Bitcoin Volatility: A Machine Learning Approach

Muhammed Ahmed
Applied Data Science
University of Georgia
Athens, GA 30602
msa65652@uga.edu

Justin Hooker
Applied Data Science
University of Georgia
Athens, GA 30602
jwh40896@uga.edu

Abstract

In this analysis we attempt to explain bitcoin's market price using a combination of 9 economic features. We perform this supervised regression task to understand the cryptocurrency's recent volatility. We believe that understanding features related to bitcoin will help us achieve higher accuracy when forecasting its future market value. In phase 1, we obtained monthly data from the FRED website [4] and 48-hour market price data from blockchain.info [2]. We found that averaging market and economic data into a uniform monthly dataset lost too many training examples. This led to severe overfitting and therefore poor generalization to unseen data. In phase 2, we obtained daily economic data from the FRED website [4] as well as daily market price data from coindesk.com [7]. The abundance of training examples allowed us develop a more stable model that captures more variance. After testing several machine learning algorithms, we achieved a 5-fold cross validation score of 95% with KNN regression on the unseen data.

Introduction

1.1 Motivation

Bitcoin is a consensus network that enables a new payment system and a completely digital money. It is the first decentralized peer-to-peer payment network that is powered by its users with no central authority. The potential for a functioning currency without a central authority is an exciting prospect for the future. Bitcoin can also be seen as the most prominent triple entry bookkeeping system in existence because all transactions are logged and readily available for anyone to see, while not revealing either party's information [1]. Bitcoin is a freely functioning currency whose price is solely determined by the market. There are only a limited number of bitcoins that can be mined which fends against inflation. Bitcoin remains trusted because of its transparency. All of its source code is available for any developer to review.

The exact functionality of Bitcoin, however, is yet to be determined. Many in the digital currency community envision a use case as an alternative to the current system reliant on debt and unstable inflationary pressure. After all, financial crises are a dime a dozen these days and it appears to be a correlated with debt levels. Figure 1 shows a map of countries indexed by their ratio of outstanding debt to the recent year's GDP (2013). GDP (Gross Domestic Product) is a cumulative sum of the value of all production factors measured that year. It is used to measure economic growth of a particular country on a macro scale. Considering GDP in relation to debt levels can provide further context to a country's financial soundness. Many of the countries with the highest debt ratios had already, or since 2013, experienced a debt crisis and required a federal bailout. These countries include Italy, Spain, Greece, Portugal, Ireland, Cyprus, and even certain industries within the United States. As part of Cyprus's negotiations with the IMF, their agreement required a levy of 15% on all deposits over a certain threshold. Furthermore, in response to fears of inflation, India imposed a ban on certain cash notes totalling 80% of current

cash in circulation while providing only a few weeks notice. These egregious acts create incentive for alternatives and provide the motivation to understand market fundamentals affecting Bitcoin's price.

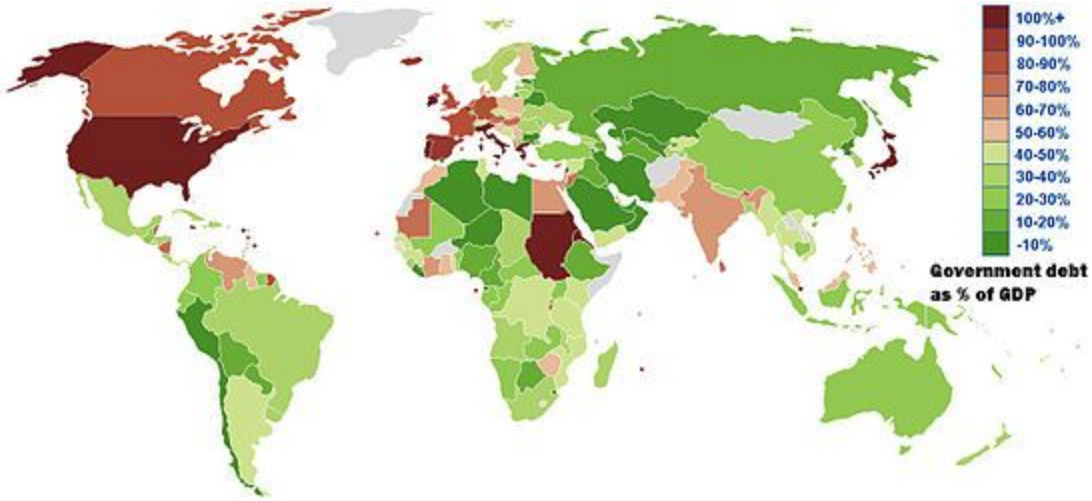

Figure 1

1.2 Volatility

Bitcoin is still in early development, both as a concept and as a functioning currency. In fact, many believe Bitcoin is better suited as a financial asset used to hedge certain market conditions. In either instance, both feel the effects of the extreme volatility in the market price. 2013 marked its meteoric rise to almost $1200 followed almost immediately by a sudden drop below $600 in the same year (figure 2). However, it has since shown a positive trend but with significant variation unfit for a functioning, mainstream currency.

