Travel Advisor

Arun Reddy Nalla, Darun Arumugham,
Kedareshwara Kartikeya Rao Pagadala

Abstract **- The research is aimed to predict travel similar places based on the reviews given by the user as ratings in google for the tourist attractions around Europe by using KNN regression machine learning algorithm. We apply our method (KNN regression) to characterize and predict behavior of the travelers with respect to the reviews given by them. Therefore, the recommendation is based on the reviews the user's interests.**

Keywords

Clustering algorithms, K-Means, KNN regression

1. INTRODUCTION

1.1 Background

With the advent of mobile devices that provide users with abundant Internet access, travel guidelines have largely shifted from passive mass media to interactive social media.
The tourism industry has experienced rapid growth in recent years and has the potential to impact a country's economy, necessitating the need for a Recommendation System.

The public perceptions of users can now influence others' decisions about their activities. Although This phenomenon is not limited to the following review methods and geographical areas. Ratings of the environment are the easiest to interpret quantitatively because they represent overall Place perception and highly developed continents may produce better results because most people, both locals and tourists, can afford mobile phones.

1.2 Problem Definition

In this challenge, we are provided a list of Attraction reviews from 24 categories We need to anticipate which places does the user like based upon the reviews given by them. The users in this dataset are all from the Europe. The 24 categories are Churches resorts beaches, parks, theatres, museums. malls, zoo restaurants pubs, local services, burger / pizza shops, hotels / other lodgings, juice bars, art galleries, dance clubs, swimming pools, gyms, bakeries, beauty and spas, cafes viewpoints, monuments, gardens.

This paper tackles how we approached the problem and created a model for predicting the user's interests.

1.3 Our Solution

The goal of the research paper is to predict what a particular place will the user likes/dislikes and predict similar places as recommendation. The Travel Review Ratings Data Set dataset is used in the project and is taken from UCI Machine Learning Repository. User ratings from Google reviews are used to populate this data set. Attraction reviews from 24 categories from across Europe are considered. Google user ratings range from 1 to 5, with the average user rating calculated for each category. Because the dataset used in the project is labelled, it is a supervised learning method. For training the model in supervised learning, we use a regression algorithm. The dataset contains various features that describe the different places reviews and characteristics. The response will be predicted by understanding and analyzing the places.

Firstly, the dataset is Analyzed of what the actual data format present in the dataset and check if there are any null or empty values in the dataset and is cleaned by clearing out the null values. Any day null values can cause irregular behavior of ML model, to reduce this behavior, we have used mode method to fill the null or empty values.

Then the dataset is explored for any similarity or pattern within the columns and is plotted according to the average ratings of different places.

Next, the data is shown in based on count of the reviews for different places. Then percentage of users giving near average or above for Average ratings for different places is shown. Then correlation between different places in the dataset is done. Now dataset is split into training and testing datasets. We have chosen KNN regression algorithm for training the model.

KNN regression: KNN regression is a non-parametric method that approximates the relationship between independent variables and continuous outcomes by averaging observations in the same neighborhood. The size of the neighborhood must be determined by the analyst or can be determined using cross-validation (as we will see later) to determine the size that minimizes the mean-squared error.

1.4 Related Works

In [1], three datasets have been taken from the travel and tourism sector. Different clustering techniques have been used to limit the data for the recommendation system. The objective is to determine which algorithm gives the best results. Based on the implementation of all clustering techniques, they concluded that K-means outperformed the others.

In another research paper [2], they have gathered the data directly from the footsteps of the user and not from the user input. It was a hybrid filtering method which is used to make the recommendation based on the behavioral pattern. In their case they took the location of the user from their footprints.

2. DATA EXPLORATION

2.1 Dataset Analysis

In the study, we took Travel Reviews Data Set - UCI Machine Learning Repository data set.

