Company Classification by Industry With KNN

Group Members:
Kirill Sukhikh - 1005371526 (kirill.sukhikh@mail.utoronto.ca)
Yinuo Zhang - 1003018940 (yinuozh.zhang@mail.utoronto.ca)
Justin Zhao - 1004745975 (justin.zhao@mail.utoronto.ca)

Introduction

Different equity investors tend to process information on which they are basing their projections of different companies differ. Some may be looking at the potential income a company is expected to generate over a certain period of time. Others look at the assets and liabilities of the companies to estimate the value. Another method of estimating a price of a stock and whether it is overvalued or undervalued is looking at the financial data, processing it and comparing it to other similar companies. The latter approach is called the Multiples Approach because it uses certain multiples such as EV/EBITDA, Price/Earnings and Debt/Equity ratios, and compares them with benchmarks of a specific industry. This approach requires a lot of judgement from an analyst because they have to pick the right benchmarks to estimate the value of a company. This creates a room for error because many large public companies have multiple streams of income, and it may be unclear what industry to use as a benchmark. For example, Tesla Inc. (TSLA) has seen a lot of growth in the past couple of years that is unprecedented to the history of vehicle manufacturing. Because of such growth and their focus on innovation in the automobile industry, many people argue that the benchmarks for Tesla should be in line with tech companies rather than with automakers. For comparison, median automobile industry EV/EBITDA is x8.51 while median technology sector EV/EBITDA is x21.1. EV/EBITDA compares the value of a company to what this company is earning. The downside of this measure is the fact that it does not account for growth since EBITDA used is given at a certain point in time. Because of that, some people believe that the higher the industry ratio, the faster it is expected to grow. In our case, the technology sector is expected to grow more than the automobile sector because the technology sector has a higher EV/EBITDA ratio.

Each industry has its unique conditions for growth that are reflected in their financial information. For example, the tech sector can justify such a high EV/EBITDA ratio due to large economies of scale that technology allows to achieve.

We will try to see whether it is possible to identify the industry a company is operating in using financial information available for the companies that are included in the S&P 500 index. We will be using supervised and unsupervised learning algorithms to see whether financial information is enough to predict the industry of a given company and combine the companies into new groups. We will be using the K-Nearest Neighbors algorithm to see whether it is possible to predict an industry a company is operating in looking at the financial information available.

There are multiple questions that we will try to answer conducting our analysis. The first question is: What financial variables could predict the industry a company operating in? We believe that there are some variables such as various ratios that combine market sentiment (stock price) and value, including earnings, of the company may be useful to predict the industry a company is operating in. What EV/EBITDA does each industry sector have and how KNN can be used to find overvalued and undervalued stocks? Using financial information available, we will try to see whether it can predict whether the stock is overvalued or undervalued for each stock given the benchmark for each group found. What financial variables can best be used to predict future growth rate? It is difficult to predict the exact future growth rate, but it is possible to compare companies that are supposed to have a similar growth rate, so we will try to find similarities within the groups and see what financial variable they are most sensitive to.

There are three papers that use similar methods as we do. The first paper by Sen et al, “*Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models*” (2018), presents a comprehensive account of applying eight different machine learning techniques, which includes both regression and classification methods, on predicting the price movements of stocks. This paper is noteworthy for two reasons: first it compares a large number of different methods which can highlight the hierarchy of accuracy between different methods. Secondly, the information they collect is on a 5 minute interval, and uses predictors that are mostly used in the technical analysis part of stock analysis. This means the results are more relevant to trading oriented applications, such as algorithmic trading, rather than long term stock forecasts. Some limitations of this paper are that the features are included in an unweighted manner, and also their data set consists of two stocks only.

The second paper we reviewed, by Cao, Lin, Li, and Zhang, *"Stock Price Pattern Prediction Based on Complex Network and Machine Learning"* (2020)*,* focuses on the SVM and KNN classifiers. They use data from the three major US stock indexes and examine daily price movement instead. One important thing to highlight from this study is that the authors extensively highlight the process of choosing the best K for the KNN classifier. They employ the traditional K-fold validation method for larger datasets, and describe how an alternative process called Leave One Out, can be useful when the datasets are smaller.

