Bitcoin Price Prediction with Historical Price, Volume and Block Data

Violeta Stoyanova, Kelly Shaffer, Eric Pitruzzella

CUNY-School of Professional Studies

Abstract

The purpose of this project is to predict the price of bitcoin using multiple variables such as individual features relating to bitcoin prices and its payment network from disparate historical data sources. Three binary classification algorithms were used, such as binomial generalized linear model (GLM), support vector machine(SVM), and random forest algorithms (RFA), to predict the sign change in Bitcoin price. After the machine learning (ML) models were build their accuracy was compared and the results were visualized. Programming was done in both Python and R. Bitcoin price prediction was done both via regression and binomially. In addition, total and average USD market price were extracted across major bitcoin exchanges as well as total output volume after change is removed from total value.

Keywords: Machine Learning, Generalized Linear Model, Support Vector Machine, Random Forest Algorithms

Bitcoin Price Prediction with Historical Price, Volume and Block Data

In January 2009, the first block on the bitcoin chain, the “genesis block”, was mined. The blockchain is a public, unalterable ledger of bitcoin transactions. As companies begin to investigate the value of blockchain to their organizations, society’s attention is increasingly drawn this relatively new world. No matter where you look, you will see mixed opinions of Bitcoin. Forbes called it the best investment of the year in 2013, and in 2014 Bloomberg called it the worst. Maybe it’s because of this mysterious reputation that we were attracted to it as our final project. Cryptocurrency exchange is an interesting, culturally relevant topic that can implement Data Science tools such as ML and demonstrate what we have learned so far in this course. ML is an especially useful tool in predicting the price of bitcoin since

Through elaborate analysis any potential indicators of the change in the price of Bitcoin could be detected granting us the ability to be able to generate predictions. Being able to predict the price of bitcoin will bring value to people who are interested in investing in bitcoin or possibly understand what determines the price of bitcoin. Originally the intention was to a model that predicts the price of bitcoin using google trends but a couple of roadblocks such as data structure, reliability and validity, were encountered. For one we realized that google trends data is being stored monthly and bitcoin data- daily and also just because people search for bitcoin information it doesn’t mean that the price is directly affected by those actions.

Methods

For the purposes of this project data was webscraped from https://blockchain.info/stats through the Python package BeautifulSoup and used for both analyses in R and Python. The extracted data was then saved and imported in a dataframe format using Pandas package. The data contained 1396 rows and 20 columns after the initial cleanup. The date variable was

converted from a string to a datetime containing data from 2010 until present and then all the NaN values were backfilled. The data was then split into test and train 75% to 25% in R and 80% to 20% in Python.

Results

Firstly, linear regression was performed on the price data to see how well does the prediction would fit the actual numbers. The linear regression model appeared to actually perform pretty well in the plot, but in reality it's just outputting the previous day's price. It's 'fitting' the data well, but has virtually no predictive power (Figure 1.) In R correlation was performed on all of the variables and the highest correlation was found between Miner Revenue and Price.

The classifiers performed pretty well given the historical data. The accuracy was of around 55% for random forest, and 60% for logistic regression and SVM. Currently the models are simply trained to predict whether the price will increase (1) or decrease (0) (Figure 2., Figure 3., Figure 4., Figure 5., Figure 6.,).

------Regression------

Trained Linear Regression in 0.0024344921112060547 sec

Score: 0.7117987969652474

Trained Random Forest Regressor in 0.1400313377380371 sec

Score: -0.6923168764033425

Trained Support Vector Regressor in 0.2955796718597412 sec

Score: -1.059409182809726

------Classification------

Trained Logistic Regression in 0.010781049728393555 sec

Score: 0.605

Trained Random Forest Classifier in 0.06097865104675293 sec

Score: 0.555

Trained Support Vector Classifier in 0.27366042137145996 sec

Score: 0.605

The ML algorithms that were build in this project were able to predict correctly the price direction- up or down with 60% accuracy. It's a good beginning for building a trading system but additional work is needed for better predictive ability and money management.

Figure 1.

