Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-021-01130-x

ORIGINAL ARTICLE

New York City taxi trip duration prediction using MLP and XGBoost

M Poongodi[1] · Mohit Malviya[2] · Chahat Kumar[3] · Mounir Hamdi[1] · V Vijayakumar[4] · Jamel Nebhen[5] · Hasan Alyamani[6]

Received: 5 January 2021 / Revised: 26 March 2021 / Accepted: 28 April 2021

Abstract New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on. Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require of him to travel from one place to another. Given the rising popularity of app-based taxi usage through common vendors like Ola and Uber, competitive pricing has to be offered to ensure users choose them. Prediction of duration and price of trips can help users to plan their trips properly, thus keeping potential margins for traffic congestions. It can also help drivers to determine the correct route which in-turn will take lesser time as accordingly. Moreover, the transparency about pricing and trip duration will help to attract users at times when popular taxi app-based vendor services apply surge fares. Thus in this research study, we used real-time data which customers would provide at the start of a ride, or while booking a ride to predict the duration and fare. This data includes pickup and drop-off point coordinates, the distance of the trip, start time, number of passengers, and a rate code belonging to the different classes of cabs available such that the rate applied is based on a regular or airport basis. Hereafter, we applied XGBoost and Multi-Layer Perceptron models to find out which one of them provides better accuracy and relationships between real-time variables. At last, a comparison of the two mentioned algorithms facilitates us to decide that XGBoost is more fitter and efficient than Multi-Layer Perceptron for taxi trip duration-based predictions.

✉ M Poongodi
dr.m.poongodi@gmail.com

Mohit Malviya
malviya.mohi2009@gmail.com

Chahat Kumar
chahatkumar3007@gmail.com

Mounir Hamdi
mhamdi@hbku.edu.qa

V Vijayakumar
vijayakumar.varadarajan@gmail.com

Jamel Nebhen
j.nebhen@psau.edu.sa

Hasan Alyamani
hjalyamani@kau.edu.sa

[1] College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

[2] Department of CTO 5G, Wipro Limited, Bengaluru, India

[3] CFO Technology, Enterprise Risk Function Technology, Bank of America, Chennai, India

[4] University of South Wales, Sydney, Australia

[5] College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University,
P.O. Box 151, Alkharj 11942, Saudi Arabia

[6] Department of Information Systems, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabia

Published online: 01 July 2021

Keywords Prediction algorithms · Xgboost · Multilayer perceptron · Entropy · Manhattan distance · Clusters · Rectified neural network · Mini Batch · Average rmse

1 Introduction

Earth is filled with an enormous population that tends to move from one place to another. Advancement in technologies had led to different ways of transportation. These include buses, autos and especially taxi services. New York City is one of the highly advanced cities of the world with extensive use of taxi services. Along with a vast population, the requirement of commonly available transportation serves the common purpose as it provides a very large transportation system. New York facilitates one of the largest subway systems in the world and comprises various green and yellow cabs which approximately count of around 13,000 taxis. Most of the population of New York depends upon public transport, and it has been estimated that 54 percent of the people do not own a car or a personal vehicle. As a matter of fact, it accounts for almost 200 million taxi trips per year.

The dataset we have used is available at Kaggle live, and its related information was collected over the years along with certain dependencies and provided to the public for further analysis. We used a collection of these datasets, which depicted around 3 years of NYC taxi trip data - about 15 lakhs records were considered, which carried the information of Taxi trip from January 2017 to January 2020.

Considering various Machine Learning models that provide reliable and improved accuracy for prediction-based use-cases, XGBoost and MLP are taken into consideration due to their novel potentiality to accumulate complex component conditions. Successful prediction of the taxi trip duration would eventually be much useful in the future to make better taxi trip duration predictions applicable to multiple cities.

1.1 XGBoost

XGBoost is short for "Extreme Gradient Boosting" which comes in association with various ensemble learning algorithms. It represents a flexible type of implementation where-in the concepts of decision trees (Gupta et al. 2020) get wholly acknowledged. Moreover, it is found to be much faster when compared to more common algorithms like Adaboost.

Further, it has recently dominated the machine learning world and gotten much attention in Kaggle competitions. Execution speed and Performance (Qureshi et al. 2020) are the two essential factors of using this algorithm in our work.

1.2 Multilayer Perceptron (MLP)

A Perceptron is considered as a linear classifier that produces a single output based on several linear functions. A multilayer perceptron (MLP) is a class of feedforward artificial neural network Sharma et al. (2020) which forms the basis for a deep learning platform. It encapsulates a deep artificial neural network that comprises more than one perceptron. This Artificial Neural Network mainly consists of nodes that use a non-linear activation function. It uses a backpropagation algorithm which gets classified under Supervised learning (Butgereit and Martinus 2019) methods. Non-linear data can also be separated using the multiple layers and the non-linear activation of the MLP, which makes it quite distinguishable (Kabán 2019) from a linear perceptron.

Thus, the contribution of this paper is as follows:

1. Since the duration of the taxi trip is highly dependent on the time at which the trip is made, the prediction becomes highly complex. In this regard, we have taken into account the time of the trip for reliable predictions. Also, we have excluded co-ordinates of locations present outside New York City because of their outlying nature. Using XGBoost equipped with K-Means clustering and given specific location, date, and time variables, we then analyzed and estimated the ride duration using real-time data which gets collected from various taxis.
2. The Multi-layer Perceptron model is used further to determine the relationship between various real-time labels and identities as taken from the data associated with different taxis.
3. Comparison between XGBoost and Multi-Layer Perceptron models is later done to determine which one of them comes off suitable and reliable for the mentioned New York City Taxi Prediction.

The remainder of the paper is organized as follows. In Sect. 2, we discussed the Related Work. New York City Taxi Duration Dataset Description is discussed in Sect. 3 followed by Methodology in Sect. 4. Simulation results and performance evaluation are provided in Sect. 5 before we summarize our concluding remarks in Sect. 6.

2 Related work

We studied a variety of different research works in the topics of Neural Networks, Multi-layer Perceptron, Bagging and Boosting, and other ML algorithms like AdaBoost and XGBoost for prediction-based methods. We tried to understand the methodology and workflow of each algorithm and how it would be beneficial to our project. The analysis of the research papers helped us to gain a number of possible insights, advantages, and disadvantages of the algorithms which could potentially provide the best solution for our problem statement. Based on the analysis, we reached a conclusion on how to work on the mentioned New York City Taxi Prediction use-case.

