Taxi Fare Prediction using Machine Learning & Deep Learning Techniques

Afsana Sharmin & Nayeem Hoque
Team 101's Project Report (Subgroup-2)

Abstract: This study explores the effectiveness of machine learning (ML) and deep learning models in predicting taxi fares using the raw NYC Taxi Trip Dataset. It emphasizes end-to-end model development starting from data preprocessing, feature engineering, and exploratory data analysis. Key features such as trip distance, pickup/dropoff zones, trip duration, and time-of-day were utilized to build predictive models. Four models Linear Regression, Random Forest, CatBoost, and a Feedforward Neural Network were trained and evaluated using metrics like R^2, RMSE, MAE, and MSE. Among them, CatBoost achieved the highest performance (R^2 = 0.8171, RMSE = 2.82), effectively capturing non-linear relationships in the data. Additionally, spatial clustering via DBSCAN revealed high-demand regions and fare trends, particularly around LaGuardia Airport and East Chelsea. Despite its strengths, the study was limited by computational resources, outliers, and the absence of external contextual data.This work contributes a scalable, data-driven solution for intelligent fare estimation and offers insights for future enhancements using geospatial and temporal modeling.

1. Introduction

Estimating the total cost of using taxi services (fare) remains a persistent challenge in metropolitan areas like New York City, where taxi services constitute a significant component of urban mobility. Traditional cost calculation methods, which primarily depend on static formulas based on distance and duration, often fall short of capturing real-world complexities such as traffic congestion, time-of-day demand fluctuations, and spatial variability between pickup and dropoff zones (Guo et al., 2024). These limitations not only undermine cost transparency for passengers but also restrict operational efficiency for service providers.

Recent advancements in Machine Learning (ML) provide a promising avenue for addressing these challenges. By leveraging large-scale historical trip data, ML models are capable of capturing intricate, nonlinear relationships among multiple influencing factors such as trip distance, temporal attributes, and geographic zones (Zhang et al., 2021). Prior research has shown that regression-based models, including Random Forest, Gradient Boosting, and Deep Learning methods, can significantly outperform traditional approaches in fare prediction accuracy (Chou et al., 2023; Naji et al., 2024). Despite this progress, existing models often emphasize a narrow set of features, primarily trip distance; while underutilizing the predictive potential of temporal variables and spatial identifiers.

This project focuses on predicting taxi fares using the raw NYC Taxi Trip Dataset. We chose to work with the raw data to apply machine learning (ML) techniques from the ground up, starting with data preprocessing. By doing so, we aim to uncover complex patterns in features like pickup/dropoff locations, trip duration, and time-related factors. While most prior research has used cleaned datasets, we wanted to see how ML can be applied directly to unprocessed data, offering a more accurate and efficient way to predict fares.To analyze the spatial and temporal distribution of taxi fares, we applied clustering techniques to identify high-demand zones and uncover fare trends. This approach helps reveal patterns in location-based and time-based fare variations directly from raw data.

Our Contributions:

1. To utilize machine learning algorithms to predict taxi fare amounts using the NYC Taxi Trip Dataset.
2. To preprocess and engineer meaningful features from the raw dataset, such as pickup/dropoff locations, trip duration, and time-based variables.
3. To train and evaluate multiple ML models and assess their prediction performance.
4. For performance analysis we have used R square, MSE, RMSE, MAE evaluation metrics
5. To analyze the spatial and temporal distribution of fares, identifying high-demand areas and fare trends using clustering techniques.

2. Background Study

Accurate fare estimation in the taxi industry has long been a topic of interest in transportation research, especially with the rise of data-driven decision-making. The availability of large-scale datasets; such as the New York City (NYC) Taxi Trip Dataset; has made it possible to apply advanced analytics and machine learning (ML) techniques to predict fares based on real-world trip information. The NYC dataset, derived from the Taxi and Limousine Commission's (TLC) official trip records, contains millions of trip entries and includes features such as pickup and dropoff locations, timestamps, trip distances, durations, and final fare amounts (Zhang et al., 2021).

Previous studies have demonstrated that trip distance is the most significant predictor of fare, showing a strong positive correlation with the fare amount (Chou et al., 2023). However, while distance explains much of the variability, it alone cannot capture fluctuations caused by temporal and spatial

factors, such as rush hour pricing or neighborhood-based surcharges. For instance, passengers traveling during early mornings or evenings may experience higher fares due to demand peaks, regardless of distance. Additionally, fares may vary depending on the starting and ending zones, reflecting localized traffic, tolls, or regulatory pricing (Guo et al., 2024). Researchers have employed various ML algorithms including Linear Regression, Polynomial Regression, Random Forest Regression, Gradient Boosting Regression, and Deep Learning models; to predict fares. While simpler models like Linear Regression offer interpretability, they often lack the ability to capture nonlinear interactions between variables. In contrast, ensemble methods and neural networks are capable of handling more complex relationships and have shown greater accuracy in real-world applications (Naji et al., 2024).

Despite these advancements, a recurring limitation is the underutilization of engineered features. Many models include raw pickup and dropoff IDs without translating them into meaningful geographic insights. Similarly, although time-related variables like hour of day and day of week can significantly influence fares, they are frequently omitted or inadequately encoded. Furthermore, trip duration, when used without filtering for anomalies, may introduce noise into the model due to inconsistencies caused by GPS errors or missing data (Abideen et al., 2021).

