



University of Strathclyde
M.Sc. Financial Technology
(2017/2018)
AC989 – Big Data Fundamentals

“Identifying key technological features behind the Ethereum Blockchain Network that have predicting power in order to predict daily prices using machine learning techniques”



Joshua Eick

Student No. 201779687

July 26, 2018

Word count – 3,285

Table of Contents

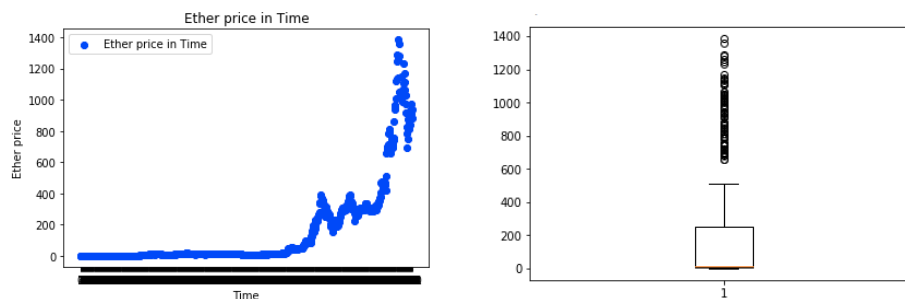
1. Introduction to the Dataset	3
Summary of the clean-up of the dataset and transformations	5
2. Identification of key challenges and problems to be addressed	5
Research Question	6
3. Summary Statistics	6
Descriptive statistics	7
Boxplot and histograms	9
Scatterplot	11
4. Description, rationale, application and findings from the unsupervised methods- K means clustering	14
5. Description, rationale, application and findings from a supervised method-Linear Regression	19
6. Reflection on methods used for analysis and Conclusion	25
7. References	27
8. Appendix	28

1. Introduction to the Dataset

The aim of this investigation is to examine the Ethereum cryptocurrency from Kaggle's dataset titled "Cryptocurrency Historical Prices" (2018), in order to discuss whether it is possible to predict Ethereum cryptocurrency prices with the features of the blockchain network. In carrying out this task both supervised and unsupervised methods will be followed. To clarify, Ethereum is the name of the network and Ether is the network's cryptocurrency. The dataset contains the prices of Ethereum from 7/30/2015 to 2/20/2018 and covers the fundamental technology variables of Ethereum. The initial impression is that the dataset is noisy, as missing values in the first months of the data can be observed. Because of the large amount of missing values in the first 30 columns of the dataset, these columns were removed, so, therefore, the analysis will start from 09/01/15, after which there are no missing values.

Ethereum's dataset contains 934 rows and 18 columns of attributes. In addition, the dataset has 18 attributes (columns) composed of floating numbers, integers and object (such as the index column called "Date (UTC)"). Therefore, it's an obvious, intuitive, requirement that the data has to be manipulated in order for it to be structured to perform well with machines' learning techniques.


Figure 1.1 –Price of Ethereum from 09/01/15 to 2/20/2018



The daily prices of the Ethereum cryptocurrency and most of the technological variables are extremely volatile with a high standard deviation, lognormal distribution, skewed to left and long tail to the right and have outliers as can be seen in figure 1.1.

Table 1.1 – Variables Log to 10

-
1. log10_eth_etherprice
 2. log10_eth_hashrate
 3. log10_eth_difficulty
 4. log10_eth_blocks
 5. log10_eth_uncles
 6. log10_eth_blocksize
 7. log10_eth_ethersupply
-



The shape of the dataset is irregular and has anomalies and several ranges of prices. For this reason the price of the Ethereum and other 6 variables are transformed to log 10 in order to reduce the standard deviation and have a symmetrical distribution. This is identified in Table 1.1.

On the other hand, the increase of the price is more predictable because it is completely new technology and is part of the mainstream in finance, so there is a certain level of bias which is shown in figure 1.1. In contrast, price decreases are less predictable and rarer. Thus, it's necessary to view the price of Ethereum as the dependent variable and other variables as independent variables because they influence the price.

From the 18 variables, 12 are selected. The selection of these independent variables is determined by three main points (a description of each of the variables selected is in Appendix 1 and the explanation for the exclusion of 6 other variables is in the Appendix 2).

The three main points are:

- 1- Key measurements of the main properties of blockchain technology.
- 2- The variables that have strong correlation with price.
- 3- Independent variables that have low multicollinearity conditions.

Summary of the clean-up of the dataset and transformations

In summary, in order to carry out a deep analysis of this complex dataset with the machine learning techniques it's necessary to make the following manipulations:

- 1- Transform the necessary variables in order to have a symmetrical distribution
- 2- Remove four independent variables from the dataset because of the multicollinearity condition, which are not relevant to investors
- 3- Drop the first 30 columns because they have too many missing values.

(A description of the dataset before clean up and transformation is in Appendix 3)

2. Identification of key challenges and problems to be addressed

- 1- Missing values in the first two months of the dataset and one column that has 934 missing values.
- 2- 9 variables are highly correlated which delivered multicollinearity condition.
- 3- 7 variables have non-symmetrical distribution
- 4- Issues of bias with the "eth_price" variable in the dataset because the cryptocurrency is relatively new. Thus the price is more likely to increase.

- 5- Difficulty at arriving at a clear picture/conclusions from these variables of the dataset because there are features of the technology of the Ether technology that are not familiar to non-expert in this blockchain technology.
- 6- Necessity to gain more insight and information from the dataset.

Research Question- The main key challenge and the aim of this research is to discover the independent variables that are most influential/correlated to the price of Ethereum and predict the prices using the machine learning techniques.

