Machine Learning Approach to U.S. Stock Investments

Ahmet Hakan Ekşi
Computer Engineering
Marmara University
Istanbul, Turkey
ahe9953@gmail.com

Canberk Köroğlu
Enivironmental Engineering
Marmara University
Istanbul, Turkey
canberkkoroglu@gmail.com

Erim Varış
Enivironmental Engineering
Marmara University
Istanbul, Turkey
ermvrs@live.com

***Abstract*— Successful predictions of stock movements have always been utterly important in the market. Billions of dollars were spent in infrastructures and R&D departments in order to predict the future outcomes of share-stocks. It is well known a successful prediction of a stock's future price could yield significant profit. By utilizing various classification and *clustering* methods we tried to approximate if stocks will be profitable in the following years.**

Keywords—U.S. Stocks, prediction, classification, clustering, Naïve Bayes, ANN, ANOVA-F, descriptive analysis, k-NN, Mutual Information, adjusted rand index.

I. INTRODUCTION

Finance markets are one of the biggest inventions for recent decades. By impacting various business areas and investors all around they stand to be a landmark in earning money with low risks. [1]

Decision-making for investment and risk management are hard to manage without prediction of stock movements. Many have done great efforts in this subject. Methodologies developed have not been sufficient in correct predictions and successful prediction still remains a formidable challenge due to many variables beside the entropic values of volatility and linearity of the stock prices. [2][3][4]

Efficient Market Hypothesis suggests all available information is provided by stock prices. Thus, meaning with the use of historical information prediction of stock prices is not possible. However, there are many fundamental research made contrary to this point. Evidences suggests one can have an edge in the stock market if skills for information processing are better compared to other analysts.[5][6][7]

II. RELATED WORK

There have been many attempts and methodologies made for stock market forecasting (prediction) to this day. Due to its profitability and also it being a long-time challenge many data analysts collaborated and experimented on this topic.

In 2016 M.M. Rounaghi, F. Nassir Zadeh conducted a research including 350 firms listed in London Stock Exchange and S&P 500 from 2007 until the end of 2013. Their study examined monthly and yearly predictions of Stock Returns on London Stock Exchange and S&P 500 by using ARMA (Auto-Regressive Moving Average) Model. Experiment resulted in success on medium and long-term predictions.[8]

In 2007 Qian Bo and Rasheed Khaled investigated predictability of the Dow Jones Industrial Average index to show that not all periods are equally random. They used the Hurst exponent to select a period with great predictability. Parameters for generating training patterns were determined by auto-mutual information and false nearest neighbor methods. Machine learning classifiers such as Artificial Neural Network, decision tree, k-nearest neighbor algorithms were used to generate patterns. Experiment achieved prediction accuracy of 65 percent. [9]

Jigar Patel, Sahil Shah, Priyank Thakkar and K. Kotecha researched predicting direction of movement of stock and stock price index for Indian stock markets. They compared four prediction models ranging from ANN, Support Vector Machine, random forest and Naïve-Bayes with two different approaches. First approach involved computation of ten technical parameters using stock trading data while second approached focused on representing said data as a trend deterministic data. Experiment resulted for first approach random forest method performed other three methods. [10]

On the contrary of Efficient Market Hypothesis which suggest a prediction of stock movement cannot pass threshold of 50 percent, most experiments concluded that even up to 65 to 70 percent accuracy can be reached with various preprocessing and classification methods.

III. APPROACH

Our dataset consists of 5 subsets which are divided by years subsequent from 2014 to 2018. All sub-data-sets include 224 instances and around 3808 to 4960 attributes. Figure 1 shows direction and pathway we took for this experiment as a simple flowchart.

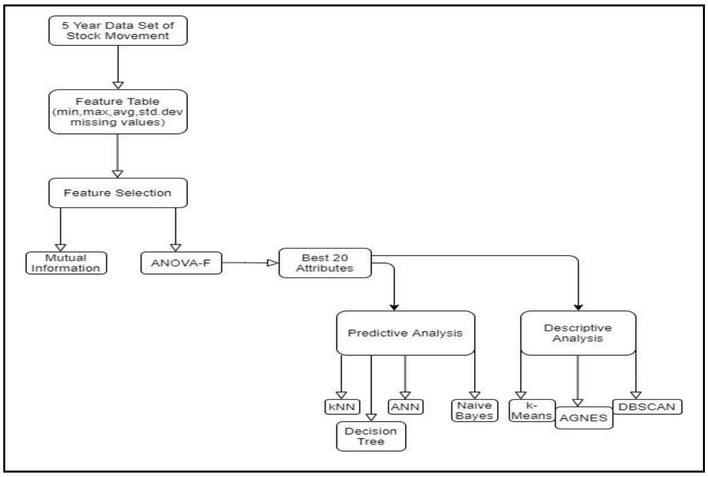


Fig.1. Simplified Flowchart of Experiment

A. *Feature Selection*

The analysis of variance can be used to describe otherwise complex relations among variables. ANOVA is a form of statistical hypothesis testing. When a test result is statistically significant it is deemed unlikely to be occurred by chance. [26][27]

F- Test on the other hand is a statistical test. It is most often used to compare statistical models fitted to a data set in order to identify the model best fitting to the population. [28]

B. *Classification Algorithms*

After researching for related work in this topic we decided to include four different classification methods in this experiment.