3. Data Preparation:

The NYC Taxi Trip Dataset (2018), sourced from the Google BigQuery public dataset and curated on Kaggle by Neil Clack, provides a comprehensive record of yellow taxi operations across New York City. This dataset is derived from the official Taxi and Limousine Commission (TLC) trip records and contains over 84 million taxi trips completed in the year 2018. There are eight predictor variables (Table 1) and one target variable (Table 2).

Table 1: Feature Description

Feature	Description
trip_distance	The elapsed trip distance in miles reported by the taximeter.
pickup_location_id	ID number representing where the trip started (pickup zone).
dropoff_location_id	ID number representing where the trip ended (dropoff zone).
day_of_week	Day of the week when the trip took place (1 = Sunday, 7 = Saturday).
hour_of_day	Hour when the trip started (1–24, based on a 24-hour clock).
month	The month when the trip took place (1 = January… 12 = December).
speed	How fast the taxi travels on average for a given trip in miles per minute.
duration	How much time needed to complete trip

Table 2: Targets Description

Target	Description
fare_amount	The base charge for the trip in US dollars.

3.1 Data Cleaning:

To properly clean the data, several variables were engineered and bound to filter out extreme outliers that may impact the effectiveness of the models. The engineered variables are the hour of the day, day of the week, month of the year, duration of the trip, and average speed for the trip. To efficiently remove extreme or illogical entries, most of the variables were bounded by upper and lower limits except for pick-up/drop-off location IDs, hour of the day, day of the week, and month of the year. The original dataset had over one million entries but after preparation, the final dataset contained 64,722 entries. From here, it was further split into training data (80%), validation data (10%), and test data (10%). Below in Figure 1 is a snapshot of the first few entries of the cleaned data.

trip_distance	fare_amount	pickup_location_id	dropoff_location_id	day_of_week	hour_of_day	trip_duration
16.97	49.5	231	138	3	13	3317
14.45	45.5	87	138	3	14	3648
11.6	42	68	138	3	14	3540
5.1	26.5	186	33	3	16	2585
11.11	45.5	163	138	3	16	4521
9.54	41	138	244	3	16	3738.0000000000005
10	38.5	138	25	3	17	3598

Figure 1: First few entries of the newly cleaned data

3.2 SHAP Analysis for feature Selection

SHAP (SHapley Additive exPlanations) is a powerful method for interpreting machine learning models by quantifying the contribution of each feature to individual predictions.It helps you understand which features matter most and why a model made a certain decision.

Figure-2 illustrates the feature analysis using SHAP values, showing the impact of each feature on the model's prediction of fare amount.

- Trip distance is the most dominant and consistent feature influencing the fare.
- Hour of day, dropoff location, and average speed contribute moderately, likely due to fare structure nuances (e.g., rush hour, airport zones).
- Trip duration, pickup location, and day of week have relatively small impacts, suggesting limited predictive power in isolation.

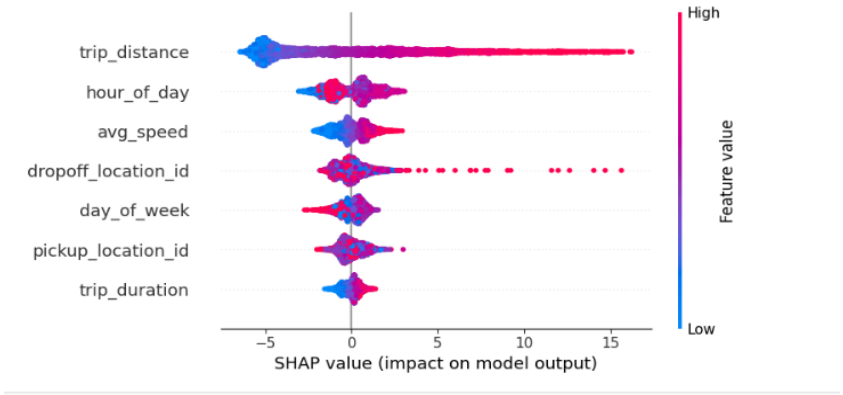

Figure 2: SHAP analysis

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach in data analysis that focuses on summarizing, visualizing, and understanding a dataset before applying machine learning models or statistical techniques. It helps identify patterns, detect anomalies, test assumptions, and check for relationships between variables.

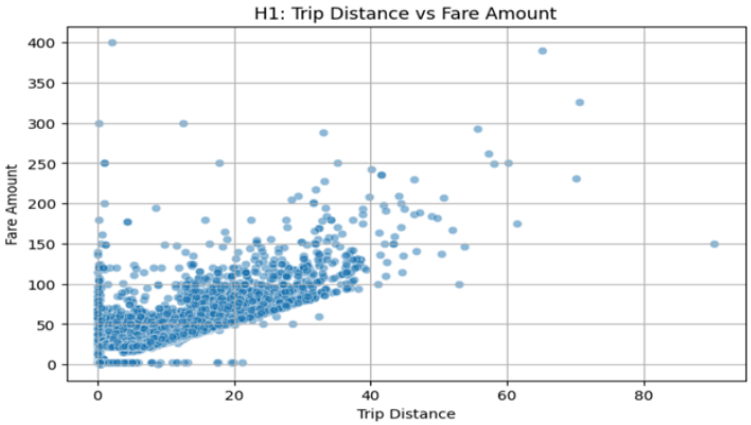

Figure 3: Trip Distance vs Fare Amount

From *Figure 3*, a clear positive trend as trip distance increases, the fare amount generally increases. The points are densely clustered along a rising band, this shows a strong linear relationship.