3. Summary Statistics

In order to choose the most appropriate method in machine learning, the dataset and key summary statistics are explored. Firstly, the figure 3.1 presents the key descriptive statistics for the 12 variables selected. A clear picture of the structure of these variables after clean up is observed, as well as transformations to log 10.

Descriptive statistics

Figure 3.1-Descriptive statistics –After clean up and transformation of the dataset

Columns = 12 (11 independent variables & 1 dependent variable)

Rows = 902 (days)

	log10_eth_etherprice	eth_tx	eth_supply	log10_eth_hashrate	\
count	902.000000	9.020000e+02	9.020000e+02	902.000000	
mean	1.370539	1.831405e+05	8.593257e+07	3.920199	
std	0.941634	2.760825e+05	7.588844e+06	0.861554	
min	-0.376751	4.777000e+03	7.291060e+07	2.522304	
25%	0.920905	3.362525e+04	7.902501e+07	3.307427	
50%	1.087071	4.667550e+04	8.626714e+07	3.792511	
75%	2.404033	2.500838e+05	9.310599e+07	4.822610	
max	3.141456	1.349890e+06	9.768027e+07	5.386574	
	log10_eth_difficulty	log10_eth_blocks	log10_eth_uncles	\	
count	902.000000	902.000000	902.000000		
mean	2.084663	3.733485	2.662358		
std	0.864871	0.071986	0.197779		
min	0.744136	3.451633	2.100371		
25%	1.433348	3.703012	2.547775		
50%	1.910997	3.770668	2.621176		
75%	3.016174	3.781091	2.699621		
max	3.481588	3.806790	3.321391		
	log10_eth_blocksize	eth_gasprice	eth_gaslimit	eth_gasused	\
count	902.000000	902.000000	9.020000e+02	9.020000e+02	
mean	3.469847	10.467501	4.656485e+06	8.223264e+09	
std	0.503308	0.191015	1.682588e+06	1.209752e+10	
min	2.860338	10.015862	5.002380e+05	1.263239e+08	
25%	3.143171	10.353704	3.141988e+06	1.082082e+09	
50%	3.220631	10.380893	4.696694e+06	1.632639e+09	
75%	3.963759	10.619307	6.706538e+06	1.265377e+10	
max	4.527385	11.973128	7.999398e+06	4.396431e+10	
	count	log10_eth_ethersupply			
	902.000000				
mean		4.431328			
std		0.084994			
min		4.171800			
25%		4.390183			
50%		4.450125			
75%		4.500861			
max		4.562910			

Heatmap - Correlation Matrix

Secondly, a heatmap of the correlation matrix is performed in order to understand the relation between variables highlighted by Igual and Seguí (2017, p.105). Figure 3.2 presents the independent variables that are highly correlated that deliver multicollinearity conditions (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.189). Hence, highly correlated variables reduced the effectiveness of the supervised models. The 11

independent variables selected are less likely to be highly correlated with other independent variables. This plot gives us an indication of how significant the variables would be in a linear model according to Igual and Seguí (2017, p.107).

Figure 3.2 – Correlation matrix against price

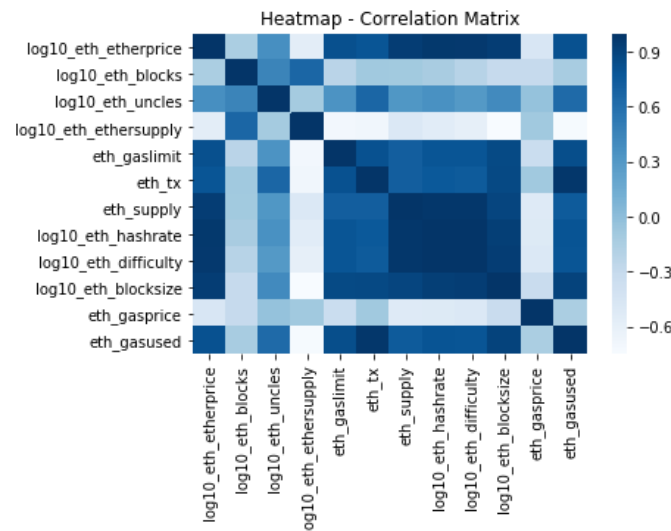


Figure 3.2 shows that the variables most negatively correlated to the price of Ethereum are “eth_gas price” and “eth_ethersupply”. The variable that has no relation with price is “eth_blocks”, which means that it would not make sense to take into account for the linear regression model. In contrast, the other 9 variables correlate positively to the price.

The heatmap correlations illustrate that the hash rate increased and the price of the Ether increased. Research by McNally, Roche and Caton (2018, p. 340) stated positive correlations between the price of Bitcoin cryptocurrency and hash rate. Thus, this dataset is consistent, logical and intuitive of what experts have stated regarding the cryptocurrency. Finally, there are nine variables with relative multicollinary conditions. In sum, this correlation matrix highlighted which variables should be further analysed because of their relationship with price.

Boxplot and histograms

The second visualizations were the boxplot and histogram, which provided vital insight into the distribution of the dataset.

1-The boxplot was selected because it displays a brief summary of the concentration of observations related to the quartiles and identified outliers.

2-The histogram was selected in order to present information about whether normal distribution was met.

