TAREA FINAL: APLICACIÓN DE LAS SERIES DE TIEMPO

Presentado por: Juan Bejarano T., Margarita De Alba, Nathalia Moreno y Jorge Rodriguez D.
Presentado a: Jonathan Acosta

Pontificia Universidad Católica de Chile
Magister en Ciencia de Datos
MCD - Series de tiempo
2023

TABLA DE CONTENIDO

CONTEXTUALIZACIÓN

Actualmente las entidades financieras se enfrentan al desafío de gestionar eficientemente la afluencia de clientes en sus sucursales físicas para llevar a cabo operaciones relacionadas con sus cuentas y carteras financieras. El inconveniente surge cuando la capacidad operativa de estas sucursales no logra satisfacer la creciente demanda de visitas. En este caso, una empresa del sector financiero (cuyo nombre se omite por confidencialidad y de ahora en adelante será referida como la *empresa*) busca anticiparse a la proyección de clientes que se espera acudan presencialmente a su sucursal en los próximos meses.

Para abordar esta problemática, la empresa ha seleccionado estratégicamente la sucursal CAMINO REAL como una sucursal clave, esto es debido a su posición central en la ciudad de Medellín y ser la oficina más solicitada a nivel nacional. Urge una intervención inmediata para adecuar la capacidad de atención a la demanda prevista para el próximo año. La situación actual se caracteriza por demoras en la atención, extensas filas y una carencia de personal suficiente. Por ende, la empresa busca anticipar la demanda futura como elemento esencial para una distribución eficaz de su personal y llevar a cabo un análisis anticipado de colas, con el objetivo de responder de manera efectiva a las necesidades operativas y a la estrategia del negocio.

La relevancia de contar con predicciones precisas en este contexto es fundamental para que la empresa pueda tomar decisiones informadas y proactivas. Al anticipar la demanda futura de clientes en la sucursal CAMINO REAL, la empresa puede ajustar estratégicamente su capacidad operativa, asignar recursos de manera eficiente y mejorar la experiencia del cliente. Contar con predicciones exactas no solo permite evitar demoras en la atención y reducir las filas, sino que también contribuye a optimizar la asignación de personal, garantizando que haya suficientes empleados disponibles para atender la demanda proyectada. Esto no solo beneficia la eficacia operativa, sino que también impacta positivamente en la percepción de los clientes, fortaleciendo la reputación de la empresa y su posición en el mercado financiero. En resumen, la precisión en las predicciones se convierte en un elemento estratégico para la toma de decisiones y el éxito a largo plazo de la entidad financiera.

Descripción del dataset

Afortunadamente, se cuenta con el registro de la cantidad de arribos diarios a la sucursal CAMINO REAL, desde septiembre de 2020 a agosto de 2023, almacenados

en el archivo *Arribos_oficina_atencion_2.xlsx* (de carácter **confidencial)**. A continuación se presenta una descripción de las variables presentes en el conjunto de datos mencionado:

- **Oficina:** variable que almacena el nombre de la oficina analizada, que en este caso particular corresponde a CAMINO REAL. Es una variable tipo texto (*character*)
- **Año (*year)*:** variable que almacena el año donde se registran los arribos que se desean analizar en la oficina. En el dataset se encuentran cuatro años en el dataset, desde 2020 hasta 2023. Es una variable numérica (*int*).
- **Mes:** variable que almacena el número del mes donde se registran arribos de personas a la oficina, Este campo está en formato numérico (*int*), un número entero entre 1 (Enero) y 9 (Agosto).
- **Fecha:** variable que indica la fecha exacta en donde se presentaron los arribos. Este campo está en formato de fecha de Microsoft Excel.
- **Arribos:** variable que almacena el número de personas que asistieron en el día indicado a esa sucursal. Es una variable numérica (*int*).

Dentro del dataset consideramos como variable objetivo *Arribos* ya que al adoptar un enfoque de series de tiempo en el contexto dado, se busca comprender la evolución de dicha variable a lo largo de los días, identificando patrones temporales que permitan prever la demanda futura. Al predecir los arribos, la empresa financiera puede tomar decisiones operativas informadas, como la asignación eficiente de personal, para mejorar la capacidad de atención y la experiencia del cliente en la sucursal.

OBJETIVOS

Objetivo general:

Implementación de un modelo óptimo de predicción de afluencia basado en series de tiempo para anticipar la demanda de clientes presenciales para los días hábiles del mes de septiembre de 2023 en la oficina CAMINO REAL.

Objetivos específicos:

1. Desarrollar modelos predictivos de series de tiempo que permitan estimar el arribo de clientes en la oficina CAMINO REAL, considerando patrones históricos y factores estacionales.
2. Evaluar los modelos de predicción usando una medida de desempeño compatible para determinar el mejor modelo posible.
3. Establecer recomendaciones prácticas basadas en las predicciones, que puedan ser utilizadas por la empresa para ajustar estratégicamente la capacidad operativa y la asignación de recursos en periodos de alta demanda.
4. Documentar y comunicar los resultados del modelo predictivo, junto con las recomendaciones, de manera clara y comprensible para que la empresa pueda tomar decisiones informadas sin requerir una integración inmediata con sistemas operativos.

