

PRONÓSTICO DEL PATRÓN DE ARRIBO DE CONTACTOS EN UNA LINEA DE BPO

Johnattan Solis Padilla

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface,
en alcances y calidad, todos los requisitos que demanda
un Trabajo de Grado de Maestría.



Director


Jurado


Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de
Magister en Ciencia de Datos.


HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias


JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago De Cali, 13 de Febrero del 2024



Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 13 de febrero del 2024

Autor: Johnattan Solis Padilla

Título del Trabajo de Grado: “PRONÓSTICO DEL PATRÓN DE ARRIBO DE CONTACTOS EN UNA LINEA DE BPO”

Director: Isabel Cristina Garcia Arboleda

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Firma del Director del Trabajo de Grado

Santiago de Cali, 09 de diciembre del 2023

Doctora

Gloría Inés Alvarez V.

Directora Maestría en Ciencia de Datos

Facultad de Ingeniería y Ciencias

Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

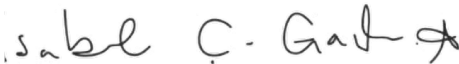
Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "PRONÓSTICO DEL PATRON DE ARRIBO DE CONTACTOS EN UNA LINEA DE BPO", el cual fue realizado por el estudiante Johnattan Solis Padilla con código *8974378* pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de Isabel Cristina Garcia Arboleda.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,



Johnattan Solis Padilla



Isabel Cristina Garcia Arboleda

C.C. _1143166001_ de _Barranquilla (ATL)

C.C. 43109746_ de _Bello(ANT)_

Documentación anexa:

Resumen del Proyecto Aplicado en formato digital (máximo 1 página).

Una copia digital (PDF) del documento del proyecto aplicado

Barranquilla, 27 de octubre del 2023

Señor(es)

Johnattan Solis Padilla

Asunto: Autorización Desarrollo Investigativo

Agradecemos su solicitud para utilizar los datos de la cantidad de llamadas en una de nuestras líneas de atención telefónica en el marco de su proyecto de maestría en Ciencia de Datos en la Pontificia Universidad Javeriana Cali. Después de revisar cuidadosamente su solicitud, nos complace otorgar la autorización para acceder y utilizar dichos datos con el fin exclusivo de llevar a cabo su investigación en torno a pronósticos de series de tiempo.

Se especifica que la propiedad de la línea de atención telefónica se mantendrá de forma anónima, y confirmamos su compromiso de no hacer referencia directa o indirecta a la identidad específica de la misma en su proyecto. Apreciamos su compromiso de cumplir con todas las leyes y regulaciones aplicables en materia de protección de datos y privacidad durante el desarrollo de su investigación.

Le recordamos la confidencialidad y la exclusividad de esta autorización; cualquier uso adicional de los datos requerirá una nueva solicitud y aprobación por parte de Teleperformance S.A. Del mismo modo, todo desarrollo en base a los datos debe ser reportado y socializado con el jefe inmediato.

Nos complace contribuir a su proyecto de maestría y esperamos que los datos proporcionados le sean de utilidad en su investigación. Estamos a disposición para discutir cualquier detalle adicional o proporcionar la información que pueda necesitar durante este proceso.



SEBASTIAN CASAS RUEDA

DIRECTOR DE CUMPLIMIENTO ORGANIZACIONAL

Teléfono: 4049081 Ext. 2512

Dirección Av. Calle 26 # 92 – 32

Bogotá, Colombia

FICHA RESUMEN PROYECTO DE TRABAJO DE GRADO

TÍTULO: PRONÓSTICO DEL PATRÓN DE ARRIBO DE CONTACTOS EN UNA LINEA DE BPO

1. ÉNFASIS: Aplicación de pronósticos
2. ÁREA DE TRABAJO: Forecasting – Series de Tiempo
3. TIPO DE PROYECTO: Aplicado
4. ESTUDIANTE(S): Johnattan Solis Padilla
5. CORREO ELECTRÓNICO: johnattans07@javerianacali.edu.co
6. DIRECCIÓN Y TELEFONO: Calle 99 # 65 64 Barranquilla - 3045499390
7. DIRECTOR: Isabel Cristina Garcia Arboleda
8. VINCULACIÓN DEL DIRECTOR: NA
9. CORREO ELECTRÓNICO DEL DIRECTOR: isabel.garcia@javerianacali.edu.co
10. CO-DIRECTOR (Si aplica): NA
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Teleperformance
12. OTROS GRUPOS O EMPRESAS: NA
13. PALABRAS CLAVE (al menos 5): Forecasting, BPO, Machine Learning, Data Modeling, Sizing.
14. FECHA DE INICIO: 11/23/2022
15. DURACIÓN ESTIMADA (En meses): 7
16. RESUMEN:

Este proyecto consiste en el desarrollo de un modelo de pronóstico para la predicción del patrón de arribo de llamadas en una línea telefónica de atención al cliente, y la implementación de este modelo en una solución de analítica avanzada. Uno de los principales retos en la administración de las líneas de atención telefónica es predecir el volumen de contactos que serán atendidos con una proyección a largo plazo, mediano plazo y en tiempo real; esto es requerido para poder generar los planes de contratación de personal, configuración de horarios y toma de decisiones en tiempo real con respecto a cuando sacar o meter agentes de la línea. Actualmente, la empresa en estudio desarrolla estas estimaciones de una forma empírica o con técnicas poco avanzadas e imprecisas. Con este proyecto se pretende establecer el modelo de pronóstico más apropiado para predecir el patrón de arribo de contactos, evaluando modelos estadísticos y de aprendizaje automático para posterior implementar una solución de analítica avanzada que le permita a la administración de la línea tener una vista del pronóstico a mediano plazo (2 meses), corto plazo (entre mes) y tiempo real (entre día) de las llamadas que ingresarán a la línea. Este proyecto será aplicado a una línea de soporte de una cuenta anónima, pero se espera que su aplicabilidad se extienda a las más de 250 líneas que opera la empresa de BPO más grande en Colombia y una de las más grandes del mundo.



Pontificia Universidad
JAVERIANA
Cali

**PRONÓSTICO DEL PATRÓN DE ARRIBO DE CONTACTOS EN UNA
LINEA DE BPO**

Johnattan Solis Padilla
Código 8974378

*Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos*

Director(a)
Isabel Cristina Garcia Arboleda

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, Febrero 13 DE 2024

TABLA DE CONTENIDO

INTRODUCCIÓN	7
1. DEFINICIÓN DEL PROBLEMA	8
1.1. Planteamiento del problema	8
1.2. Formulación del problema	8
2. OBJETIVOS DEL PROYECTO	9
2.1 Objetivo general	9
2.2 Objetivos específicos	9
2.3 Resultados esperados	9
3. MARCO DE REFERENCIA	9
3.1 Modelos de suavizamiento	10
3.1.1 Media Móvil	10
3.1.2 Suavizamiento exponencial simple	10
3.1.3 Suavizamiento Lineal de Holt	11
3.1.4 Suavizamiento exponencial estacional de Winters	11
3.2 Modelos Autorregresivos	12
3.2.1 Modelo AR	12
3.2.2 Modelo ARMA	13
3.2.3 Modelo Arima	13
3.2.4 Modelo SARIMA	14
3.3 Machine Learning y Deep Learning en Series de Tiempo	15
3.3.1 Regresiones	15
3.3.2 Red Neuronal Recurrente (LSTM)	16
3.3.3 XGBoost	16
3.4 Métricas para evaluar pronósticos	17
3.5 Antecedentes	17
4. DESARROLLO OBJETIVO 1	18
4.1 Análisis exploratorio de los datos.	18
4.1.1 Presentación de la serie de tiempo	19
4.1.2 Descomposición de la serie de tiempo	19
4.1.3 Análisis de Tendencia	20
4.1.4 Diferenciación de la serie	21
4.1.4.1 Descomposición de la serie diferenciada	21

4.1.4.2	Análisis de tendencia para la serie diferenciada	22
4.1.4.3	Análisis de estacionalidad para la serie diferenciada	22
4.1.5	Análisis de estacionariedad	23
4.1.6	Análisis de rezagos	23
4.1.7	Análisis de la distribución de la variable	24
4.2	Definición del set de entrenamiento y el set de evaluación.	25
4.3	Selección de modelos y algoritmos	25
4.4	Modelo SARIMA	26
4.4.1	Implementación Modelo SARIMA	26
4.4.2	Interpretación del Modelo SARIMA	26
4.4.3	Validación de residuales SARIMA	27
4.4.4	Pronostico del modelo SARIMA	30
4.5	Modelo Suavizamiento Exponencial Triple (SET)	30
4.5.1	Implementación Suavizamiento Exponencial Triple	30
4.5.2	Interpretación del modelo SET	31
4.5.3	Validación de residuales SET	31
4.5.4	Pronóstico del modelo SET	34
4.6	Redes neuronales LSTM (Long Short-Term Memory)	34
4.6.1	Implementación del modelo LSTM	34
4.6.2	Interpretación del modelo LSTM	35
4.6.3	Validación de residuales LSTM	36
4.6.4	Pronóstico del modelo LSTM	38
5	DESARROLLO OBJETIVO 2	39
5.1	Análisis de métricas de error	39
5.2	Selección de métricas de error	40
5.3	Evaluación de métricas de error	40
5.4	Selección del modelo	40
6	DESARROLLO OBJETIVO 3	41
6.1	Definición de la necesidad	41
6.2	Arquitectura propuesta	41
6.3	Orquestación de actividades (Pipeline)	42
6.4	Visualización de la solución	1
7	CONCLUSIONES Y TRABAJOS FUTUROS	2

7.1	Conclusiones	2
7.2	Trabajos futuros	3
8	REFERENCIAS BIBLIOGRAFICAS	3

TABLA DE ILUSTRACIONES

Ilustración 1 Importe de los datos	18
Ilustración 2 Serie de tiempo original	19
Ilustración 3 Descomposición de la serie original	20
Ilustración 4 Descomposición de la serie diferenciada	21
Ilustración 5 Graficas ACF y PACF de la serie de tiempo	22
Ilustración 6 Rezagos de la serie de tiempo	24
Ilustración 7 Histograma de volumen de llamadas por día	24
Ilustración 8 Grafica QQ para residuales	28
Ilustración 9 Grafica de dispersión de residuos	29
Ilustración 10 Pronostico de modelo Sarima	30
Ilustración 11 Grafica QQ Modelo Suavizamiento ET	32
Ilustración 12 Distribución de residuales SET	32
Ilustración 13 Gráfico de dispersión de residuos SET	33
Ilustración 14 Función de autocorrelación de residuales SET	34
Ilustración 15 Pronostico del modelo SET	34
Ilustración 16 Configuración de red neuronal LSTM	35
Ilustración 17 Gráfica de pérdida LSTM	35
Ilustración 18 Distribución de residuales LSTM	36
Ilustración 19 Grafico de dispersión de residuos LSTM	37
Ilustración 20 Gráfica de autocorrelación residuales LSTM	38
Ilustración 21 Arquitectura propuesta	42
Ilustración 22 Pipeline de actividades	1
Ilustración 23 Modelo tabular	1

INTRODUCCIÓN

Algunos de los procesos más comunes en la industria del BPO (Business Process Outsourcing) es la atención en líneas de soporte, ventas, cobranza etc. Muchas empresas a nivel mundial tercerizan estos procesos debido a la complejidad de la administración de ellos y lo económicamente rentable que resulta subcontratar estas operaciones de atención de contactos; entendamos por contactos las llamadas, chats, emails, tiques etc. Los métodos más comunes de facturación de la empresa que terceriza el proceso a la empresa primaria son: contacto gestionado, minuto de gestión, hora de gestión, entre otros. No obstante, independientemente de la modalidad de facturación adoptada, todas estas se caracterizan por estar directamente relacionadas con la producción y gestión de dichos contactos.

Las empresas que brindan servicios de BPO se enfrentan al desafío de gestionar eficazmente un alto volumen de contactos entrantes, evitando la generación de ineficiencias que pudieran resultar en tiempos improductivos por parte de sus agentes, lo que a su vez conlleva a pérdidas económicas para la empresa de BPO. Este desafío comienza con la predicción del volumen de contactos que se espera ingresen a la línea, y es precisamente este proceso el foco de investigación de este proyecto.

Durante este proyecto se pretende definir el modelo apropiado para predecir el patrón de arribo de llamas en una línea de servicio al cliente, considerando técnicas avanzadas de pronóstico, para mejorar la eficiencia en la gestión de la línea. Se estudiará el caso de una línea de servicio al cliente y se implementaran múltiples técnicas de pronósticos con el fin de evaluarlas y definir la más apropiada para este proceso; posteriormente, se implementará una solución de analítica avanzada la cual utilizará este modelo y entregará una herramienta de inteligencia de negocio que apoye la toma de decisiones en la administración de esta línea.

1. DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento del problema

El objetivo principal de las empresas de BPO es gestionar la mayor cantidad de contactos ofrecidos, pero teniendo en cuenta que tener una capacidad instalada o capacidad de atención más alta de la necesaria genera ineficiencias en el proceso, ya que la empresa de BPO debe pagar la nómina de los empleados, pero esta no puede ser mayor que la factura al cliente (empresa primaria) porque de lo contrario se estarían generando pérdidas.

El dilema se encuentra en que el volumen de contactos ofrecidos que llegan a la línea depende de variables económicas, temporadas, día de la semana, hora del día y muchas otras variables que generan incertidumbre en el volumen que ingresará a la línea y la distribución que este tendrá. Es aquí en donde se genera el reto más grande de las empresas de BPO y es saber en qué momento y con que flujo ingresarán los contactos a la línea con el fin de tener la capacidad instalada necesaria (número de agentes disponibles) para gestionar esos contactos sin llegar a generar ineficiencias.

La ciencia de datos nos ofrece diferentes herramientas para realizar predicciones tales como modelos autorregresivos, modelos de suavizamiento, regresiones, algoritmos de machine learning, entre otros. La ausencia de la implementación de estas herramientas desaprovecha la oportunidad de obtener mejores resultados en la predicción del volumen de interacciones que se tendrá en la línea y consecuentemente reduce la utilidad que se puede tener de la operación.

1.2. Formulación del problema

Teniendo en cuenta la importancia de conocer el volumen de contactos que se gestionarán en la línea la pregunta problema que pretende responder este proyecto es ¿Cómo predecir el patrón de arribo de contactos en una línea de atención telefónica?

Para darle respuesta a esta pregunta debemos hacernos un cuestionamiento de investigación que nos ayudará con este objetivo. Inicialmente debemos determinar ¿Qué métodos de pronóstico existen actualmente para predecir el patrón de arribo de llamadas?, ¿Cómo evaluar la efectividad de los modelos de pronóstico para determinar el más apropiado? y consecuente a esto ¿Cuál sería el modelo o técnica que mejor pronostica el patrón de arribo de llamadas para la línea en estudio?

2. OBJETIVOS DEL PROYECTO

2.1 Objetivo general

Predecir el patrón de arribo de llamadas en una línea de servicio al cliente, considerando técnicas avanzadas de pronóstico, para implementarlo en una solución de analítica que permita mejorar la eficiencia en la gestión de la línea.

2.2 Objetivos específicos

- Identificar los métodos o modelos de pronóstico apropiados para predecir el patrón de arribo de llamadas con el fin de implementarlos en la serie estudiada y generar un portafolio de pronósticos que nos permita implementar un plan de evaluación.
- Identificar y aplicar los métodos de evaluación de pronóstico más apropiados para el caso de estudio con el fin de evaluar su desempeño y seleccionar el modelo más apropiado.
- Desarrollar una solución de analítica avanzada que utilice el modelo definido y permita visualizar los pronósticos a largo plazo, mediano plazo y tiempo real, para que la administración de la línea pueda utilizarlo en su gestión.

2.3 Resultados esperados

El resultado esperado de este proyecto es la creación de un modelo predictivo apropiado para pronosticar el patrón de arribo de contactos en una línea de llamadas; por otro lado, este proyecto entregará una solución de analítica avanzada la cual utilizará el modelo creado y servirá como herramienta de inteligencia de negocio que ayude a la administración de la línea en la configuración de los recursos necesarios para atender el tráfico a mediano plazo, corto plazo, y en tiempo real.

