TaxiFare Forecast: Enhancing Fare Accuracy with Machine Learning

Rajib Roy (rajibroy@iisc.ac.in)
Aravind S (aravindss@iisc.ac.in)
Srividhya L (srividhyal@iisc.ac.in)
Kavipriya Ramasamy (kavipriyar@iisc.ac.in)

Abstract

Ride-hailing services use dynamic pricing strategies influenced by several factors such as trip distance, time, and traffic conditions. This project aims to build a machine learning model to predict taxi fares accurately and in real time. By applying regression models like Linear Regression, Decision Trees, and XGBoost, we analyze feature importance, optimize predictions, and integrate the model for deployment. Our final model demonstrates strong predictive accuracy, reducing customer frustration and enhancing operational efficiency.

1 Problem Definition

1.1 Background

Dynamic pricing strategies often result in unpredictable fares, frustrating customers and challenging service providers. Accurate fare prediction benefits both customers (through transparency) and businesses (via optimized pricing).

1.2 Objectives

- Build a machine learning model to predict taxi fares.
- Identify key features influencing fare predictions.
- Ensure real-time predictions for seamless customer experience.
- Improve accuracy and speed to integrate with pricing systems.

2 Data Collection and Preparation

2.1 Dataset Description

- **Source:** Public datasets like HuggingFace. (Dataset).
- **Size:** Approximately 500,000 records.
- **Format:** CSV.
- **Features:** 20+ features including Pickup/Dropoff coordinates, timestamps, passenger count, derived features like trip distance.
- **Target Variable:** *fare amount*.