The following is a description of the dataset: This dataset is built using user ratings from Google reviews. Attraction reviews from 24 categories from across Europe are considered. Google user ratings range from 1 to 5, with the average user rating calculated for each category. It has used the following variables as explanatory variables:

- User: Unique user id
- Category 1: Average ratings on churches
- Category 2: Average ratings on resorts
- Category 3: Average ratings on beaches
- Category 4: Average ratings on parks
- Category 5: Average ratings on theatres
- Category 6: Average ratings on museums
- Category 7: Average ratings on malls
- Category 8: Average ratings on zoo
- Category 9: Average ratings on restaurants
- Category 10: Average ratings on pubs/bars
- Category 11: Average ratings on local services
- Category 12: Average ratings on burger/pizza shops
- Category 13: Average ratings on hotels/other lodgings
- Category 14: Average ratings on juice bars
- Category 15: Average ratings on art galleries
- Category 16: Average ratings on dance clubs
- Category 17: Average ratings on swimming pools
- Category 18: Average ratings on gyms
- Category 19: Average ratings on bakeries
- Category 20: Average ratings on beauty & spas
- Category 21: Average ratings on cafes
- Category 22: Average ratings on viewpoints
- Category 23: Average ratings on monuments
- Category 24: Average ratings on gardens

2.2 Data Pruning

In data preprocessing, data pruning is one of the techniques used to eliminate unused parameters throughout the dataset, which may improve accuracy and solve overfitting troubles. From figure 2, it is evident that Unnamed: 25 columns have 5454 null values. Considering that it does not contain any necessary information for the analysis, we will eliminate the whole column. Additionally, Unnamed: 25 does not exist in the 24 categories listed in the UCI ML repository description.

```
Category 1         0
Category 2         0
Category 3         0
Category 4         0
Category 5         0
Category 6         0
Category 7         0
Category 8         0
Category 9         0
Category 10        0
Category 11        0
Category 12        1
Category 13        0
Category 14        0
Category 15        0
Category 16        0
Category 17        0
Category 18        0
Category 19        0
Category 20        0
Category 21        0
Category 22        0
Category 23        0
Category 24        1
Unnamed: 25     5454
dtype: int64
```

Fig. 1 NULL values count information

2.3 Data Cleaning

Data cleaning is one of the most important techniques in data preprocessing which is used to add missing data, repairing the data, removing incorrect or unwanted data from our dataset. The following data cleaning steps are performed in the dataset for better performance.

A. Changing the column labels
B. Filled incorrect values
C. Changed the datatype
D. Removed Outliers

A. Changing the column labels

The first step of every data cleaning process is to alter the column names. We have changed the column names as per the UCI ML repository. The renaming of the columns is done for ease of understanding.

B. Changing the column labels

Any day NULL values can cause irregular behavior for any machine learning model. By using. isnull() function, we have discovered that two columns have one null value each. To fill the null values of the user rating, we have implemented the mean function for each column, considering that the user missed or ignored giving the rating to that attraction place. Mean will provide the average value of that column value.

C. Changed the datatype

Among the 24 categories, the local services had the datatype object. We tried converting the object type to float but ended up getting 2\t2 error. To eliminate this error, we checked how many rows have such values. Only one row ended up having the object value, so we converted that value to float by taking the mean of rest of the column value.

```
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   churches              5456 non-null   float64
 1   resorts               5456 non-null   float64
 2   beaches               5456 non-null   float64
 3   parks                 5456 non-null   float64
 4   theatres              5456 non-null   float64
 5   museums               5456 non-null   float64
 6   malls                 5456 non-null   float64
 7   zoo                   5456 non-null   float64
 8   restaurants           5456 non-null   float64
 9   pubs/bars             5456 non-null   float64
 10  local services        5456 non-null   float64
 11  burger/pizza shops    5456 non-null   float64
 12  hotels/other lodgings 5456 non-null   float64
 13  juice bars            5456 non-null   float64
 14  art galleries         5456 non-null   float64
 15  dance clubs           5456 non-null   float64
 16  swimming pools        5456 non-null   float64
 17  gyms                  5456 non-null   float64
 18  bakeries              5456 non-null   float64
 19  beauty & spas         5456 non-null   float64
 20  cafes                 5456 non-null   float64
 21  view points           5456 non-null   float64
 22  monuments             5456 non-null   float64
 23  gardens               5456 non-null   float64
dtypes: float64(24)
memory usage: 1.0+ MB
```

Fig. 2 Information after datatype change

D. Removed Outliers

Outliers are data points that are clearly out of the ordinary. This suggests that measurements were made incorrectly, the data was incorrectly collected, or variables were left out of the data collection. From figure 3, we can see that there are lots of outliers in many features. These outliers may cause interference while building the model and the outcome. The outliers play a vital role in clustering algorithms.

