PREDICTING TAXI FARE AMOUNTS USING MACHINE LEARNING

SUBTITLE: ANALYZING KEY FACTORS THAT INFLUENCE TAXI FARES IN NYC

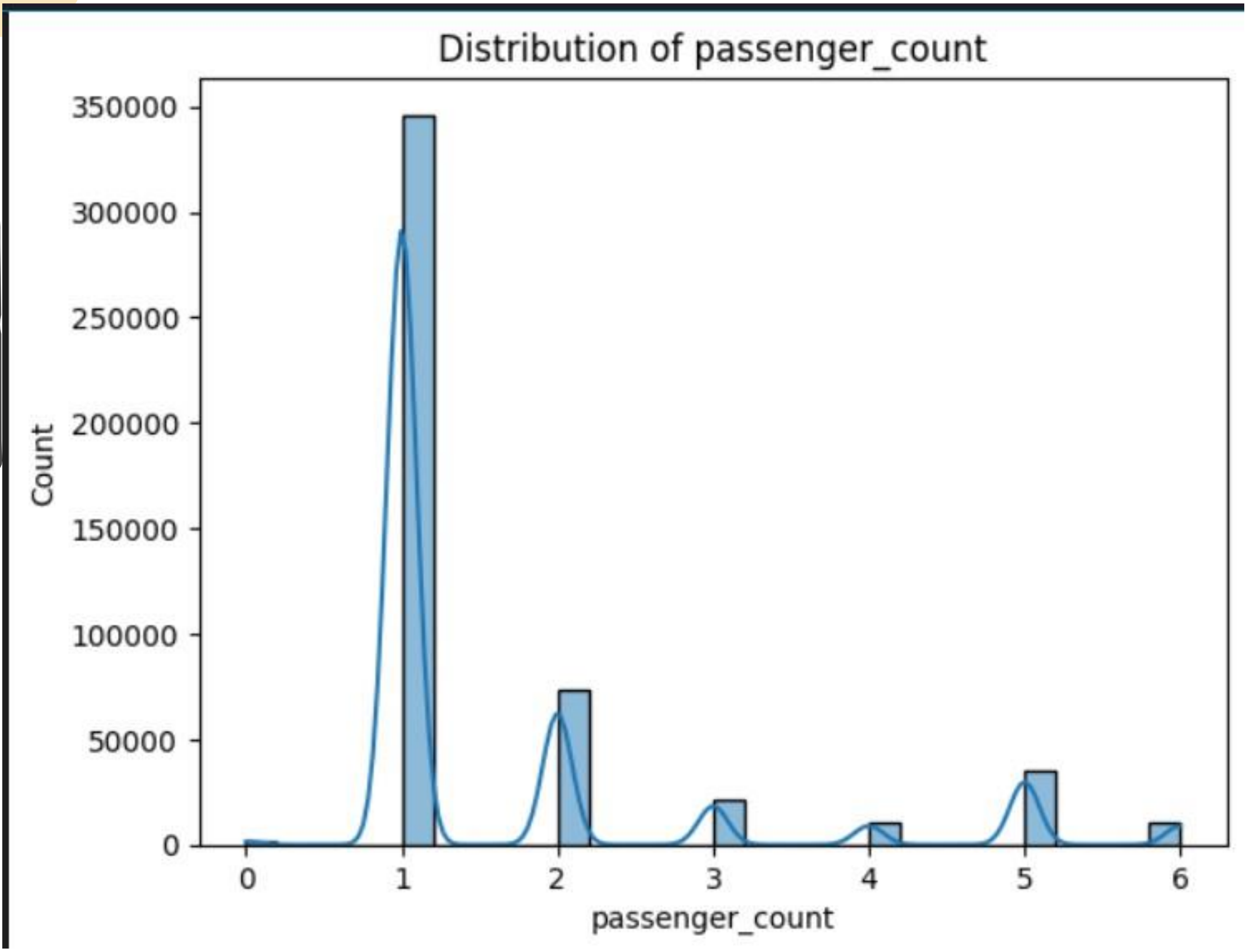
INTRODUCTION

Project Goal:
Predict taxi fares based on trip characteristics such as distance, time, and traffic conditions.

