Estimación de las ventas de La Favorita de Ecuador en el periodo previo a la pandemia COVID-19

Jessica Torres Franco
Gabriel Fernando Franco Calvo

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Yohany Pemberthy Salas, Magíster (MSc)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Torres Franco & Franco Calvo, 2023)
Referencia **Estilo APA 7 (2020)**	Torres Franco, J., & Franco Calvo, G. F. (2023). Estimación de las ventas de la Favorita de Ecuador en el periodo previo a la pandemia COVID-19 Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte V.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Agradecimientos

Agradecemos a nuestros familiares por el apoyo en este proceso y a nuestro asesor en este trabajo por su direccionamiento y guía para la culminación de este.

Tabla de contenido

Lista de tablas

Lista de figuras

Lista de ecuaciones

Siglas, acrónimos y abreviaturas

ERP	Enterprise Resource Planning
UdeA	Universidad de Antioquia
RMLSE	Root Mean Squared Logarithmic Error
MAE	Error Absoluto Medio
MSE	Error Cuadrático Medio
PCA	Análisis de Componentes Principales

Resumen

En el ámbito de las tiendas retail, la gestión eficiente del inventario es esencial para equilibrar la oferta y la demanda, evitando tanto la sobre existencia de productos perecederos como la escasez de artículos populares que puedan generar pérdida de ingresos. En este contexto, los pronósticos desempeñan un papel crucial. La aplicación de modelos de series de tiempo se presenta como una herramienta clave para estimar la demanda futura, basándose en el análisis de datos históricos de ventas y la identificación de tendencias y patrones estacionales. Este enfoque permite a las empresas ajustar su inventario de manera precisa, asegurando un stock adecuado para satisfacer la demanda en el momento oportuno. En resumen, la implementación efectiva de pronósticos, especialmente mediante modelos de series de tiempo, se convierte en una estrategia fundamental para optimizar la gestión del inventario en el mundo de las tiendas minoristas.

Palabras clave: Retail, series de tiempo, demanda.

Abstract

In the realm of retail stores, efficient inventory management is essential to balance supply and demand, avoiding both the overstocking of perishable goods and shortages of popular items that could lead to revenue loss. In this context, forecasting plays a crucial role. The application of time series models emerges as a key tool for estimating future demand, relying on the analysis of historical sales data and the identification of trends and seasonal patterns. This approach enables companies to adjust their inventory precisely, ensuring adequate stock to meet demand at the right moment. In summary, the effective implementation of forecasting, especially through time series models, becomes a fundamental strategy to optimize inventory management in the world of retail.

*Keywords***:** Retail, time series, demand.

- **Descripción del problema**

En el mundo de las tiendas retail, los pronósticos son una herramienta crucial para mantener el equilibrio entre la oferta y la demanda. Es importante gestionar cuidadosamente la cantidad de inventario que se debe comprar para evitar quedarse con productos perecederos con exceso de existencias o con una falta de artículos populares que generen una pérdida de ingresos.

Ante este desafío, es importante contar con las herramientas efectivas que permitan hacer pronósticos precisos, para poder mantener el equilibrio entre la oferta y la demanda, una posible solución es la aplicación de modelos de series de tiempo como herramienta clave para estimar la demanda futura de los productos. Este modelo se basa en el análisis de datos históricos de ventas y en la identificación de tendencias y patrones estacionales que permiten hacer pronósticos precisos sobre el consumo de los meses posteriores. De esta forma, las empresas pueden ajustar su inventario y asegurar el suficiente stock para satisfacer la demanda en el momento adecuado.

1.1. Problema de negocio

La Corporación La Favorita, una empresa minorista, busca desarrollar modelos que mejoren la precisión en las predicciones de ventas unitarias para su amplio catálogo de productos. El objetivo es perfeccionar la capacidad de anticipar la demanda de manera más efectiva, permitiendo una gestión más eficiente del inventario y optimizando la satisfacción del cliente.

1.2. Aproximación desde la analítica de datos

En el primer acercamiento al problema, se efectúa una descripción detallada de la información proporcionada por la empresa. Dada la presencia de diversas líneas de productos organizadas por familias, se busca abordar de manera más efectiva el problema planteado por la compañía al seleccionar la familia que haya tenido la mayor participación en las ventas históricas.

Dada la selección de la línea de productos se genera diferentes modelos que permitan dar solución al problema planteado por la compañía tales como series de tiempo, modelos de regresión lineal y logística, entre otros.

1.3. Origen de los datos

Se tienen datos históricos de las ventas desde el 2013 hasta el 2017, ventas antes de pandemia, realizadas en las distintas sucursales que se encuentran en todo el país de Ecuador de la compañía la Favorita. Adicionalmente se tienen datos históricos relacionados con el precio del petróleo y los días con festividades en el país para medir la influencia de esto sobre las ventas de la compañía.

1.4. Métricas de desempeño

Los modelos propuestos se evalúan con diferentes métricas de errores para determinar el modelo que mejor se adapte al problema planteado y obtenga mejores resultados en las predicciones, se tienen en cuenta las métricas tales como MAE, RMSLE (métrica propuesta por el negocio) y MSE.

Se espera que el modelo seleccionado cumpla con un RMSLE inferior al 0.20 y un R2-score superior al 85%.

2. Objetivos

2.1.Objetivo general

Optimizar la gestión de inventario para la Corporación Favorita en tiempos de pre pandemia garantizando la disponibilidad de productos y minimizando pérdidas por exceso de inventario y maximizando la eficiencia operativa a través de la implementación de modelos de pronósticos.

2.2.Objetivos específicos

- Realizar un análisis exhaustivo de la información histórica previa a la pandemia con el objetivo de desarrollar la estrategia de abordaje más efectiva, empleando modelos y herramientas adecuadas para optimizar la gestión de la situación actual y futura
- Desarrollar modelos que permitan anticipar la demanda de manera efectiva, permitiendo la gestión eficiente del inventario y optimizando la satisfacción del cliente.

3. Datos

3.1. Datos originales

Para el trabajo se toma una base de datos compartida en la plataforma Kaggle que se distribuyen en 7 archivos con un peso de 124.76 MB (comprimido en un archivo ZIP). Se tiene la siguiente información:

- **Holiday Events:** Información con las festividades en Ecuador, contiene 6 columnas y un peso de 22.31KB:
  - Date: Fecha del evento
  - Type: Tipo de fecha local, nacional o regional
  - Local name
  - Descripción: Nombre de la celebración
  - Transferred: Si el día festivo fue trasladado para ser celebrado otro día
- **Oil:** Precio diario del petróleo en ecuador, esta información es dada debido a que Ecuador es un país Petróleo dependiente y su salud financiera o económica es altamente vulnerable a choques en los precios del petróleo. El dataset tiene un peso de 20.58KB y contiene dos columnas:
  - Date: Fecha del precio de petróleo
  - Dcoilwtico: Precio del petróleo
- **Sample submission:** Un archivo de muestra con el formato correcto y con un peso de 342.15KB
- **Stores:** Dataset con información de las tiendas incluyendo información de ciudad, estado, tipo y clúster donde clúster representa una agrupación de tiendas similares de acuerdo con una clasificación proporcionada por la empresa, el dataset tiene un peso de 1.39KB
- **Train:** Dataset de entrenamiento proporcionado por la empresa, que comprende las series de tiempo que contiene las características de las tiendas, familia (tipo de producto) y una variable llamada on promotion que determina si la venta fue de un

producto con descuento y también la variable objetivo, este dataset tiene un peso de 121.8MB

- **Test:** Dataset de validación de los modelos, comprende las mismas variables que el dataset de entrenamiento y tiene un peso de 1.02MB
- **Transaction**

Para acceder a la información se realiza una carga en un repositorio de GitHub y descomprimiendo el archivo zip para acceder a los conjuntos de datos de manera individual.

3.2. Analítica descriptiva

Se procede a consolidar la información fragmentada mediante la unificación de diversos conjuntos de datos proporcionados por la Compañía La Favorita. Antes de llevar a cabo el análisis descriptivo, se realiza una revisión preliminar de los datos nulos que surgen al fusionar estos conjuntos, siguiendo los pasos detallados a continuación:

1. Unificación del Dataset de Entrenamiento y Días Feriados en Ecuador:
    a. Se observa la generación de datos nulos al combinar el dataset de entrenamiento con la información de días feriados en Ecuador.
    b. No obstante, se identifica que los valores nulos se deben a que las fechas en las cuales se registran como NaN corresponden a días "normales". Es decir, estos son días sin eventos nacionales, locales o departamentales.
    c. Estos valores son clasificados como “Normales” o “Normal Day” que permite la identificación de un día no feriado.
2. Unificación del Dataset de Entrenamiento y Precio del Petróleo:
    a. Al fusionar el dataset de entrenamiento con la información del precio del petróleo, se encuentran datos nulos.
    b. Sin embargo, estos valores nulos están asociados a fechas en las cuales el precio del petróleo en la bolsa de valores no experimenta cambios significativos. Por consiguiente, se procede a la eliminación de estos valores.

En la **Figura 1**, se presenta el comportamiento de las ventas diarias de la compañía. A primera vista, se percibe un aparente incremento de ventas hacia finales de cada año. No obstante, esta tendencia no es completamente evidente, por lo que se toma la decisión de realizar una mensualización de la serie para obtener resultados más claros y precisos.

