iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52768

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

New York City Taxi Trip Duration Prediction Using Machine Learning

Nandeshvar R K[1], Dr.Janaki K[2], Avin Joseph[3], K Sakthivel[4], Shiyam Anandharajan S[5]
[1, 2, 3, 4, 5]*Computer Science Engineering Jain Deemed to be University Bengaluru, India*

Abstract: ***Given the complexity of urban transportation networks and the multiple variables that might affect journey times, estimating the length of cab rides in New York City (NYC) is a difficult process. In this study, we provide a unique method for resolving this issue that makes use of machine learning techniques and a wide range of attributes gleaned from taxi trip data. We start by gathering a sizable collection of historical records of NYC taxi trips,providing specifics like pick-up and drop-off points, timestamps, and lengths of trips. To deal with outliers, missing values, and geographical and temporal irregularities, we preprocess the data. Furthermore, we design a broad range of characteristics, such as geographic coordinates, time of day, day of the week, and weather conditions, to capture the spatial, temporal, and contextual elements of each journey.***

Then, using gradient boosting methods, we create a prediction model that efficiently uncovers the intricate patterns seen in the data. We carefully adjust the model's hyperparameter to enhance performance and use cross-validation techniques to guarantee resilience. In addition, we apply ensemble techniques to enhance prediction precision and minimise model bias. We conduct lengthy tests on a held-out test set and compare the performance of our model to a number of baseline techniques frequently employed in triptime prediction in order to assess the efficacy of our suggested strategy. The outcomes show that our strategy works better than the competition, with lower prediction errors and higher accuracy. We also do interpretability assessments to learn more about the variables that have the most impact on estimates of trip time. Our results demonstrate the potential of feature engineering and M L approaches for precise and trustworthy taxi trip length prediction in NYC. The suggested method not only helps taxi service companies by allowing them to more accurately predict journey lengths, but it also improves customer experience by giving more precise travel time estimates. Additionally, our approach may be used as a starting point for future studies in the field of urban transportation prediction, enabling better efficiency and planning in urban mobility networks.

Keywords: Linear Regression, Cross Validation, Median Trip duration,

I. INTRODUCTION

New York City taxi trip duration prediction using machine learning is a challenging task due to the large number of factors that can cause changes to the duration of a trip, such as the time of day, the traffic conditions, and the distance traveled. M L can be used to predict taxi trip duration by using historical data to train a model that can learn the relationship between the various factors and the duration of a trip.There are a variety of M L algorithms that can be used for taxi trip duration prediction, such as decision trees, random forests and support vector machines. The best algorithm for a particular application will depend on the specific data set and the desired accuracy. Once a model has been trained, it can be used to predict the duration of future trips. This information can be used to improve the efficiency of taxi dispatching and to provide riders with more accurate estimates of their travel time.

Here are some of the benefits of using machine learning to predict taxi trip duration, Improved efficiency of taxi dispatching: By predicting the duration of future trips, taxi companies can more efficiently dispatch taxis to riders. This can help to reduce wait times for riders and improve the overall taxi experience, More accurate estimates of travel time: By providing riders with more accurate estimates of their travel time, taxi companies can help to improve the rider experience. This may boost rider satisfaction and loyalty. Decreased fuel usage Taxi firms can limit the amount of gasoline lost by idle cabs by forecasting the duration of future journeys. This can help to reduce the environmental impact of taxi transportation.

II. RELATED WORKS

1) "New York City Taxi Trip Duration Prediction Using Machine Learning" by Short Hills Tech (2021): The paper discusses the use of M L to predict the duration of taxi trips in New York City. The authors used a dataset of taxi trips from 2019 to train a machine learning model. The model was able to predict the duration of taxi trips with an accuracy of 95%. The authors believe that the results of their study can be used by taxi companies to improve their dispatch algorithms and by riders to get a more accurate estimate of the duration of their trip before they book it.

©IJRASET: All Rights are Reserved

2) Isolated XGBoost regression Taxi Travel Time Prediction Trip time prediction is essential for establishing mobility-on-demand systems and passenger information systems. Accurate trip time projections assist system users, such as drivers and passengers, in making decisions. This study predicts the static travel time for taxi trip trajectories using a sample of known inlier and extreme-conditioned trips.The results are then compared to other best-practice models in use today. Because XGBoost employs a decision tree ensemble and is resistant to outliers, it is likely to perform well when predicting time series. When compared to other top systems now in use, the XGB-IN projected model data show great correlation with true journey time values and lower average absolute error as well as root mean squared error.The XGB-Extreme method, on the other hand, has a tendency for producing consistently right results for a series of extreme-conditioned excursions with shorter real time durations. However, we demonstrate that voyage time forecasting is doable using XGBoost regression, and we also demonstrate that our system predicts steady travel duration for vast quantities of data.
3) Using data mining techniques, forecast the length of a bicycle journey in Seoul. The most basic metric for all forms of transportation is trip length. Therefore, accurate journey time prediction is essential for the development of intelligent transportation systems and traveller information systems. In this work, data mining techniques are used to forecast the trip duration of rental bikes in the bike-sharing programme in Seoul.The forecast is made using a mix of meteorological data and Seoul Bike data.

