Logo, company name

Description automatically generated

Computer Science

Spring 2023

CSCI-6401 (DataMining)

Phase 5–Data Modeling

GOOGLE MAPS

Submitted by: Submitted to:

Team Amigos Dr. Shivanjali Khare

Team Members:

Member 1: Jayanth Reddy Manda (00764288) [jmand5@unh.newhaven.edu](mailto:jmand5@unh.newhaven.edu)

Member 2: Snigdha Reddy Yeruva (00762316) [syeru2@unh.newhaven.edu](mailto:syeru2@unh.newhaven.edu)

Member 3: Jeevan Kumar Konduru (00806735) [Jkond6@unh.newhaven.edu](mailto:Jkond6@unh.newhaven.edu)

GitHub link :

Research question:

Recommendation to the customer in a path from source to destination (Ex: let us Assume google maps knows user is from India, assume user is travelling from "Newhaven railway station " to "Hartford Railway Station" via car. If the user needs coffee, we recommend best coffee available stores near user based on business data available in google maps such as rating, price levels, user rating, etc....)

Dataset:

Graphical user interface, application, table, Excel

Description automatically generated

As shown in the above image we deal with multiple attributes such as:

Latitude & Longitude: Coordinates of a business so that we able to predict if user is available in available range of coordinates.

Name: Name of different businesses at coordinates.

Vicinity: Human readable location of a business.

Type: Indicates different attributes of a store such as availability of food, coffee, restaurant.…

Rating: Indicates the rating of a business by the customer, it’s one of the key parameter for the business model, it’s one of the key parameter whether a new customer is willing to visit or not.

User rating: Number of ratings users made on business. so that if the business has more ratings with god rating it indicates good business.

List Of Data Mining Techniques used:

1. K means clustering
2. Content based Recommendation System
3. Collaborative Recommendation System
4. Linear Regression
5. Decision Tree Classifier
6. SVM
7. Logistic Regression
8. Gaussian Naïve Bayes

Details of Model parameters and Hyperparameters used in datamining techniques:

|  |  |  |
| --- | --- | --- |
| Data Mining Techniques | Model Parameters | Hyperparameters |
| K means | 'rating','price\_level','user\_ratings\_total','store\_Id','user\_id' | Features, name |
| Content Based Recommendation system | Rating, price\_level, user\_rating\_total | Features |
| Collaborative Recommendation System | User\_id, store\_Id, rating | Name, Feature, price\_level, user\_ratings\_total |
| Linear Regression | Lat, lng, user\_rating\_total, price\_level | Name, store\_Id, user\_id |
| Decision Tree Classifier | Rating, user\_rating\_total, price\_level | Features, name, store\_Id, user\_id |
| SVM | Rating, user\_rating\_total | Features, name |
| Logistic Regression | Lt, lng, user\_rating\_total, rating | Price\_level |
| Gaussian Naïve Bayes | Lt, lng, user\_rating\_total, rating | Price\_level |

**Brief Description of Hardware Used:**

1. Processor: i7

2. Hard Disk: 1000 GB

3. RAM: 16 GB

4. Operating System: Windows 11 64-bit

5. Tools: Jupyter Notebook

6. Language: Python

**K means Clustering:**

To group the data, we will be using k means clustering, it groups all the data sample based on their similarity in their names, Features, users, etc... using Euclidean distance.

Graphical user interface, text, application

Description automatically generated

The following image is the result for k means clustering with 5 cluster groups.

Text

Description automatically generated

The following image describes the cluster image based on all model parameters used in k means.

Chart

Description automatically generated

**Evaluation of K means Clustering using silhouette score:**

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

The above silhouette graph indicates that as numbers of clusters decreases the score increase which indicates positiveness of the data with minimum clusters.

**Source Code:**

“google\_maps\_clustering.ipynb” is source code for K means clustering.

**Content Based Recommendation system:**

**A Content-based recommender system tries to recommend items to users, based on their profile. The user’s profile revolves around the user’s preferences and tastes, or based on the user ratings.**

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The content based recommendation system converts all the string formatted Features data into numeric based on available features which is further multiplied to form cosine matrix to recommend certain number of store based on users previous data.

The above image indicates that Lawrence top store recommendation based on his previous history data.

Source Code:

“googlemaps\_content\_based.ipynb”

**Collaborative Recommendation System:**

It works by searching a large group of people and finding a smaller set of users with tastes like a particular user. It looks at the items they like and combines them to create a ranked list of suggestions.

We will be using Singular Value Decomposition(SVD) technique is a very popular linear algebra technique to break down a matrix into the product of a few smaller matrices.

Graphical user interface, text, application

Description automatically generated

As the number of attributes taken for SVD method is very low the RMSE value is higher which indicates there may not be proper predictions using the dataset.

As MAE indicates low,it means the ability of prediction is good with provided data.

**Source Code:**

“googlemaps\_collaborative.ipynb”

**Linear Regression:**

To forecast the value of unknown data, a data analysis technique called linear regression uses the value of related and existing data. It uses a linear equation to quantitatively represent both the known and unknown variables as dependent and independent variables.

Graphical user interface, application

Description automatically generated

Liner regression Algorithm represents the accuracy of data for prediction is approximately 70% based on used parameters.

**Histogram:**

Chart, histogram

Description automatically generated

**Decision Tree Classifier:**

A supervised learning technique used in statistics, data mining, and machine learning is decision tree learning. In this formalization, a classification or regression decision tree is used as a predictive model to make judgments about a set of observations.

Diagram

Description automatically generated

Decision Tree classifies that by following the decision tree pattern we can achieve 90% accuracy for the recommendation to the user based on his preferences using the dataset .

**SVM:**

An example of a deep learning method that performs supervised learning for the classification or regression of data sets is the support vector machine (SVM). Both the desired input and output data are provided by supervised learning systems in data mining, artificial intelligence (AI) and machine learning, which are labeled for categorization.

Table

Description automatically generated with medium confidence

The above image indicates all the Evaluation Techniques with their level of percentage.

**Logistic Regression:**

A statistical analysis method called logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Text, application

Description automatically generated

The evaluation of logistic regression indicates 38% of data may be an inappropriate based on assumed parameters for logistic regression.

**Gaussian Naïve Bayes:**

Naive Bayes is a probabilistic machine learning technique that is based on the Bayes theorem and is used for numerous classification functions.  Gaussian Naive Bayes is the extension of Naive Bayes that supports continuous data and follows to the Gaussian normal distribution.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Accuracy of a price level for the data implies to 80%. Which means the classification of price level based on its attributes leads to 80% right.

**Visualization Techniques Used:**

**Confusion Matrix:**

Chart, waterfall chart

Description automatically generated

**Zero Accuracy:**

Chart, histogram

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Scatter Plots:**

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Corelation Analysis :**

Graphical user interface, text, application, email

Description automatically generated

Square

Description automatically generated

**Conclusion:**

As a recommendation system the number of attributes taken along with different algorithm indicates different perspective. In content-based recommendation system we were able to recommend top movies based on user history. By using k means clustering we make a group of data sample based on their similarity, k means algorithm analyzes that number of clusters should be low to get more accurate recommendations based on their similarities. Similar way each datamining technique uses different assumption to analyze data. As overall data mining technique indicates 80% of the data is suitable and has ability to recommend stores based on user’s preferences.