ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

Predicción de la demanda de productos locales en una industria de higiene personal ecuatoriana

PROYECTO INTEGRADOR

Previo a la obtención del Título de:

Matemático

Presentado por:

Coraima Dennise Castillo Chele

Carlos Enrique Vega Hernández

GUAYAQUIL - ECUADOR

Año: 2023

DEDICATORIA

A mi abuelita Herminia Z., con quien tengo los recuerdos más bonitos de mi infancia. Porque sé lo mucho que anhelaba verme convertir en una profesional; con todo el amor incondicional que le tengo, este trabajo es para ella.

Coraima Castillo C.

DEDICATORIA

A mis padres, hermanos, amigos y familia, quienes a pesar que les dije todo el semestre que no lograría graduarme siguieron creyendo en mí (al menos la mayoría). A Julieta, Miranda, Mora y Bad Bunny que lograron hacerme feliz mientras sufría por este proyecto. Y a mi bombillo, que me da más penas que alegrías pero nunca dejaré de amarlo.

Carlos Vega H.

AGRADECIMIENTOS

Agradezco a mis padres por el apoyo y la paciencia durante todo este proceso, a mi hermana por no dejarme morir de hambre cuando no tenía tiempo de hacerme algo de comer, y a los que se consideren mis amigos en esta carrera.

Coraima Castillo C.

AGRADECIMIENTOS

Agradezco a (inserte nombre de mi pareja actual). A mis amigos de la carrera que hicieron más llevadera esta experiencia. Agradezco a nuestro director de proyecto por su paciencia. Finalmente, agradezco a Coqui por carrearme durante esta tesis.

Carlos Vega H.

DECLARACIÓN EXPRESA

"Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; *Coraima Dennise Castillo Chele* y *Carlos Enrique Vega Hernández,* damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Coraima Castillo C.

Carlos Vega H.

EVALUADORES

Firmado electrónicamente por:
LUZ ELIMAR MARCHAN MENDOZA

Digitally signed by CHRISTIAN EDUARDO GALARZA MORALES
Location: Guayaquil - Ecuador

Luz Elimar Marchan M. Ph.D.	**Christian Eduardo Galarza M. Ph.D.**
PROFESOR DE LA MATERIA	TUTOR DE PROYECTO

RESUMEN

La planificación de la demanda desempeña un papel crucial en el éxito de las operaciones comerciales y en la optimización de los recursos en empresas de diversos sectores. Pronosticar demanda es difícil debido a la influencia de factores como la naturaleza del producto, disponibilidad de datos y estrategias de la empresa. Incluso teniendo una base histórica, pueden existir variables relevantes al giro de negocio que se relacionen de forma no lineal. Por lo tanto, en este proyecto se propone un modelo predictivo para la demanda de productos locales de una industria de higiene personal ecuatoriana empleando datos históricos de venta, modificaciones de productos y actividades promocionales, para la reducción de desviaciones en la demanda proyectada para las cadenas de supermercados. En particular, se entrenó los modelos XGBoost y ARIMA con una base de 1 495 filas y 15 columnas, y se realizaron predicciones por zona de venta y tipo de producto. Se obtuvo una media en WMAPE del 35% en el negocio *Consumer Tissue*, y una medida en WMAPE del 13% en el negocio *Personal Care*. Finalmente, se interpretan las predicciones de los modelos finales, permitiendo minimizar el tiempo y el esfuerzo en el proceso de planificación de la demanda, lo que ofrece una reducción de costos adicionales y una eficiente gestión del inventario.

Palabras Clave: Planificación de la demanda, Ecuador, Industria de Higiene Personal, Modelos de predicción, Xtreme Gradient Boosting, Autoregressive Integrated Moving Average.

ABSTRACT

Demand planning plays a crucial role in the success of business operations and resource optimization in companies from various sectors. Forecasting demand is challenging due to factors such as the nature of the product, data availability, and company strategies. Even with a historical basis, there may be non-linear relationships with relevant variables related to the business sector. Therefore, this project proposes a predictive model for the demand of local products in an Ecuadorian personal hygiene industry using historical sales data, product modifications, and promotional activities to reduce deviations in projected demand for chains of supermarkets. In particular, XGBoost and ARIMA models were trained with a dataset of 1 495 rows and 15 columns, and predictions were made by sales zone and product type.A WMAPE mean of 35% was obtained in Consumer Tissue, and a WMAPE measure of 13% in Personal Care. Finally, the predictions of the final models are interpreted, allowing for a reduction in time and effort in the demand planning process, which results in additional cost savings and efficient inventory management.

Keywords: *Demand planning, Ecuador, Personal Hygiene Industry, Prediction Models, Xtreme Gradient Boosting, Autoregressive Integrated Moving Average.*

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

Este capítulo comienza analizando el desafío que implica predecir la demanda de una empresa utilizando métodos exclusivamente estratégicos, resaltando la importancia de un enfoque estadístico que pueda servir como base para la planificación de la demanda. Posteriormente, se presentan los fundamentos matemáticos en los que se sustentan los modelos ARIMA y de potenciación del gradiente, con el objetivo de que puedan ser adaptados al problema de la empresa en donde se desarrolla el proyecto.

El presente proyecto tiene como objeto de estudio las altas variaciones de la planificación de la demanda versus la demanda real de una industria de higiene personal ecuatoriana, donde se buscará abordar la hipótesis del problema encontrado, la cual establece que las desviaciones en la demanda planificada para artículos de producción local están influenciadas por el comportamiento de variables del mercado que no son consideradas dentro de las proyecciones en cada ciclo *Sales and Operation Planning* (S&OP).

1.1 Descripción del problema

La planificación de la demanda desempeña un papel crucial en el éxito de las operaciones comerciales y en la optimización de los recursos en empresas de diversos sectores.

La metodología empleada en cada organización depende de factores como: naturaleza del producto, disponibilidad de datos, objetivos y estrategias de la empresa, etc. En la organización donde se desarrolla el presente proyecto, el área de Planificación de Demanda lleva a cabo el proceso S&OP, el cual consta de cinco etapas, y en donde la demanda se proyecta a un lapso de cuatro meses mediante un consenso entre tres principales áreas: Planificación, Ventas y Trade Marketing.

En la etapa de colaboración del proceso S&OP, el área de Ventas realiza su proyección de demanda acorde a la media de venta histórica y a las oportunidades o posibles limitaciones que se observen en el mercado; mientras que el área de Trade Marketing se encarga de estimar los incrementales de los volúmenes de venta que generan las actividades promocionales propuestas. Sin embargo, en ocasiones existen grandes desviaciones en la demanda proyectada, debido a que no cuentan con un modelo predictivo que considere variables relevantes al giro de negocio, las que en general se relacionan de forma no lineal con la demanda. Como consecuencia, la disparidad entre la demanda real y la proyectada tiene impactos negativos para todas las áreas que conforman la cadena de suministro, teniéndose implicaciones que afectan la gestión de inventario, el nivel de servicio, la rentabilidad y los costos operativos.

Por último, respecto a las restricciones que se presentan para el desarrollo del proyecto, se detallan las siguientes:

- **Disponibilidad limitada de datos históricos.** Debido a los constantes cambios en el personal y la estructura del área de Trade Marketing de la organización propietaria de los datos del presente proyecto, la disponibilidad de datos históricos se ve afectada por la falta de un control riguroso de las actividades promocionales durante el año 2021 e inicios del

2022. Por ende, los datos históricos correspondientes a los descuentos y actividades promocionales que se emplearán para el entrenamiento del modelo estadístico pueden no ser completos ni confiables.

- **Restricciones legales en el acceso a información de la empresa.** Por motivos de seguridad y confidencialidad, el área legal de la organización ha establecido limitaciones en el acceso y la publicación de datos relacionados con descuentos, precios futuros y el nombre de la empresa. Por lo tanto, no es posible utilizar datos reales de la organización para evaluar y testear el modelo estadístico, y en su lugar, se deberán emplear datos ficticios.

1.2 Justificación del problema

El pronóstico de demanda en una organización cumple un papel fundamental en la cadena de suministro, puesto que influye en la compra de material y la planificación de la producción. Por ende, el propósito de este proyecto es abordar las limitaciones presentes en la planificación de demanda de una industria de higiene personal ecuatoriana. La falta de un modelo predictivo adecuado que considere diversos factores del mercado ha resultado anteriormente en pronósticos muy alejados de la realidad. Estas desviaciones generan consecuencias negativas en los equipos que conforman la cadena de suministro, afectando a los indicadores de desempeño que evalúan el nivel de servicio y la precisión de la demanda.

Una demanda proyectada inexacta puede llevar a un exceso o una escasez de inventario, lo que afecta la capacidad de cumplir con los pedidos de los clientes y

consecuentemente, en una disminución del nivel de servicio. Además, una gestión ineficiente del inventario puede generar costos adicionales y reducir la rentabilidad de la empresa. De esta manera, el proyecto se realiza con la finalidad de mejorar la precisión de las proyecciones de demanda y consecuentemente optimizar la gestión de la cadena de suministro.

1.3 Objetivos

1.3.1 Objetivo General

Proponer un modelo predictivo para la demanda de productos locales de una industria de higiene personal ecuatoriana empleando datos históricos de venta, modificaciones de productos y actividades promocionales, para la reducción de desviaciones en la demanda proyectada.

1.3.2 Objetivos Específicos

- Analizar los datos históricos de venta de artículos de producción local de una industria de higiene personal ecuatoriana, considerando periodos relevantes, para la identificación de patrones y tendencias de demanda.
- Validar un modelo predictivo que contribuya al pronóstico de demanda de los siguientes meses de los artículos de producción local de una industria de higiene personal ecuatoriana, utilizando los datos recopilados y las variables relevantes que impactan los requerimientos del cliente.

1.4 Marco teórico

El mercado de la higiene personal en Ecuador engloba una amplia gama de productos, que incluyen artículos elaborados con papel *tissue* (papel higiénico, toallas de cocina, servilletas), pañales de bebé, toallas femeninas, entre otros. En los últimos años, este mercado ha experimentado un crecimiento notable, reflejando la importancia que los consumidores ecuatorianos otorgan a la salud y el bienestar personal.

Según el informe de mercado realizado por De La Torre (2022), el sector de aseo personal en Ecuador registró un crecimiento del 11% en el año 2021 en comparación con el año anterior. Existen datos específicos sobre la demanda de papel tissue en Ecuador que también resaltan su relevancia en el mercado. Según Sánchez (2014), gerente de Operaciones de Inpaecsa, empresa ecuatoriana que oferta productos basados en papel tissue, la demanda anual de papel higiénico en el país alcanza las 45 mil toneladas. Además, un estudio realizado por Ipsa Group en 2011, reveló que los hogares ecuatorianos destinaban aproximadamente el 13% de sus egresos para la compra de papel higiénico.

Por otra parte, la industria de productos de higiene personal en Ecuador se caracteriza por la presencia de varias empresas que concentran el mercado. Según el Instituto Nacional de Estadísticas y Censos (INEC, 2017), hasta el año 2012, existían doce empresas dedicadas a este sector. Esto muestra el nivel de competencia y la importancia económica que esta industria tiene en el país.

Respecto al consumidor, su comportamiento a la hora de adquirir productos para la higiene personal ha sido objeto de estudio en diversas investigaciones. En un estudio de mercado realizado por Avilés (2018), se analizó específicamente el comportamiento de los

consumidores al adquirir productos de papel tissue, revelando algunos hallazgos importantes. Según los resultados, la marca de los productos de papel tissue no desempeña un papel determinante en la decisión de compra por parte de los consumidores. En cambio, el factor más influyente es el precio, ya que los consumidores están dispuestos a cambiar de marca si encuentran un precio más bajo o más unidades por un precio similar, incluso si el producto no presenta características distintivas. Además, se encontró que las promociones y descuentos ofrecidos por las marcas tienen un impacto significativo en la decisión de compra, que pueden llevar a los consumidores a eliminar productos previamente seleccionados de su carrito de compras.

Por otro lado, en la industria del aseo personal en Ecuador, a pesar del crecimiento experimentado en los últimos años, el comportamiento de la demanda del mercado está sujeto a diversos factores. Específicamente, en la empresa objeto de estudio, se ha observado que la demanda se ve influenciada por las actividades promocionales aprobadas mensualmente en las cadenas de autoservicio (i.e., supermercados) y los descuentos ofrecidos a los distribuidores y minimarkets. Asimismo, el precio y la participación de la competencia en el mercado desempeñan un papel relevante, al igual que las modificaciones realizadas a los productos, como cambios de imagen, metraje o códigos de barras. Del mismo modo, la situación socioeconómica del país puede ejercer influencia en el comportamiento de compra de los consumidores. La comprensión y consideración de estos factores interrelacionados resulta crucial para el análisis de la demanda y la planificación estratégica en la industria de la higiene personal en Ecuador.

1.4.1 La planificación de la demanda: importancia y métodos

La planificación de la demanda (*demand planning*) alude a un conjunto de estrategias y técnicas utilizadas para determinar la cantidad adecuada de productos a ser aprovisionados en uno o varios centros de almacenaje. Su objetivo es mantener un equilibrio entre la demanda y el suministro, con el fin de gestionar eficientemente los recursos y garantizar la disponibilidad de productos cuando sean requeridos (Meetlogistics, 2020). Acorde a un reporte realizado por Qurius (2010), la importancia de la planificación de la demanda en un negocio yace en comprender las dinámicas del mercado, mejorar la eficiencia del negocio, incrementar el nivel de servicio y controlar los requerimientos del mercado respetando los objetivos de rentabilidad mediante promociones eficaces.

Los métodos utilizados por cada empresa para la planificación de demanda dependen esencialmente de las herramientas disponibles para el planificador, y de la naturaleza del producto o servicio que se comercializa. Según Parra (2018), la demanda del mercado puede calcularse mediante métodos cualitativos y cuantitativos. Los métodos cualitativos son principalmente de juicio humano, buscando construir las proyecciones de venta a través de opiniones y consensos entre áreas. Uno de ellos es el método *Delphi* el cual consiste en realizar reiteradas consultas a expertos en el tema, para poder llegar a una conclusión común acerca del comportamiento del mercado en los próximos meses. En contraste, los métodos cuantitativos constan de modelos matemáticos que usan datos históricos. Estos se subdividen en dos grandes grupos: métodos de series de tiempo y métodos de pronóstico causal.

Para Chase et al. (2009), los modelos de series de tiempo emplean datos históricos de periodos específicos para la construcción de las proyecciones, es decir, se componen de

información segmentada en horizontes de tiempo cuyo periodo, inicio o fin los define el investigador. A su vez, Caba et al. (2011) manifiestan que los modelos de pronóstico causal realizan las estimaciones examinando los factores externos que pueden influir en las desviaciones de la demanda de un producto. Estos modelos consideran variables relacionadas con la variable objetivo (volumen de venta) y, una vez identificadas, se utiliza un modelo estadístico para predecir la variable de interés. Este enfoque es más robusto que los métodos de series de tiempo, que se basan exclusivamente en datos históricos para realizar pronósticos.

Con el fin de complementar los métodos adoptados (cuantitativos o cualitativos) para la planificación de demanda, muchas organizaciones lo refuerzan con el proceso *Sales and Operation Planning*. Según Parra (2018), el proceso S&OP busca alcanzar el equilibrio entre la demanda y la producción, mejorando la eficiencia, la comunicación entre las diferentes áreas de la empresa, logrando alcanzar de esta forma las estrategias y objetivos corporativos.

El proceso S&OP consta de varias fases que se componen de reuniones ejecutivas entre las áreas pertinentes, en las que se busca un consenso entre los departamentos para abordar temas como volúmenes y estimaciones de ventas, metas, planes comerciales y lanzamientos de nuevos productos. Este enfoque contribuye significativamente a la mejora de la coordinación y la toma de decisiones en la empresa. Las etapas del ciclo S&OP difieren en las compañías dependiendo de su nivel de madurez. En la empresa objeto de estudio, el proceso S&OP consta de cinco fases descritas en Logility Voyager Solutions (2010) de la siguiente manera:

1. **Product Review.** En esta etapa se revisa el portafolio de los meses siguientes, se determina cuáles son los productos a introducir en el mercado y los que quedarán obsoletos. En esta fase se consideran los volúmenes de lanzamiento para los productos

nuevos del mercado y se identifican cuáles son los artículos que serán canibalizados[1] por estos productos de lanzamiento. Además, se revisan los volúmenes incrementales por actividades promocionales.

