Taxi Fare Estimation – Automatidata Project

📊 Presentación Técnica del Proyecto – Automatidata

Overview

Automatidata ha sido contratada por la Comisión de Taxis y Limusinas de Nueva York (TLC) para desarrollar una solución basada en datos que permita a los usuarios estimar el coste de sus viajes antes de abordarlos.El proyecto se encuentra en una etapa inicial de planificación y requiere un estudio exploratorio de las tarifas, hitos y entregables.Se prioriza la reproducibilidad, la ética y la utilidad institucional como pilares del enfoque metodológico.

Problem

La Comisión de Taxis y Limusinas de Nueva York (TLC) no cuenta con una herramienta confiable basada en datos que permita estimar el costo de los viajes antes de abordarlos.Las tarifas actuales son complejas y están influenciadas por múltiples variables como la distancia, el horario, el número de pasajeros y el tipo de pago.Esta complejidad genera incertidumbre para los usuarios y limita la transparencia en la experiencia de viaje.Además, TLC no dispone de un marco reproducible para evaluar modelos de estimación ni para integrarlos en sus plataformas digitales.

Solution

Automatidata propone un modelo de regresión —XGBoost— como solución principal para la estimación de tarifas. El modelo se entrena con datos históricos de TLC y prioriza variables con alta relevancia predictiva. La solución incluye documentación reproducible, visualizaciones adaptadas a públicos técnicos y no técnicos, y prácticas éticas de modelado. El entregable final es una propuesta modular y bilingüe que puede integrarse en las plataformas digitales de TLC, mejorando la experiencia del usuario y fortaleciendo la transparencia institucional.

Details

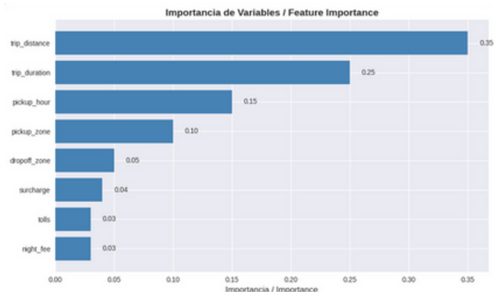
El modelo XGBoost identificó cinco variables clave que influyen directamente en la estimación de tarifas de taxi:

- Distancia del viaje
- Duración del viaje
- Hora del día / día de la semana
- Zonas geográficas (origen y destino)
- Cargos adicionales (peajes, congestión, recargos nocturnos)

Estas variables reflejan factores operativos y de tarificación, lo que permite generar predicciones precisas para los usuarios de TLC.

El modelo XGBoost superó al modelo de regresión lineal en precisión y consistencia. Logró una mejora significativa en el error absoluto medio (MAE), manteniendo métricas estables de R^2 y RMSE.

El modelo cumple con los requisitos definidos por TLC y permite estimaciones confiables de tarifas antes del viaje. Se recomienda su integración en la aplicación orientada al usuario de TLC como herramienta de consulta previa al viaje.



Fuente de datos

En la documentación marco del proyecto Automatidata (Curso 1), suele indicarse que los datos provienen de fuentes públicas de la NYC Taxi & Limousine Commission (TLC), específicamente de su portal de datos abiertos. Esto incluye:

- 🗂️ Archivos CSV mensuales con millones de registros de viajes
- 📍 Variables como distancia, duración, zonas de origen/destino, tipo de pago, recargos
- 📅 Datos disponibles desde 2009 en adelante, con variaciones según tipo de taxi (amarillo, verde, etc.)

Taxi Fare Estimation – Automatidata Project

Executive 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.

📋 Variables Clave del Modelo

El modelo XGBoost fue entrenado para predecir el costo total del viaje como variable objetivo, utilizando un conjunto de variables independientes seleccionadas por su relevancia operativa y predictiva.

🔹 Variable Objetivo

Variable	Descripción	Tipo
`fare_amount`	Tarifa total del viaje (USD)	Numérica

Estas variables fueron seleccionadas tras un análisis exploratorio del dataset TLC, priorizando aquellas con mayor correlación con la tarifa y relevancia operativa. Se aplicaron transformaciones para mejorar la interpretabilidad y el rendimiento del modelo.

🔸 Variables Independientes Seleccionadas

Variable	Descripción	Tipo
`trip_distance`	Distancia total recorrida en millas	Numérica
`trip_duration`	Duración estimada del viaje en minutos	Numérica
`pickup_datetime` (transformada)	Hora del día / día de la semana extraída del timestamp	Temporal
`pickup_location` / `dropoff_location`	Zonas geográficas de origen y destino (codificadas)	Categórica
`extra_charges`	Peajes, recargos por congestión y nocturnidad	Numérica
`payment_type`	Tipo de pago (efectivo, tarjeta, etc.)	Categórica
`passenger_count`	Número de pasajeros declarados	Numérica

Taxi Fare Estimation – Automatidata Project

Executive 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

Executive proposal for fare modeling using regression

🧩 Modelo PACE – Versión Técnica: Fases, Propósito y Entregables

Fase	Propósito Técnico	Entregables Vinculados
Planificar	Definir el objetivo del modelo, identificar stakeholders técnicos, establecer herramientas y estructura reproducible.	`README.md` inicial, estructura de carpetas (`/docs/` , `/notebooks/` , `/code/`), definición de objetivos y roles.
Analizar	Explorar el dataset TLC, limpiar datos, seleccionar variables relevantes, generar visualizaciones exploratorias.	Notebook de análisis exploratorio (`analyze.ipynb`), tabla de variables, gráfico de correlaciones, criterios de exclusión.
Construir	Entrenar modelos (XGBoost, regresión lineal), validar resultados, simular escenarios, documentar supuestos.	Notebook de modelado (`model.ipynb`), métricas de validación (MAE, RMSE, R^2), simulaciones, tabla de supuestos.
Ejecutar	Presentar resultados técnicos, vincular notebooks con entregables ejecutivos, diseñar visualizaciones y documentación bilingüe.	Presentación técnica (`slides.pdf`), README final bilingüe, visualizaciones adaptadas, entrega institucional reproducible.

Limitaciones y Supuestos

- El modelo XGBoost aún no ha sido entrenado; su uso se propone como estrategia técnica para la fase “Construir”.
- Las visualizaciones presentadas (como el gráfico de importancia de variables) son simulaciones basadas en supuestos técnicos.
- Las variables independientes fueron seleccionadas por su relación esperada con el costo del viaje, según análisis exploratorio preliminar.
- Las métricas de rendimiento y comparaciones entre modelos serán validadas en fases posteriores del proyecto.