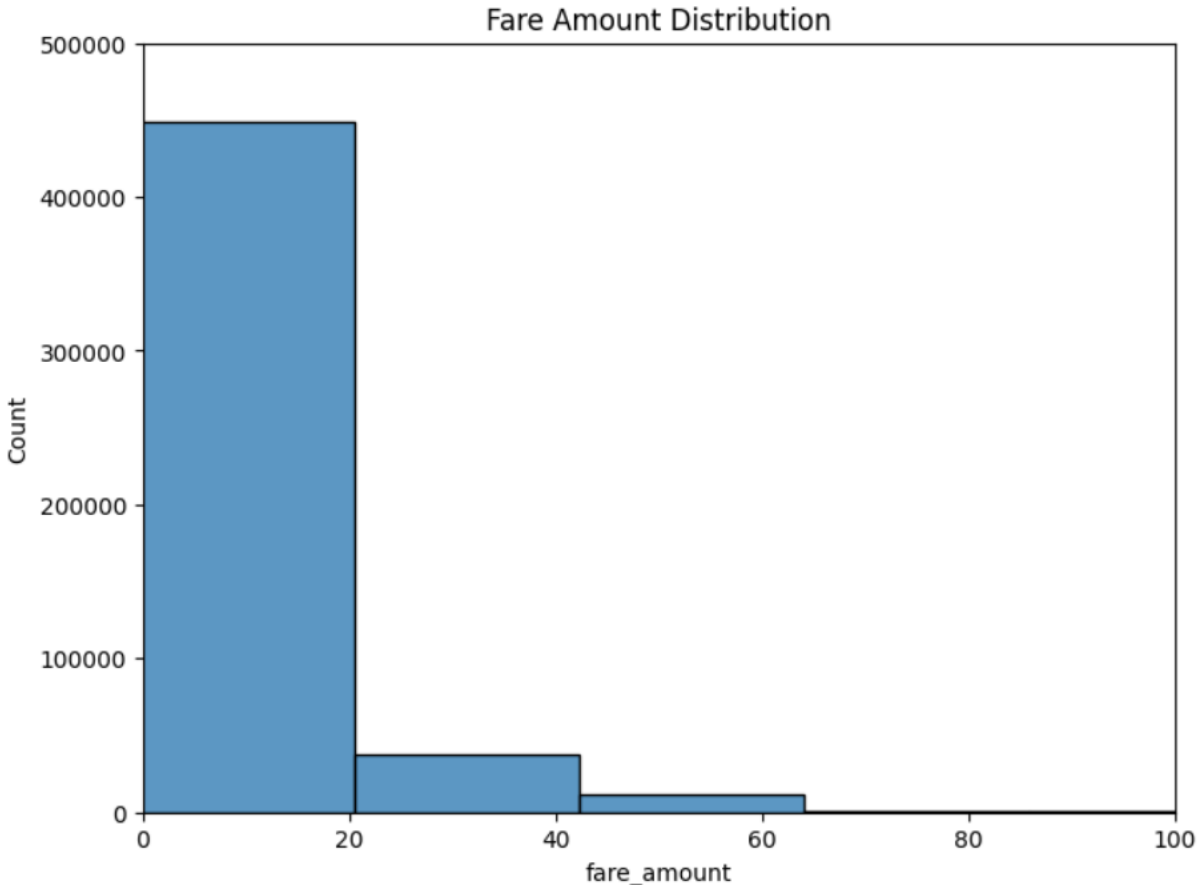
2.2 Preprocessing Steps

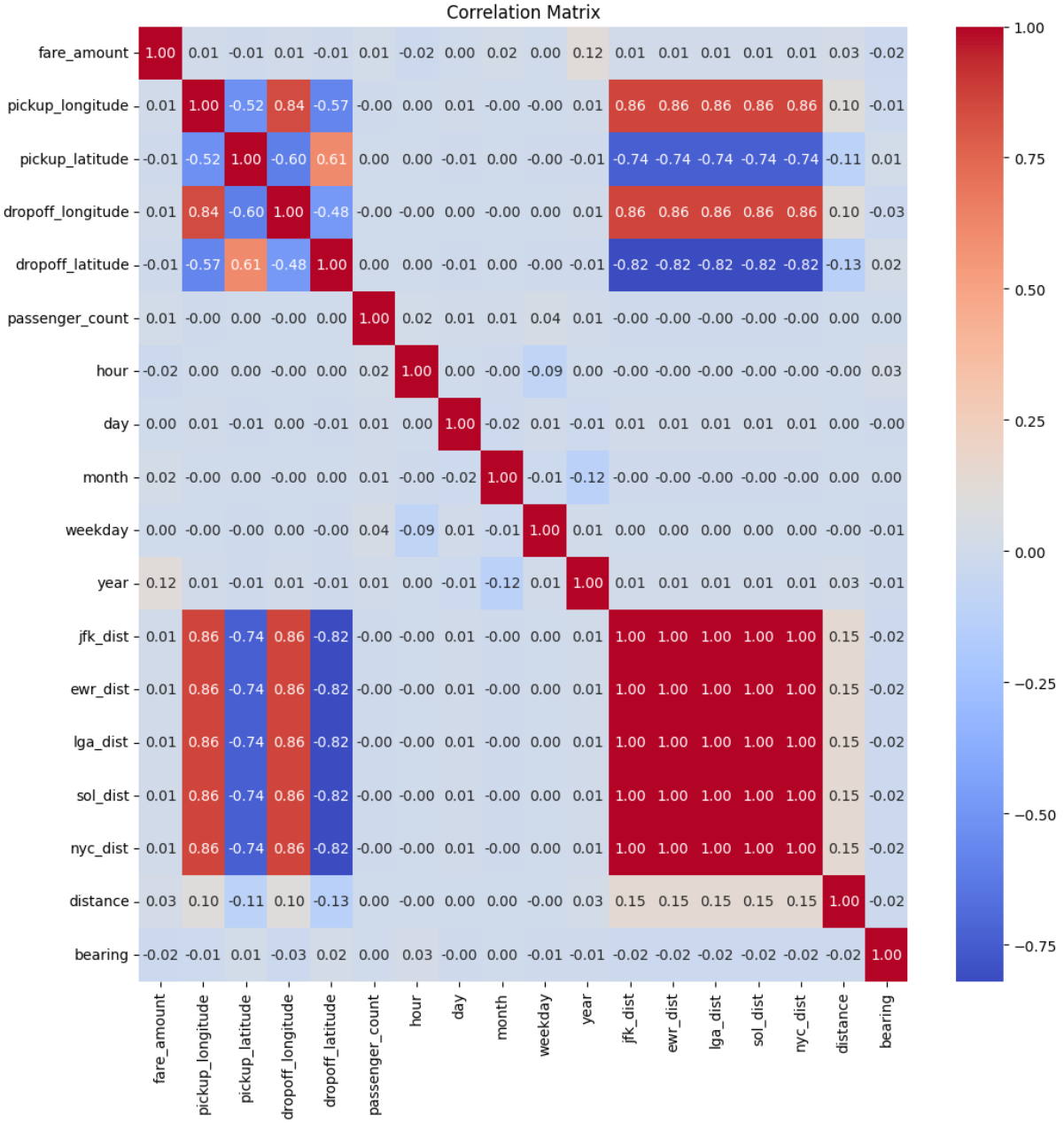
- Handled missing values and irrelevant columns.
- Engineered features (e.g., trip distance, one-hot encoding for categorical variables).

