Final Project – Stock Prediction

Role	Participation	Name
Team Member	1.0	Daehoon Gwak (dps5313@psu.edu)
Team Member	1.0	Jaehoon Ha (jxh5648@psu.edu)
Team Member	1.0	Daniel Jung (dqj5182@psu.edu)

Table of Contents

1. Introduction

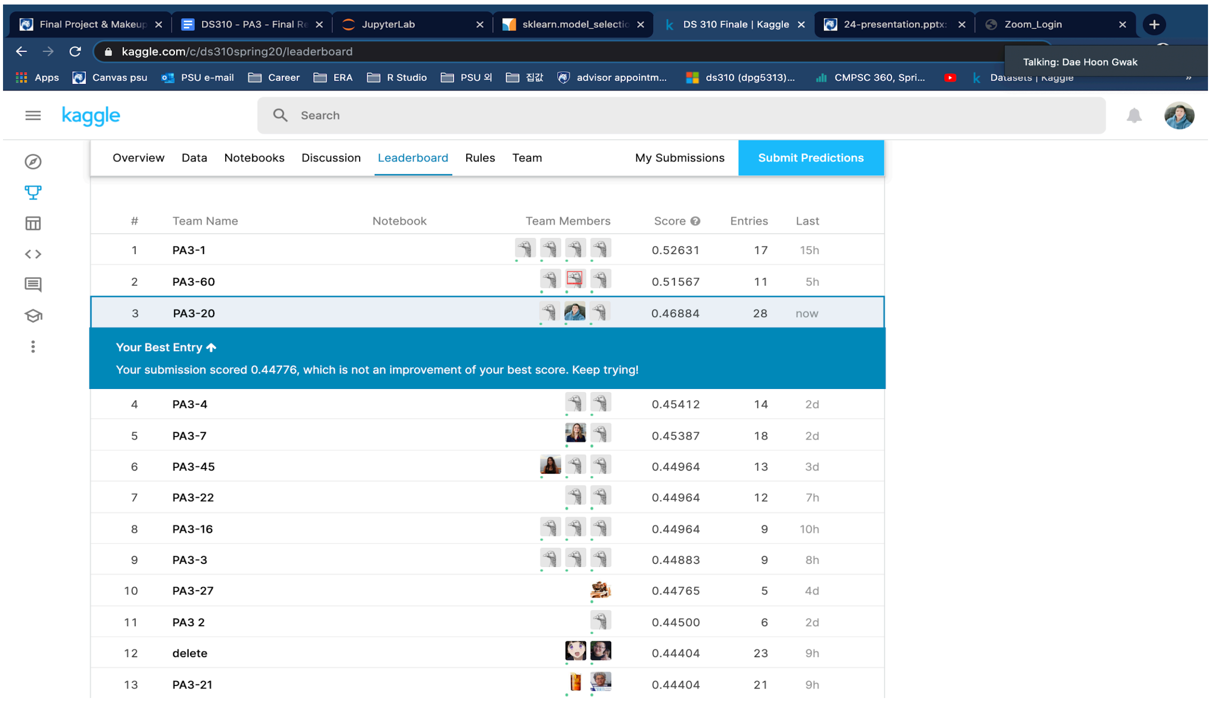
The study of stock market movement, whether it is a small company like small start-ups or a mega corporation like Coca-cola, has been conducted by many of whom seeked fortune through stock market. It is said that stock market is impossible to predict. Others argue that it can be predicted, but not in certain, with precise method of prediction based on certain features of a company. We can see from successful story of Warren buffett or Peter Lynch that this argument has some legitimate track record.

Despite this, not many people have been successful on predicting stock market and make fortune out of it. According to ChosunBiz, average retail (individual, non-institutional) investors lost 74% throughout past 10 years. This suggest that majority of people lost money through stock market. How did this happen? Considering that famous investors such as Warren buffett and Peter Lynch have been successful at making profit in the stock market more than 20% annually with just pure company's fundamental data, we can say that people were not successful at analyzing stocks or companies on the vast features provided by financial reports.

This project is proposed to solve such problem. In our dataset, there are more than 200 columns about each companies (stocks). With more than 200 columns provided, our goal is to predict whether a company's stock has been trending up or down in a given year. The models of each team will be evaluated based on f1-score that the predicted values of testing data will output.

2. Team's final rank

As of 3:17 AM on Apr 26, 2020, our group (PA3-20) is in 3rd as of this project with score of 0.46884 and total entries of 27.



#	Team Name	Notebook	Team Members	Score	Entries	Last
1	PA3-1			0.52631	17	15h
2	PA3-60			0.51567	11	5h
3	PA3-20			0.46884	28	now

Your Best Entry ↑

Your submission scored 0.44776, which is not an improvement of your best score. Keep trying!

