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 has gained a lot of attention from consumers and businesses alike. We aim to provide a prediction model built using the coalesce of tweets centring around bitcoin, number of bitcoin transactions, "bitcoin" keyword searches over the Google engine, number of unique bitcoin addresses generated and bitcoin pricing in the past year. The model uses regression and machine learning algorithms to predict the future trends in bitcoin pricing. < TODO: add regarding the experiment steps and level of accuracy acheived>

Keywords—analytics, cryptocurrency, bitcoin, bitcoin transactions, twitter, sentiment analysis, price prediction, machine learning, linear regression modelling

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. In the paper, we try and use various features which are likely to image whether the bitcoin price is likely to increase or decrease in the coming days. First feature is Twitter tweets by users all around the world centring 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 "bitcoin" on the Google Engine. Based on our research study, we found that the number of bitcoin transactions and number of unique bitcoin addresses generated on a particular day also seem to have an impact on bitcoin prices and these are our third and fourth features. Using these features and actual bitcoin pricing data on the corresponding days for validation, we ran linear regressions and came up with a prediction model. The model was tested and tend further by testing it on 6 months worth of data. The final result was a price prediction for the next day.

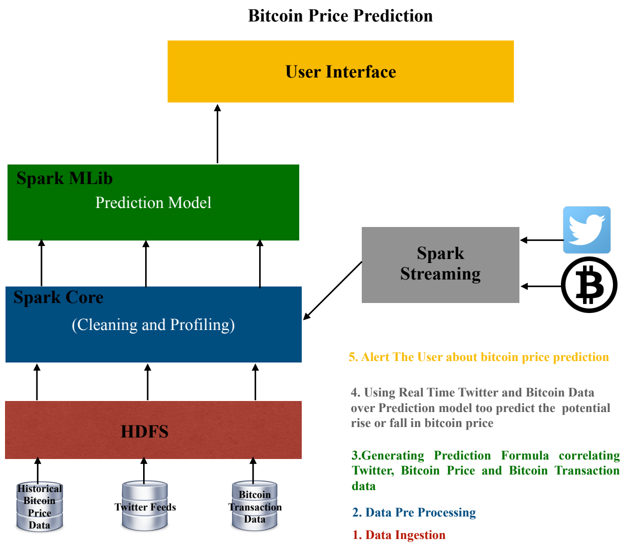
II. Motivation

The Bitcoin industry and cryptocurrency exchange markets are growing at an exponential rate. There is a growing need for efficient algorithms which can help analyse 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. <TODO: Add regarding: why twitter data, google trends data is being leveraged for bitcoin price prediction>

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. There is a direct relationship between trading volumes of Bitcoin currency and volume of queries about bitcoin on search engine has been studied. 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. This means that trading volume follows the same direction as the query search on google. Furthermore, when a time lagged Pearson's cross correlation is applied between search data and bitcoin trade data, a correlation for maximum lag of 5 days is evaluated. 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. They validated these results by applying the Granger causality test on the data and concluded that search data is a good predictor of trade volume. In the recent years, Twitter has gained tremendous popularity as a source of information for development and research in various domains. The data in form of text received from Tweeter feeds is being used in various kinds of data analytics projects. Sentiment Analysis is one of them. It can be defined as a kind of text analysis for determining the emotional tone behind a series of words, used to gain an understanding of the the attitudes, opinions and emotions expressed within an online mention. [3] presents the details of 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.

IV. DESIGN


Bitcoin Price Prediction
User Interface
Spark MLib
Prediction Model
Spark Streaming
Spark Core
(Cleaning and Profiling)
HDFS
Historical Bitcoin Price Data
Twitter Feeds
Bitcoin Transaction Data
5. Alert The User about bitcoin price prediction
4. Using Real Time Twitter and Bitcoin Data over Prediction model too predict the potential rise or fall in bitcoin price
3.Generating Prediction Formula correlating Twitter, Bitcoin Price and Bitcoin Transaction data
2. Data Pre Processing
1. Data Ingestion

The design model captures the flow of the analytics application which begins at the Data Ingestion stage where the data is taken from the three data sources which are Twitter, Google Trends and CoinDesk and stored into the HDFS. Next, this raw data is cleaned and profiled and converted into the desired format which is used to compute our bitcoin pricing prediction model. For the model formulation, Spark MLib is being used and linear regressions run on all the data sets. This prediction model is then applied on the real time data which is obtained using Spark Streaming. The output of the prediction model gives the predicted bitcoin price for the next day. This value will be passed to the user interface layer which will send a message to the application user's smartphones alerting them about the increased or decreased bitcoin price for the coming day.

V. EXPERIMENTS

TODO (In this section, you can describe: Your experimental setup, problems with: data, performance, tools, platforms, etc. Discuss your experiments, describe what you learned. Discuss limitations of the application. Discuss what you would do to expand it given time - how would you improve it, etc.)

VI. CONCLUSION

TODO (One paragraph about the value, results, usefulness of your application.)

ACKNOWLEDGMENT

TODO (This section is optional. It can be used to thank the people/companies/organizations who have made data available to you, for example. You can list any HPC people who were particularly helpful, if you used the NYU HPC. List Amazon if you used an Amazon voucher.)

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. Ahmet Onur Durahim, Mustafa Coşkun, #iamhappybecause: Gross National Happiness through Twitter analysis and big data, October 2015
4. Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, Learning Spark : Lightening Fast Data Analysis, February 2015
5. Tom White, Hadoop: The Definitive Guide, April 2015
6. < TODO: Websites >