UBER FARE PREDICTION

USING REGRESSION ANALYSIS

Presented by : Siddhi Jain

AGENDA

1. Introduction
2. Data Preprocessing and Exploration
3. Exploratory Data Analysis (EDA)
4. Regression Modeling
5. Model Evaluation and Insights
6. Recommendations

INTRODUCTION

Project Objective

The goal of this project is to predict Uber fare prices for future rides. Accurate fare prediction is crucial for enhancing customer satisfaction and operational efficiency.

Dataset Description

The dataset contains 200,000 rows and 9 columns. Key fields include:

- **Key**: A unique identifier for each trip.
- **Fare Amount**: The cost of each trip in USD.
- **Pickup Datetime**: Date and time when the meter was engaged.
- **Passenger Count**: Number of passengers in the vehicle.
- **Pickup Longitude/Latitude**: Geographic coordinates where the trip started.
- **Dropoff Longitude/Latitude**: Geographic coordinates where the trip ended.

DATA PREPROCESSING

Libraries Used

Libraries such as NumPy, Pandas, Matplotlib, and Seaborn were employed for data processing and visualization.

Data Loading

The dataset was loaded into a Pandas DataFrame (df) from the uber.csv file.

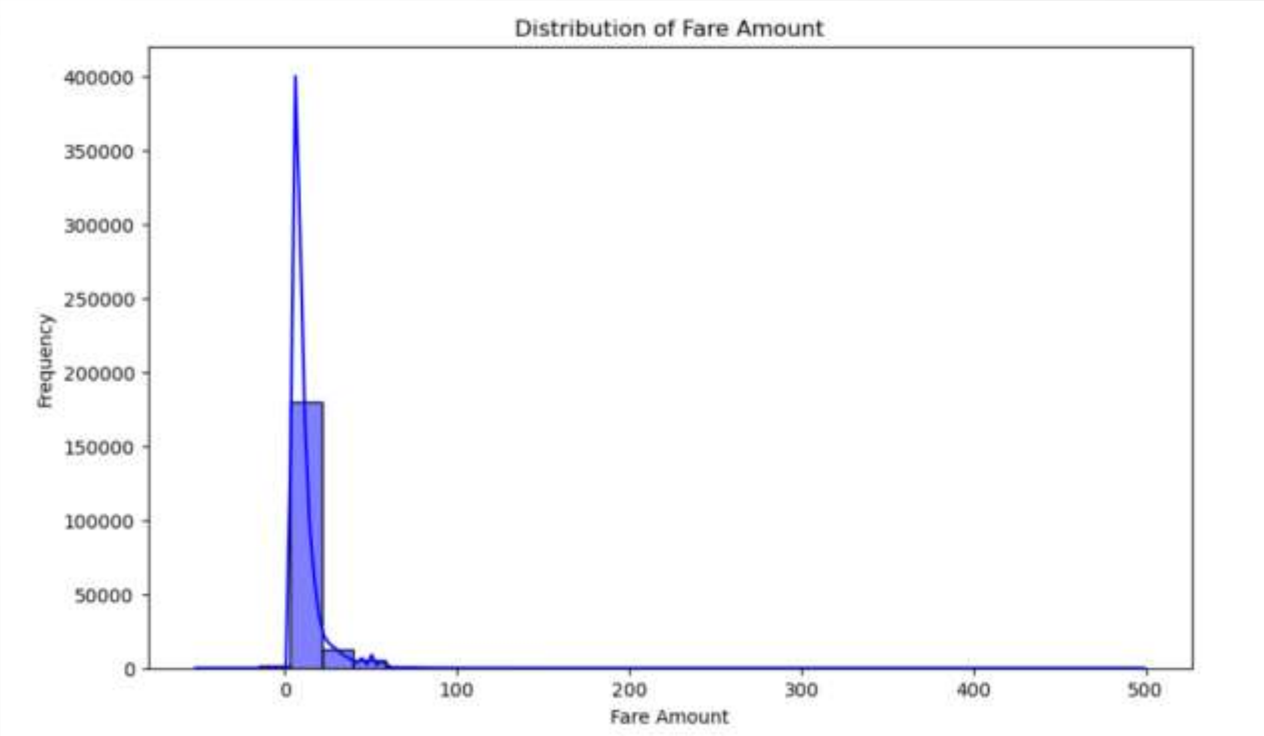
Steps Taken:

- **Initial Exploration**: Checked column names, data types, and dataset shape.
- **Handling Missing Data**: Removed rows with null values to ensure data integrity.
- **Feature Engineering**: Added a distance column using the Haversine formula to calculate the distance between pickup and dropoff points.

