Documento de Estrategia PACE

Proyecto Automatidata — TLC

Regresión Lineal

PACE: Etapa de Planificación

El proyecto se plantea en el contexto de la Comisión de Taxis y Limusinas de Nueva York (TLC), donde la estimación anticipada de la tarifa de un viaje es un insumo relevante para la planificación operativa, el análisis financiero y la comunicación con usuarios. Desde el inicio, la pregunta central fue cómo **estimar de forma confiable el monto de la tarifa antes de que el viaje ocurra**, utilizando únicamente información disponible en ese momento.

Esto llevó a una primera definición clave: el modelo no debía fijar precios exactos ni replicar reglas regulatorias, **sino proporcionar una estimación interpretable y consistente** que reflejara patrones reales del servicio. ¿Quiénes se beneficiarían directamente de este enfoque? Principalmente los equipos operativos y financieros de la TLC, pero también, de forma indirecta, plataformas de movilidad, conductores y pasajeros que dependen de expectativas de tarifa razonables antes del inicio del viaje.

Con el problema claramente delimitado, surgió la siguiente decisión: **¿qué tipo de modelo era el más adecuado para este objetivo?** Se optó por una regresión lineal múltiple como modelo base, priorizando transparencia y control metodológico sobre complejidad algorítmica. Esta elección permitió interpretar coeficientes, validar supuestos y establecer un punto de referencia sólido antes de explorar modelos más avanzados de machine learning.

Al comenzar la exploración preliminar de los datos, aparecieron nuevas preguntas que condicionaron el enfoque analítico. ¿Los datos eran lo suficientemente limpios como para modelar directamente? La respuesta fue negativa. **Se identificaron valores atípicos, registros no plausibles** —como tarifas negativas o montos extremadamente altos— **y distribuciones fuertemente sesgadas en variables clave** como distancia y monto del viaje. Además, surgió una cuestión fundamental: ¿todos los patrones observados representaban errores de datos? El análisis reveló que no; ciertas estructuras, como las tarifas fijas reguladas para trayectos específicos (por ejemplo, viajes hacia o desde el aeropuerto JFK), correspondían a reglas de negocio válidas y no debían tratarse como anomalías.

Estas observaciones iniciales condujeron a reflexionar sobre los riesgos metodológicos del proyecto. ¿Qué podía comprometer la validez del modelo desde una etapa temprana? Se identificaron dos riesgos principales: la inclusión de variables que solo se conocen después del viaje y la posibilidad de introducir fuga de datos al trabajar con variables agregadas derivadas del conjunto completo. Ambos riesgos se documentaron explícitamente desde la

planificación, estableciendo límites claros sobre qué información podía utilizarse y bajo qué supuestos.

Finalmente, se definió el alcance operativo del proyecto. ¿Cómo estructurar el trabajo para mantener trazabilidad y claridad analítica? Se optó por un enfoque incremental apoyado en Python y notebooks, priorizando la comprensión del problema y la coherencia metodológica antes de optimizar métricas. Desde esta etapa, el modelo fue concebido como un baseline interpretable, útil tanto para extraer insights como para servir de punto de partida hacia enfoques más complejos en fases posteriores.

PACE: Etapa de Análisis

Con el objetivo claramente definido, el análisis exploratorio se enfocó en **preparar un conjunto de datos coherente con el uso real del modelo**, más que en maximizar relaciones estadísticas preliminares. La primera pregunta que guió esta etapa fue: **¿qué información es válida y utilizable antes de que el viaje ocurra?**

La exploración inicial reveló la presencia de valores atípicos y registros no plausibles, como montos negativos, tarifas extremadamente altas y viajes con distancia o número de pasajeros igual a cero. Estos casos no se trataron como ruido aleatorio, sino como señales de calidad de datos que requerían depuración para evitar distorsiones en el modelo.

Durante esta etapa también surgió una decisión clave: **¿todas las variables con alta correlación con la tarifa debían conservarse?** El análisis mostró que varias variables, aunque predictivamente fuertes, solo se conocen una vez finalizado el viaje. Estas variables fueron descartadas deliberadamente en esta fase, aun a costa de reducir el desempeño aparente del modelo, priorizando coherencia metodológica sobre métricas optimistas.

Asimismo, se evaluó el impacto de las reglas de negocio presentes en los datos. Las tarifas fijas reguladas, como las asociadas a determinados trayectos, se identificaron como patrones válidos del dominio y no como errores. Reconocer esta estructura permitió evitar decisiones incorrectas durante la depuración y sentó las bases para su tratamiento posterior.

Finalmente, el análisis permitió definir el conjunto de variables definitivo con el que se trabajaría en la etapa de modelado, asegurando que cada predictor tuviera sentido operativo, estuviera disponible antes del viaje y contribuyera de forma razonable a la explicación de la tarifa.