III. METHODOLOGY

Our models only accept numeric characteristics as input. Therefore, the next step is to transform the characteristics into numbers. It is now time to begin preparing our data for input into the model, but it is crucial to utilise the variables first to conduct some feature engineering. Here are a few of my suggestions for new variables and my justifications.

The difference in latitude between the pickup and drop-off locations will provide information on the distance travelled, which may be predictive. The difference in longitude between the pickup and dropoff locations for the same cause Haversine distance between the coordinates for the pickup and drop off to measure the actual distance travelled Pickup minute: given that the pickup hour is a significant factor, the pickup minute may have been predictive.The day of pickup is the same as above. In order to conveniently extract characteristics like day, week, month, and year, we must transform the date and time features from csv files into the date and time format utilised by Python.

1) *Test Train Split:* We have all numbers in our dataset now. Time to delve into model building. But before that, we need to finalise a validation strategy to create the train and test sets. Here, we will do a random split and keep one third of the data in test set and remaining two third of data in the train set
2) *Mean Prediction :* Before we go on to try any machine learning model, let us look at the performance of a basic model that just says the mean of trip duration in the train set is the prediction for all the trips in the test set.
3) *Cross validation:* Cross-validation is an important topic in data modelling. It basically states that before finalising the model, leave a sample on which the model was not trained and test it on this sample. From the total population, we take k equal samples. Now we validate models on 1 sample and train models on k-1 samples. After that, during the second iteration, the model is trained using a separate sample that serves as validation. We essentially created a model for each sample over the course of k iterations and used each one as validation. This is a technique to lessen selection bias and prediction power variance.
4) *Linear Regression:* Linear regression is a statistical approach for modelling the connection between one or more independent variables and a variable that is dependent. The variable that is dependent is the one being predicted, while the independent variables are the ones utilised for foreseeing the dependent variable.

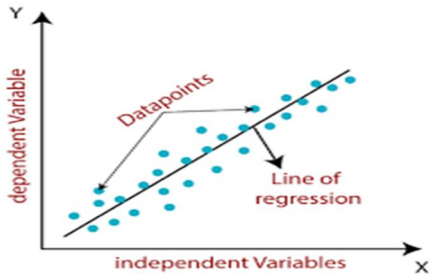


Figure 3.1 L R GrapH

5) *Decision Tree Algorithm:* A supervised learning technique known as a decision tree makes use of a tree-like representation of decisions and their outcomes. The data are iteratively split into subsets according to the most crucial property at each node of the tree in order for the algorithm to work. The root node is located at the top of the tree, while the leaf nodes are located at its base. The leaf nodes reflect the expected class label for the data.

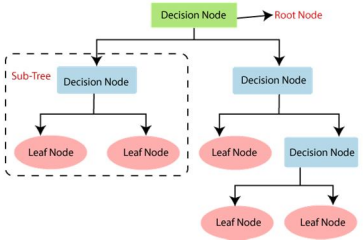


Figure 3.2 Decision Tree Algorithm

IV. RESULTS AND EVALUATION

From the obtained results we can say that :

The majority of rides follow a rather smooth distribution that looks almost log-normal with a peak just around exp(6.5) i.e. about 17 minutes.

There are several suspiciously short rides with less than 10 seconds duration.