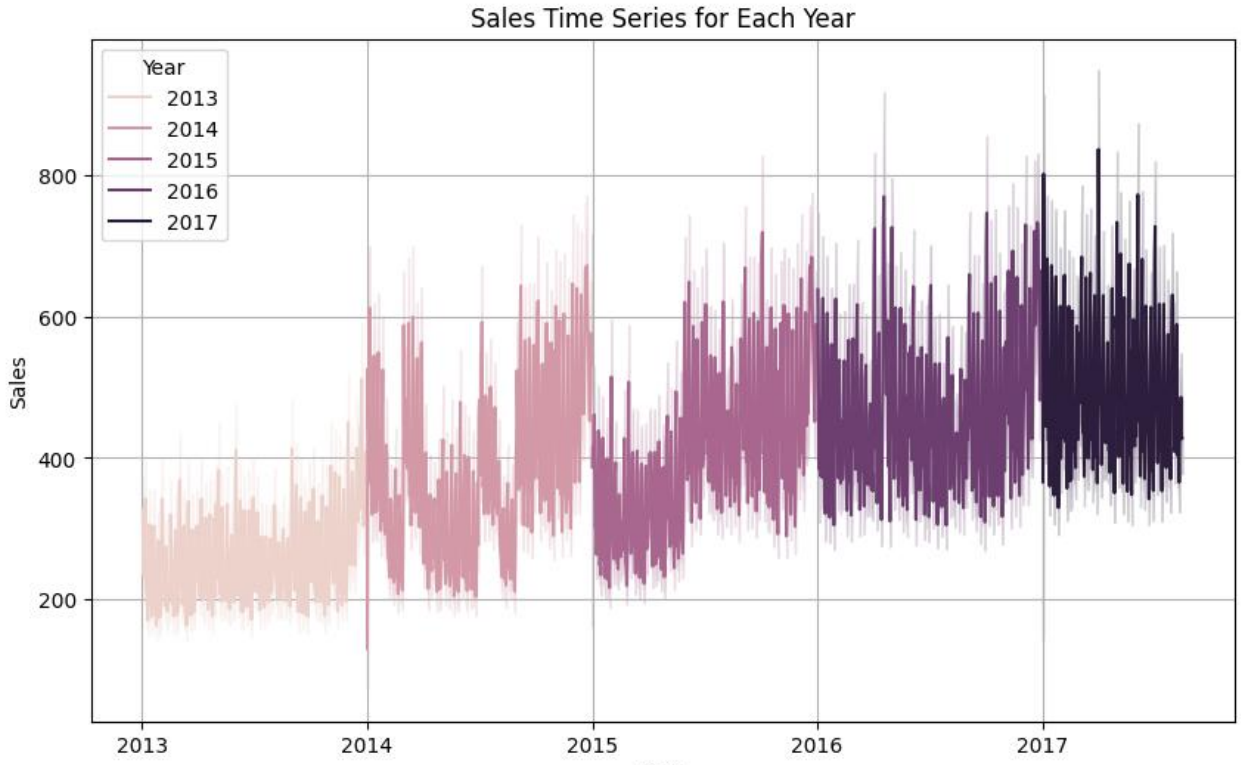

Figura 1. Serie temporal de las ventas diarias

En la **Figura 2**, se exhibe la serie temporal mensualizada, confirmando la tendencia identificada en la serie temporal diaria. Se destacan picos de ventas al final de cada año. Además, se observa que las ventas en el año 2017 son inferiores en comparación con el año anterior, tal como se puede apreciar en la **Tabla *1***, donde se presenta un resumen de ventas anuales.

Al profundizar en la investigación de las razones detrás de estas disminuciones, se revela que el año 2016 estuvo marcado por un terremoto significativo. En el año 2017, se experimentó una fase de recuperación post terremoto, sumado a la complejidad de una economía afectada por estos sucesos y a la transición gubernamental en curso en Ecuador. Estos factores contribuyeron a

las caídas en las ventas, proporcionando un contexto valioso para la interpretación de la serie temporal mensualizada.

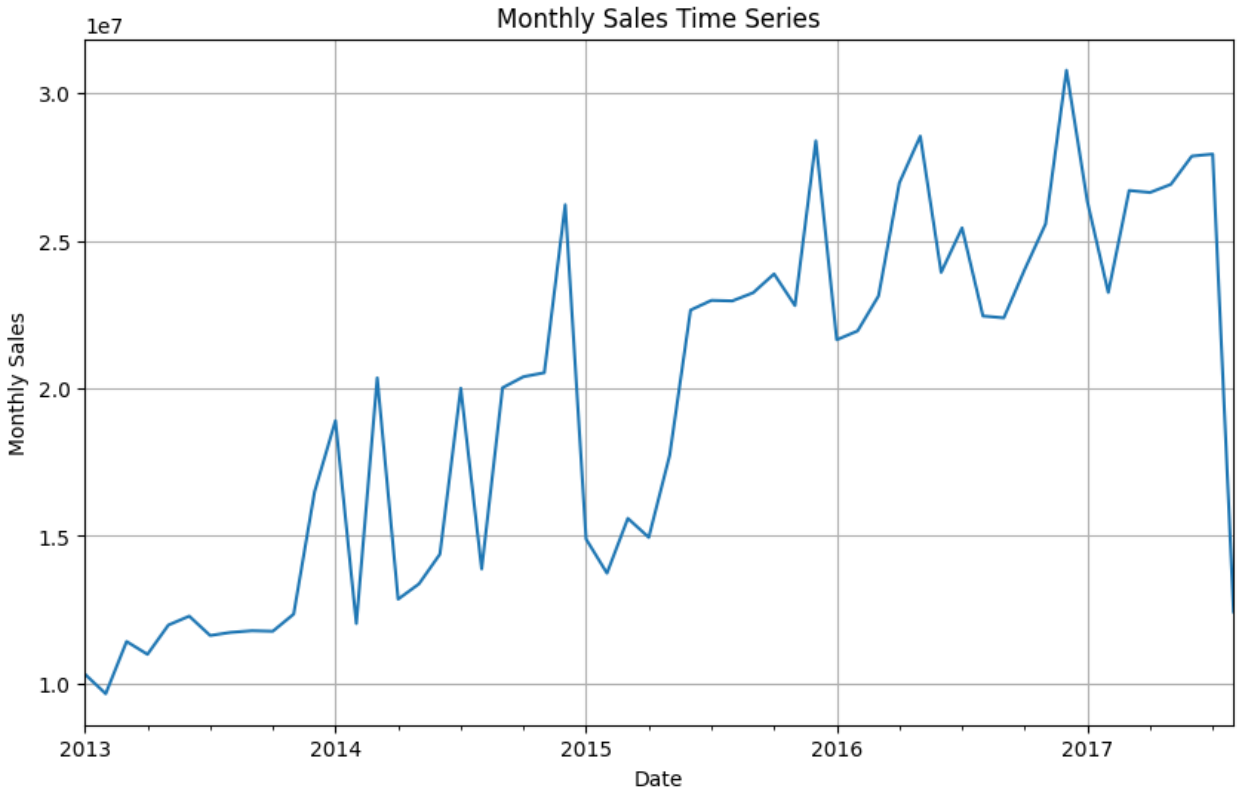

Figura 2 Serie Temporal mensualizada de las ventas

Tabla 1 Resumen descriptivo de las ventas

Año	Variación (Dinero)	Volatilidad	Ventas Max	Ventas Min	Ventas acumuladas	Variación Porcentual
2013	70,520,872	787	46,271	-	142,472,327	
2014	30,838,938	1,069	45,361	-	212,993,200	49%
2015	52,992,073	1,135	40,351	-	243,832,138	14%
	-					
2016	98,755,888	1,320	124,717	-	296,824,211	22%
2017		1,366	38,423	-	198,068,323	-33%

Dada la extensa variedad de productos en el portafolio de la compañía y la imperante necesidad de una gestión eficiente para garantizar el abastecimiento necesario y minimizar el desperdicio, se lleva a cabo un análisis detallado de cada una de las familias de productos que han contribuido significativamente a lo largo del tiempo. La idea es construir modelos específicos utilizando los datos de cada familia de productos, permitiendo así una gestión más precisa y adaptada a las características particulares de cada categoría.

Este enfoque estratégico facilitará la implementación de medidas específicas para la gestión de inventario, ajustándose a las demandas y comportamientos específicos de cada familia de productos. Al entender y modelar de manera individualizada el rendimiento de las categorías de mayor impacto, la compañía estará mejor equipada para optimizar la cadena de suministro, asegurar el cumplimiento de la demanda y reducir al mínimo cualquier exceso o desperdicio.

Este análisis más detallado y enfocado contribuirá a una toma de decisiones más precisa y eficiente en la gestión de inventarios, maximizando la rentabilidad y la satisfacción del cliente.

Como se evidencia en la **Tabla *2*** y la **Figura *3***, la familia identificada como ***Grocery I*** ostenta la mayor participación dentro de las ventas totales. Notablemente, en el año 2013, esta familia representaba el 43% de las ventas totales, pero a lo largo de los años su participación ha experimentado una disminución, llegando al 29.43%. A pesar de esta reducción, Grocery I continúa liderando en términos de participación.

En vista de esta relevancia persistente, se toma la decisión estratégica de focalizar el presente trabajo en esta familia de productos. Esta elección se sustenta en la importancia histórica y actual de Grocery I en el panorama de ventas, lo que permitirá un análisis más detallado y específico para optimizar la gestión de inventario y la toma de decisiones estratégicas asociadas a esta categoría clave.