Adicional al resultado principal de este proyecto, su desarrollo dejará un data set que podrá ser utilizado en futuras investigaciones a partir de lo encontrado durante la ejecución del proyecto. Finalmente, este estudio dejará este entregable que será un aporte al estudio de la Ciencia de Datos, específicamente a las técnicas de pronóstico.

3. MARCO DE REFERENCIA

Para el desarrollo de este proyecto se requieren conocerlos los diferentes modelos de pronóstico que existen en la actualidad, estos los agruparemos de la siguiente manera: modelos autorregresivos, modelos de suavizamiento y modelos de machine learning. Adicional a esto, veremos diferentes técnicas de evaluación de pronóstico por medio métricas de errores.

3.1 Modelos de suavizamiento

3.1.1 Media Móvil

El modelo de media móvil es una técnica de análisis de series de tiempo que se utiliza para suavizar los datos y reducir las fluctuaciones aleatorias en una serie de tiempo. El modelo de media móvil se basa en el promedio de los valores de la serie temporal en un intervalo móvil, lo que permite identificar patrones y tendencias en la serie.¹¹

El modelo de media móvil se utiliza para modelar y predecir series de tiempo que no tienen una tendencia clara o una estacionalidad marcada. En particular, los modelos de media móvil son adecuados para pronosticar valores futuros a corto y mediano plazo.¹¹

3.1.2 Suavizamiento exponencial simple

El suavizamiento exponencial simple es un método de pronóstico de series de tiempo que se utiliza para predecir valores futuros en función de los valores históricos. Este método se basa en la idea de que los valores futuros de una serie de tiempo se pueden estimar como una combinación ponderada de los valores pasados. La ponderación disminuye exponencialmente a medida que los datos se alejan en el tiempo, lo que significa que los datos más recientes tienen más peso en la predicción que los datos más antiguos.

La fórmula para el suavizamiento exponencial simple es:

$$\widehat{y}_{t+1} = \alpha y_t + (1 - \alpha) \widehat{y}_t$$

Donde:

- \widehat{y}_{t+1} es la predicción para el período $t + 1$
- y_t es el valor observado en el período t
- \widehat{y}_t es la predicción para el período t
- α es el factor de suavizado, que debe estar entre 0 y 1

El factor de suavizado α controla la cantidad de peso que se le da a los datos más recientes. Un valor alto de α significa que se les da más peso a los datos más recientes, mientras que un valor bajo de α significa que se les da más peso a los datos más antiguos.

El suavizamiento exponencial simple se utiliza comúnmente en la industria para pronosticar la demanda futura de productos y servicios. También se utiliza en finanzas para pronosticar precios futuros de acciones y otros instrumentos financieros.³

3.1.3 Suavizamiento Lineal de Holt

El modelo de suavizamiento lineal de Holt es una técnica de análisis de series de tiempo que se utiliza para suavizar los datos y reducir las fluctuaciones aleatorias en una serie de tiempo. El modelo se basa en el promedio ponderado de los valores de la serie temporal en un intervalo móvil, lo que permite identificar patrones y tendencias en la serie.

El modelo de suavizamiento lineal de Holt se utiliza para modelar y predecir series de tiempo que no tienen una tendencia clara o una estacionalidad marcada. En particular, el modelo es adecuado para pronosticar valores futuros a corto y mediano plazo.⁴

La formulación matemática del modelo de suavizamiento lineal de Holt es la siguiente:

$$\widehat{y_{t+h|t}} = l_t + hb_t$$

donde:

- $\widehat{y_{t+h|t}}$ es el valor pronosticado en el tiempo $t+h$,
- l_t es el nivel estimado en el tiempo t ,
- b_t es la pendiente estimada en el tiempo t ,
- h es el horizonte de pronóstico.

El modelo de suavizamiento lineal de Holt tiene varias ventajas, como su simplicidad y facilidad de uso. Sin embargo, también tiene algunas desventajas, como su incapacidad para modelar patrones estacionales complejos.

3.1.4 Suavizamiento exponencial estacional de Winters

El modelo de suavizamiento exponencial estacional de Winters es una técnica de análisis de series de tiempo que se utiliza para suavizar los datos y reducir las fluctuaciones aleatorias en una serie de tiempo. El modelo se basa en el promedio ponderado de los valores de la serie temporal en un intervalo móvil, lo que permite identificar patrones y tendencias en la serie.²

El modelo de suavizamiento exponencial estacional de Winters se utiliza para modelar y predecir series de tiempo que tienen una estacionalidad clara.³ En particular, el modelo es adecuado para pronosticar valores futuros a corto y mediano plazo.⁴

La formulación matemática del modelo de suavizamiento exponencial estacional de Winters es la siguiente:

$$\widehat{y_{t+h|t}} = l_t + hb_t + S_{t-m+h_{m+}}$$

donde:

- $\widehat{y_{t+h|t}}$ es el valor pronosticado en el tiempo t+h,
- l_t es el nivel estimado en el tiempo t,
- b_t es la pendiente estimada en el tiempo t,
- $s_{t-m+h_m^+}$ es el componente estacional estimado en el tiempo t+h,
- h es el horizonte de pronóstico,
- m es la longitud del ciclo estacional,
- h_m^+ es la parte entera del cociente entre h y m.

El modelo de suavizamiento exponencial estacional de Winters tiene varias ventajas, como su capacidad para modelar patrones estacionales complejos. Sin embargo, también tiene algunas desventajas, como su incapacidad para modelar patrones no lineales o no estacionarios.³

3.2 Modelos Autorregresivos

El análisis de series de tiempo a través de modelos autorregresivos es el más utilizado para la planeación de la demanda (Petropoulos et al., 2018). A continuación, se relacionan los métodos más comunes para pronosticar en series de tiempo, una breve descripción, el horizonte de tiempo que pueden pronosticar y su formulación:

3.2.1 Modelo AR

Los modelos autorregresivos (AR) son una clase de modelos que se utilizan para describir ciertos procesos que varían en el tiempo¹. En estos modelos, la variable de interés depende linealmente de sus propios valores previos y de un término estocástico.¹

La notación AR(p) indica un modelo autorregresivo de orden p. El modelo AR(p) se define como:

$$X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t$$

donde:

- X_t es la variable de interés en el tiempo t,
- c es una constante,
- ϕ_i son los parámetros del modelo,
- X_{t-i} son los valores anteriores de la variable,
- ε_t es un término de error que se asume como ruido blanco.³

Este modelo puede entenderse como un proceso que, después de enfrentar una perturbación, tarda tiempo en regresar a su equilibrio de largo plazo. Por ejemplo, si nuestra variable de interés es la producción de trigo, en cada año, la producción dependerá de las condiciones del suelo, la tecnología agrícola disponible y las condiciones climáticas.¹

Es importante mencionar que para que un modelo AR(p) sea estacionario en sentido amplio, las raíces del polinomio deben estar dentro del círculo unitario. Es decir, cada raíz debe satisfacer $|\phi| < 1$.¹

Los modelos AR son útiles para realizar pronósticos sobre variables ex-post (observaciones que conocemos completamente su valor) en determinados momentos del tiempo normalmente ordenados cronológicamente. En la vida real, encontramos previsiones mediante autorregresión en la proyección de ventas de una compañía, pronóstico sobre crecimiento del producto interior bruto (PIB) de un país, previsión sobre presupuesto y tesorería, entre otros.²

3.2.2 Modelo ARMA

Un modelo autorregresivo de media móvil (ARMA) es una extensión del modelo AR que incluye términos de media móvil (MA).³ En este modelo, la variable de interés depende linealmente de sus propios valores previos, de los valores previos del término de error y del término de error actual.³

La notación ARMA(p,q) representa un modelo autorregresivo de orden p y media móvil de orden q. El modelo ARMA(p,q) se define como:

$$X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$$

donde:

- X_t es la variable de interés en el tiempo t,
- c es una constante,
- ϕ_i son los parámetros del modelo autorregresivo,
- θ_j son los parámetros del modelo de media móvil,
- X_{t-i} son los valores anteriores de la variable,
- ε_t es un término de error que se asume como ruido blanco.³

Los modelos ARMA son útiles para modelar series temporales estacionarias.¹ En la vida real, encontramos aplicaciones en la predicción de precios de acciones, pronóstico sobre el tipo de cambio, previsión sobre el clima, entre otros.¹

3.2.3 Modelo Arima

Un modelo autorregresivo integrado de media móvil (ARIMA) es una extensión del modelo AR que incluye términos de media móvil (MA). Este modelo se utiliza para modelar series temporales estacionarias.³

En general, los modelos ARIMA son adecuados para series de tiempo que exhiben una tendencia y/o una estacionalidad clara.¹ En particular, los modelos ARIMA son útiles para modelar series de

tiempo que tienen una tendencia lineal o exponencial, así como una estacionalidad aditiva o multiplicativa.³

El horizonte de tiempo en el que funciona bien un modelo ARIMA depende de la serie de tiempo en cuestión. En general, los modelos ARIMA son adecuados para pronosticar valores futuros a corto y mediano plazo. Sin embargo, la precisión del pronóstico disminuye a medida que aumenta el horizonte de tiempo.³

La formulación del modelo ARIMA es la siguiente:

$$\Delta^d X_t = c + \sum_{i=1}^p \phi_i \Delta^d X_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$$

donde:

- $\Delta^d X_t$ es la serie temporal diferenciada de orden d ,
- c es una constante,
- ϕ_i son los parámetros del modelo autorregresivo,
- θ_j son los parámetros del modelo de media móvil,
- ε_t es un término de error que se asume como ruido blanco.³

Los modelos ARIMA son útiles para modelar y predecir series temporales estacionarias. En particular, son adecuados para series de tiempo con una tendencia lineal o exponencial y una estacionalidad aditiva o multiplicativa. Los modelos ARIMA pueden pronosticar valores futuros a corto y mediano plazo, pero su precisión disminuye a medida que aumenta el horizonte de tiempo.³

3.2.4 Modelo SARIMA

Un modelo autorregresivo integrado de media móvil estacional (SARIMA) es una extensión del modelo ARIMA que incluye términos de media móvil estacional (SMA) y diferenciación estacional. Este modelo se utiliza para modelar series temporales estacionarias con una tendencia y/o una estacionalidad clara.³

La notación SARIMA(p,d,q)(P,D,Q)_s representa un modelo SARIMA de orden (p,d,q) para la parte no estacional y de orden (P,D,Q)_s para la parte estacional. El modelo SARIMA(p,d,q)(P,D,Q)_s se define como:

$$\Delta^d \Delta_s^D X_t = c + \sum_{i=1}^p \phi_i \Delta^d \Delta_s^D X_{t-i} + \sum_{j=1}^q \theta_j \Delta^d \Delta_s^D \varepsilon_{t-j} + \sum_{k=1}^P \Phi_k \Delta_s^D X_{t-s} + \sum_{l=1}^Q \Theta_l \Delta_s^D \varepsilon_{t-s-l} + \varepsilon_t$$

donde:

- Δ_s es el operador de diferenciación estacional,
- s es la periodicidad de la serie temporal,
- Δ_s^D es el operador de diferenciación estacional de orden D ,
- Δ^d es el operador de diferenciación no estacional de orden d ,
- X_t es la variable de interés en el tiempo t ,
- c es una constante,
- $-\phi_i$ son los parámetros del modelo autorregresivo no estacional,
- θ_j son los parámetros del modelo SMA no estacional,
- Φ_k son los parámetros del modelo autorregresivo estacional,
- Θ_l son los parámetros del modelo SMA estacional,
- X_{t-i} son los valores anteriores de la variable no estacional,
- ε_t es un término de error que se asume como ruido blanco.³

Los modelos SARIMA son útiles para modelar y predecir series temporales estacionarias con una tendencia y/o una estacionalidad clara. En particular, son adecuados para series de tiempo con una tendencia lineal o exponencial y una estacionalidad aditiva o multiplicativa. Los modelos SARIMA pueden pronosticar valores futuros a corto y mediano plazo, pero su precisión disminuye a medida que aumenta el horizonte de tiempo.³

3.3 Machine Learning y Deep Learning en Series de Tiempo

Existen diferentes algoritmos de Machine learning que se han adaptado a las series de tiempo para pronosticar datos futuros, tales como las redes neuronales, regresión lineal múltiple y Xtreme Gradient Boosting también llamado XGBoost.

Pedrycz (2020) afirma que las redes neuronales artificiales son una de las técnicas con mayor precisión y más usos para pronósticos en múltiples escenarios a nivel social, monetario, bolsa de valores etc. Su adaptación a los datos históricos tiene un buen grado de precisión y buena capacidad predictiva en la presencia de ruido.

3.3.1 Regresiones

La regresión lineal múltiple es una técnica de aprendizaje supervisado que utiliza múltiples variables predictoras X_i para pronosticar una variable de respuesta Y . Esto lo hace calculado los coeficientes betas de la regresión a través del método de mínimos cuadrados. (Hastie, 2009).

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j X_j$$

La regresión lineal múltiple se utiliza para modelar y predecir relaciones entre múltiples variables. Es útil cuando se desea determinar la relación entre una variable dependiente y dos o más variables independientes.⁷ Por ejemplo, se puede utilizar para predecir el precio de una casa en función de su tamaño, número de habitaciones, ubicación y edad.⁸

El horizonte de tiempo en el que funciona bien un modelo de regresión lineal múltiple depende de la serie de tiempo en cuestión. En general, los modelos de regresión lineal múltiple son adecuados para pronosticar valores futuros a corto y mediano plazo. Sin embargo, la precisión del pronóstico disminuye a medida que aumenta el horizonte de tiempo.

3.3.2 Red Neuronal Recurrente (LSTM)

Un modelo de memoria a largo plazo (LSTM) es un tipo de red neuronal recurrente que se utiliza en el aprendizaje profundo debido a sus grandes arquitecturas las cuales pueden ser entrenadas con éxito.⁴ Las redes LSTM son una extensión de las redes neuronales recurrentes (RNN) y se utilizan para modelar y predecir series de tiempo.⁴

En particular, las redes LSTM son útiles para modelar y predecir series de tiempo que tienen dependencias a largo plazo.⁴ A diferencia de los modelos ARIMA, que se basan en la regresión lineal, las redes LSTM utilizan una arquitectura de red neuronal para modelar la serie temporal.⁵

El horizonte de tiempo en el que funciona bien un modelo LSTM depende de la serie de tiempo en cuestión. En general, los modelos LSTM son adecuados para pronosticar valores futuros a corto y mediano plazo.⁶ Sin embargo, la precisión del pronóstico disminuye a medida que aumenta el horizonte de tiempo.⁶

3.3.3 XGBoost

XGBoost es una implementación eficiente de la técnica de boosting de gradiente para problemas de clasificación y regresión.⁹ Es rápido y eficiente, y se desempeña bien en una amplia gama de tareas de modelado predictivo.⁹ XGBoost también se puede utilizar para la predicción de series de tiempo, aunque requiere que el conjunto de datos de la serie temporal se transforme primero en un problema de aprendizaje supervisado.⁹ También requiere el uso de una técnica especializada para evaluar el modelo llamada validación hacia adelante, ya que evaluar el modelo utilizando validación cruzada k-fold daría como resultado resultados optimistas sesgados.⁹

Según Zhang y Zhan (2017) “es una biblioteca de refuerzo de gradiente distribuida diseñada para ser altamente eficiente, flexible y portátil. Proporciona un impulso de árbol paralelo que resuelve muchos problemas de ciencia de datos de una manera rápida y precisa. La idea es combinar cientos de árboles simples con baja precisión para construir un modelo más preciso. Cada iteración generará un nuevo árbol para el modelo”.¹⁰

3.4 Métricas para evaluar pronósticos

Según Ha et al (2018), la mejor forma de calcular el performance de un pronóstico es evaluando que tan preciso es este con respecto a la data real. A continuación, algunos tipos de errores para medir el performance de un pronóstico, sus ventajas y desventajas:

Error	Formula	Características
MSE (Mean Square Error)	$MSE = \frac{\sum e_i^2}{N}$	Castiga los outliers y al estar elevado al cuadrado no se ve afectado por positivos y negativos.
MAE	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	No tiene en cuenta la magnitud de los errores y es menos sensible a valores atípicos.
MAPE (Mean Absolute Percentage Error)	$MAPE = \frac{\sum \left \frac{e_i}{o_i} \right }{N}$	Fácil entendimiento debido a su representación como ratio, sensible a los outliers.
RMSPE (Root Mean Square Percentage Error)	$RMSPE = \sqrt{\frac{\sum \left(\frac{e_i}{o_i} \right)^2}{N}}$	Tiene independencia de la escala de la variable, Fácil entendimiento debido a su representación como ratio, castiga los outliers y no tiene sensibilidad a ellos.