To find the outlier, all the user ratings are taken as whole by bringing all the category together and check how the users have rated the places in general. We have melted the data frame so that its unpivots the data from wide form into a long format which has variables and values as the parameters [3].

From figure 4, we can say that Rating-0 and Rating-5 is the two major outlier in the user ratings. Around 22500 user ratings can we considered as outliers, but in our case, Rating-0 and Rating-5 might be the true user ratings. Practically many users will give Rating-5 for their most favorite attraction and Rating-0 if they find the place to be disappointing. So, ignoring these data points will affect the accuracy and outcome of the model.

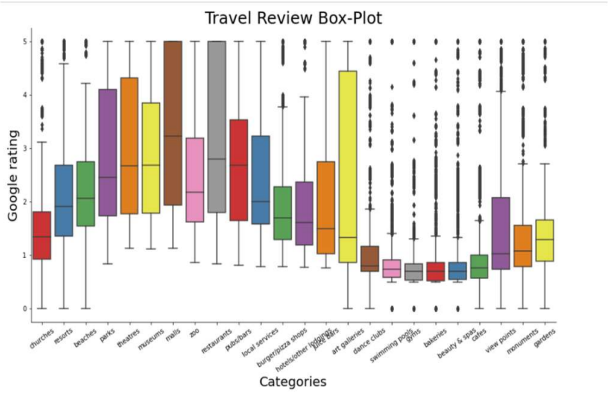

Fig. 3. Box-Plot visualization

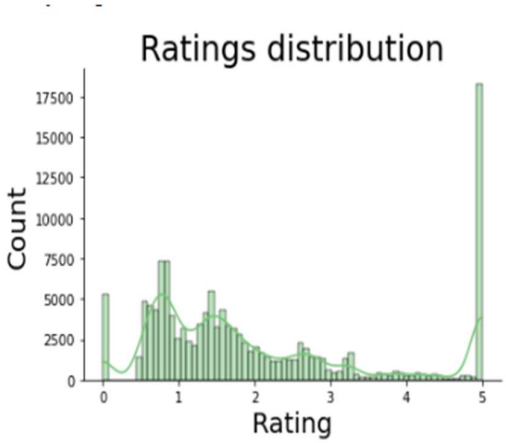

Fig. 4 Visualization which displays the Outliers

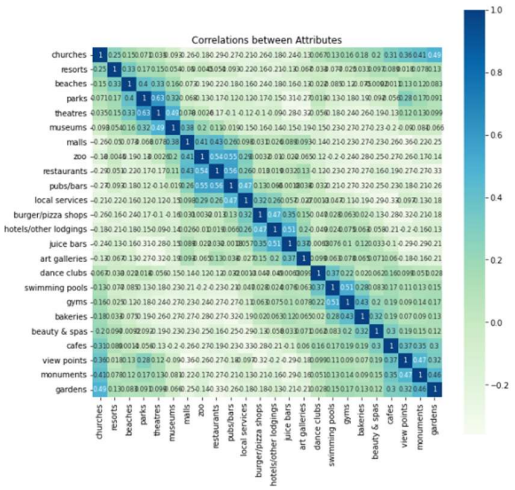

Fig.5 Correlation Matrix for the tourist places

3. METHODOLOGY

Our target variable is a most likely place a person would visit based on his previous experience of various tourist attractions. We intent to divide the users into different clusters and based on the cluster we can say that the user in that cluster should be most likely to get attracted to places in the same cluster.

