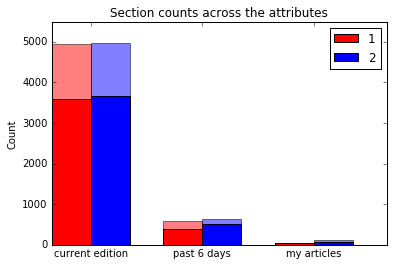
## News UK task feedback

**DATA SCIENCE**

Log data has 1161 unique uids, 827 of which have a known attribute. This means that 1053 of the users in the demographic data are not represented in our data.

I construct a left join of log data with the demographic data on uid to produce a merged data set where all 11309 entries have an attribute of either 1,2 or NAN. This is saved as: 'ds\_test\_log+demog\_out.csv'.

Our train data set can be attained by grouping the merged data set by attribute and selecting those with either 1 or 2. From this we can implement a number of methods for predicting those entries where attribute is NAN.



Methods investigated:

* **Radom forest across UID, Date, Section, Page** – This method was dismissed for a number of reasons. Firstly, the random forest methods I am familiar with cannot handle categorical data. Whilst it would be possible to reframe each of the 152 page entries as an integer, this would quickly result in a loss of meaning as each page is not related directly to another. In addition to this, the UID and Date features are essentially an index and are therefore not correlated with the attribute score, or in any way qualitatively related to demographic data.
* **Correlation with section – as** the following bar chart shows, the overwhelming majority of section types are ‘current edition’, and there is no obvious preference for attributes in any of the three section types. I therefore conclude that this feature will not influence the attribute score.
* **Aggregating Page type –** as the simplest method of predicting attribute type, per uid, I look only at the ‘Page’ data. There are 152 page types. I find the modal attribute for each page type, and assign all future pages of that type, that attribute. I save this data set as ‘ds\_test\_log+demog\_aggregate\_predictions.csv’. When the predictions are added to the known attributes (transparent compared to opaque bars), they fit with the pre-existing trends.
* **Natural Language Processing analysis of Pages –** The page contains natural language data that provides more information that could be used to classify the UID demographic attribute. I have no experience of implementing this kind of analysis and therefore any discussion of these methods is purely qualitative. A typical page has the following format: ‘article:hatred of israel and jews can’t be separated’. Statements like this could be assigned to a subject topic, such as ‘politics’ or ‘sports’ though on the face of it this would require some level of human judgement. Alternatively, all words associated with an attribute could be clustered to pull out common themes, though this would be dominated by meaningless articles and lose any structure in the page statement. Given there are only two attribute options it is unlikely that clustering will produce a clear result. Many pages are laced with sentiment, and therefore a sentiment analysis could be run of individual pages, and clusters, though I suspect that there is little link between sentiment and demography.

Testing:

I put together a rough metric for comparing the ratio of attribute 1 counts to attribute 2 counts from the training data and the predictions, for each of the three ‘sections’. Results are as follows.

Attribute count ratio for current edition from train = 0.98 and prediction = 1.03

Attribute count ratio for past 6 days from train = 0.75 and prediction = 1.57

Attribute count ratio for my articles from train = 0.73 and prediction = 0.0

Bearing in mind that 88% of the data belongs to section ‘current edition’, it is reassuring that the training set and predictions ratios are the closest match at 0.98 and 1.03. There is a wide discrepancy between ‘past 6 days’ with ratios separated by over a factor of two. Results from ‘my articles’ is influenced by the lack of unknown users. As these data are consistent with pages that are not labelled as ‘article: they can be considered as a minority special case that is unlikely to fit the overall trends.

Based on this limited assessment, aggregating existing attributes as a prediction is a quick and simple way of completing the data set and the results are consistent with the patterns observed in the existing data.

**EDITORIAL TEAM**

* Summarise the problem. How we are missing demographic data for 29% of our users.
* Summarise the method. How we can relate the known demographic data of some users to our log data in order to predict demographic traits of the missing users.
* Summarise the solution. Empathise the simplicity of the solution, looking at the all the known users who viewed a given article and assigning the most common demographic traits to those unknown user who viewed the same article.
* Summarise Testing. Known users have a known distribution across the three article sections. Our predictions match these observations, confirming that they are reliable (reference above plot at this point). Stress we most confidence in ‘current edition’ users (which make up 88% of all the data). ‘My article’ users should be considered a special case and removed from this analysis.