2. **Demand Review.** Previo a la ejecución de esta etapa, el equipo comercial realiza la estimación de los volúmenes de venta acorde al contexto de mercado. Estas proyecciones las revisa el equipo de Planificación de Demanda empleando los métodos cualitativos o cuantitativos mencionados anteriormente. Luego, se ejecuta la consolidación de los volúmenes y se valida con las áreas pertinentes (Marketing y Ventas) que las proyecciones estén alineadas con el desarrollo del negocio.

3. **Supply Review.** En esta etapa surge el denominado *plan de demanda restringido*, el cual refleja las limitaciones que afectan a la cantidad de productos que se pueden ofrecer o producir en un determinado período de tiempo. Para precisar estas restricciones, el equipo de Planificación de Abastecimiento determina cómo puede cumplir con el plan de demanda preestablecido en el Demand Review. Es decir, en esta etapa se revisan los volúmenes de demanda considerando las capacidades de las líneas de producción, la disponibilidad de insumos y las fechas de llegada de los productos importados.

4. **Financial Review.** En esta fase se revisa el plan de demanda restringido en la etapa anterior acorde a la disponibilidad de producto y se evalúa si este plan de demanda propuesto para los siguientes meses cumple con los objetivos financieros de la empresa.

5. **Executive Business Review.** En la última fase del ciclo S&OP se presenta el plan de

[1] **Canibalización.** Se refiere a la pérdida en las ventas causada por la introducción de un nuevo producto por parte de una empresa que desplaza a uno de sus propios productos más antiguos (Traders Studio, 2021).

demanda acordado por las áreas participantes en las etapas anteriores. El objetivo de esta fase es proponer planes de acción para las limitaciones que se encuentren, con el fin de alcanzar los objetivos de rentabilidad de la empresa.

En particular, en la empresa de higiene personal donde se llevará a cabo el proyecto, se emplean métodos cualitativos y cuantitativos, específicamente, el equipo de ventas realiza las estimaciones utilizando datos históricos y basándose en la media de meses anteriores. Estos métodos de planificación de demanda se complementan con la ejecución del proceso S&OP, y de esta forma, se construye un plan de demanda mediante el consenso de varias áreas.

Finalmente, para evaluar la precisión del pronóstico, se emplea la Media del Error Absoluto en Porcentaje Ponderada (WMAPE),

$$\text{WMAPE} = \frac{\sum_{i=1}^{n} |A_i - F_i|}{\sum_{i=1}^{n} A_i} \tag{1.1}$$

donde n es el total de productos, A_i corresponde a la demanda actual del i-ésimo producto y F_i es su demanda planificada.

Similarmente, también se usa el Error Porcentual Absoluto Medio (MAPE),

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \frac{|A_i - F_i|}{A_i}}{n} \tag{1.2}$$

1.4.2 Aprendizaje Automático (Machine Learning)

El aprendizaje automático o Machine Learning (ML) es un campo de estudio y una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y mejorar automáticamente a partir de datos, sin ser enseñadas explícitamente para llevar a cabo una tarea específica. En lugar de seguir instrucciones

predefinidas, las máquinas de ML utilizan patrones y ejemplos para realizar tareas y tomar decisiones.

El objetivo principal del ML es permitir que las máquinas adquieran conocimiento y realicen predicciones o tomen decisiones basadas en datos sin intervención humana constante. A través del análisis de grandes conjuntos de datos, los algoritmos de aprendizaje automático pueden identificar patrones, tendencias y relaciones ocultas en los datos. Lo que les permite realizar tareas complejas como reconocimiento de voz, detección de fraudes, clasificación de imágenes, recomendación de productos y mucho más.

El ML se basa en diversos enfoques y técnicas, como los modelos de regresión, los árboles de decisión, las redes neuronales, el aprendizaje por refuerzo y el aprendizaje no supervisado. Estos modelos son entrenados con conjuntos de datos, donde se les proporciona ejemplos y se les permite aprender a partir de ellos. Una vez entrenados, los modelos pueden generalizar su conocimiento y aplicarlo a nuevos datos para realizar predicciones o tomar decisiones.

1.4.3 Modelos para la predicción de la demanda

Dado que el objetivo de este proyecto yace en proponer un modelo predictivo de la demanda de productos locales de una industria de higiene personal ecuatoriana, a continuación, se discutirán varios modelos matemáticos ampliamente usados para realizar proyecciones de ventas. Entre los modelos predictivos a utilizar se tienen modelos tradicionales de series temporales como los modelos ARIMA, y modelos de Machine Learning basados en árboles de regresión.

Modelos Autorregresivos Integrados de Medias Móviles (ARIMA).

Según Hayes (2022), los Modelos Autorregresivos Integrados de Medias Móviles ARIMA (ARIMA por sus siglas en inglés *AutoRegressive Integrated Moving Average*) son utilizados como herramientas de pronóstico para predecir cómo actuará algún fenómeno en el futuro en función de su desempeño pasado. Sin embargo, un modelo ARIMA generalmente es inadecuado para pronósticos a largo plazo, como los que superan los seis meses, puesto que utiliza datos y parámetros pasados que están influenciados por el pensamiento humano.

Para definir formalmente los modelos ARIMA, primero es importante conocer ciertos conceptos básicos que se presentan a continuación:

- **Serie temporal.** Según Mauricio (2007), una serie temporal (o simplemente una serie) es una secuencia de n observaciones ordenadas y equisdistantes cronológicamente sobre una característica (serie univariante o escalar) o sobre varias características (serie multivariante o vectorial) de una unidad observable en diferentes momentos. González (2009) establece que una serie temporal es univariante cuando se analiza solo en función de su propio pasado, mientras que en una serie temporal multivariante se analizan varias series temporales a la vez.

  Además, González (2009) también menciona que en un modelo de series temporales univariante, la serie Y_t se descompone en dos partes, una parte sistemática y otra parte netamente aleatoria, llamada innovación: $Y_t = PS_t + a_t$, $t = 1, 2, \cdots$. La parte sistemática es la parte predecible con el conjunto de información que se utiliza para construir el

modelo. La innovación es una parte aleatoria en las que sus valores no tienen ninguna relación o dependencia entre sí, es decir, la innovación en el momento t no está relacionada ni con las innovaciones anteriores ni con las posteriores, ni con la parte sistemática del modelo.

- **Procesos estocásticos.** Mauricio (2007) define a un proceso estocástico como una familia de variables aleatorias referidas a una (proceso univariante o escalar) o a varias (proceso multivariante o vectorial) características de una unidad observable en diferentes momentos. Entre sus representaciones frecuentes:

$$Y_t : t = 0, \pm 1, \pm 2, \ldots, \text{ ó } \{Y_t\}_{t \in \mathbb{Z}}.$$

- **Procesos estacionarios.** Un proceso estocástico, $\{Y_t\}_{t \in \mathbb{Z}}$, es estacionario en sentido estricto si y solo si: $F[Y_{t_1}, Y_{t_2}, \cdots, Y_{t_n}] = F[Y_{t_1+k}, Y_{t_2+k}, \cdots, Y_{t_n+k}]$, $\forall (t_1, t_2, \cdots, t_n)$ y $k \in \mathbb{N}$, es decir, si la función de distribución o las propiedades estadísticas de cualquier conjunto finito de n variables aleatorias del proceso, no se altera si se desplaza k periodos en el tiempo. (González, 2009)

- **Procesos no estacionarios.** Un proceso estocástico,$\{Y_t\}_{t \in \mathbb{Z}}$, es no estacionario cuando las propiedades estadísticas de al menos una secuencia finita $Y_{t_1}, Y_{t_2}, \cdots, Y_{t_n}$ de n variables aleatorias son diferentes de las de la secuencia $Y_{t_1+k}, Y_{t_2+k}, \cdots, Y_{t_n+k}$, para al menos un $k \in \mathbb{N}$. (Mauricio, 2007)

- **Ruido blanco.** Es el proceso estocástico más sencillo, corresponde a una secuencia de variables aleatorias de media cero, varianza constante y covarianzas nulas; habitualmente se denota por a_t, $t = 0, \pm 1, \pm 2, \cdots$. (González, 2009)

- **Operador de retardo.** Se representa con el símbolo L y se define de la siguiente manera: $LY_t = Y_{t-1}$, $L^dY_t = Y_{t-d}$, $d \geqslant 1 \in \mathbb{N}$, donde Y_t es una variable aleatoria referida a un momento t determinado. (González, 2009)

Ahora, para entender qué son los modelos ARIMA, se debe conocer sobre los modelos autorregresivos de media móvil (ARMA por sus siglas en inglés *Autoregressive Moving Average*).

González (2009) indica que dentro de los procesos estocásticos estacionarios se sitúan los procesos lineales, que se caracterizan porque se pueden representar como una combinación lineal de variables aleatorias. En los procesos estacionarios con distribución normal y media cero, la teoría de procesos estocásticos señala que bajo condiciones generales, Y_t se puede expresar como combinación lineal de los valores pasados infinitos de Y más una innovación ruido blanco:

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + a_t, \ \forall t, t = 1, 2, \cdots \tag{1.3}$$

Además, asegura que las condiciones generales que cumple este proceso son:

- Que el proceso sea no anticipante, es decir, el valor de Y en el momento t no puede depender de valores futuros de Y o de las innovaciones de a.

- Que el proceso sea invertible, es decir, que la influencia de Y_{t-k} en Y_t ha de ir disminuyendo conforme se vaya alejando en el pasado. Esta condición se cumple si los parámetros del modelo general (1.3) cumplen la restricción: $\sum_{t=1}^{\infty} \pi_t^2 < \infty$.

Por otro lado, resulta sencillo observar que el modelo (1.3) se puede escribir de forma más

compacta en términos del operador de retardos:

$$(1 - \pi_1 L - \pi_2 L^2 + \cdots)Y_t = a_t \Rightarrow \prod_{\infty}(L)Y_t = a_t$$

$$\Rightarrow Y_t = \frac{1}{\prod_{\infty}(L)a_t} = \psi_{\infty}(L)a_t,$$

entonces,

$$Y_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \psi_3 a_{t-3} + \cdots \tag{1.4}$$

De este modo, Y_t se puede expresar como la combinación lineal del ruido blanco a_t y su pasado infinito. El modelo (1.4) cumple la condición de estacionariedad si los parámetros satisfacen: $\sum_{t=1}^{\infty} \psi_t^2 < \infty$.

Dado que en la práctica se trabaja con series finitas, los modelos no pueden expresar dependencia infinitas sin restricciones, estos tendrán que especificar una dependencia en el tiempo acotada y con restricciones. Por ende, mediante la teoría de polinomios, González (2009) aproxima el modelo (1.3) empleando el hecho de que se puede aproximar un polinomio de orden infinito mediante un cociente de polinomios finitos. De este modo,

$$\prod_{\infty}(L) \simeq \frac{\phi_p(L)}{\theta_q(L)},$$

donde $\phi_p(L)$ y $\theta_q(L)$ son polinomios en el operador de retardos finitos de orden p y q, respectivamente. Por lo tanto,

$$\prod_{\infty}(L)Y_t = a_t \simeq \frac{\phi_p(L)}{\theta_q(L)}Y_t = a_t \Rightarrow \phi_p(L)Y_t = \theta_q(L)a_t.$$

De esta forma, González (2009) establece que el modelo lineal general admite tres representaciones:

1. Representación *puramente autorregresiva* (1.3), AR(∞): el valor presente de la variable se representa en función de su propio pasado más una innovación contemporánea.

2. Representación *puramente de medias móviles* (1.4), MA(∞): el valor presente de la variable se representa en función de todas las innovaciones presente y pasadas.

3. Representación *finita*:

$$\phi_p(L)Y_t = \theta_q(L)a_t$$

$$Y_t = \underbrace{\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}}_{\text{parte autorregresiva}} + \underbrace{a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}}_{\text{parte medias móviles}}$$

   En este modelo finito, el valor de Y_t depende del pasado de Y hasta el momento $t-p$ (parte autorregresiva), de la innovación contemporánea y su pasado hasta el momento $t-q$ (parte medias móviles). Este modelo se denomina *Autorregresivo de Media Móviles* de orden (p,q), y se denota por $ARMA(p,q)$.

Por otro lado, pese a que se ha presentado a los modelos $ARMA(p,q)$ como modelos estacionarios, es relevante aclarar que si un proceso no es estacionario en media también se puede modelar dentro de la clase de modelos $ARMA(p,q)$. Se dice que un modelo $ARM(p,q)$ no es estacionario si las raíces de su polinomio AR no satisfacen la condición de estacionariedad[2].

Finalmente, para introducir a los modelos $ARIMA(p,d,q)$, González (2009) plantea un modelo $ARM(p,q)$: $\phi_p(L)Y_t = \theta_q(L)a_t$, donde el polinomio $AR(p)$ se puede factorizar en función de sus p raíces. Luego, se supone que $(p-1)$ raíces son estacionarias y una de ellas es unitaria, $L_t = 1$. Entonces, el polinomio $AR(p)$ puede reescribirse del siguiente modo:

[2] **Condición de estacionariedad** $AR(p)$**.** Un proceso autorregreviso finito $AR(p)$ es estacionario sí y solo si el módulo de las raíces del polinomio autorregresivo $\phi(L)$ está fuera del círculo unitario. González (2009)

$\phi(L) = \varphi_{p-1}(L)(1-L)$, donde el polinomio $\varphi_{p-1}(L)$ resulta del producto de los $(p-1)$ polinomios de orden 1 asociados a las raíces L_t con módulo fuera del círculo unidad.

Al sustituir en el modelo $ARMA(p,q)$ planteado, se obtiene que:

$$\varphi_{p-1}(L)\Delta Y_t = \theta_q(L)a_t \tag{1.5}$$

donde φ_{p-1} es estacionario porque todas sus raíces tienen módulo fuera del círculo unitario, y $\Delta = 1 - L$ es el que corresponde a la raíz unitaria.

El modelo (1.5) representa el comportamiento de un proceso Y_t que no es estacionario porque tiene una raíz unitaria. A un proceso Y_t con estas características se le denomina *proceso integrado de orden* 1.

En general, el polinomio $AR(p)$ del modelo (1.5) puede contener más de una raíz unitaria, por ejemplo, d, entonces se puede descomponer de la siguiente manera:

$$\phi_p(L) = \varphi_{p-d}(L)(1-L)^d$$

y sustituyendo en el modelo $ARMA(p,q)$ se tiene que: $\varphi_{p-d}(L)\Delta^d Y_t = \theta_q(L)a_t$, donde el polinomio $\varphi_{p-d}(L)$ es estacionario porque sus $(p-d)$ raíces tienen módulo fuera del círculo unidad, y el polinomio $\Delta^d = (1-L)^d$, de orden d, contiene las d raíces unitarias no estacionarias. A un proceso Y_t con estas características se le denomina *proceso integrado de orden* d y se denota por $Y_t \sim I(d)$.

