Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Pronóstico de ventas de tiendas de abarrotes en Zona Metropolitana de Guadalajara

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
Daniel García Hernández

Profesor:
Mtro. Byron Michael Motta Bonilla

Tlaquepaque, Jalisco, 29 de mayo de 2025

Pronóstico de ventas de tiendas de abarrotes en Zona Metropolitana de Guadalajara

Daniel García Hernández

Resumen

Palabras clave: serie de tiempo, pronóstico de ventas, tienda de abarrotes, variables exógenas.

Las tiendas de abarrotes de la zona metropolitana de Guadalajara compiten actualmente en un mercado en el que predominan cadenas de conveniencia, las cuales cuentan con recursos y ventajas significativas. En este contexto, resulta fundamental que los abarrotes independientes dispongan de herramientas predictivas que les permitan planificar sus ventas diarias con precisión.

Este trabajo tiene como objetivo desarrollar un modelo matemático para la predicción de las ventas diarias de las tiendas de abarrotes dentro de la zona metropolitana de Guadalajara. La finalidad es proporcionar una herramienta de pronóstico que permita a estos negocios anticipar sus ventas diarias, optimizando así la toma de decisiones estratégicas.

Para lograrlo, se propone un modelo SARIMA, debido a que las ventas diarias de este tipo de establecimientos presentan una fuerte estacionalidad semanal y estacionariedad. Estas características hacen que el modelo SARIMA sea una opción adecuada y efectiva para captar patrones temporales y proporcionar predicciones precisas, de igual manera se incorporan a este modelo el uso de variables exógenas, las cuales son variables externas que pueden afectar el comportamiento de la serie a pronosticar en este caso las ventas diarias, por medio del modelo StatsForecast de la paquetería Nixtla.

Pronóstico de ventas de tiendas de abarrotes en Zona Metropolitana de Guadalajara

Daniel García Hernández

Abstract

Key words: Time series, sales forecasting, grocery store, Exogenous variables.

Grocery stores in the Guadalajara metropolitan area currently compete in a market dominated by convenience store chains, which have significant resources and advantages. In this context, it is essential for independent grocery stores to have predictive tools that allow them to accurately plan their daily sales.

The objective of this work is to develop a mathematical model to predict the daily sales of grocery stores within the Guadalajara metropolitan area. The goal is to provide a forecasting tool that enables these businesses to anticipate their daily sales, thereby optimizing strategic decision-making.

To achieve this, a SARIMA model is proposed, as the daily sales of these establishments exhibit strong weekly seasonality and stationarity. These characteristics make the SARIMA model an appropriate and effective option for capturing temporal patterns and providing accurate forecasts. Additionally, the model incorporates the use of exogenous variables external factors that can influence the behavior of the time series being forecasted, in this case, daily sales through the StatsForecast model from the Nixtla package.

Tabla de Contenidos

Índice de figuras

Página

Índice de tablas

Página

Dedicado a mis padres Jorge García y María de la Luz Hernández por su amor incondicional, su apoyo constante y su confianza en mí a lo largo de este camino. Su ejemplo de esfuerzo y perseverancia ha sido mi mayor inspiración para lograr mis metas.

A Byron Michael Motta, gracias por su guía, paciencia y valiosos consejos durante el desarrollo de este trabajo. Su acompañamiento fue fundamental para alcanzar los objetivos planteados y para enriquecer mi formación académica.

A ITESO, por brindarme las herramientas, el conocimiento y el entorno necesario para crecer tanto profesional como personalmente. Me llevo de esta institución grandes aprendizajes, experiencias y recuerdos que marcaron mi vida.

1 Introducción

1.1 Contexto

En la actualidad, la predicción de ventas se ha convertido en una herramienta clave para la toma de decisiones estratégicas en el sector minorista. El creciente volumen de datos generados diariamente por las transacciones comerciales, el comportamiento de los consumidores y factores externos como las tendencias económicas y sociales, ha impulsado el uso de técnicas avanzadas de análisis de datos, como la ciencia de datos y el aprendizaje automático. Estas técnicas permiten a las empresas obtener patrones ocultos y realizar predicciones más precisas sobre las ventas futuras, lo que resulta en una mejor planificación y optimización de los recursos.

La Zona Metropolitana de Guadalajara es una de las áreas urbanas más grandes y dinámicas de México, con una población que según el censo económico realizado en 2020 [1] supera los 5 millones de habitantes. Esta área constituye un centro económico de gran relevancia, con un fuerte crecimiento en los sectores industrial, comercial y de servicios. Dentro de este contexto, el comercio al por menor en tiendas de autoservicio y departamentales en 2019, dentro del estado de Jalisco, representaron una producción bruta total de 30 millones de pesos, así como un ingreso total de 95 millones de pesos [2].

Las tiendas de abarrotes como se conocen en México, también conocidas como tiendas de barrio o de esquina, son aquellos negocios pequeños y medianos que ofrecen productos al por menor dentro del consumo cotidiano.

Particularmente las tiendas de abarrotes dentro de este sector, juegan un papel fundamental en la economía del estado de Jalisco, así como de la vida cotidiana de los habitantes de la Zona Metropolitana de Guadalajara debido a que como lo mencionan González y Polanco [3] son el espacio indicado para que los hogares de ingresos bajos y medios puedan adquirir diversos productos, en especial botanas, bebidas no

alcohólicas, snacks, azúcar, entre otros, a un precio justo, siendo más asequible a su presupuesto y ubicación geográfica.

Las tiendas de abarrotes en la actualidad se enfrentan a un panoráma realmente complicado, como lo mencionan González y Polanco [3] en donde las tiendas de conveniencia como lo son OXXO (76.8 porciento de penetración del mercado) o 7-eleven, cuentan con una administración profesional, buscan ubicaciones estratégicas en las zonas urbanas, emplean otros medios de pago alternativos al efectivo y algunas utilizan contratos con asociados que participan con el trabajo de estas cadenas de venta al detalle. Puede decirse que son un modelo de negocios completamente diferente al de las tiendas de abarrotes.

Dado el carácter competitivo y cambiante de este sector, es esencial que los negocios cuenten con herramientas que les permitan anticipar la demanda, ajustar su inventario y estrategia de ventas en consecuencia. La correcta predicción de las ventas no solo ayuda a minimizar el desperdicio y las pérdidas por productos no vendidos, sino que también ayuda a una tener una mejor planeación financiera al momento de manejar sus ingresos y gastos, de igual manera, poderse anticipar a una baja de ventas implementando estrategias de marketing o promociones para incentivar las ventas.

Como mencionan Nojek y compañía [4] la lección e implementación de un método adecuado de pronósticos siempre ha sido un tema de gran importancia para las empresas. Se utilizan los pronósticos en el área de compras, marketing, ventas, etc. Un error significante en el pronóstico de ventas podría dejar a una empresa sin la materia prima o insumos necesarios para su producción, o podría generarle un inventario demasiado grande. En ambos casos, el pronóstico erróneo disminuye las utilidades de la empresa.

De igual manera, Nojek y compañía [4] nos hablan acerca de de las diversas técnicas para proyectar el mercado. Dentro de las mismas se encuentran las técnicas clásicas de proyección. Una forma de clasificarlas consiste en hacerlo en función de su carácter, esto es, aplicando métodos de carácter cualitativo, modelos causales y modelos de series de tiempo.

- Los métodos de carácter cualitativo se basan principalmente en opiniones de expertos. Su uso es frecuente cuando el tiempo para elaborar el pronóstico es escaso, cuando no se dispone de todos los antecedentes mínimos necesarios o cuando los datos disponibles no son confiables para predecir el comportamiento futuro. Resulta difícil emitir un juicio sobre la eficacia de sus estimaciones finales.

- Los modelos de pronóstico causales parten del supuesto de que el grado de influencia de las variables que afectan al comportamiento del mercado permanece estable, para luego construir un modelo que relacione ese comportamiento con las variables que se estima que son las causantes de los cambios que se observan en el mercado.

- Los modelos de series de tiempo (método estadístico) se refieren a la medición de valores de una variable en el tiempo a intervalos espaciados uniformemente. El objetivo de la identificación de la información histórica es determinar un patrón básico en su comportamiento, que posibilite la proyección futura de la variable deseada.

En el presente proyecto se emplearán modelos de series de tiempo basados en métodos estadísticos, dado que se cuenta con información histórica suficiente y confiable para realizar pronósticos de ventas utilizando únicamente estos datos.

Para alcanzar este objetivo, se presenta en primera instancia la justificación del proyecto, así como la definición del problema y los objetivos, con el fin de establecer el marco conceptual y metodológico que guiará el desarrollo del trabajo.

Posteriormente, se describen los datos y se realiza un análisis exploratorio, con el propósito de comprender mejor el comportamiento de las ventas diarias de las tiendas de abarrotes dentro de la zona metropolitana de Guadalajara. Este análisis permitirá identificar el modelo estadístico adecuado para pronosticar dichas ventas.

Una vez comprendida la naturaleza de la serie de tiempo, se procede a describir los modelos estadísticos seleccionados, así como las métricas utilizadas para su evaluación, con el objetivo de determinar cuál ofrece el mejor desempeño en la predicción de las ventas diarias.

Además, se presentarán los experimentos y simulaciones realizados durante el desarrollo del proyecto que, si bien no fueron incluidos en los resultados finales, contribuyeron al proceso de análisis y selección del modelo.

Finalmente, se expondrán los resultados obtenidos con el modelo estadístico elegido, seguidos de una discusión, las conclusiones del trabajo y los posibles trabajos futuros.

1.2 Justificación

Actualmente las tiendas de abarrotes de la zona metropolitana de Guadalajara enfrentan a un panorama realmente complicado, en donde las tiendas de conveniencia como lo son OXXO o 7-eleven tienen una gran participación en el mercado debido a que son respaldadas por compañías que les brindan administración profesional, ubicaciones estratégicas, así como convenios con diferentes marcas o proveedores para obtener mayores beneficios como ofrecer pago de servicios y facturación electrónica.