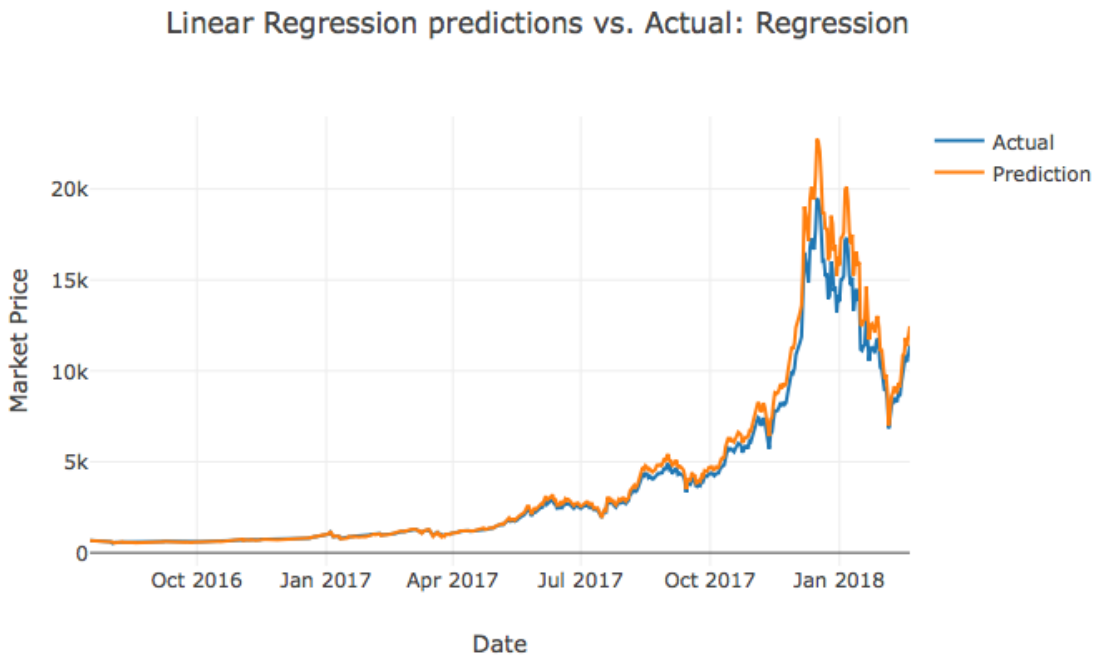


Figure 2.

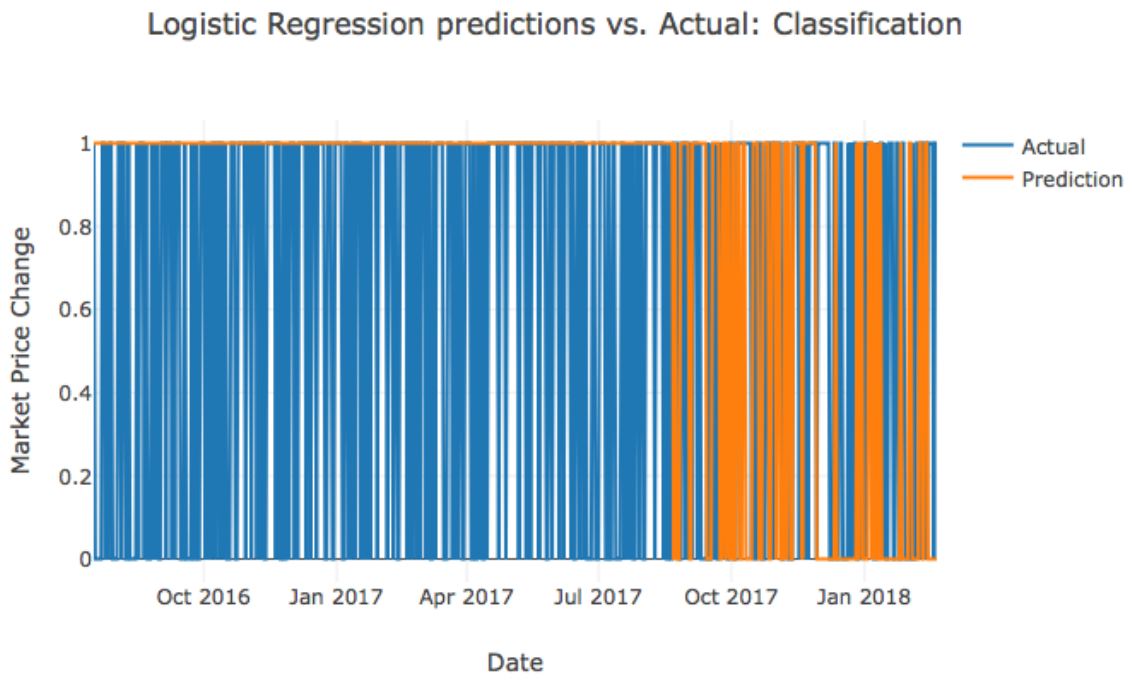


Figure 3.