We started with (Ran et al. 2020) where speed and traffic stream were taken into account as the contribution to the model. The maximum places acquired by the K-means++ model and calibrations acquired by the XGBoost model are utilized to find out the Euclidean distance(ED). The base estimation of the calculated values gets utilized as the prescient estimation of the congestion level caused by different vehicles. As indicated by the forecast trial of I15-N interstate traffic information in PeMS information base, the joined model outstripped different models and the prescient exactness of the consolidated model came up at 94.47%. Further, (Liao et al. 2019) was considered where-in a load anticipating procedure dependent on XGBoost along with comparative days was proposed. This mechanism was used to break down the basic meteorological laws and everyday types based on the heap load. The XGBoost algorithm with the loss function and Taylor extension were added to the different quantitative terms to control the unnecessary fitting and intricacy. The charge-based and temperature-based information in a specific territory was completely taken as different sets of the test. The conclusions provided that the proposed XGBoost model can anticipate the heap-based load quite adequately.

To add more, (Wang et al. 2020) presented a driving conduct wellbeing assessment SVM-based mechanism which separated out the different values of distributive features to get the ideal order of hyperplane and afterward utilized the mathematical stretch as the assessment list for driving conduct wellbeing. Simultaneously, driving conduct, crowd-based streets, proficiency, sparing of energy, and climate factors with various other loads were considered, and thereafter it partitioned various driving conduct in four types: Good, Normal, Above the threshold, and Unfit level in view of SVM and KMeans. Subsequently, the XGBoost inherent mechanism was utilized. The test inferred the normal precision of 99.21% and the normal review rate of 98.5% which eventually demonstrated the whole operation was truly viable and attainable. To comprehend the innovations in XGBoost technology, (Cao et al. 2020) threw light upon a momentary traffic stream forecast model. This technique was dependent on best and worst inclination rise such that the analysis results uncovered the predominance of the whole system by contrasting it with the previous anticipation model.

Moreover, (Yang et al. 2020) was put into consideration as it reflected LC choice procedure that enabled vehicles based on autonomous ability to settle on human-like choices. This technique joined the XGBoost algorithm alongside a profound autoencoder (DAE) network-based technology. Initially, an autoencoder gets used to assemble a strong multi-component reformation structure utilizing time arrangement information from a different category of sensors. Thereafter, the recreated log errors pertaining to the DAE get prepared with other primary and secondary information, and as such, the whole process gets examined for LCID. Thus, the preparation of information extraction was made accordingly and at this point, to address the non-symmetric and multifactorial issue of the LC dynamic cycle, a Bayesian boundary enhancement with an XGBoost calculation came into the effect. In the interim, to completely prepare the learning model with a huge scope of data information sets, a proposal of a web-based preparing methodology was furnished to refresh the model boundaries with information clusters. The exploratory outcomes delineated that the given model can precisely distinguish the LC conduct of vehicles. Moreover, when information of similar parameters was added, the whole structure accomplished preferable execution over other mainstream techniques.

In order to understand the holistic environment of XGBoost, (Montiel et al. 2020) was taken which introduced a transformation of XGB for characterization of developing information-based varied streams. Here, new information gets shown up over the long haul and the connection between the highlights and varied-classes was getting changed simultaneously. This technique made new individuals of the entity based on ensembling as new entry points which later gets opened up as set by the required changes. The greatest group size was allowed to be fixed, yet the process of learning various features didn't get stopped in light of the fact that the model was refreshed on new information to guarantee uniformity with the latest ideas. Likewise, an investigation of the utilization of drift concept identification was done to activate a component so as to refresh the group. Testing of the technique on manufactured information with drift identification was made available and later, it was differentiated against other methods of classification for information streams. The results proved out to have a powerful impact produced by the proposed idea over other previous methodology used.

To familiar with the knowledge of Multi-Layer perceptron, (Ayyappa et al. 2020) was used where-in a computerized Tumor recognition procedure was proposed which helped various doctors in recognizing cerebrum tumors. Here, a solidarity MLP based Gaussian Filtering alongside BP Neural Network was evaluated which delivered good precise outcomes while distinguishing the cerebrum tumor with an exactness pace of 93% when contrasted with different classification methods like SVM and PNN. Likewise, (Sunindyo and Satria 2020) investigated out the likelihood to utilize the CCTV film so as to perform anticipation based on regular traffic-data. The recording was prepared consequently utilizing detection and tracking of the object-based procedures to get adequate traffic information points. From that point forward, the information dedicated to traffic entry points (Suresh et al. 2021) was demonstrated by encompassing both LSTM and MLP. The efficiency of the whole structure was estimated by utilizing RMSE which in-turn provided accurate high-level information from the given data. This investigation demonstrated that prepared CCTV film is in fact a practical alternative for gridlock expectation. The best model accomplished 1.88 RMSE by measures of vehicles, transports, and trucks as an anticipated variable with a fortified MLP strategy.

To enhance the idea of MLP, (Khamees et al. 2020) was utilized to find out another methodology for preparing the MLP in light of the crow-search streamlining mechanism. The primary target of this methodology was to diminish varied shortcomings to its base level and increment the pace of the classifying process. The marked threshold of the given execution was accomplished by fabricating distinct typical datasets for the process of classification. As such, it was also done to guarantee that the nature of the outcome remains high, and additionally, this mechanism was later contrasted with other classification algorithms, for example, ACO, GA and PSO. The results showed up that the search based on crow streamline calculation was most accurate as it delivered the most elevated precision rate and tackled the improvement and optimization issue effectively.

Hereafter, (Wu et al. 2019) was availed to acknowledge another compounded variable choice mechanism for non-symmetric MLP process. The provided operation used some garrote-based conceptualization on non-negative numerical values to pack the different weights pertaining to the MLP structure. Weights that provided zero subordinate factors as input were taken out from the underlying information. Then, a factor determination was done by using optimization calculation which got carried out on extremal parameters. The new factor choice calculation was then coordinated out which combined a great determination capacity dedicated to NNG and the exact nearby capacity of EO. Lastly, two instances of informational collections and a modern debutanizer application were actualized to show the efficiency of the new structure. The outcome exhibited that the created approach presented a much greater execution alongside the variable which provided fewer input data than the other variable decision strategies.