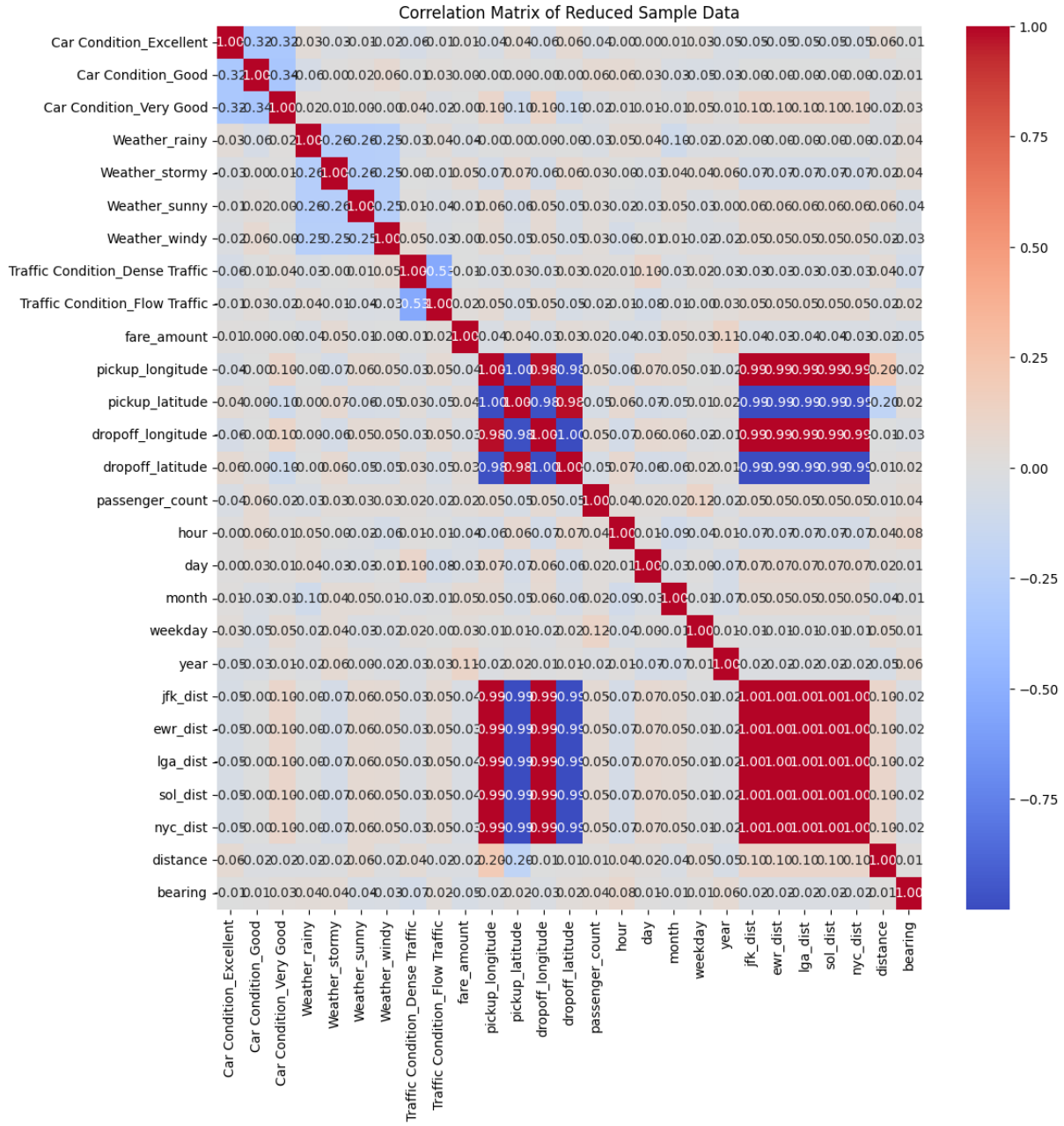
3 Exploratory Data Analysis (EDA)