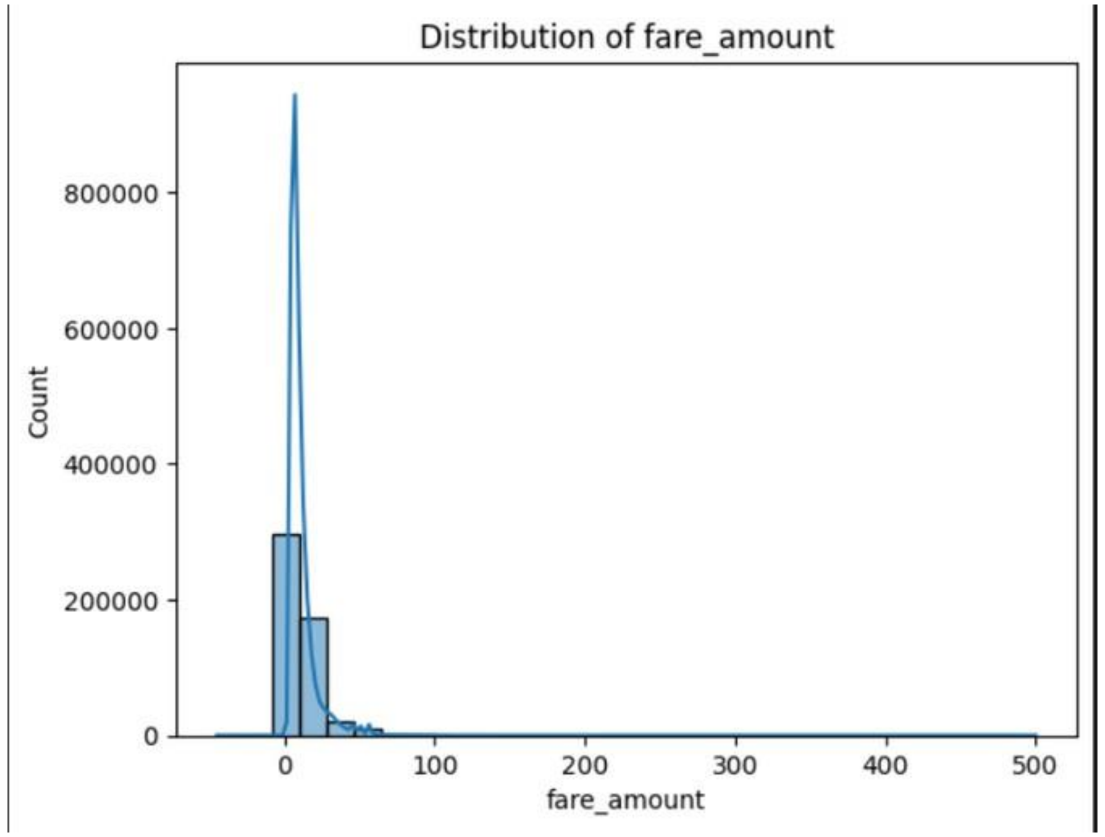
Why This Matters:
Understanding fare patterns can help improve pricing strategies and support drivers in optimizing their income.

Data Source:
The analysis is based on a large dataset of real-world taxi trips, including various trip-related features.

