**Data Science Campus : Building data science capability across government**

Alexis Fernquest, Gareth L. Jones; Data analytics apprentices; ONS Data Science Campus

**Abstract**

The third of three presentations from DSC: With the opening of the Data Science Campus ONS has established a major hub for data science capability with Government, and this talk will overview the ways that the Campus and GSS Learning Academy are helping build capacity. The talk will also present projects carried out by the Campus apprentice group, working on the UK's first data analytics apprenticeship scheme.

Key Words: Key word 1; Key word 2; Key word 3; Key word 4; Key word 5

Contact Details for Alexis Fernquest:

Telephone: 01633 45 5651

Email: Alexis.Fernquest@ons.gov.uk

Contact Details for Gareth L. Jones:

Telephone: 01633 45 6896

Email: [Gareth.L.Jones@ons.gov.uk](mailto:Gareth.L.Jones@ons.gov.uk)

**Presentation**

The first slide of the presentation was used to discuss the background on the apprentices who were presenting and to welcome the audience. Gareth, is not how you would say, a typical apprentice, in his 30’s, he is hardly the school leaver you would expect to see. Gareth has had a rich amount of experience in the law and conveyancing industry, where he took on managerial duties. Prior to this, Gareth tried the degree route, but felt it wasn’t really for him so ended up joining the work market. He has always been interested in data and how it can be used, but there were never any opportunities around for him prior to this being advertised to take this further. Gareth used this apprenticeship as a way to switch careers into an area that had always fascinated him. The other apprentice was Alexis. Alexis, who prefers to go by Lex, completed a degree in economics in Swansea University, and has always been interested in trends and numbers and the story they can tell. Upon finishing her degree, she started work in Lloyds Banking group and was drawn to analytical roles. Her previous roles have meant that she has a background in budgeting, resourcing and searching for trends in data, after 5 years in these roles she decided she would like to know more, and often wondered about going to get a masters in this area, however the idea of working part time to go back to school didn’t financially make sense. When she saw the data analytics apprenticeship advertised, this seemed like a fantastic opportunity to learn more in this area, whilst working for a reputable company like ONS. Lex explained that she’d always had a “learn by doing” method of learning and being able to put skills into practise in a work environment, ensures that the skills she have learnt so far have really embedded themselves.

After the introduction Lex went on to explain about the apprenticeship and how it works. The data analytics apprenticeship is the first of its kind in the UK. There were over 130 applications for the first cohort, where there were 8 people successful in the role into the data science campus on December 2016. Since then there has been a further 6 employed across welsh government and a further 6 due to start in the data science campus in September 2017.

The apprenticeship is effectively broken down into three parts. Part one being external courses, part two as on the job training and part three is evidence that is provided to the training providers as part of the qualification assessments. In regards to “on the job training” this is training that has been provide by ONS outside of the qualifications official training. So this includes things like R and Python application and further training, regression modelling, GitHub training, data ethics training as well as an over view of work across ONS business areas, which is where the apprentices will be going from September 2017. They have also been involved in real life project participation across ONS and other government departments.

Since starting, the apprentices have been involved in a variety of projects. The first being a practise and dummy project in order to let the apprentices get used to sources, using and analysing data. They decided to look into the sustainable development goals, with the aim to find public data and see what analysis they could provide from it. They split into two groups of 4, the first group looking at goal 7 – sustainable energy sources, looking specifically into energy demand and impacts, and the second group looked at goal 3 – non communicable diseases – and specifically looked at lung cancer and mortality rates.

Group one who were looking at clean energy, first decided to start and understand what the trend in energy was. So they applied an ARIMA forecasting model in R to have a look at the future trend. The data was taken from gridwatch which records energy usage every 20 minutes. From this they were able to spot there was a downward trend in energy consumption, which they felt was unusual due to the growing population, so they decided to look into what some of the reasons behind this may be. One of the areas they looked into was smart meter installations, so they wanted to see if there was a relationship between smart meters being installed and a reduction in energy usage – so were people becoming more “smart” with their energy use? They found there was a link, in that as there has been an increase in smart meters, energy has gone down. However, it is worth noting, that the group are aware that smart meter installations isn’t the only reason behind this trend, however as a practise project it was a good way to get used to going analysis in R and understanding the results.

Group 2 decided to focus on sustainable development goal 3, Good Health and Wellbeing and the indicator by 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being. This indicator was chosen for this training exercise as it had potential for us to develop exploratory analysis techniques as well as element of forecasting. In addition to the opportunity to develop skills this topic also gave us an opportunity to liaise with internal ONS departments. A decision was made to focus on a single non communicable disease and in our case we chose Lung Cancer. The Mortality Team within the ONS Life Events Department holds mortality information which were able to liaise with and obtain a subset relating to lung cancer deaths. Our Exploratory analysis centered around Age, Gender and location of deaths and the slide pack contains our findings with regards to gender. Our next step was to look at forecasting techniques that could be applied to the 10 year lung cancer mortality dataset obtained from the mortality team. We settled on the ARIMA model and proceeded to produce a 24month forecast to see if there were any continuing trends. Within the slide pack you can find the decomposition and forecast obtained. Both the exploratory and time-series analysis were carried out using R and the package Forecast and GGPLOT2 for visuals.

