Heng Zhou, Hongyang Zheng,

Youyou Xie, Zhengqian Xu

Professor Lisa Singh

Analytics 501 Project Part2

31 October 2018

Project Report

***Introduction***

For this project, we’d like to investigate the factors affecting stock market among 11 sectors. Stock market plays an important role in the whole economy, it is a key driver for economic trade, growth and prosperity whether in the financial markets or product markets (agriculture, real estate, manufacturing etc.). However, it can’t be ignored that the risk of stock is the highest among financial products. Investigating factors affecting stock market is not only good for company, but also a good method to forecast and regulate financial market. Our group focuses on analyzing the factors among different sectors and tries to figure out more pertinent strategy on the basis of general factors and strategy.

To investigate the questions about stock market discussed in data science problem section: the factors impacting stock trading among different sectors, we need to collect both general stock market information, specific stock ask, bid price and size in a specific stock exchange website as well as general company information such as their industry and sector in the market. Our dataset has three part: general company information for each company, general stock market information about each stock and specific stock price as well as size in a stock exchange website. For the first part, company name and stock symbol are the basic information that should be contained in our dataset. Moreover, which industry or sector each company belongs to is also critical for our project, so that we can separate these stocks into different groups and analyze problems in a bigger scale. About general stock market information, we have average total volume and latest price to observe average stock price and stock size in the market. We also have market capitalization, risk coefficient and price change percentage of each stock, which are also key element. As for the last part, we have ask stock price, ask stock size, bid stock price, bid stock size in IEX stock exchange website as stock information in the narrow scope.

One direction of this project is to analyze stock price and stock volume by industry or sector and the results will show how market factors such as market capitalization or risk coefficient influence different groups by different levels. For example, our null hypothesis is that the price and volume of financials sector stocks are more vulnerable by bigger risk coefficient than stocks in other sectors. Another direction of this project is to do clustering analysis regarding to the behavior of stocks, derive stock typology and then give stock combination which has the minimum riskiness for example.

***Exploratory Analysis***

***Part1 - Basic Statistical Analysis and Data Cleaning Insight***

As a fundamental step for exploratory analysis, this part continues to handle and discuss missing values, noisy values and outliers. Having done several further cleaning steps, we investigate the data by calculating critical statistics for each single variable such as mean(mode if it is a categorical variable), median, standard deviation, to explore the features of the data. What is more, after having a deeper insight of our dataset, we bin three numerical variables and create three new categorical variables.

***Missing Value***

As talked before in project 1, there are six variables which have missing values: latest price, iexAskPrice, iexAskSize, iexBidPrice , iexBidSize and industry. We apply two approaches to handle these missing values: dropping it or entering valid value manually.

We drop missing values for latest price, iexAskPrice, iexAskSize, iexBidPrice and iexBidSize. The reason why we think we cannot simpliy fill these missing values by using mean or median is that the value of these attributes is independent from one observation to another. For example, there are more than 5000 different stocks, and we cannot take the average of latest price and then use it to fill the missing latest price of other stocks. In addition, we cannot fill missing values by using other website data either. Since our target website is IEX, other websites do not contain information about stock trade in IEX. Having compared the disadvantages and advantages of filling these missing values, we think it is more reasonable to drop these missing values, so we can avoid our data becoming inconsistent. As for industry column, since we only have four missing values, we just searched these company names on the internet to see which industry they belong to and filled by ourselves manually.

***Further Cleaning Step***

Taking the suggestions from project 1, we do three more cleaning steps. One is to handle all 0 values for iexAskPrice, iexAskSize, iexBidPrice and iexBidSize in data1. Although 0 is not a noise value because these values can be zero when the market percent of these stocks is so small that no one sells or buys these stocks on IEX, value 0 is still not helpful for this analysis. For iexAskSize and iexBidSize, we use median of its column to fill it. For iexAskPrice and iexBidPrice, we use the value of corresponding latestPrice and add or subtract a random number from [0,1] to represent it. The reason why we fill these 0 values is not to use them, but to keep the information of other variables. If we drop all the 0, after merging the two datasets, it only left about 1000 rows data, which is not proper. Another step we take is to transfer one category variable: sector in data1 into numerical variables. We first count the number of different categories each attribute contains and then assign numbers to each. After transformation, 1-11 represents the 11 sectors respectively. One reason we do this is to make it easy for future normalization, since normalization can only take numerical variables. The below table shows the transformation.

|  |  |
| --- | --- |
| 1 | Energy |
| 2 | Consumer Cyclical |
| 3 | Communication Services |
| 4 | Real Estate |
| 5 | Technology |
| 6 | Industrials |
| 7 | Healthcare |
| 8 | Financial Services |
| 9 | Basic Materials |
| 10 | Consumer Defensive |
| 11 | Utilities |

The third cleaning step is to separate industry in data2 to two columns: industry and sub\_industry. There are some industries in our dataset2 that have sub-industry after “-”, so we decide to separate them and store in two different variables. After separating it, we then delete “;” in sub\_industry column.

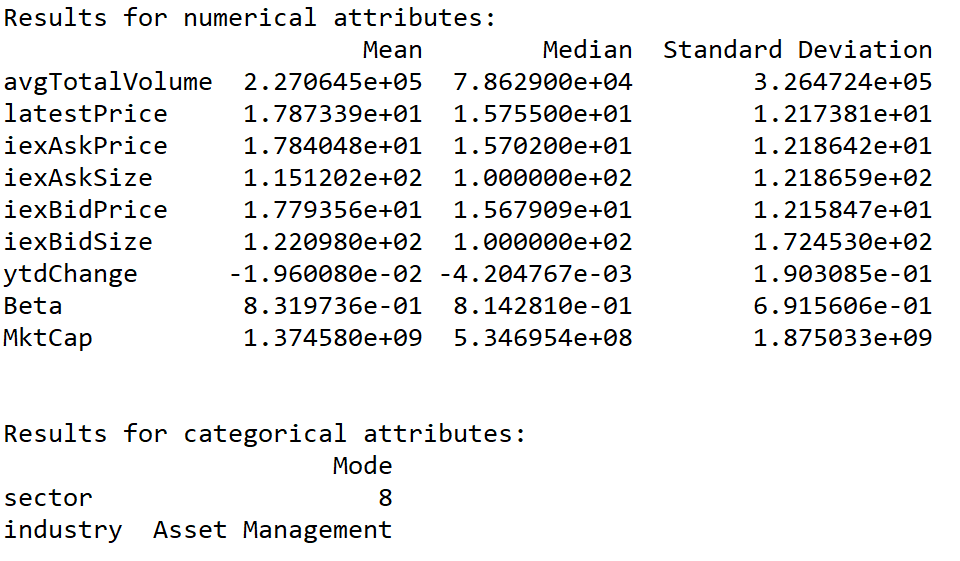
***Outliers***

To detect outliers, we first plot boxplot for every numerical attribute in all the datasets. From the plots, we can see that almost every attribute has outliers. For most attributes, we use IQR to pick outliers. The value less than Q1-1.5\*IQR or greater than Q3+1.5\*IQR will be regarded as outliers, which may be the result of error from recording and have no meaning for analyzing, so we drop them.

