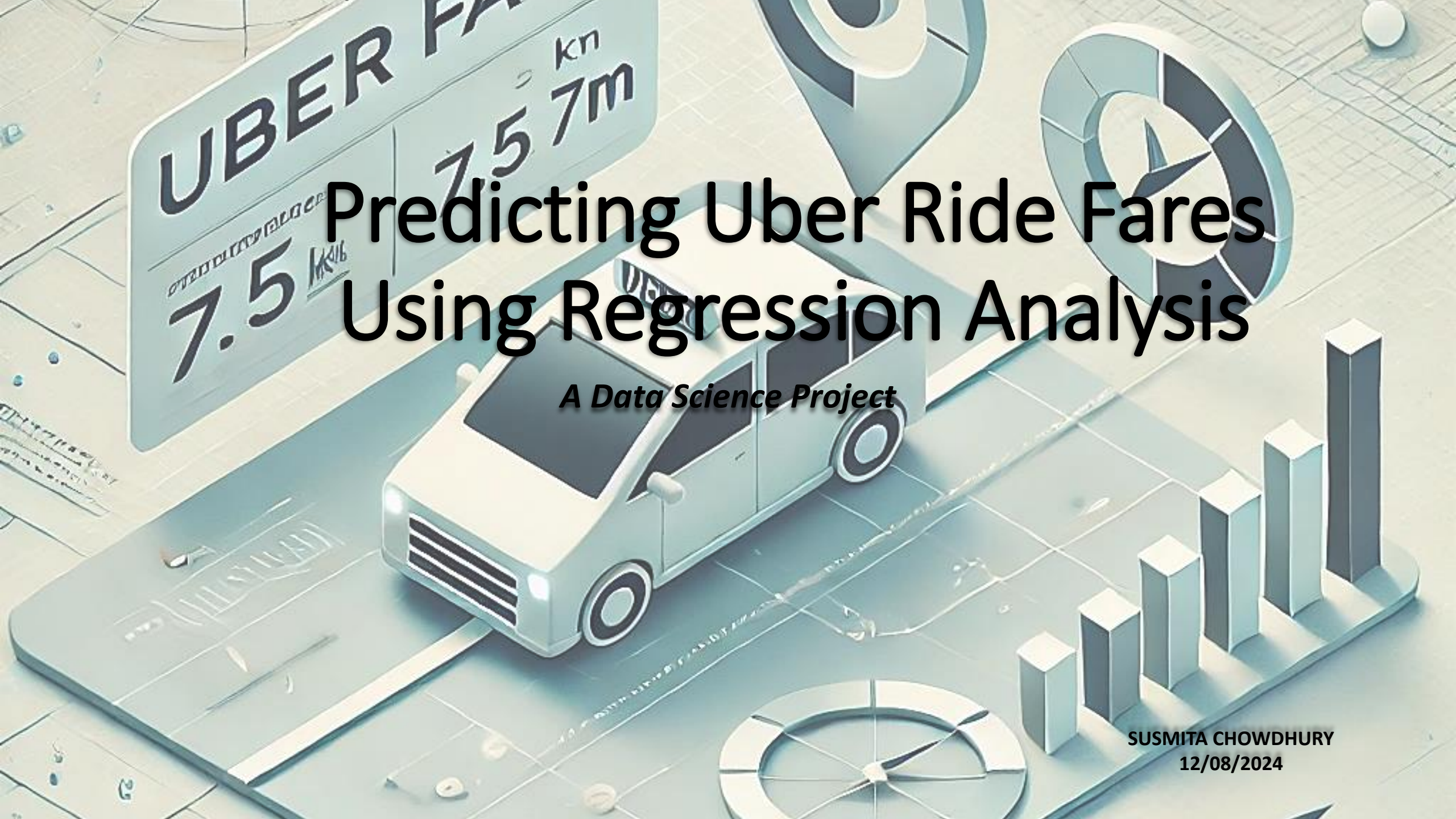


Predicting Uber Ride Fares Using Regression Analysis

A Data Science Project

SUSMITA CHOWDHURY
12/08/2024

Objective:

- The goal of this project is to predict the fare amount for Uber rides based on historical ride data using regression analysis.

