Machine Learning aplicado a la estimación de la demanda de almuerzos Unal 2023-2

Por Edwar Valenzuela Cortés

Introducción

Determinar las cantidades de comida dentro de cualquier restaurante impactan de forma significativa la rentabilidad y la operación dentro de los mismos. Existen distintas maneras de estimar estas cantidades, y una de ellas es a través del uso de los datos existentes para poder hacer estimaciones a futuro. Es por esto que este informe buscará resaltar el uso de Machine Learning para la toma de decisiones alrededor de la demanda de almuerzos dentro de la Universidad Nacional de Colombia, atravesando por los pasos claves del proyecto como la **preparación de los datos, el análisis de los mismos, la elaboración y evaluación del modelo predictivo así como las consideraciones finales respecto al resultado del modelo,** con el objetivo de que sea de utilidad a la hora de gestionar efectivamente las cantidades dentro de los restaurantes en el futuro, empezando con el siguiente semestre 2023-2.

Preparación de los datos

Inicialmente se cuenta con los registros del semestre 2022-1 que de entrada ya suponen un problema: las fechas están incorrectas ya que empiezan en el semestre 2021-1. El archivo posee un registro de 335182 transacciones de algún tipo de alimento, principalmente almuerzos, por lo que para solucionar esto se agrupa las cantidades por almuerzo por cada fecha y según el restaurante:

	FECHA	CANTIDAD	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
0	2021-03-26	3073	209.0	92.0	712.0	13.0	0.0	196.0	256.0	762.0	139.0	160.0	228.0	314.0
1	2021-03-27	3498	229.0	99.0	770.0	18.0	0.0	242.0	281.0	841.0	201.0	153.0	273.0	404.0
2	2021-03-28	3441	244.0	106.0	742.0	22.0	0.0	231.0	262.0	824.0	217.0	174.0	272.0	352.0
3	2021-03-29	3582	235.0	110.0	852.0	7.0	0.0	264.0	281.0	920.0	216.0	192.0	140.0	373.0
4	2021-03-30	2956	194.0	104.0	692.0	20.0	0.0	185.0	231.0	738.0	131.0	141.0	222.0	305.0

Al hacer esto, los datos pasan a ser de 101 filas, por lo que resulta más fácil exportar los datos a Excel para continuar con la preparación de los datos. Inicialmente, se adecuan los datos al calendario del semestre 2022-1 así como también se rellenan los valores faltantes como festivos y domingos con un valor de cero. Por otro lado, para la **corrección de los valores atípicos** como lo son los días festivos o los días donde ocurrió un evento especial (como lo son los tropeles) se usa **un promedio móvil** de 3 periodos para el reemplazo de estos valores, usando claro está, los 3 valores anteriores del día de la semana a ese valor atípico. Es decir, si un lunes era festivo, se usa el promedio de los 3 lunes anteriores para estimar el valor, **esto con el objetivo de evitar resultados sesgados en el modelo que se va a plantear.**

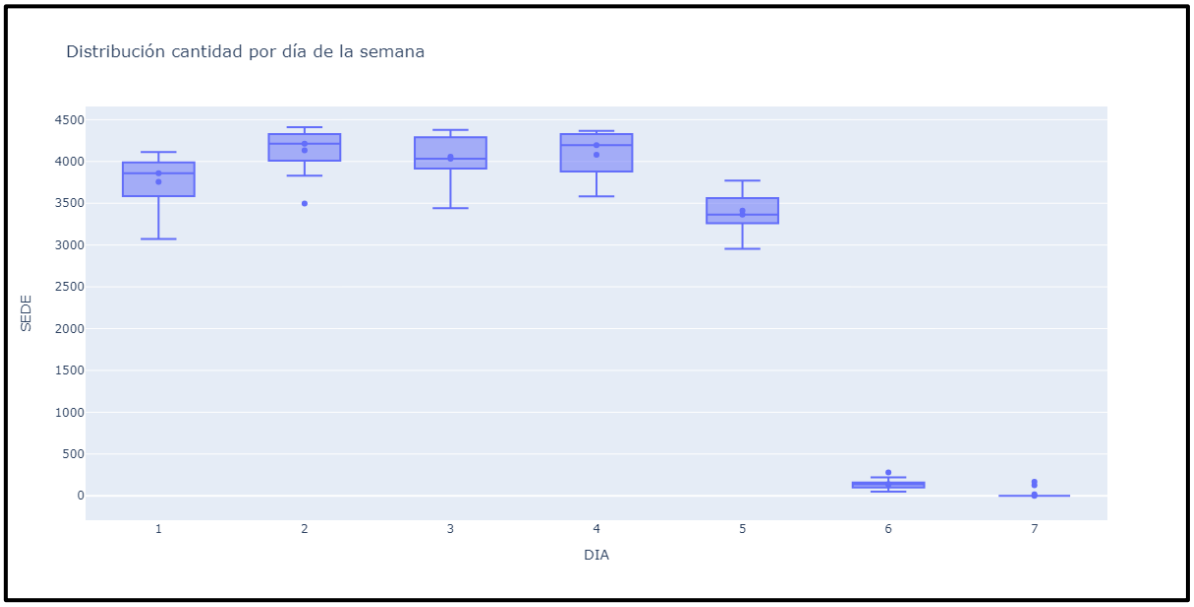
Es gracias a lo anterior, que se puede obtener finalmente los datos necesarios para la elaboración del análisis y el modelo:

FECHA	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
2022-02-28	3073.000000	209.000000	92.0	712.000000	13.000000	0	196.000000	256.000000	762.000000	139.0	160.000000	228.000000	314.000000
2022-03-01	3498.000000	229.000000	99.0	770.000000	18.000000	0	242.000000	281.000000	841.000000	201.0	153.000000	273.000000	404.000000
2022-03-02	3441.000000	244.000000	106.0	742.000000	22.000000	0	231.000000	262.000000	824.000000	217.0	174.000000	272.000000	352.000000
2022-03-03	3582.000000	235.000000	110.0	852.000000	7.000000	0	264.000000	281.000000	920.000000	216.0	192.000000	140.000000	373.000000
2022-03-04	2956.000000	194.000000	104.0	692.000000	20.000000	0	185.000000	231.000000	738.000000	131.0	141.000000	222.000000	305.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2022-06-23	3702.000000	222.000000	118.0	644.000000	26.000000	0	227.000000	309.000000	895.000000	136.0	470.000000	272.000000	388.000000
2022-06-24	3266.000000	195.000000	122.0	642.000000	26.000000	0	199.000000	245.000000	817.000000	118.0	353.000000	219.000000	337.000000
2022-06-25	220.000000	0.000000	0.0	0.000000	0.000000	147	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
2022-06-26	127.000000	0.000000	0.0	0.000000	0.000000	127	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
2022-06-27	3868.333333	242.604938	120.0	748.185185	33.419753	0	241.382716	281.580247	936.382716	151.0	416.851852	284.888889	421.185185

