This indicates that people are having mixed experiences when it comes to investing in shitcoins.

Detailed Topic Analysis

Topic Number	Description
0	Empathy, money, and people in the crypto world
1	Discussion of investing in and exiting liquidity from various cryptocurrencies
2	Discussion of dumping, pumping, and how to time the market
3	Discussion of various cryptocurrencies, including Shib, Doge, ETC, and Shiba
4	Ambivalent/uncertain sentiments expressed through negative responses like "no" and "nope"
5	Discussion of the Pepe meme and its role in the crypto world

6	Discussion of market capitalization and supply in the cryptocurrency market
7	Discussion of flipping Bitcoin and different phases of the market
8	Discussion of various cryptocurrencies and gambling





The topic modeling analysis has revealed interesting insights into the underlying themes present in the Reddit discussions related to shitcoins. The topics have been grouped based on their similarities, and it is observed that topics 1, 6, 5, and 0 are clustered together, suggesting a common theme related to people's emotions, investments, and empathy toward the market. On the other hand, topics 8, 7, and 10 seem to share similar discussions on Bitcoin, phases of the market, and the buying and selling behavior of people. The analysis also highlights that topics 2 and 9 share a common theme related to cryptocurrencies, coins, and the concept of gambling. Lastly, topics 3 and 4 appear to be related to the supply and demand of coins, market capitalization, and related concerns such as scams and alerts. Overall, these insights provide valuable information for investors and market analysts who wish to understand the key themes and discussions happening on Reddit related to shitcoins.

Predictive Models

Logistic Regression

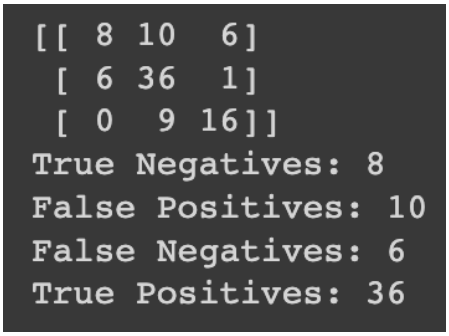
The logistic regression model has an accuracy of 0.65, which means that it correctly predicts the sentiment of 65% of the posts in the test set. The precision of 0.64 indicates that when the model predicts a particular sentiment, it is correct 64% of the time. The recall of 0.65 suggests that the model correctly identifies 65% of the posts that actually belong to a particular sentiment. The F1 score of 0.63 is the harmonic mean of the precision and recall, and it provides an overall measure of the model's accuracy.

```
Accuracy: 0.65
Precision: 0.64
Recall: 0.65
F1 Score: 0.63
```

Overall, the logistic regression model's performance is moderate in predicting the sentiment of posts related to shitcoins on Reddit. While the accuracy is not as high as the previous models, it still suggests that the model has some level of predictive power.

However, it may not be as reliable as other models such as random forest and gradient boosting, which had higher accuracy and F1 scores. Therefore, it may be worth exploring other models or optimizing the parameters of the logistic regression model to improve its performance.

With respect to the shitcoin reddit analysis, the confusion matrix shows that out of the total 92 instances in the test set:



- 8 instances that were actually negative were classified as negative (True Negatives)
- 10 instances that were actually negative were classified as positive (False Positives)
- 6 instances that were actually neutral were classified as negative (False Negatives)
- 36 instances that were actually neutral were classified as neutral (True Positives)
- 9 instances that were actually positive were classified as negative (False Negatives)
- 16 instances that were actually positive were classified as positive (True Positives)

To conclude, the logistic regression model correctly classified 44 out of 68 neutral and positive instances, which is a relatively low accuracy of 65%. The precision of the model (the proportion of true positives among all positive predictions) is 0.64, the recall (the proportion of true positives among all actual positives) is 0.65, and the F1 score (a harmonic mean of precision and recall) is 0.63. This means that the model has some difficulty distinguishing between neutral and positive sentiments, and misclassifies some of these instances as negative.