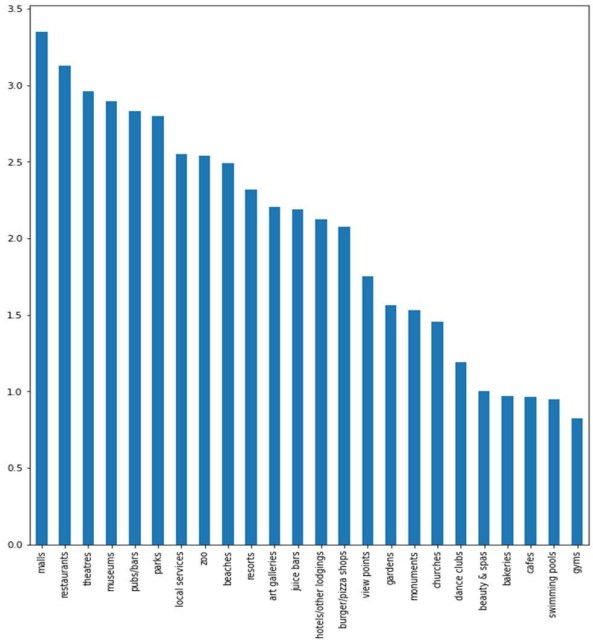

Fig.6. Bar chart on the average rating given for each category

From Fig 3.1, we can get the most and least favorite tourist attraction in Europe. In our case, Malls is the favorite attraction and gym is the lowest rated place.

3.1 K-Means algorithm

Clustering is an unsupervised learning method that divides a population or data points into groups so that data points in one group are more like data points in another group and dissimilar to data points in other groups. It is essentially a collection of objects based on their similarity and dissimilarity.

KMeans clusters continuous data using mathematical measures (distance). The closer our data points are together, the shorter the distance. We're looking at the different location ratings. So, we're going to bring all the classes together, and also look at, in general, how people rate places. No matter where they were.

For K-value, the elbow method is used in cluster analysis to determine the number of clusters in a data set. The method requires plotting the explained variation as a function of cluster count and selecting the curve's elbow as the number of clusters to use.

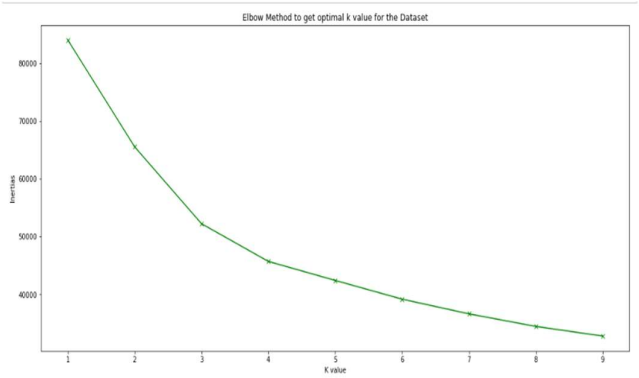

Fig. 7. Output from Elbow method to find K value

3.2 KNN Regression

In this project, we had implemented the model using k-nearest-neighbors regression. The k Nearest Neighbors (KNN) recommendation algorithm is used in most collaborative filtering recommendation processes. The KNN recommendation algorithm is simple and produces reasonably accurate results.

With the help of this model, we can record or learn what type of places a user likes to visit, based on the users' reviews. it can suggest best possible places for user using the machine learning model.

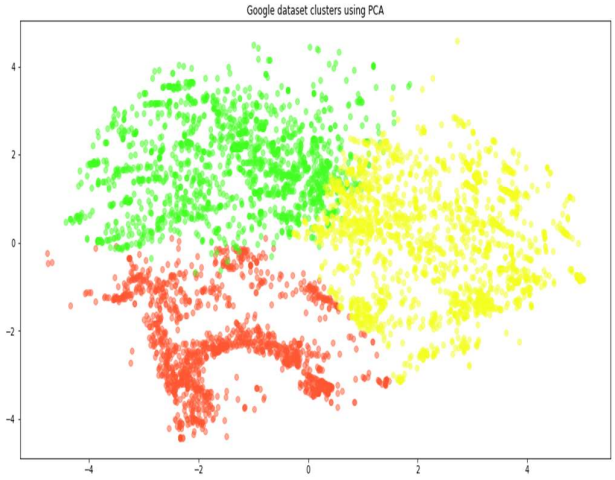

Fig. 8. Formation of three different cluster

CLUSTER	EXPLANATION
Green	Users in cluster Green like Theatres, Parks, and museums, etc. so we can recommend them places like it.
Yellow	Here in yellow Cluster, users mostly like restaurants, malls and zoo's so we can recommend users in this cluster places similar to these.
Orange	In orange cluster, Majority people like beaches, parks, and restaurants.

3.3 Evaluation

For evaluation we used ROC Curve because it is used to measure the performance of our model at various thresholds. It tells how well our model has distinguished the outputs.