Tabla 1 Métricas para evaluar pronósticos

Fuente: Velez, J. Validación De Medidas De Evaluación para el pronóstico de la tasa de cambio en Colombia (2016). Colegio de Estudios Superiores de Administración –CESA.¹²

3.5 Antecedentes

El patrón de arribo se ha estudiado en múltiples oportunidades, dentro de los estudios más recientes se encuentran los siguientes:

Mona Ebadi (2016), desarrolló un estudio utilizando un modelo de predicción basado en Elman y la red neuronal NARX y un algoritmo de propagación hacia atrás para pronosticar el patrón de arribo en un centro de llamadas. Los resultados pudieron ayudar a determinar la cantidad óptima

de agentes necesarios para reducir el tiempo de espera de los clientes en la cola de llamadas, lo que permite maximizar las ganancias y reducir los costos innecesarios.

Albrecht T, desarrolló un estudio que investiga las capacidades de los modelos de Machine Learning para pronosticar el patrón de arribo de llamadas en una línea evaluando la precisión y viabilidad de este pronóstico. En el estudio se compararon algoritmos de bosques aleatorio vs los modelos de series temporales más comunes y se determinó que el algoritmo de Machine Learning genera mejores rendimientos en la predicción. (Tobias Albrecht, 2021).

4. DESARROLLO OBJETIVO 1

Este apartado se pretende desarrollar las actividades que dan cumplimiento al objetivo 1: Identificar los métodos o modelos de pronóstico apropiados para predecir el patrón de arribo de llamadas con el fin de implementarlos en la serie estudiada y generar un portafolio de pronósticos que nos permita implementar un plan de evaluación.

A continuación, el desarrollo de cada una de las actividades requeridas para dar cumplimiento a este objetivo:

4.1 Análisis exploratorio de los datos.

Para el desarrollo de este proyecto se utilizará un patrón de arribo de llamadas de una línea de soporte telefónica, se realiza el importe de los datos en python y se obtienen los siguientes resultados:

```
      Date  Month  Year  Week  day  Volume
0  2022-05-22    5  2022    20    7     85
1  2022-05-23    5  2022    21    1    163
2  2022-05-24    5  2022    21    2    233
3  2022-05-25    5  2022    21    3    282
4  2022-05-26    5  2022    21    4    303
..      ...    ...    ...    ...    ...
318 2023-04-05    4  2023    14    3    134
319 2023-04-06    4  2023    14    4    168
320 2023-04-07    4  2023    14    5    129
321 2023-04-08    4  2023    14    6     81
322 2023-04-09    4  2023    14    7     48
```

[323 rows x 8 columns]

Ilustración 1 Importe de los datos

El set de datos cuenta con 323 registros y 6 columnas en total; cada registro representa un día de operación de la línea y los atributos en las columnas contienen la fecha, mes, año, volumen de llamadas ingresado en el día, la semana del año y el día de la semana.

Con respecto a los tipos de datos, estos ingresaron inicialmente bajo esta configuración:

Date	datetime64[ns]
Month	int32
Year	int32
Week	int32
day	int32
Volume	int32

Para el desarrollo del análisis de la serie de tiempo se debe crear un objeto de este tipo, por lo que las columnas a utilizar son “Volume” y “Date”, la primera representará los valores de la serie de tiempo y la segunda funcionará como índice de este objeto.

4.1.1 Presentación de la serie de tiempo

Los datos obtenidos muestran un registro diario de llamadas desde el 22 de mayo del 2022 hasta el 9 de abril del 2023. La frecuencia de los registros es diaria, a continuación, una primera visual de la serie de tiempo:

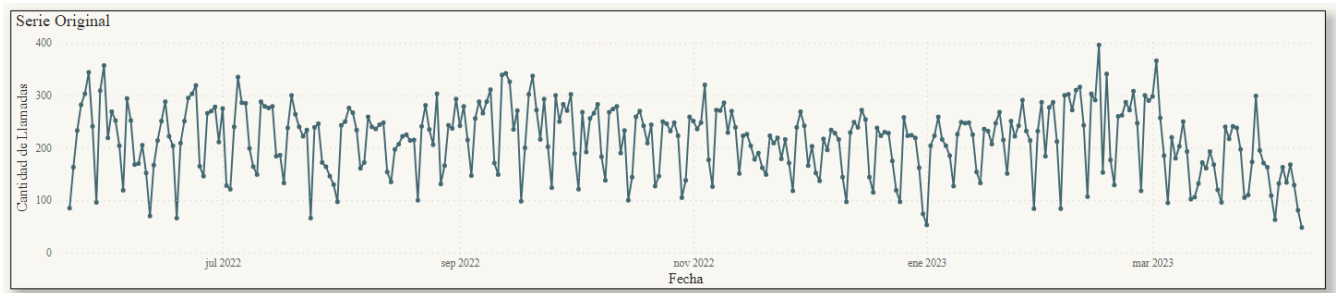


Ilustración 2 Serie de tiempo original

De forma general, se puede evidenciar que la serie de tiempo parece ser estacionaria, esto significa que a lo largo del tiempo se mantiene oscilando alrededor de un mismo valor, sin embargo, en ciertas ventanas de tiempo parece tener componente de tendencia.

Por otro lado, se puede observar que la serie presenta estacionalidad, se logran ver patrones semanales bastante marcados, esto era de esperarse dado el tipo de proceso analizado.

4.1.2 Descomposición de la serie de tiempo

Para hacer un análisis un poco más detallado de los componentes de tendencia, estacionalidad y residuos de la serie de tiempo se procede a hacer su descomposición:

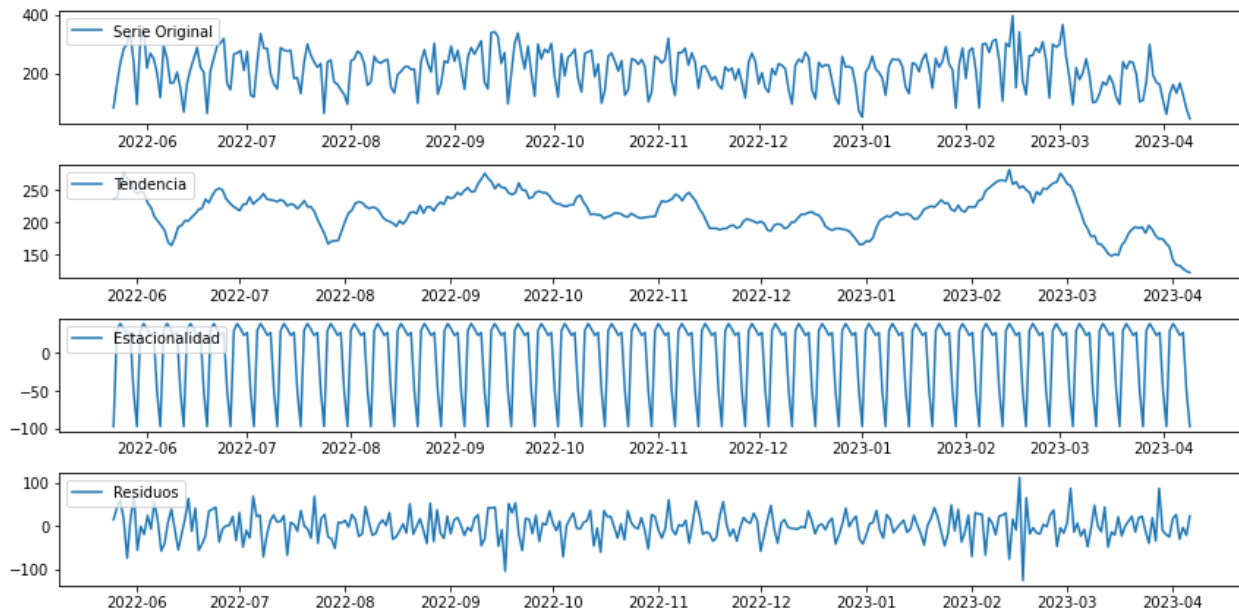


Ilustración 3 Descomposición de la serie original

A pesar de tener aislada la componente de tendencia, no se logra determinar con certeza si la serie es estacionaria. De forma global se podría considerarse estacionaria, pero al analizar ventanas de tiempo específicas, se observa la presencia del componente de tendencia. Por ejemplo, en el intervalo desde noviembre de 2022 hasta enero de 2023, la serie muestra una tendencia a la baja, mientras que, desde enero hasta marzo de 2023, presenta una tendencia al alza.

Por otro lado, se confirma la presencia de una fuerte estacionalidad en la serie, la cual se mantiene homogénea a lo largo de todo el periodo analizado. Los residuos exhiben un comportamiento satisfactorio, ya que están centrados en cero y carecen de un patrón discernible, pudiendo considerarse como aleatorios.

4.1.3 Análisis de Tendencia

Dado que la existencia de tendencia en la serie no es muy clara, se ejecutará la prueba Kendall para definir si existe o no tendencia en la serie analizada, bajo las siguientes hipótesis:

$$H_0: \text{La serie no tiene tendencia}$$
$$H_1: \text{La serie si presenta tendencia}$$

Resultados de la prueba:

Estadístico de Kendall: -0.10531110930914277

Valor p: 0.004841708117507821

Conclusión: Existe evidencia estadística de tendencia en la serie de tiempo.

A un nivel de significancia del 0.05 se concluye que hay evidencia suficiente para soportar que la serie si presenta tendencia. Por lo que se descarta que la serie en su forma original sea estacionaria.

4.1.4 Diferenciación de la serie

Dado que la serie de tiempo presenta tendencia, se le aplica un grado de diferenciación con el fin de intentar eliminar esta tendencia y analizar la autocorrelación. A continuación, la descomposición de la serie diferenciada:

4.1.4.1 Descomposición de la serie diferenciada

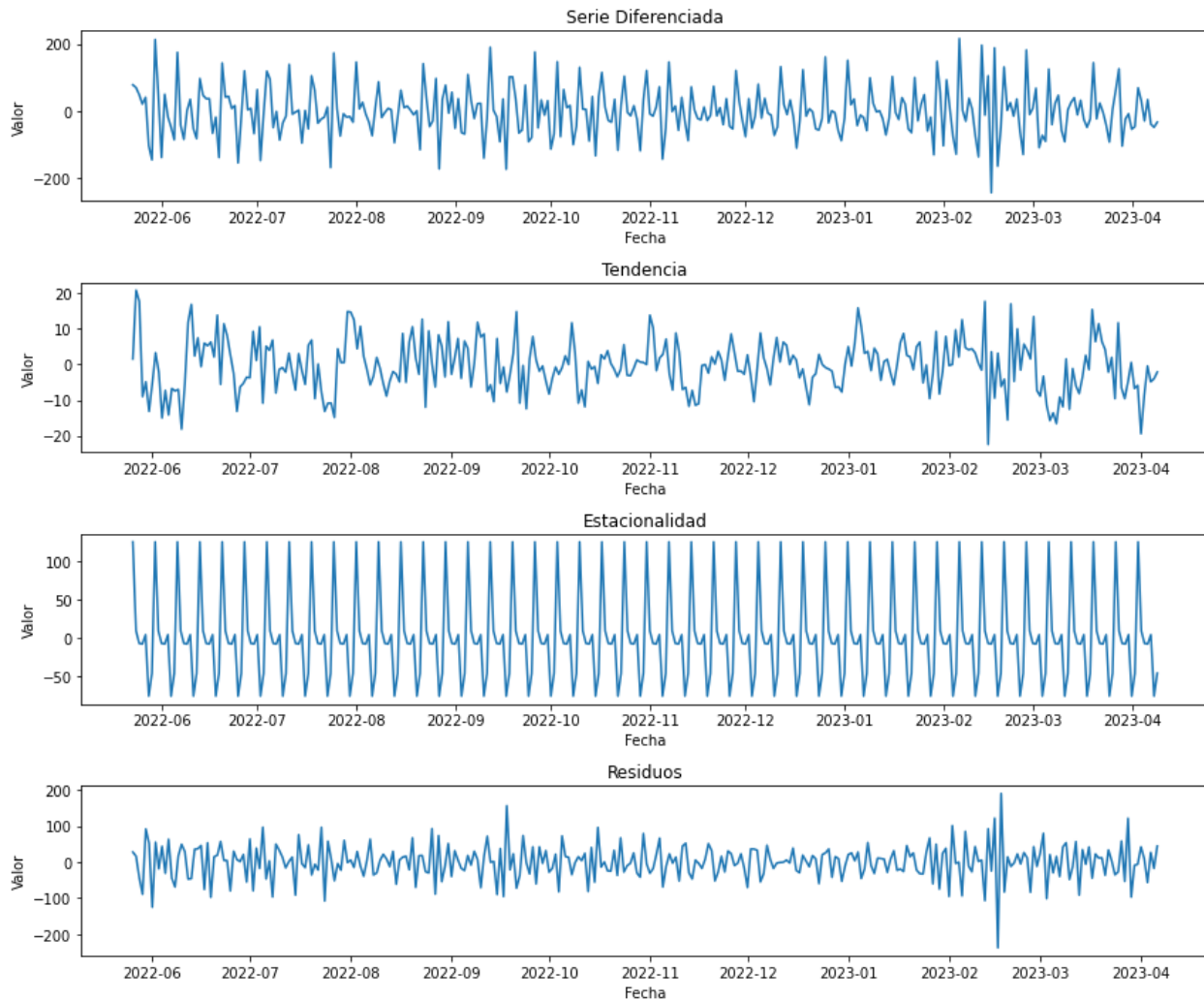


Ilustración 4 Descomposición de la serie diferenciada

Se puede observar que la serie diferenciada presenta estacionariedad, esto se validará ejecutando nuevamente la prueba de Kendall, pero ahora para la serie diferenciada; cabe destacar que la serie diferenciada sigue presentando una clara estacionalidad.

4.1.4.2 Análisis de tendencia para la serie diferenciada

Se ejecuta nuevamente la prueba de Kendall para la serie diferenciada:

H₀: La serie no tiene tendencia
H₁: La serie si presenta tendencia

Estadístico de Kendall: -0.031795639262881585

Valor p: 0.3955708933631924

Conclusión: No hay evidencia estadística de tendencia en la serie.

A un nivel de significancia del 0.05 se concluye que NO hay evidencia suficiente para soportar que la serie si presenta tendencia. Por lo que se concluye que la serie diferenciada, es estacionaria.

4.1.4.3 Análisis de estacionalidad para la serie diferenciada

La estacionalidad de esta serie diferenciada es bastante clara, pero podemos tener mayor detalle de está analizando las gráficas ACF y PACF.

La gráfica de ACF y PACF muestran la correlación entre una serie de tiempo y sus valores rezagados, pero la gráfica PACF elimina la influencia de los rezagos intermedios. En otras palabras, muestra la correlación directa entre dos puntos en el tiempo después de eliminar la correlación indirecta a través de otros rezagos.

