Pediction of Bitcoin price using 3 different approaches : Linear Regression, Auto-regressive model & Long Short Term Memory(LSTM) Neural Networks

V K Dingu Sagar, Deepak C, Subin Sahayam

***Abstract*— The paper tries to compare 3 different models for bitcoin prediction which includes the following : linear regression, auto-regressive model & LSTM neural networks. The advantages of using different models and their limitations are discussed. The scope of implementing them in Map-Reduce model to achieve better speed is also discussed.**

I. INTRODUCTION

Bitcoin is a decentralized digital currency that uses cryptographic protocol. It is not bound or backed by any government and works on a peer-to-peer system. Bitcoin is the worlds most valuable cryptocurrency introduced following the release of a white-paper published in 2008 under the pseudo name Satoshi Nakamoto. Bitcoin is very different from traditional financial markets. It operates on a decentralized, peer-to-peer and trustless system in which all transactions are posted to an open ledger called the Blockchain. This type of transparency is unheard of in other financial markets. Recently, there has been a lot of other crypto-currency (alt-coins) introduced to compliment bitcoin and what its trying to achieve

II. ABOUT BLOCKCHAIN TECHNOLOGY

A. The power of decentralization

Decentralization is the value pursued by all crypto-currencies as opposed to general fiat currencies being valued by central banks. Decentralization can be specified by the following goals: (i) Who will maintain and manage the transaction ledger? (ii) Who will have the right to validate transactions? (iii) Who will create new Bitcoins? The blockchain is the only available technology that can simultaneously achieve these three goals. Generation of blocks in the Blockchain, which is directly involved in the creation and trading of Bitcoins, directly influence the supply and demand of Bitcoins. Combination of Blockchain technologies and the Bitcoin market is a real-world example of a combination of high-level cryptography and market economies.

B. Maintaining the Security

A participant in a Bitcoin network acts as a part of a network system by providing hardware resources of their own computer, which is called a ”distributed system”. All issuance and transaction of money are conducted through P2P networks. All trading history is recorded in the Blockchain and shared by the network, and all past transaction history is verified by all network participants. The unit called ”block”, which includes recent transactions and a hash value from the previous ”block”, creates irreversible data by a hash function, and is pointed out from the next block. Figure 1 shows the general structure of Blockchain. It takes more than a certain amount of time to generate the block to make impossible to forge all or part of the Blockchain. This algorithm is called proof of work (PoW), and the difficulty is automatically set to ensure that the problem can be solved within approximately 10 minutes. PoW also provides incentives to motivate participants to maintain the value of Bitcoin by paying Bitcoin for the participant who created the block.

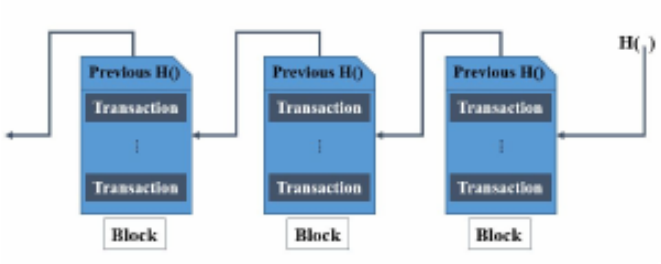

Fig. 1. Block chain

III. DATASET

Before we build the model, we need to obtain some data for it. There are many ready made datasets that details minute by minute Bitcoin prices for the last few years . Over this timescale, noise could overwhelm the signal, so well opt for daily prices.

We fetch the day to day live dataset from coinmarket-cap.com starting from 28-04-2013 till 30-04-2018.

A. Features used

The dataset contains the following columns:

- Date
- Opening Price
- Closing Price
- High
- Low
- Market Capital
- Volume

B. A Quick Plot

A quick plot of the bitcoin closing prices over different dates is shown in Figure 2

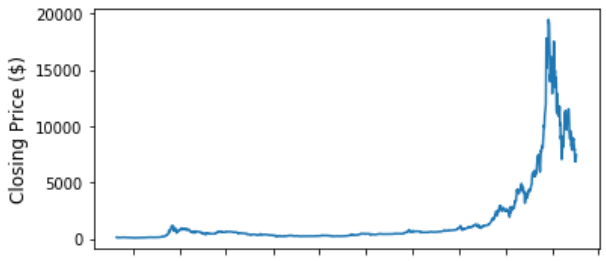

Fig. 2. Bitcoin price plot.

IV. Linear Regression

A. Brief Background

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

B. Using Linear Regression to predict Bitcoin price

A simple one dimensional linear regression model was implemented and the predictions were evaluated. But the result was very poor. As it can be seen in Figure 3, the best fitting line does not accurately represents the entire data set.