While the prediction-based algorithms becomes certainly important, Irio et al. (2021) suggested a model which transformed the directions information of the vehicles dependent on successive areas related to GPS and built an ethical-measurable surmising algorithm which in-turn was utilized in accordance with the portability expectation at an online level. Here, the surmising algorithm was dependent on Markov based secret model (HMM) such that every direction got demonstrated in terms of subset based on discrete/continuous areas. Besides, the forecast model utilized various measurable data construed up until this point and subordinated extensively on the calculation of the Viterbi mechanism that recognized the provided multiple subsets rooted on discrete/continuous areas. Along with it, the most extreme probability of numerous earlier subsets-based areas was supported to establish valuable prediction means. Additionally, a hybrid deep neural network prediction model was proposed by (Duan et al. 2019) which majorly proceeded on the idea of convolutional LSTM (ConvLSTM) techniques. Moreover, multiple certain connections between OD's stream and movement's time were investigated which later was joined for the contributions of the forecast algorithm. It also presented a lattice and street-settled technique to address ODs streams forecast around numerable street-based network degrees and tackled different issues that can't help in recognizing stream-related traffics by using grid-based representation at various statures.

In addition to above, (Zhang et al. 2020) exhibited a learning-model based on various parallel tasks such that it contained three equalized-parallel layers of LSTM for co-foreseeing pickup and drop-off taxi requests. It also helped in contrasting multiple exhibitions of expectation procedures related to single interest and co-forecast strategy requests associated with two interest-based parameters. Exploratory outcomes on provided datasets showed the imperative and extensive dependence of pickup and drop-off requests upon one another which in-turn delivered solidarity governing adequacy based on the suggested co-forecasting strategies. Furthermore, (Kankanamge et al. 2019) utilized the sophisticated idea of gathering several taxi time-based travel directions connected with static parameters. It then involved isolated-based XGBoost models with respect to regression conditions alongside the above-mentioned data. Here, a bunch of extraordinary molded excursions and distinguished inlier were discretely differentiated with the use of prevailing leading algorithms. This permitted to furnish of the impressive prediction techniques of the XGB-IN prototype such that it produced less root mean squared error and mean absolute error in accordance with the real-world time travel figures. Further, it also facilitated to provide models based on XGB-Extreme mechanisms which gave sensibly precise expectation outcomes to a bundle of maximal-configured journeys accompanied by limited real-time taxi rides.

Consequently, (Maddikunta et al. 2020) investigated a robust ML linked random forest regression model towards the prediction of IoT gadgets-based battery life. As accordingly, a few techniques related to the data pre-processing like dimensionality reduction, normalization, and transformation were utilized for the model which in-turn attained a predictive exactness of about 97% across all the various scenarios. It was also demonstrated that the evaluated model gained better performance in sustaining the battery life of IoT gadgets as compared to existing state-of-art regression-based algorithms.

A better understanding of the methodology useful for the prediction can be provided using (Poongodi et al. 2020a) where it employed maximum likelihood estimation to formulate the probabilities using the Logistic Regression Model. Here, an iterative-based regression algorithm was set to take place on all of the classes such that at least each of them was counted for various prediction structures. Later, (Poongodi et al. 2020b) was studied which encompassed a Decentralized Autonomous Organization (DAO) to create a wholly sustainable and tidy community predictive development throughout the real-time world settings. Accompanying the use of the ML algorithms, (Poongodi et al. 2020c) enhanced and improved the predictive monetary situation of all individuals connected officially with the different clusters of establishments and businesses by utilizing a model in-together which included various ML algorithms such as Hierarchical clustering, Decision tree, KNN clustering, etc. Extending different ideas, [24-25] reused or retransformed Linear SVM technology by using the prediction of any two given observations rather than the observations themselves. This accompanied to provide better and superior results for their researched use-cases. A predictive-based recommendation system was used in (Poongodi et al. 2019) where-in complex and normalized XGBoost Algorithms were used for the user credibility parameters. A number of factors based on the purchase and review history of the users were taken into consideration to develop a smooth and flexible prediction recommendation system.

In order to explore more about prediction operations, (Alazab et al. 2020) extended the use-case of the smart grid CPS mechanism by incorporating various schemas coupled with the Multidirectional Long Short-Term Memory (MLSTM) technique. This was done in order to allow the accurate prediction of the smart grid network stability matrices. Comparison between the existing best Deep Learning methods like RNN Guo et al. (2020), GRU, conventional LSTM, etc., and the suggested MLSTM procedure showed that the latter outflanks (Kashif et al. 2020) various other ML prediction-based models. At last, (Muhammad et al. 2021) was chosen which applied multiple supervised ML algorithms like SVM, naive Bayes, CNN, RNN, logistic regression, decision tree, etc. on epidemiology-based real-world labeled Coronavirus dataset so as to detect COVID-19 disease. A major part of the procedure was carried out to clean the data which benefited to find out strong correlations between independent and dependent features of the chosen dataset. Based on the critical analysis of various ML approaches, it was found that the decision tree model accomplished the best accuracy of 94.99% in comparison to other techniques.

Thus after careful analysis, we discovered several miscellaneous and mixed drawbacks in the variegated models that were hybridly used in the prediction mechanisms. Supervised Machine Learning models such as Decision tree and random forest classification/regression were found to be superior to others in terms of their sensitivity, specificity, and accuracy due to which the idea of using XGBoost is taken further for the New York City Taxi Prediction use-case. Moreover, the presence of using K-means clustering with XGBoost Model (Tang et al. 2020) over the rest of the Unsupervised ML techniques was noticed because of its convergeable, scalable and adaptable properties. Subsequently, the employment of the Multi-Layer Perceptron is involved in the second part of this research paper since it turned out to provide higher heteroskedasticity and an added advantage of solving complex and non-linear problems. Following the standards of the neural network, MLP based models aid to deduce hidden interconnections within the real-time multiplex datasets (Tang et al. 2020) which eventually supports in making out efficient and improved methods (Chinmay and Rodrigues Joel 2020) for the mentioned taxi prediction application.

3 Dataset description

New York City Taxi Duration dataset is taken from the Kaggle website which provides free access to complex challenges. This dataset helps us to predict the trip duration of a taxi ride taking into account the different factors that affect the ride duration. Along with the above-mentioned, one more dataset gets included which involves the climatic conditions of the city. Both of these datasets are combined using pre-processing techniques to create a single dataset that can be used further for accurate trip duration prediction. Some of the important attributes of the dataset are discussed below:

- **id**, which provides a unique identification to a trip.

- **vendor id**, a unique code which gets assigned to the different cab companies.
- **pickup datetime**, starting statistics of the pickup.
- **dropoff datetime**, ending statistics of the pickup.
- **passenger count**, passengers travelling in a particular trip.
- **pickup longitude**, longitudinal location of the pickup.
- **pickup latitude**, latitudinal location of the pickup.
- **dropoff longitude**, longitudinal location of the drop off.
- **dropoff latitude**, latitudinal location of the drop off.
- **store and fwd flag**, a code to identify whether the data is stored on the device and then gets forwarded to the database.
- **trip duration**, the total time of the trip in seconds.

