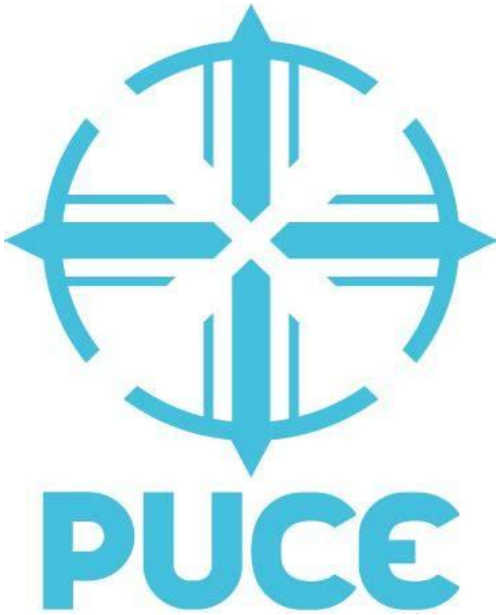
Pontificia Universidad Católica del Ecuador

Facultad De Ingeniería



Tema:

Modelo predictivo de ventas empresariales mediante técnicas de Machine Learning y análisis de series temporales utilizando datos históricos de demanda

Autor:

María Paula de la Cruz Padilla

Director:

Carlos Julio Mayorga

Proyecto de Titulación previo a la obtención del título de Magister en Sistemas de la Información Mención Data Science

Quito – Agosto 2026

Índice de Contenido

1. Introducción 8
1.1. Generalidades 9
1.2. Planteamiento del problema 10
1.3. Formulación del problema 12
1.3.1. Preguntas secundarias 12
1.4. Justificación 12
1.5. Objetivos 13
1.5.1. Objetivo General 13
1.5.2. Objetivos Específicos 13
1.6. Relación entre objetivos y preguntas de investigación 14
1.7. Alcance 14
2. Marco Teórico y Revisión de Literatura 15
2.1. Introducción 15
2.2. Predicción de ventas y forecasting empresarial 16
2.3. Series temporales 17
2.4. Machine Learning aplicado a la predicción de ventas 18
2.5. Gradient Boosting 18
2.6. Light Gradient Boosting Machine (LightGBM) 19
2.7. Ingeniería de características temporales 20
2.7.1. Variables derivadas de la fecha 20
2.7.2. Variables rezagadas — Lag Features 21
2.7.3. Promedios móviles — Rolling Mean 21
2.7.4. Media móvil exponencialmente ponderada — EWMA 22
2.8. Pronóstico recursivo en series temporales 22
2.9. Métricas de evaluación 23
2.9.1. Error Absoluto Medio — MAE 23
2.9.2. Raíz del Error Cuadrático Medio — RMSE 24
2.9.3. Error Porcentual Absoluto Medio — MAPE 24
2.9.4. Coeficiente de determinación — R^2 24
2.10. Metodología CRISP-DM 25

2.11. Enfoques alternativos utilizados en forecasting 26
3. Marco Metodológico 27
3.1. Enfoque de la investigación 27
3.2. Tipo de investigación 27
3.3. Diseño de la investigación 28
3.4. Metodología de desarrollo 28
3.5. Fuente de datos 29
3.6. Descripción de las variables originales 30
3.7. Herramientas utilizadas 31
3.8. Comprensión y exploración inicial de los datos 31
3.9. Preparación de los datos 32
3.10. Ingeniería de características 32
3.10.1. Características de calendario 33
3.10.2. Variables Lag 33
3.10.3. Promedios móviles 34
3.10.4. Variables EWMA 35
3.11. Variables utilizadas por el modelo 35
3.12. Tratamiento de valores faltantes generados por rezagos 36
3.13. División temporal de los datos 36
3.14. Justificación de la selección del modelo LightGBM 37
3.15. Construcción del modelo LightGBM 38
3.16. Configuración de hiperparámetros del modelo LightGBM 38
3.17. Evaluación del modelo 39
3.18. Generación del pronóstico para 2018 40
3.18.1. Construcción del primer día del pronóstico 41
3.18.2. Incorporación de las predicciones al historial 41
3.18.3. Validación inicial del procedimiento recursivo 42
3.18.4. Automatización del horizonte completo 43
3.19. Consolidación del pronóstico 43
3.20. Análisis posterior del pronóstico 44
3.21. Procedimiento general de la investigación 45

3.22. Consideraciones metodológicas 45
4. Resultados y Discusión 46
4.1. Introducción del capítulo 46
4.2. Comprensión de los datos 47
4.2.1. Descripción inicial del conjunto de datos 47
4.2.2. Tipos de variables y estadísticas descriptivas 48
4.2.3. Calidad de los datos 49
4.2.4. Período temporal analizado 49
4.2.5. Evolución temporal de las ventas 50
4.2.6. Evolución anual de las ventas 51
4.2.7. Comportamiento mensual de las ventas 52
4.2.8. Comportamiento de las ventas por día de la semana 53
4.2.9. Análisis de ventas por tienda 55
4.2.10. Análisis de ventas por producto 56
4.2.11. Relación entre tienda y producto 57
4.3. Preparación de los datos 58
4.3.1. Variables temporales generadas 58
4.3.2. Construcción de variables Lag 59
4.3.3. Valores faltantes generados por los rezagos 59
4.3.4. Construcción de Rolling Mean 60
4.3.5. Construcción de variables EWMA 61
4.3.6. Eliminación de valores nulos y dimensiones finales 62
4.4. Modelado 62
4.4.1. División temporal entre entrenamiento y validación 63
4.4.2. Selección de las variables predictoras 64
4.4.3. Configuración y entrenamiento de LightGBM 65
4.4.4. Importancia de las variables 65
4.4.5. Evaluación del modelo 67
4.5. Evaluación del desempeño predictivo 68
4.5.1. Comparación con el modelo baseline 68
4.5.2. Error Absoluto Medio — MAE 69

