Capstone Project – IBM

The Battle of Neighbourhoods

Opening of New Shopping complex in San Diego, California

Created By- Varun Gupta

June 2020

Introduction –

The IBM data science professional certificate has taken me through an amazing journey and towards the end creating our own project has helped me a lot. The problem statement is that a businessman wants to open a shopping mall in San Diego, California but he doesn't want to open it in a place where there is already a though competition and tons of shopping complex and nor he wants to open it in a place where there is no future scope for this kind of activity. Shopping complex are one stop destination for every kind of stuff whether it is clothing, footwear, theatre, or food corner. Shopping mall caters all these needs. This is one of those projects that need a large capital to build and no investor would want this to fail therefore it is quite important to predict whether this is the best place or not. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem –

The naïve approach to the problem should be opening the complex in area where there are already many successful malls, but this would defy the purpose of the problem. It is not necessary that a place having shopping areas is the only fit place to target by the businessman. The main aim is to determine which areas are commercially better than the rest of the neighbourhoods since this ensures that these areas are for business and then filtering among them which areas having low/moderate/high number of complex.

Data –

The main components required are the following data –

• The neighbourhoods' data which is largely extracted from Wikipedia sources.

• The co-ordinates which means the latitudes and longitudes of that place.

• The venues detail about each neighbourhood in our dataset.

Sources and methods to extract the data –

• The Wikipedia page (https://en.wikipedia.org/wiki/List_of_communities_and_neighborhoods_of_San_Diego) contains a list of neighbourhoods in San Diego, with a total of 120 different area. The technique used to extract data is with help of the python package Beautiful Soup.

• The latitude and longitude of a place are found with the help of the geocoder package in python. With this we were able to extract co-ordinates for every neighbourhood in dataset to feed in the foursquare API.

• After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category to help us to solve the business problem put forward.

This project is all in one in which we ourselves define the problem statement and collect the data, clean those and then after a bit of analysis we apply machine learning techniques that are efficient in figuring out patterns in our data.