Applied Data Science Project Presentation

Comparing areas in New York and London.

Finding a borough in London which is like Manhattan and is close to Google HQ offices near Kings Cross, London

Finding a borough in London which is like Manhattan and is close to Google HQ offices near Kings Cross, London

- A Data Scientist who currently works a Google's offices in New York has been asked to relocate to London

- They currently live in Manhattan, New York and want to live in an area of London which is similar to Manhattan when they relocate

- They also want to want to live in an area which is close to Google HQ offices near Kings Cross, London

Data acquisition and cleaning

- New York Neighbourhood data used in previous New York example. 306 Neighbourhoods in 5 Boroughs https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
- London Borough data scraped from Wikipedia. 32 Boroughs https://en.wikipedia.org/wiki/List_of_London_boroughs
- Venue type data scraped from Foursquare API. New York has 10106 venues and 428 venue types. London has 2468 venues and 263 venue types
- Venue types encoded for combined London and New York data, 468 venue types in combined data to ensure common features
- Proportion of venue types in area form features:
  - New York neighbourhood data 306 rows and 468 features.
  - London borough data 32 rows and 468 features

Data splitting and machine learning methods

Given that decide what New York Borough a London Borough was most similar to is a classification problem I chose to use a KNN Classifier, and the hyper-parameter k required establishing using a cross-validation dataset.

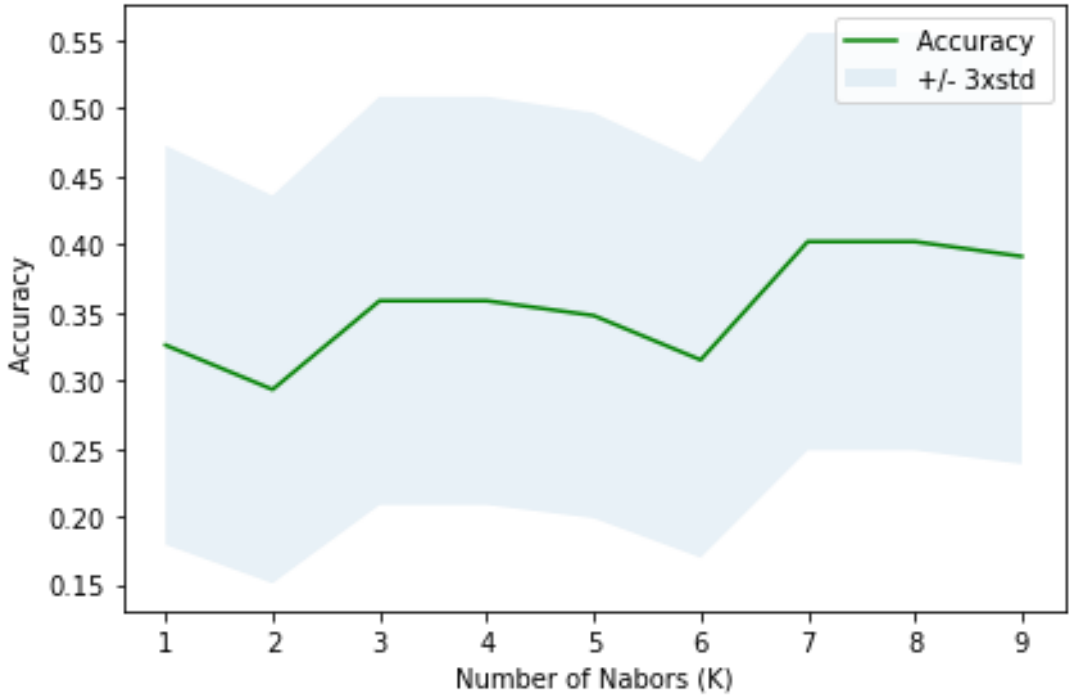
The New York data was split into a train (70%) and cross-validation set (30%)

The data as follows:

- X_train – feature data only from 70% of New York Neighbourhoods. Used for model fitting
- y_train – the Borough data from 70% of New York Neighbourhoods. Used for model fitting
- X_cv – feature data only from 30% of New York Neighbourhoods. Used for establishing best k
- y_cv – the Borough data from 30% of New York Neighbourhoods. Used for establishing best k
- X_test – feature data of all London Boroughs. Used to predict most similar New York Borough

Establishing parameter k for KNN Classifier

- The cross-validation New York data set was used to establish the best hyper-parameter k.
- A KNN Classifier model was constructed for various k values then trained using X_train and y_train.
- The accuracy of each model was then established by using the X_cv dataset to predict which Borough in New York each neighbourhood belong to.
- This was then compared to the actual Borough it was in using y_cv dataset.
- K = 7 produced the most accurate model, with a model accuracy of 40.2% on cross-validation data.
- It is worth noting that this level of accuracy is relatively poor – randomly guessing Borough would have an expected accuracy of 20%.



