**1. Introduction**

***1.1 Blockchain Technology***

Before performing our analysis, we first need to define what Blockchain technology and Bitcoin are. Blockchain technology is a distributed transaction ledger that is maintained by a peer-to-peer network. As is the case with other cryptocurrencies, Bitcoin, the currency, is secured by the bitcoin network. Other uses of blockchain include other cryptocurrencies, platforms like Ethereum and Cardano, and decentralized finance (DeFi) applications.

Blockchain technology is important because of both its intrinsic value and design and its extrinsic value. Intrinsically, blockchain technology enables currencies and financial applications to have high security and privacy. Extrinsically, the technology has attracted, at increasing rates, the investment of energy, human capital, and money into Bitcoin and altcoins alike.

We seek to answer what the best predictor of the price of Bitcoin is. We examine external factors such as news headlines and various time series models to predict Bitcoin price.

***1.2 Literature Review***

In answering which methods best predict the price of Bitcoin, we build upon extensive literature in asset price prediction, cryptocurrency economic analysis, and natural language processing methodology. A key nuance of a review of price prediction or fundamental financial analysis literature concerning cryptocurrencies is timing. Given the high volatility of Bitcoin price, there have been waves of speculation, media attention, and academic attention that are highly correlated with the rise in its price. Around 2015, the consensus in studies was that Bitcoin is a speculative bubble more appropriately classified as an asset than a currency (Bucholz et al. 2012; Kristoufek 2013; Ciaian et al. 2014; Yermack 2014; Bouoiyour et al. 2015; Bouoiyour and Selmi 2015). Additionally, Bitcoin is a proxy for the health of all other cryptocurrencies; it is the largest cryptocurrency by market capitalization, was the first successful cryptocurrency, and subsequently has been the most notable cryptocurrency. As with many other fields, with time comes more data and thus more robust results can come from the same analysis. The next important element to consider when examining the literature is whether or not the foundations of authors’ understanding of Bitcoin and the bitcoin protocol are correct. Assertion of security concerns within the bitcoin network to support claims of precarious price are not true (Bouoiyour, Selmi 2016); shaky bitcoin technical foundations also often lead to omissions of crucial components of the bitcoin protocol design such as the eventual, finite supply.

In earlier bull runs like in 2015 and 2017 alike, there was a proliferation of fundamental analysis of Bitcoin. In 2015, Cheah et. all conclude that the fundamental value of Bitcoin is zero and the rapid rise and fall of its price proves it is a speculative asset rather than a currency.

In addition to analysis of Bitcoin as a currency and asset, the use of natural language processing to predict price has been used not only with traditional financial assets but with cryptocurrencies as well. In their analysis of Bitcoin and Ethereum, currently, the largest cryptocurrencies as measured by market capitalization, Abraham et al. find that the volume of tweets is more important than a sentiment analysis of those same tweets. We extend this conclusion into our hypothesis for the data gathered from Twitter. Using a different source, Google Trends, Kristoufek et al. also attempt to quantify the relationship between the price of Bitcoin and buzz in the form of search queries. We will extend these efforts by applying similar text analysis methodologies on more current data while also combining those with other methodologies on intrinsic and market metrics.

**2. Datasets**

***2.1 Data Description***

In any given bitcoin transaction, inputs and outputs are populated by addresses, the value of the transaction fee, and information about the location of the transaction in the block. From these data, sources like exchanges create market data. These data could fall into the following categories: network usage, market, mining, fees and revenues, and supply. Outside of the bitcoin network and cryptocurrency exchanges, useful data may be found on popular financial news outlets. From these two sources, we incorporate intrinsic transaction data, market metrics, and sentiment analysis of select data from financial news outlets into our analysis. Our bitcoin data comes from Coin Metrics which organizes the world’s crypto data and makes it transparent and accessible. Description of each variable in the downloaded dataset comes from Coin Metrics as well.

