ECM3420 Learning From Data Report

Andrew Yau

December 2022

1 Introduction

The Fortune 500[1] represents the top 500 companies based in the United States of America in terms of revenue, accrued and published by Fortune magazine. Due to it displaying the top cohort of the country, it clearly indicates the sectors and industries which are the most popular and generative in revenue, and subsequently where most of the counties' funds circulate. In addition to the understanding gained about the obvious top 10 and overall net worth of the companies.

1.1 Possibilities and Research Details

Aside from the transparent discernments, the data has multiple facets of knowledge that can be derived, such as the geographical flow of funds where the value of a company can be modelled to decrease at an exponential or linear rate centered from the companies' headquarters. By overlapping this data between multiple companies graphically, is should be possible to identify locations or cities which should theoretically have a more abundant economy [2]. Therefore, if an average value within a certain distance is taken throughout the country, the cities covered by fortune 500 countries are more likely to be above average and external companies in the area can also be predicted to have a higher revenue than the national average. The overall flow of funds can be assessed by analysing this method over several years.

This report will be examining the stability and risk factors of the different sectors in the dataset, this will give and indication of what sectors are the most reliable to invest in. This is possible due to the fact that all the sectors have very different natures and business models compared to each other. Therefore, there are trends in relationships with attributes that point certain sectors, despite the range of companies and outliers within a sector. An attribute which can be derived from the dataset, debt-to-equity (D/E) ratio [3], is a major indication of risk factor, as a high ratio generally indicates how much a business' operations are maintained through debt and a negative value indicates an extremely high risk as a business' liabilities outweigh it's assets. However, the high ratio risk factor only applies to the vast majority of sectors and a small number of sectors like the Financials sector have a naturally high D/E ratio. Considering these factors, my research question will be defined as:

Are the relationships between sectors and their overall linked attributes robust and unique enough to enable the classification of corporation data to their sectors, and are those strong attributes risk factors for investment?

1.2 Main analysis objectives

- Use the Multilayer Perceptron (MLP) classification model to generate a model which can classify unlabelled data to their corresponding sectors with a sufficient degree of accuracy. This can be subsequently used to identify the strength of the correlations between sectors and their factors, as a higher accuracy representing one or more strong factors will be distinguished.
- Use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm in order to observe the clarity of the trends within each sector with unsupervised learning

2 Methodology and Dataset

2.1 Technical features and limitations of the Fortune 500 2017 Dataset

A limitation towards the methods is the fact that the number of data points in each sector is different, and the more popular and profitable sectors dominate the dataset. Therefore, sectors with more data points like Financials would have a higher bias towards them if used without preprocessing since a higher range of data is trained towards Financials and more anomalous results will skewer the training of other sectors. In order to combat this, the number of data points randomly sampled from each sector, will be the number of data points of the sector with the lowest number of data points. In addition, sectors with a frequency below 25 will be ignored as at that period, the data used to identify sectors would be too insignificant compared to the total quantity of samples, lowering accuracy as a result.

The debt-to-equity (D/E) ratio attribute and column can be derived from the (Assets - Total Share Equity) / Total Share Equity. This attribute can be subsequently trained and operated on with all other attributes.

Table 1: Fortune 500 Dataset Attributes and Descriptions

Attribute	Data Type	Description	Used
Rank	Integer	Rank of corporation according to total revenue	False
Title	String	Name of corporation	False
Website	String	Main corporation website URL	False
Employees	Integer	Total number of employees	True
Sector	String	Name of bain business sector	True
Industry	String	Specific industry in the sector	False
Hqlocation	String	Location of HQ in terms of city and state code	False
Hqaddr	String	Specific address of HQ	False
Hqcity	String	City HQ is based in	False
Hqstate	String	State HQ is based in, with a state code	False
Hqzip	Integer	HQ zip code (postal code) for package deliveries	False
Hqtel	String	HQ contact number	False
Ceo	String	CEO Name	False
Ceo-title	String	'President Chief Executive Officer & Director' title	False
Address	String	Full HQ address	False
Ticker	String	Share/Stock Code	False
Fullname	String	Official corporation name	False
Revenues	Integer	Total revenue through 2017 in millions	True
Revchange	Float	Change in revenue from 2016 in millions	True
Profits	Float	Annual profit in millions	True
Prftchange	Float	Change in profit from 2016 in millions	True
Assets	Integer	Estimated total worth of resources the corporation owns in millions	True
Totshequity	Integer	Difference between a company's total assets and its total liabilities in millions	True

2.2 Methodology for data exploration, data cleaning, feature engineering, modelling and analysis

The implementation methodology begins with data cleaning due to an understanding of the main analysis attributes. Therefore, removing null data and commas from all of the data is initially done.

