Simrit Dhinsa, Mansi Kumar, Akhil Kemburu

Objective

We wanted to create various models that could predict the market cap of Bitcoin and Ethereum by training data on different features which we believed influenced the market cap of these cryptocurrencies.

Cleaning the Data

We began by going to https://bitinfocharts.com/ where there were complete data sets. We chose to cover Bitcoin and Ethereum because their data was most complete and comprehensive. After examining the source code for the pages, we found that all the data was in the script tag. In order to parse through the data, we learned Beautiful Soup and extracted the text from within script. However, there was still an issue - the text in the tag was not in a good state as there was a lot of extraneous text that was encoded within the lists. In order to make it legible it would require some further steps. We cleaned and parsed out this text for every row and ultimately converted each row to date time format. Additionally, we used log values instead of of the raw float values - as this helps curtail large numbers and is a general standard and allowed for further normalization.

Packages used

Beautiful Soup

This python package parses through HTML and XML documentation It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

Statsmodels

This was one of the main packages we used to assist us in creating the models. Statsmodels was used for our linear regression and quantile regression models. It provides a nice function for summarizing the output of the regressions.

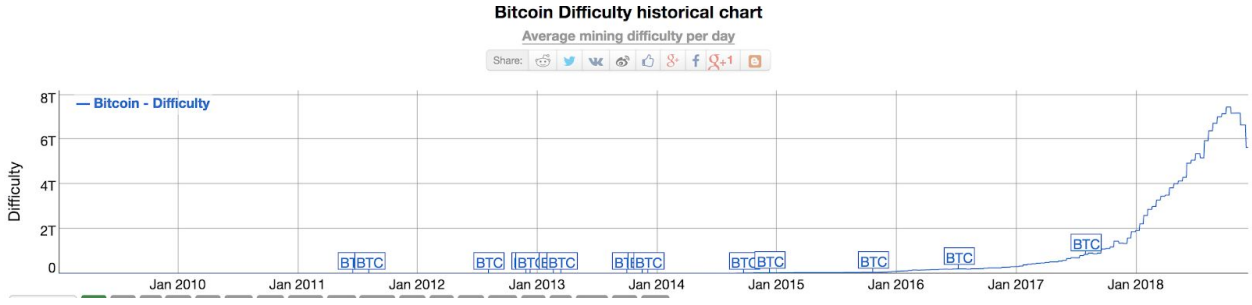
Sklearn

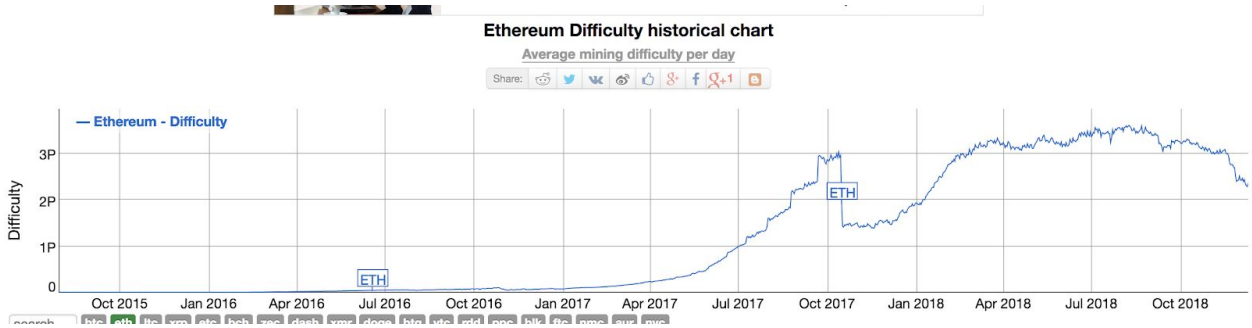
Statmodels is not compatible with Decision Tree Regressors so we had to use Sklearn. Unfortunately, sklearn does not provide a summarized output of the model.

Features

The features that we researched for this project were difficulty, hashrate, mining probability, average transaction value, and number of transactions. The details of these features will be discussed below, but we decided to use these features because we believed they are/should be correlated with changes in market capitalization. Furthermore, we also made sure to use features that were not perfectly correlated with market capitalization such as the price of the cryptocurrency.