Dado el carácter competitivo y cambiante de este sector, es esencial que las tiendas de abarrotes dentro de la zona metropolitana de Guadalajara cuenten con herramientas que les permita anticipar sus ventas diarias, con el fin de poder tener una mejor planificación financiera al momento de manejar sus ingresos y gastos, de igual manera, poderse anticipar a una baja de ventas implementando estrategias de marketing o promociones para incentivar las ventas. Por último, este tipo de herramientas pueden ser de gran utilidad para detectar patrones de compra y tendencia de consumo los cuales son de vital importancia para la toma de decisiones estratégicas.

1.3 Problema

Problema práctico: Las tiendas de abarrotes dentro de la zona metropolitana de guadalajara tienen la necesidad predecir sus ventas, así como el estimado de sus ingresos con el fin de optimizar la toma de decisiones estratégicas y obtener una ventaja competitiva frente a las grandes cadenas que dominan el mercado.

Problema científico: Obtener un modelo matemático de series de tiempo que pueda generalizar los ingresos a futuro de las tiendas de abarrotes dentro de la zona metropolitana de Guadalajara, captando su estacionalidad, así como factores externos que afectan a las ventas de las mismas, como lo son días festivos, condiciones climatológicas o comportamientos atípicos.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un modelo de predicción de ventas diarias para tiendas de abarrotes en la zona metropolitana de Guadalajara, que permita optimizar la planificación financiera y apoyar la toma de decisiones

estratégicas mediante la identificación de patrones de consumo y tendencias de compra.

1.4.2 *Objetivos específicos*

1. Obtener una base de datos que contenga información de años históricos de alguna tienda de abarrotes dentro de la zona metropolitana de Guadalajara.

2. Analizar los datos históricos de ventas diarias de las tiendas de abarrotes dentro de la zona metropolitana de Guadalajara con el fin de identificar patrones de consumo así como estacionalidad.

3. Desarrollar y evaluar distintos modelos de predicción de ventas diarias que consideren tanto factores internos como externos que puedan afectar a las ventas diarias.

4. Implementar el modelo de predicción de ventas diarias seleccionado, así como observar y evaluar sus resultados con las ventas diarias reales registradas.

2 Metodología

2.1 Descripción de los datos

El conjunto de datos analizados pertenece a *Abarrotes Robert's* la cual es una tienda de abarrotes ubicada frente al centro médico de Occidente de Guadalajara Jalisco. Esta tienda de abarrotes se eligió porque cuenta con bases de datos de cada venta realizada por mas de diez años. De igual manera, al ubicarse dentro de la zona metropolitana de Guadalajara, nos permitirá generalizar su comportamiento para el resto de tiendas de abarrotes ubicadas en la misma zona geográfica.

La base de datos analizada son los registros de cada una de las ventas realizadas desde enero del 2017 a la actualidad. Los campos con los que cuenta son:

- Fecha: Fecha en la que fue realizada la venta.
- Hora: Hora, minuto y segundo en la que fue realizada la venta.
- PrecioTotal: Monto total de la venta realizada.
- FormaPago: Puede tomar los valores 'EFE' y 'TAR' para indicar si la transacción fue pagada en efectivo o tarjeta de crédito respectivamente.

Al tener una base de datos en la que cada registro es una venta, y el objetivo del presente trabajo es poder hacer una estimación de las ventas diarias, se procede a agrupar las ventas por día, de esta manera poder obtener una base de ventas diarias.

2.2 Análisis exploratorio

Dentro del presente análisis exploratorio se comienza por observar las ventas diarias históricas de la tienda de abarrotes la cual es objeto de estudio.

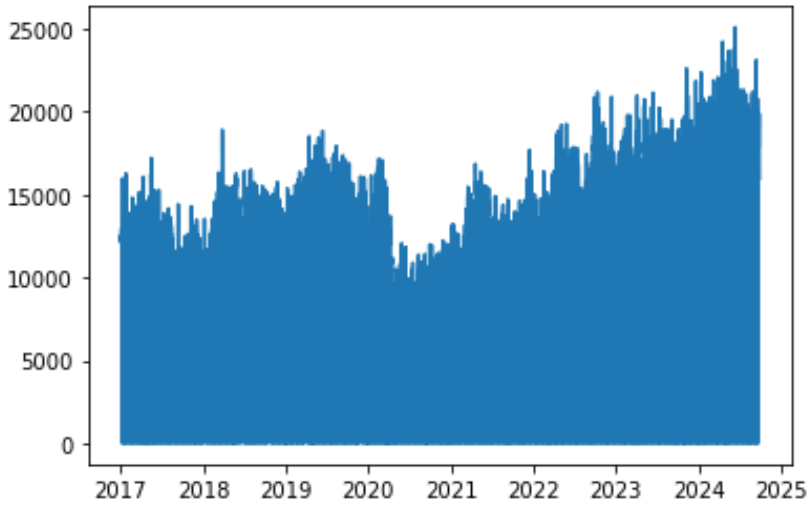

Figura 2.1: Ventas diarias históricas

En la figura 2.1 se puede observar que los datos analizados van desde enero del 2017 hasta septiembre del 2024. En donde se puede ver el impacto que tuvo la pandemia del Covid-19 a partir de finales de marzo del 2020 y duró aproximadamente un año el impacto. A partir de mediados del 2021 se alcanzaron los mismos niveles que se tenían de ventas diarias antes de la pandemia, se empezaron a incrementar dichas ventas de manera constante.

Para poder observar mejor el comportamiento de las ventas diarias de la última tendencia, se grafican únicamente las ventas del último año.

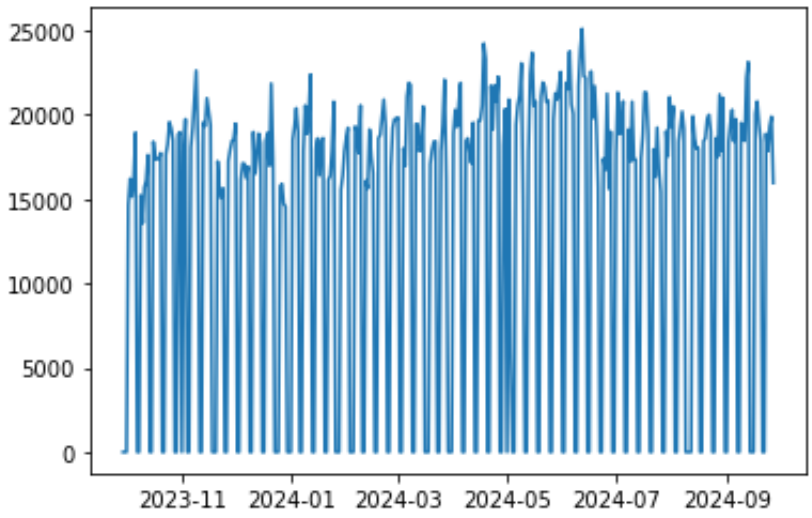

Figura 2.2: Ventas diarias último año

En el gráfico 2.2 se puede observar que los datos cuentan con una estacionalidad semanal fuertemente marcada. De igual manera, se logran ver algunos datos outliers, así como alguna estacionalidad anual, en donde hay periodos en los que las ventas diarias incrementan como lo podemos observar en temporada de verano y otros en donde disminuyen como lo puede ser el mes de diciembre.

Al continuar analizando esta estacionalidad anual, se grafican los datos históricos agrupados de manera mensual.

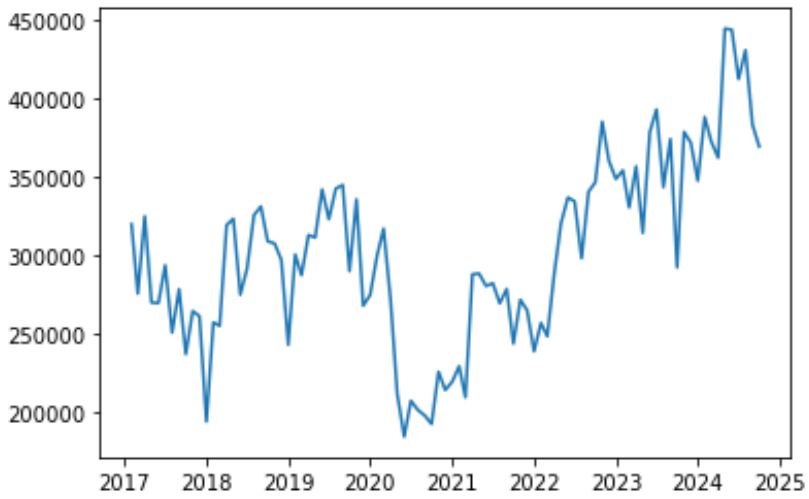

Figura 2.3: Ventas mensuales históricas

En el gráfico 2.3 se puede evidenciar aún mas el impacto que se obtuvo por la pandemia del Covid-19 y el reciente crecimiento en las ventas diarias. De igual manera, se puede observar que en todos los años, sin considerar los años impactados por la pandemia (2020, 2021), en el último mes de cada año las ventas bajan considerablemente.

A continuación se muestran las ventas diarias agrupadas de manera anual.

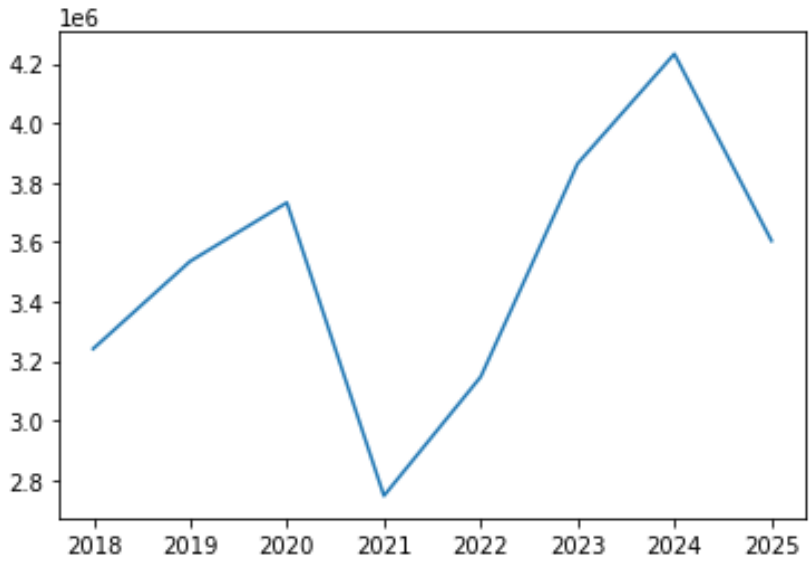

Figura 2.4: Ventas anuales históricas

En el gráfico 2.4 observamos que en los primeros tres años se muestra un crecimiento constante en las ventas anuales, seguidos por dos años donde se obtienen niveles mas bajos debido a la pandemia del Covid-19, por último, en los últimos dos años en los cuales se tiene información completa, las ventas vuelven a incrementar de manera similar a la que observábamos antes de la pandemia.

