# **[Data Mining]  Mid – Report :**

# **Analysis on Ethereum Market Cap**

***2013171033 Oh Han Seung***

***2016121114 Cho hyeon Jae***

***2019840738 XU KEVIN CHEN***

1. **Problem**
   1. **Predicting and analyzing the main factors which impact the capitalization of the Ethereum market**

Our purpose is to make it easy and reliable for those who are interested in Ethereum, the second largest cryptocurrency in the world, to understand and help their decision-making process.

There are so many factors in composing the market Cap of Cryptocurrency. So, it is very difficult for many people to understand what is Ethereum, and what factor is related in market capitalization. If There are some main factors to understand the market cap of Ethereum, we can consider these factors first, and then this will make our decision clear, spending short time to understand current state of Ethereum market.

1. **Method**

**2-1) Find the reliable data sets.**

For our project, we need Two main Data sets. One Data set is related to the structure of Ethereum data sets.

This data set should be real time data. Unless, our predicting process will be less reliable and meaningful.

Also, the other data is related the real market capitalization. Because we need to compare what we predict

From Ethereum data set With Real value of markets. Also, this dataset ‘s accuracy is very important.

Therefore, in getting and extracting these two data sets, we will try to find reasonable datasets

From reliable source. Maybe this process will account for many times of our whole project.

**2-2) Preprocess the given data and Use Data visualization.**

If reliable data sets are given, next step will be a preprocessing, cleaning the data. And show dataset’s

Structure and relation visible to understand our datasets, and this may give a good insight to find some

Prediction of our purpose.

First, for preprocessing, we need to explore our data structure. We need to understand what the variable means, and what the unit and scale of variable is and so on. After that, we check if there are some missing values on each variable. And if there are, then we also determine our policy for treating these values for each variable.

Second, for Data visualization, first we can visualize the flow of our target variable, Market Capitalization.

Also, we can express each variable which we are explore by some visualization components, like a boxplot, and so on. And finally, we can visualize each variable scatter plot and our target variables timestamp plot together to show the relationship and flow of our dataset. This will be more give prediction for purpose of project

**2-3) Check there is a chance of data reduction.**

As we know, too many variables can lead to misunderstanding of our predicting. Also, too many variables

Have a lot of time for preprocessing and using all these variables to datamining technique. For those reasons, if there is a chance of data reduction, It should be processed prior to applying datamining technique. To find data reduction, there will be any methods – visualization, data summary, using pivot table, Principal Component Analysis and so on. We can choose some methods above which can apply to our dataset’s variable with no mismatch. For selecting, we need to know how each method can be a way to data reduction.

**2-4) Data partition: training dataset, Validation Dataset, Test dataset**

**2-5) Supervised prediction task: use Linear Regression and other models.**

**2-6) Evaluate and compare model.**

**2-7) Deploy the model using new data.**

**3.(Current) result**

**3-1) Get the reliable Data set.**

Fortunately we can get, and extract reliable quality and enough quantity of dataset by two sources :

One is ‘bigquery’ from google, and the other is a official site relate to Cryptocurrency market Capitalizations

Named ‘CoinMargetCap’ . First from ‘Bigquery’ we can extract Ethereum datasets. This is reliable in that

Google operates Ethereum’s full node directly, and therefore, collects the real time data. In the sense of ‘real time’, ‘CoinMargetCap’ also gives a real time information of market capitalization of each cryptocurrency. So

We don’t need to the quality of datasets. We can just select what variable would be meaningful and

Interesting for our purpose of predicting the market capitalization. And after extracting those variables

From bigquery, then we have a dataset to use for predicting the target, market capitalization.

**3-2) Preprocess the Dataset.**

Next for preprocessing the dataset, we explored the datasets, more specifically, we opened the datafile, and,

Use some R functions like View(), head(),dim(),summary() to understand the data structure, and how

Many variables are in our dataset, and what the scale and the unit of each variable is.

And next, we detected if there are any missing-values or any outliers (using na.omit() function). But there

Are no missing-values!

**3-3) Data visualization.**

**3-4) Data reduction .**

And Then we can Checked if there is a possibility of data reduction. to determine this possibility.

First, we can do some data summary to check each variable’s structure and find this variable is meaningful

Or not. We checked each variable’s standard division, min, max, median, length, missing value’s count.

From this process, we found that there are two variables which we think may have a little impact on

Target purpose: predicting the market capitalization. These two variables are “x” variables and “tx\_date”

Variables, one is just an order of data (rows in dataset) and the other is just an date of each transaction, no meaningful for market capitalization.

So, we delete these variables. and next we checked each two variable’s correlation from our data set,

And from this process, we can find a variable which showed a low relation with market.cap variable.

So, we deleted tx\_volume\_ERC20 variable from our datasets for our accuracy of our project.

And finally, we applied PCA including normalization to our Dataset and visualize the contribution of each variables to PCs.

From this process, we can understood each variable is a meaningful for our dataset.