**Data Science Capstone Project:**

**K-Means Clustering to predict optimal Neighborhood for Construction of Residential Estate**

By

Karan Bhowmick

INTRODUCTION:

Mumbai is the Financial Capital of India. With a population of 20.4 million, with India’s ever-increasing life rate overpopulation has become a cause for concern. In a world pestered by overpopulation, availability of accommodation becomes the prime concern. To capitalize on the ever-increasing demand of Residential Areas. Real Estate has been dubbed the one of the best investment options. This Project tackles the issue of shortlisting a neighborhood with plentiful amenities, so as to receive the best prices on Residential Buildings/Estates. The location of a Residential Estate, is the deciding factor on Profitability and Demand of ownership. Thus, it is of utmost importance that the neighborhood chosen can augment returns.

Data Science and Machine Learning are two avenues that are paving the way for corporations worldwide. Predictive Analytics is taking over all departments in the industry. From Medicine to Finance, data is being used to gain insights into market trends. It is the need of the hour to make data-driven decisions to model a business.

Using Data Analytics and Machine Learning for Predictive Analytics, we can help streamline the decision-making process.

METHODOLOGY:

Data Requirements:

* Neighborhood Name
* Latitude Longitude of the Location
* Facilities around the neighborhood
* Parks, Cafes, Shops etc.

Data Procurement:

We can use BeautifulSoup Library in Python to extract the Neighborhood names from a trusted source. I used Wikipedia to scrape data on the neighborhoods in the Mumbai region of Maharashtra.

To scrape data regarding points of interest in these neighborhoods. Latitude, Longitude, Name and Category of these POIs can be extracted using Foursquare API.

For our experiment, we first use BeautifulSoup to scrape data on the names of neighborhoods in Maharashtra, as cited by Wikipedia. We later use Foursquare API to find different venues in a radius of 1km of the area.

Feature Selection:

found that access to amenities was important for 54% of home movers, up from 39% in 2014.

Thus, the houses are in accordance with the findings, Airport and Scenic outlook make a neighborhood prone to development. They have been given priority.  
For other factors, the order of precedence is as follows:  
1) Shopping Stores  
2) Transport Hub  
3) Restaurants  
4) Green Spaces  
5) Sports and Recreation

Data Analysis:

After said data procurement. I used the Pandas Library to create a data frame for the data. I then used the geocode library to assign latitude and longitude coordinates.

Then I called Foursquare API to get data on POIs in a 1km radius of the neighborhood. Next, I cleaned and formatted the new data set with Feature Selection.

Next I calculated the venue per area in the neighborhood. I then assigned weights to the features based on data from Strutt and Paker’s website.

Next I used K-Means Clustering algorithm to develop a model for the data set based on Lifestyle Score ( weight assigned to each region ).

The model was evaluated using sklearn.metrics. The R-squared value, F1-score and Mean Square Error were calculated.

Data Visualization:

We used the Folium Library to visualize the comprehensive map of Maharashtra. Join the table with location coordinates with cluster values. We then use ‘dark blue’, ‘orchid’ and ‘crimson’ for the clusters.

RESULTS:

* Accuracy Score : 0.978
* Mean Squared Error : 0.054
* R-squared Value : 0.935

CONCLUSION:

The stretch between Kalyan and Mumbai shows great growth prospect for Real Estate Investment. These conclusions are based on availability of local amenities in the mentioned neighborhoods ( within a range of 1km ).