The second dataset comprises the climatic data of the city which includes vital information such as the time of rainfall, sunlight, and various other factors which can be used for better prediction of the taxi trip.

4 Proposed methodology

Our kernel is written and developed using iPython Notebook and XGBoost model with the assistance of a mini-batch K-means clustering algorithm. The workflow of the kernel includes the following steps:

1. First of all, importation of all the necessary libraries including Sklearn library is done.
2. Both the dataset gets imported accordingly in order to analyse the various attributes of the taxi trip duration.
3. Mathematical values such as standard mean, variance and quartiles of all the features is then find out to gather multiple parameters. While calculating various constraints, careful attempts to avoid any type of mismatch gets regularly checked as needed.
4. Thereafter, the Mini batch clustering gets utilised which provides highly susceptible to outliers. Cleaning of the data to remove the outliers is accordingly employed so that the above algorithm starts to work efficiently.
5. The cleaned data is then analysed deeply for more feature extraction by finding out the correlation in the data which ensures maximum coverage.
6. Computation of three different distances i.e. manhattan, haversine and bearing distance between the pickup and drop off location gets evaluated. Manhattan distance gives the straight line distance between the specified coordinates. But since the earth is round and taking into account the straight line distance is like neglecting an important aspect of the route, so as a result, Haversine Distance gets employed extensively. Moreover, Bearing direction is used to calculate the angular distance between various point of interest.
7. Hereafter, the average of all the three distances is calculated and added to the cleaned dataset as extracted features which in-turn gets further used for critical analysis.
8. Next, Mini-batch K-means algorithm is applied to cluster points on basis of the pick-up latitude, pick-up longitude, drop off latitude and drop off longitude variables. Later, the clusters obtained are used to find out their centres and subsequently, the trips are divided according to above-mentioned clusters parameters. Additionally, these area-based clusters are added as an extra feature to the dataset.
9. As a result, addition of about 200 features gets accomplished in the form of cluster centres. Here, the added features mainly includes 100 pick up and 100 drop off clusters points.
10. Finally, the redundant columns are removed and the associated back-bone of the kernel i.e. XGBoost model gets applied to the dataset with the added parameters. Henceforth, several results are observed for the taxi-based prediction values.

A similar methodology for multi-layer perceptron is also followed which includes importing libraries and datasets (incorporating external data for improving accuracy), pre-processing the imported datasets, and so on. Rectified Neural Networks are then applied to eliminate outliers appropriately. At last, the application of linear neural networks is performed to get the desired results.

5 Results and discussion

5.1 XGBoost

As shown in Fig. 1, we plot a simple histogram of the trip duration by throwing the data into 100 bins. Binning involves taking the data's maximum and minimum points, subtracting them to get the length, dividing the calculated length by the number of bins to get the interval length, and finally grouping the data points into mentioned intervals.

Further, a Gaussian curve-based graph as shown in Fig. 2 gets plotted which aids to determine an insightful relationship between various taxi trips and the logarithm of trips duration. This also provides an intuitive pattern understanding of how taxi services work in New York City.

It is very important for us to find out whether the training and testing data are in agreement with each other or not. By this, we mean that we need to calculate the said

