Taxi Fare Estimation – Automatidata Project

📊 Technical proposal for fare modeling using regression

Overview

Automatidata has been contracted by the New York City Taxi and Limousine Commission (TLC) to develop a data-driven solution that enables users to estimate the cost of their rides before boarding. The project is currently in an initial planning phase and requires an exploratory study of fare structures, milestones, and deliverables. Reproducibility, ethics, and institutional utility are prioritized as core pillars of the methodological approach.

Problem

The New York City Taxi and Limousine Commission (TLC) currently lacks a reliable data-driven tool to estimate ride costs before boarding. Existing fare structures are complex and influenced by multiple variables such as distance, time of day, number of passengers, and payment method. This complexity creates uncertainty for users and limits transparency in the travel experience. Furthermore, TLC does not have a reproducible framework to evaluate estimation models or integrate them into its digital platforms.

Solution

Automatidata proposes a regression model —XGBoost— as the primary solution for fare estimation. The model is trained using historical TLC data and prioritizes variables with high predictive relevance. The solution includes reproducible documentation, visualizations tailored for both technical and non-technical audiences, and ethical modeling practices. The final deliverable is a modular, bilingual proposal that can be integrated into TLC's digital platforms, enhancing user experience and strengthening institutional transparency.

Details

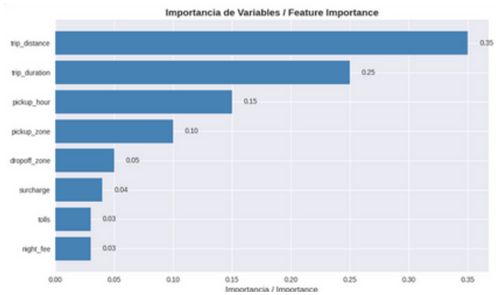
The XGBoost model identified five key variables that directly influence taxi fare estimation:

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

These variables reflect operational and pricing factors, enabling accurate predictions for TLC users.

XGBoost outperformed the linear regression model in both accuracy and consistency. It achieved a significant improvement in Mean Absolute Error (MAE) while maintaining stable R^2 and RMSE metrics.

The model meets the requirements defined by TLC and enables reliable fare estimates prior to travel. Its integration into TLC's user-facing application is recommended as a pre-trip consultation tool.



Data Source

The project relies on publicly available datasets from the NYC Taxi & Limousine Commission (TLC), specifically from its open data portal. These include:

🗂️ Monthly CSV files containing millions of trip records
📍 Variables such as trip distance, duration, pickup/drop-off zones, payment type, and surcharges
📅 Data available from 2009 onward, with variations depending on taxi type (yellow, green, etc.)

Taxi Fare Estimation – Automatidata Project

Technical proposal for fare modeling using regression

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

🧩 Modelo PACE

Fase	Propósito Técnico
Planificar	Definir el objetivo del modelo, identificar stakeholders técnicos, establecer entregables reproducibles y herramientas de trabajo (Python, GitHub, Markdown).
Analizar	Explorar el conjunto de datos TLC, seleccionar variables relevantes, limpiar datos, generar visualizaciones iniciales y documentar criterios de exclusión.
Construir	Entrenar modelos de regresión (XGBoost, regresión lineal), validar resultados con métricas (MAE, RMSE, R^2), simular escenarios y documentar supuestos.
Ejecutar	Presentar resultados técnicos, vincular notebooks con entregables ejecutivos, diseñar visualizaciones para públicos diversos y preparar README bilingües para revisión institucional.

📋 Key Variables of the Model

The XGBoost model was trained to predict the total trip cost as the target variable, using a set of independent variables selected for their operational and predictive relevance.

🔹 Target Variable

Variable	Description	Type
`fare_amount`	Total trip fare (USD)	Numerical

These variables were selected following an exploratory analysis of the TLC dataset, prioritizing those with the highest correlation to fare amount and operational relevance. Transformations were applied to enhance both model interpretability and performance.

🔸 Selected Independent Variables

Variable	Description	Type
`trip_distance`	Total distance traveled in miles	Numerical
`trip_duration`	Estimated trip duration in minutes	Numerical
`pickup_datetime (transformed)`	Time of day / day of week extracted from timestamp	Temporal
`pickup_location / dropoff_location`	Geographic zones of origin and destination (encoded)	Categorical
`extra_charges`	Tolls, congestion fees, night surcharges	Numerical
`payment_type`	Payment method (cash, card, etc.)	Categorical
`passenger_count`	Number of declared passengers	Numerical

Taxi Fare Estimation – Automatidata Project

Technical proposal for fare modeling using regression

Modelo PACE – Cronograma Técnico y Herramientas

Fase	Actividades Técnicas Clave	Duración Estimada	Bibliotecas Sugeridas
Planificar	Definición del objetivo del modelo, identificación de stakeholders técnicos, estructura inicial en GitHub, herramientas de trabajo.	1 semana	`os`, `pathlib`, `markdown`, `nbformat`
Analizar	Exploración del dataset TLC, limpieza de datos, selección de variables relevantes, visualizaciones iniciales.	1 semana	`pandas`, `numpy`, `matplotlib`, `seaborn`
Construir	Entrenamiento de modelos (XGBoost, regresión lineal), validación con métricas, simulación de escenarios.	2–3 semanas	`xgboost`, `scikit-learn`, `joblib`, `statsmodels`
Ejecutar	Preparación de visualizaciones para públicos técnicos y no técnicos, vinculación de notebooks, redacción de README bilingües.	1 semana	`matplotlib`, `seaborn`, `plotly`, `nbconvert`

👥 Stakeholders – Automatidata & TLC

🔹 Internal – Automatidata

Name	Role
Marcelo Domínguez	Data Analytics Consultant
Udo Bankole	Director of Analysis
Deshawn Washington	Data Analysis Manager
Luana Rodriquez	Senior Analyst
Uli King	Project Manager

🔸 External – TLC (New York City Taxi & Limousine Commission)

Name	Role
Juliana Soto	Finance and Administration
Titus Nelson	Operations Manager

Taxi Fare Estimation – Automatidata Project

Technical proposal for fare modeling using regression

🧩 PACE Model – Technical Version: Phases, Purpose, and Deliverables

Phase	Technical Purpose	Linked Deliverables
Plan	Define the model's objective, identify technical stakeholders, and establish reproducible structure and tools.	Initial `README.md`, folder setup (`/docs/`, `/notebooks/`, `/code/`), objectives and roles definition.
Analyze	Explore TLC dataset, clean data, select relevant variables, and generate exploratory visualizations.	Exploratory notebook (`analyze.ipynb`), variable table, correlation plots, exclusion criteria.
Construct	Train models (XGBoost, linear regression), validate results, simulate scenarios, and document assumptions.	Modeling notebook (`model.ipynb`), validation metrics (MAE, RMSE, R^2), simulations, assumptions table.
Execute	Present technical results, link notebooks to executive deliverables, design visualizations, and bilingual documentation.	Final presentation (`slides.pdf`), bilingual README, adapted visualizations, reproducible institutional package.

⚠️ Limitations and Assumptions

The XGBoost model has not yet been trained; its use is proposed as a technical strategy for the “Construct” phase.The visualizations presented (such as the variable importance chart) are simulations based on technical assumptions.Independent variables were selected based on their expected relationship with trip cost, according to preliminary exploratory analysis.Performance metrics and model comparisons will be validated in later phases of the project.