NYC Taxi Fare Prediction — Actionable Insights Report

Automatidata • NYC Taxi Analytics Project

ISSUE / PROBLEM

❖ NYC taxi fares vary widely due to inconsistent distances, durations, payment types, and data anomalies. Predicting fares accurately is challenging—especially for **long-distance and high-fare trips**, where irregular patterns create large errors.

❖ The objective was to identify key fare drivers and build a model that reliably predicts fares under real-world conditions.

IMPACT

❖ Long-distance trips (>20 miles) drive the **highest prediction errors**, reducing model accuracy.

❖ High-fare segments (80–100 & 100+) show **high variance**, creating uncertainty in revenue forecasting.

❖ Drivers lose earnings from **non-tipping customers**, as only credit card users consistently provide tips.

❖ Anomalies (unrealistic fares, durations, or distances) distort the prediction model and business decisions.

❖ Pricing inconsistency affects customer trust and fairness.

RESPONSE

❖ Designed a business-focused dashboard to monitor performance, errors, and customer behavior.

❖ Built a Random Forest fare prediction model ($R^2 = 0.974$) after removing unrealistic anomalies.

❖ Conducted deep error analysis using Power BI (Distance Bands, Fare Bands, Duration, Time of Day, Payment Type).

❖ Validated that **distance + duration** are the strongest predictors across both ML and BI layers.

KEY INSIGHTS

❖ **Long-distance trips (20+ miles) create the highest prediction errors**
These rides exhibit unstable pricing patterns, which increase model uncertainty.

❖ **Short & medium-distance trips are highly predictable**
Fares between 0–20 miles maintain low error and stable fare behavior.

❖ **High-fare trips (80–100 & 100+) show significant variance**
Due to rare occurrences and inconsistent route conditions.

❖ **Duration strongly influences prediction accuracy**
Longer trip times directly correlate with increased fare errors.

❖ **Payment behavior matters — only Credit Card users provide tips**
Additionally, the largest transaction group (~15,000 trips) makes them valuable for business strategy.

❖ **Peak hours (Morning & Evening) show higher variability**
Traffic congestion and uncertain trip durations increase fare fluctuations.

❖ **Anomalies significantly reduce model accuracy**
Incorrect distances, durations, or fare-per-mile distort both ML and BI outputs.