CS7DS3 Main Assignment

This assignment will consist of a report, with a maximum length of 20 pages. The assignment deadline is 5p.m. on Friday April 5th, 2019.

The assignment is worth 70% of the module mark, and will be evaluated under the following criteria:

- Data management (10%)
- Clarity of writing and exposition (10%)
- Modelling (60%)
- Creativity (20%)

I would like you to analyse the Yelp academic dataset: https://www.yelp.com/dataset. Specifically, you should focus your analysis on restaurant businesses in the city of Toronto. In the report I would like you to address the following questions:

1. I want you to compare the ratings of currently open Indian restaurants in the neighbourhoods of Scarborough and Etobicoke. Which neighborhood is best for this kind of food? How much better? Compare the ratings of (open) restaurants across multiple different neighbourhoods in the city. Are any neighbourhoods clearly superior to others? If so, by how much?
2. What are the factors most strongly associated with restaurants being closed? How accurately can you predict when a restaurant in the dataset will be closed?
3. Restaurants are organised by neighbourhood. There are a lot of neighborhoods in the data. Using the longitude and latitude of each restaurant, can you find other ways to organise restaurants together, e.g., using a clustering model? Can you find any interesting associations with other elements of the data using this clustering?

If you have any questions, contact me on arwhite@tcd.ie.