- Analyzed numerical and categorical distributions.
- Visualized correlations between features and the target variable.

- Assessed the impact of distance, time of day, and traffic conditions on fares.




Correlation Matrix
fare_amount 1.00 0.01 -0.01 0.01 -0.01 0.01 -0.02 0.00 0.02 0.00 0.12 0.01 0.01 0.01 0.01 0.01 0.03 -0.02
pickup_longitude 0.01 1.00 -0.52 0.84 -0.57 -0.00 0.00 0.01 -0.00 -0.00 0.01 0.86 0.86 0.86 0.86 0.86 0.10 -0.01
pickup_latitude -0.01 -0.52 1.00 -0.60 0.61 0.00 0.00 -0.01 0.00 -0.00 -0.01 -0.74 -0.74 -0.74 -0.74 -0.74 -0.11 0.01
dropoff_longitude 0.01 0.84 -0.60 1.00 -0.48 -0.00 -0.00 0.00 -0.00 0.00 0.01 0.86 0.86 0.86 0.86 0.86 0.10 -0.03
dropoff_latitude -0.01 -0.57 0.61 -0.48 1.00 0.00 0.00 -0.01 0.00 -0.00 -0.01 -0.82 -0.82 -0.82 -0.82 -0.82 -0.13 0.02
passenger_count 0.01 -0.00 0.00 -0.00 0.00 1.00 0.02 0.01 0.01 0.04 0.01 -0.00 -0.00 -0.00 -0.00 -0.00 0.00 0.00
hour -0.02 0.00 0.00 -0.00 0.00 0.02 1.00 0.00 -0.00 -0.09 0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 0.03
day 0.00 0.01 -0.01 0.00 -0.01 0.01 0.00 1.00 -0.02 0.01 -0.01 0.01 0.01 0.01 0.01 0.01 0.00 -0.00
month 0.02 -0.00 0.00 -0.00 0.00 0.01 -0.00 -0.02 1.00 -0.01 -0.12 -0.00 -0.00 -0.00 -0.00 -0.00 0.00 0.00
weekday 0.00 -0.00 -0.00 0.00 -0.00 0.04 -0.09 0.01 -0.01 1.00 0.01 0.00 0.00 0.00 0.00 0.00 -0.00 -0.01
year 0.12 0.01 -0.01 0.01 -0.01 0.01 0.00 -0.01 -0.12 0.01 1.00 0.01 0.01 0.01 0.01 0.01 0.03 -0.01
jfk_dist 0.01 0.86 -0.74 0.86 -0.82 -0.00 -0.00 0.01 -0.00 0.00 0.01 1.00 1.00 1.00 1.00 1.00 0.15 -0.02
ewr_dist 0.01 0.86 -0.74 0.86 -0.82 -0.00 -0.00 0.01 -0.00 0.00 0.01 1.00 1.00 1.00 1.00 1.00 0.15 -0.02
lga_dist 0.01 0.86 -0.74 0.86 -0.82 -0.00 -0.00 0.01 -0.00 0.00 0.01 1.00 1.00 1.00 1.00 1.00 0.15 -0.02
sol_dist 0.01 0.86 -0.74 0.86 -0.82 -0.00 -0.00 0.01 -0.00 0.00 0.01 1.00 1.00 1.00 1.00 1.00 0.15 -0.02
nyc_dist 0.01 0.86 -0.74 0.86 -0.82 -0.00 -0.00 0.01 -0.00 0.00 0.01 1.00 1.00 1.00 1.00 1.00 0.15 -0.02
distance 0.03 0.10 -0.11 0.10 -0.13 0.00 -0.00 0.00 0.00 -0.00 0.03 0.15 0.15 0.15 0.15 0.15 1.00 -0.02
bearing -0.02 -0.01 0.01 -0.03 0.02 0.00 0.03 -0.00 0.00 -0.01 -0.01 -0.02 -0.02 -0.02 -0.02 -0.02 -0.02 1.00
fare_amount
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
passenger_count
hour
day
month
weekday
year
jfk_dist
ewr_dist
lga_dist
sol_dist
nyc_dist
distance
bearing
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75