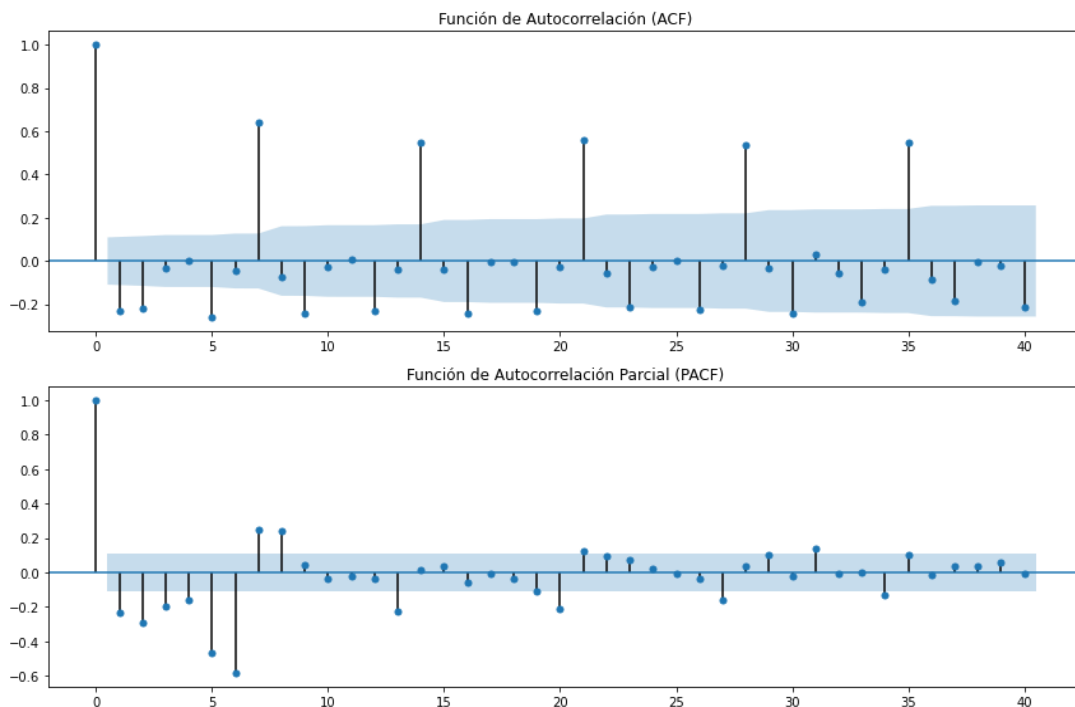


Ilustración 5 Gráficas ACF y PACF de la serie de tiempo

Por la forma en que se presentan los rezagos en la gráfica ACF se puede evidenciar que existe una fuerte correlación entre la serie y su séptimo rezago, por otro lado, en la gráfica PACF podemos ver como este rezago cada vez tiene menos autocorrelación, esto puede indicar la presencia de una media móvil.

A pesar de que no cabe duda de la presencia de estacionalidad en la serie de tiempo, la forma de onda sinusoidal de las gráficas ACF y PACF reafirman esta conclusión.

4.1.5 Análisis de estacionariedad

Dado que la prueba de tendencia mostró la existencia de esta en la serie original, se puede concluir de por sí que la serie no es estacionaria. Sin embargo, podemos reafirmar esta conclusión aplicando una prueba de estacionariedad a la serie de tiempo.

Se utilizará la prueba Dickey-Fuller para probar si existe estacionariedad en la serie estudiada:

H₀: La serie es no estacionaria: tiene raíz unitaria

H₁: La serie es estacionaria: no tiene raíz unitaria

Resultados de la prueba Dickey-Fuller:

Resultado del test Dickey-Fuller:

Estadístico de prueba: -2.2586023114912974

Valor p: 0.1856672060248603

Número de retrasos utilizados: 13

Número de observaciones utilizadas: 309

Valores críticos:

1%: -3.45169128009473

5%: -2.8709394227049154

10%: -2.5717780602423517

Conclusión: La serie de tiempo no es estacionaria (no se rechaza la hipótesis nula de presencia de raíz unitaria).

Con un nivel de significancia de 0.05 se concluye que no hay evidencia suficiente para decir que la serie es estacionaria, este resultado es coherente con la prueba de tendencia el cual mostró que la serie efectivamente si presenta una tendencia en el tiempo.

4.1.6 Análisis de rezagos

El análisis de rezagos se emplea para evaluar la correlación entre los valores actuales y los valores rezagados en n períodos. Examinemos la relación que se establece entre cada valor y sus primeros 8 rezagos. Se considera que este número es adecuado, dado que se espera identificar una relación

que refleje la estructura de los días de la semana entre sí, es decir, lunes a lunes, martes a martes, y así sucesivamente.

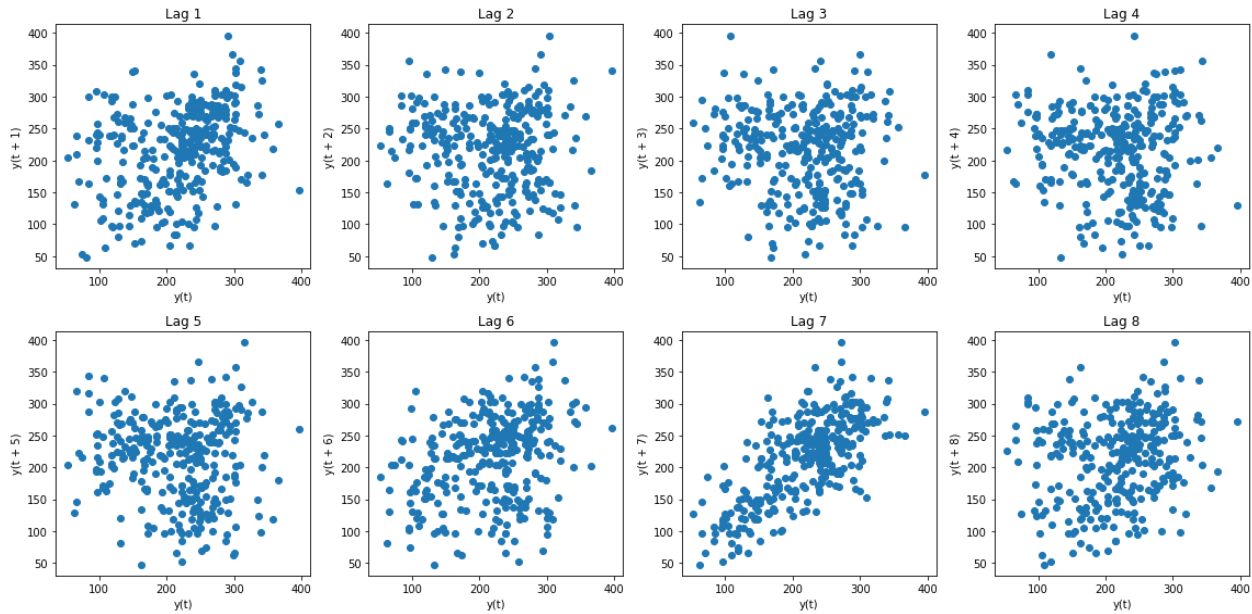


Ilustración 6 Rezagos de la serie de tiempo

Como era de esperarse, se evidencia una fuerte correlación de la serie con su séptimo rezago.

4.1.7 Análisis de la distribución de la variable

Como último punto del análisis exploratorio de los datos, se analiza la distribución del volumen de interacciones por día:

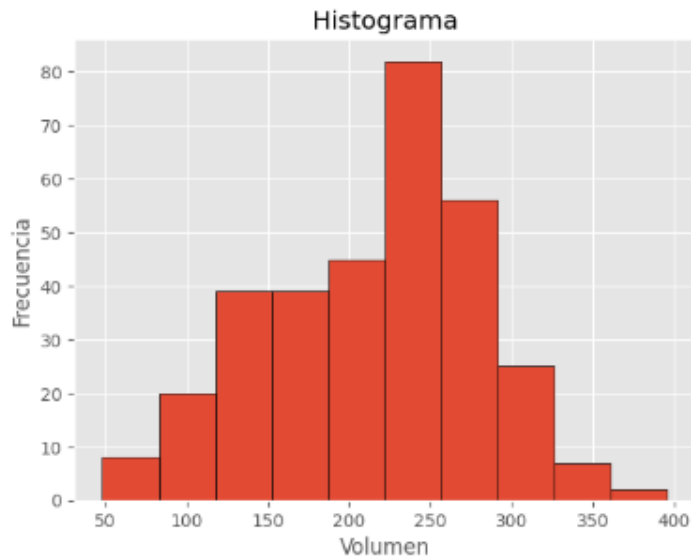


Ilustración 7 Histograma de volumen de llamadas por día

Con respecto a la distribución del volumen, del histograma podemos concluir que esta presenta una distribución aproximadamente normal, con una media 214.82 y desviación estándar de 66.3. Los valores de la variable varían entre 48 y 396 interacciones.

Se utilizará el método de IQR para validar la presencia de valores atípicos, a continuación el cálculos de límite inferior y superior:

$$LL = Q1 - 1.5(IQR) = 166 - 1.5 \cdot 99 = 17.5$$

$$UL = Q3 + 1.5(IQR) = 265 + 1.5 \cdot 99 = 413.5$$

Dado que ninguno de los valores de llamadas que ingresan por día es mayor o menor a los límites, se puede concluir que no se identifican la presencia de valores atípicos en la serie, por otro lado, y no menos importante, la serie de tiempo tampoco muestra la presencia de valores nulos.

4.2 Definición del set de entrenamiento y el set de evaluación.

Es crucial tener en cuenta que, al analizar series temporales, no se debe llevar a cabo una división aleatoria o estratificada aleatoria de los datos, ya que es fundamental preservar la estructura temporal y evitar posibles filtraciones de información futura en el conjunto de entrenamiento.

En el contexto de series temporales, la forma más común de dividir los datos en conjuntos de entrenamiento y prueba es mediante un enfoque de "punto de corte" o "split-point". Bajo este método, se elige un punto en el tiempo que actúa como límite entre el conjunto de entrenamiento y el conjunto de prueba.

Al considerar el contexto del negocio, el plan de contratación debe formularse dos meses antes del momento de dotación de personal entrenado. Por lo tanto, se espera que el modelo seleccionado sea capaz de realizar pronósticos con al menos dos meses de anticipación. Con este objetivo en mente, resulta apropiado realizar la evaluación de los modelos con pronósticos a un plazo de dos meses. Esta meta se puede lograr mediante una estrategia de división 80/20, lo que generará 258 registros para el entrenamiento de los modelos y 65 registros para la evaluación de los pronósticos.

4.3 Selección de modelos y algoritmos

Para definir los modelos o algoritmos a implementar para pronosticar el patrón de arribo de llamadas se debe tener en consideración que la serie de tiempo presenta tendencia, una fuerte estacionalidad y que se requiere un pronóstico de más de 60 pasos a futuro, esto es un horizonte bastante extenso. Adicional a esto, se considera pertinente considerar cada uno de los enfoques en esta selección, por lo que se implementará un modelo autorregresivo, uno de suavizamiento y un algoritmo de aprendizaje automático.

Teniendo todo esto presente, los modelos a implementar son los siguientes:

- 1) SARIMA (Seasonal Autoregressive Integrated Moving Average): SARIMA es una extensión del modelo ARIMA que se utiliza para modelar series de tiempo con estacionalidad. SARIMA puede ser muy eficaz para predecir valores futuros de series de tiempo que muestran patrones estacionales.
- 2) Holt-Winters: Holt-Winters es un método de suavizado exponencial que se utiliza para el pronóstico de series de tiempo con tendencia y estacionalidad. Es capaz de modelar la tendencia y la estacionalidad de manera separada y utiliza un enfoque de suavizado para predecir valores futuros.
- 3) LSTM (Long Short-Term Memory): LSTM es un tipo de red neuronal recurrente que se utiliza para el pronóstico de series de tiempo. LSTM puede manejar secuencias de tiempo largas y complejas, y puede aprender patrones a largo plazo en los datos.

4.4 Modelo SARIMA

4.4.1 Implementación Modelo SARIMA

Para definir los parámetros del modelo SARIMA se ejecutó un proceso de optimización iniciando en $m = 7$ como grado de estacionalidad, dado que el componente estacional de la serie es semanal y $d = 1$ como grado de diferenciación, dado que la serie requiere una diferenciación para llegar a ser estacionaria.

Resultado del modelo optimizado:

Statespace Model Results			
Dep. Variable:	y	No. Observations:	258
Model:	SARIMAX(2, 1, 2)x(2, 0, 1, 7)	Log Likelihood	-1300.339
Date:	Sat, 18 Nov 2023	AIC	2618.677
Time:	12:03:48	BIC	2650.619
Sample:	0	HQIC	2631.523
	- 258		
Covariance Type:	opg		

Tabla 2 Resultado del modelo Sarima

Ljung-Box (Q):	26.97	Jarque-Bera (JB):	0.38
Prob(Q):	0.94	Prob(JB):	0.83
Heteroskedasticity (H):	0.52	Skew:	0.06
Prob(H) (two-sided):	0.00	Kurtosis:	2.86

Tabla 3 Resultado de pruebas del modelo Sarima

El modelo Sarima optimizado tiene los siguientes parámetros SARIMAX(2, 1, 2)x(2, 0, 1, 7), a continuación se interpretaran estos parámetros.

4.4.2 Interpretación del Modelo SARIMA

Componentes del proceso autorregresivo (AR):

El primer número (2) indica el orden del componente AR no estacional (p), lo que significa que se incluyen los rezagos inmediatos y los dos rezagos anteriores en el modelo.

Componentes del proceso de integración (I):

El segundo número (1) indica el grado de diferenciación (d), que en este caso es 1. Esto implica que se ha aplicado una diferenciación de primer orden a la serie para lograr estacionariedad.

Componentes del proceso de media móvil (MA):

El tercer número (2) indica el orden del componente MA no estacional (q), lo que significa que se incluyen los errores inmediatos y los dos errores anteriores en el modelo.

Componentes del proceso de estacionalidad (SAR):

El cuarto número (2) indica el orden del componente AR estacional (P) para los rezagos estacionales. En este caso, se incluyen los rezagos estacionales inmediatos y los dos rezagos estacionales anteriores en el modelo.

Componentes del proceso de integración estacional (SI):

El quinto número (0) indica el grado de diferenciación estacional (D), que en este caso es 0. Esto significa que no se ha aplicado diferenciación estacional.

Componentes del proceso de media móvil estacional (SMA):

El sexto número (1) indica el orden del componente MA estacional (Q) para los errores estacionales. En este caso, se incluyen los errores estacionales inmediatos en el modelo.

El número 7 indica que la estacionalidad se repite cada 7 periodos de tiempo. Esto sugiere que hay una variación estacional que sigue un patrón semanal, ya que 7 días representan una semana.

4.4.3 Validación de residuales SARIMA

Para asegurar la viabilidad del modelo se deben garantizar los supuestos de los residuales, estos consideran que los errores son distribuyen normal, de forma aleatoria e independiente, con media 0 y varianza constante (homocedasticidad). A continuación, la validación de cada uno de estos supuestos:

- Supuesto de normalidad de en los errores

Para este supuesto se validará la siguiente hipótesis:

$$H_0: \text{Los residuos distribuyen normal}$$
$$H_1: \text{Los residuos no distribuyen normal}$$

Para validar estas hipótesis, se utilizará la grafica QQ y una prueba shapiro:

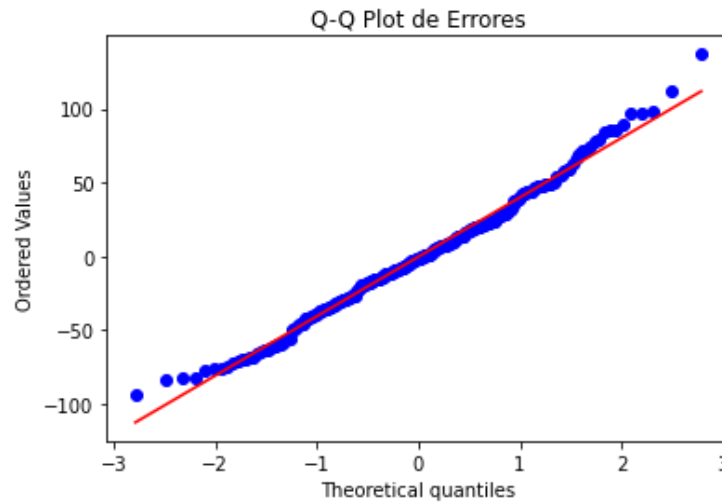


Ilustración 8 Grafica QQ para residuales

Resultado de prueba Shapiro:

Estadístico de prueba Shapiro: 0.9905636310577393

Valor p: 0.0939718633890152

De forma visual podemos se puede decir que la distribución de los errores no difiere mucho de la normal, esto se visualiza dado que los puntos no se alejan de forma considerable de la recta.

