Finndex: Forecasting and Validating Cryptocurrency Price Using Blockchain Statistics

Finn Frankis

Abstract

A multiple linear regression was modeled to forecast cryptocurrency price using blockchain factors of block count, transaction count, daily address count, Fear and Greed, and Google Trends data. The factors of address count, transaction count, and Fear and Greed were deemed the three optimal predictors using the best subsets modeling strategy, and an interaction term was incorporated between Fear and Greed and address count. A T-test analysis of the model, a check of the core assumptions – namely residuals v.s. fitted values and a normal probability plot – and a comparison of the predicted values with test values all revealed that the model has incredible predictive power over price and that it serves as an insightful pulse of cryptocurrency investor sentiment. Additionally, the model improves greatly in its predictive power upon the classic *finndex* model. As a result, the model has far-reaching implications for investors and cryptocurrency leaders alike.

Department of Statistics & Data Sciences
University of Texas at Austin
December 14th, 2020

Background and Significance

Cryptocurrency is a rapidly burgeoning technological field which has already had a dramatic impact on international finance. Two of the principles most fundamental to any blockchain are transparency and decentralization, two core tenets which afford cryptocurrency assets a security unmatched by even the most robust governments and international banking systems. As a result of these two concepts, cryptocurrency data are entirely open-source and accessible to the public; many statistics have been shown to have a strong correlation with cryptocurrency price. Forecasting currency price is crucial for currency traders, international travelers, and business owners alike.

Methods

The modeling began with five blockchain predictor variables regarding the Bitcoin cryptocurrency, namely block count (the number of blocks added to the chain on a given day), transaction count (the number of daily transactions taking place on the blockchain), address count (the number of new daily addresses registering to take part in the chain), Fear and Greed (a measure of investor sentiment factoring in technical indicators about trend), and Google Trends data (the search frequency for "Bitcoin" on Google); the single response variable was cryptocurrency price (how much one unit of the currency could be purchased for on the market). Each of these variables was retrived from open-source data repositories associated with the blockchain, and these variables spanned a timeframe of roughly nine months, from November 27th, 2019 to September 12th, 2020. The model used was a multiple linear regression model with one interaction term. Because price generally correlates linearly with each of these variables, any polynomial or non-linear terms were deemed unnecessary. To determine which combination of predictors was optimal, a best subsets prediction model was employed, inspecting Mallow's C_p, BIC, SSE, and adjusted R^2 to determine the optimal number of predictors as well as the optimal combination. After performing this simple model, a combination of interaction terms were attempted to determine if these additional terms would improve the adjusted R^2 value while limiting multicollinearity. A VIF test was also conducted to test for multicollinearity.

To validate the model, the data were divided into a training set and a test set in a roughly 70-30 ratio, respectively. The above-determined model was trained on the training set of roughly the first 6 months of data, then a series of successive forecasts were performed for the remaining 3 months. The MSE between the expected and forecasted data was computed to determine how the forecasts compare to the expected values. An additional graphical comparison was drawn between the test and predicted data.

To test whether or not the multiple linear regression model predicts more accurately than the *finndex* model with arbitrarily-selected weights, a "modified" finndex value was created which corresponds to a date index, where each day's value was calculated using the weights computed by the MLR model. This "modified" *finndex* was then normalized relative to the maximum value to make a fair comparison. For reference, the classic *finndex* model simply performs a normalized weighted average between each of the five predictor values, whereas the modified *finndex* value employs a combination of coefficients validated through hypothesis testing. The correlation between the two pairs of price and modified *finndex* and price and classic *finndex* was computed and the two values compared.

Results

Preliminary Analysis

Because the data were measured periodically over one-day increments, univariate analysis on a date level is not particularly meaningful. However, a study of the time series plots of each of the predictor variables provides some intriguing insights. Of all the predictors, only daily addresses shows a steady increasing trend. Each of the other predictors appears relatively stable, oscillating around a fixed value; each of these four other predictors show a steady drop in value corresponding with the beginning of the COVID-19 pandemic, for each of them measures cryptocurrency health or sentiment. A bivariate analysis reveals a strong correlation of transaction count, address count, and Fear and Greed with price; additionally, Fear and Greed and address count as well as transaction count are strongly correlated with one another, suggesting a potential issue of multicollinearity. Appendix A details the univariate analysis of the variables while Appendix B details the bivariate analysis.