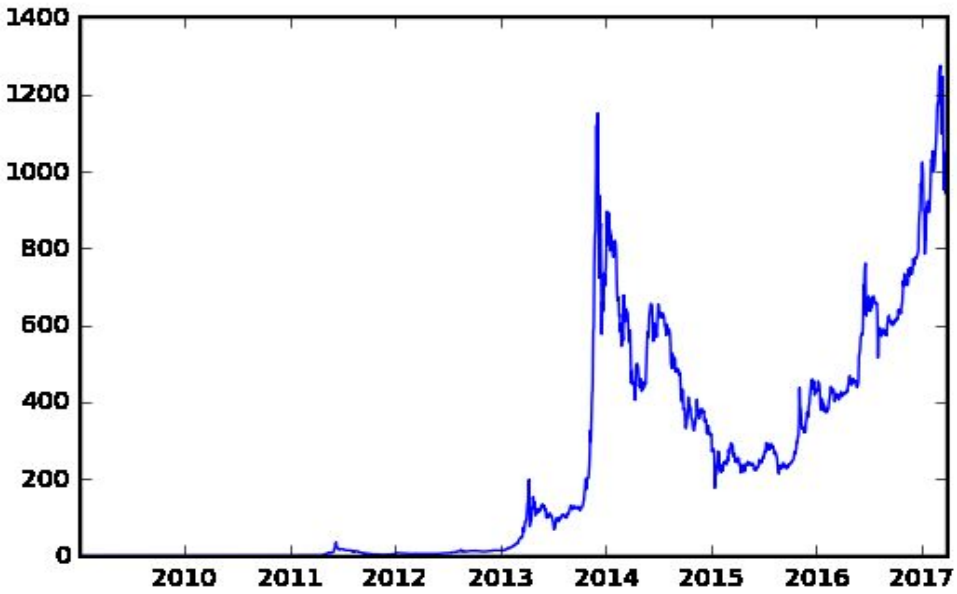

Figure 2

1.3 Relevant Work

With annual returns of 5,429%, -56%, 35%, 126% over the last four years, the year-over-year market performance of Bitcoin has certainly created a buzz in financial and academic circles alike. From what we can gather, the major focus has been geared towards the goal of explaining its price volatility in order to predict its future market value. For example, Puri [3] analyzed if public interest around Bitcoin impacts Bitcoin prices by using google searches for the keyword 'Bitcoin' and number of Bitcoin client downloads. Very few relevant works, however, have made attempts to explain bitcoin's price variation by use of machine learning algorithms pertaining to relevant economic data.

It is therefore, our goal to explore the option of Bitcoin as a digital currency and alternative to world currencies through the lens of a data scientist. To do this we first must understand its underlying fundamentals that explain the recent, extreme volatility. We hope to achieve this by examining the relationship between Bitcoin's market price and several economic variables.

Raw Data collection

1.1 Bitcoin Market Price

The Bitcoin market price, represented by the average USD market price across the major bitcoin exchanges, functions as the label in our model. The data source is http://www.coindesk.com/price [7]. The original intent was to use monthly samples since most of the relevant economic data accrued monthly. This required converting the bitcoin price to an average monthly price as part of the cleaning process. After initially exploring our data, we quickly realized monthly data was simply not enough samples over Bitcoin's timeline. We were therefore restricted to daily data, which brought about its own challenges (please see data-preprocessing section for further explanation), but there were still enough relevant features to continue with model evaluation

1.2 Economic Variables

Our economic variables serve as our features. These values come from the Federal Reserve Economic Data (FRED) website [4]. This website holds economic data and visualizations of over 470,000 US and international economic time series and adding various time series to a datalist for download.

