Cryptocurrency Price Prediction

Shivangi Gupta, Huize Mao
UC San Diego – Data Science Project Showcase

data science student society

Introduction

During the COVID-19 pandemic, cryptocurrencies like Dogecoin, Bitcoin, and Ethereum captured the attention of people worldwide. But what exactly drove these currencies to soar, and can machine learning be harnessed to predict future prices? Through this project, we delved deeper into the world of crypto assets, uncovering key factors that contribute to their dramatic price spikes. Ultimately, we developed several robust regression models capable of accurately predicting prices within our datasets (from Yahoo Finance and CoinMarketCap).

Background

Cryptocurrencies are digital assets that eliminate the need for intermediaries by using encryption to validate transactions. Bitcoin, introduced in 2009, was the first decentralized cryptocurrency, pioneering the use of blockchain technology to ensure transaction security and transparency. Ethereum, launched in 2015, expanded the functionality of blockchain with smart contracts, enabling decentralized applications beyond financial transactions. Dogecoin, created in 2013 by Billy Markus and Jackson Palmer, started as a meme-based cryptocurrency but gained significant traction. These cryptocurrencies are affected by various factors such as market demand and social media trends so their price movements highly volatile and challenging to predict.

Data Cleaning

After pre-processing, BTC had 5925 rows, ETH had 3945 rows, DOGE had 1723 rows

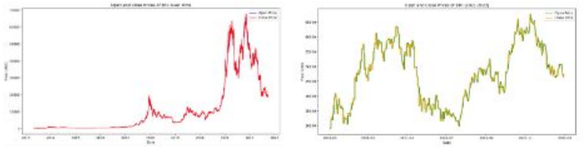
	SNo	Name	Symbol	Date	High	Low	Open	Close	Volume	Marketcap
0	1	Bitcoin	BTC	2013-04-29 23:59:59	147.488007	134.000000	134.444000	144.539993	0.0	1.603769e+09
1	2	Bitcoin	BTC	2013-04-30 23:59:59	146.929993	134.050003	144.000000	139.000000	0.0	1.542813e+09
2	3	Bitcoin	BTC	2013-05-01 23:59:59	139.889999	107.720001	139.000000	116.989998	0.0	1.298955e+09
3	4	Bitcoin	BTC	2013-05-02 23:59:59	125.599998	92.281898	116.379997	105.209999	0.0	1.168517e+09
4	5	Bitcoin	BTC	2013-05-03 23:59:59	108.127998	79.099998	106.250000	97.750000	0.0	1.085995e+09

- Standardized columns for consistency across datasets
- Merged datasets to get the most data: April 2013 - Sep 2022
- Dropped extraneous or redundant columns
- Formatted time series data using datetime module
- Standardized units and case for data points
- Normalized numerical features using StandardScaler()
- Dropped or imputed missing values

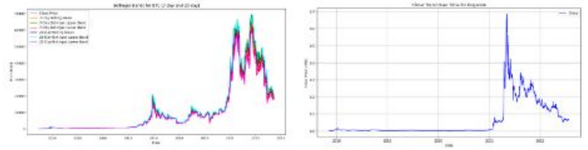
Exploratory Data Analysis

We investigated the distributions and time series trends of various variables for each of our cryptocurrencies. We noticed that 2021-2022 was a rather volatile year with exceptionally high BTC prices, so we took a closer look at that in the data. Note how the y-axis differs in scale from crypto to crypto (BTC is much higher than DOGE which is only traded in cents), reflecting their respective market prices. We also researched about how factors like market cap were crucial indicators of an asset's stability, attracting more investors and thus possibly raising prices.

The graph below on the left shows BTC's Open and Close prices between 2013-2023, and the one on the right zooms into BTC prices only between 2021-2022.



The graph below on the left shows the short and medium-term Bollinger Bands for BTC's Close prices between 2013-2023, and the right graph depicts DOGE prices between 2018 and 2022, since the data we have for Dogecoin was more recent than both BTC and ETH.

