Itinerary Planner - Capstone Project
Rohan Shukla
August 11, 2019

1. Introduction

1.1 Background

Nowadays, we frequently see the terms 'traveler for life', 'wanderlust', etc. and a lot of people wish to step out of their boundaries and go out exploring, learning new cultures, tasting different delicacies and connecting with various people around the globe. At the same time, it is not very easy to take time out of your tight schedule and plan trips to a place and prepare your travel itinerary for your travel. The itineraries available might not always match the constraints of the travelers and people often tend to miss some places on their tour.
As a part of my IBM Capstone Project, I plan to come up with an itinerary planner for all the travelers out there based on the sightseeing locations and number of days. the This itinerary planner will try to cover as many tourist destinations as possible and and can also have special days for people who would like to take a day out for shopping or other relaxation activities.
I plan to leverage the foursquare location data to get all the details of the tourist destinations in a certain city and then plan out the entire trip based on the number of days at hand and the maximum number of places a user would like to visit in a day.
The major stakeholders of this project are the tourists and travelers who wish to visit new places from time to time and wish to have planned itineraries before they start their trip. This project will help users plan a proper trip and once they have their itineraries at hand, it's easy for them to book their rooms and hotels near these places and also can plan out their remaining time for other activities so that they can enjoy their vacation to the fullest!

1.2 Problem

Data that might contribute to determining the places of visit might include the locations, neighboring food joints, other places of interest and metrics that describe what kind of place it is. This project aims to plan out an itinerary for the tourists and unburden them with the tedious tasks of finding hotels and locations near a tourist attraction.

1.3 Interest

Obviously, everyone does plan an outing during some part of the year and to get a planned itinerary for your visit at hand is a great delight. All the travellers as well as the travel companies will be interested in this as they will be able to position themselves more accurately based on the model.

2. Data acquisition and cleaning

2.1 Data sources

I have scraped almost all the location data including the nearby attractions and other hangout points nearby with travel distances between places from the FourSquare API data.
Based on the location entered by the user, I have gathered all data of the city. This data did contain some missing values which were then modified and the entire data was cleaned.

2.2 Data cleaning

To generate an itinerary as per the user requirements, the project requires a few inputs from the user as mentioned below:

1. Location details - Name to the city to be visited and the country code where the city resides
2. Duration of trip - The total number of days on the trip
3. Max visits per day - This number of places the user wishes to visit per day
4. Off days - the number of days, the user wants to take a break for other planned activities or relaxation.

All the data collected will then be preprocessed into required formats, visualized and then used to base the results.
Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping.

2.3 Feature selection

Apart from the data provided for the user, the project will require data from the FourSquare location data API. the following data fields from the Foursquare location data will be helpful in the analysis :

1. tourist destinations - this will give data about the top locations in and around the city
2. location coordinates - this data is used to calculate the distance between places
3. location reviews - this data will be used to select top locations

The Foursquare data is also used to cluster the destinations and print the locations on the city map.

3. Exploratory Data Analysis

3.1 Calculation of target variable

List of attractions based on user rating was not a feature in the dataset, and had to be calculated. I chose to create the sorted lists based on the above mentioned parameters as the target variable. User Ratings and nearby venues were chosen out of a few metrics because it is

the most interpretable, after all, that's all we need to plan. Tried applying k-means clustering to find out whether all the locations were covered using the target variable. This suggested that the chosen metric of list of attractions, was a reasonable one.

4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict the attractions a person would like to visit. Regression models can provide additional information based on the user requirements, while classification models focus on the probabilities a person might visit. The underlying algorithms are similar between regression and classification models, but different audiences might prefer one over the other. Therefore, in this study, I carried out both regression and classification modeling.

4.1 Clustering models

4.1.1 Applying standard algorithms and their problems

I applied clustering models (simple clustering and k-means clustering), support vector machines (SVM), to the dataset, using root mean squared error (RMSE) as the tuning and evaluation metric. The results all had the same problems. The predicted values had much narrow range than the actual values, and as a result, the prediction errors were larger as the actual values deviated further from zero. These results were not acceptable, because places with large euclidean distances were arguably more highly rated by the users. Having larger errors on those predictions was obviously not desirable.

4.1.2 Solution to the problems

The reason behind these problems were the uneven distribution of places. Therefore, the models tried to prioritize minimizing errors on euclidean distance with little improvement/decline when RMSE was used as the evaluation metric. My solution to this problem was to assign weights to samples based on the inverse of the abundances of target values. Using this method, all models predicted target values with similar range and distribution as the actual target values.

4.1.3 Performances of different models

Using the new approach of different sample weights, I built k-means clustering, SVM, models using weighted root mean squared error as the evaluation metric. For each model, hyperparameters were tuned using the same metric and cross validation. The number of samples in each class were about the same. I chose logarithmic loss as the metric here because the results would probably be presented with probabilities and logarithmic loss puts more emphasis on the probabilities than other metrics.

5. Conclusions

In this study, I have built a clustering model which will predict I built both regression models and classification models to categorise the places into different clusters as per the daily routine provided by the user. This returns a list of all the places to be visited on a single day along with the euclidean distance from the cluster center. THis cluster center can be used to book nearby hotels as it is somewhat equidistant from all the places of visit. The cluster centre also depends on various food outlets and shopping arcades available nearby.