Feature	Description [5]
VIX	The market's expectation of 30-day volatility.
OIL	Market price for West Texas Intermediate: benchmark for crude oil
1YT	One-Year Treasury. Represents minimum yield on low risk investment in the open market
DOLLAR	Daily index reflecting weighted average of the foreign exchange value for the USD
FFR	Central interest rate in the United States market
GOLD	Market price for set for gold bullion
SP500	500 stock index seen as a leading indicator of U.S. equities
INF	Inflation expectations measured over five year period
LIBOR	International benchmark for short-term interest rates

Data Processing

1.1 Data Preprocessing (process of creating a df from the csv, then aligning dates, dates selection, normalization)

All of the data was successfully downloaded in csv format and read using the pandas library to create a dataframe. The bitcoin price data is daily while the initial economic data was monthly. This required converting the bitcoin data to monthly by using the average price. A problem arose when we realized we would have less than 50 relevant samples to train and test from. As a result, we had to switch our economic data to daily to acquire more samples. This required interesting features to be dropped from the model such as the unemployment rate and industrial production that were only available as monthly data. Converting to daily data brought about additional preprocessing steps.

On a minor scale, the first 3-4 years of bitcoin's market price had to be excluded since it effectively hovered around zero during that timeframe. In an effort to collect the maximum number of testable samples, we decided to include all data after point which bitcoin's price hit $1. In hindsight, altering this approach could be a means to further improving our results(please see conclusion). Most significantly, the daily economic data is only recorded through weekdays while the bitcoin data is recorded through the weekend. This resulted in a misalignment of our data with the corresponding dates and index values. To resolve this issue, we added a day-of-the-week column for each sample and dropped rows whose day-of-the-week column were assigned 's'. Since this retained row index and replaced bitcoin with with 'NaN', we converted the bitcoin price column data to a list, dropping the 'NaN' values, then assigned the list to a Series and added back to our data frame.

1.2 Colinearity

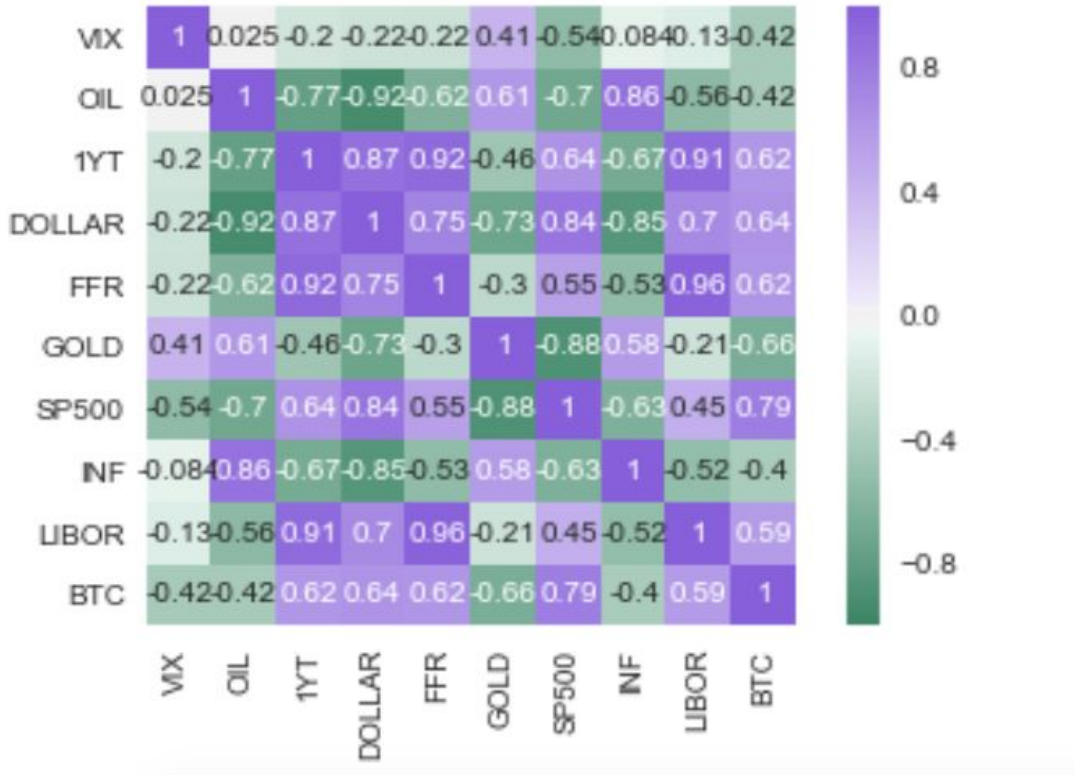

Figure 3

After thoroughly exploring our data, we became aware of two potential issues that need to be addressed before moving forward to model building and evaluation. Firstly, there appeared to be a non-linear relationship between most of our variables and the functioning of bitcoin's market price. This could potentially limit the number of available regression algorithms to fit the data. Secondly, correlation appeared high between a number of features, raising concerns of multicollinearity(please see figure 3). We expected some overlap between features as similar the Fed Funds Rate and LIBOR, but for the 1YT to exhibit 90+% correlation with the FFR and LIBOR was a shocking result that need to be addressed.

