**Data Analytics – 6000 Level Report**

**Jake Lorenzo**

**12/10/2024**

**Introduction and Abstract**

I decided to do this research project because this was my last semester in college, and I am about to be released into the working world where I will be working and earning money for myself. Over the past decade, cryptocurrency has grown from a gimmick into a real financial asset class that is worth a lot of money. I’ve heard a lot of noise about investing in Cryptocurrencies but haven’t done any real research into investing in them and the associated risks. I thought this project would be the perfect opportunity to research cryptocurrencies in-depth so that when I am ready to invest my own money, I will be knowledgeable about current market conditions and can implement an investment strategy that mitigates risk/loss and maximizes potential gains.

**Hypothesis Tests Conducted in the Project**

1. **Correlation Between Cryptocurrency Prices and Stock Market Indices**
   * **Null Hypothesis (H₀):** There is no significant correlation between cryptocurrency prices and stock market indices.
   * **Alternative Hypothesis (H₁):** There is a significant positive correlation between cryptocurrency prices and stock market indices.
2. **Impact of Interest Rate Fluctuations on the Correlation Between Cryptocurrency Prices and Stock Market Indices (Impact of Macroeconomic Factors on BTC)**
   * **Null Hypothesis (H₀):** Changes in interest rates (e.g., Federal Funds Rate) have no significant effect on the relationship between cryptocurrency prices and stock market indices.
   * **Alternative Hypothesis (H₁):** Changes in interest rates significantly influence the relationship between cryptocurrency prices and stock market indices.
3. **Relationship Between Gold and Bitcoin Prices (Test whether BTC acts similarly to Gold as a safe-haven asset during periods of high market volatility)**
   * **Null Hypothesis (H₀):** There is no significant relationship between the price of Gold and Bitcoin.
   * **Alternative Hypothesis (H₁):** There is a significant relationship between the price of Gold and Bitcoin.

**Data Description and Preliminary Analysis:**

It took me a long time to find datasets that contained data that was not only applicable but also usable. After scouring through Kaggle, Google Datasets, etc. I was able to find all the datasets that I needed from the Federal Reserve Economic Data (FRED) website. I downloaded historical datasets for Bitcoin cryptocurrency (BTC), Ethereum cryptocurrency (ETH), S&P 500 Index (S&P, SnP, or SP500), Russell 3000 Index (R3000), CRSP U.S. Total Market Index (CRSP), Federal Funds Rate (DFF) and the Gold Commodity Volatility Index (Gold\_VIX). These seven datasets provided me with information on two different cryptocurrency coins, three stock market indices that are commonly known as the best representation of the stock market, as well as the Federal Funds Rate and Gold Commodity Volatility Index which represents macroeconomic factors. Once I downloaded each dataset, I performed preprocessing to put feature columns in the correct format (Date, numerical, etc.) then aligned the datasets by Date so they were the same size. I then merged the seven datasets into one dataset with ~3000 observations on eight variables (Date + each of the seven dataset’s values). From there, I backfilled some of the values in the dataset because some of the features were missing values (i.e. stock market indices don’t trade on weekends) where I returned the most recent value to the missing value to preserve data integrity. I then created some initial time-series plots to get an idea of the data's shape and how they compare to the other features. Lastly, I performed NA removal on the merged dataset.

A graph of a bitcoin and ethereum time series

Description automatically generated

A graph of a graph showing the price of bitcoin

Description automatically generated

**Analysis**

I performed a majority of the data preprocessing during the initial phases of importing the data, prior to merging it. Any further preprocessing I performed was model-specific and will be covered in the Model Development and Application Section. I initially ran a Pearson correlation matrix between the different factors to get an idea of how they interacted with each other.

A number of numbers on a white background

Description automatically generated

BTC and ETH have a significant positive correlation which was to be expected due to them being the two largest cryptocurrency coins. SP500, R3000, and CRSP are very highly correlated as well as showing some signs of multicollinearity. This shows that maybe I don’t have to incorporate all three stock market indices into the various models as SP500 is a good enough indicator of the stock market. The Federal Funds Rate has a moderate correlation (0.3766) with BTC which will be interesting to explore further. The implications of the high correlation between the stock market indices and BTC show that those factors will serve as good predictors in the models outlined below.