Cabe resaltar, que este crecimiento constante en las ventas no necesariamente representan un crecimiento en los ingresos de las tiendas de abarrotes dentro de la Zona Metropolitana de Guadalajara, sino que este crecimiento puede estar mas influenciado por la inflación de los precios de los productos que se venden en las tiendas de abarrotes de la Zona Metropolitanax| de Guadalajara.

Una vez analizado el comportamiento de las ventas diarias por día, mes y año, procedemos a analizar las ventas por hora a lo largo del día.

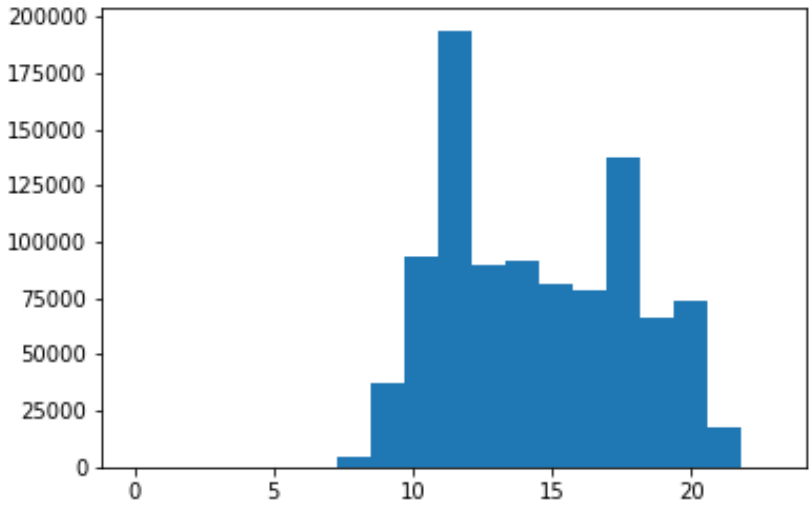

Figura 2.5: Histograma de ventas por hora

Las ventas mas fuertes a lo largo del día son a las 11 am y 6 pm como se puede visualizar en la figura 2.5, esto puede ser debido al término de las jornadas laborales del sector económicamente activo. El resto de horas se mantiene constante, en donde las ventas van disminuyendo conforme las horas van pasando.

Posteriormente se analizarán las ventas por día de la semana.

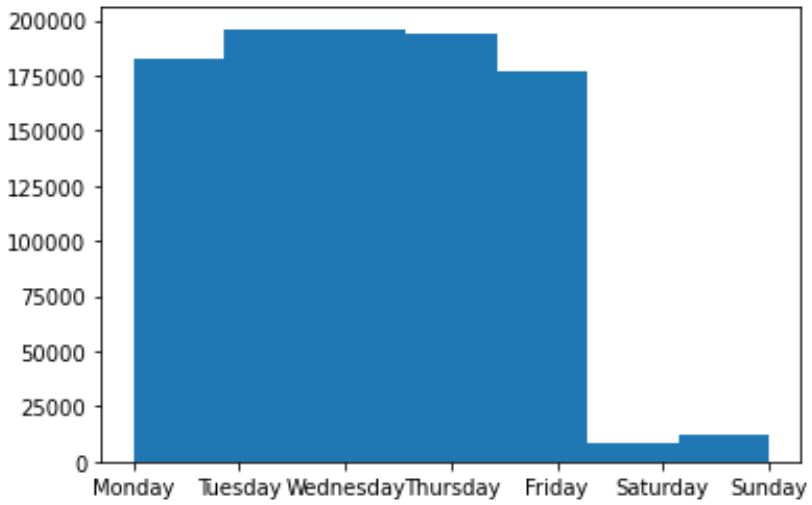

Figura 2.6: Histograma de ventas por día de la semana

En la figura 2.6 observamos que los días con mayores ventas suelen ser los días martes, miércoles y jueves, seguido por los lunes y por último los viernes. Cabe destacar que en la tienda de abarrotes la cual es objeto de estudio no abre los días sábados y domingos, por lo

que, los datos en donde obtenemos ventas en estos días podrían ser considerados datos atípicos.

Para poder entender de una mejor manera el comportamiento semanal, así como las horas del día, se realiza el siguiente gráfico.

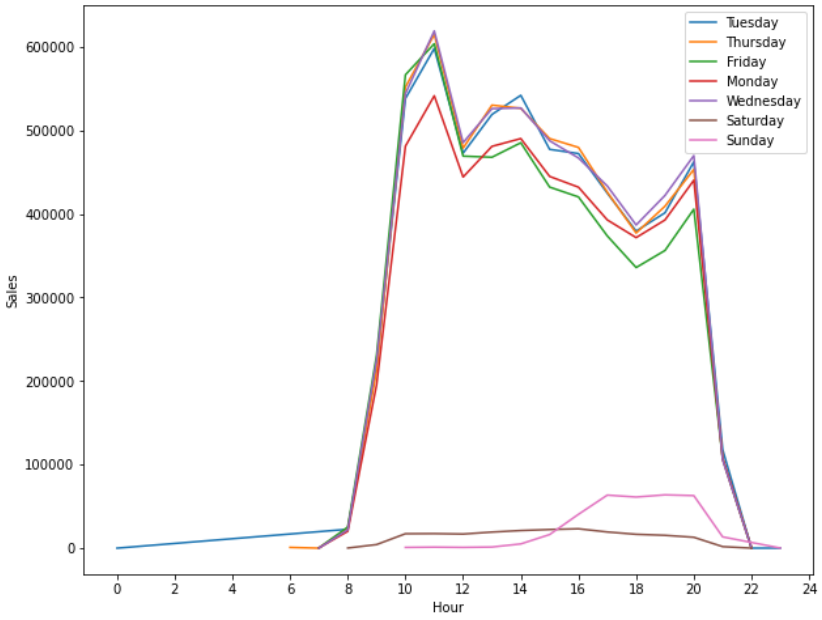

Figura 2.7: Estacionalidad por hora y día de la semana

En el gráfico 2.7 se puede observar que efectivamente las menores ventas a lo largo de la semana se encuentran en los días lunes y viernes. Cabe resaltar que las ventas mas bajas se encuentran los lunes por las mañanas y los viernes por las tardes.

2.3 *Descripción de los modelos*

El modelo utilizado fue StatsForecast de la paquetería Nixtla, el cual incorpora un modelo SARIMA. Se eligió dicho modelo debido a la alta estacionalidad y estacionariedad de la base de datos. Para poder entender por qué SARIMA es un buen modelo para un conjunto de datos con estas características, primero es primordial comprender como se compone el modelo SARIMA.

El modelo SARIMA por sus siglas en inglés Estacional Autoregresivo Integreado de promedios móviles. Basándonos en el trabajo de Miranda [5] y Villavicencio [6] procedemos a detallar los componentes de los modelos SARIMA, así como sus ventajas y desventajas.

Componente Autoregresivo (AR), los modelos autoregresivos se

basan en la idea de que el valor actual de la serie, X_t, puede explicarse en función de p valores pasados $X_{t-1}, X_{t-1}, ..., X_{t-p}$, donde p determina el número de rezagos necesarios para pronosticar un valor actual.

El modelo autoregresivo de orden p está dado por:

$$X_t = \Phi_0 + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \epsilon_t \tag{2.1}$$

En donde ϵ_t es un proceso de ruido blanco y $\Phi_0, \Phi_1, \Phi_2, ..., \Theta_p$ son los parámetros del modelo.

Componente de Medias Móviles (MA), los modelos de medias móviles o determinados por una fuente externa, son modelos que suponen linealidad, el valor actual de la serie, X_t, está influenciado por los valores de la fuente externa.

El modelo de promedios móviles de orden q está dado por:

$$X_t = \Theta_0 - \Theta_1 \epsilon_{t-1} - \Theta_2 \epsilon_{t-2} - ... - \Theta_q \epsilon_{t-q} - \epsilon_t \tag{2.2}$$

donde ϵ_t es un proceso de ruido blanco, y $\Theta_0, \Theta_1, \Theta_2, ..., \Theta_q$ son los parámetros del modelo.

Componente Integrado (I), Los modelos Autoregresivos (AR), de medias móviles (MA) o los procesos Autoregresivos de medias móviles (ARMA) a menudo requieren de alguna diferencia en la serie de tiempo, esto pasa normalmente con series económicas, las cuales no son estacionarias, porque pueden ir cambiando de nivel en el tiempo o sencillamente la varianza no es constante en el tiempo, por consiguiente se debe diferenciar la serie de tiempo d veces para hacerla estacionaria y luego a dicha serie estacionaria aplicarle un modelo AR(p), MA(q), o ARMA(p, q), conformando así un proceso ARIMA(p, d, q).

Su expresión algebraica es:

$$X_t^d = c + \Phi_1 X_{t-1}^d + ... + \Phi_p X_{t-p}^d + \Theta_1 \epsilon_{t-1}^d + \Theta_2 \epsilon_{t-2}^d + ... + \Theta_q \epsilon_{t-q}^d + \epsilon_t^d \tag{2.3}$$

donde X_t^d es la serie de las diferencias de orden d, ϵ_t^d es un proceso de ruido blanco, y $\Phi_0, \Phi_1, ..., \Phi_p, \Theta_1, ..., \Theta_q$ son los parámetros del modelo.