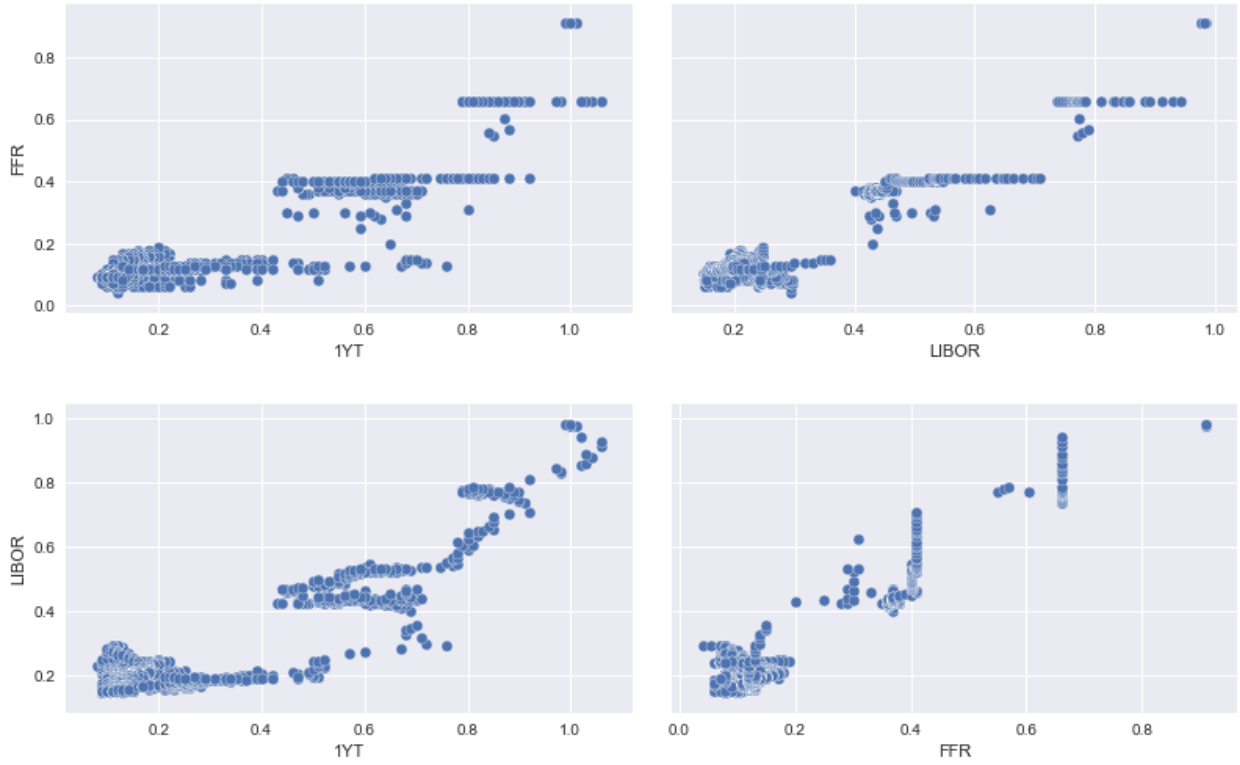

Figure 4 - Pair Plots of Correlated Features

1.3 Dimensionality Reduction

The combination of multicollinearity between features and a potentially nonlinear relationship with the label compounded the issues with our data. We decided to take an exhaustive approach by first controlling for multicollinearity through both dimensionality reduction and a regularization constraint and then control for nonlinearity by projecting our data through basis expansion.

Dimensionality reduction was accomplished through use of the PCA algorithm. We first fit the data, and plot the cumulative sum of the explained variance through each feature set. From figure 5, we could determine that reducing to three components retained enough of the variance while resolving the effects of multicollinearity. After fitting the projected dataset, testing resulted in a variance score of 0.58. This was the lowest of any tested model, lending credence to the idea of a nonlinear relationship between data and need for basis expansion instead of dimensionality reduction.