RESULTADOS

Análisis exploratorio

La base de datos *Arribos_oficina_atencion_2.xlsx* contiene 731 observaciones con 5 variables, al cargar el dataset usando pandas obtenemos la siguiente vista:

Imagen 1. Vista del dataset

	oficina	year	mes	fecha	arribos
0	CAMINO REAL	2020	9	2020-09-01	190
1	CAMINO REAL	2020	9	2020-09-02	186
2	CAMINO REAL	2020	9	2020-09-03	192
3	CAMINO REAL	2020	9	2020-09-04	222
4	CAMINO REAL	2020	9	2020-09-07	237
...	...	...	...	...	...
726	CAMINO REAL	2023	8	2023-08-25	230
727	CAMINO REAL	2023	8	2023-08-28	171
728	CAMINO REAL	2023	8	2023-08-29	199
729	CAMINO REAL	2023	8	2023-08-30	161
730	CAMINO REAL	2023	8	2023-08-31	185

731 rows × 5 columns

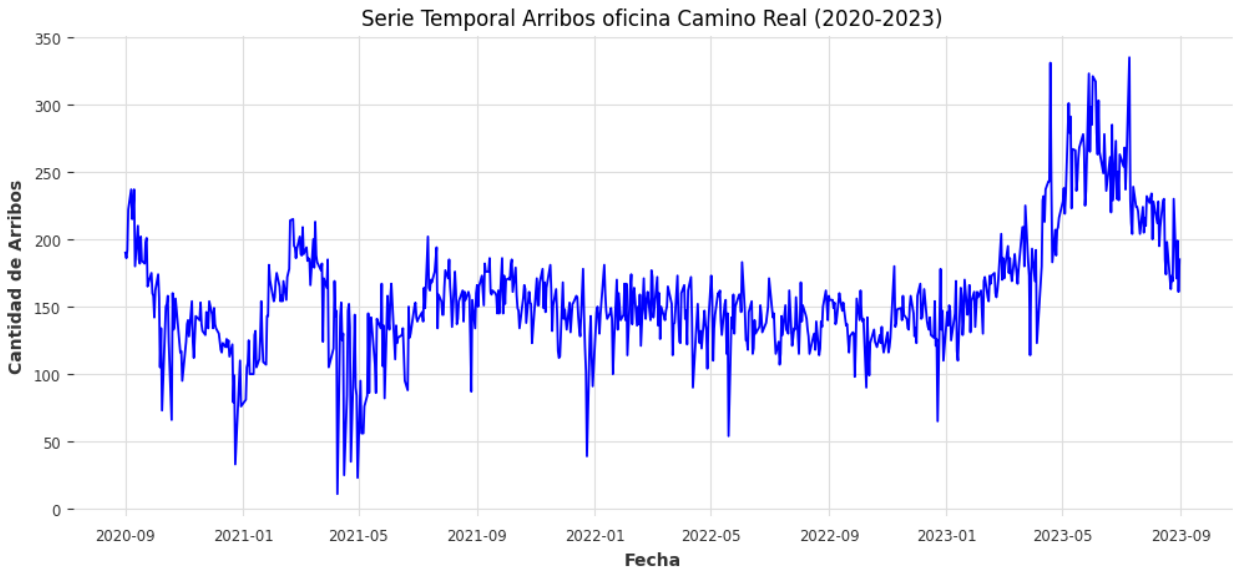
Revisamos la descripción de las variables después de cargar los datos para revisar que no haya alguna inconsistencia:

Imagen 2. Resumen de variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 731 entries, 0 to 730
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   oficina  731 non-null    object
 1   year     731 non-null    int64
 2   mes      731 non-null    int64
 3   fecha    731 non-null    datetime64[ns]
 4   arribos  731 non-null    int64
dtypes: datetime64[ns](1), int64(3), object(1)
memory usage: 28.7+ KB
```

Vemos entonces que no hay valores nulos o faltantes y los tipos de dato de cada variable son consistentes, por lo que no hay que hacer alguna transformación adicional al dataset. Con esto dicho, creamos la gráfica de la serie original:

Imagen 3. Comportamiento arribos 2020-2023



A partir de este gráfico podemos suponer que la serie no presenta un comportamiento estacional marcado ya que no hay una periodicidad definida para los picos, ya sean altos o bajos, de los arribos a la sucursal. Por otro lado, podemos suponer también que no hay una tendencia definida dentro de la serie, ya que por ejemplo tenemos una tendencia bajista entre septiembre de 2020 y enero de 2021, marzo y mayo de 2021, julio y septiembre de 2023; tenemos también una tendencia alcista entre enero y abril de 2021, enero y mayo de 2023; y un cierto comportamiento constante de los datos entre julio de 2021 y diciembre de 2023.

Detallamos ahora las funciones de autocorrelación para tener una mayor claridad sobre las afirmaciones anteriores:

Imagen 4. Función de autocorrelación serie original

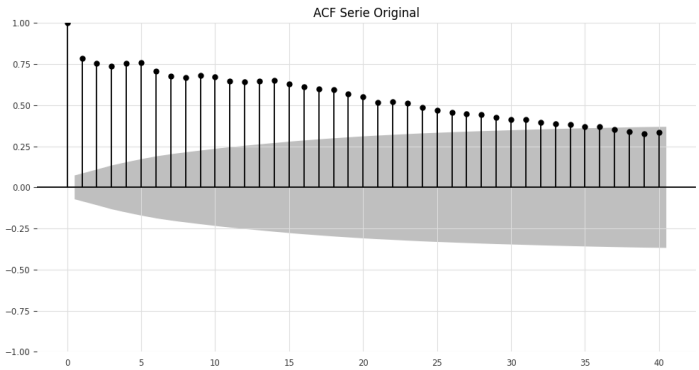
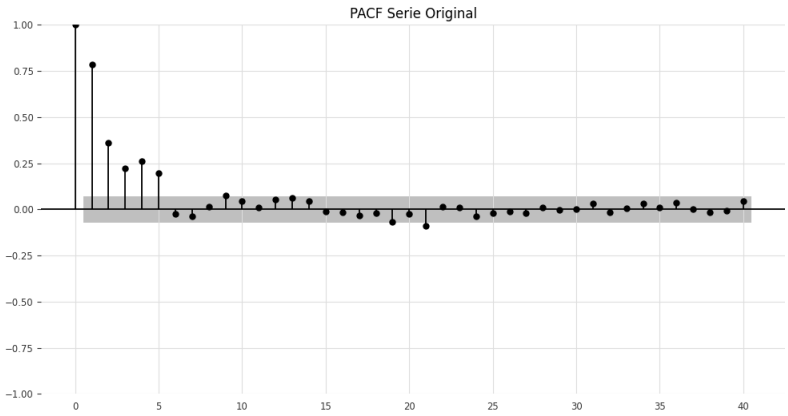


Imagen 5. Función de autocorrelación parcial serie original



Los correlogramas anteriores sugieren que la serie carece de patrones estacionales evidentes. No obstante, para validar esta suposición, llevamos a cabo la prueba de Dickey-Fuller usando el método `adfuller` de la librería `statsmodels.tsa.stattools` con el objetivo de determinar la estacionariedad de la serie.

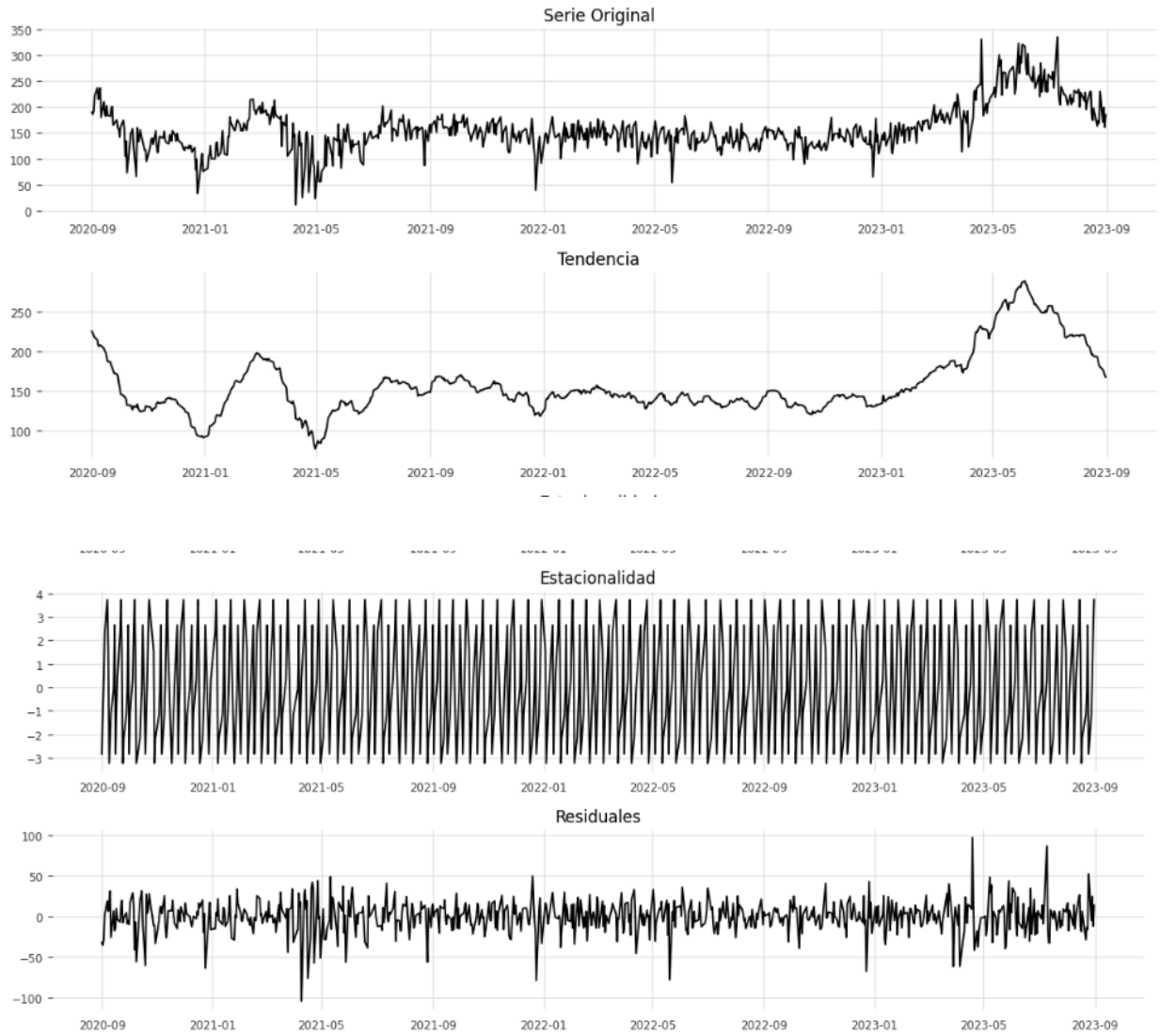
Tras llevar a cabo la prueba de Dickey-Fuller, se obtiene un *p*-valor de 0.0388 (3 c.s.) por lo que se concluye que la serie original es estacionaria, dado que el *p*-valor es inferior al nivel de significancia comúnmente establecido en 0.05.

Este hallazgo contradice aparentemente lo observado en el gráfico ACF, donde los rezagos disminuyen de manera gradual. Normalmente, una ACF que muestra rezagos que caen lentamente podría sugerir la presencia de patrones de dependencia temporal que podrían indicar no estacionariedad. Sin embargo, la prueba de Dickey-Fuller, al evaluar la presencia de raíces unitarias, sugiere lo contrario.
