Impacto del clima en las ventas del comercio electrónico en Brasil

Por: Daniel Arturo Ocampo Guzmán

Universidad de la Sabana
Maestría en Analítica Aplicada

Tutor: Rodolfo Meza Patacón

Octubre, 2024

Contenido

Resumen

Esta tesis analiza la relación entre el clima y las ventas del comercio electrónico en Brasil, un país con una notable diversidad climática. El problema central aborda cómo las fluctuaciones climáticas afectan la demanda de los consumidores, lo cual representa un desafío para las empresas que deben gestionar eficientemente su inventario y logística con base a pronósticos de venta realizados. El objetivo principal es desarrollar y evaluar una herramienta que permita a las empresas de comercio electrónico adoptar un enfoque proactivo en la predicción de ventas, alineando más estrechamente los niveles de inventario con las fluctuaciones previstas en la demanda, predichas a partir de las condiciones climáticas.

Para alcanzar este objetivo, se recopilaron datos de ventas de comercio electrónico y datos meteorológicos regionales, los cuales se analizaron utilizando modelos predictivos avanzados como Random Forest, SVM, y ARIMAX. Los resultados revelan que, aunque las variables climáticas pueden mejorar la precisión en algunos modelos, como Random Forest, estas no son siempre el factor más determinante. El Random Forest con variables exógenas (climáticas) demostró un mejor rendimiento en la predicción de ventas en comparación con su versión sin estas variables, lo que sugiere que el clima, en particular la precipitación, puede influir en la demanda. Sin embargo, otros factores no climáticos, como tendencias del mercado y campañas de marketing, podrían llegar a ser altamente influyentes en las ventas.

Este estudio contribuye a la literatura existente al demostrar que el clima puede influir en la demanda, aunque su impacto varía y debe complementarse con otros factores más significativos para mejorar la predicción de ventas en el contexto del comercio electrónico brasileño.

Introducción

Antecedentes y contexto

El comportamiento del consumidor y el desempeño de las ventas han sido fundamentales para el marketing y el análisis empresarial, centrándose predominantemente en variables internas como la segmentación de clientes(Christy et al., 2021), *customer lifetime value* (CLV)(Gadgil et al., 2023), análisis de sentimiento a partir de revisiones de pedidos(Qorich & El Ouazzani, 2023), análisis de predicción de ventas (Yin & Tao, 2021) y asignación del presupuesto para campañas de marketing (Luzon et al., 2022). A diferencia de dichos estudios, esta tesis explora la relación entre un factor externo que a menudo se pasa por alto (los patrones climáticos estacionales) y las ventas de categorías de productos, con un enfoque en los diversos climas regionales de Brasil y su influencia en el comportamiento del consumidor. Esta investigación tiene como objetivo proporcionar información sobre cómo las empresas pueden utilizar eficazmente los datos relacionados con el clima para optimizar sus estrategias de marketing.

Propósito y significado

Alineado con las prioridades comerciales estratégicas, este estudio busca permitir a las empresas emplear un enfoque proactivo para la predicción de ventas anticipando y respondiendo a las fluctuaciones en la demanda de los consumidores influenciadas por los cambios climáticos. La importancia de esta investigación radica en su potencial para proporcionar evidencia empírica que mejore la comprensión de cómo las condiciones climáticas podrían afectar las decisiones de compra de los consumidores. Este conocimiento podría ser fundamental para las empresas que operan o ingresan al dinámico mercado de comercio electrónico de Brasil, ya que podría conducir a estrategias de marketing más informadas y efectivas, lo que mejora significativamente la eficiencia operativa y la capacidad de respuesta.

Contribución a la literatura existente

Esta tesis amplia la literatura existente al proporcionar información específica sobre los efectos del clima en el comercio electrónico dentro de los climas regionales

únicos de Brasil. Busca profundizar la comprensión del análisis predictivo en las ventas, esto aporta conocimientos valiosos a los campos de la economía meteorológica y la logística empresarial. Al integrar análisis avanzados en la planificación estratégica, la investigación tiene como objetivo equipar a las empresas con las herramientas necesarias para prosperar en el competitivo panorama del comercio electrónico de Brasil, lo que potencialmente conducirá a operaciones más eficientes y una mayor satisfacción del consumidor a través de una mejor prestación de servicios.

Marco conceptual

Importancia de los análisis de marketing

El análisis de marketing es crucial para mejorar el desempeño organizacional, como lo demuestran estudios de investigadores como (Ilmudeen, 2021; Sheth, 2021). Estos estudios destacan cómo la inteligencia empresarial avanzada y el big data pueden transformar el marketing tradicional mediante procesos de toma de decisiones basados en datos, lo que representa avances significativos en el modelado predictivo para la venta de productos, así como en la segmentación de clientes. Se utilizan metodologías como el modelo RFM (recencia, frecuencia y valor monetario) y el algoritmo K-means para mejorar la eficacia del marketing y aumentar la satisfacción del cliente. En general, los análisis de marketing impulsan el crecimiento empresarial al desempeñar un papel fundamental a la hora de comprender el comportamiento del consumidor, optimizar el posicionamiento estratégico en el mercado y fomentar relaciones a largo plazo con los clientes.

Aprendizaje de maquina en análisis de marketing

El aprendizaje de maquina (ML, por sus siglas en inglés) puede ser una herramienta poderosa en el análisis de marketing, ya que ofrece una variedad de formas de mejorar la toma de decisiones y la planificación estratégica. Sus capacidades permiten a los especialistas en marketing obtener conocimientos más profundos sobre el comportamiento del consumidor, optimizar las campañas de marketing y mejorar la participación general del cliente. Algunas formas clave en las que se puede utilizar el aprendizaje automático en el análisis de marketing son:

Segmentación de clientes

Los algoritmos de aprendizaje automático pueden analizar grandes cantidades de datos para identificar distintos grupos dentro de una base de clientes en función de comportamientos, preferencias o características demográficas similares. Permite a las empresas adaptar estrategias de marketing a grupos de consumidores específicos, mejorando la precisión y eficacia de las campañas de *marketing*. Técnicas como la agrupación, entre las que se incluyen *K-means* y la agrupación

jerárquica, se utilizan comúnmente para realizar una segmentación basada en patrones de compra, comportamiento de navegación y otros datos relevantes de los clientes. Christy *et al*. (2021) exploran el modelo RFM combinado con algoritmos *K-means* y *Fuzzy C-means* para segmentar a los clientes en función de comportamientos transaccionales, lo que impacta significativamente las estrategias de *marketing* y la maximización de ingresos.(Christy et al., 2021). La segmentación no sólo proporciona información sobre el comportamiento del consumidor, sino que también ayuda a identificar clientes potenciales, impulsar las ventas y mejorar la retención de clientes.

Análisis predictivo

Uno de los usos más importantes del ML en marketing es el análisis predictivo, que implica pronosticar el comportamiento futuro de los clientes, las tendencias de ventas y los resultados de las campañas de marketing. Las técnicas de modelado predictivo han revolucionado la previsión de ventas, proporcionando a las empresas herramientas para anticipar las demandas del mercado y ajustar las estrategias en consecuencia. Yang (2024) analiza un modelo diseñado para el comercio electrónico transfronterizo que aprovecha *big data* de relevancia controlable para predecir las ventas en mercados de exportación dinámicos (Yang, 2024). De manera similar, Yin y Tao (2021) destacan la integración de la minería de datos y el aprendizaje profundo para pronosticar las ventas en plataformas de comercio electrónico, enfatizando el rendimiento superior de las redes neuronales convolucionales sobre los modelos tradicionales (Gadgil et al., 2023; Yan & Resnick, 2023). Estas metodologías subrayan la transición de la planificación empresarial intuitiva a la predictiva, lo que permite a las empresas mantener una ventaja competitiva en mercados de ritmo rápido.

Predicción del valor de vida del cliente

El ML se puede utilizar para calcular el valor de vida del cliente (CLV, por sus siglas en inglés), que estima el valor total que una empresa puede esperar de una única cuenta de cliente a lo largo del tiempo. Estos modelos predictivos son vitales para

las empresas que buscan centrar sus esfuerzos y recursos en clientes que prometen valor a largo plazo, para así optimizar los presupuestos de marketing y concentrarse en segmentos de clientes de alto rendimiento. Este enfoque estratégico es esencial para garantizar la rentabilidad a largo plazo. Por ejemplo, Gadgil *et al.* (2023) exploran el uso de técnicas avanzadas de meta aprendizaje que mejoran la predicción de CLV aprovechando modelos basados en datos, que mejoran la precisión y confiabilidad de estas predicciones. De manera similar, Yan y Resnick (2023) describen la aplicación de sistemas llave en mano en la previsión CLV, que agilizan la implementación de modelos predictivos y facilitan un rápido despliegue en entornos empresariales. Estos enfoques subrayan la importancia de integrar modelos sofisticados de ML en la planificación estratégica de negocios para maximizar el valor para el cliente de manera efectiva.

Análisis de sentimiento

El procesamiento del lenguaje natural (NLP, por sus siglas en inglés) es un subconjunto del aprendizaje automático, con el cual los especialistas en marketing pueden analizar la opinión del cliente a partir de datos de texto, como publicaciones en redes sociales, reseñas de productos y comunicaciones de atención al cliente. Este análisis ayuda a las empresas a comprender la percepción pública de su marca e identificar áreas de mejora o posibles puntos de crisis. Por ejemplo, el estudio de Qorich y El Ouazzani (2023) demuestra un modelo avanzado de análisis de sentimientos que utiliza redes neuronales convolucionales (CNN, por sus siglas en inglés) e incrustaciones de palabras para clasificar las reseñas de Amazon en sentimientos positivos o negativos. Su enfoque aprovecha el poder del aprendizaje profundo (o *deep learning)* para interpretar mejor las sutilezas de los comentarios de los clientes, y así muestra una mejora significativa en la precisión en comparación con los modelos tradicionales de aprendizaje automático. La inclusión de incrustaciones de palabras permite una mejor comprensión del lenguaje, capturando relaciones contextuales entre palabras, lo que mejora la capacidad del

modelo para discernir con precisión los sentimientos de los datos textuales (Qorich & El Ouazzani, 2023).

Optimización de campañas de marketing

Los algoritmos de aprendizaje automático pueden optimizar los esfuerzos de marketing analizando los resultados de campañas anteriores para determinar qué estrategias generaron el mejor retorno de la inversión. Esto incluye la optimización del canal (decidir en qué canales invertir más), la optimización del contenido (qué tipo de contenido funciona mejor) y la optimización del tiempo (cuándo lanzar campañas). Para mejorar este análisis, Luzon, Pinchover y Khmelnitsky (2022) introducen estrategias dinámicas de asignación de presupuesto para campañas publicitarias en redes sociales, centrándose en una asignación presupuestaria óptima a lo largo del tiempo. Su enfoque considera características únicas del marketing en redes sociales, incluida la relación entre los valores de las ofertas publicitarias y el número de usuarios recientemente expuestos, y desarrolla modelos sofisticados que pueden mejorar significativamente la eficiencia y el impacto de los gastos de campaña (Luzón et al., 2022).

Datos externos para mejorar el análisis

Aunque los datos internos suelen ser suficientes para mejorar el desempeño organizacional, las variables externas como los datos climáticos, los indicadores económicos y las estrategias de los competidores también tienen un impacto. Sin embargo, esto no suele analizarse y la investigación académica ha sido limitada sobre, por ejemplo, ¿cómo las condiciones climáticas afectan el comportamiento de los consumidores?.(Tian et al., 2018).

Análisis meteorológico en el comercio minorista

El análisis meteorológico influye significativamente en las ventas minoristas, afectando varios aspectos, desde los volúmenes de ventas hasta el comportamiento del consumidor. Según Tian, Zhang y Zhang, las condiciones climáticas como la luz solar, la temperatura y la calidad del aire pueden predecir el comportamiento de búsqueda de variedad de los consumidores en sus compras, lo que lleva a cambios

significativos en la demanda de productos durante condiciones climáticas específicas (Tian et al., 2018). Por ejemplo, los patrones de ventas se ven notablemente afectados durante los meses de primavera y verano, el viento es la condición climática más influyente, y las tiendas de las calles principales son más susceptibles a estos efectos (Rose y Dolega, 2022).

La integración de análisis meteorológicos en la gestión minorista no solo ayuda a una mejor gestión del inventario y del personal, sino que también mejora las actividades promocionales al alinearlas con los cambios previstos en el comportamiento del consumidor debido a las variaciones climáticas (Rose y Dolega, 2022).

Análisis meteorológico en agricultura

En el ámbito agrícola, la integración de los análisis meteorológicos influye significativamente en las estrategias de manejo de plagas, especialmente mediante la modulación de la aparición de plagas en diferentes regiones. Como lo detalla Courson *et al*. (2022), las variables climáticas como la temperatura y la humedad del suelo influyen en la distribución regional y la gravedad de los brotes de plagas. Por ejemplo, los inviernos más cálidos y la alta humedad del suelo favorecen la proliferación de babosas, mientras que impactan negativamente a las plagas de colza a través de diferentes mecanismos (Courson et al., 2022).

Este estudio también destaca que la integración de variables del paisaje con análisis meteorológicos mejora la precisión predictiva de la aparición de plagas. Esto sugiere que el manejo de plagas incorpora elementos tanto micro (condiciones climáticas locales) como macro (patrones climáticos regionales) para adaptar las estrategias de control de plagas de manera efectiva. Los resultados demuestran que, si bien las condiciones climáticas afectan directamente la dinámica de las plagas, la composición del paisaje, como la presencia de pastizales o setos, también ayudan en la supresión o mejora de las poblaciones de plagas (Courson et al., 2022).

El análisis meteorológico sirve como una herramienta para comprender y predecir mejor las actividades de plagas en función de las variables climáticas, y también

como un componente crítico en el diseño de prácticas agrícolas específicas de la región, que se alinean con el manejo sostenible de plagas y los objetivos de conservación ecológica.

Análisis meteorológico en el comercio electrónico

La integración del análisis meteorológico en las estrategias de marketing se ha vuelto cada vez más importante, particularmente en sectores como el comercio electrónico, donde las condiciones climáticas pueden afectar sustancialmente el comportamiento del consumidor y los resultados de las ventas. Estudios recientes, como la investigación de Steinker, Hoberg y Thonemann (2017), destacan el pronunciado impacto del clima en la dinámica de ventas en línea, enfatizando el valor de incorporar datos meteorológicos en modelos predictivos para mejorar la precisión en el pronóstico de ventas y la planificación operativa.

Su análisis reveló que variables como el sol, la temperatura y las precipitaciones influyen significativamente en las fluctuaciones diarias de las ventas, especialmente durante los fines de semana y condiciones climáticas extremas. Al integrar los pronósticos meteorológicos en los modelos de predicción de ventas, demostraron una reducción potencial de los errores de pronóstico de hasta un 50,6 % durante los fines de semana de verano, así se optimiza la asignación de fuerza laboral y se reducen los costos operativos en almacenamiento y logística. Esta integración mejora la precisión de los pronósticos de ventas y contribuye a estrategias operativas más eficientes y rentables. (Steinker et al., 2017).