Si una serie Y_t es integrada de orden d, se puede representar por el siguiente modelo:

$$\phi_p(L)\Delta^d Y_t = \delta + \theta_q(L)a_t \tag{1.6}$$

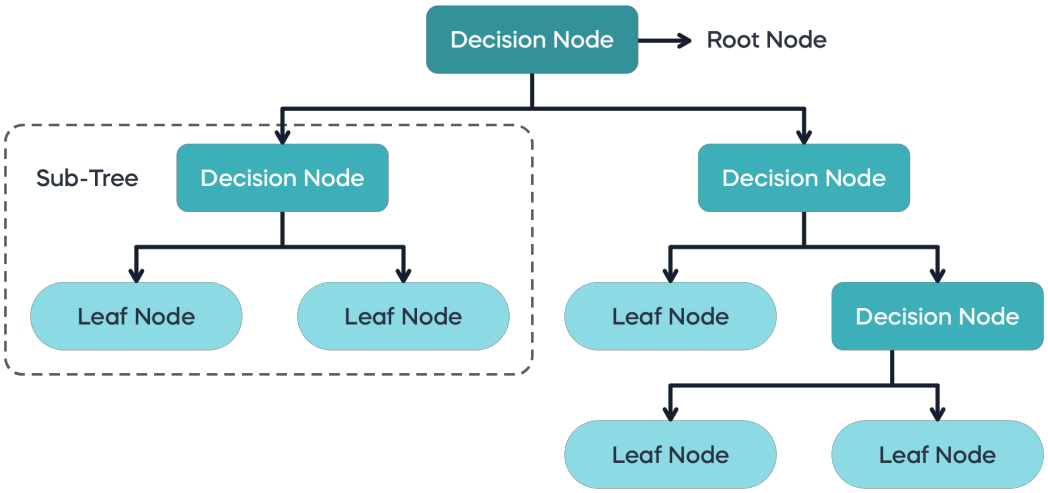
donde el polinomio autorregresivo estacionario $\phi_p(L)$ y el invertible de media móvil $\theta_q(L)$ no tienen raíces comunes.

De este modo, González (2009) presenta al modelo (1.6) como un modelo *Autorregresivo Integrado de Medias Móviles* de orden (p, d, q) o $ARIMA(p, d, q)$, donde p es el orden del polinomio autorregresivo *estacionario*, d es el de integración de la serie, y q es el orden del polinomio de medias móviles *invertible*.

Árboles de regresión.

Un árbol de decisión es un método estadístico, utilizado principalmente en ML, que permite la toma de decisiones basadas en un conjunto de suposiciones y resultados anticipados. Su base es la construcción de un árbol jerárquico como se ve en la figura 1.1, se compone por:

Figura 1.1.
Estructura de un árbol de decisión



Fuente: The 365 Team (2023)

1. Nodo: representa una pregunta o prueba sobre una variable. Estos incluyen:
    - Nodo raíz: es el primer nodo del árbol y representa el conjunto de datos completo, que luego se divide en subgrupos.
    - Nodos internos: son puntos de ramificación en el árbol que indican divisiones o particiones basadas en las características de las variables predictoras. Las ramas de estos nodos se conectan con otros nodos o nodos hoja.
    - Nodos hoja: son los nodos terminales del árbol que representan agrupaciones o subconjuntos de datos donde se generan predicciones. Cada nodo hoja tiene asociado un valor numérico que representa la predicción para ese conjunto particular de datos.
2. Rama: denota posibles soluciones o resultados.

Existen diferentes tipos de árboles de decisión que se utilizan en diversos contextos:

- Árbol de regresión (RT): Son la subfamilia de árboles de decisión que se aplica cuando la variable respuesta es continua. (Palomino Mezones, 2021)
- Árbol de clasificación: este tipo de árbol se utiliza para clasificar o categorizar datos en diferentes grupos o clases en lugar de predecir valores numéricos. Las ramas de cada nodo conducen a diferentes clases de salida, y cada nodo representa una pregunta que divide los datos según una característica.
- Árbol probabilístico: este tipo de árbol considera el grado de incertidumbre en la toma de decisiones. La decisión se determina en función de la probabilidad asociada a cada rama en lugar de seguir un único camino determinista.

- Árbol basado en reglas: a diferencia de los árboles convencionales, están compuestos por conjuntos de reglas if-then. Cada regla representa una condición y una acción adecuada. Estas reglas se organizan de manera jerárquica para reflejar diferentes combinaciones de condiciones.

Acorde a Palacios (2020), debido a su simplicidad y adaptabilidad, los árboles de decisión dentro de las ciencias económicas se enfocan en fines predictivos, análisis de riesgos, decisiones de inversión o decisiones de gestión financiera, convirtiéndolos en un objeto atractivo de investigación. En particular, indica que los RTs son un método de aprendizaje supervisado de ML usado para predecir variables continuas basado en bases de datos. Los modelos de RTs son construidos a partir de la estrategia *divide y conquistarás*, es decir, que el conjunto de datos de entrada del árbol es dividido en múltiples particiones de acuerdo con un criterio de división establecido.

Por otro lado, el análisis de árbol de clasificación y regresión, o metodología CART por sus siglas en inglés (*Classification And Regression Trees*), utiliza datos históricos para construir RTs, los cuales se emplean para predecir nuevos datos. Estos árboles CART pueden manejar tanto variables numéricas como categóricas. Entre las ventajas de CART se encuentran su robustez frente a valores atípicos (*outliers*), la invarianza en la estructura de los RTs ante transformaciones monótonas de las variables independientes y, sobre todo, su capacidad de interpretación. (Sepúlveda, 2013)

Sepúlveda (2013) indica que esta metodología se divide en tres pasos:

1. **Construcción del árbol saturado.** El árbol saturado se construye mediante particionamiento recursivo. Sea Y una variable dicotómica con valores 0 y 1, y sea τ un

nodo. En el caso de los RTs la selección natural de la impureza para un nodo τ es la varianza de la respuesta dentro del nodo:

$$i(\tau) = \sum_{i \in \tau} (Y_i - \bar{Y}(\tau))^2, \tag{1.7}$$

donde $\bar{Y}(\tau)$ es el promedio de Y_i's dentro del nodo τ. Para dividir un nodo τ en dos nodos hijos, τ_L y τ_R, se define la bondad de una división s como

$$\delta I(\tau) = i(\tau) - i(\tau_L) - i(\tau_R). \tag{1.8}$$

2. **Selección del tamaño correcto del árbol.** Posterior a la construcción del árbol T inicia el proceso de poda. La poda consiste en encontrar el subárbol del árbol saturado con la mejor calidad en cuanto a que sea lo más predictivo posible y lo menos sensible al ruido de los datos. Para esto, se determina una medida de calidad, se puede usar $i(\tau)$ para definir el costo del árbol T como

$$R(T) = \sum_{\tau \in \tilde{T}} i(\tau) \tag{1.9}$$

donde $\tilde{T}$ es el conjunto de nodos terminales de T. Luego, se define al costo-complejidad de T como

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}|, \tag{1.10}$$

donde $\alpha \geqslant 0$ es el parámetro de complejidad. La diferencia entre $R(T)$ y $R_\alpha(T)$ como una medida de la calidad del árbol reside en que $R_\alpha(T)$ penaliza un gran árbol. La idea es construir una secuencia de subárboles anidados para un árbol saturado T, minimizando el costo-complejidad $R_\alpha(T)$, y seleccionar como subárbol final el que tenga el menor costo de mala clasificación de estos subárboles.

Si se dispone de una muestra de prueba, la estimación de $R(T)$ resulta sencilla para cualquier subárbol T, ya que solo es necesario aplicar los subárboles a la muestra y seleccionar el mejor valor de α. Sin embargo, en ausencia de una muestra de prueba, es posible generar muestras artificiales mediante el proceso de validación cruzada para estimar $R(T)$ y determinar así el valor óptimo de α.

3. **Clasificación de nuevos datos empleando el árbol seleccionado.** Una vez seleccionado el árbol más óptimo, se procede a utilizarlos en nuevos datos para realizar predicciones y medir su error.

Para determinar el error de predicción de la metodología CART basta con observar la forma de los valores predichos por el RT. Si se tiene un conjunto de datos $(x_1, y_1), ..., (x_n, y_n)$, entonces:

$$y_{cart_i} = \begin{cases} r_k \; si \; x_i \in C_k; \quad k = 1, ..., l \\ \\ 0 \; en \; otro \; caso \end{cases}$$

donde,

$$r_k = \frac{\sum\{y_i | x_i \in C_k, i = 1, ..., n\}}{\#(\{y_i | x_i \in C_k, i = 1, ..., n\})}; \quad k = 1, ..., l.$$

Por lo tanto, el error de predicción por CART se define como

$$EPCART = \frac{\sum_{i=1}^{n} (y_{cart_i} - y_{verd_i})^2}{n}. \tag{1.11}$$

donde $y_{verd_i} = E(y_i)$. (Sepúlveda, 2013)

Gradient Boosting

El concepto fundamental del boosting radica en abordar problemas complejos de aprendizaje automático al combinar clasificadores débiles e imprecisos de manera cuidadosa, con el objetivo de formar un conjunto cuyas predicciones sean muy precisas. En otras palabras, un algoritmo de boosting se enfoca en construir un comité inteligente compuesto por miembros inexpertos. (Martínez Celda, 2021)

Los Gradient Boosting models (GB) son métodos de aprendizaje automático utilizados para crear modelos predictivos extremadamente precisos. Estos modelos son un subconjunto de la familia de algoritmos de boosting, que combina varios modelos más débiles en un modelo general más sólido. Los modelos predictivos se construyen uno tras otro, ajustando cada uno a los errores residuales del modelo anterior. Como resultado, la precisión y el rendimiento del conjunto de modelos mejoran constantemente, ya que cada nuevo modelo se centra en aprender y corregir los errores producidos por los modelos anteriores.

Los GB se crean utilizando el enfoque del descenso de gradiente para optimizar una función de pérdida. Este enfoque mejora continuamente la capacidad predictiva al minimizar la función de pérdida y ajustar de forma iterativa los parámetros del modelo. Son herramientas efectivas en el campo del ML porque pueden manejar características complicadas y capturar correlaciones no lineales. Se aplican con frecuencia a problemas de regresión y clasificación, y tienen una amplia gama de aplicaciones, que incluyen análisis de texto, sistemas de recomendación, pronóstico de series temporales y detección de fraudes.

Martínez Celda (2021) muestra cómo funciona un algoritmo de boosting que toma como parámetro de entrada un conjunto de muestras de entrenamiento $S = (x_1, y_1), ..., (x_n, y_n)$ donde

cada x_i es un elemento de X, y cada $y_i \in Y$ es su etiqueta asociada. Asumiendo el caso más sencillo en el cual solo hay dos clases: $Y = \{-1, +1\}$. Un algoritmo de aprendizaje A toma como entrada al conjunto de entrenamiento S y genera un clasificador $h : X \Rightarrow Y = \{-1, +1\}$ que minimice el error de entrenamiento dado por

$$\epsilon_S = \frac{1}{m} \sum_{i=1}^{m} I(y_i \neq h(x_i)), \quad (x_i, y_i) \in S \; \forall i = 1, ..., m. \tag{1.12}$$

El principio del boosting es aplicar sucesivamente el algoritmo base a versiones modificadas del conjunto de entrenamiento, de este modo en cada iteración se genera un nuevo clasificador débil h_t que posteriormente se combina formando un clasificador H. Generalmente, los clasificadores débiles se combinan a través de una suma ponderada:

$$H(x) = signo\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{1.13}$$

Los pesos $\alpha_1, ..., \alpha_T$ se calculan mediante el algoritmo de boosting y determinan la contribución del clasificador h_t dentro del comité H. El propósito de los pesos es conceder mayor importancia a los clasificadores más precisos. Los métodos para determinar los pesos α_t y la forma en que se modifica el conjunto de entrenamiento S en cada iteración $t = 1, ..., T$ conducen a los distintos tipos de algoritmos de boosting.

Extreme Gradient Boosting (XGBoost) es un método avanzado de ML. Es una variante escalable y mejorada del GB que ofrece mejoras significativas en precisión y rendimiento. Este método utiliza un algoritmo de boosting secuencial para construir un conjunto de modelos de predicción. Debido a que cada modelo se ajusta a los errores residuales del modelo anterior, la precisión del conjunto mejora de manera continua.

Con el uso de una función de pérdida ajustable y la técnica de descenso de gradiente, optimiza de manera eficiente los parámetros del modelo. Una de las ventajas distintivas de XGBoost es su capacidad para manejar características difíciles y datos de alta dimensionalidad. Además, es capaz de manejar valores perdidos, regularización y selección automática de características, lo que reduce el sobreajuste y mejora la generalización del modelo.

Manrique (2022) realizó un estudio de predicción de la demanda de smartphones en Colombia. Durante el desarrollo probó distintos modelos de regresión de ML como los son: la regresión lineal, regresión con máquinas de soporte vectorial, regresión con árbol de decisión, regresión con bosques aleatorios, regresión con redes neuronales y XGBoost Regressor. El mejor resultado que obtuvo fue con el modelo XGBoost sin hiperparametrización, alcanzando los mejores valores en Error cuadrático medio (MSE), Raíz del error cuadrático medio (RMSE), Error absoluto medio (MAE) y error porcentual absoluto medio (MAPE). El modelo de Redes Neuronales con Hiperparametrización obtuvo resultados similares, sin embargo, es más complejo y más costoso en tiempo y recursos.

1.4.4 Paradoja de Simpson

Por otro lado, para finalizar este capítulo se aborda la paradoja de Simpson, que resulta relevante para la comprensión de las decisiones que se tomen acerca de las variables asociadas a la variable de interés durante la ejecución del proyecto.

La paradoja de Simpson es un fenómeno que se presenta en el estudio de la correlación y revela que en ciertos escenarios se observa un cambio en la relación o asociación entre dos variables, ya sea cualitativas o cuantitativas, cuando se toma en cuenta el efecto de una tercera

variable. Esto sucede cuando se examina una variable dependiente en relación con otras variables independientes en un estudio o experimento. Incluso en ocasiones, una de las variables (aquella que altera el tipo o la intensidad de la correlación entre las demás) puede ser una variable desconocida o no controlada, lo que lleva al investigador a no ser consciente de este efecto y a obtener conclusiones erróneas en su estudio. (Contreras, 2012)

En resumen, la paradoja de Simpson invita a reflexionar sobre cómo ciertas relaciones pueden invertirse o cambiar al considerar subgrupos de datos, lo que puede tener un impacto significativo en las conclusiones y decisiones que se adopten en el desarrollo de la investigación.

CAPÍTULO 2

2. METODOLOGÍA

2.1 Tratamiento de los datos

Para este proyecto se utilizó la base de datos `data_venta.xlsx`, que consta de $104\,657$ filas y 34 columnas. Esta base de datos se obtuvo mediante la concatenación de dos bases de datos diferentes otorgadas por la empresa donde se desenvuelve el proyecto: base de facturaciones, la cual contiene datos de ventas y devoluciones desde enero de 2021 hasta mayo de 2023, y base de actividades promocionales. A continuación, se describen las variables más relevantes de la base de datos mencionada:

- **Fecha:** Fecha en la que se realizó la venta.
- **Material:** Código SKU[1] del producto que se solicita.
- **ubnetofac:** Cantidad de producto que se vende. Unidad de medida: bulto.
- **Negocio:** Indica el negocio al que corresponde el material. Hay dos negocios: Consumer Tissue (papel higiénico, servilletas, toallas de papel y pañuelos faciales) y Personal Care (toallas femeninas, protectores diarios, toallas húmedas y pañales).
- **Cod madre:** Código SKU de la última modificación que se realizó a un mismo producto; es

[1] SKU: conjunto de números y letras, empleado para identificar un producto en una empresa (Muñoz, 2022).

decir, corresponde al código SKU del producto que actualmente se está comercializando.

- **Sectores:** Indica el sector al que corresponde el material. Los sectores hacen referencia a la categoría del producto: papel higiénico, servilletas, toallas de papel, etc.
- **Segmento:** Indica el tipo de producto al que corresponde el material. Cada sector se subdivide en algunos segmentos o tipos. Por ejemplo, el papel higiénico se subdivide en: triple hoja, doble hoja, noble y marca propia.
- **Status madre:** Determina si el producto está activo, descontinuado o en transición. Activo: indica que el producto actualmente se está comercializando en el mercado, Transición: indica que se tiene previsto algún tipo de modificación en el producto, Descontinuado: indica que el producto ya no se encuentra en el mercado.
- **Origen material:** Indica si el producto es de producción local, es importado de otras filiales de la empresa o de proveedores terceros.
- **Zona:** Indica a qué zona de venta pertenece el cliente que solicita el producto.
- **Promoción:** Indica si en el momento que se realizó la compra, el producto estaba en alguna promoción para el cliente que lo solicitó.
- **Descuento:** Indica el descuento que tenía el producto en el momento que el cliente realizó el pedido.
- **Prop clientes:** En ciertos clientes que manejan algunos formatos o filiales de supermercados; a menudo, las promociones que se aprueban no se aplican para todos

estos formatos. Por ende, este campo indica si la promoción activa se está aplicando para todas las filiales que maneja este cliente o solo a una porción de estos.