Best Subsets & Model Generation

The best subsets modeling technique revealed that three predictors was the optimal number, for this number of predictors minimizes Mallow's C_p and BIC while maximizing adjusted R^2. Additionally, the modeling process reveals that the optimal three predictors are those of transaction count, address count, and Fear and Greed. When a model is generated with these predictors, each of their coefficients were deemed statistically significant by the result of a T-test. See Appendix C for the subsets graph as well as the results of the T-test.

Two simple linear regression models were performed between price and Fear and Greed with the data points filtered into two groups of approximately equal size based on whether the corresponding address count is greater than or less than the median address count. The respective slopes of the two groups were significantly different, which suggests that the Fear and Greed value has a significant impact on the mean effect of address count on the price value and that an interaction term between the two is necessary. When added to the model, each of the terms (interaction term included) still passes a simple T-test and the adjusted R^2 increases, suggesting that they are significant to the model. See Appendix D for analysis of interaction terms.

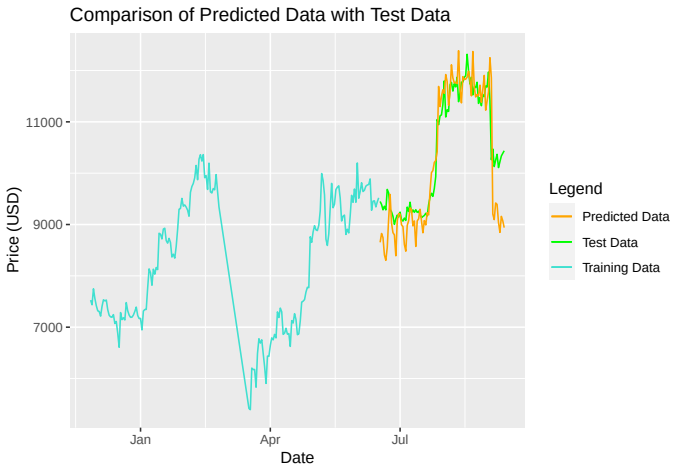
This new model was associated with unreasonable VIF values: Fear and Greed had a value of approximately 43.305 while the interaction term between Fear and Greed and address count had a value of approximately 64.575, both of which are significantly above the threshold value of 10. To address this issue of multicollinearity, the three core predictors were centered around their mean values and these centered predictors used to calculate a new interaction term, which reduced the VIF values for each of the model terms to less than 2.0.

Hypothesis Testing

To test the assumptions core to any linear regression, a residuals v.s. fitted values plot and a normal probability plot were generated, provided in Appendix E. The residuals v.s. fitted values plot shows relatively constant variance across the entire domain of fitted values, and there is no discernable pattern to the data points, which validates the linearity assumptions of constant variance between error terms. There are a few outlier points, but none were deemed significantly absurd to warrant removal from the dataset. The normal probability plot is impeccable, with no indication of heavy tails, validating the assumption that the error terms are normally distributed.

Model Validation

To validate that the generated model can accurately predict price over a wide future timeframe, the dataset were divided into a training and test set, with the model trained on the training set and the test set used to test future predictions. The model was used to predict approximately three months of future data: the $\sqrt{\text{MSE}}$ was calculated between the predicted data and expcted test data to be equal to $578.097, a value which, given that Bitcoin prices are typically given in the tens of thousands, validates the accuracy of the model. A graphical comparison is provided below, revealing that the model accurately predicts the large increase in price in late July as well as the sharp decline in early September.



Comparison to Classic Finndex

The classic *finndex* model was generated long before the onset of this project; it simply combined each of the five initial predictors through a weighted average and normalized the result by dividing by the maximum to scale each of the values from 0 to 1. The multiple linear regression in fact creates a modified *finndex* value, with each of the computed coefficients (0 for the predictors not included in the model) representing the weights of the terms in the weighted average and the interaction term representing an element of non-linearity to capture competing effects. One means of determining the success of the modeling performed in this project would be measuring whether or not the modified *finndex* value correlates more strongly with price than the classic *finndex* value. The modified *finndex* was computed by generating predictions from the model and dividing by the maximum value. Then, the linear correlation coefficients between classic *finndex* and Bitcoin price and modified *finndex* and Bitcoin price were computed, both over the date range corresponding to the test set. The correlation between the classic model and price was given to be 0.516 while the correlation between the MLR model and price was given to be 0.935.