Por último, una vez que ya vimos todos sus componentes, podemos adentrarnos en el proceso Estacional Autoregresivo Integrado y de Media Móvil ARIMA$(p, d, q)(P, D, Q)s$.

Cuando una serie de tiempo tiene intervalos de observación menores a un año, entonces es frecuente que estas tengan variaciones o patrones sistemáticos cada cierto periodo, estas variaciones sistemáticas

inferiores a un año, por ejemplo, semestral, mensual, diario, etc. Deben ser capturadas en los llamados Factores Estacionales, dentro de la estructura del modelo a construirse.

Las series de tiempo estacionales pueden ser de dos tipos: Aditivas y Multiplicativas.

Y al mismo tiempo cada una de estas series puede ser estacionaria o no estacionaria.

Usualmente se presentan con mayor frecuencia los modelos multiplicativos comparados con los modelos aditivos, de esta manera se combinan términos ordinarios del proceso y términos estacionales, así como diferencias regulares y diferencias estacionales para transformar en series estacionarias, esto es $\nabla_s^D \nabla^d X_t$. Este tipo de procesos tienen las siguientes características:

- Contiene un componente ARIMA(p, d, q) que modela la dependencia regular, que es la dependencia asociada a observaciones consecutivas.
- Contiene un componente ARIMA(P, D, Q) que modela la dependencia estacional, que está asociada a observaciones separadas por s periodos.

La estructura general de un modelo ARIMA$(p, d, q)(P, D, Q)s$ es:

$$\begin{aligned} X_t = c + \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + \Theta_1 X_{t-s} + \cdots + \Theta_p X_{t-Ps} \\ + \epsilon_t - \phi_1 \epsilon_{t-1} - \cdots - \phi_q \epsilon_{t-q} - v_1 \epsilon_{t-s} - \cdots - v_Q \epsilon_{t-Qs} \end{aligned} \tag{2.4}$$

Los parámetros son $\Phi_1, ..., \Phi_p, \Theta_1, ..., \Theta_P, \phi_1, ..., \phi_q, v_1, v_Q$ y $\epsilon_t \in N(0, \sigma^2)$

Ventajas de los modelos SARIMA:

- Incorporan estacionalidad: Son muy útiles para datos con patrones estacionales regulares, ya que incluyen componentes específicos para capturar estas fluctuaciones.
- Interpretabilidad: Los parámetros del modelo (AR, I, MA y sus contrapartes estacionales) son fácilmente interpretables y pueden ayudar a entender el comportamiento de la serie temporal.
- Predicciones precisas a corto plazo: Si los datos son estables y cumplen con los supuestos del modelo, SARIMA suele ofrecer predicciones confiables a corto plazo.
- Metodología bien establecida: Existe una amplia documentación y herramientas para ajustar y evaluar modelos SARIMA, lo que facilita su implementación.

- Flexibilidad: Pueden ajustarse tanto a series estacionarias como no estacionarias (a través de la diferenciación) y manejar patrones complejos combinando los componentes estacionales y no estacionales.

Desventajas de los modelos SARIMA:

- Supone estacionariedad: Aunque puede trabajar con series no estacionarias mediante diferenciación, las series altamente volátiles o con cambios estructurales o con cambios estructurales significativos no se modelan bien.
- No considera factores externos: No incluye variables explicativas (como clima, días festivos, promociones), por lo que no captura el efecto de eventos exógenos.
- Complejidad en el ajuste de parámetros: Elegir los valores óptimos para p, d, q, P, D, Q y s puede ser laborioso y requiere tiempo, además de experiencia.
- Sensibilidad a datos faltantes: Los datos faltantes o de baja calidad pueden afectar significativamente el rendimiento del modelo.
- Limitado en horizonte de predicción largo: Las predicciones mas allá del corto o mediano plazo tienden a ser menos confiables, especialmente si los patrones estacionales cambian con el tiempo.

StatsForecast

Según la página oficial de Nixtla, StatsForecast [7] ofrece una colección de modelos populares de pronóstico de series de tiempo univariadas optimizados para alto rendimiento y escalabilidad.

StatsForecast surge debido a que las alternativas actuales de Python para modelos estadísticos son lentas, imprecisas y no escalan bien. Por ello, crearon una biblioteca que permite realizar pronósticos en entornos de producción o como benchmark. StatsForecast incluye una amplia gama de modelos que pueden ajustar eficazmente millones de series de tiempo.

Características:

- Implementación más rápidas y precisas de AutoARIMA, AutoETA, AutoCES, MSTL y Theta en python.
- Compatibilidad fuera de la caja con Spark, Dask y Ray.

- Pronósticos probabilísticos e intervalos de confianza.
- Soporte para variables exógenas y covariables estáticos.
- Detección de anomalías.
- Sintaxis familiar de sklearn: .fit y .predict

Aspectos destacados:

- Inclusión de variables exógenas e intervalos de predicción para ARIMA.
- 20x más rápido que pmdarima.
- 1.5x más rápido que R.
- 500 veces mas rápido que Prophet.
- 4x más rápido que statsmodels.
- Compilado al alto rendimiento de código a través de numba.
- 1,000,000 de series en 30 minutos con ray.
- Reemplaza FB-Prophet en dos líneas de código y gana velocidad y precisión.
- Ajuste 10 modelos benchmark en 1,000,000 de series en menos de 5 minutos.

Pronóstico automático.

Herramientas de pronóstico automático que buscan los mejores parámetros y seleccionan el mejor modelo posible para un grupo de series de tiempo. Estas herramientas son útil para grandes colecciones de series temporales univariadas.

AutoArima en StatsForecast.

Según Nixtla [8] un autoARIMA es un modelo de series de tiempo que utiliza un procesado automático para seleccionar el ARIMA óptimo cambiando los parámetros del modelo para una serie de tiempo determinada.

El proceso de selección automática de parámetros en un modelo de autoARIMA es realizado mediante técnicas estadísticas y de optimización, tales como la Criterio de Información de Akaike (AIC) y validación cruzada, para identificar valores óptimos para la

autoregresividad, integración y promedio móvil parámetros del modelo ARIMA.

La selección automática de parámetros es útil porque puede ser difícil determinar los parámetros óptimos de un modelo ARIMA para una serie de tiempo determinada sin una comprensión completa del proceso estocástico subyacente que genera la serie de tiempo. El modelo autoARIMA automatiza el proceso de selección de parámetros y puede proporcionar una solución para el modelado y predicción de series temporales.

La librería de statsforecast.models trae la función AutoARIMA de Python la cual proporciona una implementación de autoARIMA que permite seleccionar automáticamente los parámetros óptimos para un modelo ARIMA dado una serie de tiempo.

Usando un modelo AutoARIMA() para modelar y predecir series de tiempo hay varias ventajas, entre ellas:

- Automatización del proceso de selección de parámetros: La función AutoARIMA() automatiza el proceso de selección de parámetros del modelo ARIMA, que puede ahorrar tiempo y esfuerzo del usuario eliminando la necesidad de intentar manualmente diferentes combinaciones de parámetros.
- Reducción del error de predicción: Al seleccionar automáticamente los parámetros óptimos, el modelo ARIMA puede mejorar la precisión de predicciones en comparación con modelos ARIMA seleccionados manualmente.
- Identificación de patrones complejos: la función AutoARIMA() puede identificar patrones complejos en los datos que pueden ser difíciles de detectar visualmente o con otras técnicas de modelado de series de tiempo.
- Flexibilidad en la elección de la metodología de selección de parámetros: El Modelo ARIMA puede utilizar diferentes metodologías para seleccionar los parámetros óptimos, como el Criterio de Información de Akaike (AIC), la validación cruzada y otras, lo que permite al usuario elegir el metodología que mejor se adapte a sus necesidades.

En general, utilizando la función AutoARIMA() puede ayudar a mejorar la eficiencia y exactitud del modelado y predicción de series temporales, especialmente para los usuarios que no tienen experiencia con la selección de parámetros de forma manual para modelos ARIMA.

AutoARIMA cuenta con el parámetro Level, este parámetro opcional se utiliza para la predicción probabilística. Establece el nivel (o porcentaje de confianza) de su intervalo de predicción. Por ejemplo, level=[95] significa que el modelo espera que el valor real esté dentro de ese intervalo 95 por ciento de las veces.

De igual manera, Nixtla [9] cuenta con la incorporación de variables exógenas al modelo AutoARIMA.

Las variables exógenas o los factores externos son cruciales en la predicción de series temporales, ya que proporcionan información adicional que no está directamente relacionada con la variable a pronosticar pero que podría influir en la predicción. Estas variables podrían incluir marcadores de vacaciones, gasto de marketing, datos meteorológicos o cualquier otro dato externo que se correlacione con los datos de las series temporales a pronosticar. Normalmente se recogen de fuentes externas y al incorporarlos en un modelo de predicción, pueden mejorar la precisión de nuestras predicciones.

La incorporación de estas variables exógenas al modelo no se hace únicamente en los datos históricos con los que se entrena al modelo, sino que también se agrega dicha varible en los datos o periodos a pronosticar. En este proyecto se utilizarán los días festivos, así como los días en los cuales la tienda de abarrotes la cual es objeto de estudio no abrió o no realizó ventas como variable exógena, de esta manera, al pronosticar los próximos 15 días de ventas diarias, se agrega al modelo como variable exógena todo aquel día que no sea sábado o domingo y no se piense tener ventas.

2.4 *Descripción de las métricas*

Poder medir adecuadamente los resultados de nuestro modelo es de vital impotancia, debido a que si no se hace una medición correcta de los resultados, podríamos incurrir en conclusiones equivocadas, o no contar con el modelo de predicción óptimo para nuestro pronóstico.

Es por esto, que para poder medir los resultados obtenidos por nuestros modelos que nos pronostican las ventas diarias de las tiendas de abarrotes dentro de la zona metropolitana de Guadalajara, se han seleccionado las siguientes tres métricas:

- RMSE: Raíz del Error Cuadrático Medio.
- MAE: Error Absoluto Medio.