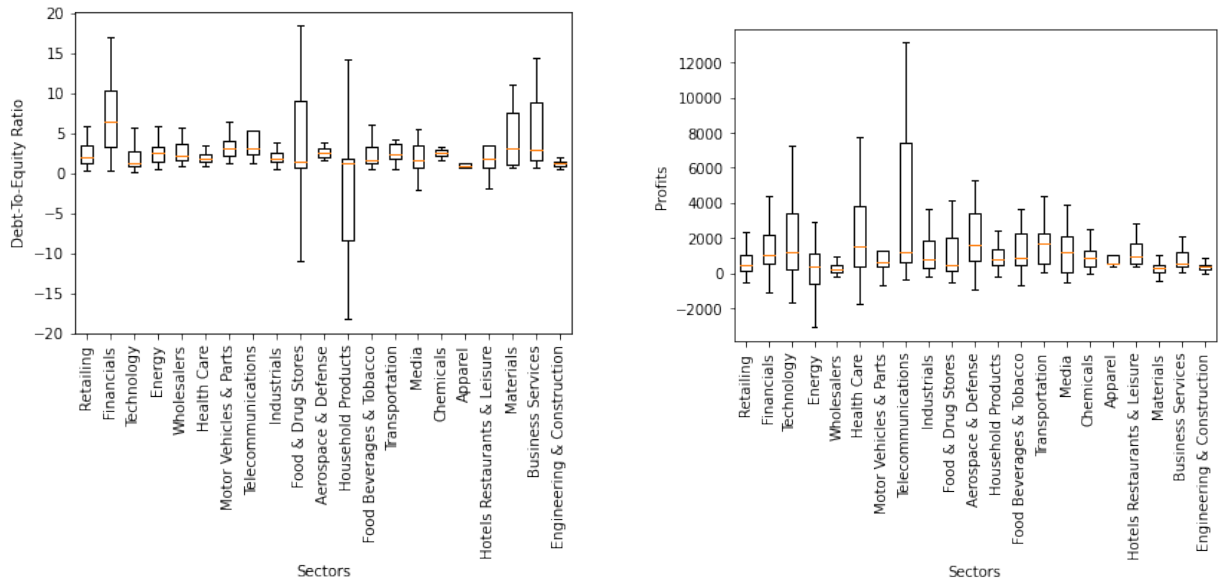

Figure 1: Boxplots for Sectors vs Debt-To-Equity Ratio and Profits

Subsequently, dimensionality reduction is used, removing all attributes excluding the sectors and numerical attributes which are intended for use.

Data exploration begins by observing the main attribute of the analysis (sectors), using statistical techniques to understand the total occurrences for each unique sector and the minimum and max value. In addition to the minimum excluding anomalies, first quartile, median, third quartile, and maximum excluding anomalies of all of the attributes for each sector. These can be applied to a data visualisation technique such as box plots seen in Figure 1, making trends and information more condensed and recognisable.

Both feature engineering and data cleaning is done whilst standardising, normalising, scaling and using PCA to transform the data for training in both MLP and DBSCAN. Feature engineering is done independently when adding the D/E ratio as it adds more weight and attributes to associate with a sector. The DBSCAN clustering is optimised by weighting the attributes according to the K-nearest neighbour distances.

MLP modelling is implemented by initialising the number of hidden layers, rows of perceptrons, that are individually sent to a logistic regression model that supplies a probability that a K latent feature is found in a specific class. These data features back propagate until the model converges to a probability of a class. In this case, all of the data left after cleaning are applied to the model when training, shifting the K latent feature probabilities of each node. The performance of the classifer was assessed by cross validating the predicted values with the correct values and subsequently calculating the percentage accuracy and overall variation from the mean.

