Uber Fare Prediction

Project title and category:

Title: Uber Fare Prediction

Category: Data Science / Predictive Analytics

Problem Statement or Motivation:

Accurate fare estimation is essential for ride-hailing platforms such as Uber, Lyft, and Ola. Customers expect transparency, while companies rely on precise fare predictions to manage demand, pricing, and profitability. However, predicting trip fares is complex because several factors influence pricing pickup and drop-off locations, trip distance, time of day, passenger count, and real-world variability such as traffic or unusual travel patterns.

Traditional rule-based fare systems struggle because they cannot adjust to dynamic conditions. This often results in inconsistent pricing and poor customer satisfaction.
To overcome this challenge, this project aims to build a machine-learning-based fare prediction system using historical Uber trip data. The goal is to deliver more accurate, logical, and reliable fare estimates that improve both user experience and business operations.

Planned Approach and Technologies:

The workflow begins with loading and exploring the dataset to understand its structure and detect missing, inconsistent, or abnormal values. The data is then cleaned by removing null entries, unrealistic fares, invalid passenger counts, and incorrect geographic coordinates.

Next, feature engineering is performed. This includes extracting meaningful time-based components (year, month, hour, day of week) and calculating travel distance using the Haversine formula for more accurate geographic representation. Visualization techniques are applied to study the relationships between distance, fare, and time to better understand fare patterns and outliers.

Once the data is prepared, multiple machine learning models will be trained and evaluated. These include:

1. Decision Tree Regressor
2. Random Forest Regressor
3. Gradient Boosting Regressor
4. AdaBoost Regressor
5. K-Nearest Neighbors Regressor
6. Support Vector Regressor (SVR)
7. Ridge Regression
8. Lasso Regression

All models will be compared to identify the most accurate one, with early testing indicating **Random Forest Regressor** delivers the highest accuracy and stability.
The entire process is implemented using Python, Pandas, NumPy, Matplotlib, and Scikit-learn, ensuring a reusable and scalable workflow.

Expected Challenges and Solutions:

- **Data Quality Issues:** Addressed through rigorous cleaning and validation.
- **Outliers:** Managed using statistical thresholds and visual analysis.
- **Feature Scaling:** Solved through normalization to support algorithms like KNN and SVR.
- **Overfitting:** Controlled with train-test splits, model tuning, and regularization.

Success Criteria:

- High R^2 score and low MAE.
- Predictions follow logical fare patterns.
- Complete, automated pipeline from ingestion to prediction.
- Model saved and deployable for real-time use.

Stretch Goals:

- Explore more advanced ensemble models.
- Build a web interface for real-time fare prediction.
- Add external data (traffic, weather).
- Create geospatial visual dashboards.
- Package the model as an API for integration.

Dataset Link:

https://www.kaggle.com/datasets/kushsheth/uber-ride-price-prediction

---