We conducted a k-nearest-neighbors regression with 5-fold cross validation. We tested K's 1-20 and found that the optimal K to minimize Mean Absolute Error was 2. However, the MAE is somewhat high at 0.42 meaning that on average our prediction is off by about 10% of the range

of possible scores. This leads to some predictability, but a different model could improve upon this result.

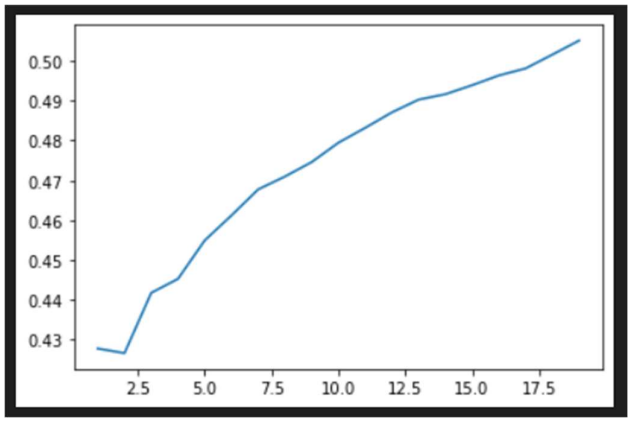

Fig. 9. ROC Curve

4. FUTURE WORK

We can use this model for developing a mobile application or web application which can record or learn what type of places a user likes to visit and based on the users reviews it can suggest best possible places for user using the machine learning model.

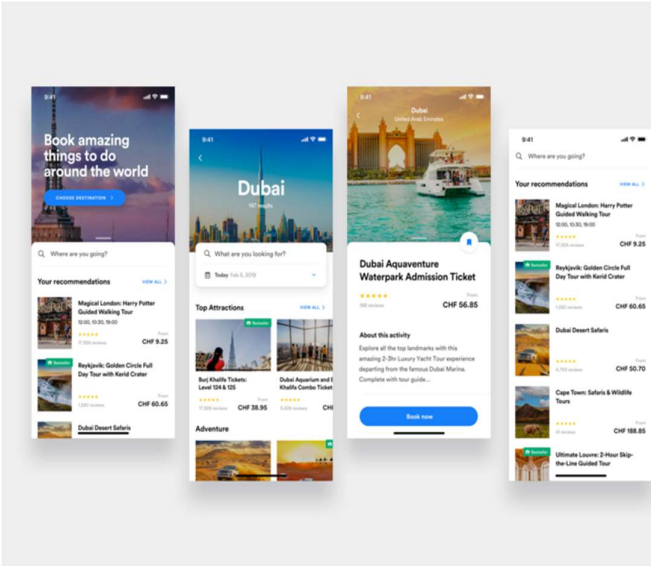

Fig. 10. Sample application snapshot

5. CONCLUSION

After completing the project, we concluded that we could use this kind of datasets of various locations and make a model which would be capable of predicting the places which can have high chances of liking by a user based on the inputs provided. Using the K means clustering we were able to divide the users into different clusters where we can say what exactly users in particular category like to visit what kind of places.

6. ACKNOWLEDGEMENT

A huge thanks to the University of Windsor, Ontario, Canada, the computer science department and to Dr. Robin Gras for the guidance.

7. REFERENCES

[1] Renjith, Shini, A. Sreekumar, and M. Jathavedan. 2018. Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism domain.

[2] Renjith, Shini, and C. Anjali. "A personalized mobile travel recommender system using hybrid algorithm." In Computational Systems and Communications (ICCSC), 2014 First International.

[3] https://pandas.pydata.org/docs/reference/api/pandas.melt.html

[4] https://dribbble.com/shots/5968159/attachments/5968159-Travel-Guide-Booking-Exploration?mode=media

[5] https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/#:~:text=KModes%20clustering%20is%20one%20of,similar%20our%20data%20points%20are.

[6] Kaggle competition

https://www.kaggle.com/code/johnmantios/travel-review-ratings-dataset

[7] https://www.techscience.com/cmc/v69n2/43849/html

[8] Kaggle competition

https://www.kaggle.com/code/hemraj12/travel-rating-reviews-analysis

[9] https://github.com/titov-vladislav/Travel-Review-Rating-Clustering