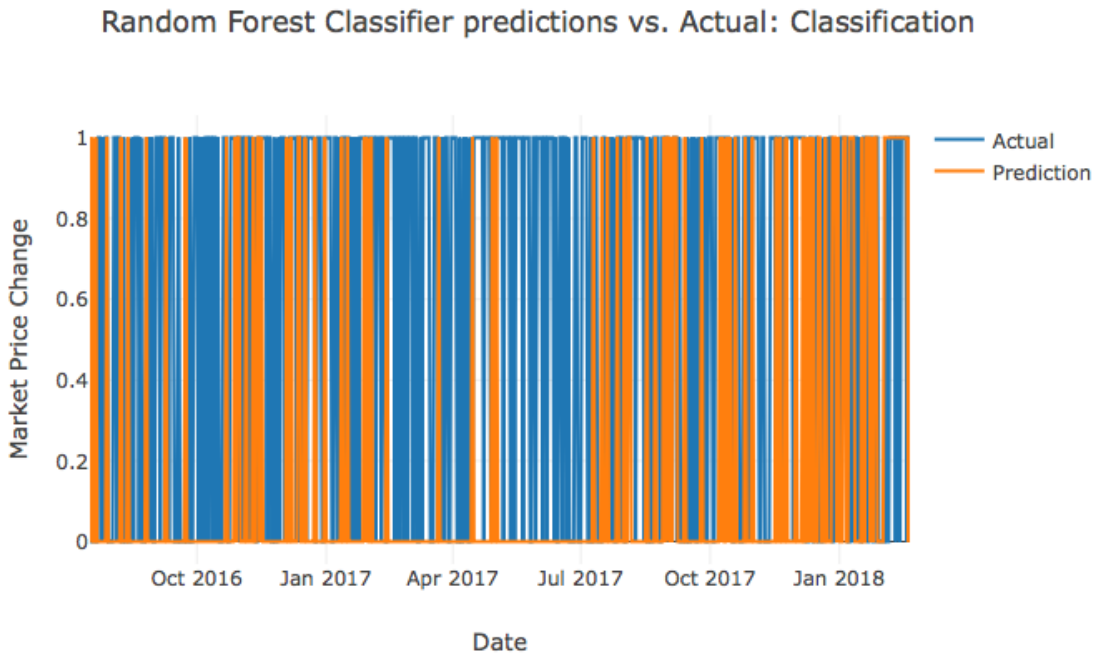


Figure 4.

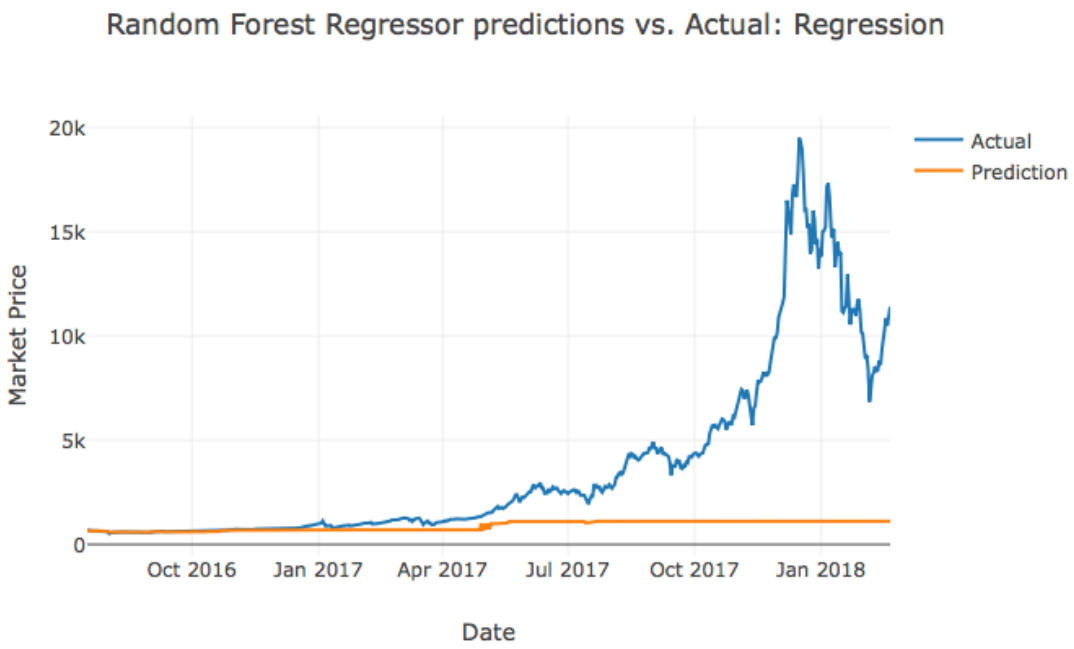


Figure 5.

