IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

New York City Taxi Trip Duration Prediction Using Machine Learning

Kapil saini[1],
Meerut Institute of Engineering Technology, India

Abstract: Accurately predicting taxi ride durations is crucial for optimizing the efficiency of taxi dispatch systems in ride-hailing services. This research aims to develop and evaluate predictive models that estimate taxi ride durations before the start of the trip, leveraging historical ride data. Various machine learning techniques, including linear regression, K-Nearest Neighbors (KNN), Ridge regression, and Lasso regression, are explored using data from New York City taxi trips.The study aims to understand the key factors influencing ride durations and build models that can provide accurate estimates. Accurate predictions can improve dispatch decisions, reduce idle times for drivers, and decrease wait times for passengers, thus enhancing operational efficiency and service quality in the ride-hailing industry.We collect a comprehensive dataset of taxi trips, enriched with external data such as weather conditions and traffic information. The data undergoes preprocessing steps like cleaning, normalizing, and feature engineering. Key variables influencing ride durations include trip distance, pickup time, day of the week, weather conditions, and traffic patterns.The modeling approach involves developing predictive models, including a benchmark linear regression model, a KNN model, and regularized linear models such as Ridge and Lasso regression. Hyperparameter tuning optimizes model performance, and cross-validation ensures robustness. Performance is evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

The linear regression model serves as a baseline, capturing some variability in ride durations but limited in modeling non-linear interactions. The KNN model improves performance by considering data point similarity, while Ridge and Lasso regression models enhance accuracy by incorporating regularization terms. Lasso also performs feature selection by setting some coefficients to zero.This study highlights the importance of feature engineering and real-time data integration. Future research should focus on integrating real-time data sources, such as live traffic updates and weather conditions, to improve prediction accuracy further. Advanced machine learning techniques, such as ensemble methods and deep learning, should be explored to capture more complex relationships in the data.Ethical considerations and data privacy are critical in the use of predictive models in ride-hailing services. Ensuring passenger and driver privacy, and maintaining transparency and fairness in algorithmic decision-making, are essential for responsible use of data-driven models.

I. INTRODUCTION

The rise of ride-hailing services such as Uber, Lyft, and Ola has revolutionized urban transportation by providing a convenient and cost-effective alternative to traditional taxis. These services rely on advanced technology to connect passengers with drivers, offering door-to-door transport with greater efficiency and flexibility. However, one of the critical challenges in managing ride-hailing services is optimizing the efficiency of taxi dispatch systems. Accurate prediction of taxi ride durations is essential for improving dispatch decisions, reducing idle times for drivers, and minimizing wait times for passengers.

Predicting taxi ride durations before the trip begins allows dispatchers to better allocate resources, ensuring that drivers are available where and when they are needed most. This capability is particularly important in densely populated urban areas where traffic conditions, weather, and other factors can significantly impact travel times. Effective ride duration prediction can enhance the overall efficiency of the ride-hailing system, leading to improved service quality and customer satisfaction.

This research aims to develop and evaluate predictive models for estimating taxi ride durations using historical ride data. We explore various machine learning techniques, including linear regression, K-Nearest Neighbors (KNN), Ridge regression, and Lasso regression, to determine which methods provide the most accurate predictions. While our study uses data from New York City taxi trips, the methodologies and findings are applicable to ride-hailing services globally. Our modeling approach involves developing several predictive models, starting with a benchmark linear regression model and extending to more sophisticated methods like KNN and regularized linear models. Hyperparameter tuning and cross-validation are employed to optimize model performance and ensure robustness. The performance of the models is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

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. Predictive modeling in transportation has evolved significantly over the years, with numerous studies exploring various methodologies for estimating travel times. Traditional statistical methods, such as linear regression, have been widely used for their simplicity and interpretability. However, the advent of machine learning has introduced more sophisticated techniques capable of capturing complex, non-linear relationships in data..
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. Linear regression has been a staple in predictive modeling due to its straightforward approach and ease of implementation. Studies have shown that linear models can provide a baseline for understanding the factors affecting ride durations. However, their performance is often limited by their inability to capture non-linear interactions among variables.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.
3) Machine Learning Techniques, Recent advancements in machine learning have led to the development of more powerful predictive models. K-Nearest Neighbors (KNN) is an instance-based learning algorithm that considers the similarity between data points to make predictions. Ensemble methods, such as Random Forests and Gradient Boosting, have demonstrated superior performance by combining multiple weak learners to form a strong predictive model. Neural networks, particularly deep learning architectures, have also shown promise in handling large, complex datasets.

