Investigation into which factors most affect academic success in Greater London wards

Victor Morland *u1803998*
03.01.2022

***Abstract*—This report will look at data relating to wealth, crime and social indicators of boroughs/wards in Greater London and compare them with academic success in those boroughs using GCSE results as a metric. This report will then try to find which if any factors strongly affect GCSE results.**

I. Introduction

With the increased inter-connectivity of our lives and especially the lives of teenagers, particularly the ability to get information themselves away from schools. The factors that affect academic success are more difficult to identify than perhaps ever before. Although it is clear that ignoring the quality of school certain socio-economic factors clearly affect the academic success of pupils. Additionally, with the rise of online learning whether that is due to recent pandemic enforced social restrictions or more generally lessons being available whether for free on websites or paid tutoring. A difference in results between wards would suggest that social factors are the biggest limitations on academic success rather than the quality of schooling. It is expected that areas with better economic indicators such as household income will have better academic success as parents/guardians will have more disposable income and/or time to support their children/wards (here ward refers to the person under the protection of a legal guardian not the geographical location of a London ward). This report will use no indicator to presume the quality of schools in an area.
The aim of this report is to see if it is the case that non-schooling related factors affect academic success, and if so, then to what degree these factors affect academic success.

II. Background

A. Measuring Academic Success

Although there are many ways to measure academic success and indeed a range of indicators would be favourable, for this report we will focus on GCSE results. We will do this for a couple reasons, first every pupil has to take GCSE results, and most schools will offer a similar range of GCSEs. This is relevant as for other qualifications there is a much wider range of chosen subjects which could affect the resulting data with some subjects seen as softer. To get one result out of each pupil taking many (often 9) GCSEs, we will use the average GCSE capped point score.
The GCSE capped point score is a sum of up to 9 GCSES (English, Math and a Science are taken then the 6 next best scores) for a pupil. This is then averaged for every pupil in the ward. Now since the grading change to GCSEs only happened in the academic year 2015/2016. GCSE results from before this change to a numeric grade are converted from the letter grade to a number. That roughly corresponds to the new grading system.
This measure for GCSE results is good as it does not hyper focus on certain GCSEs as other measures do and can also be calculated in case of further research using the new GCSE grading system.

B. Choosing indicators to compare

The indicators we want to choose to compare with academic success are preferably indicators that give a wide range of information about a borough, such as economic indicators, familial indicators, social, quality of life and others. Additionally, although this data may be harder to find, indicators that specifically relate to children at schools would be very helpful.

C. The impact of individual schools

It is important to add that as we are working on a very small geographical area, that of wards in London the quality of individual schools on academic success is something that should be at least

considered. However, pupils from the same wards will go to different schools and additionally, if no strong correlation is found between indicators then it would imply that individual schools are what impact academic success rather than any neighbourhood factors. Or that we have looked at the wrong data. Clearly from figure 3 we know that different areas have quite substantially different academic success so we know that something locally must be the cause of these differences in results whether it is the schools or socio-economic factors.

III. Data

A. Ward Profiles and Atlas

The data for this project was collected from the London datastore, using the Ward Profiles and Atlas data set[1]. This is a data set created by the Greater London Authority (GLA) it takes data from a wide range of data sets and combines them to create profiles of London wards. It has data up to 2015, and most of the data was collected between 2013-2015. Now clearly this data set has far too many features and not all are relevant to measures of academic success. Although the choice of what is and isn't relevant is not so clear for some features. The choice of what features to keep and what to remove will be up to the discretion of the author but some examples of features that will be removed are other academic indicators that are either too similar to the average GCSE capped point score, as these will likely be too heavily influenced by the individual schools which is not the aim of the report. Or academic indicators that count post GCSE qualifications such as A-level results as these will likely be irrelevant. Additionally, general data such as population will likely be removed as they would be not very relevant.
This data set is unfortunately restricted in its possible size as of course there are only 624 wards in Greater London (at least in this data set Ward boundaries change and could have changed since 2015). But hopefully this will still be large enough to gather useful conclusions. This data set cannot be meaningfully expanded as data on ward specific academic success is not measured every year, only at a borough specific level but that would lead to even less data. As there are only 32 London boroughs plus the City of London. Despite having access to a wider range of data at a borough level (and perhaps more accurate data) choosing to restrict ourselves to a ward level is likely the better option.