1- *Decision Tree*

Decision tree is a structure similar to a flowchart. Internal nodes inside the structure can be assumed as a test on an attribute. It branches as the classification takes place and every branch represents the outcome of the test. [12]

We conducted two different decision tree algorithms based on Gini index and information gain ratio.

i. Gini Index

Gini index is a popular measure of wealth inequality. It is a relative measure of variability that equals twice the distance between curve of actual distribution of wealth and the curve of total equality. [13]

ii. Gain Ratio (Mutual Information)

MI of two random variables is a measure of the mutual dependance between two variables. It quantifies the amount of information obtained about one random variable through observing the other random variable. [14][15]

2- *Naïve Bayes Classifier*

Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between features. [16][17]

These classifiers require high number of parameters in the number of variables in a learning problem. [18]

3- *Artificial Neural Network*

Artificial neural networks are computational models inspired by central nervous systems. They are used to estimate functions depending on large number of inputs. They are generally represented as interconnected neurons which feed values to each neuron in the hidden layer. Hidden layer then provides the output. Learning process generally occurs when weights among layer change in a way that predicted amounts differ from calculated amounts. [11]

Their most significant benefits are: high processing speed, pattern learning, knowledge generalization, flexibility against unexpected errors.

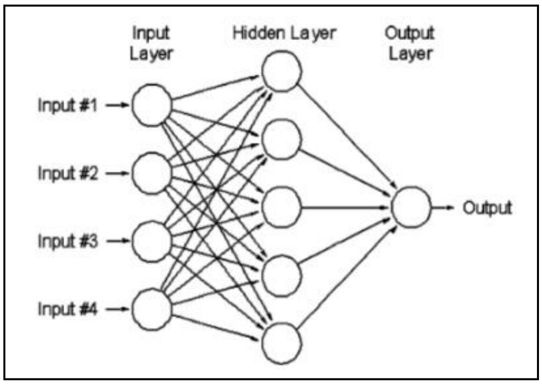


Fig.2. Back-Propagation model of artificial neural network

4- *k- Nearest Neighbor*

KNN is a non-parametric classification method. It can be used for both classification and regression. In both cases input consists of the "k" closest training examples from the data set. [19]

An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most among its k nearest neighbors. [16]

The best choice of k depends upon the data; generally, larger values of k reduce effect of the noise but make boundaries between classes less distinct. Because k is a user defined variable best value of k can be obtained by using evaluation tables and trial error. [20]

C. *Clustering Algorithms*

Unsupervised learning is a kind of machine learning strategy utilized for drawing inferences from the datasets containing input data without any labeled responses. The purpose of the unsupervised machine learning method is to determine the similarities in data points and assemble similar data together. The commonly used unsupervised learning technique is cluster analysis, which is massively utilized for exploratory data analysis to determine the hidden patterns and to group the data. [21]

i. K-Means Clustering

k-means clustering is a method of vector quantization that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers), serving as a prototype of the cluster. k-means clustering minimizes within-cluster variances. [22][23]

ii. DBSCAN

At present, the DBSCAN clustering algorithm has been commonly used principally due to its ability in discovering clusters with arbitrary shapes. The main mechanism for DBSCAN is that for any data set, at least a minimum number of objects has to be contained in the neighborhood of a given radius to be a cluster; otherwise, they are regarded as noise data. [24]

iii. AGNES

Agglomerative nesting is a hierarchical clustering. Which seeks to build a hierarchy of clusters. It is a bottom-up approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. [25]

IV. EXPERIMENTAL SETUP

A. *Feature Selection and Dimensionality Reduction*

In our datasets we had 224 instances. One instance was for class label, one for generating class label and one for stock ID. From remaining 220 features one is a categorical feature to specify stock owner company sector. Rest of the features are various financial indicators.

To choose our best 20 attributes we tried both Mutual Information and Analysis of Variance tests. Because of missing values inside these attributes were close to our threshold of 33 percent we replaced said values with the mean of the mean values of their corresponding attributes. After this process we observed for MI even our best 20 attributes had more than 33 percent of missing data. Therefore, we chose ANOVA-F value for feature selection.

B. *Classification Experiments*

To start off the classification experiments we did not go the usual way of k-fold cross validation. Reasoning behind this is k-fold cross validation is done when dataset is not sufficiently large, but our case is different. We have 5 different datasets with 225 instances and 3000-5000 attributes respectively.

Hidden layers in ANN, and user defined *k* in KNN was determined by trial error with respect to their performance evaluation tables consisting of accuracy, area under curve (AOC), F1-micro and F1-macro averages.

C. *Clustering Experiments*

For K-Means clustering we needed to choose a optimum K value. In order to choose this value we conducted an Elbow test, and with visual cues and code results we decided a K value of *6*.

DBSCAN performs remarkably when working with noisy data. [24] Therefore we wanted to include DBSCAN to our clustering experiments due to our data also having a lot of noise.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Table.1 Classification algorithm evaluation table.

Experiment	**Accuracy**	**F1-macro**	**F1-micro**	**AUC**
Decision Tree with Gini Index	0.528916	0.510808	0.528916	0.534653
Decision Tree with Gain Ratio	0.521403	0.505316	0.521403	0.531224
Naive Bayes	0.670993	0.428472	0.670993	0.603608
ANN with 1 hidden layer	0.680328	0.629554	0.680328	0.682963
ANN with 2 hidden layer	0.561475	0.555338	0.561475	0.631891
KNN-3	0.570355	0.550413	0.570355	0.592899
KNN-9	0.602687	0.579037	0.602687	0.624357
KNN-149	0.665528	0.627639	0.665528	0.670177

The performance evaluation table shows the most accurate method is ANN with 1 hidden layer. Considering Area Under Curve (AUC) which measures performance across all possible classification thresholds it suggests ANN with 1 hidden layer overperforms when compared with Naïve-Bayes which is also less accurate. Also, F1-micro (micro-averages) suggests ANN with 1 hidden layer performs better. Furthermore, F1-macro (macro-averages) indicate ANN with 1 hidden layer performs better.

