Aaron Amha

Owen Crenshaw

Zach Nguyen

Ben Xiang

CSCI-4502: Data mining

Bitcoin Price Dataset Project Proposal (2017-2023)

Introduction/Motivation

Our group selected the Bitcoin Price Dataset as our project for this course. The reason we chose this topic is because since the launch of bitcoin in 2009, it seems like the coin has vastly skyrocketed in popularity and price. To put into context in 2010 the price of bitcoin was $0.05 and today it is about $26,000. In addition, out of all digital currencies such as Etherium, Dogecoin, Binance Coin (BNB), Bitcoin seems to be the most favored with many large companies such as Expedia, Microsoft, AT&T allowing users to purchase goods and services with bitcoin (Tan). Being able to see bitcoin prices will allow us to predict its future values and its sustainability in the current market. In addition, our team members have been interested in crypto currencies, specifically bitcoin, and wanted to know more about how it operates.

While neither of our group members have any bitcoins or hold any stakes in the stock market, the broader cryptocurrency community, investors, and those interested in the future of digital currencies have a stake in Bitcoin's performance. Therefore, our analysis may provide valuable insights for anyone interested in the cryptocurrency landscape. By studying Bitcoin's historical data, through a price trend analysis, we hope to predict its future price trends and assess its sustainability in the current market. While we are students embarking on this project,

our findings may offer valuable perspectives for those interested in the world of cryptocurrencies.

Literature Review

As stated in James Royal's article, "Bitcoin's price history: 2009 to 2023,", in 2010 bitcoin "...never broke above $0.40 per bitcoin…" (Royal) but today bitcoin casually sits, "...or [trades] for around $26,000…"(Royal), each BTC. According to Tanzeel Akhtar, from Bloomberg, "The largest digital token had ended positive on Sunday for the first week in five, and increased as much as 3.7% to $27,418 on Monday" (Akhtar, 2023). According to Zheshi Chen's scholarly article, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering, "A set of high-dimension features including property and network, trading and market, attention and gold spot price are used for Bitcoin daily price prediction, while the basic trading features acquired from a cryptocurrency exchange are used for 5-minute interval price prediction. [Furthermore] Statistical methods including Logistic Regression and Linear Discriminant Analysis for Bitcoin daily price prediction with high-dimensional features achieve an accuracy of 66%, outperforming more complicated machine learning algorithms"(Chen). Although our dataset does not include some of the traditional standards in the field, we are using an API called Binance API. We are using Binance API rather than any other API because it is the current leading API for bitcoin data mining. According to many blogs such Algotrading101 and abstract, they wrote "Binance has established itself as a market leader when it comes to cryptocurrency trading. It currently ranks number one for Bitcoin volume according to coinmarketcap.com and ranks well for many other currencies."

Proposed Work

In our project, we will start by placing the dataset, 'bitcoin_2017_to_2023.csv,' in our working directory for easy accessibility in our Python code. As a team of four, we have a well-structured strategy for collaboration. We have designated meeting times to efficiently coordinate project tasks and ensure equal distribution of workload among group members. Our dataset is sourced from the Bitcoin Price Dataset on Kaggle, and it is well-organized, with easily understandable code. Our overarching objective is to conduct an in-depth exploration of the world of Bitcoin. We aim to analyze the trends that have occurred over the span of 6 years from 2017 - 2023. Additionally, we plan to identify optimal moments for buying or holding bitcoins based on these trends and patterns and,ultimately, enhance our understanding of this prominent digital currency. Furthermore, we intend to share our insights to educate both ourselves and our viewers about the dynamics of the rapidly evolving Bitcoin landscape.

To start, we will gather the dataset, ‘bitcoin_2017_to_2023.csv’ and preprocess the data provided. The dataset provided by Kaggle contains over 3 million instances worth of bitcoin price values to account for every minute from the start date in 2017 to August 1, 2023 at 1:19PM. Each minute represented contains several values for each minute consisting of, opening value, the highs, the lows, closing value, volume, the quote-asset-volume value, the number of trades, the taker-buy-base-asset-volume value, and the taker-buy-quote-asset-volume value. This quantitative data will undergo cleaning, integration, reduction, and transformation.