***2.2 Data Preparation Summar***

*2.2.1 Twitter API*

Twitter has 330 million monthly active users, 1.3 billion accounts created, 83% of the world’s leaders have Twitter accounts, approximately 23 million of Twitter’s active users are bots rather than humans, and 500 million tweets are sent each day.  The result of all of these spectacular statistics is that Twitter can be a very rich source of data on how people feel about certain topics.  This makes Twitter a great resource to collect text data on a topic like cryptocurrencies to explore relationships between that and price.

Twitter data is essential in diverse fields nevertheless, the API proposed by Twitter harshly restricts access to public data generated by users.  These restrictions have consequences that greatly slowed down the contributions of our research by limiting our scope.  First, an academic research Twitter API account was submitted for approval by Twitter Developers.  After a 2 business day turnaround approval was denied.  However, after a second submission and a 3-day business follow-up, access was granted which provided the API key, API secret key, access token, and access token secret to begin tweet retrieval with V1.1 access.  Access to the API was connected with Jupyter notebook using Python and R to pull historical tweets dating back to bitcoin’s spike in the price of Q4 2020.  However, difficulty was encountered when historical tweets dating back to Q4 2020 were not retrievable using V1.1 access due to restrictions placed by Twitter Developers.

For due diligence, we ran a decision tree model that had a very high RMSE. This confirmed that we did not have the volume data needed for this data to be useful.   Aside from Twitter Developer restrictions, the experiment expected to grab enough historical tweets dating back to Q4 2020 to collect text data on cryptocurrencies and similar “buzz words” to explore possible relationships between that and price.  However, the restrictions placed led us to pivot to another text-based data source, which was the collection of cryptocurrency news headlines.

*2.2.2 Bitcoin Data*

Two sources of data were used for the analysis. First, daily Bitcoin prices were downloaded from CoinMetrics.com, which included 43 variables with 4,437 observations. Based on the literature review, 14 variables were removed from this raw dataset. The dataset spans from the inception of bitcoin, January 3, 2009, to present, February 25, 2021. The dataset contained N/A values, found from January 3, 2009, to July 17, 2011, as there is an insufficient amount of data during this period to calculate metrics such as *ROI1year*. The dataset was filtered to remove N/A values, with the final clean dataset containing 3,511 observations across 29 variables. Within the bitcoin dataset, the following categories of bitcoin metrics were used: Addresses, Network Usage, Market, Mining, Fees and Revenues, and Supply.   Specifically:

* Addresses:
  + AdrActCnt: Addresses, active, count
* Network Usage
  + BlkWghtTot: Block, weight, total
* Market
  + CapMrktCurUSD: Capitalization, market, current supply, USD
* Mining
  + DiffLast: Difficulty, lastw
  + DiffMean: Difficulty, mean

The scope of the analysis revolves around the daily change in Bitcoin price.   Therefore, the Bitcoin dataset was mutated to calculate the daily price change.  Bitcoin is extremely volatile and as a result, there are several occurrences of significant daily price fluctuations. Below is a table that displays the top five greatest daily percent changes.

|  |  |  |
| --- | --- | --- |
| **Date** | **Price** | **Percent Change** |
| 2020-03-12 | $4,959.313 | -37.53% |
| 2017-12-07 | $17,032.293 | 23.72% |
| 2017-12-06 | $13,992.132 | 19.22% |
| 2019-04-02 | $4,903.719 | 18.49% |
| 2021-02-08 | $46,121.934 | 18.25% |

Fig. 1

Based on the distribution of the variance of price change percentage, the data was filtered to start on August 5, 2018, and end on February 1, 2021.

Chart

Description automatically generated

Fig. 2

*2.2.3 Financial News Outlets*

In place of tweets, we extracted bitcoin-related headlines from financial news sources to create a dataset on which to perform a predictive Text Mining Analysis using TDF and TF-IDF.

Due to the time constraint, the team figured that it would not be possible to extract cryptocurrency-related news data for each daily price change, which led us to pre-define a range in which to examine price change. After examining the overall price trend, the price became volatile starting from August 2017 until now (Fig. 3).