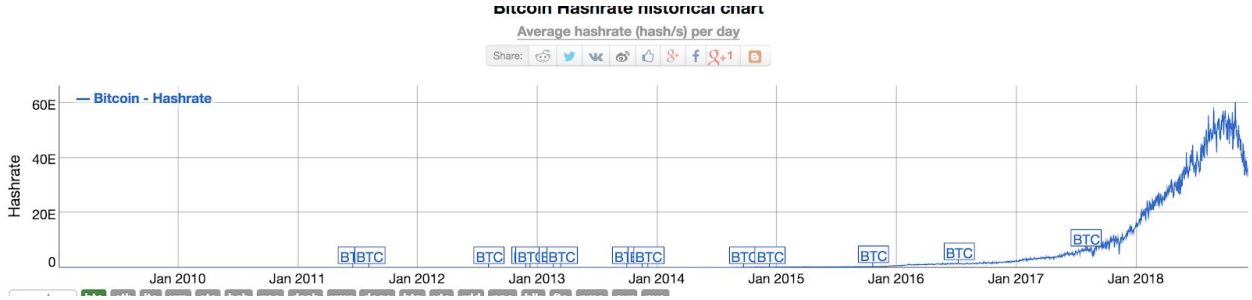
Difficulty

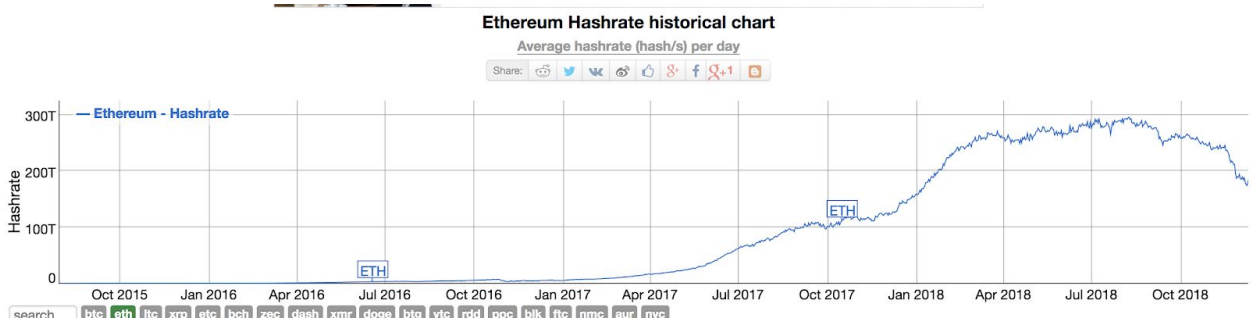




Bitcoin Difficulty illustrates the average mining difficulty per day. As illustrated the difficulty shoots up exponentially around January of 2018 and began to decrease significantly around October of 2017. The uptick in difficulty is around the same time that Bitcoin experienced an all time price high around December of 2017. For Ethereum, an uptick occurred around October of 2017. Another uptick began around January of 2018 and a more steady rise followed. The current difficulty of Ethereum is 2.359 p and Bitcoin's is 5.64 T

Hashrate

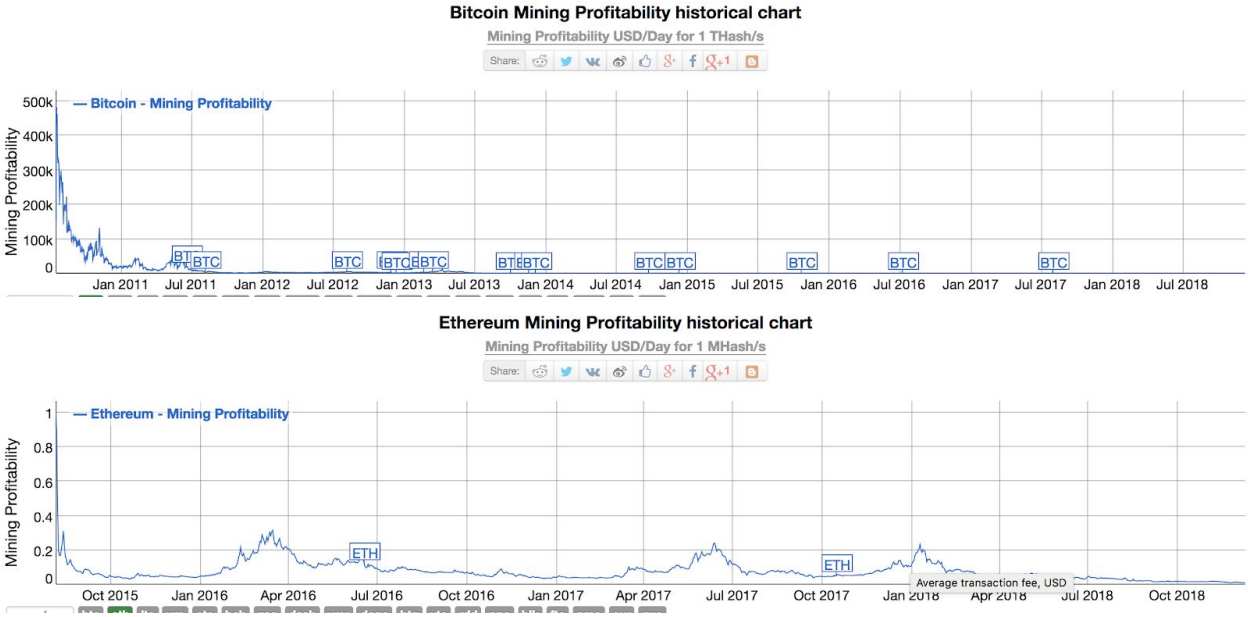




The Bitcoin hashrate indicates the number of hashes per day. The graph follows a similar upshoot as Bitcoin Difficulty. Currently, values are around 35.1E. A hash rate is the measure of