Histogram-based gradient boosting classifier

To begin the predictive analytics process, we used histogram-based gradient boosting classifier, which is a machine learning algorithm that is particularly useful for handling complex data with many variables. It works by building a series of decision trees, where each subsequent tree is built to correct the mistakes of the previous tree. This process continues until a certain stopping criterion is met.

In the context of the sentiment analysis of the Shitcoin subreddit, this model is useful because it is able to take into account multiple features of each post (such as sentiment score, word count,

character count, and average word length) and use them to predict the sentiment of the post (whether it is negative, neutral, or positive).

This output is the classification report for a machine learning model that has been trained and tested on a dataset of sentiment scores, word counts, character counts, and average word lengths. The model used is a HistGradientBoostingClassifier, which is an ensemble learning method that combines multiple decision trees to improve accuracy.

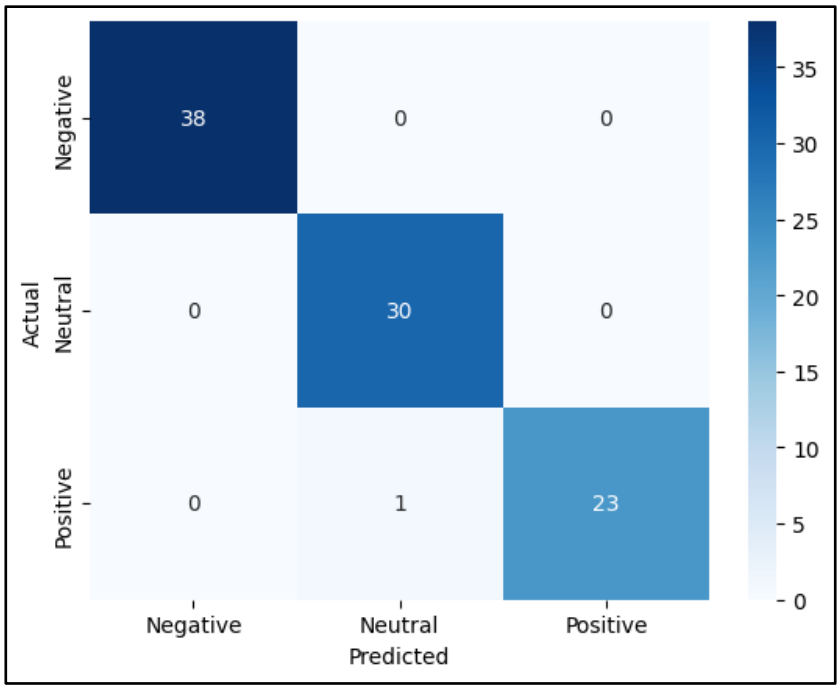
	precision	recall	f1-score	support
negative	1.00	1.00	1.00	38
neutral	0.97	1.00	0.98	30
positive	1.00	0.96	0.98	24
accuracy			0.99	92
macro avg	0.99	0.99	0.99	92
weighted avg	0.99	0.99	0.99	92

The output of the classification report provides several metrics for evaluating the performance of the model:

Precision: measures the proportion of true positive predictions (i.e., the number of correctly predicted samples) to the total number of positive predictions made by the model. A precision score of 1.0 means that all positive predictions made by the model were correct. In this case, we see that the precision scores for all three classes (negative, neutral, and positive) are very high, ranging from 0.97 to 1.0.

Recall: measures the proportion of true positive predictions to the total number of actual positive samples. A recall score of 1.0 means that the model correctly identified all positive samples. In this case, we see that the recall scores for all three classes are also very high, ranging from 0.96 to 1.0.

F1-score: combines both precision and recall into a single score. It is a weighted average of precision and recall, with a value of 1.0 being the best possible score. In this case, we see that the F1-scores for all three classes are also high, ranging from 0.98 to 1.0.



Support: indicates the number of samples in each class.

Accuracy: measures the proportion of correctly classified samples to the total number of samples. In this case, we see that the overall accuracy of the model is very high, at 0.99, indicating that the model performs very well in predicting the sentiment of the Reddit comments.

The macro average and weighted average metrics provide an overall assessment of the model's performance across all classes. In this case, we see that both the macro and weighted averages are very high, at 0.99, indicating that the model performs well across all sentiment classes.