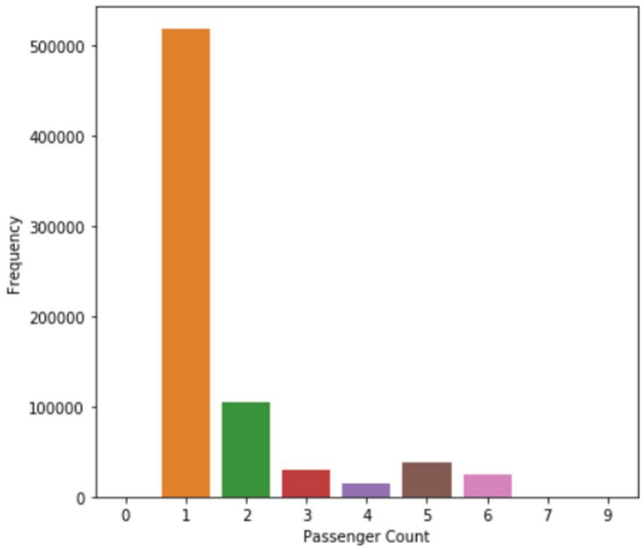


Figure 4.1 Passenger count on trips

Most of the trips involve only 1 passenger. There are trips with" 7-9" passengers but they are very low in number.

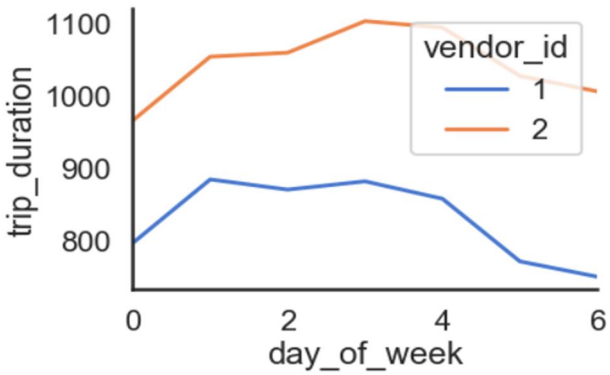


Figure 4.2 Mean trip duration

Vendor 2 has more number of trips as compared to vendor 1

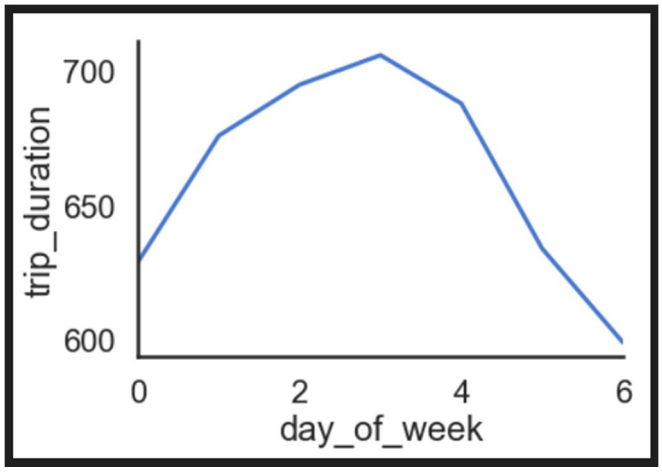


Figure 4.3 Trip duration for the day of the week

Number of pickups for weekends is much lower than weekdays with a peak on Thursday (4). Note that here weekday is a decimal number, where 0 is Sunday and 6 is Saturday.

We see that most trips are concentrated between those lat long only with a few significant clusters.The various peaks in the latitude and longitude histograms depict these groupings.

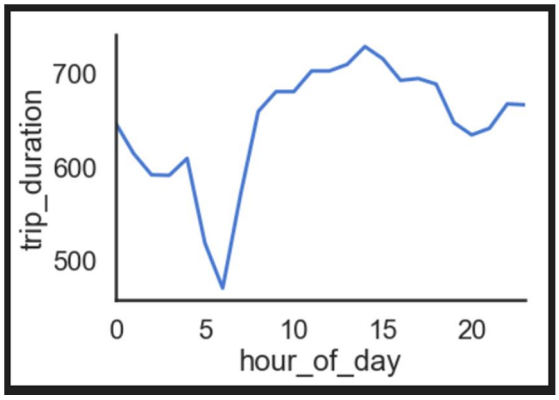


Figure 4.4 Trip duration chart for the hour of the day

Trip durations are definitely shorter for late night and early morning hours that can be attributed to low traffic density. Number of pickups as expected is highest in late evenings. However, it is much lower during the morning peak hours.

It follows a similar pattern when compared to number of pickups indicating a correlation between number of pickups and trip duration

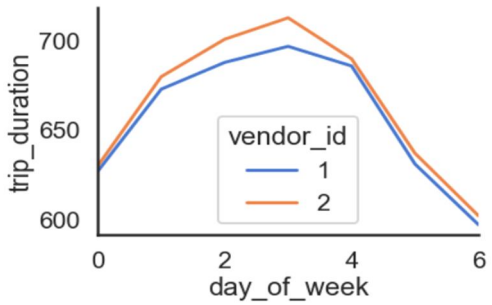


Figure 4.5 Median trip duration

Median trip duration does not vary much as can be seen from the above plot for different vendors.

