Bitcoin Price Prediction through Regression Modeling

Amala Deshpande
Computer Science, Department,
NYU Courant
New York City, USA
asd508@nyu.edu

Anshu Tomar
Computer Science, Department,
NYU Courant
New York City, USA
at3769@nyu.edu

Nikita Bhargava
Computer Science, Department,
NYU Courant
New York City, USA
nb2643@nyu.edu

***Abstract*—**

In the recent years, cryptocurrencies especially the Bitcoin have gained a lot of attention from consumers and businesses alike. We aim to provide a prediction model built using the coalesce of tweets centering around bitcoin, "bitcoin" keyword searches over the Google engine and bitcoin pricing in the past year. As per our analysis the trends in Twitter and Google Search data reflect the trends in Bitcoin prices. A regression model is generated using Tweets and Google Searches to predict the future trends in bitcoin pricing. Time lagged data was also used to predict the future prices. The model generated is able to predict the values of bitcoin price with 81.55 percent accuracy. This prediction model will be beneficial to predict the rise and fall of bitcoin price in the coming days.

Keywords—analytics, cryptocurrency, bitcoin, bitcoin transactions, twitter, price prediction, machine learning, linear regression modeling

I. INTRODUCTION

The bitcoin price trends and market are analogous to stock price and market trends. And just as in the field of stock price prediction, many machine learning and predictive algorithms are being deployed to effectively predict the rise and fall in stock prices, it is a logical that such algorithms be leveraged to predict the trends in bitcoin pricing. The bitcoin price prediction can be difficult as it is relatively new concept and hence has comparatively less historical data needed for predictive analysis. The rise of social media analytics has given us various methods to predict future changes based on social media content. One of the social media platforms that is being used extensively for analytics is Twitter. The volume of Tweets about a particular topic can give the sense of popularity of that topic among the users. Though "bitcoin" is a relatively new term, its popularity on Twitter has been rising over the last few years. Hence, Twitter platform can be exploited to derive insights about bitcoin. Machine learning techniques like Bayesian Regression and Correlation techniques have been studied in past to predict bitcoin price change. The price prediction can also lead to developing trading strategies for buying/selling of the bitcoin. The technique of Sentiment Analysis can be useful to know emotions of the people about bitcoin. Such an analysis can be done on Twitter Data and the results can be correlated with bitcoin price change.

In the paper, we try and use two features which are probable to image whether the bitcoin price is likely to increase or decrease in the coming days, to generate a predictive model. First feature is Tweets by users all around the world centering around bitcoin on a particular day. This is because the number of tweets will mirror the interest in the bitcoin technology on that day. Similar reasoning is used for using the second feature which is the number of times users searched the keyword "bitcoin" on the Google Engine. The time series graphs of these two features indicated similar trends to the Bitcoin Price variation with time. Hence, developing a prediction model based on Tweets and Google Searches gave satisfactory accuracy to the model.

Using Tweets and Google Searches as features and bitcoin price on the corresponding day as label, we ran linear regressions and came up with a prediction model. The model was generated up to a lag of three days and the accuracy of the prediction model on an average remained same. The model generated could predict the next day's Bitcoin Price, given the number of tweets and google searches of up to three days before. We incorporated the idea of lag in days as the accuracy was found to be best for lag of three days. So to predict the price for next day in our model we have taken as input the Number of Tweets and Number of Google searches of three days before.

The code for model generation and prediction is written in Scala. The Machine Learning Library of Apache Spark contains various Machine Learning Algorithms. For our purpose of model generation we used Multiple Linear Regression. This algorithm attempts to model a relationship between one dependent or outcome variable and two independent or predictor variables, by fitting a linear equation to the observed data. This equation can be used to calculate the value of dependent variable, given the corresponding independent variables. The model was tested and tend further by testing it on 1 year worth of data. The results obtained by from testing were also studied graphically to observe the accuracy. The accuracy for the model generated with three day lag data was found to be best.

