**Sangmin Lee**

**CDS 492**

**05.27.2023**

**Technical Paper**

The project’s objective is to forecast Bitcoin’s price using a variety of economic indicators and to further determine the price trend so that investors may create a macroeconomic investing strategy. Numerical indices such as CPI, PPI, VIX, and Nasdaq Composite are examples of such possible indicators. Therefore, if it is possible to estimate Bitcoin’s price using the key economic indicators described above will be the hypothesis that is tested during the project.

The project appears to be an original mode that attempts to forecast the price of Bitcoin using macro indexes. The movement of large investors, or so-called “Whales,” as well as other factors that cannot be quantified, such as investor mental and the state of world politics, have also had an impact on the price of bitcoin. For such a reason, most investors attempt to deal with Bitcoin as a result after specific occurrences take place. Therefore, the project tries to forecast the price using specific indices and provide investors with a macro-flow of the price of bitcoin. Details of the project can be found in the following GitHub repository link: <https://github.com/sfiohgihase/cds492>.

Additionally, most investors attempt to predict the price using several investing strategies, including Fibonacci Retracements and MACD, as well as the chart’s shape. However, some contend that things are just coincidence and that the investors purposefully matched the price to such techniques. When it is successful in predicting the price using those qualified indices, such as the CPI and inflation rate, it is anticipated that this disagreement will lessen. As a result, it would be an original model that exclusively attempts to predict the price of Bitcoin using certain economic indices.

McNally et al. (2018), suggested the machine learning method to expect the price of Bitcoin. They followed CRISP-DM data mining methodology and utilized Bitcoin dataset that ranged from the 19th of August 2013 until the 19th of July 2016. Also, they adapted recurrent neural network (RNN) and the long short term memory (LSTM) model to deal with the Bitcoin price. However, they estimated their model failed because of high variance task. To be specific, the high variance task of Bitcoin nature makes it difficult to transpire this into impressive validation results (McNally et al., 2018). Also, although they tried to optimize the parameter selection through Bayesian optimization, it couldn’t guarantee appropriate results.

The dataset used for this project comprises the monthly price of bitcoin, CPI, PPI, DXY, VIX, and Nasdaq composite. The historical bitcoin price dataset can be obtained by clicking the following links: <https://www.investing.com/crypto/bitcoin/historical-data>, <https://beta.bls.gov/dataViewer/view/timeseries/CUSR0000SA0;jsessionid=F006EA618FF550B6AC325CF720C971FA>. In order for the project to show the macroeconomic trend of the price, it is necessary to obtain the CPI, PPI, and other indices mentioned in month. Those datasets can be found by following links: <https://beta.bls.gov/dataViewer/view/timeseries/CUSR0000SA0;jsessionid=F006EA618FF550B6AC325CF720C971FA>, <https://beta.bls.gov/dataViewer/view/timeseries/WPSFD4>, <https://finance.yahoo.com/quote/%5EIXIC/history?period1=1230768000&period2=1672444800&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true>, <https://finance.yahoo.com/quote/%5EVIX/history?period1=1230768000&period2=1672444800&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true>, <https://finance.yahoo.com/quote/%5ENYICDX/history?period1=1230768000&period2=1672444800&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true>.

Since the project focuses on the time period between August 2010 and December 2022, each of the six datasets has 149 rows while exploring the data (EDA). The start date is August 2010 because that is when Bitcoin price records first began to be kept. There are seven columns when viewing the Bitcoin price dataset: Date, Price, Open, High, Low, Volume and Percentage Change. However, this project will only require Date and Price because its goal is to forecast the price of Bitcoin using a specific price and sets of variables. Therefore, those Open, High, Low, Volume, and percentage change will be dropped in further process. The same goes for datasets obtained from Yahoo Finance, which include columns for Date, Open, High, Low, Close, and Adjusted Close. Since the project is attempting to concentrate on specific values and prices, all columns but Date and Adjusted Close will be removed.

Those six separate datasets have been combined as one data frame by an outer join for further processing once those extraneous columns have been removed. The merged data frame has 149 rows and 7 columns, data, bitcoin price, CPI, PPI, DXY, VIX, and Nasdaq values. Also, the “Data” column consists of object type while the other columns all include float64 type. The process of filling out the null value was not necessary because the merged dataset does not contain any empty (null) values. Below are the summaries of the statistics for each column.

A picture containing text, screenshot, font, number

Description automatically generated

The price column’s standard deviation value is relatively high compared to the other columns, which indicates that each value is well outside the range of the mean.

For the precise outcomes, the project also comprises Exploratory Data Analysis. To analyze the provided dataset and create visualizations, this project makes use of the “Seaborn” and “Matplotlib” libraries. The correlation matrix would be produced initially using those libraries.

A picture containing screenshot, square, text, rectangle

Description automatically generated

It was possible to see in the matrix that the correlation between price and the Nasdaq composite was 0.9, which is obviously larger than the correlation between the other indices. Since the project has been concentrating on the relationship between price and other columns, it has neglected the correlation value between CPI and PPI, which is the highest at 0.99. Additionally, when comparing values in a scatter plot, the graph for price and Nasdaq appears to be linear whereas the graphs for the other values are less interesting.