Chart, line chart

Description automatically generated

Fig.3

Pre-defining the range helped to understand the trend, but it was not realistic to use that data for the analysis as it would take a long time to acquire appropriate news data for each data point (1304 observations). Additionally, the price change in the pre-defined range seems to be random and white noise as the mean of the daily price change is close to zero (approx. 0.003), and the values do not correlate with lag values (Fig.4 & 5).

Chart, line chart

Description automatically generated

Fig.4

Chart, timeline

Description automatically generated

Fig.5

The data was preprocessed further by only including dates with the magnitude of the price change (absolute value) that is above the 75th percentile of the range because the project's interest is to investigate which extrinsic factors derive the value of Bitcoin. In other words, we are only interested in the dates with significant changes in the price. As a result, only 326 data points were selected from the 1304 observations.

*2.2.3.1 Challenges*

While searching for appropriate text-based data, there was no single source that allows us to customize a range of news released dates except for Google News. It would be time-consuming to scrap the web data by clicking an individual website. Therefore, Phil utilized Python to automate the process with the following steps:

1. Splitting up the selected dates, setting the current row as the "begin\_date" and the following row as the "end\_date," and organizing them as a new set of ranges.
2. Utilizing a Python package (Selenium) to scrap all cryptocurrency-related news headlines on the first page of Google News.
3. Calculating the rolling average of the price change for each range.
4. Merging the average data with the scraped data.

Graphical user interface, text, application, table

Description automatically generated

Fig.6

**3. Analytical Techniques**

**3.1 Text Mining Analysis**

*3.1.1 Exploratory Data Analysis*

Before preprocessing data, we performed a brief EDA process to acquire a better understanding of the data. This process includes examining the summary statistics of the text data (characters, words, and sentences), such as metrics on count, mean, variance, and correlation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Characters** | **Words** | **Sentences** |
| Mean | 641.145 | 85.012 | 4.068 |
| Variance | 2519.291 | 74.265 | 3.501 |

Fig.7

*3.1.2 Data Preprocessing*

To deal with the unstructured data, it was essential to preprocess before moving to the analysis. The first part of the process was to create a corpus of news headline data using the VCorpus function. After constructing the corpus, the data was cleaned by functions in the tm package in R, including replacing any URL with blank space, removing punctuation and stop words, stripping any extra white space, creating a dictionary using the original corpus to stem which was used to complete the words after removing sparse terms, and finally, tokenizing the corpus using the DocumentTermMatrix function. This whole process was done to construct a term frequency matrix (TDF), and the same process was repeated for a term frequency-inverse document frequency (TF-IDF) matrix except for the last step (weighting more on the importance of the words instead of their frequency).

**Chart, bar chart

Description automatically generated**

Fig.8

*3.1.3 Modeling Process*

The preprocessed dataset was split into two sets, train (70%) and test (30%). We chose to use one regularized regression model (Lasso) and one gradient boosting model (XGBoost). We thought that it would be a good idea to compare the performance between a parametric and non-parametric model on the text-based data. We selected Lasso because the number of features increased after tokenization, and it needed to be smoothed out to prevent overfitting. Moreover, we chose to use XGBoost because it performs better than any other ensemble models when appropriately tuned.

**3.2 Time-Series Analysis**

While the main focus of our analysis was in text mining, several different time series models were used to predict Bitcoin price changes.  However, before conducting a time series analysis, we formatted the data into a time series object.  By using the TSstudio and XTS packages in R, the daily Bitcoin price change percentages were formatted into a time series object by setting the frequency to 365. We made several assumptions with time series analysis; most importantly, we assume constant variance with the response variable.

The following seven models were used to conduct time series analysis:

1. Average Model
2. Naive Model
3. Holt Model
4. Holt Damped Model
5. Drift Model
6. Auto ARIMA Model
7. SES Model

The dataset contained 1,304 days of Bitcoin price.  The data was split into a train and test sample, with the test sample containing the final 30% of the data (391 days).