Preguntas de investigación

- Teniendo en cuenta los datos disponibles sobre el clima regional y las actividades de comercio electrónico, ¿cómo influye el clima en las ventas de comercio electrónico en las diversas regiones geográficas de Brasil?
- Teniendo en cuenta los datos históricos meteorológicos y de ventas, ¿con qué precisión pueden los modelos de aprendizaje automático predecir las fluctuaciones estacionales de las ventas dentro del comercio electrónico brasileño?

Problema

Olist conecta pequeñas empresas de todo Brasil con canales de venta sin complicaciones y con un único contrato. Esos comerciantes pueden vender sus productos a través de la tienda Olist y usar sus socios logísticos para enviarlos directamente a los clientes (Kaggle, 2023). Olist, que actúa como intermediario entre vendedores y clientes en Brasil, opera en un mercado donde los patrones climáticos influyen significativamente en los comportamientos de compra de los consumidores. Las investigaciones indican que diversas condiciones climáticas afectan profundamente las ventas minoristas diarias, con impactos que varían significativamente según las diferentes épocas del año. Por ejemplo, el estudio de Rose y Dolega (2022) destaca que las condiciones climáticas como la temperatura, el viento y las precipitaciones pueden alterar las decisiones de compra de los consumidores.

Dada la diversidad del clima de Brasil, que varía desde tropical en el norte hasta templado en el sur, Olist enfrenta desafíos a la hora de predecir las fluctuaciones de la demanda causadas por los cambios climáticos estacionales. Estas fluctuaciones pueden provocar ineficiencias en la cadena de suministro, como exceso o falta de inventario, lo que podría aumentar los costos operativos y afectar la satisfacción del cliente debido a la falta de disponibilidad del producto o retrasos en las entregas. El problema se ve agravado por el papel de Olist como intermediario, que requiere mantener relaciones sólidas tanto con vendedores como con clientes.

Para abordar estos problemas, es fundamental comprender cómo los patrones climáticos estacionales impactan las ventas de comercio electrónico en varias regiones de Brasil. Este estudio busca brindar a empresas como Olist una herramienta que permita predecir las fluctuaciones impulsadas por los cambios climáticos en la demanda de los consumidores. Al predecir con precisión estas tendencias estacionales basándose en datos meteorológicos históricos, Olist puede prepararse mejor para períodos de alta o baja demanda. De esta manera, puede mejorar la eficiencia operativa y la satisfacción del cliente, al tiempo que brinda a sus vendedores un enfoque estratégico del inventario para garantizar una

disponibilidad constante de productos, reduciendo tanto situaciones de exceso como de desabastecimiento.

Objetivos:

Objetivo general

Desarrollar y evaluar una herramienta para empresas de comercio electrónico que facilite la adopción de un enfoque proactivo en la predicción de demanda. Esto permite alinear de manera más eficiente los niveles de existencias con las fluctuaciones de la demanda previstas, tomando en cuenta las variaciones climáticas en Brasil entre los años 2017-2018.

Objetivos específicos

1. Identificar y analizar patrones de ventas:
    a. Trazar y documentar los distintos patrones de ventas y sus variaciones en diferentes regiones geográficas de Brasil.
2. Investigar las correlaciones entre las condiciones climáticas y el comportamiento de ventas del comercio electrónico.
    a. Realizar análisis estadísticos para identificar y cuantificar las correlaciones entre diversas condiciones climáticas (como temperatura, precipitación y humedad) y datos de ventas de comercio electrónico.
3. Desarrollar modelos predictivos para pronosticar fluctuaciones de ventas basadas en datos meteorológicos.
    a. Utilizar técnicas de ML para crear modelos confiables que predigan cambios en el volumen de ventas, utilizando datos meteorológicos históricos y el desempeño de las ventas.

Metodología

Recopilación de datos

Recopilación de datos geográficos y meteorológicos

- Datos meteorológicos: los datos meteorológicos de Brasil se obtuvieron de un conjunto de datos públicos en Kaggle que contiene datos meteorológicos regionales que abarcan desde 2000 hasta 2021 (Holz & Araújo, 2022).
- Datos de ventas de comercio electrónico: los datos de comercio electrónico se obtuvieron de los conjuntos de datos públicos de Olist en Kaggle, que incluyen información sobre productos, pedidos, clientes, datos geográficos, métodos de pago y otros datos relacionados con las ventas entre finales de 2016 y mediados de 2018 (Olist y Sionek, 2018).

Análisis de datos

Identificación del clima estacional y patrones de venta

- Análisis estadístico descriptivo:
    - Clima: utilizar herramientas estadísticas para describir y resumir los patrones climáticos observados, identificando tendencias y anomalías estacionales en varias regiones brasileñas.
    - Comercio electrónico: aplicar herramientas estadísticas similares para describir y resumir los patrones de ventas observados, identificando tendencias estacionales y anomalías en las ventas en las regiones brasileñas.
- Combinación de conjuntos de datos
    - Los datos de ventas y clima se integran en función de la fecha, hora y geolocalización de cada orden.
    - Una vez identificada la estación más cercana, se crea un conjunto de datos que asocia cada orden con su estación meteorológica correspondiente, utilizando los campos Order_id y Station_code.

    - A partir de este conjunto de datos, se vincula la información de ventas y la información climática desde sus respectivos conjuntos de datos.
    - El resultado es un conjunto de datos consolidado que contiene la información completa de cada orden, incluyendo detalles sobre las condiciones climáticas (precipitación, humedad, temperatura) correspondientes al momento en que se realizó la orden.
- Análisis de correlación
    - Técnicas de correlación estadística: aplicar técnicas de correlación para explorar las relaciones entre diferentes condiciones climáticas y volúmenes de ventas en varias categorías de productos.

Desarrollo de modelos

Modelado

- Selección de modelos ML: identificar y seleccionar modelos predictivos apropiados para los datos. Utilizar herramientas de AutoML como NaïveAutoML (Mohr y Wever, 2023) para obtener modelos más complejos que puedan compararse con los modelos seleccionados inicialmente.
- Entrenamiento modelo: entrenar los modelos con datos meteorológicos históricos y de ventas, dividiendo los datos en conjuntos de entrenamiento, validación y prueba.

Validación del modelo

- Evaluación de desempeño: evaluar la precisión predictiva de los modelos utilizando métricas como el error absoluto medio (MAE, por sus siglas en inglés) y el error cuadrático (MSE, por sus siglas en inglés). Comparar el rendimiento entre diferentes modelos.

Implementación

Aplicación de los hallazgos

- Potencial de implementación: evaluar cómo los resultados obtenidos a través de la herramienta pueden servir como base para que las empresas de

comercio electrónico desarrollen, en el futuro, sus propias estrategias de predicción de ventas, adaptadas a las fluctuaciones de la demanda influenciadas por las condiciones climáticas.

- Validación y mejora del modelo predictivo: a partir de los resultados obtenidos, se podrán realizar ajustes y mejoras continuas a la herramienta predictiva para asegurar su precisión y adaptabilidad a diferentes escenarios climáticos y de ventas. Esto permitirá que, en el futuro, empresas puedan incluso integrarla en sus propios sistemas de gestión de inventarios con confianza, sabiendo que está respaldada por un análisis exhaustivo y riguroso.

Documentación e informes

Informe de hallazgos

- Documentación detallada: crear informes y visualizaciones detallados del análisis de datos y los resultados del modelo. Documentar todos los hallazgos, metodologías y el impacto del clima en las ventas en un formato accesible para las partes interesadas.
- Presentación: preparar una presentación para comunicar los hallazgos, metodología e implicaciones al comité académico.

Impacto esperado

Los resultados esperados de esta investigación incluyen:

- Mejora en la gestión de inventarios: la herramienta desarrollada permitirá a las empresas de comercio electrónico ajustar sus niveles de inventario de manera más efectiva según el volumen de ventas predicho, lo que optimizará métricas clave como la tasa de rotación de inventarios, reduciendo tanto el exceso de stock como los niveles de inventario promedio. Esto, a su vez, disminuirá el riesgo de desabastecimiento y ayudará a mantener un porcentaje de productos obsoletos más bajo, mejorando la eficiencia operativa y reduciendo los costos asociados al almacenamiento.

- Incremento en la adaptabilidad ante cambios climáticos: la evaluación de la herramienta proporcionará información sobre cómo las condiciones climáticas afectan la demanda, permitiendo a las empresas ajustar sus estrategias en tiempo real.
- Contribución al conocimiento académico y práctico: los hallazgos de la investigación sobre la efectividad de la herramienta y su impacto en la predicción de ventas podrán servir como base para futuros estudios y desarrollos en contextos similares.

Resultados

Recopilación de datos

Se obtuvieron conjuntos de datos con más de 99 000 pedidos realizados en la tienda de comercio electrónico de Olist entre 2016 y 2018. Los conjuntos de datos se descargaron de Kaggle y se procesaron utilizando Python 3 a través de Visual Studio Code.

Todo el conjunto de datos proviene de una base de datos relacional de proveída por Olist con información de clientes, pedidos, artículos de pedido, productos, vendedores, métodos de pago, revisiones de pedidos y datos de geolocalización. Se incluye también un *schema* con las relaciones entre los conjuntos de datos para facilitar el análisis (Ilustración 1).

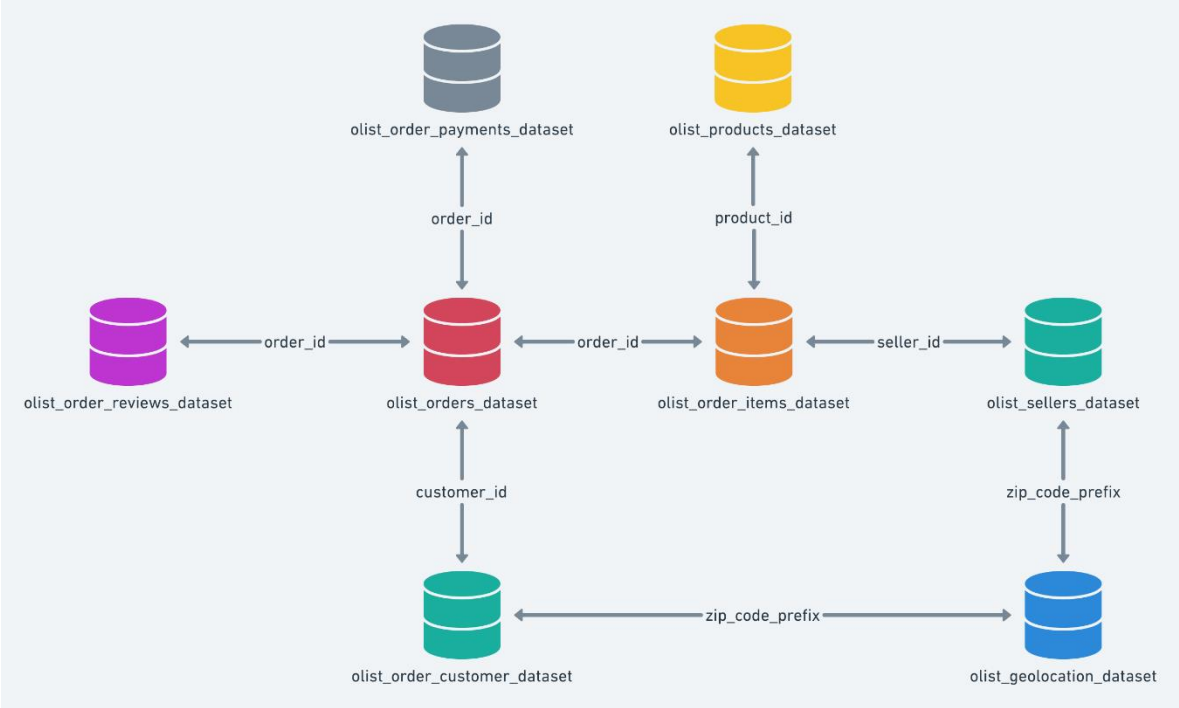


Ilustración 1. Schema base de datos relacional de Olist. Fuente: Kaggle (Olist & Sionek, 2018)

Descripción de los conjuntos de datos:

- **Ordenes:** Conjunto de datos principal. Incluye el detalle de todas las ordenes junto con las llaves utilizadas para conectarse con los demás conjuntos de datos.

Variable	Tipo de Dato	Ejemplo	Cardinalidad
order_id	object	e481f51cbdc54678b7cc49136f2d6af7	99441
customer_id	object	9ef432eb6251297304e76186b10a928d	99441
order_status	object	delivered	8
order_purchase_timestamp	datetime64[ns]	2/10/2017 10:56	98875
order_approved_at	datetime64[ns]	2/10/2017 11:07	90733
order_delivered_carrier_date	datetime64[ns]	4/10/2017 19:55	81018
order_delivered_customer_date	datetime64[ns]	10/10/2017 21:25	95664
order_estimated_delivery_date	datetime64[ns]	18/10/2017 0:00	459

- **Clientes:** contiene información de los clientes y su ubicación, se utiliza para identificar clientes únicos en el conjunto de órdenes. Cada orden genera un ID de cliente y el campo *customer_unique_id* permite identificar clientes recurrentes.

Variable	Tipo de Dato	Ejemplo	Cardinalidad
order_id	object	e481f51cbdc54678b7cc49136f2d6af7	99441
customer_id	object	9ef432eb6251297304e76186b10a928d	99441
order_status	object	delivered	8
order_purchase_timestamp	datetime64[ns]	2/10/2017 10:56	98875
order_approved_at	datetime64[ns]	2/10/2017 11:07	90733
order_delivered_carrier_date	datetime64[ns]	4/10/2017 19:55	81018
order_delivered_customer_date	datetime64[ns]	10/10/2017 21:25	95664
order_estimated_delivery_date	datetime64[ns]	18/10/2017 0:00	459

- **Geolocalización:** contiene información de códigos postales de Brasil y sus coordenadas. Se utiliza para graficar mapas.

Variable	Tipo de Dato	Ejemplo	Cardinalidad
geolocation_zip_code_prefix	int64	1037	19015
geolocation_lat	float64	23.545.621	717360
geolocation_lng	float64	46.639.292	717613
geolocation_city	object	sao paulo	8011
geolocation_state	object	SP	27

- **Artículos de las órdenes:** incluye los datos de los artículos comprados en cada orden

Variable	Tipo de Dato	Ejemplo	Cardinalidad
order_id	object	00010242fe8c5a6d1ba2dd792cb16214	98666
order_item_id	int64	1	21
product_id	object	4244733e06e7ecb4970a6e2683c13e61	32951
seller_id	object	48436dade18ac8b2bce089ec2a041202	3095
shipping_limit_date	object	19/09/2017 9:45	93318
price	float64	58.9	5968
freight_value	float64	13.29	6999

- **Productos:** información de los productos vendidos por Olist.

Variable	Tipo de Dato	Ejemplo	Cardinalidad
product_id	object	1e9e8ef04dbcff4541ed26657ea517e5	32951
product_category_name	object	perfumery	71
product_name_lenght	float64	40.0	66
product_description_lenght	float64	287.0	2960
product_photos_qty	float64	1.0	19
product_weight_g	float64	225.0	2204
product_length_cm	float64	16.0	99
product_height_cm	float64	10.0	102
product_width_cm	float64	14.0	95

- **Vendedores:** contiene la información de los vendedores que han completado órdenes.

