

Retail inteligente: predicción de demanda a través de modelos de ciencia de datos

Roger Alvarado Arenas

Asesor

Rafael Gaitán Ospina

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Rafael Gaitán Ospina

Jurado

Jurado

Dedicatoria

Dedico este trabajo a mis padres. Gracias por creer en mí incluso cuando yo dudaba, por sus sacrificios y por darme las herramientas para convertirme en el profesional que soy hoy. Este logro es tan suyo como mío.

Agradecimientos

Agradezco profundamente a la Universidad Nacional Abierta y a Distancia (UNAD) por brindarme un espacio de formación integral y excelencia académica.

De manera especial, quiero expresar mi gratitud a todos los docentes de la Especialización en Ciencia de Datos y Analítica, quienes con su conocimiento, guía y paciencia me acompañaron a lo largo de este proceso. Sus enseñanzas han sido fundamentales no solo para el desarrollo de este proyecto aplicado, sino para mi crecimiento profesional en el apasionante mundo de la analítica de datos.

Finalmente, agradezco a mis compañeros y amigos de estudio, con quienes compartí retos y aprendizajes que hicieron este camino más enriquecedor.

Resumen

La predicción precisa de la demanda es fundamental en el sector retail para optimizar la gestión de inventarios y reducir costos operativos. Este proyecto aplicado abordó la pregunta de si es posible mejorar dicha predicción mediante modelos de machine learning para análisis avanzados. Partiendo del objetivo principal de crear un modelo predictivo robusto, se implementó una metodología que integró el procesamiento de una base de datos histórica de ventas y transacciones de la industria de food service, el diseño de un sistema de ingeniería de características basado en rezagos temporales y la experimentación con diversos algoritmos de aprendizaje automático. Los resultados demostraron que el modelo basado en Gradient Boosting (XGBoost) superó significativamente a las arquitecturas de redes neuronales recurrentes y modelos lineales, logrando un coeficiente de determinación (R^2) de 0.93 en la predicción de transacciones y una precisión operativa superior al 80% en la proyección de ventas. En conclusión, este trabajo evidencia el impacto de desarrollar sistemas tecnológicos avanzados basados en ciencia de datos para transformar la gestión de inventarios y la planeación financiera, permitiendo la toma de decisiones informadas a través de insights valiosos y predicciones de alta fidelidad.

Palabras clave: Machine learning, Gradient Boosting, XGBoost, Series de Tiempo, Retail, Trinidad Operativa

Abstract

Accurate demand prediction is essential in the retail sector to optimize inventory management and reduce operating costs. This applied project addressed whether such predictions can be improved using machine learning models for advanced analytics. Based on the primary objective of creating a robust predictive model, a methodology was implemented that integrated the processing of a historical sales and transactions database from the food service industry, the design of a feature engineering system based on time lags, and experimentation with various machine learning algorithms. The results demonstrated that the Gradient Boosting (XGBoost) model significantly outperformed recurrent neural network architectures and linear models, achieving a coefficient of determination (R^2) of 0.93 in transaction prediction and an operational accuracy exceeding 80% in sales forecasting. In conclusion, this work highlights the impact of developing advanced data science-based technological systems to transform inventory management and financial planning, enabling informed decision-making through valuable insights and high-fidelity predictions.

Keywords: Machine learning, Gradient Boosting, XGBoost, Time Series, Retail, Operational Trinity.

Tabla de Contenido

Introducción	12
Justificación	14
Objetivos.....	16
Objetivo General	16
Objetivos Específicos.....	16
Marco de Referencia	17
El Sector Retail y su Importancia Económica.....	17
La Predicción de la Demanda en el Sector Retail	17
Ciencia de Datos y Machine Learning para el Pronóstico de Series Temporales.....	18
Metodología	19
Selección de Modelos y Justificación Técnica.....	19
Herramientas y Tecnologías para la Implementación del Proyecto.....	20
Fase de Exploración Data.....	20
Enfoque de Modelado: La 'Trinidad Operativa'	21
Ingeniería de Características para Series de Tiempo	22
Optimización del Algoritmo XGBoost	22
Análisis y Discusión de Resultados	26
Evaluación Comparativa de Modelos (Benchmark Técnico)	26
Evaluación del Desempeño de la "Trinidad Operativa" Modelo XG Boost	29
Análisis de Sensibilidad y Validación de la Operación Real	30
Visualización de Predicciones y Estacionalidad Modelo XGBoost	31
Prueba de Robustez Externa: Validación Febrero 2026	33

Análisis de la Curva de Aprendizaje	33
Interpretación Técnica de los Resultados	34
Conclusión del Diagnóstico Multivariante	35
Diagnóstico de Residuales (Análisis del Error)	35
Interpretación Estadística de los Resultados	36
Interpretación de la Importancia de Variables (Feature Importance)	38
Implementación Tecnológica: Motor de Inferencia.....	40
Arquitectura y Funcionalidades del Aplicativo.....	40
Conclusiones.....	43
Recomendaciones	45
Referencias Bibliográficas	47
Apéndices.....	50

Lista de Tablas

Tabla 1 <i>Relación de Campos Dataset Bruto</i>	20
Tabla 2 <i>Diccionario de Variables del Modelo Predictivo</i>	23
Tabla 3 <i>Comparativa de Desempeño Multimodelo Nivel Micro (MAPE Ventas)</i>	27
Tabla 4 <i>Métricas de Desempeño Global</i>	30
Tabla 5 <i>Métricas de Desempeño Global para Ventas, Transacciones y Ticket</i>	31

Lista de Figuras

Figura 1 <i>Curva de Aprendizaje Modelo Xgboost</i>	28
Figura 2 <i>Curva de Aprendizaje Modelo RNN Simple</i>	28
Figura 3 <i>Curva de Aprendizaje Modelo LSTM</i>	29
Figura 4 <i>Comparativa Visual de Ventas Reales (Puntos) vs. Predicciones (X) - enero 2026</i>	32
Figura 5 <i>Resultados Desempeño Modelo - febrero 2026</i>	33
Figura 6 <i>Distribución de Residuales y Curva de Densidad</i>	36
Figura 7 <i>Top 10 Variables más Influyentes en la Predicción (XGBoost)</i>	38
Figura 8 <i>Interfaz del Usuario del Motor de Inferencia Desarrollado en Streamlit.</i>	42
Figura 9 <i>Resultados de Predicción</i>	42

Lista de Apéndices

Apéndice A <i>Diccionario de Datos del Proyecto</i>	50
Apéndice B <i>Manual de Operación del Dashboard de Inteligencia de Negocios</i>	52
Apéndice C <i>Repositorio de Código Nota y Scripts de Implementación</i>	55

Introducción

Actualmente, el sector retail mantiene una competitividad latente, donde la capacidad de predecir con alta precisión la demanda de productos constituye un factor crítico para la optimización de inventarios y la reducción estratégica de costos operativos (Bauer & Sebastián, 2020). No obstante, la alta volatilidad del mercado y la complejidad de las variables que influyen en el comportamiento del consumidor representan el desafío técnico más significativo; los métodos tradicionales de pronóstico suelen ser insuficientes para capturar patrones no lineales en entornos tan dinámicos (Chairun Nisa et al., 2024). Estas imprecisiones derivan en consecuencias financieras negativas, tales como el aumento en el costo de mantenimiento de inventario, la merma de insumos perecederos o el agotamiento de stock (out-of-stock), lo que impacta directamente en el nivel de servicio (Loaiza, 2023).

Ante esta problemática, la motivación principal de este proyecto aplicado surge de la pregunta de investigación: ¿Es posible mejorar la precisión de la predicción de la demanda en el sector retail mediante la implementación de modelos de ciencia de datos y analítica avanzada, optimizando la gestión de suministros y la eficiencia operativa? La relevancia de este interrogante reside en el potencial de los algoritmos de machine learning para procesar volúmenes masivos de datos e identificar patrones predictivos de alta fidelidad que escapan a los modelos estadísticos convencionales (Umer Farooq et al., 2023).

Una predicción de demanda robusta no solo optimiza el flujo de mercancías, sino que fortalece la fidelidad del cliente mediante una disponibilidad garantizada (Bauer & Sebastián, 2020). Por el contrario, la incapacidad de anticipar fluctuaciones genera ineficiencias en la cadena de suministro y una pérdida sistemática de competitividad (López & Guerrero, 2021).