Este aparente conflicto entre el resultado de la prueba y la observación visual en el gráfico ACF plantea la necesidad de una evaluación más detallada, por lo que procederemos a realizar distintas transformaciones a la serie original.

Descomposición de la serie

Ahora procederemos con la descomposición de factores de la serie original para analizar su comportamiento:

Imagen 5. Descomposición de Factores



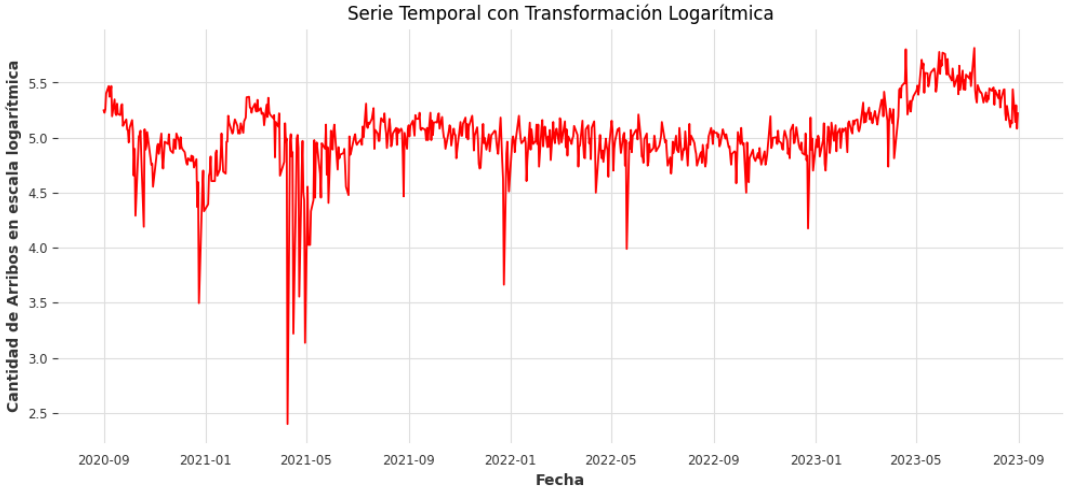
Como ya suponíamos, la serie original no muestra una descomposición de factores aceptables para tratar de lanzar modelos ingenuos. Si bien, la componente estacional parece constante y los residuos parecen ser ruido blanco, la componente de tendencia no muestra un comportamiento claro. Aún así, aplicaremos un modelo de Holt-Winters para analizar el RMSE obtenido más adelante.

Transformaciones de la serie original

Serie logarítmica

Por curiosidad académica, realizamos algunas transformaciones. A continuación, se ilustra la transformación logarítmica de la serie original

Imagen 6. Serie logarítmica

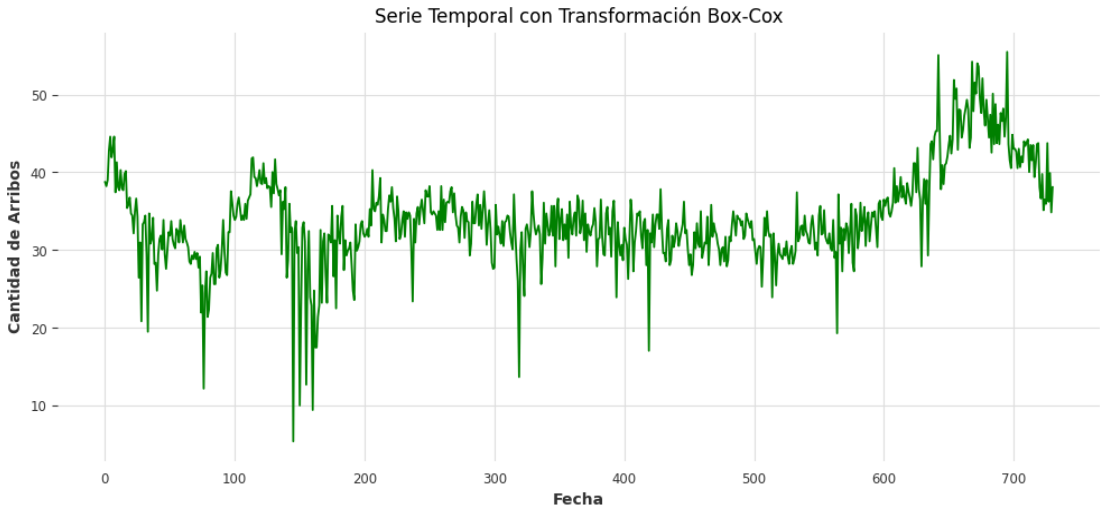


Después de hacer la transformación logarítmica de la serie vemos que se redujó la escala de valores para obtener un rango de 2.3 a 6. Sin embargo, no se logró estabilizar la varianza ni la tendencia, ya que todavía se pueden apreciar sectores de la serie donde la tendencia no es la misma. De igual manera, no se aprecian patrones estacionarios en la serie por lo que no es posible usarla en los métodos que tienen estos supuestos.

Box-Cox

De la transformación Box-Cox de la serie original obtenemos la siguiente gráfica:

Imagen 7. Transformación Box-Cox de la serie original

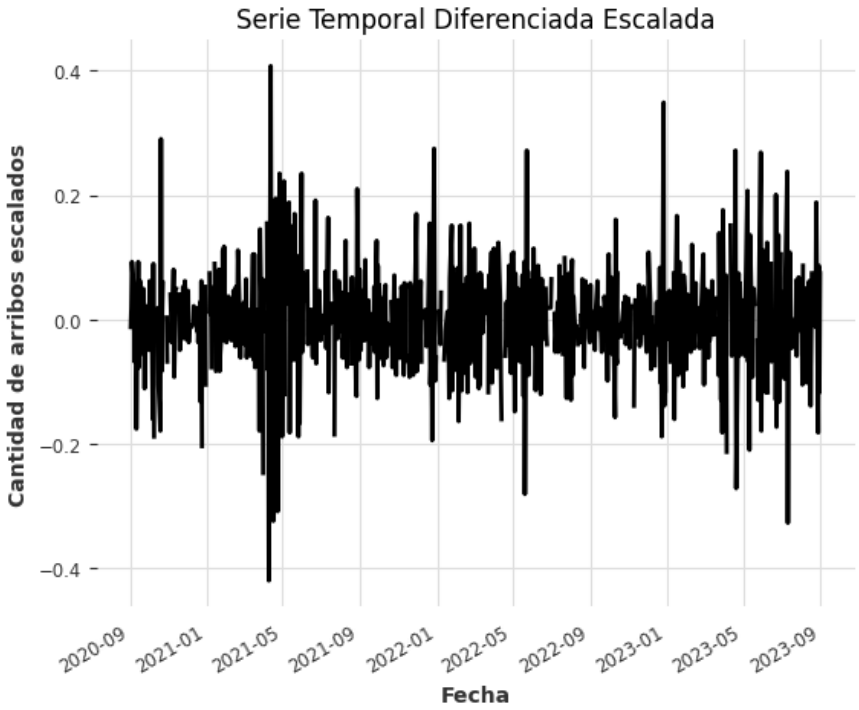


Al aplicar la transformación de Box-Cox vemos que en comparación con la serie original, la transformación se encuentra más suavizada y con un rango menor de valores. No obstante, comparte con la serie original que no presenta un comportamiento estacional marcado y no hay una tendencia definida dentro de la misma. Por tanto, al igual que la serie con transformación logarítmica, no es posible usar la serie con transformación de Box-Cox en modelos que requieran supuestos de estacionalidad y tendencia.

Serie diferenciada y escalada

Dentro de las transformaciones consideradas, primero se escalaron los datos para que estuviesen en el rango -1 y 1. Posteriormente, se procedió a realizar una diferenciación de la serie. Del proceso anterior obtenemos la siguiente gráfica:

Imagen 8. Serie diferenciada escalada