- This involves understanding the factors that influence fare pricing and developing a model that can estimate the fare for future rides.

Motivation:

- Accurate fare prediction is crucial for both passengers and the ride-sharing company to ensure transparency and fairness.

- Predictive models can help optimize pricing strategies and improve customer satisfaction.

Data Description

- Dataset Overview:

  - The dataset consists of 1000 synthetic Uber ride records.
  - Key features include:
    - *pickup_datetime:* The date and time when the ride started.
    - *pickup_latitude and pickup_longitude:* GPS coordinates of the pickup location.
    - *dropoff_latitude and dropoff_longitude:* GPS coordinates of the drop-off location.
    - *distance_km:* The calculated distance of the ride in kilometers.
    - *fare_amount:* The fare charged for the ride in USD.

- Data Source:

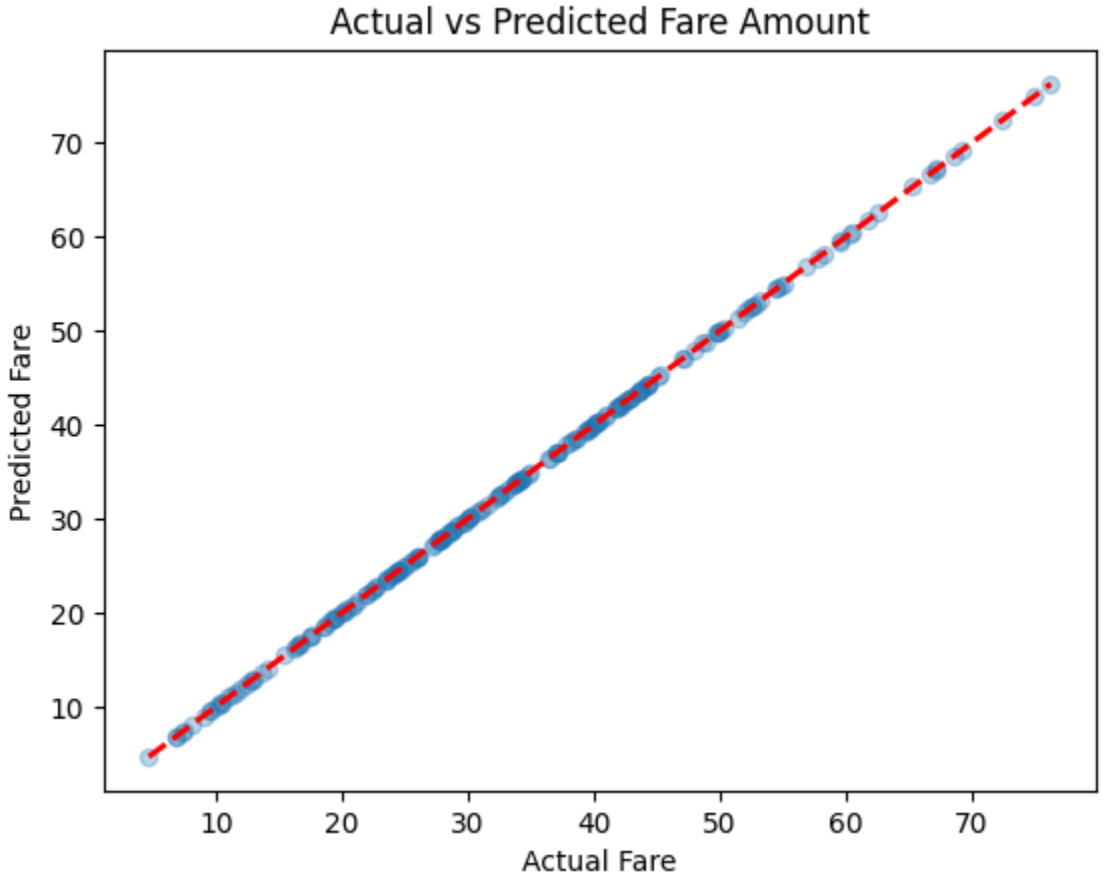
  This data was synthetically generated to mimic real-world Uber ride data for the purpose of this project.