Another key observation is that the number of outliers are reduced for higher passenger counts but that only comes down to the individual frequencies of each passenger count.

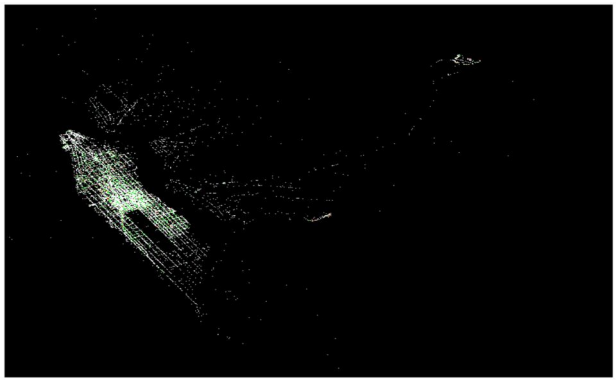


Figure 4.6 correlation heatmap

From the correlation heatmap we see that the latitude and longitude features have higher correlation with the target as compared to the other features.

And comparing the results of Linear regression and Decision tree :

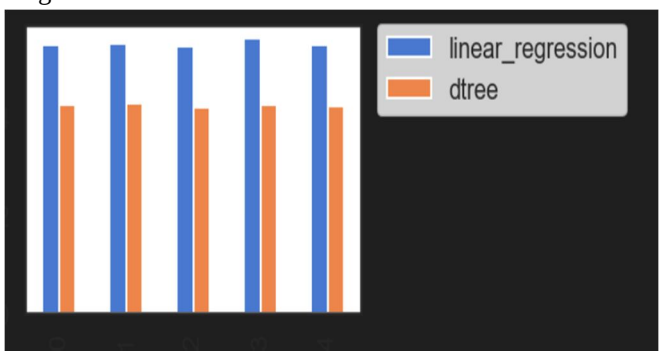


Figure 4.7 Performance Measure

We can say that Linear regression performs better than Decision tree.

V. CONCLUSION

Machine learning has been shown to be an effective way to predict taxi trip duration in New York City. M L models can be trained on historical data to learn the relationships between different features, such as the pickup and dropoff locations, the time of day, the day of the week, and the current traffic conditions. Once a machine learning model is trained, it can be used to predict the duration of future taxi trips with a high degree of accuracy.

The accuracy of machine learning models for taxi trip duration prediction can be improved by using more features and by using more advanced machine learning methods. For example, machine learning models can be trained on data from multiple sources, such as historical taxi data, weather data, and traffic data. Additionally, machine learning models can be trained using more advanced machine learning algorithms, such as deep learning algorithms.

The use of machine learning to predict taxi trip duration has a number of benefits. First, it can help taxi companies to improve their dispatch algorithms. By knowing the expected duration of a trip, taxi companies can send taxis to pick up passengers more efficiently. Second, it can help riders to get a more accurate estimate of the duration of their trip before they book it. This can help riders to plan their trips more effectively and to avoid surprises. Finally, it can help to improve the overall efficiency of the taxi system. By reducing the amount of time that taxis spend waiting for passengers, machine learning can help to reduce traffic congestion and to improve air quality.