- MASE: Error Absoluto Medio Escalado.

Para poder entender el porque de la selección de dichas métricas, es indispensable primero comprender las métricas y su medición.

RMSE

Oracle [10] nos menciona que la Raíz del Error Cuadrático Medio es la desviación estándar de los valores residuales (errores de predicción). Los valroes residuales son una medida de la distancia de los puntos de datos de la línea de regresión. RMSE es una medida de cuál es el nivel de dispersión de estos valores residuales. En otras palabras, le indica el nivel de concentración de los datos en la línea de mejor ajuste.

Fórmula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i + y_i)^2}{n}} \tag{2.5}$$

Donde $\hat{y}_1, \hat{y}_2, ..., \hat{y}_n$ son los valores de la predicción, $y_1, y_2, ..., y_n$ son los valores observados y n es el número de observaciones.

MAE

Sitiobigdata [11] menciona que el error se calcula como un promedio de diferencias absolutas entre los valores objetivo y las predicciones. El MAE es una puntuación lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio. Esta métrica penaliza a los errores grandes, por lo tanto, no es tan sensible a los valores atípicos.

Fórmula:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2.6}$$

Donde $\hat{y}_1, \hat{y}_2, ..., \hat{y}_n$ son los valores de la predicción, $y_1, y_2, ..., y_n$ son los valores observados y n es el número de observaciones.

MASE

El Error Absoluto Medio Escalado como menciona numxl [12] y Hyndman y Athanasopoulus [13] es una métrica diseñada para evaluar el rendimiento de un modelo de predicción comparándolo con un modelo base, normalmente el modelo naive (que usa el valor previo como predicción).

MASE es una métrica robusta y escalable para comparar series temporales y modelos. Debido a que es independiente a la escala de los datos, lo que permite comparar series de distintas unidades, maneja bien valores cercanos a cero o incluso negativos.

Fórmula:

$$MASE = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|} \tag{2.7}$$

Como se puede observar, la parte del denominador es el Error Absoluto Medio (MAE), por lo cual, el MASE requiere de un mayor costo computacional para su cálculo.

Interpretación:

- MASE = 1: El modelo tiene el mismo rendimiento que el modelo naive. Por lo tanto, el modelo evaluado no mejora respecto a una predicción simple basada en el último valor observado.
- MASE < 1: El modelo evaluado es mejor que el modelo naive, lo que indica que genera errores menores en promedio.
- MASE > 1: El modelo evaluado es peor que el modelo naive, ya que produce errores mayores en promedio.
- MASE cercano a 0: El modelo evaluado es muy preciso, ya que los errores son casi inexistentes en comparación con el modelo base.

2.5 Descripción de los experimentos o simulaciones

La primer simulación realizada para el pronóstico de las ventas diarias de las tiendas de abarrotes dentro de la zona metropolitana de guadalajara fue un modelo SARIMA, en donde al realizar el pronóstico se obtuvieron los siguientes resultados.

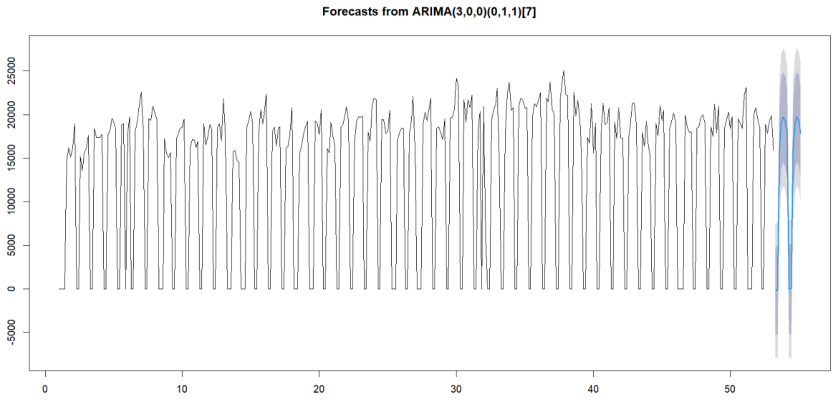

Figura 2.8: Pronóstico con SARIMA

En la figura 2.8 podemos observar el pronóstico de los próximos 15 días de ventas diarias y se puede observar que el modelo logró predecir de manera correcta la estacionalidad de la serie, así como

otorgar una pronóstico con intervalos de confianza, lo que permite obtener resultados mas confiables para la toma de decisiones.

Para poder confirmar si el modelo tiene un buen ajuste a la serie de tiempo, se realiza una prueba de backtesting, dicha prueba consiste en pronosticar un periodo pasado con el fin de poder comparar los datos reales con los pronosticados y de esta manera poder determinar si el modelo logró un ajuste correcto.

En esta prueba de backtesting se utiliza un fin de semana en donde hubo un día festivo y por lo tanto no hubo ventas, con el fin de observar si el modelo logra predecir de manera correcta este comportamiento atípico.

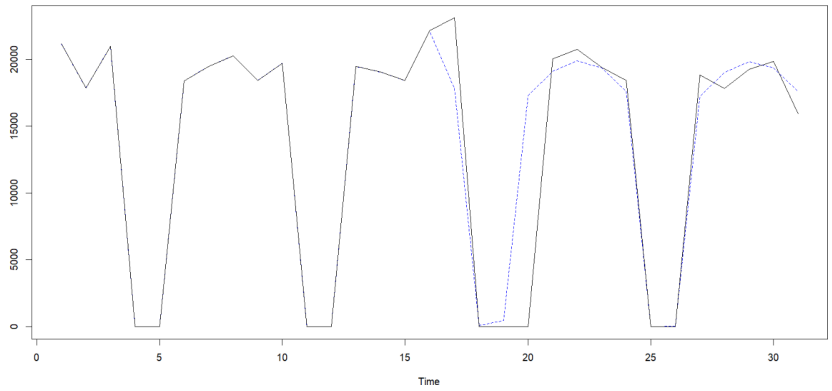

Figura 2.9: Backtesting con SARIMA

En la figura 2.9 se puede ver con una línea negra los datos reales de las ventas en los últimos 30 días, mientras que de color azul observamos el pronóstico obtenido por nuestro modelo ARIMA, en donde observamos que el modelo no logró predecir de manera correcta el día festivo, dado que pronostica ventas para dicho día como las que podríamos encontrar en un lunes ordinario, siendo que en la realidad no hubo ventas en dicho día.

Con esta simulación se pudo descubrir que para el desarrollo del proyecto se necesitaría un modelo que pudiera recibir un parámetro adicional el cual sería una variable exógena, la cual como se definió anteriormente, serán todos aquellos días entre lunes y viernes en los cuales no se cuente con ventas, para así poder predecir de manera correcta dichos casos atípicos.

Lo anterior nos guía a la segunda simulación realizada, la cual fue hecha con un modelo SARIMAX el cual es un modelo SARIMA en el que se incorpora una variable exógena con la finalidad de predecir la serie de tiempo con mayor información.

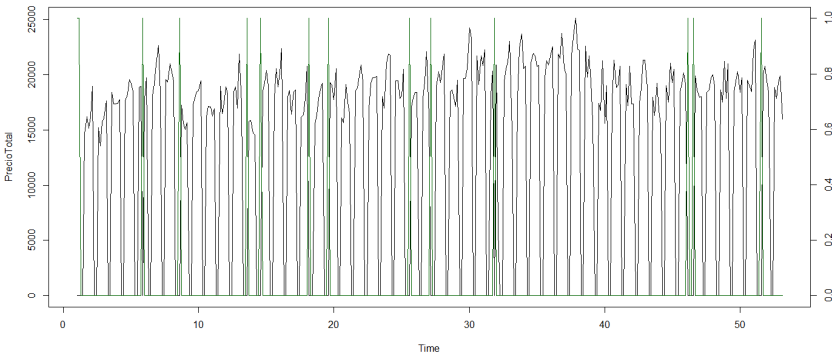

Figura 2.10: Histórico de ventas y variable exógena

En la figura 2.10 podemos observar un gráfico con el histórico de las ventas diarias y a su vez en verde podemos observar los días de lunes a viernes en los que no hubo ventas, los cuales definimos como festivos en nuestra variable exógena.

Una vez definido lo anterior, se procede a realizar el pronóstico de las ventas por los próximos 15 días incorporando dicha variable al modelo.

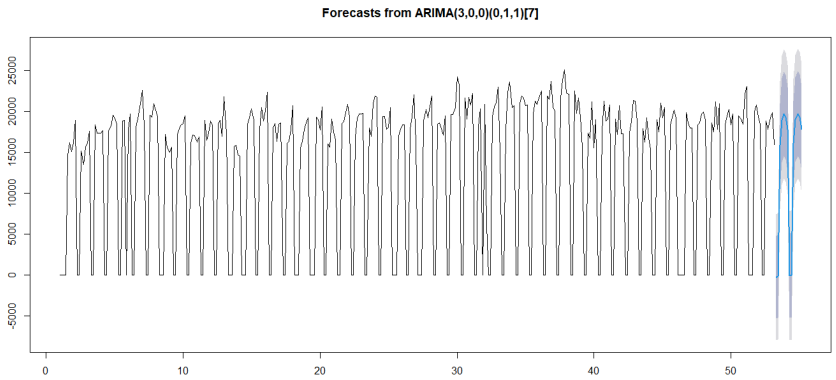

Figura 2.11: Pronóstico con SARIMAX

En la figura 2.11 podemos encontrar el pronóstico de los próximos 15 días en donde se observa una vez mas que el modelo logró predecir de manera correcta y precisa la estacionalidad de la serie y los niveles de ventas, por lo tanto, se procede a realizar una prueba de backtesting para comprobar si con la implementación de la variable exógena se logra predecir de manera correcta el comportamiento atípico del día festivo.