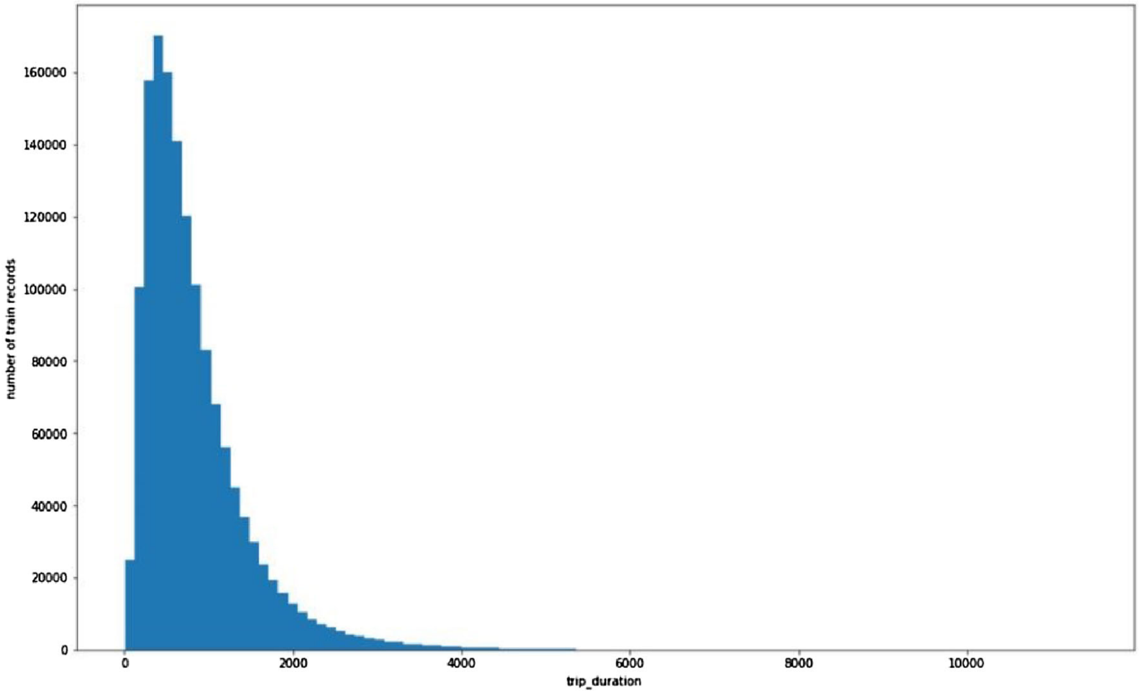


Fig. 1 Number of training records vs Trip duration

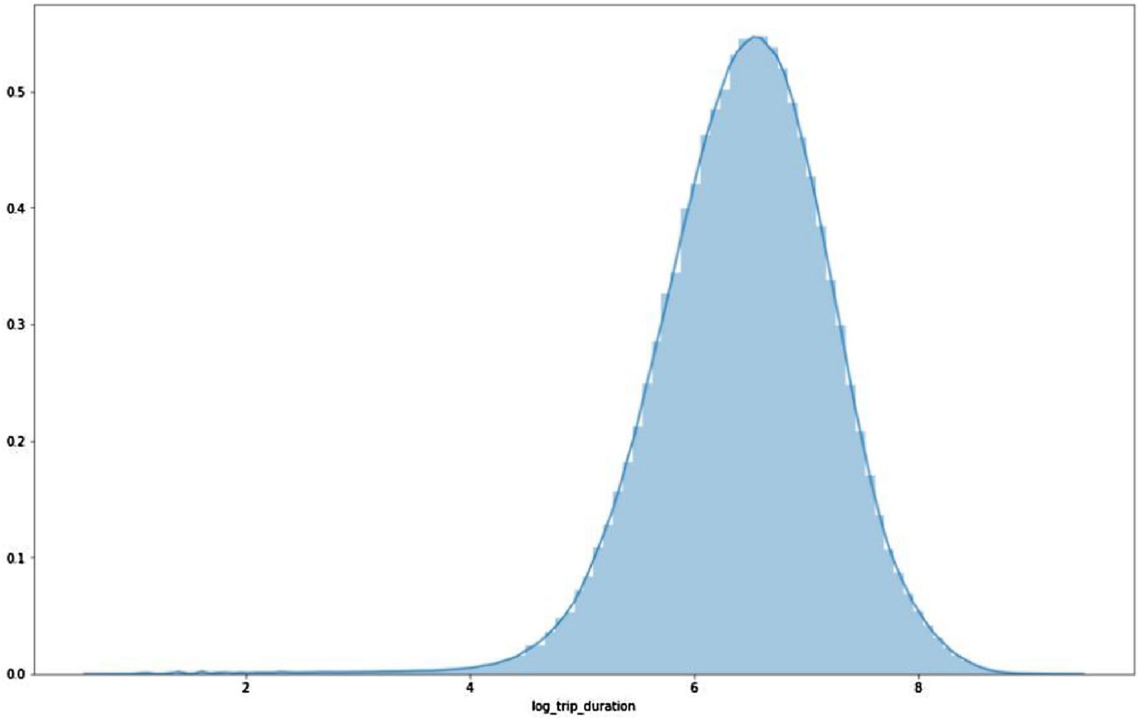


Fig. 2 Logarithmic Trip duration

parameters using a Time series graph that tell us how well are the number of trips over time-varying parameters in accordance with the training and testing dataset. As a result, we simply plot a time-series line graph of both the test and training data to not only look into identifying possible trends but to also see if both datasets follow the same pattern shape which is seen in Fig. 3.