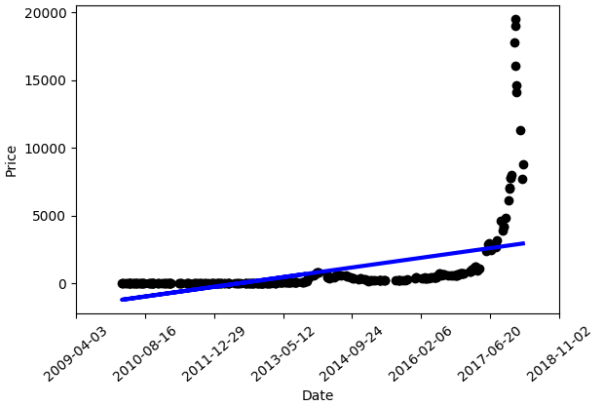

Fig. 3. Best fitting line of linear regression .

C. Limitations of this model

Linear regression tries to model a linear relationship between the dependent variables and the predicting variable.

- The exponential hike of bitcoin happened recently. Most of the training data is from the early days of bitcoin when the price was low and growing very slowly. This pollutes the model to do predictions reflecting the recent trends.
- Only one dependent variable is used i.e., Date. This may not be the only causative feature for bit-coin prices.

V. Auto-regressive model

A. Background

Auto-regressive models are based on the time series analysis. They are mainly used in stock market predictions. They are backed by the random walk hypothesis.

The random walk hypothesis is a financial theory stating that stock market prices evolve according to a random walk (so price changes are random) and thus cannot be predicted. It is consistent with the efficient-market hypothesis.

Figure 4 shows the equation of auto-regressive model. Here price at time t is predicted by the linear combinations of previous prices at t-1,t-2..etc. A stochastic noise is added to the equation to simulate random noise.

$$PredPrice_t = \phi_0 + \phi_1 * Price_{t-1} + \ldots + \phi_p * Price_{t-p} + \epsilon_t,\ \epsilon_t \sim N(0, \sigma)$$

Fig. 4.

All the constants in the equation along with the mean and standard deviation of the stochastic noise term is calculated using the training data.

B. Using this model for bitcoin prediction

Before using this model, we need to check if the daily price changes of bitcoin follow any known distribution. As shown in the Figure 5, they follow a normal distribution. Therefore we can use a normal distribution as the stochastic noise term.

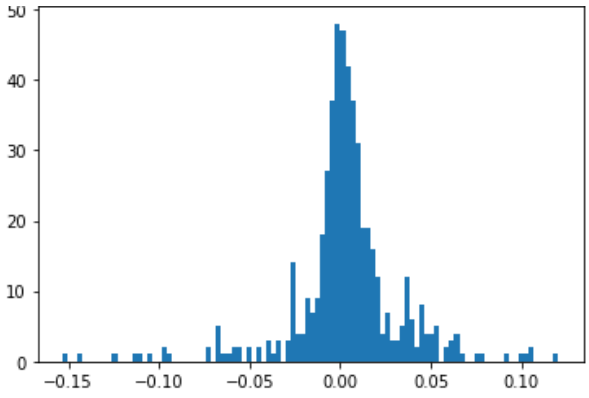

Fig. 5. Bitcoin price changes follow a normal distribution

C. Result of the model

The auto-regressive model is able to capture the general trends of bitcoin prices. This is because the model uses previous prices as inputs to predict the future price.

This is a great improvement from linear regression model. But this is not close to perfect prediction. The graph may be misleading because of the y axis scale being very big. The root mean square error calculate for this model is 633.76.

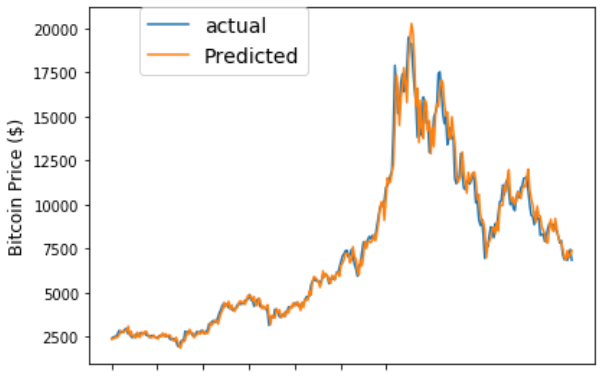

Fig. 6. Using auto-regressive model to predict bitcoin

VI. LSTM NEURAL NETWORKS

A. Background

Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for ”remembering” values over arbitrary time intervals; hence the word ”memory” in LSTM. Each of the three gates can be thought of as a ”conventional” artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM; hence the denotation ”gate”. There are connections between these gates and the cell.

