ECM3420 Report

Candidate Number: 250279

1 Introduction

In recent years, clusters of interconnected urban areas have been defined using the term 'megaregions'. These are areas comprised of multiple major metropolitan hub cities, surrounded by less populated 'sub-centers'. [1] This pattern was first identified in Gottman's 'Megalopolis', a nickname for the United States' northeastern seaboard. This region was comprised of some of the country's biggest cities which functioned like one interconnected entity. [2] Despite being a pattern seen across the world, megaregions have become instrumental in understanding the complex social and economic geography of the United States. [1, 3] Megaregions are centred on some of the biggest economic hubs, full of workers who travel within them. Thus, by understanding the trends of companies based within these megaregions, we may be able to identify these megaregions.

Megaregions have been previously identified using a combination of different methods. [1, 4, 5, 6] For example, the Regional Plan Association (RPA) methodology for identifying megaregions assigns each county in the US a value for meeting certain criteria regarding population density, population growth, and employment growth between the years 2000 and 2025. [6] Nelson and Rae posit that using data obtained from work commutes builds a clearer picture of the US' megaregions as commutes hold key importance "in the functioning of local and regional labor markets". [1]

One thing that remains consistent between these methods is the economic importance of companies and how they affect the region through their economic output or commuter reach. By using the raw economic data of companies across the United States, we may also be able to use this to quantify whether they are based within megaregions. The aim of this paper is to see whether it is possible to use a company's financial data to identify if it exists within a megaregion. This will be split into two parts, both employing two machine-learning methods. Firstly we will perform a binary classification to identify whether a company's headquarters is based within a megaregion. Then classification can be used to see whether the specific megaregion can be identified.

2 Methodology and Dataset

I am using the Fortune 500 2017 Dataset. [7] The Fortune 500 is a list compiled by Fortune magazine which ranks the 500 largest corporations in the United States by total revenue for the fiscal year. This dataset contains the largest companies which act as the drivers of growth within these megaregions. Thus we will see if a relationship is apparent within the data.

The Fortune 500 dataset is comprised of 500 rows and 23 columns. The 500 rows represent the top 500 corporations, ordered in terms of their rank. The 23 columns' attributes and what they mean can be found below in Table 1.

attribute	value type	meaning
rank	integer	the rank of the corporation
title	string	the name of the corporation
website	url	corporation's website url
employees	integer	number of employees
sector	string	business sector of corporation
industry	string	specific industry within sector
hqlocation	string	the location (in terms of city and state) of the corporation's headquarters
hqaddr	string	the number and street name of the corporation's headquarters
hqcity	string	the city of the corporation's headquarters
hqstate	string	the state of the corporation's headquarters
hqzip	integer	the zip code of the corporation's headquarters
hqtel	string	the telephone number of the corporation's headquarters
ceo	string	the name of the corporation's chief executive officer
ceo_title	string	the titles held by the corporation's chief executive officer
address	string	the concatenation of hqaddr, hqlocation, and hqzip
ticker	string	an abbreviation used to uniquely identify the shares of the corporation on the stock market
fullname	string	the corporation's full name (this may be slightly different to the 'title' attribute)
revenues	integer	the yearly revenue of the corporation (in million usd)
revchange	decimal	the percentage change of the revenue from the previous year
profit	decimal	the yearly profit of the corporation (in million usd)
prftchange	decimal	the percentage change of the profit from the previous year
assets	integer	the total value of the corporation's assets (in million usd)
totshequity	decimal	the corporation's total shareholders' equity (the amount that the owners of a company have invested)

Table 1: Table of attributes within the Fortune 500 Dataset

Using the location of the corporation's headquarters from the 'address' attribute, we can determine whether that corporation is based within a megaregion. This will allow us to create a new column to show this information. To do this I will use the currently identified megaregions in the United States that have been collated using several papers. [6]

