Sebastián Torres

Machine Learning Course ML01

Week 3 – Preprocess data

After loading the data, we first delete empty rows and columns. The last 5 columns are removed, as well as the ‘Name (Cand ID)’ Column, as it just seems to number each row. The last row is either empty or just corrupted, so it is deleted as well.

The names of the columns are changed to something easier to work with.

The ‘Date’ column has close to 100 unique points, and the data was taken for different interviews over the course of two years. This data doesn’t seem too relevant, so it will be removed.

The skillset data seems too dirty. Some points intersect partly, and some (such as ’11.30 AM) seem to just be noise. The column will be dropped as well.

The ‘Location’ and ‘Current Location’ columns are duplicates of each other. I will remove the ‘Current Location’ column, and keep ‘Location’

The native location will be removed, given the fact that there are two other location columns, more relevant to the actual information of the interviewee.

Next, I will clean up the data by merging similar points into the same category. NA has been grouped into its own category, given the large amount of NA data. The pattern of this data when plotted against the ‘Expected’ values mostly lines up with the negative values, however it could be argued that a ‘No answer’ category could be relevant for the problem. The ‘Expected’ column will have three possible points: Yes, No, and Uncertain.

The data for each column was then examined in comparison with the ‘Expected’ column, in order to find an appropriate way to fill null cells.

Finally, One-hot encoding is applied to all non-numeric columns, to obtain a usable set of data, and the response variable from the ‘Observed attendance’ column is separated from the rest of the values.