Universidad Católica Boliviana “San Pablo”

Informe Técnico: Predicción de Demanda de Movilidad Urbana (Transporte)

Machine Learning

Docente:

Paton Gutierrez Ovidio Roger

Integrantes:

- Jiménez Mendoza Manuel Franco
- Jaicel Jesus Rodrigo Velasco Turunco
- Samuel Denis Villca Castro

12 de junio de 2025

Contenido

1. Descripción del Problema

El sistema de bicicletas compartidas de una ciudad metropolitana como Londres enfrenta un desafío logístico constante: mantener el equilibrio dinámico del inventario, y al mismo tiempo, evitar que las estaciones se saturen al punto de no permitir devoluciones. Mediante el análisis de un dataset histórico de más de 17,000 registros horarios, este proyecto busca optimizar la toma de decisiones basada en datos.

Objetivos del Proyecto:

1. **Regresión:** Construir un modelo predictivo para estimar la cantidad total de alquileres de bicicletas (cnt) en un momento dado, basándose en factores climáticos y temporales.
2. **Clasificación:** Desarrollar un modelo capaz de detectar "Horas Pico" (High_Demand) para generar alertas tempranas que justifiquen la movilización de recursos logísticos.

Criterios de Éxito y Costos de Error:

- **Falso Positivo (FP):** Predecir una hora pico que no ocurre. Consecuencia: Gasto innecesario en la movilización de camiones de redistribución.
- **Falso Negativo (FN):** No detectar una hora pico real. Consecuencia: Estaciones vacías, pérdida de ingresos e insatisfacción del cliente.

2. Supuestos y Análisis Exploratorio (EDA)

Durante la exploración de los datos históricos, se confirmaron varios supuestos clave sobre la movilidad urbana:

- **Comportamiento Laboral vs. Recreativo:** Se observó que la demanda es significativamente mayor en días laborables frente a días festivos (is_holiday). Esto sugiere que el sistema se utiliza primordialmente para *commuting* (viajes al trabajo/estudio).
- **Sensibilidad Climática:** Existe una fuerte caída en la demanda durante el invierno o bajo condiciones climáticas adversas (nieve o lluvia intensa, capturadas en weather_code).
- **Multicolinealidad:** Se detectó una correlación positiva casi perfecta ($r = 0.99$) entre la temperatura real (t1) y la sensación térmica (t2) Como se puede observar en la Ilustración 1.