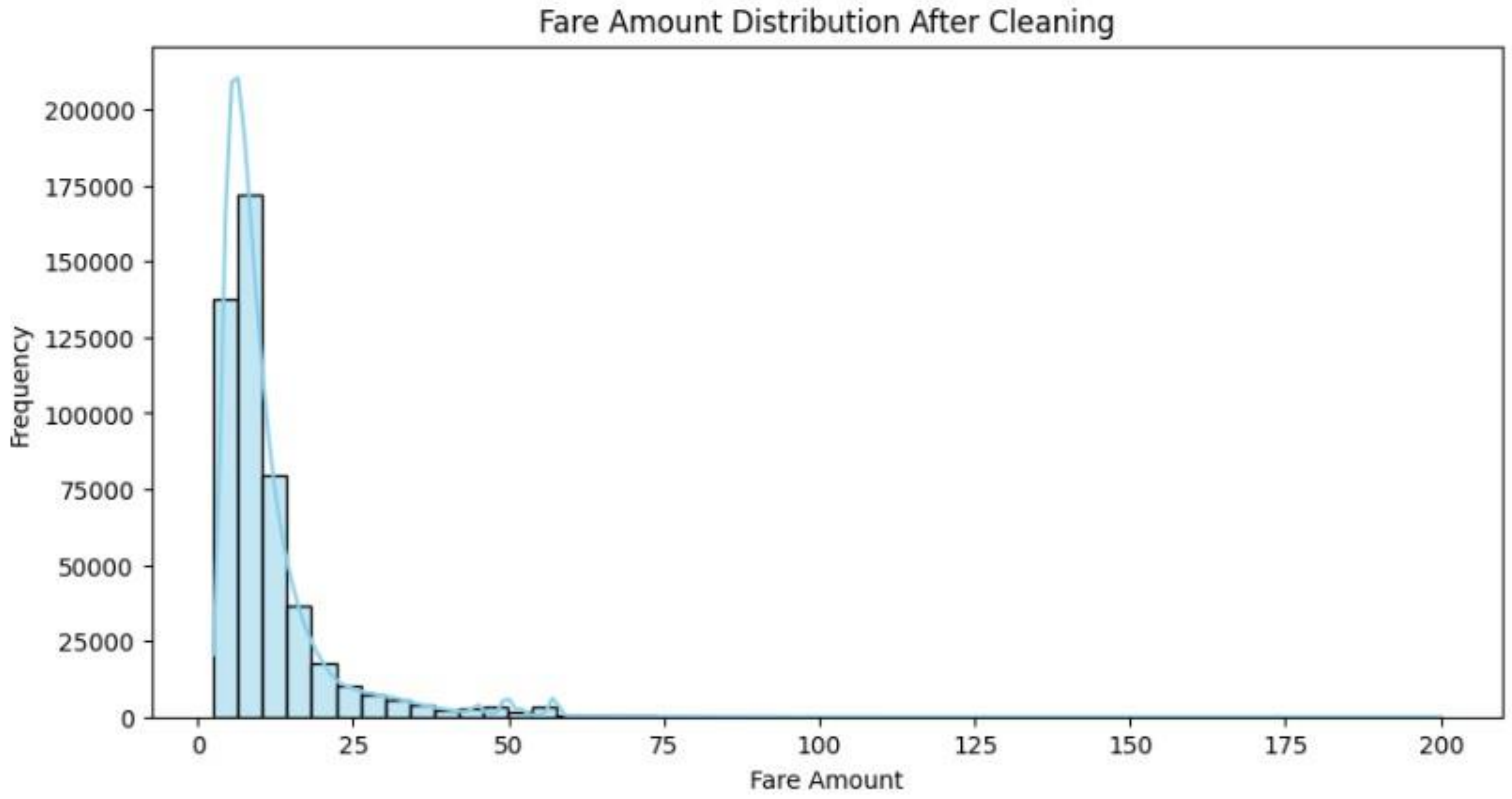
NUMERIC FEATURES DISTRIBUTION



We selected all numeric columns and plotted their distributions using histograms with KDE (Kernel Density Estimate). This helps us understand the spread and skewness of each feature.



FARE AMOUNT DISTRIBUTION AFTER CLEANING



After Cleaning

We plotted the distribution of the fare_amount column after removing outliers and handling missing values. This helps us verify that the data is more realistic and well-scaled.

FEATURE AND TARGET SELECTION

We selected the most relevant features that could influence the taxi fare

Selected Features:

distance: Distance between pickup and dropoff

hour: Hour of the day when the ride was requested

weekday: Day of the week

Weather: Weather condition during the ride

Car Condition: Quality of the car

Traffic Condition: Level of traffic.

Our prediction target is the fare_amount

MULTICOLLINEARITY CHECK USING VIF

```
VIF Table:
              Feature           VIF
14           sol_dist  1.977205e+07
15           nyc_dist  1.065316e+07
12           ewr_dist  3.596956e+06
11           jfk_dist  4.394941e+05
13           lga_dist  4.261383e+05
3   dropoff_longitude  9.470342e+00
4    dropoff_latitude  8.527962e+00
1    pickup_longitude  5.990933e+00
2     pickup_latitude  2.619573e+00
0         fare_amount  1.532729e+00
10               year  1.034852e+00
16           distance  1.028166e+00
8               month  1.016316e+00
9             weekday  1.011797e+00
6                hour  1.010956e+00
17            bearing  1.010591e+00
5     passenger_count  1.002111e+00
7                 day  1.000592e+00
```

We used Variance Inflation Factor (VIF) to check for multicollinearity among numerical features.
Features were standardized using StandardScaler before calculating VIF.

VIF values were computed for each feature.
Features with very high VIF (typically above 10) were removed to reduce redundancy.
This improves the model's performance and stability by avoiding duplicated information.