4.5.3. Raíz del Error Cuadrático Medio — RMSE 69
4.5.4. Coeficiente de determinación — R^2 70
4.5.5. Error Porcentual Absoluto Medio — MAPE 70
4.5.6. Comparación entre ventas reales y predichas 71
4.5.7. Análisis de residuos 71
4.5.8. Distribución de los residuos 73
4.5.9. Evaluación agregada por tienda 73
4.5.10. Evaluación agregada por producto 75
4.5.11. Evaluación mensual del período de validación 76
4.5.12. Síntesis del desempeño del modelo 77
4.6. Pronóstico recursivo de ventas para 2018 77
4.6.1. Construcción del horizonte futuro 78
4.6.2. Generación recursiva de las predicciones 79
4.6.3. Consolidación del pronóstico 80
4.6.4. Volumen total de ventas pronosticado 81
4.6.5. Pronóstico mensual 82
4.6.6. Pronóstico por tienda 83
4.6.7. Pronóstico por producto 85
4.6.8. Productos con menor demanda pronosticada 86
4.6.9. Comparación entre ventas reales de 2017 y pronóstico 2018 87
4.6.10. Principales resultados del pronóstico 89
4.7. Discusión de los resultados 89
4.8. Aplicación empresarial de los resultados 91
4.9. Limitaciones de los resultados 92
5. Conclusiones y Recomendaciones 93
5.1. Conclusiones 93
5.2. Recomendaciones 94
6. Referencias 96

Índice de Ilustraciones

Ilustración 1 - Fases de la metodología CRISP-DM 26
Ilustración 2 - Construcción de la matriz predictora correspondiente al primer día del horizonte de pronóstico 41
Ilustración 3 - Esquema del procedimiento de pronóstico recursivo utilizado. 42
Ilustración 4 - Ventas diarias durante el período 2013–2017 50
Ilustración 5 - Ventas totales por año 52
Ilustración 6 – Ventas totales según mes del año 53
Ilustración 7 - Ventas por día de la semana 54
Ilustración 8 - Ventas totales por tienda 55
Ilustración 9 - Ventas totales por producto durante el período 2013–2017 56
Ilustración 10 - Mapa de calor de ventas por tienda y producto 57
Ilustración 11 - División temporal del conjunto de datos para entrenamiento y validación 64
Ilustración 12 - Variables más importantes 67
Ilustración 13 - Comparación entre ventas reales y predichas 71
Ilustración 14 - Residuos del modelo LightGBM 72
Ilustración 15 - Distribución de los residuos 73
Ilustración 16 - Ventas reales y predichas por tienda 74
Ilustración 17 - Ventas reales y predichas por producto 75
Ilustración 18 - Ventas reales y predichas por mes 76
Ilustración 19 - Flujo del procedimiento de pronóstico recursivo utilizado para 2018 80
Ilustración 20 - Verificación del conjunto consolidado de predicciones correspondientes a 2018 81
Ilustración 21 - Pronóstico mensual de ventas para el año 2018 83
Ilustración 22 - Pronóstico de ventas por tienda 2018 84
Ilustración 23 - Productos con mayor volumen de ventas pronosticado para 2018 85
Ilustración 24 - Productos con menor volumen de ventas pronosticado para 2018 86
Ilustración 25 - 2017 Real vs. 2018 Pronóstico 88

Índice de Tablas

Tabla 1 - Aplicación de las fases CRISP-DM en la investigación 29
Tabla 2 - Características generales del conjunto de datos 30
Tabla 3 - Variables originales utilizadas en la investigación 30
Tabla 4 - Herramientas computacionales utilizadas 31
Tabla 5 - Variables rezagadas utilizadas 34
Tabla 6 - Variables de promedio móvil 34
Tabla 7 - Configuración de las variables EWMA 35
Tabla 8 - Composición de las variables predictoras 36
Tabla 9 - División temporal utilizada 37
Tabla 10 - Configuración de hiperparámetros de LightGBM 39
Tabla 11 - Procedimiento metodológico de la investigación 45
Tabla 12 - Estructura inicial del conjunto de datos 47
Tabla 13 - Estadísticos descriptivos de la variable ventas 48
Tabla 14 - Evaluación inicial de calidad de los datos 49
Tabla 15 - Rango temporal del conjunto histórico 50
Tabla 16 - Evolución anual de las ventas 51
Tabla 17 - Comportamiento de las ventas por día de la semana 54
Tabla 18 - Variables temporales generadas 58
Tabla 19 - Construcción de variables Lag 59
Tabla 20 - Valores faltantes generados por los rezagos 60
Tabla 21 - Construcción de Rolling Mean 60
Tabla 22 - Configuración de características EWMA 61
Tabla 23 - Dimensiones antes y después de la preparación 62
Tabla 24 - División temporal entre entrenamiento y validación 63
Tabla 25 - Variables utilizadas para el entrenamiento de LightGBM 64
Tabla 26 - Configuración y entrenamiento de LightGBM 65
Tabla 27 - Importancia de las variables 66
Tabla 28 - Desempeño general del modelo LightGBM 68
Tabla 29 - Comparación del desempeño predictivo entre el modelo baseline y LightGBM 68
Tabla 30 - Evaluación agregada por tienda 74
Tabla 31 - Estructura del horizonte de pronóstico 79
Tabla 32 - Volumen total de ventas pronosticado 82
Tabla 33 - Comparación entre ventas reales de 2017 y pronóstico para 2018 88
Tabla 34 - Resumen de resultados del pronóstico 2018 89