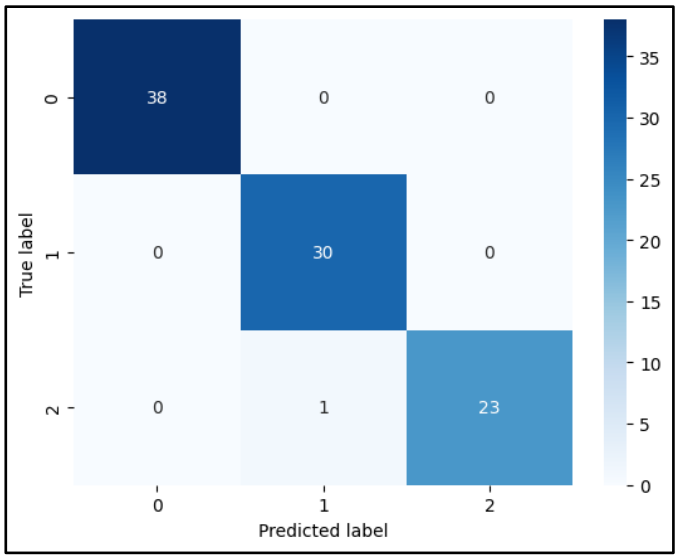
Random Forest Classifier

	precision	recall	f1-score	support
negative	1.00	1.00	1.00	38
neutral	0.97	1.00	0.98	30
positive	1.00	0.96	0.98	24
accuracy			0.99	92
macro avg	0.99	0.99	0.99	92
weighted avg	0.99	0.99	0.99	92

The output of the classification report provides us with several metrics that help us to evaluate the performance of the Random Forest Classifier. Here is an explanation of the metrics:

Precision: Precision measures how many of the predicted positive instances are actually positive. In this case, all the precision scores are 1.00, indicating that all the instances that the model predicted as positive were actually positive.

Recall: Recall measures how many of the actual positive instances the model was able to correctly identify. For example, in the negative class, the recall is 1.00, indicating that the model was able to identify all the negative instances correctly.



F1-score: The F1-score is the harmonic mean of precision and recall. It provides a more balanced view of the model's performance compared to just looking at precision or recall. In this case, the F1-scores are all high, indicating good performance across all classes.

Support: The number of samples in each class.

Accuracy: Accuracy is the proportion of correctly classified samples out of the total number of samples. In this case, the accuracy is 0.99, indicating that the model classified 99% of the samples correctly.

Macro average: The macro average calculates the mean of the scores for each class without considering the class imbalance. In this case, the macro average is 0.99, indicating good performance across all classes.

Weighted average: The weighted average calculates the mean of the scores for each class, weighted by the number of samples in each class. In this case, the weighted average is 0.99, indicating good overall performance of the model.

The output of the Random Forest Classifier is very similar to the output of the HistGradientBoostingClassifier, with both models achieving high precision, recall, and F1-scores across all classes, as well as high accuracy and macro and weighted averages. This suggests that both models are performing well on the sentiment analysis task, with little difference between them in terms of performance

Overall Insights

Based on the analysis of the Reddit Shitcoin dataset, we can see that there are some interesting insights that can be drawn. The LDA topic model analysis revealed that the most

dominant topics were related to cryptocurrency market trends, investment advice, and blockchain technology. This suggests that the discussions on the Shitcoin subreddit are largely centered around investment-related topics.

In terms of the predictive models, the Random Forest and Gradient Boosting models achieved high accuracy scores of 0.99, while the logistic regression model performed relatively worse with an accuracy of 0.65. This indicates that the Random Forest and Gradient Boosting models are more effective in predicting the sentiment of posts on the Shitcoin subreddit. The confusion matrices for the models revealed that they were particularly good at predicting negative sentiments.

Overall, the insights gained from this analysis can be used to better understand the discussions on the Shitcoin subreddit and to make more informed decisions related to cryptocurrency investments. For instance, sentiment analysis can be used by traders to identify potential trends in the cryptocurrency market and make more informed investment decisions. Additionally, the topic modeling analysis can be used by cryptocurrency companies to gain insights into the types of discussions that are taking place within their industry and to improve their products and services accordingly.