Figure 3.3 - Comparison between raw dataset and transformation

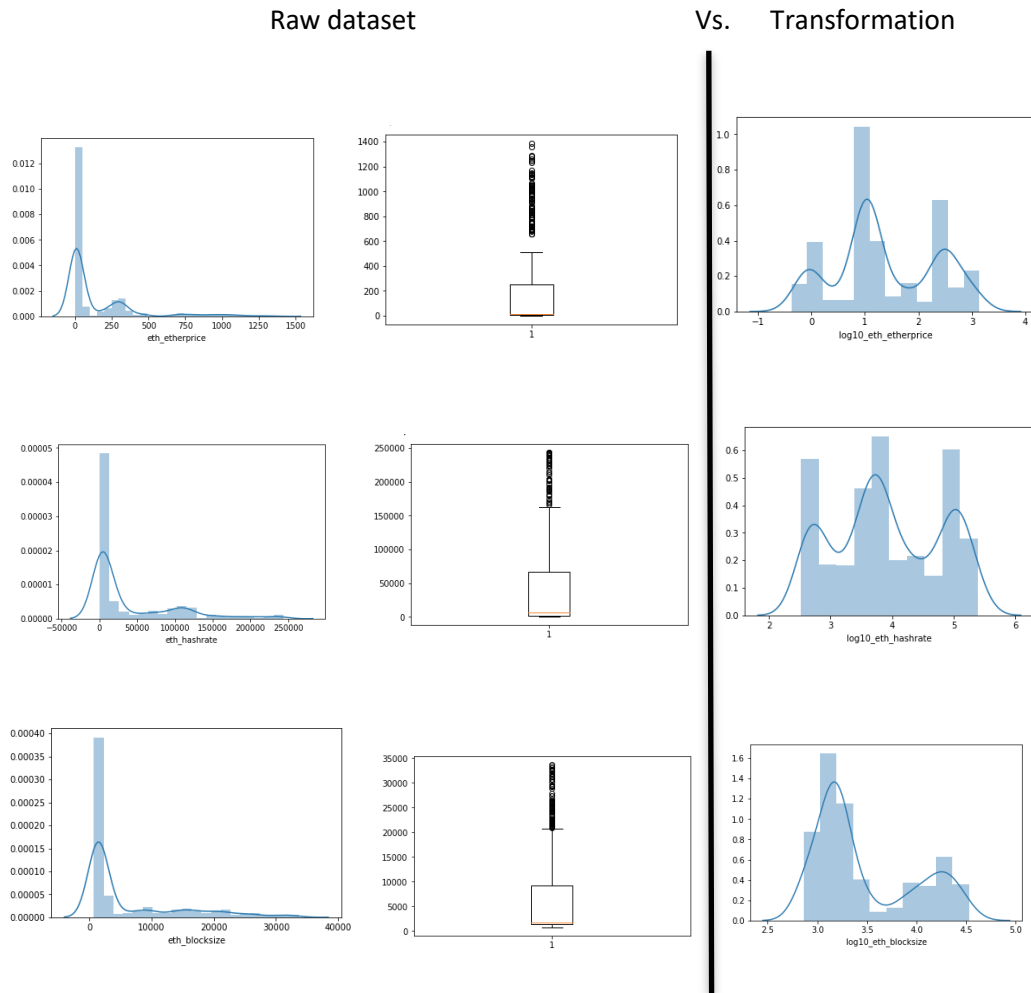


Figure 3.3 visualizes the three variables that are strongly significant from the investors' perspective.

The plotting of histogram and boxplot indicates the following deductions:

- 1- The dataset skewed to the right and has irregular distribution.
- 2- A large amount of outliers indicate that the prediction of prices will be affected by these values.
- 3- Prices of Ethereum that lie below 75% with an average of 500.

In summary, the variables are transformed to log 10 to achieve normal distribution and reduce the outliers.

Scatterplot

Another key visualization for the interpretation of this dataset is the scatterplot with regression. The scatterplot is helpful when the dataset is not clearly correlated and aims to make the relationship clearer (Halvey, 2017). Hence, a scatterplot is applied in order to visualize two assumptions for the linear regression:

1. Linear relationship with price
2. Multicollinearity condition

Figure 3.4 –Scatterplot - Price against independent variable

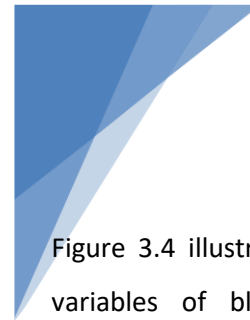
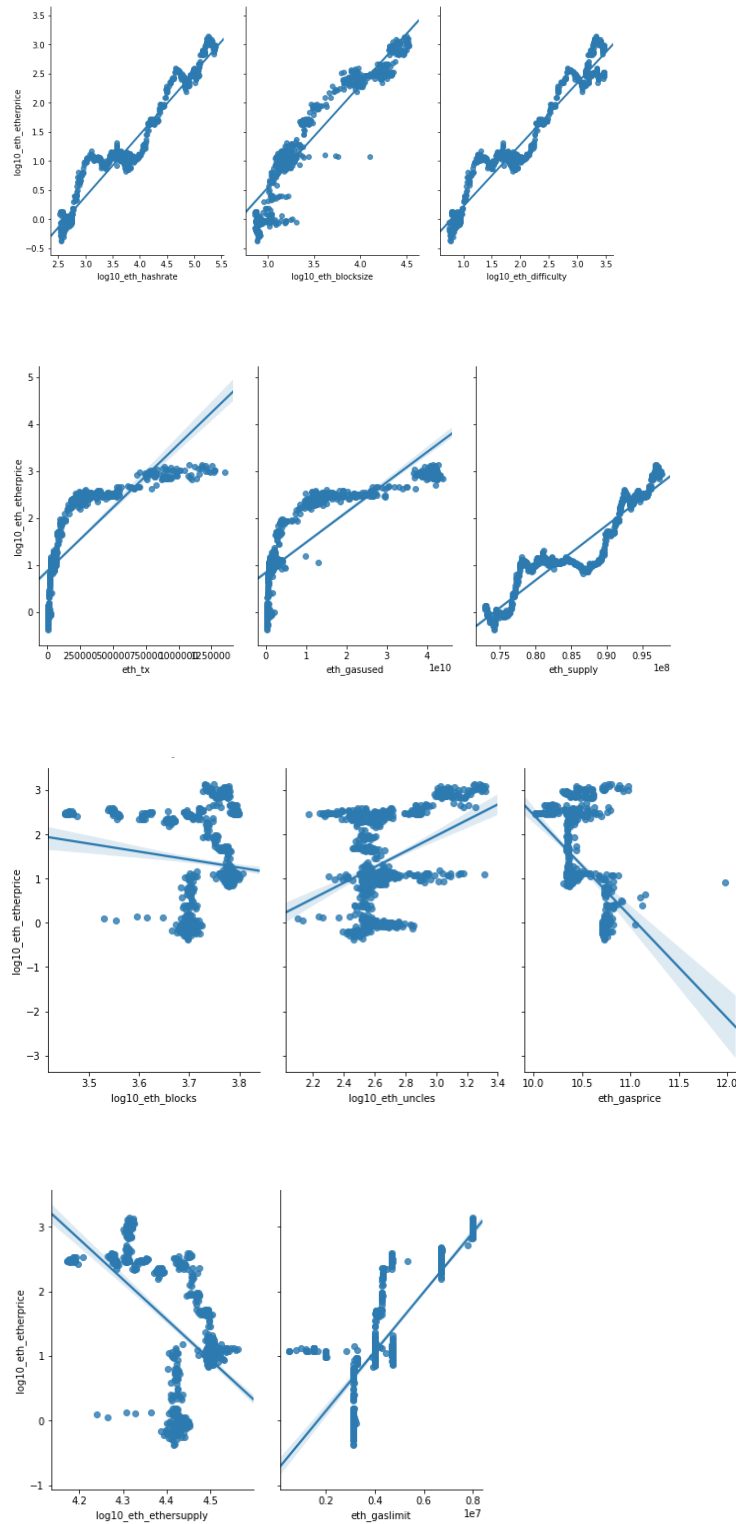
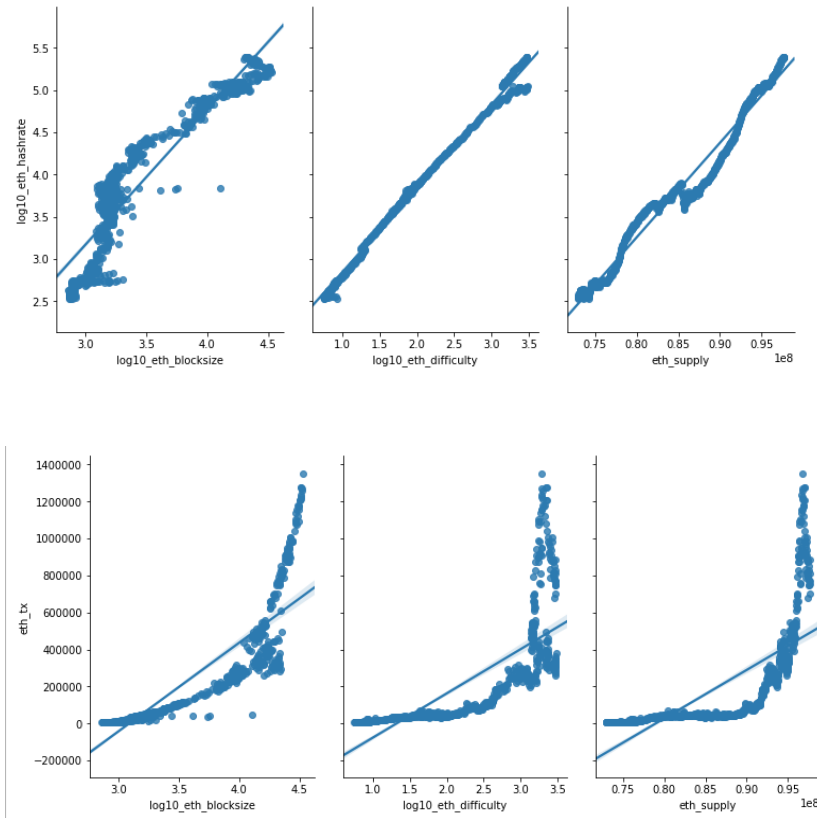


Figure 3.4 illustrates that the variables of blocks, uncles, ether supply and gas limit, gas price have poor linear relationships. Thus, these variables do not satisfy the assumption of linear regression and will be taken out from the linear regression model. In addition, one possible explanation of deviations from linearity are events like market cycles of the cryptocurrencies and the hacking of exchange rates.

Figure 3.5 - Scatterplot presenting hash rates and transactions against the most correlated independent variables



In figure 3.5 hash rates and transactions are selected in order to visualize the multicollinearity condition because these variables show a higher correlation than other independent variables. Essentially, they are clearly multicollinearity conditions, so this assumption is not met with these two variables.