Figure 3 shows our ROC curve to emphasize and support the evaluation table shown in Table 1. In this curve sensitivity and specificity are visualized. Curves coming to top left corner of the graph will be more accurate than the ones at 45 degrees. In our case ANN with 1 hidden layer is the best performing method followed by Naïve Bayes.

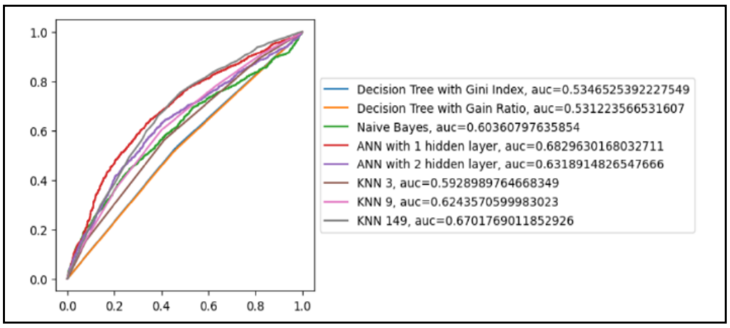


Fig.3. ROC curve of classification algorithms.

Figure 4 shows our confusion matrix which is used to also determine the performance of our best performing algorithm ANN with 1 hidden layer.

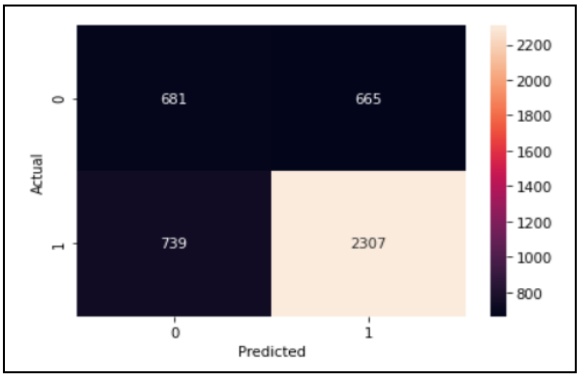


Fig.4. Confusion matrix for ANN with 1 hidden layer.

In Figure 4 it is observed that ANN with 1 hidden layer method predicted true positive cases 2307 times out of 4392 cases and true negative cases 681 times.

This suggests 68% of the algorithm is classifying correctly. Even though theoretically any algorithm can reach 100% accuracy, our achievement of 68% is good enough.

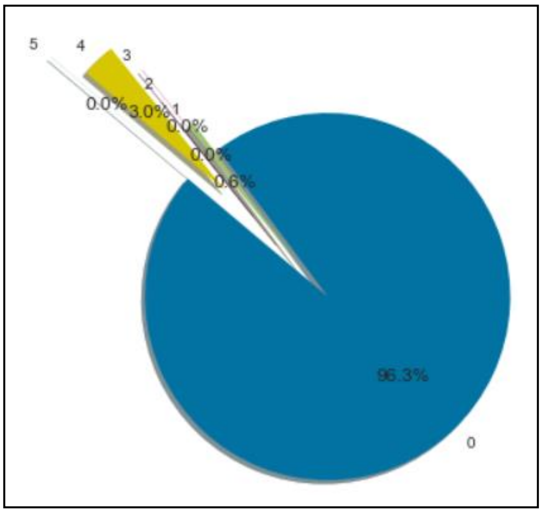


Fig.5. Clustering with K-Means K=6.

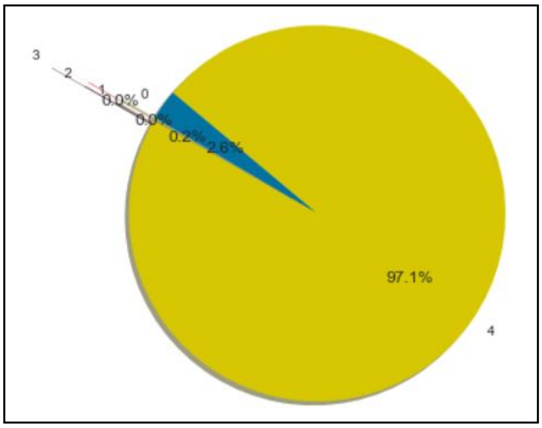


Fig.6. Clustering with AGNES 5 cluster.

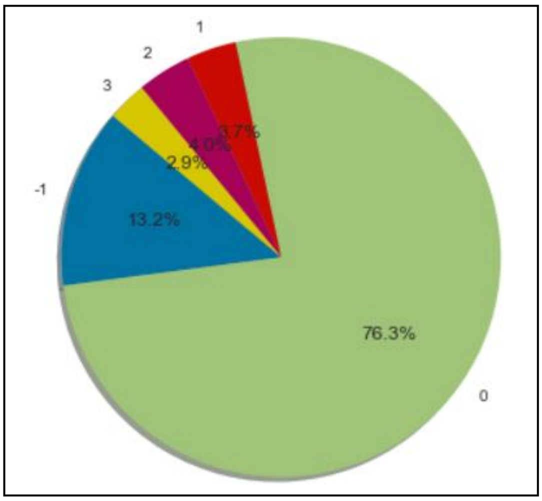


Fig.7. Clustering with DBSCAN eps=0.04 and min=72.

After evaluating our clustering algorithms, we observed from the point of Silhouette Value which is the similarity of the value to its cluster, AGNES is the best followed by K-Means. DBSCAN on the other hand performed poorly.