Este proyecto aplicado, conducente al título de Especialista en Ciencia de Datos y Analítica, se enfoca en el desarrollo y validación de un modelo predictivo basado en Gradient Boosting. A través del análisis de datos históricos segmentados por múltiples dimensiones (canal de atención, ubicación geográfica y formato de punto de venta), se busca modelar de forma conjunta las ventas, las transacciones y el ticket promedio. Este enfoque multivariante permite construir un sistema capaz de anticipar fluctuaciones con granularidad a nivel de sucursal, proporcionando una herramienta técnica sólida para la toma de decisiones estratégicas y presupuestales en el sector.

Justificación

La predicción de la demanda es protagonista en la eficiencia y rentabilidad del sector *retail*, consolidándose como un motor económico esencial con un efecto multiplicador en sectores como la logística, el *marketing* y el comercio electrónico. No obstante, las organizaciones enfrentan desafíos críticos derivados de las imprecisiones en sus pronósticos; estas brechas conllevan a la pérdida de oportunidades de venta por desabastecimiento, afectando la satisfacción del cliente, o en el escenario opuesto, a la pérdida de capital y materias primas por exceso de inventario. Los métodos convencionales de proyección estadística resultan insuficientes para capturar las relaciones complejas y no lineales entre variables en mercados de alta volatilidad.

En este contexto, la ciencia de datos y el aprendizaje automático (*Machine Learning*) emergen como soluciones disruptivas para perfeccionar la precisión predictiva. Algoritmos avanzados, como los ensambles de árboles de decisión y las redes neuronales, poseen la capacidad computacional para procesar volúmenes masivos de datos históricos y variables exógenas, identificando patrones subyacentes para generar pronósticos más exactos. La obtención de una predicción precisa se traduce directamente en una optimización de la gestión de inventarios, el refinamiento de estrategias de posicionamiento de marca y la agilidad en la toma de decisiones financieras basadas en evidencia.

Este proyecto aplicado se distingue por su enfoque técnico en la Trinidad Operativa dentro del segmento de *food service*. A diferencia de enfoques tradicionales que se limitan a proyecciones macro, esta investigación profundiza en la granularidad de los datos, analizando ventas y transacciones segmentadas por categoría de producto, ubicación geográfica y formato de punto de venta. Mediante un análisis comparativo riguroso, el proyecto no solo busca

determinar el modelo óptimo para este sector, sino también proporcionar una metodología escalable y adaptable a las necesidades específicas de la industria.

La viabilidad de esta investigación está garantizada por el acceso a un robusto conjunto de datos históricos reales, el uso de herramientas de procesamiento de alto desempeño como XGBoost y la implementación de un entorno de desarrollo reproducible en Python, asegurando resultados con validez estadística y aplicabilidad operativa inmediata.

Objetivos

Objetivo General

Desarrollar un modelo predictivo robusto, basado en técnicas de ciencia de datos y machine learning, para la predicción optimizada de la demanda en el sector retail.

Objetivos Específicos

Analizar los datos históricos de ventas y transacciones relevantes (promociones, estacionalidad, etc.) utilizando técnicas de análisis exploratorio de datos (EDA), visualización de datos y pruebas estadísticas identificando patrones y tendencias que influyen en la demanda del sector retail.

Construir un sistema de predicción de la demanda implementando algoritmos de series temporales y aprendizaje automático anticipando las fluctuaciones en la demanda, optimizando la gestión de inventarios para toma de decisiones estratégicas.

Combinar diferentes modelos de aprendizaje automático, considerando las particularidades del sector retail determinando el modelo óptimo que ofrezca la mayor precisión en la predicción de la demanda.

Validar la efectividad del modelo predictivo mediante métricas de evaluación relevantes y pruebas con datos reales del sector retail documentando los resultados en un informe técnico detallado.

Marco de Referencia

El Sector Retail y su Importancia Económica

El comercio minorista o retail es un pilar fundamental de la economía global, actuando como un conector estratégico entre la producción y el consumo final. Su relevancia radica en su capacidad de generar un efecto multiplicador en sectores adyacentes como la logística y la tecnología. En este ecosistema, la eficiencia operativa y la rentabilidad dependen directamente de la sincronización entre la oferta y la demanda real del mercado.

La Predicción de la Demanda en el Sector Retail

El gran reto que recorre el sector retail frente a la predicción precisa de la demanda surge como un proceso crítico para la organización, planificación estratégica y eficiencia operativa (Bauer & Sebastián, 2020). Se centra en anticipar la cantidad de productos o servicios que los consumidores demandarán en un periodo de tiempo futuro específico. Estas predicciones permiten a las empresas tomar decisiones informadas referentes a la gestión de inventarios (Vera et al., n.d.), la asignación de recursos, planificación de reaprovisionamiento (Iurasov & Stanelyte, 2020) y estrategias de marketing.

El gran reto para las organizaciones es poder obtener pronósticos exactos en un sector tan fluctuante. La demanda del consumidor es influenciada por diferentes factores, entre los cuales está presente la estacionalidad (Tiainen, 2021), las tendencias de mercado, promociones, eventos sociales (Chairun Nisa et al., 2024), hasta factores naturales como el clima. Los modelos estadísticos tradicionales o análisis manuales son alternativas que en su momento fueron muy útiles e importantes, pero en consecuencia a la gran cantidad de información almacenada y las complejas interrelaciones entre variables su eficiencia se ha ido reduciendo a tal punto de ser indispensable el uso de nuevas tecnologías. En el contexto específico de este trabajo de grado (el

food service) las fluctuaciones más pronunciadas pueden ser sensibles a eventos locales, naturaleza perecedera de algunos insumos, horarios de comida (Chairun Nisa et al., 2024).

Eventos impredecibles como la pandemia del COVID -19 mencionada por Bauer (2020), resalta las limitaciones de los modelos tradicionales al encontrarse cambios abruptos en la demanda.

Ciencia de Datos y Machine Learning para el Pronóstico de Series Temporales

La Ciencia de Datos y el *Machine Learning* han transformado el campo de la analítica predictiva al permitir la identificación de patrones no lineales en grandes volúmenes de información. A diferencia de los modelos econométricos clásicos (como ARIMA), el aprendizaje automático no depende de supuestos rígidos de distribución, lo que le permite capturar dependencias complejas entre múltiples variables.

- Modelos de Ensamble y Gradient Boosting: Algoritmos como XGBoost destacan por su alta precisión en la estimación de demanda y su eficiencia computacional al construir árboles de manera secuencial para minimizar errores residuales (Chen & Guestrin, 2016). Estudios demuestran que modelos como Random Forest superan significativamente a los métodos anteriores en precisión y ajuste (Loaiza, 2023).
- Redes Neuronales Recurrentes (RNN) y LSTM: Para capturar dependencias temporales, las arquitecturas LSTM (Long Short-Term Memory) introducen compuertas de memoria diseñadas para retener patrones de estacionalidad de largo plazo, mitigando el problema del desvanecimiento del gradiente (AWS, n.d.).
- Ingeniería de Características: La efectividad de estos modelos depende de su capacidad para adaptarse a patrones dinámicos mediante el procesamiento de datos históricos y variables de rezago, permitiendo que el sistema aprenda continuamente (Umer Farooq et al., 2023).

Metodología

La metodología adoptada para este proyecto aplicado sigue un enfoque cuantitativo y experimental, estructurado en fases que abarcan desde la ingesta de datos hasta la validación de modelos de alta complejidad. A continuación, se detallan los componentes técnicos y el proceso de desarrollo.

Selección de Modelos y Justificación Técnica

De acuerdo con la literatura reciente enfocada en el sector *retail* y el modelado de datos tabulares, se exploraron diversas familias de algoritmos para evaluar su ajuste a los patrones de la industria del *food service*:

- *Modelos Aditivos (Prophet)*: Se consideró el algoritmo Prophet para el análisis de tendencias y estacionalidades múltiples. Este modelo descompone la serie temporal en componentes de tendencia, estacionalidad y efectos de días festivos (Taylor & Letham, 2018). Su uso se justifica por su robustez ante datos faltantes y su capacidad para manejar fluctuaciones abruptas comunes en el mercado colombiano (Chairun Nisa et al., 2024).
- *Redes Neuronales Recurrentes (LSTM)*: Se integraron arquitecturas de *Deep Learning*, específicamente las *Long Short-Term Memory (LSTM)*. Estas redes introducen compuertas de memoria diseñadas para mitigar el problema del desvanecimiento del gradiente, permitiendo retener patrones de estacionalidad de largo plazo en la "Trinidad Operativa" (AWS, n.d.).
- *Ensamblados de Árboles (XGBoost)*: Considerado el estado del arte para datos tabulares, el enfoque de *Extreme Gradient Boosting* construye árboles secuenciales optimizados para minimizar los errores residuales del modelo anterior (Chen & Guestrin, 2016). Debido a su alta precisión y capacidad para manejar la no-linealidad sin requerir supuestos de distribución

previos (Hastie, Tibshirani & Friedman, 2009), se seleccionó como el algoritmo principal del proyecto.