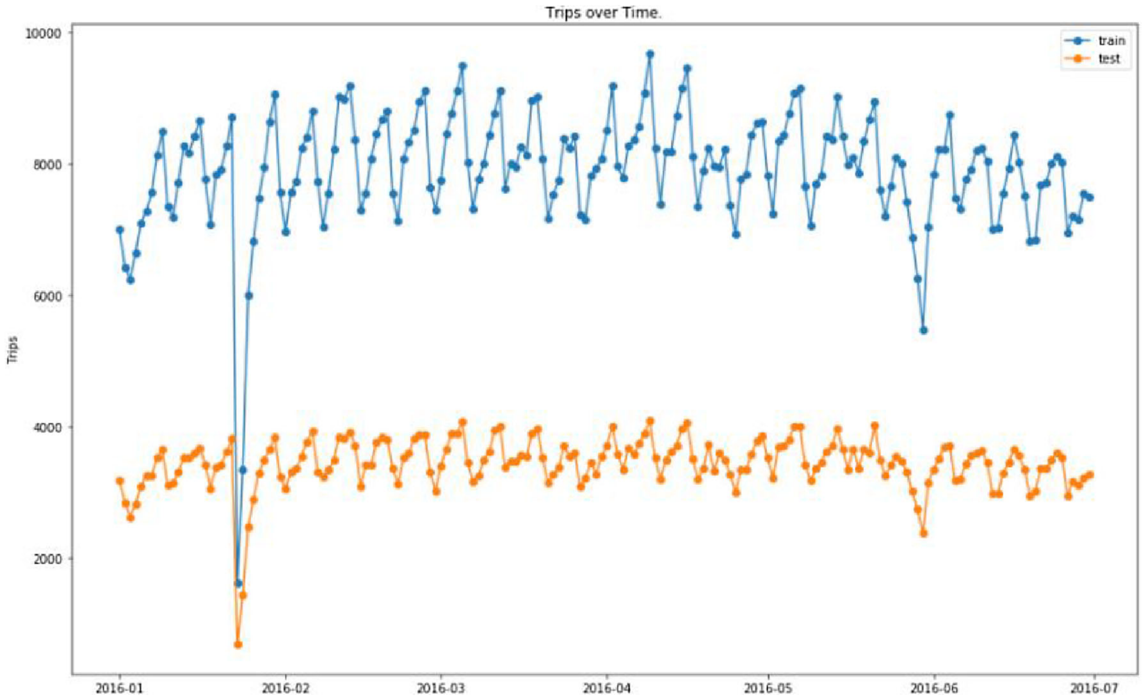


Fig. 3 Comparison of Training and Testing Datasets

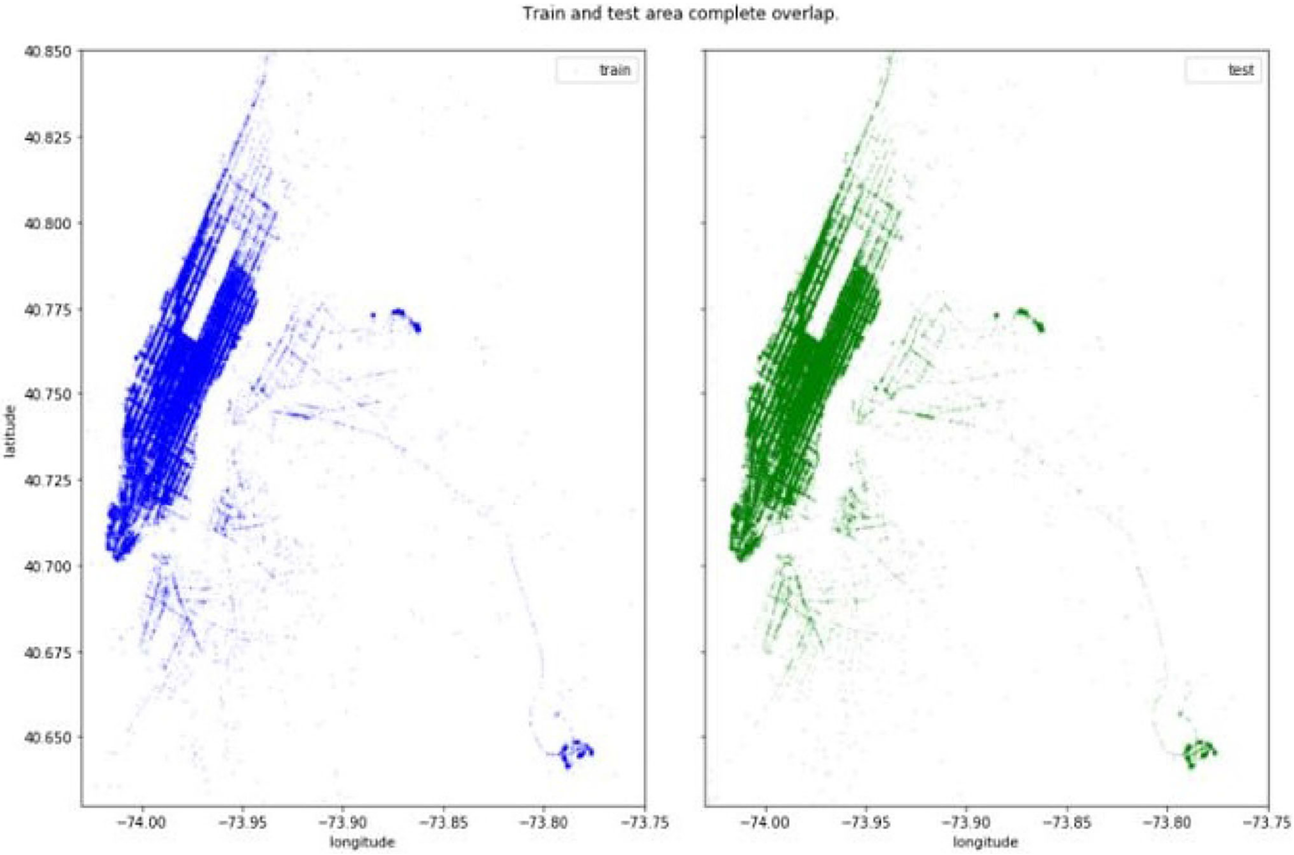


Fig. 4 Comparison of pickup and dropoff points on the map of New York

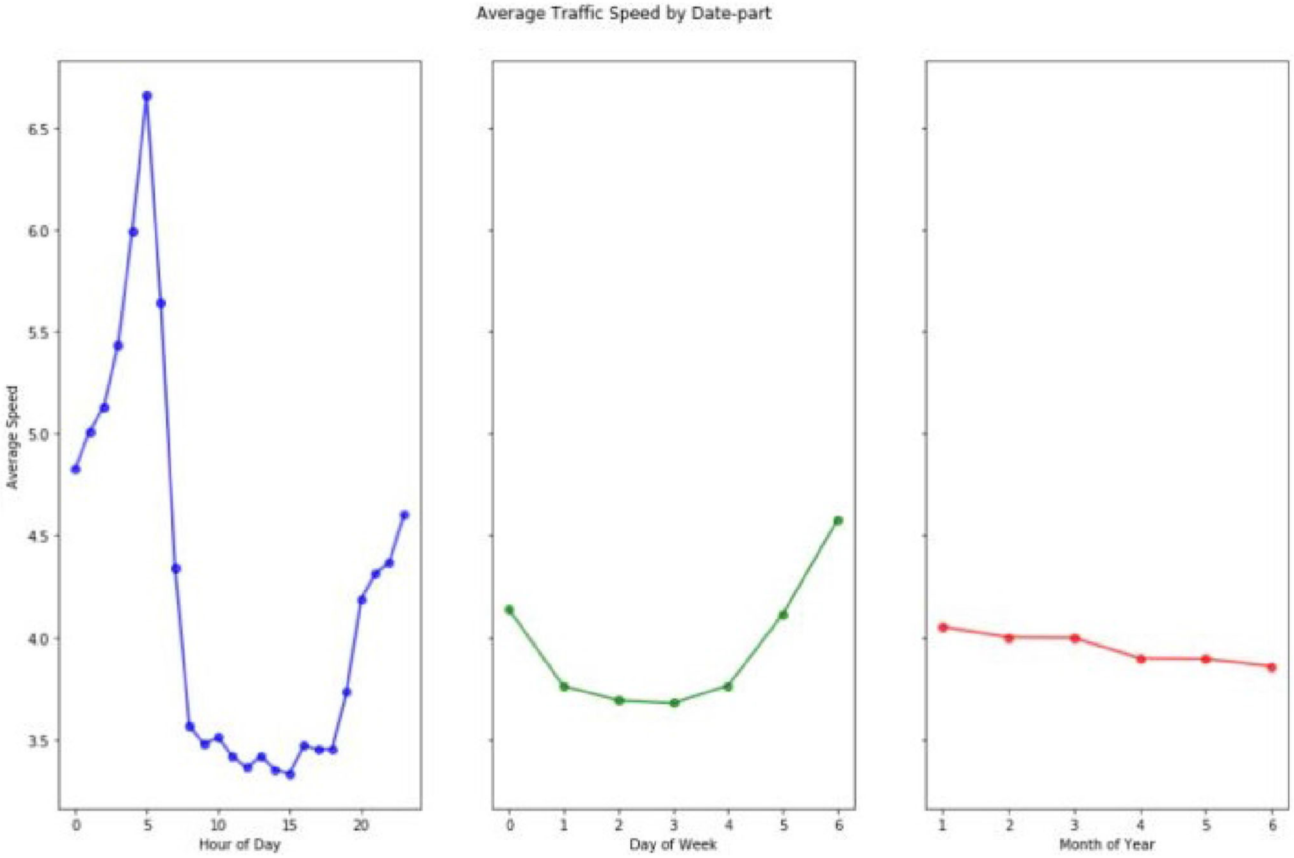


Fig. 5 Average speeds

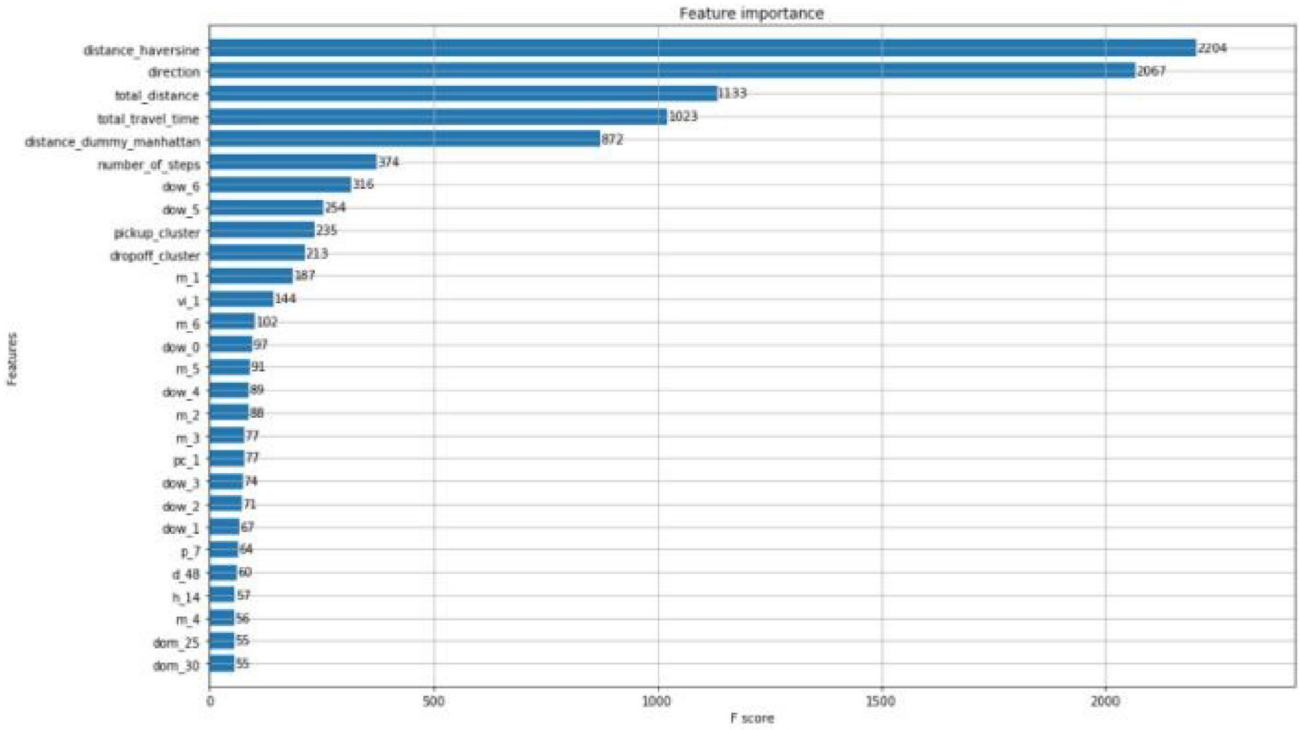
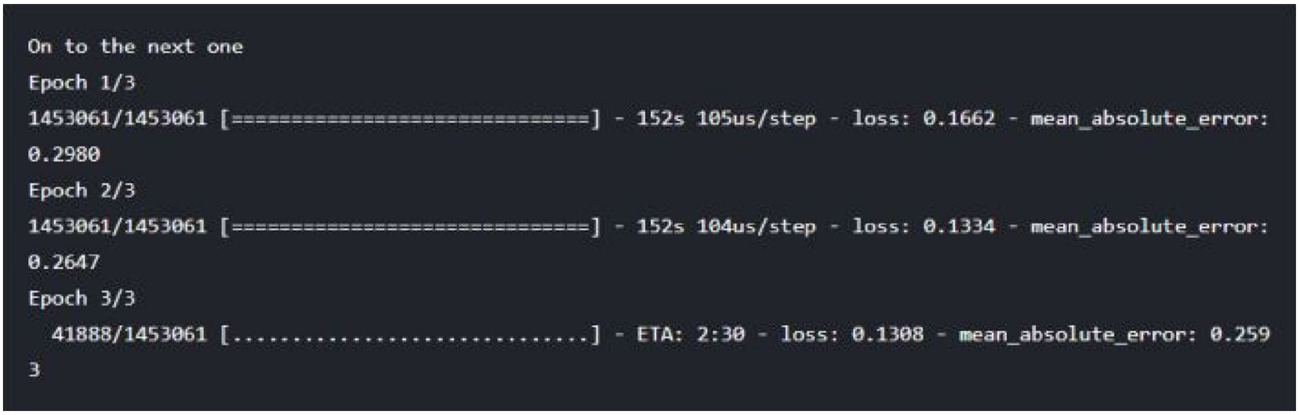


Fig. 6 Feature Importance Graph



```
On to the next one
Epoch 1/3
1453061/1453061 [==============================] - 152s 105us/step - loss: 0.1662 - mean_absolute_error:
0.2980
Epoch 2/3
1453061/1453061 [==============================] - 152s 104us/step - loss: 0.1334 - mean_absolute_error:
0.2647
Epoch 3/3
  41888/1453061 [..............................] - ETA: 2:30 - loss: 0.1308 - mean_absolute_error: 0.259
3
```

Fig. 7 Results of MLP Algorithm

Next, we utilize the New York City map border coordinates in the kernel to create the canvas where-in the coordinate points get suitably graphed. Here, a simple scatter plot is precisely used to display the actual coordinates. It helps to show whether the pick-up points in the training and testing datasets overlap each other in some manner or not. This gets shown in Fig. 4.

After that, we plot three different graphs for average speed of a taxi based on different hours of a day, different days of a week and different months in the year. This is shown in Fig. 5.

Finally, we visualise the feature importance graph as seen in Fig. 6 to see which features amongst all are most relevant and required for getting accurate results.

Successively, we run the XGBoost algorithm with the parameters shown below. Mentioned parameters can be changed as desired but before setting them out, one must study about XGBoost documentations as it greatly helps in understanding about how to fine-tune the parameters for better performance and efficiency. Accordingly, the features included are:

- max depth = 6
- learning rate = 0.09
- iteration = 250

After running the algorithm, we get to infer that the average RMSE value over 250 iterations is about 0.39 for the training dataset and 0.44 for the testing dataset.

5.2 MLP

Similarly, we employ the Multi-Layer Perceptron model on a similar dataset. It essentially requires a deep learning setup using Rectifier to eliminate outliers from the data. As such, results are shown in Fig. 7.

The training accuracy of this algorithm is observed to be around 0.2740, while the testing accuracy sets out near 0.41. This precisely shows that XGBoost is slightly better than MLP model.

6 Conclusion

We are successfully able to implement both of the algorithms on the New York City Taxi Trip Duration dataset and able to draw certain conclusions from several inferences. After implementing both of the algorithms, we come across that XGBoost is better than MLP as it shows a slightly good accuracy than the latter one. This in turn helps to conclude that XGBoost Model is more efficient and reliable in predicting the taxi trip duration as compared to MLP.

