Escuela de Negocios
Tipo de documento: Tesis de maestría

Master in Management + Analytics

Forecast de ventas de Gatorade: aplicación de modelos predictivos en la planificación de inventario

Autoría: Mauas, Matías Nadir
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

FORECAST DE VENTAS DE GATORADE: APLICACIÓN DE MODELOS PREDICTIVOS EN LA PLANIFICACIÓN DE INVENTARIO

TESIS

Matias Nadir Mauas

Mayo 2025

Tutor: Federico Favata

Resumen

Anticipar con precisión la demanda es fundamental en el sector de las bebidas deportivas, dado que factores como el clima, las preferencias de los consumidores y la dinámica en los puntos de venta pueden generar fluctuaciones significativas en las ventas. Tener herramientas de predicción eficaces facilita la optimización de la producción del concentrado para Gatorade, la administración eficiente de los inventarios y prevenir inconvenientes de vencimiento o faltantes de mercadería.

Este trabajo investiga diversas técnicas de predicción para calcular la demanda de Gatorade en Argentina, utilizando modelos de aprendizaje automático como Random Forest, XGBoost y LightGBM. Mediante el estudio de datos históricos de ventas, se examinó el impacto de varias variables en la proyección, determinando que los rezagos en las ventas recientes (t-1, t-2, t-3), la estacionalidad anual y el medio de distribución son los elementos más cruciales en el consumo. Se contrastó el rendimiento de estos modelos empleando métricas de error y métodos de optimización de hiperparámetros, lo que posibilitó incrementar la exactitud de las proyecciones.

La investigación no solo proporciona una táctica eficaz para optimizar la organización del inventario de Gatorade, sino que también evidencia la importancia de emplear modelos predictivos en la toma de decisiones de negocio. Detectar las variables fundamentales en la demanda facilita la adaptación de la producción de forma más eficaz, garantizando una distribución más equitativa del producto y optimizando los recursos en toda la cadena de abastecimiento.

Gatorade Sales Forecasting: Application of Predictive Models in Inventory Planning

Abstract

The sports drink market depends on precise demand forecasting since factors like weather, consumer preferences, and point-of-sale dynamics can affect sales. Effective prediction tools facilitate the optimization of Gatorade concentrate production, efficient inventory management, and prevention of expiration issues or stockouts.

This thesis investigates various prediction techniques to estimate Gatorade demand in Argentina, using machine learning models such as Random Forest, XGBoost, and LightGBM. By studying historical sales data, the impact of several variables on the forecast was examined, determining that recent sales lags (t-1, t-2, t-3), annual seasonality, and distribution channel are the most crucial factors in consumption. The performance of these models was compared using error metrics and hyperparameter optimization methods, which made it possible to increase the accuracy of the forecasts.

The research not only provides an effective tactic for optimizing Gatorade's inventory management, but also demonstrates the importance of using predictive models in business decision-making. Identifying key demand variables makes it easier to adapt production more effectively, ensuring more equitable product distribution and optimizing resources throughout the supply chain.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

1.1. Contexto y motivación

En un mercado tan competitivo como el de las bebidas deportivas, anticipar la demanda es clave para alcanzar el éxito y maximizar la producción. Gatorade, marca líder en Argentina y a nivel global en este rubro, enfrenta desafíos esenciales para satisfacer las necesidades del mercado, al mismo tiempo que reduce los costos y previene ineficiencias como excesos de inventario o faltantes de stock. Para poder entender estas necesidades debemos tener en cuenta factores relevantes como la conducta de los puntos de venta y el comportamiento cambiante de los consumidores finales, afectados por los contextos socioeconómicos que atraviesa el país, la estacionalidad y los múltiples elementos que puedan afectar al consumo de bebidas isotónicas. Dado este escenario, las herramientas de pronóstico de ventas son una ayuda indispensable para alinear la producción con la demanda proyectada.

Saber con anticipación cuánta demanda habrá es clave para todas las empresas productoras, y más aún si se trata de una mercadería que tiene vencimiento se debe organizar su producción y distribución de manera eficiente. En el caso de Gatorade en Argentina, su producción de concentrado es un producto fundamental dentro del portafolio de bebidas de PepsiCo, esto es todavía más importante porque el consumo varía según distintos factores. La época del año, la zona geográfica, las ocasiones de consumo y otras condiciones del mercado pueden influir en las ventas, provocando que la planificación sea un gran desafío. Para solventar este problema, me propuse crear un pronóstico acertado que no solo permita evitar los problemas de vencimientos o faltantes de stock, sino que también ayude a mejorar la logística y a garantizar que el producto esté disponible siempre que los consumidores lo necesiten, lo que repercute finalmente en un aumento de la satisfacción del cliente.

Para que una empresa de comercio pueda distribuir y operar de manera eficiente, es fundamental contar con un modelo de pronóstico que permita estimar con precisión cuántos son los bultos de producto que se venderán en cada tienda. Según Ma *et al*. (2018, pp. 5-6), prever la demanda con exactitud es clave para planificar las compras, mejorar la distribución y gestionar de manera más efectiva tanto a la fuerza laboral como los servicios posventa. Además, la capacidad de los gerentes de ventas, para estimar la cantidad probable de las ventas a nivel SKU (Unidad de Mantenimiento de Inventario) y punto de venta en el corto plazo, conduce a una mayor satisfacción del cliente, menor desperdicio, mayores ingresos por ventas y una distribución más efectiva y eficiente.

1.2. Problema

Si bien la tecnología avanza constantemente y a un ritmo cada vez más acelerado, muchas empresas aún enfrentan dificultades para obtener pronósticos precisos. Para el caso en cuestión, las variaciones estacionales y las complejidades propias del sector que enfrenta Gatorade nos lleva a plantear preguntas clave que permiten establecer un marco sólido para diseñar modelos que optimicen tanto la planeación como la operación de la marca en el mercado argentino. Estas complejidades no solo dependen de factores internos de la industria, sino también de elementos como la percepción del consumidor y las estrategias de marketing. De hecho, estudios previos han demostrado que la publicidad y la asociación de una marca con la identidad deportiva pueden influir significativamente en el consumo de bebidas deportivas, impactando así las tendencias de compra y la demanda proyectada (Ellithorpe *et al*. 2023). En este escenario, es crucial preguntarnos cuánto pueden capturar estos efectos los modelos de pronóstico y qué método proporciona la mayor exactitud en este contexto:

- ¿Qué tan precisos pueden ser los modelos de pronóstico al considerar factores como estacionalidad y tendencias?
- ¿Qué enfoques son más efectivos en este contexto, los métodos tradicionales o los modelos basados en aprendizaje automático?
- ¿Cómo pueden los pronósticos contribuir directamente a optimizar la cadena de suministro y la gestión del inventario?

El problema de esta tesis radica en poder desarrollar distintos modelos de pronóstico, con el mayor nivel de precisión posible para finalmente hallar la mejor estrategia que permita estimar el volumen de ventas de Gatorade en Argentina considerando cada región y canal de distribución en Argentina manteniendo su liderazgo. Esto será fundamental para brindar una herramienta que permita ajustar la producción de concentrado de manera anticipada, asegurando que se cumpla con la demanda en cada punto de venta sin generar excesos o faltantes de inventario, lo que beneficia a la presencia de la marca en la mente del consumidor. La idea es utilizar datos históricos mensuales, donde se buscará identificar patrones y tendencias en las ventas que permitan mejorar la proyección del pronóstico de manera ágil y como consecuencia, optimizar la cadena de suministro y la eficiencia operativa.

1.3. Objetivo

El objetivo es poder predecir las ventas con una anticipación de un mes para tener de manera anticipada la producción estimada sin dejar demanda insatisfecha y sin acumular exceso de stock. Para esto se podrá cargar mensualmente el segundo día hábil de cada mes la información de las ventas del mes anterior, de modo que, por ejemplo, cuando inicia octubre, podemos cargar la información de septiembre e intentar predecir la información de noviembre. Incluso podemos realizar una proyección para octubre mismo, pero no sería estrictamente necesario ya que ese mes ya lo habríamos estimado el mes anterior, aunque también podría ser beneficioso para realizar ajustes sobre la marcha.

La proyección se realizó a nivel mensual porque esta frecuencia representa mejor la lógica operativa del negocio. En el mercado, muchos puntos de venta realizan compras mensuales y se abastecen en una sola orden para cubrir su demanda del mes, ya que es más eficiente para sus costos de distribución. Además, los objetivos de venta son anuales y cuentan con una apertura mensual que están alineados a la planificación interna de producción y logística también estructurada por mes calendario, lo que simplifica la toma de decisiones basada en las estimaciones. Por otro lado, esta frecuencia permite captar la estacionalidad sin introducir series semanales, por ejemplo, que pueden verse afectadas por factores aleatorios o campañas puntuales generando ruido y variabilidad de manera innecesaria.

Este método mensual posibilita que la compañía modifique a tiempo los volúmenes de producción y distribución, optimizando la eficiencia en las operaciones y disminuyendo gastos relacionados con inventarios insuficientes. Como resalta Jiang *et al*. (2020), una proyección exacta de la demanda en el sector de las bebidas promueve la mejora de la logística y reduce gastos en la cadena de suministro, lo que subraya la relevancia de disponer de modelos apropiados de previsión.

Para lograr esto, se plantean los siguientes pasos:

- Analizar los patrones históricos de venta y las variables relevantes para las predicciones.
- Implementar modelos de aprendizaje automático, como Random Forest, XGBoost y LightGBM para luego comparar su desempeño con métodos tradicionales.
- Evaluar la precisión de los modelos y proponer estrategias prácticas basadas en los resultados obtenidos.

2. Revisión de la literatura

En el campo del pronóstico de ventas, múltiples investigaciones han analizado técnicas y recursos para incrementar la exactitud en la estimación de la demanda. Estas labores no solo resaltan los desafíos particulares de diversas industrias, sino que también demuestran los beneficios de aplicar técnicas avanzadas En este escenario, es crucial examinar los aportes más significativos para este estudio y comprender cómo pueden ser aplicadas al caso de Gatorade en Argentina. A continuación, se expone un resumen de las contribuciones más valiosas para este estudio.

El análisis de Fildes, Ma y Kolassa (2018) es especialmente relevante para este trabajo, dado que proporciona un repaso detallado sobre la situación actual en el pronóstico de ventas al por menor. Al examinar el efecto de las decisiones de predicción en el ámbito estratégico y operativo, los escritores subrayan la relevancia de elementos como la combinación de marketing, las ofertas y el contenido producido por los usuarios. Su foco en la adaptación de técnicas a diferentes niveles de análisis, desde mercados agregados hasta SKU concretos, proporciona un sólido marco teórico para enfrentar los retos en la organización de inventarios y tácticas de ventas. Estos principios son fundamentales en el estudio de la demanda de Gatorade, ya que la variabilidad en el consumo requiere modelos de predicción exactos y flexibles.

Jiang *et al*. (2020) estudia la manera de implementar modelos sofisticados de predicción de demanda en el sector de las bebidas alcohólicas, llegando a demostrar que técnicas en series de tiempo y en aprendizaje profundo, como las redes neuronales convolucionales, pueden ser superiores en la precisión a los procedimientos convencionales. Con la aplicación de estos modelos se posibilitó disminuir de modo considerable el error medio absoluto (MAE) en la proyección de ventas para clientes particulares, lo que nos indica lo importante que puede ser emplear métodos personalizados para productos con atributos de consumo distintivos. Estos hallazgos son particularmente relevantes para esta tesis, dado que pueden ser modificados para dividir la demanda de Gatorade por zona o medio de distribución, lo que facilitaría la mejora de la exactitud de los pronósticos y la optimización de la planificación operativa. De forma complementaria, Ford, Nava, Tan y Sadler (2020) respaldan este concepto al evidenciar que diversas combinaciones de productos y clientes pueden necesitar modelos diferentes, y que los enfoques desglosados, como el pronóstico a nivel de producto-cliente, suelen ser más exactos que los modelos utilizados en datos agregados. Estos descubrimientos son especialmente pertinentes para esta disertación, ya que pueden ser alterados para segmentar la demanda de

Gatorade por región o medio de distribución, lo que permitiría incrementar la precisión de los pronósticos y mejorar la planificación operativa.

La investigación de Mircetic *et al*. (2016) subraya la relevancia de los modelos S-ARIMA para registrar la estacionalidad y los patrones de consumo en la cadena de abastecimiento de bebidas. Estos modelos han probado su eficacia en situaciones donde las variaciones estacionales influyen de manera considerable en la demanda, un punto crucial en este estudio debido al carácter cíclico del consumo de Gatorade. Incluir un modelo S-ARIMA en el estudio podría ayudar a identificar de manera más efectiva tendencias y fluctuaciones cíclicas, potenciando la habilidad predictiva del sistema predictivo.

Rhufyano *et al*. (2022) estudió el ejemplo de una compañía de té que utilizó el método Holt-Winters para la predicción de demanda y el modelo de Cantidad Económica de Pedido (EOQ) para la organización de materiales. Su investigación evidencia que un enfoque fundamentado en pronósticos confiables puede incrementar la eficacia operacional y disminuir los gastos relacionados con la gestión de inventarios. Pese a las variaciones en la escala y la dinámica de la industria de Gatorade, los fundamentos de esta investigación pueden ser útiles, especialmente en la mejora de la cadena de suministro y la organización de inventarios fundamentada en datos predictivos. (Rhufyano, Robbani, Arifin, Mufti, & Lazuardy, 2022).