1. Introducción

La disponibilidad creciente de datos empresariales ha transformado la manera en que las organizaciones planifican sus operaciones y toman decisiones. Los registros históricos de ventas constituyen una fuente de información especialmente relevante, debido a que permiten analizar el comportamiento de la demanda, identificar patrones temporales y anticipar posibles escenarios futuros. La capacidad de generar estimaciones confiables puede contribuir a mejorar procesos relacionados con abastecimiento, inventarios, planificación comercial y asignación de recursos.

Sin embargo, la demanda no presenta necesariamente un comportamiento constante a lo largo del tiempo. Factores como la tendencia, estacionalidad, día de la semana, diferencias entre establecimientos y comportamiento específico de los productos pueden generar variaciones que dificultan la elaboración de pronósticos mediante métodos basados únicamente en promedios históricos. Por esta razón, las técnicas de Ciencia de Datos y Machine Learning representan una alternativa para construir modelos capaces de aprender relaciones complejas a partir de grandes volúmenes de información.

En este contexto, la presente investigación desarrolla un modelo predictivo de ventas empresariales mediante técnicas de Machine Learning y análisis de series temporales. Para ello se utiliza el conjunto de datos Demand Forecasting Dataset, conformado por registros diarios correspondientes a diez tiendas y cincuenta productos durante el período comprendido entre enero de 2013 y diciembre de 2017. El conjunto de entrenamiento contiene 913.000 observaciones históricas, organizadas mediante las variables fecha, tienda, producto y ventas. Esta fuente de datos constituye también la base metodológica declarada en el trabajo.

El proceso analítico se estructura mediante la metodología CRISP-DM, considerando las etapas de comprensión del problema, comprensión y preparación de los datos, modelado, evaluación y presentación de resultados. Durante la preparación de los datos se incorporan características temporales derivadas de la fecha, valores rezagados o lags, promedios móviles y medias móviles exponencialmente ponderadas —EWMA—. Estas variables permiten representar tanto la estructura temporal de las ventas como el comportamiento reciente de cada combinación tienda-producto.

Para el modelado se utiliza el algoritmo Light Gradient Boosting Machine (LightGBM), entrenado utilizando información histórica comprendida hasta diciembre de 2016 y evaluado sobre las observaciones correspondientes al año 2017. La separación temporal permite conservar el orden cronológico de los datos y simular un escenario de predicción sobre información posterior a la utilizada durante el entrenamiento.

El desempeño predictivo se analiza mediante el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE), el Error Porcentual Absoluto Medio (MAPE) y el coeficiente de determinación R^2. Adicionalmente, se examina la correspondencia entre ventas reales y predichas a nivel de tienda, producto y mes, así como el comportamiento de los residuos generados por el modelo.

Finalmente, una vez evaluado el modelo, se construye un procedimiento de pronóstico recursivo para estimar las ventas correspondientes al año 2018. En este procedimiento, cada predicción generada se incorpora progresivamente al historial disponible para construir las variables temporales necesarias en los días posteriores. Los resultados permiten analizar la evolución esperada de la demanda y proporcionar información que puede apoyar procesos empresariales de planificación y gestión de inventarios.

1.1. Generalidades

La planificación de la demanda representa una actividad fundamental para organizaciones que administran inventarios, productos o recursos asociados a procesos comerciales. Conocer únicamente el comportamiento histórico de las ventas proporciona una visión retrospectiva; sin embargo, incorporar modelos predictivos permite transformar esta información en estimaciones que contribuyen a anticipar escenarios futuros.

Las ventas empresariales conforman series de observaciones cronológicas en las que pueden coexistir diferentes patrones. Una serie puede presentar tendencias de crecimiento o disminución, comportamientos estacionales, variaciones asociadas con determinados días de la semana y diferencias vinculadas con las características de productos o establecimientos. Por esta razón, la predicción de ventas constituye un problema que requiere analizar simultáneamente la

dimensión temporal de los datos y las características relacionadas con las unidades de negocio consideradas.

