Group 4

Enhancing Urban Mobility: Predictive Modeling and Clustering of NYC Taxi Trip Data

Members: Khushboo Modha, Nicholas Bartram, Florian Caffier, Sura Ali, Phuong Nguyen

---

Project Overview

The NYC Taxi and Limousine Commission (TLC) dataset provides a rich repository of historical taxi trip records, including information on pickup and drop-off locations, timestamps, passenger counts, trip distances, payment methods, and fares. This project aims to leverage machine learning (ML) techniques to analyze and extract actionable insights from this dataset, enhancing urban transportation systems and enabling data-driven decision-making.

Objective

The primary objectives of this project are:

1. **Trip Duration Prediction:** Use regression models to predict trip durations based on pickup/drop-off locations, passenger counts, and timestamps.
2. **Demand Forecasting:** Employ time-series analysis to forecast taxi demand in specific areas, assisting drivers and fleet managers in optimizing resources.
3. **Trip Clustering:** Use clustering algorithms to identify popular routes, hotspots, and high-demand timeframes.
4. **Fare Prediction:** Develop a pricing model to estimate fares, accounting for trip characteristics and external factors (e.g., weather).

Proposed Methodology

1. Data Preparation

- **Data Cleaning:** Handle missing values, incorrect entries, and inconsistencies.
- **Feature Engineering:** Extract features such as hour of the day, day of the week, distance between pickup and drop-off locations, and weather conditions.
- **Data Transformation:** Normalize numerical features and encode categorical variables.

2. Machine Learning Models

- **Regression Analysis for Trip Duration and Fare Prediction:** Train models such as Linear Regression, Random Forest, or Gradient Boosting Machines.
- **Clustering for Trip Patterns:** Use algorithms like k-Means or DBSCAN to identify patterns in pickup and drop-off locations.
- **Demand Forecasting:** Apply time-series models such as ARIMA, LSTM, or Prophet.

3. Model Evaluation

- Use evaluation metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for regression models.
- Evaluate clustering using silhouette scores and demand forecasting with mean absolute percentage error (MAPE).

Expected Outcomes

1. A robust prediction model for taxi trip duration and fares, aiding passengers and service providers.
2. Identification of high-demand areas and times to optimize fleet allocation.
3. Insights into passenger behavior through clustering of pickup and drop-off points.
4. Visualization dashboards for stakeholders to interact with predictive and clustering results.

Tools and Technologies

- **Programming Languages:** Python (NumPy, Pandas, Scikit-learn, TensorFlow/PyTorch, Matplotlib/Seaborn).
- **Data Visualization:** Tableau, Plotly, or Dash for interactive visualizations.
- **Database Management:** PostgreSQL or MongoDB for data storage and retrieval.
- **Geospatial Analysis:** Folium, GeoPandas for mapping pickup and drop-off locations.

Impact

This project will improve urban mobility in NYC by enhancing resource allocation, optimizing pricing strategies, and providing valuable insights for transportation policy and planning. It also provides a scalable framework that can be extended to other cities.