Tabla 2 Resumen de ventas por año participación

Año	Familia	Ventas	Ventas Totales	Porcentaje de participación
2013	GROCERY I	59,200,265	142,472,327	41.55%
	BEVERAGES	22,077,089		15.49%
	CLEANING	17,760,118		12.47%
	DAIRY	7,858,863		5.52%
	BREAD/BAKERY	7,087,716		4.97%
2014	GROCERY I	66,821,696	212,993,200	31.37%
	BEVERAGES	41,695,143		19.58%
	PRODUCE	23,849,794		11.20%
	CLEANING	19,821,861		9.31%
	DAIRY	13,701,427		6.43%
2015	GROCERY I	77,080,423	243,832,138	31.61%
	BEVERAGES	49,008,269		20.10%
	PRODUCE	26,031,272		10.68%
	CLEANING	22,618,458		9.28%
	DAIRY	14,988,318		6.15%
2016	GROCERY I	88,760,925	296,824,211	29.90%
	BEVERAGES	63,456,596		21.38%
	PRODUCE	45,314,755		15.27%
	CLEANING	23,223,656		7.82%
	DAIRY	17,452,498		5.88%
2017	GROCERY I	58,283,721	198,068,323	29.43%
	BEVERAGES	44,858,817		22.65%

PRODUCE	29,816,256	15.05%
CLEANING	15,791,284	7.97%
DAIRY	11,692,401	5.90%

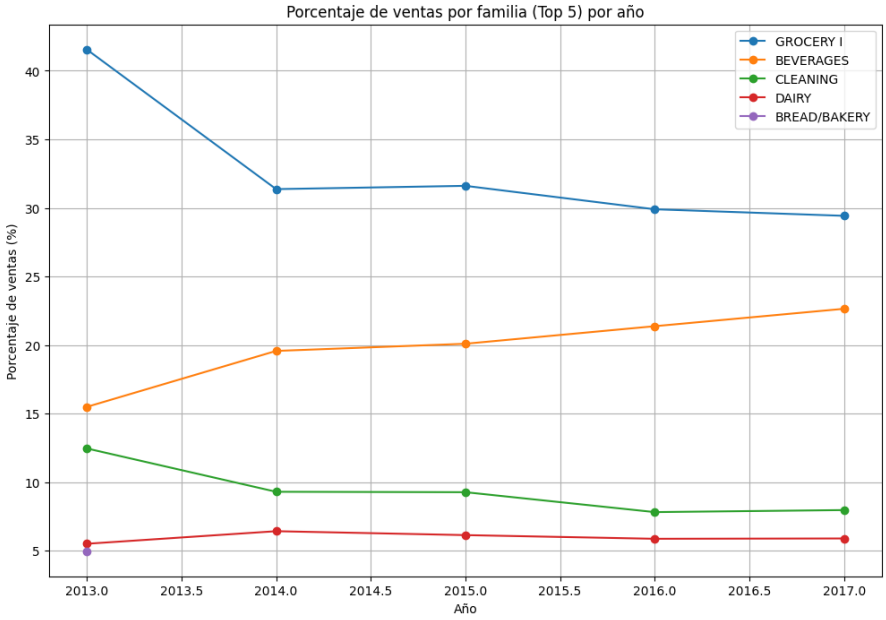

Figura 3 Evolución de participación de familia Top 5

Como se aprecia en la **Figura 4** y la **Figura 5**, se mantiene la tendencia general al analizar todos los productos en conjunto. Se observan picos hacia finales de cada año, indicando un aumento progresivo en las ventas, y se identifica una componente estacional en el comportamiento de los datos.

Este análisis integral resalta la consistencia de los patrones de ventas a lo largo del tiempo y sugiere la presencia de factores estacionales que afectan globalmente al conjunto de productos. La observación de estos comportamientos generales proporciona una visión más completa de las dinámicas de ventas, sentando las bases para estrategias efectivas de gestión y pronóstico que aborden las tendencias estacionales y los picos cíclicos.

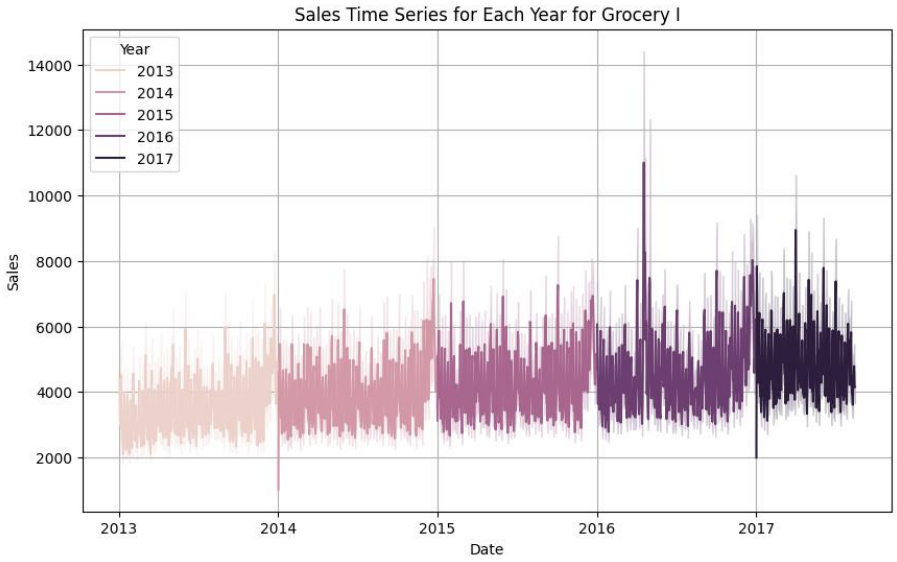

Figura 4 Series temporal de la Familia Grocery I

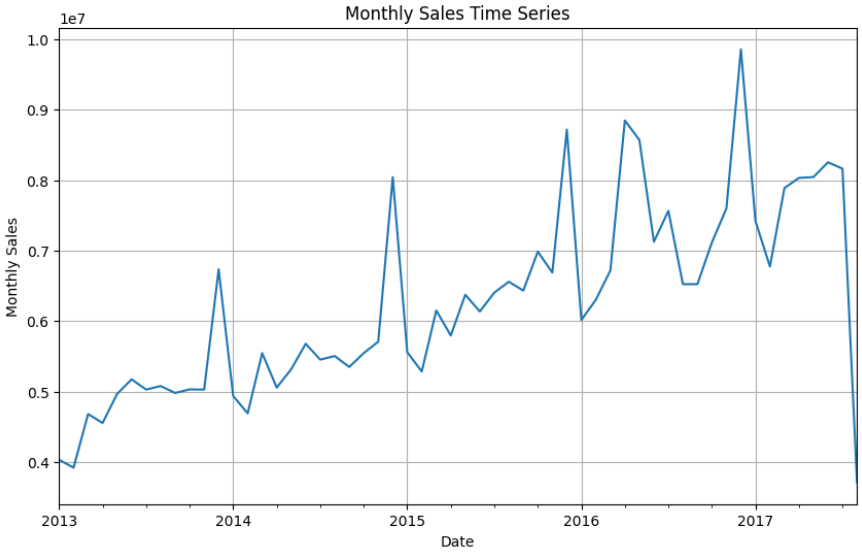

Figura 5 Serie mensualizada de ventas de Grocery I

La **Tabla *3*** revela que las ventas de la familia Grocery I experimentaron un crecimiento promedio anual del 14% durante los años 2013 hasta el 2016. Sin embargo, se observa una variación negativa del 34% en el año 2017 en comparación con el año anterior. Este descenso se explica por un período de recuperación post terremoto y otros acontecimientos que impactaron negativamente las ventas.

Esta información refuerza la decisión estratégica de elegir la familia Grocery I como el primer modelo para la estimación de la demanda. El análisis detallado de su historial de ventas proporciona una comprensión clara de las fluctuaciones y cambios en el rendimiento, permitiendo abordar específicamente los factores que contribuyen a la variabilidad observada. La elección de esta familia como punto de partida para los modelos de estimación de la demanda se respalda en la relevancia de sus datos históricos y en la capacidad de adaptarse a las circunstancias cambiantes del entorno comercial.

Tabla 3 Resumen descriptivo Ventas Grocery I

Año	Variación (Dinero)	Volatilidad	Ventas Max	Ventas Min	Ventas acumuladas	Variación Porcentual
2013	7,621,431	2,216	46,271	-	59,200,265	
2014	10,258,727	2,522	45,361	236	66,821,696	13%
2015	11,680,502	2,636	40,351	597	77,080,423	15%
2016	30,477,204	3,282	124,717	-	88,760,925	15%
2017	-	2,994	38,423	165	58,283,721	-34%

4. Proceso de analítica

4.1. Pipeline principal

El proceso de analítica se lleva en tres procesos de manera general, tal como se puede ver en la siguiente gráfica

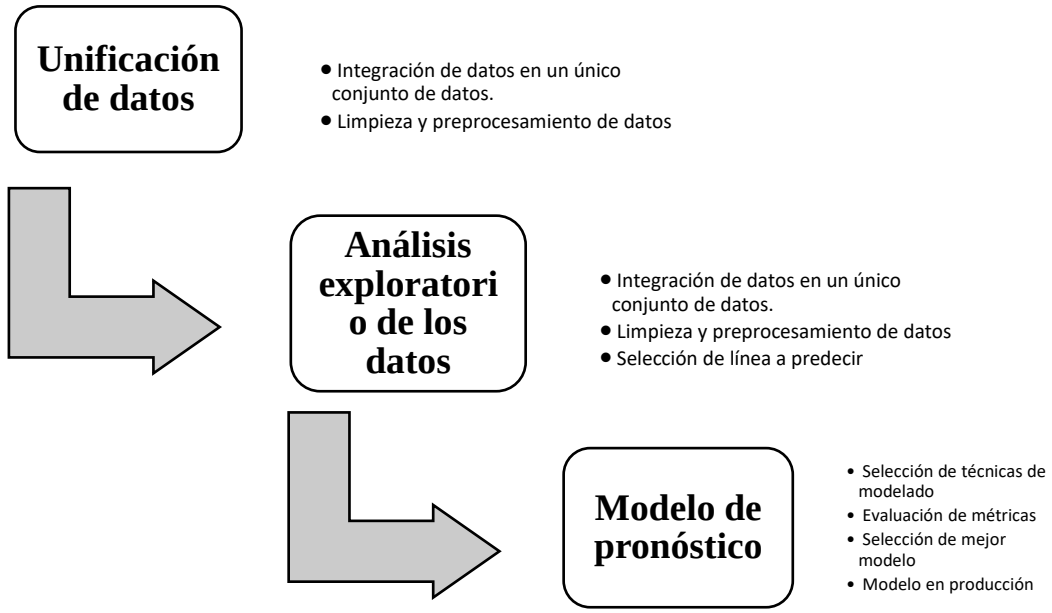

Figura 6 Pipeline del proceso de analítica

4.2. Preprocesamiento

Se procede a consolidar las bases de entrenamiento mediante la integración de información adicional, que incluye la identificación de festividades, precios del petróleo y las transacciones. Este proceso tiene como objetivo llevar a cabo una exploración estadística descriptiva que facilite la comprensión de los datos. La finalidad es proporcionar respuestas a las necesidades específicas del negocio, según lo observado en esta etapa.