- **Cambio producto:** Indica si el producto ha tenido alguna modificación. El valor es 0 si el producto (código material) ha existido de manera individual sin modificaciones previas, esta columna empieza a registrar 1 una vez que se introduce al mercado alguna modificación del producto (con un nuevo código SKU), y ahora solo se está vendiendo en el mercado para desalojar el stock que se tiene del producto anterior.

- **Proceso codificación:** Indica si el producto ha tenido alguna modificación que implique un proceso de codificación en las cadenas de autoservicios. El valor es 0 si el producto ha existido sin modificaciones relevantes y registra 1 cuando el producto si presenta modificaciones importantes, por ejemplo: cambios de EAN (código de barras), metraje, conteos.

- **85% total:** Indica si la factura pertenece a un cliente que está entre aquellos que generan el 85% de las ventas.

- **Metraje:** Indica el metraje del producto (Material) que se vende. Solo los segmentos Manzanilla y Triple Hoja tienen valor en esta variable, el resto marca 0.

- **Metraje madre:** Indica el metraje del producto madre (Cod madre) que se vende. Del mismo modo solo los segmentos Manzanilla y Triple Hoja tienen valor en esta variable, el resto marca 0.

La variable *Prop clientes* fue creada a partir de la información obtenida de la base de actividades promocionales. El objetivo de esta variable fue analizar si la manera en que el cliente

gestiona y promueve una actividad comercial, influye sobre la cantidad de producto vendido durante el periodo de duración de dicha actividad.

Asimismo, las variables *Cambio producto* y *Proceso codificación* se crearon con el fin de examinar las ventas ejecutadas netamente para liquidar el inventario de los productos que han experimentado algún tipo de modificación relevante durante su existencia, y cómo estos cambios inciden sobre la demanda del cliente.

Este estudio está enfocado en capturar la demanda real de los clientes que generan el 85% de las ventas de la empresa donde se desarrolla el proyecto. Por ende, solo se consideran aquellas facturas que marca 1 en el campo *85% total*. Adoptando las consideraciones mencionadas, la variable de interés es *ubnetofac*.

Este proyecto abarca las zonas de ventas más relevantes de la compañía, los supermercados de la Costa y de la Sierra. Dado que el estudio se centra en predecir la demanda de productos locales, se seleccionaron únicamente los pedidos de productos en los cuales la variable *Origen material* indica que son de producción local. A continuación, se examinó el estatus actual del producto en el mercado, a través de la variable *Status madre*, y se eligió exclusivamente aquellos que no han sido descontinuados.

Por otra parte, la compañía opera en zonas de venta distintas, donde el comportamiento del mercado y las actividades promocionales varían significativamente entre sí. Por esto, se consideró apropiado estimar la demanda en función de cada zona de venta. Inicialmente se buscaba predecir la demanda empleando el mínimo nivel de granularidad en los productos; sin embargo, se debían realizar 179 modelos si se requiere estimar la demanda por zona de venta y por material. Por el contrario, al tomar el máximo nivel de granularidad disponible para los

productos (sector), la cantidad de modelos requeridos se redujo a un total de 8. De este modo, se optó por un nivel intermedio de granularidad; es decir, se ejecutaron las proyecciones por zona de venta y segmento. Se realizó un primer análisis modelando los segmentos papel higiénico Doble hoja y Triple hoja para ambas zonas de venta, se determinó que necesitaban un nivel de granularidad inferior a segmento. De esta forma, para el papel higiénico Triple hoja en Costa y Doble hoja en ambas zonas, se ejecutaron proyecciones por zona de venta y metraje madre. Se consideró emplear el mismo nivel de granularidad para el papel higiénico Triple hoja en Sierra, sin embargo, las proyecciones no mejoraron. En particular, existía un problema de sesgo por variable omitida en un metraje madre. Finalmente, resultaron un total de 21 modelos.

De esta forma, para el entrenamiento de los modelos predictivos los datos se sintetizaron a las ventas mensuales de los clientes que generan el 85% de ingresos, por cada segmento o metraje madre de productos por zona comercial desde enero de 2021 hasta mayo de 2023. Así, unificada la venta por mes y realizado el tratamiento pertinente de los datos, se obtuvo una base con $1\,495$ filas y 15 columnas. Las variables de la base de datos final son: *Fecha*, *Zona*, *Negocio*, *Sectores*, *Segmento*, *Metraje madre*, *ubnetofac*, *Promocion*, *Descuento*, *Descuento std*, *Prop clientes*, *Material nunique*, *Metraje nunique*, *Cambio de producto* y *Proceso codificación*.

2.2 Análisis de las variables para el entrenamiento de los modelos

Después de la depuración de los datos, se llevó a cabo un análisis de la correlación entre la variable de interés, *ubnetofac*, y las demás variables. El análisis se ejecutó por zona de venta y sector. Se observó que algunas variables se relacionan fuertemente con la variable de interés dependiendo de la zona de venta y el sector analizados.

Figura 2.1.
Correlación entre ubnetofac y las demás variables en la Costa

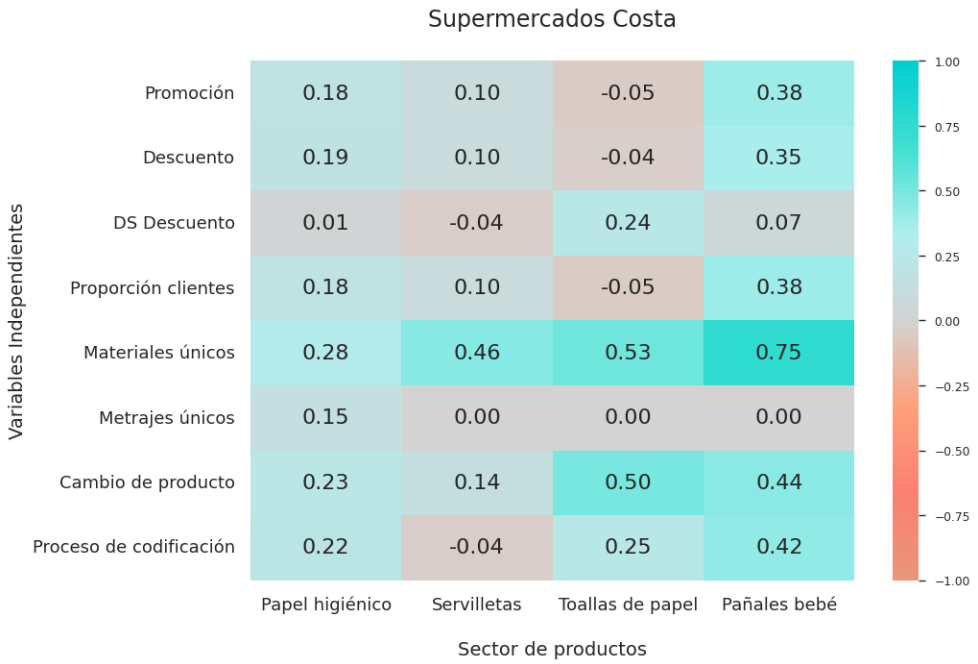
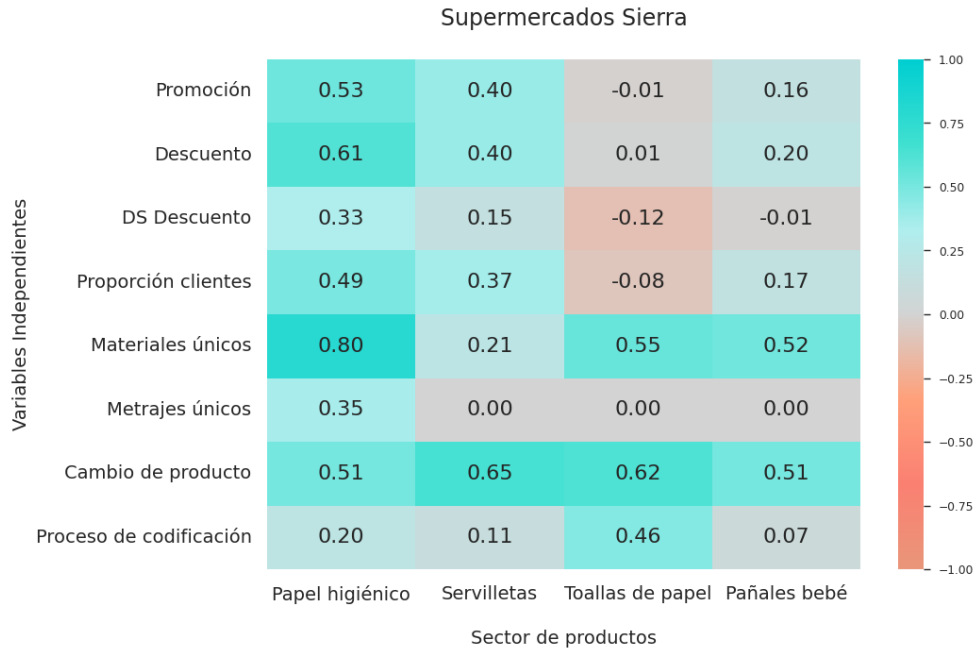


Figura 2.2.
Correlación entre ubnetofac y las demás variables en la Sierra



En ambas zonas comerciales se visualizó que en algunos casos, independientemente del sector de productos, la variable objetivo no presenta una correlación lineal fuerte con ninguna otra variable estadística. Luego, partiendo de los resultados obtenidos de este análisis, se determinó las variables que serían incluidas en el entrenamiento de cada uno de los 21 modelos.

La importancia de llevar a cabo este análisis en cada zona de venta y por sector de producto, radica en que al estudiar la correlación entre variables a nivel general se observó que en ambas zonas comerciales no existía una correlación significativa. Por lo tanto, la segmentación por zona de venta y sector de producto permitió obtener una visión más precisa y relevante de las variables que influyen en la demanda del mercado. Omitir este análisis podría haber conducido a resultados y conclusiones erróneas al momento de seleccionar las variables para el entrenamiento de los modelos predictivos.

2.3 Implementación de los modelos predictivos

Para la implementación del código se utilizó Python como lenguaje de programación, empleando la interfaz web Google Colab. Se empleó la librería **pycaret** para determinar los modelos que más se ajustan a la demanda, entre estos destacaron ARIMA y el método XG-Boost. Basándonos en las proyecciones de **pycaret**, junto con la literatura y estudios realizados en otras industrias, se concluyó que los métodos predictivos más prometedores para estimar la demanda se resumen en: modelos ARIMA y el método XG-Boost. Por ende, estos son los modelos que se usaron en el desarrollo del presente proyecto.

El entrenamiento de los modelos ARIMA se realizó mediante la librería **pmdarima**, y para los modelos XGBoost la librería **xgboost**. Posteriormente, se asignó la demanda a cada SKU con base en la participación porcentual de su código madre en los tres últimos meses. Finalmente,

para visualizar los resultados proporcionados por cada modelo predictivo se utilizaron las librerías **matplotlib** y **seaborn**.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

El presente capítulo se divide en tres secciones, en las cuales se muestra la comparación entre los modelos escogidos para la estimación de la demanda, los resultados obtenidos por zona de venta y los resultados generales. En la primera sección se detalla cuál es el modelo que realiza las mejores estimaciones por segmento y zona de venta, en donde se empleó el indicador **MAPE**. En la segunda sección se evalúan los resultados de los modelos por zona de venta, utilizando el indicador **WMAPE**. Por último, en la tercera sección se muestran los resultados generales y se discuten los hallazgos obtenidos.

3.1 Resultados modelo ARIMA y XGBoost

Para definir el modelo que mejor se ajusta a la serie temporal de cada segmento de producto y zona de venta, se procedió con el entrenamiento de los modelos ARIMA y XGBoost empleando las variables predictoras descritas anteriormente. Los datos disponibles desde enero de 2021 hasta febrero de 2023 se destinaron para el ajuste de los modelos, y los datos de marzo a mayo de 2023 se utilizaron para el testeo. Una vez ajustados los modelos, se obtuvieron las predicciones por segmento de producto y zona de venta. Luego, se realizó el cálculo del MAPE y se precisó un modelo ganador para cada uno.

Tabla 3.1.
Resultados MAPE de los modelos ganadores

Producto		Supermercados Sierra		Supermercados Costa	
		Modelo ganador	**MAPE**	**Modelo ganador**	**MAPE**
Papel higiénico	Doble hoja	XGBoost	20%	ARIMA	25%
	Triple hoja	XGBoost	12%	ARIMA+XGBoost	17%
Servilletas	Mesa	XGBoost	1%	ARIMA	30%
	Coctel	XGBoost	25%	ARIMA	20%
	Económicas	XGBoost	23%	ARIMA	5%
Toallas de papel	Premium	XGBoost	22%	XGBoost	12%
	Económicas	XGBoost	23%	ARIMA	16%
Pañales de bebé	Premium	XGBoost	17%	ARIMA	22%
	Ultra	XGBoost	21%	XGBoost	6%

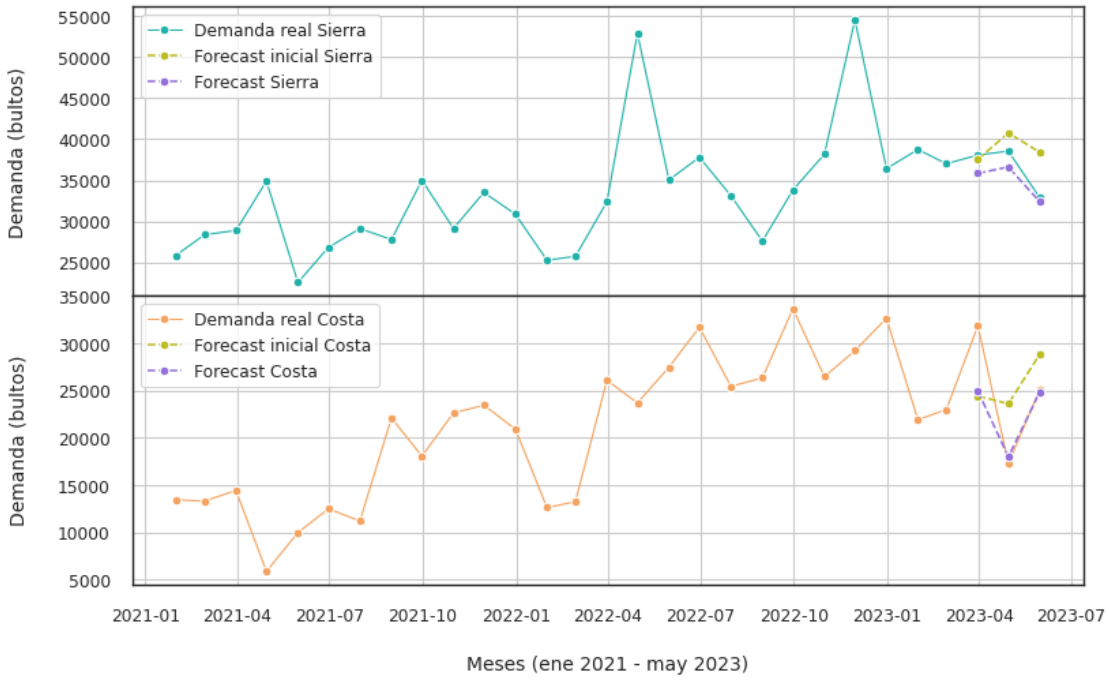
Se observó que el modelo XGBoost provee un mayor poder predictivo, puesto que es el modelo que mejor se ajusta en el 61% de los escenarios. Asimismo, se pudo afirmar que este modelo siempre resulta ser el más eficaz para la estimación de la demanda de la zona Supermercados Sierra.