4) Feature Engineering : Feature engineering plays a critical role in the success of predictive models. Identifying relevant features, such as trip distance, time of day, weather conditions, and traffic patterns, can significantly enhance model accuracy. Feature selection and transformation techniques are essential for improving the predictive power of the models.

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) Research Design: This study employs a quantitative research design to develop and evaluate predictive models for taxi ride durations. We use historical ride data from New York City to train and test our models. The research involves several stages, including data preprocessing, feature engineering, model development, and evaluation.
2) Data Collection: The dataset used in this study comprises detailed records of taxi trips, including pickup and drop-off locations, timestamps, trip distances, and fare amounts. We also incorporate external data sources, such as weather conditions and traffic information, to enrich the feature set.
3) Variables and Feature Selection: Key variables influencing ride durations include trip distance, pickup time, day of the week, weather conditions, and traffic patterns. Feature selection techniques are applied to identify the most relevant predictors, while feature transformation methods are used to normalize and scale the data.
4) Predictive Modeling Approach: We develop several predictive models, including a benchmark linear regression model, a KNN model, and regularized linear models such as Ridge and Lasso regression. Hyperparameter tuning is performed to optimize model performance. The models are trained and tested using a portion of the dataset, with cross-validation employed to ensure robustness.
5) Model Evaluation and Validation: The performance of the models is evaluated using metrics such as MAE, MSE, RMSE, and R-squared. Cross-validation techniques are used to assess the generalizability of the models. Comparative analysis of the different models helps identify the most effective approach for predicting taxi ride durations.
6) 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. To evaluate the performance of the predictive models, we split the dataset into training and testing sets. The training set is used to build and train the models, while the testing set is used to evaluate their performance. A common practice is to use 80% of the data for training and 20% for testing, ensuring that the models are tested on unseen data to assess their generalizability.
7) Mean Prediction: As a simple baseline model, we calculate the mean ride duration of the training set and use it to predict the ride duration for all trips in the testing set. This model provides a benchmark for evaluating the performance of more complex models..
8) Linear Regression: The linear regression model serves as a basic predictive model, providing initial

insights into the relationships between ride durations and various features. Although the linear model captures some variability in ride durations, its performance is limited by its inability to model non-linear interactions..

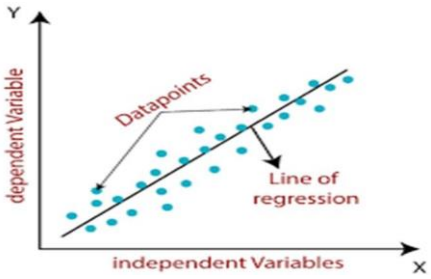


Figure 3.1 L R Graph

9) 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.

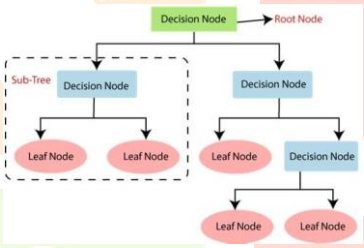


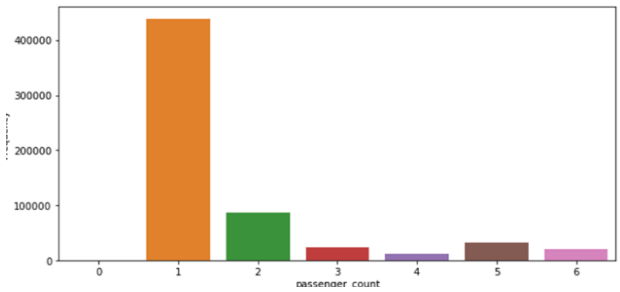
Figure 3.2 Decision Tree Algorithm

IV. RESULTS AND EVALUATION

From the obtained results we can say that :

According to our analysis results, the frequency of passenger rides is mostly one passenger, then two, then three, and so on. Figure 1 below shows the frequency distribution of the number of passengers per ride.

