Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

Uber Related Data Analysis using Machine Learning

Rishi Srinivas [1]
[1]UG Student, Department of CSE, Sathyabama Institute of Science and Technology,Chennai.India.
[1]mrchandrika2000@gmail.com,

B.Ankayarkanni[2]
[2] Associate Professor,Department of CSE, Sathyabama Institute of Science and Technology,India.
[2]ankayarkanni.s@gmail.com,

R.Sathya Bama Krishna [3]
[3]Assistant Professor, Department of CSE, Sathyabama Institute of Science and Technology, India.[3]rsathyarajeswari@gmail.com

***Abstract*-The paper explains the working of an Uber dataset, which contains data produced by Uber for New York City. Uber is defined as a P2P platform. The platform links you to drivers who can take you to your destination. The dataset includes primary data on Uber pickups with details including the date, time of the ride as well as longitude-latitude information , Using the information, the paper explains the use of the k-means clustering algorithm on the set of data and classify the various parts of New York City. Since the industry is booming and expected to grow shortly. Effective taxi dispatching will facilitate each driver and passenger to reduce the wait time to seek out one another. The model is employed to predict the demand on points of the city.**

Keywords–Artificial Neural Network, Genetic Algorithms, K-means Clustering, Recurrent Neural Network, Time delay Neural Network, Convolutional Neural Network.

I.INTRODUCTION

The Uber platform connects you with drivers who can take you to your destination or location. This dataset includes primary data on Uber collections with details that include the date, time of travel, as well as information on longitude and latitude in San Francisco and has operations in over 900 metropolitan areas worldwide. The prediction of the frequency of trips of data is by implementing a part of k-means clustering algorithm

The standard algorithm describes the maximum variance within the group as the number of square distances Euclidean distances between the points and the corresponding centroid. The use of the digital computer has since moved to technology where the program involves the use of neural networks ,Examples of RNN (Recurrent Neural Network) and TDNN (Time delay Neural Network)for importing data from uber dataset which takes the data for forecasting on a time horizon.

The ultimate aim of the project is to predict the pickup of the cab on the basis of clusters defined by the k-means clustering algorithm. This algorithm is used to divide the dataset into k-groups. where k is defined as the number of groups provided by the user. The standard algorithm describes the maximum variance within the group as the number of square distances Euclidean distances between the points and the corresponding centroid.

The important packages used in the project are pandas,numpy,seaborn,kmeans,yellowbrick and folium.

II.LITERATURE SURVEY

Past few years have seen tremendous growth in uber related data analysis using machine learning. People are coming up with various methods to analyze uber related data such as A state in which the results, k-means clustering is used to estimate the most likely collection points at a given time and to predict the best hotspots of nightlife learning trends from previous Uber pickups. The center of the taxi service decides on the space of area to be targeted for the pickup of passengers.

This can be justified by explaining that machine learning is the core of Uber and how it has impacted on tremendous growth

- Bridging the supply demand gap
- Reduction in ETA
- Route Optimization

Poulsen, L.K In this document applied an experiment of spatial analysis of Green cab and Uber to hotspots of New York to determine the competitive position of the NYCTLC. The resulted research showed that as demand of green cabs on the hotspots grew,the demand of Uber taxis on the hotspots also growed.

This research recommends that NYCTLC creates a dashboard that analyzes and displays data in real time, as we believe this will increase its competitiveness compared to Uber. Uber is a recent taxi operator in New York and is constantly devouring the market share of the yellow and green taxis of the New York Taxi and Limousine Commission (NYCTLC).The NYCTLC is an agency of the New York City Government which licenses and regulates taxis and vehicle for hire industries and also app based companies. The commission was founded on March 2 ,1971 and their headquarters are based in New York.[1].

Faghih, S.S recommends a recent modeling approach in Manhattan, New York City, to capture the demand for electronic mail services, particularly the Uber application. Uber collection data is added to the Manhattan

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432347

TAD level and at 15-minute time intervals. This aggregation allows the implementation of a modern approach to spatio-temporal modeling to obtain a spatial and temporal understanding of the demand. During a typical day, two spacetime models were developed using Uber collection data, the STAR and STAR and MSPE turns determine the output of the models. The results of the MSPE have shown that it is recommended to use the Lasso-Star system instead of the star design. A comparison between the demand for yellow and uber taxis in 2014 and 2015 in New York shows that the demand for uber has increased[2].