Watson y Herbert (2021) señalan que, cuando las empresas de bebidas enfrentan datos limitados, a menudo utilizan métodos de pronóstico más simples, como los promedios móviles trimestrales, para prever la demanda y evitar problemas como la sobreproducción o la falta de productos. En su análisis, subrayan que la predicción trimestral es beneficiosa, dado que la demanda tiende a seguir patrones estacionales que se evidencian claramente en estos modelos más básicos. A pesar de que estos procedimientos no alcanzan el nivel de sofisticación de los modelos de aprendizaje automático, resultan eficaces para administrar los inventarios y optimizar la cadena de suministro. Esta propuesta es significativa para el escenario de Gatorade, donde la demanda también fluctúa dependiendo de la estación, y modelos así podrían ser beneficiosos para optimizar la organización de la producción y distribución.

En esta misma línea, el estudio de Carbonneau, Laframboise y Vahidov (2008) muestra que los modelos más modernos, como las redes neuronales o los sistemas de clasificación, pueden predecir con más precisión que los métodos clásicos, especialmente cuando los datos no son del todo claros, como suele pasar en cadenas de distribución grandes. No obstante, también señalan que estos modelos no siempre alcanzan resultados satisfactorios si no existe una coordinación adecuada entre las áreas que gestionan los datos. Este aspecto es crucial para este estudio,

dado que en Gatorade se presentan varios canales de venta (tales como kioscos, autoservicios y restaurantes) y la calidad del pronóstico también se basará en la integración y uso de los datos de todos estos lugares. Por lo tanto, independientemente del modelo seleccionado, es crucial comprender adecuadamente el contexto y la disposición de la información dentro de la compañía.

El estudio de Edet *et al*. (2024) proporciona visiones valiosas acerca de cómo los patrones de consumo pueden afectar las estrategias de planificación de producción y distribución. En su estudio, los autores destacan cómo el consumo excesivo de bebidas con azúcar puede tener consecuencias negativas para la salud, lo que implica que la demanda de productos como Gatorade podría verse afectada por factores relacionados con la conciencia sobre la salud y la reducción de consumo de azúcares. Los modelos avanzados, tales como los métodos de aprendizaje colectivo empleados en su estudio, podrían también ser útiles para el pronóstico de demanda de Gatorade, dado que estos procedimientos facilitan la identificación de patrones complejos de consumo, modificando las proyecciones en función de variaciones estacionales, de salud pública y preferencias del usuario.

3. Metodología

3.1. Descripción de los datos

Para el desarrollo de este trabajo, se utilizó información obtenida directamente del sistema del embotellador de PepsiCo, Cervecería y Maltería Quilmes (CMQ). Este sistema proporcionó datos detallados sobre las ventas de Gatorade y adicionalmente una base con información sobre las heladeras ubicadas en los puntos de venta. A continuación, se describen las principales características de cada conjunto de datos:

Datos de Ventas:

El dataset principal está compuesto por 2.891.967 de registros distribuidos en 13 columnas. Cada fila representa la venta total de Gatorade para un punto de venta específico, identificado mediante un código único (Código B2B) y está agrupada mensualmente.

Estos datos proporcionan una visión completa del desempeño mensual de Gatorade en cada punto de venta, permitiendo analizar patrones y factores relevantes para el pronóstico a partir de las tendencias históricas.

Tabla 1. Descripción de la base de ventas

Variable	Tipo	Descripción
Año	Integer	Año de la compra
Mes	Integer	Mes de la compra
Código B2B	Integer	Código único del cliente (Llave primaria)
Cantidad de SKUs	Integer	Cantidad de SKUs distintos comprados
Región	Varchar	Región donde se ubica el cliente
Canal ajustado	Varchar	Tipo de punto de venta simplificado
Cantidad Total en HL	Float	Cantidad de hectolitros comprados por el cliente
Importe Bruto	Float	Facutación bruta de la venta
Distri/Directa	Integer	Tipo de distribución de la venta (directa o mediante distribuidor)
Canal	Varchar	Tipo de punto de venta

Datos de Heladeras:

El segundo dataset incluye información detallada sobre 64.624 heladeras que ya fueron instaladas en los puntos de venta y está compuesto por 15 columnas. Cada fila representa a una heladera incluyendo datos relacionados tanto a las características físicas de la misma como a su fecha de colocación, como el modelo o su marca con la cual está decorada ya sea Gatorade o alguna de las que maneja PepsiCo y por otro lado a los atributos del punto de venta donde está ubicada.

Tabla 2. Descripción de la base de heladeras

Variable	Tipo	Descripción
Cod. Región	Integer	Codigo de la región donde se ubica el cliente
Desc. Región	Varchar	Región donde se ubica el cliente
COD B2B	Integer	Código único del cliente (Llave primaria)
Cliente	Varchar	Dueño del punto de venta
Domicilio	Varchar	Domicilio del punto de venta
Canal Mkt	Varchar	Tipo de punto de venta
Subcanal Mkt	Varchar	Tipo de punto de venta con mayor especificación
Cod. Producto	Integer	Código de SKU para la heladera
Desc. Producto	Varchar	Descripción de la heladera
UN	Varchar	Negocio de la heladera
Logo	Varchar	Marca de la heladera
Modelo	Varchar	Tipo de heladera
Nro. Serie	Integer	Numero único para cada heladera
Fec. Colocación	Date	Fecha de colocación
Puertas	Integer	Cantidad de puertas

Para llevar a cabo el análisis, se integraron ambas bases de datos utilizando el Código B2B como clave principal. Antes de esta integración, se agruparon los datos del dataset de heladeras por punto de venta, consolidando toda la información de cada local en una sola fila. De esta manera, se logró enriquecer el dataset de ventas con atributos adicionales sobre las heladeras, proporcionando un contexto más completo que podría influir en el comportamiento de compra de Gatorade.

La combinación de estos dos conjuntos de datos permite no solo obtener una visión más amplia de las operaciones, sino también analizar si características como el tipo de heladera o su presencia en un punto de venta tienen un impacto en las ventas. Esta información es clave para detectar variables que podrían mejorar la precisión de los modelos de pronóstico y variables que no aportan valor adicional. En muchas ocasiones el consumidor final busca que el producto este frío, principalmente si lo va a consumir en el momento de compra.

El conjunto de datos históricos abarca el período de enero de 2022 a septiembre de 2024, lo que proporciona una base sólida para identificar patrones estacionales y tendencias en la demanda de Gatorade. Para estructurar el análisis y evaluar el desempeño del modelo de pronóstico, se definió la siguiente estrategia de partición:

- Datos de entrenamiento: Comprenden las ventas de 2022 y 2023. Estos registros permiten que el modelo aprenda los patrones históricos y ajuste sus predicciones en función de ellos.
- Datos de testeo: Incluyen los meses de 2024, utilizados para medir qué tan bien el modelo predice las ventas con datos más recientes.

Este enfoque garantiza que el modelo se valide con información que refleje las posibles fluctuaciones y tendencias en las condiciones actuales del mercado. Al probarlo con datos recientes, se valida su utilidad en la toma de decisiones comerciales y nos aseguramos de que pueda responder de manera efectiva con su capacidad de adaptación a un mercado cambiante y dinámico.

La variable dependiente principal en este análisis es "Cantidad Total en HL", esta mide el volumen de ventas de Gatorade en hectolitros para cada punto de venta. La predicción se realiza para el período t+1, es decir, el mes siguiente al momento en el que estamos parados ya que no tiene sentido predecir el periodo t que es en el que estamos parados al momento de realizar la predicción. Por lo tanto, el objetivo del modelo es estimar con precisión el consumo mensual de Gatorade en cada punto de venta identificado por su Código B2B, que es un identificador único.

Dado que los datos están agrupados a nivel mensual, cada registro representa el total vendido en un lapso de aproximadamente 30 días. Esto es un aspecto clave, ya que implica que el modelo no se centra en fluctuaciones diarias o semanales, sino en patrones más amplios de consumo. El enfoque temporal utilizado permite capturar tendencias, estacionalidad y otros factores que influyen en la demanda mes a mes, asegurando que la predicción se base en información consolidada y representativa del comportamiento del mercado.

Para mejorar la precisión del modelo y captar tendencias en el comportamiento de las ventas, se incorporaron variables adicionales basadas en el historial de la Cantidad Total en HL. Estas características permiten identificar patrones de consumo y efectos estacionales que influyen en la demanda de Gatorade.

Las cuatro variables generadas se enfocan en el impacto del pasado reciente y la comparación con el mismo período del año anterior:

- **Volumen en t-1**: Representa las ventas del mes inmediatamente anterior al momento donde estamos corriendo la predicción, reflejando tendencias recientes en el consumo.

- **Volumen en t-2**: Captura el comportamiento de las ventas dos meses antes, ayudando a identificar posibles ciclos de corto plazo.
- **Volumen en t-3**: Considera las ventas de tres meses atrás, lo que permite observar patrones de demanda con mayor horizonte temporal.
- **Volumen en t-11**: Corresponde al volumen de ventas del mismo mes del año anterior, clave para detectar efectos estacionales y comparaciones año contra año.

Al incorporar estas variables, el modelo puede analizar cómo evolucionan las ventas a lo largo del tiempo y responder mejor a fluctuaciones estacionales o cambios en la demanda. Este tipo de diferencia en los patrones de consumo puede influir en la precisión del modelo, ya que un punto de venta con compras recurrentes sugiere una demanda más estable y predecible. Por ejemplo, si los últimos tres rezagos son positivos se puede interpretar una reposición de stock continua con un patrón de consumo más claro, mientras que un punto de venta con compras menos estables, pero en grandes cantidades podría estar sujeto a factores externos menos predecibles, como promociones ocasionales o reposiciones de stock más irregulares.

Watson y Herbert (2021) señalan que, en el sector de las bebidas, los métodos de pronóstico basados en series temporales con rezagos que incluyen trimestres permiten una mejor interpretación de la estacionalidad y las fluctuaciones en la demanda. Esto resulta fundamental para prever la demanda de productos como Gatorade, donde la estacionalidad es variable en periodos breves, como los periodos de 3 meses. Es más efectivo implementar un rezago de tres meses (t-1, t-2 y t-3) que añadir un rezago de mayor duración (t-4), ya que la dinámica de la demanda puede fluctuar rápidamente en periodos de tiempo más largos, lo que podría disminuir la precisión de las proyecciones, dado que tendríamos en cuenta meses que ya no aportan valor.

Para capturar mejor las tendencias en el consumo de Gatorade, se consideraron distintas variables que reflejan tanto el historial de compras como las características del punto de venta.

Entre las variables históricas, se incluyen el Año y Mes, fundamentales para estructurar correctamente los datos y permitir que el modelo capture tendencias temporales. Se incorporan además valores de la Cantidad Total en HL en distintos períodos: t-1, t-2 y t-3 como se menciona previamente, que reflejan las ventas de los tres meses previos y ayudan a identificar patrones recientes, y t-11, que permite captar la estacionalidad comparando el mismo mes del año anterior al periodo t+1 que es el que buscamos estimar.

Además, la Cantidad de SKUs t-1 mide la diversidad de productos comprados por el cliente en el mes anterior, lo que puede indicar su comportamiento de compra y nivel de fidelidad hacia la marca. Además, se incluye el Importe Bruto t-1, que representa la facturación bruta total del punto de venta en el mes anterior, lo que permite evaluar la relación entre volumen de compra y valor monetario.

Por otro lado, hay un conjunto de variables que describen el contexto del punto de venta. La Región clasifica geográficamente el local (por ejemplo, NEA, NOA, Litoral), lo que permite evaluar diferencias en el consumo según la ubicación. El Canal Ajustado segmenta los puntos de venta en categorías como kioskos, autoservicios o mayoristas, que tienen ocasiones de compra muy distintas entre ellas, ya que un kiosko probablemente compre empaques más pequeños dado que vende a otro tipo de consumidor que seguramente le compra más de paso.

Un factor clave es la disponibilidad de equipos de frío, esta variable denominada EDF, indica la cantidad total de heladeras en el punto de venta, mientras que Puertas representa el número de puertas de esos equipos, lo que influye en el tamaño de refrigeración. De manera específica, EDF Gatorade mide cuántos de esos equipos están dedicados exclusivamente a la marca, lo que puede impactar en la visibilidad del producto y en la decisión de compra del cliente.

