NYC Taxi Trip Data Analysis

Big Data Analytics Project - Group 55

March 20, 2025

1 Introduction

This project uses the NYC Taxi Trip dataset[1] to explore real-world taxi rides in New York City. The dataset includes millions of records, capturing key details such as pickup and dropoff locations, timestamps, fare amounts, passenger counts, and payment methods. The goal is to apply machine learning techniques to analyze this data and gain insights.

We will work on:

- Data exploration and visualization
- Feature engineering
- Machine learning pipelines
- Clustering
- Classification or regression tasks
- Optional streaming and graph-based analysis
- Model evaluation with different metrics

2 Understanding the Data

The dataset consists of multiple months of NYC taxi trip records. Below are the main columns:

- **VendorID**: Taxi company identifier.
- **tpep_pickup_datetime** and **tpep_dropoff_datetime**: Start and end time of the trip.
- **passenger_count**: Number of passengers.
- **trip_distance**: Distance traveled in miles.
- **PULocationID** and **DOLocationID**: Pickup and dropoff location IDs.
- **payment_type**: Type of payment (e.g., cash, credit card).
- **fare_amount**: Cost of the trip before additional charges.
- **total_amount**: Final cost including tips, tolls, and other charges.

[1] Available at `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

3 Data Exploration

- Identify missing values and handle them appropriately.
- Generate summary statistics and visualizations.
- Examine distributions of trip distance, fare, and total amount.
- Analyze trends based on pickup times (e.g., rush hours, weekdays vs weekends).

4 Feature Engineering

- Extract time-based features (hour, day of the week, weekend/weekday).
- Calculate trip duration from timestamps.
- Compute distance between pickup and dropoff using geographic coordinates.
- Normalize and scale numerical features.

5 Machine Learning Pipelines

- Build a preprocessing pipeline:
  - Handle missing values
  - Encode categorical variables
  - Scale numerical features
- Create a machine learning pipeline combining preprocessing and model training.

6 Clustering Analysis

- Use K-Means or DBSCAN to cluster trips based on pickup/dropoff locations.
- Analyze different types of trips using clustering results.

7 Classification and Regression

- **Classification:** Predict payment type (cash vs. card).
- **Regression:** Predict total trip fare or trip duration.
- Train different models such as:
  - Logistic Regression, Random Forest, XGBoost (classification)
  - Linear Regression, Gradient Boosting (regression)
- Tune hyperparameters and evaluate models.

8 Evaluation Metrics

- Clustering: Silhouette Score, Davies-Bouldin Index.
- Classification: Accuracy, Precision, Recall, AUC.
- Regression: RMSE, MAE, R^2 Score.

9 Advanced: Streaming and Graph Analysis (Optional)

- Implement Spark Streaming to process real-time taxi trip data.
- Build a trip graph where nodes are locations and edges are trips.
- Use GraphFrames to find most common routes and shortest paths.

10 Summary Table

Step	**Task**
Data Exploration	Visualize trip distributions and trends
Feature Engineering	Extract time, location, and numerical features
Pipelines	Automate preprocessing and model training
Clustering	Group trips based on similarities
Classification	Predict payment type
Regression	Predict fare amount or trip duration
Evaluation	Assess models using different metrics
Streaming/Graph	Optional real-time and network analysis

Table 1: Project Overview