4. Description, rationale, application and findings from the unsupervised methods- K means clustering

Description

One of the purposes of this research is to cluster together the price of Ethereum with similar cryptocurrencies in terms of features/variables. This is an unsupervised method. Igual and Seguí (2017, p.115) state that unsupervised techniques tackle the problems of clustering, which is one of the issues to address with this dataset. Moreover, part of the aim of unsupervised methods is to find hidden structures in unlabeled datasets (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.118). In this case, similarities in independent variables are taken and clustered into price (dependent variable)

Rationale

One of the uses for K-means clustering is to group factors like customer segmentation (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.120). This research is similar to the case of customer segmentation because aim to price segmentations of Ethereum. Moreover, K-means clustering is chosen because, according to Igual and Seguí (2017, p.124), it is suitable for big datasets. Our dataset has 902 rows of daily prices of Ethereum, which is relatively large. Therefore, K-means clustering addresses the scalability issue of the dataset. In contrast, it is not considered an unsupervised method, like hierarchical clustering, in terms of the present research. Even though hierarchical clustering is believed to perform well when considering its ranking, it is not considered useful except when examining small datasets. Therefore, it was concluded that K-means clustering was a superior method for the structure of the large dataset of Ether, with this superiority being further backed up by its simplicity.

Application

Firstly, 7 clusters are chosen because they substantially increased the homogeneity and completeness of K-means.

Then the dataset is standardized/scaled so all the variables have a mean of 0 and variance of 1. As a result of this they generated the standard deviation as the unit of measurement (Halvey, 2017).

Next, the centroid of the clusters is located. The centroid is the point that corresponds to the center of mass for a cluster (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.121).

Finally, the K-means clustering inputs independent variables into clusters depending on the distance, with the independent variable selecting the clusters which have the smaller distances.

Findings

In order to visualize the results of the K-Means of the dataset and observe the differences between the variables, the scaling is removed and plotted by color. Essentially the main significant independent variables are visualized for the investors of the cryptocurrency.

Figure 4.1 - K- means clustering- Hash rate against difficulty and blocksize

*cluster colors represents a range of prices of Ethereum

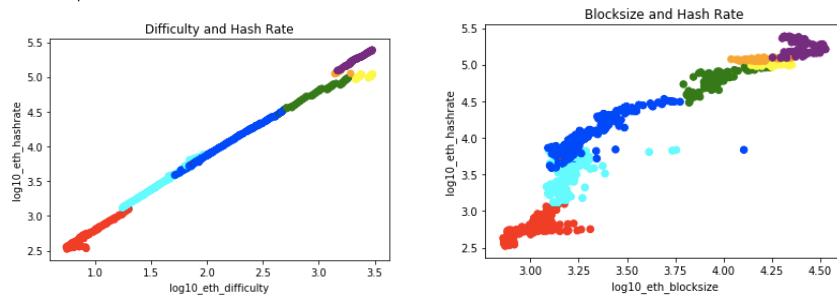


Figure 4.1. There is a clear difference in classification with hash rate and difficulty in prices. However, there is relatively low level of overlapping with blocksize and hash rate, meaning that there is less accuracy of classification. There are outliers with a possible explanation for this possibly being different attacks on the network, considering that blocksize increases significantly with the same level of hash rate.

Figure 4.2 - K-means clustering- Hash rate against supply and gas price

*cluster colors represents a range of prices of Ethereum

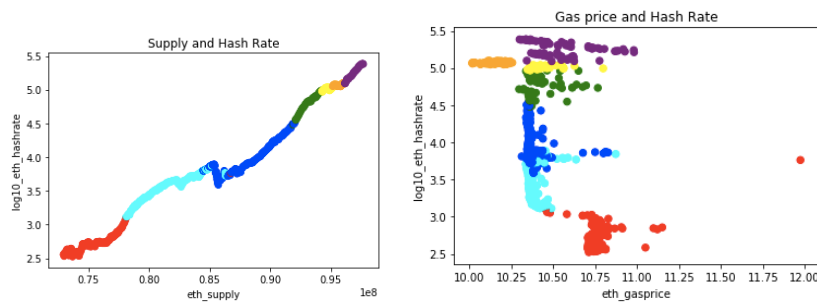


Figure 4.2 indicates that there is superior differentiation in classification with hash rate and supply in prices. In contrast, in terms of gas price there are not differences between prices. Thus, it seems that gas prices do not have an influence on the prices of Ethereum because the dots are scattered evenly over the same values and are over-fitting.

Figure 4.3 - K- means clustering- Transaction against supply and blocksize

*cluster colors represents a range of prices of Ethereum

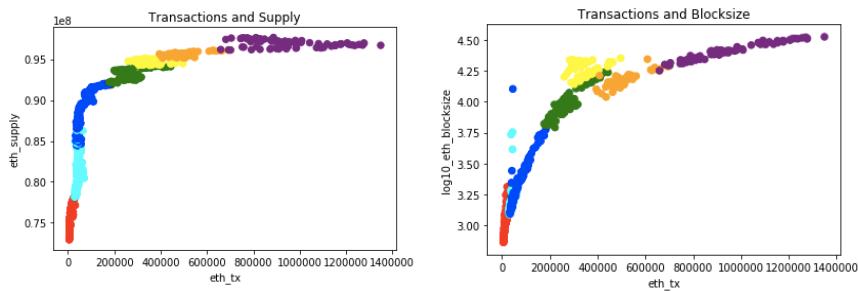


Figure 4.3 presents significant differentiation in classification with transaction and supply in prices. However, after a certain point the supply stops increasing, meaning that the supply turns out to stay constant. This is intuitive because the Ethereum is considered to make the supply fixed. The variable blocksize has a differentiation in classification but with low levels of overlapping.