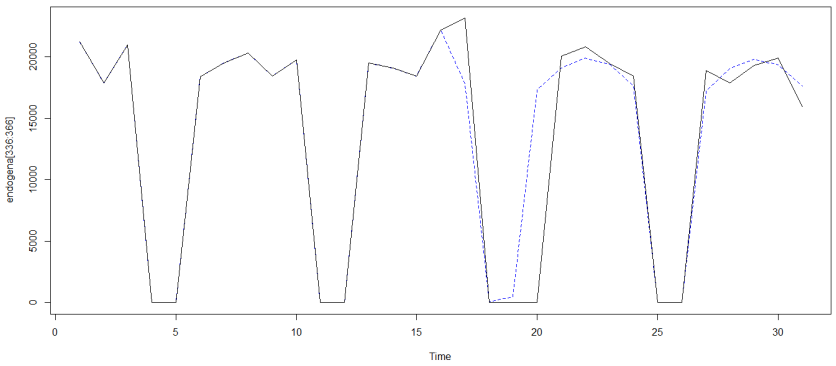

Figura 2.12: Backtesting con SARIMAX

En la figura 2.12 podemos observar que el modelo SARIMAX no logró predecir de manera correcta el lunes 16 de septiembe el cual es un día festivo en México y no hubo ventas en dicho día.

En busqueda de solucionar dicho problema, se optó por utilizar la paquetería Nixtla, la cual cuenta con múltiples modelos los cuales permiten el uso de variables exógenas, no sólo en el entrenamiento del modelo, sino también en los días a pronosticar, permitiendo así proporcionar al modelo la información de los días próximos en los cuales no se tendrán ventas y que no sean sábado o domingo.

El primer modelo a utilizar dentro de la paquetería Nixtla fue Nixtla client, el cual es un modelo que se destaca por usar días festivos como variable exógena para la predicción de una serie de tiempo, dicho modelo únicamente permite pronosticar 7 días obteniendo así los siguientes resultados.

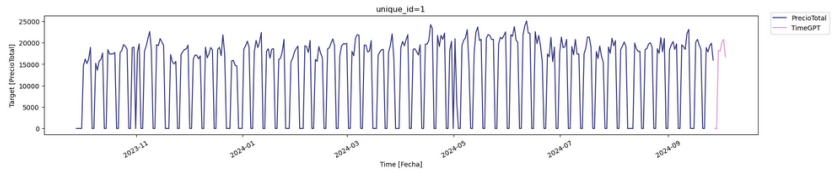

Figura 2.13: Pronóstico con Client

En la figura 2.13 se puede ver que el modelo una vez mas logró predecir de manera correcta la estacionalidad y niveles de ventas.

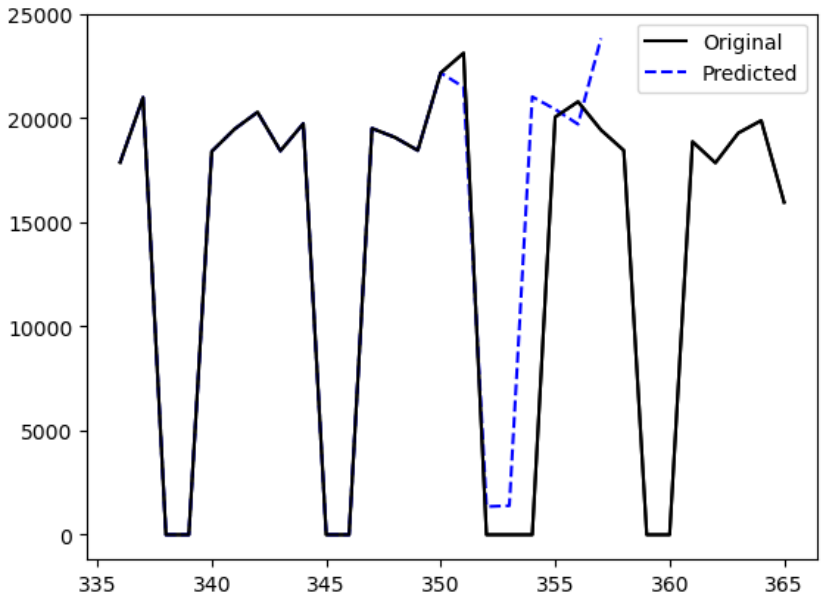

Figura 2.14: Backtesting con Client

A pesar de haber tenido un buen estimado en el pronóstico para los próximos 7 días, al momento de hacer la prueba de backtesting que se observa en la figura 2.14 vemos que no logró predecir el día lunes en el cual no hubo ventas, de igual manera el último día pronosticado contiene un error bastante grande comparado con otros modelos.

El segundo modelo utilizado dentro la paquería Nixtla fue Neural Forecast el cual ofrece una gran colección de modelos de pronóstico neuronal enfocados en su usabilidad y robustez. En esta simulación se probaron los modelos de redes "NHITS"(Interpolación Jerárquica Neuronal para Series de Tiempo) y "BiTCN"(Red Convolutiva Bidireccional Temporal) los cuales ambos permiten variables exógenas.

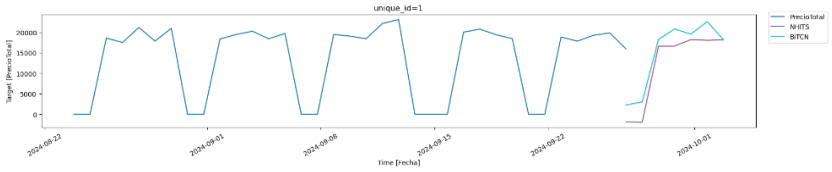

Figura 2.15: Pronóstico con Neural

En la figura 2.15 observamos los resultados por los dos modelos empleados dentro de Neural forecast, donde ambos predicen de manera correcta la estacionalidad de los fines de semana, de igual manera el modelo BiTCN a pesar de detectar que el comportamiento de sábados y domingos es diferente al resto, predice ventas de dos mil pesos para el día sábado y tres mil para el domingo, mientras que de lunes a viernes

podemos observar que el modelo BiTCN tiene una mayor variación entre las ventas diarias mientras que el pronostico de NHITS es mas constante para dichos días.

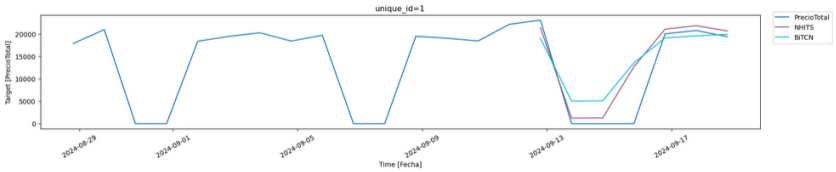

Figura 2.16: Backtesting con Neural

En la prueba de backtesting observada en la figura 2.16 observamos que ninguno de los dos modelos logró predecir de manera correcta el día festivo. Siendo que de nuevo el modelo BiTCN nos predice ventas bastante grandes para el sábado y domingo, mientras que el modelo NHITS da valores mas acercados a la realidad para dichos días.

Por último, se utilizó el modelo StastForecast de la misma paquetería Nixtla, el cual fue el modelo que se utilizó finalmente, únicamente con un ajuste debido a que al utilizar dicho modelo, pudimos observar que para los días en los que no se tienen ventas, como lo son los días festivos o fines de semana el modelo predice ventas muy bajas o incluso negativas, como lo podemos observar en la figura 2.17 y tabla 2.1.

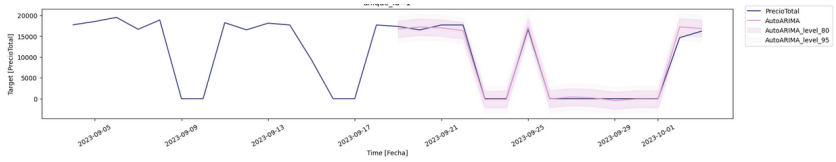

Figura 2.17: Pronóstico ventas sin ajuste

Fecha	Ventas	Forecast
2023-09-19	17,344	16,713
2023-09-20	16,502	17,204
2023-09-21	17,697	16,988
2023-09-22	17,689	16,360
2023-09-23	0	−158
2023-09-24	0	−161
2023-09-25	16,668	17,182
2023-09-26	0	−78
2023-09-27	0	405
2023-09-28	0	181
2023-09-29	0	−454
2023-09-30	0	−89
2023-10-01	0	−98
20243-10-02	14,641	17,239
2023-10-03	16,194	16,864

Tabla 2.1: Pronóstico de ventas sin ajuste

Para poder solucionar esta situación y poder contar con mejores estimaciones para las ventas, se procede a definir un umbral, en donde todas las ventas menores a dicho umbral se sustituirán con cero. El valor de dicho umbral a utilizar para los resultados del pronóstico fue 5,000 pesos. Por lo tanto, los resultados que se observarán en la sección *3.1 Resultados y discusión* ya contienen el ajuste por dicho umbral.

3 Resultados y discusión

En este capítulo se presentarán los resultados obtenidos del desarrollo de este trabajo y una discusión sobre los mismos.

3.1 Resultados y discusión

Para poder comprender los resultados obtenido primero es primordial conocer la descomposición de la serie de tiempo de las ventas diarias analizadas. La cual nos permitirá conocer mas acerca de la estacionalidad y tendencia de la misma serie.

