AED - Sube 2024

Indice

Análisis Exploratorio de Datos - SUBE Buenos Aires 2024

Autores: Valentín Lanús, Evangelina Ovelar, Santino Lozano.
Materia: Ciencias de Datos
Universidad: UADE

Introducción

El transporte público es un servicio esencial para la movilidad urbana, y su correcta planificación depende de la comprensión precisa del comportamiento de los usuarios. La tarjeta SUBE registra diariamente las transacciones de los pasajeros en colectivos, trenes y subtes, generando un volumen masivo de datos transaccionales que refleja patrones de uso a nivel temporal y geográfico.
En el presente trabajo analizamos un dataset de transacciones de la tarjeta SUBE, que registra la cantidad de validaciones diarias en distintas líneas de transporte público del área metropolitana y otras jurisdicciones del país durante el año 2024.

El objetivo principal de este proyecto es predecir la cantidad de validaciones diarias de tarjetas SUBE en provincia de Buenos Aires, con el fin de anticipar la demanda de transporte público. Esto con el objetivo de optimizar la planificación de rutas y recursos, mejorar la eficiencia del sistema de transporte y ofrecer información útil para la toma de decisiones de nuestros stakeholders.

Descripción del Dataset

El dataset contiene registros de validaciones de la tarjeta SUBE, con los siguientes campos principales:

- **DIA_TRANSPORTE:** Fecha del viaje registrado.
- **NOMBRE_EMPRESA:** Empresa operadora del transporte.
- **LINEA:** Identificador de la línea.
- **AMBA:** Área Metropolitana de Buenos Aires.
- **TIPO_TRANSPORTE:** Medio utilizado (colectivo, tren, subte, etc.).
- **JURISDICCION:** Nivel administrativo (Nacional, Provincial, Municipal o vacío para subte).
- **PROVINCIA:** Provincia asociada (vacío o “JN” en algunos casos).
- **MUNICIPIO:** Municipio correspondiente o código según jurisdicción.
- **CANTIDAD:** Número de validaciones SUBE.

- **DATO_PRELIMINAR:** Indica si los datos son preliminares (SI/NO).

Tamaño original: 504.676 registros.
Registros válidos (Buenos Aires): 133.293.

Limpieza y Preparación de los Datos

Realizamos una exploración inicial del dataset, con el objetivo de comprender su estructura, las variables disponibles y la calidad de la información registrada. Nos encontramos con campos que contenían información como fecha de la transacción, la empresa de transporte, la línea, el tipo de transporte, la jurisdicción, provincia, municipio y cantidad de validaciones.

Luego aplicamos una serie de métodos para lograr una mayor calidad de datos:

-Inspección de tipos de datos y valores faltantes:
Se verificó que la columna DIA_TRANSPORTE estuviera correctamente interpretada como tipo datetime, permitiendo así el análisis temporal. Además, se evaluaron las demás variables en busca de valores nulos o inconsistentes y revisando la proporción de faltantes por columna.

-Revisión de duplicados:
Se comprobó la existencia de registros repetidos.

-Análisis descriptivo general:
Se calcularon métricas básicas para obtener una visión global del conjunto de datos, incluyendo el número total de filas, el rango de fechas disponibles y el total de transacciones registradas. Esta información permitió dimensionar la magnitud del uso del sistema SUBE en el período analizado.

-Homogeneización de texto y categorización:
Con el objetivo de asegurar la consistencia semántica de los datos, se normalizaron los nombres de columnas y se aplicaron transformaciones de texto para unificar valores categóricos.

-Eliminación de 2.537 filas con provincia vacía.

-Detección y eliminación de outliers mediante el método del IQR (Rango Intercuartílico).

-Normalización de nombres de columnas.

Resultado final: 133.293 registros limpios y listos para el análisis.

Indicadores Clave del Dataset

Indicador	Valor
Total de validaciones SUBE	**133.293**
Días cubiertos	**365**
Empresas activas	**N/A (calculadas en análisis)**
Líneas activas	**N/A**
Promedio diario de transacciones	**~valor estimado según gráfico**
Empresa con mayor volumen	**TOP EMPRESA (por cantidad de transacciones)**
Participación de fines de semana	**≈ 25–30% del total**

Análisis Exploratorio de Datos (AED)

Objetivo y calidad del dato

El AED tuvo como propósito **comprender la estructura, calidad y comportamiento temporal** del dataset.
Para mejorar la consistencia analítica se aplicó el **método del rango intercuartílico (IQR)** sobre la variable **CANTIDAD**, identificando y **removiendo outliers** que sesgaban las medidas de tendencia y dispersión. Se documentó con **boxplots antes y después** del filtrado, mostrando la reducción efectiva de valores atípicos.