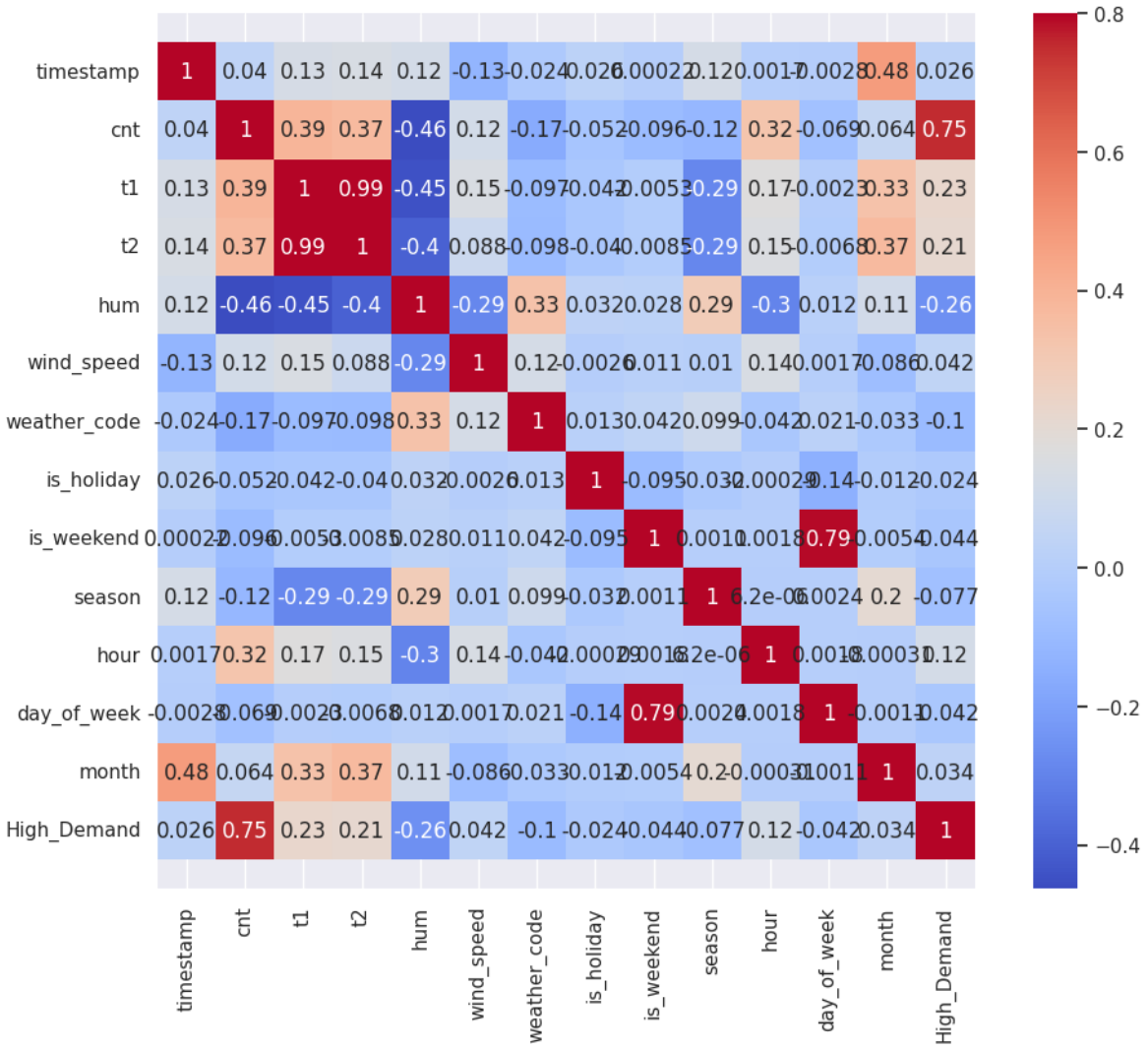

Ilustración 1 Matriz de correlación

3. Preprocesamiento de Datos Reproducible

Para garantizar la reproducibilidad y evitar la fuga de información (*data leakage*), se implementó un flujo de trabajo basado en Pipelines y ColumnTransformer de Scikit-Learn:

1. **Ingeniería de Características Temporal:** Se transformó la variable timestamp para extraer características granulares: hora (hour), día de la semana (day_of_week) y mes (month).
2. **Estratificación:** Dado el fuerte impacto de las estaciones del año, el conjunto de datos se dividió en entrenamiento (80%) y prueba (20%) estratificando por la variable season.

3. **Pipeline Numérico:** A las variables numéricas (t1, t2, hum, wind_speed) se les aplicó un SimpleImputer para manejar posibles valores nulos, seguido de un StandardScaler para estandarizar las magnitudes.
4. **Pipeline Categórico:** Las variables categóricas y temporales fueron procesadas mediante OneHotEncoder para convertirlas en un formato interpretable por los modelos matemáticos.

Importante: El preprocesador fue ajustado (fit) estrictamente sobre el conjunto de entrenamiento. Al conjunto de prueba solo se le aplicó la transformación (transform).

4. Modelado y Evaluación: Regresión (Predicción de Demanda)

Para predecir la variable continua (cnt), se establecieron dos modelos paramétricos: una Regresión Lineal Clásica (Baseline) y una Regresión Lasso (con regularización L1).

Diagnóstico de Multicolinealidad

En la Regresión Lineal, la alta correlación entre t1 y t2 provocó inestabilidad en los coeficientes, asignando pesos exagerados y opuestos (t1: 434.03, t2: -222.46). Al implementar la Regresión Lasso, la penalización L1 mitigó este problema llevando el coeficiente de t2 exactamente a 0.0, seleccionando automáticamente t1 (225.56) como la única variable de temperatura necesaria.

Tabla de Métricas

Modelo	RMSE (Raíz del Error Cuadrático Medio)	MAE (Error Absoluto Medio)	Ventaja Principal
Regresión Lineal	583.79	406.24	Modelo base estándar.
Regresión Lasso	588.70	409.98	Manejo de multicolinealidad e interpretabilidad.

Conclusión de Regresión: Aunque el RMSE sube marginalmente con Lasso, se prefiere este modelo porque es más robusto y elimina el ruido introducido por variables redundantes, generalizando mejor ante datos futuros.

5. Modelado y Evaluación: Clasificación (Detección de Horas Pico)

Se definió la variable objetivo High_Demand asignando el valor 1 si el número de alquileres (cnt) superaba el **percentil 90** de la distribución histórica. Esta decisión aísla el 10% de eventos extremos donde la capacidad de la red realmente corre riesgo de colapso, manteniendo el enfoque en anomalías logísticas.