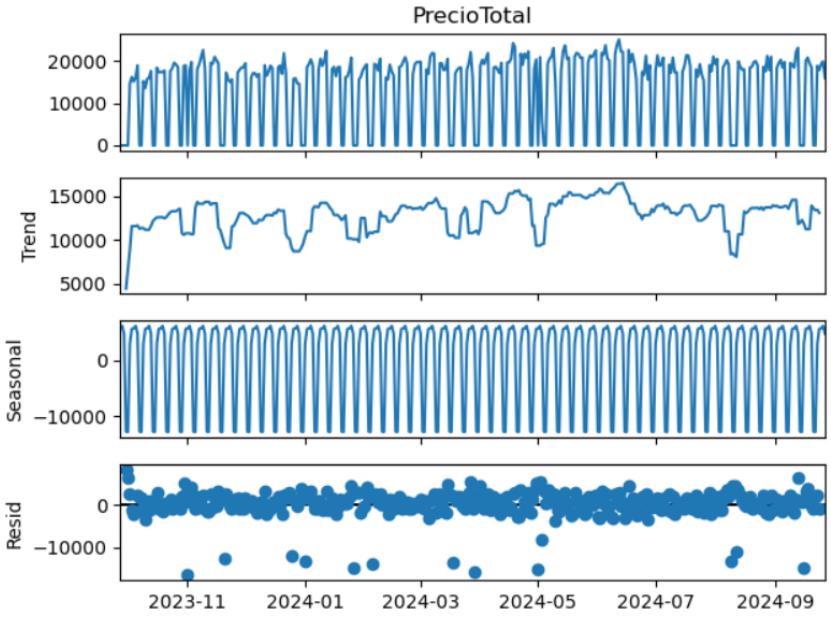

Figura 3.1: Decomposición de la serie de tiempo

En la figura 3.1 podemos observar que la serie de tiempo de las ventas diarias cuenta con una estacionalidad semanal de 7 días, así como una tendencia que varía a lo largo de los meses teniendo ventas mas altas en verano y mas bajas en invierni, pero tienden a permanecer en el mismos rango de valores.

Una vez comprendida nuestra serie de tiempo, se procede a realizar nuestro pronóstico para los próximos 15 días de ventas diarias de la tienda de abarrotes analizada dentro de la Zona Metropolitana de Guadalajara.

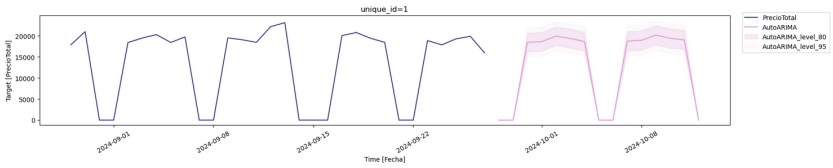

Figura 3.2: Pronóstico 15 días

Fecha	Forecast	95 low	95 high	80 low	80 high
2024-09-28	0	0	0	0	0
2024-09-29	0	0	0	0	0
2024-09-30	18,453	15,049	21,857	16,227	20,679
2024-10-01	18,618	15,195	22,042	16,380	20,857
2024-10-02	19,222	16,489	23,355	17,677	22,167
2024-10-03	19,377	15,940	22,815	17,130	21,625
2024-10-04	18,654	15,214	22,094	16,405	20,903
2024-10-05	0	0	0	0	0
2024-10-06	0	0	0	0	0
2024-10-07	18,726	15,208	22,243	16,426	21,025
2024-10-08	18,945	15,426	22,464	16,644	21,246
2024-10-09	20,155	16,636	23,675	17,854	22,456
2024-10-10	19,423	15,904	22,943	17,122	21,725
2024-10-11	19,014	15,494	22,534	16,713	21,316
2024-10-12	0	0	0	0	0

Tabla 3.1: Pronóstico 15 días

En la figura 3.2 y tabla 3.1 podemos observar el pronóstico realizado, junto con los intervalos de confianza para 95 y 80 por ciento.

Para poder determinar si el modelo propuesto es un buen modelo, se procede a llevar a cabo una prueba backtesting. En donde se pronosticarán ventas diarias pasadas y se compararán contra las reales para de esta manera poder calcular las métricas de evaluación al desempeño descritas en la sección 2.4 y determinar si tenemos un buen modelo.

En la primer prueba de backtesting se realiza el pronóstico con los últimos 15 días con los cuales se cuenta con ventas, estas fechas van desde el 13 de septiembre del 2024 al 27 de septiembre del 2024.

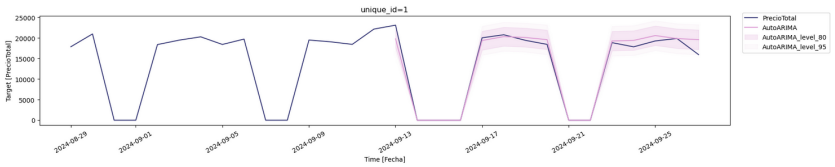

Figura 3.3: Primer Backtesting

Fecha	Ventas	Forecast	95 low	95 high	80 low	80 high
2024-09-13	23,115	19,790	16,498	23,082	17,637	21,943
2024-09-14	0	0	0	0	0	0
2024-09-15	0	0	0	0	0	0
2024-09-16	0	0	0	0	0	0
2024-09-17	20,044	19,380	15,881	22,879	17,092	21,668
2024-09-18	20,781	20,338	16,838	23,839	18,050	22,627
2024-09-19	19,425	20,120	16,619	23,621	17,831	22,409
2024-09-20	18,441	19,594	15,992	23,195	17,239	21,948
2024-09-21	0	0	0	0	0	0
2024-09-22	0	0	0	0	0	0
2024-09-23	18,862	19,272	15,657	22,887	16,908	21,635
2024-09-24	17,847	19,407	15,792	23,022	17,043	21,771
2024-09-25	19,279	20,587	16,972	24,202	18,223	22,951
2024-09-26	19,877	19,857	16,242	23,472	17,493	22,221
2024-09-27	15,953	19,599	15,951	23,247	17,214	21,984

Tabla 3.2: Primer Backtesting

En la figura 3.3 podemos observar la comparativa de los datos reales en color azul, mientras que el pronóstico de las ventas para el mismo periodo lo podemos encontrar en color rosa.

En esta prueba de backtesting podemos observar un comportamiento muy particular, debido a que el cuarto día pronosticado, a pesar de ser lunes, no hubo ventas, esto debido a que en México el 16 de septiembre se celebra el día de la Independencia, por lo tanto es un día festivo. Se puede observar que el modelo logró captar de manera correcta el comportamiento de este día debido a que no pronosticó ventas como un lunes cualquiera. De igual manera, el resto de días se logró predecir correctamente, tanto la estacionalidad como los montos de las ventas, logrando así contener dentro de los intervalos de confianza los datos reales.

Para poder medir de una manera cuantitativa el resultado del pronóstico se utilizan las métricas RMSE, MAE y MASE obteniendo los siguientes resultados.

Métrica	Valor
RMSE	1,440
MAE	882
MASE	0.07

Tabla 3.3: Métricas primer backtesting

Al observar la tabla 3.3 observamos un RMSE de 1,440 y MAE de 882 los cuales son bastante bajos considerando que el promedio de ventas del último año ha sido 18,957. De igual manera, el hecho que MASE sea 0.07 nos indica que el modelo es mucho mejor que un modelo Naive y el error es mínimo.

Para la segunda prueba de backtesting se usarán los datos del 14 de agosto del 2024 al 28 de agosto del 2024, en este periodo no hubo ningún día festivo, por lo tanto, es un periodo que podríamos considerar ordinario.

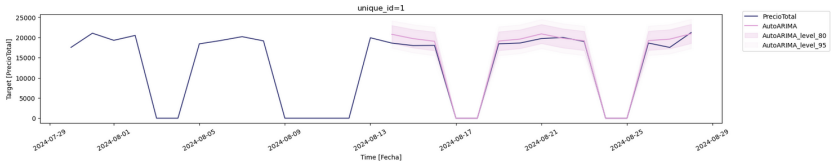

Figura 3.4: Segundo Backtesting

Fecha	Ventas	Forecast	95 low	95 high	80 low	80 high
2024-08-14	18,592	20,762	17,388	24,136	18,556	22,969
2024-08-15	17,972	19,696	16,225	23,168	17,426	21,966
2024-08-16	18,016	19,035	15,521	22,550	16,737	21,333
2024-08-17	0	0	0	0	0	0
2024-08-18	0	0	0	0	0	0
2024-08-19	18,408	19,090	15,544	22,636	16,771	21,408
2024-08-20	18,610	19,582	16,034	23,130	17,262	21,902
2024-08-21	19,715	20,871	17,263	24,478	18,512	23,229
2024-08-22	19,984	19,791	16,179	23,404	17,429	22,153
2024-08-23	18,980	19,169	15,554	22,783	16,805	21,532
2024-08-24	0	0	0	0	0	0
2024-08-25	0	0	0	0	0	0
2024-08-26	18,603	19,199	15,583	22,816	16,835	21,564
2024-08-27	17,520	19,575	15,959	23,192	17,211	21,940
2024-08-28	21,193	20,903	17,273	24,532	18,530	23,276

Tabla 3.4: Segundo Backtesting

En la figura 3.4 y tabla 3.4 podemos observar que las ventas se encuentran dentro del intervalo de confianza del 80 por ciento

de nuestro pronóstico, al igual de predecir de manera correcta la estacionalidad de 7 días.

Métrica	Valor
RMSE	1,039
MAE	736
MASE	0.07

Tabla 3.5: Métricas segundo backtesting

Al observar la tabla 3.5 observamos un RMSE bastante bajo de 1,039 así como un MAE de 736. Por último, de nuevo la métrica de MASE nos indica que nuestro modelo es bastante mejor que un modelo Naive.

En la tercer prueba de backtesting se utiliza un caso particular el cual va del 4 de agosto del 2024 al 18 de agosto del 2024. En donde no hubo ventas del viernes 9 al lunes 12 de agosto.

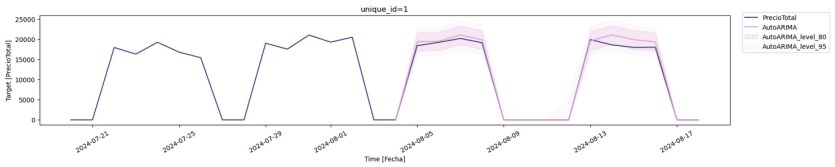

Figura 3.5: Tercer Backtesting

Fecha	Ventas	Forecast	95 low	95 high	80 low	80 high
2024-08-04	0	0	0	0	0	0
2024-08-05	18,403	19,363	15,879	22,847	17,085	21,641
2024-08-06	19,230	19,479	15,951	23,006	17,172	21,785
2024-08-07	20,182	21,049	17,499	24,598	18,728	23,369
2024-08-08	19,138	19,837	16,277	23,397	17,509	22,164
2024-08-09	0	0	0	0	0	0
2024-08-10	0	0	0	0	0	0
2024-08-11	0	0	0	0	0	0
2024-08-12	0	0	0	0	0	0
2024-08-13	19,911	19,577	15,967	23,187	17,217	21,938
2024-08-14	18,592	21,080	17,470	24,692	18,720	23,442
2024-08-15	17,972	19,890	16,279	23,502	17,529	22,252
2024-08-16	18,016	19,305	15,693	22,917	16,943	21,667
2024-08-17	0	0	0	0	0	0
2024-08-18	0	0	0	0	0	0

Tabla 3.6: Tercer Backtesting

En la figura 3.5 se observan una vez mas muy buenos resultados, donde los días que no cuenta con ventas se detectan de manera correcta por el modelo empleado dándonos así un resultado similar a la realidad.