Por otro lado, se visualizó que en la zona Supermercados Costa, el modelo ARIMA predomina como el mejor estimador de la demanda en cada segmento de producto. Se determinó que esto se debe a que las variables predictoras escogidas no influyen de manera significativa en la demanda de los clientes de esta zona de venta, por lo que el modelo XGBoost no logra capturar adecuadamente las variaciones del mercado. Esto se pudo corroborar a partir de los resultados de la figura 2.1, donde se aprecia que los sectores del negocio *Consumer Tissue* no poseen correlaciones lineales fuertes con la variable objetivo. A diferencia de la zona

Supermercados Sierra, donde cada sector de *Consumer Tissue* posee al menos una variable con una correlación mayor a 0.6.

En particular, dado que los Triple hoja a su vez se subdividen en grupos de diferentes metrajes, se resolvió que los datos de ciertos metrajes del segmento Triple hoja en la zona Supermercados Costa se ajustan mejor al modelo XGBoost, y otro grupo obtiene proyecciones más precisas mediante el modelo ARIMA. De este modo, para esta zona de venta en el segmento Triple hoja de Papel higiénico la mejor proyección se obtuvo realizando una combinación de los modelos ARIMA y XGBoost.

Figura 3.1.
Demanda real vs. Forecast inicial vs Forecast Modelado por zona de venta



La Figura 3.1 ilustra una comparativa entre la demanda real, el pronóstico inicial y el pronóstico modelado, segmentado por zonas de venta.

3.2 Resultados por zona de venta

Para el análisis de los resultados, la demanda total de la zona, negocio y sector, se obtuvieron unificando las proyecciones realizadas por segmento. Para la evaluación de resultados de los modelos se utilizó el indicador WMAPE y los objetivos planteados por la empresa para cada negocio:

- **Consumer Tissue:** Un WMAPE de no más del 23%.
- **Personal Care:** Un WMAPE de no más del 32%.

Se resaltaron en color verde los WMAPE que logran los objetivos establecidos, y en amarillo aquellos que no alcanzaron el *target* pero sí son menores que WMAPE obtenido bajo el método tradicional.

Sectores

Tabla 3.2.
Resultados WMAPE por sector de Supermercados Sierra

Sector	Marzo		Abril		Mayo	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Papel higiénico	36%	41%	48%	57%	90%	**44%**
Servilletas	32%	42%	26%	28%	16%	43%
Toallas de papel	47%	**11%**	47%	**22%**	8%	**9%**
Pañales de bebé	64%	**40%**	70%	**47%**	36%	**33%**

En el sector Toallas de papel de Supermercados Sierra se alcanzó el objetivo todos los meses.

Tabla 3.3.
Resultados WMAPE por sector de Supermercados Costa

Sector	Marzo		Abril		Mayo	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Papel higiénico	35%	**31%**	126%	**80%**	33%	**26%**
Servilletas	34%	**17%**	25%	**21%**	20%	**14%**
Toallas de papel	30%	**16%**	44%	**6%**	19%	**18%**
Pañales de bebé	12%	**9%**	18%	**16%**	13%	**13%**

En los sectores de servilletas, toallas de papel y pañales de bebé de Supermercados Costa, se cumplió la meta en cada mes. El objetivo de WMAPE se alcanzó en el 50% de los sectores estudiados. En el 29% de los casos, se superó el WMAPE inicial, mientras que en tan solo el 21% restante no se alcanzó ninguna de las metas establecidas.

Negocios

Tabla 3.4.
Resultados WMAPE por negocio de Supermercados Sierra

Negocio	Marzo		Abril		Mayo	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Consumer Tissue	36%	40%	46%	53%	78%	**43%**
Personal Care	64%	**40%**	70%	**48%**	36%	**33%**

Aunque ningún negocio de Supermercados Sierra alcanzó el objetivo establecido, el negocio Personal Care superó el WMAPE Inicial durante todos los meses.

Tabla 3.5.
Resultados WMAPE por negocio de Supermercados Costa

Negocio	Marzo		Abril		Mayo	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Consumer Tissue	34%	**27%**	100%	**57%**	31%	**24%**
Personal Care	12%	**9%**	18%	**16%**	13%	**13%**

En el negocio Personal Care de Supermercados Costa se alcanzó el objetivo todos los meses. Se logró el objetivo de WMAPE en el 25% de los negocios de estudio. Se superó el WMAPE Inicial un 58% de los casos y tan solo en el 17% restante no se logró ninguna de las dos.

3.3 Resultados generales

Figura 3.2.
Demanda real vs Forecast inicial vs Forecast Modelado para todo el canal de supermercados

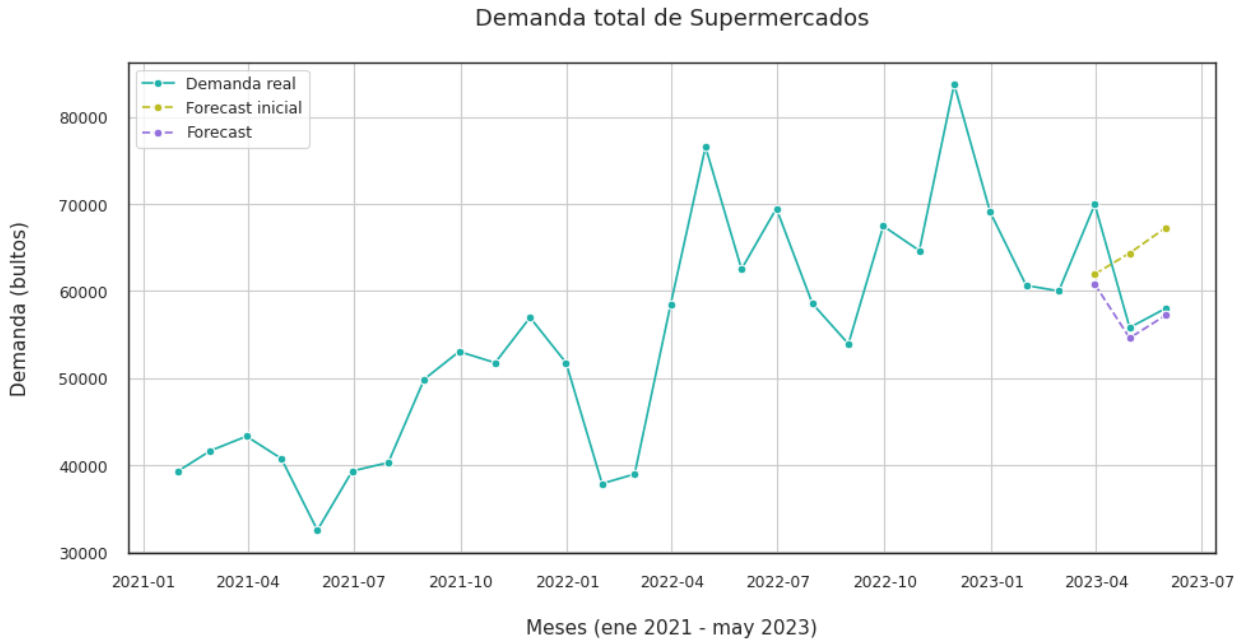


Tabla 3.6.
Resultados MAPE por negocio

Negocio	Sierra		Costa		Total Supermercados	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Consumer Tissue	24%	16%	9%	7%	8%	5%
Personal Care	9%	4%	36%	14%	17%	5%

Tabla 3.7.
Resultados WMAPE por sector

Sector	Marzo		Abril		Mayo	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Papel higiénico	33%	**28%**	56%	**53%**	51%	**34%**
Servilletas	16%	24%	20%	**16%**	12%	**21%**
Toallas de papel	37%	**8%**	38%	**15%**	6%	**7%**
Pañales de bebé	11%	**11%**	19%	**15%**	12%	**13%**

Tabla 3.8.
Resultados WMAPE por negocio

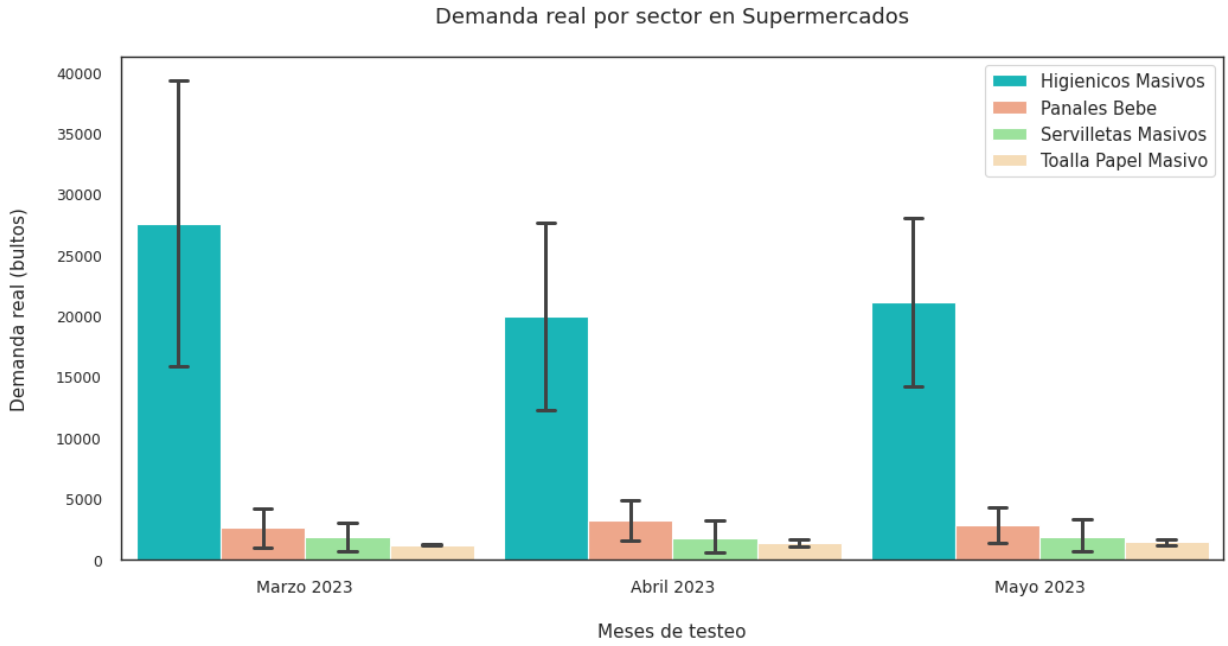
Negocio	Marzo		Abril		Mayo	
	Inicial	Modelado	Inicial	Modelado	Inicial	Modelado
Consumer Tissue	32%	**27%**	51%	**46%**	44%	**31%**
Personal Care	11%	**11%**	19%	**15%**	12%	**13%**

Se observó en la tabla 3.6 que los MAPEs obtenidos con los modelos implementados son menores a los iniciales en ambos negocios; es decir, las proyecciones realizadas por los modelos de predicción fueron más precisas que las iniciales. Sin embargo, esta mejora no fue evidente en

los WMAPEs, debido a que este indicador es más sensible a los errores de estimación por SKU. De esta manera, el método de distribución de demanda que se empleó afectó a los resultados de WMAPE en ambas zonas, provocando que no se alcancen los objetivos establecidos para el negocio *Consumer Tissue*. Además, con los resultados de la tabla 3.7 se evidenció que la categoría que resultó más afectada corresponde a Papel higiénico, lo que consecuentemente afecta el WMAPE de todo el negocio *Consumer Tissue*, puesto que la mayor parte del volumen de venta corresponde a este sector (véase figura 3.3). Aún con las limitaciones expuestas, el WMAPE del negocio *Consumer Tissue* presentó una mejoría importante respecto al WMAPE inicial.

Por otra parte, los resultados obtenidos en el negocio *Personal Care* sí alcanzaron los objetivos planteados, puesto que en cada mes proyectado los WMAPE estuvieron por debajo del 32%.

Figura 3.3.
Demanda real de los meses de testeo por sector de producto



CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En este proyecto se investigó el comportamiento de las ventas históricas de una empresa de productos de higiene personal y se examinaron ciertas variables que influyen en ellas. Se evaluó cómo los modelos ARIMA y XGBoost describen estas ventas con el propósito de predecir la demanda futura. Por último, se analizaron los resultados con el objetivo de determinar si la integración de estos modelos en el proceso de planificación de la demanda permite optimizar procesos. A continuación, se presentan las conclusiones y recomendaciones más relevantes del proyecto.

Conclusiones

- De acuerdo con la gestión y el análisis realizado sobre los datos históricos de ventas, se concluyó que las variables relacionadas con promociones y modificaciones de productos tienen un impacto significativo en la demanda del mercado, pues afectan su tendencia; principalmente en la zona Supermercados Sierra. Esto resalta la necesidad de emplear métodos cuantitativos para detectar y comprender estas variaciones.

- Los resultados sugieren que los modelos ARIMA y XGBoost se ajustan apropiadamente a los datos históricos de ventas de la industria de higiene personal ecuatoriana. Estos modelos lograron reducir los errores en la predicción total de la demanda para los meses analizados.

- En particular, el modelo XGBoost se ajusta de manera adecuada al 61% de los segmentos estudiados, por lo que se propone como el modelo predictivo idóneo para la predicción de la demanda en la empresa donde se desenvuelve el proyecto.
- Se espera que este trabajo funcione como un incentivo para que otras industrias incorporen el análisis y el modelado de datos en sus procesos de planificación de la demanda. Asimismo, se espera que los modelos propuestos puedan contribuir de manera significativa a este esfuerzo.

Recomendaciones

- Examinar el mercado de Supermercados Costa para identificar variables que puedan explicar de manera más efectiva el comportamiento de las ventas, con el fin de mejorar el rendimiento del modelo XGBoost en esta área de ventas.
- Se sugiere para trabajos futuros explorar más en las variables disponibles y su significancia según el tipo de mercado donde se realizarán las proyecciones. Para de esta forma, mejorar el rendimiento del modelo XGBoost.
- Estudiar métodos más eficientes para la distribución de la demanda total a niveles de granularidad inferiores a los propiamente proyectados, con el fin de mitigar el impacto negativo en WMAPE y alcance de objetivos.

BIBLIOGRAFÍA

Avilés, W. (2018). Análisis del comportamiento de compra de papel higiénico en canal moderno en la ciudad de Guayaquil. Master's thesis, Departamento de Marketing. Universidad Católica de Santigo de Guayaquil.

Caba, N., Chamorro, O., and Fontalvo, T. (2011). *Gestión de la producción y operaciones*. Corporación para la Gestión del Conocimiento Asesores del 2000, Barranquilla, Colombia.

Chase, R. B., Jacobs, F. R., and Alquilano, N. J. (2009). *Administración de Operaciones: Producción y cadena de suministros*. Mc Graw Hill, 12 edition.

Contreras, J. M., B. C. C. G. . G. M. M. (2012). La paradoja de simpson. *Suma*, 71:19–26.

De La Torre, M. (2022). Crecimiento del consumo masivo en Ecuador. https://www.kantar.com/latin-america/inspiracion/consumidor/2022-ec-hasta-cuando-el-consumo-masivo-crecera-en-el-ecuador.

González, M. P. (2009). *Análisis de series temporales: Modelos ARIMA*. Departamento de Economía Aplicada III. Facultad de Ciencias Económicas y Empresariales. Universidad del País Vasco, España.

Hayes, A. (2022). Autoregressive integrated moving average (arima) prediction model. *Investopedia*. https://www.investopedia.com/terms/a/

autoregressive-integrated-moving-average-arima.asp#:~:text=The%20ARIMA%20model%20is%20used,predict%20an%20asset's%20future%20performance.

Logility Voyager Solutions (2010). Successful sales and operations planning in 5 steps. *Logility*. https://www.supplychainbrain.com/ext/resources/secure_download/KellysFiles/WhitePapersAndBenchMarkReports/Logility/S-OP+5+Steps+to+Success+Logility.pdf.