Variable	Tipo de Dato	Ejemplo	Cardinalidad
seller_id	object	3442f8959a84dea7ee197c632cb2df15	3095
seller_zip_code_prefix	int64	13023	2246
seller_city	object	campinas	611
seller_state	object	SP	23

- **Clima de Brasil:** Se componen de datos climáticos por hora de más de 600 estaciones meteorológicas del Instituto Meteorológico Nacional (INMET) de Brasil entre los años 2000 y 2021. Los datos se dividen en cinco conjuntos de datos principales, uno por región (norte, noreste, sur, sudeste, centro oeste) y se compone de variables climáticas como: precipitación, humedad,

temperatura, radiación solar, velocidad del viento y altura, entre otras. También incluye las coordenadas de cada estación meteorológica.

Variable	Tipo de Dato	Ejemplo	Unidad	Cardinalidad
precip	float64	0.0	ml	348
at_pres	float64	1013.0	mb	3071
max_pres	float64	1014.0	mb	3069
min_pres	float64	1013.0	mb	3058
rad	float64	2846.0	Kj/m2	5014
temp	float64	28.5	°c	924
dewp	float64	18.4	°c	905
max_temp	float64	29.4	°c	921
min_temp	float64	27.4	°c	921
max_dewp	float64	19.5	°c	918
min_dewp	float64	18.1	°c	893
max_hum	float64	60.0	%	94
min_hum	float64	53.0	%	97
hum	float64	54.0	%	97
wind_dir	float64	300.0	radius degrees (0-360)	361
wind_gust	float64	9.6	m/s	332
wind_speed	float64	5.1	m/s	192
region	object	SE		1
state	object	ES		4
station	object	LINHARES		160
station_code	object	A614		150
lat	float64	-19.356.944		307
lon	float64	-40.068.611		314
Height	float64	40.0		247
date_time	datetime64[ns]	28/09/2017 16:00		37944

Cada conjunto de datos se cargó y procesó utilizando Python 3 a través de Visual Studio Code.

Análisis de datos

Herramientas tecnológicas

El desarrollo de este proyecto se apoyó en diversas herramientas y librerías tecnológicas para facilitar el procesamiento, análisis y modelado de datos. A continuación, se describe el conjunto de herramientas utilizadas, agrupadas por categorías.

- **Lenguaje de Programación**
  - Python 3.12.3

- **Entorno de Desarrollo**
    - Jupyter Notebook
    - Visual Studio Code
- **Librerías y Frameworks**
    - NumPy
    - Pandas
    - SciPy
    - Statsmodels
    - Scikit-learn
    - Matplotlib
    - Seaborn
    - Missingno
    - Holoviews
    - Geoviews
    - Datashader
    - Bokeh

Extracción de conjuntos de datos de ventas

Cada conjunto de datos se cargó como dataframe en Python usando pandas.

Transformación del conjunto de datos de ventas

Para el ejercicio se utilizaron los conjuntos de datos de Ordenes, artículos de las órdenes y productos para obtener solo un conjunto de 'ventas' para el análisis. El conjunto de datos inicial resulto en un *dataframe* de 22 columnas y 113 425 filas. Para el propósito de la investigación, se redujeron las columnas a solo:

- 'order_id': identificador único para cada pedido
- 'customer_id': identificador del cliente
- 'order_purchase_timestamp': hora y fecha del pedido
- 'product_category_name': categoría de producto de los artículos incluidos en el pedido

Esto con el fin de reducir el ruido y descartar columnas que no estaban relacionadas con el propósito de la investigación como: Fecha de aprobación de pedido, fecha de envío, fecha de entrega, precio, costo de envío, fotos, dimensiones y nombre del producto, entre otras.

Exploración de conjuntos de datos de ventas

El conjunto de datos original tiene 70 categorías de productos; sin embargo, varias de ellas pueden considerarse subcategorías que pueden agruparse en una sola categoría. Por ejemplo, están 'fashion_female_clothing, fashion_bags_accessories, fashion_childrens_clothes', 'fashion_male_clothing', 'fashion_shoes', 'fashion_sport' y 'fashion_underwear_beach', todos los cuales se pueden agrupar en una categoría de producto más grande llamada 'Fashion' (moda).

Siguiendo esta lógica, se creó un nuevo mapeo de categorías y el número de categorías de productos se redujo de 70 a 24, lo que facilitará el análisis en el futuro.

Distribución de categorías de productos

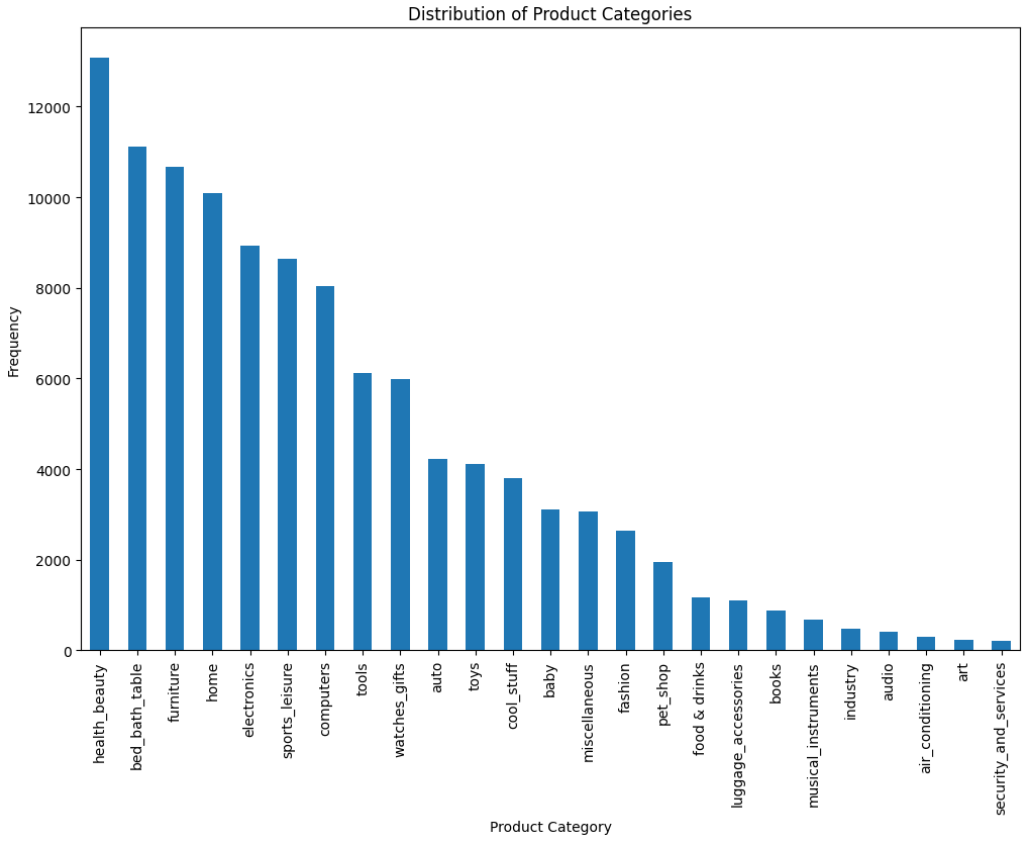


Ilustración 2. Distribución de categorías de productos. Productos vendidos organizados desde la categoría más vendida 'health_beauty' a las menos vendidas 'art' y 'security_and_services'. Fuente: Elaboración Propia

Como se puede ver en la ilustración 2, salud y belleza, cama, baño y mesa, y muebles son las categorías con la mayor cantidad de unidades vendidas. Aunque el modelo inicial se desarrollara agrupando todas las categorías para tener un total de ventas más robusto, se evaluarán estas 3 categorías con el fin de asegurar que el modelo funciona en diferentes granularidades.

Tendencias de ventas

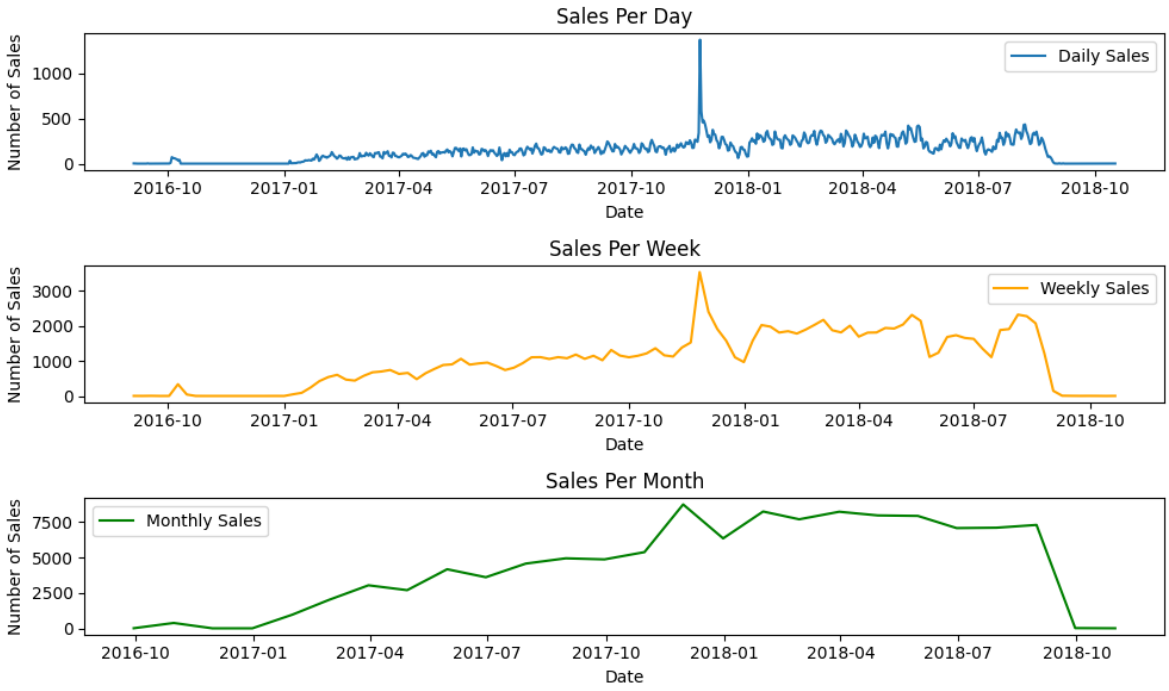


Ilustración 3. Ventas diarias, semanales y mensuales. Fuente: Elaboración Propia

La ilustración 3 contiene información de la tendencia de ventas en granularidad diaria, semanal y mensual. Al revisarla, quedan claras dos cosas:

1. Las ventas anteriores a enero de 2017 y posteriores a agosto de 2018 están incompletas o son inexistentes.
2. Durante noviembre de 2017 hubo un aumento en las ventas, aproximadamente 4 veces mayor que el día anterior.

Debido a esto, se decidió filtrar los datos para incluir valores solo entre enero de 2017 y agosto de 2018, esto permitió una mejor comprensión de las tendencias y el comportamiento del negocio. Tras un análisis más detallado, el aumento en las ventas durante noviembre de 2017 se debió a las ventas del *'Black Friday'* del último fin de semana del mes. Este evento generó un aumento en las ventas y probablemente se habría replicado si los datos también hubieran cubierto todo 2018.

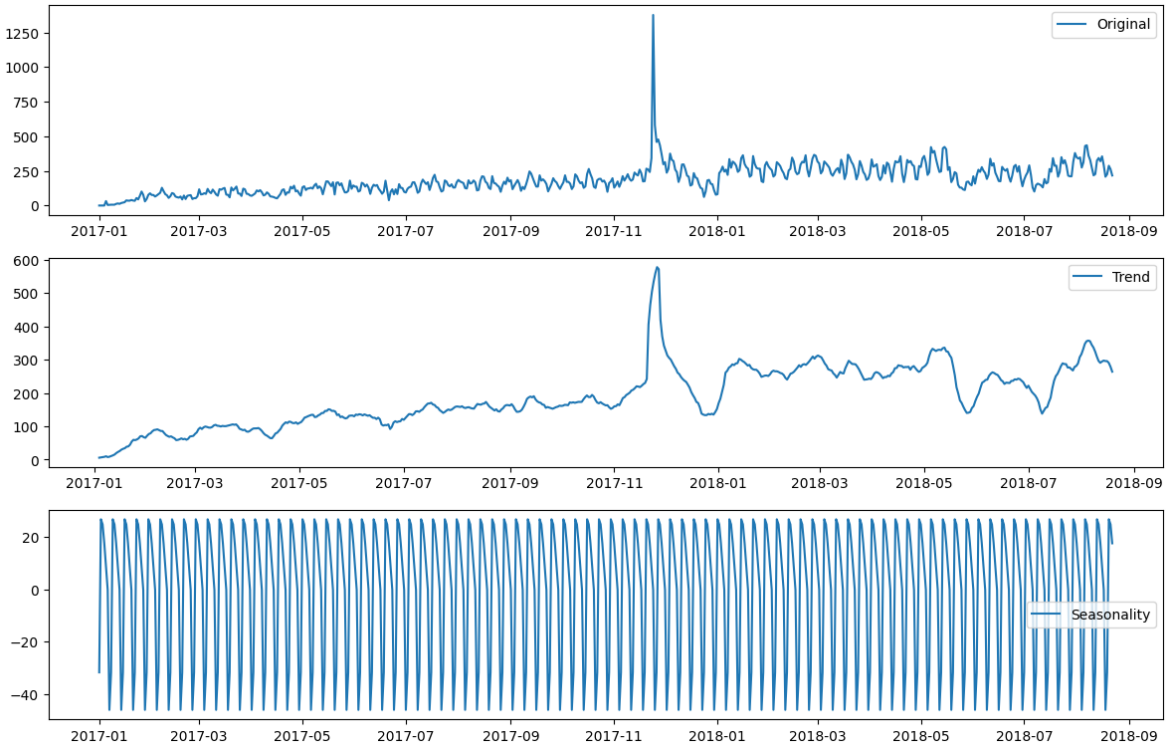


Ilustración 4. Ventas diarias con filtro para remover los valores incompletos como los identificados antes de enero 2017 y posteriores a agosto 2018. Fuente: Elaboración Propia

Después de filtrar el conjunto de datos eliminando los meses que no tenían suficiente información, podemos visualizar una tendencia más clara en las ventas y revisar la estacionalidad de las ventas como se ve en la Ilustración 4.