Métrica	Valor
RMSE	962
MAE	587
MASE	0.05

Tabla 3.7: Métricas tercer backtesting

Al observar la tabla 3.7 se pueden ver las métricas más bajas hasta el momento, obteniendo un RMSE de 962, así como un MAE de 587.

Por último, se hace una prueba mas de backtesting, en donde se utiliza un caso donde en toda una semana sólo se tuvieron ventas el día lunes 25 de septiembre del 2023, del día martes 26 de septiembre al domingo primero de octubre el negocio permaneció cerrado.

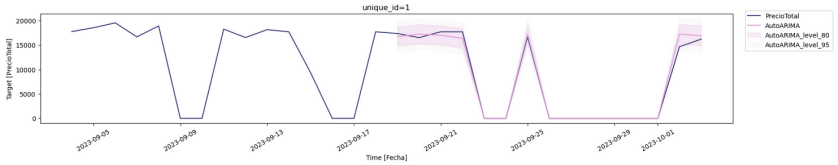

Figura 3.6: Cuarto Backtesting

Fecha	Ventas	Forecast	95 low	95 high	80 low	80 high
2023-09-19	17,344	16,713	13,666	19,760	14,720	18,705
2023-09-20	16,502	17,204	14,143	20,265	15,203	19,206
2023-09-21	17,697	16,988	13,915	20,060	14,978	18,997
2023-09-22	17,689	16,360	13,277	19,442	14,344	18,375
2023-09-23	0	0	0	0	0	0
2023-09-24	0	0	0	0	0	0
2023-09-25	16,668	17,182	14,079	20,284	15,165	19,210
2023-09-26	0	0	0	0	0	0
2023-09-27	0	0	0	0	0	0
2023-09-28	0	0	0	0	0	0
2023-09-29	0	0	0	0	0	0
2023-09-30	0	0	0	0	0	0
2023-10-01	0	0	0	0	0	0
20243-10-02	14,641	17,239	14,087	20,391	15,178	19,300
2023-10-03	16,194	16,864	13,692	20,036	14,790	18,938

Tabla 3.8: Cuarto Backtesting

Una vez mas en la figura 3.6 podemos observar como el modelo empleado logró predecir de manera correcta las ventas, tanto los días en los cuales se tienen ventas como los días en los que no, pronosticando de manera correcta las ventas dentro del intervalo de confianza al 95 por ciento.

Métrica	Valor
RMSE	841
MAE	477
MASE	0.07

Tabla 3.9: Métricas cuarto backtesting

En las métricas de evaluación al desempeño de la tabla 3.9 podemos observar un RMSE de 841 y un MAE mas bajo que en cualquier prueba anterior realizada siendo este de únicamente 477.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

Al analizar la serie de tiempo correspondiente a las ventas diarias de una tienda de abarrotes ubicada dentro de la Zona Metropolitana de Guadalajara, se observó un comportamiento relativamente estable a lo largo del año. Esta estabilidad responde a la naturaleza del negocio, enfocado en la venta de productos de primera necesidad, lo cual implica una demanda constante por parte de los consumidores. Esta característica se traduce en una serie altamente estacionaria y con patrones estacionales bien definidos, condiciones que son fundamentales para el desarrollo de modelos predictivos confiables.

Gracias a esta estructura en los datos, fue posible obtener buenos resultados utilizando modelos estadísticos relativamente simples, como el implementado con la librería StatsForecast, basado en un modelo SARIMA. La incorporación de variables exógenas, como los días festivos y aquellos considerados atípicos en los que, por diversas razones, no se registraron ventas, permitió mejorar el rendimiento del modelo, sobre todo en la predicción de eventos atípicos. Estas mejoras se reflejaron en métricas de evaluación más robustas y en una mayor precisión en los escenarios donde el comportamiento de la serie se desvía de lo ordinario.

Los resultados obtenidos representan una herramienta valiosa para los propietarios de tiendas de abarrotes, ya que permiten anticipar la demanda diaria con mayor exactitud, facilitando así la toma de decisiones estratégicas. Este tipo de soluciones permite cerrar la brecha tecnológica que existe entre pequeños negocios tradicionales y las grandes cadenas de tiendas de conveniencia, que suelen contar con recursos tecnológicos más avanzados. De esta manera, se contribuye a generar condiciones de competencia más equitativas en el sector comercial local.

Además, el enfoque propuesto tiene el potencial de escalarse y adaptarse a otras tiendas de abarrotes en la región o incluso en otras

zonas geográficas con características similares. Esto abre la puerta a futuros desarrollos donde puedan integrarse nuevas variables, como condiciones climáticas, campañas de promoción, precios, o incluso información económica de la zona, con el fin de enriquecer aún más los modelos y ofrecer soluciones más integrales.

En resumen, este trabajo demuestra que mediante el uso adecuado de técnicas de ciencia de datos y modelado estadístico es posible generar valor tangible para pequeños comercios, optimizando sus operaciones y empoderándolos con información estratégica que anteriormente solo estaba al alcance de grandes empresas.

4.2 *Trabajo futuro*

- Como trabajo futuro, se tiene planeado poner en producción todo el sistema desarrollado para la tienda de abarrotes objeto de estudio. Esto se implementará utilizando la nube pública de Amazon Web Services (AWS). El primer paso será migrar el sistema de punto de venta y su base de datos a la nube. Posteriormente, se desplegará el modelo de pronóstico de ventas diarias y se desarrollará un dashboard interactivo. Este tablero permitirá visualizar tanto las ventas históricas como los pronósticos generados, además de incluir indicadores clave de rendimiento (KPIs) y otras visualizaciones relevantes que faciliten la toma de decisiones y brinden un mayor entendimiento del negocio.
- Se recomienda realizar pronósticos de las ventas semanales y mensuales, para de esta manera poder otorgar mayor información a las tiendas de abarrotes de la zona metropolitana de Guadalajara para la toma de decisiones y planeación financiera.
- Desarrollar un modelo de predicción de ventas diarias diferenciando el método de pago, es decir, para aquellas transacciones efectuadas con dinero en efectivo y para aquellas realizadas con tarjeta de crédito o débito, con el fin de aportar información valiosa que mejore la planeación financiera de las tiendas de abarrotes de la Zona metropolitana de Guadalajara.

Bibliografía

[1] Data México. "Guadalajara". economia.gob.mx. Accedido: Abr. 19, 2025. [Online]. Disponible en: https://www.economia.gob.mx/datamexico/es/profile/geo/guadalajara-991401?populationType=totalPopulation.

[2] Data México. "Comercio al por Menor en Tiendas de Autoservicio y Departamentales". economia.gob.mx. Accedido: Abr. 19, 2025. [Online]. Disponible en: https://www.economia.gob.mx/datamexico/es/profile/industry/retail-trade-in-supermarkets-and-department-stores?measuresIndicator=Total%20Income&optionsSelector1=Total%20Income&yearEconomicCensus=option2

[3] González, R.F., Polanco, M., "Análisis de la elección del consumidor entre tiendas de conveniencia y tiendas de abarrotes en Colima. Uso del modelo de regresión multinominal logit", Paradigma económico. Revista de economía regional y sectorial, vol. 7, núm. 2, julio-diciembre, 2015, pp. 27-46.

[4] Nojek, S., Britos, P., Rossi, B. y García R. "Pronósticode ventas: comparación de predicción entre redes neuronales y método estadístico", Revista electrónica de ciencia administrativa, vol. 2, núm. 1, mayo 2003, pp. 1-18.

[5] Miranda, C., "Modelización de Series Temporales modelos clásicos y SARIMA", Trabajo de grado Maestría, Departamento de Estadística e Investigación Operativa, Universidad de Granada, Granada, España, 2021.

[6] Villavicencio, J., "Introducción a Series de Tiempo", Universidad Nacional Pedro Ruíz Gallo, Lambayeque, Perú, 2018.

[7] Nixtla. "StatsForecast". nixtla.io. Accedido: Abr. 20, 2025. [Online]. Disponible en: https://nixtlaverse.nixtla.io/statsforecast/index.html

[8] Nixtla. "AutoARIMA". nixtla.io. Accedido: Abr. 20, 2025. [Online]. Disponible en: https://nixtlaverse.nixtla.io/statsforecast/src/core/models.html#autoarima

[9] Nixtla. "Exogenous Regressors". nixtla.io. Accedido: Abr. 20, 2025. [Online]. Disponible en: https://nixtlaverse.nixtla.io/statsforecast/docs/how-to-guides/exogenous.html

[10] Oracle. ”RMSE (Error cuadrático medio)”. oracle.com. Accedido: Abr. 20, 2025. [Online]. Disponible en: https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_RMSE.htm#PFUSU-GUID-FD9381A1-81E1-4F6D-8EC4-82A6CE2A6E74

[11] Sitiobigdata. ”Aprendizaje Automático ml: Métricas de regresión”. sitiobigdata.com. Accedido: Abr. 20, 2025. [Online]. Disponible en: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/

[12] numxl. ”MASE - Error de Escala Absoluta de Media”. numxl.com. Accedido: Abr. 20, 2025. [Online]. Disponible en: https://support.numxl.com/hc/es/articles/115001223523-MASE-Error-de-Escala-Absoluta-de-Media

[13] Hyndman, R., Athanasopoulos, G., ”Forecasting Principles and Practice”, Tercera edición, OTexts: Melbourne, Australia, 2021.