II. MOTIVATION

The Bitcoin industry and cryptocurrency exchange markets are growing at an exponential rate. They are being used as "digital gold" by millions of investors keen on exploiting the good returns scenario offered by these cryptocurrencies. CoinBase which is one of the leading secure online platform for buying, selling, transferring, and storing digital currencies added 1 million bitcoin users last month. Furthermore, The Wall Street Journal reported there to be around 300,000 bitcoin transactions per day. All these point towards Bitcoin having an exponential growth. Users of bitcoin are becoming more and more interested in buying bitcoins and trading them at the right time to yield high profits. Thus, there is a growing demand for tools and sources which can accurately predict the future bitcoin price. Consequently, there is a growing need for efficient algorithms which can help analyze trends in bitcoin pricing and potentially predict future pricing. A prediction tool that can achieve this prediction with high accuracy will surely be in high demand among the bitcoin user base.

III. RELATED WORK

In [1], Satoshi Nakamoto introduces the concept of Bitcoin - "A Peer-to-Peer Electronic Cash System" which essentially is introduced as a peer-to-peer distributed time-stamp server that generate computational proof of the chronological order of publicly maintained transactions. The public history/logs is impractical for an attacker to change as long as majority of CPU power is controlled by honest nodes. These nodes work with little coordination with each other and can leave and rejoin the network any time as long as they update their blockchain upon re-entering the network. The system solves the problems of third-party involvement in transactions and that of double-spending (i.e. risk that a digital currency can be used twice) by proposing a peer-to-peer network that records a public history of transactions using proof-of-work and hashing. It has been studied that there is a direct relationship between trading volumes of Bitcoin currency and volume of queries about bitcoin on search engine.

In [2], Martina Matta et al studied Google web search data about bitcoin and investigated if it can be helpful in anticipating trading volumes of bitcoin currency. They found that when the Pearson's correlation is applied on trade volume data and search data, the result equals 0.6. Furthermore, when a time lagged Pearson's cross correlation is applied between search data and bitcoin trade data, the results showed that maximum correlation was for positive delays than negative delays and particularly +3 days(0.68). Through this, they concluded that search data can predict bitcoin trade volumes in 3 days. In [3], a profitable quantitative strategy is developed to buy/sell bitcoin. This was achieved by predicting future bitcoin price change from historical bitcoin data. Bayesian Regression technique inspired by latent source model was used for this purpose. The quantitative strategy was built by assuming that past price movements can be used to predict future changes. At each time instance, average price movement (Δp) over 10 seconds interval was predicted using Bayesian regression. The trading strategy was based on a comparison of Δp with a threshold. Hence, it can be seen that bitcoin price change can be predicted from bitcoin historical data. Also, trading strategies can be developed if future bitcoin price change is predicted. Trading strategies can directly help the traders in important decision making.

Sharing knowledge is a vital part of learning and enhancing skills. Among other mediums of sharing knowledge and information, social media has gained a lot of popularity in recent years. The data produced by social media acts as a collective indicator of thoughts and ideas regarding every aspect of the world. Twitter, an online social networking website and microblogging service, has become an important tool for businesses and individuals to communicate and share information with a rapid growth and significant adoption. Twitter data has proved to be beneficial in analyzing public opinions and views on a number of activities/events going on around the world. Twitter has gained tremendous popularity as a source of information for development and research in various domains also. The data in form of text received from Twitter feeds is being used in various kinds of data analytics projects. [4] attempts to analyze the impact of online social media on Bitcoin price and whether the result could be used by investment professionals. Volume of daily tweets on bitcoin and user's opinion about bitcoin were used as two factors that could influence bitcoin price. The study aims to discover if the social chatter can be used to make qualitative predictions about Bitcoin market, attempting to establish whether there is any correlation between the volume of tweets and bitcoin price, and between sentiment of tweets and Bitcoin's price. The results suggested a significant relationship with future Bitcoin's price and volume of tweets exists on a daily level. [5] attempts to perform Sentiment Analysis on the publicly available data through Tweeter's APIs to calculate the Gross National Happiness (GNH) of a Middle East country, Turkey. The results of the study were compared with the survey results published by Turkish Statistical Institute in previous year. Both methods yielded similar outcomes when the results for the whole country were taken into account.

