Proyecto Final Ciencia De Datos:
Modelos predictivos para la afluencia en el Metrobús de la CDMX

Millaray Sarmiento Escobar
Tutores: David Alexis García Espinosa, Luis Eduardo Flores Luna,
Ángel Andrés Moreno Sánchez, Derek Saúl Morán Pérez
Universidad Nacional Autónoma de México, Ciudad de México, México

Resumen

La movilidad en la Ciudad de México representa un desafío constante debido a su alta densidad poblacional y la demanda creciente de sistemas de transporte. Este trabajo presenta modelos predictivos basados en XGBoost para estimar la afluencia diaria de pasajeros en las líneas 1 y 2 del Metrobús.

Palabras clave: movilidad urbana, modelos predictivos, XGBoost, afluencia de pasajeros, transporte público, ciencia de datos

1. Introducción

La ciencia de datos ha adquirido especial relevancia en los últimos años debido a su enfoque multidisciplinario para transformar, organizar, modelar, analizar, visualizar y comunicar datos útiles. Los modelos de aprendizaje automático como redes neuronales o árboles de decisión han permitido anticipar afluencias, ajustar frecuencias de unidades y mejorar la experiencia del usuario.

2. Contexto

La Ciudad de México tiene más de 9 millones de habitantes y más de 20 millones en el área metropolitana. El Metrobús, sistema de transporte público de tránsito rápido, ha sido una solución clave para enfrentar la alta demanda de movilidad.

Dentro de la red del Metrobús, las líneas 1 y 2 son de especial relevancia:

- **Línea 1 (Indios Verdes-El Camionero):** Conecta el norte y sur de la ciudad, atravesando importantes alcaldías como Gustavo A. Madero, Cuauhtémoc, Benito Juárez, Coyoacán y Miguel Hidalgo. Algunas de las estaciones más transitadas son Indios Verdes, Colonia del Valle y Perisur. Esta línea moviliza aproximadamente a 480,000 personas diariamente.
- **Línea 2 (Tepalcates-Tacubaya):** Recorre el oriente y poniente de la ciudad, pasando por las alcaldías de Iztapalapa, Iztacalco, Benito Juárez, Cuauhtémoc y Miguel Hidalgo. Destacan estaciones como Coyuya, Etiopía y Tacubaya. Tiene una afluencia diaria aproximada de 180,000 personas.

Este estudio enfoca su análisis en estas dos líneas por su alta demanda y cobertura territorial, siendo representativas del comportamiento general del sistema de Metrobús.

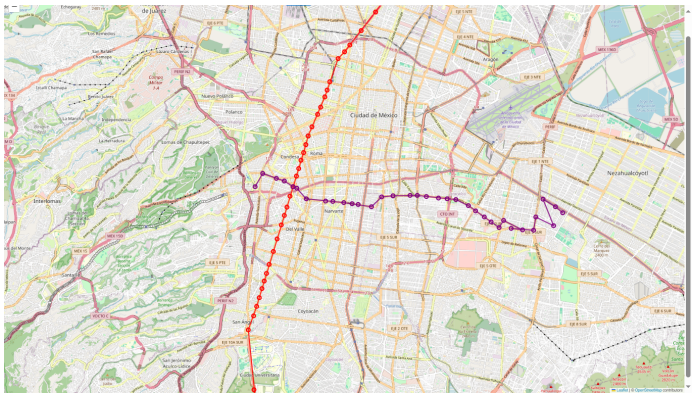

Figura 1: Mapa de la CDMX con estaciones de Metrobús, línea 1(roja) y la línea 2 (morada)

3. Objetivo

El objetivo del presente trabajo es aplicar técnicas de ciencia de datos para generar dos modelos predictivos basados en el algoritmo XGBoost que estimen la afluencia diaria de personas en la línea 1 y 2 del metrobús de la Ciudad de México durante los meses de febrero a mayo 2025 con los datos históricos de febrero a mayo del 2024.

4. Desarrollo

4.1. Preparación de los datos

El conjunto de datos inicial tenía registros históricos diarios de febrero a mayo de 2024 de las líneas 1 y 2 del Metrobús de la CDMX, con variables como fecha, año, tipo de pago (prepago y gratuidad), temporal_fecha y afluencia.

Se analizaron los datos y no existió presencia de datos nulos por lo cual se procedió a agrupar las afluencias en base a la forma de pago (prepago y gratuidad) para obtener un registro diario total al cual se le nombró afluencia_total, eliminando las columnas redundantes (afluencia y tipo de pago). Adicionalmente, se excluyó el registro correspondiente al 29 de febrero de 2024 para mantener la consistencia temporal y evitar sesgos en los modelos predictivos.

Posteriormente, los datos se separaron por línea (1 y 2) y se añadieron las variables climatológicas, incluyendo temperatura del aire, precipitación, humedad específica y humedad relativa.

4.2. Análisis exploratorio de los datos (EDA)

Se generaron gráficos de líneas para visualizar la afluencia a lo largo del tiempo, detectando mayor demanda en la Línea 1. La afluencia crece mensualmente. También se realizaron diagramas de caja para ambas líneas, observando valores atípicos en días de semana, decremento del flujo de personas los fines de semana.

