**Bitcoin Price Analysis**

Yulong Yang, Yaqin Tang

**ABSTRACT**

In this paper, we discuss different methods for predicting price variation of Bitcoin, a recently popularized virtual, cryptographic currency. First, we apply two traditional way of doing stock market analysis – Bayesian regression and support vector machines (SVM). And compare the performance on Bitcoin and Google’s stock price, of which the relative error rate of the former one is larger than the latter. Based on this, we use moving average trend classifier to predict the trend of Bitcoin and try two new methods – Google trend analysis and news sentimental analysis, in order to improve the accuracy of the classifier. Results shown that SVM outperforms Bayesian curve fitting in terms of large number of datasets, and those traditional way of doing stock market analysis cannot be directly applied to Bitcoin price analysis due to its large and unpredictable price variation.

***Keywords***: Bitcoin, Bayesian regression, SVM, Google trend analysis, news sentimental analysis

1. **INTRODUCTION**

Bitcoin is a new currency that was created in 2009 by an unknown person using the alias Satoshi Nakamoto. Transactions are made with no middle men – meaning, no banks! There are no transaction fees and no need to give your real name. More merchants are beginning to accept them: You can buy web hosting services, pizza or even manicures. Bitcoins can be used to buy merchandise anonymously. In addition, international payments are easy and cheap because Bitcoins are not tied to any country or subject to regulation. Small businesses may like them because there are no credit card fees. Some people just buy Bitcoins as an investment, hoping that they’ll go up in value. People can send Bitcoins to each other using mobile apps or their computers. It’s similar to sending cash digitally. Bitcoins are stored in a “digital wallet,” which exists either in the cloud or on a user’s computer. The wallet is a kind of virtual bank account that allows users to send or receive Bitcoins, pay for goods or save their money. Unlike bank accounts, Bitcoin wallets are not insured by the FDIC. Though each Bitcoin transaction is recorded in a public log, names of buyers and sellers are never revealed – only their wallet IDs. While that keeps Bitcoin users’ transactions private, it also lets them buy or sell anything without easily tracing it back to them [1].

No one knows what will become of Bitcoin. However, since Bitcoin is such a new type of currency and is of great difference than traditional stock market with such a worldwide concern, it’s an interesting topic to research if we can do some analysis on Bitcoin price prediction.

Moreover, the Bitcoin market Cap is shown below in figure 1, it’s worth an effort to dig into this problem.

Figure 1 Bitcoin Market Cap

1. **BAYESIAN REGRESSION**
   1. **Data Preprocessing**

As can be seen from figure 2, we get the Bitcoin historical data from the *COINBASE* website in US dollar. Then sort the data based on time and choose the closed price of each day as the data we use in Bayesian curve fitting.

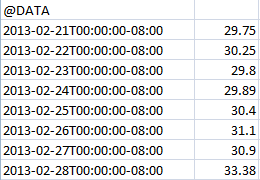
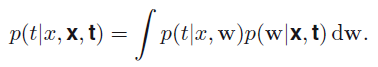


Figure 2 Dataset Example

* 1. **Algorithms**

In the curve fitting problem, we are given the training data **x** and **t**, along with a new test point x, and our goal is to predict the value of t, we therefore wish to evaluate the predictive distribution p(t|x,**x**,**t**). Here we shall assume that the parameters α and β are fixed and known in advance. A Bayesian treatment simply corresponds to a consistent application of the sum and product rules of probability, which allow the predictive distribution to be written in the form [3]

 (Eq. 1)

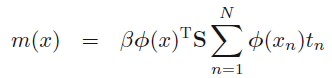
Here p(t|x,w) is given by Eq. 2, and we have omitted the dependence on α and β to simplify the notation. Here p(w|**x**,**t**) is the posterior distribution over parameters, and can be found by normalizing the right-hand side of Eq. 3. For problems such as the curve fitting, this posterior distribution is a Gaussian and can be evaluated analytically. Similarly, the integration in Eq. 1 can also be performed analytically with the result that the predictive distribution is given by a Gaussian of the form in Eq. 4.