But for iexAskSize, from the boxplot, we can see that most iexAskSize value are under 5000, some values are too much larger than 5000 and IQR is not good for this situation. So we set the maximum value for iexAskSize is 5000 and regard those values that are larger than 5000 as outliers, then drop them.

***Basic Statistics***

We compute statistics of 11 attributes after further cleaned data. Since our data contains both numerical attributes and categorical attributes, we split them to compute the statistics. At the beginning, we pick the attributes we want to analyze and put them in two new dataframes, one contains numerical attributes and another contains categorical attributes. In the process of computing, we ignored ‘nan’ created from combining different attributes with different length. The table of results is displayed below:



*For numerical attributes:*

The mean value of avgTotalVolume is 227064.5, which refers to the 30 day average volume on all markets, the median is 78629 and standard deviation is 326472.4, which is very large and larger than mean. Thus the average volume of stocks differ to a great extent in 30 days between companies in the whole market and most of them fall within (326472.4-227064.5) above and below . The latest price has a mean value 17.87339, median is 15.755 and standard deviation is 12.17381. The standard deviation is not small, but smaller than mean, which means the latest price of stocks differs between companies and most of them fall between (17.87339-12.17381, 17.87339+12.17381). The iexAskPrice has a mean 17.84048, median 15.702 and standard deviation is 12.18642. It’s standard deviation is not small, but smaller than mean, indicating that ask price of stocks on IEX does differ between different companies. But most of them fall between (17.84048-12.18642, 17.84048+12.18642). The iexAskSize has a mean 115.1202, median is 100 and standard deviation is 121.8659. It has a big standard deviation, which is larger than mean, indicating it has a big variation of ask size of stocks on IEX between different companies and most of them fall within (121.8659-115.1202) above and below. The iexBidPrice has a mean 17.79356, median is 15.67909 and standard deviation is 12.15847. It has a big standard deviation, but smaller than mean, indicating that bid price of stocks on IEX does differ between different companies. But most of them fall between ( 17.79356- 12.15847, 17.79356+ 12.15847). The iexBidSize has a mean is 122.0980, median is 100 and standard deviation is 172.4530. It has a big standard deviation, which is larger than mean, indicating it has a big variation of bid size of stocks on IEX between different companies and most of them fall within (172.4530-122.0980) above and below. The ytdChange has a mean -0.0196, median is -0.004204767 and standard deviation is 0.1903085. It has a very small standard deviation and also bigger than mean, which means that the the price change percentage from start of year to previous close of stocks does not have much difference between different companies and most of them fall within (0.1903085- (-0.0196)) above and below. The Beta has a mean 0.8319736, median is 0.8142810 and standard deviation is 0.6915606. It has a small standard deviation and smaller than mean, which indicated that the variation of stock's volatility of stocks is small between companies and most of them fall between ( 0.8319736-0.6915606, 0.8319736+0.6915606). The MktCap has a mean is 1374580000, median is 534695400 and standard deviation is 1875033000. It’s standard deviation is very large and larger than mean. Therefore, the market capitalization of stocks in different companies differ a lot and most of data fall within (1875033000-1374580000) above and below.

*For categorical attributes:*

The mode of sector is 8, representing financial service. It means that in our data, financial service sector has maximum weight. The mode of industry is asset management, which indicates asset management weights at most in our data. The mode of industry kind of correlates to the mode of sector, since asset management is a kind of financial service.

***Binning the Data***

In this process, we bin three numerical variables and create three new columns accordingly: Beta in data2, avgTotalVolume in data1 and ytdChange in data1.

First, considering Beta is a coefficient to represent risk level of a stock, we bin it and create a column called ‘Risk’ to store the risk level of stocks. We found that when beta=1, the risk level of a stock can be considered as same as the average risk level of the whole market, so we assign ‘normal’ to represent it; when 0<beta<1, the risk level of a stock can be considered less than the average risk level of the whole market, so we assign ‘small’ to represent it; when beta>1, the risk level of a stock can be considered greater than the average risk level of the whole market, so we assign ‘big’ to represent it; when beta<0, the risk level of a stock can be considered in the opposite direction of the average risk level of the whole market. That is, when the whole stock market becomes riskier, this specific stock with beta<0 is safer, so we assign ‘negative’ to represent it. We create bins according to these criterions by ourselves and then bin Beta.

Second, avgTotalVolume is an indicator, since high average total volume reflects high level of interest in a stock at its current price, so we create a new variable called Volume\_level. We bin it into 5 categories: A, B, C, D, F, representing the interest level from highest to lowest. We use min, 25% quantile, mean, 75% quantile, max as the bins for this variable. After binning, we can know that stocks with risk level A gaining much interest in the market.

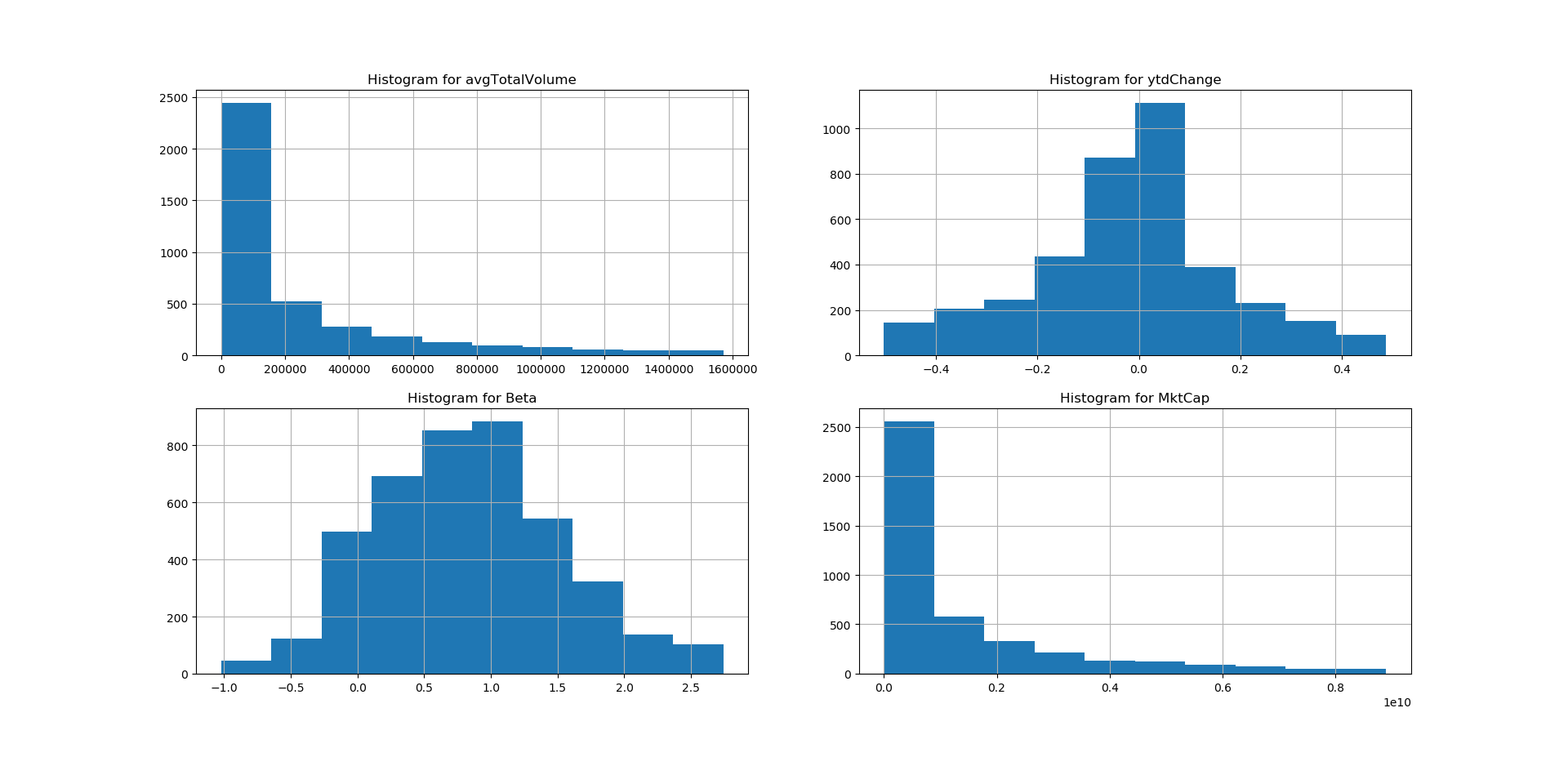
The third variable ytdChange represents the percentage change of a stock price from last year to this year. Therefore, creating a variable called Change\_level to represent the change level is reasonable. We bin it into 5 categories: ++, +, |, -, --, representing the change level from highest to lowest. We use min, 25% quantile, mean, 75% quantile, max as the bins for this variable. After binning, we can know that price of stocks with change level | do not change much during last year and price of stocks with change level ++ increases dramatically during last year.