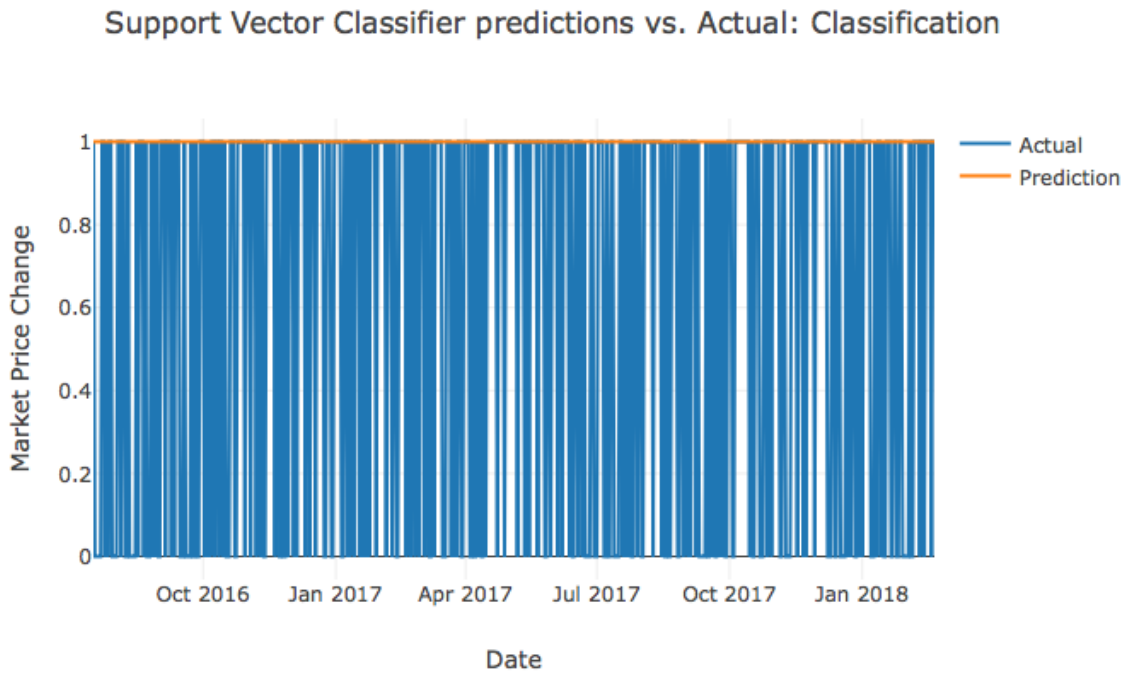
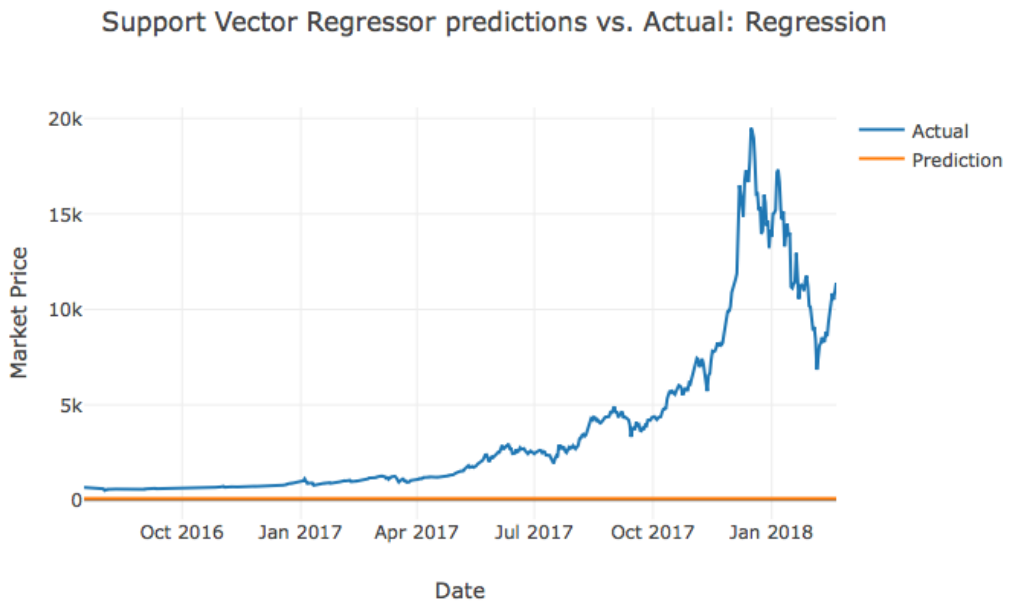


Figure 6.


Support Vector Regressor predictions vs. Actual: Regression
20k
15k
10k
5k
0
Market Price
Oct 2016
Jan 2017
Apr 2017
Jul 2017
Oct 2017
Jan 2018
Date
Actual
Prediction