IBM Data Science Capstone Report (Background and Data)

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

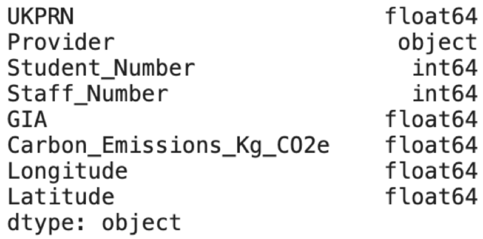
The variables we want to look at are split across three of these statutory returns: The Estates Maintenance Record (EMR), the student record and the staff record. All three of the records contain the UKPRN and the provider name, which can be used to merge the tables. The data locations are evidence in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