Ghuhaexplained the grouping of the sequences calculated and observed by using a small amount of memory and time was necessary for applications that needed to develop a data flow model to involve large data sets and consider categorizing the data in the form of clusters[3].

Ahmed, M., has shown that by using detailed data on taxis at the travel level and on the rental vehicle and data on complaints about the level of new complaints at the level of incidents, we study how Uber and Lyft enter damaged the quality of taxi services in New York City. The overall effect of the organizations based on the scenario and in particular of the riding administrations was enormous and widespread. One of these effects is the expansion of the rivalry between Uber and Lyft over the quality of taxi administration. They use a new set of complaint data to measure (the lack of) quality of service that we have never been analyzed before. Focus on the quality dimensions generated by most of the complaints we demonstrate. The increased competition for these shared travel services has had an intuitive impact on the behavior of taxi drivers[4].

Wallsten, S, stated that the results of New York and Chicago are consistent with the possibility that taxis react to the new challenge by improving quality. In New York, the rise of Uber is linked to the reduction of objections to travel to the city. They discuss the competitive effect of sharing taxis in the taxi industry using the complete data set of the New York City Taxi and Limousine Commission for more than one billion taxi trips in complaints and details of New York, New York and Chicago Google Trends on the success of Uber's largest shared travel service.[5].

Sotiropoulos, D.N, represented that this document addresses the problem of grouping, by using a new approach to genetic algorithms that is highly scalable in large volumes of textual details, developing a coding scheme based on centroids. We apply k means clustering algorithm in this document. Clustering is the unsupervised machine learning algorithm used to solve grouping problems based on similarities. This technique has aroused interest in a wide range of scientific fields, which address clustering methods, to solve complex classification problems.[6].

Faghih, S.S said that the demand for electronic mail services is growing rapidly, particularly in large cities. Uber is the first and most famous email company in the United States and New York City. A comparison between the demand for yellow and Uber taxis in New York in 2014 and 2015 shows that the demand for Uber has increased. To study the forecast performance of the models, you choose to choose data for a typical day. Our goal in this document is to describe how these models can be used for forecasting Uber demand. The Uber data contains information about the position and time of the pick-ups and returns of each trip during a day. According to the available data, the Uber historical data of April 2014[7].Kumar, states that, k-means clustering is used to estimate the most likely collection points at a given time and to predict the best hotspots of nightlife learning trends from previous Uber pickups [8,9].

L.Liu, C.Andris, and C.Ratti planned for a strategy to disclose cabdrivers working patterns by inspecting their unbroken anatomy track[10].R-H Hwand focuses on GPS and the locality to pick up passengers , A venue to venue plot model referred to as an OFF-ON model [11].

III. PROPOSED METHOD

Based on the problems of forecasting errors and risk of overfitting due to large datasets. The data analyzed and sent to the company is resulted as inefficient and ineffective. Thus to overcome the problem we are going to predict the pickup of cab from a coordinated cluster of points predicted by using applied k-means clustering algorithm.

The k-means clustering algorithm adopted will effectively dispatch taxis to the cluster.This facilitates each driver and passenger to attenuate the wait-time to search out one another. Drivers don’t have enough info concerning wherever passengers and different taxis area unit and shall move.Therefore, a cab center will organize the taxicab fleet and with efficiency give out consistent request to the whole town.

The system uses the latitude and longitude of the cab scheduled and also the day of the travel and the month.An unsupervised learning model is trained with this dataset and the model is employed to predict the pickup of the cab on the cluster.The proposed method for the project is explained on 7 steps.

