Uber Fare Prediction Project Report

Objective

To build a regression model to predict fare amounts for Uber rides based on ride features like distance, time, and passenger count. This analysis will aid in optimizing pricing strategies, improving operations, and enhancing customer and driver satisfaction.

Steps Undertaken

1. Data Understanding & Preprocessing

- **Dataset Columns**:
  - `pickup_longitude`, `pickup_latitude`, `dropoff_longitude`, `dropoff_latitude`, `fare_amount`, `passenger_count`, `pickup_datetime`.
- **Challenges Addressed**:
  - Removed outliers in fare amounts and distances.
  - Derived new features:
    - **Trip Distance (`distance_km`)**: Calculated using Haversine formula.
    - **Time Features**: Extracted `pickup_hour` and categorized rides into time slots (morning, afternoon, evening).

---

2. Exploratory Data Analysis (EDA)

- **Insights**:
  - Longer distances lead to higher fares (strong correlation).
  - Fares are higher during peak hours (morning and evening).
  - Passenger count minimally impacts fare but may affect shared ride dynamics.

---

3. Model Building and Evaluation

Three regression models were tested:

1. **Linear Regression**:
   - MSE: 102.80
   - R^2: 0.00026 (poor fit).
2. **Decision Tree Regression**:
   - MSE: 146.94
   - R^2: -0.42 (underfitting).
3. **Random Forest Regression** (Selected Model):
   - MSE: 80.89
   - R^2: 0.21 (best fit).

---

4. Model Fine-Tuning

- Used **Grid Search CV** to optimize hyperparameters for the Random Forest model:
  - Best Parameters: `n_estimators=200`, `max_depth=10`.
  - Improved performance with reduced overfitting.

---

5. Predictions on New Data

- Predictions were made for new ride data using the fine-tuned Random Forest model.
- Demonstrated scalability for real-world application.

Key Recommendations

1. Pricing Strategies

- Introduce **dynamic pricing** for peak hours and long-distance rides.
- Offer discounts on short trips to attract more local commuters.

2. Driver Incentives

- Provide bonuses for accepting long rides and for working during peak hours.

3. Fleet Management

- Optimize driver placement based on high-demand pickup/dropoff zones identified in the data.
- Promote shared rides for trips with multiple passengers to maximize profitability.

4. Future Improvements

- Incorporate additional data like traffic, weather, and ride cancellation reasons for better accuracy.

- Periodically retrain the model with updated data for continuous improvement.

Conclusion

The project successfully developed a predictive model for Uber fares, with actionable insights for pricing, incentives, and operational efficiency. Implementing these recommendations will enhance revenue generation and customer satisfaction while ensuring optimal resource utilization.