Estacionalidad

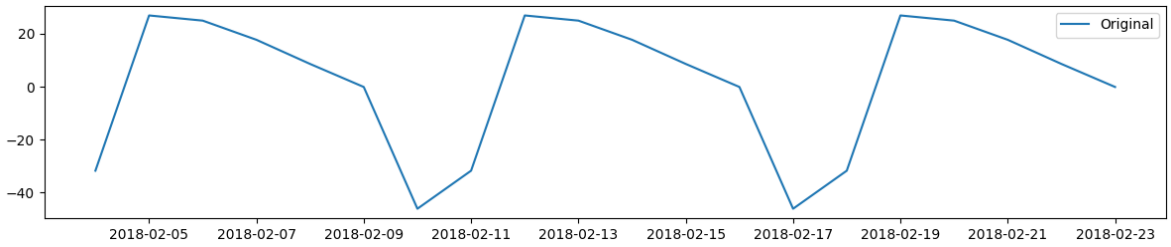


Ilustración 5. Se implementa un filtro que permite ver la gráfica de estacionalidad más de cerca, permitiendo identificar los picos y bajos en la estacionalidad. Fuente: Elaboración Propia

Después de agregar un filtro en el gráfico de Estacionalidad (Ilustración 5), podemos ver que, de hecho, existe estacionalidad en las ventas del negocio de Comercio electrónico. En concreto, existe una estacionalidad semanal, con mayores ventas los lunes, martes y miércoles y menores ventas los fines de semana. Esta estacionalidad probablemente se debe a la naturaleza del negocio, la mayoría de los negocios de comercio electrónico dependen de equipos/empresas de logística que no trabajan los fines de semana, lo que hace que los clientes que realizan pedidos los fines de semana esperen hasta el siguiente día hábil programado para recibir el producto pagado, mientras que los clientes que realicen pedidos los primeros días de la semana recibirán beneficios como la entrega el mismo día. Por otro lado, las tiendas físicas podrían mostrar una estacionalidad inversa, ya que la mayoría de los clientes acuden durante los fines de semana y salen de los días laborables con menores ventas.

Identificación de patrones estacionales con ACF y PACF

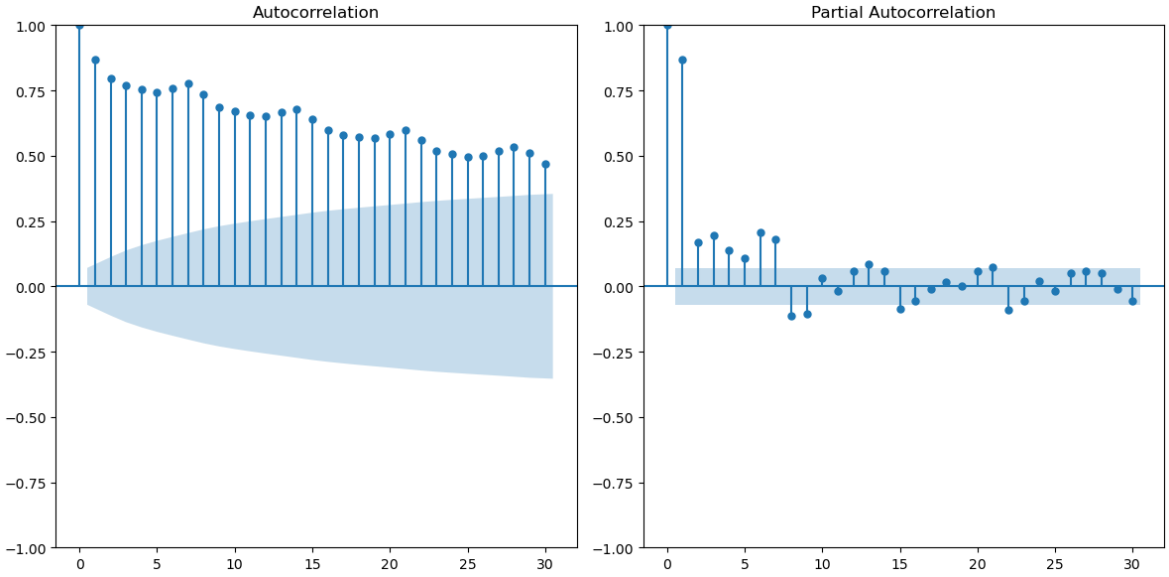


Ilustración 6. Gráficos ACF y PAC. Fuente: Elaboración Propia

ACF: función de autocorrelación, por sus siglas en inglés; PACF: función de autocorrelación parcial, por sus siglas en inglés.

Las gráficas presentadas en la Ilustración 6, muestran las funciones de autocorrelación (ACF, por sus siglas en inglés) y autocorrelación parcial (PACF, por sus siglas en inglés). La ACF muestra la correlación entre sus propios valores retrasados y, en este caso, muestra una caída lenta, lo que podría sugerir que la serie temporal no es estacionaria y podría tener una raíz unitaria. Se requiere una prueba ADF para comprobar si hay una raíz unitaria presente. Si es así, no se puede modelar con ARIMA a menos que esté diferenciado. Por otro lado, la PACF muestra que los primeros 9 rezagos podrían ser estadísticamente relevantes para el análisis.

Prueba Dickey-Fuller Aumentada

La prueba Dickey-Fuller Aumentada (ADF) es una prueba estadística para determinar si una serie de tiempo es estacionaria o no. Las hipótesis por probar en una prueba ADF son:

- Hipótesis nula: existe una raíz unitaria, ARIMA no se puede modelar.
- Hipótesis alternativa: no hay raíz unitaria, se puede modelar ARIMA.

El nivel de significancia se fijó en 0,05

```
ADF Statistic: -2.288260598383136
p-value: 0.17578212009196947
Critical Values:
   1%: -3.439075747702915
   5%: -2.8653910653234655
   10%: -2.568820711931304
```

Ilustración 7. Resultados de la prueba ADF. Fuente: Elaboración Propia

ADF: Dickey-Fuller Aumentada, por sus siglas en inglés.

Con un valor *p* de 0,176 (superior al nivel de significancia de 0,05) en la prueba ADF, no tenemos suficiente evidencia estadística para rechazar la hipótesis nula. Esto significa que no podemos modelar un ARIMA directamente con nuestro conjunto de datos actual, ya que la serie no es estacionaria. En este caso, dado que la estacionalidad ya se verificó anteriormente (Ilustración 5), un SARIMA (ARIMA estacional, por sus siglas en inglés) podría ser una mejor opción. Sin embargo, antes de continuar con el modelo, debemos abordar los valores atípicos en el conjunto de datos.

Manejo de valores atípicos

Como se mencionó anteriormente, en noviembre de 2017 se observó un pico en las ventas, alcanzando aproximadamente 4 veces las ventas máximas promedio. Este dato atípico es originado por variables externas a las que se están investigando. Teniendo esto en cuenta, se utiliza una función basada en el filtro Hampel, que ayuda a detectar y eliminar valores atípicos del conjunto de datos y asegura una corrección adecuada, minimizando el impacto de este dato en las tendencias generales de ventas.

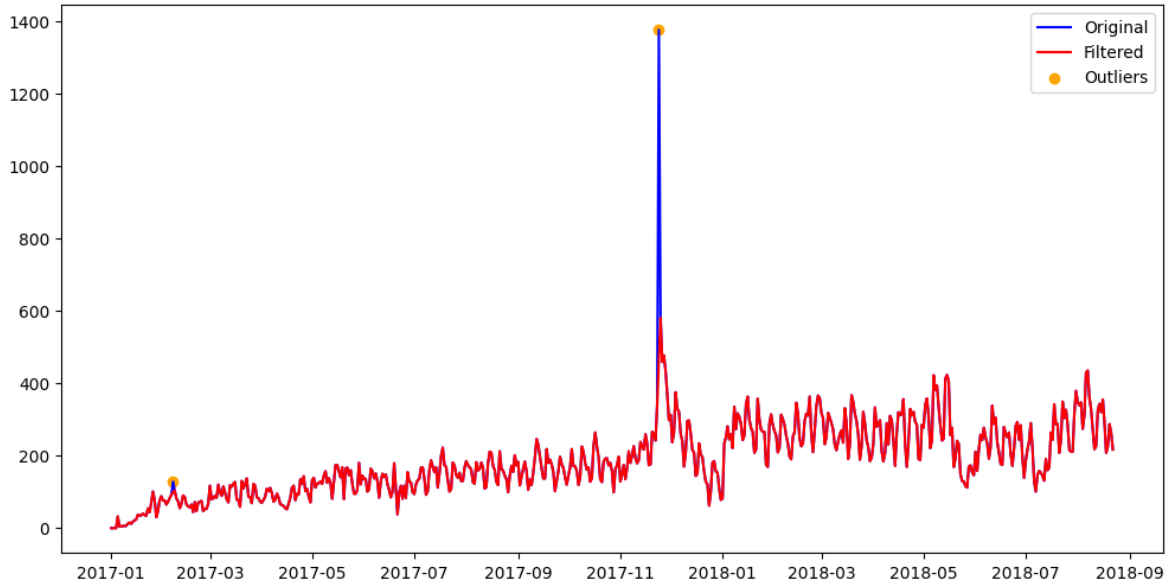


Ilustración 8. Comparación ventas originales (azul) frente a ventas con filtro de hampel (Rojo). Fuente: Elaboración Propia

Modelado SARIMA

El modelo SARIMA se ajusta y prueba utilizando el conjunto de datos filtrado creado mediante el filtro Hampel.

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                       order_id   No. Observations:                  599
Model:            SARIMAX(2, 0, 1)x(1, 0, 1, 7)   Log Likelihood               -2942.454
Date:                          Mon, 30 Sep 2024   AIC                           5896.908
Time:                                  17:28:32   BIC                           5923.279
Sample:                              01-01-2017   HQIC                          5907.174
                                   - 08-22-2018
Covariance Type:                           opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3245      0.158      8.361      0.000       1.014       1.635
ar.L2         -0.3664      0.141     -2.604      0.009      -0.642      -0.091
ma.L1         -0.6037      0.147     -4.116      0.000      -0.891      -0.316
ar.S.L7        0.9953      0.005    199.066      0.000       0.985       1.005
ma.S.L7       -0.8963      0.027    -32.992      0.000      -0.950      -0.843
sigma2      1056.2143     40.503     26.078      0.000     976.831    1135.598
===================================================================================
Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):               713.93
Prob(Q):                              0.79   Prob(JB):                         0.00
Heteroskedasticity (H):               2.79   Skew:                             0.87
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.06
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Ilustración 9. Resultados SARIMA. Fuente: Elaboración Propia

La ilustración 9 muestra los resultados del modelo SARIMA aplicado al conjunto de datos luego de filtrar los datos atípicos.

- Importancia del modelo: los coeficientes del modelo para los términos de auto regresión (AR) y media móvil (MA, por sus siglas en inglés), tanto estacionales como no estacionales, son altamente significativos, lo que indica que estos componentes son cruciales para explicar la varianza en la serie temporal.
- Ajuste del modelo: el Criterio de Información de Akaike (AIC, por sus siglas en inglés) y el Criterio de Información Bayesiano (BIC, por sus siglas en inglés) son herramientas estadísticas que permiten evaluar y comparar la calidad de diferentes modelos. Aunque sus valores no son directamente interpretables, son útiles para comparar modelos, siendo que valores más bajos indican un mejor ajuste.

Análisis residual

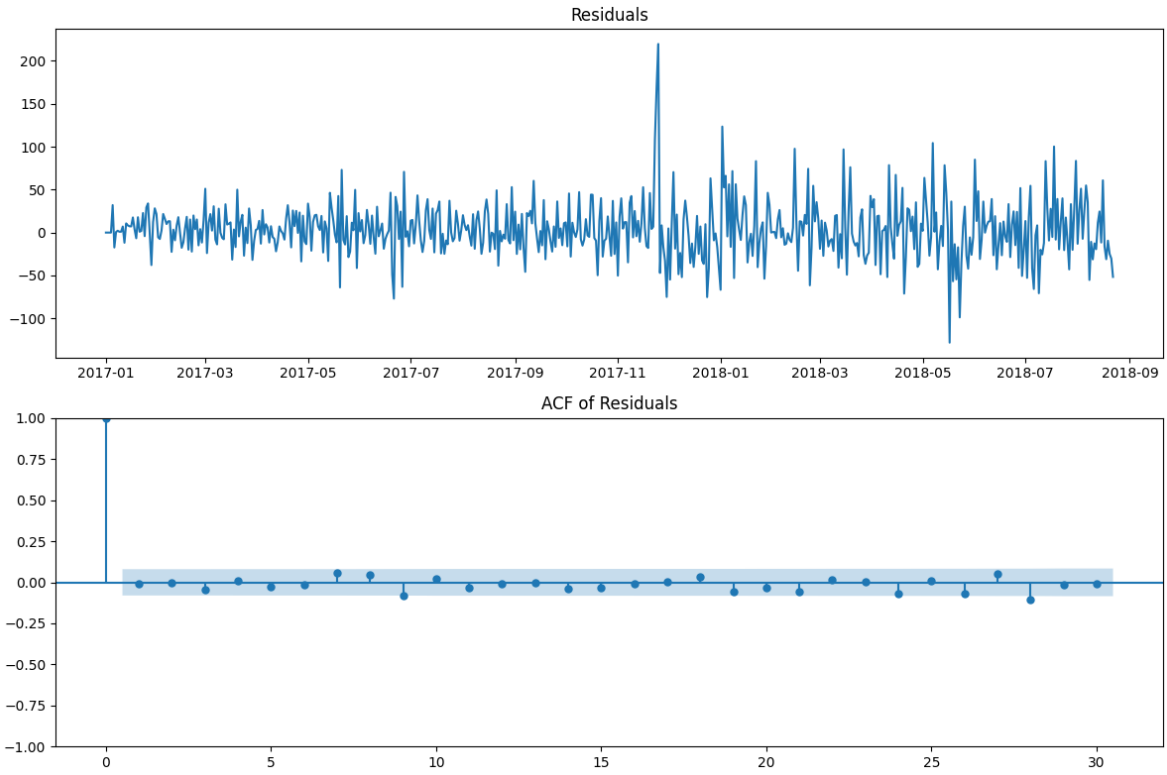


Ilustración 10. Graficas de los residuales del modelo confirman que el modelo se ha ajustado bien. La mayoría de los rezagos (lags) se encuentran dentro del intervalo de confianza. Fuente: Elaboración Propia

La **Ilustración 10** permite visualizar el comportamiento de los residuos del modelo ARIMA y su respectiva prueba ACF para evaluar si están adecuadamente distribuidos. En este caso, podemos confirmar que el modelo se ha ajustado bien debido a que la mayoría de *lags* en el ACF se encuentran dentro del intervalo de confianza.

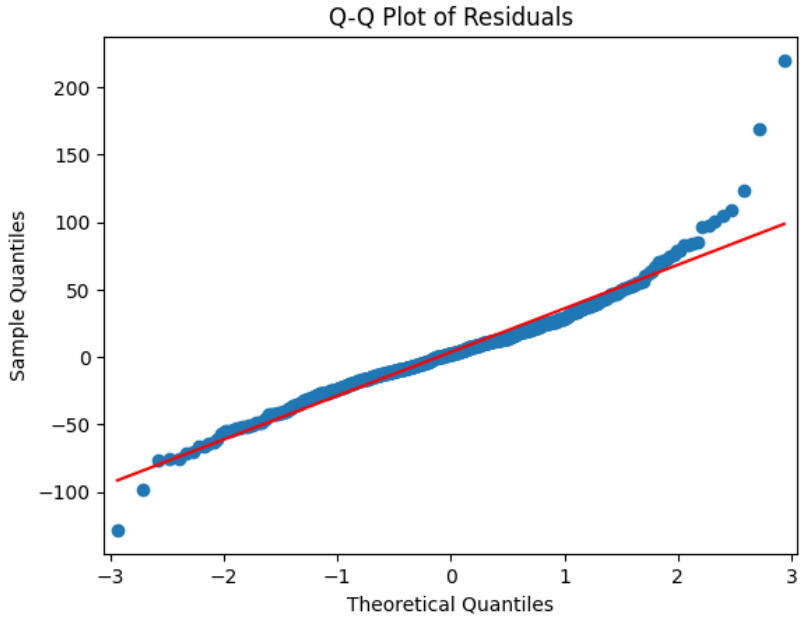


Ilustración 11.Gráfico QQ residuales. Fuente: Elaboración Propia

- Prueba de Ljung-Box: el valor *p* (0,450663) es mucho mayor que 0,05, lo que indica que no podemos rechazar la hipótesis nula. Esto sugiere que no existe una autocorrelación significativa en los residuos hasta el retraso 10. Esta es una buena señal, ya que indica que los residuos son aproximadamente ruido blanco.
- Prueba de Jarque-Bera: el valor *p* es extremadamente pequeño ($1{,}1256 \times 10^{-152}$), mucho menor que 0,05, lo que indica que rechazamos la hipótesis nula de normalidad. Esto sugiere que los residuos no se distribuyen normalmente. El estadístico de prueba alto y el valor *p* bajo indican desviaciones significativas de la normalidad.