 (Eq. 2)

 (Eq. 3)

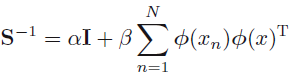
 (Eq. 4)

Where the mean and variance are given by

 (Eq. 5)

 (Eq. 6)

Here the matrix S is given by

 (Eq. 7)

Where I is the unit matrix, and we have defined the vector ф(x) with elements фi(x) = x^i for I = 0, …, M. We see that the variance, as well as the mean, of the predictive distribution in Eq. 4 is dependent on x. the first term in Eq. 6 represents the uncertainty in the predicted value of t due to the noise on the target variables and was expressed already in the maximum likelihood predictive distribution. However, the second term arises from the uncertainty in the parameters w and is a consequence of the Bayesian treatment.

1. **SVM FOR REGRESSION**
   1. **Data Preprocessing**

Choosing a suitable forecasting horizon is the first step in financial forecasting. From the trading aspect, the forecasting horizon should be sufficiently long so that the over-trading resulting in excessive transaction costs could be avoided. From the prediction aspect, the forecasting horizon should be short enough as the persistence of financial time series is of limited duration. As suggested by Thomason [5] a forecasting horizon of five days is a suitable choice for the daily data. As the precise values of the daily prices is often not as meaningful to trading as its relative magnitude, and also the high-frequency components in financial data are often more difficult to successfully model, the original closing price is transformed into a five-day relative difference in percentage of price (RDP). The input variables are determined from four lagged RDP values based on five-day periods (RDP-5, RDP-10, RDP-15, and RDP-20) and one transformed closing price which is obtained by subtracting a 15-day exponential moving average (EMA15) from the closing price. The subtraction is performed to eliminate the trend in price as the maximum value and the minimum value is in the ratio of about 2: 1 in all the five data sets. The optimal length of the moving day is not critical, but it should be longer than the forecasting horizon of five days. EMA15 is used to maintain as much of the information contained in the original closing price as possible since the application of the RDP transform to the original closing price may remove some useful information. The output variable RDP+5 is obtained by first smoothing the closing price with a three-day exponential moving average, because the application of a smoothing transform to the dependent variable generally enhances the prediction performance of neural networks.

* 1. **Algorithms**

Given a set of training data {(x1, y1), …, (xl, yl)}, where each xi belongs to X and X belongs to R^n. (X denotes the input space of the sample) and corresponding target value yi belongs to R for I = 1, …, l (where l corresponds to the size of the training data), the objective of the regression problem is to determine a function that can approximate the value of y for an x not in the training set [4].

The estimating function f is taken in the form:

f(x) = ( w\*ф(x)) + b (Eq. 8)

where w belongs to R^m, b belongs to R is the bias and ф denotes a non-linear function from R^n to high dimensianl space R^m (m>n). The objective is to find the value of w and b such that values of f(x) can be determined by minimizing the risk.

 (Eq. 9)

Where L is the extension of insensitive loss function originally proposed by Vapnik [6] and defined as:

 (Eq. 10)

For those αi and αi\* for which the xi’s corresponding to 0 < αi < C and 0 < αi\* < C are called support vectors. f(x) is computed as in Eq. 11.



 (Eq. 11)

It is to be noted that we do not require function ф to compute f(x) which is one the advantage of using the kernel.

1. **MOVING AVERAGE TREND CLASSIFIER**

The moving average algorithm is to calculate the average value within a window size and then move to the next time period for the fixed size. Combining the moving average and any short-term price prediction algorithm, we are able to classify Bit-coin price trend weekly and monthly, which works as follows:

* sample the historical data weekly, which means the sample rate is 7;
* predict each day's price for next week;
* calculate moving average this week go back N ave\_now and next week go back N ave\_future;
* if ave\_now > ave\_future, predict 0: decrease, otherwise 1: increase;
* from results of each day next week, vote for the trend next week;
* Similarly vote for monthly trend with weight [0.6; 0.25; 0.1; 0.05].