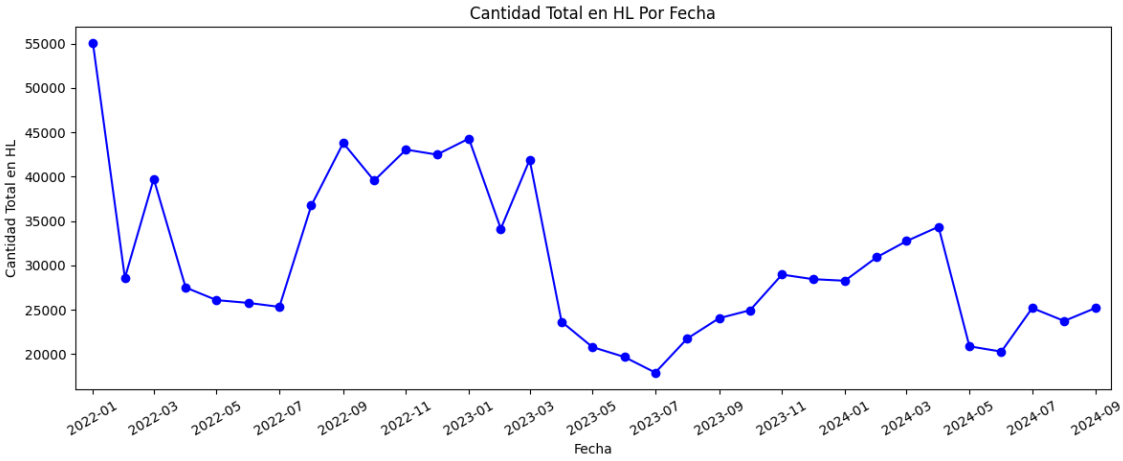
Finalmente, se considera la variable Distri/Directa, que diferencia entre ventas directas y aquellas realizadas a través de distribuidores. Esta distinción es importante porque puede haber diferencias en los patrones de compra dependiendo del modelo de distribución utilizado.

3.2. **Análisis descriptivo**

Al integrar toda la información en un solo dataset, se obtiene una base de datos estructurada que permite un análisis exploratorio más profundo. En este dataset se observan distintos tipos de variables que son del tipo numéricas, como la cantidad de SKUs, el importe bruto y los hectolitros vendidos, también las que son categóricas, como la región y el canal de distribución, por último, una variable de fecha generada a partir de la combinación de las columnas de año y mes, lo que facilita el análisis de tendencias temporales o estacionales en las ventas.

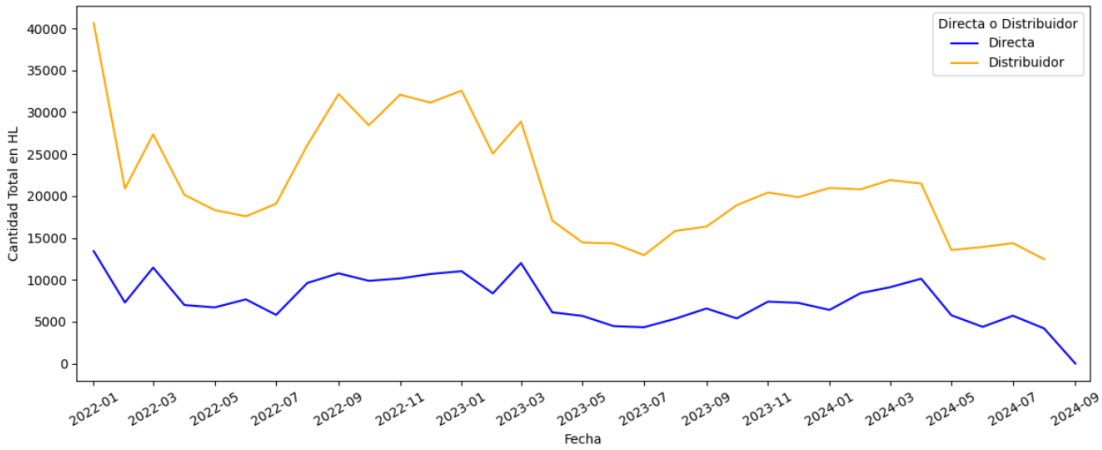
En cuanto a su tamaño, el dataset ocupa alrededor de 308.9 MB, lo que lo hace manejable para el análisis y la construcción del modelo de pronóstico. Con esta base completa, el primer paso lógico es observar la evolución del volumen de ventas de hectolitros a lo largo del tiempo, ya que esto nos pone en contexto de lo que queremos predecir y puede dar señales de patrones estacionales, tendencias de crecimiento o caídas en el consumo. Primero es importante realizar un análisis exploratorio para conocer con profundidad el dataset.

Figura 1. Evolución de la venta real de Gatorade en Hectolitros vendida en Argentina



Se puede observar una clara estacionalidad de la venta en los últimos y primeros meses de cada año, especialmente entre noviembre y marzo, lo cual tiene sentido ya que el calor y el sudor incentivan al consumo de bebidas isotónicas, mientras que en meses donde hace frío como junio y julio parece que el consumo baja ya que la necesidad de hidratación es menor. También hay una tendencia negativa, en 2022 el consumo era mucho mayor al de 2023 y aún más al de 2024, esto podría estar relacionado a una caída de la industria en general dado el momento económico que atraviesa el país actualmente y hace ya algunos años.

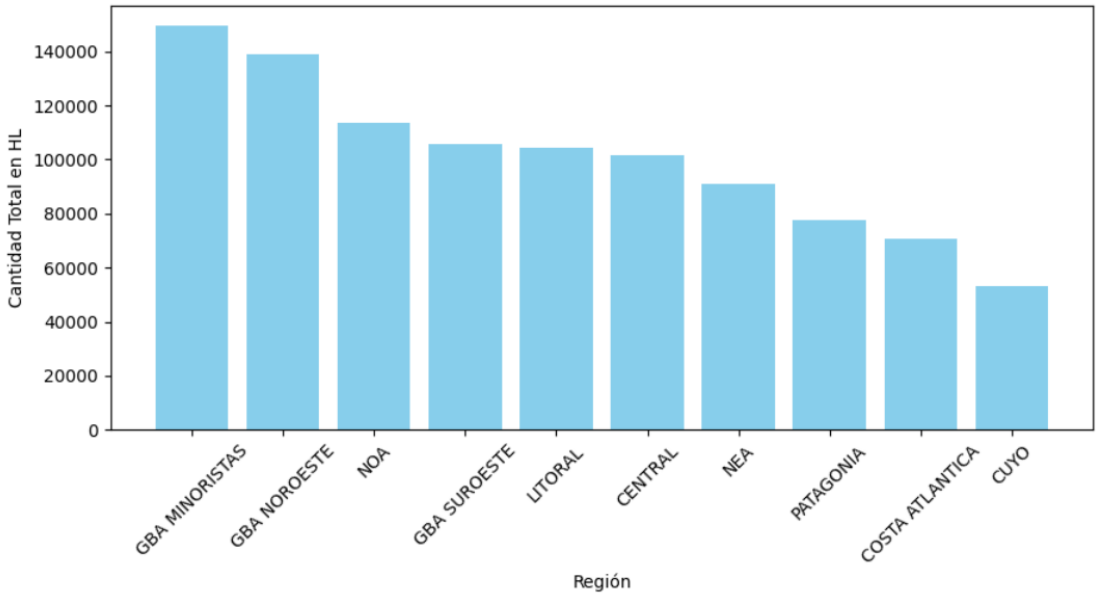
Figura 2. Evolución de la venta real de Gatorade en Hectolitros según el tipo de venta (Directa o Distribuidor)



Este gráfico muestra el volumen total de ventas mensuales separado por tipo de venta (directa o a través de distribuidores). Se observa claramente que ambas modalidades comparten una tendencia y estacionalidad similares, con curvas extremadamente simétricas, aunque la venta a través de distribuidores se mantiene consistentemente como la predominante. Además, esta

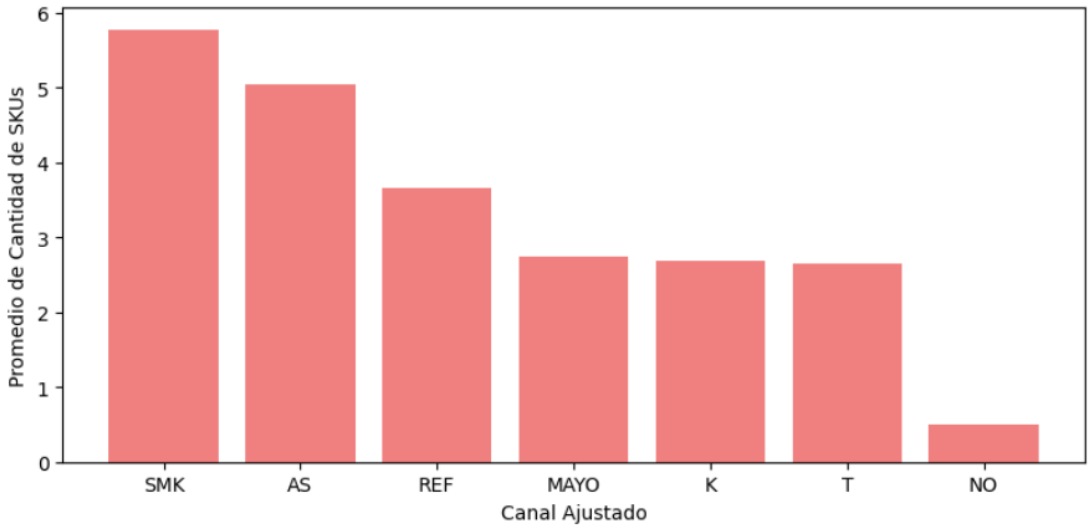
división resulta coherente con el gráfico anterior, que muestra el volumen total de ventas mensuales, correspondiente a la suma de ambas líneas.

Figura 3. Cantidad total de Hectolitros vendidos por región



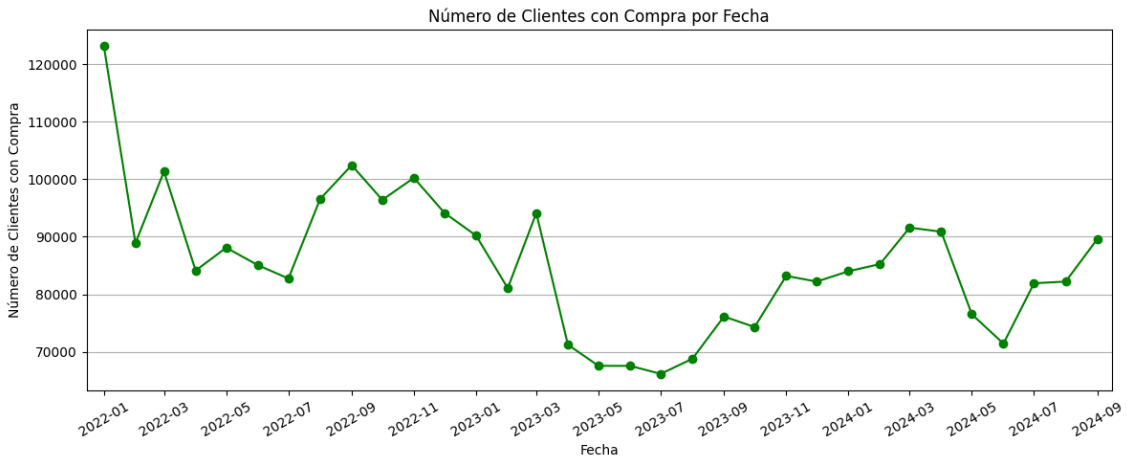
En este grafico podemos ver la cantidad total de hectolitros vendidos desde 2022 agrupado por la región que más se vendió a la que menos. También podemos observar que el Gran Buenos Aires está abierto en 3 regiones la primera (GBA Minoristas) es la que se vende por venta directa y las otras 2 son mediante Distribuidores (GBA Noroeste y GBA Suroeste). Aunque se haya hecho la división se puede observar que siguen siendo las de mayor relevancia ocupando los puestos 1,2 y 4 en el gráfico. Por otro lado las de menor relevancia son todas las del Sur (Patagonia, Costa Atlántica y Cuyo).

Figura 4. Promedio de la cantidad total de SKUs que compran los puntos de venta según su canal ajustado



Aquí observamos el promedio de SKUs vendidos por punto de venta según su canal. Es bastante razonable ver que en Supermercados y autoservicios lideren esta visualización ya que suelen vender más e incluso tienen venta de tanto calibres grandes como pequeños. En la categoría Supermercado tenemos pocos datos ya que en este proyecto usamos la venta real y esta incluye la venta directa y de distribuidores, pero no la que retiran los Supermercados por sus propios medios. Luego tenemos Refrigerados que incluye restaurantes y lugares de comida rápida y más abajo Mayoristas, Kioscos y Tradicionales (Almacenes) que se les venden alrededor de 3 SKUs, típicamente Gatorade 500cc de Manzana, Cool Blue y Frutos tropicales. También tenemos algunos Clientes que no tienen un canal ajustado, pero además de ser pocos también son de muy baja relevancia ya que son claramente los de menor cantidad de SKUs.