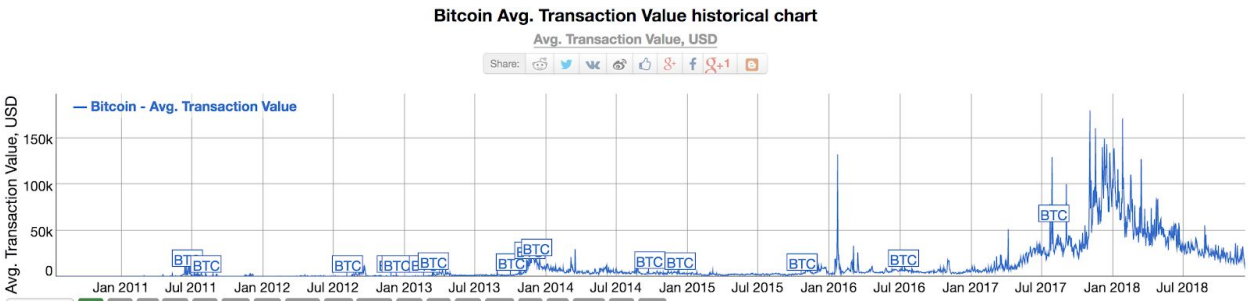
miner's performance. In other words, it is the hash function's output or it is the speed at which a miner solves the Bitcoin code. It makes sense that the graph would spike as the hash rate will increase as more people are on the platform. Ethereum's upticked around January of 2018 and plateaued in the following months and is now decreasing. Currently, it's value is 184.5 T.

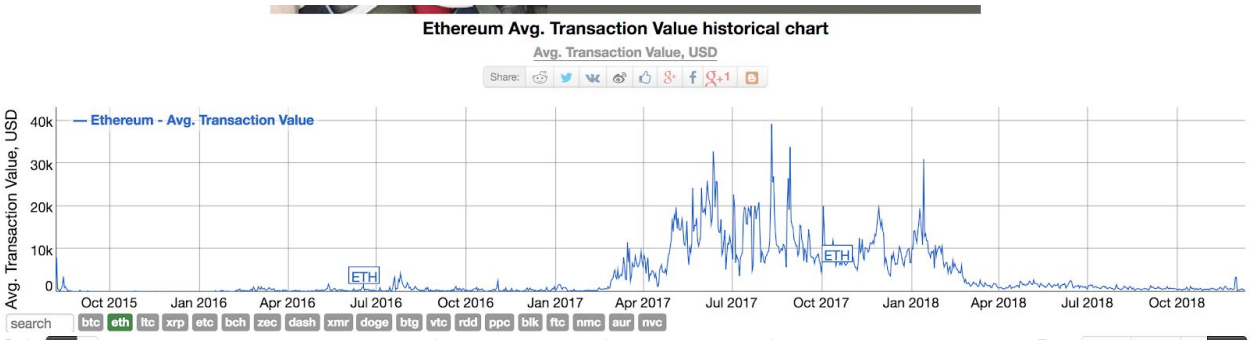
Mining Profitability



This graph indicates that mining profitability decreases with time. This makes sense as the more miners there are and the more users the profitability will decrease as other fees increase to compensate the miners. Today, mining profitability is near 0. Ethereum has followed similarly with a current value near 0