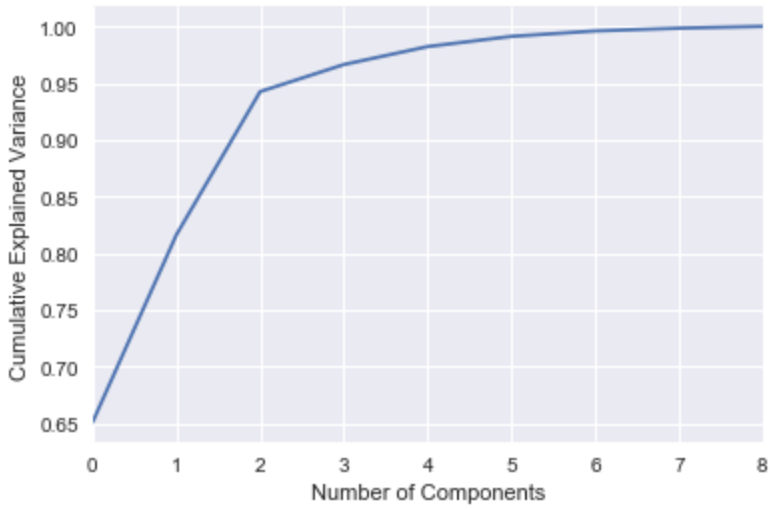

Figure 5

1.4 Features

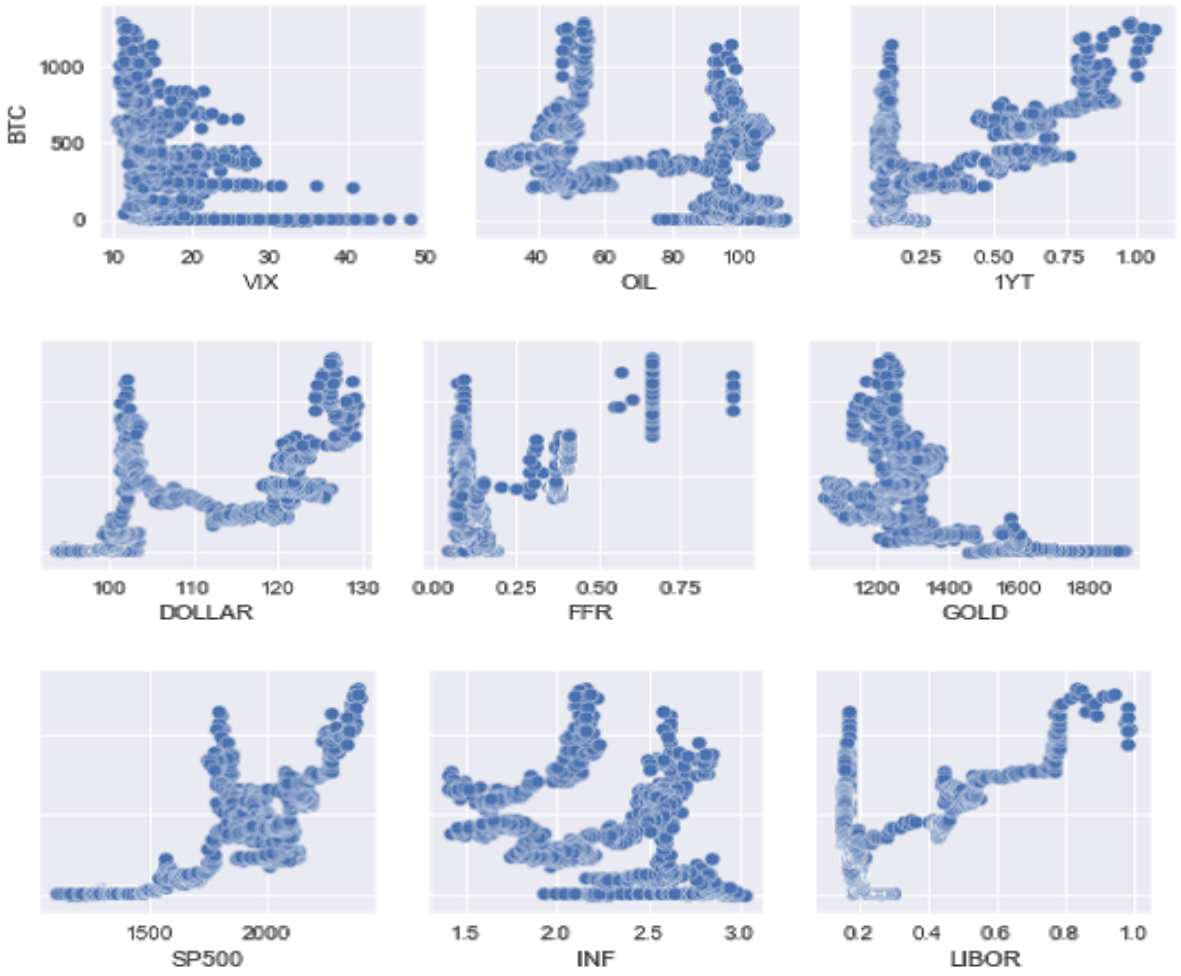

Figure 6 - Pair plots in relation to Bitcoin market price

Methods

1.1 Score and Error (any equations we may have utilized)

We used the R^2 score of predicted X and the mean squared error (MSE) of predictors to interpret the accuracy of our algorithms.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2$$

Where $\widehat{Y}$ is a vector of predictions and Y is a vector of observed values of n examples.

$$R^2 = 1 - \frac{regression\ sum\ of\ squares}{residual\ sum\ of\ squares}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Where $\widehat{Y}$ is a vector of predictions, $\overline{Y}$ is a vector of the mean of observed values and Y is a vector of observed values.

1.2 Model Selection (list of models we chose to use and why - Supervised Regression Models)

Due to high levels of multicollinearity among features (pre-PCA), we chose to employ ridge regression. Its regularization term penalizes model complexity and overfitting to the training data.

$$\widehat{\beta}^{ridge} = minimize_{\beta}\ ||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

In the process of building of our ridge regression model we created 50 different penalty terms between [50, -10] and then used the ridge cross-validation model from sckit-learn to fit the training data and return the penalty term with the lowest cross-validation error. The result was a constraint of 1.02. Next, we call the ridge constructor from the Ridge regression module, fit the training data and