For **data cleaning**, we must focus on outliers and duplicate values which may exist in the dataset. By removing these values and occurrences, we prevent excess data from ruining the data’s accuracy and potential trend patterns. Therefore, it is essential that we smooth any noisy data, remove any outliers, and resolve any inconsistencies. By focusing on either regression or clustering, we should be able to easily determine and remove the noisy data.For **data**

integration, we will need to amalgamate, or combine data from different sources to create a more insightful analysis. We can use either schema integration or object matching for this part. For **data transformation**, we will need to modify the data by normalizing and adjusting it to work well in any potential machine learning algorithms we may use. Lastly, we will need to use **data reduction** to make the dataset easier to understand while still maintaining the overall essence of the data.

Overall, our objective with this data mining experience is to predict Bitcoin cryptocurrency prices moving forward provided the historical trends recorded over the past 6 years. With significant outliers, we will be able to see what affects the market and how similar effects in the future may demonstrate correlation. These results may be useful not only for ourselves but others who may hold interest in cryptocurrencies, especially Bitcoin.

Evaluation

The metrics that we hope to use are mainly price. We aim to extract data using the Binance API and Matplotlib finance library. These APIs will help us visualize Bitcoin prices in a way that we can determine why they are dropping. There are no official equations to determine this but by using inbuilt tools such as the ones suggested above, we can make certain graphs to calculate metrics. We can utilize market caps to lead ourselves into comparative analysis. One equation that will be useful for calculating the price of this cryptocurrency is the BTC total market cap divided by the number of existing coins. By doing this division, we are left with the cost of each bitcoin. We can then compare this value with other cryptocurrencies and see how the trends of these values change over time. Furthermore, during our evaluation we will clean the data to make it into a form that is useful and informative. While outliers in some data are bad, the

outliers in our data set will tell us more about the price metric and why that metric is acting in such a manner. To validate the data we can develop a sophisticated program to determine the outcome depending on certain features of the market. Additionally, validation of the data will consist of determining any false positives, and the best way to determine these is through cross-validation. For outlier validation, we can use visualizations to determine whether any values recorded fall far from the mean or average value of the distribution during that day, week or even month depending on consistency. This will be able to test on previous data as well as data that is new. For predictive data, we will utilize statistical testing to see if the consistencies are true.

Milestones

In this project, we have outlined three key milestones to guide our progress:

Our initial milestone involves establishing the foundational structure of our project. We will set up the code's basic structure, ensuring that all necessary libraries are imported. Additionally, we will meticulously organize folders and files for future efficiency and success. The second milestone represents a critical phase of our project. Here, we aim to visualize Bitcoin trends through various types of plots, including scatter plots, box plots, histograms, and more. Beyond the technical aspect, one of our primary goals is to educate viewers about Bitcoin, its workings, and its trends. This milestone will be the most coding-intensive and holds paramount importance in our project's overall development. Our final milestone focuses on preparing the project for presentation. During this phase, we will refine the dataset and enhance its visual elements to ensure clarity and ease of understanding for our audience. This milestone marks the culmination of our project efforts.

In the context of our class, this project aligns with the core principles taught, emphasizing the recognition of patterns and trends in datasets to enhance our comprehension. Such understanding is valuable for businesses in assessing their profitability and operations. By choosing Bitcoin as our subject matter, we aim to solidify our knowledge of cryptocurrency dynamics and better prepare for the future. While we have not specified exact dates for these milestones, we are committed to completing the first and third milestones promptly. Our primary focus will be on efficiently progressing through the second milestone to meet our project deadline.

