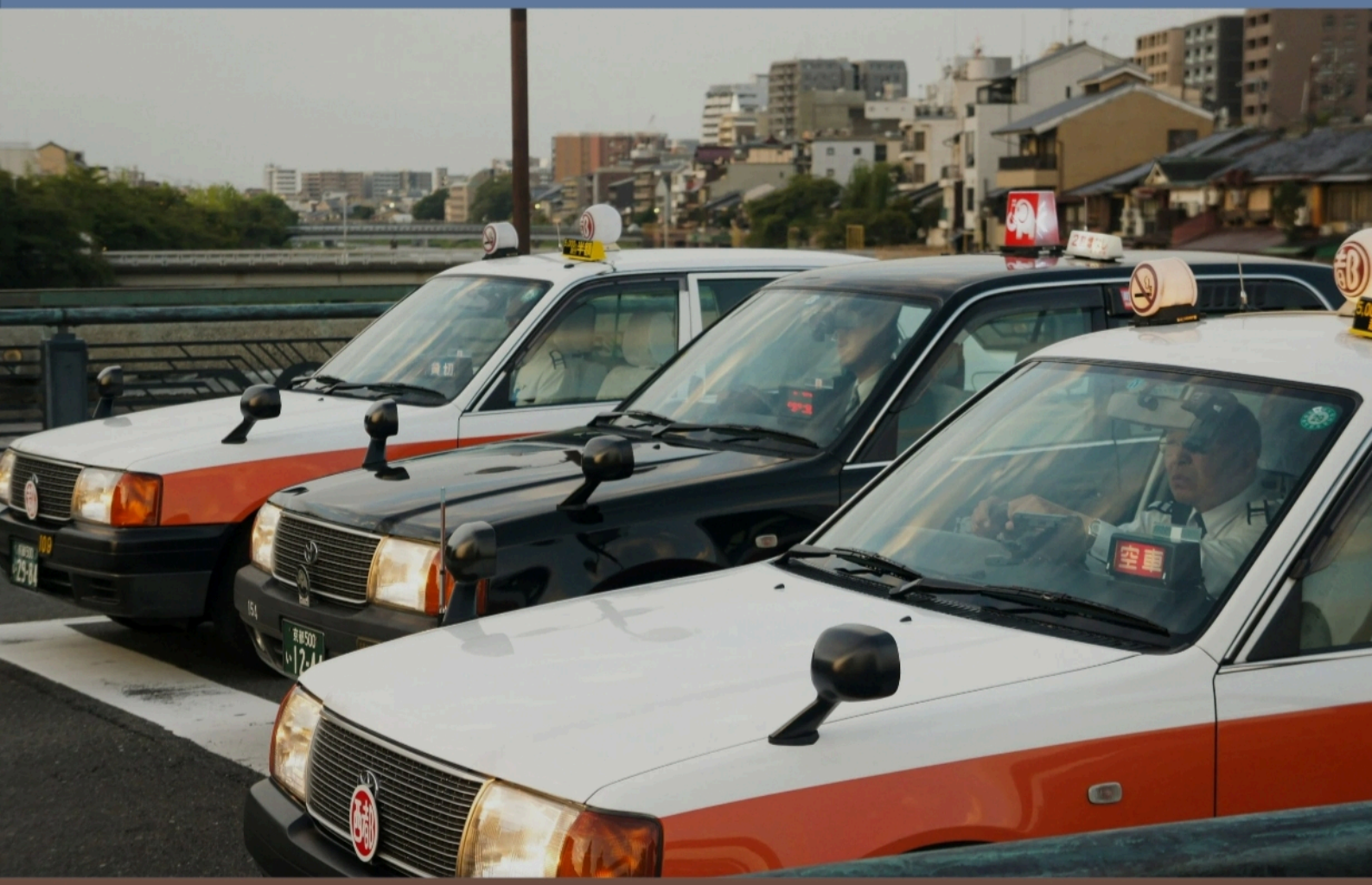
Uber Fare Prediction

Using Machine Learning to Predict Uber Fares



Harshal Bhendare

Table of Contents

Harshal Bhendare

Introduction

This report presents a comprehensive analysis of fare prediction for Uber, using various statistical and machine learning models. The primary objectives were to identify the factors influencing fare amounts, predict Uber fare prices accurately, and provide actionable recommendations to improve business strategies such as dynamic pricing, driver incentives, and service quality.