EXPLORATORY DATA ANALYSIS

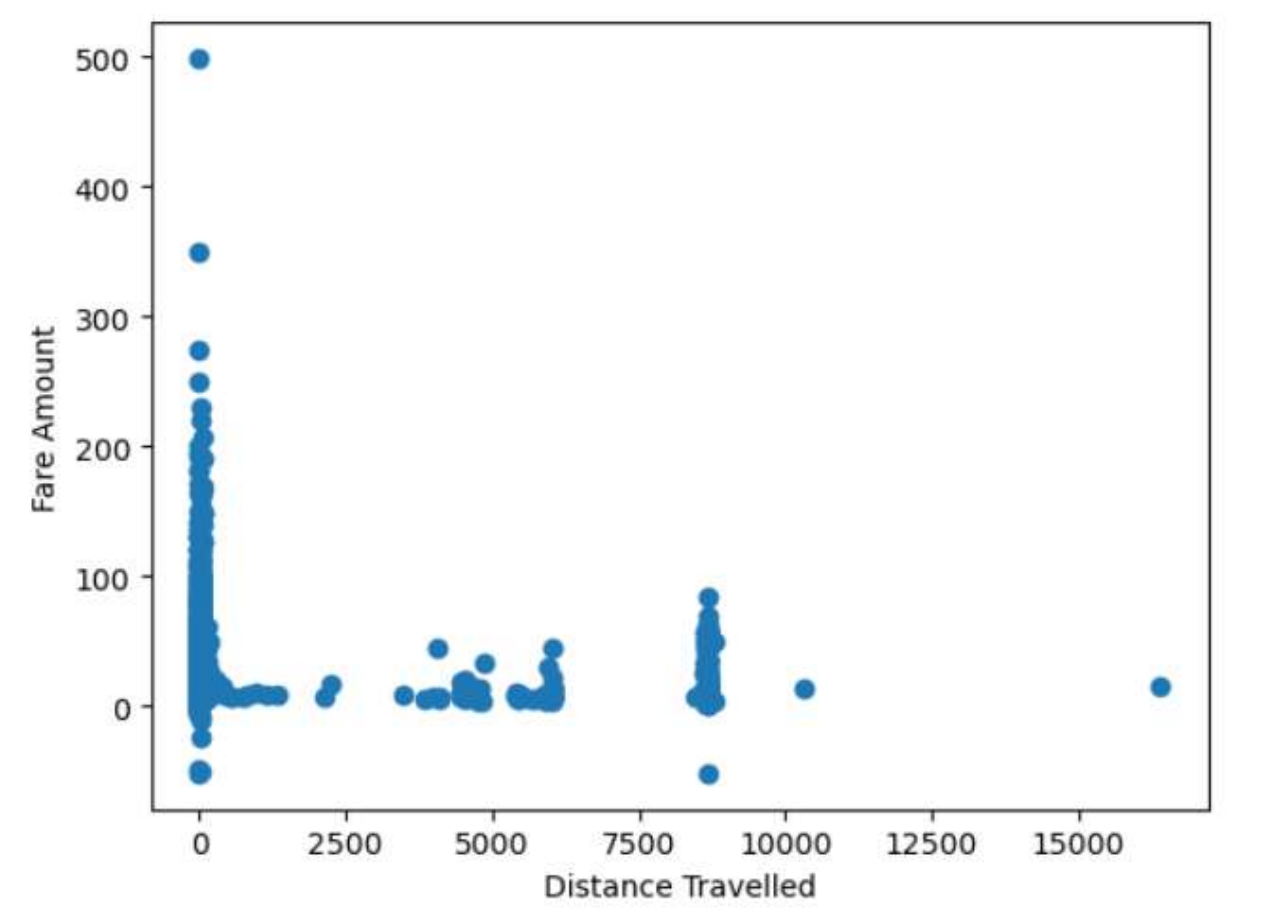
DISTRIBUTION OF FARE AMOUNT

- The histogram shows that the majority of fare amounts are concentrated between 0 and 20, indicating that most trips are short-distance or low-cost.