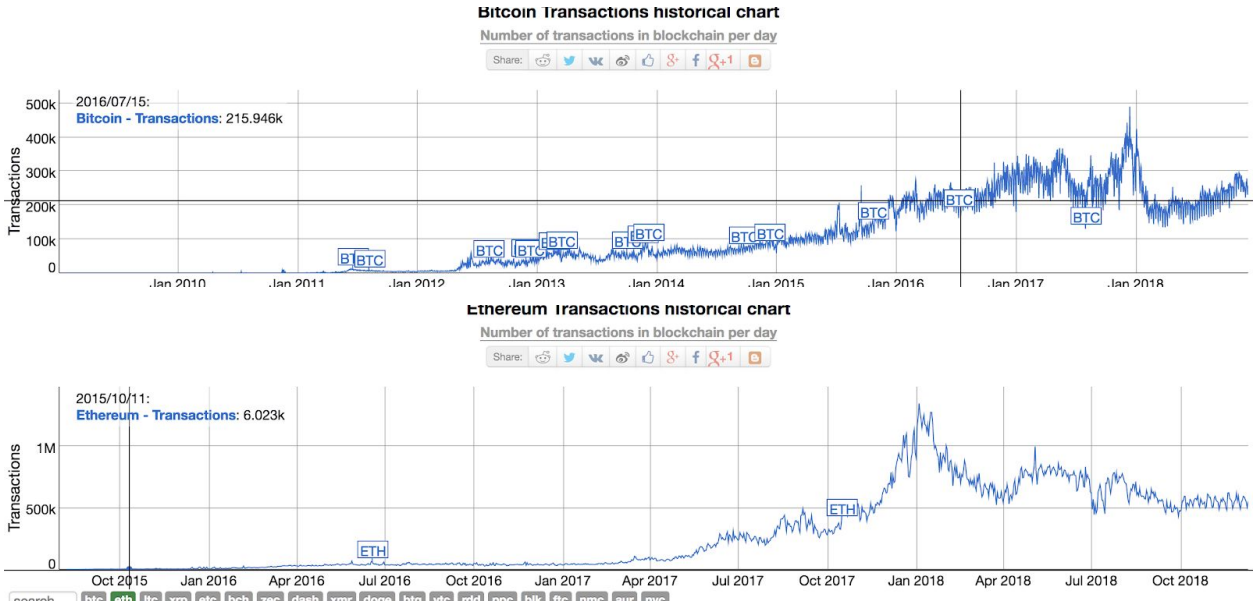
Avg. Transaction Value





This graph indicates various spikes with the largest being around January of 2018. A transaction is a transfer of Bitcoin value that is broadcast to the network and collected into blocks. The largest spike took place around the Bitcoin bubble peak - with an average value around 180.44 thousand dollars. Currently, the value is around 14.7 thousand dollars. Ethereum had a rapid and volatile spike between May of 2017 and February of 2018. ETH's current value is 418 dollars

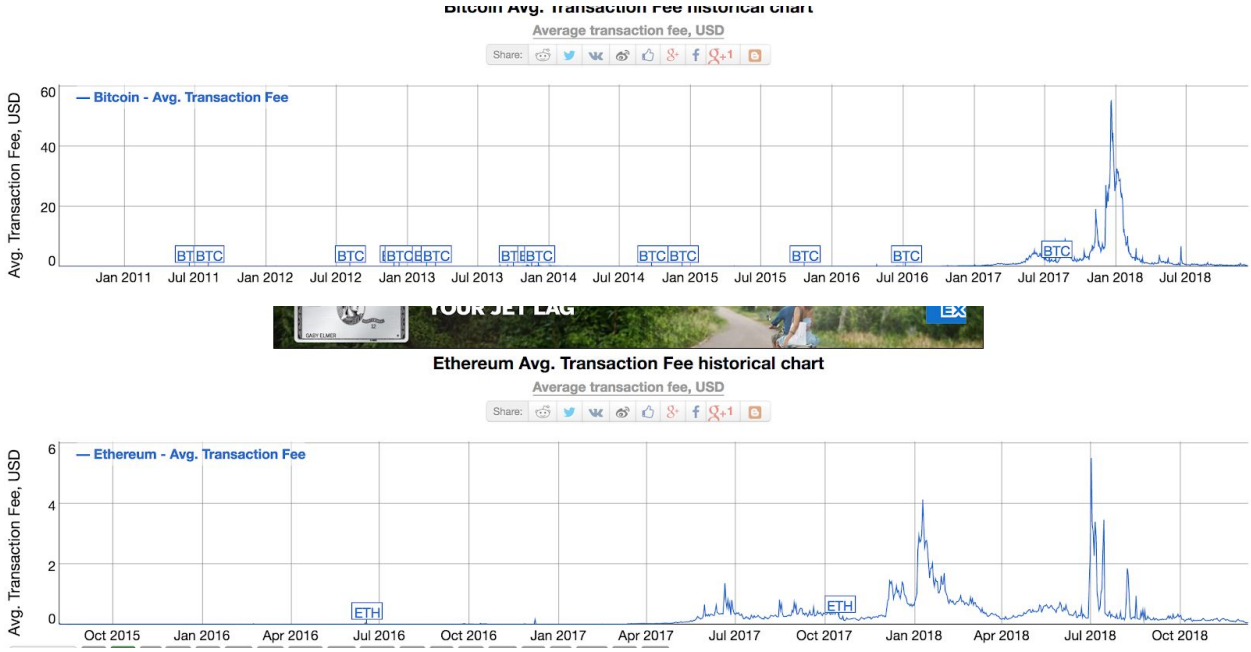
No. of Transactions



The number of transactions has followed a strong linear relationship. Again, the peak being in late 2017 early 2018 with a small uptick since. The number of transactions indicates the number

of bitcoin transactions per day between users. The current value is around 258.2 thousand transactions per day. Ethereum transactions per day is currently at 555 thousand

Avg. Transaction Fee



Transaction fees are the amount remaining when the value of all outputs in a transaction are subtracted from all inputs in a transaction; the fee is paid to the miner who includes that transaction in a block. The spike occurs around December of 2017, January of 2018. Currently, the value is around .42 USD. Ethereum has a current average transaction fee of .07 USD

For each model, we took a look at how each of these features impact the market cap of their respective cryptocurrencies. After iterating through the features, we narrowed the regressions down to only include the features that had a statistically significant effect on the market cap. We determined this by observing how the MSE would change as we added/eliminated variables, as well as the P>|t| columns of the summary outputs. This indicates the p-value of

the variable if it was tested against 0 (which indicates no effect). Using a p-value of .05, all the variables that were included had p-values less than this value.

