Bitcoin Linear Regression:
A Correlation Exploration

Brian McMahon
2 February 2018

Objective

1. Introduce a predictive model of the price of Bitcoin
2. Explore significant underlying features of the model
3. Provide key insights and takeaways

Model Overview

1. Predictive model for the price of Bitcoin
2. Standard linear regression
    a. As opposed to time series analysis; factors into cross-validation assumptions)
3. Three features with high correlation to price of Bitcoin; regularization was evaluated but deemed not necessary

Feature Exploration

Target features not significantly influenced by the price of Bitcoin

	logBTCPrice
logBTCVol	-0.430067
logGold	-0.296716
logNoTxns	0.408713
logAvgBlkSz	0.580606
logUniqueAddresses	0.770760
logETHPrice	0.936719
logNasdaq	0.958115
logTxFees	0.958548
logHashRate	0.967749
logInterest	0.977783
logCrypto Market Cap	0.987122
logCostperTxn	0.987860
logBTCPrice	1.000000

	logBTCPrice	logNasdaq	logInterest	logTxFees
logNasdaq	0.958115	1.000000	0.945212	0.948694
logTxFees	0.958548	0.948694	0.957145	1.000000
logInterest	0.977783	0.945212	1.000000	0.957145
logBTCPrice	1.000000	0.958115	0.977783	0.958548

3 Features contribute to R^2 of ~97%

Dep. Variable:	logBTCPrice	**R-squared:**	0.968
Model:	OLS	**Adj. R-squared:**	0.968
Method:	Least Squares	**F-statistic:**	4647.
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	0.00
Time:	11:40:55	**Log-Likelihood:**	127.12
No. Observations:	465	**AIC:**	-246.2
Df Residuals:	461	**BIC:**	-229.7
Df Model:	3		
Covariance Type:	nonrobust		

Accuracy 👍

Fit 👍

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-17.4446	2.346	-7.435	0.000	-22.055	-12.834
logInterest	0.5879	0.031	19.129	0.000	0.528	0.648
logNasdaq	2.6151	0.286	9.142	0.000	2.053	3.177
logTxFees	0.0826	0.020	4.060	0.000	0.043	0.123

Variable Contribution 👍

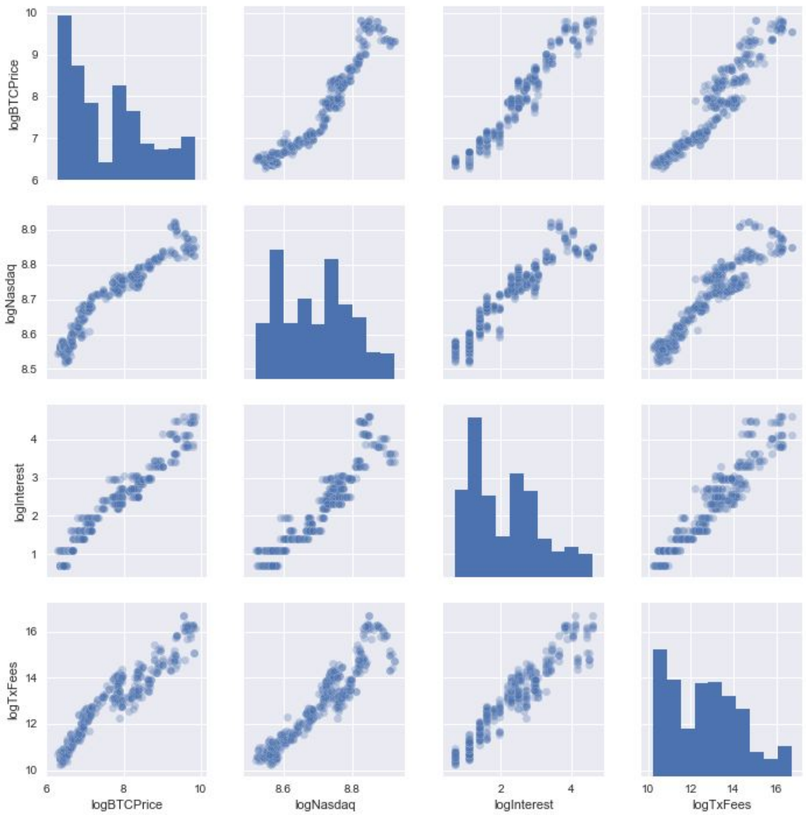
Omnibus:	7.400	**Durbin-Watson:**	0.203
Prob(Omnibus):	0.025	**Jarque-Bera (JB):**	7.008
Skew:	-0.255	**Prob(JB):**	0.0301
Kurtosis:	2.680	**Cond. No.**	4.31e+03