***Part2 - Histograms and Correlations***

Another key step for general exploratory analysis is to plot histogram, scatter plot and calculate the correlation between variables. By these methods, we can explore the relationship between two or multiple variables and answer questions about distribution and linear relationship. For example, does this variable look like a normal distribution? Does the two variables have a linear relationship? If it does, is the correlation is positive or negative? Most of these questions can be answered in this step, and the results can be used for future traditional statistical hypothesis testing and more.

***Histograms:***

In this process, we make a histogram for four variables:avgTotalVolume, ytdChange in data1 and Beta, MktCap in data2.The graph is shown below.



First, for ‘avgTotalVolume’ variable, it is obvious to see that the distribution is skew and its domain ranges from 0 to 1600000. It is not hard to find that most of its value is less than 200000 with little of it is greater than 1000000 since higher average total volume reflects higher [liquidity](https://en.wikipedia.org/wiki/Liquidity) in the market. Therefore, most of stocks are of low liquidity and the proportion becomes fewer with the increase of avgTotalVolume.

Second, for ‘ytdChange’ variable, it is almost normally distributed and ranges from -0.5 to 0.5. ytdChange represents the percentage change of a stock price from last year to this year. Since ytdChange of most of stocks are between -0.2 to 0.2, the fluctuation of stock market of this year is not very large.

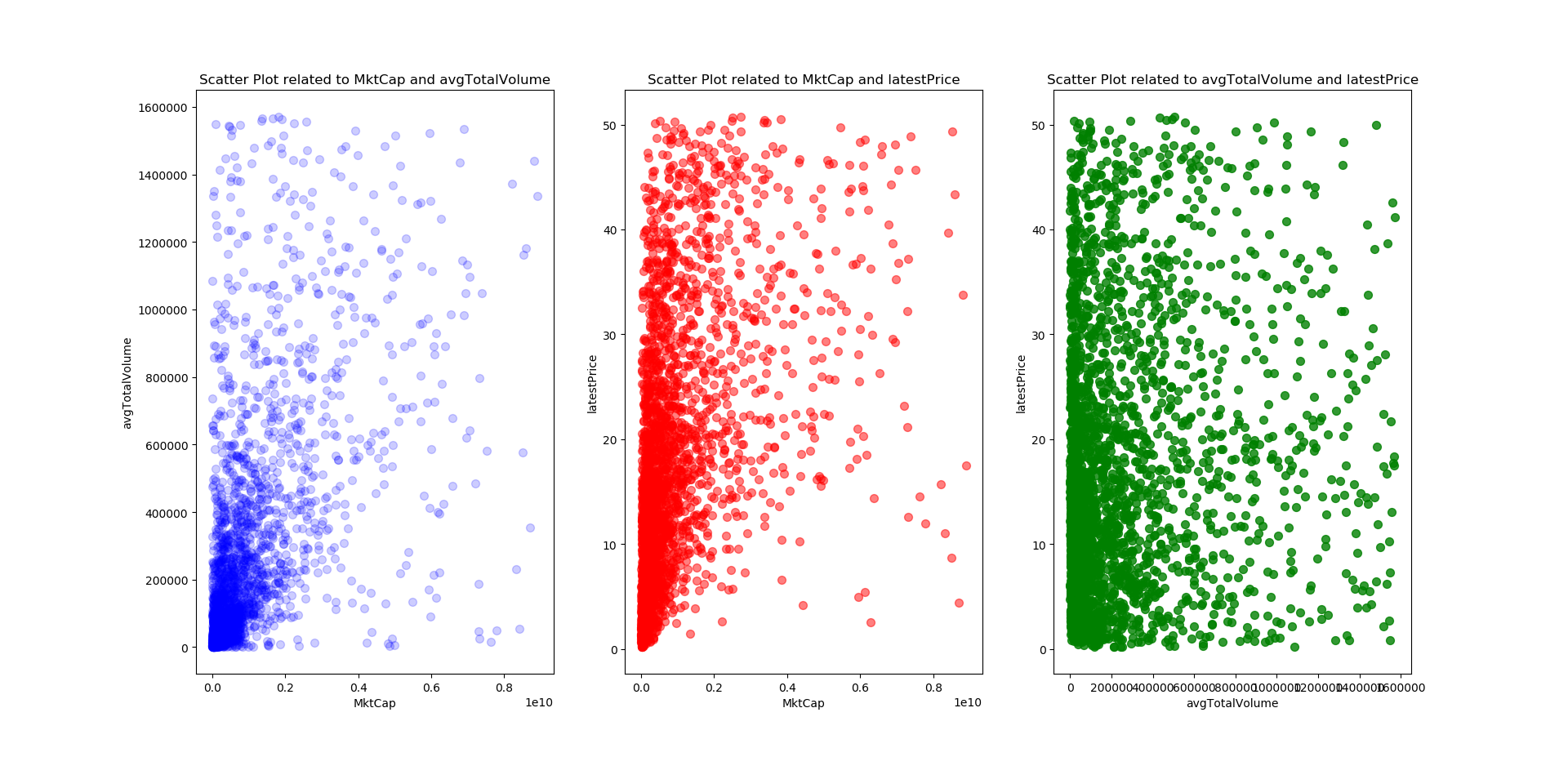
Third, for ‘Beta’ variable, it is also like a normal distribution which ranges from -1.0 to 2.5. Beta is a good indicator of risk level of a stock with 1.0 to be its normal risk level. From the graph, it is clear that most of its value are between 0 to 1.5, which means most stocks are of average risk level. Only a little is greater than 2.0, which represents for high risk. And 5% is less than 0, which goes the opposite direction of the average risk level of the whole market.