As a part of the future work, the Multi-layer Perceptron model could be auto-tuned to further learn and determine which features need to get joined to detect numerous interactions between them as needed. Moreover, variabilities and quantities related to the various location features might also be computed in the upcoming research in order to localize the traffic-based effects on the taxi prediction coordinates. Speed limitations-based features could later be incorporated alongside to comprehend better analysis of the datasets. Further, New York Central Park and the associated weather conditions could also be closely taken care of as New Yorkers might take a taxi when they are near Central Park or when the weather condition is severe, but not when they are near Central Park and it is raining, since they may not visit the park in bad weather. At last, enhancements to the K-Means Clustering algorithm could be provided by encompassing additional features such as distance to the closest metro station, number of bars and eateries in a given zone, etc. so as to exploit comparative qualities belonging to various zones. This would also ensure the rightful evaluation of various clusters in which each data point falls such that it fills in as an extra vital element for our models.

References

Alazab M, Khan S, Krishnan SSR, Pham Q, Reddy MPK, Gadekallu TR (2020) A Multidirectional LSTM Model for Predicting the Stability of a Smart Grid. IEEE Access 8:85454–85463

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

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). July, p, Coimbatore, India

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

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, Netherlands

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

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

Guo Z, Shen Y, Bashir AK, Imran M, Kumar N, Zhang D, Yu K (2020) Robust spammer detection using collaborative neural network in internet of thing applications. IEEE Internet of Things J 1–1. https://doi.org/10.1109/JIOT.2020.3003802

Gupta A, Sharma S, Goyal S, Rashid M (2020) "Novel XGBoost Tuned Machine Learning Model for Software Bug Prediction', 2020 International Conference on Intelligent Engineering and Management (ICIEM). United Kingdom, London

Irio L, Ip A, Oliveira R, Luís M (2021) An adaptive learning-based approach for vehicle mobility prediction. IEEE Access 9:13671–13682

Jeyachandran A, Poongodi M (2018) Securing Cloud information with the use of bastion algorithm to enhance confidentiality and protection. Int J Pure Appl Math 118(24)

Kabán Ata (2019) "Compressive Learning of Multi-layer Perceptrons: An Error Analysis", In: 2019 International Joint Conference on Neural Networks (IJCNN). Budapest, Hungary

Kankanamge KD, Witharanage YR, Withanage CS, Hansini M, Lakmal D, Thayasivam U (2019) "Taxi trip travel time prediction with isolated XGBoost Regression", In: 2019 Moratuwa Engineering Research Conference (MERCon). Moratuwa, Sri Lanka, pp. 54–59

Kashif BA, Suleman K, Rabadevi B, Deepa N, Alnumay WS, Gadekallu TR, Maddikunta PKR (2020) "Comparative analysis of machine learning algorithms for prediction of smart grid stability", Int Trans Electr Energy Syst, Feb

Khamees M, Ahmed WS, Abbas SQ (2020) "Train the Multi-Layer Perceptrons Based on Crow Search Algorithm", In: 2020 1st. Information Technology To Enhance e-learning and Other Application (IT-ELA), Baghdad, Iraq, July

Koo J, Faseeh QNM, Siddiqui IF, Abbas A, Bashir AK (2020) IoT-enabled directed acyclic graph in spark cluster. J Cloud Comput 9(1):1–5

Liao X, Cao N, Li M, Kang X (2019) "Research on Short-Term Load Forecasting Using XGBoost Based on Similar Days", In: 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). Changsha, China

Maddikunta PKR, Srivastava G, Gadekallu TR, Deepa N, Boopathy P (2020) Predictive model for battery life in IoT networks. IET Intel Transport Syst 14(11):1388–1395

Montiel J, Mitchell R, Frank E, Pfahringer B, Abdessalem T, Bifet A (2020) "Adaptive XGBoost for Evolving Data Streams", In: 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, United Kingdom

Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chinmay C, Mohammed IA (2021) Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN Comput Sci

Poongodi M, Ashutosh Sharma, Vijayakumar V, Vaibhav Bhardwaj, Parkash Sharma Abhinav, Razi Iqbal, Rajiv Kumar (2020c) Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. Comput Electr Eng 81:106527

Poongodi M, Vijayakumar V, Chilamkurti N (2020a) Bitcoin price prediction using ARIMA model. Int J Int Technol Secured Trans 10(4):396–406

Poongodi M, Vijayakumar V, Rawal B, Bhardwaj V, Agarwal T, Jain A, Ramanathan L, Sriram VP (2019) Recommendation model based on trust relations & user credibility. J Intell Fuzzy Syst 36(5):4057–4064

Poongodi M, Hamdi M, Vijayakumar V, Rawal BS (2020b) and ", 2020 IEEE 3rd 5G World Forum (5GWF). Bangalore, India, pp 1–6

Ran D, Jiaxin H, Yuzhe H (2020) "Application of a Combined Model based on K-means++ and XGBoost in Traffic Congestion Prediction", In: 2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA). Zhangjiajie, China

Sharma R, Schommer C, Vivarelli N (2020) "Building up Explainability in Multi-layer Perceptrons for Credit Risk Modeling", In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Australia, Sydney

Sunindyo WD, Satria ASM (2020) "Traffic Congestion Prediction Using Multi-Layer Perceptrons And Long Short-Term Memory", In: 2020 10th Electrical Power. Electronics, Communications, Controls and Informatics Seminar (EECCIS), Malang, Indonesia

Suresh P, Sundresan P, Mujahid T, Ganthan N, Chinmay C, Saju M, Zeeshan B, Mohammad TQ (2021) ANN base novel approach to detect node failure in wireless sensor network, CMC-Computers. Tech Science Press, Materials & Continua

Tang Q, Xia G, Zhang X, Long F (2020) "A Customer Churn Prediction Model Based on XGBoost and MLP", In: 2020 International Conference on Computer Engineering and Application (ICCEA). Guangzhou, China

Wang X, Lou XY, Hu SY, He SC (2020) "Evaluation of Safe Driving Behavior ofTransport Vehicles Based on K-SVM-XGBoost", In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). Shenzhen, China

Wu X, Li Y, Wu H, Zhang F, Sun K (2019) "A hybrid variable selection algorithm for multi-layer perceptron with nonnegative garrote and extremal optimization", In: 2019 19th International Conference on Control, Automation and Systems (ICCAS). Jeju, Korea (South)

Yang B, He Y, Liu H, Chen Y, Ji Z (2020) “A Lightweight Fault Localization Approach based on XGBoost”, 2020 IEEE 20th International Conference on Software Quality, Reliability and Security (QRS). Macau, China

Zhang C, Zhu F, Wang X, Sun L, Tang H, Lv Y (2020) Taxi demand prediction using parallel multi-task learning model. IEEE Trans Intell Trans Syst 1–10. https://doi.org/10.1109/TITS.2020.3015542

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.