A. System Architecture

B. Raw Data

C. Data Importing

D. Data Visualization

E. Testing Data

F. Predicted Scheduling Of Cab Using Algorithm

G. Algorithm

A. System Architecture

The system architecture for the given module is as follows:

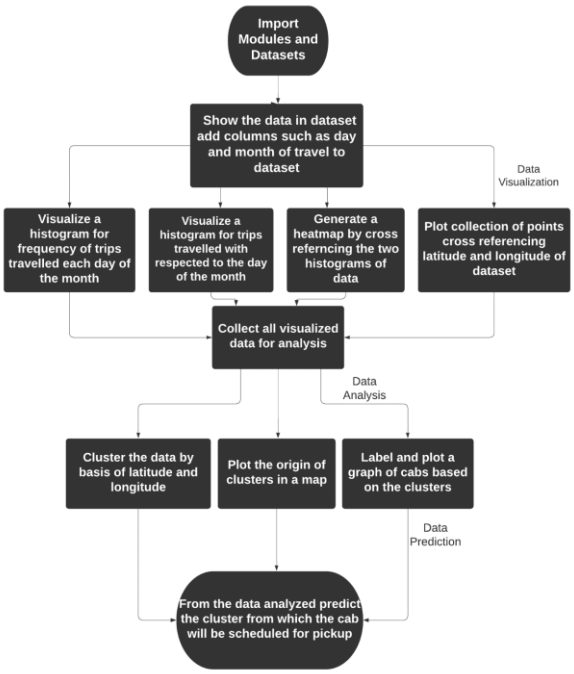


Fig 1. System Architecture

B. Raw Data(Dataset)

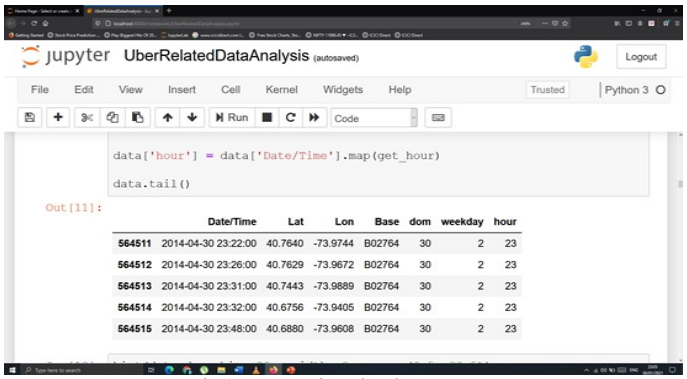
The definition of raw data comes from the concept of data not processed and is obtained from the dataset or is sometimes made by the end product of data processing. The steps required in raw data are extraction, organization and sometimes analysis.

	A	B	C	D
1	Date/Time	Lat	Lon	Base
2	04/01/2014 00:11	40.769	-73.9549	B02512
3	04/01/2014 00:17	40.7267	-74.0345	B02512
4	04/01/2014 00:21	40.7316	-73.9873	B02512
5	04/01/2014 00:28	40.7588	-73.9776	B02512
6	04/01/2014 00:33	40.7594	-73.9722	B02512
7	04/01/2014 00:33	40.7383	-74.0403	B02512
8	04/01/2014 00:39	40.7223	-73.9887	B02512
9	04/01/2014 00:45	40.762	-73.979	B02512
10	04/01/2014 00:55	40.7524	-73.996	B02512
11	04/01/2014 01:01	40.7575	-73.9846	B02512
12	04/01/2014 01:19	40.7256	-73.9869	B02512
13	04/01/2014 01:48	40.7591	-73.9684	B02512
14	04/01/2014 01:49	40.7271	-73.9803	B02512
15	04/01/2014 02:11	40.6463	-73.7896	B02512
16	04/01/2014 02:25	40.7564	-73.9167	B02512
17	04/01/2014 02:31	40.7666	-73.9531	B02512

Fig 2. Raw dataset (csv file)

C.DataImporting

A huge amount of trip data will be collected fromUberfor training and testingdata. From the collected dataset the latitude and latitude will be clustered and classified based on the frequency of trips travelled by the cab during the day. When these criteria are considered, and data preprocess will be done on these datasets.



	Date/Time	Lat	Lon	Base	dom	weekday	hour
564511	2014-04-30 23:22:00	40.7640	-73.9744	B02764	30	2	23
564512	2014-04-30 23:26:00	40.7629	-73.9672	B02764	30	2	23
564513	2014-04-30 23:31:00	40.7443	-73.9889	B02764	30	2	23
564514	2014-04-30 23:32:00	40.6756	-73.9405	B02764	30	2	23
564515	2014-04-30 23:48:00	40.6880	-73.9608	B02764	30	2	23

Fig 3. Processing the dataset

D.Data Visualization

Data visualization is defined as to evaluate the performance of a model by using graphs and metrics that calculate performance.Data visualization can be mainly used to categorize the data into new levels such that the algorithm used can be generalized to an observation of each output variable derived by an observed input variable.

