Executive Summary

Introduction

In 2019, New York City recorded over 103 million taxi trips, presenting a rich dataset that reflects complex urban mobility patterns. With the transportation sector contributing approximately $15 billion annually to the city’s economy, optimizing fare predictions is critical for improving operational efficiency and user satisfaction. This project focuses on predicting taxi fare amounts using machine learning and deep learning techniques. It aligns with the smart city initiative by utilizing real-time data for sustainable and intelligent mobility solutions.

Problem Statement

Accurate taxi fare prediction remains a challenge in a dynamic environment like NYC, where trip characteristics such as distance, traffic, and time of day constantly fluctuate. Inaccurate predictions can lead to customer dissatisfaction and inefficiencies in fare pricing. This project aims to address these issues by developing models that use historical and contextual data to estimate fares more accurately.

Objectives

1. Predict taxi fare amounts using historical trip data.
2. Classify trips into fare categories for simplified analysis.
3. Evaluate machine learning and deep learning models for fare estimation.
4. Analyze features influencing fare variability.
5. Validate model performance using statistical metrics.
6. Explore clustering for passenger behavior segmentation.
7. Offer practical recommendations for real-world application.

Methodology

The study applies k-Nearest Neighbors, Linear Regression, Decision Trees, SVR, Random Forest, Gradient Boosting, and Deep Learning models to fare prediction. Data preprocessing included feature engineering such as time-based and spatial variables. Clustering with K-Means and DBSCAN was also performed to identify usage patterns.

Key Findings

- Linear Regression achieved RMSE of 2.93 and R^2 of 0.958, demonstrating strong accuracy.
- Random Forest and Gradient Boosting showed robust performance, with Random Forest slightly outperforming in regression tasks.
- Deep Learning model had the lowest MAE (1.48), excelling in low-fare prediction but slightly underestimating high fares.
- Classification models like Logistic Regression and Random Forest achieved accuracy over 94% and F1 scores above 0.90.
- K-Means clustering (k=3) provided optimal separation per silhouette scores, highlighting trip pattern clusters.

Recommendations

- Incorporate weather, traffic, and event data to enhance prediction accuracy.
- Apply data balancing techniques to address fare class imbalances.
- Consider LSTM or Transformer architectures for capturing temporal dynamics.
- Develop a real-time dashboard for fare prediction to assist drivers and passengers.
- Use geospatial clustering to refine location-based analysis.

Conclusion

This project demonstrates that combining traditional ML, deep learning, and clustering can effectively model taxi fare dynamics. These insights support data-driven decisions in urban mobility planning and can improve pricing, dispatch, and customer experience. Overall, the models performed best for short trips and lower fare ranges, where most taxi activity is concentrated.