Figure 4.4 - K- means clustering- Hash rate against blocks

*cluster colors represents a range of prices of Ethereum

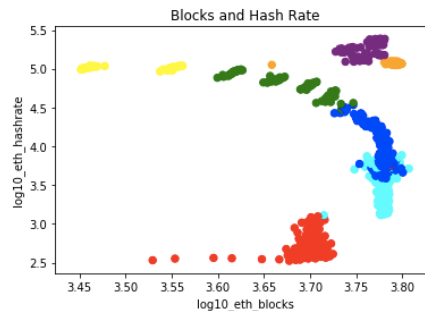


Figure 4.4. In terms of blocks there is no differentiation with prices and there is overfitting. Again, hash rate indicates a clear difference in classification with prices.

Figure 4.5 - metrics of evaluations of K- means clustering

```
metrics.completeness_score(target,model.labels_)  
0.9685858671553719  
  
metrics.homogeneity_score(target,model.labels_)  
0.26588694127566553
```

The evaluations metrics in figure 4.5 strongly suggest that the quality of the performance of k-means clustering was significant because completeness was high and homogeneity was relatively low. The completeness encompasses all variables of a given class that are assigned to the same cluster, while homogeneity means that each cluster contains only variables of a single class (Halvey, 2017). Hence, there are indications that variables of this dataset are classified correctly achieving 97%. However the clusters were not clearly classified with 27%. In fact, if the number of clusters selected is increased, the classification accuracy of the cluster also increased.

The following are some possible reasons why the classification of K-means clustering could be improved:

1. Another possible explanation that this k-mean clustering homogeneity behaves poorly is because of the distance between the assigned values within a cluster, which is not meaningful (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.134).
2. Igual, L. and Seguí, S. (2017, p.124) highlighted that K-means algorithms respond poorly because of the irregular structure/shapes of the dataset, which can be the case for this dataset, which is reflected by the plots.

In summary, K-means clustering is specifically suitable as an unsupervised method for this dataset because it is capable of classification accuracy and delivers insightful information about the dataset.

5. Description, rationale, application and findings from a supervised method-Linear Regression

Description

The supervised method applied for the analysis of the dataset, which aimed at predicting the price of Ethereum, is linear regression. The statistical model of linear regression aims to explain the influence that a set of variables (features of Ethereum) has on the outcome of the dependent variable, in this case, price (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.162). Moreover, linear regression chooses the best fitting line to a set of observations minimizing residuals, the difference between predicted (linear model) and actual outcomes (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.163). Therefore, the model will predict the price of Ethereum based on the features of Ethereum.

Rationale

Linear regression is used in finance cases. For example, the model can take features of the home living area in order to predict the home prices of that area (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.162). Likewise this research is similar because it takes the features of Ethereum cryptocurrency in order to predict the prices. It can be concluded, then, that the model of linear regression is a suitable fit for this analysis, given that it is widely agreed that it is useful for the finance industry in predicting quantitative data. Furthermore, Halvey argues that it is a “good first technique for modeling quantities” (Halvey, 2017). While Igual and Seguí (2017, p.113) highlight that regression can also find significant variables for the model, something this research aims to do.

Application - (Building the model)

The linear regression makes large assumptions about the dataset, such as the following:



- 1-Linear relationship between dependent and independent variables - Meet the assumption for 7 variables.
- 2-Residuals between the model and the original are normally distributed.
- 3-Variance of the residuals is constant with a mean of 0.
- 4-Residuals are independent from each other.
- 4-Multicollinearity- Most of the variables were found to have multicollinearity condition.
- 5-Outliers-There are outliers causing skews. If logged to 10, the variables that have outliers, because of high sensitivity, can improve performance of linear regression.
- 6-Sample size – This assumption is met because it is a large dataset.

According to our visualization of the scatterplot this assumption is met with the variables selected. In fact, only 7 independent variables meet this assumption, as illustrated in figure 5.1. The gas price variable, in particular, has no linear regression, however, if we transform to log 10 then a linear relationship can be achieved (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.173).

The linear regression is applied with the 7 independent variables and 1 dependent variable selected. Afterwards, the dataset is split between the training set (80%) and the test set (20%), in order to validate the model forwarded by Igual and Seguí (2017, p.175).

In addition, Igual and Seguí (2017, p.107) argue that split data is for evaluating how strong the predictive power of the model is. Understanding deeply the relationship between these variables through the visualizations and almost all the assumptions of the linear regression have been met it is performed the model.

Findings

Figure 5.1-Intercept and Coefficients of the linear regression

```
In [206]: print(lm.intercept_)
-4.165507455235088
```

	Independent variables	Coefficients
0	eth_tx	3.323350e-07
1	eth_supply	-6.763398e-08
2	log10_eth_hashrate	2.335760e+00
3	log10_eth_difficulty	-9.079045e-01
4	log10_eth_blocksize	7.006717e-01
5	eth_gasprice	1.676819e-01
6	eth_gasused	-1.958767e-11

Table 5.1 illustrates that the model does not have a coefficient that equals 0, meaning that no variables were discarded during predictions (Igual and Seguí 2017, p.108). The greater the coefficient the more significant the variable to predict price is. As a result, the variables with the most unique statistical significance are hash rate, difficulty and blocksize. Therefore, these variables in general have a strong power to predict price.

First, evaluating the logistic regression method it is necessary to consider the Mean Square Error, R squared and variance. Mean Square Error is the standard deviation of the differences between predicted values and actual values and with cross-validation helps make an unbiased MSE.