References

- Akhtar, Tanzeel. "Bitcoin (BTC) Rises above $27,000 for First Time since August." *Bloomberg.Com*, Bloomberg, 18 Sept. 2023, www.bloomberg.com/news/articles/2023-09-18/bitcoin-climbs-above-27-000-for-the-first-time-since-august#xj4y7vzkg.
- Chen, Zheshi, et al. "Bitcoin Price Prediction Using Machine Learning: An Approach to Sample Dimension Engineering." *Journal of Computational and Applied Mathematics*, North-Holland, 13 Aug. 2019, www.sciencedirect.com/science/article/abs/pii/S037704271930398X
- Davda, Jignesh. "Binance Python API – a Step-by-Step Guide - Algotrading101 Blog." *Quantitative Trading Ideas and Guides - AlgoTrading101 Blog*, 4 Apr. 2023, algotrading101.com/learn/binance-python-api-guide/#:~:text=The%20Binance%20API%20is%20a,to%20send%20and%20receive%20data.
- Kraayenbrink, Jonathan. "Bitcoin Price Dataset (2017-2023)." *Kaggle*, Kaggle, 24 Aug. 2023, www.kaggle.com/datasets/jkraak/bitcoin-price-dataset/data.

➢ Royal, James. “Bitcoin’s Price History: Tracking the Volatile Rise of the World’s Biggest Cryptocurrency.” Edited by Brian Beers, *Bankrate*, Bankrate, 14 June 2023, www.bankrate.com/investing/bitcoin-price-history/.

➢

➢ Tan, Eli. “What Can You Buy with Bitcoin?” *CoinDesk Latest Headlines RSS*, CoinDesk, 29 May 2023, www.coindesk.com/learn/what-can-you-buy-with-bitcoin/.

➢

ACM Template

Link: https://www.overleaf.com/3383688747ygbhdtgwqsqp#c5c642

Bitcoin Price Dataset Project Proposal (2017-2023)

AARON AMHA, OWEN CRENSHAW, ZACH NGUYEN, and BEN XIANG

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

ACM Reference Format:
Aaron Amha, Owen Crenshaw, Zach Nguyen, and Ben Xiang. 2023. Bitcoin Price Dataset Project Proposal (2017-2023). 1, 1 (October 2023), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Our group selected the Bitcoin Price Data set as our project for this course. The reason we chose this topic is because since the launch of bitcoin in 2009, it seems like the coin has vastly skyrocketed in popularity and price. To put into context in 2010 the price of bitcoin was $0.05 and today it is about $26,000. In addition, out of all digital currencies such as Etherium, Dogecoin, Binance Coin (BNB), Bitcoin seems to be the most favored with many large companies such as Expedia, Microsoft, AT&T allowing users to purchase goods and services with bitcoin (Tan). Being able to see bitcoin prices will allow us to predict its future values and its sustainability in the current market. In addition, our team members have been interested in crypto currencies, specifically bitcoin, and wanted to know more about how it operates.

While neither of our group members have any bitcoins or hold any stakes in the stock market, the broader cryptocurrency community, investors, and those interested in the future of digital currencies have a stake in Bitcoin's performance. Therefore, our analysis may provide valuable insights for anyone interested in the cryptocurrency landscape. By studying Bitcoin's historical data, through a price trend analysis, we hope to predict its future price trends and assess its sustainability in the current market. While we are students embarking on this project, our findings may offer valuable perspectives for those interested in the world of cryptocurrencies.

2 LITERATURE REVIEW

As stated in James Royal's article, "Bitcoin's price history: 2009 to 2023,"in our introduction, in 2010 bitcoin "...never broke above $0.40 per bitcoin..." (Royal)was .05 but today bitcoin casually sits, "...or [trades] for around at $26,000..."(Royal), each BTC. According to Tanzeel Akhtar, from Bloomberg, "The largest digital token had ended positive on Sunday for the first week in five, and increased as much as 3.7% to $27,418 on Monday" (Akhtar, 2023). According to Zheshi Chen's scholarly article, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering, "A set of high-dimension features including property and network, trading and market, attention and gold spot price are used for Bitcoin daily price prediction, while the basic trading features acquired from a cryptocurrency exchange

Authors' address: Aaron Amha; Owen Crenshaw; Zach Nguyen; Ben Xiang.