Last, for ‘MktCap’ variable, the distribution is similar to ‘avgTotalVolume’. The distribution is skew and most of it is less than 1010 with a little greater than 3\*1010. MktCap is an [indicator](https://en.wikipedia.org/wiki/Proxy_(statistics)) of public opinion of a company's [net worth](https://en.wikipedia.org/wiki/Net_worth). Therefore, it should be related to avgTotalVolume to some extent since higher MktCap may lead to higher avgTotalVolume.

***Correlations:***

In this process, we find correlation between three variables:avgTotalVolume, latest price in data1 and MktCap in data2. The result table is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MktCap | avgTotalVolume | latestPrice |
| MktCap | 1.000000 | 0.560204 | 0.471481 |
| avgTotalVolume | 0.560204 | 1.000000 | 0.151997 |
| latestPrice | 0.471481 | 0.151997 | 1.000000 |

Among these three variables, the correlation between MktCap and avgTotalVolume is the highest. As we have mentioned in the previous part, MktCap represents public opinion of a company's [net worth](https://en.wikipedia.org/wiki/Net_worth) and avgTotalVolume reveals liquidity of the stock in the market. Therefore, they may be related to each other to some extent since higher MktCap may lead to higher avgTotalVolume. Next, the correlation coefficient of latest price and MktCap is 0.47, which means they are a little dependent. It is not difficult to understand it since MktCap and price of stock is likely to affect each other in some cases though it is not always true. Last, for avgTotalVolume and latest price, there is almost no relation between them since high avgTotalVolume does not necessarily mean high price of the stock. In conclusion, the correlation coefficient of these three variables reflects the actual situation of the stock in the market.

***Part3 - Clustering Analysis***

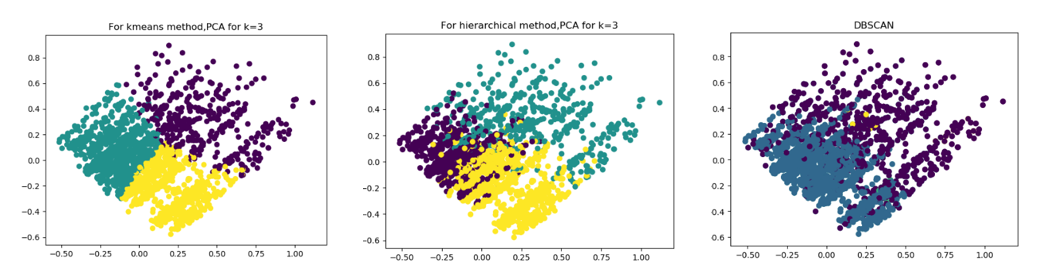
Clustering analysis is used to derive groups of a set of objects such that all the objects in one group share some similarities and the objects in one group is more similar to the objects in the same group rather than the objects in a different group. The goal is to find the optimal number of clusters of the variable that we are interested in.

In this step, we use three clusters analyses on my data: k-means, hierarchical clustering and dbscan. Then we use Calinski-Harabasz to assess the quality of the clusters. Finally, we convert the high dimensional data to two dimensions by using PCA method. A large value of Calinski-Harabasz is an indication of a good clustering. For the k-means, we try seven numbers for k to see which one is better for clustering. From the result we can see that as the number of clusters increases, the Calinski-Harabasz score decreases evenly. The score began to decay after number of clusters exceeds 2. When the cluster is 2, the average Calinski\_Harabaz\_score is 949.94, the highest score has been reached. However, grouping in two clusters might not be ideal to you, when that happens the local maximum measures. This is why we choose three clusters with the Calinski\_Harabaz\_score of 892.86, which indicates that those two clusters separate each other better than that of other number of clusters.

For hierarchical clustering, we draw the dendrogram of the last 20 steps to see we should divide into how many clusters, from the figure we can see that 3 clusters does well, so we divide it into three clusters. When clusters is 3, the average calinski\_harabaz\_score is 707.84. This quiet high score indicates that the 3 cluster separate each other better than that of other number of clusters.

For dbscan, we set a series of eps and min\_sample to see which parameter is the best. When eps = 0.14 and min\_samples = 6, the Calinski-Harabasz is 179.25, which is the highest score among the other eps and min\_samples. And it divides the data into four clusters. However, the Calinski\_Harabaz\_score in dbscan performs much lower than k-means and hierarchical clusters, this is because Calinski-Harabasz is generally higher for convex clusters than other concepts of clusters like density-based clusters.

The PCA method converts high dimensional data to 2 dimensions. From the figure, we can see that the clustering directly and clearly. K-means and hierarchical clusters performs better than dbscan. Although dbscan has four clusters, it showed two large clusters in the figure and only small points from different clusters in it, which indicated it does not perform well in dbscan.