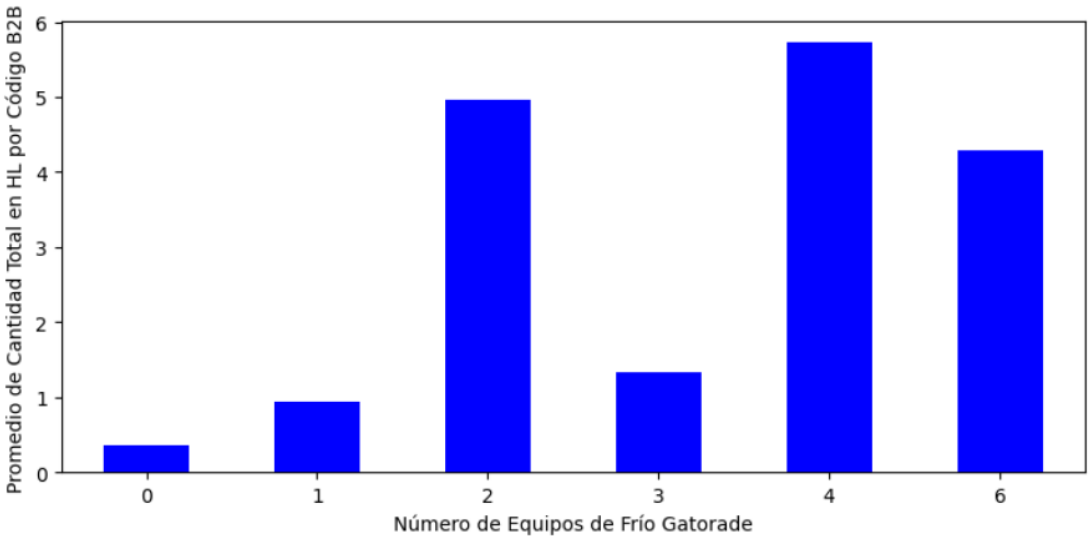
Figura 5. Evolución de la cantidad de puntos de venta con compra de Gatorade



En este caso, podemos ver la evolución de clientes con compra mensuales. Para definir cliente con compra, se lo considera si su volumen en hectolitros fue mayor a cero.

Vemos que tiene una correlación muy alta con el grafico de la evolución mensual de volumen, donde hablábamos de la estacionalidad que genera una mayor venta cuando hace calor y viceversa en épocas de frío y una tendencia negativa dada también una caída que viene teniendo la industria de bebidas en Argentina. Vemos también que en general la marca tiene un promedio de 90.000 clientes con compra por mes, aunque tiene un máximo de 120.000 en Enero de 2022 y un mínimo de 65.000 aproximadamente en Julio de 2023.

Figura 6. Promedio de la venta mensual en Hectolitros según la cantidad de equipos de frío de Gatorade que tenga el punto de venta



En este grafico podemos apreciar la incrementalidad promedio que genera tener un equipo de frío ploteado de Gatorade, vemos que el consumo mensual por código B2B en HL parece ser optimo cuando hay 2 EDF llegando a 5 hectolitros de consumo promedio por mes, es un consumo muy grande para un solo punto de venta, considerando que aquellos sin equipo de frío consumen menos de 0.5 hectolitros en promedio. También cabe mencionar que si bien 2 EDF trae una enorme incrementalidad, hay una tendencia positiva entre 3 y 6 heladeras ploteadas de Gatorade.

3.3. **Modelos y técnicas**

Para predecir el volumen de ventas de Gatorade en los distintos puntos de venta, opté por modelos que se ajustan a la naturaleza cambiante de los datos. Esto me permitió obtener estimaciones más precisas y planificar mejor la producción. Como las ventas tienden a fluctuar

debido a factores estacionales y tendencias a largo plazo, exploré tres enfoques distintos, cada uno con beneficios específicos:

- **Random Forest Regressor**: Este modelo es útil porque puede detectar relaciones complejas entre las variables sin necesidad de hacer suposiciones estrictas sobre los datos. Además, maneja grandes volúmenes de información sin problemas y, si se configura correctamente, evita el sobreajuste, es decir, que el modelo se adapte demasiado a los datos y pierda capacidad de generalización.

- **XGBoost**: Es un algoritmo diseñado para trabajar con datos estructurados y detectar patrones difíciles de identificar. Su principal ventaja es su capacidad de hacer predicciones precisas, incluso cuando los datos contienen ruido o múltiples variables que interactúan entre sí. Además, permite ajustar sus parámetros para optimizar su rendimiento.

- **LightGBM**: Comparado con XGBoost, este modelo es más rápido y consume menos memoria, lo que lo hace ideal cuando se trabaja con grandes volúmenes de datos. Su estructura permite entrenar modelos de manera eficiente, logrando un buen balance entre velocidad y precisión, lo que resulta útil en el análisis de series temporales con múltiples factores en juego.

Estos modelos fueron elegidos porque ayudan a descubrir patrones ocultos en los datos y proporcionan predicciones confiables, lo que es clave para tomar mejores decisiones en la gestión de la demanda. El estudio de Edet, Ekong y Attih (2024) evidencia la eficacia de los métodos de aprendizaje colectivo para efectuar proyecciones en el sector de la industria de las bebidas, particularmente en la valoración de los efectos en la salud del consumo de bebidas gaseosas. Aunque su aplicación está centrada en un área diferente, el uso de estas técnicas predictivas refuerza la idea de que modelos como Random Forest, XGBoost y LightGBM pueden mejorar la precisión de las predicciones en un contexto relacionado con el análisis de la demanda de productos como Gatorade.

Aunque existen métodos tradicionales de análisis de series de tiempo, como S-ARIMA o Holt-Winters, en este estudio se decidió emplear modelos de aprendizaje automático por su mayor adaptabilidad para incluir múltiples variables explicativas. En contraposición a los métodos convencionales, que generalmente se enfocan solo en el desarrollo histórico de la variable objetivo, los modelos seleccionados posibilitan considerar elementos extra como el canal de venta, la región, la cantidad de SKUs, entre otros. Esto es particularmente beneficioso en un

ambiente de negocios complejo como el de las bebidas, donde la demanda puede verse afectada por diversos factores más allá del comportamiento previo. Además, estos modelos facilitan la identificación de patrones no lineales y una mejor adaptación a datos con variaciones o características específicas del mercado.

Antes de entrenar los modelos, era necesario llevar a cabo un preprocesamiento para asegurar la limpieza y transformación de los datos con el objetivo de que pudieran ser correctamente interpretados y utilizados en los modelos de predicción. Este preprocesamiento incluyó varias etapas clave.

Cantidad Total en HL: La variable presentaba puntos como separadores de miles y comas como separadores decimales, lo que generaba errores en la conversión a valores numéricos. Para corregirlo, eliminé los puntos y reemplacé las comas por puntos antes de convertir la columna a tipo numérico.

Transformación de la variable objetivo: Dado que la variable de interés (“Cantidad Total en HL”) presenta alta variabilidad, se aplicaron transformaciones logarítmicas y diferenciales para estabilizar la serie.

Generación de features: Se crearon variables adicionales como “rezagos de ventas” (últimos tres meses y el mismo mes del año anterior), indicadores de tendencia y efectos estacionales.

Conversión de ‘Importe Bruto’ en variable categórica: Dado que el importe bruto es una variable numérica con una alta variabilidad, se agrupó en cuatro categorías (‘A’, ‘B’, ‘C’ y ‘D’) para facilitar su interpretación y reducir la influencia de valores extremos en el modelo. Esta transformación reduce el riesgo de sobre ajustar los datos y mitiga los valores extremos. Si bien implica la pérdida del carácter ordinal de esta variable, se compensa con una representación más robusta y generalizable para los fines predictivos del modelo.

Codificación de variables categóricas: Se aplicó la técnica de One Hot Encoding a las variables categóricas, como ‘Región’, ‘Canal ajustado’ e ‘Importe Bruto’, convirtiéndolas en un formato numérico compatible con el modelo. Este método transforma cada posible valor de una variable en una nueva columna binaria, cuyo valor es 1 si el registro pertenece a esa categoría y 0 en caso contrario.

Meses en formato de texto: La columna "Mes" almacenaba los nombres de los meses como texto, lo que dificultaba su uso en cálculos numéricos y ordenamientos. Para resolverlo, creé un diccionario de mapeo y convertí los valores a tipo entero.

Generación de la variable Fecha: Para facilitar el análisis temporal, creé la variable "Fecha" combinando las columnas de "Año" y "Mes", convirtiéndola a formato datetime para su uso en series de tiempo.

Corrección de valores atípicos y datos faltantes: Se eliminaron valores extremos y se imputaron datos nulos mediante la interpolación de series temporales y métodos basados en la distribución de los datos.

Estandarización de variables numéricas: Para evitar que las diferencias en escala entre las variables afecten el modelo, se aplicó estandarización a variables como la cantidad de SKUs vendidos, la cantidad total en hectolitros y el número de puertas en los puntos de venta.

Para mejorar el rendimiento de los modelos, llevé a cabo un ajuste en la optimización de hiperparámetros utilizando dos enfoques:

- **Random Search** una estrategia más flexible que explora un rango más amplio de valores con un menor costo computacional.
- **Grid Search** que permite explorar de manera exhaustiva combinaciones específicas dentro de un conjunto limitado de valores.

Random Forest

En el caso del modelo Random Forest, se ajustaron los siguientes hiperparámetros:

- n_estimators: cantidad de árboles en el bosque.
- max_depth: profundidad máxima de los árboles.
- min_samples_split: mínimo de muestras para dividir un nodo.
- min_samples_leaf: mínimo de muestras en una hoja.
- max_features: número máximo de características consideradas en cada división de los árboles.

XGBoost

Para el modelo XGBoost, se optimizaron los siguientes hiperparámetros:

- n_estimators: número de árboles en el ensamble.
- max_depth: profundidad máxima de los árboles.
- learning_rate: tasa de aprendizaje utilizada en la actualización de pesos.

- subsample: fracción de muestras utilizadas para entrenar cada árbol, lo que ayuda a reducir el sobreajuste.
- colsample_bytree: proporción de características utilizadas en cada árbol.
- gamma: reducción mínima en la función de pérdida requerida para realizar una partición adicional.

LightGBM

Para LightGBM, se ajustaron los siguientes hiperparámetros:

- num_leaves: número máximo de hojas en cada árbol, lo que controla la complejidad del modelo.
- max_depth: profundidad máxima de los árboles.
- learning_rate: tasa de aprendizaje utilizada para la actualización de los pesos.
- feature_fraction: fracción de características utilizadas en cada iteración del entrenamiento.
- bagging_fraction: proporción de datos utilizados en cada iteración, lo que introduce aleatoriedad y mejora la generalización.
- min_data_in_leaf: número mínimo de muestras en cada hoja para evitar sobreajuste.

Los modelos fueron entrenados mediante un método de validación temporal deslizante (time series cross-validation), lo que permitió evaluar su capacidad para predecir en distintos periodos a futuro. Se aplico un método iterativo, en el que las predicciones a corto plazo se incorporan progresivamente para prever situaciones a mayor plazo.

3.4. Validación del modelo

Para evaluar el desempeño de los modelos de pronóstico desarrollados, se optó por implementar una estrategia de validación basada en los datos históricos. La idea es utilizar los datos de ventas de Gatorade correspondientes a los años anteriores, es decir que para este caso serían los de los años 2022 y 2023 para el entrenamiento de los modelos, de modo que los datos de 2024, que son los más recientes disponibles, se reservaron exclusivamente para la fase de evaluación.

Este método concuerda con los principios que plantea Mircetic *et al*. (2016) acerca de las series de tiempo en la industria de bebidas, destacando la importancia de utilizar datos históricos en la calibración de modelos para mejorar la exactitud de las predicciones previniendo problemas como el sobreajuste y conduciendo a una cadena de suministro eficiente. Esta separación nos

permite reflejar el comportamiento del modelo en condiciones realistas de la predicción, ya que al estar en una serie temporal buscamos predecir los datos más actuales por lo tanto esos son los que el algoritmo no debe conocer en la validación, de esta manera evitamos el sobreajuste y al mismo tiempo una medida más precisa de su capacidad para generalizar. Por otro lado, se aplicó un muestreo del 20% de los datos en el entrenamiento para que el tiempo de ejecución no sea excesivo facilitando la carga computacional.

La validación se llevó a cabo utilizando métricas ampliamente reconocidas en la literatura de pronóstico de series temporales, tales como MAE, MSE, RMSE y MAPE:

Error Absoluto Medio (MAE):

Esta medida permite identificar cuánto se equivocan en promedio, las estimaciones del modelo con respecto a los valores reales. No importa si el error es hacia arriba o hacia abajo, solo se considera la diferencia en términos absolutos. Esta es una de las formas más sencillas e intuitivas de saber por cuánto se está desviando el modelo en cada predicción.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1)$$

Donde:

- n es el número total de observaciones.
- y_i es el valor real en el punto i.
- $\hat{y}_i$ es el valor predicho en el punto i.
- $|y_i - \hat{y}_i|$ es el valor absoluto del error en cada observación.

Error Cuadrático Medio (MSE):

Usamos esta métrica para calcular el promedio de los errores, pero en vez de hacerlos en términos absolutos se realiza elevándolos al cuadrado. Esto genera que los errores grandes en la predicción, ya sea por arriba o por abajo, tengan un mayor impacto en el resultado, lo que de alguna manera ayuda a detectar si el modelo está cometiendo fallos significativos en algunas predicciones.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

Donde:

- n es el número total de observaciones.
- y_i es el valor real en el punto i.
- $\hat{y}_i$ es el valor predicho en el punto i.
- $(y_i - \hat{y}_i)$ es el error cuadrático de cada observación.