Por otro lado, con un nivel de significancia de 0.05, no hay evidencia para rechazar la hipótesis nula, por lo que no se puede concluir que la distribución de los residuos no sea normal.

- Supuesto de media cero

Para este supuesto se validará la siguiente hipótesis:

$$H_0: \text{Media de residuales} = 0$$
$$H_1: \text{Media de residuales} \neq 0$$

Para la validación de esta hipótesis se desarrollará una prueba t, a continuación los resultados de esta prueba:

Estadístico de prueba t: -0.17090242230111946,

Valor p: 0.864434966211699

Con un nivel de significancia de 0.05 no se rechaza la hipótesis nula por lo que no hay evidencia significativa para decir que la media de los residuales es diferente de cero.

- Supuesto de homocedasticidad

El supuesto de homocedasticidad hace referencia a que la varianza de los errores es constante en el tiempo, para la validación de este supuesto se analizará la gráfica de dispersión de residuos y calculando el indicador Heteroskedasticity (H). Hipótesis para validar:

H₀: Los residuos son homocedasticos
H₁: Los residuo son heterocedasticos

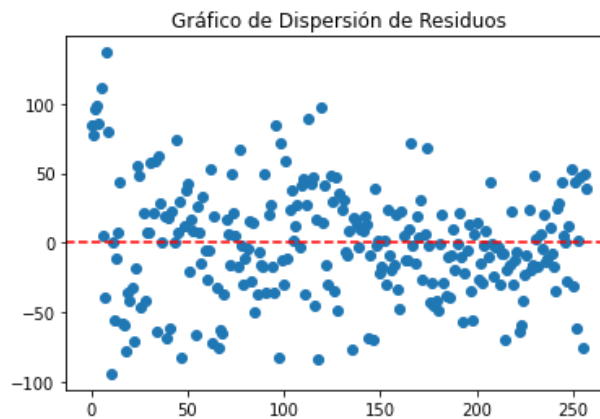


Ilustración 9 Grafica de dispersión de residuos

Heteroskedasticity (H): El valor de 0.52 es un indicador de la presencia de heteroscedasticidad en los residuales, lo que significa que la varianza de los residuales no es constante a lo largo del tiempo. Un valor bajo de H sugiere que los residuales tienen una varianza constante. En este caso, el valor de 0.52 indica que los residuales tienen una varianza relativamente constante.

De forma visual también se puede validar que la varianza en el tiempo es constante, tal vez se presenta una variación al inicio de la serie de tiempo, pero en general se puede decir que la varianza es la misma en el tiempo.

- Supuesto de independencia

Para este supuesto se validará la siguiente hipótesis:

H₀: Los residuales son independientes
H₁: Los residuales son dependientes

Para la validación de esta hipótesis, se desarrollará una prueba Ljung-Box, el resultado de esta prueba es 26.97 para el estadístico Q, que se utiliza para evaluar la independencia de los residuos. Un valor alto de Q sugiere que los residuales no son independientes, lo cual es indeseable en el análisis de series temporales. Sin embargo, el p-valor asociado (Prob(Q)) es 0.94, lo que indica que no hay evidencia significativa de falta de independencia en los residuales. En este caso, no hay evidencia de autocorrelación significativa en los residuales.

4.4.4 Pronostico del modelo SARIMA

Con el modelo previamente ajustado se calcula el pronostico en el set de pruebas y se genera el siguiente resultado:

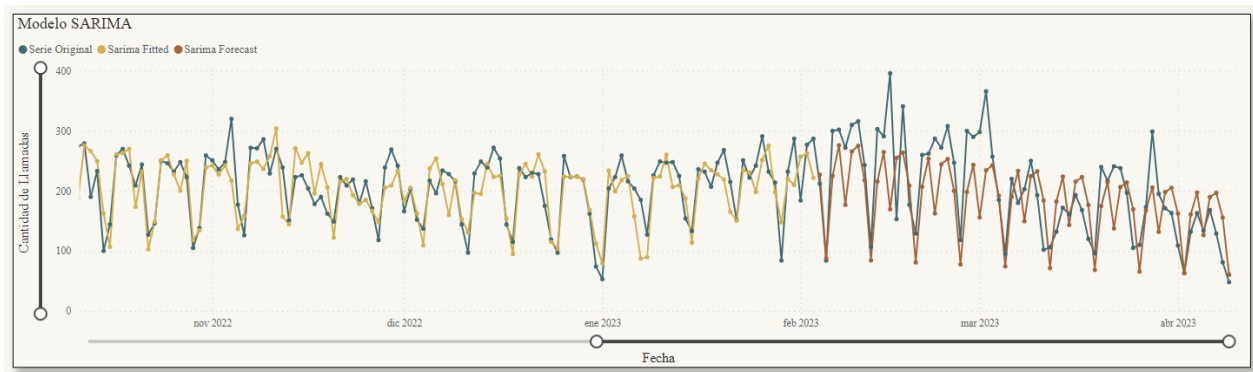


Ilustración 10 Pronostico de modelo Sarima

De forma general se ve un buen ajuste de las predicciones del modelo, se puede ver como se siguen los patrones de estacionalidad y los puntos pronosticados no están muy distantes de los reales. El modelo presentó un muy buen ajuste durante el entrenamiento y logra mantener el componente de tendencia durante la prueba.

4.5 Modelo Suavizamiento Exponencial Triple (SET)

El suavizamiento exponencial es especialmente útil cuando se trabaja con datos que muestran tendencias y patrones suaves, pero no poseen patrones estacionales o cíclicos pronunciados. Dado que nuestra serie presenta una fuerte estacionalidad, no se espera que este método genere un buen ajuste a la serie. Analicemos su comportamiento:

4.5.1 Implementación Suavizamiento Exponencial Triple

Para la implementación del modelo de suavizamiento exponencial triple, no se deben optimizar parámetros, pero si se debe especificar si la tendencial y estacionalidad de la serie son aditivas.

ExponentialSmoothing Model Results			
Dep. Variable:	endog	No. Observations:	258
Model:	ExponentialSmoothing	SSE	348649.679
Optimized:	True	AIC	1881.887
Trend:	Additive	BIC	1920.969
Seasonal:	Additive	AICC	1883.379
Seasonal Periods:	7	Date:	Sat, 18 Nov 2023
Box-Cox:	False	Time:	12:24:45
Box-Cox Coeff.:	None		

Tabla 4 Resultamos modelo de suavizamiento ET

A continuación, la interpretación del ajuste de este modelo:

4.5.2 Interpretación del modelo SET

El modelo de suavizamiento exponencial triple utilizado para el análisis tiene como variable dependiente la serie temporal. Este modelo se ha optimizado y se caracteriza por tener una tendencia aditiva, lo que implica que la relación entre las variables dependientes e independientes se representa mediante la suma de componentes. Asimismo, el componente estacional del modelo es también aditivo, lo que sugiere que las variaciones estacionales se suman de manera directa a la tendencia general.

Por otro lado, el modelo considera periodos estacionales de longitud 7. Además, no se ha aplicado la transformación de Box-Cox en este caso, ya que el parámetro Box-Cox se mantiene en "False" y su coeficiente asociado es "None"; cabe resaltar que la transformación de Box-Cox se utiliza en modelos de suavizado exponencial triple cuando se busca estabilizar la varianza de la serie temporal. Dado que la varianza de la serie muestra ser constante, esto no fue requerido.

4.5.3 Validación de residuales SET

Para asegurar la viabilidad del modelo se deben garantizar los supuestos de los residuales, estos consideran que los errores son distribuyen normal, de forma aleatoria e independiente, con media 0 y varianza constante (homocedasticidad). A continuación, la validación de cada uno de estos supuestos:

- Supuesto de normalidad de en los errores

Para este supuesto se validará la siguiente hipótesis:

$$H_0: \text{Los residuos distribuyen normal}$$

$$H_1: \text{Los residuos no distribuyen normal}$$

Para validar estas hipótesis, se utilizará la gráfica QQ:

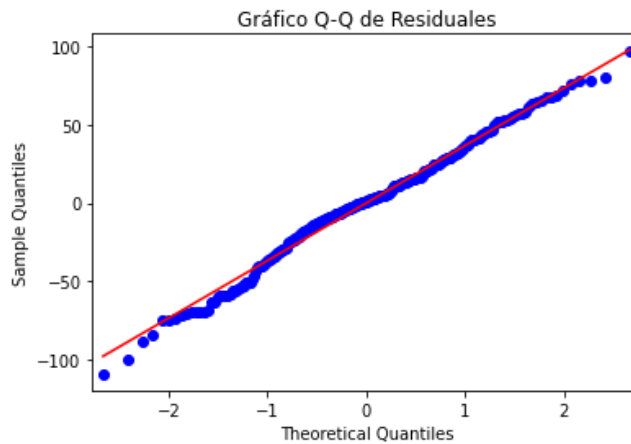


Ilustración 11 Grafica QQ Modelo Suavizamiento ET

De forma visual podemos decir que la distribución de los errores no difiere mucho de la normal, esto se visualiza dado que los puntos no se alejan de forma considerable de la recta.

- Supuesto de media cero

Para este supuesto se validará la siguiente hipótesis:

$$H_0: \text{Media de residuales} = 0$$

$$H_1: \text{Media de residuales} \neq 0$$

Para la validación de esta hipótesis se desarrollará una prueba t y un análisis de la distribución de los residuales, a continuación, los resultados de esta prueba:

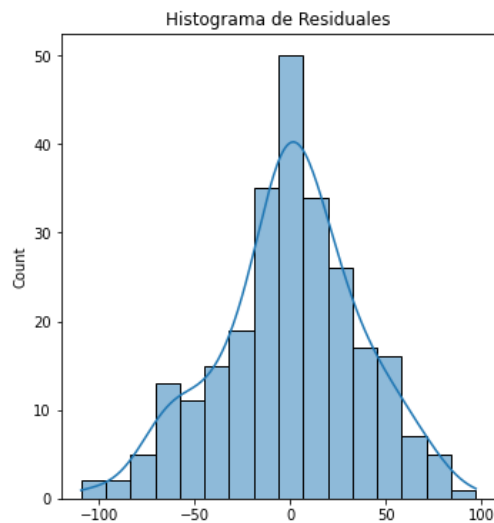


Ilustración 12 Distribución de residuales SET

Estadístico de prueba t: 0.0008143517303851033

Valor p: 0.9993508731513379

De forma visual podemos observar que la distribución de los residuales está centrada en cero, adicional a esto, con un nivel de significancia de 0.05 no se rechaza la hipótesis nula por lo que no hay evidencia significativa para decir que la media de los residuales es diferente de cero.

- Supuesto de homocedasticidad

El supuesto de homocedasticidad hace referencia a que la varianza de los errores es constante en el tiempo, para la validación de este supuesto se analizará la gráfica de dispersión de residuos. Hipótesis para validar:

H₀: Los residuos son homocedasticos

H₁: Los residuo son heterocedasticos

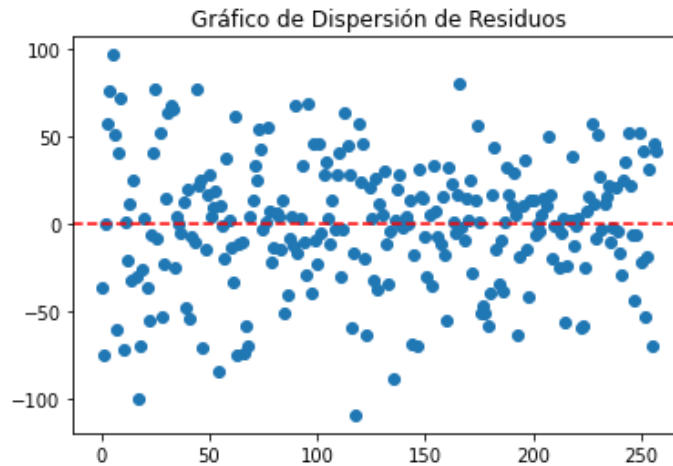


Ilustración 13 Gráfico de dispersión de residuos SET

De forma visual se puede validar que la varianza en el tiempo es constante, por lo que no se rechaza la hipótesis nula y se concluye que se cumple el supuesto de homocedasticidad en los residuales.

- Supuesto de independencia

Para este supuesto se validará la siguiente hipótesis:

H₀: Los residuales son independientes

H₁: Los residuales son dependientes

Para la validación de esta prueba se utilizará la gráfica de función de autocorrelación de los residuales:

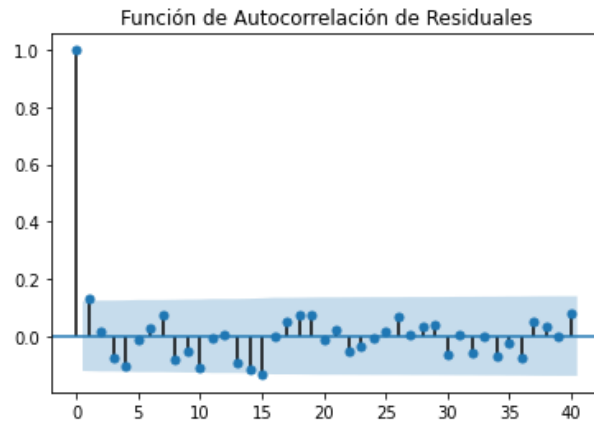


Ilustración 14 Función de autocorrelación de residuales SET

Dado que ninguno de los rezagos de la serie de los errores es significativo, esto se valida visualmente porque ninguno sobrepasa la franja azul, se puede concluir que los residuales no tienen dependencia entre sí, por lo que se puede considerar validado el supuesto de independencia.

4.5.4 Pronóstico del modelo SET

Con el modelo previamente ajustado se calcula el pronóstico en el set de pruebas y se genera el siguiente resultado:

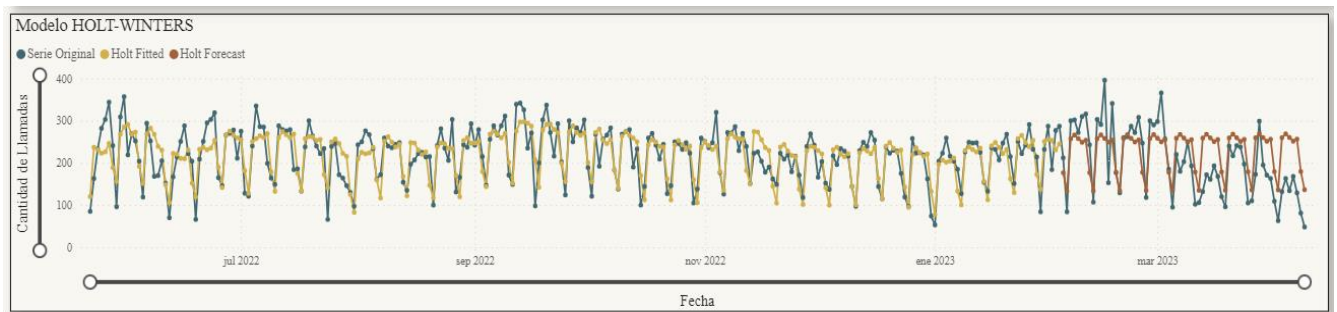


Ilustración 15 Pronóstico del modelo SET

De forma general se ve un buen ajuste de las predicciones del modelo, se puede ver como se siguen los patrones de estacionalidad y los puntos pronosticados no están muy distantes de los reales. El modelo presentó un muy buen ajuste durante el entrenamiento, pero al parecer no logra sostener la tendencia en el set de pruebas.