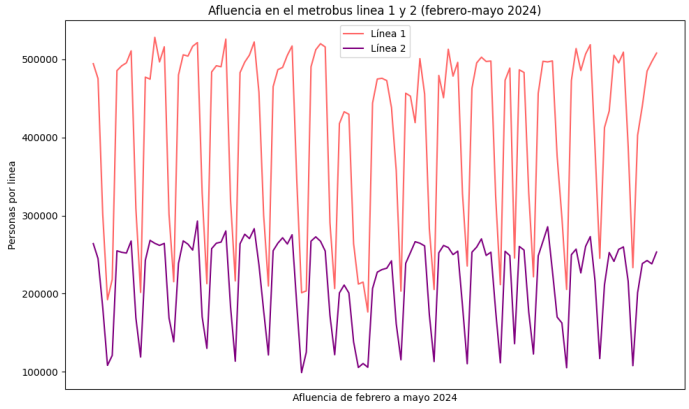

Figura 2: Afluencia de febrero a mayo 2024 para la línea 1 y 2 del metrobús

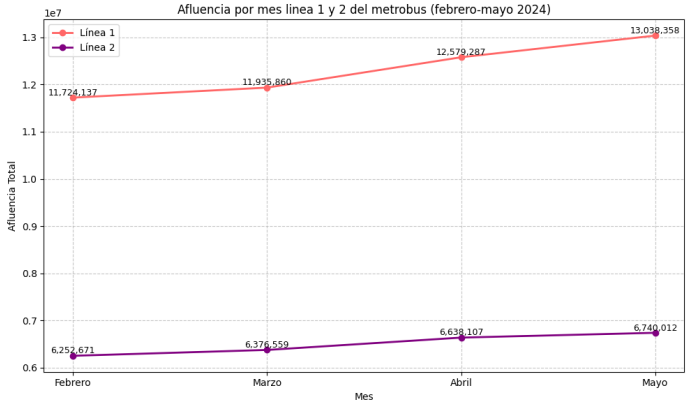

Figura 3: Afluencia de febrero a mayo 2024 para la línea 1 y 2 del metrobús

Para la línea 1 encontramos un media de 410,647 y una mediana de 473,087, con valores más cercanos al tercer cuartil (496,815) ,asimismo contamos con una desviación estándar de 113,024.

Para la línea 2 encontramos un media de 216,727 y una mediana de 243,921, con valores más cercanos al tercer cuartil 261,362,adicionalmente tenemos una desviación estándar de 57,508. Ambas líneas tienen distribuciones sesgadas a la izquierda , indicando que la mayoría de los días tienen afluencia cercana al tercer cuartil, con algunos días con valores bajos , lo que que reduce la media. La desviación estándar en especial para la Línea 1 muestra la variabilidad en la afluencia

Estadístico	Línea 1	Línea 2
Número de registros (count)	120	120
Media de pasajeros (mean)	410,647	216,728
Desviación estándar (std)	113,024	57,509
Mínimo de pasajeros (min)	176,589	98,966
Percentil 25 %	316,837	172,914
Mediana (50 %)	473,088	243,922
Percentil 75 %	496,815	261,363
Máximo de pasajeros (max)	528,289	293,031

Tabla 1. Estadísticos descriptivos de afluencia diaria para las líneas 1 (Indios Verdes-El Caminero) y 2 (Tacubaya-Tepalcates) del Metrobús CDMX. Período analizado: febrero a mayo de 2024.

Los diagramas de caja y bigotes muestran outliers inferiores y en algunos días outliers superiores, por tanto hay días que la afluencia es baja en comparación con la demanda regular, lo que explica porque en los descriptivos la mediana es superior a la media

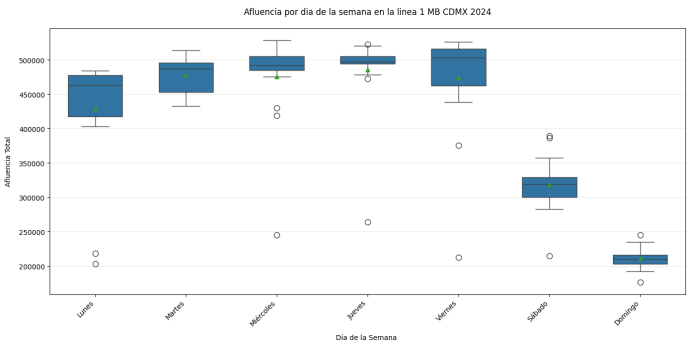

Figura 4: Afluencia por día de la semana para la línea 1 del MB CDMX 2024

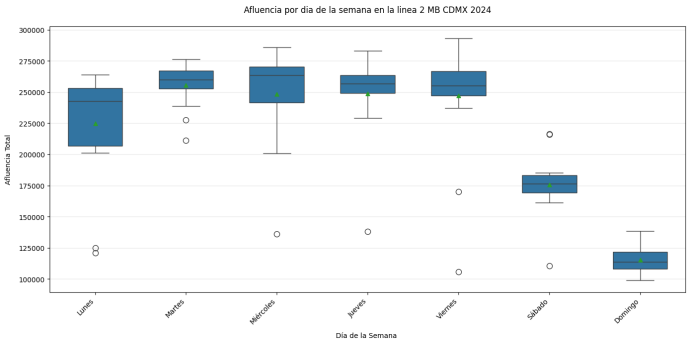

Figura 5: Afluencia por día de la semana para la línea 2 del MB CDMX 2024

Por último se obtuvieron las mapa de calor entre las variables, en donde para ambas líneas se encontraron correlaciones muy ba-