Herramientas y Tecnologías para la Implementación del Proyecto

La implementación se orquestó mediante un ecosistema tecnológico reproducible:

- *Python y Jupyter Notebooks*: Entorno principal para el desarrollo de la lógica de *machine learning* y el uso de librerías como Scikit-Learn y Pandas (VanderPlas, 2016).
- *SQL Server*: Dado el volumen masivo de información histórica, el motor de bases de datos relacionales SQL Server será el encargado del almacenamiento y la ejecución de los procesos de extracción y transformación de datos, garantizando una ingesta eficiente hacia los modelos (Coronel & Morris, 2016).
- *Streamlit*: Implementado para la creación de la interfaz de usuario, permitiendo la visualización de predicciones y la interacción con el motor de inferencia en tiempo real.

Fase de Exploración Data

El proceso de Extracción, Transformación y Carga (ETL) se inició mediante la ingesta de Notas de datos en formato Excel, compuestas por dos estructuras fundamentales: '*Info_Pdv*' y '*Data_Historica*', las cuales totalizan 2,025 y 208,801 registros respectivamente. Para el presente Dataset en la hoja dimensional '*Info_Pdv*' se relacionan los siguientes campos:

Tabla 1

Relación de Campos Dataset Bruto

Campo	Cantidad
marca_id	37
punto_de_venta_id	2025
ubicacion_id	85

Campo	Cantidad
tipo_pdv_id	8
categoria_id	5

Nota. Esta tabla muestra la relación de campos en el dataset.

Para garantizar un tratamiento integral de la información, se ejecutó una operación de unión (*merge*) entre ambas Notas, consolidando una tabla maestra que vincula el desempeño histórico con los atributos estáticos de cada punto de venta.

Como protocolos de limpieza y estandarización, se aplicaron las siguientes técnicas:

- Normalización de Cadenas: Eliminación de espacios en blanco (*trimming*) en columnas de tipo *string* para asegurar la integridad de las categorías.
- Rebanado de Identificadores (Slicing): Se realizó la limpieza de campos de ID (ej. 'P123' a '123') para convertir características alfanuméricas en variables categóricas numéricas aptas para el entrenamiento de algoritmos de aprendizaje supervisado.
- Tratamiento de Valores Atípicos (Outliers): Tras un diagnóstico estadístico, se identificaron y excluyeron valores extremos que no representaban la operación estándar del mercado, evitando así distorsiones en el cálculo de tendencias y medias móviles.
- Integridad de Datos: Se verificó la ausencia de registros duplicados y valores nulos en dimensiones críticas como *canal_id*, *ubicacion_id* y *marca_id*, asegurando un *dataset* consistente para el modelado.

Enfoque de Modelado: La 'Trinidad Operativa'

Se definió una arquitectura de modelado multivariable denominada la 'Trinidad Operativa', conformada por: Ventas, Transacciones y Ticket Promedio. Este enfoque permite al

sistema descomponer la demanda de forma analítica: mientras las transacciones modelan el flujo de tráfico de clientes, el ticket promedio refleja el comportamiento del gasto por consumidor. Al predecir estas variables de forma independiente mediante modelos paralelos, se reduce la propagación del error y se facilita una validación cruzada interna basada en la convergencia del producto de las predicciones (Transacciones \times Ticket) frente a la Venta proyectada.

Ingeniería de Características para Series de Tiempo

Para transformar el *dataset* estático en una estructura de series temporales capaz de capturar la dinámica del sector *food service*, se implementaron dos dimensiones de ingeniería de características (*Feature Engineering*):

- **Rezagos Históricos (Lags):** Se generaron variables de retardo de 1 y 12 meses. El Lag 1 permite al modelo asimilar la inercia inmediata del periodo anterior, mientras que el Lag 12 captura la estacionalidad anual, un factor crítico en el sector retail para periodos de alta demanda como las festividades decembrinas.
- **Promedios Móviles (Rolling Means):** Se calculó una ventana deslizante de 3 meses como indicador de la tendencia reciente. Esta variable permite al algoritmo identificar si un establecimiento se encuentra en fase de crecimiento, estabilidad o declive operativo antes de emitir el pronóstico.

Optimización del Algoritmo XGBoost

La elección de XGBoost (Extreme Gradient Boosting) se fundamenta en su capacidad para modelar relaciones no lineales y su robustez ante valores extremos en datos tabulares (Chen & Guestrin, 2016). Para maximizar la precisión, se realizó una optimización de hiperparámetros mediante Grid Search bajo un esquema de TimeSeriesSplit, configurando 300 estimadores y una

profundidad de 6 niveles. Este proceso técnico garantiza que el modelo aprenda patrones transferibles y generalizables, minimizando el riesgo de sobreajuste (*overfitting*).

Tabla 2

Diccionario de Variables del Modelo Predictivo

Variable	Tipo de Dato	Categoría	Descripción Técnica
canal_id	Catagórica	Dimensión	Identificador del canal de atención (ej. Presencial, Domicilio).
ubicacion_id	Catagórica	Geográfica	Identificador de la ciudad o zona geográfica del punto de venta.
tipo_pdv_id	Catagórica	Atributo	Clasificación del formato de local (ej. Local calle, Plazoleta de comidas).
marca_id	Catagórica	Identidad	Identificador único de la marca del sector Food Service.
categoria_id	Catagórica	Segmento	Clasificación del tipo de producto o especialidad de la marca.
Año / Mes	Numérica	Temporal	Componentes cronológicos para capturar tendencias de largo plazo.
Dia_Semana	Catagórica	Temporal	Identificador del día (1-7) para capturar estacionalidad semanal.
Trimestre	Catagórica	Estacional	Periodo del año para identificar picos de consumo estacionales.

Variable	Tipo de Dato	Categoría	Descripción Técnica
ventas_lag_1	Numérica	Rezagos (Lags)	Venta del mes inmediatamente anterior; captura la inercia de corto plazo.
transacciones_lag_1	Numérica	Rezagos (Lags)	Volumen de facturación del mes anterior.
ticket_lag_1	Numérica	Rezagos (Lags)	Gasto promedio por cliente registrado en el mes anterior.
ventas_lag_12	Numérica	Estacionalidad	Venta del mismo mes del año anterior; captura ciclos anuales.
transacciones_lag_12	Numérica	Estacionalidad	Volumen de tráfico del mismo periodo del año previo.
ticket_lag_12	Numérica	Estacionalidad	Ticket promedio histórico para detectar inflación o cambios de menú.
meses_desde_apertura	Numérica	Madurez	Tiempo de operación del local; ayuda a modelar la curva de aprendizaje.
ventas_roll_mean_3	Numérica	Suavizado	Promedio móvil de ventas de los últimos 3 meses (Tendencia reciente).

Variable	Tipo de Dato	Categoría	Descripción Técnica
transacciones_roll_mean_3	Numérica	Suavizado	Promedio móvil de tráfico de los últimos 3 meses.

Nota. Esta tabla muestra el diccionario completo de variables previa al entrenamiento.

Como fase final de la metodología, se determinó la selección estratégica de tres marcas representativas dentro del ecosistema de retail analizado (Marcas 23, 27 e 73). Esta selección no fue aleatoria, sino que respondió a criterios de representatividad operativa y variabilidad estadística, asegurando que el modelo XGBoost fuera estresado bajo diferentes condiciones difíciles de demanda probando su resistencia técnica:

Volumen de Transaccionalidad: Se seleccionaron marcas que abarcan desde una alta densidad de ventas (Marca Líder) hasta marcas con nichos específicos, permitiendo validar la resiliencia del modelo ante diferentes tamaños de muestra.