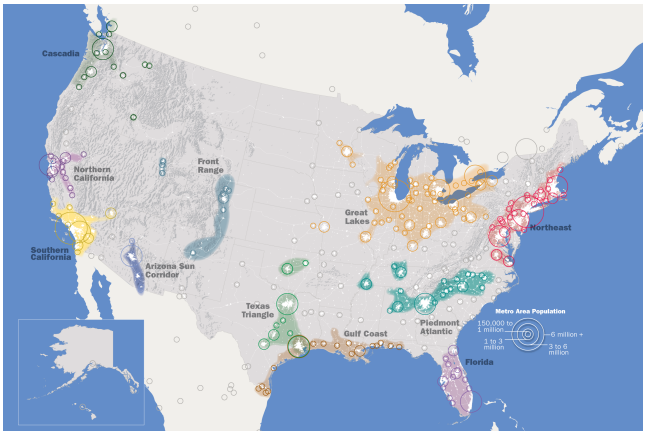

Figure 1: RPA Map of Emerging US Megaregions

By splitting the data into these regions, we do notice significant biases. Firstly, as the Fortune 500 dataset contains the biggest companies, they tend to be located within megaregions. This means that our dataset is biased when performing binary classification as the data trends towards always identifying megaregions. When performing the individual identification classification of megaregions, we see a large bias towards the Megalopolis (Northeast) megaregion. Megalopolis

has 158 companies (31.6%) in the Fortune 500, whilst only having a population of 17% of the US in 2010. [8] This is because not only is it geographically the largest megaregion but it also contains large commerce hubs like New York and Boston. Intending to reduce this bias, I will be separating Megalopolis into two halves, as it has been shown by Nelson and Rae that there are distinct clusters of commuters in this region. (See Figure 2) [1]

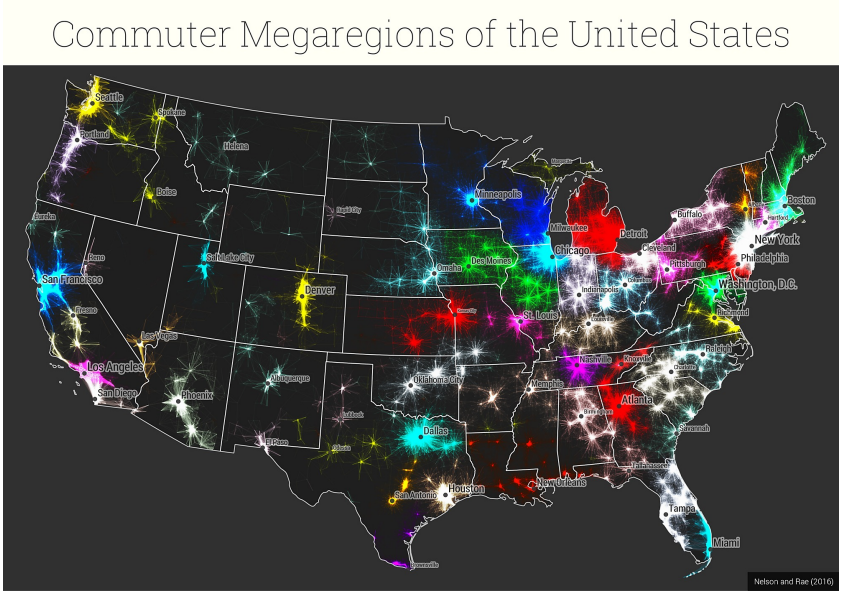

Figure 2: United States Commutes and Megaregions data for GIS by Nelson and Rae

I have also decided to combine some of the geographically close megaregions. Firstly Northern California and Southern California have been combined to make a singular California megaregion. In addition as the Gulf Coast and Texas Triangle megaregions, converge on Houston, Texas I have decided to combine these as well. Finally, I have combined the Arizona Sun Corridor with the neighbouring Front Range as both have a limited number of Fortune 500 companies (4 and 10 respectively). This is the final breakdown of the megaregions, the number of companies in them, and what they are comprised of:

Megaregion	Number Of Fortune 500 Companies	Comprised Of Companies from:
Great Lakes	99	Illinois, Indiana, Iowa, Michigan, Ohio, Upstate New York (Buffalo, NY and Corning, NY), Wisconsin
Upper Megalopolis	87	Connecticut, Massachusetts, New Hampshire, New York, Rhode Island
Lower Megalopolis	71	Delaware, Maryland, New Jersey, Pennsylvania, Virginia
Texas Gulf	57	Louisiana, Oklahoma, Texas
California	53	California
Piedmont	29	Alabama, Georgia, Mississippi, South Carolina, Tennessee
Florida	17	Florida
Cascadia	14	Idaho, Oregon, Washington
Front Range Arizona	14	Arizona, Colorado, New Mexico, Utah
None	59	Any other location...