jas con la afluencia,en la mayoría de los casos con valores menores a .1 en relación a la afluencia

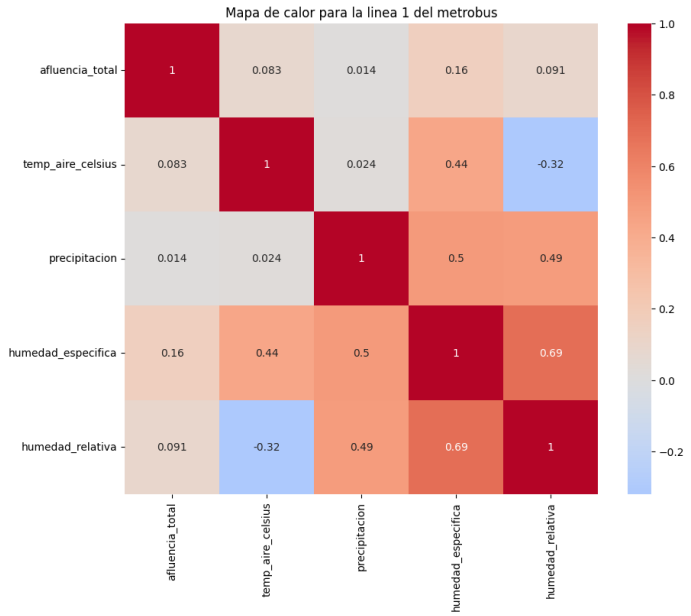

Figura 6: Correlación entre las variables para la línea 1 del metrobús CDMX (febrero-mayo 2024)

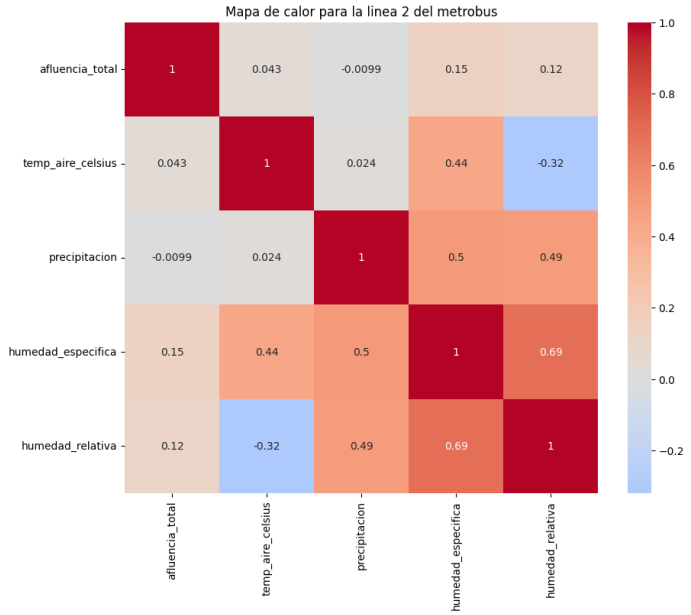

Figura 7: Correlación entre las variables para la línea 2 del metrobús CDMX (febrero-mayo 2024)

4.3. Ingeniería de características

Se incorporaron nuevas variables:

- `dia_semana`: codifica del 0 (lunes) al 6 (domingo)
- `dia_mes`: día del mes (1-31)
- `categoría_día`: días entre semana, fines de semana y feriados

4.4. ACF Y PACF

Se generaron gráficas de autocorrelación (ACF) y autocorrelación parcial (PACF) en un periodo de 60 días. Se detectaron patrones estacionales para los lags 1, 6, 7, 14 y 21, y autocorrelaciones negativas para los lags 2, 3 y 4.