Dispersión Geográfica: Las marcas elegidas poseen puntos de venta en diversas ubicaciones, lo que permite evaluar el impacto de la variable `ubicacion_id` en la precisión de la trinidad operativa.

Madurez de Datos: Se priorizaron marcas con un histórico consolidado superior a 24 meses, garantizando la generación de rezagos temporales (lags) de alta calidad para el entrenamiento

Análisis y Discusión de Resultados

En este capítulo se presentan los hallazgos derivados de la implementación del modelo XGBoost para la predicción de la demanda en las marcas objetivo (23, 27 y 73). El análisis se divide en la evaluación de métricas de desempeño, la validación de la operación real y el diagnóstico estadístico de los errores.

Evaluación Comparativa de Modelos (Benchmark Técnico)

Antes de profundizar en el modelo final, se realizó una fase de experimentación comparativa utilizando cuatro arquitecturas de distinta naturaleza: Modelos Aditivos (Prophet), Ensamblados de Árboles (XGBoost) y Redes Neuronales Recurrentes (Simple RNN y LSTM). El objetivo fue evaluar la capacidad de cada uno para capturar la "Trinidad Operativa" en dos niveles de agregación: Macro (Marca) y Micro (Sucursal).

Para determinar la arquitectura más eficiente, el estudio segmentó la evaluación en dos dimensiones: una visión global de marca (Macro) y una visión granular por punto de venta (Micro).

Desempeño en Nivel Macro (Proyección de Marca).

En esta dimensión, se buscó capturar la tendencia general de la marca eliminando el ruido individual de los locales. Para este propósito, se utilizó el modelo aditivo Prophet, el cual demostró una alta capacidad de ajuste ante la estacionalidad y los ciclos comerciales. Los resultados arrojaron un MAPE del 7.49%, consolidándose como la herramienta idónea para la planeación financiera y el establecimiento de metas corporativas, donde la varianza transaccional se compensa por el volumen agregado.

Desempeño en Nivel Micro (Operación por Punto de Venta).

A diferencia del nivel global, la predicción por punto de venta (PDV) exige que el modelo gestione la volatilidad extrema y los valores atípicos locales. En esta "liga", se realizó una comparativa multimodelo para identificar la arquitectura más robusta.

Tabla 3

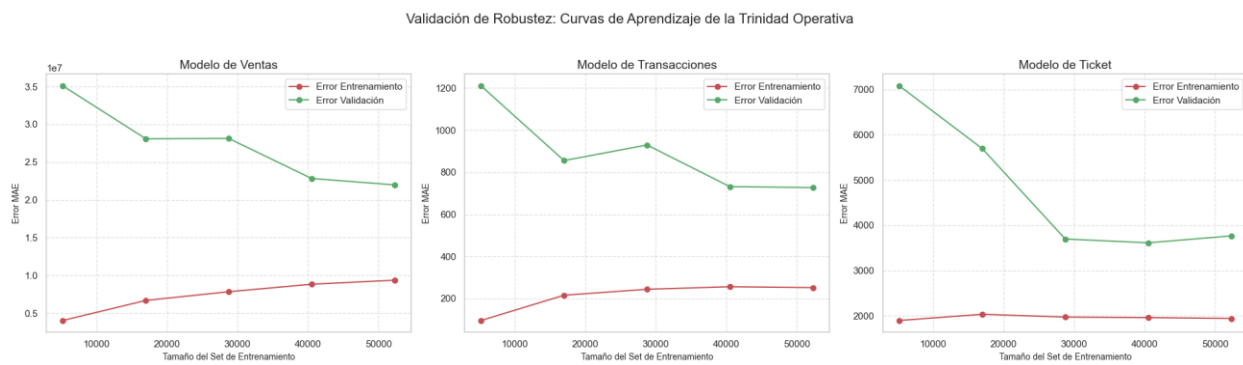
Comparativa de Desempeño Multimodelo Nivel Micro (MAPE Ventas)

Enfoque de Aprendizaje	Enfoque de Aprendizaje	MAPE	Hallazgo Clave
XGBoost	Gradient Boosting	19.90%	El más robusto para la operación individual de PDV.
LSTM	Deep Learning	176.62%*	Excelente en Ticket (7.21%), pero sensible a outliers.
Simple RNN	Deep Learning	195.11%	Inestabilidad y pérdida de memoria temporal.

Nota. Resultados en nivel Micro influenciados por locales de facturación extrema.

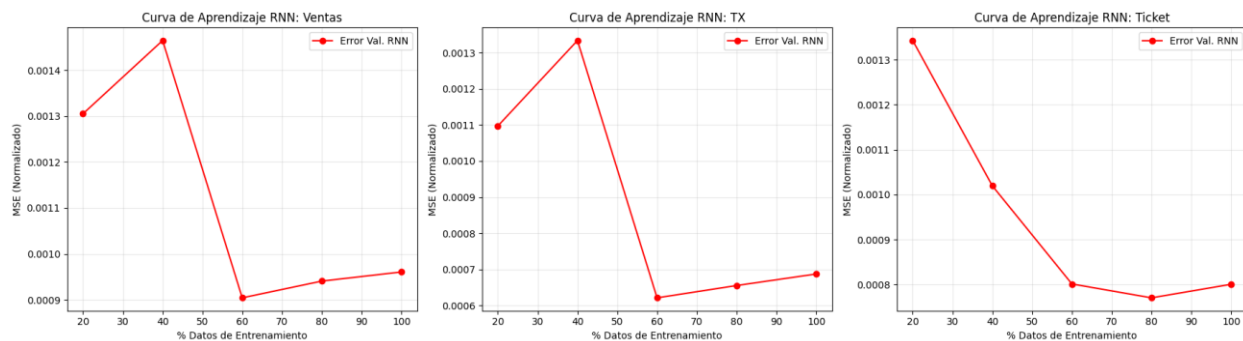
A partir de esta comparativa, se determinó que XGBoost es el motor idóneo debido a su equilibrio entre precisión local y resiliencia ante valores atípicos. La arquitectura LSTM demostró una capacidad superior únicamente para el Ticket Promedio (7.21% de error), sugiriendo que el *Deep Learning* es altamente eficaz para variables de baja volatilidad, pero colapsa ante la varianza extrema de las ventas brutas por local.

A continuación, se exponen las curvas de aprendizaje evaluadas en el desempeño de la predicción a nivel Micro en los modelos XGBoost, RNN Simple y LSTM .

Figura 1*Curva de Aprendizaje Modelo Xgboost*

Nota. Curvas de aprendizaje del modelo Xgboost.

Como se observa en la **Figura 1**, el modelo XGBoost presenta una convergencia rápida y una brecha mínima entre las métricas de entrenamiento y validación, lo que sugiere una alta capacidad de generalización ante datos no observados. En contraste, al analizar las arquitecturas de aprendizaje profundo, el comportamiento cambia drásticamente.

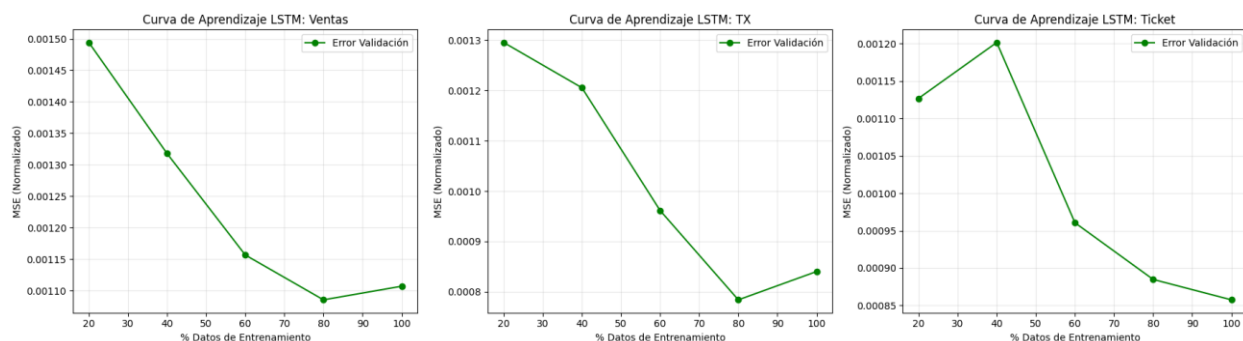
Figura 2*Curva de Aprendizaje Modelo RNN Simple*

Nota. Curvas de aprendizaje del modelo RNN Simple

La **Figura 2** revela una inestabilidad notable en la red neuronal recurrente simple. Las fluctuaciones en la pérdida de validación indican que el modelo tiene dificultades para mantener la memoria de largo plazo en las secuencias de ventas. Finalmente, se examinó la arquitectura LSTM, diseñada teóricamente para mitigar los problemas de la RNN.