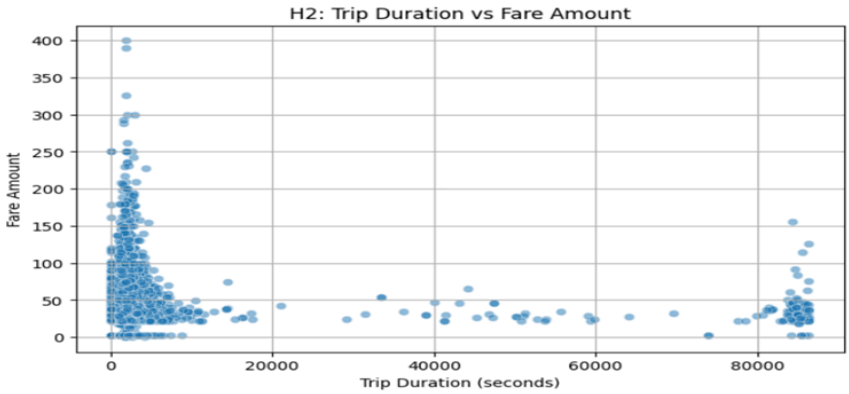

Figure 4: Trip duration vs Fare amount

From *Figure 4*, most points are densely packed in the lower-left corner (short durations and low fares).There is no clear upward trend, unlike with trip_distance. Some very long durations (e.g., > 80,000 seconds ≈ 22 hours) have low fares, which is unusual and likely due to: GPS or logging errors, missing data.

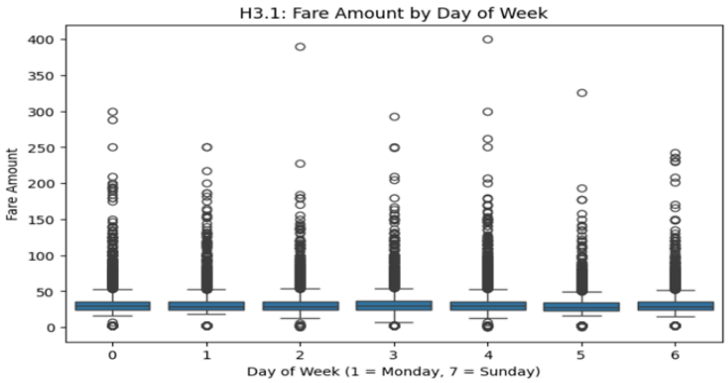

Figure 5: Fare Amount vs Day of Week

From *Figure 5*, the boxplot for Fare Amount by Day of Week shows that while the median fare remains relatively consistent across all days, there are outliers (high fares) present on each day, particularly during weekends (Days 5 and 6). This suggests that, although fares are generally stable, premium or long-distance trips occur more frequently on weekends.

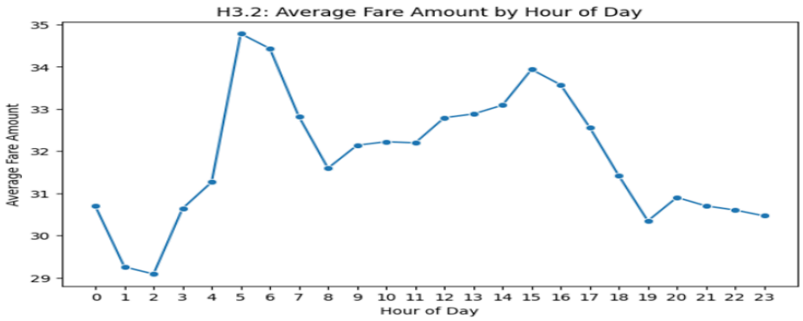

Figure 6: Average Fare Amount vs Hour of Day

From *Figure 6*, the line plot for Average Fare Amount by Hour of Day shows a sharp spike in fares between 4-6 AM, likely due to early morning rides or higher demand. After that, fares gradually decrease throughout the day, with a slight increase in the evening (around 5-7 PM), possibly reflecting rush hour pricing.

4. Methodology:

In this section, we will discuss the machine learning and deep learning techniques that we have applied for our taxi fare prediction.We will also discuss various performance metrics.