DBSCAN works by clustering data points together according to a measure of distance between the data points. In this case K-nearest neighbour was used along with a minimum number of clusters defined. Data points in low density regions are marked as anomalous points. The attributes I decided to use were the D/E ratio and profits, this is due to these graphs being the most significant attributes when observing the generated box plots (Figure 1) since the 6 sectors on the left side of the graph are the used attributes and clear differences in values are visible. I excluded the assets and total share equity for plotting as they are dependent on the D/E ratio and vice versa. The performance was assessed according to factors such as homogeneity, V-measure and the Silhouette Coefficient.

3 Results

The final processed MLP results are displayed as confusion matrices in Figures 2, 3 and 4.

The final processed DBSCAN results are displayed as confusion matrices in Figures 5, 6 and 7.

3.1 Summary of training and modelling results

The resulting predictions and modelling of the MLP model had a decreasing performance at a linear rate in terms of the mean prediction accuracy, starting at a satisfactory 80% which dropped by approximately 20% with every 2 additional sectors added to the model. The standard deviation and layer units were almost constant for all tests. The sectors which were identified the most accurately were the Energy, Financials and Retailing sectors.

The results obtained from the DBSCAN were inadequate in terms of performance as they were mostly valid but not correct, since the only reason this result occurred is that almost all of the data point are essentially in one cluster so they are not measured to be incorrect, leaving the V-measure at 0. This was to be expected due to the lack of data points, whilst the data points had a small number average variations. The algorithm only progressed in terms of identifying more clusters being present graph.

3.2 Model Comparison and Recommendation about the Accuracy and Explainability

Both approaches were evaluated using different determined hyper parameters such as evaluating a different number of sectors, standardisation methods etc. MLP also had the number of hidden networks whilst DBSCAN had different numbers of nearest neighbours. MLP may have been modified by selecting a few attributes to train rather than all of them, enabling a more thorough assessment of individual attributes. DBSCAN may have been improved by attempting to modify variables like epsilon or using a different method to standardise the data.

Overall, the MLP model's performance far exceeded the DBSCAN due to the clarity, and ability to interpret why a sector was easier to identify than the other sectors when comparing to the box plots, such as the Energy being due to it's lower revchange and the Financials having a higher D/E ratio than other sectors. There were also insufficient data points that would enable the DBSCAN to find new trends beyond supervised methods.

4 Discussion

4.1 Key findings and insights

The MBL data in Figure 4 has proven that small data samples of 29 for each sector is robust enough for sectors to be identified after accounting for all possible attributes and Figure 2 depicts a benchmark accuracy of 80% using 114 data points. Therefore, if the same model has more data applied to it, a higher overall performance should be viable.

The data from the standard deviation model in Figure 4 shows that the Retailing sector has a dominating relationship with one or more attributes which overlaps with the Wholesale sector which can be identified to be due to the number of employees as they an almost exact box plot match to each other than employees. This shows that it is possible for the significance of the overall attribute identification to perfectly match all companies to their corresponding sector bar anomalous data

Due to the scatter graph nature of the DBSCAN model, a slight linearly decreasing trend is visible, depicting that the profit decreases as the D/E ratio increases, proving that a higher value is generally a risk factor in stability for investment.

4.2 Limitations of the modelling approaches used and possible future general improvements

The MLP model is restricted by it's training method as the process is fundamentally latent, forcing explanations for accurate results to be unclear, requiring the analysis of external data or backtracking

