Uber Trip Analysis Project Report

1 Project Overview

This project aims to build machine learning models to:

1. Predict **ride price** (Regression)
2. Classify **cab type** (Classification)
3. Detect whether **surge pricing** is applied (Binary Classification)

We used a dataset containing Uber trip information including ride metadata (time, location, distance), cab type, pricing, and weather data.

2 Dataset Summary

- **Total Rows:** 693,071
- **Numerical Columns:** 40+
- **Categorical Columns:** 10
- **Target Variables:** price, cab_type, surge_multiplier

Sample Features:

- Time: hour, day, month
- Trip Info: distance, source, destination
- Weather: temperature, humidity, windSpeed, summary, etc.

3 Preprocessing Steps

1. **Missing Values:** Dropped rows with missing `price` values.
2. **Feature Selection:** Selected relevant numerical and categorical features.
3. **Encoding:** Used `LabelEncoder` on categorical features; saved encoders as `.pkl` files.
4. **Feature Engineering:** Added a new binary column `is_surge` (1 if surge_multiplier 1).

4 Target Variables

Table 1: Overview of Target Variables Used in the Uber Trip Analysis Project

Task	Target Variable	Type	Description
Price Prediction	`price`	Regression	Predicts the estimated cost (in dollars) of the Uber/Lyft trip.
Cab Type Classification	`cab_type`	Multi-class Classification	Predicts the type of cab (e.g., Uber, Lyft, Shared, etc.).
Surge Detection	`is_surge` (created)(0/1)	Binary Classification (0/1)	Predicts whether surge pricing is applied (`1`) or not (`0`).

5 Models Used

Task	Model	Metrics
Price Prediction	RandomForestRegressor	MSE, RMSE
Cab Type Classification	RandomForestClassifier	Accuracy, Confusion Matrix
Surge Detection	RandomForestClassifier	Accuracy, Precision, Recall

6 Model Evaluation

Price Prediction

- **MSE:** 7.45
- **RMSE:** ≈ 2.73
- **Interpretation:** On average, predicted prices deviate by $2.73 from actual values.

Cab Type Classification

- **Accuracy:** ∼99%
- **Classes:** Lyft, Uber, Shared, etc. (LabelEncoded)

Surge Detection

- **Accuracy:** ∼97%
- **Binary Classification:** 0 = No Surge, 1 = Surge Applied

7 Model Persistence

All trained models and encoders are saved using `joblib`:

- price_model.pkl
- cabtype_model.pkl
- surge_model.pkl
- source_encoder.pkl, destination_encoder.pkl, etc.

8 Sample Prediction

- **Input:**
  - hour = 14, day = 15, month = 12
  - distance = 2.5
  - temperature = 45.0, humidity = 0.6
  - windSpeed = 5.0
  - source = Haymarket Square, destination = North Station
  - name = Shared, weather = Clear
- **Predictions:**
  - Predicted Price: $7.24
  - Predicted Cab Type: Uber
  - Surge Applied: No

9 Directory Structure

```
uber_trip_analysis/

 models/
    price_model.pkl
    cabtype_model.pkl
    surge_model.pkl

 encoders/
    source_encoder.pkl
    destination_encoder.pkl
    .
    .

 data/
   uber_data.xlsx
    about_uber_data.txt

notebook/
    EDA.ipynb
    Model Training.ipynb
    prediction.ipynb
 README.md
```

10 Conclusion

This project demonstrates the practical use of machine learning for:

- Predicting ride prices from features
- Classifying cab types
- Detecting surge pricing

It can benefit riders, drivers, and companies in ride fare estimation, service selection, and demand forecasting.