La unificación de estas fuentes de datos complementarias permite obtener una visión más completa y contextualizada, proporcionando insights valiosos para la toma de decisiones estratégicas. La exploración estadística descriptiva resultante contribuirá a entender las relaciones entre las variables, identificar patrones significativos y orientar el análisis de cara a las metas y requerimientos comerciales.

4.3.Modelos

Una serie de tiempo es una serie de puntos de datos ordenados en el tiempo. Durante un evento en una serie de tiempo, las medidas son organizadas típicamente en tiempos sucesivos (Adhikari, 2013). Gracias a esto, existe la posibilidad de una correlación entre las observaciones. En gran medida, el análisis de las series de tiempo tiene como objetivo explicar esta correlación y las principales características de los datos, usando modelos estadísticos y métodos descriptivos apropiados (Paul S, 2009). En una serie de tiempo, el tiempo es a menudo la variable independiente y el objetivo suele ser hacer un pronóstico para el futuro. Se pueden extraer diversas características de las series de tiempo, como las tendencias y variaciones estacionales que pueden ser modeladas de forma determinista con funciones matemáticas del tiempo (Paul S, 2009).

Sea Y_t una serie temporal en la que t denota el momento en que se toma la observación, donde $t \in z^+$. El objetivo es construir un modelo que describa la evolución de la serie a través del tiempo, para esto se asume que los datos se pueden expresar como una función de una componente de tendencia T_t estacional s_t y un error E_t (Jonathan, 2009)

Ruido Blanco

Un proceso ϵ_t se denota ruido blanco de media 0 y varianza σ^2 sí satisface

$$E(\epsilon_t) = 0, Var(\epsilon_t) = \sigma^2 < \infty, Cov(\epsilon_t, \epsilon_{t-k}) = 0$$

Para todo $k \neq 0$. En particular, una sucesión de variables aleatorias independientes e idénticamente distribuidas, con media 0 y varianza σ_ϵ^2 representa un caso especial de un proceso de ruido blanco, y que se denota por $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. Si además ϵ_t se distribuye normalmente, la serie se denomina ruido blanco gaussiano.

Hay tres formas de comprobar si la serie temporal se asemeja al ruido blanco:

- Trazando la serie temporal
- Comparando la media y la desviación estándar a lo largo del tiempo
- Examinando los gráficos autocorrelación

Modelo AR(p)

Los modelos de autorregresivos de orden finito *p*, son una representación de un proceso aleatorio, en el que la variable de interés depende de sus observaciones pasadas. En general, para denotar el modelo autorregresivo AR se usa AR(p). Así, un modelo (AR) de orden *p* se puede escribir como

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

Ecuación 1 Modelo AR de orden *p*

Para constantes $\phi_0, \cdots, \phi_p$ y $\epsilon_t \sim RB(0, \sigma^2)$.

Modelo MA(q)

Los modelos de medias móviles de orden finito *q*, son una aproximación común para las series de tiempo univariadas. El modelo de medias móviles especifica que la variable de salida depende linealmente del valor actual y varios de los anteriores. En general para denotar un modelo de medias móviles MA se usa MA(*q*). Así, un modelo MA de orden *q* se puede escribir como

$$Y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p \in_{t-q}$$

Ecuación 2 Modelo MA de orden *q*

Para constantes $\theta_0, \cdots, \theta_p$ y $\epsilon_t \sim RB(0, \sigma^2)$.

Modelo ARMA(p,q)

Un proceso ARMA(*p*,*q*) es un modelo que combina las propiedades de memoria larga, de los AR(*p*) con las propiedades de ruido débilmente autocorrelacionado en los MA(*q*), y que tiene suficiente flexibilidad y parsimonia para representar una variedad grande de procesos estacionarios en covarianza

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=0}^{q} \theta_0 \in_{t-j}$$

Ecuación 3 Modelo ARMA de orden *p* y *q*

Donde $\epsilon_t \sim RB(0, \sigma^2)$.

Modelos de regresión lineal con mínimos cuadrados ponderados

Modelos de regresión lineal con errores con comportamiento heterocedasticidad pueden modelarse a través de un método llamado mínimos cuadrados ponderados (WLS, por sus siglas en inglés), donde los parámetros a estimar son obtenidos minimizando la suma ponderada cuadrática de los residuos donde los pesos son inversamente proporcionales a la varianza de los errores. Se

usa el método WLS para estimar los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ son obtenidos minimizando la siguiente ecuación,

$$\sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_0 x_{i1} - \ldots - \beta_p x_{ip})^2$$

Ecuación 4 Mínimos cuadrados ponderados

Donde $w_i = {}^{1}/_{\sigma_i^2}$ son pesos inversamente proporcionales a la varianza de los residuales. Cuando una observación tenga un peso pequeño, el método castigará severamente a la hora de determinar los valores de β.

4.4.Métricas

La métrica utilizada para la evaluación del desempeño del modelo es el **RMSLE** (Root Mean Squared Logarithmic Error) propuesta por el negocio, sin embargo, se utilizarán métricas adicionales para verificar el desempeño de los distintos modelos.

RMSLE:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(1+\hat{y}_i) - log(1 - \hat{y}_i))^2}$$

Ecuación 5 Root Mean Squared Logarithmic Error

MAE: Error Absoluto Medio (MAE) es una métrica usada para evaluar la precisión de un modelo de precisión. Mide la diferencia entre el valor predicho y el valor real

$$MAE = \frac{1}{n}\sum |y_{pred} - y_{actual}|$$

Ecuación 6 Error Absoluto Medio

Donde,

- y_pred: Valores predichos por el modelo
- y_actual: Valores reales
- n: Número de observaciones en el dataset

MSE: Error cuadrático medio (MSE) es una métrica usada para evaluar la precisión del modelo de predicción, midiendo la diferencia entre el valor predicho y el valor real

$$MSE = \frac{1}{n}\sum(y_{pred} - y_{actual})^2$$

Ecuación 7 Error Cuadrático Medio

Donde,

- y_pred: Valores predichos por el modelo
- y_actual: Valores reales
- n: Número de observaciones en el dataset

Para el cálculo de las métricas se utiliza las funciones de la librería Sklearn:

- sklearn.metrics.mean_absolute_error
- sklearn.metrics.mean_squared_error

Para el cálculo de la función RMSLE, se crea una función que recibe como argumentos los valores predichos por el modelo y los valores actuales para realizar el respectivo cálculo de la métrica tal como se muestra en la siguiente imagen:

```python
# Función para calcular el RMSLE (Root Mean Squared Logarithmic Error):
#
# predicted  : Array que contiene los valores predichos por el modelo
# actual     : Array con los valores reales
#
# Ejemplos:
#    rmsle([2,8,7], [3,5,12]])
# ================================================================================
def RMSLE(predicted, actual):
    # Apply logarithm to predicted and actual values
    predicted_log = np.log1p(predicted)
    actual_log = np.log1p(actual)

    # Calculate the squared differences between predicted and actual logarithmic
    # values
    squared_diff = (predicted_log - actual_log) ** 2

    # Calculate the mean squared difference
    mean_squared_diff = np.mean(squared_diff)

    # Calculate the RMSLE by taking the square root of the mean squared
    # difference
    RMSLE_resultado = np.sqrt(mean_squared_diff)
    return RMSLE_resultado
```

Figura 7 Función creada para cálculo de métrica

5. Metodología

5.1. Baseline

La fase inicial implica un análisis exploratorio de los datos iniciales para adquirir conocimientos sobre las variables del conjunto de datos y determinar la mejor aproximación a los modelos. Como resultado de este análisis, se decide construir un modelo inicial focalizado en una única familia de productos.

Como se evidencia en la **Figura 8**, las primeras familias con mayores ventas en la compañía están vinculadas principalmente a alimentos, junto con la categoría de limpieza. Dada la significativa contribución de las categorías alimentarias a las ventas totales, se refuerza la necesidad de desarrollar modelos específicos en lugar de generales. Esta estrategia se considera la más efectiva para generar predicciones de demanda precisas y adaptadas a las particularidades de cada categoría.

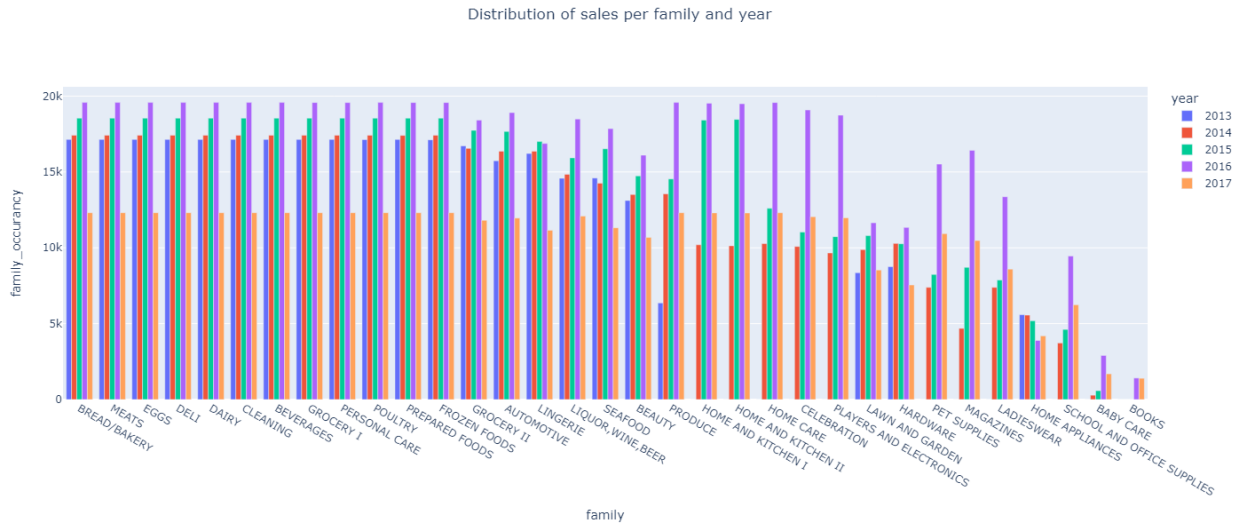

Figura 8 Distribución de ventas por familia y año

5.2. Validación

Para llevar a cabo la validación y entrenamiento de los modelos, se procede a realizar una partición del conjunto de datos, asignando el 75% de los datos para el proceso de entrenamiento de los modelos, mientras que el 25% restante se reserva para la fase de prueba. Aunque la compañía proporciona datasets previamente particionados, se ha tomado la decisión de realizar una nueva partición específicamente en el conjunto de entrenamiento. Además, se opta por considerar el conjunto de datos de prueba original como un conjunto adicional destinado para la evaluación posterior de los modelos entrenados.