Data Cleaning and Preprocessing

- **Handling Missing and Erroneous Values**:
  - Rows with missing fare amounts were removed.
  - Unreasonable fare values (e.g., negative or excessively high fares) were filtered out.
  - Latitude and longitude values were constrained within reasonable geographic boundaries (e.g., New York City area).
- **Feature Engineering**:
  - **Distance Calculation**: Using the Haversine formula, we calculated the distance between the pickup and drop-off points.
  - **Hour Extraction**: Extracted the hour of the day from the *pickup_datetime* feature to capture time-of-day effects on fare pricing.

Exploratory Data Analysis

- **Visualization of Fare Distribution**:
  - A histogram showing the distribution of fare amounts, revealing the most common fare ranges and identifying any outliers.
- **Correlation Analysis**:
  - A heatmap showing the correlation between different features (e.g., distance, hour of the day, fare amount).
  - Key insight: The distance traveled has a strong positive correlation with the fare amount.
- **Distance vs. Fare Scatter Plot**:
  - A scatter plot demonstrating the relationship between ride distance and fare, with a clear trend that longer distances generally lead to higher fares.



Model Selection

❖ **Why Regression?**:

•Regression analysis is well-suited for predicting continuous outcomes, such as fare amounts, based on input features.

•Linear Regression was chosen for its simplicity, interpretability, and effectiveness in capturing linear relationships between features.

❖ **Other Models Considered**:

•Polynomial Regression: Considered for potential non-linear relationships, but the linear model performed sufficiently well.

•Decision Trees/Random Forests: Not selected due to the simplicity of the problem and the sufficient performance of linear regression.

Model Training and Evaluation

- **Training Process**:
  - The dataset was split into training and testing sets (80% training, 20% testing).
  - Features were standardized to improve model performance.

- **Model Evaluation Metrics**:
  - **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions.
  - **Mean Squared Error (MSE)**: Penalizes larger errors more heavily, providing insight into model accuracy.
  - **R-squared**: Indicates how well the model explains the variance in the fare amounts.

- **Results**:
  - **MAE**: [Provide value] USD
  - **MSE**: [Provide value] USD
  - **R-squared**: [Provide value]

Model Performance Visualization

- **Actual vs. Predicted Fares**:
  - A scatter plot comparing actual fare amounts to those predicted by the model.
  - The plot includes a diagonal reference line to indicate perfect predictions.

- **Residual Analysis**:
  - A histogram or scatter plot showing the distribution of residuals (the difference between actual and predicted fares).
  - Helps identify any patterns in prediction errors, such as systematic over- or under-prediction.

Key Findings

- **Important Factors**:
  - Distance is the most significant factor influencing fare amount, as expected.
  - Time of day also plays a role, with fares tending to be higher during peak hours.

- **Model Insights**:
  - The linear regression model performed well, explaining a significant portion of the variance in fare amounts.
  - Simple models like linear regression can still be highly effective for certain predictive tasks, especially with well-engineered features.

Conclusion and Future Work

- **Conclusion**:
  - The project successfully built a model to predict Uber ride fares using historical data and regression analysis.
  - The model can be useful for fare estimation, pricing strategy optimization, and improving transparency for riders.

- **Future Work**:
  - Incorporate additional features like traffic conditions, weather, and ride type (e.g., UberX, UberXL) for improved predictions.
  - Experiment with more complex models like Gradient Boosting Machines or Neural Networks to capture non-linear relationships.
  - Deploy the model in a real-time environment for dynamic fare predictions.

References

Data:

- Mention the synthetic data generation process and any real-world datasets you may have referenced.

Tools and Libraries:

- Python (pandas, NumPy, Scikit-learn, Matplotlib, Seaborn)
- Any other tools or libraries used in the project.

Custom image by ChatGPT

"Thank you!"