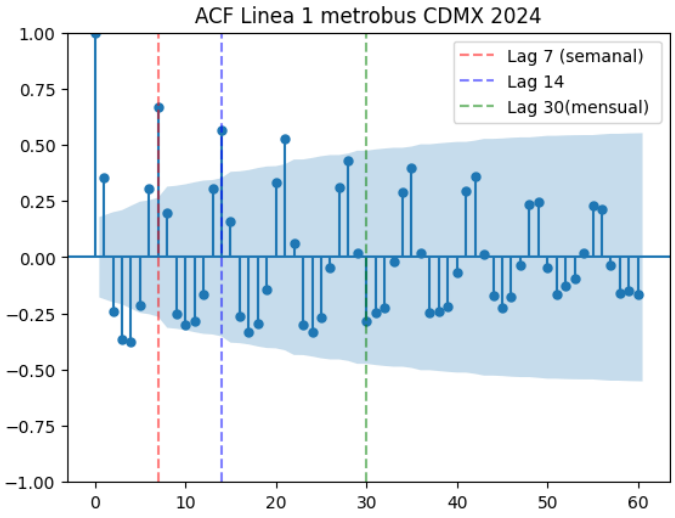

Figura 8: ACF para la línea 1 del metrobús de la CDMX (2024)

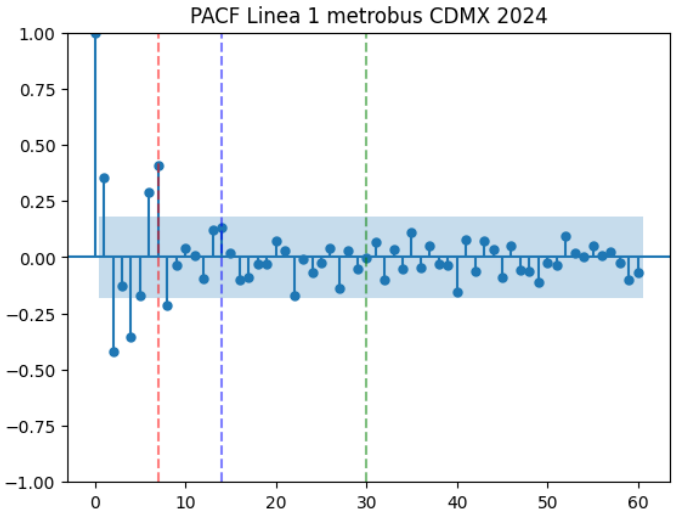

Figura 9: PACF para la línea 1 del metrobús de la CDMX (2024)

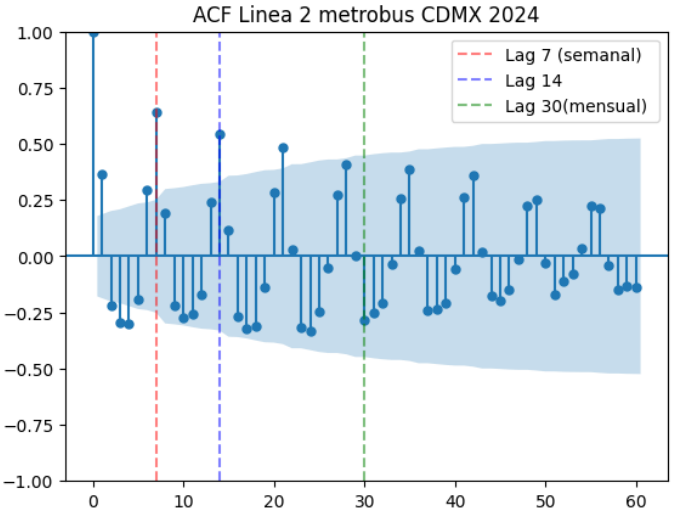

Figura 10: ACF para la línea 2 del metrobús de la CDMX (2024)

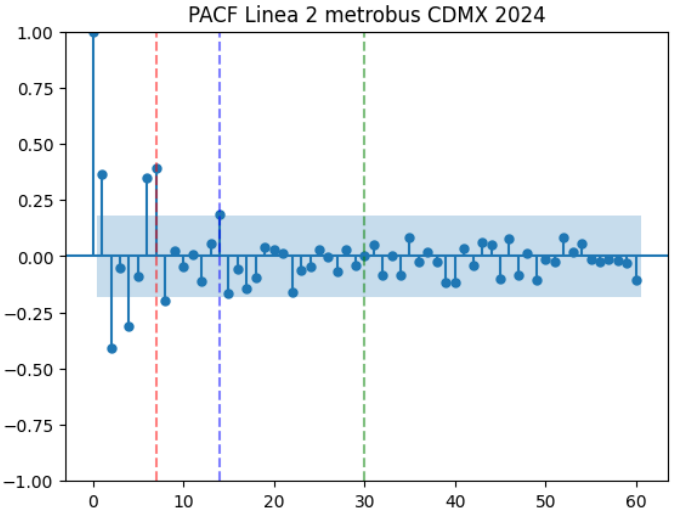

Figura 11: PACF para la línea 2 del metrobús de la CDMX (2024)

4.5. División de los datos

El conjunto de datos se dividió de la siguiente forma:

- **Entrenamiento:** 15 de febrero – 9 de mayo
- **Validación:** 10 – 19 de mayo
- **Prueba:** 20 – 31 de mayo