evaluate on the test data. The resulting variance score returned was 0.62, which was moderately better than PCA-transformed linear regression, but again provides evidence of nonlinearity. As mentioned previously, the linear regression model, post-PCA transformation accounted for 58% of total variation.

Next, we attempt to project our data through basis expansion using polynomial regression. We first fit our projected, training data to our polynomial model using the make_pipeline constructor for the LinearRegression and PolynomialFeatures modules. This involved iterating over degrees of range 5, fitting our model according to that expansion and returning the variance score from the testing set. The results follow from figure 7. Our model's predictive power appears to flatten out around degree 5, so a 3rd our 4th degree expansion using polynomial regression accounts for more than 80% of explained variance from our model.

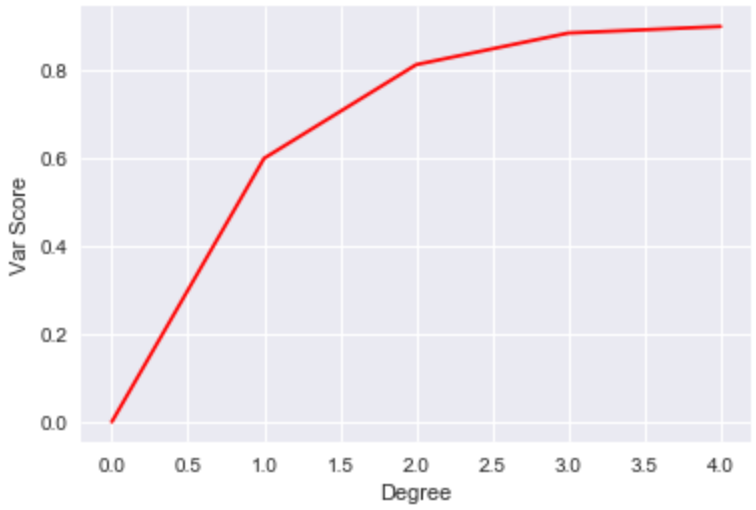

Figure 7 - Polynomial Regression

The final algorithm we fit to our data was K-Nearest-Neighbors. It was interesting to note that KNN has a regression module as part of scikit-learn and given its distinction as nonparametric, we thought it would be interesting to use it as part of our model evaluation. Our first task was to determine the 'k' parameter for the KNN algorithm. We chose to iterate from 1 to the square root of the total number of samples, which was approximately equal to 40, fit our projected, training data to the algorithm and return the cross-validation score that retains the highest accuracy.