Table 2: Table of Megaregions, the number of Fortune 500 companies within them, and the US States they are comprised of.

The other values I will be using from the dataset will be 'employees', 'revenues', 'profits', and 'assets'. These give a good indicator of the financial might of a corporation. It should be noted that many of these companies have stores nationwide and so the 'employees' value does not reflect the population of the megaregion. In theory, I expect that richer companies will be based within megaregions, though I don't think the sample size is large enough to see this in full effect. I will use these values to classify the megaregion using two methods of classification in this project: Weighted K-Nearest Neighbours (KNN) and a Multi-Layer Perceptron (MLP).

2.1 Weighted K-Nearest Neighbours

Weighted K-Nearest Neighbours is a supervised machine learning algorithm for classification problems which obtains a better idea of similarity between values by using a distance metric.

KNN algorithms assume that similar data points will exist close together. Using this, KNN can determine the class of a new point by the majority class of its k nearest neighbours. For example, a k value of 5 would mean a data point will be determined by its 5 closest values. KNN algorithms are susceptible (especially in instances with low k values) to the influence of outliers. Regular KNN cannot factor in the prominence of classes, so outliers can have a noticeable effect on accuracy. In addition, KNN algorithms are very susceptible to class imbalance. If the dataset contains significantly more data points of a certain class, it will be biased towards that class.

Weighted KNN slightly offset this bias by assuming that the closer the neighbour is, the more influence it has on the classification compared to further away points. Due to its heavy reliance on the distance metric, it may give better results with Normalised data thanks to better scale. This distance metric is usually Euclidean distance but can be changed to use other metrics such as Manhattan distance. Weighted KNN performs instance-based learning, constructing a hypothesis directly from the training data. As a result, the complexity grows with the size of the training data. This means that if there is too much training data, the overall accuracy will be lower and will further lean into any class imbalances apparent in the dataset.

2.2 Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a supervised learning algorithm. It allows us to classify data that does not follow a linear classification pattern, where the decision boundaries between classes tend to be more complicated. MLPs perform individual logistic regressions on data's 'K latent features', characteristics that are usually unobservable. This yields a probability for the classification of the data. MLPs utilise 'Deep learning', where the model has multiple layers of these latent processes. This allows us to train the model using n layers. This means that it can learn decision boundaries far more complicated than those generated through other machine learning algorithms. It is able to factor in more values to reach a conclusion in order to see better patterns not visible to algorithms like Logistic Regression or KNN that only work in two dimensions.

3 Results

Firstly we will look at how our binary classification looks with some of the attributes of the Fortune 500 dataset when plotted on scatter graphs. There is a significant bias within the data towards megaregions.

