Neelanshu Thapar

Bitcoin Price Value and Direction Prediction using Statistical Learning Methods

Introduction

Problem

This project belongs to the field of financial analysis and can be applied to the price prediction of a variety of assets. The focus will be on the cryptocurrency Bitcoin (BTC) -- which has price data that is ubiquitous and easy to access. On a fundamental level mining complexity, purchase power, and demand for Bitcoin tends to be what influences price action. Unfortunately, these fundamental factors are not well quantified or directly accessible like they are in the stock market, making predicting price action based on fundamental data complicated. This reality forces all analysis to be done using technical data like market capitalization and volume. Statistical Learning techniques may prove to be a viable approach to making use of such data. Two primary focuses in the field of Bitcoin research are the prediction of price values and price direction. The volatile nature of Bitcoin prices makes such predictions quite difficult. The market has been previously described as highly disordered with a high degree of randomness.[3] Given this nature, powerful models will have to be deployed to produce meaningful results, potentially beyond the scope of models covered in this class.

Questions

This paper tries to answer a few questions surrounding Bitcoin price prediction. The primary response variables will be price value, and price direction. Research groups tend to quantify these variables differently, as they are features that have to be generated. Mallqui et.al. defines price values as the Closing, High, and Low prices for a given trading period.[2] They define price direction as the difference in Closing prices at trading period t+1 compared to t. This paper defines the price value differently, namely as the High price of trading period t+1 shifted down to trading period t, and the price direction in the same way as Mallqui et.al. This research seeks to quantify the accuracy of price value prediction using linear regression. It then seeks to classify the direction of bitcoins price at a future time point using observations from a prior time

point. Bitcoins rise to dominance as an asset is a meaningful reason to engage in this research, but more important than that is the capacity to abstract successful approaches. The underlying theme here is the prediction of price vectors for highly volatile assets -- something that is of critical importance in finance.

State of the art

Both factors are often assessed in a myriad of ways. Some groups have opted for simple classification and regression approaches, with meaningful success in granular price prediction at 1 hour time scales using linear regression and Support Vector Machine (SVM) regression.[1] Classification attempts by the same group yielded poor accuracy, with the highest being the result of neural network usage.

More sophisticated approaches have involved starting with a more protracted data conditioning phase. In the work of Mallqui et.al. 86 attributes were used to start and the usage of attribute selection techniques like Information Gain, Correlation, and Principal Component Analysis narrowed down the number of predictors. From there Artificial Neural Networks -- specifically Multilayer Perceptrons -- and Recurrent Neural Networks were used to predict price values. Techniques like Logistic Regression, Classification Trees, and SVMs were used to classify price direction. SVM algorithms obtained the best results[2].

Overall state of the art work in this space takes on similar approaches, notable differences in the success rates arise from different predictor variable choices and time intervals for analysis.

Methods

Dataset

Data is sourced from Kaggles Cryptocurrency Historical Prices repository posted by user SRK. The dataset has 2682 observations over a period of 7 years (12-27-2013 to 02-26-2021) logged on a daily basis. The data is BTC specific and has 11 attributes. Numeric attributes include BTC daily Open, High, Low, Close values as well as trading Volume and overall Market Capitalization (Marketcap). Character attributes include Serial Number for the given day, Name of the currency, Symbol (ticker) for the currency, and Date of transaction. Character attributes were excluded as predictors in this paper's analysis as they contribute little to no information about BTC price direction or values.

Preprocessing

Predicting price change direction and values required the generation of response variables from the predictors. To represent price change the difference in closing prices at time point t+1 and t was assessed. Time intervals are defined as 1 trading day, which in the case of BTC coincides with calendar days. The price change response variable was named "gain" and has 2 classes: 0 to represent a negative trading day and 1 to represent positive.

The price value response variable, "tomorowHigh" was generated by shifting the daily High price of BTC down from time point t+1 to time point t. The dataset required preprocessing to convert values from 0's to NA and then drop the corresponding rows. This process sized down the dataset to 2619 observations. The majority of the 0 values were found in the Volume predictor variable, representing rare non-trading days, therefore they were irrelevant observations. There were no original NA values.