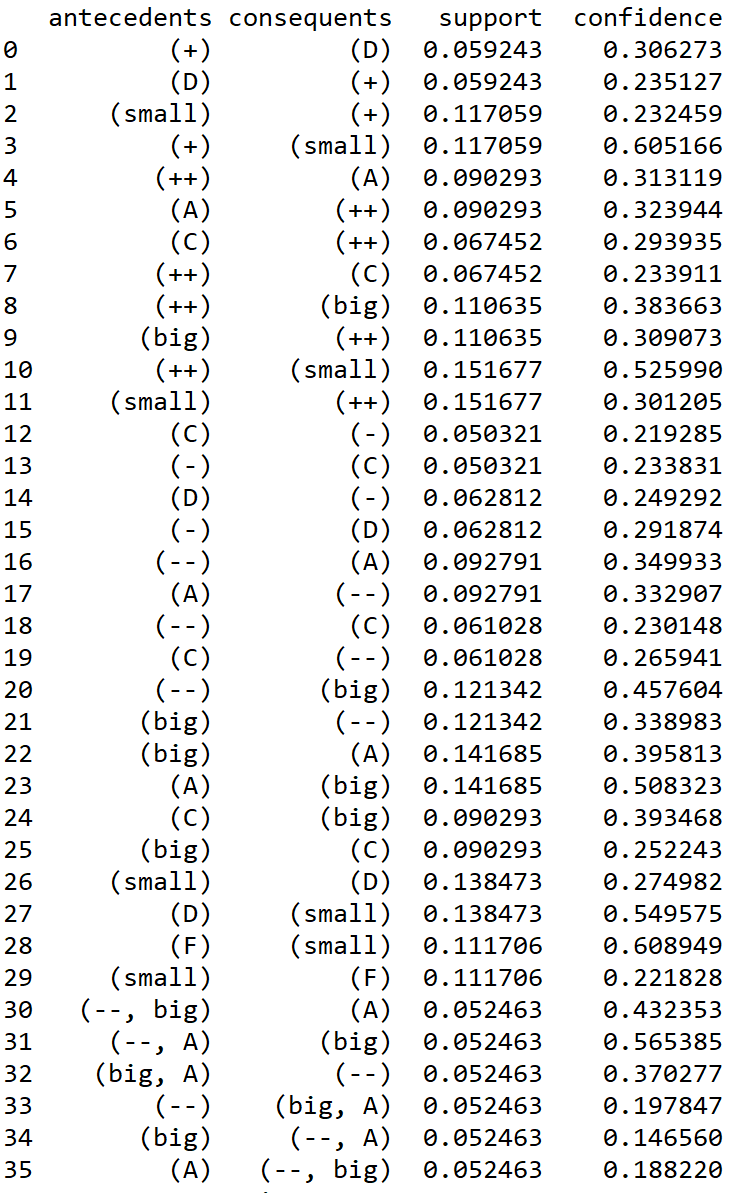
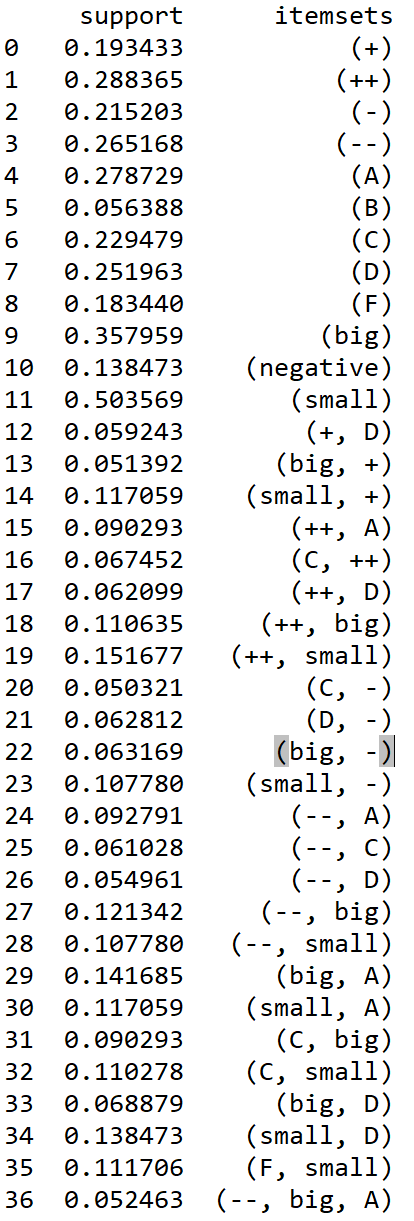


***Part4 - Association Rules/ Frequent Itemset Mining Analysis***

When people shop, after they buy item A, they may always buy item B(or more items) together, such as movie ticket and popcorn. This kind of associated relationship is an association rule for item A and B. To find association rules, we first need to find all frequent itemsets, which are itemsets appearing among all transactions with a probability that is greater than a threshold(we say minimum support here). For example, if we set minsup=0.3, item A appears 80 times among 100 transactions, its support is 80/100=0.8>0.3, so A can be considered as a frequent itemset. Another important concept is confidence which is equal to the ratio between the number of transactions containing both A and B(or more items) and the number of transactions containing A. For example, item A and B appear together 40 times conditioned on 80 transactions containing A, support for set (A,B) is 40/100=0.4 and confidence is 40/80=0.5.

Investigating association rules which can be used to predict the behaviors of objects is also useful in stock market. For example, if we know a stock having a big risk level, what we can say about its percentage change of price over last year? To answer question like that, we use three variables that were created in binning step: Risk in data2, Change\_level and Volume\_level in data1. We then converted this subset data into transaction data form, which is a traditional form for association rules analysis. There are 13 items for each transaction that a stock can be: ++, +, |, -, --, A, B, C, D, F, big, small, negative.

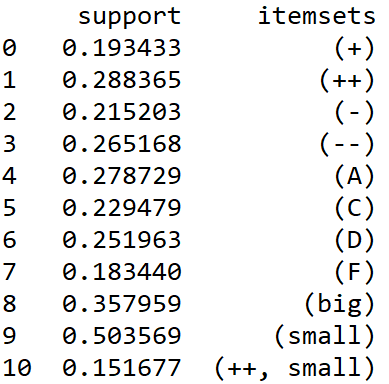
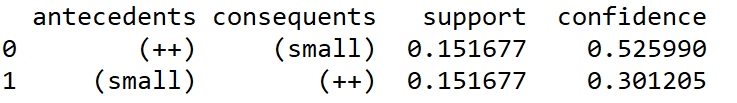
There are lots of algorithm for finding association rules between item sets, the one we employed is Apriori. Given minimum support level, Apriori algorithm will first find all single item (we say length=1) that has a support level greater than the threshold, then it will filter all length=2 items that has a support level greater than the threshold based on last step frequent single items and so on. Noticed that when we set minimum support greater than or equal to 0.20, there is no association rules can be found among the frequent itemsets, so we set minimum support levels below 0.20. The four different minimum support levels are: 0.05, 0.10, 0.15, 0.20(there is no such association rule). For example, when minisup=0.05, the frequent itemsets is shown in the left and association rules is shown in the right.



When minisup=0.15, the frequent itemset is shown in the below left graph and association rules is shown in the right. Noticed that when we increase the minisup, the itemset that satisfies this minisup is less. The most frequent pattern is clear when we set minisup=0.15 because only two rules left:

1. small→ ++ with support = 0.151677 and confidence = 0.525990
2. ++ → small with support = 0.151677 and confidence = 0.301205

This pattern indicates that among all stocks, stock that both have a small risk level and very big percentage change of price over last year counts for 15.1677%. When a stock with small risk level, there is 30.1205% probability that the price stock is also increase dramatically. On the other hand, if the price of stock increases dramatically over the last year, there is more than 50% probability that this stock has a small risk level. This pattern is not surprising, since stock with an increasing price often indicates that the company grows healthy in a long term, which with a big probability that this company has a low-risk stock.



***Predictive Analysis***

***Part5 - Hypothesis Testing***

The final part consists of two steps: traditional statistical hypothesis testing and machine learning predictive analysis. Traditional statistical hypothesis assumes the generative process of data and often assumes the data have a normal distribution. We choose ‘Beta’ and ‘ytdChange’ for t-test and ANOVA test because their distributions look more normal than other attributes based on histogram and correlation part. Based on these presumptions, we can operate hypothesis testing like whether the mean of a variable is same among different sectors/industries, or to see whether two or more variables are independent. Machine learning can be used to build predictive model and then use the model to predict the classification of a new observation. For example, if we know the lastestPrice, ytdChange, Volume\_level, we can predict the risk level of that stock. In the second part, we can also investigate hypothesis testing to test whether two machine learning method is equally efficient regarding to predict classification.

***Traditional Statistical Hypothesis Testing***

**Hypothesis 1:**

Null hypothesis: the mean of beta for Oil & Gas - Integrated is same as the mean of beta for Oil & Gas - Midstream.