Next, I performed a Granger Causality Test to test the following: stock market indices influence on BTC, Federal Funds Rate influence on stock market indices, Federal Funds Rate influence on BTC, Gold Volatility Index influence on BTC, and if the Gold Volatility Index influences stock market indices. From those tests, we were able to get the following results:

A screenshot of a test results table

Description automatically generated

In terms of potential errors, uncertainties, or bias in the data; the datasets I downloaded from the FRED website were very usable and seemed to be highly accurate. However, in terms of the data itself, the high volatility of cryptocurrencies (BTC and ETH) could add noise to the data which could affect modeling. The frequency at which the Federal Funds Rate (DFF) is updated is much lower than the other features which could affect the various models. The exclusion of other macroeconomic factors could limit the extent to which it can be tested how much macroeconomic factors influence the price of BTC. Because all the data being tested is continuous it is important to think about a time-lagged effect between features. Lastly, while BTC and ETH are the two largest and most well-known cryptocurrencies, the exclusion of other smaller crypto coins could limit the extent to which this project will shed light on the cryptocurrency market.

**Model Development and Application**

**Random Forrest Model:**

This model required me to create a training and testing set to run the model on. The most accurate model I created used Ethereum, all three stock market indices, and both macroeconomic factors as independent variables to predict the price of Bitcoin..

rf\_model <- randomForest(BTC ~ ETH + SP500 + R3000 + CRSP + DFF + Gold\_VIX, data = train\_data, ntree = 500, importance = TRUE)

A graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing a graph showing

Description automatically generated

#MAE: 360.6366

#MSE: 509703.8

#RMSE: 713.9354

A highly accurate model with high predicted power. As seen in the plot above there are several high-leverage values that are still affecting the model. Further data munging would be required to remove these and refit the model.

**KNN Regression Model:**

The KNN regression model I created required some data munging before the model could correctly fit the data. I normalized all the data except the target variable (BTC) and the Date column then created a testing and training dataset from the normalized data. I first used the square root of the total observations in the dataset as the first value of K before creating a loop to iterate through different values of k to return the most accurate. K = 3 was the most accurate value for K and returned a model with the following attributes.

KNNpred <- knn.reg( train = normalized\_train[, 2:8], test = normalized\_test[, 2:8], y = normalized\_train$BTC, k = k)$pred

A graph showing a price of bitcoin

Description automatically generated

#MAE: 19.72659

#MSE: 1496.12

#RMSE: 38.67971

The KNN regression model was one of the most accurate models that I created which makes sense because all the data being tested was continuous numerical data which a KNN regression model performs particularly well on.

**Linear Model:**

I created a variety of different linear models to test the relationships between different predictors and features. The most accurate model I created used all the feature variables to predict BTC and returned the following error statistics:

#MAE: 4058.785

#MSE: 30906671

#RMSE: 5559.377

After plotting the above model, I saw there were still high-leverage values remaining in the dataset that were affecting the accuracy of the model. Once I identified and removed the high-leverage values I refitted the linear model and got some surprising results. While removing the high-leverage points from the dataset did improve the error statistics it did not improve the model as drastically as I expected. There are still elements of noise or outliers that are still impacting the model.

A graph of a line

Description automatically generated with medium confidence

#MAE: 3793.237

#MSE: 28736906

#RMSE: 5360.682

**Principal Component Analysis (PCA) Combined with KNN Regression:**

I ran a PCA analysis to combine the different predictors and see if there were similar trends between them that could help with the analysis.

A number of numbers on a white background

Description automatically generated

A graph of a graph

Description automatically generated

After looking through the information revealed from the PCA analysis, I decided to refit the KNN model from above to the PCA refined data. I used only the first two principal components to train the model. The most accurate model was the KNN PCA Regression with both crypto coins, three stock market indices, and the Federal Funds Rate. The addition of the gold volatility index into the model worsened the error statistics.

A graph with blue dots

Description automatically generated

#PCA MAE: 661.6412

#PCA MSE: 2522305

#PCA RMSE: 1588.177

**SVM (Linear, Radial, Tuned, and Polynomial Kernels):**

I created SVM models with Linear, Radial – Tuned, and Polynomial Kernels. The most accurate model used a linear kernel. I also tried a radial kernel as well as tuning the model to see if a change in Cost or Gamma values would increase the model’s performance. Tuning the model drastically increased the complexity of the model but did not yield any significant improvements. The SVM model with a polynomial Kernel had the same performance as the SVM with a radial Kernel. Overall, changing the Kernels did not significantly influence the model. SVM is not the best model to evaluate this dataset given its decent error statistics and high computational demand.

A graph showing a graph of red dots

Description automatically generated

#Test MAE(With All): 5296.489

#Test MSE(With All): 51051693

#Test RMSE(With All): 7145.047

**XGBoost:**

To properly implement an XGBoost model I had to convert the training and testing data to matrix form. After running the first iteration of the model, I decided to tune the hyperparameters which vastly improved the model’s performance. The most drastic changes were noticed after changing the following hyperparameters, learning rate, and the maximum depth of the trees. I also explored how many nRounds to run the model and found that the MAE of the model did not decrease after 1,500 nRounds.