Manrique, J. □. (2022). *Predicción de la demanda de Smartphone de introducción al mercado Colombiano mediante modelos de Machine Learning*. Fundación Universitaria Konrad Lorenz, Colombia.

Martínez Celda, B. (2021). *Gradient boosting (Potenciación del gradiente) en aprendizaje estadístico*. Facultad de Ciencias. Universidad de Valladolid, España.

Mauricio, J. A. (2007). *Análisis de Series Temporales*. Universidad Complutense de Madrid, España.

Meetlogistics (2020). Planificación de la demanda: Fundamentos. https://meetlogistics.com/demand-planning/planificacion-de-la-demanda-fundamentos/.

Muñoz, A. (2022). ¿qué es un sku? todo lo que debes saber. https://blog.saleslayer.com/es/que-es-un-sku-todo-lo-que-debes-saber#:~:text=Un%20SKU%20es%20un%20conjunto,usamos%20como%20Referencia%20de%20Almac%C3%A9n.

Palacios, C. A. (2020). *Análisis y predicción de las tendencias de venta en el mercado usando árboles de regresión*. Colegio Ciencias e Ingenierías. Universidad San Francisco de Quito, Ecuador.

Palomino Mezones, M. D. (2021). *Árboles de regresión para el análisis de rating de avisos publicitarios del sector automotriz.* Escuela Profesional de Estadística. Facultad de Ciencias Matemáticas. Universidad Nacional Mayor de San Marcos, Perú.

Parra, J. (2018). Planeación de la demanda en una empresa de venta directa. Master's thesis, Facultad de Ingeniería. Universidad Militar Nueva Granada.

Qurius (2010). Planificación de la demanda. *Microsoft Dynamics*. https://www.choisirmonerp.com/documents/pdf/ES/ES/Planificacion%20Demanda.pdf.

Sepúlveda, J. F. D., . M. (2013). *Comparación entre árboles de regresión CART y regresión lineal*. Comunicaciones en Estadística. Universidad Santo Tomás, Colombia.

The 365 Team (2023). Introduction to decision trees: Why should you use them? https://365datascience.com/tutorials/machine-learning-tutorials/decision-trees/.

Traders Studio (2021). Canibalización del mercadocanibalización del mercado. https://traders.studio/canibalizacion-del-mercado/.

APÉNDICES

APÉNDICE A

A.1 Proyecciones iniciales obtenidas por segmento y zona de venta

A continuación, se presentan los primeros resultados obtenidos con el entrenamiento de los modelos ARIMA y XGBoost para cada segmento y zona de venta estudiada.

A.1.1 Supermercados Sierra

Figura A.1.
Papel higiénico doble hoja en Supermercados Sierra

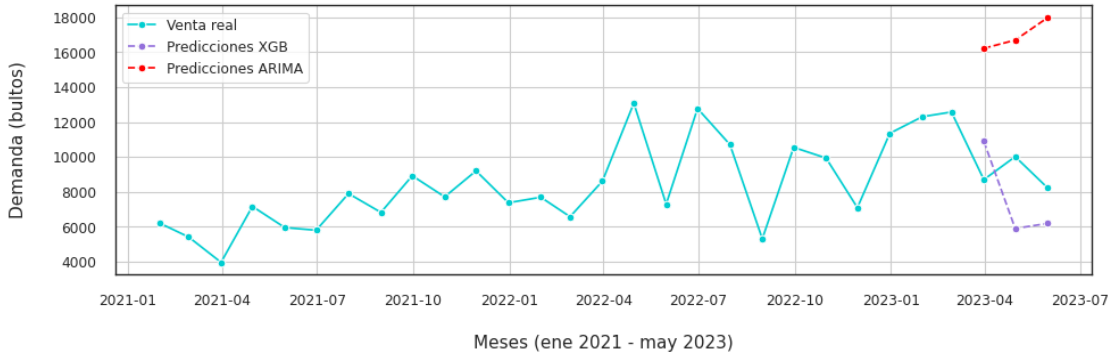


Figura A.2.
Papel higiénico triple hoja en Supermercados Sierra

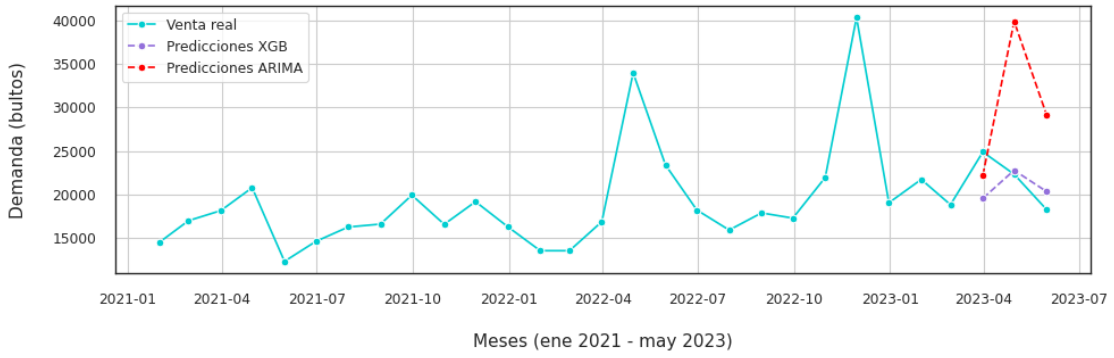


Figura A.3.
Servilletas Mesa en Supermercados Sierra

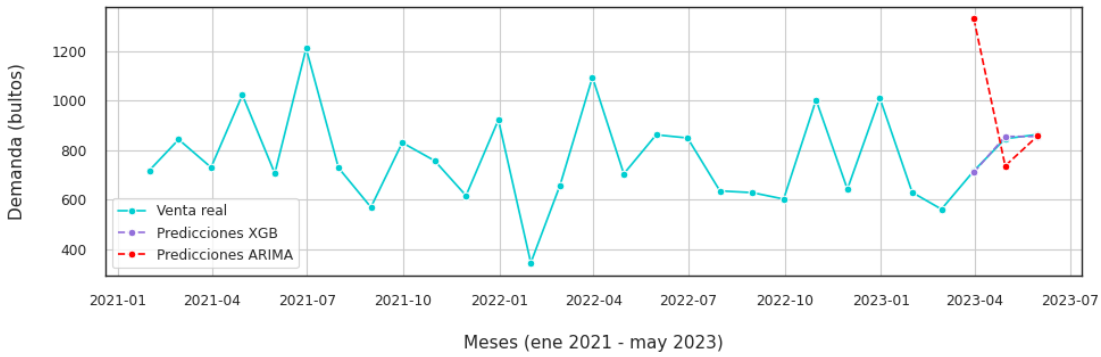


Figura A.4.
Servilletas Coctel en Supermercados Sierra

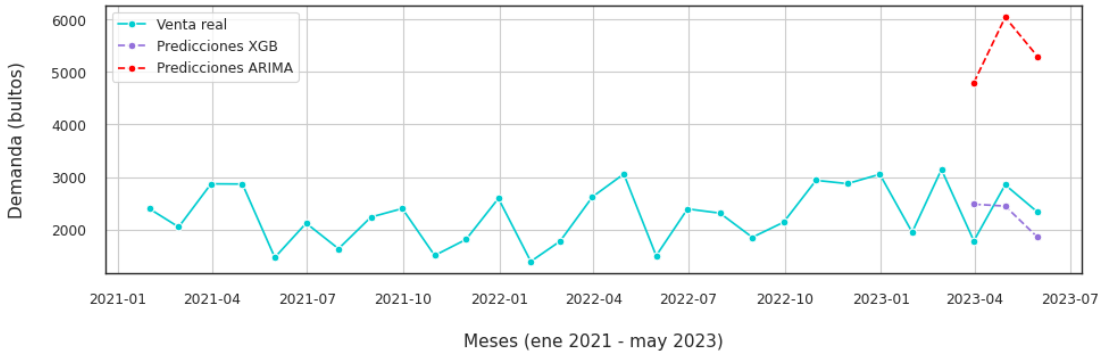


Figura A.5.
Servilletas Económicas en Supermercados Sierra

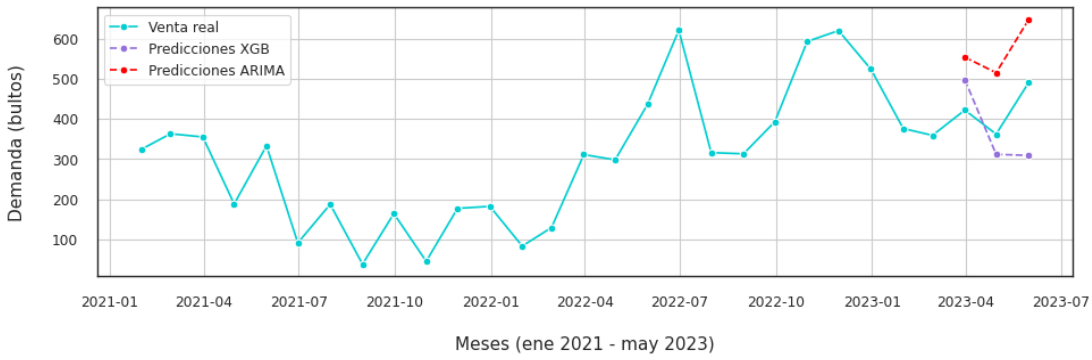


Figura A.6.

Toallas de papel Premium en Supermercados Sierra

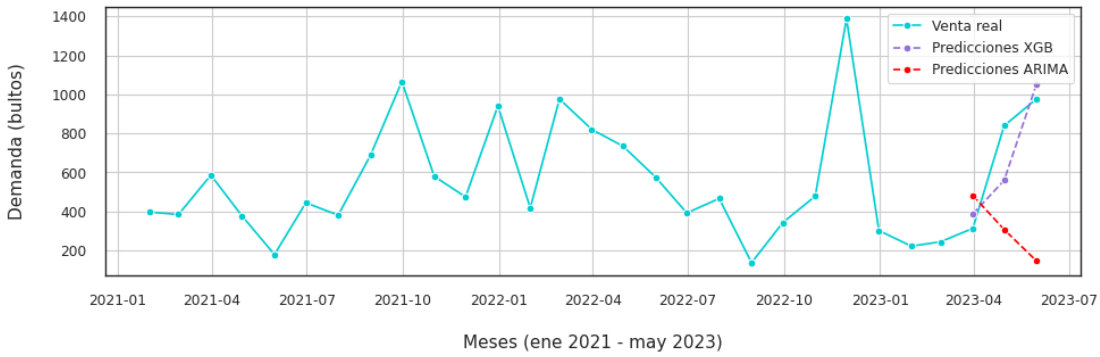


Figura A.7.

Toallas de papel Económicas en Supermercados Sierra

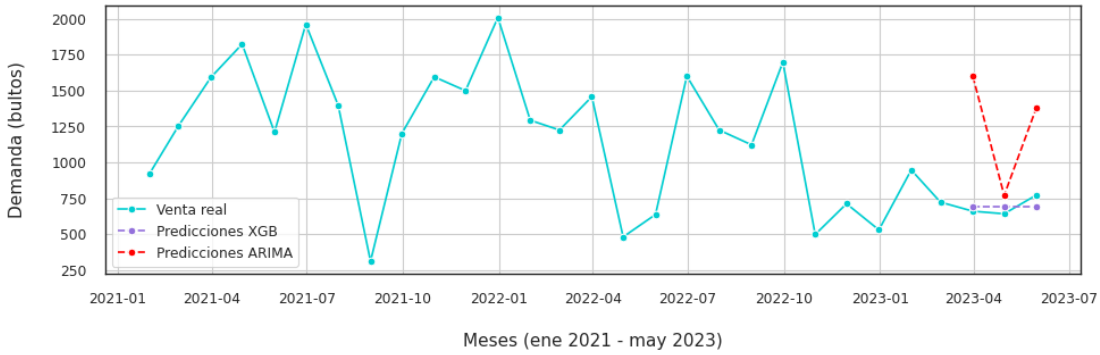


Figura A.8.

Pañales de bebé Premium en Supermercados Sierra

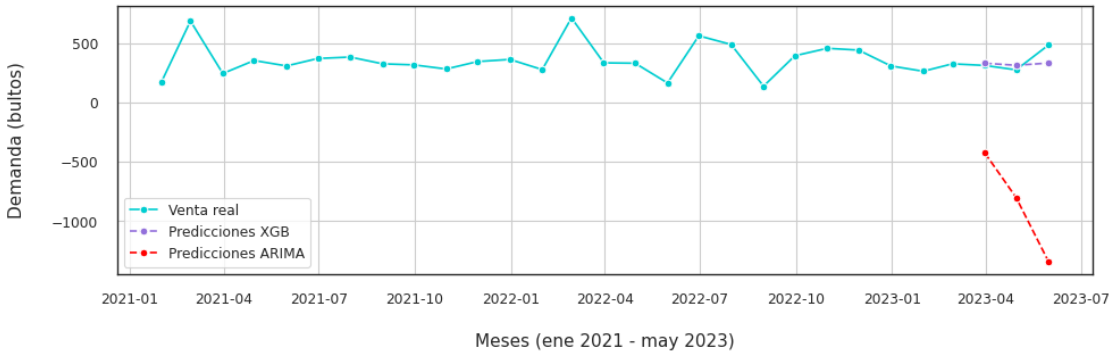
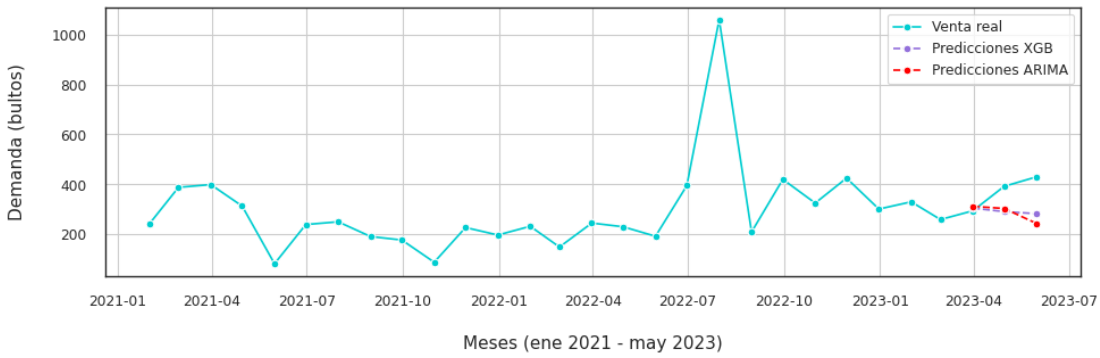