Alternative hypothesis: the mean of beta for Oil & Gas - Integrated is not same as the mean of beta for Oil & Gas - Midstream.

We use t-test and t-test result is:



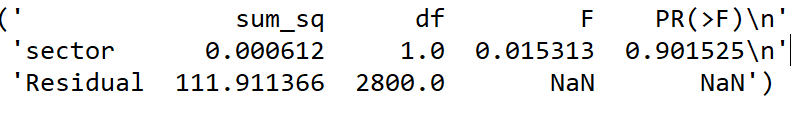
Assume confidence level is 5%, we have p-value =0.97057375586539107, which is larger than 0.05, we fail to reject null hypothesis. Therefore, the mean of beta for Oil & Gas - Integrated is same as the mean of beta for Oil & Gas - Midstream.

**Hypothesis 2:**

Null hypothesis: the mean of ytdChange for 11 sectors are same.

Alternative hypothesis: the mean of ytdChange for at least one sector of 11 sectors are not same as others.

We use ANOVA test and the result:



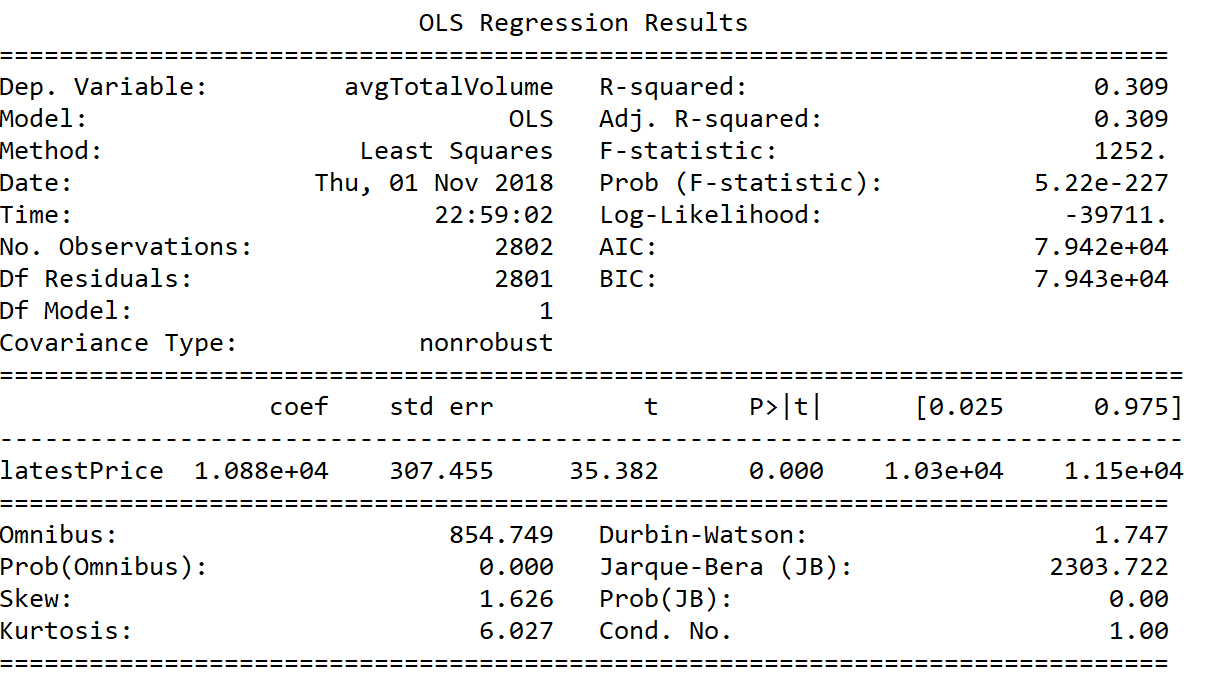
Assume confidence level is 5%, we have p-value = 0.901525, which is larger than 0.05, we fail to reject null hypothesis. Therefore, the mean of ytdChange for 11 sectors are same.

**Hypothesis 3:**

Null hypothesis: latestPrice and avgTotalVolume are independent.

Alternative hypothesis: latestPrice and avgTotalVolume are not independent.

We use linear regression and the result:



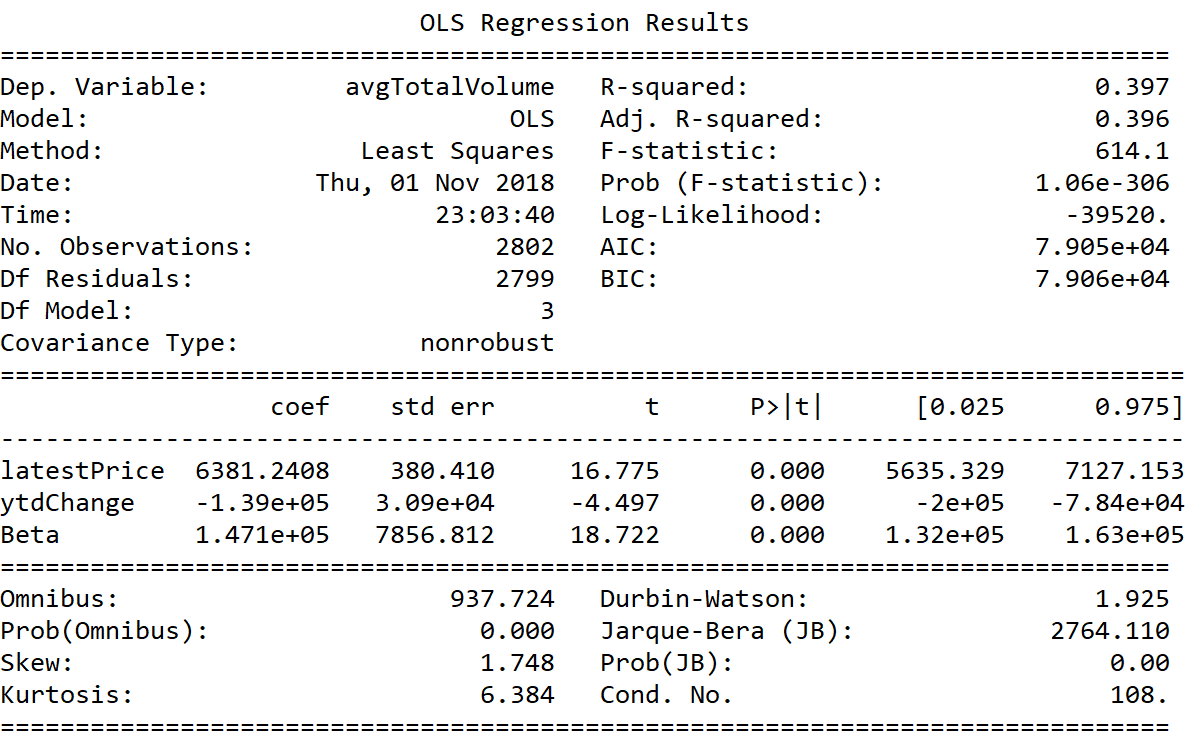
Assume confidence level is 5%, from linear regression, we have p-value =5.22e-227, which is smaller than 0.05, we reject null hypothesis. Therefore, latestPrice and avgTotalVolume are not independent. From OLS Regression results, we have t =35.382, which has p-value =0.000, smaller than 0.05, we reject null hypothesis. Therefore, latestPrice and avgTotalVolume are not independent.