Comparative Analysis

The comparative study between the predictive analysis of the crypto market and the sentiment analysis of the Reddit Shitcoin dataset provides some valuable insights into the interplay between data-driven insights and sentiment-driven decision making. While the predictive analysis of the crypto market focuses on identifying factors that can predict the occurrence of a "shitcoin," the sentiment analysis of the Reddit Shitcoin dataset focuses on understanding the sentiment and topics of discussion among investors in the crypto market.

The insights gained from the predictive analysis can help investors make more informed investment decisions by identifying undervalued coins with high growth potential and avoiding risky investments. On the other hand, the insights gained from the sentiment analysis can help investors understand the sentiment of the market and adjust their investment strategies accordingly.

Given that the crypto market is heavily sentiment-driven, the insights gained from the sentiment analysis can be particularly valuable in guiding investor decision making. For instance, if the sentiment analysis indicates that there is widespread negative sentiment towards a particular coin, investors may choose to avoid investing in that coin. Alternatively, if the sentiment analysis indicates that there is positive sentiment towards a particular coin, investors may choose to invest in that coin.

To conclude, the comparative study highlights the importance of both data-driven insights and sentiment-driven decision making in the crypto market. By combining these two approaches, investors can gain a more comprehensive understanding of the market and make more informed investment decisions.

Limitations

Data Limitations: The analysis relies heavily on the data that is available from CoinGecko and Reddit. There may be some gaps in the data or inaccuracies that could affect the accuracy of the analysis.

Sample Size: The dataset used in the analysis may not be representative of the entire crypto market or the sentiment of all Reddit users. The analysis is limited by the number of posts and comments that were included in the dataset.

Timeframe: The analysis is based on data up until a certain point in time, and the crypto market is known for its volatility. Therefore, the findings may not be applicable to future time periods or may need to be updated regularly.

Generalizability: The findings of the analysis may not be applicable to other cryptocurrencies or investment strategies. The study focuses on the prediction of "shitcoins" and sentiment analysis of the Shitcoin subreddit, and the insights may not be directly transferable to other areas of the crypto market.

Lack of Context: The analysis may not take into account the broader economic or political context that can affect the crypto market. For instance, regulatory changes or macroeconomic events may have a significant impact on the performance of cryptocurrencies.

Conclusion

In conclusion, this project provides valuable insights into the crypto market through the analysis of both quantitative and qualitative data. The analysis of the CoinGecko dataset provides us with a model to predict the occurrence of "shitcoins" based on factors such as fully diluted valuation to market capitalization ratio, and the correlation analysis highlights the importance of market capitalization and trading volume. Meanwhile, the Reddit sentiment analysis provides insights into the discussions taking place on the Shitcoin subreddit, with topic modeling revealing the most dominant topics and predictive models achieving high accuracy scores. However, the project has several limitations, including the limited scope of data used and the potential for bias in the sentiment analysis. Despite these limitations, this project demonstrates the potential for data analysis to provide investors with the tools they need to make more informed decisions in the crypto market. By taking into account both quantitative and qualitative data, investors can gain a more complete picture of the market and achieve greater success in their investments.

Future Work

There are several avenues for future work in this area of research. Firstly, it would be interesting to expand the sentiment analysis to include other social media platforms such as Twitter and Facebook, which are also popular sources of information for cryptocurrency investors. This would provide a more comprehensive picture of the sentiment surrounding different cryptocurrencies and could help to identify potential trends in the market.

Another area of future work could be to explore the use of other machine learning algorithms, such as deep learning models, to improve the accuracy of the predictive analysis. Additionally, incorporating more data sources such as news articles or economic indicators could provide a more robust model for predicting the value of different cryptocurrencies.

Finally, it would be valuable to conduct a longitudinal study to investigate how the sentiments and opinions of cryptocurrency investors change over time, and how this affects the performance of different cryptocurrencies in the market. Such a study could provide insights into

the long-term trends of the cryptocurrency market and could help investors to make more informed decisions about their investments.