#	Team Name	Notebook	Team Members	Score	Entries	Last
4	PA3-4			0.45412	14	2d
5	PA3-7			0.45387	18	2d
6	PA3-45			0.44964	13	3d
7	PA3-22			0.44964	12	7h
8	PA3-16			0.44964	9	10h
9	PA3-3			0.44883	9	8h
10	PA3-27			0.44765	5	4d
11	PA3 2			0.44500	6	2d
12	delete			0.44404	23	9h
13	PA3-21			0.44404	21	9h

3. Team`s solution

1) Data pre-processing

The two data given were “train_data.csv” and “test_data.csv”. The “train_data.csv” is a csv file containing testing data with 224 columns. These columns were company’s fundamental data features such as company’s name, revenue, profit, expenses. On the last column, it also included column named “Class”, which gave data about whether the stock price has gone up or gone down in that year. For each stock, the value 1 of “Class” column represented that the stock has gone up and the value 0 of “Class” column represented that the stock as gone down.

a. Balance the number of each classes

For processing the data, our group have conducted three methods. Firstly, we decided to balance the number of rows with “Class” value of 0 and the number of rows with “Class” value of 1. The below code gathered rows that has “Class” value of 0 in “df_train0” dataframe and rows that has “Class” value of 1 in “df_train1” dataframe. We could see from the shape function in pandas library that there were more rows with “Class” value of 0 than rows with “Class” value of 1. Therefore, we sampled from the “df_train0” dataframe so that we can make the number of rows of “df_train0” same as “df_train1”. Then, we concatenated “df_train0” dataframe and “df_train1” data frame into “df_train” data frame with pd.concat function.

b. Factorize “Sector” column

The second method is to factorize sector columns into integers. We have decided to include “Sector” column in our data set because we could see from our correlation analysis that it was the single most important feature in our data set when it comes to predicting stock price. However, most of the clustering algorithms do not work on categorical string values. Therefore, we decided to factor them into integers and make dictionary on what factorized values mean which sectors. As we have factorized “Sector” column of “df_train” dataframe, we also need to factorize “df_test” which is the testing data for this project. With the “sector_dict” dictionary that we have previously made, we have replaced “Sector” column values into factorized sector values as specified in the dictionary.

These were our group’s two main data pre-processing methods taken place. All pre-processing after these were minor adjustments that facilitated clustering that will happen in further code. For the minor adjustment, we have deleted company name column, “Class” column, and “Sector” column (non-factorized) from training data and deleted company name column from testing data.

c. Data Imputation

The last main data pre-processing method is the imputation. As there were tons of missing values in both training dataset and testing dataset, we needed certain imputation method to be taken place. We first started with simple imputation by calculating mean values for each column. However, we thought there must be a better way of imputation and we used ‘KNNImputer’ since we would use KNN clustering for the model. KNNImputer provides K-Nearest Neighbor approach and it imputes the mean value from the nearest neighbor in the training data.

2) Feature engineering

Our group has decided to use all of features given for training and testing data. We have tried from removing features with low variance to removing features with low correlation with “Class” variable and made different models with such selected features. However, in all of the cases, we could see that these feature selection led to lower f1-score. Therefore, we concluded to retain all features given. To be more specific, according to evaluation matrix analysis conducted for both K-means classifier based on original data and K-means classifier based on feature reduced data, we can see that there is no notable improvement with both of the models outputting same Rand index with value 0.5148, F1-score of 0.51487, and Purity of 0.5896. Therefore, we can say that reducing feature does not improve data modeling for this data set.

3) Model building

a. Comparison

Our group have tried various clustering method such as K-means clustering, Hierarchical clustering, density based clustering, and also linear regression model. We could see from values such as Rand Index, F1-score, and precision that K-means clustering was the best clustering method for predicting stock price movement.

b. Hyperparameter setting and tuning

We first used default value of the K Means Clustering function, where n_neighbors is 5, weight equals uniform, and so on. However, the accuracy base on the default setting showed around 72% which is fair, but not better. So, we imported 'GridsearchCV' to implement grid search . This helps to find the optimal classifier with cross-validation.

We controlled two parameters; n_neighbors and weights, since they are the most important factor for turning. We set the range 1 to 30 for n_neighbors, and weights equal to either 'uniform' or 'distance'. Finally, we found that the optimal values are: n_neighbors = 8, weights = 'distance'. On top of that, we imported 'startifiedShufflesplit' to control cross-validation value.

c. Performance evaluation

After tuning, we fit the model and found the accuracy equals to 98.18, which was higher than the default settings.

4. Conclusion

In class, we learned many different ways of analyzing dataset using clustering methods. Therefore, as mentioned before, our group tried various clustering techniques, including K-means, Hierarchical, Density-based clustering, and so on. Our group met at least twice a week via zoom in order to find the best clustering method for this project and ended up with using K-means clustering. From this analysis, our group learned two basic lessons. First, we learned the features that we choose are weighty matters in data analyzation. It was highly important task to choose appropriate features for the analysis. Secondly, we learned that each data clustering method has clear advantages and disadvantages. Even if we choose the best features for analysis, if we do not use suitable data clustering method, we will get unsatisfiable outcome. Therefore, from this project, our group learned not only the features affect the outcome but also the method of clustering.

Reference

Lee, J., Kim, J. (2017, Mar 6). 10 year stock investing earnings… Foreigner 78%, Retail investor -74%.

Retrieved from https://biz.chosun.com/site/data/html_dir/2017/03/06/2017030602584.html

Scikit Learn. (2019). Imputation of missing values.

Retrieved from https://scikit-learn.org/stable/modules/impute.html

Scikit Learn (2019). sklearn.model_selection.GridSearchCV.

Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html