Due to the large numerical discrepancy between BTC price attributes (Open, High, Low, Close) and market based attributes (Marketcap and Volume), the data was normalized using minmax normalization as follows:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Equation 1: min-max normalization equation

To allow for a validation set approach, the original data was randomly split 80/20 with 80% corresponding to the training set, and 20% corresponding to a hold out or testing set. It is important to note that the response variables tomorrowHigh, and Gain were separated and treated independently, effectively creating 2 similar datasets. This was done to allow for true independent assessment of linear regression and classification based techniques respectively.

Descriptive Statistics

Basic descriptive statistics were run on the underlying data including correlations between all predictor variables as well as response variable tomorrowHigh. Correlation coefficients help quantify how strong the relationship between two variables is. Variable inflation factor (VIF) values were also assessed for each predictor. The VIF indicates how much of the variance in a regression coefficient is inflated by multicollinearity. These two statistics combined can help understand how independent the predictor variables truly are, a key component of successful regression.

Linear Regression

To arrive at an appropriate model to predict tomorrowHigh the predictors were assessed one by one using a linear regression. Linear regression attempts to model the relationship between predictor and response variables in a less flexible way than other techniques, an

approach that often helps avoid overfitting. It is a common technique that assumes independence of predictor variables. Linear regression was conducted on each predictor variable independently, and then all predictor variables at once.

Classification via Logistic Regression

Logistic regression is a generalized linear technique that can classify response variables. This paper uses logistic regression to classify the direction of price change for BTC into 2 classes: 0 (negative price change), and 1 (positive price change). This approach assumes the classes are linearly separable.

Classification via Decision Trees

Classification trees are a specific form of a decision tree where observations are categorized based on a series of nodes from root to leaves. Binary recursive partitioning splits the data at each level iteratively until useful partitions can no longer be found. State of the art papers employ this approach, so it is included here.

Classification via Random Forest Algorithm

Random forest is a generalized algorithm that uses an ensemble of decision trees constructed with bagging that can be used to both classify and assess variable importance.

Classification via Support Vector Machines

Support vector machines are some of the most powerful tools that produce one or multiple hyperplanes that are used to categorize new data. In the work of Mallqui et.al. this was one of the most powerful approaches to classification of price direction. One of the aims of this paper is to replicate said accuracy. Both linear and radial kernels are used here -- linear is typically applied to linear problems while radial is useful for nonlinear classification.

Classification via Linear and Quadratic Discriminant Analysis

Linear Discriminant Analysis or LDA uses a linear combination of predictors to arrive at an accurate classification. It is superior in a multi-class classification problem, but will likely behave similarly to logistic regression in this context. It operates under the assumption that output classes have a normal distribution and similar variances. Quadratic discriminant analysis is similar to LDA but does not make the same assumptions about classes.

Results

Descriptive Statistics

```
      SNo             Name              Symbol             Date          
 Min.   : 243.0   Length:2619        Length:2619        Length:2619       
 1st Qu.: 897.5   Class :character   Class :character   Class :character  
 Median :1552.0   Mode  :character   Mode  :character   Mode  :character  
 Mean   :1552.0                                                           
 3rd Qu.:2206.5                                                           
 Max.   :2861.0                                                           
      High               Low                Open               Close        
 Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000  
 1st Qu.:0.004577   1st Qu.:0.005093   1st Qu.:0.004986   1st Qu.:0.004951  
 Median :0.046267   Median :0.045543   Median :0.045864   Median :0.045885  
 Mean   :0.089152   Mean   :0.088659   Mean   :0.088357   Mean   :0.088624  
 3rd Qu.:0.141142   3rd Qu.:0.142363   3rd Qu.:0.140701   3rd Qu.:0.140743  
 Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000  
     Volume            Marketcap          gain       
 Min.   :0.0000000   Min.   :0.000000   0:1210  
 1st Qu.:0.0001208   1st Qu.:0.004005   1:1409  
 Median :0.0039920   Median :0.040960           
 Mean   :0.0278978   Mean   :0.084914           
 3rd Qu.:0.0422301   3rd Qu.:0.133921           
 Max.   :1.0000000   Max.   :1.000000           
```

Figure 1: General descriptive statistics

There is a serious multicollinearity issue with extreme values for every predictor except Volume. This has been demonstrated via the use of both standard correlation coefficients and variance inflation factors. Correlation Coefficients are represented on a color scale corresponding from -1 (red) to 1 (blue).