4 Proposed Methodology

4.1 Models Used

- Linear Regression.
- Decision Trees.
- Gradient Boosting (XGBoost).

4.2 Justification

- Regression models are well-suited for predicting continuous variables.

- Gradient Boosting captures complex feature interactions.
- Decision Trees and Linear Regression provide baseline comparisons.

4.3 Tools and Libraries

Python libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, joblib, json, Streamlit.

5 Results and Model Comparison

Model	Parameter	Training Set	Test Set
Linear Regression	**Mean Absolute Error (MAE)**	4.8943	4.9329
	Mean Squared Error (MSE)	64.1171	64.3020
	R^2 Score	0.3466	0.3512
Decision Tree	**Mean Absolute Error (MAE)**	0.0096	2.6084
	Mean Squared Error (MSE):	0.0003	37.1167
	R^2 Score	0.9999	0.6255
XGBoost	**Mean Absolute Error (MAE)**	1.6706	1.7997
	Mean Squared Error (MSE):	12.6767	19.2302
	R^2 Score	0.8708	0.8059

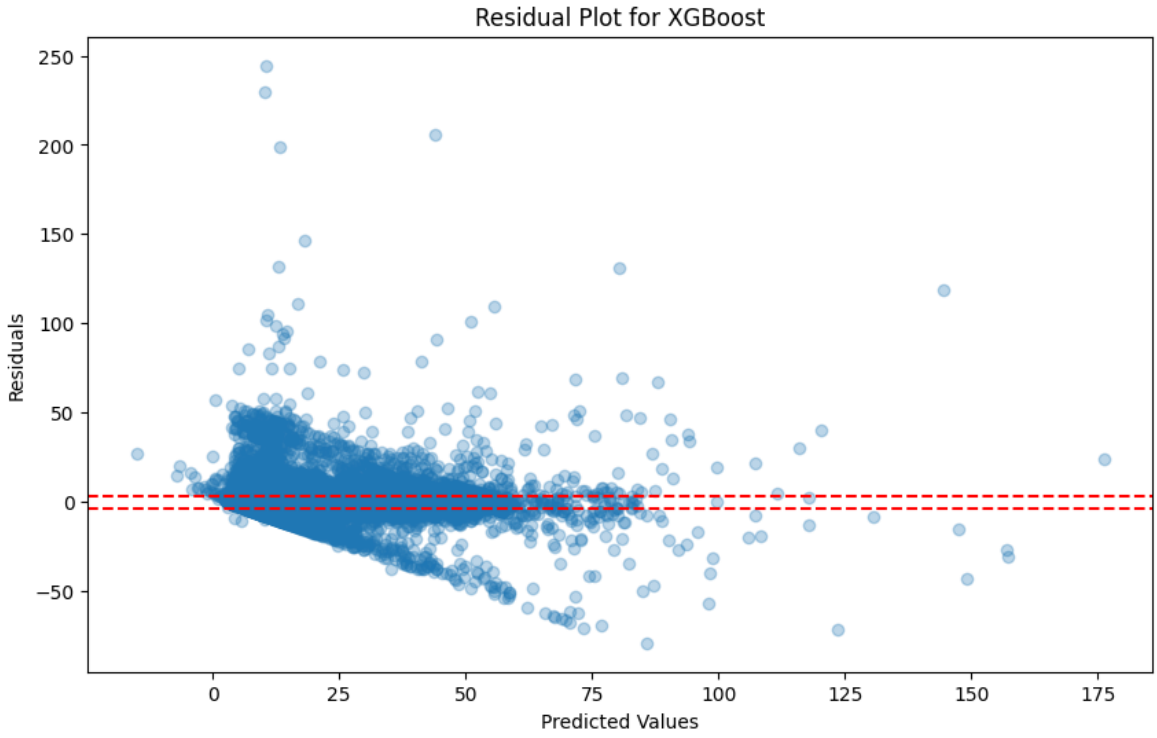
	Actual Fare	Predicted Fare
0	4.5	5.892794
1	16.9	16.286190
2	5.7	5.037998
3	7.7	9.294933
4	5.3	5.323117