Resultado de SARIMA

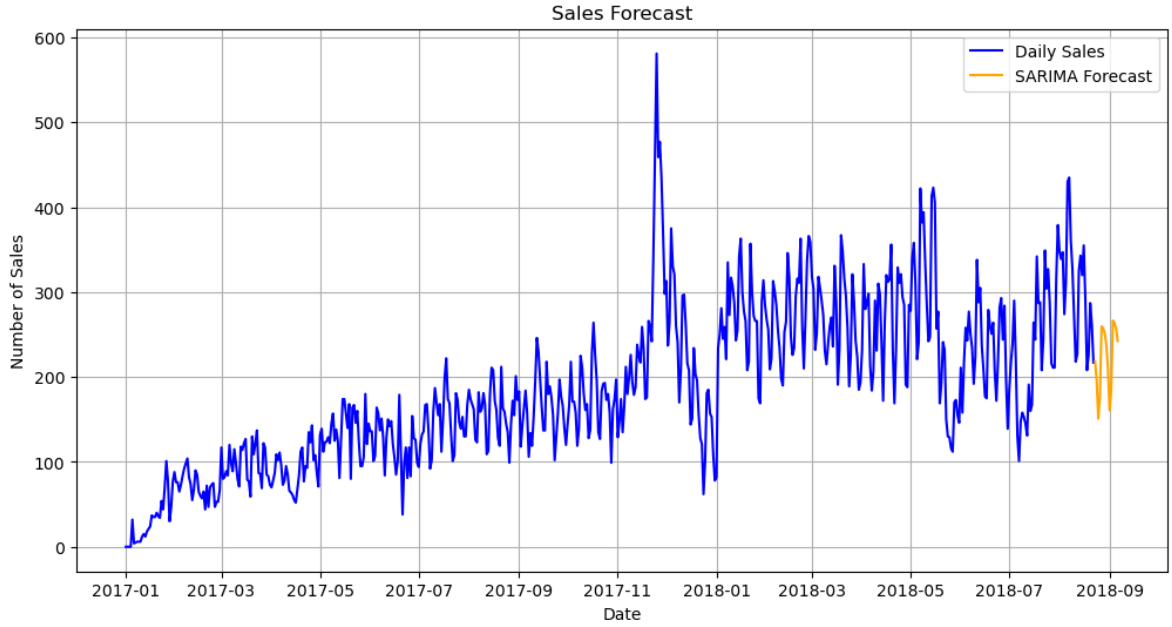


Ilustración 12. Pronostico ventas SARIMA. Fuente: Elaboración Propia

El modelo logró comprender y predecir las ventas futuras en el conjunto de datos, que funciona como modelo de referencia. Sin embargo, a pesar de su capacidad para capturar la estacionalidad y los patrones autorregresivos, el modelo presenta ciertas limitaciones que podrían afectar su capacidad predictiva. Los resultados del test de Jarque-Bera indican que los residuos no siguen una distribución normal, lo que podría influir negativamente en la precisión de las predicciones en casos extremos. Además, la presencia de heterocedasticidad sugiere que las variaciones en los datos no son constantes a lo largo del tiempo, lo que podría reducir la fiabilidad del modelo al predecir datos futuros bajo condiciones de alta volatilidad. Es necesario compararlo con otros modelos de ML para confirmar si es el mejor modelo para predecir las ventas. Para ello, crearemos grupos de prueba y validación y compararemos las capacidades de predicción de los modelos con otros modelos como *Support Vector Machine* (SVM) y Random Forest.

Validación de modelos

Proceso de Validación Cruzada

La validación cruzada se realizó utilizando el enfoque de validación cruzada de series temporales (*Time Series Cross Validation*), dado que los dato involucran una secuencia temporal y es importante respetar su orden cronológico. Para este ejercicio, se utilizó el método *TimeSeriesSplit* de la librería *sklearn.model_selection* y se utilizaron 5 *splits* o subconjuntos que sirven para entrenar los modelos.

Selección de modelos

Los modelos seleccionados son SARIMA, Random Forest y Support Vector Machine.

- SARIMA: este modelo se selecciona porque está diseñado específicamente para manejar series de tiempo que presentan patrones estacionales. Proporciona un modelo base robusto y es una buena referencia inicial para series de tiempo sin influencias externas.
- RANDOM FOREST: Se caracteriza por su capacidad de modelar relaciones no lineales como las encontradas en el conjunto de datos y el manejo de datos multivariados como clima y ventas. En el caso de Random Forest, se ha demostrado que tiene un gran desempeño sobre otros modelos estadísticos y de *ML* a la hora de realizar pronósticos de datos relacionados con series de tiempo (Gaertner, 2024)
- SUPPORT VECTOR MACHINE: Se destaca por su capacidad para manejar problemas de clasificación y regresión, incluso cuando las relaciones entre las variables no son lineales.

Adicional a esto, una vez se integren los conjuntos de datos de clima y ventas, se reemplazará el modelo SARIMA por un modelo SARIMAX, el cual permite la integración de variables externos y se utilizará NAIVE AUTOML (Mohr & Wever, 2023) para generar un modelo adicional que incluya preprocesamiento y ajuste de parámetros más detallado.

Entrenamiento y evaluación de modelos

Para cada subconjunto generado por la validación cruzada se dividen los datos en conjuntos de entrenamiento y de prueba respetando el orden temporal. Los modelos seleccionados se entrenan con los datos de entrenamiento y se generan predicciones con los datos de prueba. Finalmente, se evalúa el rendimiento del modelo utilizando el error cuadrático medio (MSE, por sus siglas en inglés) y el error absoluto medio (MAE, por sus siglas en inglés).

Resultados de los Modelos Seleccionados

Ilustración 13. Resultados de validación de los modelos seleccionados: SARIMA, RANDOM FOREST, SUPPORT VECTOR MACHINE. Fuente: Elaboración Propia

La Ilustración 13 presenta los resultados de las predicciones de los modelos seleccionados. Al comparar las predicciones del modelo SARIMA con los datos reales, se observa que su precisión es inferior a la deseada, lo que se evidencia al calcular el error cuadrático medio (MSE, por sus siglas en inglés) y el error absoluto medio (MAE, por sus siglas en inglés).

Modelo	MSE	MAE
Random Forest	630.49	18.52
SVM	4843.84	52.35
Sarima	8518.27	73.20

Ilustración 14. Puntuación de los modelos

MAE: error absoluto medio, por sus siglas en inglés; MSE: error cuadrático medio, por sus siglas en inglés.

En comparación con SARIMA y SVM, el modelo Random Forest tiene el mejor ajuste para el conjunto de datos de ventas. Esto se demuestra no sólo a través de la visualización de su pronóstico, sino también a través de las métricas de puntuación (MSE y MAE) utilizadas (Ilustración 14). Esto nos da un modelo base de predicción de ventas que será comparado con los modelos que tratarán de predecir las ventas con base al clima.

Análisis geoespacial de ventas

Dentro del conjunto de datos '*Geolocation*' hay 19 015 códigos postales diferentes, cada uno con un promedio de 52,6 coordenadas por código postal. Sin embargo, hay un código postal con 1146 coordenadas, lo cual puede indicar que existen resultados anormales o *outliers*. Para asegurar que todos los datos se encuentren dentro del territorio de Brasil, se filtran los datos en un rectángulo hecho por los límites de Brasil.

Ilustración 15 Distribución de Códigos postales en Brasil. Fuente: Elaboración Propia

La Ilustración 15, muestra el resultado de graficar el mapa de Brasil una vez se excluyen las coordenadas que no pertenecen a Brasil.

Ingresos por Pedidos

A la información geoespacial se le incluyen los resultados de las ventas para ubicar la distribución de los pedidos alrededor de todo Brasil. Se calcula el ingreso por pedidos y se grafican según su ubicación geográfica.

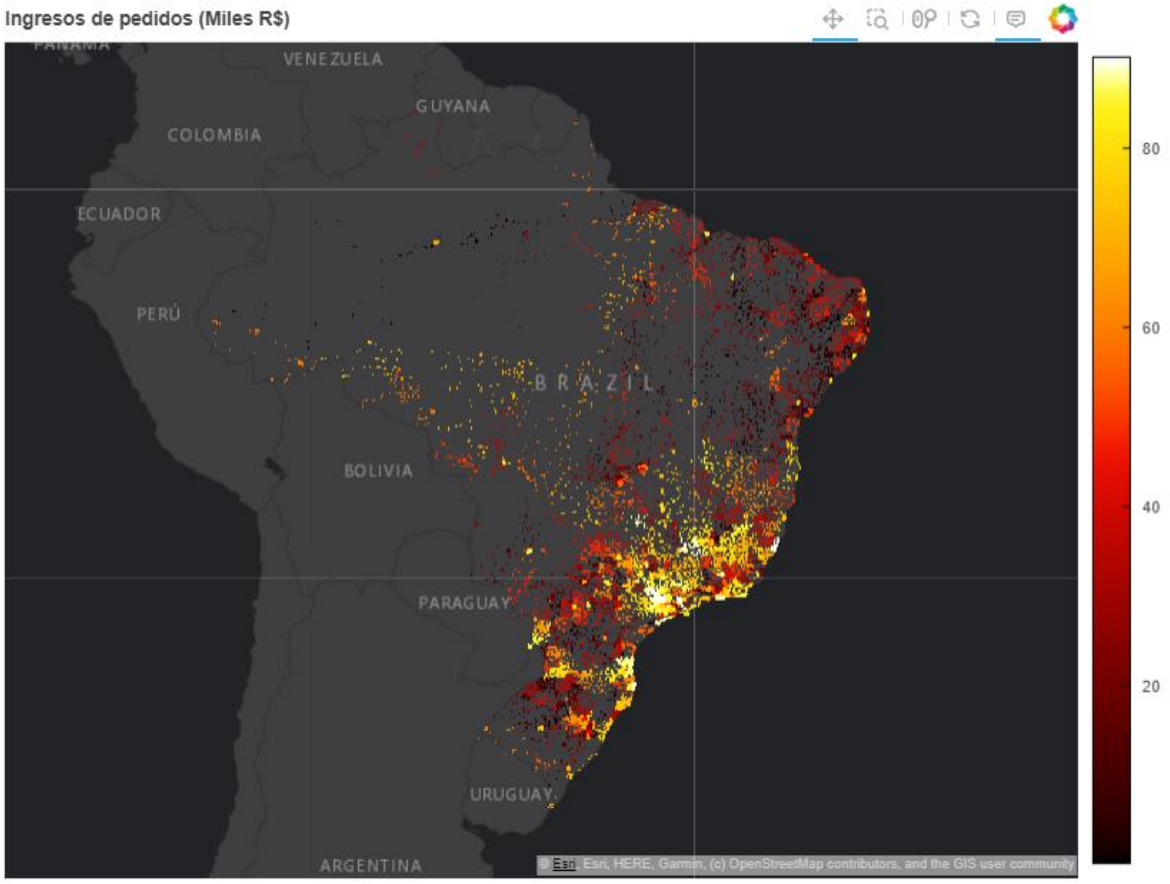


Ilustración 16. Distribución geográfica de los ingresos (en miles de reales). Fuente: Elaboración Propia

La Ilustración 16 muestra los ingresos por pedidos, parece evidente que la mayoría de los pedidos fueron realizados en la región sudeste de Brasil.

Filtro ventas sudeste

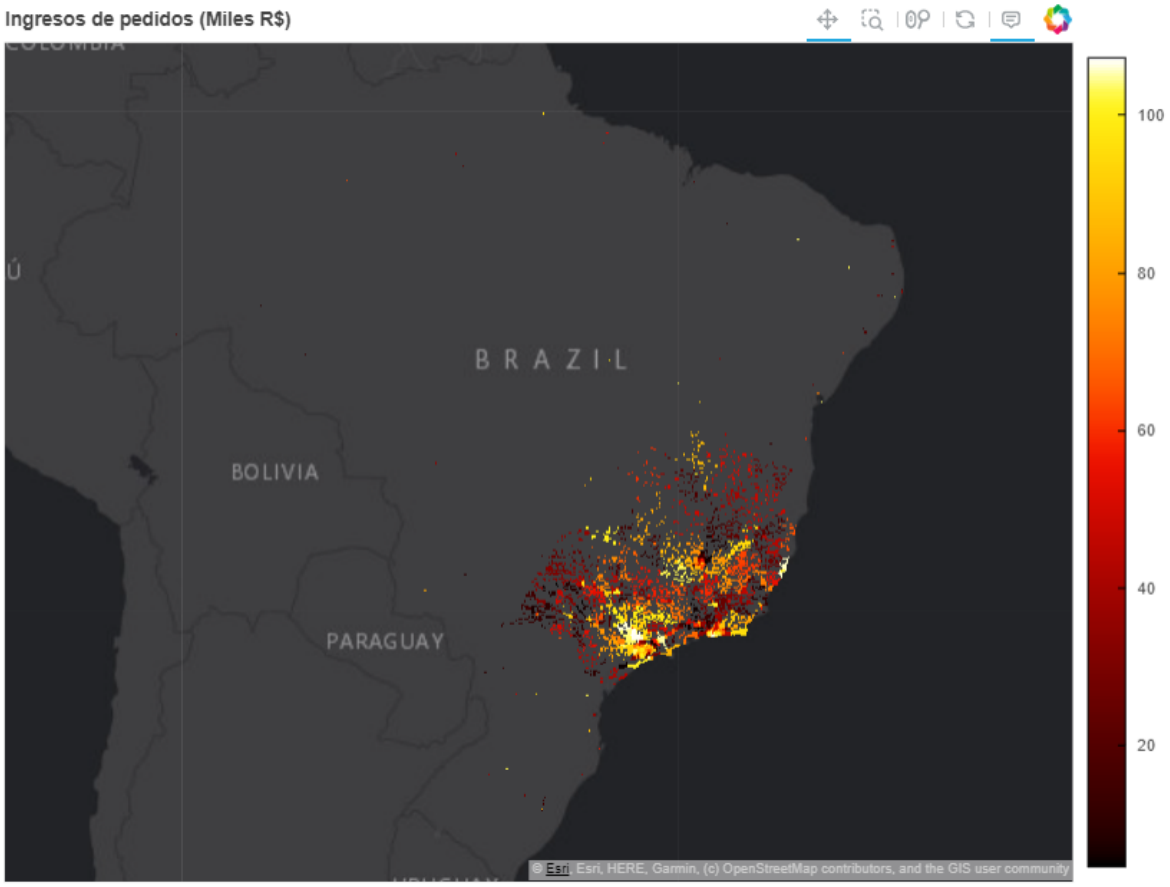


Ilustración 17. Distribución geográfica de los ingresos (en miles de reales) enfocado en la región sudeste de Brasil. Fuente: Elaboración Propia

Al limitar la visualización de las ventas a únicamente la región sudeste de Brasil **(Ilustración 17)**, se identifica que, de las 97 917 órdenes recibidas, 67 171 pertenecen a la región sudeste del país, representando un 68,6 % de las ordenes totales.