A picture containing text, diagram, map

Description automatically generated

More details would be seen in the link: <https://github.com/sfiohgihase/cds492>.

The goal of this project is to predict the Bitcoin price by regression methods. The regression approach is selected due to its effectiveness in modeling the relationships between independent variables and a dependent variable. In this project, the independent variables would be Consumer Price Index (CPI), Producer Price Index (PPI), VIX (Volatility Index), DXY (U.S. Dollar Index), and the Nasdaq Composite Index. By incorporating these variables into the regression model, it is expected to capture the potential relationships and use them to predict the future price movements of Bitcoin.

There would be several advantages when using regression methods. First, regression methods allow for the examination of the relationships between multiple independent variables and the dependent variable, enabling a comprehensive understanding of the factors that may influence Bitcoin’s price. Furthermore, regression models provide quantitative insights by assigning coefficients to each variable, indicating the magnitude and direction of their impact on the predicted outcome. Through this, it is expected to identify the key factors of Bitcoin price.

However, there are limitations to consider. The methodology assumes linearity between the variables and the Bitcoin price, which may not fully capture the complex dynamics and nonlinearity of the relationship. It also assumes relies on historical data and assumes that past relationships will continue to hold in the future, which may not always be the case due to changing market conditions. Despite these limitations, the regression methodology with the selected variables is expected to provide a data-driven approach to predict the Bitcoin price, offering valuable insights into the potential factors of its fluctuations.

This project adopted Principal Component Analysis (PCA), Bayesian Ridge, and polynomial regression methods. PCA is a useful method for dimensionality reduction, which efficiently addresses problems like multicollinearity. Also, it is able to enhance interpretability by downscaling a high-dimensional dataset into a lower dimensional space. Since the CPI and PPI are highly correlated as seen above, it was necessary to remove those variables and extract useful features, resulting in more consistent and understandable regression coefficients. Finally, by revealing underlying structures and patterns in the data, PCA helps interpret the regression model by providing a greater understanding of the relationships between variables. As a result, it is expected to improve the robustness, effectiveness, and interpretability of regression analysis by using PCA as a preprocessing step.

The project also uses polynomial regression as an approach. By capturing non-linear correlations between variables, it is a helpful tool that enhances conventional linear regression. Regression would perform better with the complex data patterns and increase the model’s predictive accuracy by using polynomial terms. Since there are several columns with different characteristics, it was necessary to find out the regression methods that can deal with those complexities. The model may also account for the curvature and fluctuation in the data by integrating polynomial terms, which results in a more accurate and nuanced evaluation of the correlation between the economic indices and the price of bitcoin.

The Bayesian Ridge model is the other approach used in the project. This model incorporates Bayesian inference, allowing for the inclusion of prior knowledge about the data and the uncertainty associated with the estimates. By adopting a Bayesian approach, the model accounts for the unpredictability and noise in the data to produce more reliable and precise predictions. The project’s ultimate purpose, as previously indicated, is to deal with a variety of economic indices to estimate the price of bitcoin. However, there are other numerous indices that affect price that could not be included in the project. To deal with those complex indices, it was necessary to adopt the regression methods that have flexibility as well as prediction accuracy over complex dataset.

Below chart is the result of the Bayesian Ridge Regression with the training data.

A picture containing text, screenshot, diagram, line

Description automatically generated

Although the training dataset cannot follow all outliers, it generally follows the actual points and seems to have produced results that are quite evident. Also, the r-square and MSE values, which demonstrate the model’s dependability, appear to be appropriate. It implies that neither the model is over-fitted nor under-fitted.

A picture containing text, screenshot, diagram, line

Description automatically generated

The Bayesian Ridge Regression test set shows an increase in r-square while a decrease in MSE, indicating that the model is becoming more accurate. Furthermore, the yellow points, which represent the projected data, largely coincide with the real data. However, given the limited dataset available for model training and the potential for the model to learn the dataset’s noise due to its complexity, the overfitting issue should be taken into account.

A picture containing text, screenshot, diagram, line

Description automatically generated

The graph above displays how the training dataset’s polynomial regression model performed. Although it appears to be similar to the prior Bayesian Ridge Regression model, features 1 and 2 differ from each other. Additionally, it displays a higher r-square value and a lower MSE value compared to the Bayesian Ridge model, which is projected to perform better. However, the training model does not accurately predict those points when the actual datapoints experience abrupt changes, such as those between Features 1 and 2.

A picture containing text, screenshot, diagram

Description automatically generated

Finally, the graph above compares the actual and predicted data with test dataset using Polynomial Regression. There is no significant change in the r-square values between the training and test datasets. However, the test dataset’s MSE value is higher than the training dataset’s. The dataset’s small size and outliers could be the factor. In order to display much more accurate projections for the future project, it is required to keep more complete datasets with various economic indices.

**References**

McNally, S., Roche, J., & Caton, S. (n.d.). *Predicting the price of Bitcoin using machine learning*. IEEE Xplore. <https://ieeexplore.ieee.org/document/8374483>