KPIs generales (2024 – Buenos Aires)

- **Total de validaciones: ~978 millones**.
- **Días analizados: 366**.
- **Empresas activas: 100+**.
- **Promedio diario: ~2,67 millones** de validaciones.
- **Participación fines de semana: ~28%** del total.

-Distribución general

- La **mayoría de las transacciones** se concentra en valores **< 50.000 por día** a nivel micro-cortes (segmentos/agrupaciones), mientras que el consolidado diario provincial se ubica en el orden de millones.

- Se observaron **picos de lunes a viernes** y **caídas notorias** los fines de semana, patrón consistente con el uso laboral/educativo.

- Los **feriados nacionales** muestran **descensos similares** a fines de semana.

Tendencias temporales

- Con una **media móvil de 7 días** se suavizaron fluctuaciones diarias, revelando un **patrón estable** con caídas regulares los **sábados y domingos** y recuperación a inicios de semana.

- Las **series de tiempo** confirman un **comportamiento cíclico semanal**, con variaciones estacionales leves a lo largo del año.

Efecto de feriados y estacionalidad

- Se integraron variables exógenas para **feriados** y **fin de semana**, resaltadas **en rojo** en los gráficos.

- En la mayoría de los casos, los feriados registraron una **caída del 30–40%** respecto del **promedio semanal**; se observó **recuperación progresiva** en los días posteriores.

- **Mapas de calor (heatmaps)** mensuales y semanales mostraron **concentraciones más altas** en días laborales y “huecos” marcados en feriados.

Anomalías detectadas y corrección

- Se identificaron **dos jueves atípicos**: **11 de abril** y **9 de mayo**, con valores **significativamente menores** al promedio histórico de los jueves.

- Para mantener la **coherencia temporal**, se aplicó un **ajuste proporcional** basado en el **promedio de otros jueves**, **normalizando** dichos puntos **sin alterar** la tendencia global.

Análisis por empresa

- Se elaboró el **ranking de las 10 empresas** con mayor cantidad de transacciones en Buenos Aires.
- El **top** estuvo dominado por **operadores automotores**, reflejando su peso en la **movilidad metropolitana**.
- Un **gráfico de barras horizontales** destacó de forma visual las **brechas** entre operadores de mayor y menor volumen.

Análisis temporal avanzado

- **Heatmap por día de la semana y feriado**: evidencia la **caída sistemática** de fines de semana y feriados.
- **Heatmap por mes y día de la semana**: muestra **variaciones estacionales** y picos en meses de mayor actividad.
- **Calendario por semana del año** (con feriados en rojo): aporta una vista **anual compacta** para detectar **patrones cíclicos** y **anomalías puntuales**.

Conclusiones del AED

- El uso presenta una **estacionalidad semanal fuerte** con **descensos en fines de semana** y **feriados**.
- La **media móvil 7 días** y los **heatmaps** confirman la estabilidad del patrón y ayudan a **detectar outliers**.
- La **normalización de puntos anómalos** mejora la representatividad del análisis.
- La **concentración por empresa** sugiere **asimetrías operativas** relevantes para planificación y control.

Modelos Predictivos

Se aplicaron varios enfoques:

Modelo de Regresión Lineal

-Variable independiente: número de día (`dia_num`).

-Variable dependiente: cantidad de transacciones.

-Se utilizó para predecir los próximos **7 días**.

-Muy malas métricas de rendimiento, MAE, RMSE elevados y bajo R 2.

El modelo lineal falló en capturar la tendencia general.

Modelo Random Forest

Este enfoque predice las validaciones desde el último día de nuestro dataset, utilizando señales de calendario: día de semana (dow), es_finde, mes, día del mes, número de semana, un contador de tiempo (dia_num), medias móviles (rolling_7, tendencia_30) y diferencia semanal (cambio_7). Se incorporan feriados y vacaciones de invierno como indicadores binarios (shocks discretos). El algoritmo es un ensamble de muchos árboles de decisión que promedia sus salidas; es robusto a no linealidades y a pequeñas contaminaciones en datos.