Análisis climático

Dado que el núcleo del negocio se encuentra en la región sudeste de Brasil, el análisis climático se centrará en esta área. La región sudeste de Brasil está compuesta por cuatro estados: Espírito Santo, Minas Gerais, Río de Janeiro y São Paulo. El conjunto de datos original incluye millones de mediciones tomadas entre

el año 2000 y 2021, por lo que fue necesario filtrarlo al rango de fechas utilizado en el conjunto de datos de ventas. El conjunto de datos final incluye más de 5 millones de mediciones tomadas entre el 1 de enero de 2017 y el 30 de abril de 2021, registradas en 148 estaciones meteorológicas ubicadas en distintos puntos de la región. Las medidas principales a analizar son la temperatura (°C), precipitación (mm/hora) y la humedad (%).

Temperatura

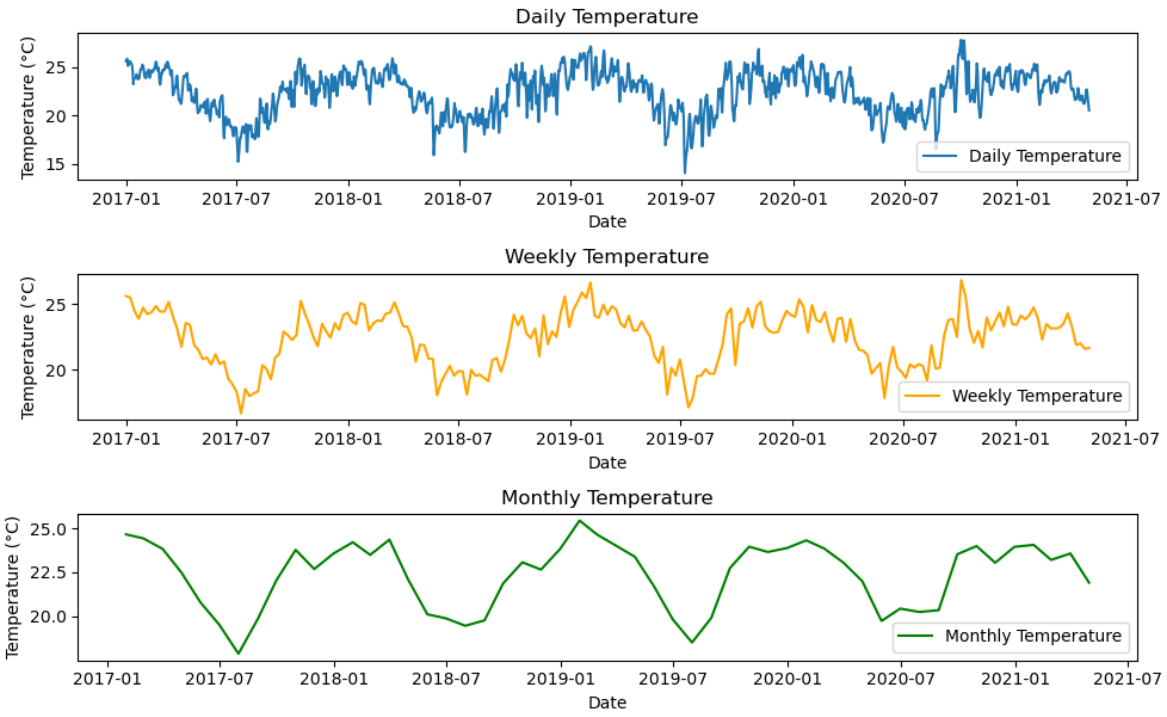


Ilustración 18. Temperatura diaria, semanal y mensual de la región. Fuente: Elaboración Propia

La temperatura media en la región sudeste de Brasil es de 22,4°C, pero el rango de temperaturas varía desde -7,6°C hasta 42,9°C, lo cual indica una amplia variabilidad climática.

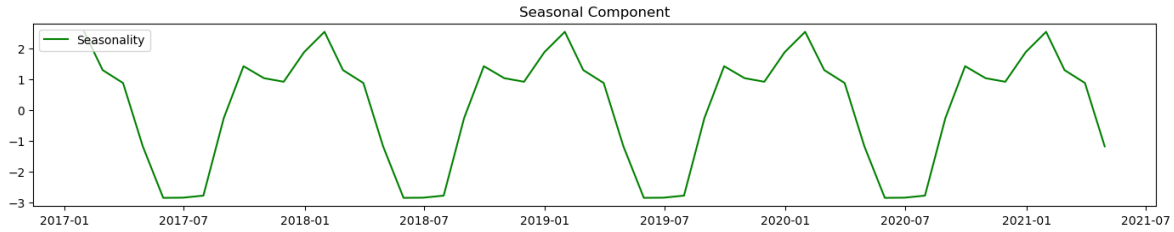


Ilustración 19. acercamiento a la gráfica de descomposición estacional para observar con mayor claridad la estacionalidad

Al descomponer el conjunto de datos y evaluar la estacionalidad **(Ilustración 19)**, se confirma que hay estacionalidad en el clima, lo cual explica la variabilidad del clima junto con las variaciones geográficas entre los diferentes estados.

Precipitación

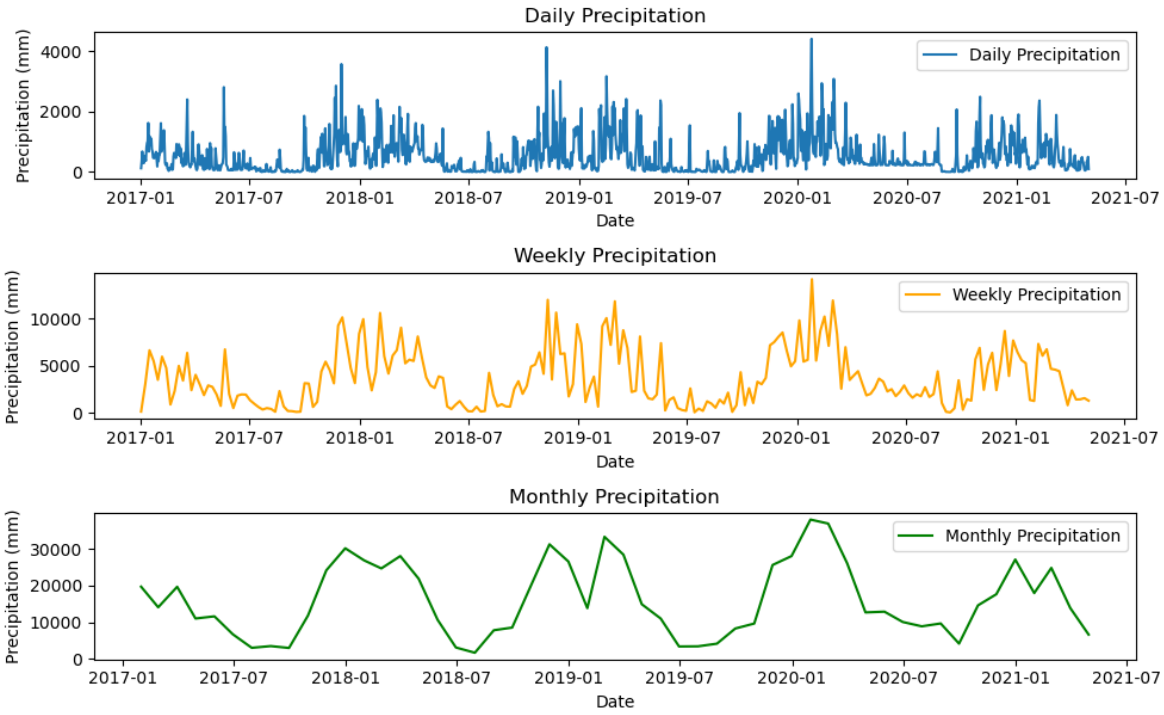


Ilustración 20. Graficas de Precipitación diaria, semanal y mensual. Fuente: Elaboración Propia

Al igual que la temperatura, la precipitación en la región sudeste es estacional y especialmente alta en los meses de verano (diciembre a marzo) donde alcanza

niveles de 30 000 mm al mes (en toda la región). En los meses de invierno (julio a septiembre) la precipitación no suele superar los 10 000 mm al mes.

Humedad

La humedad media de la región es del 71,7 %, lo que sugiere un ambiente generalmente húmedo. Los estados de Espírito Santo y Rio de Janeiro mantienen una humedad promedio entre el 70 y el 80 %, pero los estados de Sao Paulo y Minas Gerais, tienen una mayor variabilidad, llegando incluso a una humedad media del 50 % en los meses de invierno. Esto probablemente se deba a la ubicación geográfica de los estados, ya que Rio de Janeiro y Espirito Santo son estados principalmente costeros mientras que Minas Gerais y Sao Paulo están ubicados principalmente en el interior del país.

Clima General

El clima en la región sudeste generalmente es cálido y húmedo, con variaciones significativas en los principales parámetros climáticos debido a una geografía diversa y estaciones bien definidas a lo largo del año. También se registran eventos extremos, los cuales pueden estar relacionados con fenómenos meteorológicos significativos como tormentas, olas de calor o días muy soleados.

Esto nos permite explorar cómo estos cambios meteorológicos significativos afectan las ventas de un eCommerce como Olist.

Análisis Impacto Clima en Ventas

Hasta el momento se tiene información de más de 60 000 órdenes en la región sudeste de Brasil y una comprensión general del clima. Teniendo estos datos en cuenta es momento de definir ¿cómo influye el clima en las ventas de comercio electrónico en las diversas regiones geográficas de Brasil? ¿con qué precisión pueden los modelos de aprendizaje automático predecir las fluctuaciones estacionales de las ventas dentro del comercio electrónico brasileño?

Preprocesamiento

Los conjuntos de datos se integran identificando la estación meteorológica activa más cercana al lugar donde se realizó cada orden, mediante el cálculo de distancias entre coordenadas geográficas. Posteriormente, se asignan las mediciones climáticas correspondientes a la hora de compra, redondeando está a la hora siguiente. Es decir, si la orden fue a las 10:06 *a.m.*, se toman las mediciones marcadas a las 11:00 *a.m.* debido a que estas cubren el horario de 10:00 a 11:00 *a.m.* Una vez que el conjunto de datos incluye tanto la información de ventas como las condiciones climáticas, los datos se agregan a una granularidad diaria para optimizar su procesamiento. Las órdenes y la precipitación se suman (es decir, el total de órdenes y precipitación por día), mientras que la temperatura y la humedad se promedian. El conjunto de datos final consta de 594 filas y 5 columnas.

Correlación Ventas – Clima

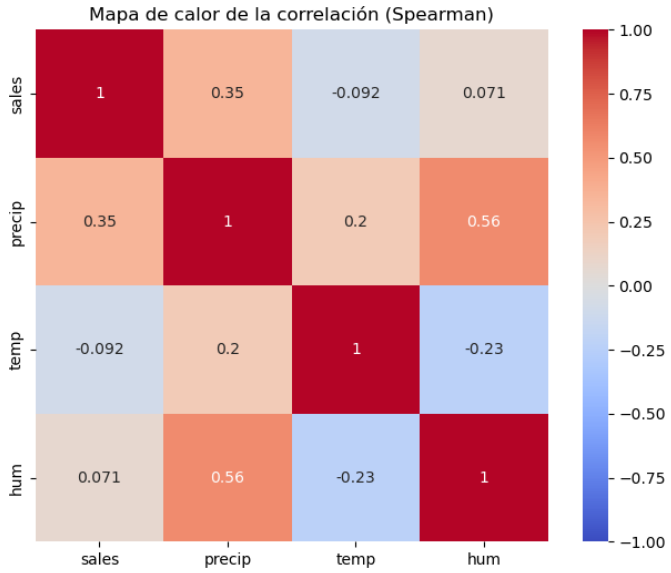
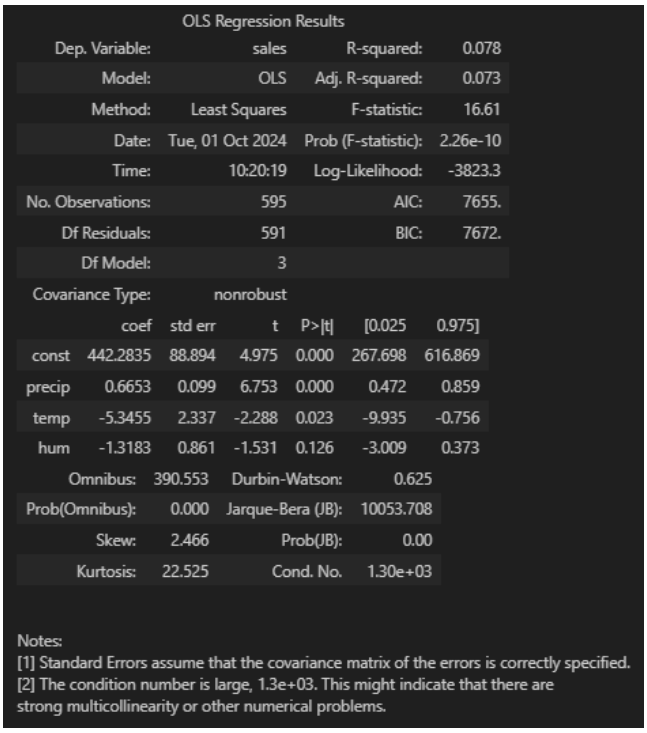


Ilustración 21. Matriz de Correlación de Ventas y variables de clima. Fuente: Elaboración Propia

Para evaluar la correlación entre ventas y clima, se utiliza la correlación de Spearman debido a que las variables no son lineales y su distribución no es normal. Al evaluar la matriz de correlación de Spearman (Ilustración 20) entre las ventas y las variables climáticas, observamos una relación moderada entre las ventas y la precipitación (0,347), mientras que la relación con la temperatura (-0,092) y la humedad (0,071) es muy débil. Esto sugiere que, aunque la precipitación tiene un cierto impacto en las ventas, no existe una relación fuerte con la temperatura o la humedad. Estos resultados indican que otros factores, más allá del clima, podrían ser más determinantes para las ventas.