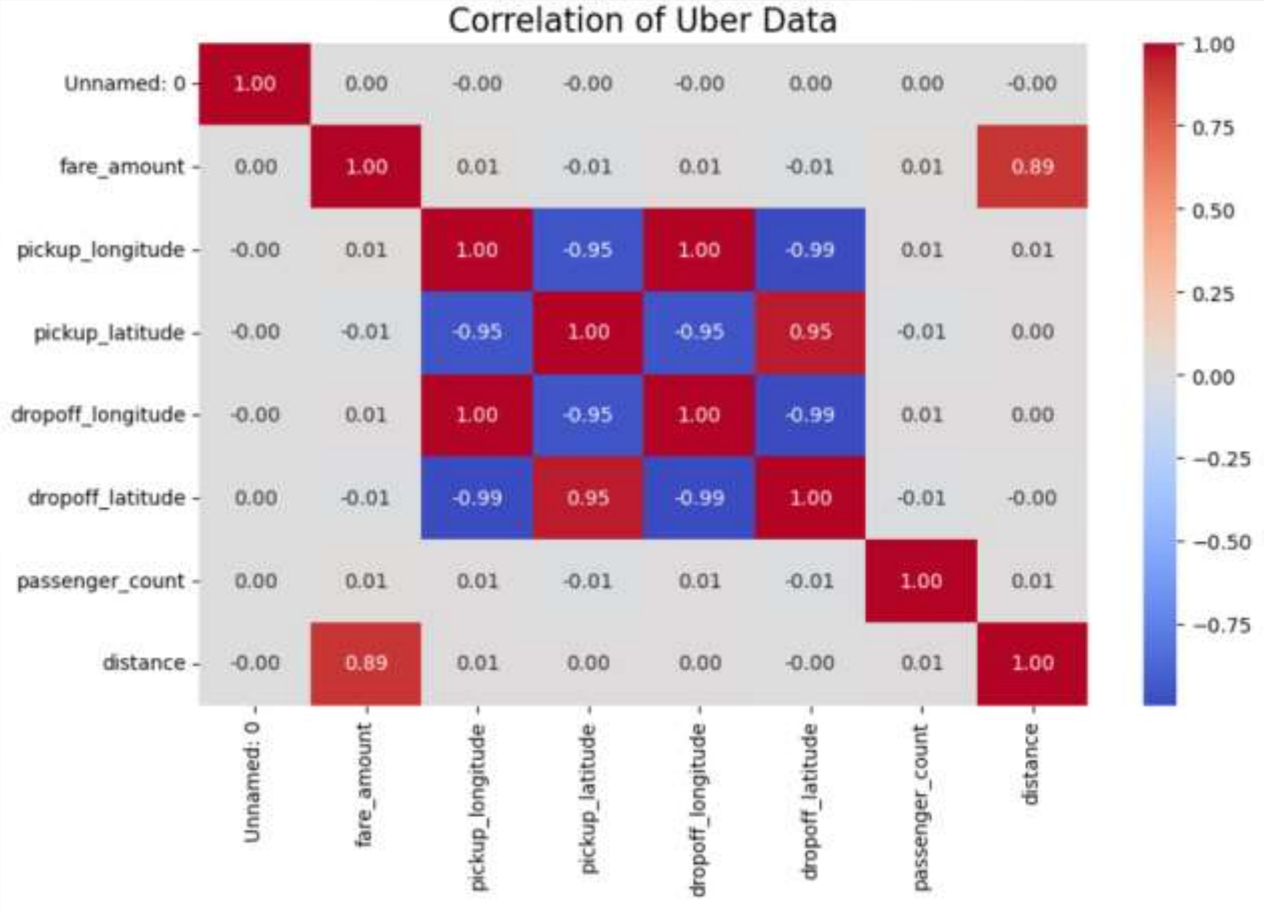
LINE PLOT: FARE AMOUNT BY DISTANCE

- We can get rid of the trips with very large distance that are outliers as well as trips with 0 distance.
- Removed row with non-plausible fare_amount and distance travelled.
- Final dataset contains 193,481 rows and 10 columns.



CORRELATION MATRIX

- Strong Positive Correlation Between **fare_amount** and **distance** (0.89).
- This indicates that as the distance increases, the fare amount also increases, which is logical for ride fare prediction.