Discussion & Conclusions

In summary, the MSE analysis, the graphical comparison, and the hypothesis testing serve to validate the model, supporting the hypothesis that a mutiple linear regression approach has a valuable place in the world of cryptocurrency. In more detail, given that the T-test results on the multiple regression model all produce probabilities significantly less than 0.05, it can be said with confidence that each of the chosen predictors has a slope significantly different with zero and thus is a strong predictor of price.

It must be noted that even though the model is displayed with the date on the x-axis, date does not serve as a predictor, and no time series analysis was performed. Rather, the state of each of the three predictors – address count, transaction count, and Fear and Greed – would need to be known to make a future prediction. These three predictors follow more well-defined trajectories than price, so their states can be more easily predicted. More generally, however, this form of modeling does not require any date-centric element (though this dataset is date-indexed): the model can be used to simulate a wide array of future economic conditions, and the effect on price can be carefully studied.

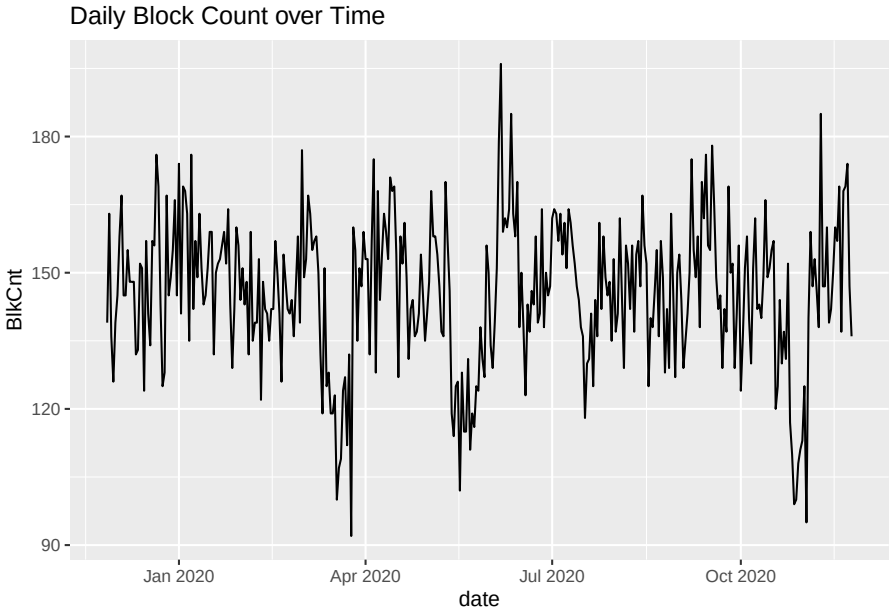
The implications of these results to the broader field of cryptocurrency are far-reaching. For instance, the fact that the new *finndex* model correlates much more strongly with price than the classic model suggests that it is a much stronger indicator of long-term sentiment. This discovery builds upon previous research by highlighting which predictors serve as more valuable indicators of cryptocurrency sentiment. Additionally, the comparison of the predicted data with test data shown in Figure 1 shows many points in time in which the modified *finndex* value in fact serves as a leading indicator of price – for instance, it predicts the meteoric rise beginning in mid-July a few days before it actually occurs – which serves as a testament to the predictive power of this new model over the price.

This predictive power expands the applications of the model dramatically. For instance, a currency trader seeking to time their investment decisions could apply this model to inform their long-term plan. Even that this model is a leading indicator of price by a few days is beyond sufficient, for currency trading occurs at a very rapid rate, with decisions often made in a matter of seconds. Furthermore, because the model provides a valuable indicator of cryptocurrency sentiment, it can be employed by cryptocurrency developers and leaders seeking to evaluate new cryptocurrency projects. One of the greatest new developments in cryptocurrency, which serves at times as both a blessing and a curse, is the proliferation of *altcoins*, other coins which employ a diverse array of new ideas. The sheer number of altcoins leads to marked difficulty in evaluating them; this numerical model allows for an effective means of efficiently comparing large number of altcoins. One valuable next step which would validate this assumption that the model could be applied to a wide range of altcoins would be to train the model on the same metrics over an array of time frames on an array of altcoins, calculating the MSE to determine if its predictive power and sentiment analysis applies to coins other than Bitcoin.

