

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.

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

- **Linear Regression:**

- MAE: 4.93
- MSE: 64.30
- R^2 : 0.35

- **Decision Tree:**

- MAE: 2.61
- MSE: 37.12
- R^2 : 0.63

- **XGBoost:**

- MAE: 1.80
- MSE: 19.23
- R^2 : 0.81

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.”