**Neural Network Deep Learning**

**New York Taxi Trip Duration**

Project Proposal

Kishor Kumar Andekar ( 700744713 )

Niteesh Kumar Pondugula ( 700763258 )

Sai Preetham Nagaswaram

https://link.springer.com/article/10.1007/s13198-021-01130-x

## 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.

## 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR9)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR16)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR26)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR4)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR12)) from a linear perceptron.

Thus, the contribution of this paper is as follows:

1. 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.

1. 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.

1. 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#Sec4), we discussed the Related Work. New York City Taxi Duration Dataset Description is discussed in Sect. [3](https://link.springer.com/article/10.1007/s13198-021-01130-x#Sec5) followed by Methodology in Sect. [4](https://link.springer.com/article/10.1007/s13198-021-01130-x#Sec6). Simulation results and performance evaluation are provided in Sect. [5](https://link.springer.com/article/10.1007/s13198-021-01130-x#Sec7) before we summarize our concluding remarks in Sect. [6](https://link.springer.com/article/10.1007/s13198-021-01130-x#Sec10).

## 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR25)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR17)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR30)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR5)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR32)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR19)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR3)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR27)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR28)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR15)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR31)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR10)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR7)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR33)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR13)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR18)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR22)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR24)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR21)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR23)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR1)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR8)), GRU, conventional LSTM, etc., and the suggested MLSTM procedure showed that the latter outflanks (Kashif et al. [2020](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR14)) various other ML prediction-based models. At last, (Muhammad et al. [2021](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR20)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR29)) 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](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR29)) which eventually supports in making out efficient and improved methods (Chinmay and Rodrigues Joel [2020](https://link.springer.com/article/10.1007/s13198-021-01130-x#ref-CR6)) for the mentioned taxi prediction application.

Problem Statement:

The problem we aim to address is accurately predicting the duration of taxi trips in New York City. Efficiently estimating trip duration is crucial for both taxi drivers and passengers, as it helps in planning routes, managing schedules, and improving overall service quality. For example, predicting trip durations in New York City accurately can be challenging based on various factors such as traffic congestion, pickup location, drop-off location, time of day, weather conditions, and route complexities.etc.

-----------

Second approach:

**Title: Predicting New York City Taxi Trip Durations**

**Abstract:**  
This paper presents a predictive modeling approach to estimate the durations of taxi trips in New York City. The proposed model leverages a stacked ensemble technique combining XGBoost and a Feedforward Neural Network (FNN) to effectively capture the complex relationships between various features and trip durations. The model's performance is evaluated using standard regression metrics on a real-world dataset, demonstrating its effectiveness in accurately predicting taxi trip durations.

**Keywords:** New York City, taxi trips, trip duration prediction, predictive modeling, stacked ensemble, XGBoost, Feedforward Neural Network (FNN).

**I. Introduction**  
**A. Background:**  
Transportation plays a crucial role in urban environments, with taxis being a popular mode of transportation in cities like New York City (NYC). Accurately predicting taxi trip durations is essential for improving service quality, optimizing routes, and enhancing user experience.

**B. Problem Statement:**  
The variability in factors such as traffic congestion, weather conditions, and route complexities makes predicting taxi trip durations in NYC a challenging task. This paper addresses this problem by developing a predictive model that can estimate trip durations with high accuracy.

**C. Objectives:**  
The primary objective of this study is to develop a robust predictive model capable of accurately estimating taxi trip durations in NYC. Specifically, we aim to:

* Investigate the effectiveness of a stacked ensemble approach combining XGBoost and a Feedforward Neural Network (FNN) for trip duration prediction.
* Evaluate the performance of the proposed model using real-world taxi trip data.

**II. Related Work**  
This section provides an overview of existing literature and research efforts related to taxi trip duration prediction, including traditional regression methods, machine learning techniques, and deep learning approaches.

**III. Data Collection and Preprocessing**  
**A. Data Sources:**  
The dataset used in this study consists of historical taxi trip records obtained from [source]. It includes features such as pickup and drop-off locations, timestamps, and trip distances.

**B. Data Preprocessing:**  
Prior to model training, the dataset undergoes preprocessing steps, including feature engineering, handling missing values, and normalization.

**IV. Model Architecture**  
**A. XGBoost (Base Model 1):**  
This subsection describes the architecture and training process of the XGBoost model used as the first base model in the stacked ensemble approach.

**B. Feedforward Neural Network (Base Model 2):**  
Here, the architecture and training procedure of the Feedforward Neural Network (FNN) employed as the second base model are discussed.

**C. Stacked Ensemble (Meta Model):**  
The meta-model, which combines predictions from the base models to generate the final prediction, is explained in this subsection.

**V. Experimental Results**  
**A. Evaluation Metrics:**  
This subsection details the evaluation metrics used to assess the performance of the proposed model, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

**B. Experimental Setup:**  
Information regarding data splitting, model training, and hyperparameter tuning is provided in this subsection.

**C. Results:**  
The results of the experiments, including performance metrics and model comparisons, are presented and analyzed in this subsection.

**VI. Discussion**  
This section discusses the implications of the experimental results, including the strengths and limitations of the proposed model, as well as potential areas for improvement.

**VII. Conclusion**  
In conclusion, this paper presents a predictive modeling approach for estimating taxi trip durations in New York City. The stacked ensemble model combining XGBoost and a Feedforward Neural Network demonstrates promising performance in accurately predicting trip durations, highlighting its potential for improving transportation services in urban environments.

**VIII. References**  
[List of references cited throughout the document using IEEE citation style.]