There are various other cryptocurrencies introduced in the market today, though Bitcoin maintains being the most popular. One such cryptocurrency is Namecoin. In [6], Tao-Hung Chang and Davor Svetinovic have carried out an analysis of Namecoin and done an extensive comparative analysis of Namecoin and Bitcoin. Namecoin follows the structure of bitcoin but unlike bitcoin. it can store data within its own blockchain transaction database. It is also the first cryptocurrency that can act like a DNS server. Namecoin like Bitcoin has a limit of 21 million Namecoins. Furthermore, the essential differences included the fact that unlike Bitcoin Networks that seem to follow the Densification Law through all the years of their existence, the Namecoin network only followed the Densification Law in its first introductory year. This is a good indicator that bitcoin will remain the most popular cryptocurrency in the future as well even if the market sees a rise of other competitive cryptocurrencies.

IV. DESIGN

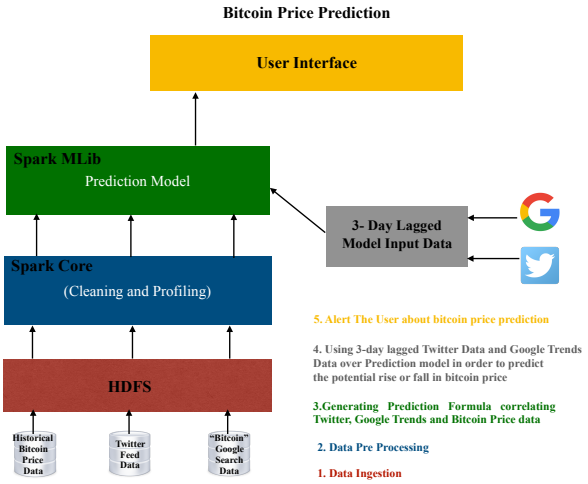

Figure 1: High Level Design Flow Diagram

The design model captures the flow of the analytics application which is divided into 4 phases: (1) Data Ingestion (2) Cleaning and profiling of the collected data (3) Prediction Model Generation 4) GUI development to display the Result.

At the Data Ingestion phase , the data is taken from the three data sources namely Twitter[3], Google Trends[4] and CoinDesk[5]. [3] gives the total number of daily tweets related to bitcoin. [4] provides the total number of google searches on bitcoin. [5] provides daily bitcoin price. In second phase, the raw data is cleaned, profiled, converted into the desired format in Apache Spark environment. The third phase, Prediction Model Generation, uses Linear Regression algorithm provided by the Spark MLlib[3] to generate the prediction model. A scatter plot was used to determine the strength of the relations between (i) bitcoin price and number of tweets , and (ii) bitcoin price and number of google searches. The correlation found between them encouraged us to use this model. After third phase, the model becomes ready to predict potential rise/ fall in next day's price of bitcoin using today's number of tweets and google searches related to bitcoin. To send this result to the user of the application, a GUI is designed in the final phase of the application development where the result will be output for use (to help in decision making whether or not to buy a bitcoin). For the present work, the output screen would be the on the desktop running the model which can be extended further to connect users mobiles/ iPads etc.

V. EXPERIMENTS

This section describes the steps taken to develop the Bitcoin Predictor Application. Firstly, the Bitcoin, Twitter and Google Trends data was collected from different sources. This data was cleaned and formatted as per the needs of our application. Then analysis was done to find trends amongst different data parameters. Further, the results of analysis were used to generate prediction model using Linear Regression. The model generated could competently predict the price change in Bitcoin.