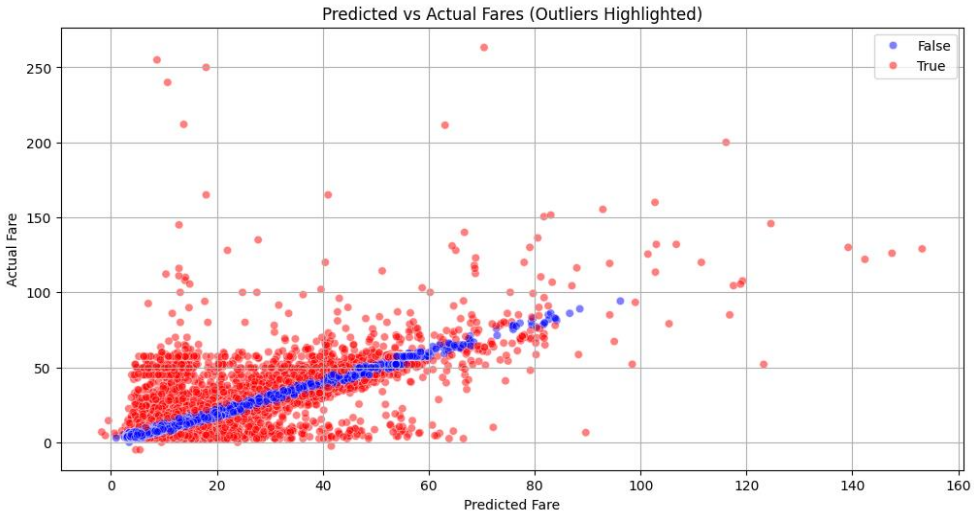
Observation: XGBoost outperformed all other models, explaining 81% of the variance in fare predictions.