Los métodos de análisis de series temporales permiten estudiar la relación existente entre observaciones presentes y pasadas. Paralelamente, los algoritmos de Machine Learning permiten modelar relaciones no lineales y utilizar múltiples variables predictoras para representar fenómenos complejos. La integración de ambos enfoques puede realizarse mediante la generación de características temporales que posteriormente son utilizadas como variables de entrada para algoritmos de aprendizaje supervisado.

Dentro de esta investigación, la información histórica se analiza considerando cada combinación de tienda y producto como una secuencia temporal de ventas. El conjunto utilizado contiene diez tiendas y cincuenta productos, generando un total de 500 combinaciones tienda-producto observadas diariamente durante cinco años. Esta estructura permite evaluar tanto el comportamiento agregado de la demanda como las diferencias existentes entre establecimientos y productos.

El desarrollo se realiza utilizando Python en Google Colab, conjuntamente con bibliotecas destinadas al procesamiento de datos, visualización y Machine Learning. El modelo predictivo principal utiliza LightGBM, algoritmo basado en árboles de decisión potenciados mediante gradient boosting. Para incorporar información histórica se generan variables de rezago, promedios móviles y medias móviles exponenciales, permitiendo que el algoritmo utilice el comportamiento reciente de las ventas para estimar valores posteriores.

El proyecto adopta CRISP-DM como marco metodológico debido a que permite organizar el desarrollo de soluciones de Ciencia de Datos mediante fases interrelacionadas. Este enfoque facilita documentar desde la comprensión inicial de los datos hasta la evaluación del modelo y la interpretación de los pronósticos generados.

1.2. Planteamiento del problema

Las organizaciones que manejan productos e inventarios requieren estimaciones de demanda que permitan planificar adecuadamente sus operaciones. Una diferencia considerable entre las

ventas esperadas y las ventas efectivamente registradas puede provocar decisiones ineficientes de abastecimiento, acumulación innecesaria de inventario, desabastecimiento y utilización inadecuada de recursos.

Uno de los principales desafíos asociados con la predicción de ventas corresponde a la naturaleza dinámica de la demanda. El comportamiento comercial puede variar entre períodos, productos y establecimientos, de manera que utilizar únicamente valores promedio o tendencias generales puede resultar insuficiente para representar correctamente los patrones presentes en los datos.

En el conjunto utilizado para esta investigación se dispone de registros diarios correspondientes a diez tiendas y cincuenta productos durante el período 2013–2017. El análisis preliminar evidencia diferencias importantes entre establecimientos y productos, así como patrones temporales asociados con el año, el mes y el día de la semana. Por tanto, la predicción de ventas requiere un modelo capaz de considerar simultáneamente estas dimensiones y de incorporar información proveniente del comportamiento histórico de cada serie.

Los algoritmos de Machine Learning permiten abordar este tipo de problemas mediante la construcción de características que representen la estructura temporal de los datos. En particular, variables rezagadas, promedios móviles y medias móviles exponencialmente ponderadas permiten suministrar al algoritmo información sobre valores pasados y tendencias recientes sin utilizar información futura durante el entrenamiento.

A partir de esta problemática, la investigación plantea desarrollar y evaluar un modelo basado en LightGBM capaz de predecir ventas diarias utilizando variables asociadas con tienda, producto, fecha y comportamiento histórico de la demanda. Posteriormente, el modelo será empleado para generar un pronóstico recursivo correspondiente al año siguiente al período histórico disponible.

En consecuencia, el problema central de investigación puede expresarse mediante la siguiente pregunta:

1.3. Formulación del problema

¿En qué medida un modelo de Machine Learning basado en LightGBM e ingeniería de características temporales permite predecir la demanda de ventas a partir de información histórica de tiendas y productos?

1.3.1. Preguntas secundarias

1. ¿Qué patrones y tendencias pueden identificarse en los datos históricos de ventas mediante técnicas de análisis exploratorio de datos?
2. ¿Qué nivel de desempeño predictivo presenta el modelo LightGBM al incorporar características temporales derivadas de los datos históricos de ventas?
3. ¿Cómo influye la calidad de los datos históricos en la precisión de los modelos predictivos?
4. ¿De qué manera los modelos predictivos pueden contribuir a mejorar la planificación financiera y operativa empresarial?

1.4. Justificación

La predicción precisa de ventas representa uno de los principales desafíos dentro de las organizaciones, debido a que una estimación incorrecta de la demanda puede generar consecuencias negativas como sobreinventarios, pérdidas económicas, escasez de productos y dificultades en la planificación financiera. En muchos casos, las empresas continúan utilizando métodos tradicionales de pronóstico que presentan limitaciones frente a la variabilidad y complejidad del comportamiento del mercado.

La importancia de esta investigación radica en la aplicación de técnicas de Machine Learning y análisis de series temporales para mejorar la precisión en los pronósticos de ventas empresariales. Estas herramientas permiten analizar grandes volúmenes de datos históricos e identificar patrones y tendencias que facilitan la generación de predicciones más confiables y eficientes.

Desde el punto de vista académico, el estudio contribuye al fortalecimiento de conocimientos relacionados con Data Science, análisis predictivo y minería de datos, integrando métodos estadísticos y algoritmos de aprendizaje automático en un caso práctico empresarial. Asimismo, permite evidenciar la aplicabilidad de modelos predictivos dentro de procesos de toma de decisiones estratégicas.