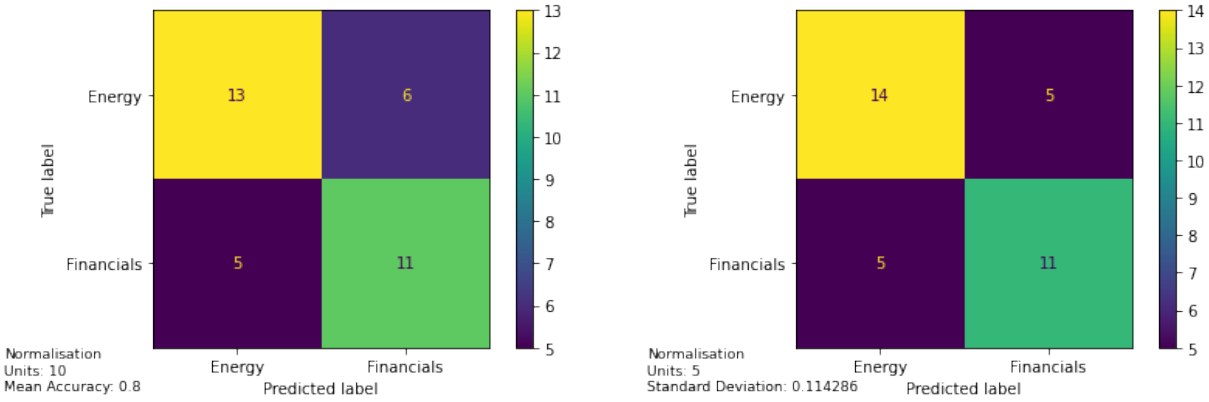

Figure 2: Two Sectors Confusion Matrices

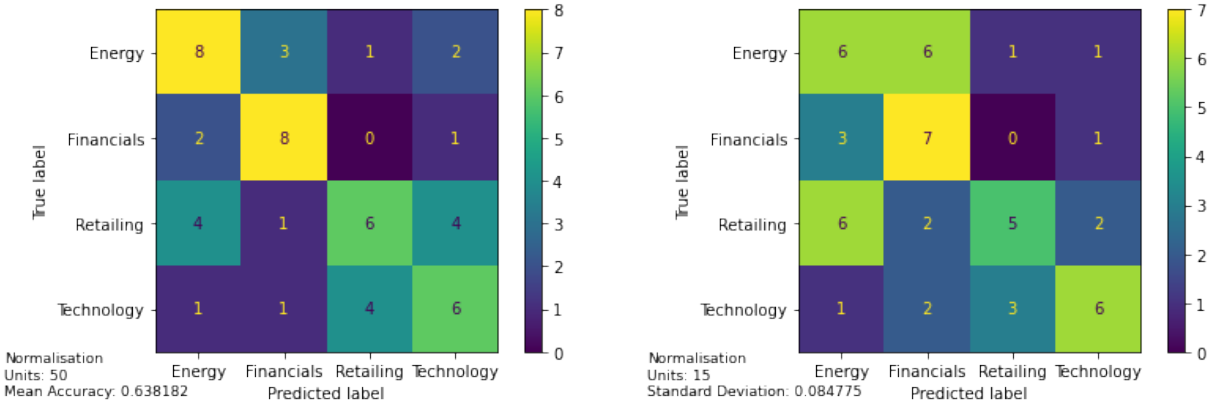

Figure 3: Four Sectors Confusion Matrices

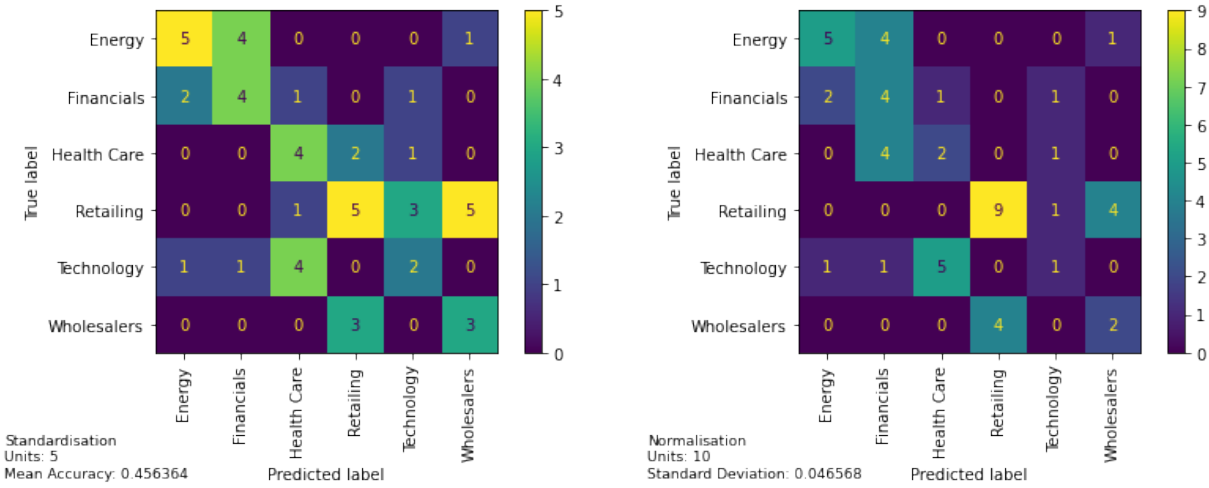

Figure 4: Six Sectors Confusion Matrices

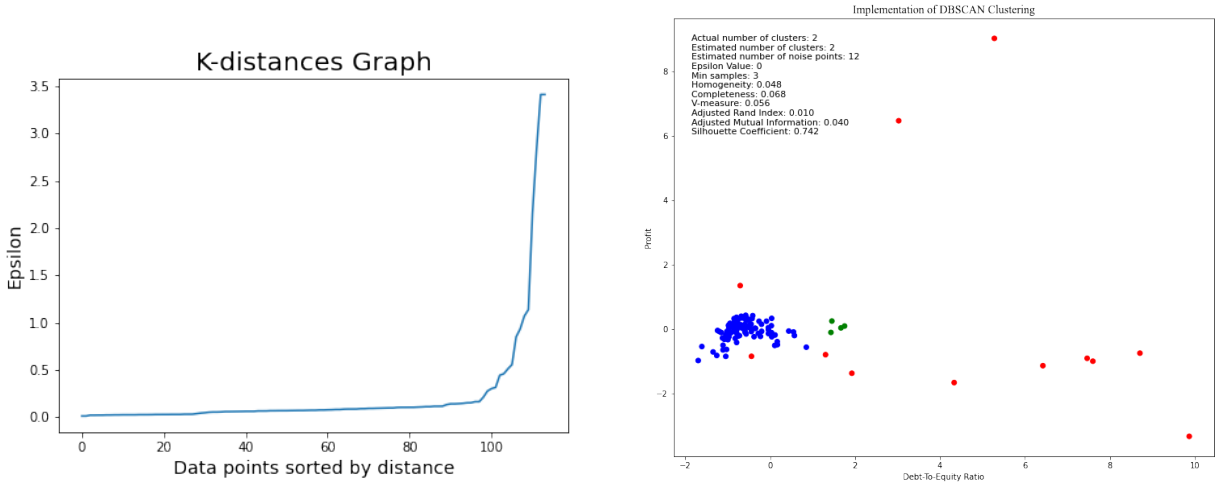

Figure 5: Two Sectors K-distances and corresponding DBSCAN

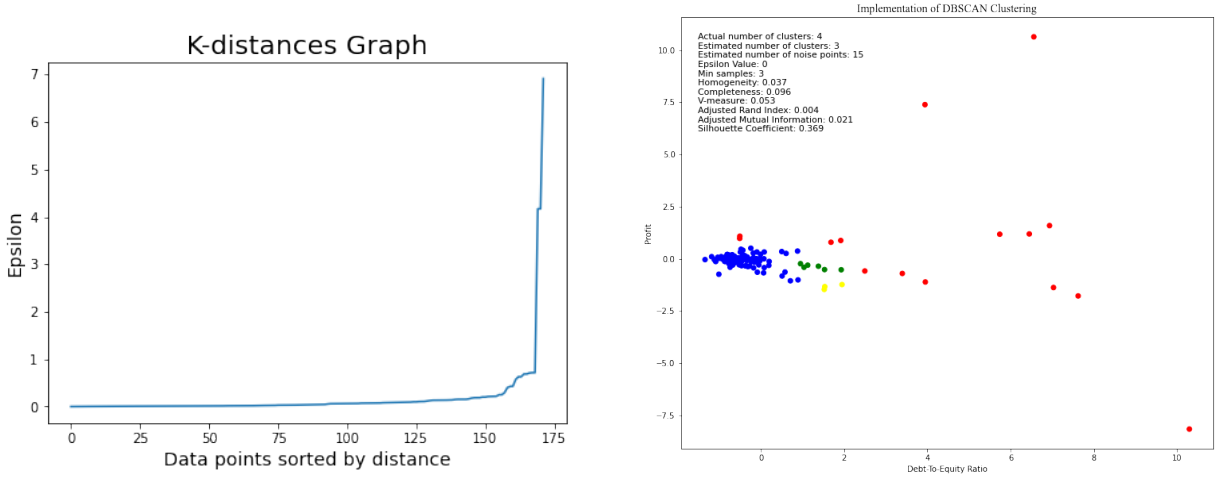

Figure 6: Four Sectors K-distances and corresponding DBSCAN

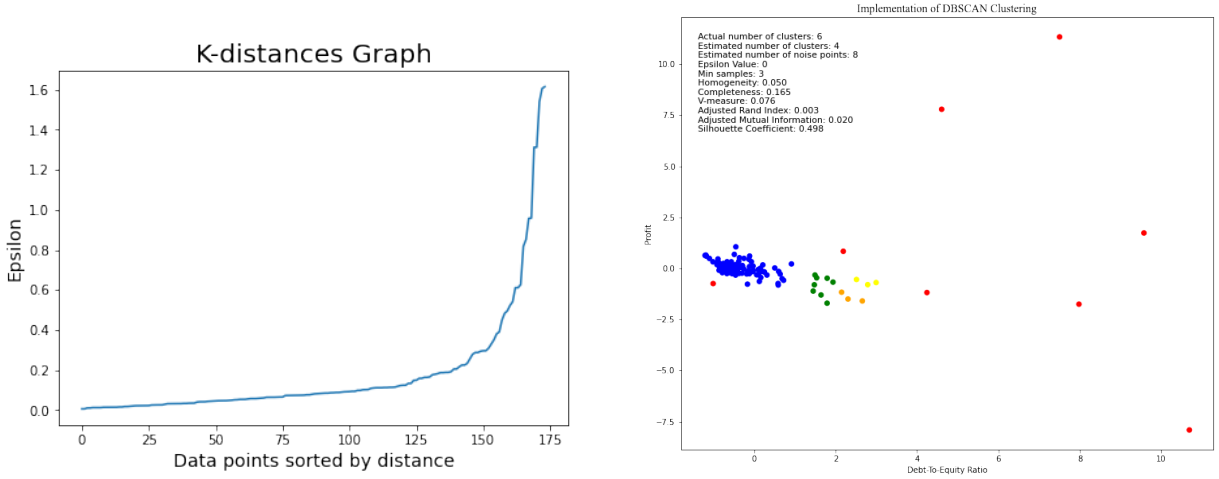

Figure 7: Six Sectors K-distances and corresponding DBSCAN

to understand the results. DBSCAN is severely weak to overlapping data points, as the data becomes inseparable and unidentifiable when placed under the banner of a single cluster. For dense data sets, this causes the majority of data to be identified incorrectly unsupervised.

The model is limited as the Fortune 500 data used is only from 2017 which restricted predictions via linear regression and the majority of the data was unusable. Therefore this data set would greatly benefit from being augmented by the following years of the Fortune 500 data. In addition, since the data identifies the stock code, the stock data for each corporation can be imported and manipulated together in conjunction, allowing regression analysis to be done more progressively. More data points will naturally stabilise the means of results, revealing more representative identifying features of every sector.

References

[1] A. Hayes, "What is a fortune 500 company? how companies are ranked," Nov 2022. [Online]. Available: https://www.investopedia.com/terms/f/fortune500.asp

[2] L. Belanger, "Visualize the fortune 500," May 2022. [Online]. Available: https://fortune.com/franchise-list-page/visualize-the-fortune-500-2022/

[3] J. Maverick, "Industry vs. sector: What's the difference?" Sep 2021. [Online]. Available: https://www.investopedia.com/articles/investing/083115/why-do-debt-equity-ratios-vary-industry-industry.asp