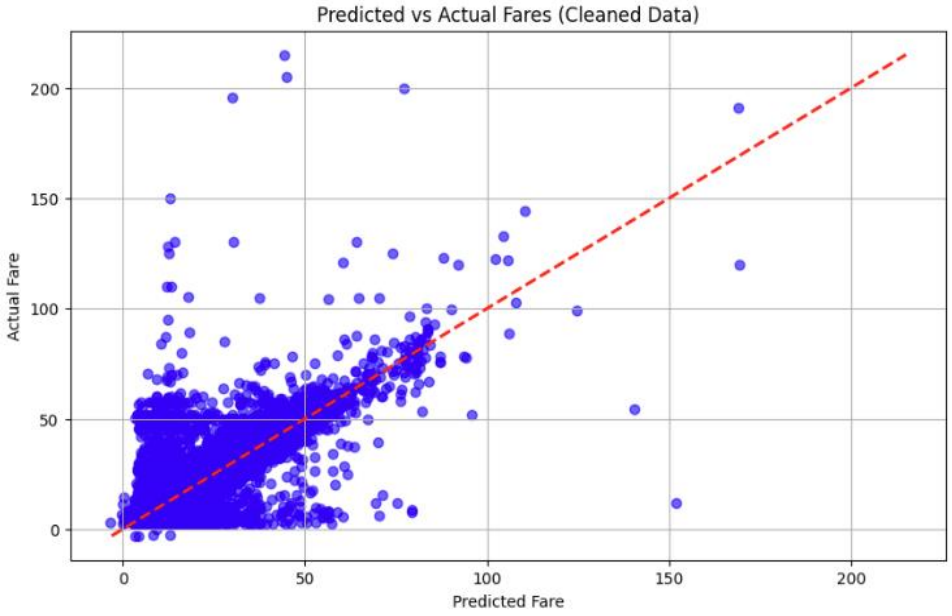
6 Model Fine-Tuning

6.1 Outlier Treatment

- Used the Interquartile Range (IQR) method to identify and remove extreme outliers.
- Retained relevant outliers to maintain dataset integrity.
- Impact: Improved generalization and reduced noise in the dataset.





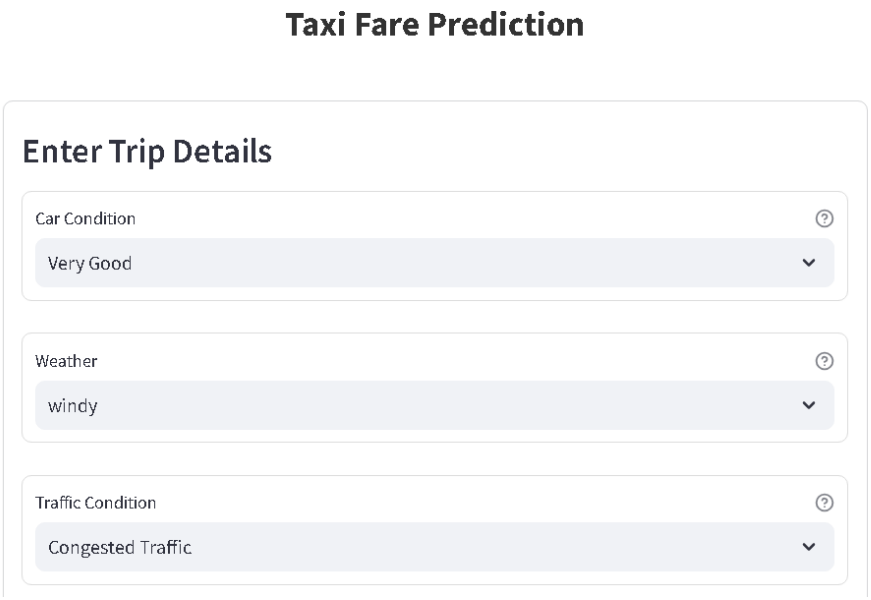


6.2 Hyperparameter Tuning

- Used Grid Search with cross-validation for hyperparameter optimization.
- Optimal parameters:
  - Learning rate: 0.1
  - Max depth: 7
  - Estimators: 200
- Results: MAE = 1.60, MSE = 13.95, R^2 = 0.84.

7 Deployment

- Exported the model using joblib.
- Built a web application using Streamlit.
- Deployed via Streamlit Community Cloud.
- URL: TaxiFare Forecast App



8 Contributions

- Aravind S S: Decision Trees, Streamlit app.
- Rajib Roy: XGBoost, Cross-validation, Deployment.
- Srividhya L: Feature Engineering, Linear Regression.
- Kavipriya Ramasamy: Data Collection, Data Preprocessing.

9 References

1. HuggingFace Datasets (ElvisGitau/Uber-Fare-Predict)
2. GitHub Repository: TaxiFare Forecast.
3. Scikit-learn Documentation.
4. Chen, T. and Guestrin, C., “XGBoost: A Scalable Tree Boosting System.”