PACE: Etapa de Construcción

Con el conjunto de variables ya depurado y validado durante el análisis, la etapa de construcción se centró en **materializar el modelo de regresión lineal bajo condiciones controladas**. La primera decisión fue separar los datos en conjuntos de entrenamiento y

prueba, con el objetivo de evaluar la estabilidad del modelo y su capacidad de generalización.

El modelo se entrenó utilizando únicamente las variables definidas en la etapa anterior, sin incorporar nuevas transformaciones que alteraran su interpretabilidad. La ingeniería de características basada en promedios históricos por par origen–destino se aplicó como una forma de capturar información estructural del sistema, siendo conscientes de que su cálculo sobre el conjunto completo introducía un grado de fuga de datos. Este compromiso metodológico fue aceptado de manera explícita y documentado como parte del alcance del proyecto.

Durante la construcción se observó que, tras eliminar las variables con fuga directa de información, el desempeño del modelo se mantuvo consistente entre entrenamiento y prueba. Esta consistencia fue interpretada como una señal de estabilidad, más relevante que una mejora marginal de métricas.

El modelo final se mantuvo intencionalmente simple, sin técnicas avanzadas de regularización o ajuste fino. La prioridad fue construir un **baseline interpretable**, que permitiera explicar el efecto de cada variable y sirviera como punto de referencia claro para enfoques más complejos en etapas posteriores.

PACE: Etapa de Ejecución

Con el modelo ya ejecutado, la atención se centró en verificar si realmente cumple su propósito: **estimar de forma anticipada la tarifa de un viaje con un nivel de error razonable y comportamiento estable**. Los resultados muestran un **R^2 cercano a 0.84 en entrenamiento y 0.87 en prueba**, lo que significa que el modelo explica aproximadamente entre **84 % y 87 % de la variabilidad de la tarifa** usando solo información disponible antes del viaje. En términos simples, si dos viajes son distintos en distancia, duración u horario, el modelo logra capturar la mayor parte de esa diferencia en el monto final. Además, el **margen de error promedio** se sitúa alrededor de **2.2 dólares**, lo que implica que, por ejemplo, para un viaje con una tarifa real de **14 dólares**, el modelo suele estimar un valor entre **12 y 16 dólares**, un rango razonable para una estimación previa.

Al revisar casos individuales, se observa este comportamiento de forma clara. En viajes cortos y de bajo monto, las diferencias entre la tarifa real y la estimada suelen ser pequeñas, mientras que en trayectos más largos el error puede ampliarse, aunque sin mostrar un sesgo sistemático. Esto refuerza la idea de que el modelo no está “adivinando”, sino siguiendo patrones consistentes del sistema. La estabilidad entre entrenamiento y prueba indica, además, que el modelo **generaliza bien** y no depende de particularidades de un subconjunto específico de datos.

Uno de los objetivos centrales de esta etapa fue **entender qué factores influyen en el costo por milla recorrida**. El análisis de los coeficientes confirma que la *distancia promedio* es el factor con mayor impacto, seguida por la *duración promedio* y, en menor medida, el *horario punta*. Para interpretar esto de forma intuitiva, es importante considerar

que las variables fueron estandarizadas. Al llevar el efecto de la *distancia promedio* a su escala original, se obtiene que **cada milla adicional recorrida incrementa la tarifa en aproximadamente 2 dólares**, manteniendo constantes los demás factores. Esto significa que un viaje que pasa de 5 a 10 millas no solo es más largo, sino que, en promedio, **cuesta unos 10 dólares más**, coherente con la lógica operativa del servicio.

Durante esta etapa surgió una reflexión clave: **¿cómo capturar información estructural del sistema, como patrones recurrentes entre zonas de origen y destino?** La respuesta está en reconocer que estos patrones existen y que ignorarlos empobrece la representación del comportamiento real del sistema. Por ello se incorporaron campos agregados como la *distancia promedio* y la *duración promedio* históricas por combinación de zonas, que permiten reflejar rutas habituales y condiciones típicas de operación. Naturalmente, esto llevó a una segunda pregunta: **¿estos nuevos campos no implican fuga de información?** En sentido estricto, podrían introducirla si se usaran sin cuidado; sin embargo, en este contexto analítico se consideran aceptables porque el objetivo es construir un modelo explicativo y didáctico. Además, el efecto de esta decisión se reconoce explícitamente y se entiende como un compromiso consciente, no como una omisión metodológica.

En conjunto, la ejecución confirma que el modelo alcanza un **equilibrio adecuado entre precisión, claridad e interpretabilidad**. El nivel de explicación del $\mathbf{R^2}$, el **margen de error cercano a 2 dólares** y una estructura de costos intuitiva —como el incremento aproximado de **2 dólares por milla adicional**— permiten comunicar los resultados de forma comprensible a públicos no técnicos. Aunque existen limitaciones y simplificaciones asumidas de manera consciente, el modelo cumple su función como **línea base sólida**, útil para estimar tarifas antes del viaje y para sentar las bases de enfoques más avanzados en trabajos futuros.