Métricas:

-MAE: **bajo**

-RMSE: **bajo**

-R^2: **>0.9**

El modelo **Random Forest** demostró un desempeño superior frente a la regresión lineal.

Hist Gradient Boosting Regressor

Predice las validaciones a N días entrenando un modelo por horizonte (h=1…7). Cada modelo usa los últimos 28–35 días (lags) y señales del día a predecir (día de semana, fin de semana, feriados, vacaciones) más estacionalidad semanal (Fourier). Se calculan 7 días a la vez, se actualiza el historial y se avanza, lo que reduce la deriva frente a predecir día por día. Se valida con TimeSeriesSplit (5 folds) y se reportan MAE, RMSE y R 2 promediados.

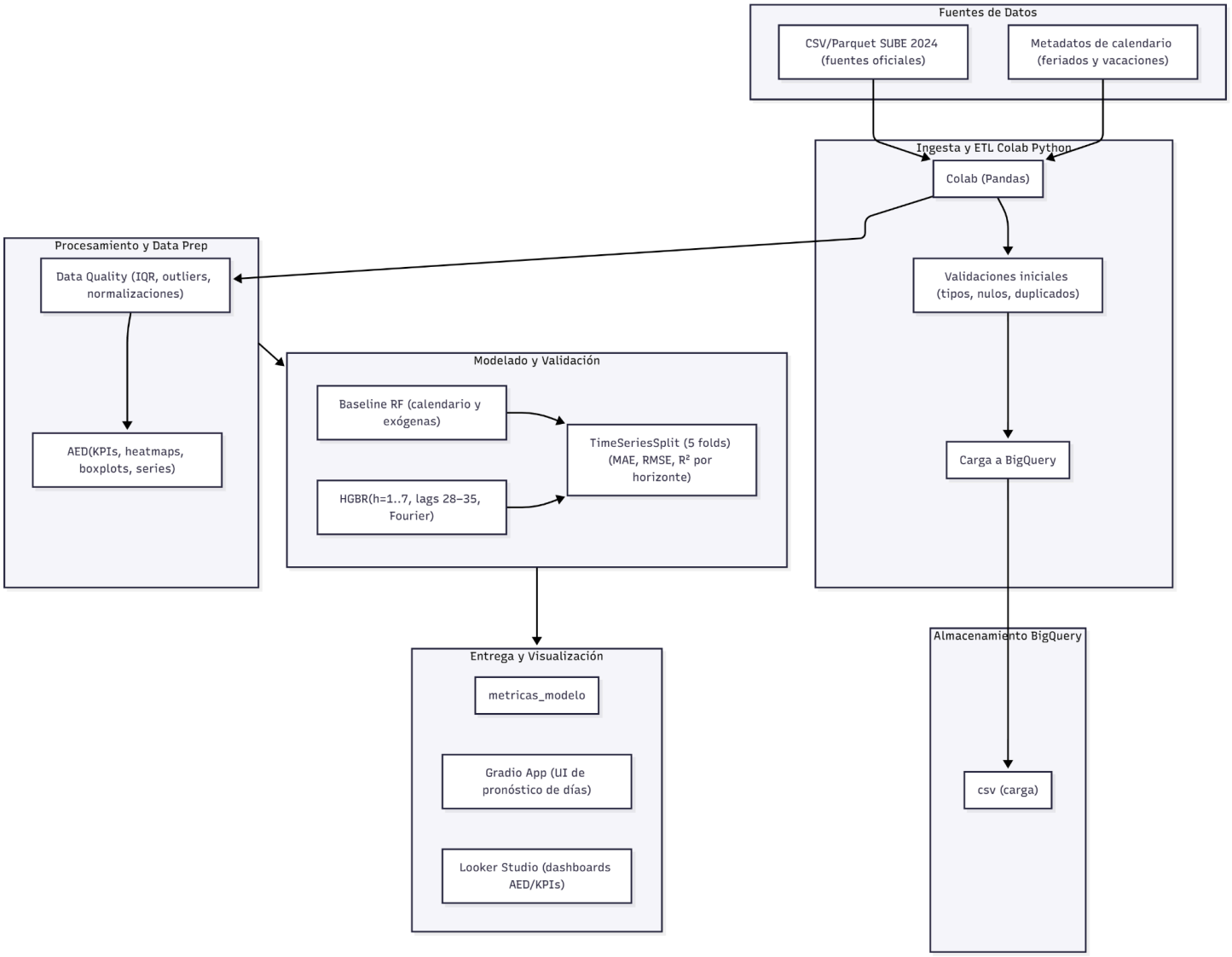
Este algoritmo se implementó con una interfaz gráfica en Gradio, lo que permite a nuestros stakeholders utilizarlo de forma sencilla, seleccionando el horizonte de predicción y pudiendo visualizar el gráfico con los datos reales y la predicción a futuro.