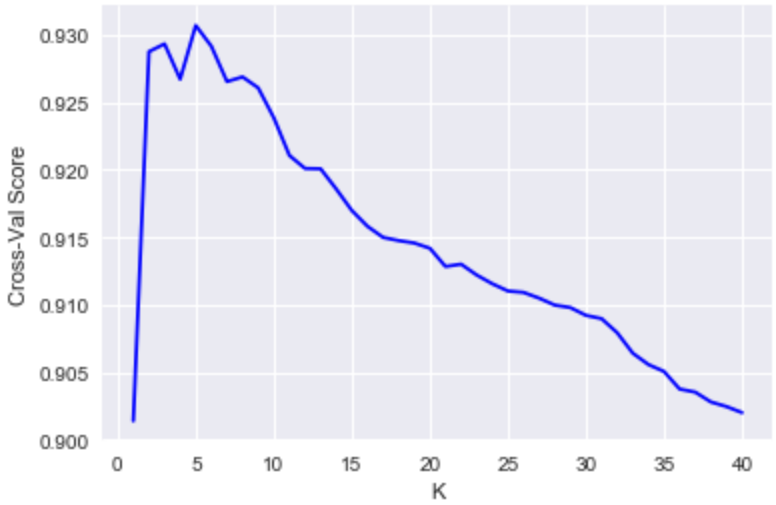

Figure 8 - KNN Cross-Validation Score

From figure 8, it is clear that a 'k' parameter of 5 returns the lowest error with a cross-validation score just over 0.93. Next we train the algorithm with a k parameter equal to 5 and evaluate the test sample. This process yielded a variance score equal to an astounding 0.95, implying 95% of total variance in our label is explained by the KNN model.

Results

1.1 Predictions

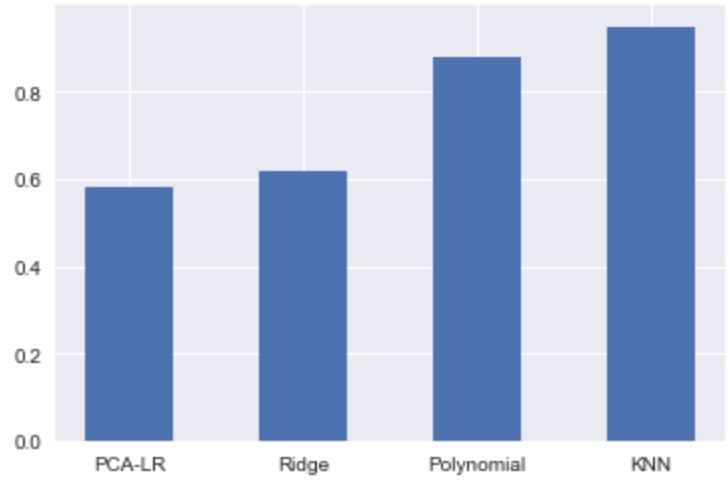

Figure 9 - Cross Validation Variance Score

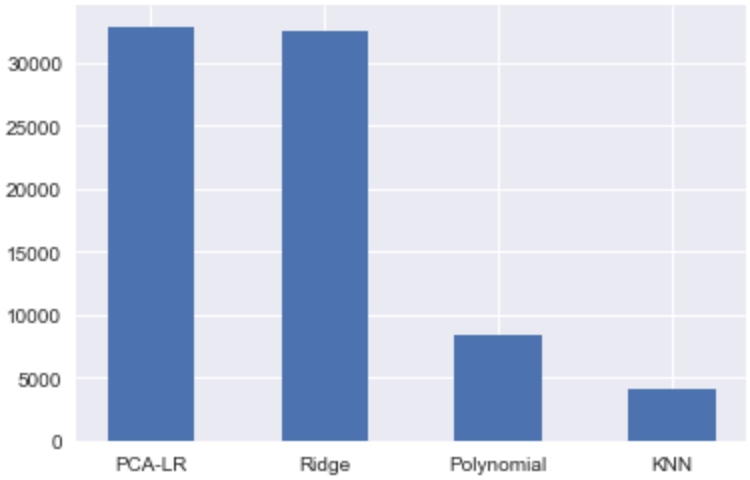

Figure 10 - MSE

1.2 Analysis

It is clear from the results of evaluation that the two models which performed the best involved either a functional increase in dimensionality through basis expansion or a non-parametric structure without linear inference. From these results, there exists convincing reason to infer some non-linear relationship between the features and label, or else it is too early in bitcoin's development to infer relationships and use predictive analytics, at least with a linear interpretation. Therefore, we cannot conclude with any reasonable certainty the exact relationship between bitcoin's price variation and the relevant features without further investigation. As a result, we also cannot yet determine if bitcoin's variation will stabilize to the point that it can be considered a legitimate alternative as a functioning currency. It would be interesting to restructure our label as a classification problem in an effort to predict an increase or decrease in market price.