The features of this data set that will be kept are the ones that are deemed most relevant these will be:

- % BAME - 2011
- % English is First Language of no one in household - 2011
- % children in year 6 who are obese- 2011/12 to 2013/14
- Employment rate (16-64) - 2011
- Median Household income estimate (2012/13)
- % Households Social Rented - 2011
- % dependent children (0-18) in out-of-work households - 2014
- Crime rate - 2014/15
- % area that is open space - 2014
- Average Public Transport Accessibility score - 2014

And of course the average GCSE capped point score.

IV. Software

A collection of different software tools will be used in the making of the report, here is a list of the most important pieces of software used to process and analyse the data. The data set can be installed in either Excel files or Comma Separated Value (CSV) files.

A. Microsoft Excel

Microsoft Excel is useful for downloading data that can only be downloaded as an Excel file and for seeing the raw data in an easy to view way. It is very easy to use and can be used for basic data manipulation.

B. Command Line Tools

Used for very simple data analytics, mainly used for searching for values and for file manipulation.

C. Weka

Provides a graphical user interface to look at data and will be used to quickly check various classification algorithms and get visual results from the data.

D. Python

This, will be the main tool used for this report. It will be used to generate the graphs and figures in this report and will be the tool mainly used for data cleaning and analysis. Its flexibility and the author's knowledge of specific python libraries such as pandas and sklearn, which can both be used for data analysis is the key reason Python will be used. But other programming languages such as R could easily be used instead.

V. Hypothesis

It is reasonable to assume that economic indicators will have a strong effect on academic success. Therefore, we will hypothesize that the most significant factor to strong academic success will be the Median Household income estimate. And that Median Household income estimate can be used as a predictor for academic success.

VI. Data Cleaning

A. Missing Values

The first step is to check for missing data, any empty cells for wards for any of the features or cells that are designated as empty using a designator such as '-' or 'N/A'. Fortunately, the data set is mostly complete with the only missing values being found in the column for % of children in year 6 who are obese 2011/12 to 2013/14. As there seems to be no particularly good way to calculate these values, the value for the whole borough that the ward is in was used. For example, Camden Hampstead Town had no value for this column, therefore, the value for Camden as a whole was used which is 21.4. This data for the individual boroughs was also contained in the data set.

B. Removing Unnecessary Data

There is a mix of unnecessary data in the downloaded data set. First of all there are a lot of features that are concluded to be irrelevant to academic success. These columns were removed using python using the pandas library by only keeping the columns described in the above Data section.
Additionally, the data set came with information on the boroughs in London on top of data for the wards. Since the aim of this report is to focus on individual wards these rows were removed from the data set. This was accomplished in Python by checking for the area codes that matched the boroughs.

C. Dealing with Outliers

Fig. 1. Employment Rate Box Plot

As we are dealing with very small areas it is likely that outliers in the data will occur that are not truly representative of the ward. To check how many outliers may exist in this data set box-plots were created for the columns as shown in figure 1. These provide a great visual way for seeing the distribution of data in a column. The standard way, and the way we have decided to identify outlier points is if they are more than three standard deviations away from the mean. To do this the z-score was calculated which is calculated using the equation:

$$\frac{\text{value} - \text{mean}}{\text{std}}$$

for each cell in a feature. Now if the z-score is greater than three, we would consider it an outlier. Then all outlier values are replaced with the mean value. Now this is of course not done for the GCSE capped point score. But additionally, since we do expect different wards to have substantially different scores for some features the formula was not applied to '% area that is open space - 2014' and 'Average Public Transport Accessibility score - 2014'. As these two features are very static and easy to calculate and so, should be more trustworthy. Figure 2 shows the employment rate Box Plot after outliers have been replaced. As you can see there are no more outliers.

