IBM Data Science Capstone Report

Patrick Marshall 1.12.20

Introduction

Universities in the UK often have issues identifying how to achieve sustainability goals. One of the most common issues encountered, is how to set carbon emission targets, that are ambitious enough to reach a sustainable aim, whilst also remaining realistic and in line with sector wide competition.

Senior executives at the University are responsible for setting these targets in yearly planning rounds, which are then reported to governance boards, such as councils, or external committees. Within these governance structures, senior executives will be questioned behind the rationale and methodology used to create targets, in order to provide assurance that these are well calculated, and appropriate.

The purpose of this project is to provide two case study universities with recommendations of how to set carbon emission targets for the coming year, based on benchmarking of other Universities carbon emissions, across the UK.

To do this, we will need to create a model which can predict carbon emissions output, based on publicly available sector data, on historical carbon outputs, and other data fields which may influence this.

Data

Each year, UK Universities have to return certain data sets to the regulatory body, in order to meet conditions of registration as a higher education body. These data sets are then analysed, and certain fields are made publicly available via HESA – the Higher Education Statistics Agency.

This data undergoes rigorous internal checks prior to submission to the sector body, and is then checked again to provide assurance on the accuracy, and clarity of the data submitted. This means that should our model work, we can be confident in the legitimacy of the data used to train it.

The data we identified for our variables are:

- Scope 1 and 2 Carbon emissions kg/Co2e (y variable)
- Gross Internal Area (GIA)
- Staff Number

- Student Number

The variables we want to look at are split across three of these statutory returns: The Estates Maintenance Record (EMR), the student record and the staff record. All three of the records contain the UKPRN and the provider name, which can be used to merge the tables. The data locations are evidence in the table below.

Data Source	UKPRN (unique provider number)	Provider Name	Student Number	Staff Number	Gross Internal Area	Scope 1 and 2 Carbon Emissions kg/Co2e
EMR Record	x	x			x	x
Student Record	x	x	x			
Staff Record	x	x		x		

So that we can get the data in the format that we need to do the tests, we need to cleanse it through the following steps:

- Filtering out any academic years that are non 2018/19, so we're only working with the most recent data.
- Filtering out any data that is irrelevant for the student numbers, so that we only have the total figures – the records also contain further breakdown by course type, mode of study etc.
- Filtering out any data that is irrelevant for the staff numbers, so that we only have total FTE equivalent figures – the records contain further breakdown by contractual arrangement, headcount etc.
- Filtering out any data in the EMR record that is not needed, for example the record also contains scope 3 carbon emissions, which we have chosen not to focus on due to institutional variance in how these figures are reported.

We then needed to correct any data fields that were not correctly configured for our aims (ie. Changing strings to floats), and merge these different datasets into one, so that we ended up with a dataframe, with the following attributes:

```
UKPRN                      float64
Provider                    object
Student_Number               int64
Staff_Number                 int64
GIA                        float64
Carbon_Emissions_Kg_CO2e   float64
Longitude                  float64
Latitude                   float64
dtype: object
```

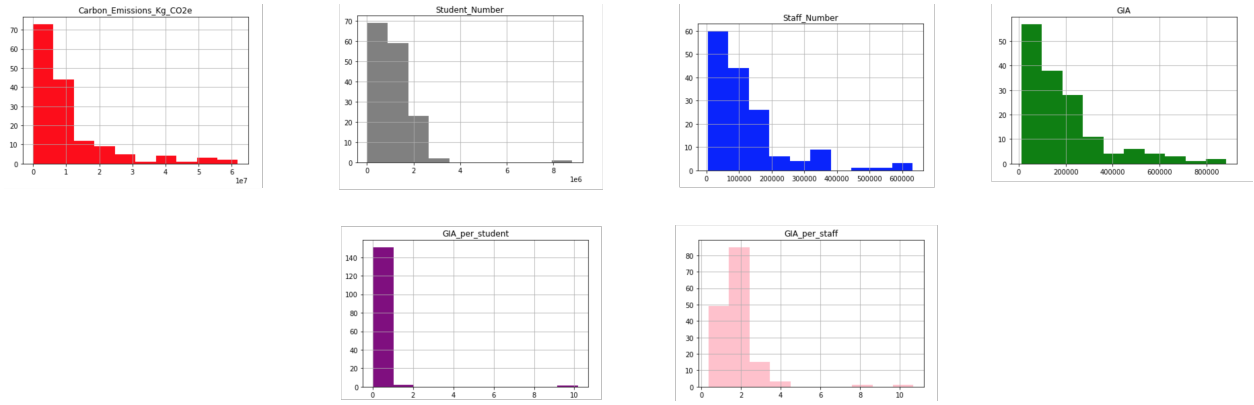
We will use the foursquare API to get information on how many venues are in the immediate proximity of the University. This will then be used as our 'urbanisation metric', to indicate how built up an area is.