From the point of ARI (adjusted Rand index) all of the algorithms performed poorly. This can be related to most financial indicators being relevant only to themselves and ARI does not see local relevancy between these attributes. Therefore, considers the values in the clusters randomly placed.

Normalized Mutual Information values suggest high number of correct clusters for the instances. It is observed that K-Means algorithm is more reliable. Although the NMI values are too low to discriminate.

Evaluation table for clustering algorithms can be found in the appendix under *Delivery 5 Table.5.*

VI. CONCLUSION

To conclude our experiments were success from the point of Efficient Market Hypothesis which suggests predictions on stock movements cannot exceed 50 percent. We have crossed that treshold and reached 68 percent accuracy in our predictions.

When we chose our test set as 2018 which was in our dataset we simply tried to put ourselves as analyst in the last quarter of 2017 and wanted to be rich by predicting next years stock directions. To some extend our goal was achieved.

Finally the challenge of prediction in stock markets will still be significant. Efficient data collection without faulty or missing data will possibly increase accuracy which the algorithms perform.

REFERENCES

[1] Deepak Kumar, Pradeepta Kumar Sarangi, Rajit Verma, A systematic review of stock market prediction using machine learning and statistical techniques, Materials Today: Proceedings, 2021

[2] Wei-Jia Wang, Yong Tang, Jason Xiong, Yi-Cheng Zhang, Stock market index prediction based on reservoir computing models, Expert Systems with Applications, Volume 178, 2021

[3] Y.S. Abu-Mostafa and A.F. Atiya Applied Intelligence, 6 (1996), pp. 205-213

[4] P. Yu and X. Yan Neural Computing & Applications, 32 (2020), pp. 1609-1628

[5] E. Fama The Journal of Business, 38 (1) (1965), pp. 34-105

[6] L. Pedersen Princeton University Press, 2015

[7] Nan Jing, Zhao Wu, Hefei Wang, A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction, Expert Systems with Applications, Volume 178, 2021

[8] Mohammad Mahdi Rounaghi, Farzaneh Nassir Zadeh, Investigation of market efficiency and Financial Stability between S&P 500 and London Stock Exchange: Monthly and yearly Forecasting of Time Series Stock Returns using ARMA model, Physica A: Statistical Mechanics and its Applications, Volume 456, 2016

[9] Qian Bo, Rasheed Khaled, Stock market prediction with multiple classifiers, 2007

[10] Jigar Patel, Sahil Shah, Priyank Thakkar, K Kotecha, Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques, Expert Systems with Applications, Volume 42, Issue 1, 2015, Pages 259-268

[11] Javad Zahedi, Mohammad Mahdi Rounaghi, Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange, Physica A: Statistical Mechanics and its Applications, Volume 438, 2015, Pages 178-187

[12] Kamiński, B.; Jakubczyk, M.; Szufel, P. "A framework for sensitivity analysis of decision trees". Central European Journal of Operations Research. 26 (1): 135–159, 2017

[13] Edward Furman, Yisub Kye, Jianxi Su, Computing the Gini index: A note, Economics Letters, Volume 185, 2019

[14] Kreer, J. G. "A question of terminology". IRE Transactions on Information Theory. 3, 1957

[15] Wolpert, D.H.; Wolf, D.R. "Estimating functions of probability distributions from a finite set of samples". Physical Review E. 52 (6): 6841–6854, 1995

[16] Piryonesi, S. Madeh; El-Diraby, Tamer E. "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146, 2020

[17] Hastie, Trevor. The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. Tibshirani, Robert., Friedman, J. H. (Jerome H.), 2001

[18] Russell, Stuart; Norvig, Peter. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall, 1995-2003

[19] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). The American Statistician. 46 (3): 175–185, 1992

[20] Everitt, Brian S.; Landau, Sabine; Leese, Morven; and Stahl, Daniel "Miscellaneous Clustering Methods", in Cluster Analysis, 5th Edition, John Wiley & Sons, Ltd., Chichester, UK, 2011

[21] Satish Chander, P. Vijaya, 3 - Unsupervised learning methods for data clustering, Editor(s): D. Binu, B.R. Rajakumar, Artificial Intelligence in Data Mining, Academic Press, 2021, Pages 41-64

[22] Michael McCool, Arch D. Robison, James Reinders, Chapter 11 - K-Means Clustering, Editor(s): Michael McCool, Arch D. Robison, James Reinders, Structured Parallel Programming, Morgan Kaufmann, 2012, Pages 279-289

[23] Forgy, Edward W. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics, 1965, Pages 768–769

[24] Qidan Zhu, Xiangmeng Tang, Ahsan Elahi, Application of the novel harmony search optimization algorithm for DBSCAN clustering, Expert Systems with Applications, Volume 178, 2021

[25] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005, Pages 321-352

[26] Ronald A. Fisher. Metron. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample, Pages 3–32, 1921

[27] Michael Meyners, Anne Hasted, On the applicability of ANOVA models for CATA data, Food Quality and Preference, Volume 92, 2021

[28] Box, G. E. P. "Non-Normality and Tests on Variances". Biometrika. 40 (3/4): Pages 318–335, 1953