4.1: ML Models:

Linear Regression:

Linear regression is a supervised learning technique used to predict a continuous outcome (target variable) based on one or more predictor variables (independent variables). In this method, the relationship between the target and predictor variables is modeled as a straight line, where the coefficients of the predictors indicate the magnitude and direction of their impact on the target variable. These coefficients help determine how changes in the predictors influence the predicted outcome. Here, we have modelled the target variable fare amount against the eight variables

$$Fare\ amount = b0 + b1 * trip\_distance + b2 * pickup\_location\_id + b3 * dropoff\_location\_id + b4 * day\_of\_week + b5 * hour\_of\_day + b6 * month + b7 * speed + b8 * duration$$

Random Forest:

Random forests are an ensemble learning method used for both regression and classification tasks.For regression, the models within a random forest are decision trees. The process of training a random forest involves creating many decision trees through bootstrapping, where random samples from the training data are selected with replacement to train each tree. The features used to split nodes are also randomly chosen for each tree, ensuring diversity in the models.To make a prediction, the input data is passed through each decision tree, and the final prediction is the average of all individual tree predictions. Random forests are particularly effective for regression because they handle large datasets and multiple features well, and are less prone to overfitting compared to other methods

CatBoost:

CatBoost is a gradient boosting algorithm designed to handle categorical features efficiently. It works by building an ensemble of decision trees in a sequential manner, where each tree corrects the errors made by the previous one. CatBoost specifically optimizes the handling of categorical data by using a technique called ordered boosting, which helps prevent overfitting and ensures better generalization on datasets with categorical features. It also automatically deals with missing values and requires minimal data preprocessing, making it a powerful tool for structured data tasks.

Feed Forward Network:

A Feed Forward Neural Network (FNN) can be used for taxi fare prediction by modeling complex relationships between input features (e.g., trip distance, pickup/dropoff locations, time of day) and the target variable (taxi fare). The network consists of an input layer that receives features, one or more hidden layers that capture non-linear interactions using activation functions, and an output layer that predicts the fare. During training, the FNN adjusts weights to minimize prediction errors, and after training, it can predict fares for new trips based on the learned patterns. This approach captures non-linear dependencies, making it more accurate than linear models for predicting taxi fares.

4.2:Metrics and Measures for Model Performance:

To measure the model performances for the four models created we will look at some common metrics.

R-Squared:

A regression model's goodness of fit can be measured using R-Squared. It is a number that indicates how well the model fits the data and is between 0 and 1, with 1 representing the best fit and 0 representing no fit.The lower the MSE the more accurate the model, while a higher MSE indicates a less accurate model.

Mean Squared Error (MSE):

MSE is the squared difference between the predicted and actual values, which is the average of the squared differences.

Mean Absolute Error (MAE):

MAE is the average absolute error between actual and predicted value. It is useful to us as it gives us a confidence interval on how far off our prediction is as it is measured in the same unit as our target variable.

RMSE (Root Mean Squared Error):

RMSE represents the square root of the average of the squared differences between the predicted values and the actual values.Lower RMSE indicates a better fit

5. Experimental Results Analysis:

5.1: Linear Regression:

The model achieved a R^2 score of 0.6601, meaning it explained around 66% of the fare variance. With an RMSE of 3.85 and MAE of 2.89, the model's predictions were, on average, off by about $3 to $4. This performance is acceptable for mid-range fares but inadequate for capturing more complex fare patterns.

```
R² Score: 0.6601
RMSE: 3.85
MAE: 2.89
MSE: 14.80
```

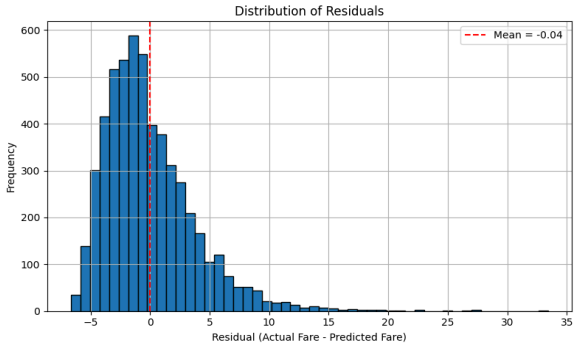

Figure 7: Distribution of residuals (Actual Fare – Predicted Fare) for Linear Regression.

As shown in Figure 7, the residuals are right-skewed with a mean near -0.04, suggesting a slight underprediction bias. The long tail on the positive side indicates the model struggles with high-fare trips, likely due to its inability to model nonlinear interactions among features. Figure 8 illustrates a scatter plot of actual vs. predicted fares. While the model aligns well for lower and average fares, it underestimates higher fare values, resulting in wider dispersion. This confirms that Linear Regression, while useful for initial modeling, is limited in handling the complexities of real-world taxi fare prediction, justifying the use of more advanced models in subsequent analyses.

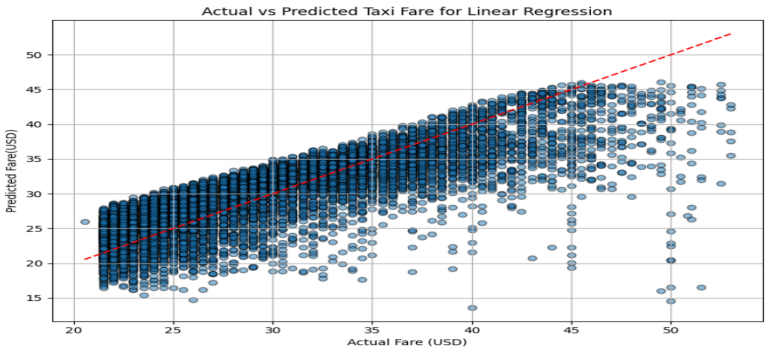

Figure 8: Actual vs. Predicted Taxi Fare for Linear Regression.

5.2 Random Forest:

The model achieved an impressive R^2 score of 0.8028, meaning it explained over 80% of the variance in fare predictions. It also outperformed the baseline model (linear regression) with an RMSE of 2.93 meaning the model is off by $2.93 on average, MAE of 2.05 indicating that most of the model's predictions are relatively close to the true fare values, and MSE of 8.58, indicating smaller prediction errors and improved consistency.

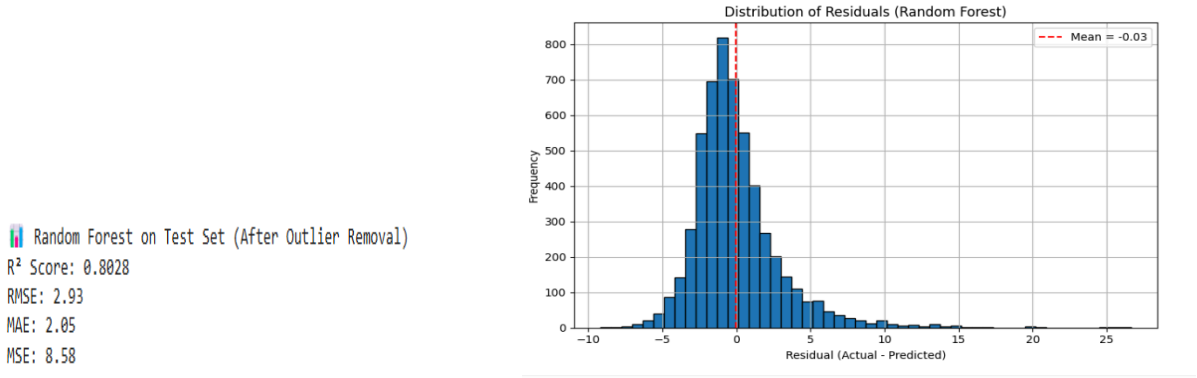

Figure 10: Distribution of Residuals for Random Forest Model.

The distribution of residuals in Figure 10 supports this observation. The histogram is nearly symmetrical and centered around zero, with a mean residual of -0.03, indicating minimal bias. The narrower and more balanced distribution compared to the linear model confirms Random Forest's capability to produce stable predictions across the fare range, even with complex feature interactions. As shown in Figure 11, the actual vs. predicted fare plot reveals a tighter clustering around the ideal diagonal line compared to Linear Regression, especially for low- to mid-range fares. This signifies better alignment between predictions and actual values. However, a degree of spread remains visible for higher fare values,suggesting occasional underprediction for expensive trips.

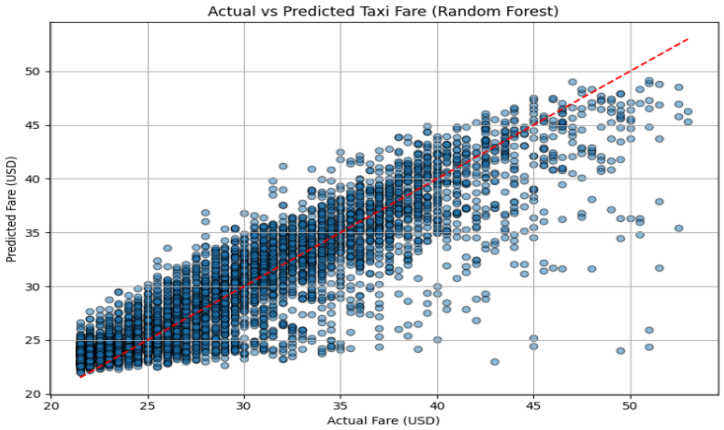

Figure 11: Actual vs. Predicted Taxi Fare for Random Forest.

5.3 Catboost:

On the test set, the model achieved a R^2 score of 0.8171, the highest among all models used, meaning it explained over 81% of the variance in fare predictions. The RMSE of 2.82 meaning the model is off by $2.82 on average and MAE of 1.95 indicate that the predictions were both accurate and consistent and very close to the actual fare values, with the MSE at 7.96 confirming low overall error.

CatBoost Evaluation on Test Set:
R² Score: 0.8171
RMSE: 2.82
MAE: 1.95
MSE: 7.96

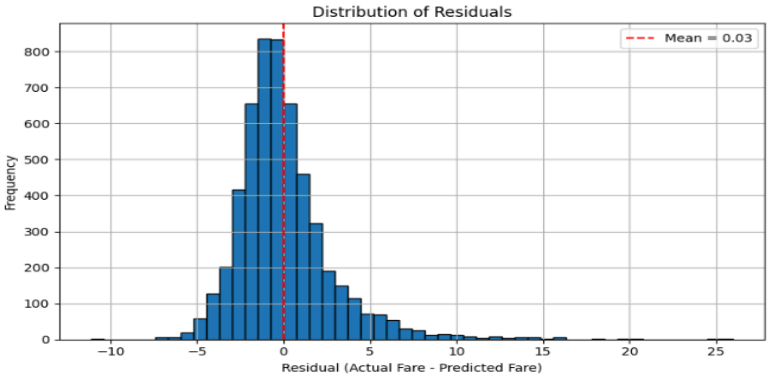

Figure 12: Distribution of Residuals for CatBoost Model.