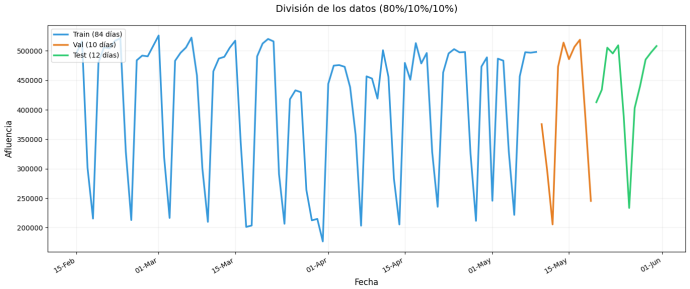

Figura 12: :División temporal de los datos para el entrenamiento del modelo

4.6. Búsqueda de hiperparámetros

Se usó `RandomizedSearchCV` para encontrar los mejores hiperparámetros y se eligieron los siguientes:

Línea 1

- **Combinación**: [lag7_linea1, lag14_linea1, ma7_linea1]
- **RMSE en validación**: 50,340.94

Línea 2

- **Combinación**:[lag7_linea2, lag14_linea2,ma14_linea2]
- **RMSE en validación**: 33,645.65

Parámetro	Línea 1	Línea 2
subsample	0.8	0.7
reg_lambda	3	4
reg_alpha	2	2
n_estimators	200	200
min_child_weight	5	5
max_depth	2	3
learning_rate	0.015	0.015
gamma	0.2	0.2
colsample_bytree	0.7	0.8

Tabla 2. Hiperparámetros del modelo para las líneas 1 (Indios Verdes-El Caminero) y 2 (Tacubaya-Tepalcates) del Metrobús CDMX.

4.7. Variables utilizadas en el modelo

Las **features** incluidas en el modelo final fueron:

Tipo de variable	Ejemplos
Temporal	mes, año, día_semana, día_mes
Climatológica	temp_aire_celsius, precipitación, humedad_relativa
Categórica	tipo_día_Entre Semana, tipo_día_Feriado
Series temporales	lag7_línea1, lag14_línea1, ma7_línea1

Tabla 3. Variables utilizadas en el modelo predictivo para la línea 1 del metrobús de la CDMX.

Nota: Para el modelo de la línea 2, las variables de entrada son similares pero ajustadas a esa línea, lag7_linea2 ,lag14_linea2 ,ma14_linea2.

4.8. Modelado de datos

El modelo implementado para las predicciones fue **XGBoost** (Extreme Gradient Boosting), un algoritmo basado en ensambles de árboles de decisión. Su funcionamiento consiste en construir secuencialmente árboles donde cada nuevo árbol corrige los errores cometidos por el anterior. La predicción final se obtiene como una combinación ponderada de los árboles generados.

XGBoost fue seleccionado por su rendimiento comprobado en tareas de predicción en series temporales, así como por su capacidad para capturar relaciones no lineales entre variables, siempre y cuando se realice un adecuado ajuste de hiperparámetros.

Un estudio realizado por **Ariyo et al. (2014)** utilizó XGBoost para predecir precios bursátiles en Estados Unidos, obteniendo una eficacia del 87 % en la predicción de precios de acciones. Este rendimiento lo posiciona como una alternativa confiable frente a otros algoritmos de aprendizaje automático.

Además, **Liang Zou et al. (2022)** utilizaron XGBoost para predecir el flujo de pasajeros en un sistema de autobuses inteligentes. Probaron distintos modelos:

- Modelo 1: XGBoost con variables como número de autobuses, flujo de pasajeros previos, día de la semana, etc.
- Modelo 2: XGBoost sin la variable de número de autobuses.
- Modelo 3: Regresión KNN.
- Modelo 4: Red neuronal Bp.
- Modelo 5: Red LSTM.

Los modelos fueron evaluados mediante las métricas **MAPE**, **RMSE** y **MAE**, así como con criterios de eficiencia computacional. El primer modelo con XGBoost resultó el más preciso en la mayoría de estaciones analizadas, demostrando su superioridad en la predicción del flujo de pasajeros en sistemas de transporte.

4.9. Generación de lags y medias móviles para 2025

Para generar las variables de entrada necesarias para el modelo (como los lags y medias móviles), se tomaron los valores reales de afluencia del 18 al 31 de enero de 2025 y se aplicó una disminución artificial del 2 % diario para simular comportamiento futuro. A partir de estos valores sintéticos se generaron los primeros 14 lags y medias móviles (`ma7`, `ma14`) requeridos por los modelos entrenados.

5. Resultados

Se generaron gráficos de líneas para visualizar la afluencia diaria predicha durante febrero a mayo de 2025. Se observó que la línea 1 presenta mayores valores, oscilando entre 250,000 y 500,000 pasajeros diarios, mientras que en la línea 2 se mantuvieron entre 150,000 y 240,000.