A. Data Collection

In the Data Collection phase we have collected the datasets from different online sources in comma separated value format. All three datasets were collected for the time period December 2016 to November 2017.

(i) Bitcoin Data

Blockchain.info is an online system that provides detailed information about Bitcoin market. Launched in August 2011, this system shows data on recent transactions, plots on the Bitcoin economy and several statistics. It allows users to analyze different Bitcoin aspects like Total Bitcoins in circulation, Number of Transactions, Total output volume, USD Exchange Trade volume, Market price (USD) []. We decided to use this website to collect Bitcoin Price data and Number of daily Bitcoin Transactions for the above mentioned time period.

(ii) Google Search Data

Google Trends is a public website based on Google Search. It shows how often a term is searched relative to the total number of searches across the world. The x-axis of the graph represents the date and y-axis represents the number of google searches on a particular date relative to the total number of searches globally. We collected number of daily Google searches regarding Bitcoin from this website for the above mentioned time period. The Google trends does not give the data in form of actual number of searches but a normalized form of data. In a time period, the day with maximum number of google searches is given the value 100 and rest of the days are given values correspondingly. We normalized the data again as per our time period need.

(iii) Twitter Data

The Twitter Data was collected from the website *bitinfocharts.com*. This data gave the number of tweets about bitcoin for one year.

B. Data Formatting

The collected data was loaded into HDFS and formatted using following steps:

- The date field was normalized to be in one format for all three datasets.
- From the Twitter Dataset number of tweets for each day were calculated.
- The google data was normalized by bringing the entire date to a scale of 1000 with the day having the maximum tweets set to 1000
- The bitcoin price was Time Lagged to use the current day data to predict next day's Bitcoin price.

• The formatted datasets were joined on Date column and resulting dataset contained Date, Actual Bitcoin Price, Time Lagged Bitcoin Price, Number of tweets, Number of Google Searches and Number of Bitcoin Transactions.

C. Analysis

In this work, we considered three features that could affect the Bitcoin price. The features are number of Tweets about Bitcoin, number of Google Searches about Bitcoin and number of Bitcoin Transactions per day. To understand the correlation among these features and price of Bitcoin, scatter plots were used. The plots demonstrated fair positive correlation between number of tweets , number of google searches and Bitcoin's price. However, the price and number of bitcoin transactions did not fit well into each other. Figure 2 below shows no correlation between Number of Transactions and bitcoin price. Thus, we concluded that this feature cannot be used for bitcoin price prediction.

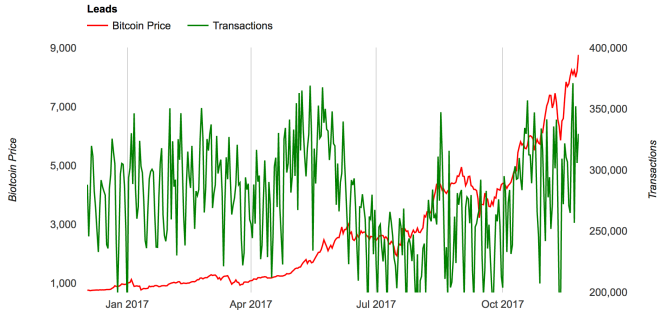

Figure 2: Bitcoin price vs Transactions

Figure 3 shows that the Bitcoin Price increases steadily along with the number of tweets, except for a few outliers. Thus, we chose Twitter data as one of the features that could prove useful in price prediction.

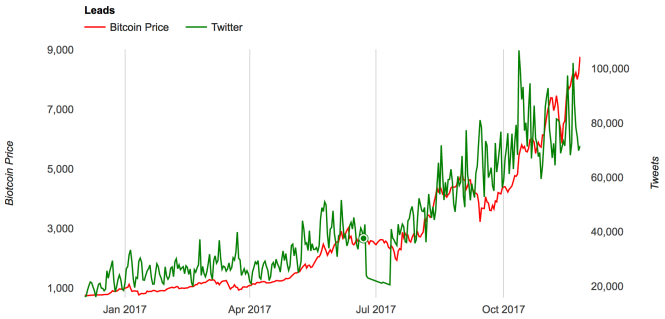

Figure 3:Bitcoin Price vs Tweets

Figure 4 also shows positive correlation between Bitcoin Price and Google Searches. Along with few outliers, the Bitcoin Price is in correspondence with the number of Google Searches for most of the data points.