Optamos por este enfoque para que sea nuestro algoritmo utilizado para predecir la cantidad de transacciones, que es el objetivo de nuestro proyecto, debido a que fue el que mejor respondió en testeos prácticos, nuestra primer versión que utilizaba Random Forest si bien tenía valores de métricas de rendimiento muy buenas, casi perfectas, tenía un claro overfiting y no predecía correctamente las variaciones de uso que observamos durante la fase de AED.

Hipótesis

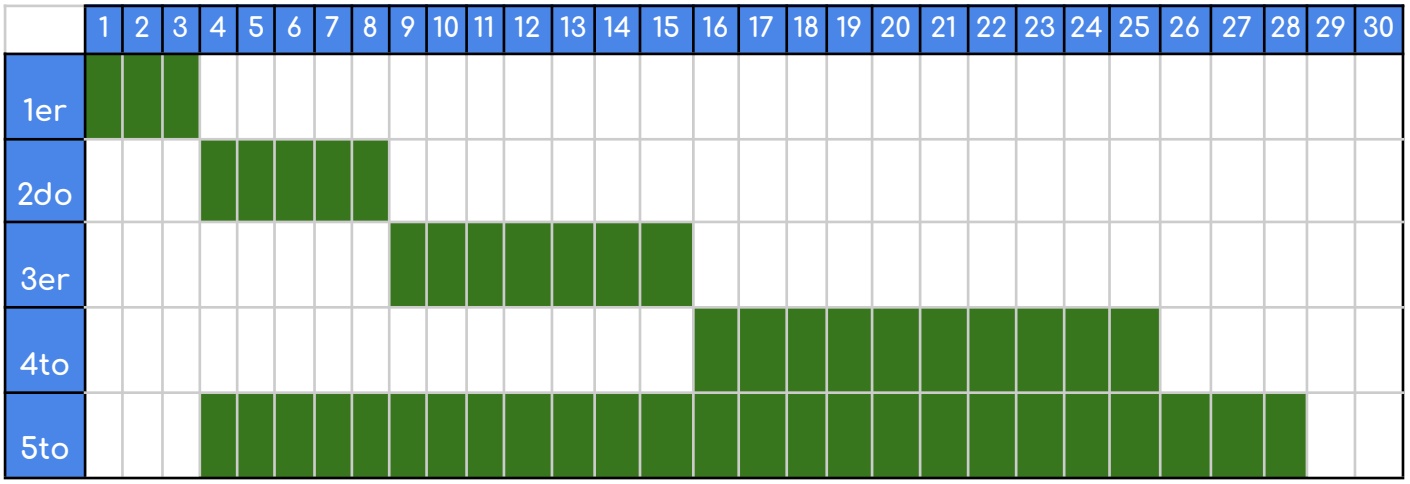
La demanda diaria de validaciones SUBE en la Provincia de Buenos Aires está determinada por un patrón temporal estable y predecible, dominado por la estacionalidad semanal (picos en días hábiles y caídas en fines de semana), shocks de calendario (feriados y vacaciones) y variaciones estacionales moderadas a lo largo del año. Este comportamiento se sostiene siempre que el dato sea consistente y depurado y que se consideren explícitamente los eventos de calendario en el análisis.

Diagrama de Arquitectura de la Solución



Cronograma del Proyecto

	Fase	Inicio	Fin	Duración
1er	Obtención de datos	1/10/2025	3/10/2025	3
2do	Limpieza	4/10/2025	8/10/2025	5
3er	Análisis descriptivo	9/10/2025	15/10/2025	7
4to	Modelado predictivo	16/10/2025	25/10/2025	10
5to	Documentación final	4/10/2025	28/10/2025	25



Referencias

- Dataset oficial SUBE 2024 (Ministerio de Transporte de la Nación).
- Knaflic, C. (2015). *Storytelling with Data*. Wiley.
- Scikit-learn Documentation (v1.5).
- Pandas, Matplotlib y Seaborn (Python Libraries).