Raíz del Error Cuadrático Medio (RMSE):

Es la raíz cuadrada del MSE, una medida que calcula el promedio de los errores al cuadrado y da más peso a los errores grandes. Al aplicar la raíz cuadrada, el resultado se expresa en las mismas unidades que los datos originales, lo que facilita su interpretación y nos ayuda a evaluar qué tan preciso es el modelo en la práctica.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Donde:

- n es el número total de observaciones.
- y_i es el valor real en el punto i.
- $\hat{y}_i$ es el valor predicho en el punto i.
- $(y_i - \hat{y}_i)$ es el error cuadrático de cada observación.

Error Porcentual Absoluto Medio (MAPE):

Expresa el error en términos relativos al valor real, lo que facilita la comparación del rendimiento entre distintos modelos y diferentes escalas de datos.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \quad (4)$$

Donde:

- n es el número total de observaciones.
- y_i es el valor real en el punto i.
- $\hat{y}_i$ es el valor predicho en el punto i.

- $\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100$ Es el error absoluto en porcentaje para cada observación.

Sin embargo, esta métrica termino no siendo adecuada en este caso debido a la presencia de valores cercanos a cero en la variable objetivo (Puntos de venta que en el mes no compran o compran muy poco). El MAPE se calcula dividiendo el error absoluto por el valor real, lo que genera resultados excesivamente altos cuando los valores reales son muy pequeños, distorsionando la interpretación de la precisión del modelo. Por este motivo, se descartó el MAPE en favor de métricas más robustas para este conjunto de datos.

Se implementó una validación temporal, en la que los modelos fueron ajustados con datos hasta un determinado mes y luego se realizaron predicciones sobre los meses subsiguientes. Este enfoque refleja fielmente un escenario de aplicación real, donde las decisiones de producción y distribución se basan en estimaciones de demanda futura.

Además, se realizó un análisis comparativo entre los diferentes modelos implementados, evaluando su rendimiento en función de las métricas mencionadas. Se analizaron las distribuciones de error para identificar sesgos sistemáticos y detectar posibles ajustes necesarios en los hiperparámetros de los modelos. Finalmente, los resultados fueron contrastados con enfoques de referencia, como un modelo de promedio móvil y un modelo de regresión base, para determinar la ganancia de precisión obtenida con los métodos propuestos.

Los hallazgos de esta fase de validación permiten no solo seleccionar el modelo más preciso para la predicción de ventas de Gatorade, sino también establecer una base sólida para futuras mejoras y refinamientos en el proceso de forecast. Los resultados específicos de MSE, RMSE y MAE se presentan en la sección de Resultados, donde se analiza su impacto en la calidad del modelo.

4. Resultados

Se probaron un total de tres modelos de aprendizaje automático: Random Forest, XGBoost y LightGBM, con el objetivo de predecir la demanda de Gatorade. Para compararlos, todos los modelos fueron evaluados en función de las métricas MAE, MSE y RMSE y para finalmente determinar cuan precisos pueden llegar a ser.

Para optimizar el desempeño de los modelos predictivos, se emplearon técnicas de ajuste de hiperparámetros que permiten encontrar las combinaciones más adecuadas para cada algoritmo. En particular, se utilizaron dos métodos comunes: Grid Search consiste en evaluar exhaustivamente todas las combinaciones posibles dentro de un rango predefinido de valores, garantizando un análisis sistemático y detallado de las configuraciones. Por otro lado, Random Search selecciona de forma aleatoria un número determinado de combinaciones dentro del espacio de parámetros, lo que puede ser más eficiente en términos computacionales cuando el espacio es muy amplio, sin comprometer la calidad de la búsqueda. Ambos métodos ayudan a mejorar la precisión y solidez de los modelos al facilitar un ajuste minucioso de sus parámetros internos.

Asimismo, se evaluó la importancia de las variables utilizadas por cada modelo. Dado que cada algoritmo determina esta importancia de manera distinta, se decidió respetar sus métodos internos. En el caso de Random Forest y LightGBM, se empleó el enfoque clásico basado en la estructura de los árboles, considerando la frecuencia con la que cada variable participa en las divisiones. Sin embargo, la forma de reportar esta importancia varía, ya que en Random Forest se expresa como un porcentaje de contribución total, mientras que en LightGBM se presenta una frecuencia absoluta. Por otro lado, para XGBoost se aplicaron dos criterios diferentes: gain, que mide cuánto mejora la función objetivo al realizar divisiones con una determinada variable, y weight, que indica la frecuencia con la que dicha variable aparece en los árboles. Esta doble perspectiva permitió obtener una comprensión más detallada y profunda del rol que desempeña cada predictor en el modelo de XGBoost.

4.1. Random Forest

Resultados del Modelo Random Forest con Búsqueda Aleatoria de Hiperparámetros

Se ajustaron los hiperparámetros del modelo Random Forest mediante RandomizedSearchCV con validación cruzada de 3 folds, optimizando según MAE. El modelo final entrenado con los mejores parámetros se evaluó en el conjunto de prueba, obteniendo los siguientes resultados:

- Error Absoluto Medio (MAE): 0.23
- Error Cuadrático Medio (MSE): 4.75
- Raíz del Error Cuadrático Medio (RMSE): 2.18

Para analizar la relevancia de cada variable en la predicción de la demanda de Gatorade, se calcularon las importancias de las características del modelo Random Forest. Este análisis permitió identificar cuáles tenían mayor impacto en la estimación de la demanda de Gatorade, proporcionando información valiosa para futuras optimizaciones.

Tabla 3. Nivel de importancia de variables del Random Forest con Random Search.

Variable	Importancia
Cantidad Total en HL (t-1)	45.89%
Cantidad Total en HL (t-2)	18.57%
Cantidad Total en HL (t-3)	7.98
Cantidad de SKUs	7.88%
Mes	3.66%
Región_GBA MINORISTAS	2.88%
Año	2.42%
Cantidad Total en HL (t-11)	1.92%
Canal ajustado_MAYO	1.59%
Canal ajustado_T	1.23%

En la estrategia Random Forest con Random Search, la variable más influyente en la predicción de la cantidad total en hectolitros es la venta del periodo (t-1), con un peso del 45.89%. Esto indica que el comportamiento de las ventas continua con una fuerte tendencia de continuidad, donde el dato más reciente es el mejor predictor del volumen actual. Además, las ventas de los dos meses previos (t-2 y t-3) también tienen una alta importancia siendo las variables que le siguen en mayor relevancia, con 18.57% y 7.98% respectivamente, lo que refuerza lo importante que es contar con los datos históricos que principalmente son recientes en la estimación del forecast. A su vez, la variable correspondiente a las ventas de un año atrás (t-11) muestra una menor influencia (1.92%), aunque su presencia sugiere que el modelo capta cierta estacionalidad en la demanda.

Además del historial de ventas, la cantidad de SKUs adquiridos tiene un impacto considerable en la predicción, con una importancia del 7.88%. Esto se puede conectar a la diversidad de

productos comprados por un cliente está relacionada con patrones de compra más estables. La variable mes, con un peso del 3.66%, también muestra la existencia de fluctuaciones estacionales en la demanda, mientras que el año (2.42%) sugiere la presencia de tendencias a más largo plazo, posiblemente influenciadas por cambios en el mercado o la estrategia comercial.

Por otro lado, la región donde se realiza la venta también tiene cierto impacto, en particular la región "GBA MINORISTAS", que representa un 2.88% de la importancia total. Esto sugiere que la demanda puede variar según la ubicación y el tipo de cliente. Finalmente, las variables relacionadas con el tipo de punto de venta, como "Canal ajustado_MAYO" (1.59%) y "Canal ajustado_T" (1.23%), tienen un peso menor, aunque siguen aportando información relevante para la predicción, siendo en este caso los mayoristas y los almacenes también llamados tradicionales en la base de datos.

Resultados del Modelo Random Forest con Búsqueda en Cuadrícula de Hiperparámetros

Para mejorar la precisión del modelo Random Forest, se ajustaron sus hiperparámetros mediante Grid Search con validación cruzada. Los valores óptimos encontrados se usaron para entrenar el modelo final, que luego fue evaluado con el conjunto de prueba obteniendo las siguientes métricas:

- MAE (Error Absoluto Medio): 0.23
- MSE (Error Cuadrático Medio): 4.93
- RMSE (Raíz del Error Cuadrático Medio): 2.22

El modelo Random Forest permite analizar la importancia de cada variable en la predicción. Las diez variables más influyentes fueron:

Tabla 4. Nivel de importancia de variables del Random Forest con Grid Search.

Variable	Importancia
Cantidad Total en HL (t-1)	33.92%
Cantidad Total en HL (t-2)	27.49%
Cantidad Total en HL (t-3)	14.56%
Cantidad de SKUs	4.43%
Cantidad Total en HL (t-11)	4.03%
Mes	2.99%

Canal ajustado_MAYO	2.24%
Importe Bruto_D	1.73%
Región_GBA MINORISTAS	1.59%
Año	1.28%

Para la estrategia Random Forest con Grid Search, la variable más determinante en la predicción de la cantidad total en hectolitros es la del periodo (t-1), con una importancia del 33.92%. Sin embargo, en comparación con la estrategia de Random Search, el modelo asigna un mayor peso a la venta de periodos atrás (t-2), que alcanza el 27.49%, mientras que la de t-3 representa el 14.56%. Esto refuerza la idea de que las ventas pasadas son el principal factor explicativo de la predicción, con un enfoque más equilibrado en el historial reciente.

El impacto de la cantidad de SKUs vendidos es menor en este modelo, con un 4.43%, lo que sugiere que, aunque sigue siendo una variable importante, su aporte es más limitado en comparación con la importancia de las ventas previas. La variable correspondiente a la venta de t-11 obtuvo un peso del 4.03%, indicando cierta influencia de patrones estacionales en la predicción. La variable mes, con un 2.99%, mantiene un rol en la tarea de captar los efectos estacionales, aunque con una menor relevancia respecto a los datos históricos de ventas.

El canal de venta de mayoristas también tiene cierta influencia, con "Canal ajustado_MAYO" aportando un 2.24% a la predicción, lo que indica que la estrategia comercial puede afectar el volumen de ventas. Además, la variable "Importe Bruto_D" aparece con una importancia del 1.73%, lo que sugiere que el valor de las transacciones también aporta información al modelo. Por otro lado, la región "GBA MINORISTAS" y el año tienen un impacto menor, con 1.59% y 1.28% respectivamente, lo que implica que la localización geográfica y las tendencias anuales tienen un peso secundario en comparación a las variables que se mencionan previamente.

4.2. **XGBoost**

Resultados del Modelo XGBoost con Búsqueda Aleatoria de Hiperparámetros

Se ajustaron los hiperparámetros del modelo XGBoost mediante RandomizedSearchCV con validación cruzada de 3 folds para optimizar su rendimiento. El modelo final entrenado con la mejor configuración fue evaluado en el conjunto de prueba, obteniendo los siguientes resultados:

- Error Absoluto Medio (MAE): 0.24
- Error Cuadrático Medio (MSE): 6.70
- Raíz del Error Cuadrático Medio (RMSE): 2.59

Para analizar la relevancia de cada variable en la predicción de la demanda de Gatorade, se calcularon las importancias de las características del modelo XGBoost utilizando dos métricas:

Weight: Representa cuántas veces una variable fue utilizada en los árboles de decisión del modelo.

Gain: Mide la contribución promedio de una variable a la reducción de la pérdida en el modelo.

Tabla 5. Importancia de variables del XGBoost con Random Search (Weight).

Variable	**Importancia**
Cantidad de SKUs	907
Mes	883
Cantidad Total en HL (t-2)	603
Cantidad Total en HL (t-3)	494
Cantidad Total en HL (t-1)	487
EDF	379
Puertas	319
Año	308
Cantidad Total en HL (t-11)	307
Región_GBA MINORISTAS	236

Analizando la estrategia XGBoost con Random Search, se observa que la variable más relevante según el criterio weight es la Cantidad de SKUs, con un valor de 907, esto indica que fue la variable más usada al momento de generar reglas de decisión. Le sigue muy de cerca el Mes (883), lo que demuestra que la estacionalidad juega un papel crucial en la predicción.

Las variables de rezago de la Cantidad Total en HL también tienen una importancia significativa, con t-2 (603), t-3 (494) y t-1 (487) entre las cinco más influyentes. Esto refuerza la idea de que los valores históricos de ventas son claves para la predicción.

Otras variables como EDF (379) y Puertas (319) también aparecen como factores relevantes, mostrando que tener equipo de frío en el punto de venta y con cuantas puertas es información

muy útil. Mientras tanto, el Año (308) y la Cantidad Total en HL (t-11) (307) sugieren que tanto la tendencia a largo plazo como los patrones anuales pueden aportar información valiosa.

Finalmente, la Región_GBA MINORISTAS (236) tiene la menor importancia dentro del top 10, lo que indica que, la región del gran buenos aires es la que más está aportando en la estimación, aunque su impacto es menor en comparación al resto de variables mencionadas dentro del top 10 de weight.

Tabla 6. Importancia de variables del XGBoost con Random Search (Gain).