Después de aplicar la diferenciación vemos que se eliminaron tanto las tendencias presentes en la serie original y se observa un alto grado de estacionalidad; además, si bien la diferenciación no redujo totalmente la varianza de la serie original, ayudó en gran medida a estabilizarla. Lo anterior, hace que la serie diferenciada escalada sea una gran candidata para aplicar un modelo SARIMA para predecir valores futuros.

Para tener total contundencia sobre las afirmaciones anteriores, realizamos la prueba de Dickey-Fuller y obtuvimos las gráficas de la ACF y PACF de la serie diferenciada escalada. De la prueba mencionada obtuvimos un *p*-valor de 0, lo que confirma de sobremanera que la serie analizada es estacionaria. Veamos ahora los gráficos considerados:

Imagen 9. Función de autocorrelación serie diferenciada escalada

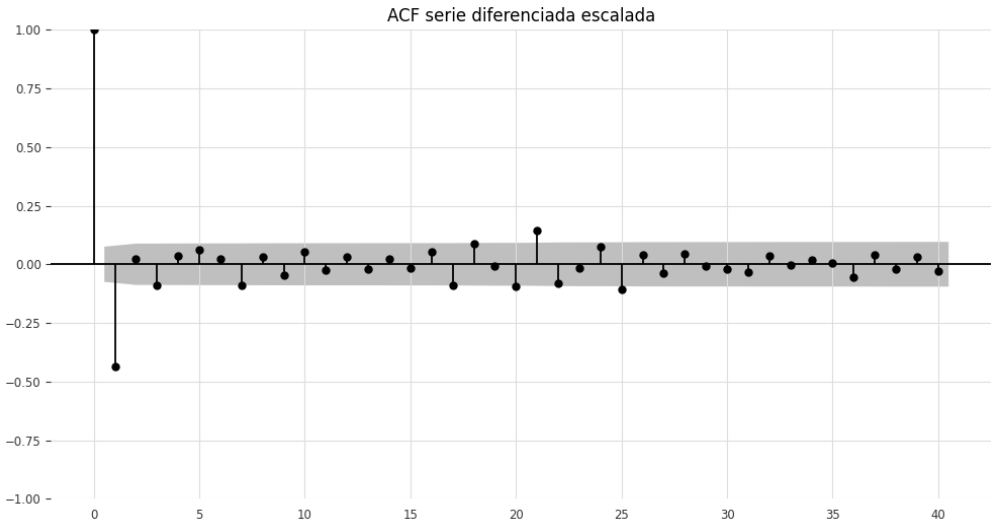
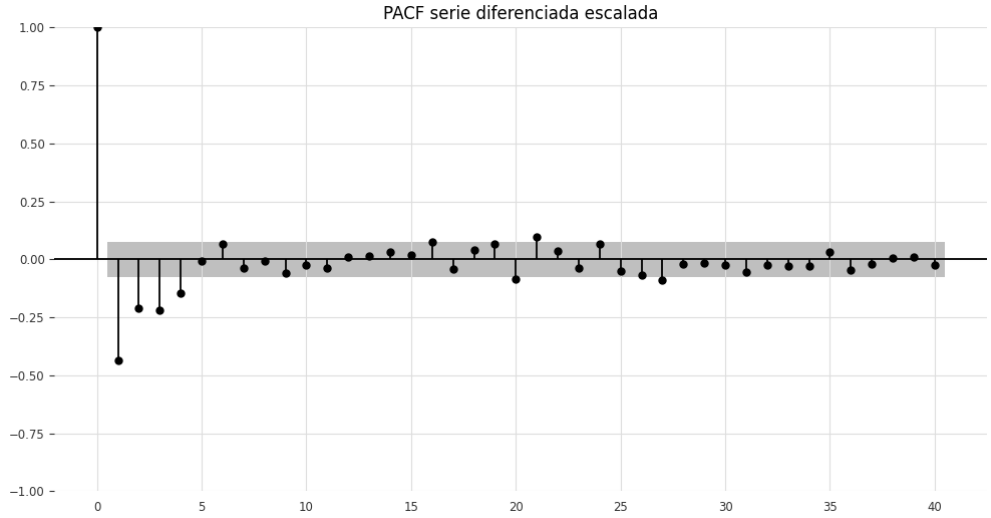


Imagen 10. Función de autocorrelación parcial serie diferenciada escalada



De los correlogramas anteriores podemos resaltar que el rezago 21 puede ser considerado como significativo en ambos gráficos. Esto podría deberse a que como los valores son diarios en días hábiles, tal vez corresponda a un ciclo mensual, de alguna estacionalidad en una fecha particular del mes.

También, podemos inferir que probablemente la serie puede ser descrita con un modelo AR(1), dado el rezago significativo que tenemos en el gráfico ACF. Así mismo, también podría ser descrita con un modelo MA(3) ó MA(4) debido a los rezagos significativos en el gráfico PACF. Finalmente, dado que la serie corresponde a datos de días hábiles, podemos llegar a suponer un modelo con un período igual a 5. Estas conclusiones serán bastante útiles a la hora de diseñar un posible modelo SARIMA más adelante.

Estudio de Modelos

En esta sección evaluaremos posibles modelos que podemos aplicar a la serie de tiempo analizada para poder realizar predicciones. Como criterio de comparación usaremos la raíz del error cuadrático medio (*RMSE)* ya que es una métrica fácilmente interpretable en la misma escala que los datos originales. Además, penaliza de manera efectiva los errores grandes y su sensibilidad a los valores extremos lo hace adecuado para capturar anomalías importantes dentro de la serie considerada.

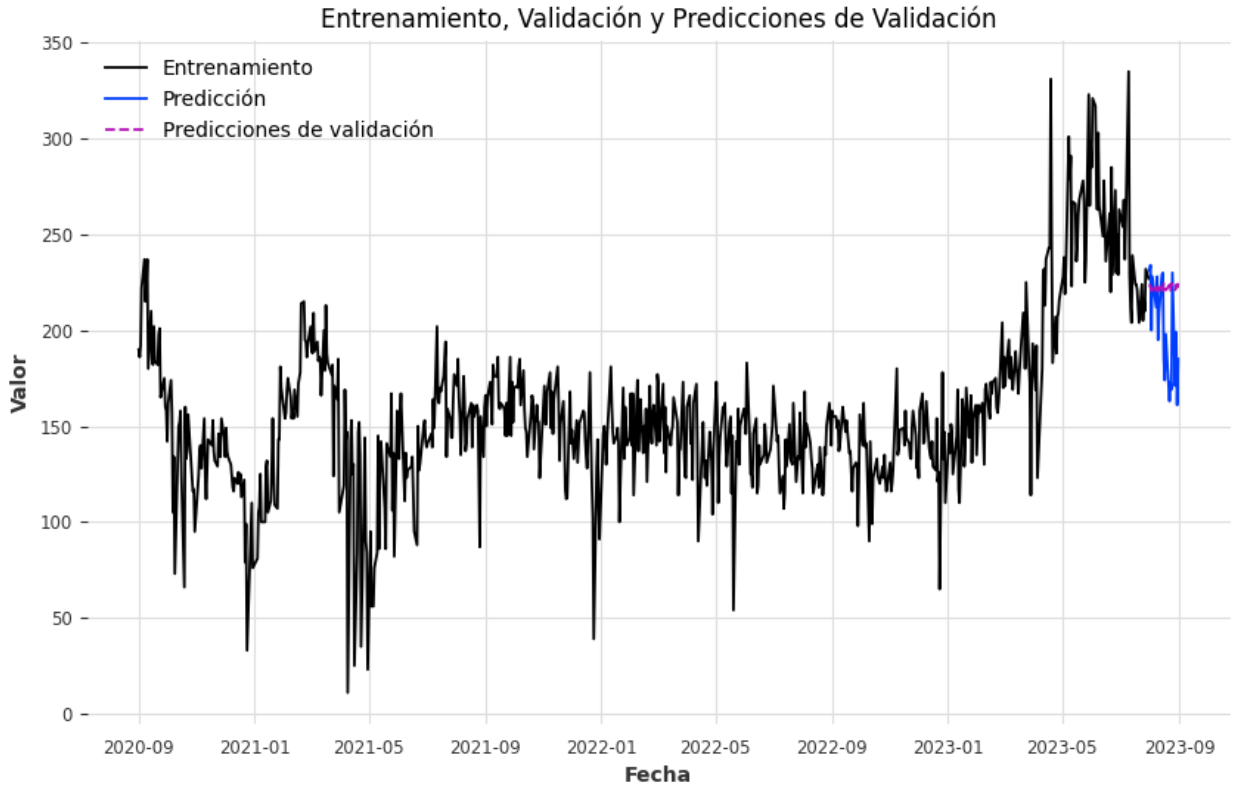
Comenzaremos el estudio de modelos considerando meramente por curiosidad un modelo ingenuo y posteriormente pasaremos a modelos más avanzados: SARIMA.

Modelo ingenuo