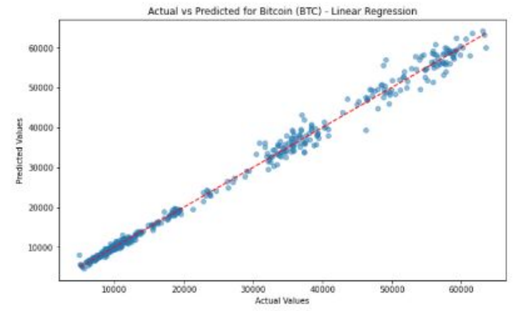


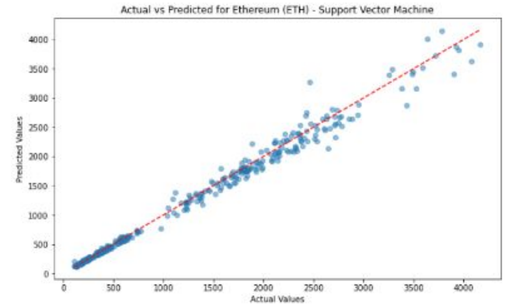
Feature Engineering

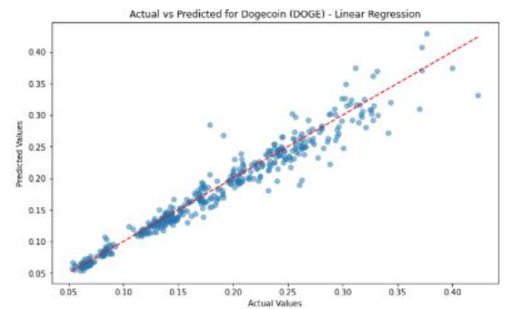
We wanted to capture 2 key features with our models: price change and volatility. Price change is the absolute change in closing prices from the previous day, and our volatility indicators were Bollinger Bands, which are upper and lower bands based on a moving average of a 7 day and 20 day window (covering both short term and medium term trading trends) and standard deviation. Thus, we created the following new features: Price_Change, Rolling_Mean_7, Bollinger_Upper_7, Bollinger_Lower_7, Rolling_Mean_20, Bollinger_Upper_20, and Bollinger_Lower_20 to capture price trends and volatility. Since we did time-series forecasting, we created a target column with future closing prices by shifting our target variable (closing price).

Modeling

We evaluated various machine learning models to predict the closing prices of Bitcoin (BTC), Ethereum (ETH), and Dogecoin (DOGE). Our approach involved training and testing multiple regression models (75%/25% split), including Linear Regression, KNeighbors Regressor, Support Vector Machines, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor. The evaluation metrics considered for each model were Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score.







Results

Using a test size of 0.25, we achieve the following performance metrics for our best models:

Cryptocurrency	Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R^2 Score
Bitcoin (BTC)	Linear Regression	1,312,054.70	1,145.45	0.9951
Ethereum (ETH)	Support Vector Machine	9,017.01	94.96	0.9898
Dogecoin (DOGE)	Linear Regression	0.0003	0.0176	0.9491

- BTC: The plot shows a strong linear relationship between actual and predicted values, with most points closely following the red dashed line, also reflected in the high R^2 score. BTC has the highest y-axis values, reflecting its large price scale which ranges from a few thousand to over $60,000.
- ETH: The plot demonstrates a strong linear relationship, similar to Bitcoin, indicating that the SVM model accurately captures the price patterns of Ethereum. The price values are smaller than Bitcoin, as Ethereum usually ranges from a few thousand to over $4,000.
- DOGE: The plot shows a good linear fit, though there is more dispersion compared to the other two cryptocurrencies. The values are much smaller, reflecting Dogecoin's lower price in USD (a range of a few cents to tens of cents).

Takeaways

We can see that Bitcoin's Linear Regression model shows an exceptionally high MSE of 1,312,054.70 and RMSE of 1,145.45, but it explains 99.51% of the variance (R^2 score). Dogecoin's best model, the Linear Regression model, has a very low MSE of 0.0003 and RMSE of 0.0176, with an R^2 score of 0.9491, suggesting a robust linear relationship with much less variability compared to Bitcoin. Ethereum's SVM model has an MSE of 9,017.01 and RMSE of 94.96, with an R^2 score of 0.9898, explaining a lot of the variability in the data. Overall, R^2 scores are high for all, indicating good fit, but Bitcoin's high MSE may result from its greater volatility. Dogecoin's lower MSE may reflect its smaller price scale and fewer data points. Linear models work well for Bitcoin and Dogecoin, while Ethereum's SVM hints at more complex patterns.

Acknowledgements

Cryptoassets
https://rpc.cfainstitute.org/-/media/documents/article/rf-brief/rfbr-cryptoassets.pdf
Bitcoin Explained and made Simple
https://www.youtube.com/watch?v=s4g1XFU8Gto
Number of identity-verified Cryptoasset Users
https://www.statista.com/statistics/1202503/global-cryptocurrency-user-base/
The History of Bitcoin
https://money.usnews.com/investing/articles/the-history-of-bitcoin
A Brief History of Ethereum
https://www.theblock.co/learn/245716/a-brief-history-of-ethereum
History of Dogecoin
https://dogecoin.com/dogepedia/articles/history-of-dogecoin/