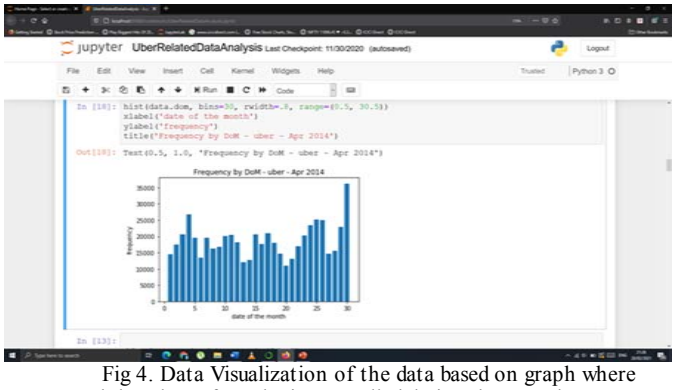


Fig 4. Data Visualization of the data based on graph where graph has data of total trips travelled during the month .

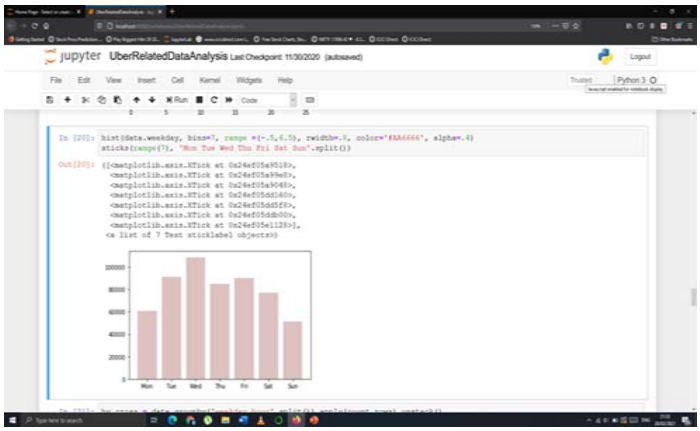


Fig 5.Visualizing data on graph on frequency of trips travelled during the day.

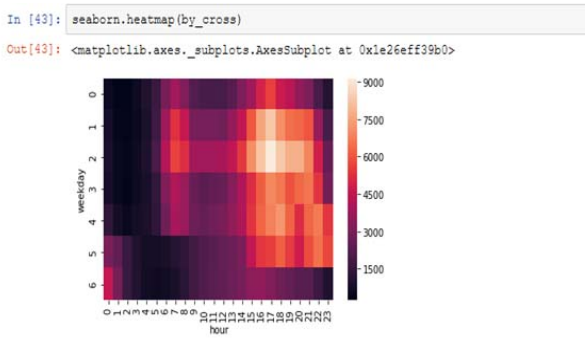


Fig 6 Data Visualization of Heatmap of frequency of cabs travelled during the hour.

E. Testing data

The main step after visualizing data in an algorithm is to test the data,the test set can be defined as a set of observations which is used to evaluate the performance of a model by using performance metrics.The program that uses the test set must be able to generalize and effectively perform with the dataset to yield the predicted data accurately such that the program is effective in nature. Moreover when the program memorizes the dataset it is termed overfitting hence to balance overfitting we use regularization which is applied to the model to reduce it.

F.Predicted Scheduling of Cab using Algorithm

The scheduling of the cab can be predicted on the basis of the location given by the user and the proposed method finds the nearest hotspot which is defined as a cluster of points analyzed by k-means clustering and gives info to the cab on the hotspot nearest to the location of the user and is booked to pickup the user.