Normal Distribution 👍

Minimal autocorrelation 👍

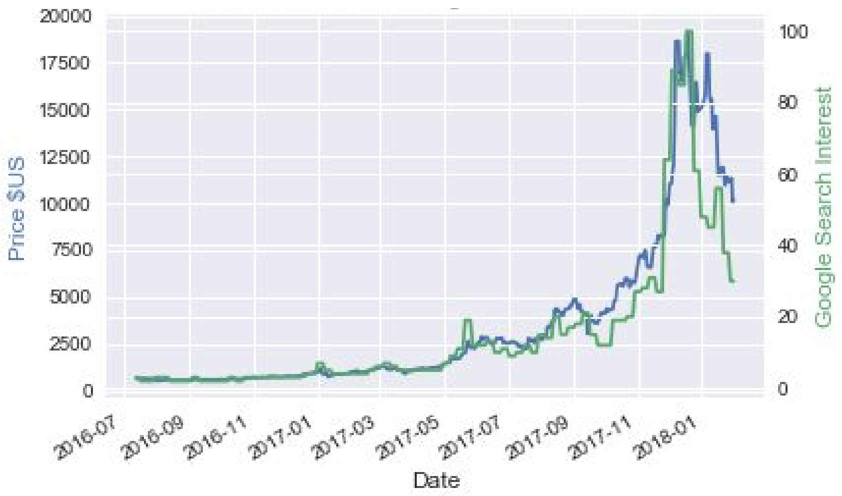
Significant multicollinearity may exist 👎👎👎

Correlation (and Multicollinearity)