APPENDIX

Delivery #2

Table.2. Structure of our data

Feature name	Description	Type	Min	Max	Avg.	Std.Dev.	Entropy	# of values	missing values %
Stock	company stock code	text						22077	0
Revenue	financial indicator	numeric_continuous	-627616000	1,88689E+12	5161618858	31973144008	9,672183349	20906	5,3
Revenue Growth	financial indicator	numeric_continuous	-12,7693	42138,6639	3,622214228	312,6481703	8,712478398	19989	9,46
Cost of Revenue	financial indicator	numeric_continuous	-2986887895	1,58153E+12	3258565393	25830920898	8,324618032	20306	8,02
Gross Profit	financial indicator	numeric_continuous	-12808000000	4,6216E+11	1970452467	8735750257	9,662553969	20870	5,47
R&D Expenses	financial indicator	numeric_continuous	-109800000	28837000000	103333292,3	767606165,7	4,607590975	19939	9,68
SG&A Expense	financial indicator	numeric_continuous	-140159420,3	1,85683E+11	869927885,8	3804283410	9,766505565	20408	7,56
Operating Expenses	financial indicator	numeric_continuous	-5495511688	3,05605E+11	1368669853	5662983943	9,93257468	20375	7,71
Operating Income	financial indicator	numeric_continuous	-19339000000	1,56554E+11	589697890,7	2976453997	9,91550949	20976	4,99
Interest Expense	financial indicator	numeric_continuous	-1710953646	31523000000	97789390,14	499654291,6	7,713733971	20358	7,79
Earnings before Tax	financial indicator	numeric_continuous	-21772000000	87205000000	492500290,6	2484345450	9,922069467	20713	6,18
Income Tax Expense	financial indicator	numeric_continuous	-7,38E+11	8,49E+11	130002043,9	7962080411	8,709526051	20489	7,19
Net Income - Non-Controlling int	financial indicator	numeric_continuous	-1587227414	6430813535	13390062,13	143753327,2	4,106210439	19818	10,23
Net Income - Discontinued ops	financial indicator	numeric_continuous	-15914500000	8368000000	-3240313,279	242497021,3	2,302210594	19818	10,23
Net Income	financial indicator	numeric_continuous	-23045000000	2,33997E+11	388672668,9	2643759192	9,88869576	20512	7,09
Preferred Dividends	financial indicator	numeric_continuous	-161000000	2741588000	4673906,343	53287995,57	2,236780508	19818	10,23
Net Income Com	financial indicator	numeric_continuous	-23045000000	2,33997E+11	387125353,8	2633920242	9,881889132	20686	6,3
EPS	financial indicator	numeric_continuous	-101870898,1	8028004,014	-10657,47956	896097,6629	7,508129002	20776	5,89
EPS Diluted	financial indicator	numeric_continuous	-101870898,1	6624003,312	-10735,82333	895348,7469	7,496316581	20785	5,85
Weighted Average Shs Out	financial indicator	numeric_continuous	0	1,11292E+11	263176678,5	2046155695	9,889783661	20583	6,77
Weighted Average Shs Out (Dil)	financial indicator	numeric_continuous	0	1,11292E+11	266492638,3	2136719667	9,896613459	20140	8,77
Dividend per Share	financial indicator	numeric_continuous	0	10100,664	1,215391588	72,20095121	4,410088667	19818	10,23
Gross Margin	financial indicator	numeric_continuous	-74,3191	31	0,487843946	0,945601415	7,740068322	20878	5,43
EBITDA Margin	financial indicator	numeric_continuous	-24207	3090,87	-8,88059069	239,6253606	7,829068728	19630	11,08
EBIT Margin	financial indicator	numeric_continuous	-24242	1056,4658	-7,258787199	217,7949818	8,811856665	20403	7,58
Profit Margin	financial indicator	numeric_continuous	-24414	3090,87	-9,307057101	243,2054675	7,617884372	19633	11,07
Free Cash Flow margin	financial indicator	numeric_continuous	-23256	689,8297	-6,293523163	204,8696797	8,803824835	19786	10,38
Sector	specified sector of selected company stock	nominal	Communication Services	Financial Services			2,187700308	22077	0
PRICE VAR NEXT YEAR	annual stock value variation for following year in percent	numeric_continuous	-100,3972195	2418600,915	269,8892659	19346,1736	9,990984519	22077	0
Class	indicator for decision making with respect to price var value being positive or negative	boolean						22077	0

This is a simplified table for understanding the dataset we used. This is only a part of our 225 instance dataset.

Delivery #3

Table.3. Feature Selection by Mutual Information

Attribute	Score	Missing Values %
Total non-current assets	0.022112280957419017	27,277256873669
EBIT Margin	0.022054091974503587	7,582551977171
eBITperRevenue	0.01920992042646863	12,963717896453
Gross Profit	0.019154132780264455	5,467228337183
payoutRatio	0.018948078374790178	10,825746251755
Total non-current liabilities	0.018644065311503644	27,340671286860
Earnings before Tax	0.01823924284012657	6,178375685102
ebitperRevenue	0.018153038022910906	12,963717896453
EPS	0.017993695863804504	5,893010825746
Earnings Before Tax Margin	0.017921437192584877	5,412873125878
Free Cash Flow margin	0.01782413204082678	10,377315758482
Profit Margin	0.016968018418720643	11,070344702632
netProfitMargin	0.01651437783387477	12,963717896453
EBITDA Margin	0.016334752774368688	11,083933505458
Total shareholders equity	0.016233635705729332	6,214612492639
dividendYield	0.016230110792283314	14,938623907234
Consolidated Income	0.016169867868345955	7,097884676360
Other Assets	0.016166675280912646	30,601983965213
Net Profit Margin	0.016102373873498665	7,799972822394
Net Income	0.01600560547987806	7,088825474476