The last paper by Chen and Hao, *“A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction”* (2017), takes a more sophisticated approach on the KNN classification of stock prices. The data they used come from two main stock indices in China. The important contribution from their paper is the application of feature weighting. They did not simply include each feature equally weighted into their models, but rather calculated the optimal weights based on the metric of information gain.

All three papers present meaningful insight and methodology that could contribute to our proposed study.

Data And Methodology

We used the data provided by datahub.io (Data Link) on the financial metrics of each S&P 500 company. These metrics included the following:

- Stock Price
- Price/Earnings Ratio
- Dividend Yield
- Earnings-Per-Share
- Market Cap
- EBITDA
- Price/Sales
- Price/Book

Additionally, we have a sector for each company in the list, which will be our target variable.

We also construct one additional variable, **volatility,** as the difference between the 52 week high and 52 week low, as a crude measure of the range of movement the price of the stock exhibits.

We first started with the KNN model. We split out dataset into train and test datasets, and after that, we found the optimal number of clusters using an Elbow Method for the first 30 clusters. It did not make sense to look for more as each cluster past 30 would not be as meaningful.

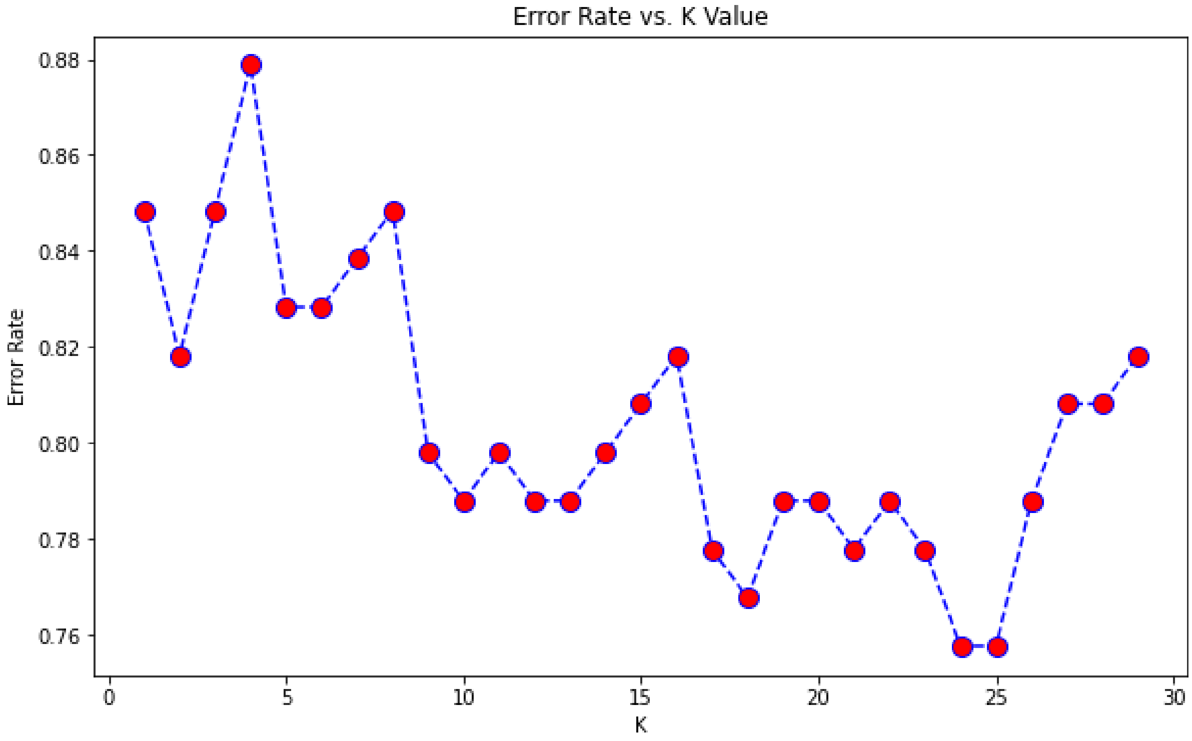

Figure 1. Error Rate for Different K-Values

We can see in Figure 1 that the optimal K-Value is 10 since after 10 the error rate does not decrease as much. The Error Rate refers to the proportion of errors to the total number of observations. Even though the most optimal K-Value is 10, the error rate is quite high, which may be due to limited data available and omitted variables for training the model.