Variable	Importancia
Canal ajustado_T	3527
Cantidad Total en HL (t-3)	3398
Región_GBA NOROESTE	2559
Canal ajustado_MAYO	2546
Región_GBA MINORISTAS	2499
Importe Bruto_D	2489
Cantidad de SKUs	2282
Cantidad Total en HL (t-1)	2209
Cantidad Total en HL (t-11)	1858
Región_CUYO	1773

Si observamos XGBoost con Random Search, utilizando gain como métrica de importancia, la variable más relevante es Canal ajustado_T (3527), lo que indica que el canal tradicional es decir los almacenes aportan la mayor reducción de la pérdida en la construcción de los árboles de decisión. También Canal ajustado_MAYO (2546) que representa a los mayoristas, esta entre las variables con mayor gain. Esto nos dice que estos canales de venta tienen una influencia significativa en la predicción de la demanda para Gatorade.

Las variables de rezago en ventas también tienen una fuerte contribución, con Cantidad Total en HL (t-3) (3398) y Cantidad Total en HL (t-1) (2209) entre las más importantes. Esto refuerza la idea de que los valores históricos de ventas desempeñan un papel clave en la estimación de la demanda futura.

En el caso de factores geográficos también destacan en la importancia del modelo. Región_GBA NOROESTE (2559), Región_GBA MINORISTAS (2499) y Región_CUYO (1773) aparecen entre las

principales variables, lo que sugiere que la ubicación de los puntos de venta es información relevante para explicar la variabilidad en las ventas.

Resultados del Modelo XGBoost con Búsqueda en Cuadrícula de Hiperparámetros

Para optimizar el rendimiento del modelo XGBoost, se ajustaron sus parámetros mediante Grid Search, evaluando 96 configuraciones. El modelo final se entrenó con los mejores hiperparámetros y se aplicó early stopping para evitar sobreajuste. Los resultados en el conjunto de prueba fueron:

- Error Absoluto Medio (MAE): 0.25
- Error Cuadrático Medio (MSE): 6.69
- Raíz del Error Cuadrático Medio (RMSE): 2.59

Se analizaron las características más relevantes según dos métricas:

Weight: Representa cuántas veces se utilizó una variable en los árboles de decisión.

Gain: Mide la contribución de cada variable en la reducción del error.

Tabla 7. Importancia de variables del XGBoost con Grid Search (Weight).

Variable	Importancia
Cantidad Total en HL (t-1)	154
Cantidad Total en HL (t-3)	138
Cantidad Total en HL (t-2)	107
Cantidad Total en HL (t-11)	106
Canal ajustado_MAYO	47
Región_GBA MINORISTAS	47
EDF	37
Cantidad de SKUs	28
Año	27
Importe Bruto_D	24

Comenzando con el análisis para la estrategia de XGBoost con Grid Search, utilizando weight como métrica de importancia, lo primero que podemos ver es que las variables de rezago en ventas son las más influyentes a la hora de realizar una predicción. Cantidad Total en HL (t-1) (154), Cantidad Total en HL (t-3) (138) y Cantidad Total en HL (t-2) (107) son las tres variables

más usadas en la construcción de estos árboles, lo que nos refuerza la importancia de la demanda histórica en la estimación del volumen de ventas a futuro. Además, incluso un rezago más lejano, como el de un año atrás (t-11), con un valor de 106, tiene un impacto importante, lo que sugiere que la estacionalidad anual también juega un papel relevante en la predicción.

Entre los factores del tipo de punto de venta y geográficos, destacan Canal ajustado_MAYO (47) y Región_GBA MINORISTAS (47), por lo que el canal mayorista de venta y zonas geográficas tienen una incidencia destacable en las decisiones del modelo.

Otras variables como EDF (37), Cantidad de SKUs (28), Año (27) e Importe Bruto_D (24) tienen menor participación en la partición de los datos dentro de los árboles, pero de todos modos aportan información importante para la predicción. Se puede decir que, si bien la historia de ventas es la fuente principal de información, factores adicionales como el surtido de productos, el canal de venta, el equipo de frío y el contexto temporal también influyen en la demanda estimada.

Tabla 8. Importancia de variables del XGBoost con Grid Search (Gain).

Variable	Importancia
Región_GBA NOROESTE	12664
Cantidad Total en HL (t-3)	11802
Región_GBA MINORISTAS	10805
Cantidad Total en HL (t-1)	7843
Cantidad Total en HL (t-11)	7049
Canal ajustado_K	6744
Importe Bruto_D	4586
Cantidad Total en HL (t-2)	4384
Canal ajustado_T	4074
Mes	4039

En la estrategia de XGBoost con Grid Search, utilizando gain como métrica de importancia, se observa que las variables geográficas y los rezagos en las ventas son las que más aportaron a la reducción del error en las divisiones de los árboles.

Región_GBA NOROESTE (12.664) y Región_GBA MINORISTAS (10.805) obtuvieron la mayor ganancia en la construcción del modelo, por lo que indica que la ubicación geográfica de los clientes, principalmente en GBA, es un factor clave en la predicción del volumen de venta.

Los rezagos en ventas siguen siendo fundamentales, con Cantidad Total en HL (t-3) (11802) como la segunda variable más importante, seguida de Cantidad Total en HL (t-1) (7843) y Cantidad Total en HL (t-11) (7049). Esto sigue demostrando que la información de la demanda pasada, tanto reciente como de períodos previos más alejados, es altamente relevante para estimar las futuras ventas. Entre los factores del tipo de cliente, se destacan Canal ajustado_K (6744) y Canal ajustado_T (4074), que en este caso vendrían a ser los kioskos y los almacenes, se puede ver que la segmentación por tipo de canal de venta afecta a la distribución de la demanda.

4.3. **LightGBM**

Resultados del Modelo LightGBM con Búsqueda Aleatoria de Hiperparámetros

Llegando al último modelo, para optimizar LightGBM se ajustaron sus hiperparámetros mediante Random Search, buscando la configuración que minimizara el error absoluto medio (MAE). El modelo final entrenado con los mejores parámetros fue evaluado en el conjunto de prueba, con los siguientes resultados:

- Error Absoluto Medio (MAE): 0.25
- Error Cuadrático Medio (MSE): 7.36
- Raíz del Error Cuadrático Medio (RMSE): 2.71

Luego buscamos analizar la relevancia de cada variable en nuestra estimación para entender cuales están siendo de gran utilidad y cuáles son las de menor impacto. Los resultados mostraron que las variables más influyentes fueron:

Tabla 9. Importancia de variables del LightGBM con Random Search.

Variable	Importancia
Cantidad Total en HL (t-2)	1553
Cantidad Total en HL (t-1)	1509
Cantidad Total en HL (t-3)	1099
Cantidad Total en HL (t-11)	644
Cantidad de SKUs	629
Mes	553
Puertas	358
EDF	279
Año	129

Importe Bruto_D	50

Para el modelo LightGBM con Random Search, la importancia de las variables muestra una fuerte dependencia de los rezagos en las ventas para la estimación de la demanda.

Las tres variables con mayor importancia son Cantidad Total en HL (t-2) (1553), Cantidad Total en HL (t-1) (1509) y Cantidad Total en HL (t-3) (1099), esto confirma que los datos de ventas recientes tienen un peso significativo en la estimación del volumen futuro. La inclusión de Cantidad Total en HL (t-11) (644) sugiere que las tendencias anuales también desempeñan un papel en la predicción, posiblemente reflejando patrones estacionales.

Además, Cantidad de SKUs (629) se posiciona como una variable relevante, es decir que la diversidad de productos vendidos a un cliente puede estar asociada con los niveles de compra más estables.

Entre las variables relacionadas a las heladeras dentro del punto de venta, Puertas (358) y EDF (279) muestran cierta influencia en la predicción, lo que sugiere que la infraestructura del punto de venta puede afectar los volúmenes de compra y la estabilidad, ya que para tener equipo de frío se requiere una gran inversión inicial que asegura que el cliente se compromete a repagar.

Resultados del Modelo LightGBM con Búsqueda en Cuadrícula de Hiperparámetros

Para mejorar el rendimiento del LightGBM, se optó por utilizar Grid Search para ajustar los hiperparámetros del modelo. Esta estrategia permitió explorar de forma sistemática diversas combinaciones y seleccionar aquella que ofrecía el menor error absoluto medio. Con los valores óptimos identificados, se entrenó el modelo final, el cual fue evaluado utilizando el conjunto de prueba. A continuación, se presentan los resultados obtenidos:

- Error Absoluto Medio (MAE): 0.27
- Error Cuadrático Medio (MSE): 8.44
- Raíz del Error Cuadrático Medio (RMSE): 2.90

Se analizaron las importancias de las variables en la predicción de la demanda de Gatorade. Los resultados indicaron que los factores más influyentes fueron:

Tabla 10. Importancia de variables del LightGBM con Grid Search.

Variable	Importancia

Cantidad Total en HL (t-1)	428
Cantidad Total en HL (t-3)	415
Cantidad Total en HL (t-2)	375
Cantidad de SKUs	136
Puertas	118
Cantidad Total en HL (t-11)	96
Mes	92
EDF	28
Año	25
Canal ajustado_MAYO	23

En el modelo LightGBM con Grid Search, la importancia de las variables mantiene una estructura similar a la observada en el mismo modelo, pero con Random Search, manteniendo un fuerte énfasis en los rezagos de ventas.

Las tres variables más influyentes son Cantidad Total en HL (t-1) (428), Cantidad Total en HL (t-3) (415) y Cantidad Total en HL (t-2) (375), confirmando que los valores históricos recientes de volumen son los principales predictores de las ventas futuras. La presencia de Cantidad Total en HL (t-11) (96) también es relevante y el modelo sugiere que existe un componente estacional.

Por otro lado, Cantidad de SKUs (136) sigue teniendo un peso considerable siendo la cuarta variable más importante para el modelo, lo que indica que la variedad de productos adquiridos por los clientes influye en el volumen de ventas. Además, Puertas (118) y Mes (92) aportan información adicional, reflejando que la infraestructura del punto de venta y la temporalidad tienen impacto en la predicción.

Las variables EDF (28), Año (25) y Canal ajustado_MAYO (23) tienen una menor importancia en comparación con el resto, si bien pueden aportar información, su impacto en la predicción es limitado en relación con las demás características.

5. Conclusiones

En esta sección se comparan los resultados finales de los modelos evaluados en términos de precisión y relevancia de variables. Se analizan las métricas de error obtenidas para concluir que estrategia ofrece la mejor capacidad predictiva y cuáles factores tienen mayor impacto en la predicción del volumen de ventas.

Tabla 11. Comparación de métricas de validación por estrategia

Estrategia	MAE	RMSE	MSE
Random Forest (Random Search)	0.23	2.18	4.74
Random Forest (Grid Search)	0.23	2.22	4.93
XGBoost (Random Search)	0.24	2.59	6.70
XGBoost (Grid Search)	0.25	2.59	6.69
LightGBM (Random Search)	0.25	2.71	7.36
LightGBM (Grid Search)	0.27	2.90	8.44

En cuanto al desempeño general de los modelos, el Random Forest obtuvo los mejores resultados en todas las métricas, si bien en el MAE la mejora es mínima sigue siendo la mejor alternativa en todos los casos principalmente en la validación de MSE.

Por otro lado, XGBoost mostro un rendimiento intermedio con un MAE de 0.24-0.25 y un RMSE de 2.59 se ve claramente que funciona peor que Random Search pero mejor que LightGBM y además no tiene prácticamente diferencia entre optimizar hiperparámetros de forma aleatoria o en forma de grilla.

Por último, LightGBM tuvo el peor desempeño con los valores más altos siempre llegando en el caso de la estrategia con Grid Search a un MSE de hasta 8.44. Lo que indica que en este caso el modelo no se ajusta tan bien a los datos comparado al resto de alternativas.

Respecto a la optimización de hiperparámetros se observa que Random Search siempre funciona por encima o igual que Grid Search exceptuando el caso de XGBoost con la métrica MSE (6.70 contra 6.69) donde la diferencia es mínima. Lo que sugiere que utilizar una búsqueda aleatoria trae mejores resultados.

Finalmente podemos concluir que la mejor estrategia observada es Random Forest con Random Search logrando menor cantidad de errores en la predicción con un MAE de 0.23, RMSE de 2.18 y el MSE de 4.74. Del lado opuesto se tiene la estrategia de LightGBM con Grid Search que obtuvo la mayor cantidad de errores, llegando a un MAE de 0.27, RMSE de 2.90 y el MSE de 8.44.

Tras evaluar los modelos Random Forest, XGBoost y LightGBM con estrategias de optimización de hiperparámetros mediante Random Search y Grid Search, ya identificamos patrones clave en el desempeño y ahora vamos a hacerlo para entender la relevancia de las variables en la predicción de ventas de Gatorade.

Tras un análisis detallado observado en los resultados en la importancia de variables en cada modelo ha permitido identificar cuales son los factores son más relevantes para la predicción de ventas. Es evidente que los rezagos de ventas (Cantidad Total en HL en diferentes períodos anteriores) surgieron como la advertencia más fuerte en todas las estrategias, con especial énfasis en los valores de t-1, t-2 y t-3, tal y como habíamos observado en el análisis exploratorio que el volumen de ventas recientes es el mejor predictor del futuro.