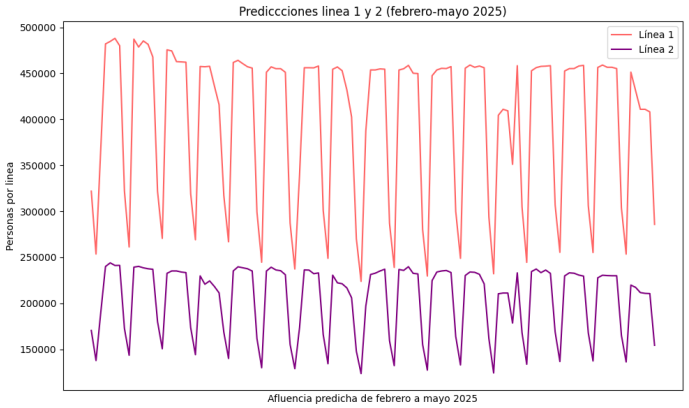

Figura 13: Afluencia predicha febrero-mayo 2025 para líneas 1 y 2 del Metrobús.

Por mes visualizamos un aumento gradual para ambas líneas, a excepción para el mes de Abril para la línea 1,la cual disminuye de 12,017,597 para el mes de marzo, a 11,973,452 en el mes siguiente, mantienen rangos desde los 11 a los 12 millones para la línea 1 y para la otra línea se mantiene entre los 5 a 6 millones de personas que abordan el metrobus.

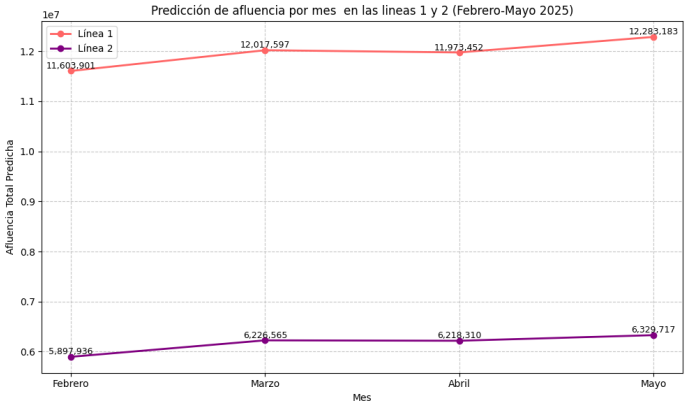

Figura 14: Afluencia de pasajeros en la línea 1 y 2 del metrobús de la CDMX por mes, de febrero a mayo 2025

En relación a los estadísticos descriptivos, para la Línea 1, se observa una media de 398,984 personas, mientras que su mediana alcanza los 453,625, lo que indica una diferencia significativa entre ambos valores. Esta diferencia indica la presencia de días con una afluencia baja, los cuales terminan reduciendo el promedio general.Adicionalmente la desviación estándar de 83,897 confirma una variación considerable en el número de personas, reflejando una alta dispersión en los datos diarios.

Por otro lado, la Línea 2 presenta una media de 205,604 y una mediana de 229,834, valores cercanos al tercer cuartil (235,111), lo que indica una distribución con cierto sesgo hacia las afluencia más altas, pero con una variabilidad menor en comparación con la Línea 1. Cuenta con una desviación estándar de 38,415, al igual que en la otra línea, existen días con una afluencia baja, lo que afecta la media en general.

Estadístico	Línea 1	Línea 2
Número de registros (count)	120	120
Media de pasajeros (mean)	398,984	205,604
Desviación estándar (std)	83,897	38,415
Mínimo de pasajeros (min)	223,882	123,662
Percentil 25 %	314,610	168,800
Mediana (50 %)	453,625	229,834
Percentil 75 %	457,329	235,111
Máximo de pasajeros (max)	488,018	244,015

Tabla 4. Estadísticos descriptivos para la afluencia predicha en las líneas 1 (Indios Verdes-El Caminero) y 2 (Tacubaya-Tepalcates). Período: feb.-mayo 2025.

Analizando por dia de la semana la linea 1 tiene valores cercanos a los 455,000 de lunes a viernes, con menor afluencia los fines de semana, especialmente domingos (250,000). Mayor variabilidad en fines de semana.

La linea 2 cuenta con una mediana cercana a 230,000 entre semana, con outliers en jueves y domingos. Disminución significativa el domingo (mediana 135,000).

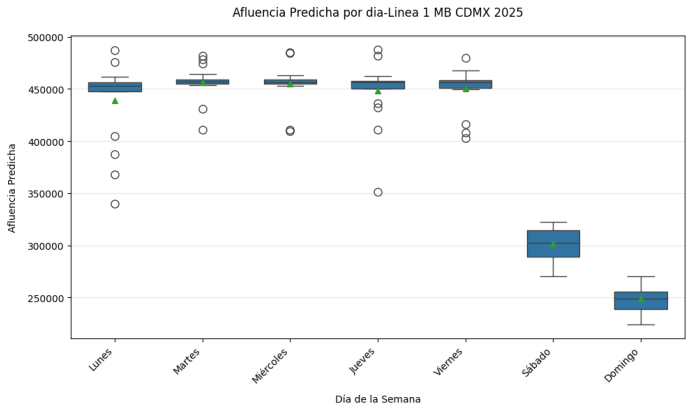

Figura 15: Diagrama de caja y bigotes de la afluencia predicha para la línea 2 del metrobús CDMX (febrero- mayo 2025)