Dentro de los modelos ingenuos usaremos el modelo de Holt-Winters sobre la serie original escalada a través del método `ExponentialSmoothing` de la librería `statsmodels.tsa.holtwinters`. Después de aplicar el modelo se obtuvo un *RMSE* en la muestra de entrenamiento de 23.8 (3 c.s.) y 33.5 (3 c.s.) en la muestra de validación. Al ser el *RMSE* de validación más alto que el *RMSE* de entrenamiento sugiere que el modelo puede tener dificultades para generalizar a nuevos datos; incluso, dichos valores altos del *RMSE* indican que el modelo no se ajusta bien al comportamiento de la serie, como era de esperarse para un modelo ingenuo.

Veamos lo anterior gráficamente:

Imagen 11. Modelo de Holt-Winters sobre la serie escalada



Resulta evidente que el método de Holt-Winters no logra capturar con precisión la variabilidad de la serie original, como era de esperarse. Este resultado justifica la decisión de usar la diferenciación de la serie y la aplicación de modelos más avanzados, buscando mejorar la capacidad predictiva y ajuste a la estructura de la serie temporal.

SARIMA

Siendo conscientes del pobre desempeño de los modelos ingenuos en nuestra serie original procedemos a usar métodos más avanzados, para ello consideraremos varios modelos *SARIMA*. Teniendo en cuenta los supuestos de estacionalidad y homocedasticidad, debemos usar la serie diferenciada en nuestros modelos *SARIMA* ya que es la cumple (o se acerca más, respecto a la homocedasticidad) en vez de la serie original.

Para la comparación y elección del mejor modelo usaremos el método `auto_arima` de la librería `pmdarima`, este método nos permite seleccionar automáticamente los hiperparámetros del mejor SARIMA teniendo en cuenta un espacio de parámetros proporcionado. Si bien `auto_arima` usa el *AIC* para seleccionar el mejor modelo dentro del espacio de parámetros, usaremos el *RMSE* para decidir finalmente entre los

distintos modelos proporcionados a partir de una iteración sobre el hiperparámetro *s*. A continuación vemos el resultado de las iteraciones:

Imagen 12. Comparación modelos SARIMA

s	p	d	q	P	D	Q	RMSE
3	0	0	1	1	0	0	0.093957
4	0	0	1	0	0	0	0.094381
5	0	0	1	1	0	0	0.094558
6	0	0	1	0	0	0	0.094381

De acuerdo con las múltiples iteraciones realizadas, el modelo que provee el mejor *RMSE* es el *SARIMA*(0,0,1)(1,0,0,3), con un *RMSE* de 0.0940 (3 c.s.). Ahora bien, no necesariamente es el que ofrece el mejor *AIC*, dado que no revisamos esta métrica. Como nuestra métrica principal es el *RMSE*, escogemos dicho modelo.

Analizemos ahora detenidamente el valor de cada hiperparámetro del modelo seleccionado:

- *p=0*: no hay términos autorregresivos en el componente no estacionario.
- *d=0*: no se realiza diferenciación en el componente no estacionario.
- *q=1*: se considera un término en la media móvil en el componente no estacionario.