Figure 4.1 Passenger count on trips



Most taxi trips involve only one passenger. Although there are some trips with 7-9 passengers, they are quite rare

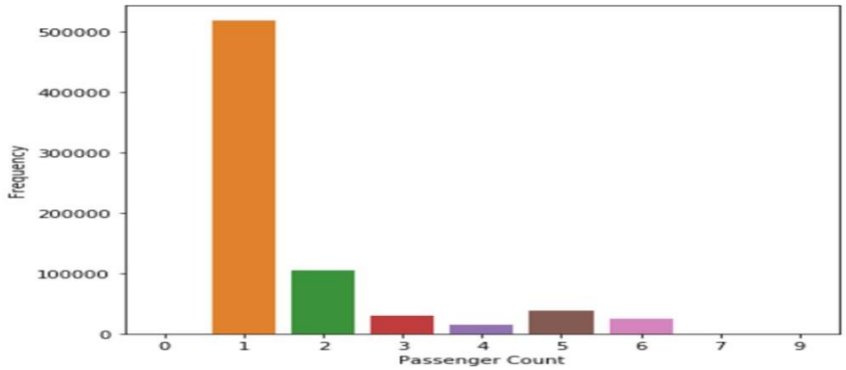


Fig.4.2

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

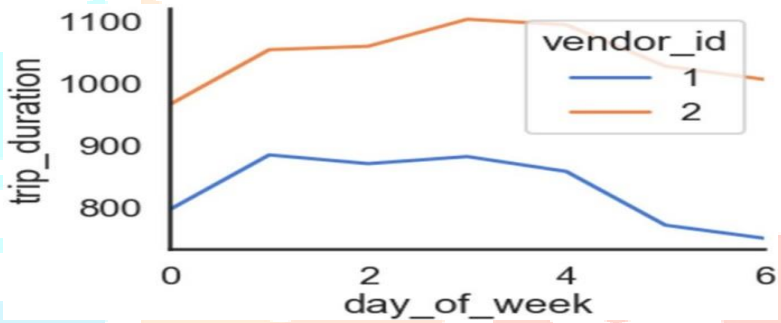


Fig.4.3

The number of pickups on weekends is significantly lower compared to weekdays, with the highest number of pickups occurring on Thursday (4). For reference, the weekday is represented as a decimal number, where 0 corresponds to Sunday and 6 corresponds to Saturday.

Upon examining the distribution of trips by latitude and longitude, it is evident that most trips are concentrated within specific geographic coordinates, with several notable clusters. These clusters are highlighted by the peaks observed in the histograms of latitude and longitude, indicating areas of high trip density.

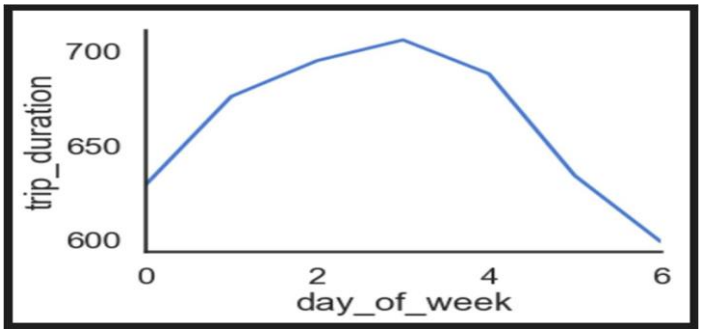


Figure 4.4 Trip duration for the day of the week

Trip durations are notably shorter during late night and early morning hours, which can be attributed to lower traffic density during these times. The number of pickups peaks in the late evenings, as expected, but is significantly lower during the morning rush hours.

This pattern aligns with the trend observed in trip durations, suggesting a correlation between the number of pickups and trip duration.

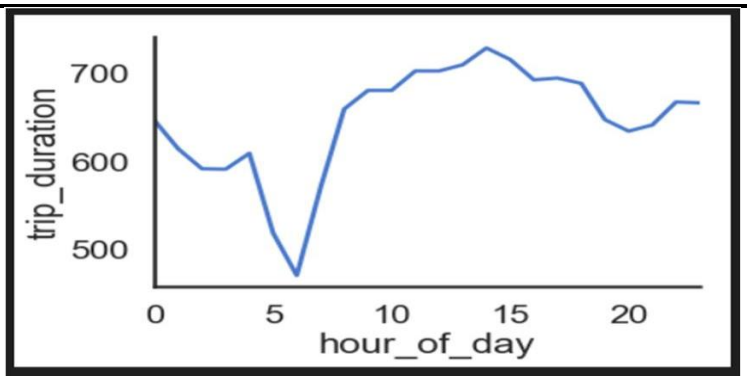


Figure 4.5 Trip duration chart for the hour of the day

Another key observation is that the number of outliers decreases with higher passenger counts. However, this is primarily due to the lower frequency of trips with a high number of passengers..