The expression long short-term refers to the fact that LSTM is a model for the short-term memory which can last for a long period of time. An LSTM is well-suited to classify, process and predict time series given time lags of unknown size and duration between important events.

B. Using LSTM Neural Networks to predict bitcoin prices

Preprocessing Steps

- Removing Open price,daily highs and lows from the data set
- Removing date from dataset (we no longer need the literal date since we are giving the previous outputs as inputs)
- Adding new column : close-off-high. It represents the gap between the closing price and price high for that day, where values of -1 and 1 mean the closing price was equal to the daily low or daily high, respectively
- Adding new column : volatility. It is the difference between high and low price divided by the opening price.

The window size of 10 is selected for the LSTM network. This means to get the price of the next day, we need to give previous 10 days prices as input to the neural network.

C. RESULT OF LSTM NEURAL NETWORK

The LSTM neural network was trained on the training set and evaluated on the test set.

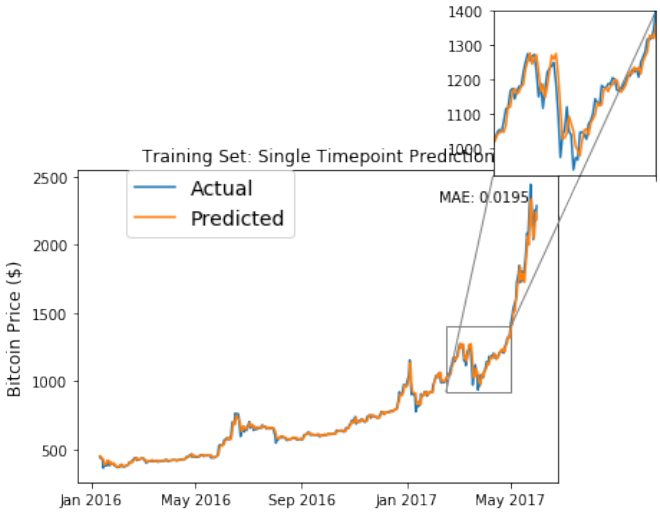

Fig. 7. LSTM on training dataset

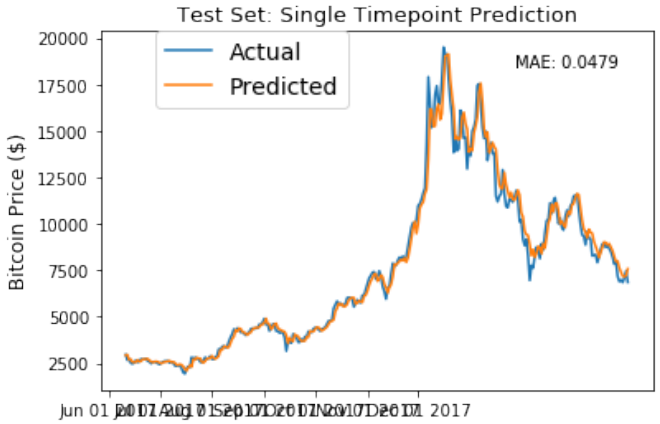

Fig. 8. LSTM on test dataset

VII. SOURCE CODES

All sequential codes are done using python 3 with the help of jupyter notebook IDE. Linear regression is implemented in the map reduce model. The LSTM model can be easily implemented using the ready made distributed version of keras library. All Source codes are available at the following github link:

https://github.com/dingusagar/Data-Mining-Project—Bitcoin-Prediction

VIII. CONCLUSIONS

We experimented three different machine learning models and saw the limitations of different models in predicting the price of bitcoin. Bitcoin is really volatile these days and it is very hard to come up with a perfect model for price prediction.

A. *Comparing different models*

- Linear Regression : linear regression does a very poor job. The best fitting line is not able to capture the price trends of bitcoins. One scope of improvement is increasing the number of features.
- Auto-regressive model : It does a decent job. The RMSE is 633.76. These types of models are suited for time-series analysis because they use the previous trends to predict the future trend.
- LSTM Neural Networks : This is an expensive deep learning model. But it does a pretty good job for this problem. LSTMs are popular because of their ability to remember important short-term trends and forget unimportant trends over a long period. The Mean absolute error is found to be 0.0479 which is great. But the main limitation here is change of accuracies on changing the window size.

REFERENCES

[1] An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks, Huisu Jang and Jaewook Lee.

[2] Bayesian regression and Bitcoin, Massachusetts Institute of Technology,Devavrat Shah & Kang Zhang

[3] The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. Xin Li, Chong Alex Wang,Department of Information Systems, College of Business, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong