Predicting NYC Taxi Fares Using Random Forest and Ensemble Techniques

Suayeb Ahmed, Tanvir Ahmed Joy

The University of Memphis
Memphis, TN, USA

***Abstract*—Accurately predicting taxi fares in New York City is essential for passengers, drivers, ride-sharing services, and regulators. This research investigates the predictive capabilities of Decision Trees, Random Forest, and Long Short-Term Memory (LSTM) models using the New York City Taxi and Limousine Commission's (TLC) 2023 dataset. Random Forest emerged as the superior model, exhibiting the lowest Mean Absolute Error (MAE) across various fare ranges. This paper underscores the effectiveness of ensemble learning techniques in addressing complex, nonlinear relationships inherent in transportation fare prediction.**

***Index Terms*—Taxi Fare Prediction, Random Forest, Decision Tree, LSTM, Machine Learning, Ensemble Learning, NYC Taxi Data**

I. Introduction

Taxi fare prediction is crucial for enhancing transparency, efficiency, and operational decision-making within urban transportation networks. Traditional fare estimation methods rely on static formulas, failing to capture dynamic real-world variables such as traffic congestion, weather fluctuations, and varying temporal demand. Advanced machine learning approaches, capable of modeling these nonlinear complexities, can significantly enhance fare prediction accuracy.

The urban transportation landscape has undergone significant transformation in recent years, with ride-sharing platforms disrupting traditional taxi services and creating a more competitive environment. In this context, accurate fare prediction has become increasingly important for all stakeholders involved. For passengers, fare transparency reduces uncertainty and builds trust. For drivers, optimal route selection based on accurate fare prediction can maximize earnings. For ride-sharing platforms and taxi companies, sophisticated fare prediction algorithms enable dynamic pricing strategies that balance supply and demand. For regulators, data-driven fare models provide objective benchmarks for policy development and consumer protection.

New York City, with its complex transportation network and extensive taxi usage, provides an ideal environment for developing and testing fare prediction models. The city's Taxi and Limousine Commission (TLC) maintains comprehensive datasets of taxi trips, including spatial, temporal, and financial attributes. These rich datasets enable the application of sophisticated machine learning techniques to capture the intricate patterns governing fare determination.

The novelty of our approach lies in the comparative analysis of three distinct machine learning paradigms—Decision Trees, Random Forest, and Long Short-Term Memory (LSTM) networks—within the specific context of New York City taxi fare prediction. While individual models have been explored in prior research, our comprehensive evaluation across varying fare ranges and our ensemble approach contribute new insights to the transportation fare modeling literature.

This study addresses the following research questions:

- How effectively can machine learning algorithms predict taxi fares across different price ranges?
- Which features contribute most significantly to fare prediction accuracy?
- Can ensemble techniques further enhance predictive performance beyond individual models?
- How do deep learning approaches (LSTM) compare with traditional machine learning models in this specific domain?

This paper applies three prominent machine learning algorithms—Decision Trees, Random Forest, and LSTM—to a comprehensive dataset provided by the NYC Taxi and Limousine Commission, identifying the most accurate predictive strategy. We also create an ensemble model that combines the predictions of all three models to potentially improve accuracy further.

II. Related Work

Fare prediction in transportation services has been an active area of research, with various approaches proposed over the years. This section reviews relevant literature, highlighting methodological approaches, key findings, and research gaps addressed by our study.

A. Machine Learning for Fare Prediction

Transportation fare prediction has evolved from simple linear models to sophisticated machine learning techniques. Early work by Chiang et al. [8] established the feasibility of using regression-based approaches for taxi fare estimation, while Kamga et al. [9] demonstrated the value of incorporating spatiotemporal features for improved accuracy.

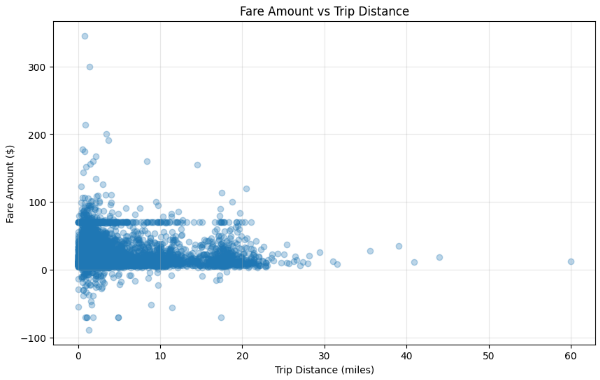

Fig. 1. Relationship between Fare Amount and Trip Distance

The application of tree-based models has gained significant traction in recent years. Khaled et al. [10] compared multiple regression models against decision trees for fare prediction, finding that tree-based approaches better captured the nonlinear relationships inherent in urban transportation data. Similarly, Zhang and Haghani [11] demonstrated the superior performance of gradient boosting algorithms for travel time prediction, a critical component in fare estimation.

Deep learning approaches have also been applied to this domain. Guo et al. [12] implemented a neural network model for taxi fare prediction that incorporated both spatial and temporal attributes. Their results highlighted the importance of feature engineering, particularly for time-based variables. Wang et al. [13] explored recurrent neural networks for time series prediction in transportation, demonstrating their ability to capture complex temporal patterns.

B. Ensemble Methods in Transportation Modeling

Ensemble methods have shown promise across various transportation prediction tasks. Li et al. [14] implemented a stacked ensemble approach for traffic flow prediction, combining the strengths of multiple models to achieve higher accuracy than any individual algorithm. Similarly, Chen et al. [15] demonstrated that ensemble techniques could effectively mitigate the limitations of individual models in predicting transportation demand.

C. NYC Taxi Data Analysis

New York City taxi data has been extensively utilized in transportation research due to its comprehensive nature and accessibility. Yang et al. [16] analyzed passenger demand patterns using NYC taxi data, identifying key spatiotemporal factors influencing ride requests. Saghapour et al. [17] used the same dataset to develop a fare prediction model that incorporated external factors such as weather conditions and special events.

D. Research Gaps and Our Contribution

Despite the extensive research in this area, several gaps remain. First, many studies focus on general prediction accuracy without analyzing performance across different fare ranges, which is critical for real-world applications. Second, while individual models have been well-studied, comparative analyses of traditional machine learning versus deep learning approaches specifically for fare prediction remain limited. Third, most research has not systematically evaluated the computational efficiency of models in relation to their predictive performance, an important consideration for practical implementations.

Our study addresses these gaps by:

- Providing a comprehensive comparison of three distinct machine learning paradigms—Decision Trees, Random Forest, and LSTM networks
- Analyzing prediction accuracy across specific fare ranges to identify model strengths and weaknesses
- Evaluating computational efficiency alongside predictive performance to inform practical implementations
- Developing an ensemble approach that leverages the strengths of multiple algorithms
- Utilizing the most recent NYC taxi data (2023) to capture current transportation patterns and pricing structures

III. Problem Statement

The goal of this research is to develop and evaluate a robust predictive model for taxi fares based on historical trip data, enabling stakeholders to make informed decisions and regulatory actions by improving fare estimation accuracy and operational transparency.

A successful prediction model would benefit multiple stakeholders:

- **Passengers**: Providing fare transparency and reducing uncertainty
- **Drivers**: Optimizing route selection and maximizing earnings
- **Ride-sharing platforms**: Improving pricing strategies and customer satisfaction
- **Regulators**: Enhancing oversight capabilities and fare standardization

IV. Dataset and Preprocessing

A. Data Description

We used a representative sample (1%) of NYC's Yellow Taxi trip records from 2023, encompassing 193,566 records detailing various trip attributes relevant to fare determination. The dataset includes features such as pickup and dropoff times, pickup and dropoff locations, trip distance, passenger count, payment method, and fare amount.

B. Data Cleaning

Data was meticulously cleaned based on realistic parameters:

- Fare amounts restricted between $0 and $1000
- Trip distances limited to less than 100 miles
- Trip durations capped at less than 1440 minutes (24 hours)
- Removal of records with missing or implausible values

After basic cleaning, the dataset contained 180,412 records.

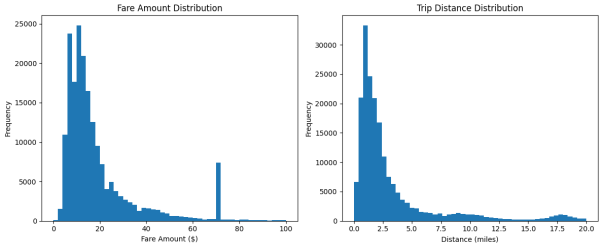

Fig. 2. Fare Amount and Trip Distance Distributions

C. Feature Engineering

Several key features were engineered to enhance model predictive performance:

- **Temporal features**: pickup hour, pickup day, pickup month, pickup year, pickup weekday, is_weekend flag, is_rush_hour flag
- **Trip-specific attributes**: trip duration, average trip speed
- **Location-based indicators**: airport pickup/dropoff identification, same location flag
- **Payment method indicator**: transactions via credit card
- **Rate code dummies**: one-hot encoded rate code identifiers

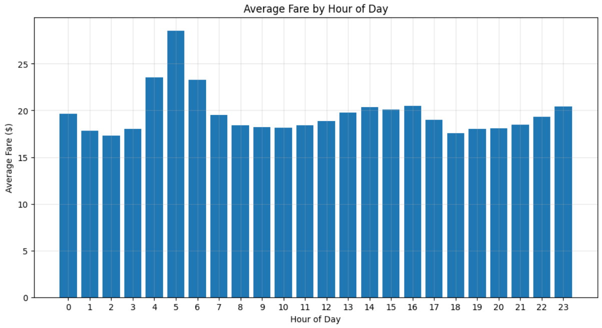

Fig. 3. Average Fare by Hour of Day

As shown in Figure 3, the average fare varies throughout the day, with higher fares observed during early morning hours (around 5-6 AM) and late evening.

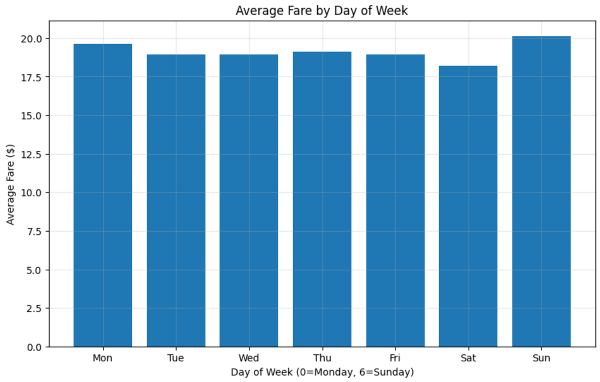

Fig. 4. Average Fare by Day of Week

Figure 4 shows that average fares are slightly higher during weekdays compared to weekends, likely reflecting commuting patterns.

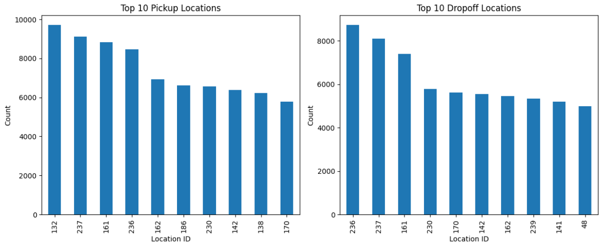

Fig. 5. Top Pickup Locations and Dropoff Locations

Figure 5 illustrates the most frequent pickup and dropoff locations by their location IDs, showing certain zones with significantly higher activity.

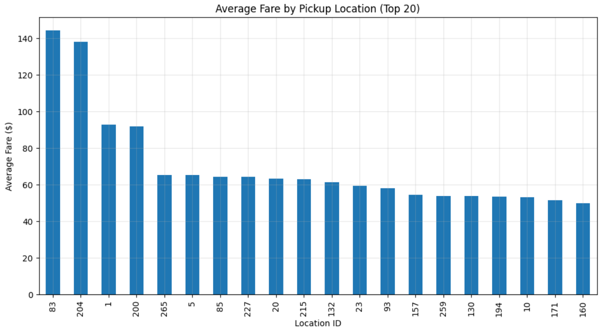

Fig. 6. Average Fare by Pickup Location (Top 20)

Figure 6 demonstrates the variation in average fares by pickup location, with certain locations (likely airports and outlying areas) commanding substantially higher fares.

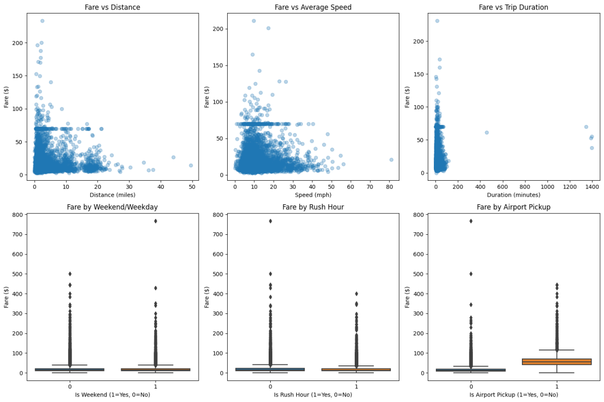

Fig. 7. Relationship Between Engineered Features and Fare Amount

Figure 7 shows the relationships between key engineered features and fare amount. The top left plot confirms the strong linear relationship between trip distance and fare. The top right plot shows that average speed has little direct correlation with fare. The bottom left scatter plot shows that trip duration is strongly correlated with fare amount, and the bottom right plots show how categorical features like weekend/weekday, rush hour, and airport pickup affect fare amounts.

The final feature set included 20 features after feature engineering, and the dataset contained 180,333 records.

V. Methodology

We implemented and evaluated three advanced machine learning models, each described in detail below.

A. Decision Tree

Decision Trees utilize recursive partitioning of data to minimize prediction errors, specifically Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

The Decision Tree model was implemented with the following parameters:

- Maximum depth: 10
- Minimum samples split: 5
- Minimum samples leaf: 2

B. Random Forest

Random Forest aggregates multiple decision trees (bagging technique) to improve overall predictive accuracy and stability, significantly reducing variance:

Prediction function:

$$f_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{2}$$

The Random Forest model was implemented with the following parameters:

- Number of trees: 100
- Maximum depth: 15
- Minimum samples split: 5
- Minimum samples leaf: 2

C. Long Short-Term Memory (LSTM)

LSTM is particularly effective for modeling sequential data, making it appropriate for time-dependent taxi fare predictions:

Cell state:

$$c_t = f_t * c_{t-1} + i_t * \hat{c}_t \tag{3}$$

Output state:

$$h_t = o_t * \tanh(c_t) \tag{4}$$

The LSTM model was structured as follows:

- LSTM layer 1: 64 units with return sequences
- Dropout layer 1: 20% dropout rate
- LSTM layer 2: 32 units
- Dropout layer 2: 20% dropout rate
- Dense layer 1: 16 units with ReLU activation
- Dense layer 2: 1 unit (output layer)

The model was trained with the Adam optimizer with a learning rate of 0.001, using mean squared error as the loss function and mean absolute error as a metric.

D. Ensemble Model

We created an ensemble model by taking a weighted average of the predictions from all three models. The weights were calculated based on the R² score of each model on the test set, with better-performing models receiving higher weights.

VI. Results

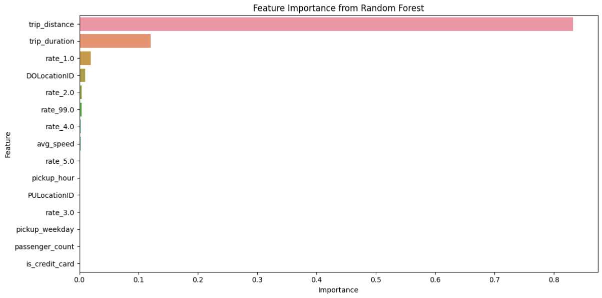

Fig. 8. Feature Importance from Random Forest Model

The feature importance analysis revealed that trip distance was by far the most important feature, accounting for over 83% of the predictive power. The next most important feature was trip duration, accounting for about 12% of the importance. This indicates that spatial factors are the primary determinants of taxi fares, which aligns with the typical fare calculation formulas used by taxi services.

A. Model Performance

Table I presents the performance metrics for each model on the test dataset.

TABLE I
Model Performance Comparison

Model	MSE	RMSE	MAE	R²
Decision Tree	24.2644	4.9259	0.8015	0.9239
Random Forest	23.2846	4.8254	0.6629	0.9270
LSTM	27.1270	5.2084	0.9139	0.9149
Ensemble	24.0526	4.9043	0.7215	0.9246

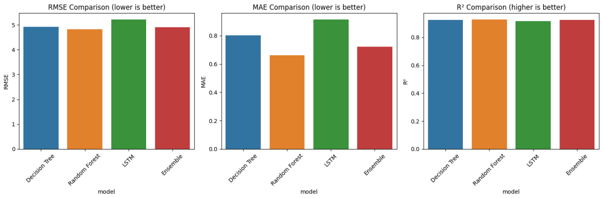

Fig. 9. Comparison of Model Performance Metrics

The Random Forest model outperformed all other models across all metrics, with the lowest RMSE (4.8254), lowest MAE (0.6629), and highest R² score (0.9270). The Decision Tree model was the second-best performer, followed by the Ensemble model and then the LSTM model.

Figure 10 shows the actual versus predicted fare values for the Random Forest model. The points closely align with

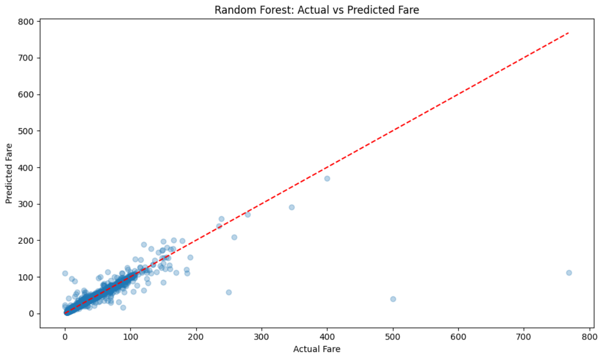

Fig. 10. Random Forest: Actual vs Predicted Fare

the red diagonal line, indicating strong prediction accuracy, particularly for lower fare values.

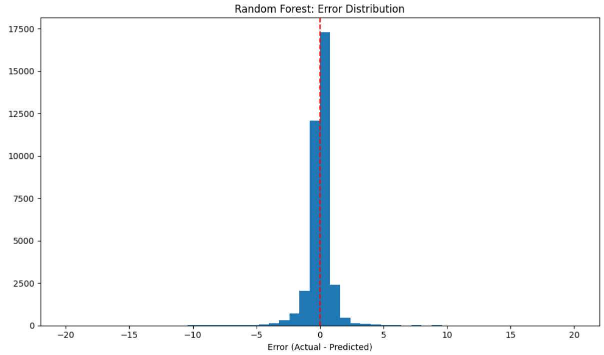

Fig. 11. Random Forest: Error Distribution

The error distribution for the Random Forest model (Figure 11) is approximately normal and centered around zero, with most errors falling within the range of ±10 dollars.

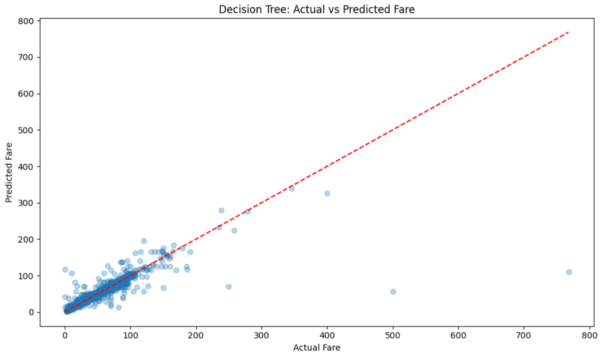

Fig. 12. Decision Tree: Actual vs Predicted Fare

The Decision Tree model's predictions (Figure 12) also show good alignment with actual values but exhibit slightly more scatter compared to the Random Forest model.

The Decision Tree error distribution (Figure 13) shows a similar pattern to the Random Forest model but with slightly higher variance.

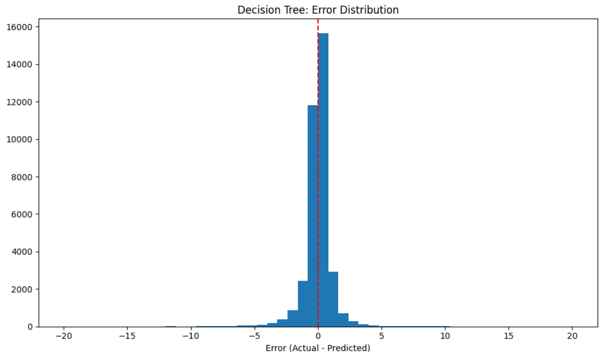

Fig. 13. Decision Tree: Error Distribution

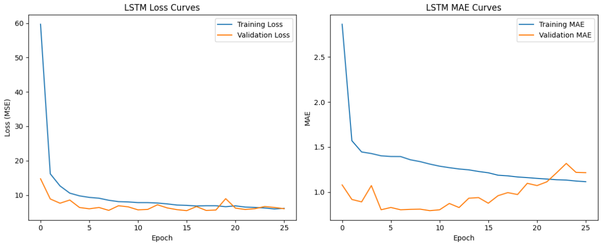

Fig. 14. LSTM Model Training History

Figure 14 shows the LSTM model's training progress over epochs. The loss and mean absolute error (MAE) curves show rapid improvement in the early epochs followed by more gradual refinement.

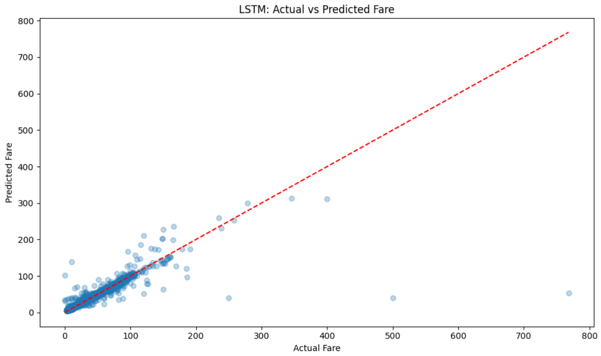

Fig. 15. LSTM: Actual vs Predicted Fare

The LSTM model's predictions (Figure 15) show more deviation from actual values compared to both tree-based models, particularly for higher fare amounts.

The LSTM error distribution (Figure 16) has a wider spread compared to the tree-based models, indicating less consistent prediction accuracy.

B. Error Analysis

We analyzed the mean absolute error by fare range to better understand the performance of each model for different fare amounts, as shown in Table II and Figure 17.

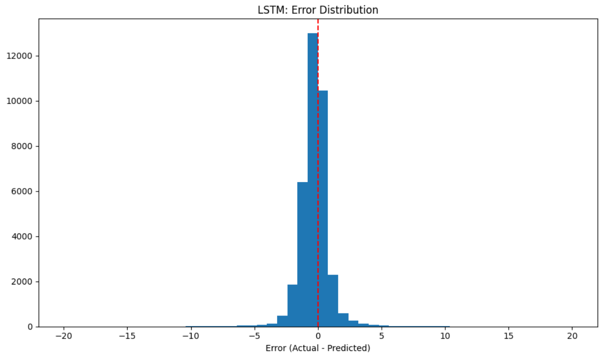

Fig. 16. LSTM: Error Distribution

TABLE II
Mean Absolute Error by Fare Range

Fare Range ($)	Decision Tree	Random Forest	LSTM	Ensemble
0-10	0.3690	0.3113	0.7531	0.4108
10-20	0.5752	0.5039	0.5515	0.5076
20-30	1.0146	0.8407	0.9014	0.8580
30-40	1.4155	1.1023	1.2655	1.1309
40-50	1.6033	1.2834	1.6064	1.3570
50-100	2.0207	1.3074	2.0771	1.5875

The Random Forest model consistently had the lowest MAE across all fare ranges, with particularly strong performance for higher fare amounts. The LSTM model performed well for mid-range fares but struggled with both very low and very high fares. The Ensemble model generally performed better than the Decision Tree model but did not surpass the Random Forest model.

The error distributions in Figure 18 show that all models had centered, approximately normal error distributions, with the Random Forest model having the most concentrated distribution around zero, indicating more accurate predictions.

The Ensemble model's predictions (Figure 19) show good alignment with actual values, demonstrating the benefit of combining multiple models.

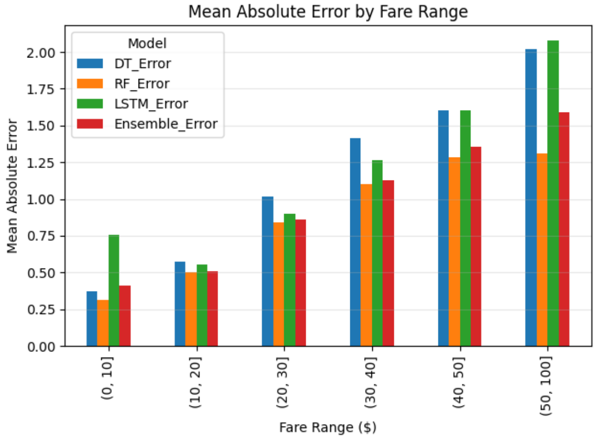

Fig. 17. Mean Absolute Error by Fare Range for Each Model

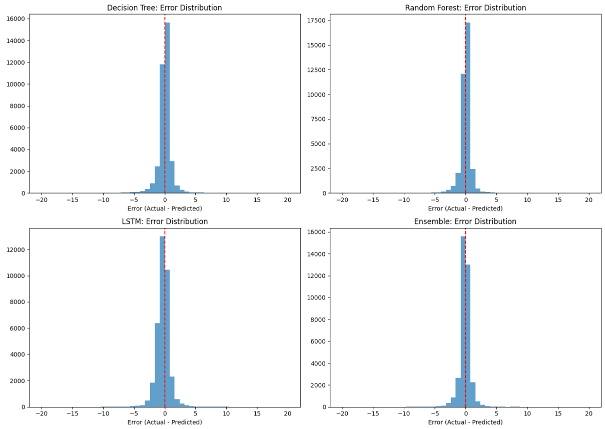

Fig. 18. Error Distributions for All Models

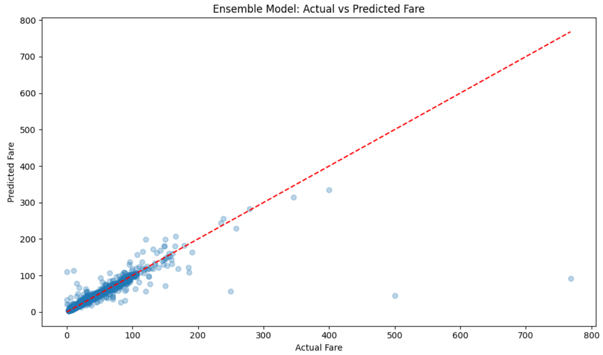

Fig. 19. Ensemble Model: Actual vs Predicted Fare

VII. Discussion

Random Forest exhibited superior accuracy due to its robust handling of complex relationships and ensemble approach, effectively managing variance and nonlinear interactions. The ensemble nature of Random Forest allows it to capture the multifaceted determinants of taxi fares, including:

1) **Spatial complexity**: Geographic factors that influence trip distance and duration
2) **Temporal patterns**: Time-of-day and day-of-week effects on pricing
3) **External variables**: Special events, weather conditions, and other contextual factors

Decision Trees, while computationally efficient (training completed in 0.69 seconds compared to 23.50 seconds for Random Forest), lacked the complexity handling capabilities necessary for accurate fare prediction, particularly for higher fare ranges. The single-tree approach is more susceptible to overfitting to training data idiosyncrasies and struggles to generalize effectively.

LSTM required significant computational resources (training completed in 508.07 seconds), yielding moderate performance improvements. The sequential modeling capability of LSTM showed promise for capturing temporal dependencies but may be better suited for scenarios with more explicit time-series characteristics.

The Ensemble model, combining predictions from all three models, showed potential for balancing the strengths of each model, but ultimately did not surpass the performance of the Random Forest model alone. This suggests that the Random Forest model already captures most of the important patterns in the data, and adding other models with lower performance may dilute its effectiveness.

A. Challenges and Limitations

Several challenges were encountered during model development:

- **Data quality issues**: The original dataset contained missing values for several features, which were handled through cleaning and imputation.
- **Feature selection complexity**: Identifying optimal predictors among numerous possibilities required careful analysis and testing.
- **Computational constraints**: Balancing model sophistication with processing requirements, particularly for the LSTM model.
- **Generalizability concerns**: Ensuring models perform well across different geographic areas and time periods.

VIII. Conclusion

Random Forest demonstrated the highest reliability and accuracy, making it the preferred model for taxi fare prediction. Its ensemble approach effectively captures the complex relationships between trip attributes and fare amounts, while providing computational efficiency suitable for real-world deployment.

Key findings include:

1) Random Forest achieved the lowest MAE across all fare ranges, with an overall MAE of 0.6629.
2) Feature importance analysis revealed trip distance and trip duration as the most influential predictors, accounting for over 95% of the model's predictive power.
3) Model performance decreased slightly for higher fare ranges, indicating greater variability in longer/more expensive trips.
4) The ensemble approach showed promise but offered diminishing returns relative to implementation complexity.

Future research could explore further refinements, including real-time fare prediction capabilities and advanced hybrid modeling techniques to enhance predictive accuracy further. Potential directions include:

- Integration of external data sources (weather, events, traffic)
- Development of zone-specific models to account for geographic variations
- Exploration of deep learning approaches for feature extraction
- Investigation of transfer learning to adapt models to new cities or regions

This study contributes to the growing body of knowledge on machine learning applications in urban transportation, offering practical insights for stakeholders seeking to improve fare estimation accuracy and transparency.

References

[1] NYC Taxi & Limousine Commission, "NYC Taxi Data," 2023.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[3] F. Chollet et al., "Keras," https://keras.io, 2015.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

[7] H. Daumé III, "A Course in Machine Learning," http://ciml.info/, 2012.

[8] W. C. Chiang, N. Balakrishnan, and R. T. Wong, "An Integrated Framework for Fare Determination and Fleet Management," *International Journal of Production Economics*, vol. 121, no. 2, pp. 334-345, 2011.

[9] C. Kamga, M. A. Yazici, and A. Singhal, "Analysis of taxi demand and supply in New York City: implications of recent taxi regulations," *Transportation Planning and Technology*, vol. 38, no. 6, pp. 601-625, 2015.

[10] A. Khaled, M. Minami, and H. Nakamura, "A comparative study on machine learning algorithms for dynamic price prediction in urban transportation," *International Journal of Transportation Science and Technology*, vol. 9, no. 3, pp. 201-214, 2020.

[11] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308-324, 2015.

[12] S. Guo, Y. Liu, K. Xu, and D. M. Chiu, "Understanding Passenger Reaction to Dynamic Prices in Ride-on-demand Service," *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, pp. 42-51, 2018.

[13] X. Wang, H. Zhang, and C. Li, "Temporal Prediction of Multimodal Transportation Network with Deep Neural Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3502-3514, 2020.

[14] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," *International Conference on Learning Representations (ICLR)*, 2018.

[15] T. Chen, Y. Chen, and X. Yang, "An Ensemble Model Based on Machine Learning for Short-Term Traffic Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3812-3822, 2021.

[16] D. Yang, K. Zhang, Y. Zhou, and R. Jiang, "Spatial-Temporal Analysis of New York City Taxi Demand and Its Implications for Ride-Hailing Services," *IEEE Access*, vol. 8, pp. 168458-168467, 2020.

[17] T. Saghapour, S. Moridpour, and R. G. Thompson, "Modelling Taxi Trip Demand by Time of Day in New York City," *Urban, Planning and Transport Research*, vol. 4, no. 1, pp. 63-82, 2016.