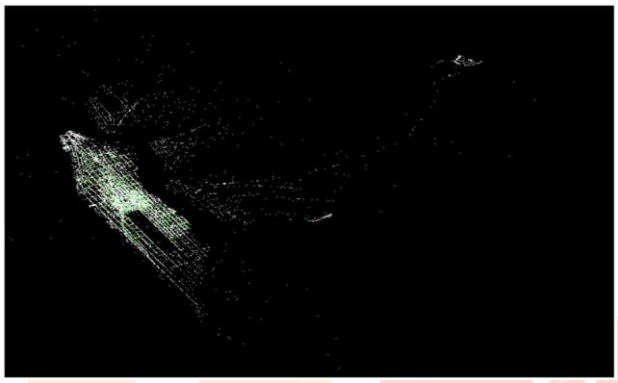


Figure 4.6 correlation heatmap

The correlation heatmap reveals that the latitude and longitude features exhibit a stronger correlation with the target variable compared to other features.
And comparing the results of Linear regression and Decision tree :

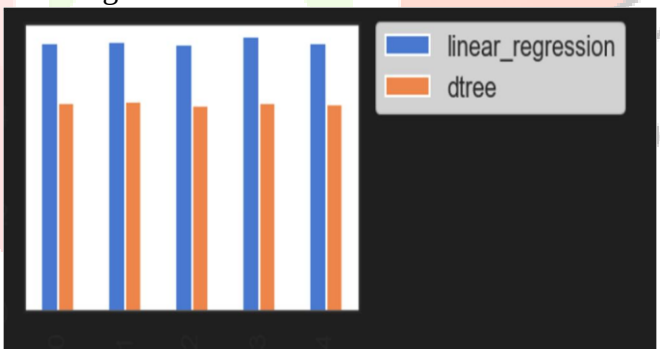


Figure 4.7 Performance MeasureWe can conclude that Linear Regression outperforms the Decision Tree model.

V. CONCLUSION

Machine learning has proven to be an effective method for predicting taxi trip durations. By training machine learning models on historical data, it is possible to uncover relationships between various features such as pickup and dropoff locations, time of day, day of the week, and current traffic conditions. Once trained, these models can predict the duration of future taxi trips with a high degree of accuracy.

The predictive accuracy of machine learning models can be further enhanced by incorporating a greater variety of features and employing more advanced machine learning techniques. For example, integrating data from multiple sources, including historical taxi data, weather data, and traffic data, can provide a richer dataset for the models to learn from. Additionally, using sophisticated machine learning algorithms, such as deep learning, can lead to more precise predictions.

There are several benefits to using machine learning for predicting taxi trip durations. Firstly, it can help taxi companies optimize their dispatch algorithms. By accurately estimating trip durations, companies can more efficiently allocate taxis to pick up passengers. Secondly, it offers riders a more accurate estimate of their trip duration before booking, enabling them to plan more effectively and avoid unexpected delays. Finally, by reducing the time taxis spend waiting for passengers, machine learning can contribute to overall system efficiency, decreasing traffic congestion and improving air quality.

In summary, machine learning is a promising approach for predicting taxi trip durations. By leveraging advanced techniques and comprehensive datasets, machine learning models can significantly enhance the efficiency of taxi systems and improve service quality for riders.

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] 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

[6] 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

[7] 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.

[8] Roy, B., & Rout, D. (2022). Predicting Taxi Travel Time Using Machine Learning Techniques Considering Weekend and Holidays. In Lecture Notes in Networks and Systems (Vol. 417 LNNS, pp. 258–267). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-96302-6_24

[9] Ramani, S., Ghiya, A., Aravind, P. S., Karuppiah, M., & Pelusi, D. (2022). Predicting New York Taxi Trip Duration Based on Regression Analysis Using ML and Time Series Forecasting Using DL. In Lecture Notes in Networks and Systems (Vol. 458, pp. 15–28). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-19-2894-9_2

[10] Liu, Z., Xia, X., Zhang, H., & Xie, Z. (2021). Analyze the impact of the epidemic on New York taxis by machine learning algorithms and recommendations for optimal prediction algorithms. In ACM International Conference Proceeding Series (pp. 46–52). Association for Computing Machinery. https://doi.org/10.1145/3475851.3475861

[11] Poongodi, M., Malviya, M., Kumar, C. et al. New York City taxi trip duration prediction using MLP and XGBoost. Int J Syst Assur Eng Manag **13** (Suppl 1), 16–27 (2022). https://doi.org/10.1007/s13198-021-01130-x