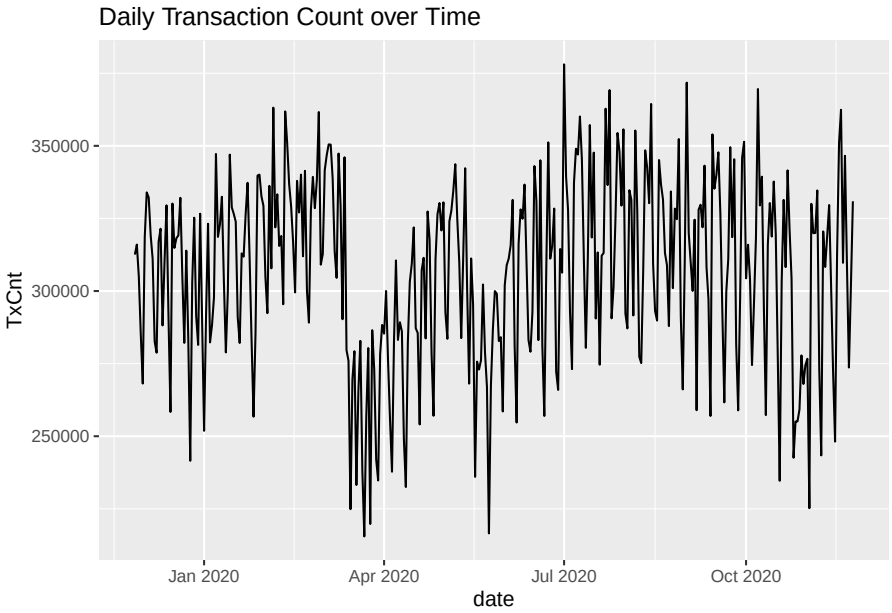
References

Appendix A: Univariate Analysis

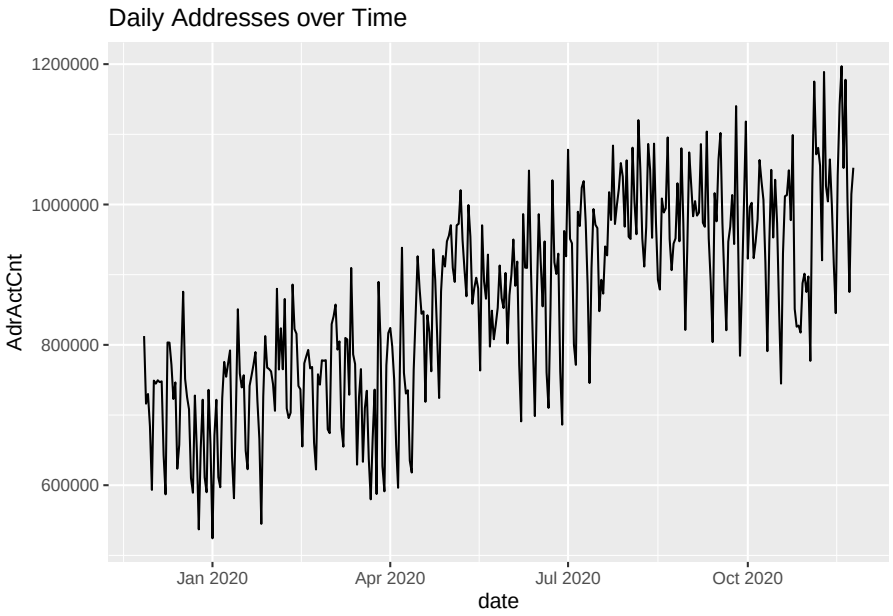
To assess the impact of each of the predictors on price, a study of each of their respective time series provides valuable insights.