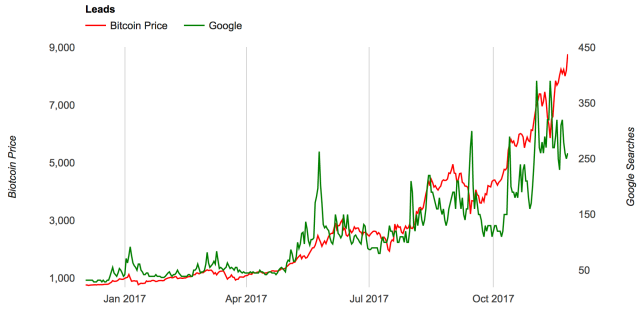

Figure 4: Bitcoin Price vs Google Searches

The above graph also shows positive correlation between Bitcoin Price and Google Searches, though there are there are outliers, but for most parts the Bitcoin Price change is in correspondence with Google Searches

D. Model Generation

Apache Spark provides a number of Machine Learning algorithms to use within its environment. We used Multiple Linear Regression model for the prediction model generation. Linear regression follows the approach for modeling the relationship between a dependent variable and one or more independent variables. Depending on the nature of the collected data, this model fitted best for the prediction. The RDD (Resilient Distributed Data) of the collected data was split into a Labeled Point and Vector of features where price being the labeled point and number of tweets and google searches being the features. The model was generated with two approaches. In one approach the intercept for the model was taken to be false. In other approach the intercept was true. The approach with intercept generated better model.The algorithm was run using step size of 0.0001 and hundred iterations and calculated the intercept and weights of the features. This generated model was tested on different data points to train it further.

E. Results

The model was generated in three ways- with a lag of one day, two days and three days. The Bitcoin prices were predicted for a year by the model generated. The Bitcoin price predicted by the model was compared with the actual bitcoin price. The result shows a positive correlation between the actual bitcoin price and predicted bitcoin price for all three ways. Hence, the model yielded satisfactory results. We used the three day lag data for model generation and prediction as that gave better accuracy.

The exact characteristics of the model are :
weights: [w1 - 11.97, w2 - 20.054],

intercept: 1.076 which can be represented in mathematical formula as-

$$Bitcoin\ Price = w1 * NumberOfTweets + w2 * NumberOfGoogleSearches + intercept$$

Hence, given the Number of Tweets and Google Searches for a particular day, the model can predict the bitcoin price for the next day.

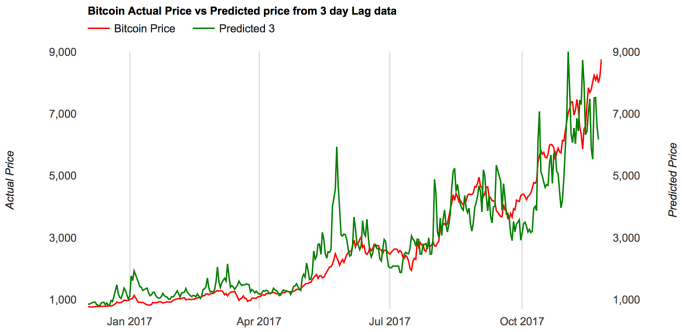

Figure 5: Actual Bitcoin Price vs Predicted Bitcoin Price

Figure 5 represents a graph between the actual bitcoin price and predicted price with 3 day lag for last one year. The graph indicates a positive correlation.