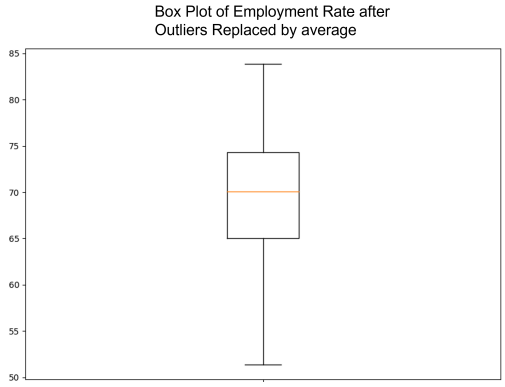

Fig. 2. Employment Rate Box Plot with Outliers Replaced

D. Grouping Academic Success into Classes

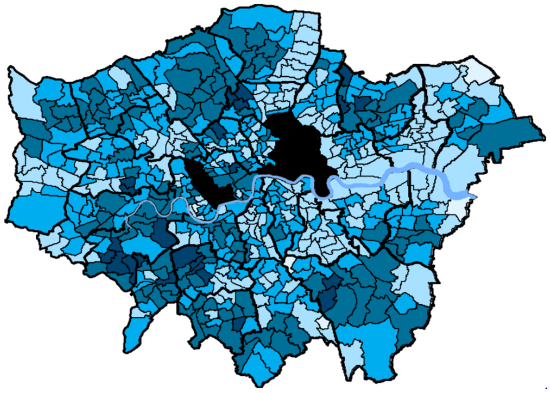

Fig. 3. Academic Success by Ward

To be able to create a classifier we first need to group up the data from the average GCSE capped scores. Using a box-plot of the scores the following groups have been created to get a roughly normal distribution (see Figure 4) (as shown in the following table). The academic success using the grouped values by ward has been shown in Figure 3. With darker blues representing higher academic success. As we can see there is indeed a large variation in academic success by ward in Greater London.

Classification	Values
Very Low - 1	≤ 290
Low - 2	≤ 315
Average - 3	≤ 340
High - 4	≤ 370
Very High - 5	> 370

The class distributions are shown in figure 4.

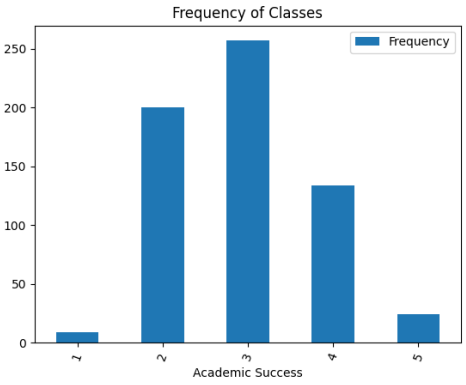

Fig. 4. Class Frequency

E. Normalising Data

As we have many features which use different measurements it is important to normalise the data as otherwise features which have a large variance between values will be dominant in our classification. Additionally, as we have so many features it might be wise to perform Principal Component Analysis (PCA) during the data analysis, therefore, to save us time in that case we will normalise the features now. We will do this with min-max normalisation so each data value will be between 0 and 1. The equation for this is:

$$\text{New Value} = \frac{\text{value} - \text{min}}{\text{max} - \text{min}}$$

VII. Correlation Analysis

Our first step should be to look for correlations between the features and academic success. For this we do not use the normalised values or the grouped versions for academic success for presentation purposes of the graphs. We will only plot graphs for features that seem relevant (that have a non-weak correlation). We first start with our hypothesis on median household income. We will see that economic indicators and features that specify information on the children in the ward in particular are most strongly correlated with academic success. Additionally, we see that open space and transport accessibility have no correlation with academic success and therefore will no longer be used as part of our data set.