5.3. Iteraciones y evolución

En la primera iteración se realiza un modelo de regresión con los datos en bruto (los datos fueron escalados previamente) es decir, con todas las variables de tal forma que permita evaluar la significancia de las variables en modelo y tomar aquellas que tienen un valor-p mayor al 0.05.

Tal como se puede observar en la **Figura *9***, se presenta los resultados del primero modelo de regresión para la evaluación de la significancia de las variables, esta iteración nos permite pasar de 38 variables a 28.

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.883
Model:	OLS	Adj. R-squared:	0.883
Method:	Least Squares	F-statistic:	4506.
Date:	Fri, 17 Nov 2023	Prob (F-statistic):	0.00
Time:	03:30:39	Log-Likelihood:	-1.7257e+05
No. Observations:	20925	AIC:	3.452e+05
Df Residuals:	20889	BIC:	3.455e+05
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1382.5485	741.072	-1.866	0.062	-2835.106	70.009
onpromotion	5.9778	0.268	22.313	0.000	5.453	6.503
oil_price	-6.2946	0.318	-19.781	0.000	-6.918	-5.671
transactions	2.5904	0.023	112.978	0.000	2.545	2.635
state_store__Azuay	-3150.5955	4415.185	-0.714	0.475	-1.18e+04	5503.510
state_store__Bolivar	-2603.7595	4700.794	-0.554	0.580	-1.18e+04	6610.161

Figura 9 Resultados parciales primer modelo de regresión

En la iteración siguiente, se implementa una técnica de reducción de dimensionalidad de los datos con el propósito de mejorar los resultados, no mediante la eliminación de variables, sino mediante la creación de nuevas variables sintéticas que capturan el 98% de la variabilidad de la información. En lugar de retener las 38 variables originales, se emplea el Análisis de Componentes Principales (PCA, por sus siglas en inglés), generando un conjunto de datos reducido a 17 variables.

La **Figura *10*** ilustra que las primeras tres variables explican el 36% de la variabilidad total. A partir de la variable número 15, se observa un cambio y no se evidencian diferencias significativas en las contribuciones más allá de este punto. Sin embargo, con el objetivo de retener al menos el 98% de la variabilidad, se seleccionan las primeras 17 componentes principales, marcando así el umbral de nuestro objetivo.

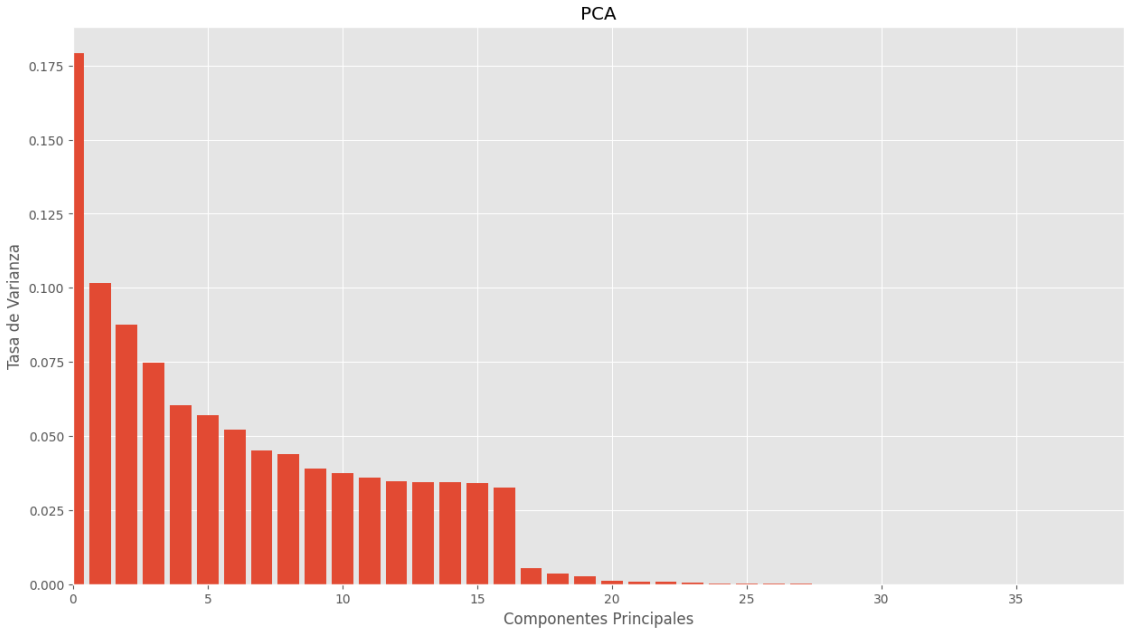

Figura 10 Componentes principales PCA

Las dos primeras iteraciones se llevan a cabo aplicando los modelos de regresión y XGBoost. Además, se introduce un modelo de series de tiempo, que representa el enfoque tradicional para abordar este tipo de problemas. En este contexto, se lleva a cabo un análisis de la serie temporal tanto en su forma diaria como mensual, como se detalla en la **Figura 11**.

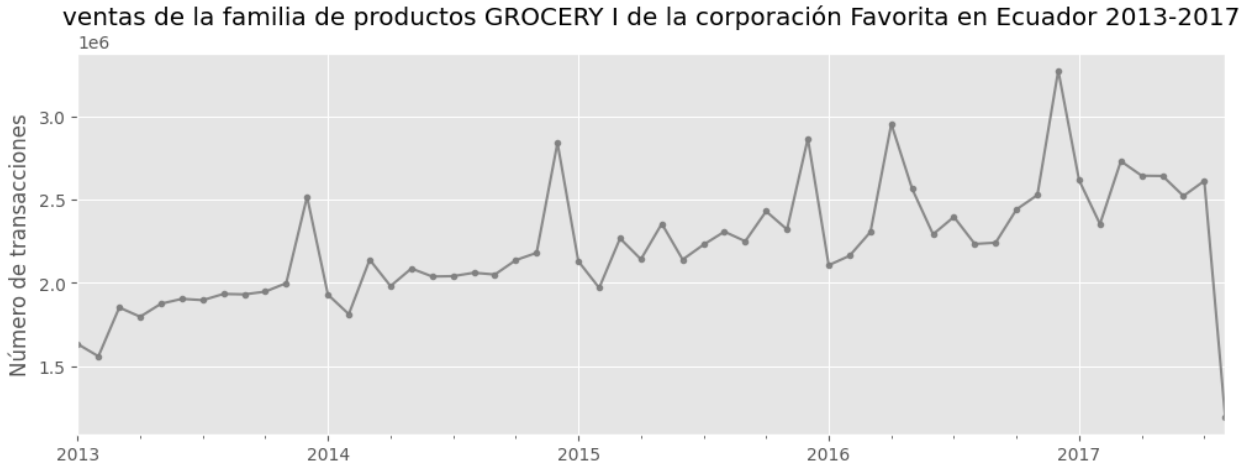

Figura 11 Serie mensualizada

La iteración de los modelos de series temporales implica la creación de una grilla que facilita la evaluación de diversos hiperparámetros, como el orden AR, la integración, MA,

estacionalidad, integración estacional, MA estacional y el periodo de estacionalidad semanal. Debido a la considerable cantidad de hiperparámetros, se lleva a cabo la evaluación de un total de 144 modelos. Posteriormente, se seleccionan los dos modelos con los valores más bajos en términos de AIC, BIC y HQIC para su consideración adicional.

En la **Figura *12*** y **Figura 13**, se detalla el resultado del modelado de la serie temporal, donde se han seleccionado específicamente dos modelos para su análisis detallado. Estos modelos se identifican como el N°106 SARIMAX (2,0,2)(1,1,1)[52] y el N°119 SARIMAX (2,0,1)(0,1,0)[52].

Las Figuras muestran la comparación entre los valores reales de la serie temporal y las predicciones medias generadas por los modelos seleccionados. Esta representación visual permite evaluar la capacidad de los modelos para ajustarse a la realidad de los datos observados. La observación clave es que, según las gráficas presentadas, ambos modelos logran ajustarse de manera efectiva a los valores reales, lo que sugiere que estos modelos SARIMAX son apropiados para describir y prever el comportamiento de la serie temporal en cuestión.

En la serie original, se identifica un pico que los modelos no logran captar de manera precisa. Es decir, las predicciones se sitúan notablemente por debajo de los valores reales en ese punto específico de la línea temporal. Esta discrepancia podría atribuirse a la influencia de un fenómeno externo no reflejado en los datos disponibles, lo cual escapa a la capacidad de los modelos para captar. A pesar de esta limitación en la predicción de este evento singular, es importante destacar que, en general, el modelo demuestra un ajuste considerablemente mejor en el resto de la serie temporal.

Este análisis visual refuerza la selección de estos modelos específicos entre los 144 evaluados, ya que demuestra su capacidad para capturar patrones y tendencias en los datos observados. Estos resultados respaldan la validez y la utilidad de los modelos SARIMAX (2,0,2)(1,1,1)[52] y SARIMAX (2,0,1)(0,1,0)[52] para la modelización y predicción de la serie temporal bajo consideración.