Prueba de Significancia



OLS Regression Results

Dep. Variable:	sales	R-squared:	0.078
Model:	OLS	Adj. R-squared:	0.073
Method:	Least Squares	F-statistic:	16.61
Date:	Tue, 01 Oct 2024	Prob (F-statistic):	2.26e-10
Time:	10:20:19	Log-Likelihood:	-3823.3
No. Observations:	595	AIC:	7655.
Df Residuals:	591	BIC:	7672.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	442.2835	88.894	4.975	0.000	267.698	616.869
precip	0.6653	0.099	6.753	0.000	0.472	0.859
temp	-5.3455	2.337	-2.288	0.023	-9.935	-0.756
hum	-1.3183	0.861	-1.531	0.126	-3.009	0.373

Omnibus:	390.553	Durbin-Watson:	0.625
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10053.708
Skew:	2.466	Prob(JB):	0.00
Kurtosis:	22.525	Cond. No.	1.30e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.3e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Ilustración 22. Resultado prueba de significancia. Fuente: Elaboración Propia

Se utilizó la regresión por mínimos cuadrados ordinarios (OLS, por sus siglas en inglés) **(Ilustración 22)** para modelar la relación entre las ventas y tres variables climáticas: precipitación, temperatura y humedad. El modelo presentó un coeficiente de determinación R^2 del 7,8 %, lo que indica que las variables climáticas no logran capturar adecuadamente las variaciones en las ventas. De los resultados obtenidos, los coeficientes de la precipitación y la temperatura fueron estadísticamente significativos, mostrando un efecto positivo en las ventas por cada unidad de incremento en la precipitación y un efecto negativo en función del aumento de la temperatura. Por otro lado, la humedad no alcanzó significancia estadística. Además, se identificaron problemas en los residuos del modelo, tales como una autocorrelación positiva considerable, con un estadístico de Durbin-Watson de 0,625 y un *p*-valor extremadamente bajo ($5{,}17\times10^{-294}$) en la prueba de Ljung-Box, lo que sugiere violaciones de los supuestos de homocedasticidad. La prueba de

Breusch-Pagan también indicó un *p*-valor cercano a 0,061, sugiriendo heterocedasticidad, aunque los factores de inflación de la varianza (VIF, por sus siglas en inglés) descartaron la multicolinealidad significativa entre las variables explicativas.

Exploración de modelos alternativos

Debido a la autocorrelación y a los indicios de heterocedasticidad en los residuos, es fundamental evaluar modelos alternativos que aborden estos problemas. En este contexto, el modelo ARIMAX, que integra variables climáticas, se considera apropiado para corregir la autocorrelación y la heterocedasticidad identificadas en el análisis previo. De igual manera, otros enfoques podrían revelar tendencias o comportamientos no lineales en la relación entre el clima y las ventas. Para explorar esta posibilidad, se seleccionaron y evaluaron varios modelos adicionales, como Random Forest, SVR y un pipeline recomendado por herramientas de AutoML. Además, se incorporaron rezagos de 1, 3, 7, 15 y 30 días para todas las variables, junto con sus respectivas medias móviles. Estos rezagos permiten analizar si los efectos de las condiciones climáticas no son inmediatos, sino que se manifiestan con cierto retraso, proporcionando una visión más integral de la relación entre clima y comportamiento de ventas. Los resultados de estos modelos se compararon utilizando el MAE y el MSE, ya que ambas métricas ofrecen una medida sólida de la precisión del modelo: el MAE por su interpretación directa de las desviaciones promedio y el MSE por su capacidad para penalizar errores grandes, lo que permite una evaluación más rigurosa.

ARIMAX

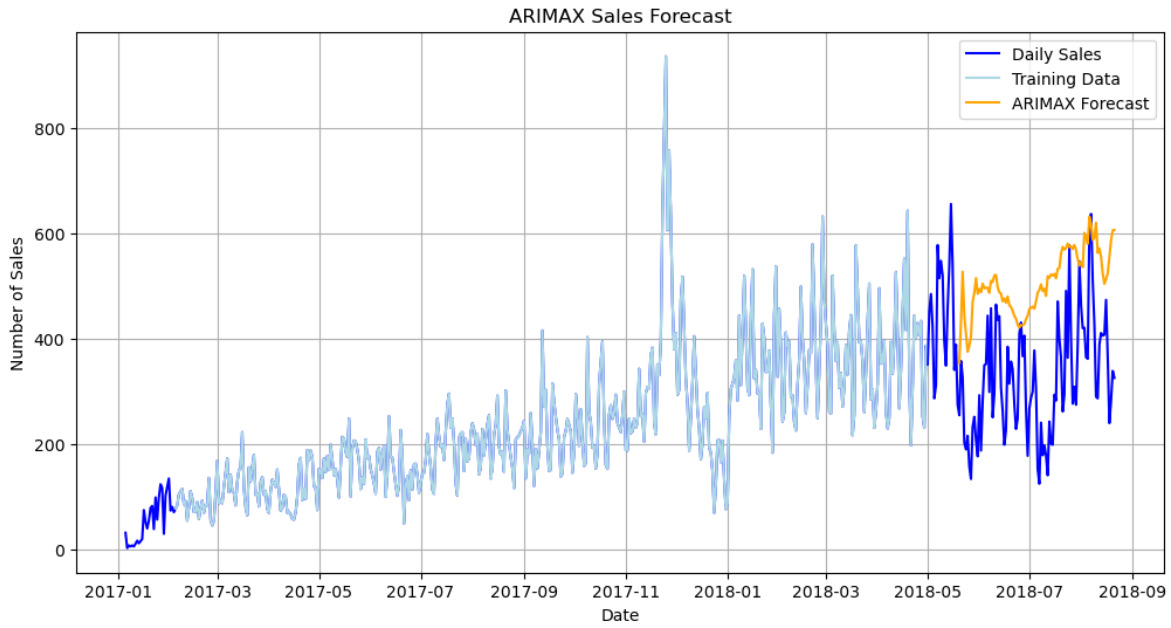


Ilustración 23. Gráfico con predicciones hechas por modelo ARIMAX. Fuente: Elaboración Propia

El modelo SARIMAX aplicado a la variable de ventas muestra que los coeficientes de los términos AR (autorregresivos) son altamente significativos, lo que sugiere que las ventas diarias están fuertemente influenciadas por sus valores pasados hasta en cinco días anteriores. En particular, la precipitación tiene un efecto positivo significativo sobre las ventas, lo que indica que, a mayor precipitación, se observa un incremento en las ventas, mientras que la temperatura y la humedad no resultaron ser factores estadísticamente significativos.

- **Importancia del modelo:** los coeficientes del modelo para los términos AR son altamente significativos, lo que indica que estos componentes son cruciales para explicar la varianza en la serie temporal. La influencia de las condiciones meteorológicas, como la precipitación, también es relevante en la modelización de las ventas.
- **Métricas de error:**
    - MSE: 157,59
    - MAE: 129,68

Estos valores de error muestran que el modelo ARIMAX tiene una precisión moderada, con una desviación significativa entre los valores predichos y los reales, como puede verse en la ilustración 18. El modelo presenta algunos problemas de

diagnóstico, como la no normalidad de los residuos y heterocedasticidad, lo que podría indicar que aún hay espacio para mejorar el ajuste y la predicción del modelo.

Random Forest

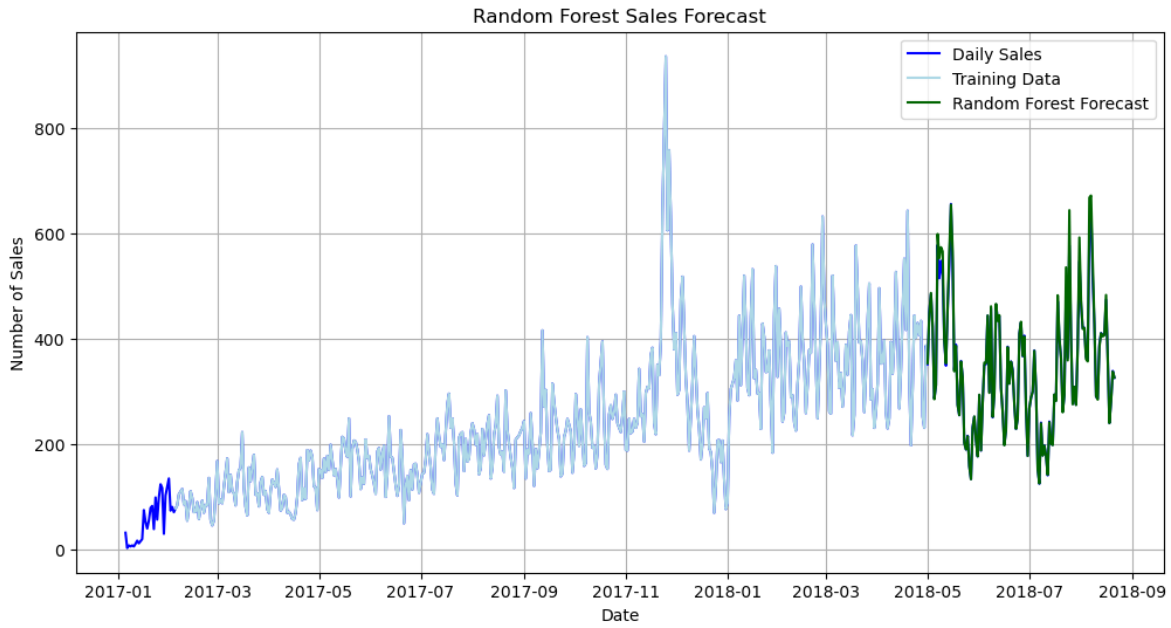


Ilustración 24. Gráfico con predicciones del modelo Random Forest frente a datos originales del conjunto de datos. Fuente: Elaboración Propia

El modelo Random Forest aplicado a la variable de ventas sugiere que las ventas diarias están influenciadas por sus valores pasados y las condiciones meteorológicas. Las características más relevantes suelen ser las ventas pasadas y los promedios móviles de estas, mientras que las variables meteorológicas juegan un papel menos importante.

- **Importancia del modelo:** aunque no ofrece coeficientes estadísticos como el SARIMAX, Random Forest identifica de manera efectiva las relaciones entre las variables y predice las ventas diarias con precisión. Las ventas anteriores son cruciales para el modelo, y las condiciones meteorológicas parecen tener una menor influencia comparativa.
- **Métricas de error:**
    - MSE: 185,50
    - MAE: 5,61

Los valores de error de Random Forest son muy bajos, lo que indica que este modelo es muy preciso en la predicción de ventas diarias. Esto puede evidenciarse

en la Ilustración 23, donde se ve la comparación entre los valores del conjunto de datos original y las predicciones hechas por el modelo.

SVM (Support Vector Machine)

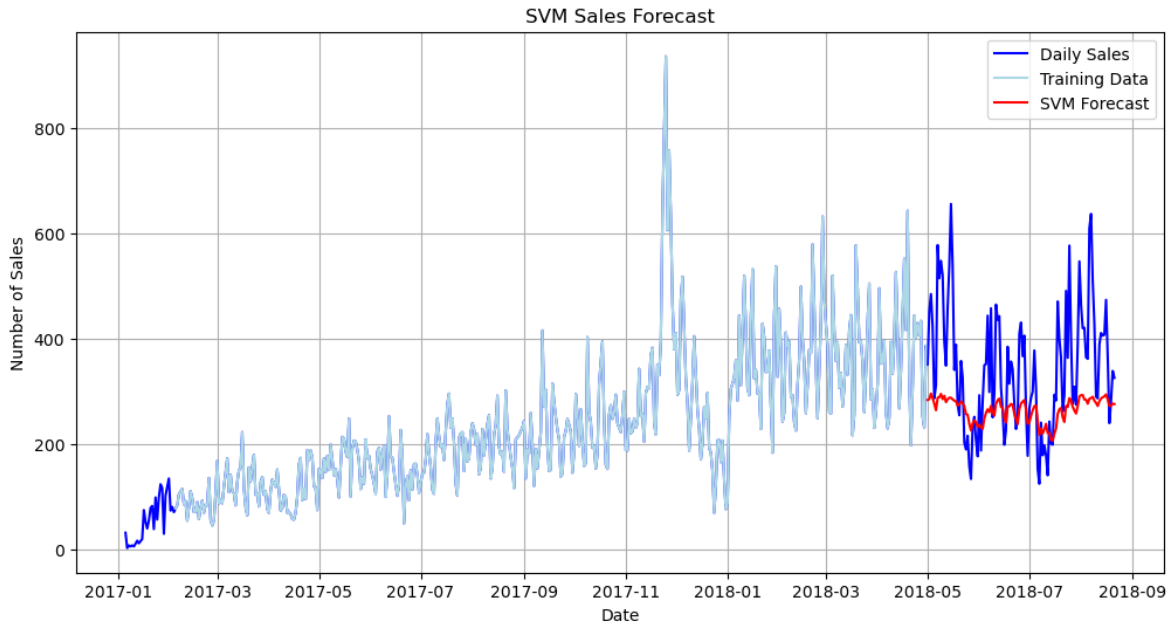


Ilustración 25. Gráfico con predicciones del modelo SVM frente a datos originales del conjunto de datos. Fuente: Elaboración Propia

El modelo SVM aplicado a la predicción de ventas diarias muestra que, si bien es capaz de capturar algunas relaciones no lineales en los datos, tiene dificultades para modelar los patrones temporales autorregresivos presentes en las ventas. Este modelo tiene el peor desempeño entre los probados.

- **Importancia del modelo:** el modelo SVM no asigna directamente importancia a las variables como lo haría un modelo autorregresivo. En este caso, la SVM no fue efectiva para capturar las dinámicas de las ventas en función de sus valores pasados y el clima.
- **Métricas de error:**
    - MSE: 16 139,60
    - MAE: 96,78

Estos resultados muestran que SVM tiene los mayores errores entre los modelos probados, lo que sugiere que SVM podría no ser la mejor opción para series temporales con fuertes componentes autorregresivos, como es el caso de las ventas diarias.

Naive Pipeline (PCA + Extratrees)

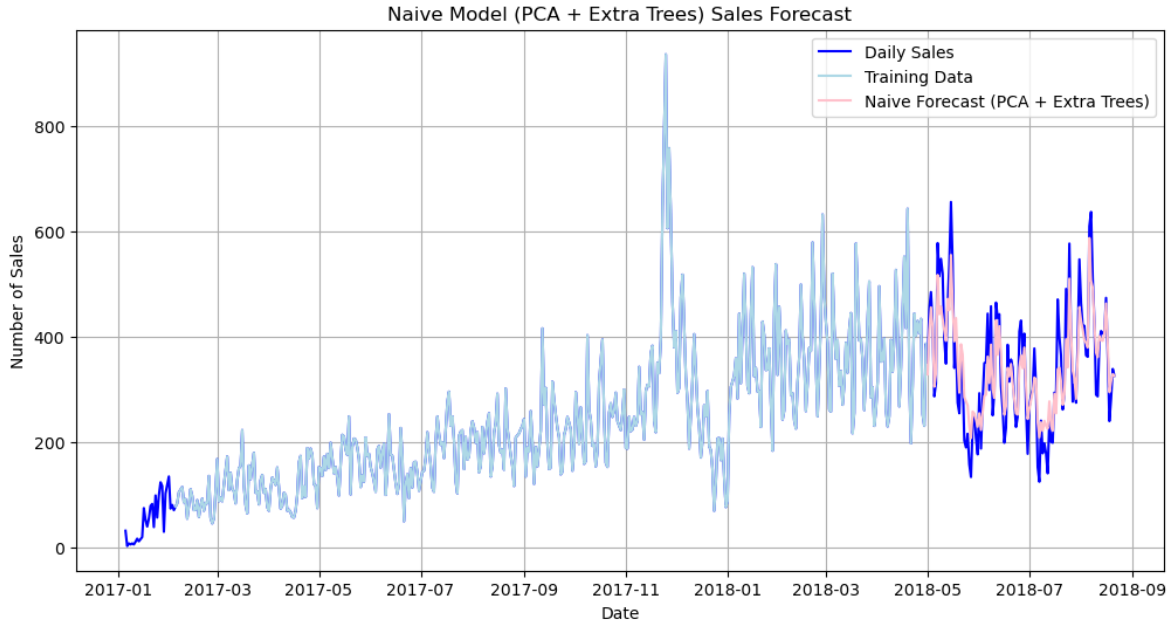


Ilustración 26. Gráfico de predicciones modelo Naive. Fuente: Elaboración Propia

El modelo Naive que utiliza PCA para la reducción de dimensionalidad antes de aplicar el algoritmo Extra Trees muestra que al simplificar el espacio de características se logra una buena precisión en la predicción de las ventas diarias. Este modelo captura de manera efectiva las relaciones más importantes entre las ventas y las variables meteorológicas sin sobrecargar el modelo con información redundante.