4.6 Redes neuronales LSTM (Long Short-Term Memory)

4.6.1 Implementación del modelo LSTM

Para la implementación de este modelo se debe implementar una normalización de los datos, dado que esto es requerido por el algoritmo de LSTM. Por otro lado, no se ejecuta algún proceso de optimización de parámetros.

El modelo LSTM se construye utilizando la biblioteca Keras. La capa LSTM tiene 50 unidades con la función de activación ReLU y toma como entrada secuencias de longitud 7 con una sola característica en cada paso de tiempo. Se añade una capa densa con una sola unidad para la salida.

El modelo se compila utilizando el optimizador 'adam' y la función de pérdida 'mean_squared_error' (error cuadrático medio).

```
[101] ✓ 1 sequence_length = 7 # 7 días para capturar la estacionalidad semanal
2
3 X_train, y_train = create_sequences(train_data_scaled, sequence_length)
4 X_test, y_test = create_sequences(test_data_scaled, sequence_length)

[102] ✓ 1 # Crea el modelo LSTM
2 model = Sequential()
3 model.add(LSTM(50, activation='relu', input_shape=(sequence_length, 1)))
4 model.add(Dense(1))
5 model.compile(optimizer='adam', loss='mean_squared_error')

1 # Entrena el modelo
2 history = model.fit(X_train, y_train, epochs=100, batch_size=32)
3
```

Ilustración 16 Configuración de red neuronal LSTM

4.6.2 Interpretación del modelo LSTM

La curva de pérdida (loss) que se presenta durante el entrenamiento proporciona información sobre cómo está evolucionando el rendimiento del modelo a medida que se ajustan los pesos durante cada época. En este caso, la pérdida disminuye a lo largo de las épocas, lo cual es un indicador positivo.

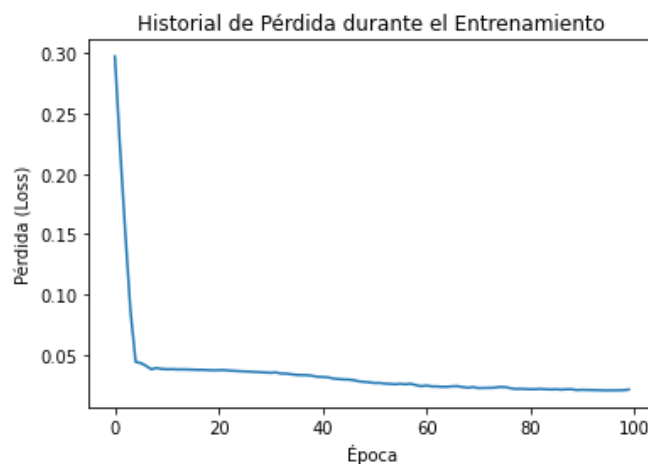


Ilustración 17 Gráfica de pérdida LSTM

Para este modelo el Loss tiene el siguiente comportamiento:

Epoch 1: La pérdida inicial es 0.297.

Epoch 2: La pérdida disminuye a 0.221.

Epoch 3: La pérdida sigue disminuyendo a 0.151.

Epoch 4: La pérdida disminuye significativamente a 0.086.

Epoch 5 a 14: La pérdida continúa disminuyendo gradualmente, alcanzando un valor de alrededor de 0.037 en la Epoch 14.

Una curva de pérdida decreciente es un buen indicador durante el entrenamiento, ya que sugiere que el modelo está aprendiendo de manera efectiva y ajustando sus pesos para mejorar la precisión en los datos de entrenamiento.

4.6.3 Validación de residuales LSTM

Para asegurar la viabilidad del modelo se deben garantizar los supuestos de los residuales, estos consideran que los errores distribuyen normal, de forma aleatoria e independiente, con media 0 y varianza constante (homocedasticidad). A continuación, la validación de cada uno de estos supuestos:

- Supuesto de normalidad de en los errores

Para este supuesto se validará la siguiente hipótesis:

H₀: Los residuos distribuyen normal
H₁: Los residuos no distribuyen normal

Para validar estas hipótesis, se utilizará la grafica de distribución de los residuales:

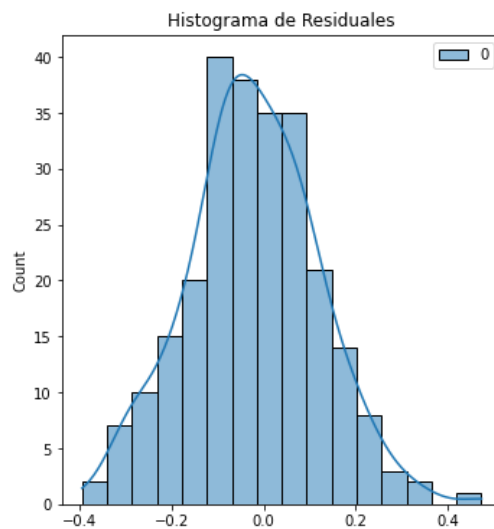


Ilustración 18 Distribución de residuales LSTM

Visualmente, se puede evidenciar que la distribución de residuales esta ligeramente sesgada, los errores no siguen una distribución normal.

- Supuesto de media cero

Para este supuesto se validará la siguiente hipótesis:

$$H_0: \text{Media de residuales} = 0$$

$$H_1: \text{Media de residuales} \neq 0$$

Para la validación de esta hipótesis se desarrollará un análisis visual de la grafica de distribución de residuales LSTM.

La distribución de los residuales esta sesgada y centrada antes del cero, por este motivo, podemos rechazar la hipótesis nula y concluir que la media de los residuales es diferente de cero.

- Supuesto de homocedasticidad

El supuesto de homocedasticidad hace referencia a que la varianza de los errores es constante en el tiempo, para la validación de este supuesto se analizará la gráfica de dispersión de residuos. Hipótesis para validar:

$$H_0: \text{Los residuos son homocedasticos}$$

$$H_1: \text{Los residuo son heterocedasticos}$$

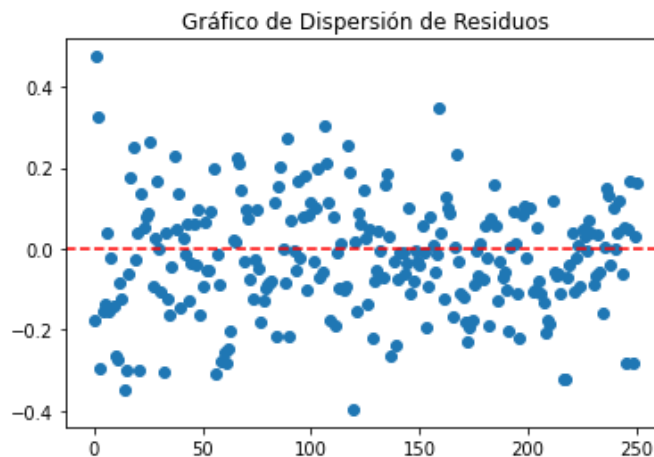


Ilustración 19 Grafico de dispersión de residuos LSTM

De forma visual se puede validar que la varianza en el tiempo es constante, por lo que no se rechaza la hipótesis nula y se concluye que se cumple el supuesto de homocedasticidad en los residuales.

- Supuesto de independencia

Para este supuesto se validará la siguiente hipótesis:

H₀: Los residuales son independientes

H₁: Los residuales son dependientes

Para la validación de esta prueba se utilizará la gráfica de función de autocorrelación de los residuales:

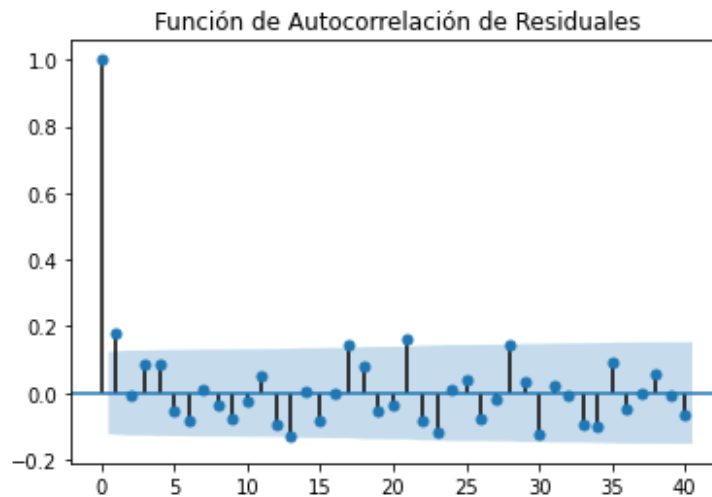
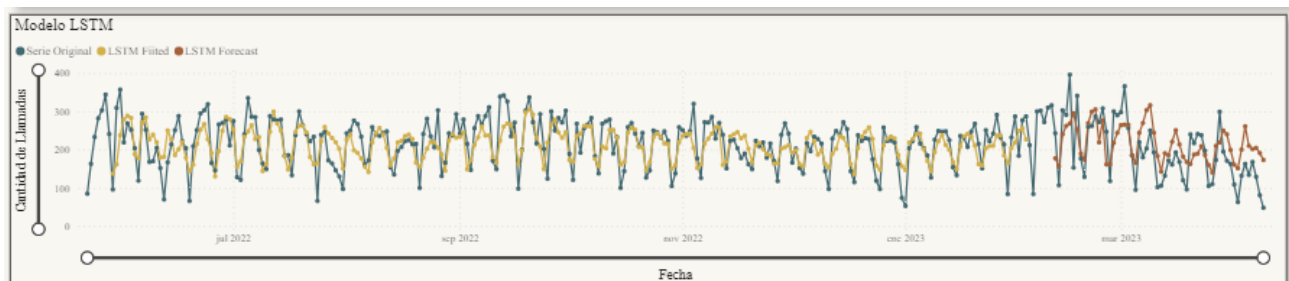


Ilustración 20 Gráfica de autocorrelación residuales LSTM

Dado que alguno de los rezagos sobrepasa la línea de significancia, no se puede concluir que los errores son independientes.

4.6.4 Pronóstico del modelo LSTM

Con el modelo previamente ajustado se calcula el pronóstico en el set de pruebas y se genera el siguiente resultado:



De forma general se ve un buen ajuste de las predicciones del modelo, se puede ver como se siguen los patrones de estacionalidad y los puntos pronosticados no están muy distantes de los reales. El

modelo presentó un muy buen ajuste durante el entrenamiento y logra sostener la tendencia en el set de pruebas.

5 DESARROLLO OBJETIVO 2

Este apartado pretende desarrollar las actividades que dan cumplimiento al objetivo 2: Identificar y aplicar los métodos de evaluación de pronóstico más apropiados para el caso de estudio con el fin de generar métricas de desempeño y seleccionar el modelo más apropiado

A continuación, el desarrollo de cada una de las actividades requeridas para dar cumplimiento a este objetivo:

5.1 Análisis de métricas de error

De la investigación realizada se resumen las siguientes métricas de error las cuales se consideran pueden funcionar para evaluar el caso de estudio:

Error Cuadrado Medio (MSE): El MSE calcula la diferencia cuadrada promedio entre los valores pronosticados y los valores reales. Se considera una medida de dispersión de los errores. Cuanto menor sea el valor de MSE, mejor será la precisión del pronóstico. Esta métrica penaliza más los errores grandes debido al término de diferencia al cuadrado. El MSE es recomendable utilizarlo cuando no hay valores extremos que puedan afectar de manera significativa la evaluación.

Error Absoluto Medio (MAE): El MAE calcula la diferencia absoluta promedio entre los valores pronosticados y los valores reales. Es una medida de precisión del pronóstico que no tiene en cuenta la dirección del error. Cuanto menor sea el valor de MAE, mejor será la precisión del pronóstico. A diferencia del MSE, el MAE no penaliza más los errores grandes. El MAE es útil cuando los errores positivos y negativos tienen la misma importancia y se desea una medida más interpretable en las mismas unidades que los datos originales.

Raíz del Error Cuadrado Medio (RMSE): El RMSE es la raíz cuadrada del MSE. Al calcular la raíz cuadrada, el RMSE proporciona una medida de error en la misma unidad que los datos originales. Es una medida popular ya que es fácil de interpretar y se utiliza ampliamente para comparar diferentes modelos de pronóstico. El RMSE también penaliza más los errores grandes y es adecuado cuando se desea una medida en las mismas unidades que los datos originales.

Error Porcentual Absoluto Medio (MAPE): El MAPE calcula el porcentaje promedio de diferencia absoluta entre los valores pronosticados y los valores reales. Es una medida relativa que evalúa la precisión del pronóstico en términos porcentuales. El MAPE es útil cuando se desea evaluar la precisión relativa del pronóstico en comparación con los valores reales. Sin embargo, puede producir resultados inestables o infinitos cuando hay valores cercanos a cero en los datos reales.

Coefficiente de Determinación (R²): El R² es una medida utilizada en modelos de regresión para medir la proporción de la variabilidad de la variable dependiente que puede ser explicada por el modelo. Cuanto más cercano a 1 sea el valor de R², mejor será el modelo de pronóstico. El R² es

una medida útil para evaluar la capacidad de pronóstico en modelos de regresión lineal, pero puede no ser apropiado para otros tipos de pronóstico, como series de tiempo.

5.2 Selección de métricas de error

Como se mencionó en la descripción y objetivo de este proyecto, el pronóstico de este caso de estudio es usado para planificar toda la operación de las líneas de atención de contactos; las desviaciones altas del tráfico de contacto real con respecto al pronóstico generan afectaciones en el nivel de servicio, es por esto por lo que uno de los objetivos de nuestra métrica del desempeño del pronóstico es que debe ser sensible a los errores atípicos.

Teniendo esto en cuenta, se considera que la métrica más apropiada para la evaluación del desempeño del pronóstico es el RMSE. Sin embargo, todas se generarán los resultados para el MSE, MAE, RMSE y MAPE.

5.3 Evaluación de métricas de error

Se evalúan las diferentes métricas de error para cada uno de los modelos, tanto en el set de entrenamiento como en el set de pruebas, se obtienen los siguientes resultados:

Modelo		MSE	MAE	RMSE	MAPE
Sarima	TRN	2810.74	38.54	53.02	19.22
Sarima	TEST	3996.03	49.63	63.21	25.96
Holt	TRN	1351.40	28.31	36.76	15.17
Holt	TEST	4545.41	56.72	67.42	37.10
LSTM	TRN	1949.29	35.17	44.15	19.34
LSTM	TEST	4307.52	57.73	65.63	39.10

Tabla 5 Métricas de error

De forma general podemos ver que, para todas las métricas, los resultados en entrenamiento son mejores que los resultados en pruebas; este es un comportamiento normal y esperado. El modelo de HoltWinters muestra los mejores resultados en el set de entrenamiento, sin embargo, para el conjunto de pruebas, el modelo Sarima es el que muestra los mejores resultados. Este resultado favorable para el modelo Sarima se logra conseguir principalmente porque el modelo Sarima es capaz de mantener la estacionalidad a largo plazo.

A nivel operativo este error en el pronóstico representa un grado de imprecisión en la cantidad de agentes a ser programados para atender el volumen de contactos que ingresan a la línea, sin embargo, este error proveniente de un modelo predictivo da menor incertidumbre al negocio del error producido la estimación empírica del volumen que ingresarán.

5.4 Selección del modelo

Por tener el mejor resultado a nivel de métricas de error en el set de pruebas y por tener un estricto cumplimiento de los supuestos del error, el modelo seleccionado para el pronóstico del patrón de arribo de llamadas en esta línea de atención telefónica es el modelo Sarima.