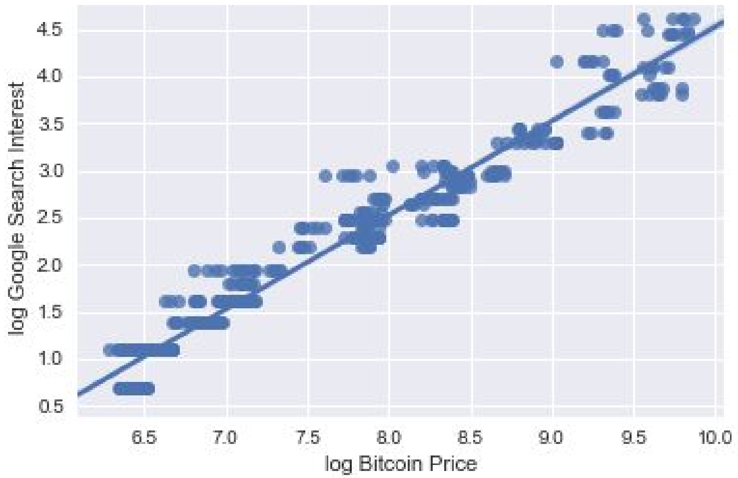


Google Search Interest at R^2 of 95%

Correlated Relationship



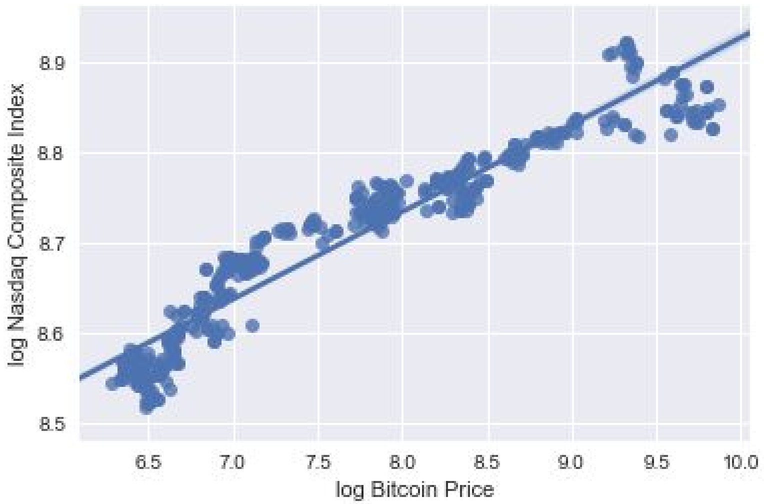
Single Feature Regression



Nasdaq Index at R^2 of 92%

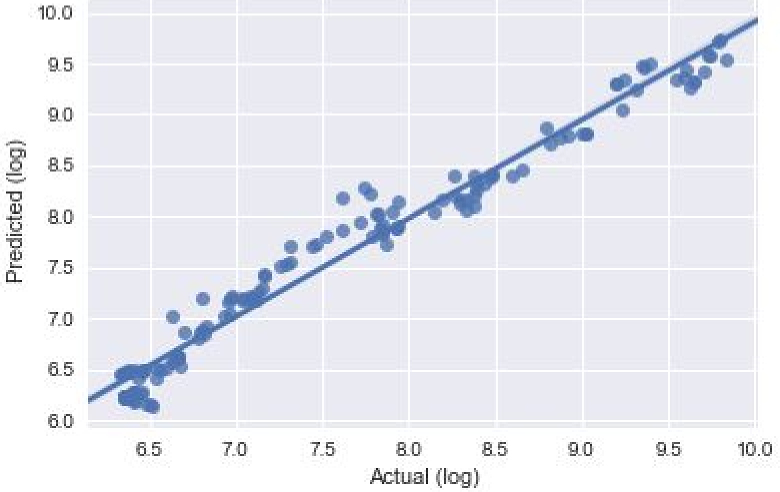
Correlated Relationship

Single Feature Regression

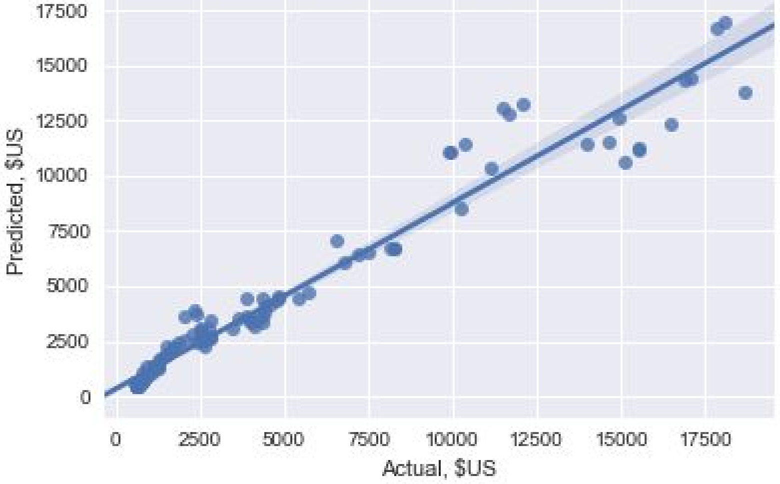


Linear Regression

Predicted vs. Actual (log)



Predicted vs. Actual



Key Insights / Takeaways

Bitcoin price a function of:

1. Google Search Interest
2. Nasdaq Composite Index
3. Network Transaction Fees

Google Search Interest may be leading indicator

Nasdaq Composite on “bull run” for history of Bitcoin; what happens when the “bear” comes?