**Hypothesis 4:**

Null hypothesis: avgTotalVolume, latestPrice, ytdChange and Beta are independent.

Alternative hypothesis: at least one variable of latestPrice, ytdChange, Beta is not independent from avgTotalVolume.

We use linear regression and the result:



Assume confidence level is 5%, from linear regression, we have p-value =1.06e-306, which is smaller than 0.05, we reject null hypothesis. Therefore, at least one variable of latestPrice, ytdChange, Beta is not independent from avgTotalVolume.

***Machine Learning Predictive Analytics***

The first hypothesis is that the classification result of decision trees and random forest is the same. Therefore, we set the null hypothesis is that the classification result of decision trees and random forest is the same. The alternative hypothesis is that the classification result of random forest is better than the result of random forest.

Decision Tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features and to divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label. It can handle both categorical and numeric data.

Random Forest is an ensemble learning method for classification. We make a single decision tree by randomly selecting subsets of the training data. Then, we repeat this process for many times to make a forest of a decision trees. To classify a new record, each tree offers a classification and the result is based on the majority classification based on all the trees.

In our hypothesis, we select 5 attributes to do the predictive analytics. More specifically, we use 'sector', 'avgTotalVolume', 'latestPrice', 'industry', 'MktCap', to predict 'Risk', which has three labels: big, small and negative. Since we should convert all attribute into numeric, so we make a numeric label for ‘industry’. Then, we randomly select 20% of the data as the test data for further validation. For the training data, we use 10-fold cross validation to validate our classification.

Here is the result:

(a) For training data:

The cross\_val\_score for decision tree is: 0.478339 (std: 0.028573)

The cross\_val\_score for random forest is: 0.535034 (std: 0.039406)

(b) For test data:

For decision trees:

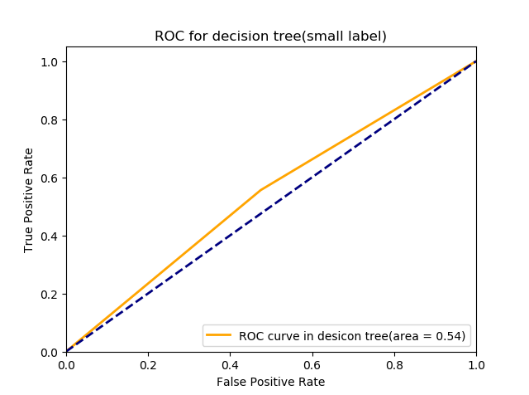
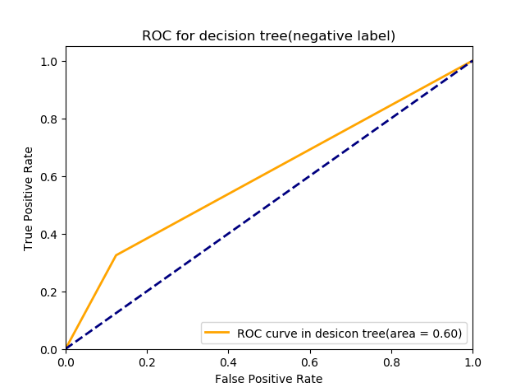
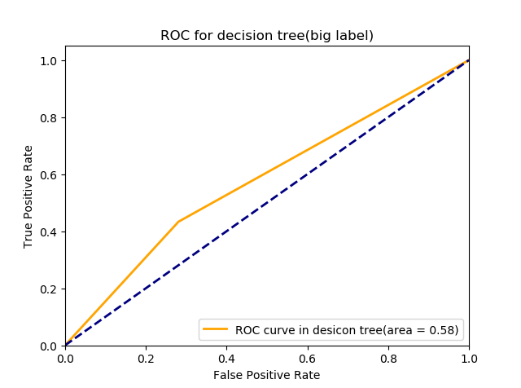
1.the accuracy: 0.48841354723707664

2.the confusion matrix is:

|  |  |  |  |
| --- | --- | --- | --- |
|  | big | negative | small |
| big | 83 | 19 | 78 |
| negative | 16 | 23 | 38 |
| small | 93 | 43 | 168 |

3.classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| big | 0.43 | 0.46 | 0.45 | 180 |
| negative | 0.27 | 0.30 | 0.28 | 77 |
| small | 0.59 | 0.55 | 0.57 | 304 |
| avg / total | 0.50 | 0.49 | 0.49 | 561 |

Roc Curve:

For random forest:

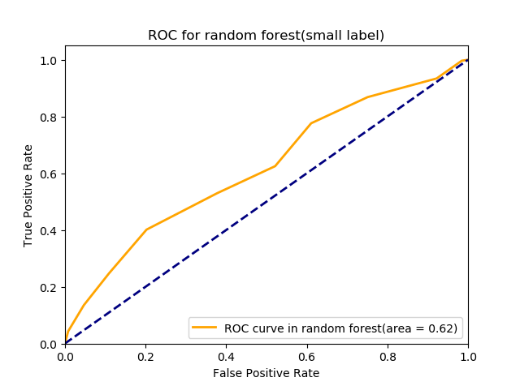
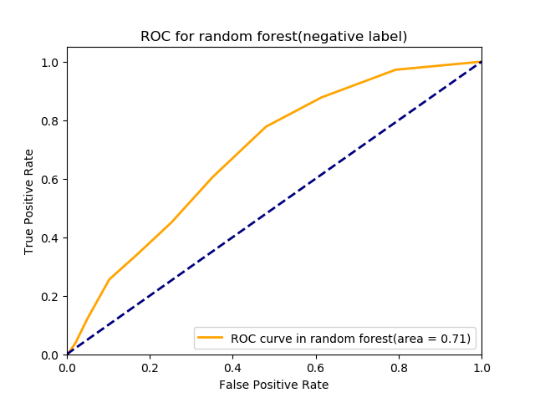
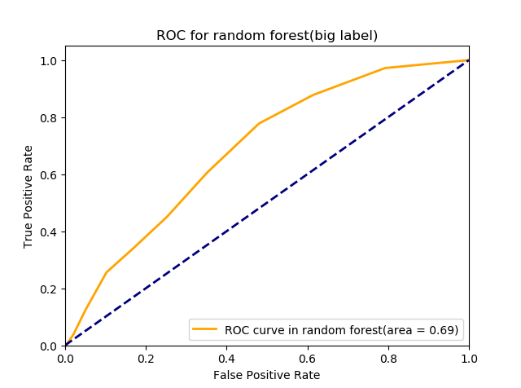
1.the accuracy for random forest: 0.5329768270944741

2.the confusion matrix for random forest is:

|  |  |  |  |
| --- | --- | --- | --- |
|  | big | negative | small |
| big | 99 | 9 | 72 |
| negative | 18 | 25 | 34 |
| small | 99 | 30 | 175 |

4.classification report for random forest:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| big | 0.46 | 0.55 | 0.50 | 180 |
| negative | 0.39 | 0.32 | 0.35 | 77 |
| small | 0.62 | 0.58 | 0.60 | 304 |
| avg / total | 0.54 | 0.53 | 0.53 | 561 |



On the training data, the mean accuracy rate of decision trees is 0.478, less than that of random forest which is 0.535. Meanwhile, the standard deviation of decision trees is 0.028573, lower than that of random forests which is 0.039406. On the test data, the accuracy of decision trees is 0.488, lower than that of random forest which is 0.533. Therefore, it is clear that the accuracy of random forest is higher than the decision tree for both training data and test data. Hence, we reject the null hypothesis that the classification result is the same for these two methods.