The Average Model was used as it accounts for short-run autocorrelation.  Essentially, it suggests that the next day's price is the average of the historical prices.  The Naive model forecasts prices based exclusively on historical observations of price.  Specifically, the Naive model does not consider any underlying causal relationships with price.  Moreover, the Holt model, which turned out to be the best model to predict Bitcoin price, which uses linear exponential smoothing to forecast Bitcoin prices.  The SES model (simple exponential smoothing) uses the weight averages, where the weights decrease exponentially with the most recent observation receiving the heaviest weight.  This model is appropriate for forecasting Bitcoin prices as there is no clear trend or seasonal pattern like Bitcoin price. Finally, the Auto Arima model, (Autoregressive integrated moving average) combines unit root tests, auto minimization of AICc and MLE. To evaluate the different time series models, each model went under accuracy analysis by analyzing performance metrics such as RMSE and MAE.

**4. Results**

**4.1 Text Analysis**

In Fig. 9 below, the accuracy metrics for each model are displayed.

Graphical user interface, table

Description automatically generated

**4.2 Time Series**

In Fig. 10 below, the accuracy metrics for each model are displayed.  The Holt Model had the smallest RMSE and tied with the Average Model, Auto Arima, and SES Model for the smallest MAE.  Graphical user interface, application, table, Excel

Description automatically generated

In Fig. 11 below, the seven different model forecasts are represented.  We observe the Holt Model as the only model projecting a strong increase in Bitcoin price.

Chart, line chart, histogram

Description automatically generated

Fig. 11

**5. Discussion**

Our models were unable to predict the massive increase in the price of bitcoin in late 2020 and early 2021. Given the truncated period of our dataset, as well as factors contributing to the speculative nature of people purchasing Bitcoin, this is not surprising.

The results from the text analysis of news headlines show that while there is a connection between term frequency and price of bitcoin, there is even more analysis that can be done around specific terms. Future research can explore specific terms on various sources such as Twitter.

The goal of the time series analysis was to utilize various mathematical models to predict future Bitcoin price.  After incorporating seven different models, the Holt’s Model had the lowest RMSE of all the models tested.  Although Holt's Model RMSE was ~0.0419, this doesn’t mean this method will accurately predict future Bitcoin prices.  From working with Bitcoin time series data, it is clear that there is not a clear, reliable time series model to accurately predict Bitcoin price.  The nature of Bitcoin is extremely volatile and historically there have been multiple occurrences of daily price changes of over 20% (Fig. 1).

Overall, Bitcoin seems like it is here to stay. Intrinsically, the rapid increase in price and retainment of value are strong arguments but the growing activity in mining and transaction metrics are some of the strongest reasons for why Bitcoin is a viable project. Next, growing institutional support and acceptance are paramount for future growth of the network. Moreover, future analysis could analyze the effect of institutional buy-in on the price of Bitcoin. As landmark developments such as the first Bitcoin ETF are possibly rolled out in 2021, future studies would have necessary data to analyze this type of connection. Other important events include regulation extension and creation. Lastly, there is an influx of intellectual and financial resources into the network. While our analysis focuses on Bitcoin, which was developed to be digital cash, there are thousands of other crypto projects with varying use cases and dozens with market capitalizations in the billions of dollars.

**Citations**

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review,* *1*(3).

Bouoiyour, J., Tiwari, A. K., Selmi, R., & Olayeni, O. R. (n.d.). What drives Bitcoin price? *Economics Bulletin,* *36*(2).

Bouoiyour, J. and Selmi, R. (2015) “What Does Bitcoin Look Like?” Annals of Economics and Finance 16 (2), 449–492.

Bouoiyour, J., Selmi, R. and Tiwari, A-K. (2015) “Is Bitcoin Business Income or Speculative Bubble? Unconditional vs. Conditional Frequency Domain Analysis” Annals of Financial Economics 10 (2), 1–23.

Buchholz, M., Delaney, J., Warren, J. and Parker, J. (2012) “Bits and Bets, Information, Price Volatility, and Demand for Bitcoin” Economics 312.

Kristoufek, L. (2013) “Bitcoin meets Google Trends and Wikipedia: Quantifying the

relationship between phenomena of the Internet era” Scientific Reports 3 (3415), 1–7.

Schilling, L., & Uhlig, H. (2018). Some simple bitcoin economics. *Journal of Monetary Economics*. doi:10.3386/w24483