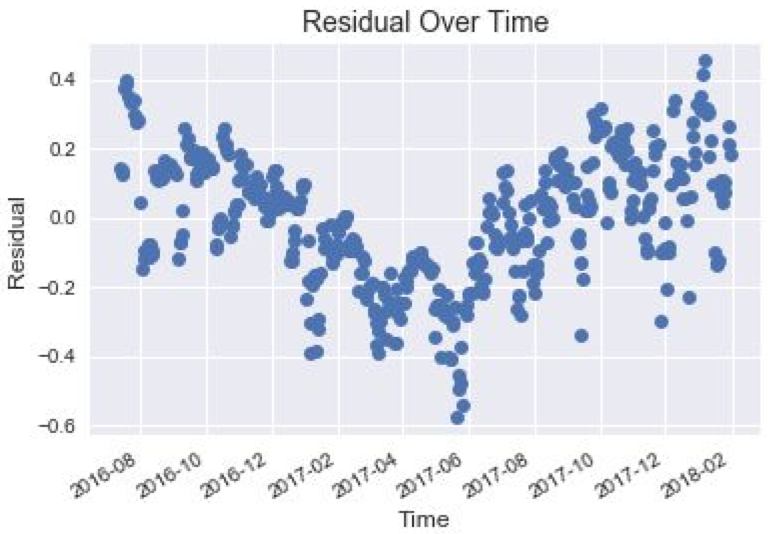
Next Steps

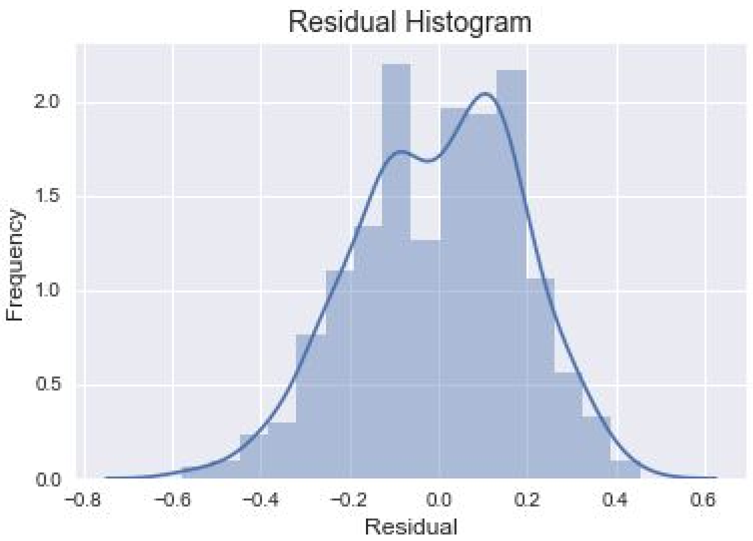
- Further explore seasonality in residual data
- Reconfigure model as a time series analysis, complete with price prediction for a certain timeline (requires adjustment of cross-validation)
- Explore social media sentiment (e.g. Twitter) as leading indicator

Questions?

