Taxi Fare Estimation – Automatidata Project

Executive proposal for fare modeling using regression

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 costo de sus viajes antes de abordarlos. El proyecto se encuentra en una etapa inicial de planificación y requiere un marco estratégico que defina tareas, hitos y entregables.

Problem

- Esta propuesta presenta la estructura del proyecto, los actores clave y el enfoque metodológico basado en el modelo PACE. Aunque la TLC posee una gran cantidad de datos históricos, estos aún no han sido transformados en soluciones prácticas. Además, los ejecutivos de la agencia necesitan visualizaciones claras que les permitan comunicar los resultados a audiencias no técnicas.

Solution

- Automatidata propone desarrollar un modelo de regresión utilizando Python, basado en los datos históricos de la TLC. El proyecto se organizará según el modelo PACE, dividiendo las tareas en etapas claras: planificación, análisis exploratorio, construcción del modelo y ejecución de visualizaciones. Se entregará una propuesta técnica y ejecutiva que responda a las necesidades tanto del equipo interno como de los stakeholders externos.

Details

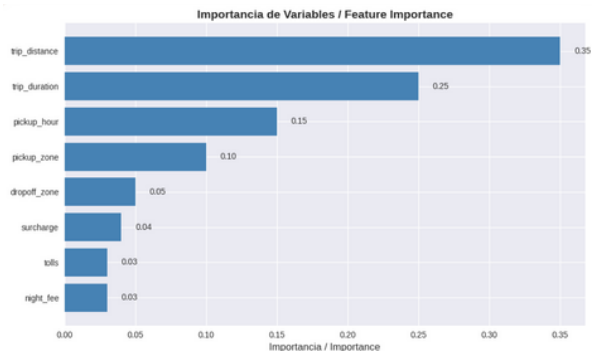
El modelo XGBoost identificó cinco variables clave que influyen directamente en la estimación de tarifas de taxi:

- 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.



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

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.

Modelo PACE – Resumen de Fases

Fase (ES)	Phase (EN)	Propósito / Purpose
Planificar	**Plan**	Definir objetivos, identificar stakeholders, establecer entregables y herramientas
Analizar	**Analyze**	Explorar datos, seleccionar variables relevantes, generar visualizaciones iniciales
Construir	**Construct**	Entrenar modelos, validar resultados, simular escenarios
Ejecutar	**Execute**	Documentar reproduciblemente, presentar resultados, integrar soluciones

El modelo PACE guía el desarrollo ético y estratégico de proyectos de ciencia de datos, asegurando claridad, reproducibilidad y alineación institucional.

Tiempos estimados por fase

Fase	Actividades clave	Tiempo estimado
Planificar	Definición de objetivos, identificación de stakeholders y entregables	1 semana
Analizar	Exploración de datos, selección de variables, visualizaciones iniciales	2 semanas
Construir	Entrenamiento del modelo, validación, simulaciones	2–3 semanas
Ejecutar	Documentación reproducible, presentación, integración institucional	1 semana

El proyecto se estima en 6 a 7 semanas, sujeto a ajustes según disponibilidad de datos y validación técnica.

Taxi Fare Estimation – Automatidata Project

Executive proposal for fare modeling using regression

Participantes clave

Rol	Responsabilidad
Mentor y consultor ético	Supervisión estratégica, documentación reproducible, narrativa ética
Analista de datos	Exploración de datos, selección de variables, modelado predictivo
Desarrollador técnico	Integración del modelo, visualizaciones, soporte técnico
Stakeholders institucionales (TLC)	Validación de objetivos, toma de decisiones, alineación institucional