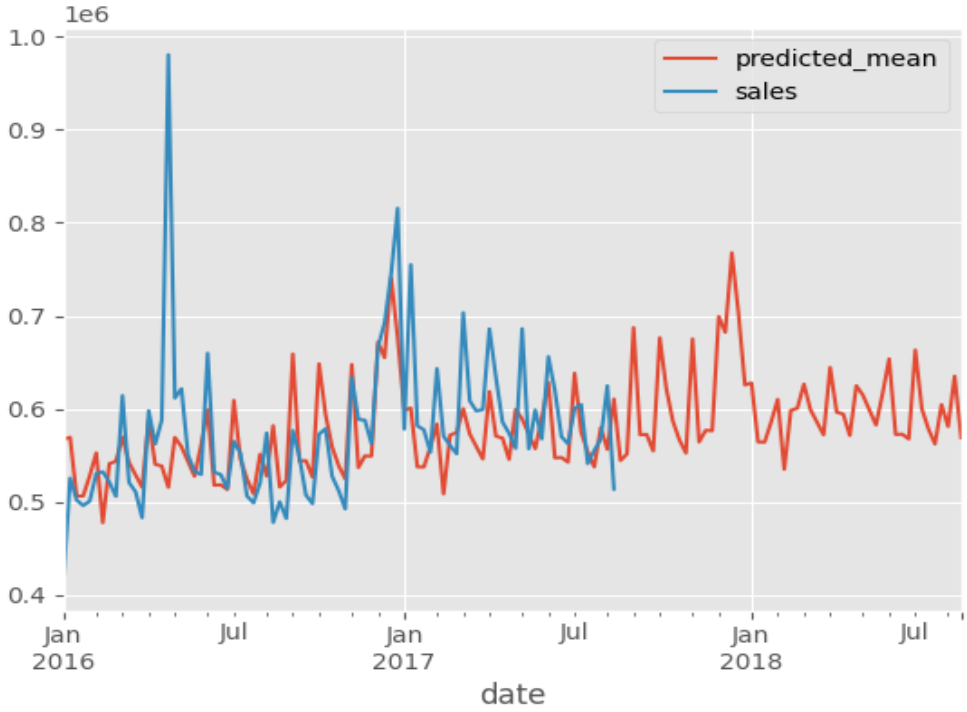

Figura 12 Predicción media del modelo N°106

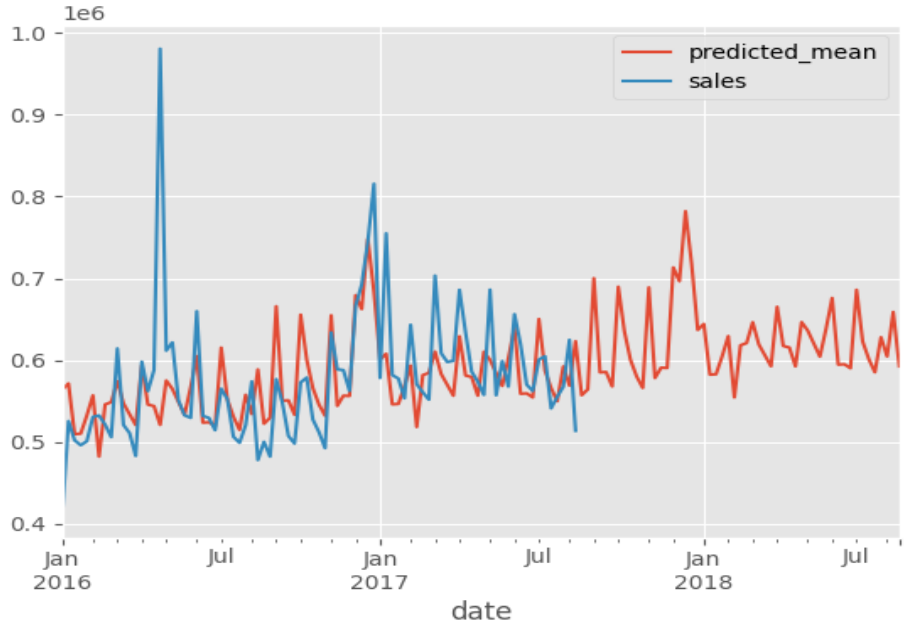

Figura 13 Predicción media del modelo N°119

5.4 Herramientas

Para llevar a cabo el proyecto se usan las librerías disponibles en el lenguaje Python, divididas en dos:

1. Lectura, procesamiento y visualización de datos
    i. Pandas
    ii. Plotly
    iii. Numpy
    iv. Scipy
    v. Matplotlib, plotly y seaborn
    vi. Git
2. Modelación
    i. Sklearn

6. Resultados y discusión

6.1. Métricas

Los resultados obtenidos a partir de las primeras iteraciones permiten llevar a cabo una comparación exhaustiva y un análisis del rendimiento de cada uno de los modelos propuestos. Como se destaca en la **Tabla *4***, es evidente que los modelos SARIMAX son los que muestran los mejores desempeños presentando un menor MSE 55,882 para el modelo SARIMAX 106 y 56,311 para el modelo SARIMAX 109 además de un RMSLE del 0.11 y 0.11 respectivamente, y es notable que ambos presentan resultados similares en términos de las métricas evaluadas.

Además, al comparar los modelos SARIMAX con el modelo XGBoost (entrenado con los datos en su forma original), se observa que este último muestra un RMSLE cercano a los modelos SARIMAX, con un valor de 0.12. Sin embargo, es importante destacar que el MSE asociado al modelo XGBoost es notablemente más elevado, alcanzando los 546,513, en comparación con los modelos SARIMAX mencionados anteriormente.

Esto sugiere que, aunque el modelo XGBoost logra aproximarse en términos de precisión de las predicciones, su desempeño se ve afectado por un mayor error cuadrático medio. Estos hallazgos indican que, en este contexto particular, los modelos SARIMAX pueden superar al XGBoost en términos de ajuste a los datos y capacidad predictiva, resaltando la importancia de seleccionar el modelo más apropiado según las características específicas del conjunto de datos y los objetivos de la predicción.

Tabla 4 Métricas de los modelos

	Regresión		**XGBoost**			
	Datos completos	**PCA**	**Datos completos**	**PCA**	**SARIMAX (2,0,2)(1,1,1)[52]**	**SARIMAX (2,0,1)(0,1,0)[52]**
MAE	608.21	910.34	399.05	469.48	-	-
MSE	1,243,662.1	2,168,746.2	546,515.3	676,267.4	55,882.09	56,311.87
RMSLE	0.18	0.25	0.12	0.15	0.1125	0.1129
R2-Score	0.82	0.7	0.92	0.9	-	-

6.2. Evaluación cualitativa

Desde los resultados obtenidos (consultar **Tabla *4***), se destaca que los modelos SARIMAX exhiben los mejores desempeños en términos de menor error, lo que sugiere su capacidad para ajustarse de manera precisa a la serie temporal en consideración. Además, se observa que el modelo XGBoost, al utilizar todas las variables disponibles, también presenta un desempeño notablemente bueno.

Para el modelo de regresión en la primera iteración, la **Figura 14** revela una tendencia preocupante: los valores predichos están consistentemente por debajo de los valores reales. Este patrón sugiere una posible situación de underfitting, indicando que el modelo no está capturando adecuadamente la complejidad de los datos y, como resultado, no logra realizar predicciones precisas.

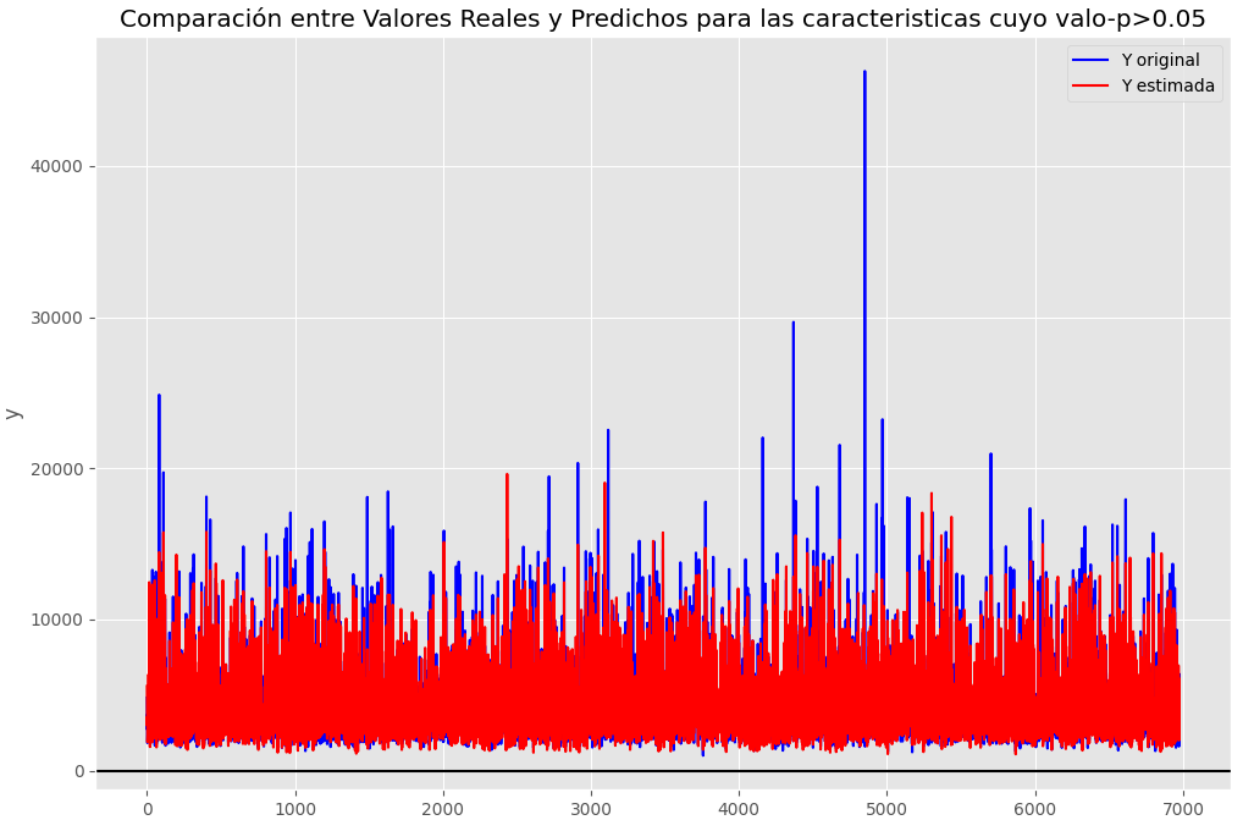

Figura 14 Comparación valores reales vs Predicho Regresión

Para profundizar en esta evaluación, la **Figura 15** presenta una comparación visual entre los valores predichos y los valores reales. Aquí, se observa claramente que los puntos de datos se sitúan significativamente por debajo de la línea de 45 grados, que representa la igualdad entre las predicciones y los valores reales. Este fenómeno subraya aún más la discrepancia entre las predicciones del modelo y la realidad observada. La ubicación sistemática de los puntos por debajo de la línea de 45 grados indica que el modelo está subestimando consistentemente los valores reales.

