**Project Report**

**BIA656**

**Adv Data Analytics & Machine Learn**

**Predicting Bitcoin Price Using Machine Learning Models**

**(Group 3)**

**Instructor: Dragos Bozdog**

**Group Member: Email:**

**Wei Yang**  [wyang17@stevens.edu](mailto:wyang17@stevens.edu)

**Xinyi Li** [xli176@stevens.edu](mailto:xli176@stevens.edu)

**Ying Lu** ylu58[@stevens.edu](mailto:zyang69@stevens.edu)

**Abstract**

This essay is to talk about how to use machine learning to predict Bitcoin price by using Bitcoin data from January 2016 to July 2020. First, we built two ARIMA models with different orders and use grid search to find the best parameters and apply it to predict the Bitcoin values. Then we built multilayer perceptron model and different recurrent neural network structures like: simple RNN, GRU, LSTM and their combination to predict the price, comparing the performance of these model. Finally, we use the factor analysis to analyze several prices for gold, oil, and other digital currencies at the same period and build regression models to find the relationship between the Bitcoin price and factors.

**Key words**: Bitcoin, ARIMA, Regression, RNN

**1. Introduction**

Bitcoin, created in 2008, is a digital currency which is decentralized and free from the control of a central bank or a manager. Additionally, Bitcoin can be transferred from an individual to another individual directly. Bitcoin transactions are confirmed by network nodes and cryptography and recorded by using blockchain. Mining is a reason why bitcoin is invented, and bitcoin can be used for exchange of merchandise, and services. Hileman (2017) claims that in 2017, ranging from 2.9 to 5.8 million people use cryptocurrency wallet, and majority of them utilized bitcoin. Although many financial agencies launched warning about bitcoin considering its drawbacks which may trigger illegal activities due to price volatility, bitcoin have been widely used as a kind of investment, according to U.S. Securities and Exchange Commission. Moreover, bitcoin has the potential to accelerate the revolution of finance. In 2019, the Intercontinental Exchange commences trading the future of bitcoin via its exchange, so-called Bakkt. Olga (2019) indicates that Bakkt also declares the launch of the options of bitcoin in 2019.

According to Mai (2018), it is somehow possible to forecast the price of bitcoin using social media. Thus, this essay is to discuss the factors which influence the price of Bitcoin and explore the relationship between factors within textual data from related social media text and bitcoin value. Furthermore, this essay will also manifest the outcome of the sensitivity analysis and discuss the relativity, interdependency between bitcoin and other three digital currencies: BCH, ETH, LTC.

**2. Data Source**

We get BTC (Bitcoin) price from January 2016 to July 2020 in Yahoo Finance. It includes six columns: High, Low, Open, Close, Volume and Adj Close.

A picture containing text, tiled, white, public

Description automatically generatedAccording to the graph, we know that from 2016 to 2018, Bitcoin is on a rising trend, especially between 2017 and 2018. And bitcoin price reached its highest level in four years in January 2018. Since then, its price has also shown a significant reduction. A picture containing screen, orange, water, sitting

Description automatically generatedAfter that, the price has high volatility, but is always maintained at the middle level.

We add two new columns named ‘return’ and 'log\_return’ which is periodic return and log return to create more features. We can see it from the graph as follows.

We also get price data for gold, oil and another digital concurrency to analyze the correlation.

**3. Model**

3.1 ARIMA model

Our first attempting is ARIMA model. As you can see ACF and Partial ACF graphs.

A screenshot of a cell phone

Description automatically generated

From the tailing ACF, we discover that the data is strongly autocorrelated. And combined with the graph of Partial ACF, they show the serials are highly correlated with the previous one state. When we selected autocorrelation order with the BIC information criterion, we also find that AR (1) achieve the highest BIC score of 11.337 compared with other AR models.

A screenshot of a cell phone

Description automatically generated

On this daily BTC dataset, AR(1) model seems to enough to explain the relationship in a large scale. But further manually improving and tuning is need for better performance and accuracy. That’s why we remove the outliers. We can see how the A screenshot of a social media post

Description automatically generatedresiduals changes before and after A screenshot of a cell phone

Description automatically generatedclean up as follows.

Furthermore, the residual of this model fails to pass the Normal test. So, we decide to use grid search to find better parameters combination for this dataset. Grid search for p is from 0 to 8, for d is from 0 to 4 and for q is from 0 to 4. With the brute force searching, we get the best parameters [5, 0, 3] for initializing the ARIMA model structure. The coefficient for each part is listed as follows. Also, we adjusted the bias term to make the residual of this model equal to 0.

A screenshot of a cell phone

Description automatically generatedFinally, with the best parameters, this ARIMA model with rolling based prediction step by step performs better than AR(1). This model only has mean absolute error 292 on the test set.

In a nutshell, from what I learn about Bitcoin market, Bitcoin is highly fluctuating in a single day, even in a minute. But it can be treated as a cyclic in one week from history data and from PACF graph as well. And ARIMA model structure has its drawbacks. it is quite explainable but constraint on the time serials framework, which is not that flexible as machine model.

