BITCOIN PREDICTION

Raghavaraju Nithisha, Reetikaa Reddy Munnangi, Duggempudi Vijaya Deepika Reddy.

INTRODUCTION

In the changing world of money, cryptocurrencies, especially Bitcoin, are becoming really important. As more people rely on digital money, we can't ignore the role of social media in how cryptocurrencies like Bitcoin are valued. This project looks into the interesting connection between Twitter and changes in Bitcoin prices, trying to find if there's a link. The reason we're doing this is because we believe that what people say, talk about, and share on Twitter can influence how people feel about the market, and that might affect Bitcoin prices. By carefully studying and understanding these complex connections, the project hopes to add a lot to our understanding of how cryptocurrencies work.

we aim to conduct a thorough analysis of Twitter data, employing advanced tools and methodologies. The scope encompasses identifying patterns, trends, and potential indicators within the vast information shared on Twitter. By comprehensively studying these connections, we seek to provide valuable insights into the relationship between social media and the Bitcoin prices variations.

Analyzing the correlation between sentiments expressed on Twitter and changes in Bitcoin prices is a complex task that involves studying the collective mood and opinions of Twitter users in relation to cryptocurrency markets.

- Positive sentiments on Twitter may indicate a bullish market perception (Upward trajectory). Users expressing optimism, excitement, or confidence in Bitcoin could suggest a positive outlook, potentially leading to increased demand and higher prices.

- Conversely, negative sentiments may signal a bearish market perception (Downward trajectory). Tweets expressing skepticism, fear, or caution could contribute to a more pessimistic market sentiment, potentially leading to lower demand and decreased prices.

The relationship between the frequency of Bitcoin-related discussions on Twitter and subsequent price variations is a topic of interest in the realm of cryptocurrency analysis.
While a correlation may exist, it's crucial to recognize the limitations and potential noise associated with social media data.

This project, driven by more than academic interest, it holds practical implications for investors, financial analysts, and stakeholders navigating the cryptocurrency market. If successful, the findings could offer a valuable tool for decision-making, providing insights into the often unpredictable world of cryptocurrency investments.

WORK DIVISION:

Reetikaa Reddy Munnangi : Data Analysis

Collect and Aggregate Relevant Datasets

- Identify potential sources: Determine where relevant datasets can be sourced, whether it's through public repositories, APIs, or proprietary databases.
- Acquire datasets: Download or access the identified datasets, ensuring they align with the project's objectives.
- Aggregate data: Combine datasets if needed, ensuring compatibility and consistency in variables across different sources.

Ensure Data Quality and Completeness

- Data cleaning: Identify and rectify errors, inconsistencies, or missing values in the dataset.
- Validate data: Cross-check data against known benchmarks or conduct exploratory data analysis to identify outliers or anomalies.
- Handle missing data: Implement strategies to address missing data, such as imputation or removal, while considering the impact on analysis.
- Preprocessing: Normalization or scaling and Transform, or manipulate the data.

Raghavaraju Nithisha : Machine Learning Model Implementation

Implement the Machine Learning Model

- Select appropriate algorithms: Choose machine learning algorithms suited to the specific nature of the problem and dataset.
- Feature engineering: Identify and preprocess relevant features to enhance model performance.
- Model construction: Build the machine learning model using selected algorithms and features.

Fine-Tunning

- Iterative refinement: Systematically adjust hyperparameters based on performance evaluation.
- Validation checks: Employ validation techniques to assess the impact of parameter adjustments.
- Performance metrics: Evaluate the model using appropriate performance metrics to guide parameter fine-tuning.

Duggempudi Vijaya Deepika Reddy : Testing

Evaluate the Model

- Test dataset preparation: Set aside a separate dataset not used during training or fine-tuning for testing purposes.

- Execute tests: Run the model on the test dataset, ensuring a representative evaluation of its performance across different scenarios.
- Monitor for overfitting: Check for signs of overfitting, where the model performs well on the training data but poorly on unseen test data.

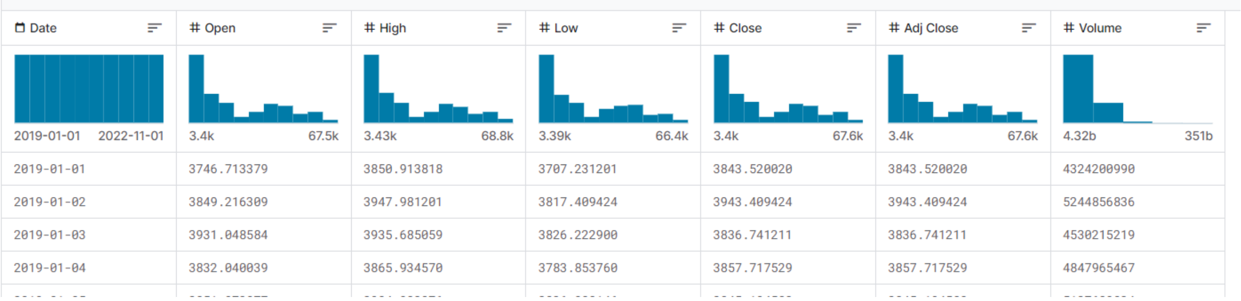
Relevant Metrics

- Select performance metrics: Choose metrics relevant to the nature of the problem, such as accuracy, precision, recall, F1 score, or area under the ROC curve.
- Calculate metrics: Evaluate the model's performance using the selected metrics based on the predictions made on the test dataset.
- Analyze results: Interpret the metric values to gain insights into the model's strengths and areas for improvement.
- Identify strengths and weaknesses: Highlight areas where the model excels and areas that may require improvement based on test result

PROJECT IMPLEMENTATION:

Data Loading and Cleaning:

The data was sourced from Kaggle, encompassing Twitter users such as Elon Musk, Binance, SBF FTX, and CZ Binance, along with numerical data having following data.



Date	# Open	# High	# Low	# Close	# Adj Close	# Volume
2019-01-01 2022-11-01	3.4k 67.5k	3.43k 68.8k	3.39k 66.4k	3.4k 67.6k	3.4k 67.6k	4.32b 351b
2019-01-01	3746.713379	3850.913818	3707.231201	3843.520020	3843.520020	4324200990
2019-01-02	3849.216309	3947.981201	3817.409424	3943.409424	3943.409424	5244856836
2019-01-03	3931.048584	3935.685059	3826.222900	3836.741211	3836.741211	4530215219
2019-01-04	3832.040039	3865.934570	3783.853760	3857.717529	3857.717529	4847965467

Fig1: Dataset

Date : The specific point in time when financial data is recorded.
Open : The starting price of a financial instrument for a given time period.
High : The highest value reached by a financial instrument during a specified time period.
Low : The lowest value reached by a financial instrument during a specified time period.
Close : The concluding price of a financial instrument for a given time period.
Adj Close : The closing price adjusted for factors like dividends or stock splits.
Volume : The total number of shares or units traded during a specified time period.

The Twitter data is concatenated into a single dataframe. To enhance the quality of the textual content, the code introduces functions for cleaning tweet text, systematically removing retweets, URLs, mentions, and symbols. Subsequently, sentiment analysis is conducted using the VADER sentiment analysis tool. VADER assigns a compound score to each tweet, encapsulating its overall sentiment. This compound score, ranging from -1 to 1, categorizes tweets as positive, negative, or neutral based on their emotional tone. The sentiment insights derived from this analysis serve as a valuable foundation for understanding public sentiment surrounding Bitcoin and may be instrumental in exploring correlations between Twitter sentiment and cryptocurrency price movements.

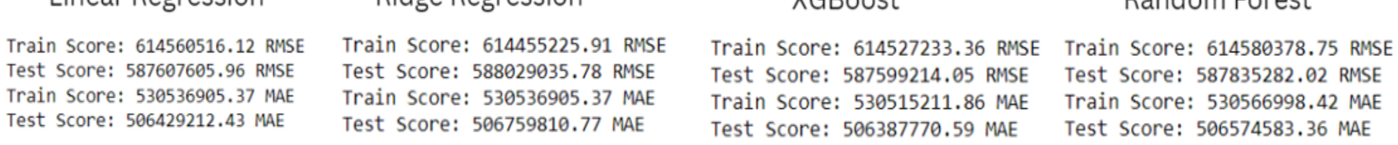
Data Preprocessing and Exploration:

In the initial stages of data analysis, comprehensive data preprocessing was conducted to ensure the quality and reliability of the dataset. Missing values were addressed through appropriate techniques, and outliers were identified and examined using box plots to maintain data integrity. To gain insights into the relationships among different features, a correlation analysis was performed and visually represented through a heatmap, providing a clear overview of inter-feature dependencies.