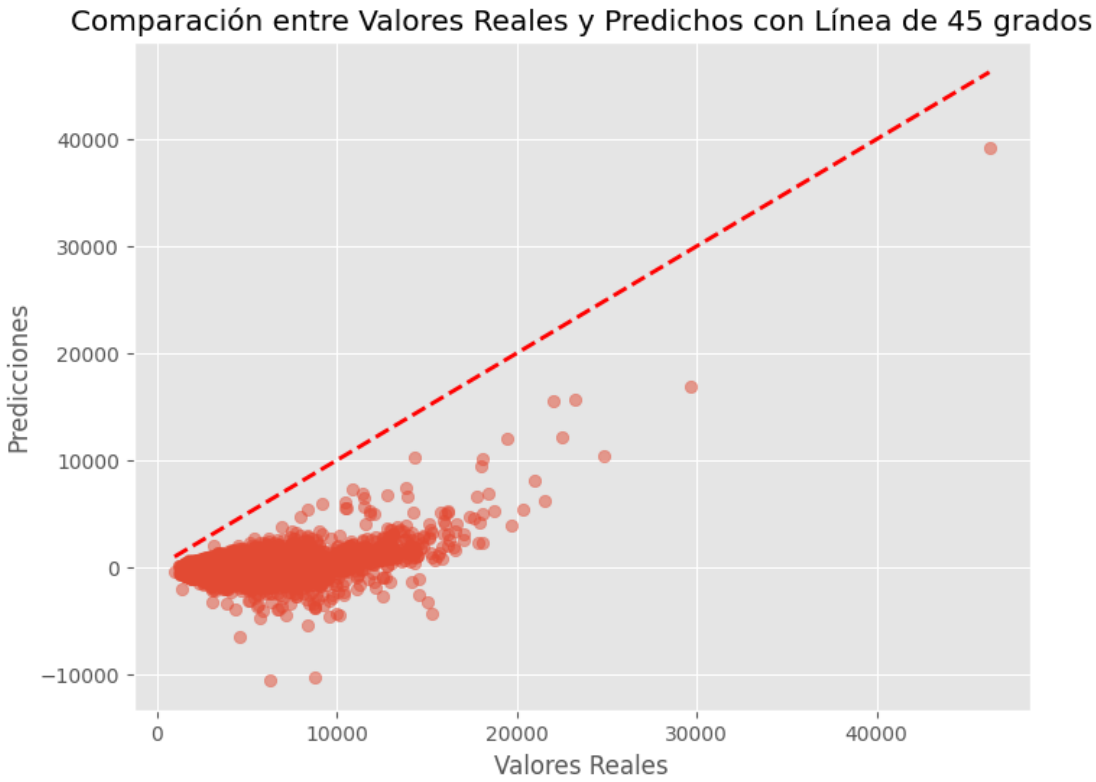

Figura 15 Comparación entre valores reales y predichos con línea de 45° - Modelo de regresión sin PCA

Para abordar el problema de underfitting identificado en la primera iteración, se proponen dos enfoques estratégicos. En primer lugar, se implementará una regresión lineal con una reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA). En segundo lugar, se explorará la implementación de un modelo basado en XGBoost. XGBoost es conocido por su capacidad para manejar problemas de underfitting y overfitting mediante la optimización de árboles de decisión y técnicas de regularización.

A pesar de la implementación de la reducción de dimensionalidad mediante PCA y la utilización de un modelo de regresión lineal, la persistencia del comportamiento subóptimo es evidente en la **Figura 16**. Aunque se observa una mejora con respecto a la primera iteración, aún se evidencia la falta de correlación significativa entre los valores predichos y los valores reales. Este resultado indica que, a pesar de la simplificación introducida por la reducción de dimensionalidad, la regresión lineal no es suficiente para modelar la complejidad inherente en los datos de demanda.

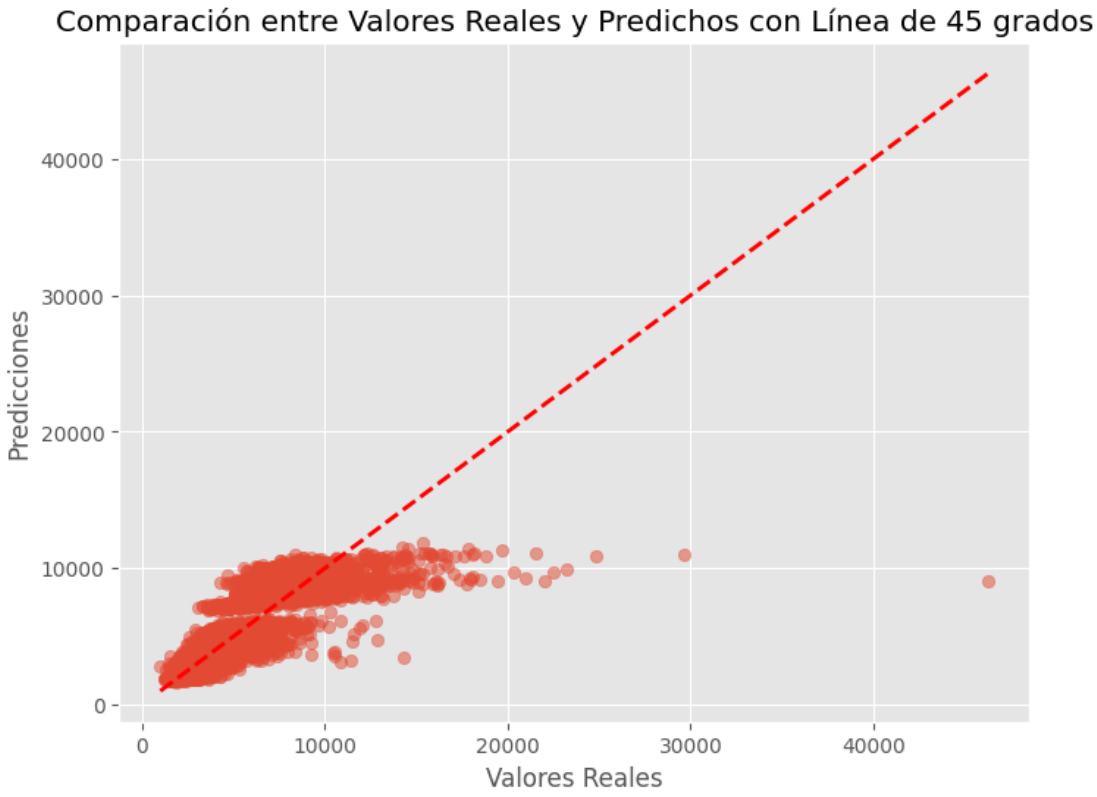

Figura 16 Comparación entre valores reales y predichos con línea de 45° - Modelo de regresión con PCA

La implementación de dos modelos XGBoost, uno con los datos en bruto y el otro con reducción de dimensionalidad PCA, representa un enfoque prometedor en la mejora del rendimiento del modelo. En particular, el primer modelo, que utiliza los datos en bruto, muestra un comportamiento alentador como se evidencia en la **Figura 17**. Aquí, los puntos que representan los valores predichos están cercanos a la línea de 45 grados, indicando una correlación significativa entre las predicciones y los valores reales. Este resultado sugiere que el modelo XGBoost, al trabajar con datos sin procesar, ha logrado capturar patrones complejos y no lineales presentes.

En cuanto al segundo modelo, se observa la misma correlación entre los valores predichos y valores reales, sin embargo, al realizar las métricas de medición, el modelo 1 sin PCA tiene métricas menores en RSMLE y MSE (Ver **Tabla *4***).