Predictor	VIF
High	2356.352703
Low	945.703104
Open	1015.421132
Close	3263.830378
Volume	3.301768
Market Cap	870.408985

Figure 2: Variance Inflation Factor for each predictor

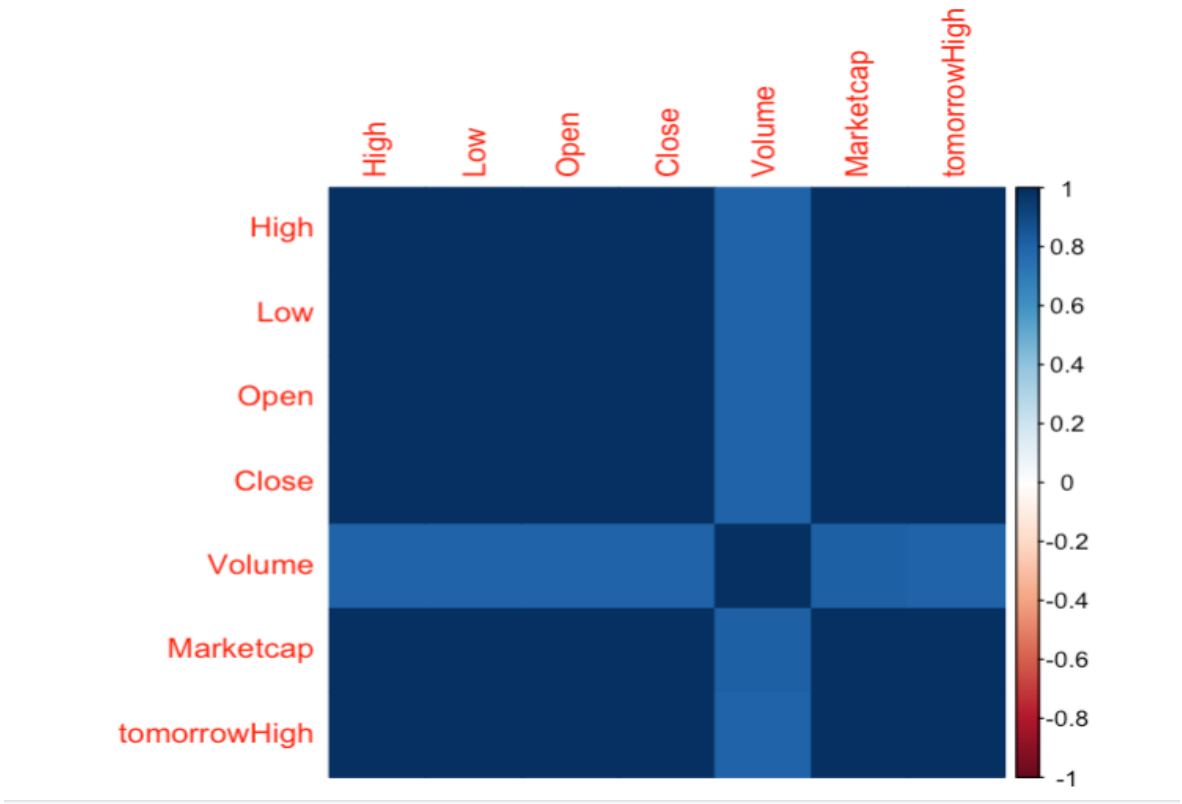

Figure 3: Correlation plot comparing Predictors to each other and to the response tomorrowHigh. Values correspond to color gradient scale from -1 to 1 (red to blue). Gain was excluded because it is categorical and cannot be correlated.

Linear Regressions

The Mean Squared Error values for linear regression approaches were extremely high, and indicate an inability of the model to produce meaningful predictions.