3.2 Machine Learning models

Before using machine learning, we have to do some feature engineering. First, We add such features on our dataset. We think some of this financial indicators may be useful for prediction. They are EMA (exponential moving average), ROC (Return on capital) and so on. This is few rows of our whole dataset.

And 'Close', 'Open', 'High' and 'Low' are our target labels. We split the training and testing dataset with 80 and 20 %. Each sample has 15 days information of bitcoin price and other financial indicators. We hope to use 15 days data to predict theA screen shot of a social media post

Description automatically generated high, low, open and close on the next day. When training the model, we set batch size equal 8 and iterate the updating process with 20 epochs.

3.2.1 MLP

First machine learning model we use is Multilayer Perceptron. Multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). It has an input layer, at least one hidden layer and an output layer. Except for the input nodes, each node is a neuron which make a nonlinear activation function and use back propagation for training models. We try different structure of this kind of dense model which means different numbers of hidden layers and nodes as you can see as A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedfollows.A screenshot of a cell phone

Description automatically generated

And we try to optimize the mean absolute error on the training data. We set L2 normalization on each layer to reduce the over fitting problem. This graph is the A close up of a map

Description automatically generatedvalidation loss changing with epochs. It converts very well.

3.2.2 RNN

RNN is also a type of artificial neural network in which the connections between nodes form a directed graph along the time series. Unlike MLP, RNN has an internal memory that keeps the data sequence unchanged, so each sample can be dependent on the previous sample.A picture containing object, clock

Description automatically generated

We also try different Recurrent Neural network structures like: simple RNN, GRU, LSTM and rest of combination. Their structures are listed as follows.A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated Simple RNN GRN

LSTM

When we make prediction, we also add two dense layers from the result of RNNs. Because it can avoid overfitting problem.

3.3 Conclusion

Finally, we summarize the prediction result of all the models. We can see the LSTM and GRU achieve a good result with low error in each target features. ARIMA model’s mean absolute error is 292.665 which has the higher error than machine A picture containing phone

Description automatically generatedlearning model.

**4. Factor Analysis**

4.1 Correlation

There are many factors that will affect the price of Bitcoin. We tested the following factors: Oil price, Gold price, Curcy price, ETH price, LTC price, BCH price, Financial transaction indicator ATR, CCI, PVT. The correlation between Bitcoin and these factors are given below.

According to the figure, LTC price has the highest correlation with Bitcoin price. In addition, BCH, ETH price and these Financial transaction indicators all have significant positive relationship with Bitcoin Price.

4.2 OLS Model

Ordinary least-squares (OLS) models assume that the analysis is fitting a model of a relationship between one or more explanatory variables and a continuous or at least interval outcome variable that minimizes the sum of square errors, where an error is the difference between the actual and the predicted value of the outcome variable. In this case, an OLS model with multiple predictor variables is delivered. By setting the BTC price as dependent variable and others as independent variable, the result is given below:

R-squared of the model is 0.851, which means the model can predict the dependent variable at the level of 0.851

BTC price = 1.068 + 0.74\*(TR)+ 2.02\*(ATR) + 8.91\*(gold price) + 82.29\*(oil price) – 2.05\*(Curcy price) – 1.17\*(ETH price) + 22.14\*(LTC price) + 1.43\*(BCH)

4.3 Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The coefficient of the model is list below

BTC price = 7.38\*(TR)+ 2.03\*(ATR) + 8.92\*(gold price) + 8.21\*(oil price) – 2.05\*(Curcy price) – 1.16\*(ETH price) + 2.21 \*(LTC price) + 1.43\*(BCH)

4.4 Variable Selection

When fitting models, it is possible to increase the likelihood by adding parameters but doing so may result in overfitting. BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model. In this section, BIC Criterion are used to select the best factors which have significant impact on the price of Bitcoin. After selection, gold price, oil price, LTC price, BCH price and two financial transaction indicator s are remaining in the equation. The new coefficient of the model built by these significant factors is as follow:

BTC price = 0.69\*(TR)+ 2.51\*(ATR) + 11.76\*(gold price) + 86.72\*(oil price) + 14.89 \*(LTC price) + 0.54\*(BCH)

**5. Future work**

In the future, we may extracted bitcoin related social media text, and so on. And we will do sentiment analysis for a certain period. Because we hope to Quantify sentiment as another important feature for BTC price. Second, we may further evaluate and improve the models in smaller granularity dataset. Afterwards, we will also implement trading policy code for quantitative trading. Last but not least, we will explore more important features and explore some event driven analysis from case study.

**References**

1. Hileman, G& Rauchs, M. "Global Cryptocurrency Benchmarking Study" .Cambridge University. Archived from the original on 10 April 2017. Retrieved 14 April 2017.
2. Mai, Feng, Zhe Shan, Qing Bai, Xin Wang, and Roger HL Chiang.2018, "How does social media impact Bitcoin value? A test of the silent majority hypothesis." Journal of Management Information Systems 35, no. 1 (2018): 19-52.
3. Olga, K, 2019. "Bakkt Plans First Regulated Options Contracts on Bitcoin Futures".https://www.bloomberg.com/news/articles/2019-10-24/bakkt-plans-first-regulated-options-contracts-on-bitcoin-futures