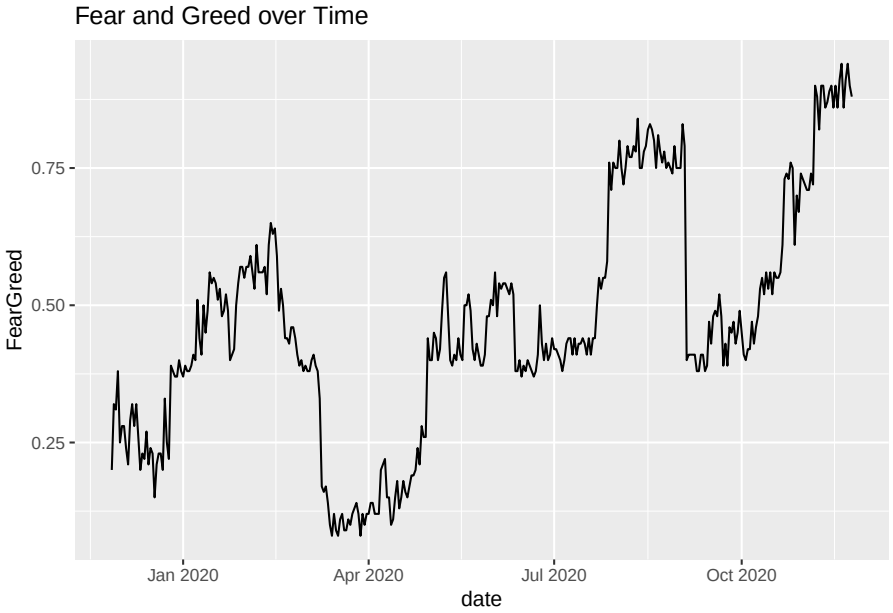
The plot of block count over time shows that block count is relatively stable, with no indication of seasonality or significant trend. It appears to range between 90 and 200 blocks per day, oscillating around a value of approximately 140 blocks per day. The dip beginning in mid-March, corresponding to the onset of the COVID-19 pandemic, is pronounced.



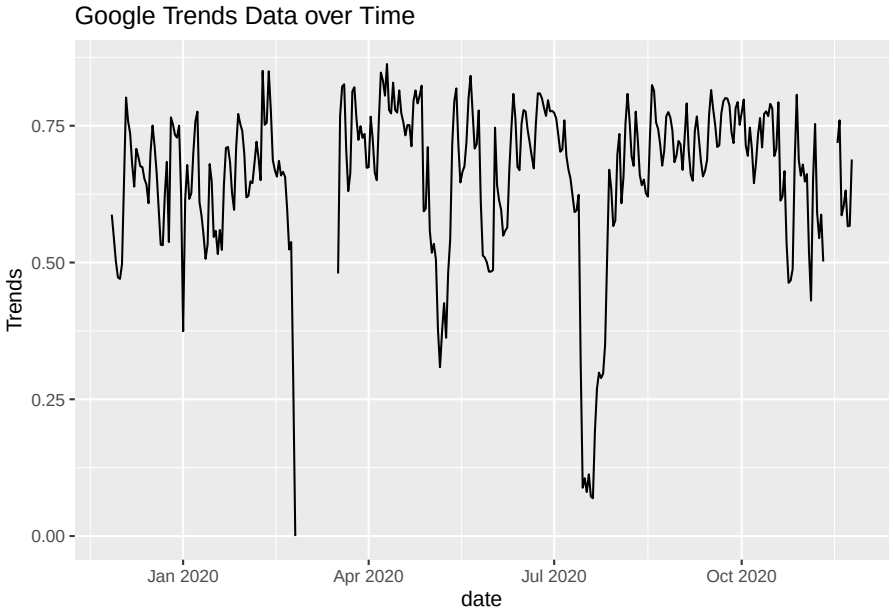
The daily transaction count, despite significantly dropping with the onset of the COVID-19 pandemic, appears to steadily increase throughout the late spring and early summer of 2020; however, it begins to decline from early August until mid-October, at which point it again skyrockets. The value appears to range between 225000 and 375000 transactions per day.



The daily address count shows a continually-increasing trend throughout the period, dipping very slightly at the beginning of the COVID-19 pandemic but almost immediately rebounding. Over the relevant time scale, it increases from rougly 500,000 to nearly 1.2 million addresses per day.