Linear Regression

Linear Regression is one of the more basic models we used in trying to break apart and better understand our data. As implied by the name it is a linear approach to modelling data. Linear regression is one of the most widely used data modelling regression types not because most data is linear but rather because of its ability to give better scope into how to examine data and spot trends. We decided to use linear regression as it is the simplest model to use and thus serves as a great foundation. As we plotted the data points against various factors we saw what was comparable to a downward linear relationship.

Overfitting

When developing this model, we began to worry that the model was overfitting the data. The definition of overfitting is that the model picks up on a lot of noise and variation from the training data, but does not accurately represent or fit the predicted data. We found this to be inaccurate because the predicted data had an MSE lower than .03 for both the Bitcoin and Ethereum predictor (shown below). This showed us that the model we had built was extremely accurate.

In addition, we made sure that none of the features that we were perfectly correlated with the variable we were trying to predict, market capitalization. The pictures below detail what the correlation coefficients between each feature and the market capitalization. In this case, we see that mining profitability and market capitalization were highly correlated for both the bitcoin and ethereum models, but mining profitability alone did not yield the model with the lowest

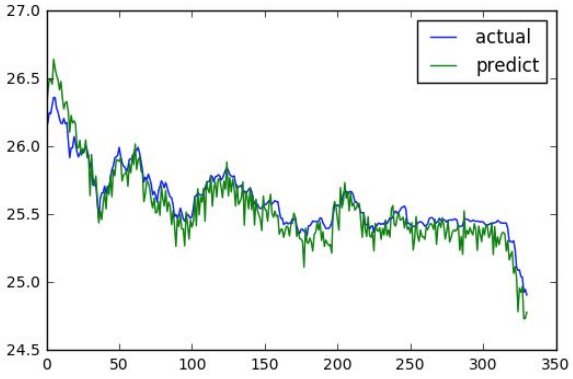
MSE. In fact, by including these specific set of features for each model were we able to achieve the lowest MSE. In addition, each of these features had a p-value of 0.00 which indicates that they are statistically significant to each model.

```
Correlation coefficient for ethereum linreg between difficulty & market cap: -0.405238
Correlation coefficient for ethereum linreg between hashrate & market cap: -0.314496
Correlation coefficient for ethereum linreg between mining prob. & market cap: 0.972790
Correlation coefficient for ethereum linreg between no. of transactions & market cap: 0.807776
Correlation coefficient for ethereum linreg between avg. transaction fee & market cap: 0.776994
```

```
The correlation coefficient between bitcoin difficulty and bitcoin market cap -0.837482
The correlation coefficient between bitcoin hashrate and bitcoin market cap -0.797565
The correlation coefficient between bitcoin market prof. and bitcoin market cap 0.931353
```

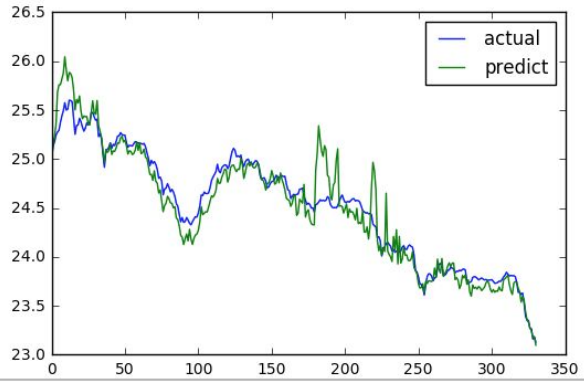
```
*****************************BTC Lin Reg ****************************
                            OLS Regression Results
==============================================================================
Dep. Variable:     Bitcoin Market Cap   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 5.016e+06
Date:                Mon, 10 Dec 2018   Prob (F-statistic):               0.00
Time:                        20:31:23   Log-Likelihood:                -404.66
No. Observations:                2725   AIC:                             815.3
Df Residuals:                    2722   BIC:                             833.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Bitcoin Difficulty      0.7043      0.006    124.128      0.000          0.693     0.715
Bitcoin Hashrate        0.4546      0.002    260.851      0.000          0.451     0.458
Bitcoin Mining Prob     1.0536      0.006    166.356      0.000          1.041     1.066
==============================================================================
Omnibus:                      238.287   Durbin-Watson:                   0.173
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              304.582
Skew:                           0.818   Prob(JB):                     7.26e-67
Kurtosis:                       3.070   Cond. No.                         68.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
R-squared of BTC running linreg is 0.862733
MSE of BTC running linreg is 0.012450
```



```
*****************************ETH Lin Reg ******************************
                            OLS Regression Results
==============================================================================
Dep. Variable:     Ethereum Market Cap   R-squared:                       0.998
Model:                             OLS   Adj. R-squared:                  0.998
Method:                  Least Squares   F-statistic:                 4.528e+05
Date:                 Mon, 10 Dec 2018   Prob (F-statistic):               0.00
Time:                         20:31:23   Log-Likelihood:                -1791.8
No. Observations:                2725   AIC:                             3592.
Df Residuals:                    2721   BIC:                             3615.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
=========
                                 coef    std err          t      P>|t|      [95.0% Co
nf. Int.]
------------------------------------------------------------------------------
---------
Ethereum Hashrate                0.7690      0.006    121.036      0.000         0.757
    0.781
Ethereum Mining Prob             0.6790      0.034     19.883      0.000         0.612
    0.746
Ethereum No. of Transactions     0.1052      0.017      6.375      0.000         0.073
    0.138
Ethereum Avg. Transaction Fee    0.2443      0.010     24.450      0.000         0.225
    0.264
==============================================================================
Omnibus:                     8674.031   Durbin-Watson:                   1.948
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        757021989.016
Skew:                          50.151   Prob(JB):                         0.00
Kurtosis:                    2583.172   Cond. No.                         70.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
Ethereum Features: Date, Ethereum Hashrate, Ethereum Mining Prob, Ethereum No. of Transa
ctions, Ethereum Avg. Transaction Fee
R-squared of ETH running linreg is 0.923339
MSE of ETH running linreg is 0.029414
```



