



EAST WEST UNIVERSITY

Department of Computer Science and Engineering

Course Title: Statistics for Data Science (CSE303)

Section: 07

Project Report

Project Title: Taxi Price Regression

Submitted By:

MD Ryhan Munna ID: 2022-3-60-166

Mahir Shahriar ID: 2022-3-60-197

Khaled Mahmood Sifat ID: 2022-3-60-217

Date of Submission: 20/Jan/2025

The GitHub repository link: https://github.com/MMahir007/Taxi_Trip_Pricing_Project.git

Submitted To

Puja Chakraborty

Lecturer

Department of Computer Science and Engineering

East West University

Introduction

Background

The dataset represents taxi trip data that includes various factors influencing pricing, such as trip distance, passenger count, traffic conditions, weather, and time variables. Predicting trip prices is crucial for taxi operators and ride-hailing companies to optimize pricing strategies, improve customer satisfaction, and ensure operational efficiency. Machine learning models, such as Linear Regression, are commonly used to identify relationships between the trip price and the influencing factors, enabling predictive capabilities and data-driven decision-making.

Motivation

Accurate prediction of taxi trip prices has significant benefits:

1. For Companies: Optimizing fare calculations to stay competitive while ensuring profitability.
2. For Customers: Providing fair pricing based on trip details and real-time conditions like traffic or weather.
3. For Urban Planning: Understanding trends in taxi usage can aid city planners in optimizing transportation networks and traffic management.
4. Quality Assurance: By training a model on the given dataset, the aim is to validate the data's accuracy and ensure it aligns with real-world pricing patterns.

Using Linear Regression, a foundational machine learning technique, allows us to assess the integrity and predictive capability of the dataset while ensuring a simple, interpretable model.

Problem Statement

The primary goal is to predict the trip price (Trip_Price) of taxi rides based on several features such as trip distance, time of day, day of the week, traffic conditions, and more. However, challenges arise due to:

1. Missing Data: Several columns have missing values, which could impact model performance.
2. Data Distribution: Ensuring that the split between the training (80%) and testing (20%) datasets provides an accurate representation of real-world scenarios.
3. Model Validity: By fitting a Linear Regression model, the aim is to:
 - Verify the dataset's quality and ensure it correlates logically with the trip pricing.
 - Identify any potential issues, such as overfitting, underfitting, or lack of feature significance.

This approach provides a foundation for improving predictions and enhancing future implementations using more complex models or additional data.

Dataset

General Overview:

1. Key Factors:

- Target Column: Trip_Price.
- Number of Row: 1000
- Number of Column: 11
- Dataset also holds some missing data points in each column.

2. Column Names and Types:

- Trip_Distance_km (float64): Distance of the trip in kilometers.
- Time_of_Day (object): The time of day (e.g., Morning, Afternoon, Evening).
- Day_of_Week (object): The day of the week (e.g., Weekday, Weekend).
- Passenger_Count (float64): Number of passengers.
- Traffic_Conditions (object): Traffic level (e.g., Low, High).
- Weather (object): Weather conditions (e.g., Clear, Rainy).
- Base_Fare (float64): Initial base fare for the trip.
- Per_Km_Rate (float64): Cost per kilometer.
- Per_Minute_Rate (float64): Cost per minute of trip duration.
- Trip_Duration_Minutes (float64): Duration of the trip in minutes.
- Trip_Price (float64): Total price for the trip.

Methodology

1. **Data Understanding:** Project requires analyzing the dataset and finding the targeted parameter to be predicted. It also needs to know the relation between each column with the targeted one.

2. Data Preprocessing:

- The dataset holds some missing data points, so that those data points should be retrieved as precisely as possible. And that is why we replaced the null values with median value and mode for respectively continuous numerical data and categorical data.
- We have performed Label Encoding method to convert categorical data to numerical data.

3. **Data Splitting:** We have split the whole dataset 80% data for training and 20% data for testing the model.

4. **Model Development:** Linear Regression Model is used for the analysis. In the model we have fitted the featured columns (All columns except Trip_Price) in X and the targeted column (Trip_Price) in Y and trained the model. The impact of every individual feature of the dataset on the targeted column (Trip_Price) is observed significantly

5. Model Evaluation

The model's performance is evaluated based on the following key performance metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions without considering their direction.
- **Mean Squared Error (MSE):** Penalizes larger errors more heavily, providing insight into extreme deviations.
- **R-squared (R^2):** Indicates the proportion of variance in the target variable explained by the model.

Experimental Result

Polynomial Regression

- MAE Linear Regression: 5.5723054998993256
- MSE Linear Regression: 70.73577003473478
- RMSE Linear Regression: 8.410455994459205

Linear Regression

- MAE Linear Regression: 9.869196279929023
- MSE Linear Regression: 191.62808455244746
- RMSE Linear Regression: 13.842979612512888

Conclusion

The results demonstrate that Polynomial Regression significantly outperforms Linear Regression:

Polynomial Regression:

MAE: 5.572
MSE: 70.736
RMSE: 8.410

Linear Regression:

MAE: 9.869
MSE: 191.628
RMSE: 13.843

This indicates the data has a nonlinear relationship that Polynomial Regression successfully captures, leading to more accurate predictions. The reduction in MAE, MSE, and RMSE highlights its superiority in minimizing both average and large prediction errors.

Future Direction

1. Tune Polynomial Degree: Test higher degrees to enhance performance while avoiding overfitting.
2. Regularization: Apply Ridge or Lasso to handle overfitting as model complexity increases.
3. Feature Engineering: Explore interactions or transformations to improve input representation.
4. Model Comparison: Evaluate advanced models like SVR, Decision Trees, or Neural Networks.
5. Residual Analysis: Check residuals for unmodeled patterns to refine the model.
6. Cross-Validation: Validate the model on multiple folds to ensure consistency.
7. Dataset Expansion: Increase data size and diversity for better generalization.

Final Thoughts

The significant improvement of Polynomial Regression suggests its suitability for modeling nonlinear relationships. Further work should focus on tuning and testing advanced methods to enhance performance and robustness.