Sin embargo, los modelos presentan diferencias en cómo ponderan el resto de variables:

Random Forest otorga mayor importancia a Cantidad de SKUs y variables temporales como el Mes y el Año, por lo que este modelo capta patrones generales de estacionalidad y variedad de productos de manera efectiva en la estimación.

XGBoost, dependiendo de la métrica de importancia utilizada ya sea Weight o Gain), destaca la influencia de ciertas regiones como en el Gran Buenos Aires y canales de venta como los mayoristas y los kioskos entre otros. Por lo tanto, se puede decir que es más sensible a segmentaciones geográficas y características del tipo de punto de venta.

LightGBM sigue una estructura similar al Random Forest, aunque con menor variabilidad en la importancia de las variables, priorizando fuertemente a las ventas pasadas y otorgando menos relevancia a factores como los geográficos y los canales de venta. Además, resulta interesante destacar que tiene en cuenta la estructura de frío, tanto para las heladeras como la cantidad de puertas de las mismas que posee el punto de venta.

Los resultados obtenidos nos vuelven a confirmar lo importante que es tomar en cuenta la estacionalidad y la tendencia en la demanda de Gatorade, ya que en todos los modelos las variables más influyentes fueron los rezagos de ventas de meses anteriores. Esto tiene sentido, pues el consumo de bebidas isotónicas no se mantiene estable durante el año, sino que sigue

un patrón marcado por las estaciones. Por ejemplo, durante el verano están los picos más altos de consumo debido a las altas temperaturas que provocan un mayor sudor al deportista y el aumento de la actividad física generan una mayor demanda, reflejando así una clara tendencia estacional.

Las diferencias en la importancia de las demás variables reflejan cómo cada modelo capta distintos aspectos del negocio. Mientras que Random Forest prioriza la cantidad de SKUs y variables temporales, XGBoost parece ser más sensible a diferencias entre regiones y canales de venta. LightGBM, en cambio, mantiene una estructura más homogénea en la ponderación de variables, aunque con menor peso en factores externos.

Desde una perspectiva práctica, la elección del modelo depende del equilibrio entre precisión y eficiencia computacional. Si el objetivo es maximizar la exactitud de la predicción, Random Forest con Random Search se presenta como la mejor opción. Sin embargo, XGBoost y LightGBM pueden ser útiles si se busca entender mejor la segmentación de la demanda o trabajar con modelos más ligeros para implementaciones en producción.

Como conclusión, se recomienda la aplicación del modelo Random Forest optimizado por Random Search como herramienta de guía en la planificación de producción de Gatorade. Esta estrategia podría implementarse en el segundo día hábil de cada mes para calcular la demanda prevista por región, lo que permitiría ajustar con mayor exactitud los volúmenes de producción y distribución. De este modo, se maximiza la utilización de recursos logísticos, mientras se reducen los riesgos de exceso de inventario o faltante de stock y se fortalecen las decisiones relacionadas a las acciones comerciales o la distribución del presupuesto para campañas de la marca. En situaciones de alta estacionalidad o cambios de tendencia, disponer de una estimación sólida también permite anticipar desviaciones y reaccionar con mayor agilidad.

6. Bibliografía

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European journal of operational research, 184(3),* 1140-1154. https://doi.org/10.1016/j.ejor.2006.12.004

Edet, A., Ekong, B., & Attih, I. (2024). Machine Learning Enabled System for Health Impact Assessment of Soft Drink Consumption Using Ensemble Learning Technique. *International Journal Of Computer Science And Mathematical Theory*, *10(1),* 79-101.

Ellithorpe, M. E., Bleakley, A., Hennessy, M., Jordan, A., Stevens, R., & Maloney, E. (2023). Athletes drink gatorade: DMA advertising expenditures, ad recall, and athletic identity influence energy and sports drink consumption. *Health Communication*, *38(13),* 3031-3039.

Ford, J., Nava, C., Tan, J., & Sadler, B. (2020). Automated Machine Learning Framework for Demand Forecasting in Wholesale Beverage Alcohol Distribution. *SMU Data Science Review, 3(3), 7. https://scholar.smu.edu/datasciencereview/vol3/iss3/7/*

Jiang, L., Rollins, K. M., Ludlow, M., & Sadler, B. (2020*).* Demand forecasting for alcoholic beverage distribution. *SMU Data Science Review*, *3(1), 5.* https://scholar.smu.edu/datasciencereview/vol3/iss1/5

Ma, S., Kolassa, S., & Fildes, R. (2018). Retail forecasting: research and practice. *Lancaster University Management School. Lancaster, UK.* https://mpra.ub.uni-muenchen.de/89356/

Mircetic, D., Nikolicic, S., Maslaric, M., Ralevic, N., & Debelic, B. (2016). Development of S-ARIMA model for forecasting demand in a beverage supply chain. *Open engineering*, *6(1).*

Rhufyano, A. F., Robbani, M. F., Arifin, H. R., Mufti, J. S., & Lazuardy, A. (2022). Optimizing efficiency through application of the materials requirement planning and demand forecasting: a case study of a small and medium-sized enterprise in the tea beverage industry. In *Proceedings of the 5th European International Conference on Industrial Engineering and Operations Management* (pp. 1152-1162).

Watson, B., & Herbert, S. (2021). Forecasting demand for an overseas beverage company. *Proceedings of the Annual General Donald R. Keith Memorial Conference. Society for Industrial and Systems Engineering.*

7. Apéndices

Apéndice A. Registros iniciales del conjunto de datos

Datos de ventas:

Tabla 12. Muestra de la venta real de Gatorade por punto de venta a mes cerrado

Año	Mes	Código B2B	Cantidad de SKUs	Región	Canal ajustado	Cantidad Total en HL	Importe Bruto	EDF	Puertas	EDF Gatorade	Distri/Directa	Canal
2024	Febrero	3,99782E+13	2	NEA	K	0,12	20.126,8	1	1	0	1	Kioscos/Maxikioscos
2024	Febrero	3,99782E+13	2	NEA	K	0,41	71.360,4	0	0	0	1	Kioscos/Maxikioscos
2024	Febrero	3,99782E+13	1	NEA	T	0,05	7.928,9	0	0	0	1	Tradicional
2024	Febrero	3,99782E+13	1	NEA	T	0,09	15.857,9	0	0	0	1	Tradicional
2024	Febrero	4,03492E+13	3	GBA NOROESTE	K	0,14	21.773,7	0	0	0	1	Kioscos/Maxikioscos
2024	Febrero	4,03492E+13	3	GBA NOROESTE	T	0,45	60.014,0	0	0	0	1	Tradicional
2024	Febrero	4,03492E+13	2	GBA NOROESTE	T	0,14	19.394,5	0	0	0	1	Tradicional
2024	Febrero	4,03492E+13	6	GBA NOROESTE	AS	0,71	86.057,0	1	3	0	1	Autoservicio
2024	Febrero	4,03492E+13	4	GBA NOROESTE	AS	1,2	137.595,0	1	2	0	1	Autoservicio
2024	Febrero	4,03492E+13	3	GBA NOROESTE	K	0,5	64.588,05	0	0	0	1	Kioscos/Maxikioscos

Datos de heladeras:

Tabla 13. Muestra de la colocación de equipos de frío de PepsiCo

Primeras 7 columnas

Cod. Región	Desc. Región	COD B2B	Cliente	Domicilio	Canal Mkt	Subcanal Mkt
500	CENTRAL	04755100000019	BELLANDI CARLOS	RAFAEL OBLIGADO 173	Tradicional	Almacen
500	CENTRAL	04755100000031	CANIGIANI GEORGINA	BV ROCA 746	Kioscos/Maxikioscos	Kiosco/Maxikiosc
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso
500	CENTRAL	04755100000058	ESTACION CENTRO S R L	A DADONE 850 P/D: OF 10	COMIDA	Comida al paso

Siguientes 8 columnas

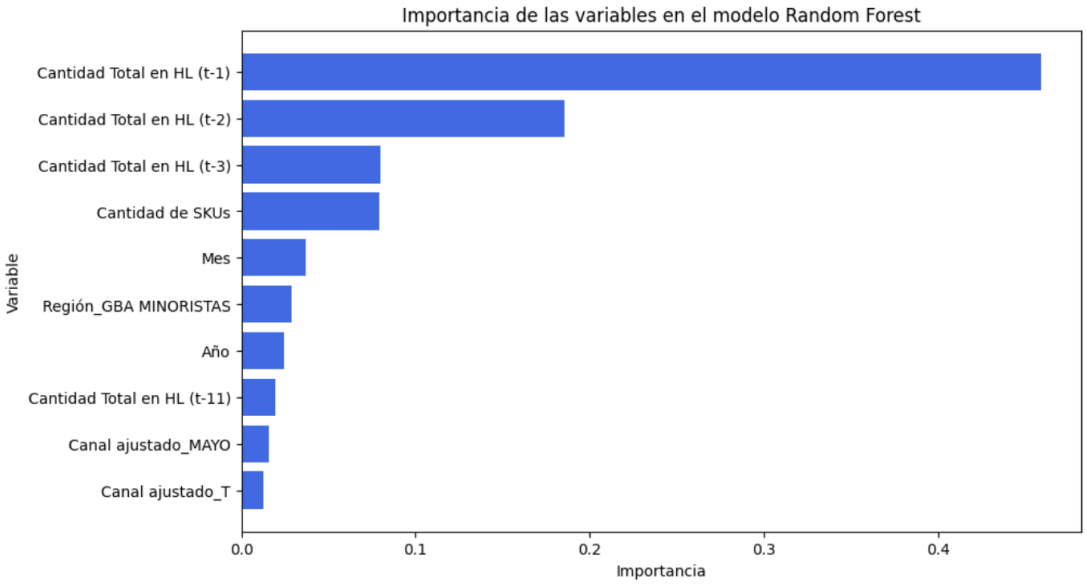
Cod. Producto	Desc. Producto	UN	Logo	Modelo	Nro. Serie	Fec. Colocación / Estado	Puertas
4106	HELADERA GATORADE VERT MEDIANA USADA	UNG	GATORADE	VM	651953	12/19/08	1
83306	HELADERA PEPSI VG REG BC	UNG	PEPSI	VG	90804049	11/21/16	1
4062	HELADERA PEPSI VERT GRANDE USADA	UNG	PEPSI	VG	26841102260027	6/25/21	1
82484	HELADERA PEPSI VERT GDE B/C	UNG	PEPSI	VG	00632864	1/24/22	1
82484	HELADERA PEPSI VERT GDE B/C	UNG	PEPSI	VG	00680098	6/8/22	1
83306	HELADERA PEPSI VG REG BC	UNG	PEPSI	VG	00435458	6/25/21	1
83306	HELADERA PEPSI VG REG BC	UNG	PEPSI	VG	00458351	3/26/14	1
83306	HELADERA PEPSI VG REG BC	UNG	PEPSI	VG	00458371	1/24/22	1
83306	HELADERA PEPSI VG REG BC	UNG	PEPSI	VG	00458487	9/29/21	1
83306	HELADERA PEPSI VG REG BC	UNG	PEPSI	VG	00458643	3/26/14	1

Apéndice B. Mejores hiperparámetros y variables más relevantes del Random Forest

Los mejores hiperparámetros encontrados para Random Forest con Random Search fueron:

- Número de árboles (n_estimators): 100
- Profundidad máxima (max_depth): 30
- Mínimo de muestras para dividir un nodo (min_samples_split): 8
- Mínimo de muestras en hoja (min_samples_leaf): 2
- Número de características consideradas (max_features): sqrt

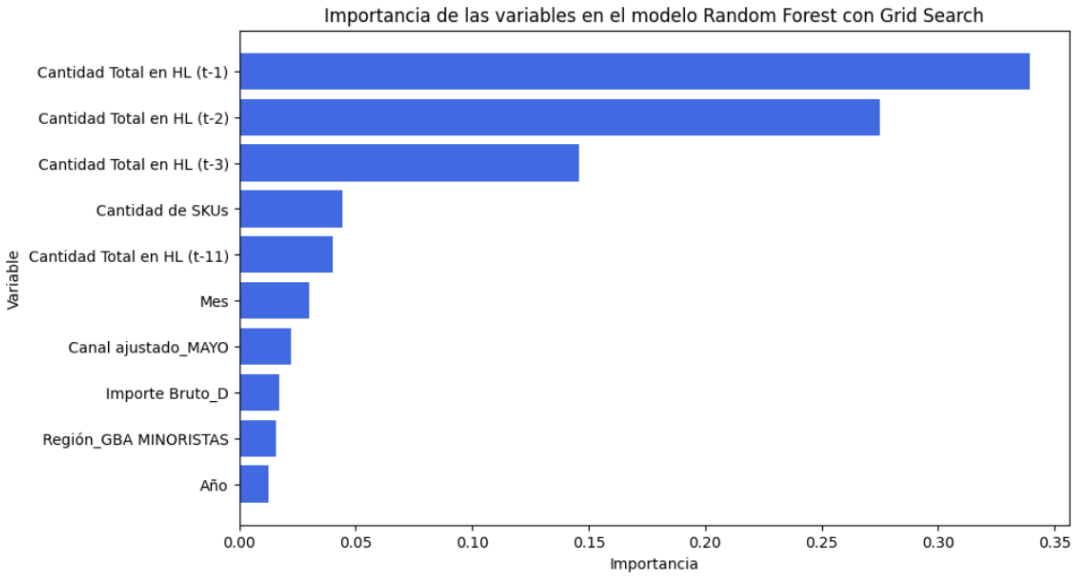
Figura 7. Variables más relevantes del Random Forest con Random Search



La búsqueda en Grid Search con la mejor combinación de hiperparámetros fue:

- Profundidad máxima del árbol (max_depth): 20
- Número máximo de características consideradas en cada división (max_features): 'sqrt'
- Mínimo de muestras en una hoja (min_samples_leaf): 2
- Mínimo de muestras para dividir un nodo (min_samples_split): 2
- Cantidad de árboles en el bosque (n_estimators): 100

Figura 8. Variables más relevantes del Random Forest con Grid Search.