6 DESARROLLO OBJETIVO 3

Este apartado pretende desarrollar las actividades que dan cumplimiento al objetivo 3: Desarrollar una solución de analítica avanzada que utilice el modelo definido y permita visualizar los pronósticos a largo plazo, mediano plazo y tiempo real, para que la administración de la línea pueda utilizarlo en su gestión

A continuación, el desarrollo de cada una de las actividades requeridas para dar cumplimiento a este objetivo:

6.1 Definición de la necesidad

Con el fin de poder aprovechar al máximo el pronóstico realizado con el modelo seleccionado, la administración de la operación de esta línea de atención telefónica requiere una analítica que pueda mostrar el pronóstico a nivel de mes, día e intervalos del día. Adicional a esto, también se requieren ver resultados de gestión del comportamiento de la línea, lo cual incluye la visualización de los siguientes indicadores:

- Llamadas entrantes
- Llamadas contestadas y abandonadas
- Llamadas contestadas y abandonadas en menos de 20 segundos
- Tiempo promedio de gestión de las llamadas (AHT)
- Ocupación de la línea
- Velocidad promedio de respuesta y abandono

La analítica por desarrollar debe contar con una arquitectura capaz de actualizarse cada 30 minutos.

Por otro lado, una de las limitantes que ha tenido la línea para analizar estos datos, es que la empresa no tiene posesión de esta información; actualmente los datos del comportamiento de la línea persisten por 6 meses en la herramienta de comunicación que usan los operadores, dicha herramienta es tercerizada con un proveedor y la única forma de acceder a los datos de los últimos 6 meses es a través de APIS.

Sabiendo esto, otro de los requerimientos que se deben tener con este desarrollo es que los datos persistan de forma permanente en bases de datos de dominio interno, de esta forma garantizamos el acceso a todo el histórico y no tenemos la limitación de 6 meses que nos da el aplicativo del proveedor.

6.2 Arquitectura propuesta

Para el desarrollo de este requerimiento la empresa cuenta con un portafolio de herramientas de Microsoft, de las cuales se consideran necesarias para el desarrollo de esta analítica las siguientes:

- Azure SQL Data Base: Base de datos SQL como herramienta de almacenamiento de datos.
- Azure Synapse Analytics: Herramienta de orquestación de integración.
- Power BI: Herramienta de visualización

- Power BI Portal: Herramienta para embeber la visualización en la web y garantizar seguridad y segmentación en el acceso.

De forma general, la arquitectura propuesta consiste en utilizar azure synapse analytics para orquestar las actividades de integración y transformación de los datos, esto incluye el procesamiento de los datos con el modelo de pronóstico previamente seleccionado utilizando Apache Spark, desde synapse se almacenarán los datos en una base de datos en la nube de Azure. Posterior a esto, se utilizará la herramienta de power BI para desarrollar la visualización de las diferentes métricas y pronósticos; este power BI será embebido en el power BI portal, la utilización de power BI portal se da para poder garantizar acceso a los diferentes power BIs de la organización de forma segmentada y centralizada.

A continuación, la esquematización de la arquitectura a implementar en este desarrollo:

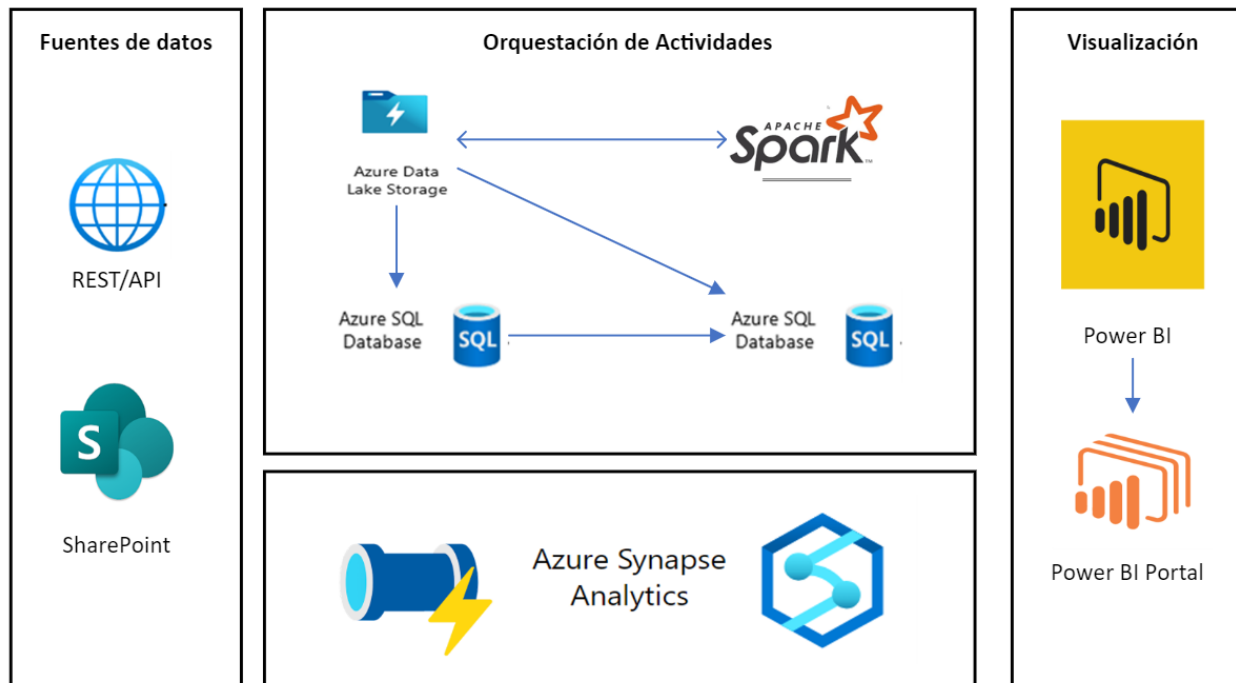


Ilustración 21 Arquitectura propuesta

6.3 Orquestación de actividades (Pipeline)

En la herramienta de integración Synapse Analytics, se orquestan todas las actividades requeridas para poder extraer, transformar y almacenar los datos en el repositorio final donde persistirán de forma histórica. La orquestación de actividades se refiere al proceso de programar de forma secuencial o paralela la ejecución de diferentes tareas que en conjunto generan un estado final esperado del sistema.

Los tipos actividades de integración utilizadas para hacer todo el procesamiento de este proyecto

son los siguientes:

- Copy Data: Extraer datos de una fuente origen y los escribe en una fuente destino.
- Look Up: Busca información en una fuente origen y los entrega como valores libres al pipeline para que este los use como variables o parámetros.
- Notebook: Ejecuta un notebook de Synapse el cual se puede construir en lenguaje Python, R o Escala. La ejecución de este notebook es necesaria para correr el modelo de ciencia de datos.
- Execute Pipeline: Ejecuta un pipeline externo.
- Script: Ejecuta sentencias SQL sobre una base de datos fuente.
- Set Variable: Asigna valores a variables dentro del pipeline.
- Data Flow: Ejecuta actividades especiales, para el caso de este desarrollo se utilizó un apisonamiento para normalizar archivos json a tabla.

El proceso comienza con la extracción de datos de las APIs, que los proporcionan en formato JSON. Los datos se guardan inicialmente en una base de datos de "staging" donde llegan sin un formato específico, todo está en formato de "string". Para poder guardar los datos de tipo JSON en una base de datos relacional, estos requieren ser normalizados (convertidos a tablas); un desafío que se enfrentó en este punto fue la compleja jerarquía del JSON que entregaba una de las APIs, esto hacía que la actividad de copiado no pudiese convertirlo a tabla. Para solucionarlo, usamos una técnica llamada "flattening" en una actividad de tipo "Data Flow", que nos permitió convertir este JSON complejo en tabla estructurada que pudiera ser almacenada en la base de datos de staging.

Una vez que los datos están en la base de datos de staging, son movidos a un almacén de datos donde se aplican las transformaciones necesarias para darle a cada columna el formato adecuado y limpiar los datos de nulos o registros basura.

Una vez se tienen los datos limpios en el almacén, se orquesta una actividad que ejecuta un "notebook" que lee los datos del almacén y procesa el modelo de pronóstico; después, guarda la serie pronosticada de vuelta en el almacén. Y de esta forma terminaría el proceso de integración y procesamiento de los datos, obteniendo el histórico y el pronóstico listos para ser consumidos por una herramienta de visualización.

Este proceso de orquestación ocurre de dos maneras: durante la noche, extrayendo los datos del día anterior y recalculando el pronóstico para los próximos 62 pasos y durante el día, se ejecuta este proceso cada 30 minutos, desde las 8:00 am hasta las 7:00 pm, para capturar información del día actual y permitir que la analítica muestre el comportamiento de la línea telefónica en tiempo real, facilitando así la toma de decisiones operativas para el negocio.

A continuación, el pipeline de actividades orquestadas para esta solución:

Maestría en Ciencia de Datos
 Facultad de Ingeniería y Ciencias

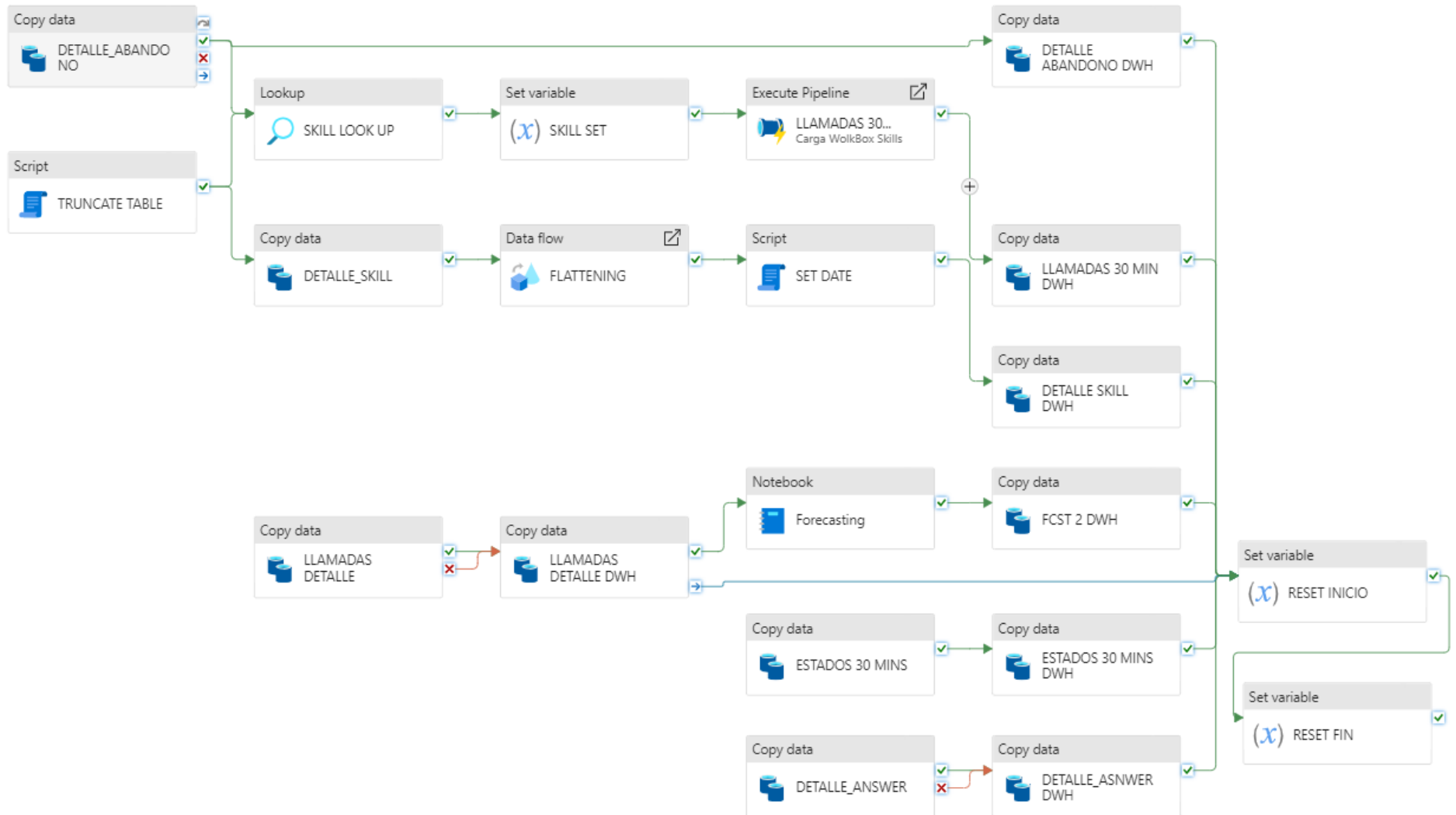


Ilustración 22 Pipeline de actividades

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

6.4 Visualización de la solución

Para la visualización de los datos, se utilizó power BI, en donde se importaron las tablas de la base de datos (almacén) al igual que algunos archivos de Excel en Sharepoint, los cuales almacenan información de metas que se utilizarán en la construcción de la solución de BI.

Teniendo todos los datos en el PBI se desarrolló un modelo tabular, este modelo cuenta con 7 tablas transaccionales las cuales registran el comportamiento de la línea a diferentes niveles, por otro lado, se tienen diferentes dimensiones que pueden o no interactuar con todas las tablas transaccionales y por último, algunas tablas de soporte que ayudan a controlar particularidades del modelo. A continuación, la descripción de cada una de las tablas utilizadas en el modelo:

Tablas de hechos:

Detalle_Abandono: Registra el detalle de las llamadas abandonadas y tiempos de abandonos.

Detalle_Skill: Registra el detalle del comportamiento de la línea por skill.

Answer_20sg: Registra el detalle de las llamadas contestadas en menos de 20 segundos.

Detalle_Answered: Registra el detalle de las llamadas contestadas por cada agente.

Llamadas_30mins: Registra el detalle de la línea por intervalos de 30 minutos.

Llamadas_Detalle: Registra el detalle de cada llamada, tipologías y características de la llamada.

Estados_30mins: Registra el detalle de los estados de los agentes conectados a la línea.

Dimensiones:

DimFecha: Controla las fechas a considerar en el modelo

DimAgent: Controla los agentes a considerar en el modelo

DimSkill: Controla los skills a considerar en el modelo

Tablas de soporte:

Métricas: Contiene todas las métricas calculadas sobre el modelo.

Distribución: Contiene la distribución por intervalos del pronóstico para ese día.