These algorithms should provide insights on how financial data correlates between the firms of similar sectors. Using KNN, we can see how this algorithm can predict the sector a company is operating in only using financial data available. If our prediction error is less than the prediction error of random guessing, then we may conclude that the data has some power in predicting the sector. On the other hand, if our prediction error is equal to or higher than the prediction error of random guessing, then financial variables may not be useful in predicting the industry a company operates in.

Variable Selection

In terms of selecting the variables to include in the KNN classification, we run 5 different specifications:

Model 1: Using all variables

Model 2: Using the price ratios: Price/Earnings, Price/Sales, Price/Book

Model 3: Using stock related variables: volatility and dividend yield

Model 4: Using size variables: EBITDA and Market Cap

Model 5: Using all variables from Model 2 to 4

The logic behind this is we want to explore which kind of variables would exhibit the strongest with-industry similarities. The model with the highest predictive power would represent the type of variables that is most predictive of the industry.

It is worth mentioning that the model with the most variables may not be the most predictive - irrelevant variables that do not depend on industry could create spurious proximities between companies from different industries, causing error.

Monte-Carlo Simulation

Since the process of dividing testing and training data involves statistical uncertainty, we run each specification in repetitions of 100 and obtain their average predictive score in order to obtain a less noisy conclusion. For each of the 5 models, we run 100 repetitions of n=1,2,3,4,5,6,7,8,9,10, resulting in a total of 5000 trials, 1000 per each model.

Results

Model	**Average Score**
1 - All variables	0.200161
2 - Price Ratios	**0.28145**
3 - Stock Related	0.185376
4 - Size Variables	0.205094
5 - Model 2 through 4	0.203235

Based on 1000 trials, the 5 KNN models achieved average predictive scores of 0.20, 0.28, 0.19, 0.21, 0.20 respectively. Since we have 11 different industries included in the dataset, uninformed random-guessing would achieve on average an accuracy of 9.2%.

Even though the results obtained do not provide conclusively strong predictions, it does improve significantly from random guessing. This suggests that the variables chosen do have some power in predicting the industry the company is in. The strongest predictors come from the price ratios, which are often used as valuation multiples. Notably, adding other variables on top of the price ratios actually decreases the predictive power, as seen by the lower score in model 5 and model 1.

This suggests that the variation in factors such as size, stock price movement, dividend yield etc. are quite large for firms within each industry, that including these variables in the algorithm in fact decreases the accuracy of predictions.

Conclusion

Even though it seems that the financial variables do not play a significant role in predicting a sector a company operates in, it still provides a slight decrease in prediction error compared to random guessing. It might be possible to decrease prediction error further by adding new observations and refining the variables used in classification. It might be also possible to enhance our model by adding time series on each variable for each model since different companies may be at a different stage of maturity. It is likely that the financial variables have limitations in predicting the industry a company operates in. Market sentiment on various sectors that may be captured using Natural Language Processing, may also enhance the prediction accuracy greatly.

References

- Mehtab, Sidra, Jaydip Sen, and Abhishek Dutta. "*Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models.*" Communications in Computer and Information Science, n.d., 88–106. doi:10.1007/978-981-16-0419-5_8.
- Hongduo Cao, Tiantian Lin, Ying Li, Hanyu Zhang, *"Stock Price Pattern Prediction Based on Complex Network and Machine Learning"*, Complexity, vol. 2019, Article ID 4132485, 12 pages, 2019. https://doi.org/10.1155/2019/4132485

- Yingjun Chen, Yongtao Hao, *“A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction”, Expert Systems with Applications: An International JournalVolume 80Issue CSeptember 2017 pp 340–355 https://doi.org/10.1016/j.eswa.2017.02.044*