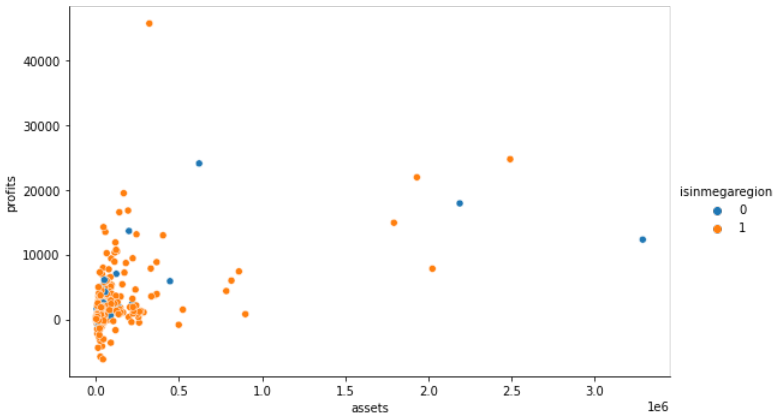

Figure 3: profits to assets Binary Scatter Graph

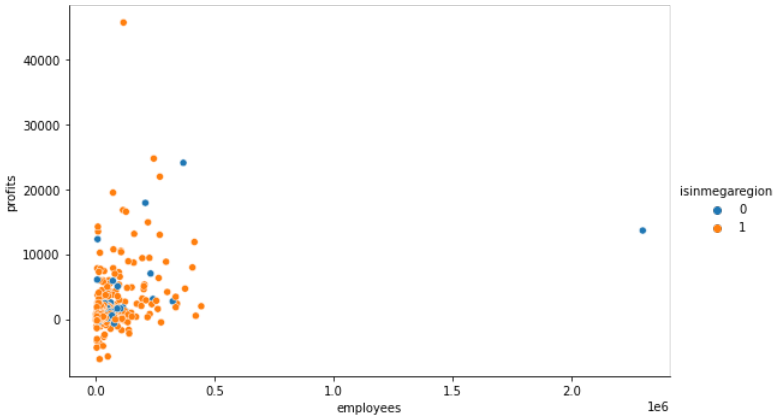

Figure 4: profits to employees Binary Scatter Graph

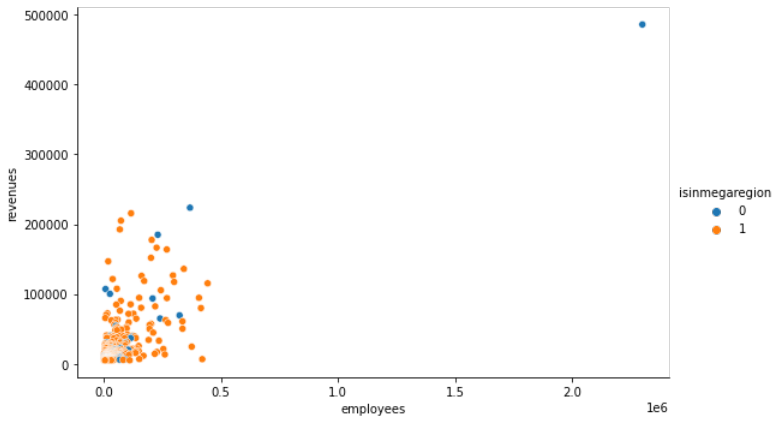

Figure 5: revenues to employees Binary Scatter Graph

Now we will look at the same plots for identifying the megaregions.