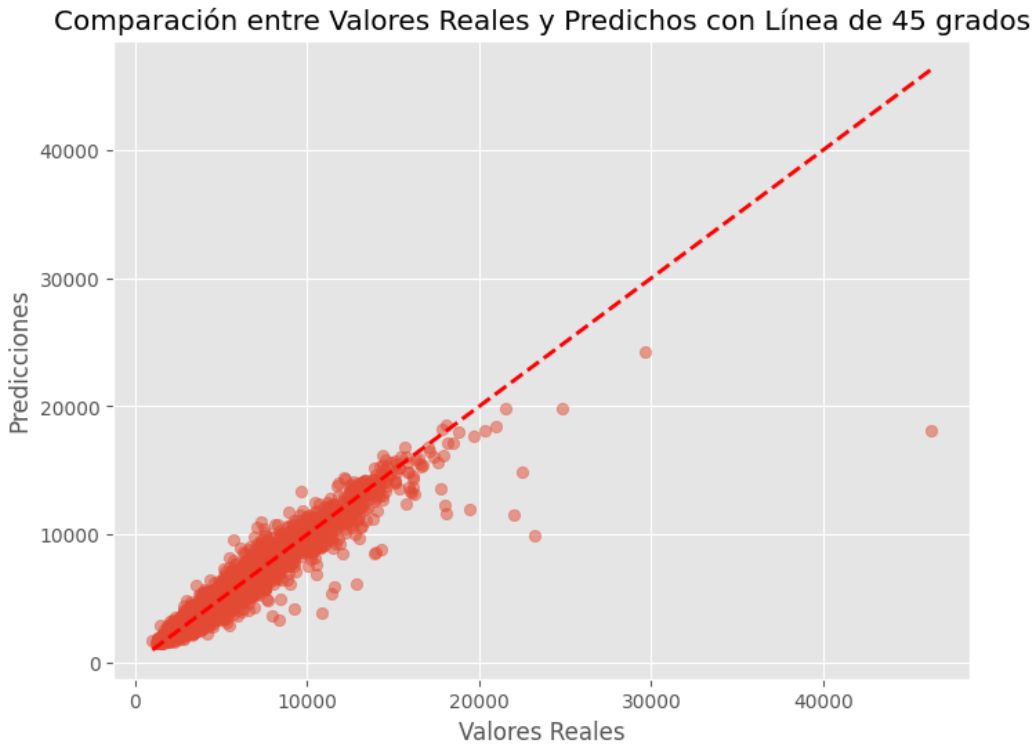

Figura 17 Comparación entre valores reales y predichos con línea de 45° - Modelo XGBoost

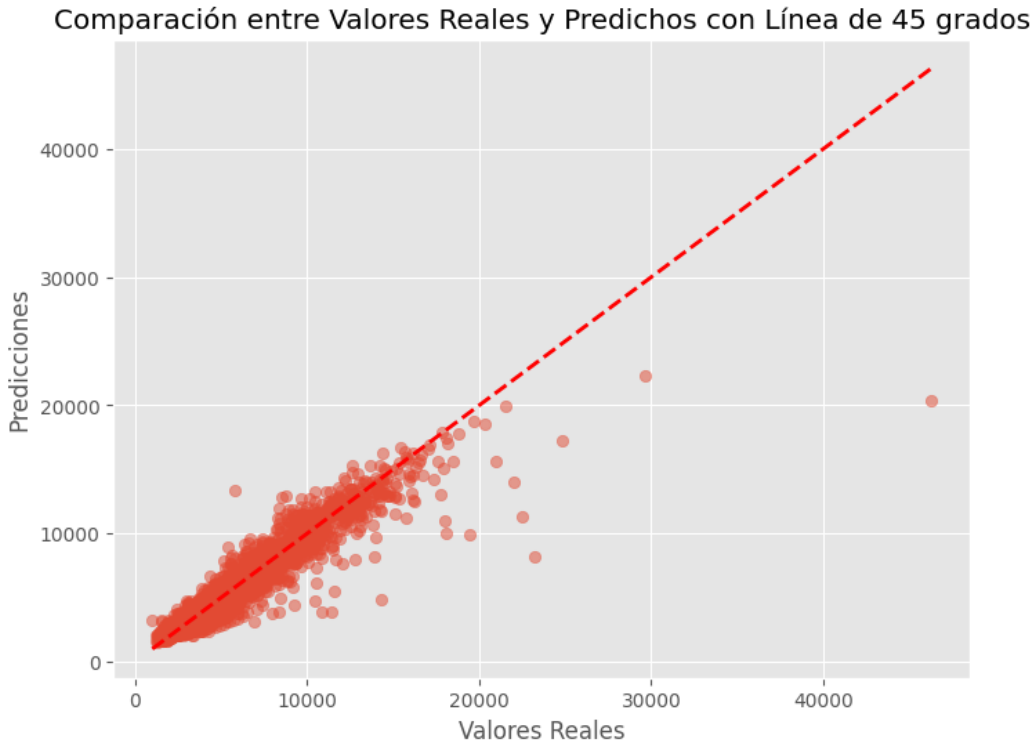

Figura 18 Comparación entre valores reales y predichos con línea de 45° - Modelo XGBoost

6.3. Consideraciones de producción

El modelo de predicción tiene la capacidad de ser aplicado en dos áreas fundamentales en una fase inicial. En primer lugar, puede ser aprovechado por el área de compras de productos de la compañía, permitiendo anticipar las necesidades y proporcionar información valiosa a los proveedores tanto en términos de planificación de compras como de disposición de productos. En segundo lugar, el área logística puede beneficiarse al utilizar el modelo para el desarrollo de planes logísticos, facilitando la gestión eficiente del envío de mercancías a las diversas tiendas. La versatilidad del modelo lo convierte en una herramienta estratégica para optimizar procesos y mejorar la toma de decisiones en ambas áreas.

Adicionalmente para mejor aplicabilidad del modelo propuesto, se propone una conexión completa con servicios en la nube tales como Google Cloud, AWS y Azure donde existan distintos ambientes que garanticen un mínimo viable tal como se menciona a continuación:

1. Conexión a la información de ventas de la compañía a través del ERP, garantizando una integración fluida con streams de datos para la actualización del modelo en tiempo real.
2. Implementación del modelo en un ambiente de producción, garantizando escalabilidad, es decir, implementación de número de clúster adecuados que soporten la carga de trabajo esperada.
3. Implementación de un sistema de monitoreo continuo para supervisar el rendimiento del modelo en producción generando informes de métricas relacionadas al modelo de tal forma que permita tomar medidas preventivas. Para la configuración de métricas se deben definir umbrales y herramientas posibles son CloudWatch de AWS y Azure Monitor de Microsoft Azure
4. Generación de tableros de información que permitan analizar tendencias y realidades no solo de predicciones si no también de las mismas métricas.

7. Conclusiones

Los resultados derivados de las primeras iteraciones posibilitan una comparación minuciosa y un análisis exhaustivo del rendimiento de cada uno de los modelos propuestos. Según se resalta en la **Tabla *4***, emerge claramente que tanto los modelos SARIMAX como XGBoost exhiben los mejores desempeños en comparación con otros en consideración. Resulta significativo destacar que ambos modelos muestran similitudes notables en términos de las métricas evaluadas.

Específicamente, los modelos SARIMAX demuestran un rendimiento destacado, evidenciado por sus bajos valores de Error Cuadrático Medio (MSE) y el índice de Error Cuadrático Medio de Raíz (RMSLE). Al mismo tiempo, el modelo XGBoost también se destaca al presentar resultados comparables en términos de precisión, aunque se observa que sus métricas pueden diferir ligeramente de las de los modelos SARIMAX.

Esta observación resalta la robustez y consistencia de los modelos SARIMAX en la capacidad para ajustarse y prever la serie temporal en cuestión. La elección de estos modelos parece respaldada por su rendimiento superior en comparación con otras alternativas consideradas en las iteraciones. La similitud en los resultados de ambos modelos SARIMAX y XGBoost también sugiere que la elección entre ellos podría depender de consideraciones adicionales o de preferencias específicas, ya que ambos ofrecen un rendimiento competitivo.

La elección entre SARIMAX y XGBoost para llevar a producción debe ser cuidadosamente considerada y alineada con las necesidades específicas de la compañía. Ambos modelos han demostrado ser eficaces en las primeras iteraciones, presentando ventajas similares en términos de rendimiento, pero divergen en aspectos clave, como la interpretabilidad.

En resumen, los análisis comparativos derivados de las primeras iteraciones indican claramente que los modelos SARIMAX y XGBoost destacan como los más eficaces en términos de rendimiento, brindando una base sólida para la elección de estos modelos específicos en la modelización y predicción de la serie temporal bajo investigación.

8. Recomendaciones

Como sugerencia para trabajos futuros, se plantea la generación de modelos de pronóstico adicionales para las demás familias de productos, ya que en este trabajo se aborda únicamente una familia. La individualización de pronósticos por cada categoría permitirá cumplir de manera más precisa con los objetivos estratégicos de la compañía.

En específico, para la categoría de libros, se propone la recolección de una mayor cantidad de información, considerando que este producto se introdujo en 2016. Dada la influencia de eventos externos, como el terremoto en 2016 y la transición gubernamental en 2017, obtener más datos facilitará la elaboración de pronósticos más sólidos y detallados.

Adicionalmente, se destaca la necesidad de analizar el impacto de la pandemia en las ventas. Un enfoque integral debería abordar cómo la compañía se adaptó a estos cambios y cómo evolucionó el comportamiento del consumidor en el contexto de las tiendas retail, tanto durante la pandemia como en la fase post pandemia. Este análisis proporcionará información valiosa para entender las transformaciones en los patrones de consumo y ajustar estrategias comerciales de manera efectiva.

Además, se plantea la necesidad de realizar validaciones adicionales de forma que permita evaluar la sensibilidad de los modelos y su robustez en diferentes escenarios y condiciones además de considerar ajustes adicionales en los hiperparámetros de cada modelo para maximizar su rendimiento.

Referencias

Adhikari, R. and R. Agrawal (2013). An Introductory Study on Time Series Modeling and Forecasting. Lap Lambert Academic Publishing GmbH KG

Alexis Cook, DanB, inversion, Ryan Holbrook. (2021). Store Sales - Time Series Forecasting. Kaggle. https://kaggle.com/competitions/store-sales-time-series-forecasting

Chatterjee, S., & Hadi, A. S. (2006). Regression Analysis by Example. Wiley-Interscience.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. Chapter 7: Model Assessment and Selection.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts. Chapter 2: Time series graphics. Available at: https://otexts.com/fpp2/accuracy.html

Pinedo Chapa, Joely Mireilli. Propuesta de un modelo de pronósticos de demanda y gestión de inventarios para la planeación de demanda en prendas de vestir juvenil. Edu.pe. Recuperado el 21 de abril de 2023, de : https://cutt.ly/IwYgRdyQ https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/623528/Pinedo_CJ.pdf?sequence=5

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.