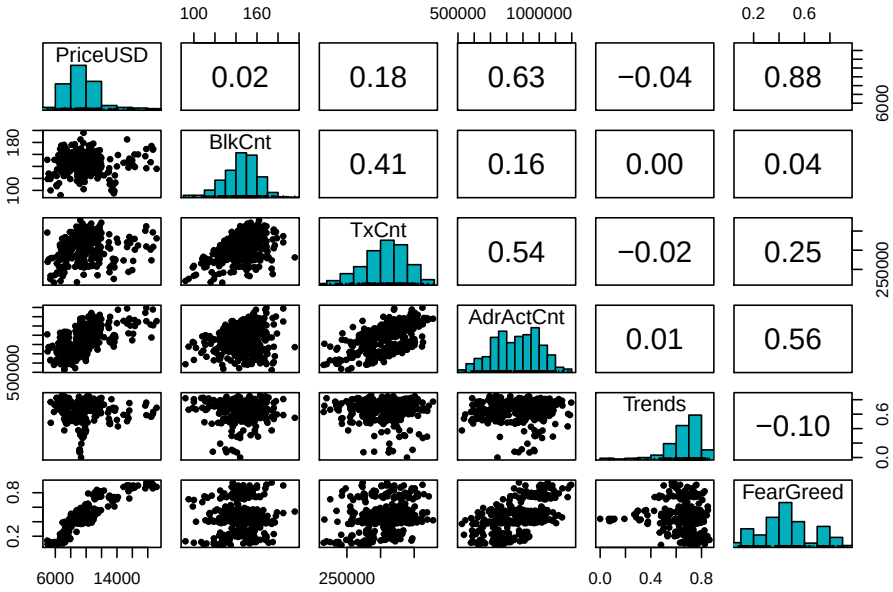


The Fear and Greed value appears generally much more stable than the other predictors, following a much more gradual trend rather than a trend marked by significant oscillation. Its trend is generally increasing over the course of the given timeframe, dipping to a low of around 0.18 with the COVID-19 pandemic but reaching a significant high of nearly 0.80, a valuable notable in that Fear and Greed is always a percentage ranging from 0 to 1.



The Google Trends data appears generally consistently high (mostly between 0.45 and 0.80; it is expressed as a percentage from 0 to 1), apart from a few points in time, namely mid-March 2020 and mid-July 2020, in which it drops significantly. In addition, there are multiple time frames characterized by missing data. One of the largest contiguous blocks of missing data occurs in the middle of the training data (mid-March 2020), suggesting that Google Trends might not be an insightful predictor.

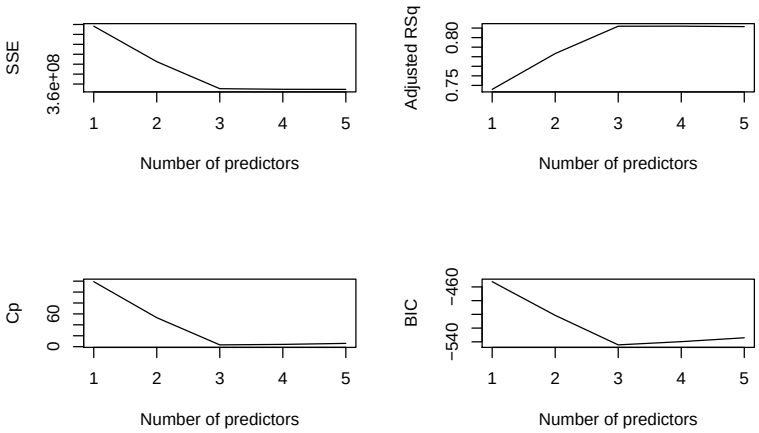
Appendix B: Bivariate Analysis



The scatterplot matrix can be studied to investigate potential multicollinearity issues between predictors and to determine which predictors are most strongly correlated with the response value of price. Block count and trends predictors both have very weak correlations with price, thus they might not be worth including in the model. With regards to multicollinearity, block count and transaction count are moderately correlated in a positive direction, as are transaction count with address count and address count with Fear and Greed. If the model has a high VIF, and particularly if interaction terms are added, it may be valuable to center the predictors around their mean values or to remove the predictors which are strongly correlated with one another to reduce multicollinearity. The correlation matrices do not indicate any strong non-linear patterns (all the weak correlations are caused by slope $= 0$, not by a non-linear pattern), so a polynomial regression model with higher-order terms is likely not necessary.