	Validation set approach	CV-5	CV-10	LOOCV
High	78220.03	156119.4	157101.3	156924.8
Open	262559.3	300697.5	300791.2	299904
Low	160134.6	241693.2	241957.7	242253.9
Close	41265.3	96690.5	97072.3	96807.64
Market Cap	67307.56	180673.3	178650.2	178825.6
Volume	8672714	20341264	20113868	20085902
Multiple Linear Regression	39225.84	101432.6	97693.67	97409.15

Figure 4: Mean Squared Error for linear regressions between tomorrowHigh and High, Open, Low, Close, Market Cap respectively. Multiple Linear Regression conducted between tomorrowHigh and all Predictors at once.

Variable Importance

The variable importance as assessed by Mean Decrease in Accuracy taken from the Random Forest classification indicated that Volume was the most relevant variable to the response variable tomorrowHigh. This was followed by Market Capitalization and Closing price.

Attribute	Mean Decrease Accuracy
High	4.207587
Low	5.003132
Open	4.445751
Close	5.069013
Volume	6.636005
Marketcap	5.611051

Figure 5: Variable importance per predictor measured via output of random forest classification. The Mean Decrease in accuracy corresponds with variable importance.

Classifications

Overall classification was not particularly successful. The majority of approaches had a classification error rate close to 0.5, which in a 2 class problem is the worst possible error rate demonstrated by a meaningful learning approach. The logistic regression fared slightly better with a validation set error rate of 0.463 but a cross validation set error rate closer to 0.25 across all folds. The classification tree could not be cross validated due to the multicollinearity of predictor variables limiting the growth of the tree.

Logistic Regression Validation Set Approach	0 -- ground truth	1 -- ground truth
0 -- predicted	8	7
1 -- predicted	236	273

Figure 6: Confusion matrix for logistic regression approach. 0 denotes negative gain, 1 denotes positive gain

Based on the confusion matrix shown in figure 5 we can see that the base logistic regression model had a high specificity of 0.975 but a low sensitivity at 0.0327. The majority of true positives were therefore recognized, showing a model tendency towards correct prediction of positive price action, but losses were poorly anticipated.

