Data Science Capstone Report

Introduction

Given the nature of London, the capital city of England, and its ever increasing density of population and general busyness, there appear to be few possible moments of genuine peace, tranquillity, and calm in the outdoor environment. So often, the man-made skyscrapers, buildings, and general infrastructure contrasts the historical living conditions for people in the United Kingdom, and expensive weekend holidays are consequently taken to visit what some may refer to as 'the great outdoors'. However, to take advantage of the occasional bubbles of parks that are offered in London, most people visit these green areas as an opportunity to walk through greenery to be in the fresh air, escape the grey architecture of the city, and often peruse these areas at lunch breaks during the working week, whilst still remaining in London. Parks act as an opportunity to escape urban life.

For this project, we will take the hypothetical situation where a business owner wants to select an area in London for their new office, where the number of nearby high-quality parks will be the swaying factor in their decision making process.

The CEO wants to provide staff with numerous opportunities to improve wellbeing via facilitating the ability to visit various parks during lunch breaks. Would it be preferential for the business owner to invest in a property in North or South London to maximise the ability for his happy employees to visit as many parks as possible during lunch breaks? Additionally, which parks would the employees derive most satisfaction from? Consequentially, in which area should the company look to purchase their new office space?

Data

To help answer these questions in an effective manner, we would need to gather data indicating the difference between North and South London, the neighbourhoods/boroughs and parks situated in North and South London as well as their latitude and longitude coordinates, and also data that relates specifically to the given parks (venue data).

We will be able to utilise Foursquare API to acquire data, which will provide coordinates and venue data, from which it will be possible to slice, dice, drill, drop, cleanse, and conquer the data, ultimately using it to integrate fact-based argumentation into our finalised recommendation.

It will be possible to visualise our data in a variety of ways, yet also harness the power of machine learning to develop clusters which will be effective for geospatial analysis.

Methodology

Having developed a search query and results to analyse the data related to parks in London, it was possible to assign relevant elements of the JSON file to a dataframe, which allowed us to more effectively approach the data for analysis. Having decided to drop columns which were not relevant for our data analysis, we were given a more condensed dataframe consisting of highly relevant data which would help contribute towards our final recommendations. Since our Foursquare API data includes all data within a specified radius containing the word ‘Park’, we would have to ensure that we filter down to only include rows which are categorised as Parks in the dataframe.

Following this, exploratory data analysis allowed us to identify the number of parks within the vicinity, yet also provided a basis for visualising the geospatial latitude and longitude data onto a map via Folium, which provided an easily accessible document for assessing the spread of data in London. Afterwards, it was possible to pull the data for ranking of the best parks in London (Hyde Park, and St James’ Park) which were ranked as 9.7 and 9.6 respectively.

In order to use an unsupervised machine learning clustering model, more specifically the Knn Clustering algorithm, our dataframe had to be condensed to include purely numerical variables, which meant that categorical variables had to be omitted prior to the algorithmic operation. It was then possible to normalise this dataset via Standard Scaler Pre-processing, and transform the data into clustered sets. The Knn algorithm then provided us with 5 clusters, from which it was possible to assign each individual row/park to their respective cluster. The data indicated that Cluster 2 contained the most parks in the vicinity, from which it was later possible to visualise the cluster on the map of London in an effective manner.

Results

The analysis and visualisation performed initially showed that the majority of parks were located in areas North of the river, therefore signifying that the wisest investment opportunity would occur in this area. Further analysis then indicated that Cluster 2 provided the highest number of Parks in the vicinity, at a number of 7, whilst other clusters were only able to provided 2 parks.

Discussion

Since Cluster 2, which is situated North of the River, had the most nearby parks, it would be recommend to approach an investment opportunity in a location within this area. Analysis has indicated that Hyde Park and St James’ Park were amongst the most highly ranked venues within our criteria, and as a result the purchase of a property near here would be thoroughly recommended, since we can expect employees to derive the most satisfaction from the most highly rated parks.

Conclusion

The CEO should invest in a property within Cluster 2 for the following reasons:

1. Cluster 2 provides the most opportunity for employees to visit parks during lunch breaks
2. The 3 highest rated parks in London are situated in Cluster 2
3. Alternative Clusters do not provide as many opportunities to visit parks, nor do they hold the status of an area containing the most highly rated parks.