1. **GOOGLE TREND ANALYSIS**

This algorithm is inspired by previous work [7]. The work predicted the change of Dow Jones index based on the Google search volume of a certain keyword such as “debt”. The intuition is, people would likely gather information before they invest in anything. If the search volume of financial-related keywords increases, it is likely that people are more concerned about their financial assets, and therefore they are likely to short, making the price/index of financial market decrease.

The original work collected longitudinal time of DJ index, and corresponding Google search volume with timestamp. They defined is the search volume of week t, and

where is 3 weeks. depicts the average search volume from week t-1 to week . And then the algorithm looked at the difference between this average volume and the search volume of week t:

They predicted that if the difference is positive, meaning search volume increases this week, and there the price/index would fall next week (week t+1); otherwise price/index would rise.

This algorithm worked well on DJ index, and we were wondering if it had the same effect on Bitcoin price too. Theoretically, Bitcoin is internet-centric, and therefore people should rely on internet searches to know it more than traditional financial products.

We incorporate this algorithm by substituting DJ index with Bitcoin price, and search volume with Google trend search interest index. We obtained Bitcoin close price from coinbase [8], a US-based Bitcoin trading market. Google trend search interest index (SI index) is a discrete value ranging from 0 to 100. It depicts the search interest of a certain keyword at a certain moment. The value is normalized so that the absolute search volume does not affect it. We were unable to obtain the Google search volume data because Google does not provide it anymore. Both data we collected are in weekly format.

1. **NEWS SENTIMENTAL ANALYSIS**

We also looked at how Bitcoin price would be affected by relevant news. It is well-known that Bitcoin price fell drastically when negative news such as Bitcoin trading market was shut down occurred. Therefore it is possible to examine the relationship between the two to gain new insights on the pattern of Bitcoin price.

For this part, we didn’t derive out a full-ledge algorithm. Instead, this is an on-going work which we still need more effort to convert into a working algorithm. The basic idea is, if negative news drags down Bitcoin price, and positive news brings it up, we could predict the price trend based on the polarity of news that occurred recently then. Things to consider including how to detect the polarity of news, how recent news should be in order to be effective on the price, etc.

To tackle the first issue, we got the help of text sentiment analysis. The analysis and algorithm it used itself is beyond the scope of this report. To put it simply, it scores or predicts the polarity of a piece of texts based on analyzing the text. We hope if we could obtain the news polarity from text sentiment analysis, it is possible to use it in the prediction of price change as we mentioned above.

1. **EXPERIMENTS AND RESULTS**

The experiments are conducted through all the historical datasets starting from November 2011 to November 2014. And below figure 3 and 4 show the results for Bayesian curve fitting and supporting vector separately.



Figure 3 Comparison of actual price and predict price by Bayesian curve fitting



Figure 4 Comparison of actual price and predict price by support vector

Blow in table 1 is the evaluation for both methods.

Table 1 performance evaluation

|  |  |  |
| --- | --- | --- |
| Algorithms | MSE | RMSE |
| Bayesian | 26.76 | 0.08 |
| SVM | 21.39 | 0.07 |
| Google stock | 16.79 | 0.03 |

From the table, we can see that the relative mean square error for Google stock is smaller than any of the algorithms applied to Bayesian or SVM, the reason could because Bitcoin price is fluctuating a lot and we calculate the variance of Bitcoin is 27 times that of Google stock price in figure 5.