Harshal Bhendare

Data Overview and Initial Insights

The dataset included key features such as fare amounts, passenger count, trip distance, and geographical coordinates (pickup and dropoff locations). Initial exploration revealed several important aspects:

- **Fare Amount:**

The average fare was $11.36, with potential outliers (minimum fare of -$52.00 and maximum of $999.99) that could skew analysis.

- **Passenger Count**:

Average passenger count was 1.68, with unrealistic values like 0 and 208, which indicated data quality issues.

- **Trip Distance:**

The average trip distance was 2.39 miles, with distances ranging from 0 to 33.96 miles.

General Observations:

Outliers in fare and passenger count suggested the need for cleaning or further investigation. Descriptive statistics and visualizations (e.g., histograms) helped understand the distribution of key features.

Correlation Analysis (Heatmap Insights)

Fare Amount and Distance:
A strong positive correlation (around 0.8) was observed between fare amount and trip distance, confirming that distance is a primary driver of fare.

Passenger Count and Fare:
A weak but positive correlation (0.02) suggested that the number of passengers might have a slight effect on fare.

Geographical Coordinates:
Pickup and dropoff coordinates showed correlations among themselves but had weak correlations with fare.

Overall, the strongest determinant of fare amount was distance, while time variables (year, month, day, hour) and geographic features played a less significant role.

Modeling and Predictions

Ordinary Least Squares (OLS) Regression

OLS regression was used to model the relationship between fare amounts and several predictor variables. Key insights:
R-squared: 0.779, indicating that 77.9% of the variance in fare amounts could be explained by the model's features.
Coefficients: Distance had the highest positive coefficient, reaffirming its strong influence on fare. Other features like year, month, day, hour, and geographical coordinates were statistically significant but had smaller effects.
P-values: Distance, time variables, and coordinates were statistically significant, while passenger count was less significant.
Key Recommendation: Based on OLS analysis, dynamic pricing strategies should be primarily based on distance, with some adjustments for time and geographic factors.

Random Forest Model and Hyperparameter Tuning

A Random Forest model was later employed to improve prediction accuracy. The model was optimized through hyperparameter tuning:
Number of Trees (n_estimators): Initial model used 10 trees, with tuning likely optimizing this for better performance.
Out-of-Bag Score: Enabled to estimate model performance on unseen data.
Random Forest Results:
R-squared (R^2): The final R^2 score for the Random Forest model was 0.7991, indicating that 79.91 % of the variance in Uber fare amounts was explained by the model's features.
Mean Squared Error (MSE): The final MSE value was 5.91, representing the average squared difference between predicted and actual fare amounts.
Root Mean Squared Error (RMSE): The RMSE, which gives the error in the same units as the predicted variable (fare amount), was 2.43. This shows the average deviation from the actual fare in dollars, suggesting that on average, the predicted fare is within $2.43 of the actual fare.

Potential Use Cases and Business Impact

The results of these models provide several use cases to optimize both customer experience and operational efficiency for the ride-sharing company:

Fare Estimation for Riders:
Real-time fare estimates can be provided, improving transparency and decision-making for riders.

Dynamic Pricing:
Adjust fares based on demand, time of day, location, and distance to optimize revenue while ensuring fair pricing.

Driver Revenue Optimization:
Use predictive insights to inform drivers of optimal times and locations to maximize their earnings.

Business Planning:
Analyze fare trends and patterns to make informed decisions on operations, marketing, and fleet management.

Recommendations

Based on the insights from both models and statistical analysis, the following recommendations are proposed:

Dynamic Pricing Optimization:

Implement dynamic pricing based on time factors like hour, day, month, and year to capture demand fluctuations.

Adjust fare rates during peak hours or holiday seasons to maximize revenue while maintaining customer satisfaction.

Targeted Driver Incentives:

Since distance is the strongest predictor of fare, offer incentives to drivers for long-distance trips, especially during low-demand periods, to ensure better service coverage.

Geographical Pricing Adjustments:

Adjust pricing based on pickup and dropoff locations. Higher base fares may be justified in high-demand areas, while promotional pricing can encourage more rides in underutilized zones.

Service Quality Enhancements:

Use data insights to improve the rider experience during high-demand periods by optimizing arrival time estimates and communication between riders and drivers.

Predictive Maintenance:

Periodically retrain the model with updated data and monitor performance to maintain high prediction accuracy. Ensure that the system can detect anomalies, such as potential fraud.