Table.4. Feature Selection by ANOVA-F value

Attribute	Score	Missing Values %
Dividend payments	40,99933856905	9,76581963129
Gross Margin	32,31890334987	5,43099152965
Earnings before Tax	31,54958855350	6,17837568510
Free Cash Flow	31,14306858038	7,42401594420
Operating Income	27,75962964513	4,98709063731
Stock-based compensation	27,30370658113	7,84979843276
SG&A Expense	23,46423370946	7,55990397246
Cash and short-term investments	22,72453457323	11,21076233184
R&D Expenses	22,33696073442	9,68428681433
Issuance (buybacks) of shares	22,10822550551	8,79648502967
EBITDA	20,25353857580	7,94492005254
Gross Profit	19,58500849234	5,46722833718
EBIT	16,07228249774	6,45468134257
Retained earnings (deficit)	15,51009740358	5,28604429950
ebitperRevenue	15,12235226494	12,96371789645
Goodwill and Intangible Assets	14,64329427410	6,50903655388
Total current liabilities	14,41323496096	11,45083118177
Short-term investments	14,14247371009	12,28427775513
Operating Expenses	13,79281169365	7,70938080355
Net Income Com	13,71337691993	6,30067491054

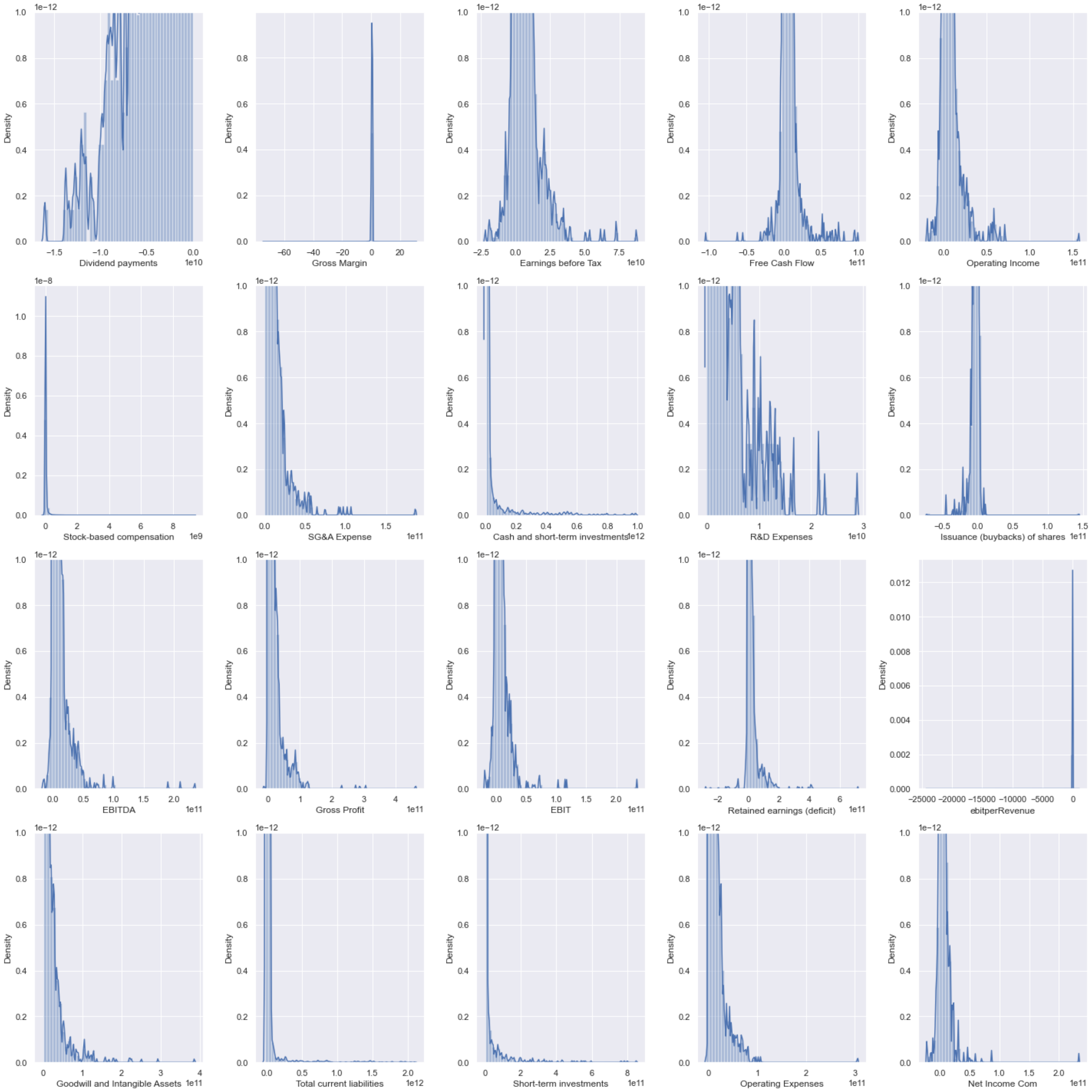


Fig.8. Density (Distribution) Charts

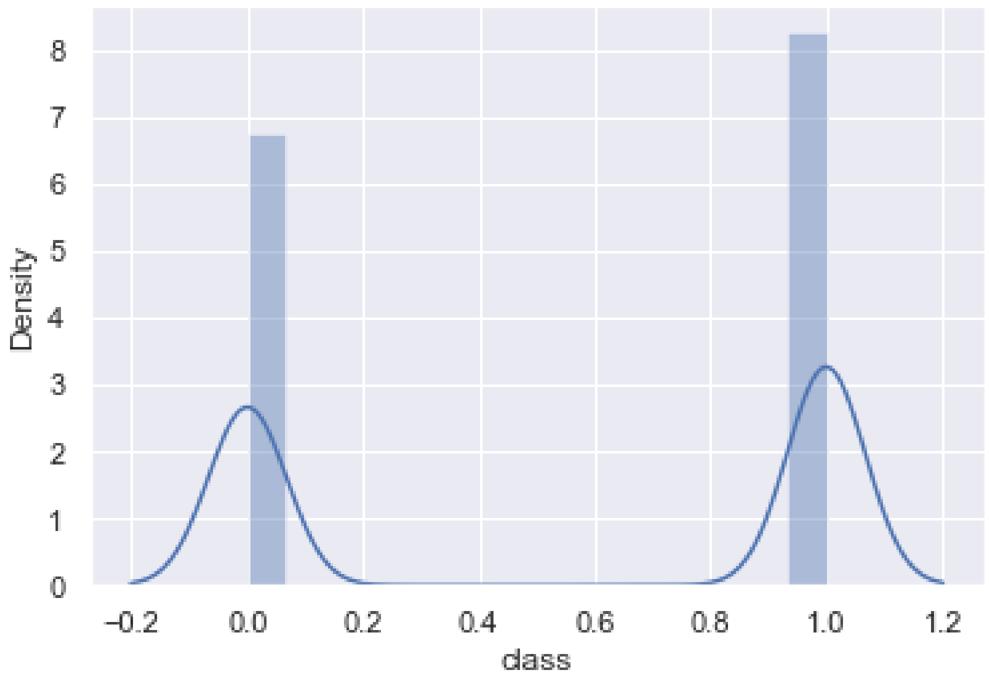


Fig.9. Class Label Chart

This is our class labels density chart. As you can see from the chart our stocks prices most likely to increase next year.

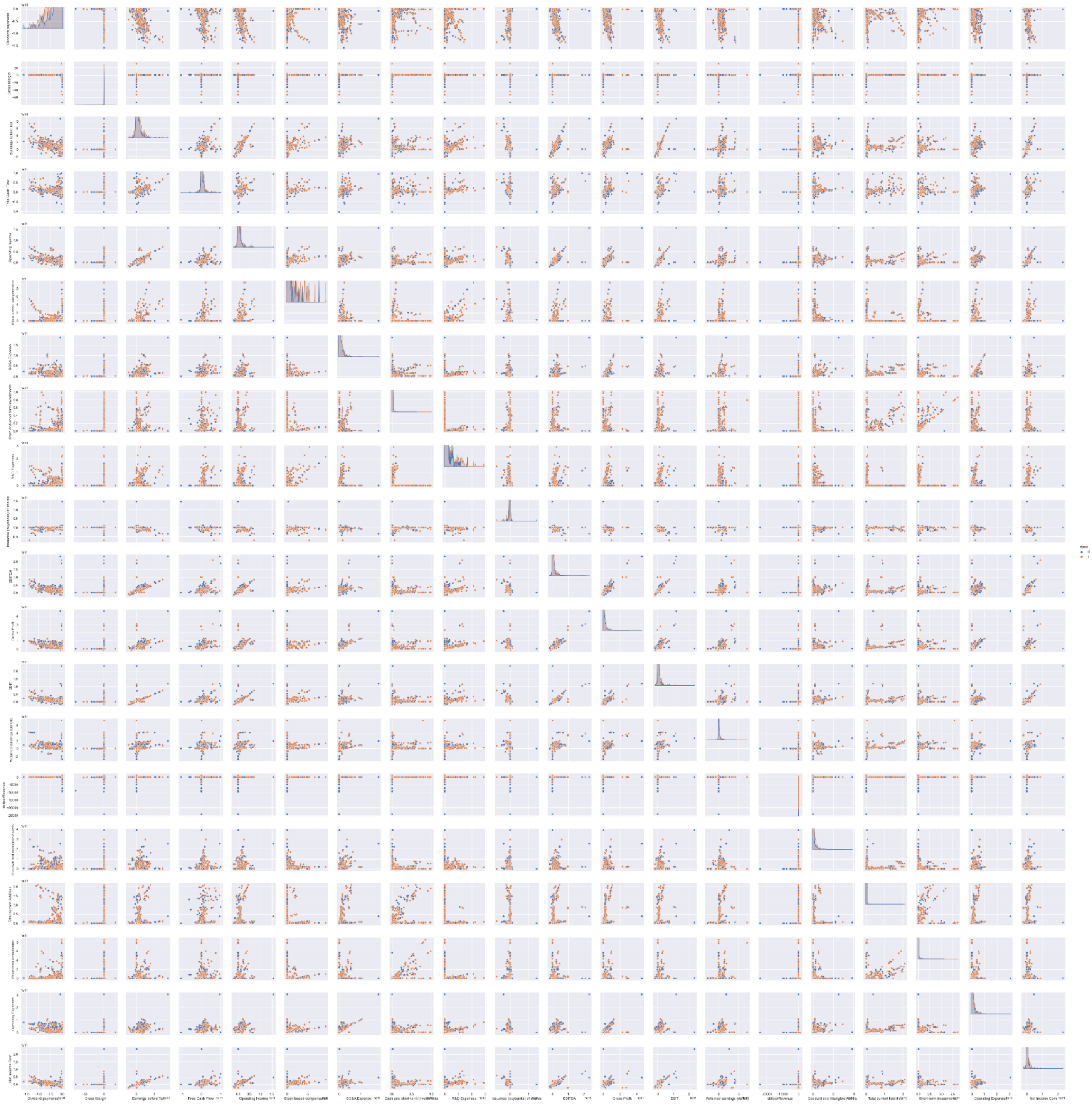


Fig.10. Scatter Plot Matrix

In Figure 10 relationships between financial indicators are observed. Interactions between these indicators are mostly positively correlated for e.g. (3,6) or negatively correlated for e.g. (17,17) due to slopes being either negative or positive.