Appendix

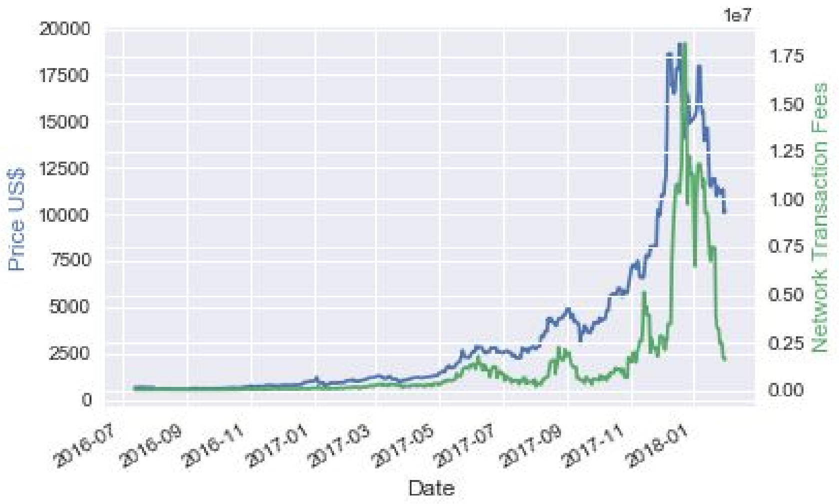
Residuals


Residual Over Time
0.4
0.2
0.0
−0.2
−0.4
−0.6
Residual
2016-08
2016-10
2016-12
2017-02
2017-04
2017-06
2017-08
2017-10
2017-12
2018-02
Time

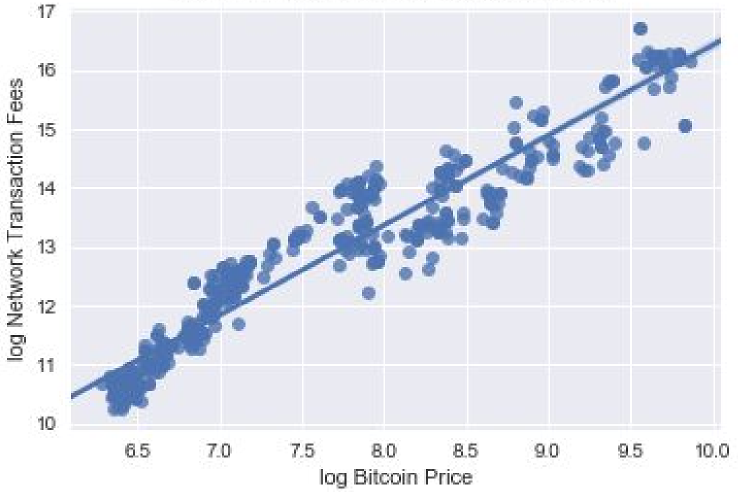

Residual Histogram
2.0
1.5
1.0
0.5
0.0
Frequency
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
Residual