120 rows × 13 columns

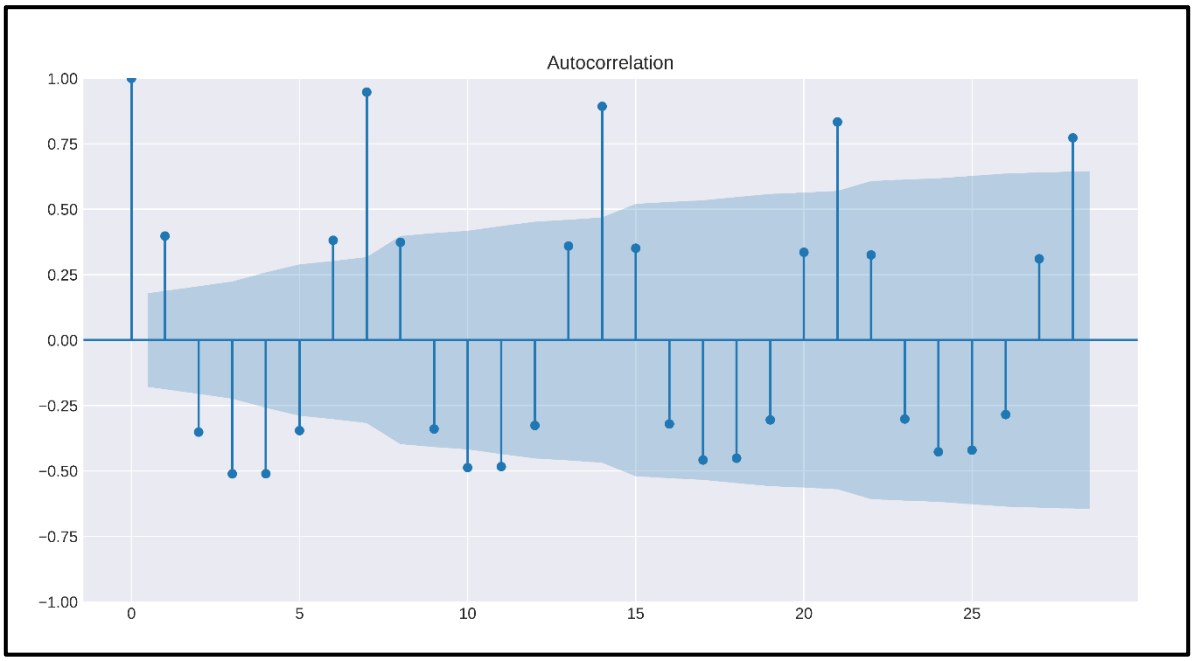
Análisis de los datos

1. Diagrama de cajas: En el siguiente gráfico se puede observar la distribución de los datos según el día de la semana, siendo lunes “1”, martes “2”, miércoles “3” hasta llegar al domingo “7”. Se puede inferir **que los martes, miércoles y jueves son los días en los que ocurre una mayor demanda en comparación a los lunes y viernes.** Por otro lado, **los sábados suele haber una menor demanda y los valores del domingo son atípicos** ya que solo hubo demanda de almuerzos gracias a eventos especiales al final del semestre, sin embargo, estos valores atípicos no afectarán el modelo final.

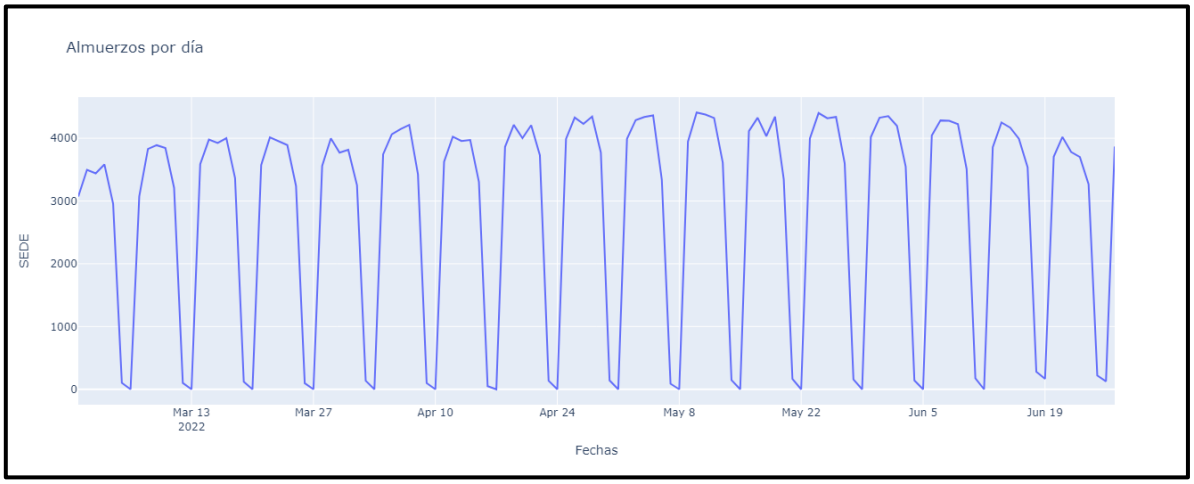


2. Gráfico de autocorrelación: Con el gráfico de autocorrelación se puede observar mejor las variables predictoras que permiten una mejor elaboración de un modelo. En este caso, **los múltiplos**

de 7 tienen una correlación alta en comparación a los otros valores. Es decir, podemos hacer una estimación de las cantidades usando 7,14,21,28, etc. días atrás para estimar la cantidad futura de un día.

