Predicting NYC Taxi Trip Duration Using Machine Learning

1. Introduction

This project aims to develop a machine learning model that predicts taxi fare amounts in New York City using publicly available yellow taxi trip data. Fare prices are influenced by many factors such as distance, pickup time, passenger count, and even weather conditions. The objective is to apply a complete data science pipeline to analyze the factors affecting fares and create a model that can make accurate fare predictions.

Dataset:

The dataset used is from NYC's official Taxi & Limousine Commission (TLC), specifically the *"yellow-tripdata-2024-01.csv"* file. It contains over 1 million rows and multiple features such as timestamps, location IDs, fare amounts, and passenger counts.

We split the data into training (80%) and testing (20%) sets:

- **Total samples:** 1,048,575
- **Train set:** 80% (838,860 samples)
- **Test set:** 20% (209,715 samples)
  A fixed random state was used to ensure reproducibility of results.

2. Methodology

2.1 Data Preprocessing:
- Removed records with missing or invalid fare values (e.g., total_amount ≤ 0)
- Filtered out extreme outliers (e.g., fares > 200 USD)
- Converted pickup timestamps to datetime format
- Extracted new features from datetime: hour, day, month, weekday
- Integrated weather data from external source

2.2 Exploratory Data Analysis:
- Histograms and boxplots to explore fare distributions

- Scatter plots to observe relationships
- Heatmaps for feature correlation

2.3 Feature Engineering:
- Created 'transaction_hour', 'weekend', 'precipitation' variables

2.4 Modeling:
- Models: Decision Tree, Random Forest, Gradient Boosting
- Evaluation: MAE, RMSE, R^2
- Hyperparameter tuning using GridSearchCV

3. Results & Findings

Model	MAE	RMSE	R^2
Decision Tree	~4.90	~6.08	~0.78
Random Forest	~2.79	~3.68	~0.93
Gradient Boosting	~2.03	~2.82	~0.96

Key Observations:
- Gradient Boosting had the best overall performance
- Trip distance, hour, and weather are key predictors

4. Conclusion

- NYC taxi fares can be predicted accurately using ML
- Gradient Boosting gave best performance
- Future Work: add geospatial data, test on more months

5. References

- NYC TLC Trip Data: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Scikit-learn documentation: https://scikit-learn.org/
- Pandas, Seaborn, Matplotlib libraries