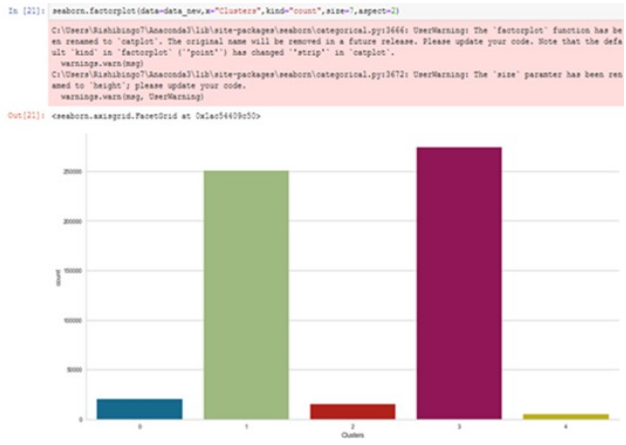


Fig.7. Visualization of total trips travelled by cabs on a day on the basis of clusters.

G.Algorithm

The algorithm used is involved on the concept of k-means clustering algorithm. which belongs to unsupervised learning. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

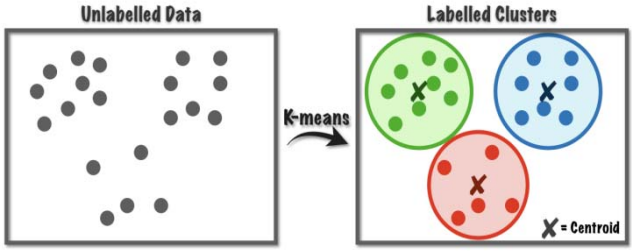


Fig.8. K means clustering

The clustering algorithm is categorized into three steps:

- Take mean value
- Find nearest number of mean and put in cluster
- Repeat the mean value and number of means in the cluster till we get same mean

Clustering is the method of grouping objects into groups based on similarities. This algorithm is used to divide a given data set into k groupsHere, k represents the number of groups and must be provided by the user. The idea behind the grouping of k-means is to identify the clusters in such a way to reduce total variation within the cluster. The standard algorithm describes the maximum variance within the group as the number of square distances Euclidean distances between the points and the corresponding centroid. The grouping can be classified into two groups.

Hard grouping and soft grouping. Through a hard grouping, each object or data point belongs to a cluster. For example, all locations clustered in the dataset belong to a district.In the soft grouping, a data point can belong to more than one group with a certain probability or probability value. In connectivity-based clustering, the main idea behind this cluster is that the data points closest to the data space are more related than those of the data point further back. Groups are created by linking data points based on their length. Grouping based on centroids, in this type of grouping, the groups are represented by a central vector or a centroid. This centroid may not necessarily be a member of the data set. This is an iterative grouping algorithm in which the notion of similarity derives from the proximity of the data point to the center of the cluster.

```
In [1]: import matplotlib.pyplot as plt
        from sklearn.cluster import KMeans
        from yellowbrick.cluster import KElbowVisualizer

In [8]: kmeans = KMeans(n_clusters = 5, random_state = 0)
        kmeans.fit(clus)

Out[8]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
               n_clusters=5, n_init=10, n_jobs=None, precompute_distances='auto',
               random_state=0, tol=0.0001, verbose=0)

In [9]: centroids = kmeans.cluster_centers_
        centroids

Out[9]: array([[ 40.79813773, -73.87204835],
               [ 40.76302051, -73.97574403],
               [ 40.6599309 , -73.77672246],
               [ 40.71968154, -73.99233502],
               [ 40.70048892, -74.20152276]])
```

Fig.9 Importing k-means module and calculating the centroids of the dataset

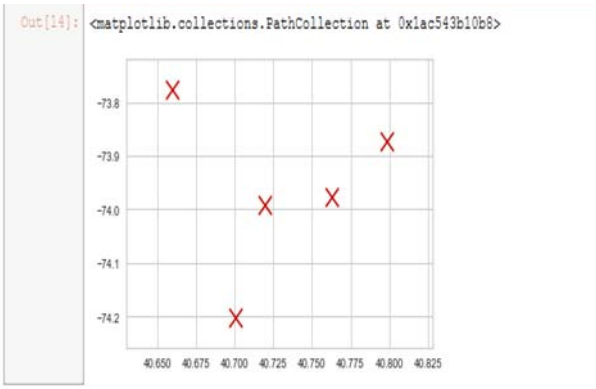


Fig 10. Plotting the centroids on x-y graph of latitude and longitude.