***Classification***

K-nearest Neighbor Classifiers: a non-parametric method used for classification and regression. This method identify k nearest neighbors by computing distance to other training records and used class labels nearest neighbors to determine the class label of unknown record. This method does not build models explicitly and unlike eager learners such as decision tree induction and rule-based systems. Using this method to classify unknown records are relatively expensive.

Naïve Bayes is a simple technique for constructing classifiers, which is based on common principle that all value of a particular feature is independent of the value of any other feature.

Assuming independence among attributes when class is given:

P(|

Compute the posterior probability

P(C|=

Choose value of C that maximizes P(C|

Equivalent to choosing value of C that maximizes

P(C|

This method is isolated noise points and robust to irrelevant attributes. It can handle missing values by ignoring the instance during probability estimate calculations.

SVM is a supervised machine learning with associated learning algorithms that analyze data used for classification and regression analysis. It uses known and labeled data to “train”. It uses different “kernels” options such as linear, gamma, sigmoid and Gaussian. SVM can perform feature transformation into higher (or infinite Hilbert Space) dimensional space so that input vectors are separable by hyperplane. Most “important” training points are support vectors. This becomes a quadratic programming problem that is easy to solve using standard methods.

In our hypothesis, we use 5 attributes to do the predictive analytics. More specifically, we use 'sector', 'avgTotalVolume', 'latestPrice', 'industry', 'MktCap', to predict 'Risk'.

In this classification, we need to guarantee all the attributes are numbers. Therefore, the first step we need to change the 'industry' column into numeric. Then we treat the 'Volumn\_level' as the class and divided into 2 class label. A subset of features are 'sector','lastprice','ytdchange', 'Beta', 'MktCap', 'industry'. After that, we transform the numeric into Min-Max Normalization in order to have a same scale.

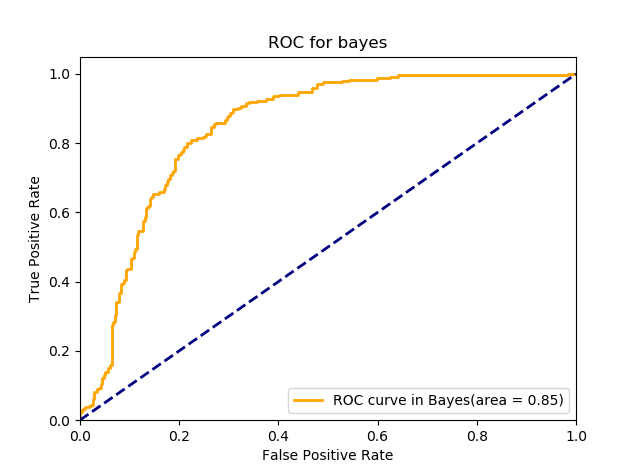
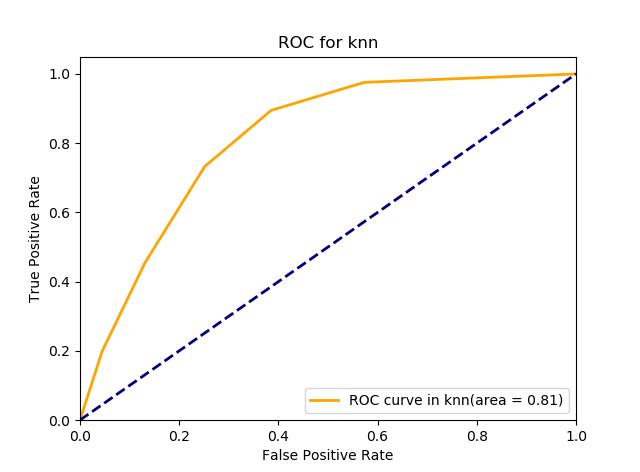
From the result, we can see that knn has the higher accuracy ranking than a decision tree does. On the training data, the mean accuracy rate of K-Neighbors is 0.720647, higher than that of Naïve Bayes which is 0.670206 and SVM which is 0.565819. Meanwhile, the standard division of K-Neighbors is 0.026021, lower than that of Naïve Bayes which is 0.034487 and SVM which is 0.046344.On the test data, the accurate\_score of knn is 0.74153, higher than that of Naïve Bayes which is 0.661319 and SVM which is 0.559714.

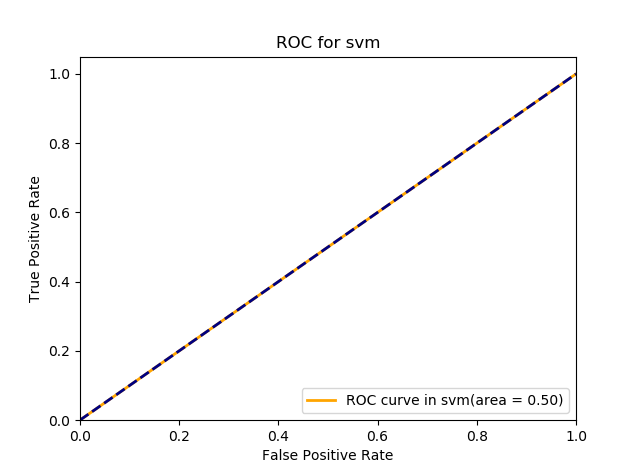
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| knn | precision | recall | f1-score | support |
| High | 0.78 | 0.75 | 0.76 | 314 |
| Low | 0.70 | 0.73 | 0.71 | 247 |
| avg / total | 0.74 | 0.74 | 0.74 | 561 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bayes | precision | recall | f1-score | support |
| High | 0.97 | 0.41 | 0.57 | 314 |
| Low | 0.57 | 0.98 | 0.72 | 247 |
| avg / total | 0.79 | 0.66 | 0.64 | 561 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| svm | precision | recall | f1-score | support |
| High | 0.56 | 1 | 0.72 | 314 |
| Low | 0.00 | 0.00 | 0.00 | 247 |
| avg / total | 0.31 | 0.56 | 0.40 | 561 |

From the graph, we can see that one of the class in svm is zero, which means that the svm classification performs bad. The Naïve Bayes, for each class, the precision and recall performs differently, so the f1-score did not high. The knn performs best and the f1-score is quite high.





The ROC in knn is 0.81, which is similar to bayes 0.85. They performs better than ROC curve in svm which is 0.5.