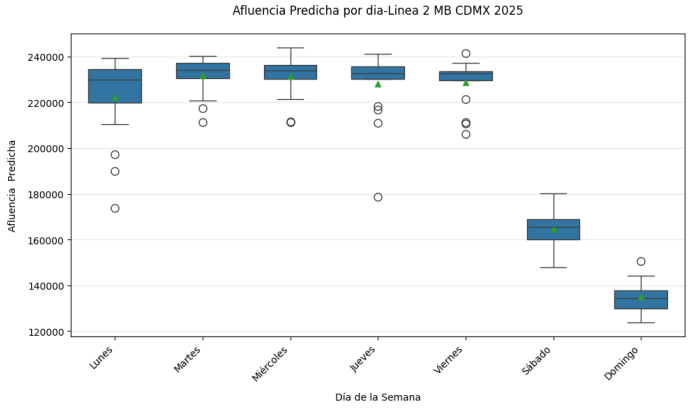

Figura 16: Diagrama de caja y bigotes de la afluencia predicha para la línea 2 del metrobús CDMX (febrero- mayo 2025)

A partir de lo anterior podemos concluir que el modelo muestra un comportamiento estable, con valores situados dentro de un rango estrecho con una variabilidad baja en las predicciones. Esto sugiere que el modelo un ajuste óptimo para los patrones regulares en la demanda

Sin embargo, esta misma estabilidad podría indicar una limitación en la captura de patrones atípicos, en específico para los fines de semana. La ausencia de outliers en la mayoría de esos días sugiere que el modelo podría estar subestimando comportamientos atípicos(como días festivos) esto es consistente con los hiperparámetros establecidos,con una profundidad máxima (max_depth) 2 y 3 que si bien genera predicciones estables, limita la capacidad del modelo para identificar desviaciones significativa en la afluencia

6. Propuesta de solución

A partir de las predicciones obtenidas es posible plantear estrategias para el mejoramiento del servicio para ambas líneas del metrobús, la demanda es mayor en ambas líneas durante los días laborales, especialmente en la Línea 1, donde la afluencia diaria es mayor a los 450,000. Debido a esto una de las acciones más efectivas sería aumentar la cantidad de metrobuses en circulación a lo largo de estos días, esto ayudaría a evitar la saturación y por tanto a reducir los tiempos de espera para los usuarios mejorando su experiencia en este STC.

Para la Línea 2, aunque la afluencia es menor, también es recomendable aumenta la frecuencia de los metrobuses, ya que la mediana diaria se mantiene por encima de los 220,000 en caso de que el presupuesto lo permita, sería conveniente probar un piloto de un mes aumentando en un 20 por ciento la frecuencia de los metrobuses, y a partir de ahí generar los ajustes pertinentes.

Por otro lado, los fines de semana muestran una disminución en la cantidad de pasajeros, en especial los domingos.Debido a esto sería útil reducir la cantidad de metrobuses en circulación los sábados y domingos, seria conveniente ajustar la frecuencia del servicio para evitar el uso innecesario de recursos y optimizar los costos operativos.

7. Evaluación

La evaluación del modelo se realizó mediante las métricas RMSE (Error Cuadrático Medio), MAE (Error Absoluto Medio) y MAPE (Error Porcentual Absoluto Medio), La evaluación se realizó en dos etapas:

- **Primera fase (2024):** Usando los datos reales de prueba del 20 al 31 de mayo de 2024.
- **Segunda fase (2025):** Comparando las predicciones del modelo con los datos reales disponibles de febrero a abril de 2025.

	2024		2025	
Métrica	**Línea 1**	**Línea 2**	**Línea 1**	**Línea 2**
RMSE	49,624	23,082	69,037	31,753
MAE	43,839	19,122	54,585	20,691
MAPE	10.09 %	9.62 %	15.10 %	15.36 %

Tabla 5. Métricas de evaluación para la Línea 1 y 2 en 2024 y 2025.

Los resultados de las métricas señalan un incremento en el error porcentual (MAPE) de 10 a 5 por ciento aproximadamente, empeorando un 5 por ciento aproximadamente,lo que más resalta en las métricas es el RSME lo que sugiere la presencia de errores grandes de 2024 a 2025 por lo anterior se realizaron gráficas para tener una idea más clara de los errores presentes en el año 2025.

Para ambas líneas observamos caídas en la afluencia los fines de semana, en relación a la línea 1 y como se mencionó en los resultados el modelo cuenta con un desempeño más bajo en presencia de valores extremos teniendo un peor desempeño a la predicción de valores más altos,es notable el error de predicción a partir del día 15 de abril, lo que posiblemente aumentó las métricas