Apéndice C. Mejores hiperparámetros y variables más relevantes del XGBoost

Los mejores hiperparámetros encontrados para XGBoost con Random Search fueron:

- Fracción de características utilizadas (colsample_bytree): 0.6
- Tasa de aprendizaje (learning_rate): 0.05
- Profundidad máxima (max_depth): 7
- Número de árboles (n_estimators): 500
- Fracción de muestras utilizadas (subsample): 0.8
- Parámetro de regularización (gamma): 0.2

Figura 9. Importancia de variables en XGBoost con Random Search según Weight.

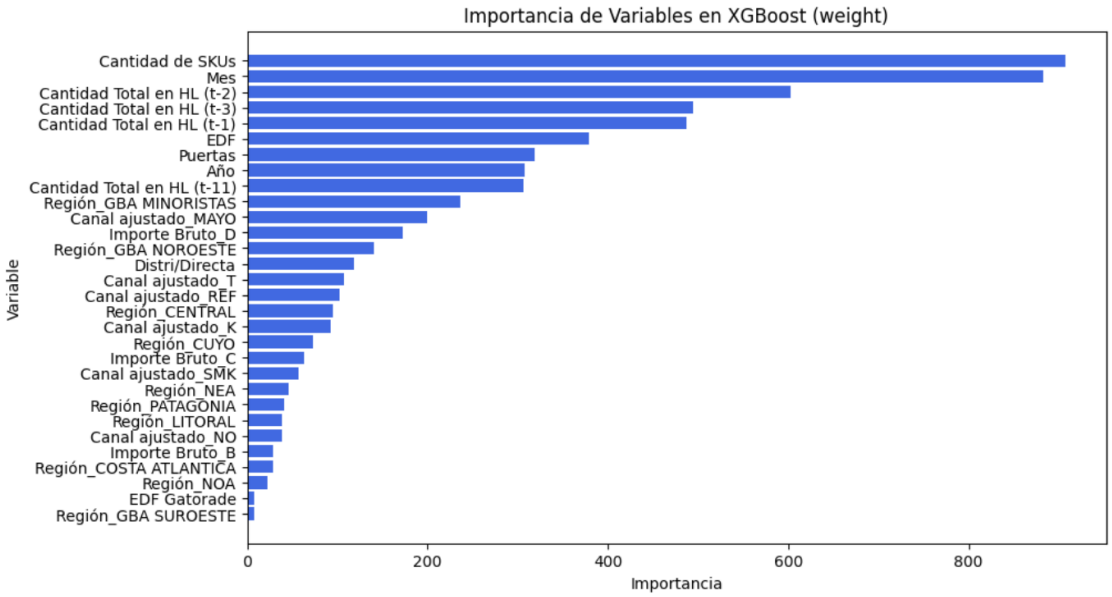
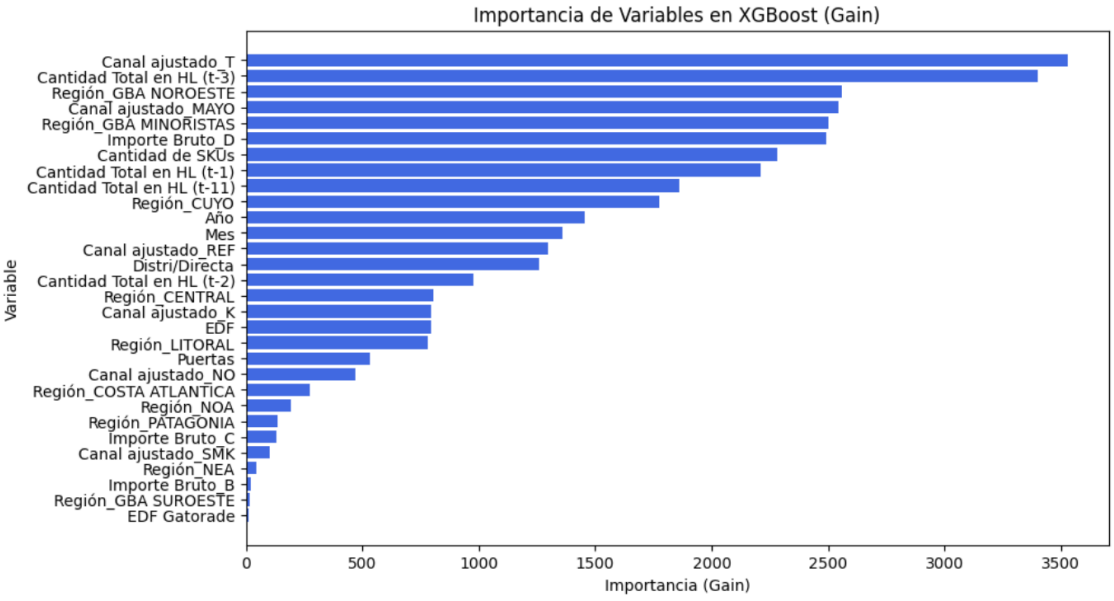


Figura 10. Importancia de variables en XGBoost con Random Search según Gain.



Los mejores hiperparámetros encontrados para XGBoost con Grid Search fueron:

- Fracción de características utilizadas (colsample_bytree): 0.8
- Tasa de aprendizaje (learning_rate): 0.05
- Profundidad máxima (max_depth): 3
- Número de árboles (n_estimators): 100
- Fracción de muestras utilizadas (subsample): 0.8

Figura 11. Importancia de variables en XGBoost con Grid Search según Weight.

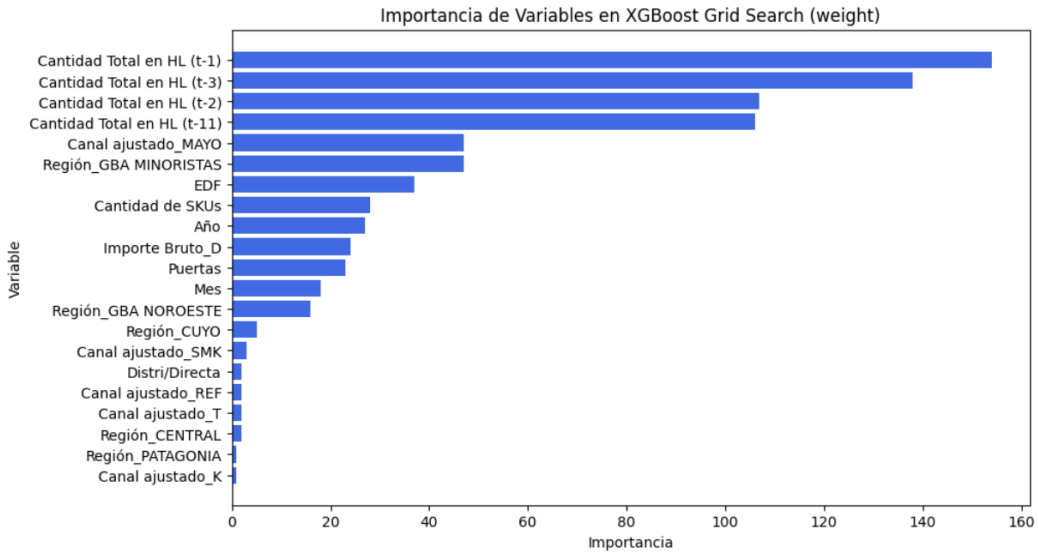
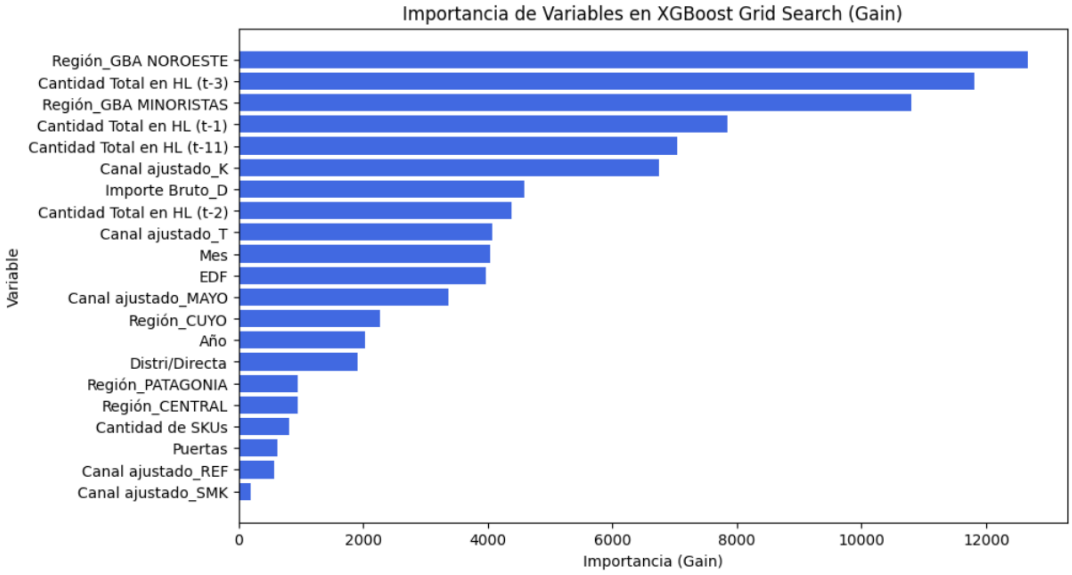


Figura 12. Importancia de variables en XGBoost con Grid Search según Gain.



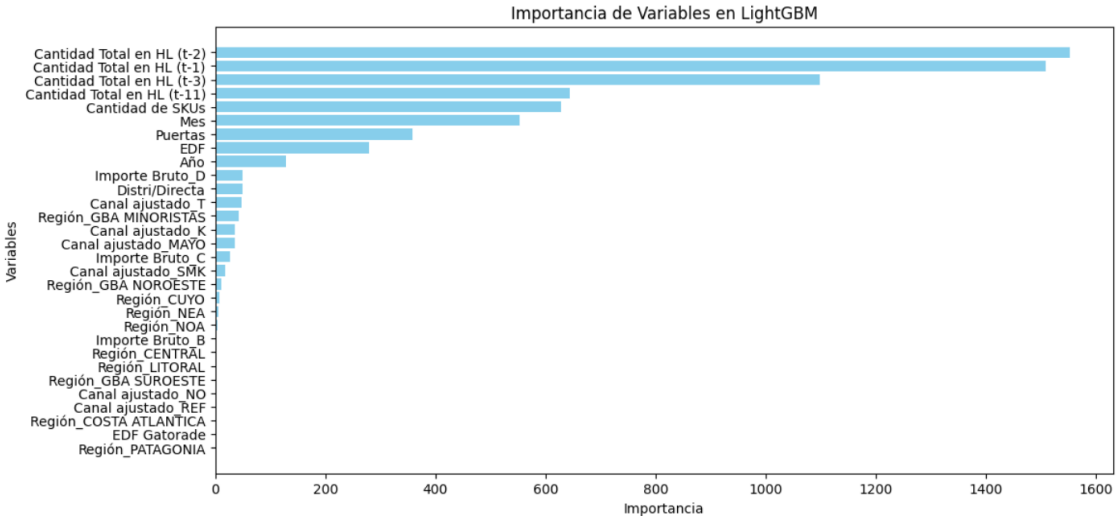
Apéndice D. Variables más relevantes y mejores hiperparámetros del LightGBM

Los mejores hiperparámetros encontrados para LightGBM con Random Forest fueron:

- Número de árboles (n_estimators): 400
- Tasa de aprendizaje (learning_rate): 0.05
- Profundidad máxima (max_depth): 10

- Número máximo de hojas (num_leaves): 20
- Fracción de datos por árbol (subsample): 1.0
- Fracción de características por árbol (colsample_bytree): 1.0

Figura 13. Variables más relevantes del LightGBM con Random Search.



Los mejores hiperparámetros encontrados para LightGBM con Grid Search fueron:

- Número de árboles (n_estimators): 500
- Tasa de aprendizaje (learning_rate): 0.1
- Profundidad máxima (max_depth): 8
- Número máximo de hojas (num_leaves): 31
- Fracción de datos por árbol (subsample): 0.8
- Fracción de características por árbol (colsample_bytree): 0.8

Figura 14. Variables más relevantes del LightGBM con Grid Search.