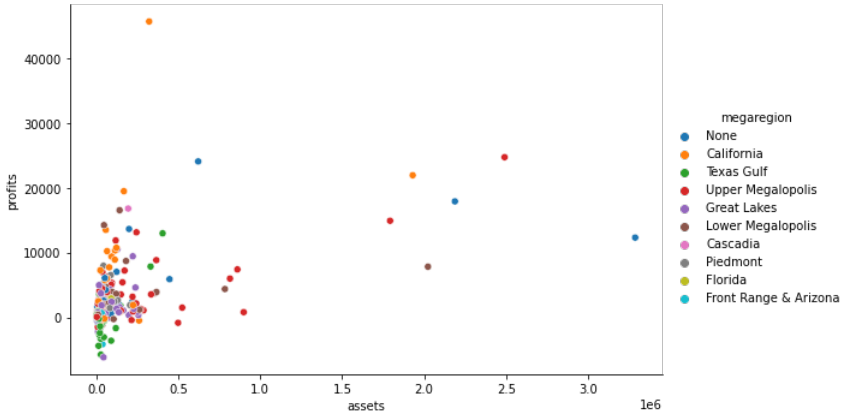

Figure 6: profits to assets Identity Scatter Graph

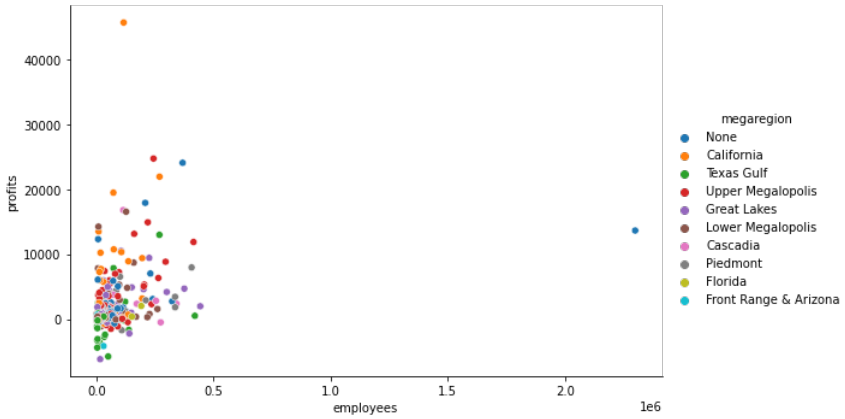

Figure 7: profits to employees Identity Scatter Graph

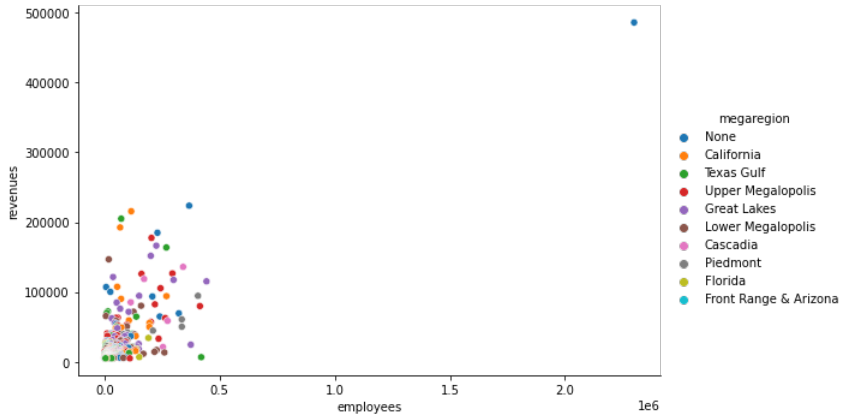

Figure 8: revenues to employees Identity Scatter Graph

There are no clear clusters of related data. This means that the Weighted KNN may struggle with accuracy.

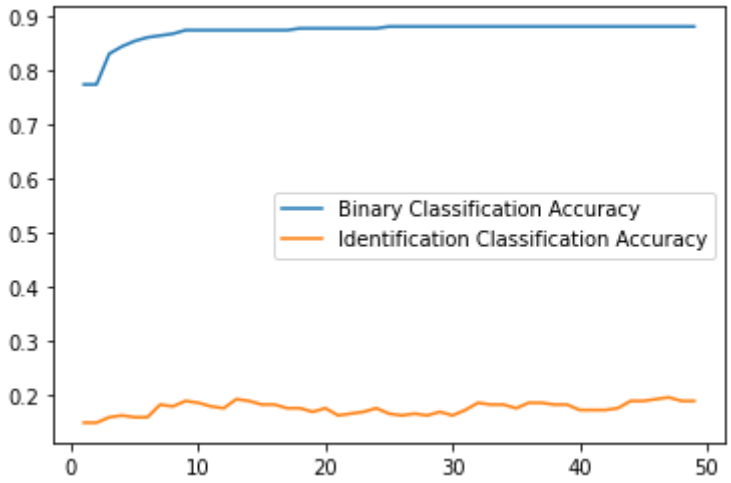

Figure 9: Accuracy for values of k = 1 to 50

Increasing the K value has little effect on the Identification accuracy. Whilst the accuracy appears to increase for the Binary accuracy, it is actually guessing every value is a megaregion due to the enormous bias. Normalising does not have an effect on either accuracy. Changing the distance metric to Manhattan does change some predictions but the overall accuracy doesn't change.

distance type	accuracy	confusion matrix
manhattan	0.86	[0, 39, 3, 258]
euclidian	0.8633333333333333	[0, 39, 2, 259]

Table 3: Comparison of different distance metric calculations

In comparison, the Multi-Layer Perceptron has had similar accuracy. However, normalising the Identification Classification has had beneficial effects. However, normalising the binary classification has resulted in the model predicting a megaregion almost every time due to the high bias of this outcome.