En el ámbito empresarial, esta investigación puede aportar beneficios significativos en la optimización de inventarios, reducción de costos operativos y mejora de la planificación financiera. Además, el uso de modelos predictivos puede ayudar a las organizaciones a anticiparse a cambios en la demanda y responder de manera más eficiente a las necesidades del mercado.

Finalmente, el proyecto es viable debido a la disponibilidad de datasets históricos de ventas y herramientas tecnológicas como Python, bibliotecas de Machine Learning y plataformas de análisis de datos que facilitan el desarrollo e implementación de modelos predictivos.

1.5. Objetivos

1.5.1. Objetivo General

Desarrollar y evaluar un modelo predictivo de ventas empresariales basado en Machine Learning e ingeniería de características temporales, utilizando datos históricos de demanda para generar pronósticos que apoyen procesos de planificación empresarial.

1.5.2. Objetivos Específicos

1. Analizar y preparar los datos históricos de ventas, identificando su estructura, calidad, tendencias y patrones temporales relevantes para el proceso predictivo.
2. Construir variables predictoras temporales mediante características derivadas de la fecha, valores rezagados, promedios móviles y medias móviles exponencialmente ponderadas.
3. Desarrollar un modelo predictivo basado en LightGBM utilizando una división temporal entre los conjuntos de entrenamiento y validación.
4. Evaluar el desempeño del modelo mediante las métricas MAE, RMSE, MAPE y R^2, complementando el análisis mediante la comparación de valores reales y predichos y el estudio de los residuos.
5. Generar un pronóstico recursivo de ventas para un período posterior al conjunto histórico y analizar su comportamiento a nivel temporal, de tienda y de producto.
6. Interpretar los resultados obtenidos desde una perspectiva empresarial para identificar su utilidad en procesos relacionados con planificación de demanda e inventarios.

1.6. Relación entre objetivos y preguntas de investigación

Las preguntas de investigación formuladas se encuentran directamente alineadas con el problema planteado y los objetivos del estudio, ya que buscan analizar cómo las técnicas de Machine Learning e ingeniería de características temporales pueden mejorar la precisión en el pronóstico de ventas empresariales.

La pregunta principal se relaciona con el objetivo general al enfocarse en el desarrollo de un modelo predictivo que apoye la toma de decisiones estratégicas mediante el análisis de datos históricos. Por su parte, las preguntas secundarias permiten profundizar en aspectos específicos como el análisis de patrones temporales, la evaluación del desempeño predictivo del modelo y la interpretación de su utilidad dentro de procesos de planificación empresarial.

Esta relación garantiza coherencia metodológica entre el problema identificado, las preguntas de investigación y los objetivos planteados para el desarrollo del estudio.

1.7. Alcance

La presente investigación se enfoca en el desarrollo y evaluación de un modelo predictivo de ventas utilizando información histórica correspondiente al conjunto público Demand Forecasting Dataset, disponible en la plataforma Kaggle. El conjunto contiene registros diarios de ventas de diez tiendas y cincuenta productos entre el 1 de enero de 2013 y el 31 de diciembre de 2017.

El alcance del estudio comprende la exploración y preparación de los datos, la identificación de patrones temporales, la construcción de variables predictoras, el entrenamiento de un algoritmo de Machine Learning y la evaluación de su capacidad predictiva utilizando información posterior al período de entrenamiento.

La variable objetivo corresponde a la cantidad diaria de unidades vendidas. Como variables explicativas se consideran los identificadores de tienda y producto, características derivadas de la fecha, valores históricos rezagados, promedios móviles y medias móviles exponencialmente ponderadas.

El entrenamiento del modelo se realiza utilizando observaciones históricas hasta el 31 de diciembre de 2016, mientras que las ventas correspondientes al año 2017 se utilizan como conjunto

de validación temporal. Este esquema permite evaluar el comportamiento del modelo sobre un período futuro no utilizado durante el entrenamiento.

Posteriormente, se genera un pronóstico correspondiente al año 2018 mediante un procedimiento recursivo. Las estimaciones son analizadas de manera agregada por mes, tienda y producto, con el propósito de identificar posibles cambios en los niveles de demanda.

La investigación se limita a la información disponible en el dataset seleccionado. Por esta razón, no incorpora variables externas como precio, promociones, condiciones macroeconómicas, ubicación geográfica, comportamiento de competidores o características demográficas de los consumidores. En consecuencia, los resultados deben interpretarse dentro del contexto del conjunto de datos y no como una representación universal del comportamiento de todas las empresas.

Asimismo, el estudio se concentra en la evaluación de un modelo LightGBM enriquecido mediante ingeniería de características temporales. Los modelos estadísticos y de Deep Learning son considerados dentro de la revisión de literatura como enfoques relevantes para forecasting, pero no forman parte de la experimentación principal desarrollada en este trabajo.

2. Marco Teórico y Revisión de Literatura

2.1. Introducción

La predicción de ventas constituye un campo de estudio relevante dentro de la Ciencia de Datos debido a su aplicación directa en procesos de planificación, abastecimiento, gestión de inventarios y toma de decisiones empresariales. El desarrollo de modelos predictivos permite utilizar información histórica para identificar patrones de comportamiento y estimar posibles niveles futuros de demanda.