IV.RESULTS AND DISCUSSION

The program predicts the pickup location of the cab based on the centroids plotted using applied by k-means clustering for appropriate cab scheduled for pickup.The results discussed are based on the following figures below

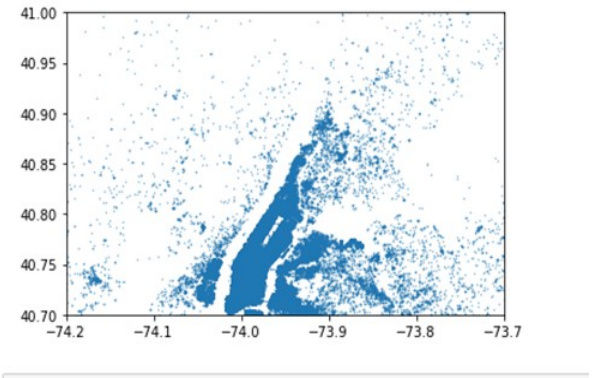


Fig.11 Plotting the collection of points through which cab has travelled during the course of the month.

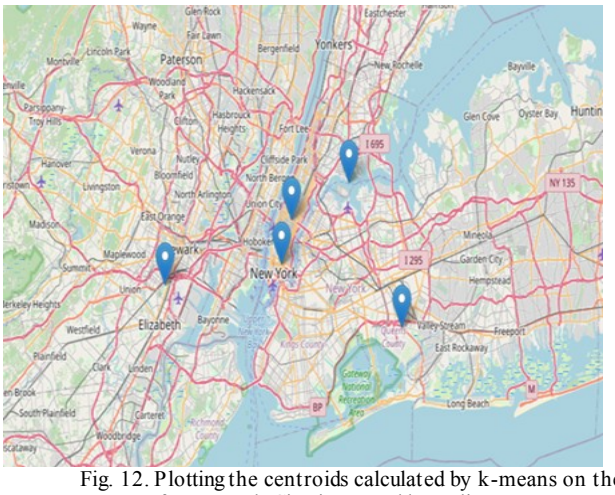


Fig. 12. Plotting the centroids calculated by k-means on the map of New York City imported by Folium.

```
In [18]: label = kmeans.labels_
         label

Out[18]: array([1, 3, 3, ..., 1, 3, 3])

In [20]: data_new = data
         data_new['Clusters'] = label
         data_new
```

Out[20]:

	Date/Time	Lat	Lon	Base	Clusters
0	4/1/2014 0:11:00	40.7690	-73.9549	B02512	1
1	4/1/2014 0:17:00	40.7267	-74.0345	B02512	3
2	4/1/2014 0:21:00	40.7316	-73.9873	B02512	3
3	4/1/2014 0:28:00	40.7588	-73.9776	B02512	1
4	4/1/2014 0:33:00	40.7594	-73.9722	B02512	1
5	4/1/2014 0:33:00	40.7383	-74.0403	B02512	3
6	4/1/2014 0:39:00	40.7223	-73.9887	B02512	3
7	4/1/2014 0:45:00	40.7620	-73.9790	B02512	1
8	4/1/2014 0:55:00	40.7524	-73.9960	B02512	1
9	4/1/2014 1:01:00	40.7575	-73.9846	B02512	1
10	4/1/2014 1:19:00	40.7256	-73.9869	B02512	3
11	4/1/2014 1:48:00	40.7591	-73.9684	B02512	1
12	4/1/2014 1:49:00	40.7271	-73.9803	B02512	3
13	4/1/2014 2:11:00	40.6463	-73.7896	B02512	2
14	4/1/2014 2:25:00	40.7564	-73.9167	B02512	1
15	4/1/2014 2:31:00	40.7666	-73.9531	B02512	1
16	4/1/2014 2:43:00	40.7580	-73.9761	B02512	1

Fig.13. Labelling the dataset on the basis of clusters

```
In [24]: new_location = [(40.86, -75.56)]
         kmeans.predict(new_location)

Out[24]: array([4])

In [25]: clocation.head()
```

Out[25]:

	Latitude	Longitude
0	40.798138	-73.872048
1	40.763021	-73.975744
2	40.659931	-73.776722
3	40.719682	-73.992335
4	40.700489	-74.201523

Fig. 14. Predicting the pickup of the cab from the particular cluster and showing the cluster coordinates.