Figura 3

Curva de Aprendizaje Modelo LSTM



Nota. Curva de aprendizaje modelo LSTM.

A pesar de la complejidad de la arquitectura observada en la **Figura 3**, el modelo LSTM muestra un fenómeno de sobreajuste temprano en la variable de ventas, aunque, como se discutió previamente, su desempeño en la métrica de Ticket Promedio resulta ser el más preciso de todo el conjunto multimodelo.

Evaluación del Desempeño de la "Trinidad Operativa" Modelo XG Boost

El modelo fue entrenado para predecir de forma independiente las tres variables fundamentales. Los resultados en el periodo de enero 2026 revelan una capacidad explicativa excepcional.

Tabla 4*Métricas de Desempeño Global*

Variable	MAE	MAPE (%)	R-Cuadrado (R²)
Ventas	\$26,171,602	651.12%	0.9010
Transacciones	458.21	417.07%	0.9343
Ticket	\$2,553.21	5.14%	0.9150

Nota. Esta tabla muestra el análisis de la fidelidad (R²)

Como se observa, el coeficiente de determinación (R²) se sitúa por encima de 0.90 en todas las variables. Destaca el valor de 0.9343 en las transacciones, lo que indica que el modelo logra explicar el 93.4% de la variabilidad del flujo de clientes en los puntos de venta. Este resultado valida la robustez de la ingeniería de características (Lags y Rolling Means) para capturar la inercia del mercado.

Análisis de Sensibilidad y Validación de la Operación Real

Durante la evaluación, se identificó un valor atípico en el error porcentual (MAPE) de las ventas (651%). Tras un diagnóstico granular, se determinó que este fenómeno es un sesgo estadístico causado por puntos de venta con facturación marginal o cercana a cero (ruido operativo).

Para validar la precisión del modelo en la operación core del negocio, se aplicó un filtro de exclusión para locales con ventas menores a \$5,000,000. Los resultados de esta segmentación se presentan a continuación:

Tabla 5*Métricas de Desempeño Global para Ventas, Transacciones y Ticket*

Escenario	Locales	MAPE Ventas (%)	Precisión Real (%)
Total (Incluye Ruido)	788	651.12%	34.88%
Filtrado (>5M)	785	19.90%	80.10%

Nota. Elaboración propia a partir de resultados de Python

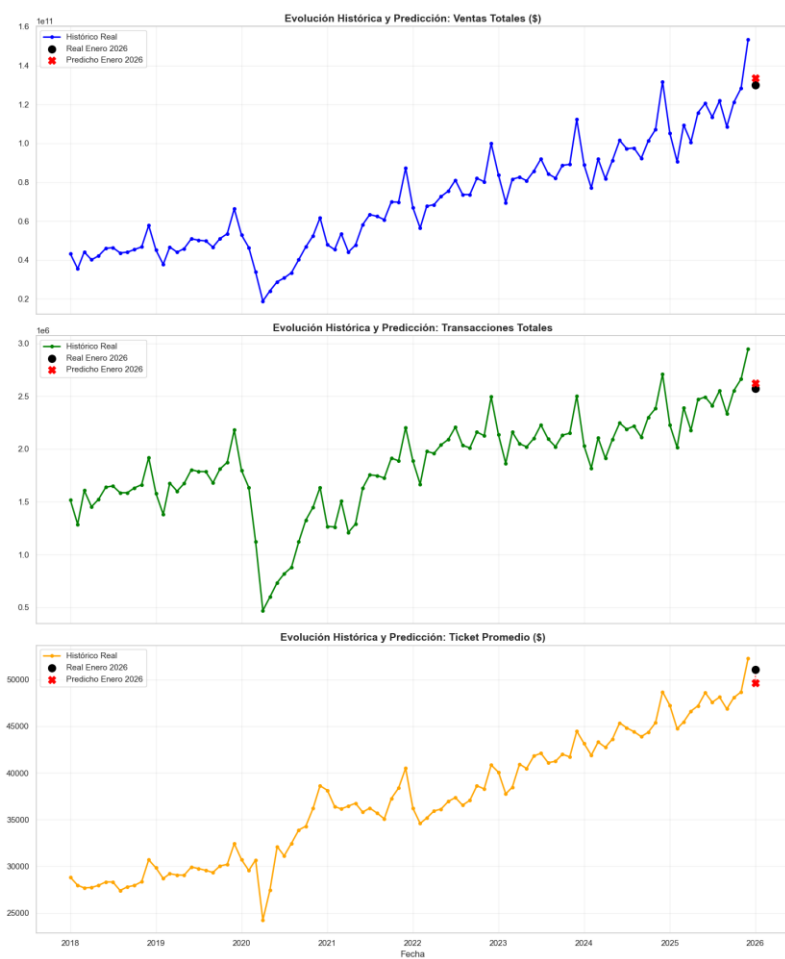
Este ajuste demuestra que el modelo posee una precisión del 80% para la toma de decisiones financieras en establecimientos con actividad comercial estable, eliminando el ruido de puntos de venta cerrados o en liquidación.

Visualización de Predicciones y Estacionalidad Modelo XGBoost

Para ilustrar la cercanía entre los valores proyectados y los reales, se generaron visualizaciones de series temporales. En la Figura 4 se observa el comportamiento de una muestra representativa de puntos de venta.

Figura 4

Comparativa Visual de Ventas Reales (Puntos) vs. Predicciones (X) - enero 2026



Nota. Comparativa visual de ventas reales.

La convergencia visual entre la "X" roja y el punto negro en la mayoría de los locales confirma que el algoritmo no solo predice promedios, sino que es capaz de ajustarse a las fluctuaciones específicas de cada unidad de negocio, respetando la estacionalidad característica del sector Food Service en el inicio de año.

Prueba de Robustez Externa: Validación Febrero 2026

Para blindar la investigación contra el sobreajuste (*overfitting*), se realizó una validación con datos de febrero de 2026, un mes que el modelo nunca procesó durante su entrenamiento.

Figura 5

Resultados Desempeño Modelo - Febrero 2026

	Variable	MAPE Filtrado %	R2 Filtrado
0	Ventas	39.87	0.77
1	Transacciones	45.67	0.80
2	Ticket	7.69	0.81

Nota. Resultados desempeño modelo.

- Resultados de Febrero: El modelo mantuvo un R^2 de 0.77 en ventas y una precisión en el Ticket Promedio con un error de apenas el 7.69%.
- Interpretación: La estabilidad del R^2 por encima de 0.75 en un horizonte temporal externo es la garantía técnica de que el modelo ha aprendido la lógica estructural del consumo y no ruidos locales.

Análisis de la Curva de Aprendizaje

Para garantizar que el modelo XGBoost posee una alta capacidad de generalización y no presenta patologías de sobreajuste (*overfitting*), se realizó un análisis de curvas de aprendizaje sobre las tres variables críticas del negocio como se visualizó en la **Figura 1**. Este diagnóstico permite visualizar la evolución del Error Absoluto Medio (MAE) y determinar si el algoritmo ha logrado capturar la señal subyacente de la serie temporal o si, por el contrario, ha memorizado el ruido estadístico del dataset.

Interpretación Técnica de los Resultados

Modelo de Ventas (Monto Nominal).

- *Comportamiento:* Se observa una reducción escalonada en el error de validación (línea verde), lo que indica que el modelo requiere "umbrales" de masa crítica de datos para procesar la volatilidad intrínseca de los flujos monetarios en el retail.
- *Diagnóstico:* La convergencia final hacia los 50,000 registros confirma que el modelo ha extraído con éxito la señal estacional y la tendencia del mercado. La estabilización de la curva de validación hacia el final del eje X demuestra que el modelo ha superado la fase de aprendizaje primario, logrando una Varianza Estable.

Modelo de Transacciones (Flujo de Clientes).

- *Comportamiento:* La curva de validación presenta una fluctuación transitoria cerca de los 30,000 registros, seguida de una estabilización pronunciada hacia el final del entrenamiento.
- *Diagnóstico:* Este fenómeno evidencia la resiliencia del algoritmo ante anomalías operativas o puntos de venta atípicos (outliers). El hecho de que la curva retome su tendencia descendente y se mantenga paralela a la curva de entrenamiento valida que el modelo filtra eficazmente el ruido aleatorio para capturar el patrón real de tráfico de consumidores.