Manuscript submitted to ACM

are used for 5-minute interval price prediction. [Furthermore] Statistical methods including Logistic Regression and Linear Discriminant Analysis for Bitcoin daily price prediction with high-dimensional features achieve an accuracy of 66%, outperforming more complicated machine learning algorithms" (Zheshi Chen)." Although our dataset does not include some of the traditional standards in the field, we are using an API called Binance API. We are using Binance API rather than any other API because it is the current leading API for bitcoin data mining. According to many blogs such Algotrading101 and abstract, they wrote "Binance has established itself as a market leader when it comes to cryptocurrency trading. It currently ranks number one for Bitcoin volume according to coinmarketcap.com and ranks well for many other currencies."

3 PROPOSED WORK

In our project, we will start by placing the dataset, 'bitcoin_2017_to_2023.csv,' in our working directory for easy accessibility in our Python code. As a team of four, we have a well-structured strategy for collaboration. We have designated meeting times to efficiently coordinate project tasks and ensure equal distribution of workload among group members. Our dataset is sourced fromthe Bitcoin Price Dataset on Kaggle, and it is well-organized, with easily understandable code. Our overarching objective is to conduct an in-depth exploration of the world of Bitcoin. We aim to analyze the trends that have occurred or recured over the span of 6 years from 2017 - 2023.specific timeframes., Additionally, we plan to identify optimal moments for buying or holding bitcoins based on these trends and patterns, and, ultimately, enhance our understanding of this prominent digital currency. Furthermore, we intend to share our insights to educate both ourselves and our viewers about the dynamics of the rapidly evolving Bitcoin landscape.

To start, we will gather the dataset, 'bitcoin_2017_to_2023.csv' and preprocess the data provided. The dataset provided by Kaggle contains over 3 million instances worth of bitcoin price values to account for every minute from the start date in 2017 to August 1, 2023 at 1:19PM. Each minute represented contains several values for each minute consisting of, opening value, the highs, the lows, closing value, volume, the quote-asset-volume value, the number of trades, the taker-buy-base-asset-volume value, and the taker-buy-quote-asset-volume value. This quantitative data will undergo cleaning, integration, reduction, and transformation.

For **data cleaning**, we must focus on outliers and duplicate values which may exist in the dataset. By removing these values and occurrences, we prevent excess data from ruining the data's accuracy and potential trend patterns. Therefore, it is essential that we smooth any noisy data, remove any outliers, and resolve any inconsistencies. By focusing on either regression or clustering, we should be able to easily determine and remove the noisy data.For **data integration**, we will need to amalgamate, or combine data from different sources to create a more insightful analysis. We can use either schema integration or object matching for this part. For **data transformation**, we will need to modify the data by normalizing and adjusting it to work well in any potential machine learning algorithms we may use. Lastly, we will need to use **data reduction** to make the dataset easier to understand while still maintaining the overall essence of the data.

Overall, our objective with this data mining experience is to predict Bitcoin cryptocurrency prices moving forward provided the historical trends recorded over the past 6 years. With significant outliers, we will be able to see what affects the market and how similar effects in the future may demonstrate correlation. These results may be useful not only for ourselves but others who may hold interest in cryptocurrencies, especially Bitcoin.