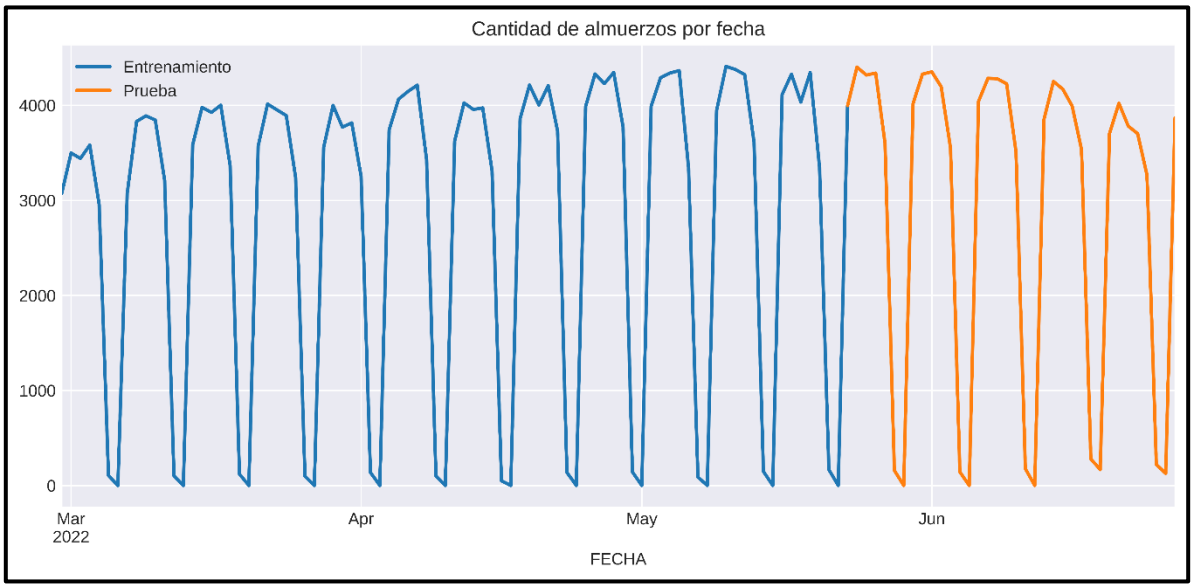


3. Gráfico general de los datos: Estas son las cantidades por fecha a lo largo del semestre, después de haber corregido los valores atípicos.



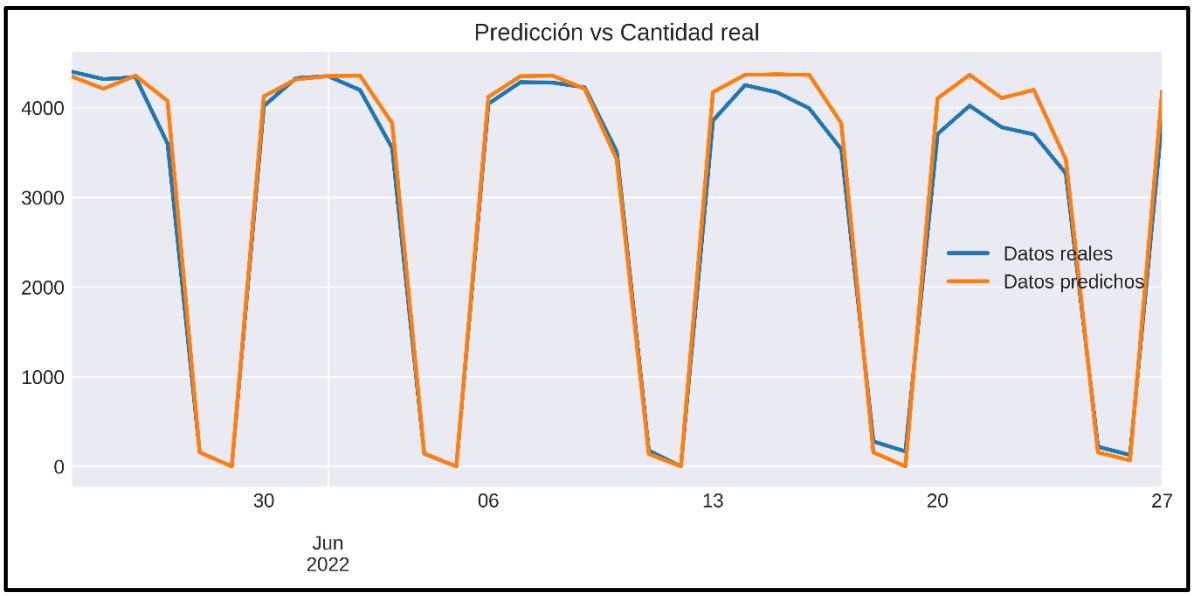
Elaboración y evaluación del modelo

1. Conjunto de entrenamiento y prueba: Para la evaluación del modelo, se usará **un conjunto de datos de entrenamiento y otro de prueba.** Haciendo uso de las prácticas base de las ciencias de datos, **se tomará como entrenamiento el 70% de la cantidad de los datos y de prueba el 30% de los datos** (2022-02-28 hasta 2022-05-23 como entrenamiento y 2022-05-23 hasta 2022-06-27 como prueba), como se puede ver en el gráfico.