REGRESSION MODEL

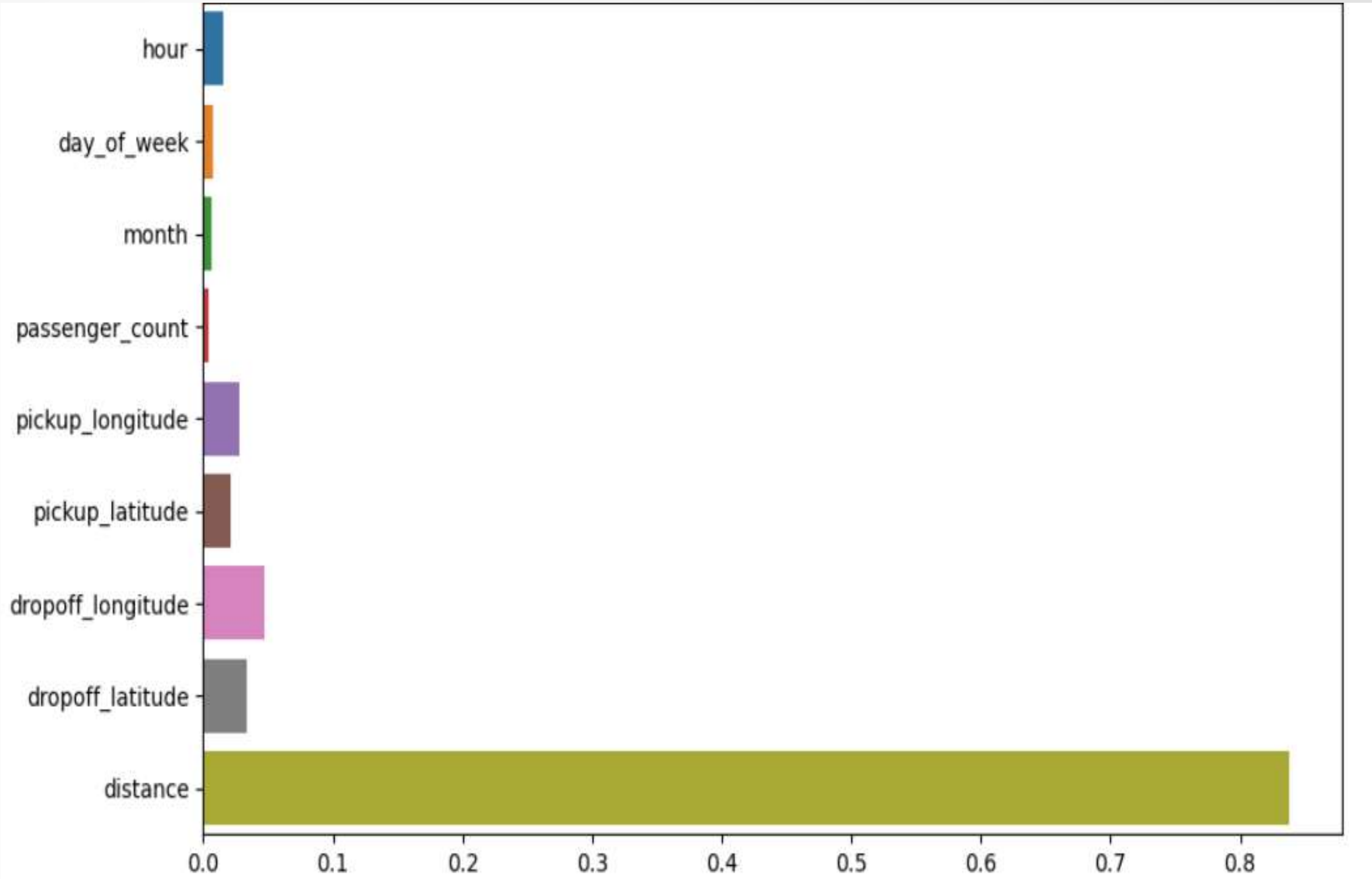
REGRESSION MODEL DEVELOPMENT

Feature Selection:

- The derived distance feature showed a strong correlation with fare_amount, making it a key predictor.

Model selection:

- **Random Forest Regression**: Selected for its ability to handle both linear and non-linear relationships, making it well-suited for this dataset.



Model Evaluation Metrics

Mean Absolute Error (MAE): 2.32 (indicates low average error in predictions).

Mean Squared Error (MSE) : 18.50 (shows minimal large prediction errors).

R^2 Score: 0.794 (explains 79.4% of variance in fare_amount).

Recommendations

Pricing Strategies

- **Dynamic Pricing** : Implement a tiered system where longer rides offer discounted per-mile rates to attract more customers.
- **Short-Trip Discounts :** Offer loyalty points or discounts for frequent short-distance riders to encourage repeat usage.

Operational Improvements

- **Surge Pricing** : Introduce higher driver payouts and surge pricing during peak hours to ensure availability and meet demand.
- **Outlier Detection** : Deploy automated systems to flag and review outliers in real-time, improving fare accuracy.

Customer Experience Enhancements

- **Fare Transparency** : Display predicted fare ranges before trip confirmation to build customer trust.
- **Shared Ride Incentives** : Offer pricing incentives for shared rides, benefiting customers and maximizing driver earnings.

Conclusion

- The project demonstrates high accuracy in predicting Uber fares, highlighting its potential to support real-world ride-hailing platforms.

- Practical recommendations are provided to enhance both customer satisfaction and operational efficiency.

- Future work could integrate additional contextual features, such as weather and traffic data, to make the model even more robust and adaptable across diverse scenarios.

THANKYOU