4 EVALUATION

The metrics that we hope to use are mainly price. We aim to extract data using the Binance API and Matplotlib finance library. These APIs will help us visualize Bitcoin prices in a way that we can determine why they are dropping. There are no official equations to determine this but by using inbuilt tools such as the ones suggested above, we can make certain graphs to calculate metrics. We can utilize market caps to lead ourselves into comparative analysis. One equation that will be useful for calculating the price of this cryptocurrency is the BTC total market cap divided by the number of existing coins. By doing this division, we are left with the cost of each bitcoin. We can then compare this value with other cryptocurrencies and see how the trends of these values change over time. Furthermore, d During our evaluation we will clean the data to make it into a form that is useful and informative. While outliers in some data are bad, the outliers in our data set will tell us more about the price metric and why that metric is acting in such a manner. To validate the data we can develop a sophisticated program to determine the outcome depending on certain features of the market. Additionally, validation of the data will consist of determining any false positives, and the best way to determine these is through cross-validation. For outlier validation, we can use visualizations to determine whether any values recorded fall far from the mean or average value of the distribution during that day, week or even month depending on consistency. This will be able to test on previous data as well as data that is new. For predictive data, we will utilize statistical testing to see if the consistencies are true.

5 MILESTONES

In this project, we have outlined three key milestones to guide our progress:

Our initial milestone involves establishing the foundational structure of our project. We will set up the code's basic structure, ensuring that all necessary libraries are imported. Additionally, we will meticulously organize folders and files for future efficiency and success. The second milestone represents a critical phase of our project. Here, we aim to visualize Bitcoin trends through various types of plots, including scatter plots, box plots, histograms, and more. Beyond the technical aspect, one of our primary goals is to educate viewers about Bitcoin, its workings, and its trends. This milestone will be the most coding-intensive and holds paramount importance in our project's overall development. Our final milestone focuses on preparing the project for presentation. During this phase, we will refine the dataset and enhance its visual elements to ensure clarity and ease of understanding for our audience. This milestone marks the culmination of our project efforts.

In the context of our class, this project aligns with the core principles taught, emphasizing the recognition of patterns and trends in datasets to enhance our comprehension. Such understanding is valuable for businesses in assessing their profitability and operations. By choosing Bitcoin as our subject matter, we aim to solidify our knowledge of cryptocurrency dynamics and better prepare for the future. While we have not specified exact dates for these milestones, we are committed to completing the first and third milestones promptly. Our primary focus will be on efficiently progressing through the second milestone to meet our project deadline.

6 REFERENCES

- Akhtar, Tanzeel. "Bitcoin (BTC) Rises above $27,000 for First Time since August." *Bloomberg.Com*, Bloomberg, 18 Sept. 2023, www.bloomberg.com/news/articles/2023-09-18/bitcoin-climbs-above-27-000-for-the-first-time-since-august#xj4y7vzkg.

- Chen, Zheshi, et al. “Bitcoin Price Prediction Using Machine Learning: An Approach to Sample Dimension Enginehttps://www.overleaf.com/project/6540c702f9251f19bdf17742ering.” *Journal of Computational and Applied Mathematics*, North-Holland, 13 Aug. 2019, www.sciencedirect.com/science/article/abs/pii/S037704271930398X
- Davda, Jignesh. “Binance Python API – a Step-by-Step Guide - Algotrading101 Blog.” *Quantitative Trading Ideas and Guides - AlgoTrading101 Blog*, 4 Apr. 2023, algotrading101.com/learn/binance-python-api-guide/#:~:text=The%20Binance%20API%2
- Kraayenbrink, Jonathan. “Bitcoin Price Dataset (2017-2023).” *Kaggle*, Kaggle, 24 Aug. 2023, www.kaggle.com/datasets/jkraak/bitcoin-price-dataset/data.
- Royal, James. “Bitcoin’s Price History: Tracking the Volatile Rise of the World’s Biggest Cryptocurrency.” Edited by Brian Beers, *Bankrate*, Bankrate, 14 June 2023, www.bankrate.com/investing/bitcoin-price-history/.
-

Tan, Eli. “What Can You Buy with Bitcoin?” *CoinDesk Latest Headlines RSS*, CoinDesk, 29 May 2023, www.coindesk.com/learn/what-can-you-buy-with-bitcoin/.