Figura A.9.
Pañales de bebé Ultra en Supermercados Sierra



A.1.2 Supermercados Costa

Figura A.10.
Papel higiénico doble hoja en Supermercados Costa

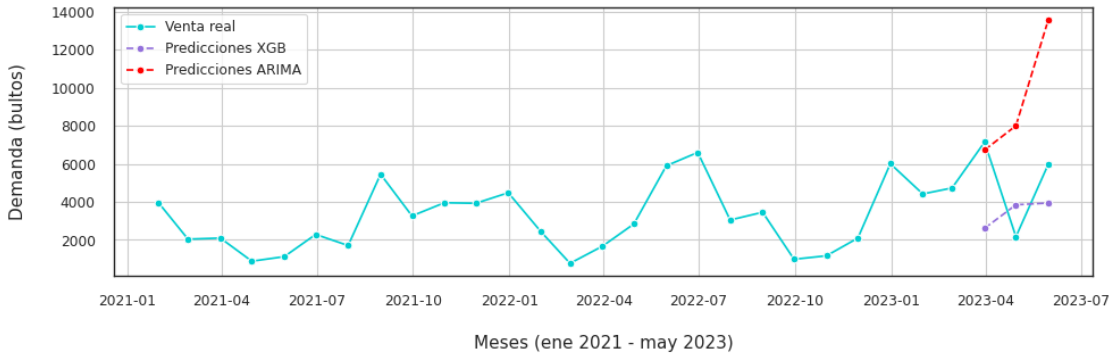


Figura A.11.
Papel higiénico triple hoja en Supermercados Costa

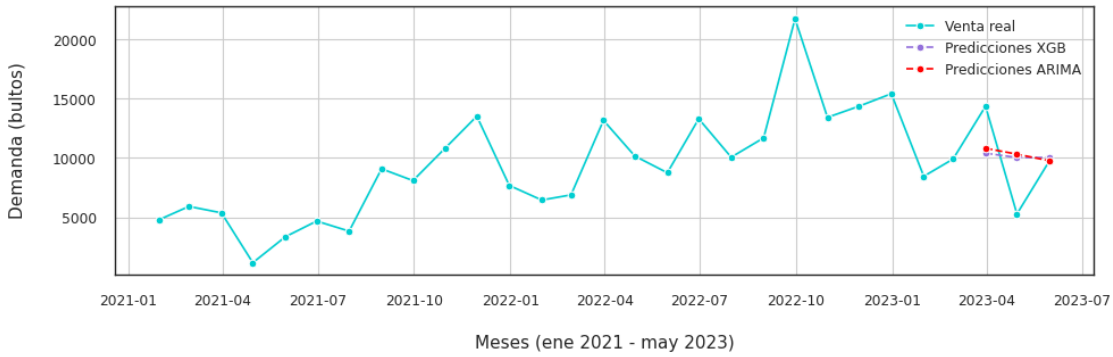


Figura A.12.
Servilletas Coctel en Supermercados Costa

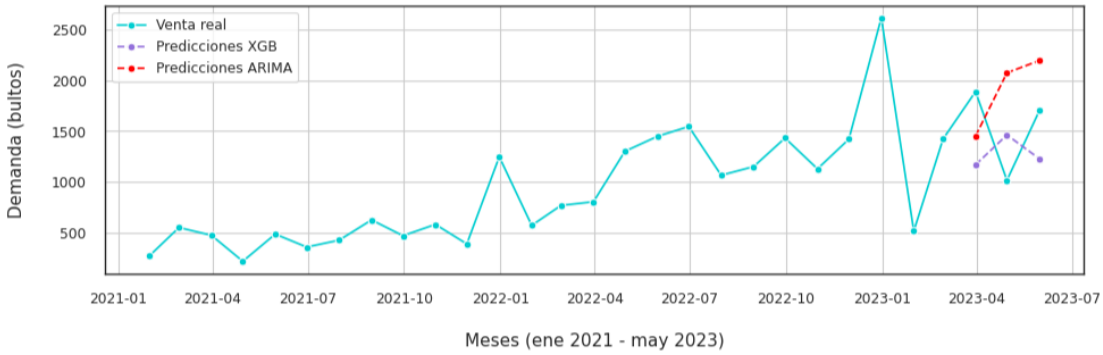


Figura A.13.
Servilletas Mesa en Supermercados Costa

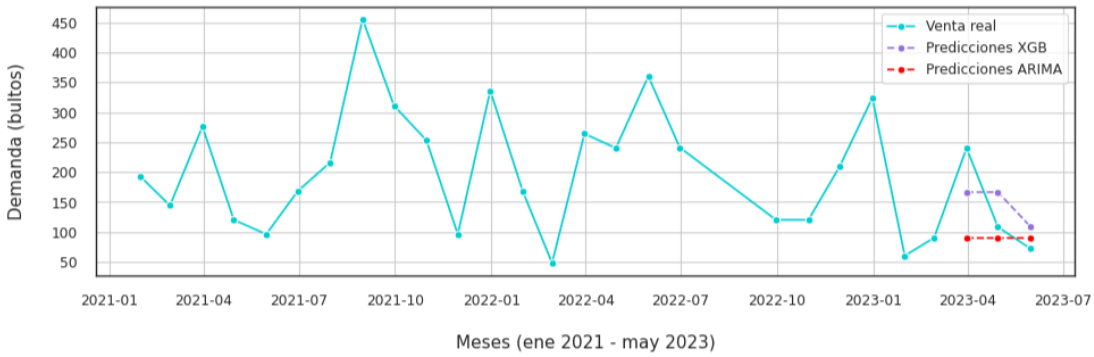


Figura A.14.
Servilletas Económicas en Supermercados Costa

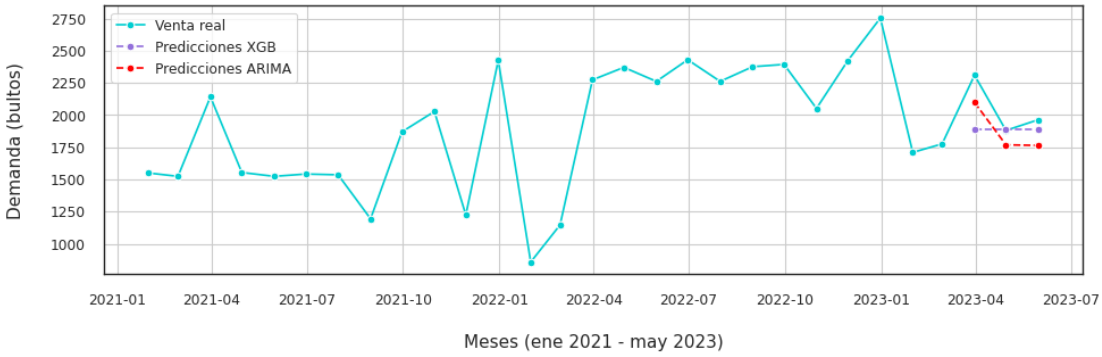


Figura A.15.
Toallas de papel Premium en Supermercados Costa

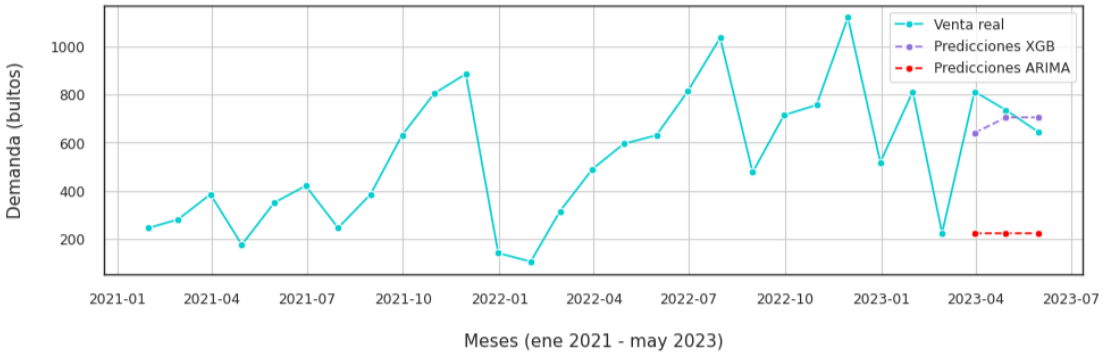


Figura A.16.
Toallas de papel Económicas en Supermercados Costa

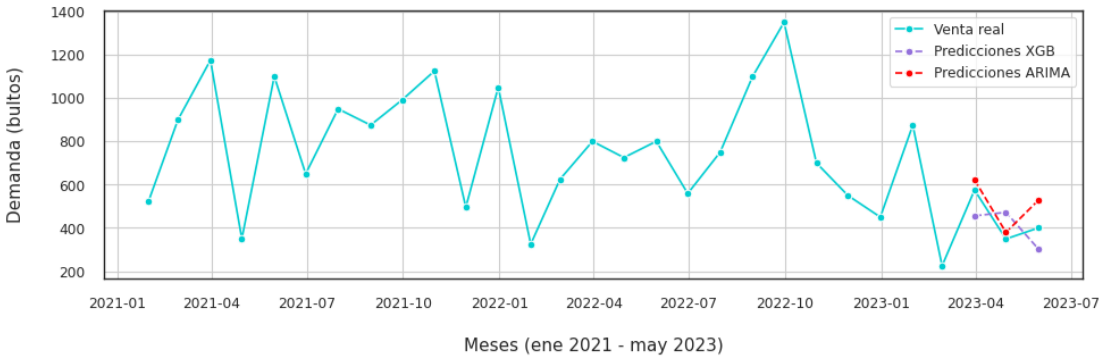


Figura A.17.
Pañales de bebé Premium en Supermercados Costa

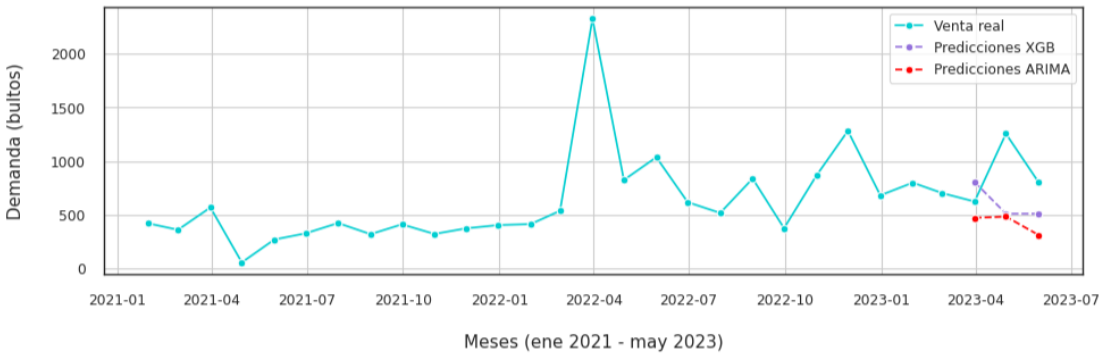
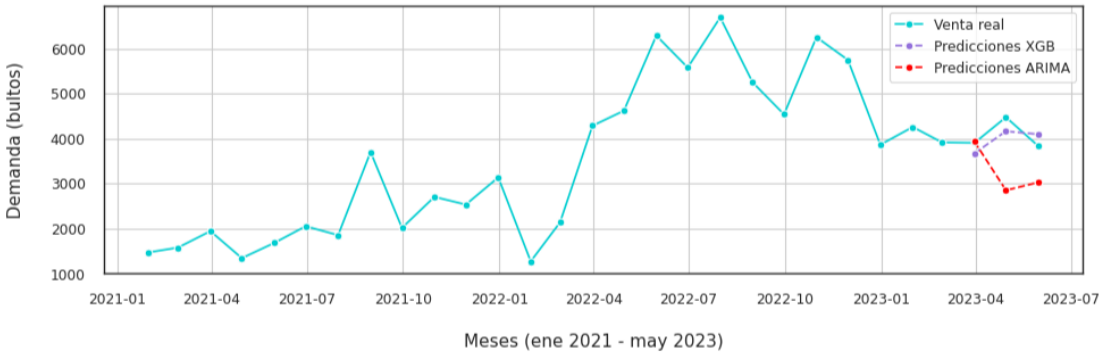


Figura A.18.
Pañales de bebé Ultra en Supermercados Costa



A.2 Proyecciones mejoradas tras el ajuste de parámetros en los modelos predictivos

Tras los resultados presentados, se estudió más de cerca el comportamiento de los segmentos que a simple vista no se logran ajustar a ninguno de los dos modelos propuestos. De este modo, se obtuvieron los parámetros adecuados para que al menos uno de los dos modelos se ajuste apropiadamente a cada serie temporal.

A.2.1 Supermercados Sierra

Figura A.19.
Papel higiénico doble hoja 15m en Supermercados Sierra

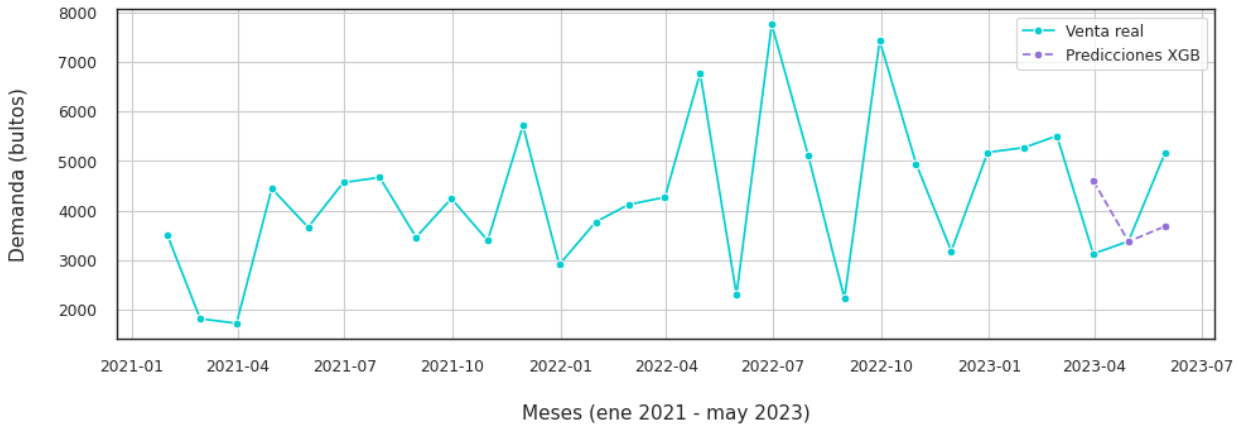


Figura A.20.
Papel higiénico doble hoja 32m en Supermercados Sierra

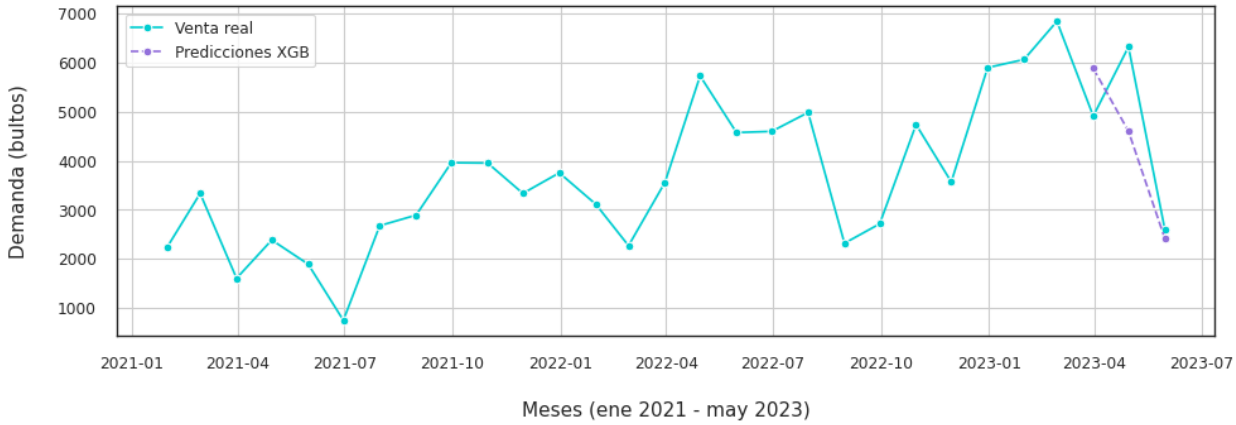
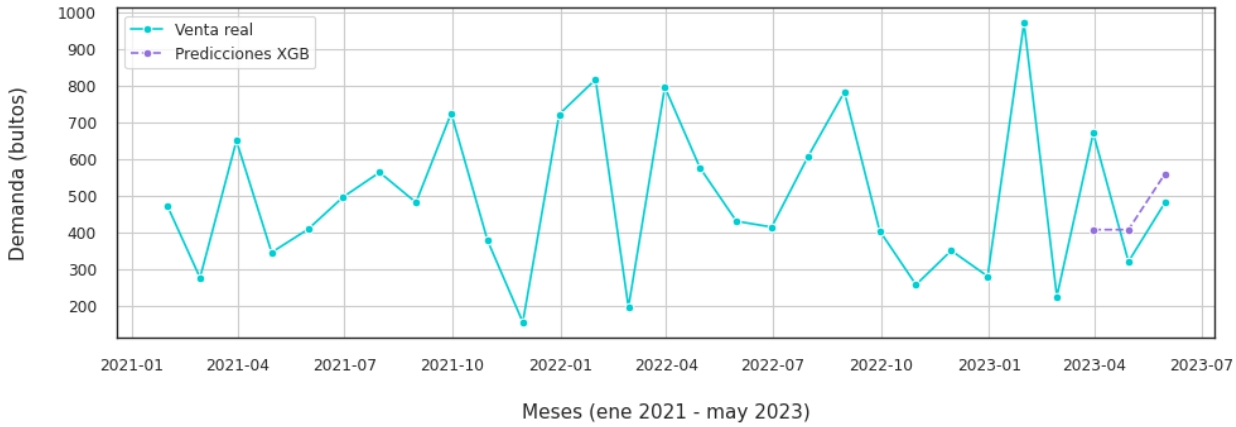