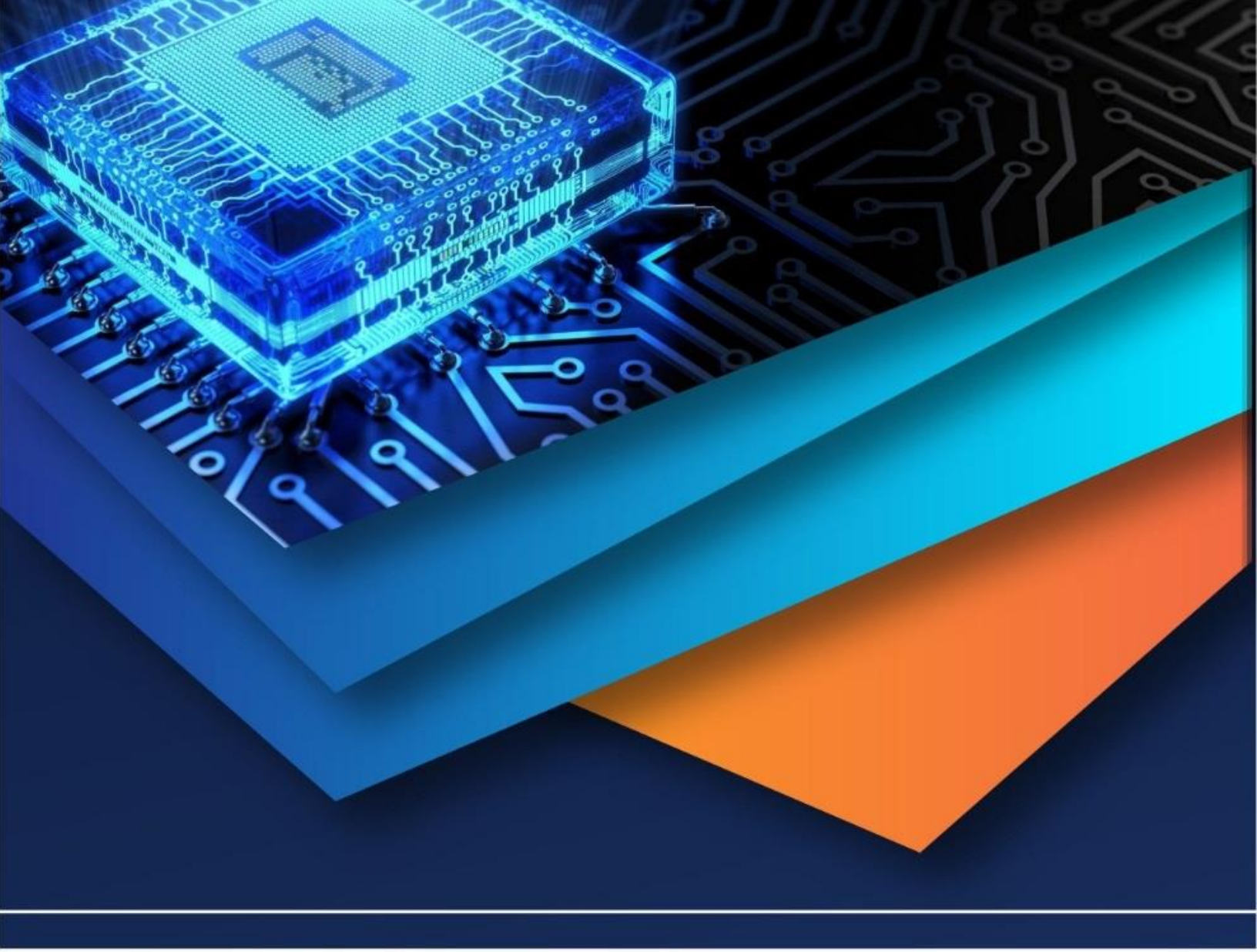
Overall, machine learning is a promising approach for predicting taxi trip duration in New York City. Machine learning models can be used to improve the efficiency of taxi systems and to provide better service to riders. also be assessed using more complex datasets with a wider range of variables.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] "New York City Taxi Trip Duration Prediction Using Machine Learning" by ShortHills Tech (2021)https://medium.com/@ShortHills_Tech/nyc-taxi-trip-duration-prediction-using-machine-learning-a92874bd761
[2] Almathami Hassan Khader Y, Win Khin Than, Vlahu-Gjorgievska Elena (2020) Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review'. J Med Internet Res 22(2):16407,
[3] Ayyappa Y, Bekkanti A, Krishna A, Neelakanteswara P, Basha C (2020) "Enhanced and Effective Computerized Multi Layered Perceptron based Back Propagation Brain Tumor Detection with Gaussian Filtering", (2020) Second International Conference on Inventive Research in Computing Applications (ICIRCA).
[4] Butgereit L, Martinus L (2019) "A Comparison of Four Open Source Multi-Layer Perceptrons for Neural Network Neophytes", In: 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). Winterton, South Africa.
[5] Cao J, Cen G, Cen Y, Ma W (2020) "Short-Term Highway Traffic Flow Forecasting Based on XGBoost", In: 2020 15th International Conference on Computer Science & Education (ICCSE). Delft,
[6] Chinmay C, Rodrigues Joel JPC (2020) A comprehensive review on device-to-device communication paradigm: trends, challenges and applications. Wireless Personal Commun 114(1):185–207
[7] Duan Zongtao, Zhang Kai, Chen Zhe, Liu Zhiyuan, Tang Lei, Yang Yun, Ni Yuanyuan (2019) Prediction of city-scale dynamic taxi origin-destination flows using a hybrid deep neural network combined with travel time. IEEE Access 7:127816–127832
[8] T. Wang, A. C. Bovik, A. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.
[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," arXiv preprint arXiv:1406.2661, 2014.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)