NYC Taxi Fare Prediction— Model Selection Report

Automatidata • NYC Taxi Analytics Project

ISSUE / PROBLEM

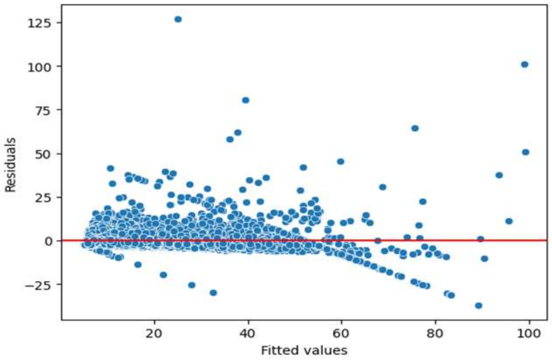
The objective was to identify a suitable baseline approach to understand fare drivers and validate whether fare increases are statistically meaningful before selecting a predictive model.

RESPONSE

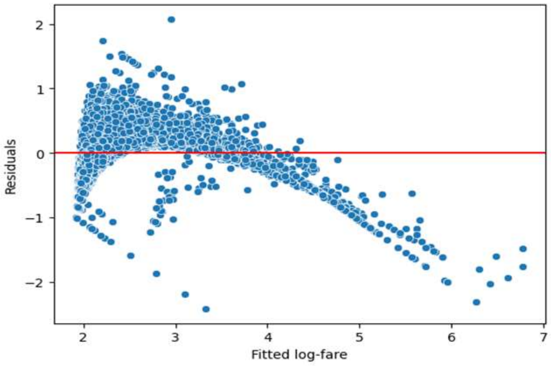
Ordinary Least Squares (OLS) regression was used **only as a diagnostic tool**, not as a production model. OLS helped evaluate linear relationships, statistical significance, and hypothesis validity between fare amount and key predictors.

IMPACT

OLS validated the pricing signal in the data and justified moving toward more flexible models capable of capturing non-linear fare behavior.

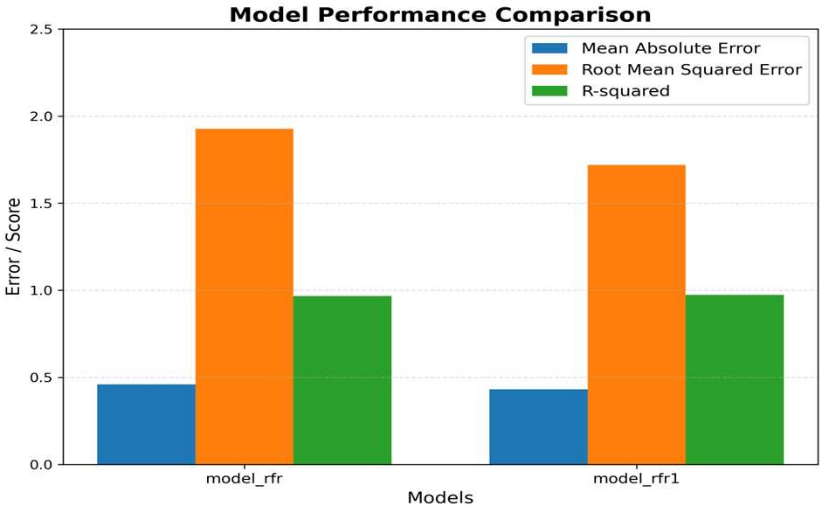


Residual plots from the OLS model show clear curvature and increasing variance, indicating violations of linearity assumptions. Applying a log transformation reduced variance, but non-linear patterns remained.



A model performance comparison shows that Random Forest significantly outperforms the linear baseline.

This confirms that fare pricing follows a non-linear structure, justifying the final model choice.



KEY INSIGHTS

- ❖ Trip distance and trip duration are statistically significant predictors of fare amount.
- ❖ All p-values were below 0.05, and confidence intervals did not include zero, confirming the relationships are not due to chance.
- ❖ Coefficients provide interpretable estimates of fare increase per mile and per minute.