A. Median Household Income

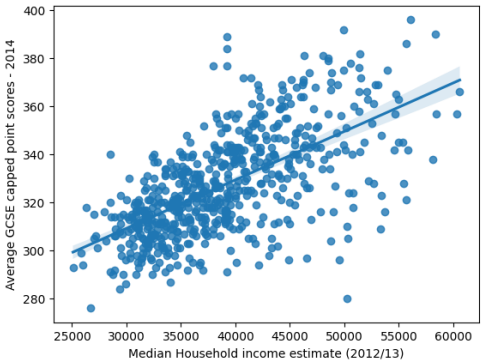

Fig. 5. Median Household Income

First we look for correlations with our hypothesis, we get a value of 0.6279. Which is a decent correlation and shows our hypothesis has at least some legs. In that if a ward has a higher median household income then it is likely to have a better academic success.

B. % BAME - 2011

This has a very weak negative correlation of -0.3311. This shows that coming from a non-white racial background does seem to correlate with a lower score in GCSEs however minorly.

C. % English is First Language of no one in household - 2011

This has a very weak negative correlation of -0.1597. This implies that not having English as a first language has little impact on academic success.

D. % children in year 6 who are obese - 2011/12 to 2013/14

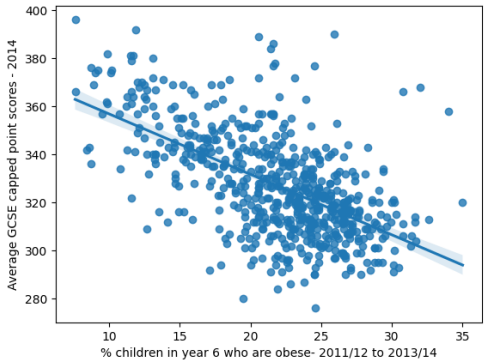

Fig. 6. Obesity

This has a negative correlation of -0.5908. This is the strongest non-economic correlation found. This could be due to a variety of factors, it could also be assumed that childhood obesity is a result of economic conditions as well. But perhaps other reasons such as bullying, or lack of confidence could explain this.

E. Employment rate

This has a positive correlation of 0.4840. Another somewhat strong economic indicator that correlates well with academic success supporting our hypothesis.

F. % Household Social Rented - 2011

This has a negative correlation of -0.5784. This is another economic indicator that supports our hypothesis that wards where families are more economically stable have more academic success.

G. % dependent children (0-18) in out-of-work households - 2014

This has a negative correlation of -0.7111. This is the strongest correlation we have found. This is perhaps not surprising as it is an economic indicator that focuses on children rather than on an entire household.

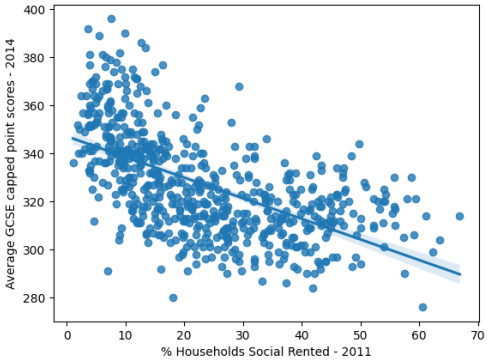

Fig. 7. Households socially rented

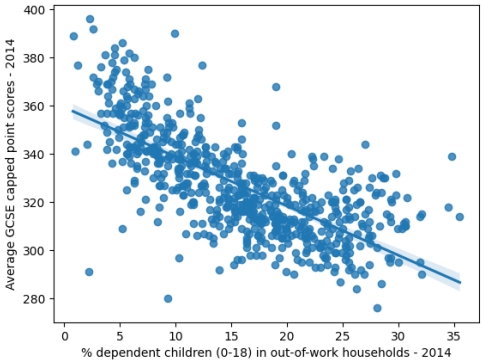

Fig. 8. Dependent children in out-of-work households

H. Crime rate - 2014/15

This has a negative correlation of -0.2692. So it seems the crime rate in a ward has little in common with academic performance.

I. % area that is open space - 2014

This has a very weak negative correlation of -0.0110. It seems barely relevant at all which is quite surprising.