Date	Actual Bitcoin Price	Tweets	Google Searches	Bitcoin Lagged Price by 3 days	Predicted price calculated with 3 day lag
12/3/16	764.97	16186	32	758.81	846.34
12/4/16	766.46	16345	32	763.9	848.32
12/5/16	750.71	18558	32	766.75	875.85
12/6/16	758.81	20236	32	770.41	896.73

Figure 6: Table Summary of the data

The table in Figure 6 gives a snapshot of the data used. The 5th column depicts the time lag in bitcoin price by three days. The regression model was generated using the 5th column as the label and the tweets and google searches as features. For the above data the predicted values were calculated from the model generated as shown in the 6th column The accuracy was calculated by comparing the Predicted price (6th column) and the Actual Bitcoin Price on that day (2nd column)

The formula used for accuracy calculation is -

$$Accuracy = \sum_{12/3/17 < t < 11/25/17} 1 - (|(A)_t - (P_t)| \ / \ A_t)$$

where,
A_t *- Actual Predicted price on date t*
P_t*- Predicted price from model with lag of 3 days for date t*

This formula gave the accuracy of around 81.55 percent when calculated for predicted values from the three models considered. Further the model was found to predict rise and fall in the bitcoin price accurately for the test data. Hence, though there could be other factors that effect the Bitcoin price but Tweets and Google searches can be used as satisfactory predictors for the Bitcoin Price rise or fall.

VI. CONCLUSION

In this paper, we attempted to predict bitcoin price on a particular day based on the number of tweets regarding bitcoin and number of google searches with keyword “bitcoin” two days back. We were able to successfully predict whether there will be a rise or fall in the bitcoin price on a particular day with a 81.55% accuracy rate. The difference in price can be accounted to the fact that we need to segregate the positive and negative tweets and google searches and then use only the positive tweets and/or google searches as features for predicting a more accurate future price trend for bitcoin. We observed that there is a direct correlation between the frequency of tweets and google searched relating to bitcoin and the bitcoin price by running regressions on these aforementioned datasets spanning a year. We exploited this fact to get out result. An important observation was that there is no correlation between bitcoin price and number of bitcoin transactions.

Our application essentially leveraged Multiple Linear Regression machine learning algorithm to construct a prediction model which takes in absolute number of bitcoin related tweets and scaled number of google searches about bitcoin and outputs the predicted bitcoin price against dollar. In addition to the above mentioned features, we experimented with other features as well - including number of bitcoin transactions per day. An important observation was that there is no correlation between bitcoin price and number of bitcoin transactions. Hence, we concluded to not use this as a feature for out price prediction model. In conclusion, our application can effectively predict the future rise and fall in bitcoin prices and can prove to be useful to our user base of bitcoin traders.

VII. FUTURE WORK

In our work we used volume of Twitter and Google Search data to predict the change in Bitcoin Price, which could generate a model that predicted change in bitcoin price. Sentiment Analysis of tweets could be used to further enhance the accuracy of the model. As we could prove that Twitter data correlates with the Bitcoin Price variation, if sentiments of the Tweets are considered then Bitcoin Price can be predicted with more accuracy.

References

1. Satoshi Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, October 2008
2. Martina Matta, Ilaria Lunesu, Michele Marchesi, The Predictor Impact of Web Search Media on Bitcoin Trading Volumes, August 2016
3. Devavrat Shah, Kang Zhang, Bayesian regression and Bitcoin, October 2014
4. Martina Matta, Ilaria Lunesu, Michele Marchesi, Bitcoin Spread Prediction Using Social And Web Search Media , June 2015
5. Ahmet Onur Durahim, Mustafa Coşkun, #iamhappybecause: Gross National Happiness through Twitter analysis and big data, October 2015
6. Tao-Hung Chang and Davor Svetinovic, Data Analysis of Digital Currency Networks: Namecoin Case Study, 2016
7. Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, Learning Spark : Lightening Fast Data Analysis, February 2015
8. Tom White, Hadoop: The Definitive Guide, April 2015