Figure 5.2- MSE, R square & Variance

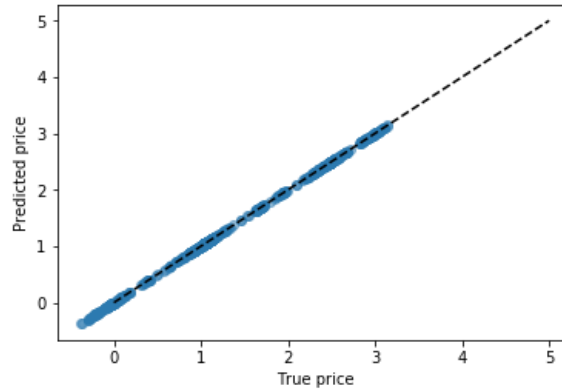
```
In [200]: print(metrics.mean_squared_error(Y_test, predict))
0.027621491633743493

In [201]: print(metrics.r2_score(Y_test, predict))
0.9675194480489353

In [202]: predict.var()
Out[202]: 0.8274917878386445
```

Figure 5.2 shows that MSE is 3%, which is relatively small, and suggests at superior predicting power. Secondly, the R square is evaluated, which measures the variation of the data that is explained by the model (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.170). Therefore, the model explains 97% of the variance in the perceived stress according to Pallant (2010, p.160). In sum, the model explained the data's predictions well, but had a high variance with .83.

Figure 5.3 – Predicted Price vs. True Price



It is believed that the features of Ethereum have strong predictive power and that we can predict price using the linear regression model as is illustrated in figure 5.3. The predicted power was sound because there is a relatively large degree of variance. This is beneficial as the prices of cryptocurrency are very volatile and is big issue for making predictions.

Finally, in order to complete the evaluation process, the validation process is followed to observe the performance of the model. The cross-over has been contemplated already, however, to add to the validation a K-Fold Cross is accomplished in order to see how the model behaves with different subsets of the dataset.

Table 5.1- K-Fold Cross

K-Fold Cross				
Test set	MSE	R Squared	Variance	
10%	0.0284	0.9604	0.7085	
20%	0.0255	0.9712	0.8563	
40%	0.0295	0.9676	0.8872	
60%	0.0306	0.9644	0.8248	
80%	0.0315	0.9642	0.8879	
Average	3%	97%	.83	

As a result of K-fold cross, it can be observed in table 5.1 that MSE and R Squared are constant. However the variance increased as in test sets increased, indicating that prices are significantly volatile. Thus, K-fold cross indicates that the performance of the model is acceptable, unbiased prediction of the model and the model holds in the general dataset.

There are several reasons to improve the performance of the linear regression model, such as:

1-The multicollinearity conditions among independent variables. As a consequence of these the coefficient estimates can take different directions. The best ways to approach this condition is to remove the variable or make functions for the correlated variables, for the future projects (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.189).

2-Most of the investors of cryptocurrencies are speculators, which means they do not strongly consider fundamental variables, but rather market sentiments. Thus, input

variables like market sentiment will increase the prediction power (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.178).

Despite these reasons, it can be assumed that it is difficult to predict returns from the cryptocurrency, because there are many factor that affect returns which are highly volatile. This can be summed up by the phrase, “correlation does imply causation” (Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015, p.189). In this case, there are many other independent variables can influence the behavior of prices of the cryptocurrency.

In conclusion, it is believed that predicting prices for cryptocurrency of Ether with linear regressions is an appropriate supervised method, however, more independent variables need to be correlated to returns, such as market sentiment variables, in order to increase the performance of the model.

6. Reflection on methods used for analysis and Conclusion

It is necessary to reflect on the unsupervised methods selected in this research, which is the K-means clustering, that gives guidance to better understand the classification with clusters regarding each variable. For instance, the visualization of the k-means with the colors helps us to interpret how independent variables behave with prices, which is something that is very difficult to differentiate with a big dataset. Specifically, the k-mean for quantitative variable seems to be advantageous because it can specify the clusters (in this case, groups of prices) and lead to revealing substantial inside information. Therefore, the main reason why the k-means clustering provides an extra significant value is because it is designed for a dataset that has multiple groups, or can classify several groups, and as a result will simplify and structure the data. In summary, the unsupervised method is of unique statistical significance that lends itself to clusters of data, that can help us better understand unstructured, difficult and large datasets.

The supervised method selected is linear regression, which is designed to predict and understand the variables significant to the dependent variable. The main reason linear regression was selected was because the model is designed to predict dependent variables and observe clearly

the unique significance of the relationship between variables. However, the linear regression has a high amount of assumptions that need to be considered. Thus, this is a disadvantage of this method, however, if you manage and explore the dataset well to meet the assumptions, the model can still be applied acceptably. In fact, it is widely agreed that linear regression is the best fit to predict quantities.

In summary, the supervised and unsupervised methods utilized in this research are useful for classification of data and delivers insightful information about prices of the Ethereum cryptocurrency. In the unsupervised method the K-means clustering is superior to Hierarchical clustering in this case because of the scalability issue of the dataset, as Hierarchical clustering doesn't handle large datasets well. Principally, the unsupervised methods have to be used in combination with supervised methods to extract more information from the dataset, and can act as prefaces for supervised methods. The results show that the combination of K-mean clustering and linear regression was a strong method to apply because of the accuracy level, how it manages prices and supports more insight into the classification of the perspective of quantitative variables. In other words, it can be concluded that the combinations of these two models in quantitative variables perfectly fit if the assumptions are meet.

In conclusion, the evidence above demonstrates that predicting the prices of Ethereum, with the fundamental variables of the blockchain technology is a superior method for doing so, however, it is necessary to be cautious because of issues of volatility. Investors with an appetite for high risks, can consider these variables (SME is 3% and R squared of 97%) when weighing up investment decisions.

7. References

Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. (2015). Indianapolis, IN: John Wiley & Sons, Inc.

Halvey, M. (2017). Lecture 2: Exploring and Managing Data. University of Strathclyde, Department of Computer and Information Sciences. (Accessed: 16 June 2018).