Appendix C: Best Subsets and Regression

The best subsets regression model-building strategy was employed, analyzing the below subsets plot showcasing C_p, BIC, SSE, and adjusted R^2 for the best subset model as a function of the number of predictors used.



These plots clearly reveal that 3 predictors is optimal, for 3 predictors minimizes Mallow's C_p and the BIC and maximizes adjusted R^2.Additionally, the SSE does not decrease by a significant amount for more than 3 predictors. Now, it must be determined must determine which 3 predictors would be optimal. The output of the model in R is shown below.

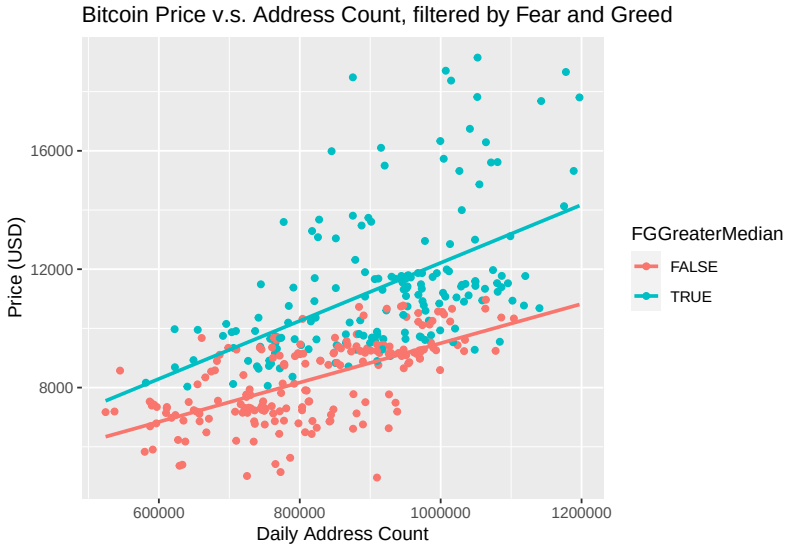
```
##          BlkCnt TxCnt AdrActCnt Trends FearGreed
## 1  ( 1 ) " "    " "   " "       " "    "*"
## 2  ( 1 ) " "    " "   "*"       " "    "*"
## 3  ( 1 ) " "    "*"   "*"       " "    "*"
## 4  ( 1 ) " "    "*"   "*"       "*"    "*"
## 5  ( 1 ) "*"    "*"   "*"       "*"    "*"
```

This output indicates that the best model with 3 predictors uses transaction count, address count, and Fear and Greed, which supports the results shown in the scatterplot matrices. The truncated output of a regression with each of these three predictors is shown below.