Conclusion

1.1 Phase 1

Initially, in order to perform analysis we obtained 1000 monthly features from the Federal Reserve Economic Data (FRED) website. FRED holds economic data and visualizations of over 470,000 US and international economic time series [4]. We obtained the labels from blockchain.info in the form of 48-hour interval bitcoin market data points. This is the average USD market price across major bitcoin exchanges [2]. Because the data had different periods we averaged the 48-hour bitcoin data into monthly data in order to relate to the monthly economic data. This lead to a severe reduction of training examples. We were unable to produce an accurate bitcoin price prediction in our test case because the dimensionality of the dataset was far too large given the number of training examples. For this reason, our model had overfit to the training set and therefore did not generalize well on the test set. Our new objective was to seek daily datasets for both bitcoin market prices and economic variables.

1.2 Phase 2

The daily economic data we obtained is only recorded through weekdays while the bitcoin data is recorded through the weekend. This resulted in a misalignment of our data with the corresponding dates and index values. To resolve this issue, we added a day-of-the-week column for each sample and dropped rows whose day-of-the-week column were assigned 's'. Since this retained row index and replaced bitcoin with with 'NaN', we converted the bitcoin price column data to a list, dropping the 'NaN' values, then assigned the list to a Series and added back to our data frame. Using daily data allowed for a greater sample, leading to more reliable results and stronger testing generalization. Model evaluation based on algorithms such as PCA-transformed linear regression, ridge regression, polynomial regression, and K-Nearest Neighbors, resulted in KNN performing the best in terms of MSE and variance score. We can therefore, conclude that bitcoins relationship to the relevant economic features is nonlinear, or at least too early in its development to model based on a linear interpretation.

Looking Forward

Choosing smaller time intervals for more training examples could potentially boost our results. The analytics website okcoin.com offers bitcoin exchange data that is updated every 10-minutes via their API. This approach would give us 144x more the training examples daily data we currently have. All of that new data could provide us insight on patterns that occur throughout the day that were not previously expressed.

It is important to reiterate that our goal in this analysis was to explain bitcoin prices using economic variables. Moving forward however, it is our biggest aspiration to understand the bitcoin trend well enough to project its future market value. This means using features that indicate a rise or fall in bitcoin price days or weeks before they happen. This type of regression will require a deeper understanding of the cryptocurrency's economic indicators.

Another factor affecting the market price is Bitcoin's sensitivity to political news. In the past month, Japan has announced it is recognizing bitcoin as a legal form of tender, which caused a surge in bitcoin prices[6]. This represents a change that could not have been captured by our model in its current state. Even if we mastered its projection capabilities, we would not be able to foresee this because political data is not contained in our data set. One solution to this could be the implementation of natural language processing on relevant online news sources that contain the keyword 'Bitcoin' and determining whether the document has a positive or negative sentiment about the currency. This positive or negative sentiment could be stored for each day as a new feature in our data set then used for future predictions.

An interesting alternative to consider involves changing the scope of the project by turning it into a classification problem. We could do this by transforming the label 'price' into a binomial variable. Giving it a value of 1 for a price increase and 0 if not. In this way we can predict an increase or decrease in the next observed sample, potentially simplifying the problem.

References

[1] What is bitcoin? https://bitcoin.org/

[2] Blockchain Data. https://blockchain.info/

[3] Varun Puri. "Decrypting Bitcoin Prices and Adoption Rates using Google Search." http://scholarship.claremont.edu/cgi/viewcontent.cgi?article=2379&context=cmc_theses

[4] Financial Reserve Economic Data. https://fred.stlouisfed.org

[5] Investopedia. http://www.investopedia.com/

[6] "Japan's Bitcoin Law Goes into Effect Tomorrow."

http://www.coindesk.com/japan-bitcoin-law-effect-tomorrow/

[7] Prices. http://www.coindesk.com/price