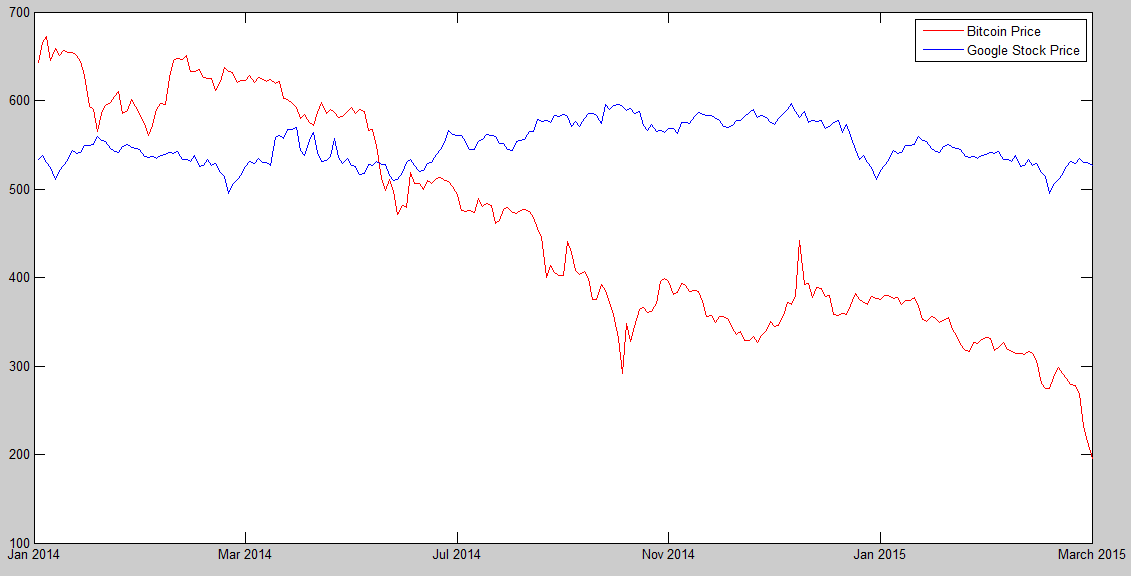


Figure 5 comparison of actual price for Bitcoin and Google stock

Also, based on the analysis for Bayesian and SVM, we evaluate the moving average trend classifier and the confusion matrix is shown below, and for Bayesian, the TPR = 0.7091; FPR = 0.5694; for SVM, the TPR = 0.7455; FPR = 0.5278.

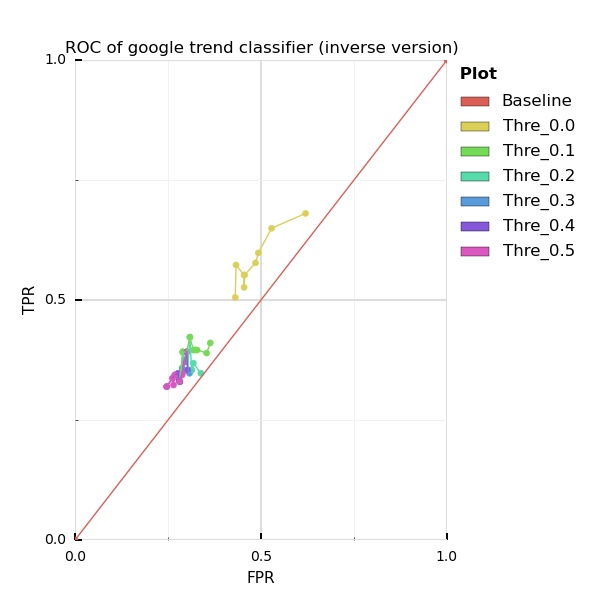
|  |  |  |  |
| --- | --- | --- | --- |
| Bayesian  Based | Predicted Class | | |
| Actual Class |  | C\_IN | C\_DE |
| C\_IN | 78 | 32 |
| C\_DE | 41 | 31 |

|  |  |  |  |
| --- | --- | --- | --- |
| SVM  Based | Predicted Class | | |
| Actual Class |  | C\_IN | C\_DE |
| C\_IN | 82 | 28 |
| C\_DE | 38 | 34 |

We ran the algorithm on the data of 2012 to 2015, because prior to 2012 the SI is too small to have any effect. There were 173 data points in total. We modeled it into a binary classification problem, that is, for each week t, predict whether its price would rise or fall based on previous data.