The residual distribution in Figure 12 further supports this performance. The histogram is sharply centered near zero, with a mean residual of -0.03, indicating minimal prediction bias. The residuals follow a nearly normal distribution, demonstrating balanced performance across the full range of fare values.As shown in Figure 13, the actual vs. predicted fare plot displays a strong alignment of predicted values along the ideal line. Compared to previous models, this plot shows tighter clustering and reduced spread, especially for fares above $35, highlighting CatBoost's effectiveness in capturing complex fare dynamics.

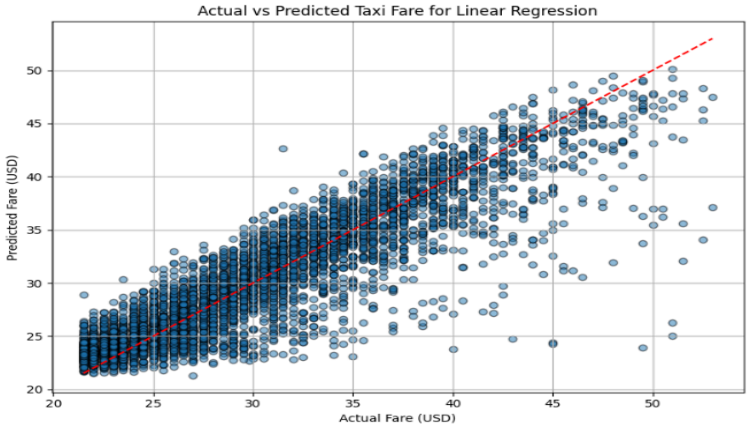

Figure 13: Actual vs. Predicted Taxi Fare for CatBoost Model.

5.4 Feed Forward Network:

The model performed well, achieving an R^2 score of 0.7494, meaning it explained about 75% of the variance in the fare data. With an RMSE of 3.30 and an MAE of 2.46, the predictions were typically off by $2 to $3, which is a solid result for most fares.

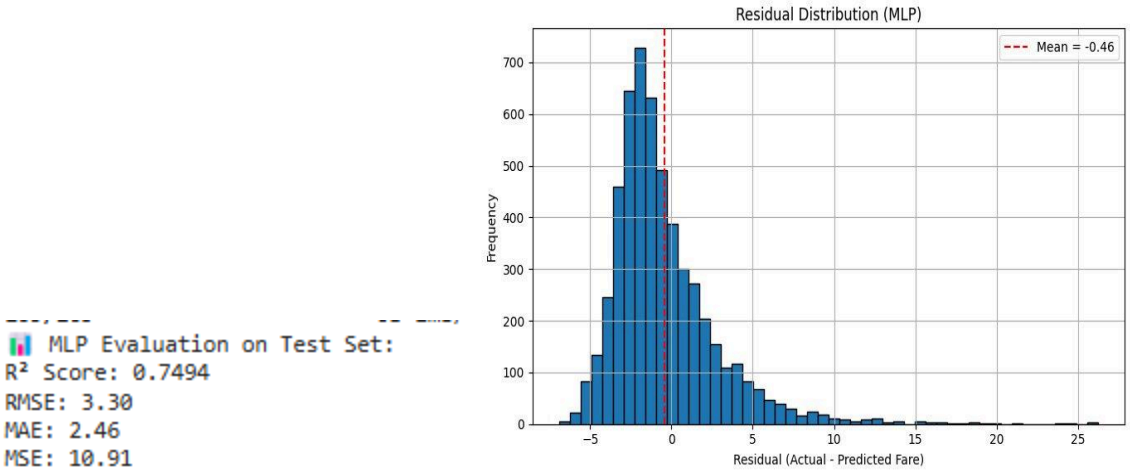

Figure 14:Distribution of Residuals for FNN Model.

In the residual distribution (Figure 14), we can see that most of the errors are fairly small, but the model tends to slightly underpredict fares, with a mean residual of -0.46. The residuals are mostly centered around zero, but there's a small leftward skew, indicating that the model generally gives slightly lower fare predictions. However, there are still some larger discrepancies, especially when the actual fares are much higher than the predicted ones.

Looking at the scatter plot of actual vs. predicted fares (Figure 15), we can see that the model does a good job predicting the fares for most trips. However, for higher fares, there's more variability in the predictions, with some points deviating more significantly from the true fare. This suggests that while the Feed Forward Network performs well in general, it still faces challenges with predicting extreme fare values, likely due to factors like traffic or longer trips that the model isn't fully capturing.

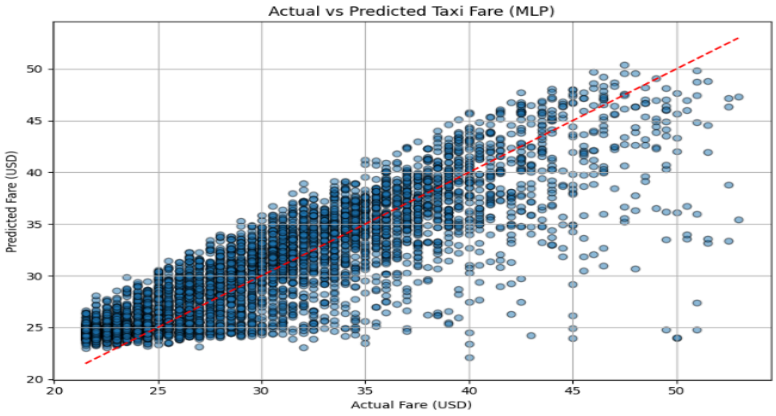

Figure 15: Actual vs. Predicted Taxi Fare for FNN Model.