J. Average Public Transport Accessibility score - 2014

This has a very weak negative correlation of -0.0633. Similarly to open space transport accessibility has near to no effect on academic success.

VIII. Classification

As the data set is not too large there is no need for PCA, additionally when using PCA on the data set we will lose information and therefore the quality of the classifier will be hurt. But in the process of using PCA we can find a variance ratio for principal components. It was discovered that in fact the first two principal components are responsible for over 75% of the classification information in the data set.

A. Accuracy of Classifiers

A number of different classifiers were used on the data set with a 75-25 split used for training and testing. After trying Logistic Regression, SVM, Random Forest Classifier, MLP (Multiple Layer Perceptron) the SVM classifier had the best result by a considerable margin with an accuracy of 0.6859 or **68.59%**. Now this is not a great result and probably cannot be used for prediction, but it does imply that academic success is associated with local social and economic factors. The list of accuracy's for the different classifiers are in figure 9.

```
Logistic Regression Accuracy: 0.6154
SVM Accuracy: 0.6859
Random Forest Accuracy: 0.6282
MLP Accuracy: 0.5769
```

Fig. 9. Accuracy's Of Classifiers from Python Code using sklearn

B. Kappa Statistic

More specific information on the SVM classifier shown in this figure including the confusion matrix and the Kappa statistic:

```
SVM Accuracy: 0.6859
Confusion Matrix:
[[ 0  1  0  0  0]
 [ 0 35 25  3  0]
 [ 0  5 44  3  0]
 [ 0  1  7 28  0]
 [ 0  0  0  4  0]]
Kappa Statistic: 0.5330482590103849
```

Fig. 10. Information on the SVM classifier

The Kappa statistic compares the number of correctly classified instances to the probability of random agreement between the attributes. Where a value of 1 shows that the variables are in complete agreement while a value of 0 would imply that all the data matches are in fact random. As our Kappa

statistic (Figure 10) is just above 0.5 it does imply that there is some agreement however it is not very strong so we will now look at another metric for analysing our classifier.

C. ROC curves

A ROC curve compares the True Positive Rate (TPR) to the False Positive Rate (FPR). The area under the ROC curve will range between 0.5 and 1. With 0.5 showing all matches are by chance. Since our classifier is not binary we will have multiple ROC curves for each class in our data set.

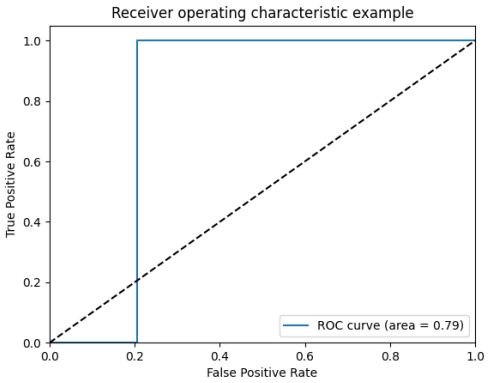

Fig. 11. ROC Curve for academic success 1 - Very Low

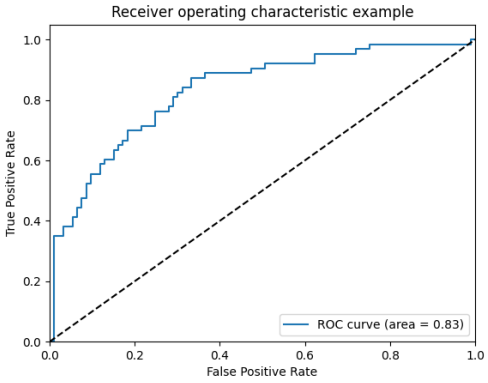

Fig. 12. ROC Curve for academic success 2 - Low

For the first and last class there are very few data values in them which is why the ROC curves have so little information. For the classes 1,2,4 and 5 the ROC curve areas are around 0.8 which is rather good and would imply that there does indeed a correlation between local social and economic indicators and academic success.