Machine Learning Model Implementation :

The project involved in exploring various machine learning models such as, Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, XGBoost Regression, and LSTM Neural Network.

Linear Regression served as an initial benchmark, capturing straightforward relationships between input features and Bitcoin prices. Ridge and Lasso Regressions were then employed to introduce regularization, aiding in mitigating potential overfitting issues and enhancing the models' generalization to unseen data. Random Forest Regression and XGBoost Regression, both ensemble methods, were chosen for their ability to capture complex non-linear relationships and patterns present in the data. However, the observed results were very bad with very high mean squared error as shown below. Metrics Rootmean square error and Mean absolute error is used to analyse the model.



Linear Regression

```
Train Score: 614560516.12 RMSE
Test Score: 587607605.96 RMSE
Train Score: 530536905.37 MAE
Test Score: 506429212.43 MAE
```

Ridge Regression

```
Train Score: 614455225.91 RMSE
Test Score: 588029035.78 RMSE
Train Score: 530536905.37 MAE
Test Score: 506759810.77 MAE
```

XGBoost

```
Train Score: 614527233.36 RMSE
Test Score: 587599214.05 RMSE
Train Score: 530515211.86 MAE
Test Score: 506387770.59 MAE
```

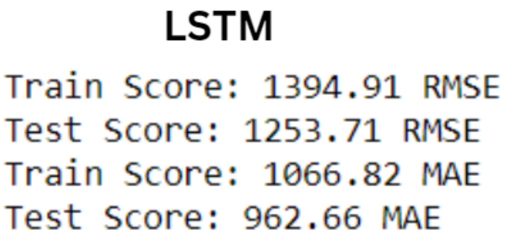
Random Forest

```
Train Score: 614580378.75 RMSE
Test Score: 587835282.02 RMSE
Train Score: 530566998.42 MAE
Test Score: 506574583.36 MAE
```

Fig 2: RMSE and MAE scores

However, The LSTM (Long Short-Term Memory) had the standout performer in this scenario. The inherent strength of LSTMs lies in their ability to effectively model sequential data and capture long-range dependencies, which is particularly valuable when dealing with time-series data, such as Bitcoin price movements and related Twitter sentiment over time. The LSTM's success can be attributed to its capability to recognize and leverage temporal patterns and intricate dependencies within the Twitter data, allowing it to better adapt to the dynamic nature of cryptocurrency markets. The neural network's architecture excels in learning from historical patterns, making it well-suited for tasks where understanding the temporal evolution of data is crucial.

Error Score for LSTM are shown below.



LSTM

```
Train Score: 1394.91 RMSE
Test Score: 1253.71 RMSE
Train Score: 1066.82 MAE
Test Score: 962.66 MAE
```

Fig 3 : RMSE and MAE score for LSTM

LOSS VS EPOCH CURVE

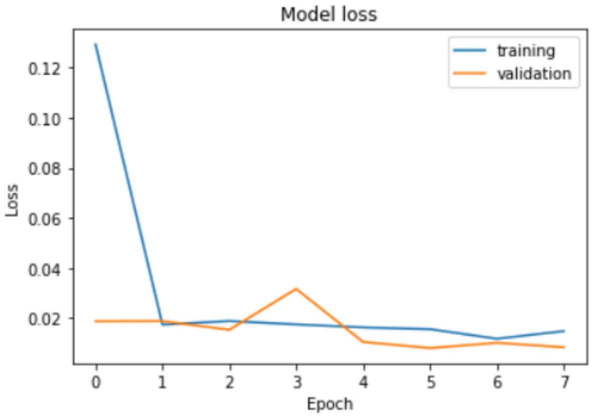


Fig 4 : Loss vs Epoch Curve

The good performance of the LSTM in Bitcoin price prediction using Twitter data underscores the significance of considering time-dependent features and the intricate interplay between social media sentiment and cryptocurrency market dynamics.