2. El modelo y su evaluación: Para la elaboración del modelo se usó el regresor **XGBoost**, el cual es un algoritmo de aprendizaje automático basado en árboles de decisión, específicamente diseñado para problemas de regresión. Esto a través del algoritmo Gradient Boosting para construir un conjunto de árboles de decisión de forma secuencial, **donde cada árbol se enfoca en corregir los errores del modelo anterior** (ArcGIS Pro 3.1). Usando las variables predictoras de **28, 35 y 7** (es decir, los datos de hace 28, 35 y 7 días), se logra construir el modelo a través de la función **ForecasterAutoreg**. Habiendo hecho lo anterior y haciendo uso de bucles, se logra construir **no solo un modelo a nivel de sede, si no también para cada uno de los restaurantes**.

En el siguiente gráfico podemos ver un ejemplo de la efectividad del modelo haciendo uso de la función **backtesting_forecaster,** realizando predicciones de **7 días**, tomando como referencia los datos **a nivel de sede** y comparando los **datos reales** con los **datos predichos**, **habiendo hecho uso solamente de las primeras 12 semanas para proyectar las restantes:**



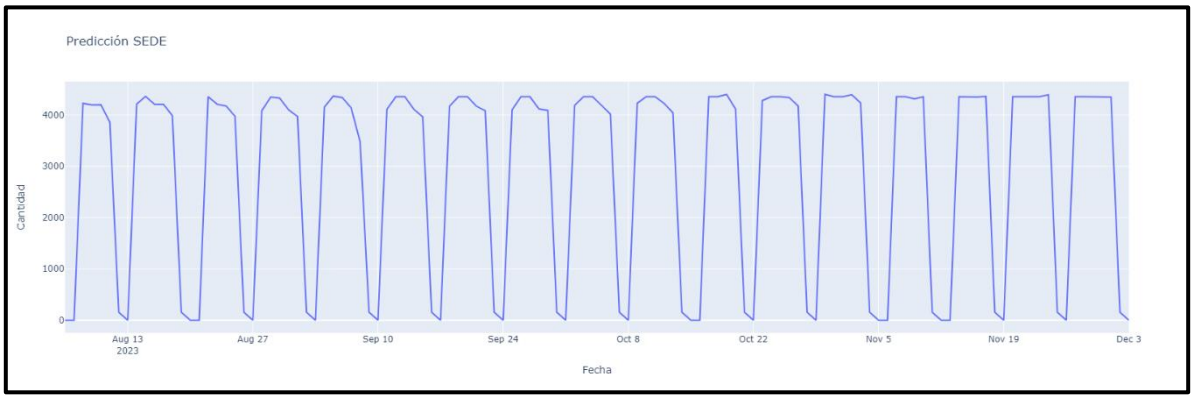
Obteniendo como **error absoluto 152 almuerzos** y un **error porcentual de 5.27% a nivel de sede,** habiendo hecho uso de la métrica que la misma función de backtesting nos ofrece. Esto se obtiene para cada uno de los restaurantes:

	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
Error absoluto	152.0	15.0	3.0	45.0	7.0	21.0	16.0	11.0	40.0	18.0	17.0	22.0	23.0

	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
Error porcentual	5.27	8.45	2.87	7.99	32.95	100.0	9.0	5.05	5.93	14.37	5.71	10.01	7.39

Estimación del semestre 2023-2

1. Estimación: Para elaborar la estimación final, se hace uso de la función **predict**, se asigna como fecha inicial el **2023-08-06**, se hace el conteo necesario (120 días) para ajustar los datos al rango del semestre (hasta **2023-12-02**) y se **reemplaza** los valores de los días **festivos y domingos por cero**, permitiendo esto obtener una estimación aún más precisa. El resultado **(tomando como ejemplo el modelo de sede)** de la estimación se puede en el siguiente gráfico:



Así mismo, se obtienen las cantidades estimadas totales para cada uno de los restaurantes para el semestre 2023-02:

	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
Cantidades estimadas	342461.0	21263.0	9782.0	63495.0	1666.0	0.0	22032.0	24098.0	73910.0	13414.0	32502.0	25584.0	36382.0

2. Resultado final: Todo lo mostrado a lo largo del documento se aplica para cada uno de los restaurantes, permitiendo obtener como **resultado final un marco de datos convertible en archivo Excel con las estimaciones para cada uno de los restaurantes para el semestre 2023-2**, como se puede ver en la siguiente imagen:

	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS	FESTIVO	DIA
2023-08-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	False	7
2023-08-07	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	True	1
2023-08-08	4226.0	297.0	119.0	794.0	18.0	0.0	284.0	299.0	912.0	192.0	438.0	334.0	453.0	False	2
2023-08-09	4198.0	297.0	119.0	794.0	13.0	0.0	252.0	290.0	892.0	191.0	434.0	333.0	455.0	False	3
2023-08-10	4200.0	263.0	119.0	776.0	15.0	0.0	254.0	319.0	895.0	177.0	434.0	333.0	457.0	False	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2023-11-29	4357.0	271.0	119.0	835.0	20.0	0.0	285.0	296.0	946.0	158.0	396.0	309.0	453.0	False	3
2023-11-30	4357.0	267.0	119.0	815.0	22.0	0.0	281.0	325.0	895.0	184.0	393.0	328.0	457.0	False	4
2023-12-01	4351.0	273.0	125.0	630.0	20.0	0.0	270.0	297.0	912.0	131.0	396.0	359.0	453.0	False	5
2023-12-02	157.0	0.0	0.0	138.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	False	6
2023-12-03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	False	7

120 rows × 15 columns

```
# Descargar archivo excel con los valores estimados del semestre:
datos_finales.to_excel('Cantidades estimadas del segundo semestre de 2023.xlsx', index=True)
```

Conclusiones y consideraciones finales

1. La semana de receso se está tomando como una **semana normal**

2. Las estimaciones se pueden ver afectadas por el impacto de las cantidades REALES que haya en el siguiente semestre del restaurante **LIBRERIA** (notar que en el **archivo excel** este fue el único restaurante que no se debe tener en cuenta). Por lo tanto, **las estimaciones más confiables siguen siendo las de nivel de sede**, a no ser que haya una mejor proporción de datos para el restaurante LIBRERÍA **o este restaurante no tenga un impacto significativo** en el siguiente semestre

3. Los datos pueden funcionar mejor para **una estimación aproximada del presupuesto** requerido para **sede o cada uno de los restaurantes, más no para estimar las cantidades a elaborar a lo largo del total del semestre**, **ya que las estimaciones se pueden ver sesgadas debido a la longitud de la predicción,** pues se pueden sobreestimar las cantidades a lo largo del semestre.

4. Se podrían plantear o probar otros predictores, pero se recomienda evitarlo **debido a la posibilidad de sobre ajuste**.

5. Algunas estimaciones como las de los restaurantes **FLOR DE LOTO, PARQUE INFANTIL o PLAYA ROJA,** pueden estar sesgadas gracias al nivel de error proporcionado por el modelo y también por la distribución de sus datos.

6. Sería útil probar el modelo con datos como los del semestre 2022-2 o 2023-1 para comprobar la efectividad aún más del mismo. Se podría usar el nivel de error **5.27% o los de cada restaurante** para estimar aún mejor el nivel de sesgo probable de las estimaciones.

7. Las **cantidades totales estimadas** por el modelo se asemejan a las **cantidades reales de los datos procesados (del 2022-2)**, cuestión que muestra en cierta medida la efectividad del modelo, como se puede ver en las siguientes imágenes:

	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
Cantidades reales	336869.0	21383.0	10151.0	69307.0	2024.0	720	21813.0	25072.0	78053.0	16267.0	30993.0	25462.0	36289.0

	SEDE	CAMPAMENTO	CAMPANARIO	CLINICAS	FLOR DE LOTO	LIBRERIA	MARIPOSARIO	MIRADOR	PALMERAS	PARQUE INFANTIL	PARQUE LUNA	PLAYA ROJA	VENTANAS
Cantidades estimadas	342461.0	21263.0	9782.0	63495.0	1666.0	0.0	22032.0	24098.0	73910.0	13414.0	32502.0	25584.0	36382.0

Bibliografía:

-ArcGIS Pro 3.1. Cómo funciona el algoritmo XGBoost. Obtenido de:

https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm

-El modelo de machine learning se elaboró en base a un proyecto de forecasting para la serie temporal de la demanda eléctrica (MW) del estado de Victoria (Australia):

Predicción (forecasting) de la demanda eléctrica con Python by Joaquín Amat Rodrigo and Javier Escobar Ortiz, available under a Attribution 4.0 International (CC BY 4.0) at **https://www.cienciadedatos.net/documentos/py29-forecasting-demanda-energia-electrica-python.html**