Following completion of the training projects the apprentices were each allocated to a Data scientist within the Campus. The data scientist would both line manage and mentor and the apprentices would now work on live projects.

One project the apprentices have worked on has been in collaboration with the Cabinet Office in obtaining classified personas from the survey data. The aim of the project was to see if personas can be developed from a subset of survey data relating to s specific economic activity. For this initial stage of the project we decided to use the Labour Force Survey (LFS) and to apply clustering techniques with the aim to classify personas. We decided upon the LFS as the this data set contained variables relating to the respondent and their household which believed would provided valuable insight into economic activity we were focused around. Our first challenged was how we were going to deal with the mix of data types which is more commonly found in survey data. To tackle this problem we used a package with R called PCAMix. This package separates the quantitative and qualitative data then carried out Principle Component Analysis on the quantitative and Multiple Component Analysis on the qualitative data. The package then combines the results of this analysis into a scores matrix representing influence each variable and observation has on the principle component. It is this numerical scores matrix which is then passed to t he clustering techniques which in this case was KMeans. Using the elbow method we were able to cluster the data into four, map back the clusters assignments to the original dataset which thus allowed us to carry out exploratory analysis on each cluster subset. The personas which were developed as expected, for example one cluster contained respondents who were married couples aged 30 – 40 with children owning their home by means of mortgage, whilst a second cluster contained married couples, with children not living at home owning their house out right. From this we determined the results to be positive however the personas were still quite generic.

The next steps in this project are to increase the data set used in the clustering. A challenge we had is that survey data contains quite a large proportion of null values where clustering does not deal well with this. For this first stage we decided to take the brute option and subset our data to complete records. This considerably reduced the data set that was available for clustering. In phase two of this project we will look into ways of dealing with these null values in order to increase the dataset with a view to obtain further refined clusters.

One of the other projects that the apprentices have been working on is the Question bank. Four of the apprentices have been working on the Question bank since April 2017. The projects aim is to capture all the business surveys in a machine readable format so that they can perform text analytics and analysis on them later on. The aim is to have all these stored digitally in one place. The benefits of this project is to provide harmonisation across departments, as well as being able to link these questions to admin data and to offer a resource to the wider ONS. The surveys have now all been scraped into a JSON format and the next step is to import these into python. Once this has been completed they can carry out cleaning on the data in order to do text analytics and visualise these outputs.

So we have discussed the projects we have been working on, and felt it made sense to reflect on some of the skills we have learnt and put into practise here, such as programming in R and python by cleaning data, analysing it and presenting it. We have also had exposure to stakeholder management, report writing and web scraping just to mention a few.

Throughout this presentation, the projects the apprentices have been working on and how the apprenticeship works has been discussed, but it is also worth talking about some other initiatives the apprentices have been involved in. Four of the apprentices are currently STEM (science, engineering, technology and maths) ambassadors, part of this role has been to go out to local schools and give talks on the apprenticeship to raise awareness of it and to give options to people of school age. They have also been involved in data science awareness days with schools, so making them aware of what data science is and why it’s such an interesting field to go into. They have also been involved in blog posts as part of national apprenticeship week, and monthly blogs to discuss their projects on the data science campus blog page. They were also involved in the launch of the data science campus in March 2017, part of this was showcasing their work and meeting with internal and external stakeholders to discuss the apprenticeship and the work they had been involved in. They were also involved in intruder testing, where they could put their hacking skills to use inside the VML and report back any findings, to ensure data that is published is safe and secure. They have also been involved in charitable endeavours – such as data dive, which took part in May 2017 in ONS, where they helped local charities to do analysis and provide results of those analysis to them for free.

The aim of the presentation was to show how the apprenticeship contributes towards building capability of data science within government. What we have shown is that in a short period of time Data Analytics Apprentices have been able to provide value towards ongoing projects whilst developing a new skillet within government working. The next stage of the apprenticeship is to go out into wider ONS department. There the apprentice will apply their newly gained skills on various projects whilst help facilitate data science learning within the ONS. The main points we would like you to take form this presentation is that capability in data science can been developed swiftly, value can be added to ongoing processes at speed and through collaborative working capability in data science can be developed from within.

This presentation has mostly been about the apprenticeship, as that is where Gareth and Lex’s area of expertise is, however there are many other pathways available through the learning academy in ONS, so if you would like more information please contact them on [Learning.Academy@ons.gov.uk](mailto:Learning.Academy@ons.gov.uk).