We also need location data for the UK providers to allow us to map the institutions. This can be found at http://learning-provider.data.ac.uk/

Methodology

Understanding the Data

The first thing we needed to do was create a base layer of understanding for our data, so that we could decide how to progress our work. To do this, we created histograms to show how the data varied, shown below.



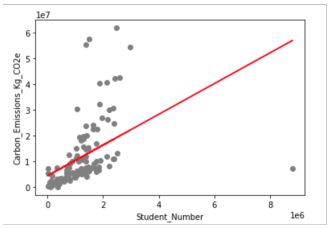
Looking at our data, we can see that there is a spread of data across our variables, though this tends to slope off at the higher ends of the scale. We thought it would be worthwhile exploring the effect of a couple of 'manufactured variables' which we got by manipulating the data. These were GIA per student, and GIA per staff.

We also mapped the providers to visualise how they were split across the UK. To do this we created a base map of the UK, and defined a function to map each unique provider as a separate point on top of this.

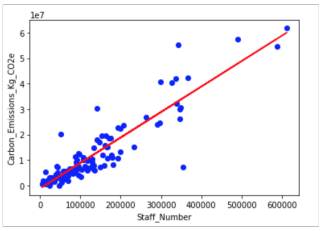
Linear Regression models

We then needed to analyse the relationship between our separate variables (staff number, student number, GIA, GIA per student and GIA per staff), and our target variable (carbon emissions). To do this, we ran separate linear regression tests, evaluation these based on the strength of the r2 value. The results of these are shown here:

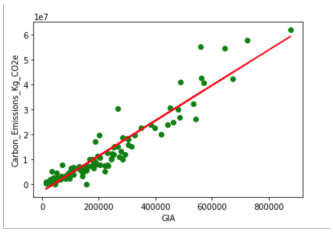
0.36
R2

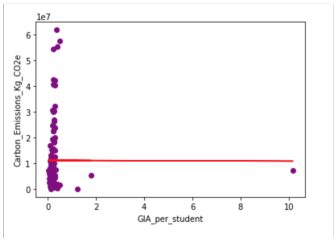


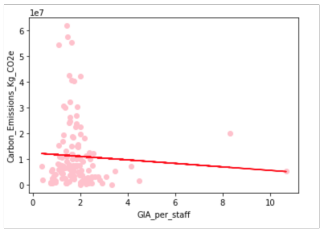
0.87
R2



0.94
R2







As we can see, there is a clear correlation between staff number, student number, and GIA on carbon emissions, with the relationships between student number and GIA being the strongest out of these. However, for our calculated metrics, GIA per student, and GIA per staff, there was no evident correlation. So we'll drop these from further investigation.

We then went on to use these variables in a multiple regression model, to assess whether using the three variables together could create an accurate model, that was better informed than if we used a single linear regression model. An alternative to this would have been to use the three linear regression models separately, and then calculate the mean of the three predictions to provide our estimated carbon target. Our multiple regression model provided a variance score of 0.93, so we can be confident that the model is strong. Therefore, we decided to opt for the MLR model to create a more informed prediction for our case study universities.