- **Importancia del modelo:** el uso de PCA ayuda a reducir el ruido en los datos y enfocar el modelo en los patrones más importantes. Las ventas pasadas y los promedios móviles juegan un papel clave, y aunque las variables meteorológicas están incluidas, su influencia no es tan destacada como en el SARIMAX.
- **Métricas de error:**
    - MSE: 2748,26
    - MAE: 41,94

Aunque el modelo Naive muestra valores de error más bajos que SVM, Random Forest lo supera en términos de precisión. Este modelo ofrece una buena combinación de simplicidad y precisión, y no presenta problemas de diagnóstico significativos como heterocedasticidad o no normalidad de los residuos.

Resultados

	Model	MSE	MAE
0	Random Forest	630.485637	18.524444
1	SVM	4843.835695	52.355335
2	SARIMA	8518.273028	73.197384

Ilustración 27. Resultados modelos ML para venta sin variables climáticas. Fuente: Elaboración Propia

	Model	MSE	MAE
0	Random Forest	185.502763	5.606903
1	Naive Pipeline	2527.789924	40.524115
2	SVM	16139.600887	96.780341
3	ARIMAX	157.589038	129.677153

Ilustración 28. Resultados de modelos ML para ventas con variables climáticas. Fuente: Elaboración Propia

Las Ilustraciones 27 y 28 muestran los resultados MSE y MAE obtenidos en la evaluación de los modelos utilizando la base de datos con y sin variables climáticas, los resultados indican que el modelo Random Forest logra el mejor rendimiento global. Este modelo obtuvo los valores más bajos de MAE y MSE en ambos grupos de datos, lo que sugiere que las ventas pueden ser predichas de manera efectiva utilizando únicamente las características intrínsecas de la serie temporal. Sin embargo, el Random Forest con variables exógenas (climáticas) demostró un desempeño superior en comparación con el mismo modelo sin estas variables.

Este mejor rendimiento se debe a que las variables climáticas, como la precipitación, proporcionan información adicional que complementa los patrones de ventas pasadas. Las condiciones externas, como el clima, pueden tener un impacto significativo en el comportamiento de los consumidores, y al incluir estas variables en el modelo, se reduce la incertidumbre en las predicciones. Por ejemplo, el clima puede influir en la demanda de productos o en la afluencia a tiendas físicas, lo que las variables de ventas pasadas no pueden captar por sí solas. Así, el Random

Forest con variables exógenas logró capturar mejor las fluctuaciones en las ventas, mostrando un menor error de predicción.

El modelo propuesto por Naive AutoML (Mohr & Wever, 2023) también fue destacable, aunque su rendimiento no superó al de Random Forest en el conjunto de datos sin variables climáticas. Esto refuerza la idea de que las ventas pueden ser modeladas de manera efectiva utilizando las características intrínsecas de la serie temporal, pero cuando se incorporan variables exógenas relevantes como el clima, se pueden lograr mejoras adicionales en el rendimiento del modelo.

Finalmente, modelos como SVM y ARIMAX presentaron mayores errores en ambos conjuntos de datos, lo que sugiere que no fueron capaces de capturar adecuadamente las dinámicas temporales o las influencias climáticas en las ventas. En particular, ARIMAX, a pesar de la inclusión de variables climáticas, mostró un MAE significativamente alto, indicando una menor precisión en la predicción de ventas.

Predicciones por Categoría

Con el objetivo de validar que el modelo predictivo seleccionado funcione en diferentes granularidades, se realizan las pruebas de validación en las 3 categorías más vendidas y se comparan con las ventas agrupadas para validar que el modelo final (*Random Forest* con variables exógenas) sea apto para diferentes granularidades del conjunto de ventas.

Resultados modelo de predicción de Ventas

Prueba	Modelo	Métrica	Ventas Agrupadas	Health & beauty	Bed, bath & table	furniture
Predicción Ventas	SARIMA	MSE	8.518,27	180,90	79,50	51,02
		MAE	73,20	11,12	7,41	5,46
	Random Forest	MSE	630,49	38,69	14,55	6,12
		MAE	18,52	4,59	2,74	1,89
	SVM	MSE	4.843,84	147,32	27,83	12,24
		MAE	52,36	9,25	3,90	2,63

Resultados modelo de predicción de Ventas con variables climáticas

Prueba	Modelo	Métrica	Ventas Agrupadas	Health & beauty	Bed, bath & table	furniture
Predicción Ventas con variables climáticas	SARIMAX	MSE	157,59	18,99	19,20	24,11
		MAE	129,68	15,34	15,66	19,02
	Random Forest	MSE	185,50	8,96	0,56	0,39
		MAE	5,61	0,92	0,27	0,20
	SVM	MSE	16.139,60	245,22	129,78	136,50
		MAE	96,78	11,97	8,60	8,66
	Naive Pipeline	MSE	2.556,28	107,81	79,99	67,65
		MAE	40,38	7,98	7,04	6,35

Tanto en el modelo de predicción utilizando solo las ventas, como en el modelo utilizando ventas y variables climáticas, el Random Forest consistentemente supera en desempeño a todos los demás modelos.

Conclusiones

La investigación realizada logra desarrollar y evaluar un modelo predictivo con base en Random Forest con variables climáticas que puede ser utilizado como herramienta para la predicción de ventas. También se ha revelado que las variables climáticas sí pueden ayudar a mejorar la precisión de los modelos predictivos en el comercio electrónico brasileño, especialmente en modelos como el Random Forest, que lograron capturar mejor las fluctuaciones de ventas, tanto totales como por categoría. Se han identificado patrones de ventas y se ha examinado la correlación entre condiciones climáticas y comportamiento de ventas, los resultados muestran que la inclusión de datos climáticos aporta valor a la predicción de fluctuaciones en la demanda, particularmente en ciertos modelos. El Random Forest con variables exógenas demostró un mejor rendimiento que el modelo sin dichas variables, lo que sugiere que las condiciones climáticas, como la precipitación, pueden influir en las ventas.

En relación con la primera pregunta de investigación, se ha observado que, aunque el clima regional puede no ser un factor determinante en todas las circunstancias, sí puede tener una influencia en las ventas de comercio electrónico en Brasil cuando se utiliza un modelo predictivo adecuado, como el Random Forest. Este resultado sugiere que la relación entre las condiciones climáticas y las ventas puede aprovecharse de manera eficaz si se emplean las técnicas de modelado correctas.
Respecto a la segunda pregunta de investigación, los modelos de ML, como Random Forest, lograron predecir las fluctuaciones estacionales de las ventas con alta precisión tanto utilizando datos históricos de ventas como variables climáticas. La inclusión de datos meteorológicos mejoró la precisión del Random Forest, lo que demuestra que las condiciones climáticas pueden ser un factor clave para la planificación empresarial en el comercio electrónico brasileño.

Implicaciones para el pronóstico de ventas

Estos hallazgos sugieren que las condiciones climáticas pueden ser una variable relevante en la predicción de ventas, especialmente en ciertos modelos, pero no

necesariamente en todos los contextos. Además de las variaciones climáticas, otros factores como las tendencias de mercado, campañas de marketing y eventos económicos también pueden tener un impacto significativo en las fluctuaciones estacionales de las ventas. Por lo tanto, para alcanzar el objetivo general de esta investigación – Desarrollar y evaluar una herramienta para empresas de comercio electrónico que facilite la adopción de un enfoque proactivo en la predicción de demanda–, es recomendable que las empresas enfoquen sus herramientas predictivas en una combinación de factores. Esto incluye variables exógenas como el clima, pero sin depender exclusivamente de ellas.

En cuanto a los objetivos específicos:

- **Identificación y análisis de patrones de ventas:** se ha logrado trazar y documentar las variaciones en las ventas en diferentes regiones de Brasil.
- **Correlaciones entre clima y ventas:** se ha realizado un análisis detallado de las correlaciones entre condiciones climáticas y ventas, los resultados han demostrado que estas correlaciones pueden ser útiles cuando se integran adecuadamente en modelos predictivos avanzados, como Random Forest, pero no son determinantes para todos los modelos de predicción de ventas.
- **Desarrollo de modelos predictivos:** se han desarrollado modelos predictivos que muestran alta precisión, especialmente en el caso del Random Forest con variables climáticas. Esto sugiere que los futuros esfuerzos deben continuar explorando la inclusión de datos exógenos relevantes, como el clima, pero también considerar otras fuentes de datos para optimizar los resultados.

Reflexiones finales y futuras direcciones

A pesar de los valiosos hallazgos obtenidos en esta investigación, se reconoce que el proceso podría perfeccionarse mediante una mayor refinación en la selección y el preprocesamiento de las variables climáticas. Dentro del proceso de modelado de clima y ventas, se tomó la decisión de incluir el filtro Hampel para eliminar datos atípicos, de la misma forma en que se hizo al analizar las ventas por sí solas. Sin

embargo, en este proceso no se consideró el impacto que podría tener la eliminación de los atípicos de las ventas mientras se mantenían los valores originales del clima. Por esta razón, es posible que un enfoque más riguroso permitiría integrar de manera más efectiva los datos climáticos en los modelos predictivos, mejorando aún más su precisión. Además, contar con una base de datos más robusta, que incluya factores externos adicionales como campañas de marketing o eventos económicos y cubra un mayor rango de tiempo, podría permitir evaluar el comportamiento del modelo antes situaciones como el Black Friday y otros eventos que generen un aumento en ventas sin necesidad de evaluarlos sino como parte de la estacionalidad natural del modelo.

Es importante incluir también que el mundo y el comportamiento del consumidor ha cambiado significativamente en los 7 años que tienen los datos de ventas, en el año 2020 la pandemia del COVID-19 hizo más relevante que nunca el acceso a compras en línea. Teniendo esto en cuenta, es posible que un análisis similar con ventas más recientes pueda llevarnos a identificar correlaciones más profundas entre las ventas de plataformas de E-commerce y el clima.

De cara al futuro, se alienta a los investigadores a profundizar en esta área, explorando metodologías alternativas y ampliando los conjuntos de datos con nuevas fuentes de información. Estos esfuerzos no solo podrían desentrañar relaciones más complejas entre el clima y el comportamiento de las ventas, sino también identificar otros factores clave que optimicen aún más los modelos predictivos. Esto sería especialmente beneficioso para las empresas que buscan mejorar su capacidad de respuesta ante las fluctuaciones del mercado, permitiéndoles tomar decisiones más informadas y estratégicas.

Referencias

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. https://doi.org/https://doi.org/10.1016/j.jksuci.2018.09.004

Courson, E., Petit, S., Poggi, S., & Ricci, B. (2022). Weather and landscape drivers of the regional level of pest occurrence in arable agriculture: A multi-pest analysis at the French national scale. *Agriculture, Ecosystems & Environment*, *338*, 108105. https://doi.org/https://doi.org/10.1016/j.agee.2022.108105

Gadgil, K., Gill, S. S., & Abdelmoniem, A. M. (2023). A meta-learning based stacked regression approach for customer lifetime value prediction. *Journal of Economy and Technology*, *1*, 197–207. https://doi.org/https://doi.org/10.1016/j.ject.2023.09.001

Gaertner, B. (2024). Geospatial patterns in runoff projections using random forest based forecasting of time-series data for the mid-Atlantic region of the United States. *Science of The Total Environment*, *912*, 169211. https://doi.org/https://doi.org/10.1016/j.scitotenv.2023.169211

Holz, J., & Araújo, A. (2022). *Climate Weather Surface of Brazil - Hourly*. Kaggle.

Ilmudeen, A. (2021). Big data analytics capability and organizational performance measures: The mediating role of business intelligence infrastructure. *Business Information Review*, *38*(4), 183–192. https://doi.org/10.1177/02663821211055321

Luzon, Y., Pinchover, R., & Khmelnitsky, E. (2022). Dynamic budget allocation for social media advertising campaigns: optimization and learning. *European Journal of Operational Research*, *299*(1), 223–234. https://doi.org/https://doi.org/10.1016/j.ejor.2021.08.019

Mohr, F., & Wever, M. (2023). Naive automated machine learning. *Machine Learning*, *112*(4), 1131–1170. https://doi.org/10.1007/s10994-022-06200-0

Olist, & Sionek, A. (2018). *Brazilian E-Commerce Public Conjunto de datos by Olist [Data set]*. Kaggle.

Qorich, M., & El Ouazzani, R. (2023). Text sentiment classification of Amazon reviews using word embeddings and convolutional neural networks. *The Journal of Supercomputing*, *79*(10), 11029–11054. https://doi.org/10.1007/s11227-023-05094-6

Rose, N., & Dolega, L. (2022). It's the Weather: Quantifying the Impact of Weather on Retail Sales. *Applied Spatial Analysis and Policy*, *15*(1), 189–214. https://doi.org/10.1007/s12061-021-09397-0

Sheth, J. (2021). New areas of research in marketing strategy, consumer behavior, and marketing analytics: the future is bright. *Journal of Marketing Theory and Practice*, *29*(1), 3–12. https://doi.org/10.1080/10696679.2020.1860679

Steinker, S., Hoberg, K., & Thonemann, U. W. (2017). The Value of Weather Information for E-Commerce Operations. *Production and Operations Management*, *26*(10), 1854–1874. https://doi.org/10.1111/poms.12721

Tian, J., Zhang, Y., & Zhang, C. (2018). Predicting consumer variety-seeking through weather data analytics. *Electronic Commerce Research and Applications*, *28*, 194–207. https://doi.org/10.1016/j.elerap.2018.02.001

Yan, Y., & Resnick, N. (2023). Correction to: A high-performance turnkey system for customer lifetime value prediction in retail brands. *Quantitative Marketing and Economics*. https://doi.org/10.1007/s11129-023-09275-8

Yang, Y. (2024). A Dynamic Export Product Sales Forecasting Model Based on Controllable Relevance Big Data for Cross-Border E-Commerce. *Applied Mathematics and Nonlinear Sciences*, *9*(1). https://doi.org/doi:10.2478/amns.2023.2.00049

Yin, X., & Tao, X. (2021). Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. *Scientific Programming*, *2021*, 2179692. https://doi.org/10.1155/2021/2179692