Halvey, M. (2017). Lecture 3: Modelling Methods. University of Strathclyde, Department of Computer and Information Sciences. (Accessed: 18 June 2018).

Halvey, M. (2017). Lecture 4: Linear and Logistic Regression. University of Strathclyde, Department of Computer and Information Sciences. (Accessed: 16 June 2018).

Halvey, M. (2017). Lecture 5: Unsupervised Methods. University of Strathclyde, Department of Computer and Information Sciences. (Accessed: 16 June 2018).

Igual, L. and Seguí, S. (2017). Introduction to Data Science. Cham: Springer International Publishing.

Kaggle.com. (2018). Cryptocurrency Historical Prices | Kaggle. Available at: https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory#ethereum_dataset.csv [Accessed 10 Jun. 2018].

Ledolter, J. (2013). Business analytics and data mining with R. Hoboken, NJ: Wiley.

McNally, S., Roche, J. and Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning. Dublin Business School. p.340.

Pallant, J. (2011). SPSS Survival Manual - A step by step guide to do data analysis using SPSS. 4th ed. UK: Mc Graw Hill.

8. Appendix

Appendix 1 - The definition of the 13 variables selected according to Kaggle (2017)

1. eth_etherprice : price of ethereum
2. eth_tx : number of transactions per day
3. eth_supply : Number of ethers in supply
4. eth_hashrate : hash rate in GH/s
5. eth_difficulty : Difficulty level in TH
6. eth_blocks : number of blocks per day
7. eth_uncles : number of uncles per day
8. eth_blocksize : average block size in bytes
9. eth_gasprice : Average gas price in Wei
10. eth_gaslimit : Gas limit per day
11. eth_gasused : total gas used per day
12. eth_ethersupply : new ether supply per day

Appendix 2 - The explanation for the exclusion of the 6 variables

Firstly, the “etch_address” is removed from the analysis because its average is highly correlated with other independent variables, and can be viewed as an insignificant variable for consideration by investors of cryptocurrency. Removing this variable reduces the multicollinearity condition. Moreover, the variable “eth_register” has 640 missing values from 934 rows, which means that 68% of the values are missing. Because this figure is above 50% and the variable is not considered to be of great significance to investors, it is removed from the dataset. Furthermore, the “UnixTimeStamp” is removed from the dataset because from an investor’s perspective it is not considered important for the returns of the cryptocurrency. The variable “eth_blocktime” is also removed because of its insignificance for investors but also because it does not have a strong

correlation with price. In fact, the “market cap” variable is excluded because is the similar variable to price of Ethereum.

Appendix 3 - Description of the dataset before clean up and transformation

	index	UnixTimeStamp	eth_etherprice	eth_tx	eth_address \
count	902.000000	9.020000e+02	902.000000	9.020000e+02	9.020000e+02
mean	482.500000	1.479923e+09	149.233681	1.831405e+05	3.597997e+06
std	260.529269	2.254281e+07	263.606243	2.760825e+05	6.123891e+06
min	32.000000	1.440979e+09	0.420000	4.777000e+03	1.916800e+04
25%	257.250000	1.460441e+09	8.335000	3.362525e+04	1.551858e+05
50%	482.500000	1.479902e+09	12.220000	4.667550e+04	8.186625e+05
75%	707.750000	1.499364e+09	253.532500	2.500838e+05	4.120962e+06
max	933.000000	1.519085e+09	1385.020000	1.349890e+06	2.704779e+07

	eth_supply	eth_marketcap	eth_hashrate	eth_difficulty \
count	9.020000e+02	902.000000	902.000000	902.000000
mean	8.593257e+07	14202.534737	38453.011138	585.946983
std	7.588844e+06	25524.425823	58207.456778	871.317166
min	7.291060e+07	31.181760	332.892500	5.548000
25%	7.902501e+07	703.079146	2029.680550	27.123750
50%	8.626714e+07	1014.689755	6201.704600	81.470000
75%	9.310599e+07	23590.910542	66467.770725	1037.943750
max	9.768027e+07	134210.789000	243542.069800	3031.012000

	eth_blocks	eth_uncles	eth_blocksize	eth_blocktime	eth_gasprice \
count	902.000000	902.000000	902.000000	902.000000	9.020000e+02
mean	5480.635255	520.227273	6202.440133	16.001419	3.328354e+10
std	777.714591	323.825055	8115.024048	3.212932	3.430675e+10
min	2829.000000	126.000000	725.000000	13.430000	1.037198e+10
25%	5046.750000	353.000000	1390.500000	14.130000	2.257898e+10
50%	5897.500000	418.000000	1662.000000	14.495000	2.403773e+10
75%	6040.750000	500.750000	9199.500000	16.867500	4.162047e+10
max	6409.000000	2096.000000	33681.000000	30.310000	9.400000e+11

	eth_gaslimit	eth_gasused	eth_ethersupply
count	9.020000e+02	9.020000e+02	902.000000
mean	4.656485e+06	8.223264e+09	27484.438297
std	1.682588e+06	1.209752e+10	4873.837707
min	5.002380e+05	1.263239e+08	14852.500000
25%	3.141988e+06	1.082082e+09	24557.539062
50%	4.696694e+06	1.632639e+09	28191.953125
75%	6.706538e+06	1.265377e+10	31685.546875
max	7.999398e+06	4.396431e+10	36551.875000

Appendix 4 - Python version- 3.6

Appendix 5 – Package and library used

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
from sklearn import cluster
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.cross_validation import train_test_split
from sklearn.metrics import roc_auc_score
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import roc_curve
import seaborn as sns
from sklearn.preprocessing import scale
from time import time
```