Delivery #5

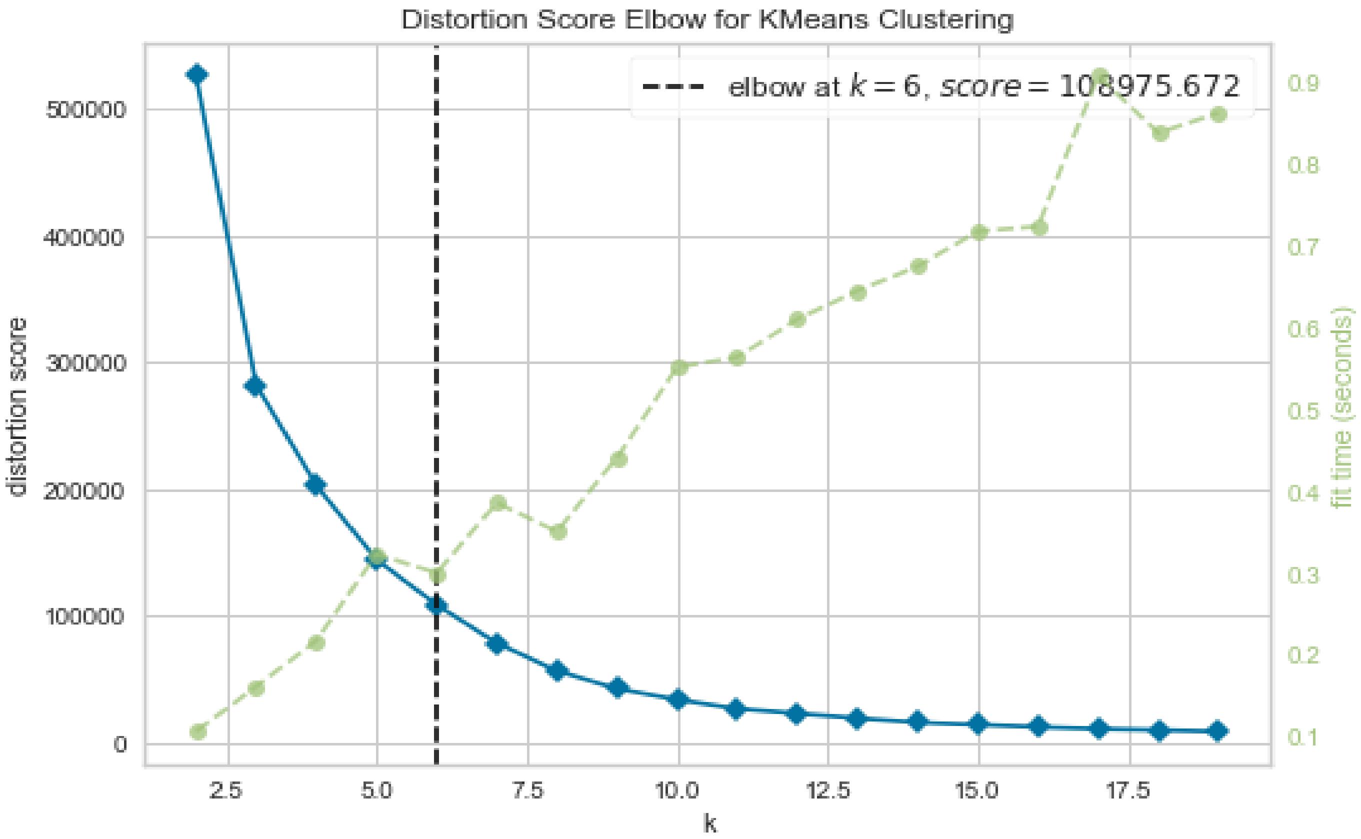
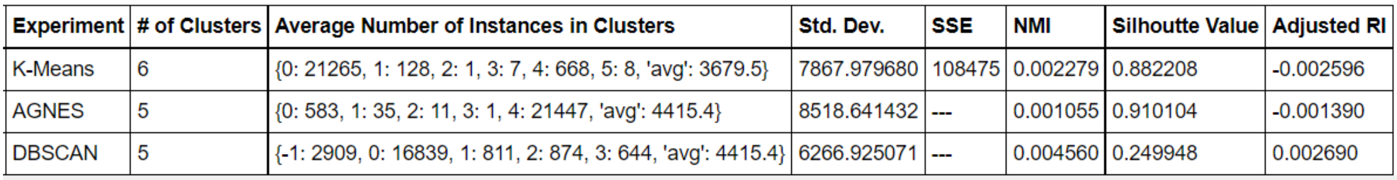


Fig.11. Elbow Method for K-Means

Elbow method is used to find out exact number of clusters to both not divide information unnecessarily and add much information possible. By looking at the blue-line we can see the elbow(bend) is exactly at point k=6. Also point of infliction suggest that point k=6 is the best fitting value for K-Means clustering.

Table.5. Evaluation Table for clustering algorithms



Experiment	# of Clusters	Average Number of Instances in Clusters	Std. Dev.	SSE	NMI	Silhoutte Value	Adjusted RI
K-Means	6	{0: 21265, 1: 128, 2: 1, 3: 7, 4: 668, 5: 8, 'avg': 3679.5}	7867.979680	108475	0.002279	0.882208	-0.002596
AGNES	5	{0: 583, 1: 35, 2: 11, 3: 1, 4: 21447, 'avg': 4415.4}	8518.641432	---	0.001055	0.910104	-0.001390
DBSCAN	5	{-1: 2909, 0: 16839, 1: 811, 2: 874, 3: 644, 'avg': 4415.4}	6266.925071	---	0.004560	0.249948	0.002690