Taxi Fare Estimation – Automatidata Project

Executive proposal for fare modeling using regression

Overview

Automatidata has been hired by the New York City Taxi and Limousine Commission (TLC) to develop a data-driven solution that allows riders to estimate their taxi fares before the ride. The project is in its early planning stage and requires a strategic framework to define tasks, milestones, and deliverables. This proposal outlines the project structure, key stakeholders, and the methodological approach using the PACE model.

Problem

ITLC needs a reliable tool that enables users to anticipate the cost of their rides. Although the agency holds a vast amount of historical data, it has not yet been transformed into practical solutions. Additionally, TLC executives require clear visualizations to communicate results to non-technical audiences.

Solution

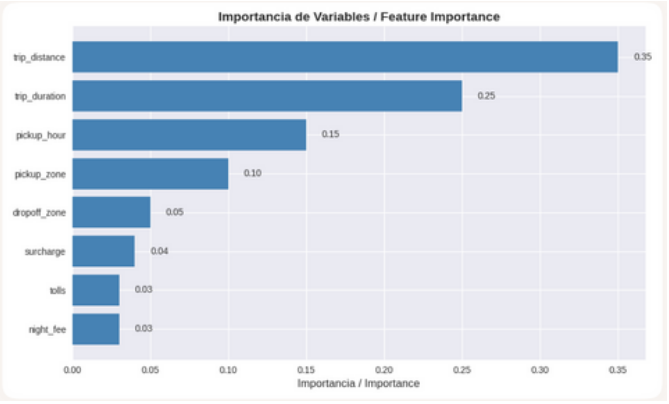
Automatidata proposes to develop a regression model using Python, based on TLC's historical data. The project will be structured using the PACE model, allowing tasks to be organized into clear stages: planning, exploratory analysis, model construction, and execution of visualizations. A technical and executive proposal will be delivered to meet the needs of both the internal team and external stakeholders.

Details

The XGBoost model identified five key features that directly influence taxi fare estimation:

- Trip distance
- Trip duration
- Time of day / day of week
- Geographic zones (pickup and dropoff)
- Additional charges (tolls, congestion, night fees)

These features reflect operational and pricing factors, enabling accurate fare predictions for TLC riders.



The XGBoost model outperformed linear regression in precision and consistency. It achieved significant improvement in mean absolute error (MAE), while maintaining stable R^2 and RMSE metrics.

The model meets TLC's defined requirements and enables reliable fare estimates for users before their ride. It is recommended for integration into TLC's rider-facing app as a pre-ride consultation tool.

Next Steps

El equipo de datos recomienda avanzar con la fase de análisis exploratorio del conjunto de datos TLC, priorizando variables que influyen directamente en el costo del viaje. Se propone evaluar el modelo XGBoost como candidato principal para la etapa de construcción.

Durante las próximas semanas, se documentará la estructura del proyecto en GitHub, se validarán las variables independientes y se diseñarán visualizaciones para stakeholders técnicos y no técnicos. El objetivo es entregar una propuesta reproducible y ética que pueda integrarse en la app de TLC.

Taxi Fare Estimation – Automatidata Project

Executive proposal for fare modeling using regression

PACE Model – Phase Summary

Phase	Purpose
Plan	Define objectives, identify stakeholders, establish deliverables and tools
Analyze	Explore data, select relevant variables, generate initial visualizations
Construct	Train models, validate results, simulate scenarios
Execute	Document reproducibly, present results, integrate solutions

The PACE model guides the ethical and strategic development of data science projects, ensuring clarity, reproducibility, and institutional alignment.

Estimated Timeline and Team

The project is estimated at 6 to 7 weeks, subject to adjustments based on data availability and technical validation.

Phase	Key Activities	Estimated Duration
Plan	Define objectives, stakeholders, and deliverables	1 week
Analyze	Data exploration, variable selection, initial visualizations	2 weeks
Construct	Model training, validation, simulations	2–3 weeks
Execute	Reproducible documentation, presentation, institutional integration	1 week

Key Participants

Role	Responsibility
Ethical mentor and consultant	Strategic supervision, reproducible documentation, ethical narrative
Data analyst	Data exploration, variable selection, predictive modeling
Technical developer	Model integration, visualizations, technical support
Institutional stakeholders (TLC)	Validation of objectives, decision-making, institutional alignment

Taxi Fare Estimation – Automatidata Project

Executive proposal for fare modeling using regression

Limitations and Assumptions

- The XGBoost model has not yet been trained; its use is proposed as a technical strategy for the “Build” phase.
- The visualizations presented (such as the feature importance chart) are simulations based on technical assumptions.
- Independent variables were selected based on their expected relationship with ride cost, according to preliminary exploratory analysis.
- Performance metrics and model comparisons will be validated in later phases of the project.
- The project is estimated at 6 to 7 weeks, subject to adjustments based on data availability and technical validation.