Model Evaluation :

We evaluate the performance of each model by measuring root mean squared error (RMSE) and mean absolute error (MAE) on both the training and test sets. The results are presented for each model, offering valuable insights into their effectiveness in predicting Bitcoin prices.

The generated plot provides a comprehensive visualization of Bitcoin price predictions derived from our model. The x-axis represents dates, reflecting the chronological order of the data. The y-axis denotes Bitcoin values, capturing both actual and predicted values. The 'Train' curve corresponds to the model's predictions on the training data, while the 'Test' curve extends into the future, showcasing the model's performance on previously unseen data. The 'Actual' line represents the true Bitcoin values, combining both training and testing periods. The visual comparison between predicted and actual values enables a quick assessment of the model's accuracy and its ability to capture underlying trends.

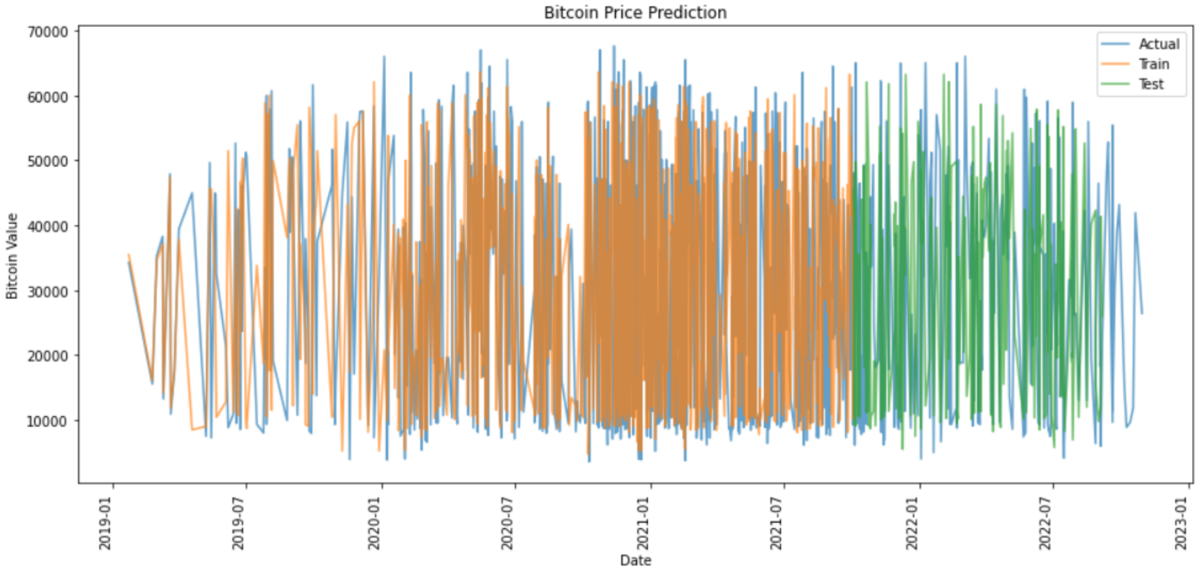


Fig 5 : Train and Test data Evaluation

The code also demonstrates how to utilize the trained LSTM model for predicting Bitcoin prices on new data, providing practical guidance for extending the model's applicability to real-world scenarios.

CONCLUSION :

In summary, our project set out to explore the connection between Twitter activity and predicting Bitcoin prices. We recognized the growing importance of cryptocurrencies like Bitcoin and believed that social media, especially Twitter, could help predict how the market feels about them. We aimed to solve challenges in predicting Bitcoin price changes, empowering people like investors and regulators. We followed a step-by-step process, gathering Twitter data, analyzing sentiments using Natural Language Processing, and creating features like sentiment scores and tweet volume.

While we faced challenges like dealing with unreliable data and the unpredictable nature of financial markets, we remain optimistic about the potential impact of our project.Moving forward, we recognize the importance of considering challenges in interpreting data and understanding how market sentiments can change. The collaboration of our team members, each contributing their expertise to collecting data, building models, and testing, highlights the multifaceted nature of our project and its potential to provide valuable insights in the evolving world of cryptocurrency analytics.