Tradicionalmente, los problemas de pronóstico han sido abordados mediante métodos estadísticos de series temporales. Sin embargo, el crecimiento en el volumen de datos y la disponibilidad de herramientas computacionales han favorecido la incorporación de técnicas de Machine Learning capaces de representar relaciones no lineales e integrar múltiples variables explicativas.

En problemas de ventas, el componente temporal adquiere especial relevancia debido a que las observaciones no son independientes entre sí. El comportamiento registrado en períodos anteriores puede aportar información para estimar períodos posteriores. Por esta razón, técnicas como los valores rezagados, promedios móviles y medias móviles exponencialmente ponderadas permiten transformar la información histórica en variables que pueden ser utilizadas por algoritmos de aprendizaje supervisado.

En este capítulo se presentan los fundamentos conceptuales relacionados con forecasting de ventas, series temporales, Machine Learning, Gradient Boosting, LightGBM, ingeniería de características temporales, evaluación de modelos predictivos y metodología CRISP-DM. Adicionalmente, se revisan otros enfoques utilizados en problemas de forecasting con el propósito de contextualizar metodológicamente la solución desarrollada.

2.2. Predicción de ventas y forecasting empresarial

El forecasting puede entenderse como el proceso mediante el cual se utilizan observaciones históricas y métodos analíticos para generar estimaciones sobre el comportamiento futuro de una variable. Dentro del ámbito empresarial, una de sus principales aplicaciones corresponde a la predicción de ventas, donde el objetivo consiste en estimar la demanda futura de productos o servicios.

La utilidad del pronóstico no se limita a conocer anticipadamente una determinada cantidad de ventas. Sus resultados pueden utilizarse como insumo para procesos de planificación de inventarios, abastecimiento, asignación de recursos y establecimiento de estrategias comerciales. Por esta razón, la calidad de las estimaciones puede influir directamente sobre diferentes decisiones operativas.

La dificultad del problema radica en que las ventas pueden presentar variaciones asociadas con múltiples factores. Entre ellos se encuentran los patrones temporales, diferencias entre productos y establecimientos, estacionalidad y comportamiento histórico de la demanda. En consecuencia, los métodos de forecasting deben ser capaces de representar adecuadamente la estructura existente en los datos.

Hyndman y Athanasopoulos (2021) señalan que la selección de un método de pronóstico depende, entre otros aspectos, de las características de los datos disponibles, el horizonte de

predicción y los patrones que presenta la serie. Esto implica que no existe necesariamente una única técnica apropiada para todos los problemas de forecasting.

En esta investigación, la predicción se desarrolla a partir de información histórica diaria y considera simultáneamente las dimensiones temporal, tienda y producto. El propósito es aprovechar el comportamiento previo de las ventas para construir un modelo capaz de generar estimaciones posteriores.

2.3. Series temporales

Una serie temporal corresponde a un conjunto de observaciones registradas secuencialmente a lo largo del tiempo. A diferencia de otros conjuntos de datos, el orden cronológico constituye una característica fundamental, debido a que una observación puede mantener relación con valores registrados anteriormente.

Capa Santos (2017) plantea que el análisis de series temporales permite estudiar la evolución de fenómenos dinámicos e identificar patrones que pueden ser utilizados para generar pronósticos. Dentro del contexto empresarial, las ventas constituyen un ejemplo de variable temporal, ya que su comportamiento puede analizarse diariamente, semanalmente, mensualmente o mediante otros intervalos.

Entre los principales componentes que pueden observarse en una serie temporal se encuentran la tendencia, estacionalidad, ciclos y variaciones aleatorias. La tendencia representa movimientos generales de crecimiento o disminución; la estacionalidad corresponde a patrones que se repiten periódicamente; mientras que el componente aleatorio representa variaciones que no pueden explicarse directamente mediante patrones sistemáticos.

Para desarrollar modelos predictivos sobre series temporales es necesario preservar el orden cronológico de las observaciones. En consecuencia, la separación entre entrenamiento y validación debe realizarse respetando la dimensión temporal, evitando utilizar información futura para construir predicciones correspondientes a períodos anteriores.

En el presente estudio, los registros correspondientes al período 2013–2016 son utilizados para el entrenamiento del modelo, mientras que el año 2017 funciona como período de validación. Esta

estructura permite simular un escenario en el que se utilizan datos históricos para predecir un período posterior.

2.4. Machine Learning aplicado a la predicción de ventas

El Machine Learning o aprendizaje automático corresponde a un conjunto de métodos computacionales que permiten identificar patrones a partir de datos y utilizar dichos patrones para realizar predicciones sobre nuevas observaciones. En problemas de aprendizaje supervisado, el modelo aprende una relación entre un conjunto de variables explicativas y una variable objetivo conocida.

Dentro de la predicción de ventas, los algoritmos de Machine Learning ofrecen la posibilidad de utilizar simultáneamente información temporal, características de productos, establecimientos y variables construidas a partir del comportamiento histórico. Esto representa una diferencia importante frente a enfoques que modelan exclusivamente la evolución de una única serie.

El problema desarrollado en esta investigación corresponde a una tarea de regresión supervisada, debido a que la variable objetivo —ventas— es cuantitativa. El algoritmo recibe diferentes características predictoras y aprende una función capaz de aproximar la cantidad de ventas asociada con cada observación.