Figura A.21.
Papel higiénico doble hoja 42m en Supermercados Sierra



A.2.2 Supermercados Costa

Figura A.22.
Papel higiénico doble hoja de 32m en Supermercados Costa

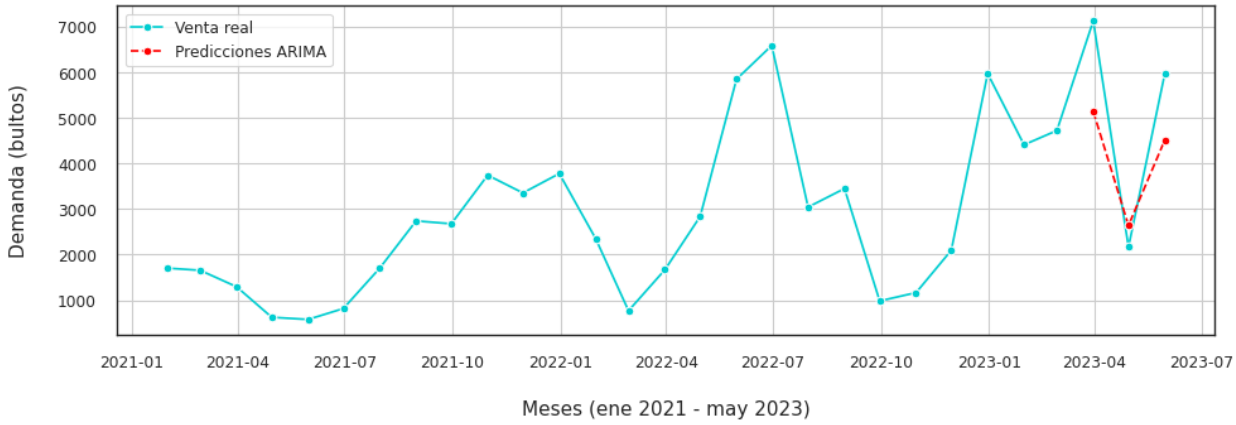


Figura A.23.
Papel higiénico triple hoja de 32m en Supermercados Costa

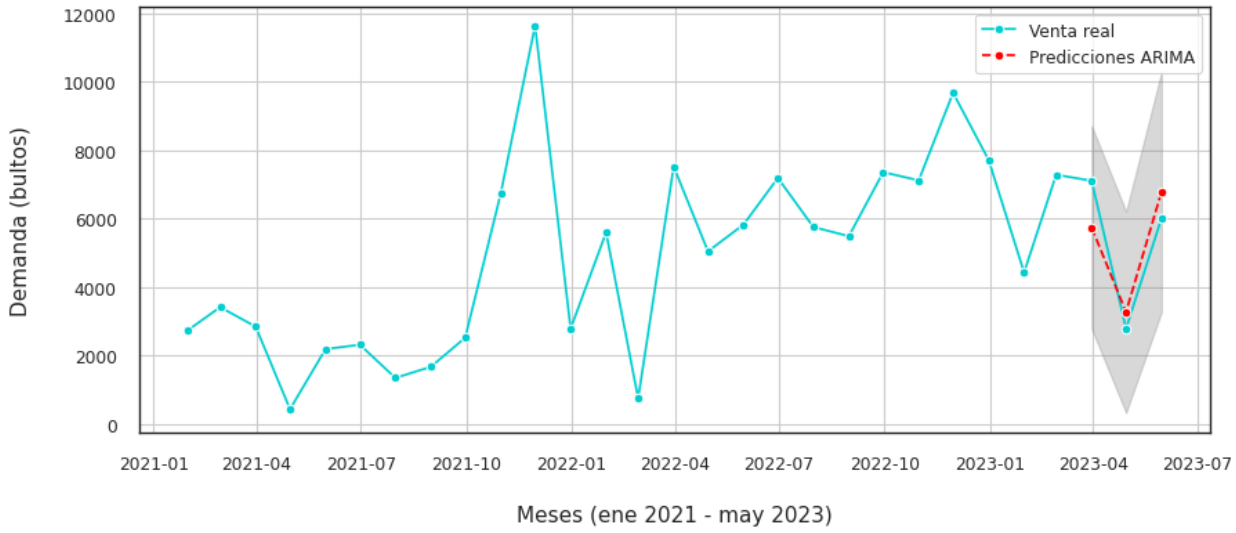


Figura A.24.
Papel higiénico triple hoja de 20m en Supermercados Costa

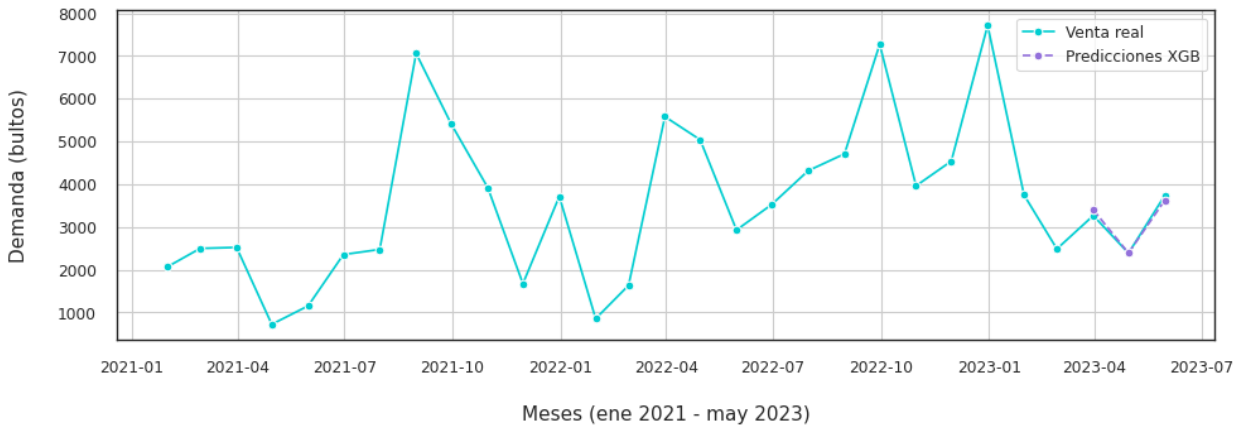


Figura A.25.
Servilletas Coctel en Supermercados Costa

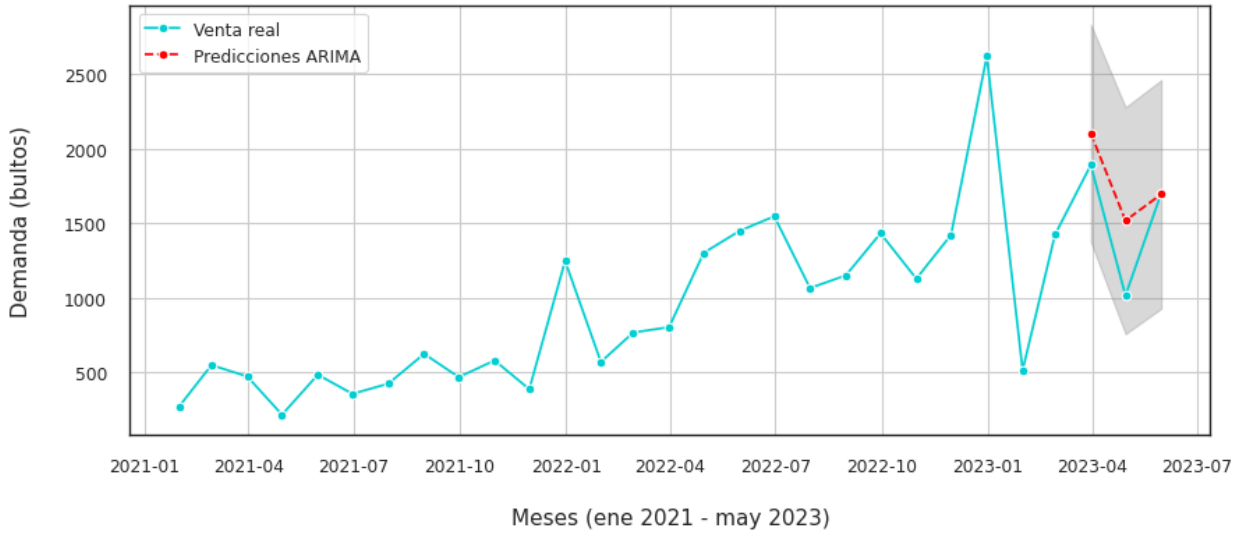


Figura A.26.
Servilletas Mesa en Supermercados Costa

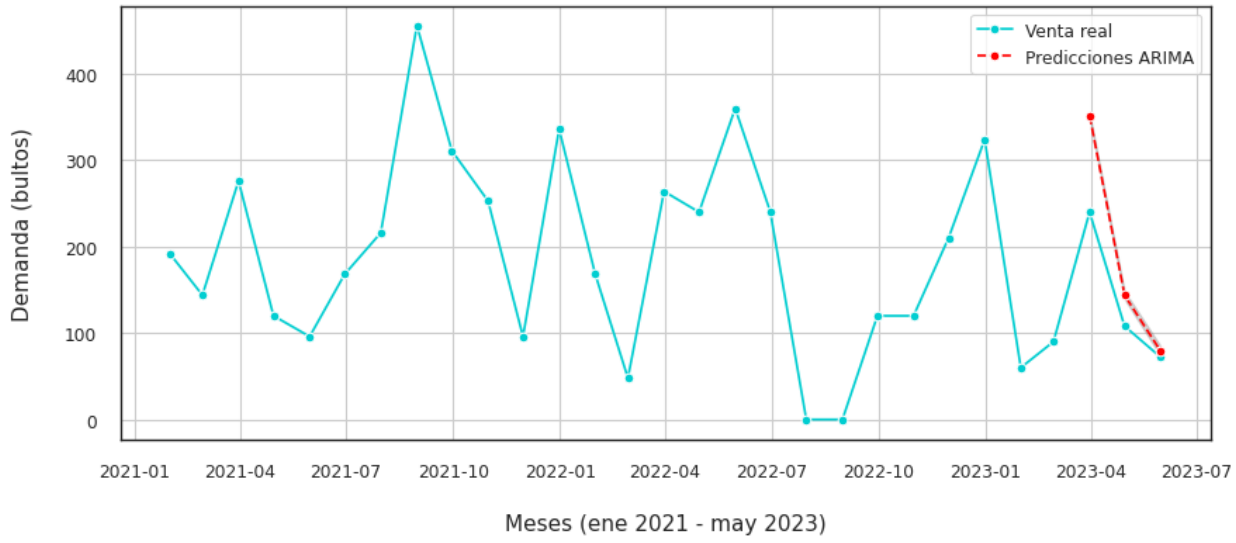


Figura A.27.
Servilletas Económicas en Supermercados Costa

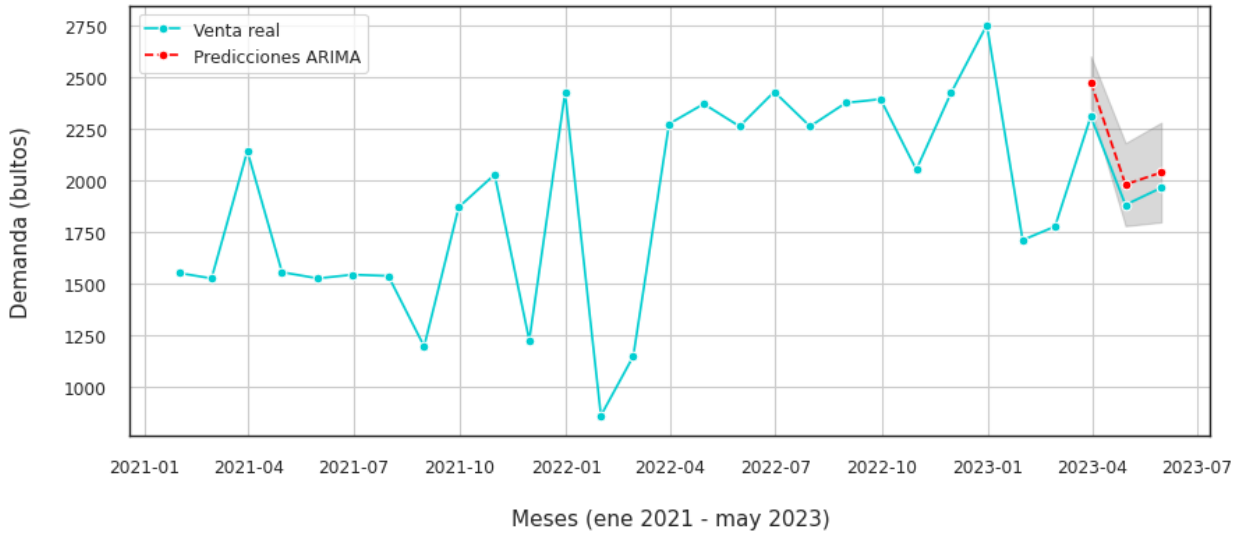


Figura A.28.

Toallas de papel Económicas en Supermercados Costa

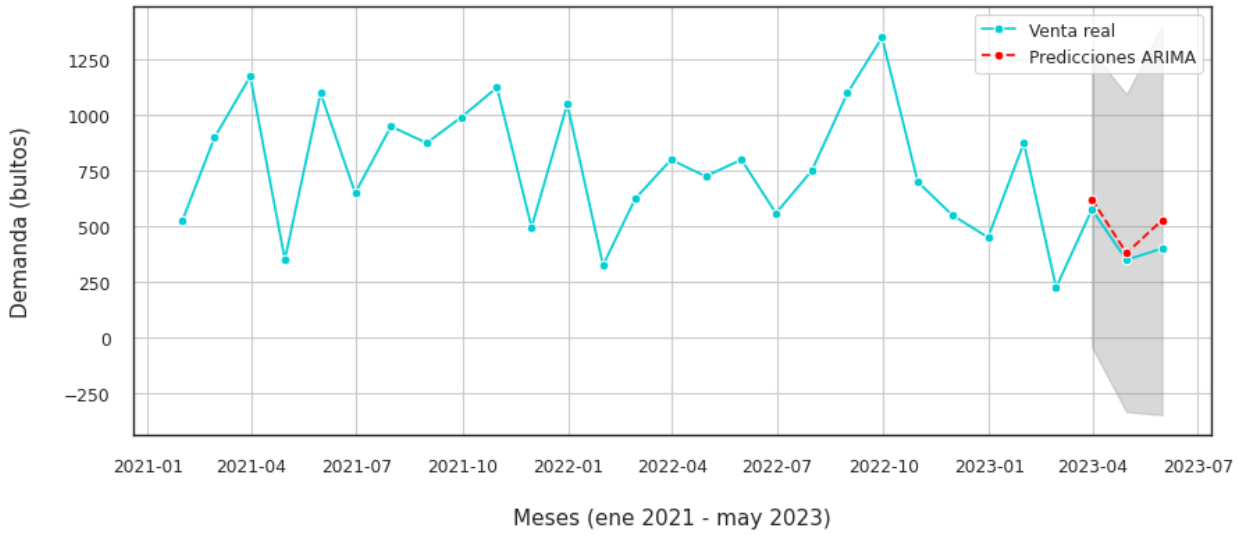


Figura A.29.

Pañales de bebé Premium en Supermercados Costa