Clustering the data to create benchmark groups

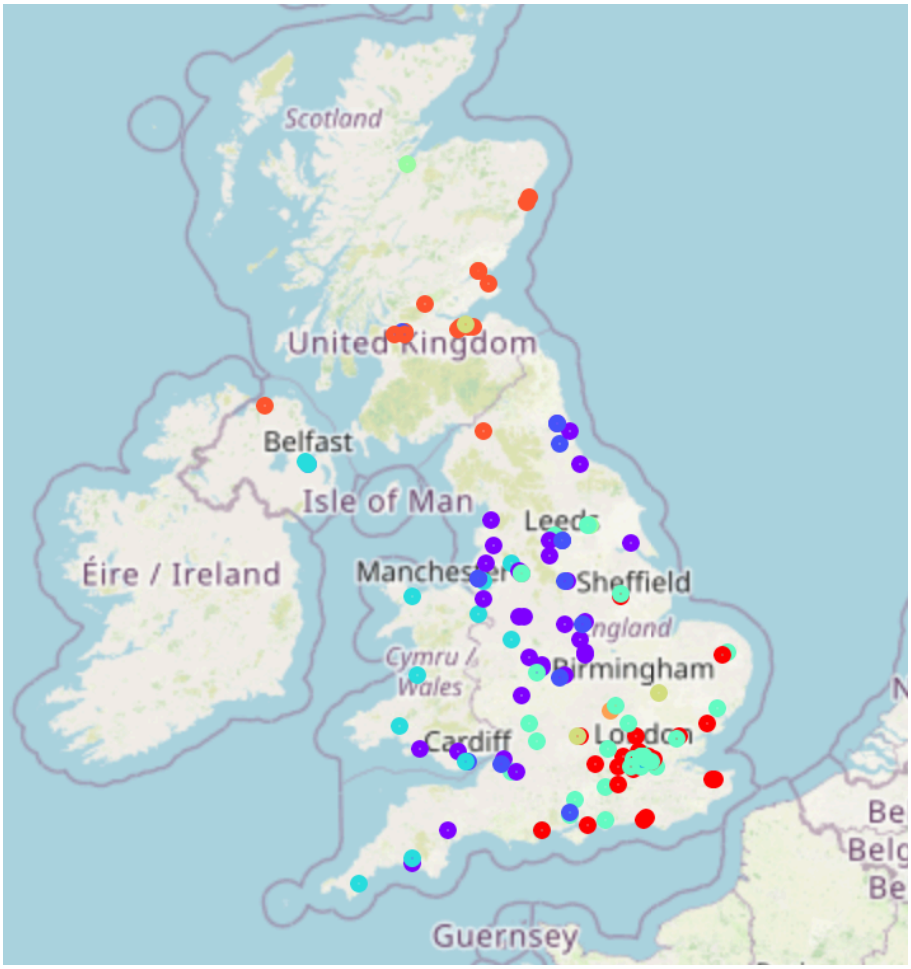
In order to provide stakeholders with a picture of how the projected estimate carbon emissions fit within the sector, we decided to create benchmark groups of Universities, which would offer a range of carbon targets to aim for, based on the minimum and maximum emissions within the separate groups.

To do this, we used K-means clustering. We then allocated each institution a cluster. After reviewing the make up of these clusters, we could separate these into different groups, based on commonalities in our data, and sector knowledge of the providers in each. These groups were:

0	Teaching Universities 1
1	Holding Group University
2	Research Giants
3	Teaching Universities 2
4	Oddities
5	Teaching Universities 3
6	Scottish Universities
7	The Open University
8	The Russell Group
9	Specialists and Conservatoires

To allow us to allocate the predictions for our Universities to relevant groups, we calculated the mean of each group, to create a figure, where we could assign the predictions to the group with the most similar carbon emissions.

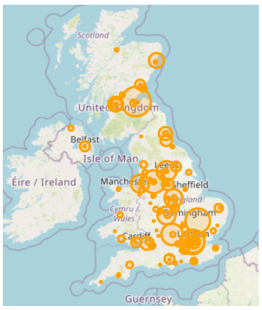
We also then mapped these clusters (using the function mentioned above), to analyse how clusters were split across the UK. This is shown below.



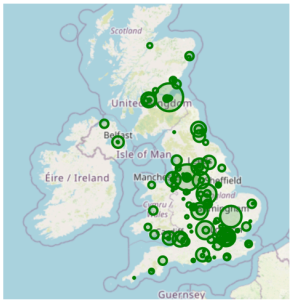
Mapping the results

Using the function mentioned above, we mapped the Universities to visualise regional variations in the data across the UK. We did this by altering the radius of the points, so that it was dependent on the variable we were mapping, to create a bubble plot. These maps are shown below.

Carbon Emissions

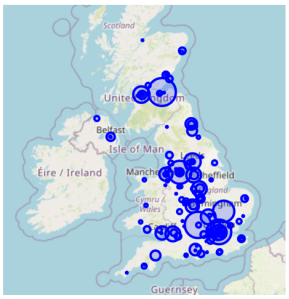


GIA



Student Number

Staff Number

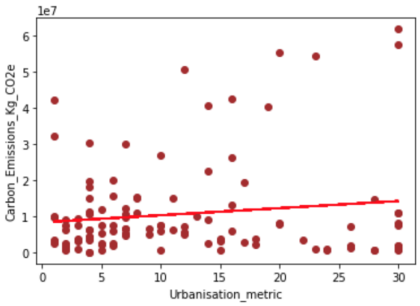


From this we identified the following regional variations in our data.

Urbanisation Metrics

We wanted to test if the urbanisation (how built up the area its located in is) of a University affects carbon emission output. In order to do this, we used the foursqaure API, to explore the immediate proximity, and count the nearby venues. We had to iterate this process a few times, varying the radius to explore until we got a relatively even split of figures across the University. We then used this figure for each university as the urbanisation metric.

In order to incorporate this metric into our prediction, we conducted a single linear regression test, against the carbon emissions. This is shown below. However, this didn't provide us with a clear correlation, or a strong R^2 number.