Se entrenó un modelo de **Regresión Logística**. Dado el desbalance de clases (90% Normal / 10% Alta Demanda), se configuró el modelo con pesos balanceados (class_weight='balanced').

Diagnóstico con Umbral por Defecto (0.50)

Al utilizar el umbral de probabilidad estándar (0.50), la Matriz de Confusión arrojó los siguientes resultados sobre el conjunto de validación:

- **Verdaderos Positivos (Picos detectados):** 345
- **Falsos Negativos (Picos no detectados):** 17
- **Falsos Positivos (Falsas alarmas):** 323

Análisis: El modelo prioriza fuertemente el Recall (Exhaustividad), lo cual genera un exceso de falsas alarmas (323). Operativamente, esto sería muy costoso.

6. Interpretación y Decisiones de Negocio

Para solucionar el exceso de falsas alarmas, se analizó la **Curva Precision-Recall (PR)**. El Área Bajo la Curva (PR-AUC) fue de 0.699, demostrando una buena capacidad discriminativa.

Recomendación del Umbral Operativo:

Buscando maximizar el F1-Score (el equilibrio armónico entre Precisión y Exhaustividad), se identificó matemáticamente que el **umbral óptimo es 0.74**.

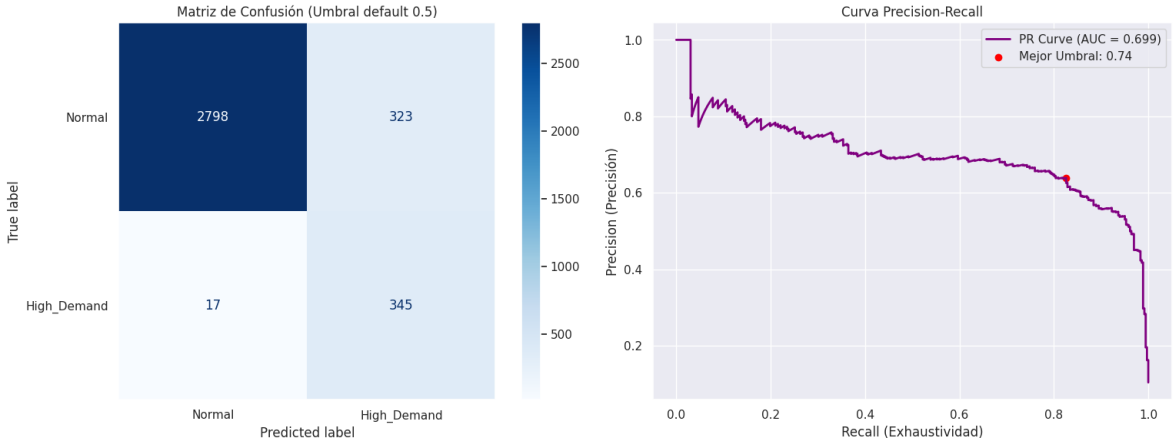

Ilustración 2 La Curva Precision-Recall

Trade-off y Justificación:

Al subir el umbral de decisión a 0.74, le exigimos al modelo estar un 74% seguro antes de emitir una alerta de "Hora Pico".

- **Beneficio Logístico:** Se reducirá drásticamente la cantidad de Falsos Positivos, ahorrando miles de libras esterlinas en envíos innecesarios de camiones de redistribución.
- **Riesgo Asumido:** Aumentarán ligeramente los Falsos Negativos. Ocasionalmente, alguna estación podría quedarse sin bicicletas, pero este costo de insatisfacción es asumible frente a los masivos ahorros logísticos.

7. Breve Discusión de Limitaciones

A pesar de los resultados robustos, el presente estudio tiene las siguientes limitaciones:

1. **Falta de granularidad espacial:** El modelo predice la demanda a nivel de ciudad (o red global). No nos indica *qué* estación específica se vaciará. Modelos futuros deberían incluir datos geoespaciales por estación.
2. **Modelos Lineales Limitados:** Las relaciones entre la demanda y variables como la hora del día no son estrictamente lineales. Modelos basados en árboles (como Random Forest o XGBoost) podrían capturar mejor los patrones no lineales y cíclicos (como los picos de demanda a las 8 AM y 5 PM).
3. **Factores exógenos omitidos:** Eventos atípicos como huelgas de transporte público (metro), eventos masivos (conciertos) o cierres de calles no están contemplados en el dataset actual, lo cual podría explicar variaciones abruptas no predecibles por el modelo actual.