Transaction Fees at R^2 of 92%

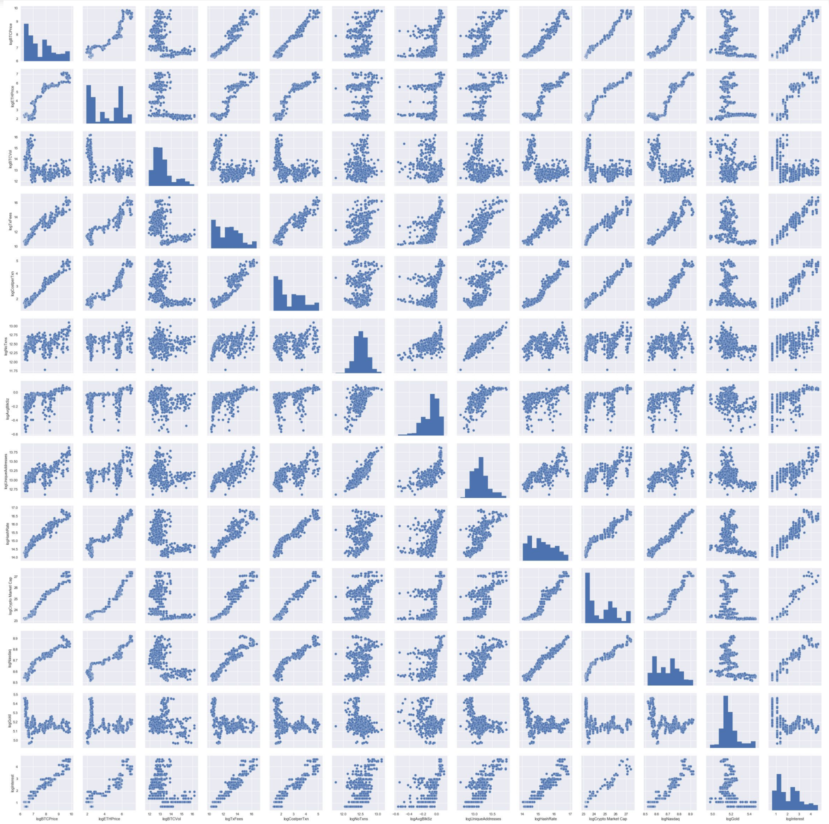
Correlated Relationship



Single Feature Regression



Feature Universe



3 Features contribute to R^2 of ~97%

Dep. Variable:	logBTCPrice	**R-squared:**	0.968
Model:	OLS	**Adj. R-squared:**	0.968
Method:	Least Squares	**F-statistic:**	4647.
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	0.00
Time:	11:40:55	**Log-Likelihood:**	127.12
No. Observations:	465	**AIC:**	-246.2
Df Residuals:	461	**BIC:**	-229.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-17.4446	2.346	-7.435	0.000	-22.055	-12.834
logInterest	0.5879	0.031	19.129	0.000	0.528	0.648
logNasdaq	2.6151	0.286	9.142	0.000	2.053	3.177
logTxFees	0.0826	0.020	4.060	0.000	0.043	0.123

Omnibus:	7.400	**Durbin-Watson:**	0.203
Prob(Omnibus):	0.025	**Jarque-Bera (JB):**	7.008
Skew:	-0.255	**Prob(JB):**	0.0301
Kurtosis:	2.680	**Cond. No.**	4.31e+03

Model Scorecard	
Accuracy	R^2 of 96.8%, explaining randomness left in model (SSE) as proportion to variation in data (SST)
Fit	**F-statistic** P-value very small; data too extreme to fit model by chance alone High **Log-Likelihood, low AIC and BIC** indicate good fit
Variable contribution	**T-test** indicates high contribution of variables
Normal Distribution	**Omnibus and Jarque-Bera** significant at < 0.05, indicating no exact normal distribution of ε
Autocorrelation	**Durbin-Watson** indicates slight positive autocorrelation (common in time series data)
Multicollinearity	**Condition Number high, implying matrix may not have a unique, well defined solution**

Google Search Interest at R^2 of 95%

Dep. Variable:	logBTCPrice	**R-squared:**	0.954
Model:	OLS	**Adj. R-squared:**	0.954
Method:	Least Squares	**F-statistic:**	9643.
Date:	Tue, 30 Jan 2018	**Prob (F-statistic):**	3.37e-311
Time:	17:04:46	**Log-Likelihood:**	46.288
No. Observations:	463	**AIC:**	-88.58
Df Residuals:	461	**BIC:**	-80.30
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.5065	0.023	236.676	0.000	5.461	5.552
logInterest	0.9640	0.010	98.199	0.000	0.945	0.983

Omnibus:	22.440	**Durbin-Watson:**	0.259
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24.987
Skew:	-0.504	**Prob(JB):**	3.75e-06
Kurtosis:	3.530	**Cond. No.**	6.21

Model Scorecard	
Accuracy	R^2 of 95.4%, explaining randomness left in model (SSE) as proportion to variation in data (SST)
Fit	**F-statistic** P-value very small; data too extreme to fit model by chance alone High **Log-Likelihood, low AIC and BIC** indicate good fit
Variable contribution	**T-test** indicates high contribution of variables
Normal Distribution	**Omnibus and Jarque-Bera** significant at < 0.000, indicating no exact normal distribution of ε
Autocorrelation	**Durbin-Watson** indicates slight positive autocorrelation (common in time series data)
Multicollinearity	**Condition Number** low implying unique, well-defined solution

Nasdaq Index at R^2 of 92%

Dep. Variable:	logBTCPrice	**R-squared:**	0.918
Model:	OLS	**Adj. R-squared:**	0.918
Method:	Least Squares	**F-statistic:**	5145.
Date:	Tue, 30 Jan 2018	**Prob (F-statistic):**	3.18e-252
Time:	17:08:24	**Log-Likelihood:**	-90.077
No. Observations:	463	**AIC:**	184.2
Df Residuals:	461	**BIC:**	192.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.3836	1.156	-65.188	0.000	-77.656	-73.111
logNasdaq	9.5430	0.133	71.730	0.000	9.282	9.804

Omnibus:	21.858	**Durbin-Watson:**	0.060
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23.621
Skew:	0.533	**Prob(JB):**	7.42e-06
Kurtosis:	3.297	**Cond. No.**	744.

