Case Study: NYC Taxi Fare Prediction – Automatidata Project

By Akash Raj

---

1. Project Overview

Automatidata aimed to build a reliable **NYC Taxi Fare Prediction System** that improves pricing transparency for passengers and provides deeper operational insights for the transportation network.

This project integrates:

- **Python** – Data cleaning, feature engineering, and ML model
- **Machine Learning (Random Forest Regression)** – Predicting taxi fares
- **Tableau –** Model Evaluation on Testing Data
- **Power BI** – Final BI dashboard, actionable insights, and decomposition
- **Business Recommendations** – To improve accuracy, operations, and revenue

---

2. Business Problem

NYC taxi fares vary widely due to inconsistent trip durations, route differences, anomalies in the dataset, and behavioural factors like tipping.
Automatidata needed:

1. A **predictive model** with minimal error
2. A robust **data-cleaning workflow**
3. Visual analytics to identify **business patterns**
4. A BI layer for **decision-making**

---

3. Tools Used

Python

- Data preparation
- EDA (Deep Analysis for Validation)
- Outlier treatment
- Hypothesis testing
- Feature engineering
- Model building
- Residual analysis

Tableau

Tableau was used to visually evaluate model performance on the testing dataset.
Instead of relying only on numerical metrics, Tableau helped verify whether the model predictions aligned with real-world fare patterns.

Key Tableau Contributions

- Actual vs Predicted scatterplots to validate how closely predictions follow the ideal diagonal trend.
- Fare distribution visuals to confirm the Random Forest model learned real fare behaviour, especially in common fare ranges.
- Distance vs Fare analysis to ensure the model captured the linear relationship between distance and price for normal trips.
- Detection of outlier clusters, where long-distance or high-fare trips deviated from predictions.
- Error pattern comparison to confirm that the model's largest errors matched the same patterns later seen in Power BI (20+ miles, high fares).

Purpose of Tableau in the pipeline

✓ Early visual validation
✓ Model sanity-check before deployment
✓ Understanding prediction stability
✓ Highlighting segments where the model underperforms

Tableau acted as the model evaluation layer, helping confirm that the model was behaving correctly before scaling the analysis into Power BI.

Power BI

- Final operational dashboards
- Decomposition Tree
- Absolute Error Analysis
- Interactive tooltips
- Payment type insights
- Time-of-day analysis
- Fare Band and Distance Band behaviour
- "Deep Dive" mode for granular trip-level exploration

Together, Tableau + Power BI ensured both **model evaluation** and **actionable business reporting**.

---

4. Model Development

Model: Random Forest Regressor

Results after anomaly cleaning:

- **R^2 : 0.974**
- **MAE : 0.43**
- **RMSE : 1.71**

Hypothesis Testing

- All predictors have **$p < 0.001$ → significant relationships**
- Confidence intervals confirm no coefficient crosses zero
- Fare prediction is strongly influenced by:
  - Trip distance
  - Duration

---

5. Key Insights

1️⃣ Long-distance trips (>20 miles) cause the highest prediction errors

- Absolute error spikes up to **27+**

2️⃣ Medium-distance trips (5–20 miles) are the most predictable

- Most stable and frequent trip segment
- Ideal for reliable model predictions

3️⃣ High-fare trips (80–100 and 100+ fare band) show huge variance

- Both Tableau and Power BI show unstable fare jumps

4️⃣ Duration strongly influences prediction

- Longer durations → exponential increase in prediction error
- Caused by data anomalies, traffic, route choices, and tolls

5️⃣ Credit Card users dominate trips (~15K)

- Strongest and most profitable customer segment

6️⃣ Time-of-day impacts prediction

- Morning (5–12) and Evening (5–9) show the highest unpredictability, linked to peak-hour traffic variation

---

6. Key Actions

1️⃣ **Promote Credit Card Payments**

- Only credit card users tip
- Better revenue for drivers + platform

2️⃣ **Implement QC anomaly detection**

- Flag unrealistic duration/distance
- Automatically block bad records from training future models

3️⃣ **Stabilize long-distance trip pricing**

- Introduce **flat fares** or **upfront pricing** for rides >20 miles

4️⃣ **Strengthen Peak Hour Operations**

- Add more drivers in the Morning & Evening
- Reduces ride delays + improves customer satisfaction

5️⃣ **Target Middle-Fare Customers**

- Fare band 20–60 is the most stable, predictable segment
- Ideal for loyalty campaigns & subscription models

---

7. Business Impact

✓ More transparent & reliable fare predictions
✓ Reduced anomaly impact through QC rules
✓ Insights to increase tipping rates & driver earnings
✓ Strong operational visibility with Power BI dashboards
✓ Model evaluation testing using Tableau

---

8. Final Deliverables

- Clean dataset
- Random Forest model
- Residual analysis
- Tableau Model evaluation dashboards
- Power BI full business dashboard
- Actionable insights report
- Case study documentation