V.CONCLUSION AND FUTURE WORK

The conclusion of the project is to project a basic outline of trips travelled with respect to latitude and longitude of locations and pinpoint the locations travelled with respect to the frequency of trips travelled by a uber cab during the day and also based on the cross analyzing of the dataset based on the latitude and longitude of the point travelled by the cab which is then analyzed by deploying k-means clustering which classifies the locations on the basis

of centroids and then orders the frequency of trips based on labels or clusters. By the location given by the user, the algorithm predicts the cluster nearest to the location so that cab can be assigned to the user for pickup.

The merit of the project is that it explains the functioning of how cabs are assigned to passengers based on an unsupervised algorithm and also explains the key concepts of machine learning. The limitations of the project are that the algorithm deployed may be inefficient for huge data for over 10 years.

The future work suggests that the system will provide the location to the user. The algorithm then records the time, latitude, longitude of the trip and assigns it to a cluster nearest to the passenger location where a cab is scheduled for pickup. We can also predict the passenger count on each district to deploy more cabs to the clustered coordinates using convolutional neural networks (CNN)

REFERENCES

[1] Poulsen, L.K., Dekkers, D., Wagenaar, N., Snijders, W., Lewinsky, B., Mukkamala, R.R.andVatrapu, R., 2016, June. Green Cabs vs. Uber in New York City. In 2016 IEEEInternational Congress on Big Data (BigData Congress) (pp. 222-229). IEEE.

[2] Faghih, S.S., Safikhani, A., Moghimi, B. and Kamga, C., 2017. Predicting Short-Term UberDemand Using Spatio-Temporal Modeling: A New York City Case Study. arXiv preprintarXiv:1712.02001.

[3] Guha, S. and Mishra, N., 2016. Clustering data streams. In Data stream management (pp.169-187). Springer, Berlin, Heidelberg.

[4] Ahmed, M., Johnson, E.B. and Kim, B.C., 2018. The Impact of Uber and Lyft on TaxiService Quality Evidence from New York City. Available at SSRN 3267082.

[5] Wallsten, S., 2015. The competitive effects of the sharing economy: how is Uber changingtaxis. Technology Policy Institute, 22, pp.1-21.

[6] Sotiropoulos, D.N., Pournarakis, D.E. and Giaglis, G.M., 2016, July. A genetic algorithmapproach for topic clustering: A centroid-based encoding scheme. In 20167th InternationalConference on Information, Intelligence, Systems & Applications (IISA) (pp. 1-8). IEEE

[7] Faghih, S.S., Safikhani, A., Moghimi, B. and Kamga, C., 2019. Predicting Short-TermUber Demand in New York City Using Spatiotemporal Modeling. Journal of Computing inCivil Engineering, 33(3), p.05019002.

[8] Shah, D., Kumaran, A., Sen, R. and Kumaraguru, P., 2019, May. Travel Time EstimationAccuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR✱.In Companion Proceedings of The 2019 World Wide Web Conference (pp. 130-136). ACM.

[9] Kumar, A., Surana, J., Kapoor, M. and Nahar, P.A., CSE 255 Assignment II PerfectingPassenger Pickups: An Uber Case Study.

[10] L.Liu, C.Andris, and C.Ratti , "Uncovering cabdrivers behaviour patterns from their digital traces",Compu. Environ.UrbanSyst.,vol.34,no.6,pp.541-548,2010

[11] R.H.Hwang,Y.L.Hsueh , and Y.T.Chen,"An effective taxi recommender system model on a spatio-temporal factor analysis model,"Inf.Sci.,vol.314,pp.28-40,2015.

[12] Vigneshwari, S., and M. Aramudhan. "Web information extraction on multiple ontologies based on concept relationships upon training the user profiles." In Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, pp. 1-8. Springer, New Delhi, 2015.

[13] L. Rayle, D. Dai, N. Chan, R. Cervero, and S. Shaheen, "Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco," Transport Policy, vol. 45, 01 2016.

[14] O. Flores and L. Rayle, "How cities use regulation for innovation: the case of uber, lyft and sidecar in san francisco," Transportation research procedia, vol. 25, pp. 3756–3768, 2017.

[15] H. A. Chaudhari, J. W. Byers, and E. Terzi, "Putting data in the driver's seat: Optimizing earnings for on-demand ride-hailing," in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018, pp. 90–98.