Festivos: Contiene los festivos nacionales

Int: Contiene los intervalos de operación del día

TiempoReal: Contiene la formulación del día actual

A continuación, el modelo de datos desarrollado:

Maestría en Ciencia de Datos
 Facultad de Ingeniería y Ciencias

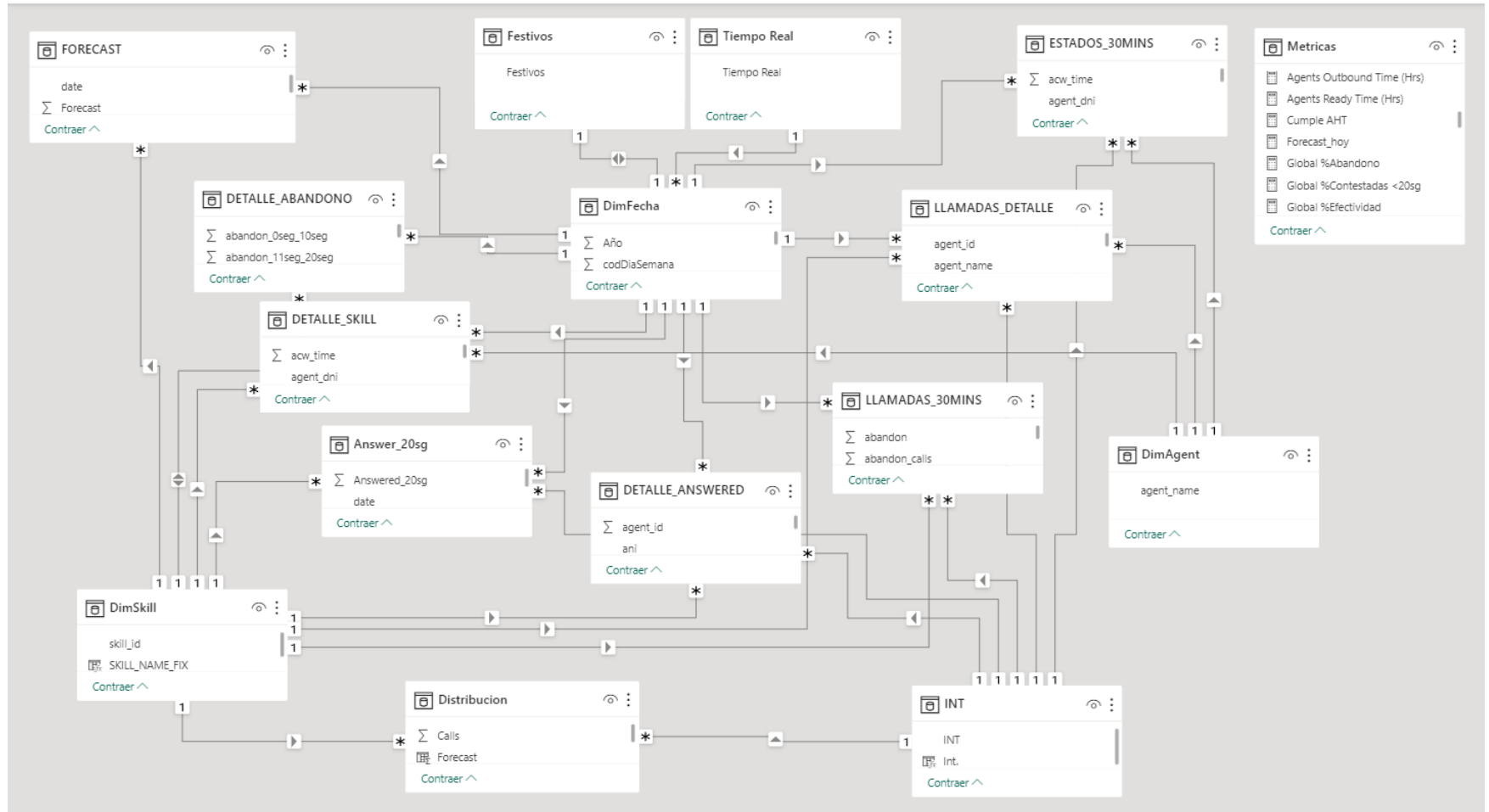
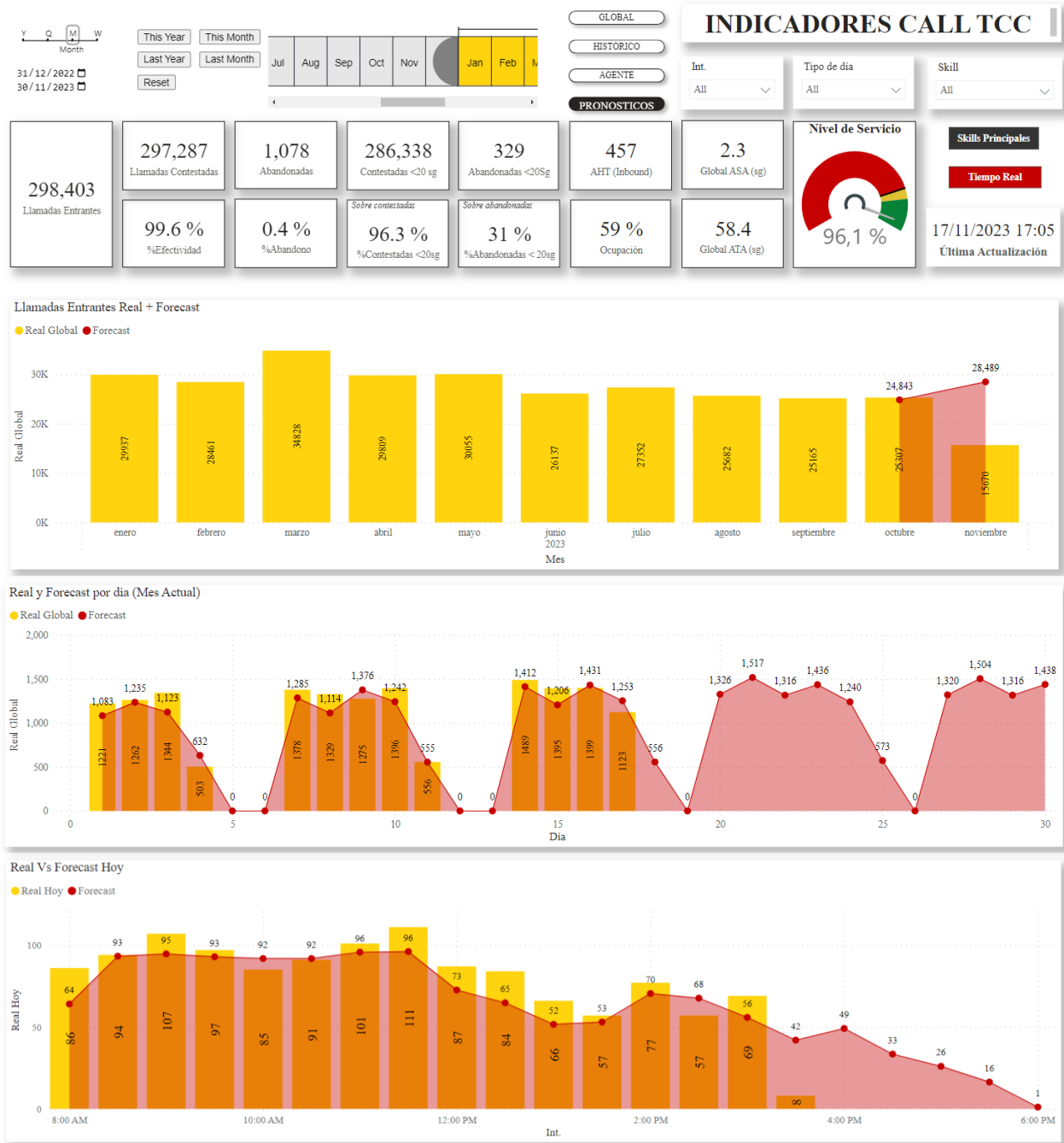


Ilustración 23 Modelo tabular

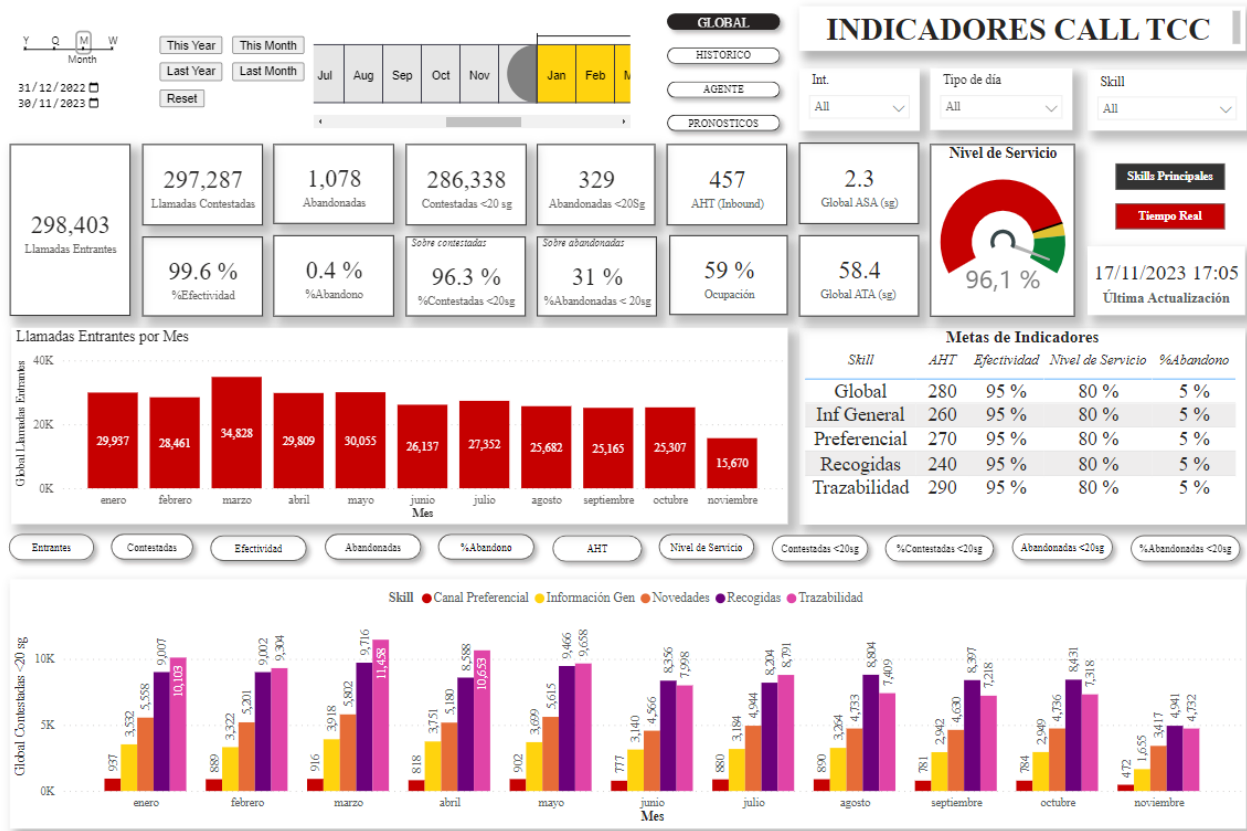
Maestría en Ciencia de Datos Facultad de Ingeniería y Ciencias

Con este modelo se desarrolló la siguiente solución de BI:



En este tablero se pueden observar los resultados de las métricas principales de la línea y el pronóstico en 3 niveles distintos: a nivel mensual, por día para todo el mes y por cada intervalo del día en curso. El pronóstico de las llamadas que van a ingresar por intervalo fue calculado utilizando el pronóstico del día y distribuyéndolo en cada intervalo siguiendo la distribución promedio de las últimas 4 semanas de las llamadas por intervalo para ese día.

Por otro lado, esta solución cuenta con otras páginas que se enfocan en representar las diferentes métricas de la línea en diferentes niveles y ventanas de tiempo:



7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 Conclusiones

Para el pronóstico del patrón de arribo de llamadas de esta línea se identificó que la técnica más apropiada fue un modelo estacional autorregresivo integrado y de media móvil (SARIMA), este modelo logró aprender la estacionalidad y tendencia de la serie, y mantenerla durante el horizonte de tiempo necesario para el proceso en cuestión. A pesar de que durante el proceso de entrenamiento otros modelos, como el de redes neuronales, presentaron un mejor ajuste, durante el periodo de prueba fue el modelo SARIMA el que mejor se logró ajustar en la ventana de tiempo evaluada.

El modelo SARIMA mostró un error de 53.02 en el conjunto de entrenamiento y 63.21 en el conjunto de pruebas (RMSE). A nivel operativo, este margen de error en el pronóstico representa una imprecisión en la planificación de la cantidad de agentes necesarios para manejar el volumen de contactos entrantes a la línea. Sin embargo, este error, derivado del modelo predictivo, resulta en una menor incertidumbre en comparación con la estimación empírica actualmente utilizada. Por otro lado, la operación continuará programando a los agentes, superando el margen de error del

pronóstico y adaptando la asignación de recursos en tiempo real. Esto se basa tanto en el comportamiento actual de la línea como en la planificación de actividades externas, como reuniones o entrenamientos, que no están directamente relacionadas con la operación. En resumen, el margen de error del pronóstico, asumido como una precaución adicional, se aprovecha para facilitar el desarrollo de actividades fuera de la línea, al tiempo que se garantiza la gestión eficiente del volumen de interacciones entrantes.

Con esta solución de analítica, la administración de la línea podrá gestionar con mayor precisión el aprovisionamiento de personal requerido para atender el volumen de llamadas entrantes, reduciendo el riesgo de caer en sobrecostos. También podrá generar una mejor programación de los horarios de los agentes y decidir cuándo sacar o meter agentes de la línea durante el día dependiendo de la disponibilidad y pronóstico de llamadas a ingresar.

7.2 Trabajos futuros

En cuanto a los trabajos futuros, se busca aumentar la robustez del pronóstico de la serie de tiempo mediante el aumento de complejidad del modelo y la mejora de la solución analítica propuesta. Se sugieren dos líneas de mejoras para el desarrollo actual:

7.2.1 Desarrollo de Modelos de Pronóstico Multivariados

Se propone la implementación de modelos de pronóstico multivariados que consideren festividades y eventos planificados de la línea. Este enfoque podría conducir a un ajuste más preciso de la serie de tiempo, al tener en cuenta múltiples variables que podrían influir en las predicciones.

7.2.2 Automatización con Portafolio de Modelos

El desarrollo actual se centra en un único modelo de pronóstico en producción. Una mejora sustancial sería incorporar automatización que siempre considere un portafolio de modelos y determine, durante el procesamiento de los datos, cuál modelo utilizar. Esta estrategia puede resultar beneficiosa ante posibles cambios en el comportamiento de la serie a lo largo del tiempo, permitiendo una adaptación dinámica y precisa.

8 REFERENCIAS BIBLIOGRAFICAS

1. De Arce, R. (1998). Introducción a los modelos autorregresivos con heterocedasticidad condicional (ARCH). Programa de Doctorado en Modelización Económica. Instituto LR Klein.
2. Preece, M. R. M., & Martínez, F. V. (2014). Análisis del riesgo de mercado de los fondos de pensión en México un enfoque con modelos autorregresivos. *Contaduría y administración*, 59(3), 165-195.
3. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

4. Jason Brownlee. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. Machine Learning Mastery1
5. Jason Brownlee. How to Develop LSTM Models for Time Series Forecasting. Machine Learning Mastery2
6. Aya Abdelsalam Ismail et al. Input-cell attention reduces vanishing saliency of recurrent neural networks. Advances in Neural Information Processing Systems3
7. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (Vol. 5). McGraw-Hill/Irwin.
8. Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (Vol. 7). Upper Saddle River, NJ: Pearson.
9. Brownlee, J. (2018). Time Series Prediction with XGBoost in Python. Machine Learning Mastery
10. Zhang, T., & Zhan, J. (2017). GPU-acceleration for Large-scale Tree Boosting. arXiv preprint arXiv:1706.08359.
11. Brownlee, J. (2018). How to Use Moving Average Methods for Time Series Forecasting with Python. Machine Learning Mastery.
12. Velez, J. VALIDACIÓN DE MEDIDAS DE EVALUACIÓN PARA EL PRONÓSTICO DE LA TASA DE CAMBIO EN COLOMBIA (2016). Colegio de Estudios Superiores de Administración –CESA.
13. Fuente: Alarcón, M. (2009, septiembre). Clasificación del Método de pronóstico de Torres (Segunda Parte). Poliantea.
14. Paredes, J. (1994). Planificación de la producción. México: Editorial Mc Graw Hill.
15. Makridakis, S. (1997). Métodos de pronóstico. México: Editorial Limusa.
17. Pérez López, C. (2005). Métodos estadísticos avanzados con SPSS. Madrid: Editorial Thompson.
18. Pindyck, R. (2001). Econometría y modelos de pronóstico. México: Editorial Mc Graw Hill.
19. Shyi-Ming Pedrycz, Witold Chen. Deep Learning: Algorithms and Applications. Springer, 2020.
20. Fernandez, R. Series Temporales Avanzadas: Aplicación de Redes Neuronales para el Pronóstico de Series de Tiempo (2021). Departamento de Estadística e Investigación Operativa. Universidad Nueva Granada.
21. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
22. Evaristo, r. Aplicación de técnicas de machine learning para la predicción de los precios de la acción APPLE (2022). Revista de investigación de sistemas e informática 15(1): 13-22 (2022). DOI: <https://doi.org/10.15381/risi.v15i1.23737>

23. Albrecht, T. Raushc, T. Daniel, N. Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. Journal of Business Research, Volume 123, February 2021, Pages 267-278.
24. Ebadi, M. Hosseini, M. Karlsson, S. 2016. Forecasting incoming call volumes in call centers with recurrent Neural Networks. Journal of Business Research. El SEVIER, Volume 69, Issue 11, November 2016, Pages 4811-4814.