- *P=1*: se considera un término autorregresivo en el componente estacionario.
- *D=0:* no se realiza diferenciación en el componente estacionario.
- *Q=0*: no hay términos de la media móvil en el componente estacionario.
- *s=3*: La serie de tiempo escalonada diferenciada modelada tiene una estacionalidad con período 3.

En síntesis, el mejor modelo *SARIMA* encontrado por `auto_arima` tiene un componente no estacionario que incluye un término en la media móvil y un componente estacionario que incluye un término autorregresivo con una estacionalidad de período 3, esto último sugiere la presencia de una estacionalidad en la serie original que se repite cada 3 períodos, posiblemente relacionada con ciclos mensuales.

Predicciones

Luego de encontrar el modelo *SARIMA*, se procede a realizar la predicción de los 21 días hábiles del mes de septiembre de 2023. Es importante recordar que la serie que estamos analizando fue escalada y diferenciada con un lag igual 1, por lo que las predicciones requieren invertir dichas transformaciones para ser analizadas en la serie original.

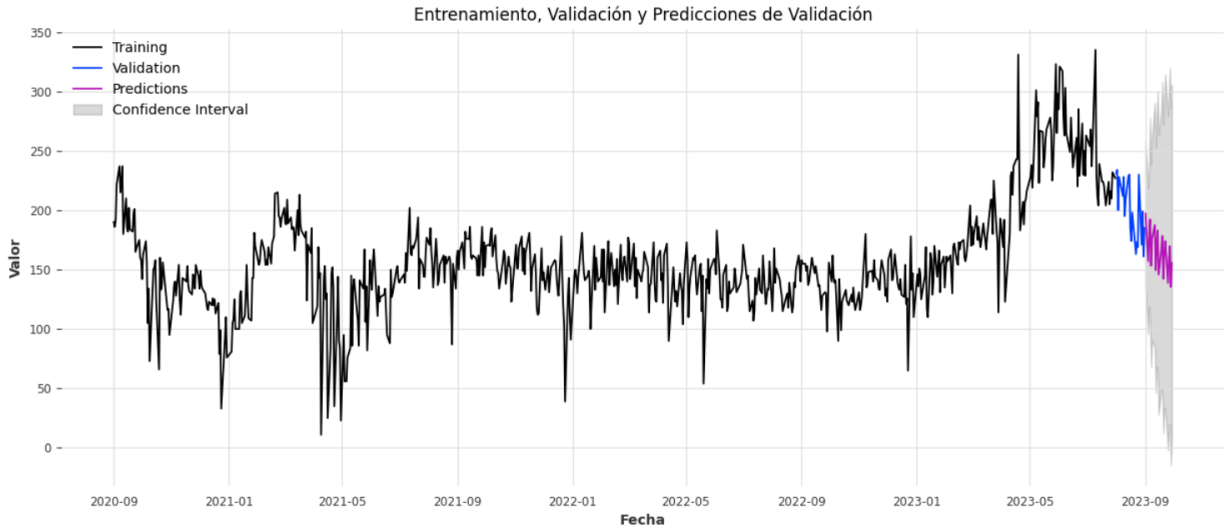
Para realizar las predicciones necesarias usamos la serie escalada sin diferenciar, dado que al usar la serie escalada diferenciada para las predicciones y luego realizar las transformaciones inversas, el resultado de las predicciones no era coherente. Por tanto, conscientes de que las predicciones no iban a ser igual de precisas, se obtuvo las siguientes predicciones para el mes de septiembre usando la serie escalada:

Imagen 13. Predicciones para el mes de septiembre

time	Arribos Predichos
2023-09-01	197.199069
2023-09-04	157.057630
2023-09-05	180.426850
2023-09-06	192.305298
2023-09-07	153.218875
2023-09-08	175.973895
2023-09-11	187.540148
2023-09-12	149.481012
2023-09-13	171.637974
2023-09-14	182.900237
2023-09-15	145.841390
2023-09-18	167.416012
2023-09-19	178.382275
2023-09-20	142.297426
2023-09-21	163.305013
2023-09-22	173.983056
2023-09-25	138.846605
2023-09-26	159.302062
2023-09-27	169.699460
2023-09-28	135.486481
2023-09-29	155.404318

En la siguiente imagen se presentan gráficamente las predicciones junto con el intervalo de confianza:

Imagen 14. Serie original junto con las predicciones e intervalos de confianza



De acuerdo con los resultados obtenidos, el modelo predice que en septiembre de 2023 se continuará con la tendencia bajista de visitas en la oficina de camino real, que comenzó alrededor del mes de mayo 2023. Es posible evidenciar que el modelo propone que las predicciones tenderán hacia la media de la serie original. También resulta importante describir que, debido a la alta varianza de la serie original, los intervalos de confianza también presentan el mismo comportamiento.

CONCLUSIONES

En la presente tarea se implementó un modelo óptimo de predicción de afluencia basado en series de tiempo para anticipar la demanda de clientes presenciales para el mes de septiembre de 2023 en la oficina CAMINO REAL de la empresa considerada. Para tal fin se realizó en primera medida un análisis exploratorio de la serie de arribos, llegando la conclusión que esta serie no era estacional, tenía una alta variabilidad y no tenía una tendencia marcada.

Se propusieron entonces varias transformaciones de la serie para encontrar una que satisficiera los criterios necesarios para usar modelos avanzados, resaltando la serie diferenciada, que mostraba ser estacionaria y con baja variabilidad. Posteriormente se desarrollaron modelos predictivos de series de tiempo que permitían estimar el arribo de clientes en la oficina, considerando patrones históricos y factores estacionales: el primero a través de un modelo ingenuo, Holt-Winters y posteriormente, encontrando el mejor modelo *SARIMA* a través de variadas iteraciones.

Los modelos propuestos se evaluaron usando como métrica de desempeño la raíz del error cuadrático medio, dando como mejor modelo un *SARIMA*(0,0,1)(1,0,0,3), con un *RMSE* de 0.0940 (3 c.s.). De las predicciones mostradas bajo este modelo y usando la serie original escalada, se espera que haya mínimo 136 y máximo 197 clientes en todo el mes de septiembre, con un promedio de 166 en el mes.

Con esta información, se espera que la empresa pueda anticiparse a la demanda de clientes que visiten la oficina, siempre teniendo presente que los modelos cuentan con un margen de error, como lo son los intervalos de confianza mostrados. Aun así, estos modelos pueden ser de gran utilidad para el gerente de oficina en su objetivo de maximizar la eficiencia en el personal dispuesto a la atención de los clientes.

Esta tarea nos ayudó a reforzar la idea que las series de tiempo en la vida real no muestran los patrones estacionales y de tendencia tan marcados como se ve en la literatura. Como anticipamos, los modelos ingenuos no funcionaron bien frente a una

serie que no muestra tendencia y estacionalidad clara, aunque métodos más avanzados si lo hicieron después de transformar la serie para verificar sus supuestos.

Es un reto entonces enfrentarse a series que no tienen una apariencia similar a las de los textos académicos, aunque con esta tarea pudimos evidenciar que contamos con las herramientas básicas necesarias para poder enfrentarnos a problemas complejos de la vida real que involucran series de tiempo.