Una consideración particularmente importante en problemas temporales corresponde a la prevención de fuga de información o data leakage. Las características utilizadas para predecir una determinada fecha deben construirse exclusivamente a partir de información disponible antes de dicha fecha. Por este motivo, las variables históricas utilizadas en esta investigación se generan mediante desplazamientos temporales que impiden utilizar directamente información futura.

2.5. Gradient Boosting

El Gradient Boosting constituye una técnica de aprendizaje automático basada en la combinación secuencial de múltiples modelos simples, generalmente árboles de decisión. A diferencia de métodos que construyen modelos independientes, el procedimiento de boosting genera nuevos árboles con el propósito de reducir progresivamente los errores producidos por las etapas anteriores.

El resultado final se obtiene mediante la combinación de los diferentes modelos construidos durante el proceso de entrenamiento. Esta estrategia permite representar relaciones no lineales entre las variables y trabajar con estructuras de datos complejas.

En problemas de regresión, el objetivo consiste en minimizar una función de pérdida asociada con la diferencia entre los valores observados y las predicciones. Durante las sucesivas iteraciones, el algoritmo concentra el aprendizaje en aquellos patrones que todavía no han sido adecuadamente representados.

Esta familia de algoritmos resulta especialmente relevante para el presente estudio debido a que las ventas dependen simultáneamente de variables temporales, identificadores de tienda y producto e información histórica derivada de cada combinación.

2.6. Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) es un algoritmo basado en árboles de decisión y perteneciente a la familia de métodos de Gradient Boosting. Su funcionamiento consiste en construir secuencialmente árboles que buscan reducir los errores residuales producidos durante las iteraciones anteriores.

Una de sus características relevantes corresponde a la eficiencia computacional para trabajar con conjuntos de datos de gran tamaño. Esta propiedad resulta particularmente conveniente en investigaciones donde se dispone de un número elevado de observaciones y múltiples características predictoras.

A diferencia de algoritmos basados en un único árbol de decisión, LightGBM combina múltiples árboles para obtener una estimación final. Esto permite representar relaciones no lineales e interacciones entre variables sin requerir especificar previamente una relación matemática determinada entre ellas.

En el contexto de esta investigación, LightGBM resulta apropiado debido a la estructura del conjunto analizado, conformado por 913.000 observaciones históricas y múltiples características relacionadas con fecha, tienda, producto y comportamiento previo de las ventas.

El modelo recibe como variables predictoras las características temporales y las variables obtenidas mediante ingeniería de características. La variable objetivo corresponde a la cantidad

diaria de ventas. Posteriormente, su capacidad predictiva se evalúa utilizando información correspondiente a un período posterior al utilizado durante el entrenamiento.

2.7. Ingeniería de características temporales

La ingeniería de características o feature engineering corresponde al proceso mediante el cual se generan nuevas variables a partir de la información disponible con el propósito de representar patrones relevantes para un modelo predictivo.

En problemas de series temporales, esta etapa permite convertir información histórica en características que pueden ser interpretadas por algoritmos tradicionales de Machine Learning. En lugar de utilizar únicamente la fecha como identificador temporal, pueden construirse variables relacionadas con calendario y comportamiento histórico.

En esta investigación se utilizan cuatro grupos principales de características: variables derivadas de la fecha, valores rezagados (lags), promedios móviles (rolling means) y medias móviles exponencialmente ponderadas (EWMA).

2.7.1. Variables derivadas de la fecha

Las variables de calendario permiten representar patrones relacionados con diferentes períodos temporales. A partir de una fecha pueden obtenerse características como año, mes, día del mes, día de la semana y trimestre.

Estas variables permiten que el modelo identifique comportamientos asociados con períodos específicos. Por ejemplo, si determinados meses presentan sistemáticamente niveles diferentes de demanda, el algoritmo puede aprender esta relación utilizando la variable correspondiente al mes.

En el presente estudio se generan características temporales a partir de la fecha original con el propósito de proporcionar al modelo información adicional sobre la posición cronológica de cada observación.

2.7.2. Variables rezagadas — Lag Features

Las variables rezagadas representan valores históricos de una variable correspondientes a períodos anteriores. Matemáticamente, para una serie Y_t, una variable rezagada k de períodos puede representarse como:

$$Lag_k = y_{t-k}$$

De esta manera, para predecir las ventas correspondientes al período, el modelo puede utilizar información registrada en t - k.

Esta estrategia resulta especialmente útil cuando existe dependencia entre las ventas actuales y las observadas anteriormente. Además, la utilización adecuada de desplazamientos permite evitar que el modelo acceda al valor que intenta predecir.

En esta investigación, los rezagos se calculan independientemente para cada combinación tienda-producto, preservando así el historial particular de cada serie.

2.7.3. Promedios móviles — Rolling Mean

Los promedios móviles permiten resumir el comportamiento reciente de una serie utilizando una ventana determinada de observaciones anteriores.

De manera general:

$$RM_t = \frac{1}{w} \sum_{i=1}^{w} y_{t-i}$$

Donde w representa el tamaño de la ventana utilizada.