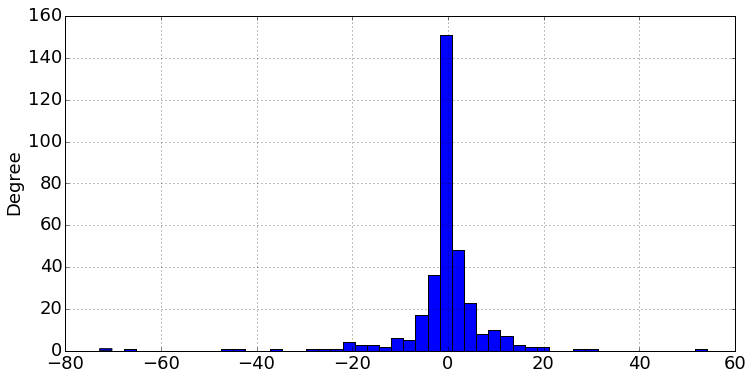
The overall correction of our experiment was 42.4%, which is very bad, worse than random guess. One of reasons we found was that the process of information gathering might not be the same for Bitcoin. Since it is a new concept and product, the information gathering activity should be coming from people’s interest on Bitcoin, instead of their concerns. Therefore, when search interest increases, we should predict the price to rise in the nearby future.

We ran the modified algorithm again on the same data set, and obtained a better result. The ROC is plotted as below.

****

From the figure we could see that TPR/FPR is approaching 0.7 in best case. While TPR is relatively high, so does FPR.

Lastly but not least, we analyzed the relationship between Bitcoin price and news. We obtained the same Bitcoin price data as mentioned above, and 341 pieces of news headlines from New York Times that had “Bitcoin” in the headline. The news covered 2012 to 2015, as theres no news met our requirement prior to 2012. We looked at the change in Bitcoin price 12 hours after every piece of news occurred. We plotted the distribution of such price change as below.



It is clear that most of news had little or no effects on the price as the price change were mostly 0. However, we did find a change of e.g. $20 occurred after certain news came out. Moreover, we had news that dragged down the price nearly $80 after 12 hours.

To summary, (1) news did have effect on the price of Bitcoin; (2) based on our study, it is unknown whether the change we observed were all because of the news, as it is possible there are other factors (3) most of news have not effect on Bitcoin price (neutral news).

1. **CONCLUSION**

In this paper, we compared the accuracy of different methods for predicting the price and trend for Bitcoin price, and results shown that SVM outperforms Bayesian curve fitting in terms of large number of datasets and these traditional methods to analyze Bitcoin price is not as good as analysis for the stock market; More care needs to be taken to analyze Bitcoin market because of its large and unpredictable price variations affected by all kinds of sources.

**REFERENCES**

[1] CNN money: http://money.cnn.com/infographic/technology/what-is-bitcoin/

[2] Devavrat Shah, Kang Zhang, “Bayesian Regression and Bitcoin”, Lab for Information and Decision Systems, Department of EECS, MIT.

[3] Pattern Recognition and Machine Learning Bishop, page 21 – 32.

[4] Shom Prasad Das, Sudarsan Padhy, “Support Vector Machines for Prediction of Futures Prices in Indian Stock Market”, International Journal of Computer Applications, Vol. 41-No.3, 2012.

[5] Thomason M., “The practitioner methods and tool”, Journal of Computational Intelligence in Finance 1999; pp.35–45.

[6] V. Vapnik., 1995, The Nature of Statistical Learning Theory. Springer, N.Y. ISBN 0-387-94559-8.

[7] Preis, Tobias, Helen Susannah Moat, and H. Eugene Stanley. "Quantifying trading behavior in financial markets using Google Trends." *Scientific reports* 3 (2013).

[8] Coinbase, https://www.coinbase.com.