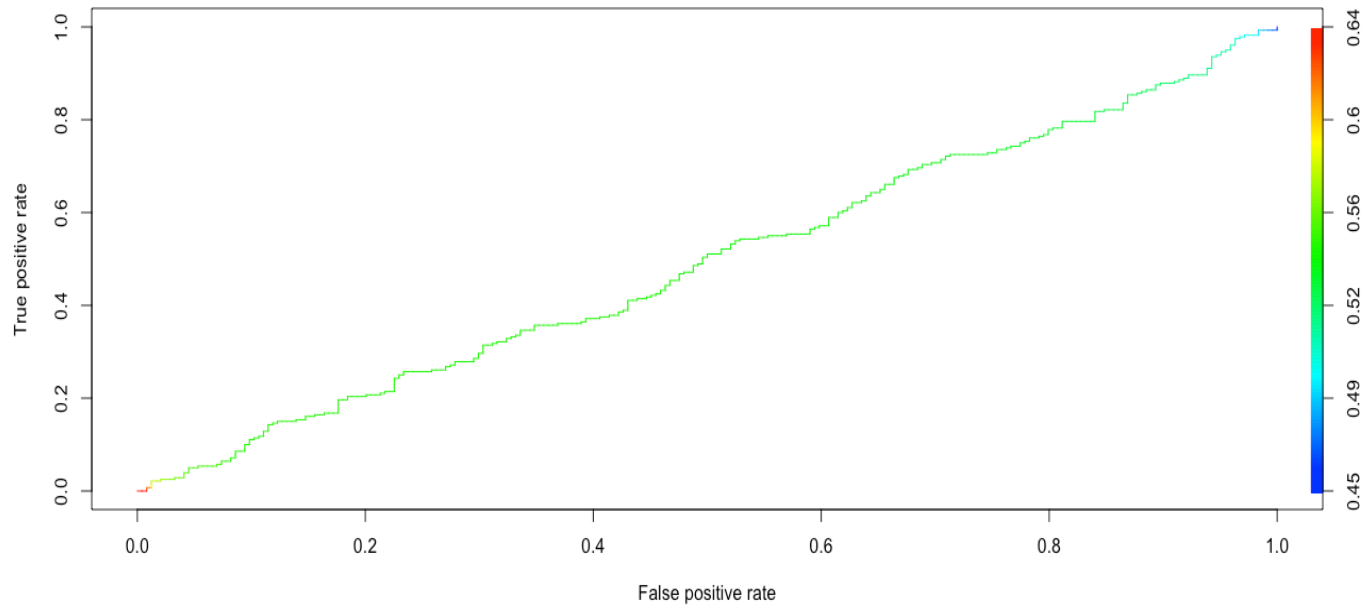

Figure 7: Receiver operating curve for logistic regression using the validation set approach, plot false positive vs. true positive rate

	Validation Set Approach	CV-5	CV-10	LOOCV
Logistic regression	**0.4637405**	**0.2497433**	**0.2498004**	**0.2496797**
RandomForest	0.5248092	0.4768996	0.4795724	0.4822451
SVM -- linear	0.4656489	0.4620167	0.4620046	0.4620084
SVM -- radial	0.4618321	0.4620014	0.4620075	0.4620084
Classification Tree	0.4611	N/A	N/A	N/A
LDA	0.4637405	N/A	N/A	N/A
QDA	0.4866412	N/A	N/A	N/A

Figure 8: Success of classification using different approaches as assessed by Classification error rate. Cross Validation (CV) and Validation set approaches both assessed.

Conclusions

Overall the answers to key questions in this paper were muddled by an unfortunate choice of dataset. The predictors chosen demonstrated serious collinearity issues, which can adversely affect attempts at regression. In terms of variable importance, Volume was a clear stand out among the predictors. Whether or not this is the ground truth is currently unclear. Volume ranked high on the Random Forest generated Variable Importance assessment (see Figure 4), but it also demonstrated the least collinearity via Correlation and VIF assessment. It may be that simply having a lower collinearity than other predictors made it stand out in the generation of classification trees. Volume's principal importance was not corroborated by performance in the linear regression experiments. In fact, Volume as the sole predictor produced the highest MSE values of any predictor. All attempts at linear regression were woefully inaccurate, in line with the inability to apply regression to predictors that are multicollinear.

In terms of classification, logistic regression was surprisingly successful in producing meaningful accuracy for cross validated approaches. The cross validation accuracy did not differ significantly between different folds. This result is not consistent with other research in the field, and may not be generalizable. The classification tree yielded poor accuracy and generated a single node centered around the Volume predictor. This is likely due to the inability of other predictors to provide meaningful information as the tree grows. The single node tree was incompatible with cross validation implementation. LDA and QDA were included to add robustness to the analysis, but had similarly poor performance, and so will not be commented on extensively.

Ultimately this research was not particularly insightful but did allow for significant learning. There are a few key differences in established literature and the approach taken here that could be implemented to improve the performance of these models. The first is including significantly more attributes. Mallqui.et.al started with approximately 86 attributes that represented various facets of the global economy including stock indexes, commodity exchanges, technical data extrapolated from BTC exchanges etc.[2] The limited choice of attributes used here was a significant issue. Furthermore, in literature an ensemble of learners are often used to achieve the most significant results. Although this paper used Random Forest (an ensemble) -- it did not generally implement this approach. Finally it is clear that the majority of new research in this area is neural network based, even going as far as incorporating deep learning.[4] Meaningful future work will likely necessitate the use of these powerful techniques.

References

1. Greaves, Alex, and Benjamin Au. "Using the bitcoin transaction graph to predict the price of bitcoin." (2015).
2. Mallqui, Dennys CA, and Ricardo AS Fernandes. "Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques." *Applied Soft Computing* 75 (2019): 596-606.
3. Lahmiri, Salim, Stelios Bekiros, and Antonio Salvi. "Long-range memory, distributional variation and randomness of bitcoin volatility." *Chaos, Solitons & Fractals* 107 (2018): 43-48.
4. Lahmiri, Salim, and Stelios Bekiros. "Cryptocurrency forecasting with deep learning chaotic neural networks." *Chaos, Solitons & Fractals* 118 (2019): 35-40.