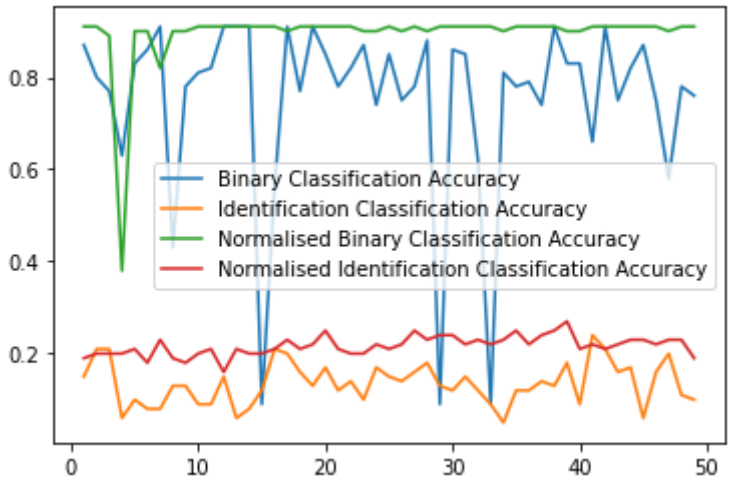

Figure 10: Accuracy for number of layers n = 1 to 50

Figure 11 shows the main comparison of the accuracies of both models.

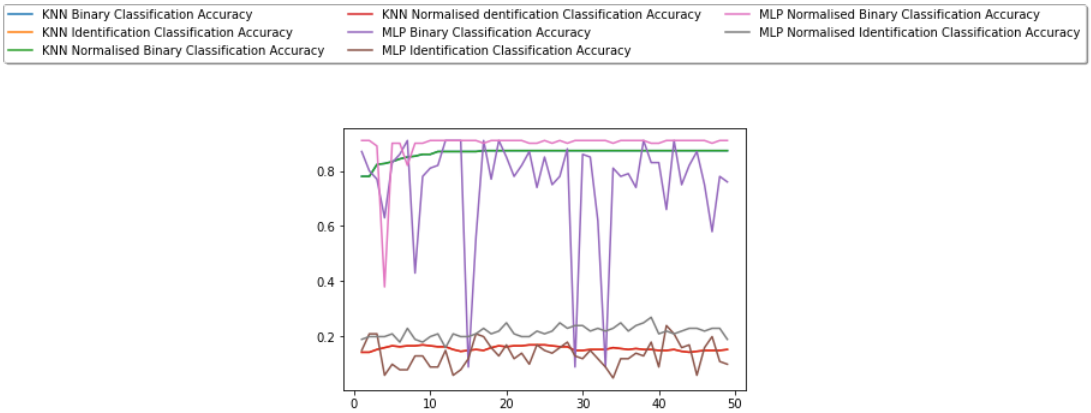

Figure 11: Accuracy for number of layers n = 1 to 50

4 Discussion

From the models I have built, it does not seem possible to use the Fortune 500 dataset to classify megaregions. Due to a lack of clear clusters within the identification classification and an overall bias that leads to predicting megaregions with the binary classification, Weighted KNN is not particularly accurate. MLP had limited success, but due to similar values amongst the top 500 companies, I don't think the attributes have much effect on the overall classification. Using the same methods on a larger or more diverse dataset (such as the Fortune 1000) may lead to a better result. Other models may produce better results such as a Convolution Neural Network.

References

[1] G. Dash Nelson and A. Rae, "An economic geography of the united states: From commutes to megaregions," *PLOS ONE*, vol. 11, no. 11, pp. 1–23, 11 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0166083

[2] J. Gottmann, *Megalopolis: The Urbanized Northeastern Seaboard of the United States*. The MIT Press, 03 1964. [Online]. Available: https://doi.org/10.7551/mitpress/4537.001.0001

[3] R. Florida, "Mapping the mega-regions powering the world's economy," Feb 2019. [Online]. Available: https://www.bloomberg.com/news/articles/2019-02-28/mapping-the-mega-regions-powering-the-world-s-economy

[4] R. Florida, T. Gulden, and C. Mellander, "The rise of the mega-region," *Cambridge Journal of Regions Economy and Society*, vol. 1, 05 2008.

[5] J. Kotkin and M. Schill, "A map of america's future: Where growth will be over the next decade," *New Geogr*, 2013.

[6] Y. Hagler, "Defining us megaregions," *America*, vol. 2050, pp. 1–8, 2009.

[7] Fortune, "Fortune 500 - 2017," 2017. [Online]. Available: https://data.world/aurielle/fortune-500-2017

[8] United States Census Bureau, "Decennial census by decade, 2010," 2010. [Online]. Available: https://www.census.gov/programs-surveys/decennial-census/decade.2010.html