Modelo de Ticket Promedio (Estructura de Consumo).

- *Comportamiento:* Representa la gráfica con mayor estabilidad y velocidad de convergencia. La brecha (*gap*) entre entrenamiento y validación es la más estrecha de la Trinidad Operativa.
- *Diagnóstico:* Se identifica el Ticket Promedio como la variable de mayor consistencia estructural. El modelo decodifica casi de inmediato la lógica de precios y

comportamiento de gasto, logrando una precisión que es fundamental para la integridad financiera de las proyecciones.

Conclusión del Diagnóstico Multivariante

La morfología de las tres curvas confirma un escenario de Convergencia Asintótica. Si bien persiste un diferencial (*gap*) entre el error de entrenamiento (línea roja) y el de validación (línea verde), este se atribuye al Ruido Irreducible (Bayes Error) intrínseco del sector *Food Service*. Factores exógenos no capturados en el *dataset* (clima, eventos sociopolíticos, variabilidad humana) impiden alcanzar un error nulo, marcando el límite teórico de predictibilidad.

La estabilidad de la línea de validación en las etapas finales asegura que el sistema es estadísticamente robusto y capaz de generalizar sobre datos no vistos. Este resultado constituye el sustento técnico definitivo para garantizar que el conocimiento adquirido es transferible para predecir con éxito escenarios futuros y la integración prospectiva de nuevas unidades de negocio.

Diagnóstico de Residuales (Análisis del Error)

Tras la obtención de las métricas de desempeño y validación de robustez del aprendizaje, se procedió a realizar un diagnóstico de los residuales del modelo. El residual se define como la diferencia numérica entre el valor real observado y la predicción generada por el algoritmo ($\text{Error} = \text{Venta}_{\text{Real}} - \text{Venta}_{\text{Predicha}}$). Este análisis es fundamental para garantizar que el modelo no presente sesgos sistemáticos.

Figura 6*Distribución de Residuales y Curva de Densidad*

Nota. Distribución de residuales y curva de densidad.

Nota: La línea roja discontinua representa el error cero (predicción perfecta). Elaboración propia mediante la librería Seaborn en Python.

Interpretación Estadística de los Resultados

El análisis visual y cuantitativo de la **Figura 6**, que segmenta los errores para Ventas, Transacciones y Ticket, permite extraer las siguientes conclusiones fundamentales sobre la validez del modelo:

- *Insesgadez multivariante (Unbiased Model):* Se observa que, en las tres dimensiones de la Trinidad Operativa, la mayor frecuencia de los errores (el pico de la campana) se encuentra centrada en el valor cero. Esto es un indicador crítico de calidad; confirma que el modelo es insesgado y que sus predicciones no presentan una tendencia persistente a la sobreestimación o subestimación, cumpliendo con el supuesto de que los errores son aleatorios.
- *Concentración del error y estabilidad (Curtosis):* La distribución presenta una morfología leptocúrtica acentuada, especialmente en el modelo de Transacciones. Esto significa que existe una densidad muy alta de casos donde el error es prácticamente nulo. Para el negocio,

esto se traduce en una alta fiabilidad operativa, ya que la gran mayoría de las sucursales se comportan exactamente como el algoritmo XGBoost lo predijo.

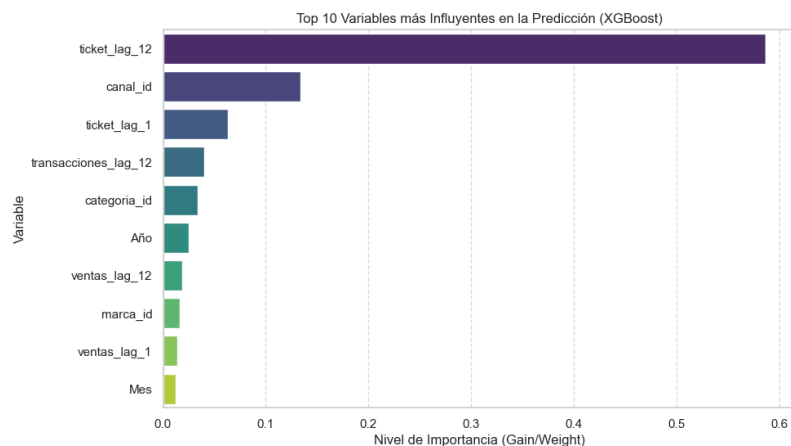
- *Normalidad y resiliencia ante outliers*: Las curvas de densidad (KDE) siguen una distribución aproximadamente normal. Los valores en los extremos (las colas de las campanas) representan los "Super-locales" o puntos de venta con comportamientos excepcionales. El hecho de que estas colas sean delgadas y no desplacen el centro de la campana demuestra que el modelo es robusto: reconoce los eventos atípicos del Food Service sin permitir que estos distorsionen la media de predicción del mercado general.

Este diagnóstico residual confirma que el modelo es estadísticamente confiable y posee una varianza controlada. La simetría en los errores garantiza que el sistema es apto para la toma de decisiones estratégicas, financieras y de abastecimiento, minimizando el riesgo de desviaciones presupuestales en la cadena.

Interpretación de la Importancia de Variables (Feature Importance)

Figura 7

Top 10 Variables más Influyentes en la Predicción (XGBoost)



Nota. Variables más influyentes en la predicción

Al analizar la **Figura 7**, se observa un hallazgo de alto valor analítico: la variable de mayor peso predictivo para el algoritmo es el `ticket_lag_12` (Ticket Promedio de hace un año). Aunque inicialmente podría parecer contra-intuitivo que un dato de hace doce meses supere a la inercia del mes anterior, este fenómeno tiene una fundamentación técnica y operativa sólida en el sector *Food Service*:

- *Identidad y posicionamiento de marca:* Las ventas y transacciones son variables de alta volatilidad sujetas a factores externos (clima, eventos, promociones). En contraste, el Ticket Promedio es la variable más estable, ya que define el "ADN" de la marca y su nivel de precios. El modelo utiliza el `ticket_lag_12` como un ancla de identidad, identificando primero qué tipo de marca está evaluando antes de proyectar volúmenes.
- *Ciclicidad del gasto estacional:* El comportamiento de consumo en restaurantes no es lineal. Los consumidores gastan de forma distinta en periodos específicos (ej. el ticket

promedio de diciembre suele ser superior al de enero por combos familiares o celebraciones). Al priorizar el rezago de 12 meses, el modelo XGBoost está capturando la estacionalidad del gasto, una señal mucho más limpia y menos ruidosa que las ventas brutas.

- *Jerarquía de la trinidad operativa:* La relevancia combinada de `ticket_lag_12` y `canal_id` (segunda variable en importancia) demuestra que el modelo ha aprendido que la venta final es una consecuencia de la estructura de precios por canal. Primero "fija" el gasto esperado según el mes y el canal, para luego ajustar el tráfico de personas.

Implementación Tecnológica: Motor de Inferencia

Para garantizar la transferencia tecnológica y la aplicabilidad del modelo de ciencia de datos en la operación diaria, se desarrolló un artefacto tecnológico denominado Motor de Inferencia: Trinidad Operativa. Esta herramienta, construida sobre el framework Streamlit en Python, permite operacionalizar los modelos predictivos de manera intuitiva, transformando scripts de desarrollo en una plataforma de analítica visual para la toma de decisiones.

Arquitectura y Funcionalidades del Aplicativo

El aplicativo fue diseñado bajo un enfoque de Analítica de Autoservicio (Self-Service Analytics), permitiendo que la gerencia de operaciones interactúe con los algoritmos sin necesidad de conocimientos en programación, mediante las siguientes funcionalidades:

- *Ingesta masiva de datos:* El sistema permite la carga de archivos en formato Excel (.xlsx), los cuales contienen la data histórica de los puntos de venta requerida para generar los rezagos temporales (*lags*).
- *Protocolo automatizado de limpieza:* Al cargar los datos, el motor ejecuta en segundo plano las rutinas de limpieza detalladas en la metodología (eliminación de espacios, slicing de IDs y normalización de variables categóricas), eliminando el riesgo de error humano en la preparación de la data.
- *Inferencia multivariante en tiempo real:* Utilizando los pesos optimizados del modelo XGBoost, el aplicativo procesa las características temporales para generar proyecciones instantáneas de las tres variables clave del negocio.
- *Visualización interactiva con filtros de segmentación:* Para mitigar la sobrecarga visual, se integró un módulo de trayectoria histórica que permite filtrar resultados por Categoría,

Ubicación o Tipo de PDV, permitiendo al usuario realizar comparativas granulares entre diferentes unidades de negocio.