Using KNN Classifier model on London feature data to find most similar NY Borough

- The KNN Classifier with k=7 was used to classify each London Borough as the New York Borough it was most similar
- All but two boroughs (Barnet and Hillingdon) were most similar to Manhattan.
- This didn’t help much to reduce the Boroughs to consider for the Data Scientist to live in when they relocate to London.
- It also suggests that most London Boroughs are most similar to Manhattan.
- Alternatively, this could be due to overfitting caused dimensionality of training data as the number of features was larger than the number of examples.

	Borough	Latitude	Longitude	NY Similar Borough
0	Barking and Dagenham	51.5607	0.1557	Manhattan
1	Barnet	51.6252	-0.1517	Brooklyn
2	Bexley	51.4549	0.1505	Manhattan
3	Brent	51.5588	-0.2817	Manhattan
4	Bromley	51.4039	0.0198	Manhattan
5	Camden	51.5290	-0.1255	Manhattan
6	Croydon	51.3714	-0.0977	Manhattan
7	Ealing	51.5130	-0.3089	Manhattan
8	Enfield	51.6538	-0.0799	Manhattan
9	Greenwich	51.4892	0.0648	Manhattan
10	Hackney	51.5450	-0.0553	Manhattan
11	Hammersmith and Fulham	51.4927	-0.2339	Manhattan
12	Haringey	51.6000	-0.1119	Manhattan
13	Harrow	51.5898	-0.3346	Manhattan
14	Havering	51.5812	0.1837	Manhattan
15	Hillingdon	51.5441	-0.4760	Brooklyn
16	Hounslow	51.4746	-0.3680	Manhattan
17	Islington	51.5416	-0.1022	Manhattan
18	Kensington and Chelsea	51.5020	-0.1947	Manhattan
19	Kingston upon Thames	51.4085	-0.3064	Manhattan
20	Lambeth	51.4607	-0.1163	Manhattan
21	Lewisham	51.4452	-0.0209	Manhattan
22	Merton	51.4014	-0.1958	Manhattan
23	Newham	51.5077	0.0469	Manhattan
24	Redbridge	51.5590	0.0741	Manhattan
25	Richmond upon Thames	51.4479	-0.3260	Manhattan
26	Southwark	51.5035	-0.0804	Manhattan
27	Sutton	51.3618	-0.1945	Manhattan
28	Tower Hamlets	51.5099	-0.0059	Manhattan
29	Waltham Forest	51.5908	-0.0134	Manhattan
30	Wandsworth	51.4567	-0.1910	Manhattan
31	Westminster	51.4973	-0.1372	Manhattan

Establishing distance from Borough to new office location

After reducing the London data set to Boroughs which are similar to Manhattan, using the longitude and latitude to calculate the distance to the office and then sorting by distance office revealed the top 3 locations to consider living in when relocating to London. Camden, Islington or Westminster.

	Borough	Latitude	Longitude	NY Similar Borough	Distance to office (km)
4	Camden	51.5290	-0.1255	Manhattan	0.475216
15	Islington	51.5416	-0.1022	Manhattan	1.890084
29	Westminster	51.4973	-0.1372	Manhattan	4.074319

Conclusion and future directions

- Built New York Borough classifier given area features using K Nearest Neighbour Classifier and k=7 (found using cross-validation dataset)
- Relatively poor 40.2% accuracy on cross-validation dataset as random guessing would give expected accuracy of 20% accuracy
- Using classifier all boroughs (except Barnet and Hillingdon) in London are similar to Manhattan, which did little to reduce possible areas. Meaning top 3 areas to consider driven by distance from office.
- Model may have been overfit due to the curse of dimensionality
- Improvements in future models may include reducing the number of features using a method such as Principle Component Analysis (PCA)