5.5 Comparison among Models:

The table-3 below compares the performance of four different models for taxi fare prediction based on key evaluation metrics: R^2, RMSE, MAE, and MSE.CatBoost outperforms the other models in terms of prediction accuracy and error metrics, followed closely by Random Forest. FNN and Linear Regression lag behind, with FNN being better than Linear Regression but still not as effective as the other models.

Table -3: Comparison among ML models

Models	R Square	RMSE	MAE	MSE
Linear Regression	0.6601	3.84	2.89	14.08
Random Forest	0.8028	2.93	2.05	8.58
CatBoost	**0.8171**	**2.82**	**1.95**	**7.96**
FNN	0.7494	3.30	2.46	10.91

5.6 Comparison with Related Works:

We found some research articles where different datasets were used for predicting taxi fares. In [9], the authors applied Linear Regression, achieving an R-Squared value of 0.4, and Random Forest, which had an R-Squared value of 0.5. Another article, [10], also used a different dataset, reporting better results with CatBoost (R-Squared = 0.945) and Random Forest (R-Squared = 0.937).For our dataset, the results show that CatBoost achieved an R-Squared value of 0.8171, while Random Forest reached an R-Squared value of 0.8028.

6. Spatio-Temporal Analysis of Taxi Fares (Monday):

The goal of this analysis is to identify the most densely populated clusters of taxi pickups and dropoffs on Mondays, then explore how fare amounts are distributed within these busy zones. This helps uncover patterns in passenger demand and fare fluctuations across different parts of the city. To achieve this, we use DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a clustering algorithm that groups together points that are closely packed while treating isolated points as noise. Unlike traditional clustering methods, DBSCAN doesn't require a predefined number of clusters, making it ideal for identifying natural groupings in spatial data. By applying this method, we can clearly highlight areas with high taxi activity and gain insight into how fares behave in those regions.

The dataset includes geographic information for pickup and dropoff locations, identified by taxi zone IDs. By merging this spatial data—specifically, the centroids of each zone—with fare details, we can better understand how location influences fare amounts. The analysis focuses exclusively on rides that occurred on Mondays, allowing for a more targeted spatio-temporal exploration of taxi activity and fare patterns on that day of the week.

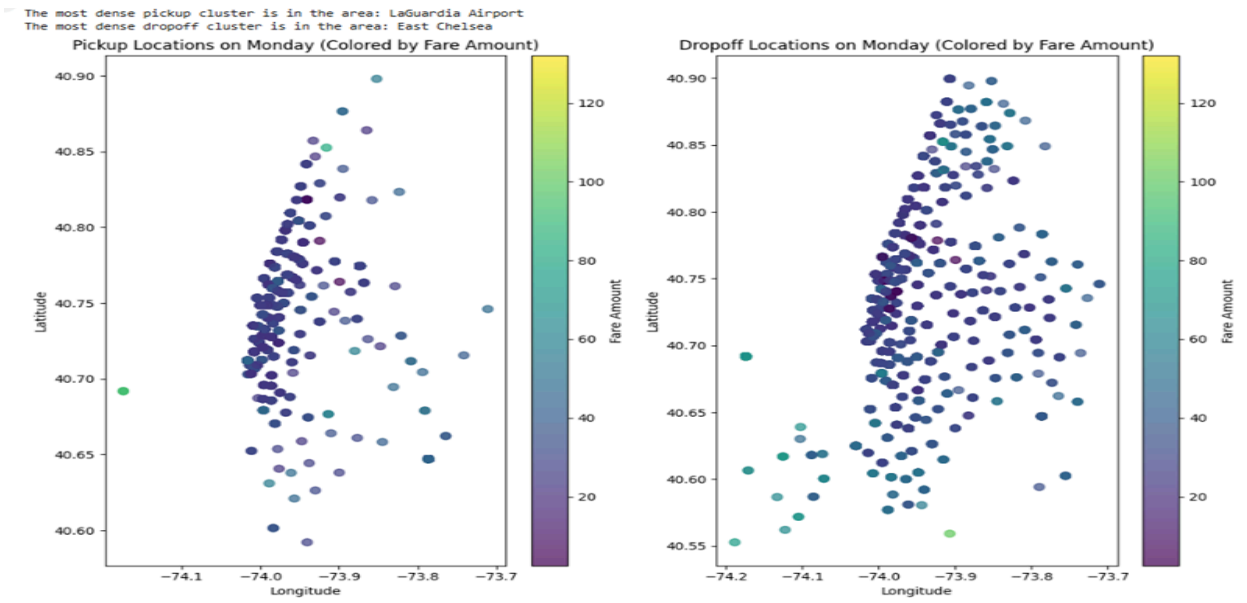

Figure 16: Spatio-Temporal Distribution of Taxi Pickup and Dropoff Locations

From figure 16 the results highlight two key insights. The densest pickup cluster is centered around LaGuardia Airport, pointing to a high volume of passengers and potentially longer trips originating from this area. In contrast, the densest dropoff cluster is found in East Chelsea, suggesting it's a popular destination where many taxi rides conclude.Fare patterns also emerge from the analysis. Higher fares tend to cluster in central or high-demand areas like LaGuardia or other major hubs, possibly due to longer trip distances or traffic-related delays. In comparison, lower fares are more common in residential neighborhoods, where trips are typically shorter.

