PROJECT REPORT: PREDICT THE FARE AMOUNT OF FUTURE RIDES USING REGRESSION ANALYSIS

Submitted by - Aparna K

JANUARY 12, 2025

Acknowledgement

- **MentorMind, upGrad**

Introduction

This project focuses on predicting the fare amount for Uber rides using regression analysis. By leveraging historical ride data, a linear regression model was trained to estimate fare amounts based on key features such as trip distance, time of day, and day of the week. The project provides insights into the most influential factors affecting fares and demonstrates how predictive models can be used to enhance operational efficiency and customer experience.

Objective:

- Build a regression model and accurately develop a predictive model.
- Identify key variables such as distance,time, pickup/dropoff locations and so on.
- Perform data preprocessing, feature engineering, and evaluate various regression models.
- Deploy and Automate Predictions.

Significance:

- Improved Customer Experience
- Optimizing resource allocation
- Improving profitability in ride-hailing services
- Enabling transparent pricing

Methodology

1. **Data Preparation:**

   Data Collection: Historical ride data was used, including features such as:

   a. Distance (in miles)
   b. Time of day (e.g., peak or off-peak hours)
   c. Day of the week (e.g., weekday or weekend)

   Preprocessing Steps:

   d. Handling missing values.
   e. Encoding categorical variables.
   f. Normalizing and scaling numerical features.
   g. Splitting the dataset into training (80%) and testing (20%) sets.

2. Model Development

Model Used: Linear Regression

Training Process:

- The model was trained using the training dataset.
- Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) were calculated.

3. Model Performance Evaluation

To evaluate the trained regression models, we calculated key metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Below are the results for Linear Regression, Decision Tree, and Random Forest models:

Linear Regression

- **Train Accuracy**: 0.6515
- **Test Accuracy**: 0.6544
- **MAE (Test)**: 5133.47
- **MSE (Test)**: 9109266550.84
- **R^2 (Test)**: -87568947.71

Observation: Indicates good generalization, as there is no significant gap between the accuracies. Accuracy is moderate but consistent, making this model a reliable choice if interpretability is essential.

Decision Tree Regression

- **Train Accuracy**: 0.7196
- **Test Accuracy**: 0.6855
- **MAE (Test)**: 2.57
- **MSE (Test)**: 34.99
- **R^2 (Test)**: 0.66

Observation: Slight overfitting is present but not excessive. The higher accuracy on both training and testing sets suggests the decision tree captures the patterns in the data well, though it may not generalize as consistently as linear regression.

Random Forest Regression

- **Train Accuracy**: 0.4765
- **Test Accuracy**: 0.4774
- **MAE (Test)**: 2.47
- **MSE (Test)**: 33.48
- **R^2 (Test)**: 0.68

Observation: Both scores are significantly lower compared to the other models, indicating underfitting. The random forest model is not capturing the data's complexity, likely due to inadequate tuning of hyperparameters like the number of estimators or tree depth.

Feature Importance Analysis

Regression models provide insights into the contribution of each feature to the target prediction. Below are the key features and their importance derived from the Random Forest model:

Top Features Impacting Fare Amount:

1.Distance between pickup and dropoff locations : It primarily determines the fare. It ensures accurate fare predictions by reflecting the core calculations, such as rates per mile or kilometer.

Implications: Ensures accurate fare predictions by reflecting core fare calculations (e.g., per mile/kilometer rates). Longer distances equate to higher fares.

2.Pickup and dropoff locations : This feature is important because geographic variations influence pricing. Certain areas, such as airports or city centers, may include surcharges or experience higher demand.

Implications: Identifies hotspots and regional trends, helping optimize fleet placement and route planning.

3.Time-Based Features (Hour, Day, and Month): High impact due to fare variations during peak hours, weekends, and holidays.

Implications: Captures temporal demand fluctuations, allowing optimization of surge pricing and better ride availability.

4.Passenger Count : Moderate, as group sizes or shared rides might influence fare types or vehicle requirements.

Implications: Differentiates ride types (e.g., standard vs. shared), influencing both fare calculation and service offerings.

Predictions on New Data

The trained Random Forest model was applied to new ride data to predict fare amounts. Below is an example workflow:

1. **Input Data**: Include features such as distance, time of day, pickup/drop-off locations, and weather conditions.
2. **Apply Model**: Use the trained Random Forest model to predict fare amounts.
3. **Output**: Generate fare predictions for new rides.

Use Cases:

- **Dynamic Pricing**: Predict fares based on real-time conditions (e.g., traffic, weather) to optimize revenue.
- **Operational Efficiency**: Help drivers plan routes and maximize earnings.
- **Customer Transparency**: Provide upfront fare estimates to customers for better service satisfaction.

Recommendations

Based on the regression model insights and predictions, the following recommendations are proposed for the ride-sharing company:

1. Integrate the model into Uber’s fare estimation system for real-time predictions.
2. Explore additional features, such as traffic patterns and weather conditions, to improve model accuracy.
3. Regularly retrain the model with updated ride data to ensure reliability.
4. Use feature importance analysis to refine pricing strategies and develop customer-centric promotions.

Conclusion

The regression model effectively predicts ride fares, offering consistent performance across datasets. Key insights into fare determinants, such as distance and time of day, provide actionable strategies for dynamic pricing and operational enhancements. The approach demonstrates significant potential for improving customer satisfaction and business profitability.