Fare Amount Prediction Project

Introduction

This project focuses on predicting Uber fare amounts using a dataset containing details of Uber trips,
such as pickup and dropoff locations, timestamps, passenger count, and actual fare amounts.
Data preprocessing, feature engineering, visualization, and machine learning models were employed
to build an accurate fare prediction system. The purpose of this analysis is to understand the relationship
between trip features and fare amounts, and to create a model that can accurately predict fare amounts
based on these features.

Dataset Information

The dataset used in this analysis is publicly available and contains information about Uber trips,
including the pickup and dropoff locations (latitude and longitude), timestamps, the number of passengers,
and the corresponding fare amounts. The dataset required extensive preprocessing, including the handling
of missing values and the removal of outliers.

Data Preprocessing

Data preprocessing is a critical step in any data science project. It ensures that the data is clean
and ready for analysis. In this project, the following preprocessing steps were applied:

- Handling missing values: Rows with missing values in the critical columns were dropped.

- Outlier removal: Trips with fare amounts greater than $200 or with unrealistic trip distances were removed.

- Feature extraction: New features, such as 'pickup_hour', 'pickup_day', and 'distance_km' (calculated using

the Haversine formula), were created to enrich the dataset for better predictions.

Feature Engineering

Feature engineering plays a significant role in improving the performance of machine learning models.

For this project, additional features were extracted from the 'pickup_datetime' column:

- 'pickup_hour': The hour at which the trip began, which can influence fare due to traffic conditions.

- 'pickup_day': The day of the week, which can capture weekend or weekday effects.

- 'distance_km': Calculated using the Haversine formula, which gives a more accurate distance between the

pickup and dropoff locations based on their latitude and longitude.

Data Visualizations

Several visualizations were created to better understand the data:

- A histogram was plotted to display the distribution of fare amounts, showing that most fares were below $50.

- A scatter plot was used to show the relationship between fare amounts and trip distances, indicating

that fare increases with distance.

- A boxplot revealed variations in fare amounts across different days of the week, suggesting that

fares

might be higher on certain days, such as weekends.

These visualizations provided valuable insights into the dataset and informed the feature engineering process.

Correlation Analysis

A correlation heatmap was generated to examine the relationships between numerical variables in the dataset.

Key insights include:

- A strong positive correlation between fare amount and trip distance.

- A weaker but still significant correlation between fare amount and the pickup hour, possibly reflecting

peak-hour effects.

Understanding these correlations helped in selecting the most important features for model training.

Model Training and Evaluation

Two machine learning models were trained to predict fare amounts:

1. Linear Regression: This model establishes a linear relationship between the independent variables and

the target variable (fare amount).

2. Gradient Boosting Regressor: A more advanced ensemble model that iteratively improves predictions

by correcting the errors of weaker models.

The dataset was split into training and testing sets, and both models were evaluated using metrics like

Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and

R-squared (R^2) score.

Model Performance

Performance of the models on the test data:

- Linear Regression:
  - MSE: 22.88
  - RMSE: 4.78
  - MAE: 3.45
  - R^2: 0.67
- Gradient Boosting Regressor:
  - MSE: 15.89
  - RMSE: 3.99
  - MAE: 2.87
  - R^2: 0.78

Gradient Boosting significantly outperformed Linear Regression, as seen from its lower error metrics and

higher R-squared value, making it the better model for this prediction task.

Feature Importance

For the Gradient Boosting model, feature importance analysis revealed the following:

- The most important feature was the trip distance ('distance_km'), which had the highest impact on fare prediction.
- Pickup and dropoff coordinates also played a significant role, followed by the pickup hour and passenger count.

This analysis helps in understanding which factors most influence fare amounts and can guide further feature

selection and engineering efforts.

Predictions on New Data

The trained Gradient Boosting model was used to predict fare amounts for new Uber trips based on the features
of the trips, such as pickup and dropoff coordinates, passenger count, and trip distance. The model performed
well, providing reasonable fare estimates for the new data, closely matching actual fare trends.

Conclusion

In this project, we successfully built and evaluated two machine learning models to predict Uber fare amounts.
Through feature engineering, data visualization, and model training, we identified the key factors affecting
fare amounts and developed a reliable prediction system. The Gradient Boosting model outperformed the
Linear Regression model, and its feature importance analysis highlighted the importance of trip distance
and coordinates in fare prediction. This project demonstrates the effectiveness of machine learning
in real-world applications such as ride fare prediction.