Model Scorecard	
Accuracy	R^2 of 91.8%, explaining randomness left in model (SSE) as proportion to variation in data (SST)
Fit	**F-statistic** P-value very small; data too extreme to fit model by chance alone Low **Log-Likelihood / High AIC and BIC** may indicate poor fit
Variable contribution	**T-test** indicates high contribution of variables
Normal Distribution	**Omnibus and Jarque-Bera** significant at < 0.000, indicating no exact normal distribution of ε
Autocorrelation	**Durbin-Watson** indicates slight positive autocorrelation (common in time series data)
Multicollinearity	**Condition Number high, implying matrix may not have a unique, well defined solution**

Transaction Fees at R^2 of 92%

Dep. Variable:	logBTCPrice	**R-squared:**	0.918
Model:	OLS	**Adj. R-squared:**	0.918
Method:	Least Squares	**F-statistic:**	5213.
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	1.86e-254
Time:	19:16:46	**Log-Likelihood:**	-91.837
No. Observations:	466	**AIC:**	187.7
Df Residuals:	464	**BIC:**	196.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0698	0.107	-0.654	0.513	-0.280	0.140
logTxFees	0.6011	0.008	72.199	0.000	0.585	0.617

Omnibus:	3.705	**Durbin-Watson:**	0.214
Prob(Omnibus):	0.157	**Jarque-Bera (JB):**	3.662
Skew:	0.181	**Prob(JB):**	0.160
Kurtosis:	2.759	**Cond. No.**	101.

Model Scorecard	
Accuracy	R^2 of 91.8%, explaining randomness left in model (SSE) as proportion to variation in data (SST)
Fit	**F-statistic** P-value very small; data too extreme to fit model by chance alone Low **Log-Likelihood / High AIC and BIC** may indicate poor fit
Variable contribution	**T-test** indicates high contribution of variables
Normal Distribution	**Omnibus / JB >0.05 may indicate exact normal distribution of ε**
Autocorrelation	**Durbin-Watson** indicates slight positive autocorrelation (common in time series data)
Multicollinearity	**Condition Number high, implying matrix may not have a unique, well defined solution**

Methodology

- Explored various features which may have correlation with the price of Bitcoin
- Train-Test split executed using standard Linear Regression methodology rather than as a Time Series, per project requirements
  - The difference stems from the “shuffling”, or randomizing, of data in determining split; time series data would not be randomized
  - As such, the price of bitcoin is predicted at a moment in time, rather than predicted on a time series basis

Scraping

- Twitter sentiment analysis
- Coinmarketcap price history
- Bitcoin futures

Future Implementations

- Adjust model to predict price via time series
- Explore valuation of other cryptocurrencies using bitcoin metrics
- Compile sufficient real-time data to execute twitter sentiment analysis over reasonable timeframe (ie 1 week+)

Other analyses to explore:

- Usage by country
- Bitcoin trading by exchange
- Bitcoin trading by currency
- Use of leverage
- SEASONALITY

Process

- Brainstorm possible interesting correlations, including:
    - Bitcoin-specific metrics (ie transaction fees, volume, block size, hash rate)
    - Bitcoin futures
    - Google search interest
    - Social media sentiment (via Twitter)

Sources / Resources

- www.coinmarketcap.com
- www.quandl.com
- https://trends.google.com/trends/
- https://marcobonzanini.com/2015/03/09/mining-twitter-data-with-python-part-2/
- http://cs229.stanford.edu/proj2015/029_report.pdf
- http://text-processing.com/
- https://trends.google.com/trends/explore?q=bitcoin,ethereum