This step is important to ensure that the model does not rely on indirectly repeated features, which makes it more stable

MODEL BUILDING USING PIPELINE

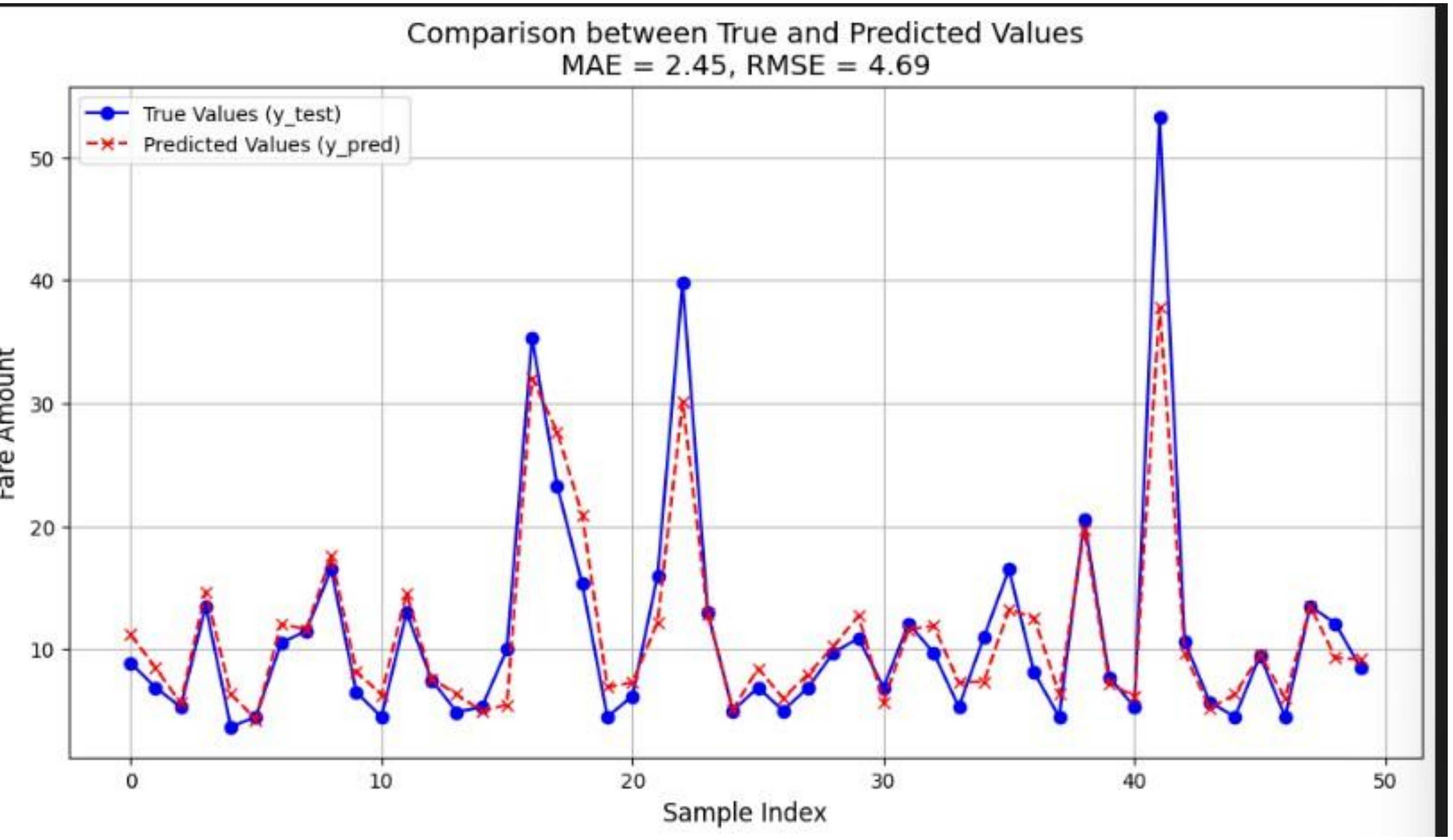
Built a Pipeline to streamline preprocessing and model training in one step.

Used:

preprocessor: for scaling and encoding data.

RandomForestRegressor: as the prediction model.

MODEL PREDICTION VS ACTUAL



Comparison between True & Predicted Fare Amounts

MAE = 2.15 RMSE = 3.20

Blue Line: True Values

Red Dashed Line: Predicted Values

The model performs well with slight deviations.

MODEL EVALUATION RESULTS

Random Forest:

RMSE: 3.15

R^2: 0.82

XGBoost:

RMSE: 3.04

R^2: 0.84

The R^2 score of 0.82 means that 82% of the variance in the target variable is explained by the Random Forest model.

Higher R^2 indicates a better fit between the predicted and actual values.

THANK YOU