A graph of a graph showing the price of bitcoin

Description automatically generated

#XGBoost Test MAE: 1065.575

#XGBoost Test MSE: 4964896

#XGBoost Test RMSE: 2228.205

The XGBoost model performed well with high accuracy scores. XGBoost models are very good with scaling large datasets as well as using hyperparameter tuning for (Lasso and Ridge) for regularization to reduce overfitting.

**Neural Networks (FNN, RNN, LTSM, GRU, CNN, and Hybrid):**

Feed-Forward Neural Network (FNN)

Recurrent Neural Network (RNN)

Long Short-Term Memory Neural Network (LTSM)

Gated Recurrent Unit Neural Network (GRU)

Convolutional Neural Network (CNN)

Hybrid Neural Network (CNN + RNN)

The above were the different types of neural networks that I created to test on the data. All the neural networks performed rather poorly on the dataset with similar error statistics, except for the Convolutional Neural Network (CNN) which had the highest accuracy scores out of all the models run.

A graph of data and data

Description automatically generated with medium confidence

#Mean Absolute Error (MAE): 15.15001

#Mean Squared Error (MSE): 468.1019

#Root Mean Squared Error (RMSE): 21.63566

**Conclusion and Discussion**

**A screenshot of a computer

Description automatically generated**

Above are the error statistics for the best-performing model of each type. The most accurate models are highlighted in green. I initially predicted that a linear model would be the most accurate due to the continuous nature of the values being tested. However, I think the extreme volatility of the cryptocurrencies added noise to the data which in turn affected the accuracy of the model. SVM models significantly increased the computational complexity of analyzing the data without providing any additional accuracy. The KNN regression model was very accurate which makes sense due to the data being continuous and KNN using the nearest neighbors to predict values. There were not a lot of features of different types present in the data, so PCA was ineffective as well as KMeans clustering. The Convolutional Neural Network (CNN) was the most accurate model created. I would recommend using the KNN regression model or the CNN model if trying to predict Bitcoin prices on the factors listed above.

**Hypothesis 1: Correlation Between Cryptocurrency Prices and Stock Market Indices**

* **Findings**: Significant time-lagged and positive correlation between cryptocurrency prices (BTC, ETH) and stock indices (S&P 500, Russell 3000).
* **Conclusion**: Null hypothesis rejected. Cryptocurrency prices are positively correlated with stock market indices.

**Hypothesis 2: Impact of Interest Rate Fluctuations on Crypto-Stock Relationship**

* **Findings**: Changes in the Federal Funds Rate influence the relationship between cryptocurrency prices and stock market indices but not enough to be deemed statistically significant.
* **Conclusion**: Inconclusive (interest rates have limited effect) other macroeconomic factors are required for further analysis

**Hypothesis 3: Relationship Between Gold and Bitcoin Prices**

* **Findings**: Significant inverse relationship between Gold and Bitcoin prices. The correlation coefficient is -0.35, and Gold's regression coefficient is negative (p < 0.01).
* **Conclusion**: Null hypothesis rejected. Bitcoin acts as a risk-on asset, inversely correlated with Gold (typical safe-haven asset during periods of high volatility or economic downturn).

If I had to rerun this experiment, I would choose a different problem to answer. I think that a classification or clustering problem would have been more interesting to compute. If I had to do the same experiment again, I would look for more datasets of different macroeconomic factors and dive deep into seeing which factors influence BTC the most. The Federal Funds Rate did not give enough information on its own to fully test out that hypothesis.

**References**

***CBOE Gold ETF volatility index*. FRED. (2024a, December 12).** [**https://fred.stlouisfed.org/series/GVZCLS**](https://fred.stlouisfed.org/series/GVZCLS)

***Coinbase Bitcoin*. FRED. (2024, December 10).** [**https://fred.stlouisfed.org/series/CBBTCUSD**](https://fred.stlouisfed.org/series/CBBTCUSD)

***Coinbase Ethereum*. FRED. (2024b, December 10).** [**https://fred.stlouisfed.org/series/CBETHUSD**](https://fred.stlouisfed.org/series/CBETHUSD)

***CRSP US total market index*. Center for Research in Security Prices. (2024, November 25).** [**https://www.crsp.org/indexes/crsp-us-total-market-index/**](https://www.crsp.org/indexes/crsp-us-total-market-index/)

***Federal funds effective rate*. FRED. (2024c, December 2).** [**https://fred.stlouisfed.org/series/FEDFUNDS**](https://fred.stlouisfed.org/series/FEDFUNDS)

***Russell 3000 historical data (RUA)*. Investing.com. (n.d.).** [**https://www.investing.com/indices/russell-3000-historical-data**](https://www.investing.com/indices/russell-3000-historical-data)

***S&P 500*. FRED. (2024c, December 12).**

[**https://fred.stlouisfed.org/series/SP500**](https://fred.stlouisfed.org/series/SP500)