We then factored this metric into our multiple linear regression model from earlier (again, shown below). This caused our correlation to be slightly weaker, and the variance score was slightly less (0.84), indicating that the overall accuracy of the model dropped slightly. However, because the variance score was still relatively high, we decided to use this model to predict carbon emissions output, as it was informed by a wider variety of data, and it would model predictions on factors that otherwise may remain unaddressed.

Summary

By conducting these varying tests and visualisations, we created a model that could predict carbon emissions based on a Universities student number, staff number, GIA and urbanisation metric. Once we had predicted this, we could then compare this with our benchmark groups((created by K-means clusters), to see how the prediction looked against other similair institutions. We could then use this information to offer a ranged target for carbon emissions for the University. Using the regional variances identified through our mapping, and our urbanisation metrics, we could offer further suggestions on which end of the scale the University should be setting its carbon emissions targets at.

Results

University 1

Based on our model we can draw the following results:

The model predicts that the University would produce 10413210.43849994 Kg/CO2e Carbon Emissions. This falls within cluster 1 of our groups - which ranges from 3229459.5110000004 to 19674461.337 kg/CO2e Carbon Emissions. These Universities generally are within metropolitan non-coastal areas, which is representative of our chosen case study.

As a result, we make the following recommendation:

The predicted carbon output is in the middle of the range to aim for, and based on the likeliness to other Universities in the cluster, is a realistic target. The University should aim to record this amount of scope 1 and 2 carbon emissions at a maximum, and aim lower where possible.

University 2

Based on our model we can draw the following results:

The model predicts that the University would produce 595366.28194308 Kg/CO2e Carbon Emissions. This falls within cluster 5 of our groups - which ranges from 456340.662 to 7656865.655 Kg/CO2e. These Universities are generally more rural areas some of which are coastal.

As a result, we make the following recommendation:

The predicted carbon output is at the lower end of the range to aim for, and based on the likeliness to other Universities in the cluster, and the case studies position as a rural, coastal campus, this is a realistic target. The University should aim to record this amount of scope 1 and 2 carbon emissions at a maximum, and aim lower where possible.

Discussion

Benefits

A high variance score means we can be confident in the accuracy of the model. Despite the variance score dropping as a result of including the urbanisation metric, this change was not to dramatic, and our stakeholders would prefer a balanced trade off between an accurate and informed model (which this allowed us to achieve)

Clustered groups allows us to provide a range to stakeholders for target setting. By looking at the minimum and maximum of these sector specific groups, we could analyse how the prediction fit in relation to the sector. This would be especially helpful for our stakeholders, who like to position strategic initiatives on sector competition.

Recommendations can be made to steer direction within this range, based on regional variances identified through our mapping. This allowed us to create contextualised recommendations, providing our stakeholder with better illustrated predictions.

The visualisation of this data helps stakeholders who may not be as data literate.

Limitations

This report is based on 2018/19 data due to a data publishing lag from HESA.

At the time of writing this is 2 years out of date. As this is an area that should be rapidly changing, this data is not the best to use. To remedy this, we could create a trend metric to try and predict change over these 2 years and then build from this. W

We used the total scope 1 and 2 emissions as one of our variables, to increase the accuracy of our model, we could have looked at contributions to this (e.g. carbon emissions from

transport) to create a base variable that we know effects our target. This would have made our model more reliable

We allocated clusters to our predictions visually, based on the nearness of the prediction to the mean of one group. This could have been strengthened by looking at classification strategies such as K nearest Neighbour.

Barriers

The Open University influenced the strength of some results, particularly on student numbers, due to its unique operating model of very high student numbers, and very low GIA. If we were to run the project again, we would filter out the Open University, on the basis that its unique position in the sector does not offer a fair comparison on how to set target practice.

The foursquare API can only call information within 100,000 metres, and has a maximum count of 30. We originally wanted to use the ONS definition of a built up area, which has a radius of 200,000 miles. Because of these two limitations, we had to iterate our urbanisation metric, until we got a visually fair split.

We only counted for scope 1 and 2 emissions. There are difficulties across the sector in recording scope 3, so we can't trust the accuracy of available data.

Conclusion

Our original aim was to provide two case study universities with recommendations of how to set carbon emission targets for the coming year, based on benchmarking of other Universities carbon emissions, across the UK.

To do this we created a model which predicts carbon emissions output, based on publicly available sector data, such as staff numbers, student numbers and urbanisation, on historical carbon outputs, and regional variances which may influence this, such as urbanisation.

We can be confident in the accuracy of our model, and that any supplementary recommendations we made to stakeholders were robust and defensible.