Decision Tree Regressor

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A

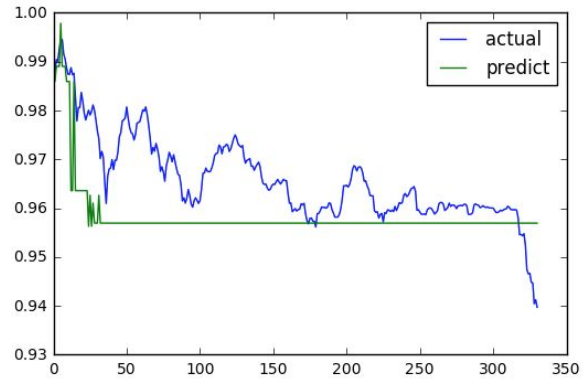
decision node has two or more branches each representing values for the attribute tested. Leaf nodes represent a decision on the numerical target. A big pro to using a decision tree regression model is its ability to take into account both categorical and numerical data. We decided to use the Decision Tree Model for a variety of reasons. First and foremost, Decision tree regression is great for data that has high variance. As the prices indicate Bitcoin and Ethereum fluctuate heavily and have high volatility. In this aspect, a decision tree model seemed apt. Moreover, after using a linear regression we wanted to implement a model that was not in the linear or polynomial family. This allowed for a more nuanced evaluation of the data. The figures below show the outputs of our decision tree models.

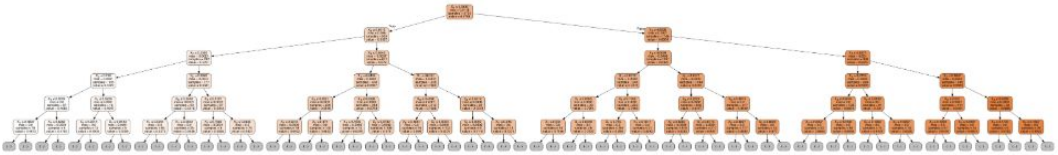
A reason why the Decision tree may not have been the closest fit to the data is as follows; it fit our training data exceedingly well, however decision tree models are very sensitive to any small variation and this was evident when we used it on the predicted data. This explains the various spikes as displayed above.

```
*****************************BTC Dec Tree ******************************
The fit of this BTC data with a Decision Trees model is 0.999998
Bitcoin range 0.054907
R-squared of BTC running Decision Trees is -1.794196
MSE of BTC running Decision Trees is 0.000107
```

```
/Users/simritdhinsa/anaconda/lib/python3.5/site-packages/ipykernel/__main__.py:14: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexi
ng.html#indexing-view-versus-copy
```

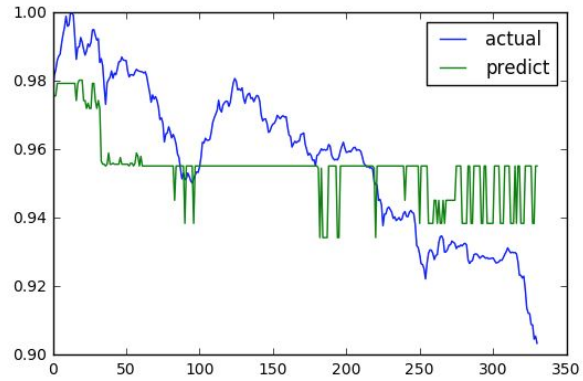


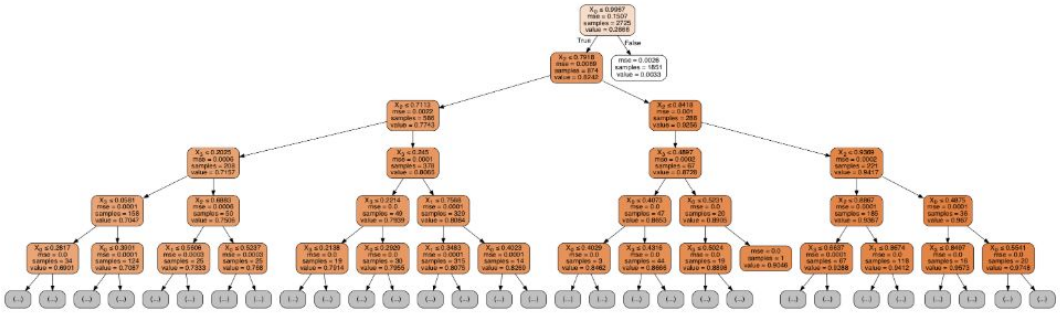


```
*****************************ETH Dec Tree ******************************
The fit of this ETH data with a Decision Trees model is 0.988425
R-squared of ETH running Decision Trees is -2.439173
MSE of ETH running Decision Trees is 0.000322
```

```
/Users/simritdhinsa/anaconda/lib/python3.5/site-packages/ipykernel/__main__.py:64: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexi
ng.html#indexing-view-versus-copy
```