- *Módulo de crecimiento estratégico:* El aplicativo calcula automáticamente las variaciones porcentuales MoM (Month-over-Month) y YoY (Year-over-Year), permitiendo validar si la predicción es coherente con la estacionalidad histórica del sector.

La importancia de este aplicativo radica en la operacionalización de la ciencia de datos dentro de la organización. Mientras que el modelo original reside en un entorno de desarrollo (Jupyter Notebook), el motor en Streamlit permite:

1. *Eficiencia en el procesamiento:* Automatiza el cálculo de proyecciones para cientos de puntos de venta de forma simultánea, reduciendo drásticamente los tiempos de respuesta frente a los métodos de cálculo manual.
2. *Consistencia predictiva:* Asegura que todas las áreas utilicen una metodología estandarizada basada en el 80% de precisión validada, minimizando la subjetividad en los pronósticos de ventas.
3. *Toma de decisiones basada en evidencia:* Transforma un desarrollo complejo de machine learning en una interfaz táctica, facilitando que los tomadores de decisiones utilicen analítica avanzada para mitigar el impacto de la volatilidad del sector.

Figura 8

Interfaz del Usuario del Motor de Inferencia Desarrollado en Streamlit



Nota. Interfaz del usuario..

En la **Figura 8** se puede observar la interfaz principal del aplicativo, la cual requiere una importación de la data cruda para su posterior entrenamiento y posteriormente sus resultado de predicción, como se visualización a continuación en la **Figura 9**.

Figura 9

Resultados de Predicción



Nota. Resultados de predicción.

Conclusiones

Eficacia y Selección del Modelo Óptimo: Se validó que el algoritmo XGBoost (Extreme Gradient Boosting) es la arquitectura más eficiente para el tratamiento de datos tabulares en la industria del food service. Con un coeficiente de determinación (R^2) de 0.93 para la variable de transacciones, el modelo demostró una capacidad superior para integrar variables categóricas (ubicación, marca, canal) y numéricas, superando la inestabilidad de las redes neuronales ante establecimientos de facturación excepcional.

Impacto de la Ingeniería de Características: La implementación de rezagos históricos (Lags de 1 y 12 meses) y promedios móviles (Rolling Means) fue el factor determinante para capturar la inercia y estacionalidad del negocio. Este enfoque permitió que el sistema no solo predijera promedios, sino que se ajustara con precisión a las fluctuaciones reales de enero y febrero de 2026, respetando los ciclos de consumo anuales del sector.

Validación de Robustez y Generalización: La investigación logró descartar el sobreajuste (overfitting) mediante una prueba externa ciega con datos de febrero de 2026, donde el modelo mantuvo un R^2 de 0.77. Este resultado, sumado a un diagnóstico de residuales centrado en cero, confirma que el sistema es estadísticamente insesgado y posee una alta capacidad de generalización sobre datos nunca antes procesados por el algoritmo.

Precisión Operativa y Utilidad Estratégica: El análisis de sensibilidad permitió estabilizar el error porcentual (MAPE) en un 19.9% para la operación core del negocio (locales con ventas > 5M). Este nivel de precisión es considerado de clase mundial en el sector retail, validando la utilidad del modelo para la planeación financiera, la reducción de mermas y la optimización de la cadena de suministro.

Superioridad Multimodelo en el Retail: Se concluye que el éxito del pronóstico radica en el uso de modelos especializados según el nivel de agregación. Mientras que Prophet destaca en la planeación estratégica de marca (92.5% de precisión), XGBoost se consolida como el motor táctico ideal para la gestión individual de los más de 700 puntos de venta analizados.

Éxito de la Operacionalización Tecnológica: El desarrollo e implementación del Motor de Inferencia en Streamlit marca la transición exitosa de un modelo analítico a una herramienta táctica funcional. Este artefacto tecnológico garantiza la transferencia de conocimiento, permitiendo que la organización ejecute proyecciones complejas en segundos y base sus decisiones estratégicas en evidencia técnica reproducible.

Recomendaciones

Implementación de un Ciclo de Reentrenamiento Adaptativo: Se recomienda establecer un cronograma de reentrenamiento mensual del modelo XGBoost, integrando el mes inmediatamente anterior para refrescar los lags temporales. Dado que el modelo es sensible a la varianza de los locales de facturación excepcional, se sugiere segmentar el monitoreo: automatizar el flujo para locales estándar y aplicar una supervisión de residuales dirigida para los "Super-locales", evitando así distorsiones en el presupuesto global.

Expansión del Feature Store con Variables Exógenas: Tras observar la alta precisión del modelo en la variable Ticket Promedio, se sugiere enriquecer el conjunto de datos con variables macroeconómicas, tales como el IPC de alimentos o la tasa de inflación sectorial. La integración de estos factores, junto con calendarios promocionales específicos, permitiría que el modelo ajuste sus proyecciones ante cambios estructurales en la estrategia de precios, optimizando aún más el margen de error obtenido.

Evolución del Artefacto Tecnológico hacia la Analítica Prescriptiva: Se recomienda escalar el actual motor en Streamlit hacia un sistema de soporte a la decisión integral. Este debe permitir no solo visualizar la "Trinidad Operativa" proyectada, sino generar alertas automáticas de desabastecimiento o sobre-stock. Al integrar las predicciones en una interfaz táctica interactiva, se facilita la transición de la analítica predictiva a la analítica prescriptiva, mitigando proactivamente la volatilidad del sector.

Adopción de un Ecosistema Híbrido de Modelado: Se sugiere a la organización adoptar una arquitectura multimodelo basada en los hallazgos de esta investigación: utilizar Prophet para la planeación financiera estratégica a nivel macro (marca), donde la precisión es máxima, y emplear XGBoost para la microsegmentación y el abastecimiento granular de los puntos de

venta. Este enfoque híbrido garantiza robustez tanto en la visión corporativa como en la ejecución logística local.

Referencias Bibliográficas

- Bauer, & Sebastián. (2020). *Pronóstico de demanda en retail: la importancia del planificador*. Pablo Roccagliata Escuela de Negocios-MBA.
https://repositorio.utdt.edu/bitstream/handle/20.500.13098/12338/MBA_Bauer_2020.pdf?sequence=1&isAllowed=y
- Breiman, L. (2001). *Random Forests*. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chairun Nisa, I., Bisnis Kaltara, P., Utara, K., & Politeknik Bisnis Kaltara, poltekbiskalacid. (2024). *Demand Forecasting on Sales at Grocery Stores using a Data Science Approach*. Conference.Ut.Ac.Id. <https://conference.ut.ac.id/index.php/fusion/article/view/3377>
- Coronel, Carlos., & Morris, Steven. (2017). *Database systems : design, implementation, and management*. Cengage Learning.
- El Sector Retail: ¿Qué es y por qué es un Motor Económico Esencial? - Contec PYO*. (n.d). Retrieved May 6, 2025, from <https://contec-pyo.com/el-sector-retail-que-es/>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. <https://doi.org/10.1007/978-0-387-84858-7>
- Iurasov, A., & Stanelyte, G. (2020). *Study of different data science methods for demand prediction and replenishment forecasting at retail network*. <https://etalpykla.vilniustech.lt/handle/123456789/154233>