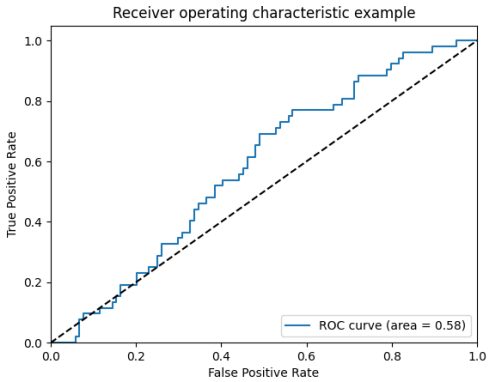

Fig. 13. ROC Curve for academic success 3 - Average

Interestingly though for class 3 - Average (Figure 13), the ROC curve area is very low. This perhaps suggests that for areas with average academic success changing local social and economic conditions has no effect on academic performance.

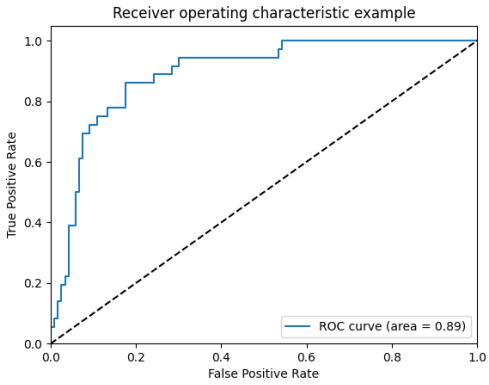

Fig. 14. ROC Curve for academic success 4 - High

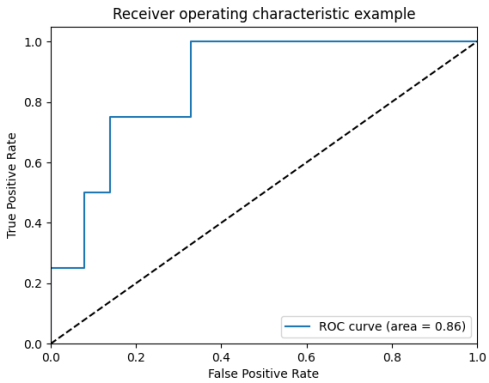

Fig. 15. ROC Curve for academic success 5 - Very High

IX. Conclusion

It can be concluded that as hypothesised there is indeed a relationship between economic indicators and academic success within wards in Greater London. This could show, that to level up academic success it is not just up to individual schools to improve teaching but also for councils to have more complete measures to help pupils achieve their potential.

However, unfortunately the ability to predict pupil performance in a ward using this data is not possible as the accuracy is indeed far too low. This is understandable as we must assume that individual schools still have a strong impact on pupil academic success. Moreover, factors that to the author originally seemed important such as rate of crime, amount of open space, and public transport accessibility had very little relevance (very low correlation) with academic success. The most important factors were definitely economic factors, which can be used to infer that even though all pupils go to school for free (only state schools are considered in this data set), they are still restricted by their economic class.

Due to the strict limitation on the number of wards in London the data set is still comparatively small with only 624 records. This could be increased by including data from other years however more research would have to take place to capture the relevant yearly data for each ward. As unfortunately this data (and especially more recent data) is not currently available in the London Datastore.

X. Possible Extensions

A. Covid

In particular comparative research on how academic success has changed under recent social restrictions with the Covid pandemic would be of great interest. Especially over the years 2020-2022 where many classes have been taught online with exams not taking place (except for 2022 where the decision on whether exams will take place is not yet known) it would be very interesting if certain factors have become more or less important to academic success and indeed whether it has further decreased the importance of individual schools on academic success or in fact if the opposite has occurred. A useful feature to add to the data set for this extension would be broadband speeds.

B. Extending to the UK

This would be a large undertaking as such specific information at a local level is difficult to find and additionally as the population is less dense outside of London there are fewer schools in each area. However, comparing which factors affect academic success in the UK compared to London would be an interesting extension.

References

[1] Greater London Authority. *Ward Profiles and Atlas*. URL: https://data.london.gov.uk/dataset/ward-profiles-and-atlas. (accessed: 01.01.2022).