Quantile Regression

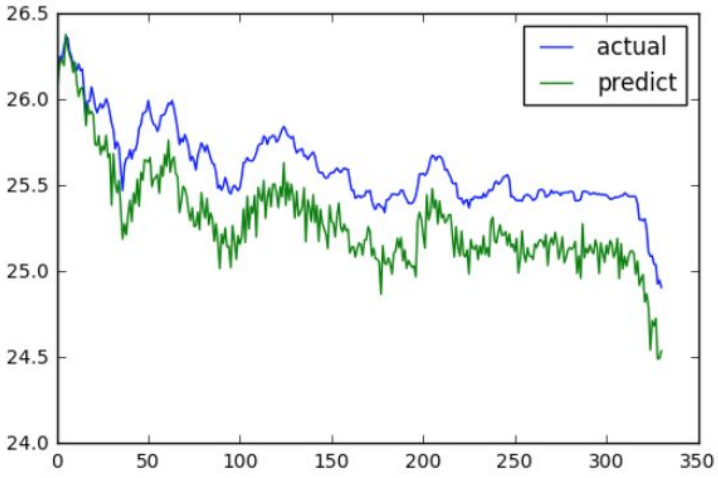
Essentially, quantile regression is the extension of linear regression and we use it when the conditions of linear regression are not applicable. Quantile regression is desired if conditional quantile functions are of interest. One advantage of quantile regression, relative to the ordinary least squares regression, is that the quantile regression estimates are more robust against outliers in the response measurements. However, the main attraction of quantile regression goes beyond that. Different measures of central tendency and statistical dispersion can be useful to obtain a more comprehensive analysis of the relationship between variables. As our

linear regression did quite well in our initial model, we decided to extend it through the quantile regression for Bitcoin. This proved to be very successful and allowed for great matching and prediction of the data.

```
*****************************BTC Quant Reg ******************************
                          QuantReg Regression Results
==============================================================================
Dep. Variable:     Bitcoin Market Cap   Pseudo R-squared:               0.8969
Model:                       QuantReg   Bandwidth:                     0.09602
Method:                Least Squares   Sparsity:                       0.7988
Date:                Mon, 10 Dec 2018   No. Observations:                2725
Time:                        20:33:31   Df Residuals:                    2722
                                        Df Model:                           3
================================================================================================
                           coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------------------
Bitcoin Difficulty       0.7083      0.008     85.883      0.000          0.692      0.724
Bitcoin Hashrate         0.4521      0.003    179.288      0.000          0.447      0.457
Bitcoin Mining Prob      1.0535      0.009    114.353      0.000          1.035      1.072
================================================================================================
Bitcoin range 1.455371
R-squared of BTC running Quantile Regression is 0.871327
MSE of BTC running Quantile is 0.011620
```