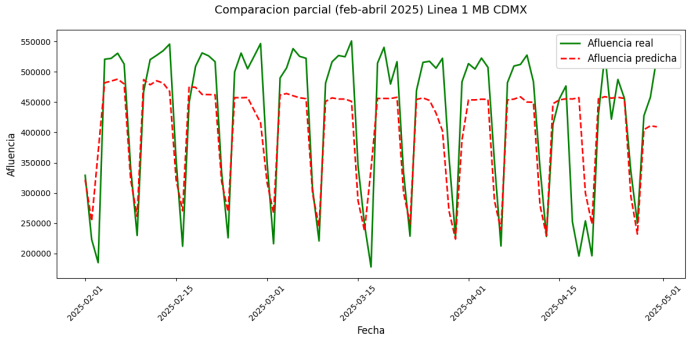

Figura 17: Comparación parcial entre valor predicho vs real línea 1 MB CDMX 2025

En el caso de la línea 2 vemos que predice de manera más precisa los picos altos de afluencia sin embargo para los valores inferiores(como lo son los fines de semana), predice valores más altos de lo que realmente son, analizando más la gráfica vemos que la predicción se aleja bastante de la realidad a partir del 15 de abril, de manera más extrema que en el de la línea anterior.

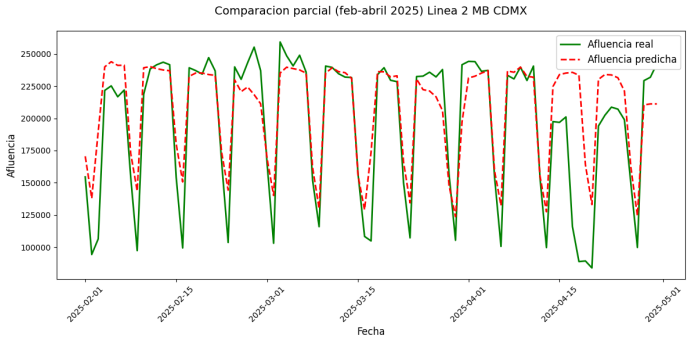

Figura 18: Comparación parcial entre valor predicho vs real línea 2 MB CDMX 2025

Uno de los desafíos más importantes del modelado fue que para la generación de las features como los lags y las MA era necesario la afluencia, lo cual es lo que se buscaba predecir, la forma de solucionarlo, fue tomar valores existentes del 18 al 31 de enero 2025, crear los primeros 14 lags como base inicial, lo que permitió generar valores sintéticos de afluencia con una disminución diaria del 2 por ciento

8. Conclusión

Se desarrollaron dos modelos predictivos para estimar la afluencia diaria en las líneas 1 y 2 del Metrobús de la CDMX utilizando el algoritmo XGBoost. Los modelos fueron entrenados con datos históricos de febrero a mayo de 2024 y aplicados para predecir los mismos meses del año 2025.

- Se identificaron patrones semanales claros, con mayor afluencia de lunes a viernes y disminuciones durante los fines de semana.
- La línea 1 presentó una media de afluencia mayor (aproximadamente 410,000 pasajeros), mientras que la línea 2 tuvo una media de 216,000.
- Las predicciones para 2025 oscilaron entre 250,000–500,000 para la línea 1 y 150,000–240,000 para la línea 2, con una tendencia mensual creciente.
- La evaluación del modelo se realizó mediante el Error Porcentual Absoluto Medio (MAPE) mostró un incremento del 5 % aproximadamente en 2025 respecto al año anterior
- El modelo mostró estabilidad en sus predicciones, pero con dificultad para capturar eventos atípicos como feriados no programados o alteraciones excepcionales.

8.1. Propuestas de mejora

Con base en los resultados y observaciones obtenidas, se proponen las siguientes mejoras al modelo:

1. **Incluir variables socioeconómicas y demográficas:** como población flotante, zonas comerciales o actividades laborales por zona, que podrían mejorar la precisión del modelo.
2. **Optimización de features:** realizar selección de variables basadas en su importancia relativa para reducir ruido y sobreajuste.
3. **Mejorar la generación de lags y medias móviles:** utilizar simulaciones más robustas o técnicas de predicción en cascada para evitar errores acumulativos al generar valores sintéticos.
4. **Reestructurar el flujo de trabajo del modelo:** usar funciones organizadas o pipelines de procesamiento para mejorar la mantenibilidad y reproducibilidad del código.

5. **Considerar modelos multivariados:** como XGBoost multisalida o redes neuronales multivariadas que permitan modelar ambas líneas simultáneamente.

6. **Ampliar horizonte de predicción:** aplicar el modelo para todo el año y a otras líneas del Metrobús e incluso del Metro para lograr una solución generalizable.

9. Referencias

- Anu Priya et al. (2024). Optimización de las operaciones del metro de Delhi con XGBoost.
- Gobierno de la Ciudad de México (2025). Afluencia diaria de Metrobús CDMX. Datos Abiertos CDMX.
- INEGI (2020). Censo de Población y Vivienda 2020.
- Ortiz, A. E. (n.d.). Optimización de rutas en bases de datos espaciales.
- Sistema de Transporte Público Metrobús (s.f.). Fichas técnicas.
- Sroka, Ł. (2024). Simulation analysis of artificial neural network and XGBoost algorithms.
- Zou, L. et al. (2022). Passenger flow prediction using smart card data and XGBoost.

10. Anexo

Repositorio de Github donde se encuentra el código