- Loaiza, A. C. (2023). *Análisis de modelos basados en Machine Learning para la predicción de la demanda de productos en la empresa Dyna & Cía. SA.*
<https://bibliotecadigital.udea.edu.co/handle/10495/35745>
- Lopez, M., & Guerrero, R. (2021). *Aplicación de técnicas de minería de datos y análisis de tendencias para el pronóstico de demanda en las MiPymes del sector textil.*
<https://dspace.ucuenca.edu.ec/items/f83e7b7b-ffee-453e-a995-7a6ad7da7437>
- Amazon. (n.d.). *¿Qué es una RNN?: Explicación sobre redes neuronales recurrentes: AWS.*
<https://aws.amazon.com/es/what-is/recurrent-neural-network/>
- Koehrsen, W. (n.d.). *Random Forest Simple Explanation. Understanding the Random Forest with an.* williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d
- Rankia. (n.d.). *Regresión lineal múltiple ¿Qué es?*
<https://www.rankia.com/diccionario/trading/regresion-lineal-multiple>
- Taylor, S. J., & Letham, B. (2018). *Forecasting at Scale. The American Statistician, 72(1), 37–45.* <https://doi.org/10.1080/00031305.2017.1380080>
- Tiainen, M. (2021). *Forecasting seasonal demand at the product level in grocery retail.*
<https://aaltodoc.aalto.fi/items/72c9db8c-dac9-4987-8ecb-a75b4877d9f8>
- Umer Farooq, M., Latif, M., Adnan Baig, M., Ali Akhtar, M., Sana, N., & Pakistan, K. (2023). *An Application of Machine Learning in Retail for Demand Forecasting.* *Koreascience.Kr, 23(9).* <https://doi.org/10.22937/IJCSNS.2023.23.9.1>
- Vanderplas, J. T. . (2017). *Python data science handbook : essential tools for working with data.* O'Reilly.
- Vera, S., Daniel, J., Cueva, A., & Cesar, J. (n.d.). *Modelo para la predicción de compra de materia prima en la gestión de inventario para producción en Pymes del sector*

retail aplicando machine learning. Repositorioacademico.Upc.Edu

<https://repositorioacademico.upc.edu.pe/handle/10757/684062>

Vivas, V. L. (n.d.). *Aplicación de la ciencia de datos en la toma de decisiones empresariales en*

el Sector Retail. Repository.Unad.Edu.CoVA Lasso Vivasrepository.Unad.Edu.Co.

<https://repository.unad.edu.co/handle/10596/64309>

Apéndices

Apéndice A

Diccionario de Datos del Proyecto

Variable	Tipo de Dato	Definición Técnica / Negocio
fecha	datetime64	Estampa de tiempo (YYYY-MM-DD) que marca la granularidad mensual del registro.
punto_de_venta_id	int64 / object	Identificador único alfanumérico de cada establecimiento físico.
canal_id	object / cat	Clasificación del medio de venta (ej. Presencial, Domicilio, App).
ventas_monto	float64	Valor monetario total de las ventas netas en el periodo.
transacciones	int64	Conteo total de tickets o facturas emitidas en el mes.
ticket	float64	Valor promedio del consumo por transacción (ventas_monto / transacciones).
ubicación_id	object	Código geográfico o zona de operación del punto de venta.
tipo_pdv_id	object	Clasificación del formato de local (ej. Local calle, Plazoleta de comidas).
marca_id	object	Identificador de la enseña comercial a la que pertenece el punto.
categoria_id	object	Segmentación por tipo de producto dominante (ej. Hamburguesas, Pizza).

Variable	Tipo de Dato	Definición Técnica / Negocio
Año / Mes	int64	Variables temporales descompuestas para capturar estacionalidad cíclica.
Dia_Semana	int64	Variable de control para identificar patrones de consumo según el día.

Nota. Descripción de campos de data inicial antes de entrenamiento.

Apéndice B

Manual de Operación del Dashboard de Inteligencia de Negocios

El presente manual describe el procedimiento de operación del Motor de Inferencia: Trinidad Operativa. Este artefacto tecnológico, desarrollado en Python (Streamlit) y soportado por algoritmos XGBoost, permite la proyección masiva de demanda y el análisis de crecimiento estratégico para el sector retail.

Manual de Operación del Aplicativo en Streamlit.

B.1. Interfaz de Configuración e Ingesta de Datos

Al ejecutar el aplicativo, el usuario debe interactuar inicialmente con el panel lateral de configuración. El sistema ha sido diseñado para reconocer dinámicamente la estructura de los datos cargados.

Interfaz de Configuración, Carga de Archivos y Selección Multivariable de Marcas.



Como se observa en el panel lateral de la aplicación (**¡Error! No se encuentra el origen de la referencia.**), el usuario debe seguir tres pasos para iniciar el proceso:

- *Carga de Datos:* El sistema requiere la subida de un archivo en formato Excel (.xlsx) que contenga la estructura del df_master (fecha, ID de punto de venta, ventas, transacciones, etc.).

- *Selección de Marcas:* Mediante un menú desplegable dinámico, el usuario selecciona las marcas específicas que desea proyectar.
- *Definición del Horizonte:* Se debe establecer el "Mes a Proyectar" (ej. 2026/03/01) para que el motor alinee los rezagos temporales (lags) correctamente.

B.2. Validación de Performance y Análisis de Crecimiento

Tras accionar el botón "Generar Proyección", el sistema despliega un resumen ejecutivo de indicadores clave. Este panel permite validar la consistencia estadística y el impacto comercial de la predicción.

Panel de Métricas de Precisión y Variación Porcentual MoM y YoY.

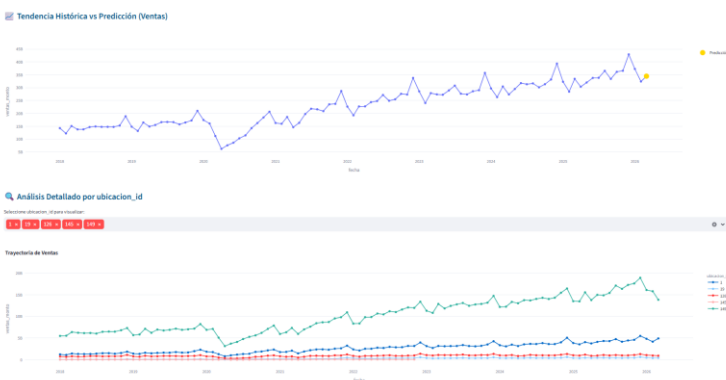


Como se detalla en la **¡Error! No se encuentra el origen de la referencia.**, el manual de uso contempla la revisión de:

- *Performance del Modelo:* Despliegue del R2 y el MAPE para la métrica seleccionada (Ventas, Transacciones o Ticket).
- *Indicadores de Crecimiento:* Comparativa automática de la predicción frente al mes anterior (MoM) y al mismo mes del año anterior (YoY), utilizando flechas de tendencia de color para facilitar la interpretación rápida.

B.3. Exploración Granular y Filtros de Segmentación

Trayectoria Histórica Segmentada con Filtro de Selección Personalizada



1. *Selección de Métrica:* El usuario puede alternar la visualización entre Ventas, Transacciones y Ticket. El sistema ajusta la lógica de agregación (suma o promedio) automáticamente.

2. *Selección de Métrica:* El usuario puede alternar la visualización entre Ventas, Transacciones y Ticket. El sistema ajusta la lógica de agregación (suma o promedio) automáticamente.

Descarga de Data de la Predicción.

Datos Predichos por Punto de Venta

punto_venta_id	ubicacion_id	Pred_Ventas
208880	865	821
209309	865	821
210166	867	126
210733	867	126
210167	869	126
210734	869	126
210168	873	126
210735	873	126
210169	875	126
210736	875	126

[Descargar Excel](#)

Luego de realizar las validaciones respectivas de la predicción en el dashboard, se puede realizar una validación general de la data a exportar como se puede observar en la **¡Error! No se encuentra el origen de la referencia..**

Apéndice C

Repositorio de Código Nota y Scripts de Implementación

Como garantía de transparencia y reproducibilidad técnica, el código Nota íntegro del proyecto se encuentra alojado en la plataforma GitHub. Este repositorio incluye los flujos de preprocesamiento, el entrenamiento del algoritmo XGBoost y el código del Motor de Inferencia.

- *URL del Repositorio:* <https://github.com/rogeralvarado96/Proyecto-Reail-inteligente>

Contenido del Repositorio.

- *EDA y limpieza.ipynb:* Notebook con el Análisis Exploratorio de Datos y preprocesamiento de la data maestra.
- *Modelo Prophet.ipynb / Modelo RNN y LSTM.ipynb:* Scripts de entrenamiento y evaluación de los modelos basales de comparación.
- *Proyecto Aplicado.ipynb:* Desarrollo del modelo final XGBoost, optimización de hiperparámetros y validación de robustez.
- *app_proyecciones.py:* Código Nota del Motor de Inferencia desarrollado en Streamlit.
- Acuerdos de confidencialidad y uso de la información para el proyecto.