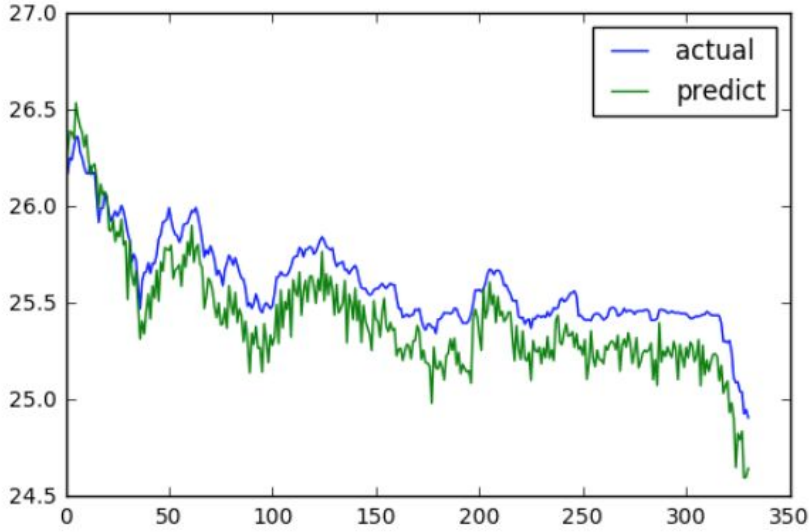
Q=.25

MSE = .1014



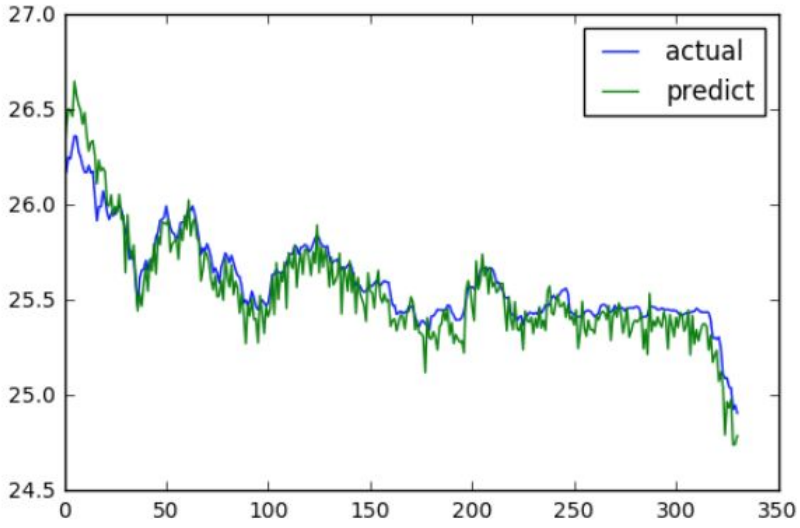
Q = .5

MSE = .0422



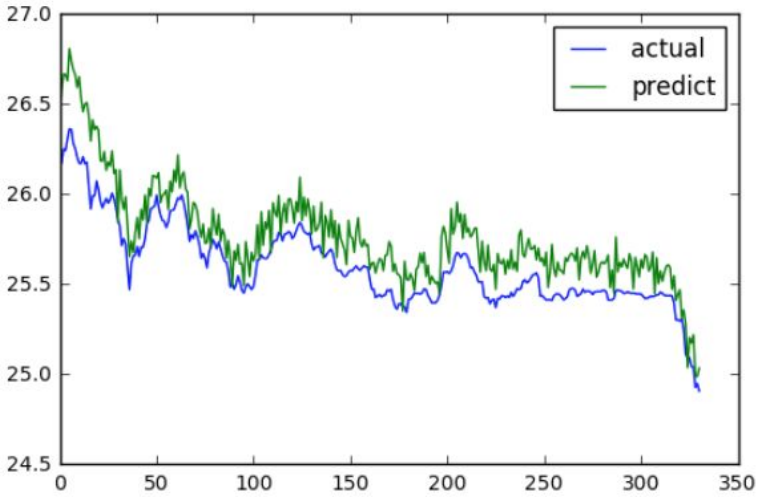
Q = .6

MSE = .0116



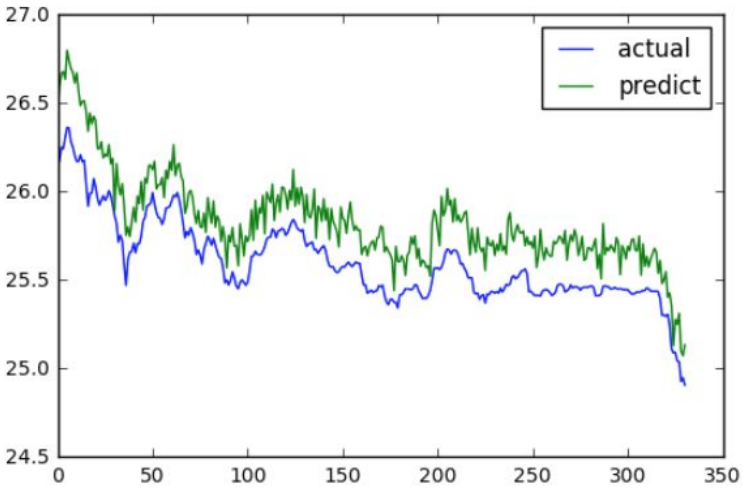
Q = .75

MSE = .0362



Q = .99

MSE = .0609



Conclusion

In reference to the models, the models based in linear regression performed the best. Quantile followed by linear regression were the best fit models and provided the best understanding of the data. Quantile Regression allowed for a pseudo- R squared near .9 which indicated very promising results and confirmed our intuition that it was a good model. Linear Regression, was more stellar than Quantile and provided an excelled framework for predictions. It's R-Squared was 1 and indicated that the data could be fitted to a linear model - which was why we further experimented Quantile. Lastly, the decision tree model was the least predictive of the data. It was not the best fit but did provide us the added authority to claim that linear and linear related models were what needed to be used to predict the data.

On a more holistic level this project allowed us to better visualize how machine learning functions within the fintech space. It allowed for us to bridge the understanding of the two units covered in the class. One, the use of blockchain and two, market evaluation. By bridging these two units and applying various algorithmic models to better predict market cap, we were able to understand various models as well as their applicability to blockchain currencies - such as Bitcoin and Ethereum. This also provided a lot of groundwork for how to better anticipate price or market cap change. To elaborate, by being able to understand which factors had the most impact on market cap/price it gave us better insight into how this might be applied if evaluating other blockchain currencies.