The visualization on the right illustrates this spatial pattern for Monday rides. Pickup and dropoff points are color-coded by fare, with darker purples indicating higher fares and lighter shades representing lower ones. This visual effectively links geographic clustering with fare distribution, making it easier to spot trends at a glance.

7. Limitation:

Despite the promising findings, this study faced several limitations. Computational constraints restricted the use of deep learning models, such as Feedforward Neural Networks, limiting their depth and the extent of hyperparameter tuning due to limited hardware resources. Additionally, outliers in trip duration and fare amount remained despite data cleaning and may have skewed the model's learning, particularly in predicting high-fare trips. The analysis also lacked external influencing factors such as weather conditions, special events, traffic congestion, or ride-sharing competition, all of which can significantly impact fare variability. Moreover, since only a small portion of the full dataset was used, the analysis may not effectively capture patterns related to long-duration or less frequent trips. Finally, the spatial clustering focused solely on Monday data, which limits the generalizability of the insights to other weekdays or weekends when travel behaviors might differ.

8. Conclusion:

This study showcased the effectiveness of machine learning techniques particularly CatBoost and Random Forest in predicting NYC taxi fares using a raw dataset from 2018. By incorporating time and location-based features, the models were able to capture key patterns in the data, though predicting high-fare and long-duration trips remained a challenge. Linear Regression served as a solid baseline, and Feedforward Neural Networks showed potential, but ensemble methods (Catboost) consistently delivered superior accuracy and robustness.

In addition to predictive performance, the analysis uncovered valuable spatial trends. High-demand areas like LaGuardia Airport and East Chelsea emerged as key zones influencing fare distribution. Overall, the project demonstrates the power of data-driven approaches for understanding urban mobility and lays the groundwork for building more advanced, context-aware forecasting models in the future.

To improve taxi fare prediction, future work could incorporate GPS coordinates and use geospatial or graph-based models like GCNs to better capture spatial relationships. Advanced deep learning models (LSTM, BiLSTM, Transformers) can help capture temporal patterns, especially for long trips. Integrating external data such as weather, traffic, and events would add valuable context. Predicting fare components individually using multi-output models may improve interpretability and accuracy. Lastly, building a real-time prediction system with a web or mobile interface would enhance practical usability.

Note: Both group members have equally contributed to all aspects of this project, including data preprocessing, model development, analysis, and report writing.This is our own Project.

References

[1] Abideen, Z. U., Sun, H., Yang, Z., & Fahim, H. (2021). Regional‐based multi‐module spatial–temporal networks predicting city‐wide taxi pickup/dropoff demand from origin to destination. *Expert Systems*, *39*(2). https://doi.org/10.1111/exsy.12883

[2] Chou, K. S., Wong, K. L., Zhang, B., Aguiari, D., Im, S., Lam, C., Tse, R., Tang, S.-K., & Pau, G. (2023). Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation. *Applied Sciences*, *13*(18), 10192–10192. https://doi.org/10.3390/app131810192

[3] Clack, Neil. "NYC Taxi Trip Data - Google Public Data." *Kaggle*, 3 June 2022, www.kaggle.com/datasets/neilclack/nyc-taxi-trip-data-google-public-data/data?select=original_cleaned_nyc_taxi_data_2018.csv

[4] Guo, Y., Chen, Y., & Zhang, Y. (2024). Enhancing Demand Prediction: A Multi-Task Learning Approach for Taxis and TNCs. *Sustainability*, *16*(5), 2065–2065. https://doi.org/10.3390/su16052065

[5] Hasan, Xue, Q., & Li, T. (2024). A Distributed VMD-BiLSTM Model for Taxi Demand Forecasting with GPS Sensor Data. *Sensors*, *24*(20), 6683–6683. https://doi.org/10.3390/s24206683

[6] Naji, H. A. H., Xue, Q., Zhu, H., & Li, T. (2021). Forecasting Taxi Demands Using Generative Adversarial Networks with Multi-Source Data. *Applied Sciences*, *11*(20), 9675. https://doi.org/10.3390/app11209675

[7] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., & Li, Z. (2018). Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11836

[8] Zhang, J., Chen, H., & Fang, Y. (2021). TaxiInt: Predicting the Taxi Flow at Urban Traffic Hotspots Using Graph Convolutional Networks and the Trajectory Data. *Journal of Electrical and Computer Engineering*, *2021*, e9956406. https://doi.org/10.1155/2021/9956406

[9] Banerjee, P., Kumar, B., Singh, A., Ranjan, P., & Soni, K. (2020). Predictive analysis of taxi fare using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 6(2), 2456-3307. https://doi.org/10.32628/CSEIT2062108

[10] Amadxarif, Z. (n.d.). *Predicting New York City taxi fares with supervised machine learning* (BSc Mathematics). Queen Mary University of London.