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.190e+03  5.430e+02   9.558  < 2e-16 ***
## TxCnt       -1.447e-02  2.086e-03  -6.936 1.88e-11 ***
## AdrActCnt    5.557e-03  5.480e-04  10.140  < 2e-16 ***
## FearGreed    8.982e+03  3.194e+02  28.120  < 2e-16 ***
##
## Residual standard error: 1051 on 360 degrees of freedom
## Multiple R-squared:  0.8239, Adjusted R-squared:  0.8225
## F-statistic: 561.5 on 3 and 360 DF,  p-value: < 2.2e-16
```

Appendix D: Interaction Terms

It would be reasonable to assume that, because Fear and Greed is a published value indicating investor sentiment, new Bitcoin addresses would be less likely to make a significant impact on the price if Fear and Greed is low. Therefore Fear and Greed likely has an impact on the effect of new addresses on price, and an interaction term may be necessary. This theory can be verified by plotting Price v.s. Address Count for two groups of data and performing a simple linear regression for each, one with all entries where Fear and Greed is above its median value and one with all entries where Fear and Greed is below its median value. The result of this plot is shown below.



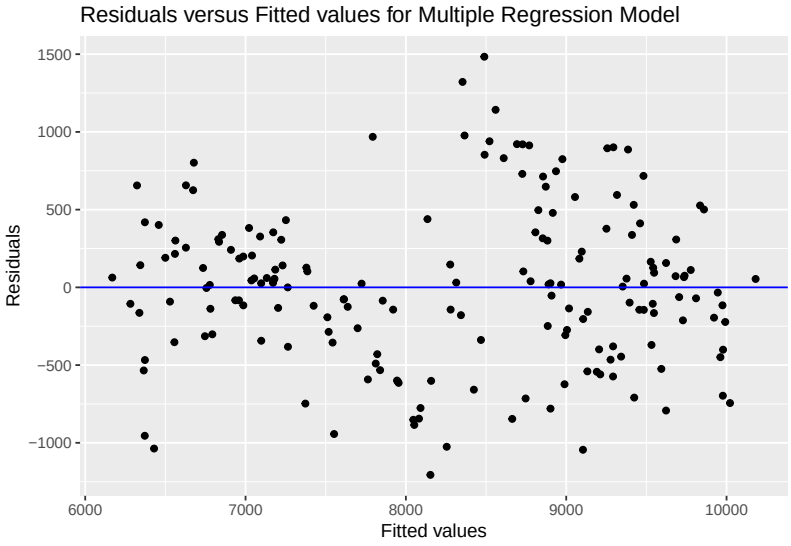
The slope of the line fitted to the data where Fear and Greed was less than the median is significantly less than that where Fear and Greed is greater than the median. Therefore address count has a more dramatic effect on price when Fear and Greed is higher, supporting the notion that Fear and Greed impacts buying decisions and suggesting the need for an interaction term. The truncated R output of the regression model with the incorporation of an interaction term, after the terms are centered to address multicollinearity, is shown below.

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.609e+03  6.262e+01 153.443  < 2e-16 ***
## TxCnt.c                -1.314e-02  2.100e-03  -6.258 1.11e-09 ***
## AdrActCnt.c             5.419e-03  5.428e-04   9.984  < 2e-16 ***
## FearGreed.c             8.847e+03  3.182e+02  27.809  < 2e-16 ***
## AdrActCnt.FearGreed.c   6.141e-03  1.916e-03   3.205  0.00147 **
## ---
##
## Residual standard error: 1037 on 359 degrees of freedom
## Multiple R-squared:  0.8288, Adjusted R-squared:  0.8269
## F-statistic: 434.5 on 4 and 359 DF,  p-value: < 2.2e-16
```

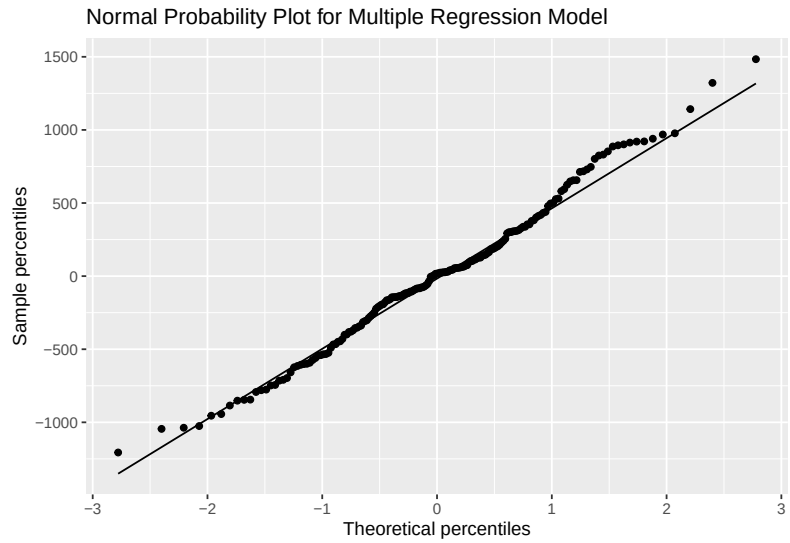
The p-value for each of the terms is significantly less than 0.05, suggesting that each of these terms makes a valuable impact on the model and that their corresponding slopes can be claimed with confidence to be significantly different from zero.

Appendix E: Hypothesis Testing

The following residuals v.s. fitted values plot and normal probability plot serve to test the core assumptions of the model, namely the constant variance of the error terms, a normally distributed set of errors, and an uncorrelated set of errors with no clear pattern.



The residuals v.s. fitted values plot shows a clear constant variance and no clear pattern to the data. There are a few potential outlier values for fitted values betewen 8000 and 9000 USD; however, these are not significantly different than the other residual values.



The normal probability plot closely follows the expected line across the range of theoretical percentiles, revealing that the errors do closely follow a normal distribution and validating the corresponding assumption.