Su incorporación permite suavizar fluctuaciones de corto plazo y proporcionar al modelo una representación del nivel reciente de demanda. En problemas de ventas, esta característica puede resultar útil cuando las observaciones presentan variabilidad diaria considerable.

Al igual que las variables rezagadas, los promedios móviles utilizados en esta investigación se calculan considerando únicamente información anterior a la fecha que se desea predecir.

2.7.4. Media móvil exponencialmente ponderada — EWMA

La media móvil exponencialmente ponderada, conocida como Exponentially Weighted Moving Average (EWMA), constituye una técnica que resume observaciones históricas asignando mayor importancia a los valores recientes.

De manera general, puede expresarse como:

$$EWMA_t = \alpha y_{t-1} + (1-\alpha) EWMA_{t-1}$$

Donde ∝ representa el factor de suavizamiento.

A diferencia de un promedio móvil simple, donde las observaciones incluidas en una ventana reciben el mismo peso, EWMA permite que la influencia de los registros disminuya progresivamente conforme aumenta su antigüedad.

Su incorporación en el modelo permite representar cambios recientes en el comportamiento de la demanda y complementar la información proporcionada por los rezagos y promedios móviles.

2.8. Pronóstico recursivo en series temporales

Cuando el horizonte de predicción comprende múltiples períodos futuros, puede utilizarse una estrategia de pronóstico recursivo. Este procedimiento consiste en generar inicialmente la predicción correspondiente al primer período futuro y posteriormente incorporar ese resultado como parte de la información disponible para construir las características requeridas en el siguiente período.

El proceso puede representarse de manera simplificada como:

$$\hat{y}_{t+1} = f(X_{t+1})$$

Posteriormente:

$$\hat{y}_{t+2} = f(X_{t+2}, \hat{y}_{t+1})$$

y el procedimiento continúa sucesivamente hasta completar el horizonte de pronóstico.

Esta estrategia resulta necesaria cuando las variables predictoras incluyen información rezagada y no existen valores reales para los períodos futuros. En consecuencia, las predicciones anteriores pasan progresivamente a formar parte del historial utilizado para generar nuevas características.

Una consideración importante es que los errores producidos en las primeras predicciones pueden propagarse hacia períodos posteriores. Por esta razón, la calidad del modelo durante la validación temporal constituye un elemento relevante antes de utilizarlo para generar horizontes extensos.

En esta investigación, el procedimiento recursivo se emplea para construir las estimaciones diarias correspondientes al año 2018 a partir del historial disponible hasta diciembre de 2017.

2.9. Métricas de evaluación

2.9.1. Error Absoluto Medio — MAE

El MAE mide el promedio de las diferencias absolutas entre los valores reales y los valores predichos:

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

Un valor menor representa, en términos generales, una menor diferencia promedio entre observaciones y predicciones. Su interpretación resulta intuitiva debido a que mantiene la unidad original de la variable analizada.

2.9.2. Raíz del Error Cuadrático Medio — RMSE

El RMSE se calcula mediante:

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

Al elevar los errores al cuadrado antes de obtener su promedio, esta métrica penaliza con mayor intensidad las diferencias grandes entre valores reales y predichos.

2.9.3. Error Porcentual Absoluto Medio — MAPE

El MAPE expresa el error de predicción en términos porcentuales:

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

Su principal ve Sin embargo, debe utilizarse con preca ro, debido a que el denominador pued

2.9.4. Coeficiente de determinación — R²

El coeficiente R² permite analizar qué proporción de la variabilidad observada en la variable objetivo es explicada por el modelo:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Los valores próximos a 1 indican un mayor ajuste respecto de la variabilidad observada. Sin embargo, esta métrica debe interpretarse conjuntamente con medidas de error y no como único criterio para determinar la calidad predictiva del modelo.

2.10. Metodología CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) constituye una metodología orientada al desarrollo estructurado de proyectos de minería de datos y Ciencia de Datos. Su organización permite relacionar el problema que se desea resolver con las diferentes actividades requeridas para transformar datos en resultados analíticos.

La metodología comprende seis fases principales:

1. **Comprensión del negocio:** definición del problema, objetivos y necesidades que justifican el desarrollo de la solución analítica.
2. **Comprensión de los datos:** recopilación, descripción, exploración y evaluación de la calidad de la información disponible.
3. **Preparación de los datos:** selección, limpieza, transformación y generación de las características necesarias para el modelado.
4. **Modelado:** selección, configuración y entrenamiento del algoritmo predictivo.
5. **Evaluación:** análisis del desempeño obtenido y determinación de la capacidad del modelo para responder al problema planteado.
6. **Despliegue:** utilización, comunicación o presentación de los resultados obtenidos para apoyar procesos posteriores de toma de decisiones.

Una característica relevante de CRISP-DM es que sus fases no necesariamente conforman un proceso estrictamente lineal. Los resultados obtenidos durante una etapa pueden evidenciar la necesidad de regresar a actividades anteriores para realizar ajustes en los datos, características o modelo.

En esta investigación, CRISP-DM constituye la estructura central utilizada para organizar el proceso analítico, desde la comprensión inicial del problema de forecasting hasta la evaluación del modelo LightGBM y la generación del pronóstico correspondiente a 2018.

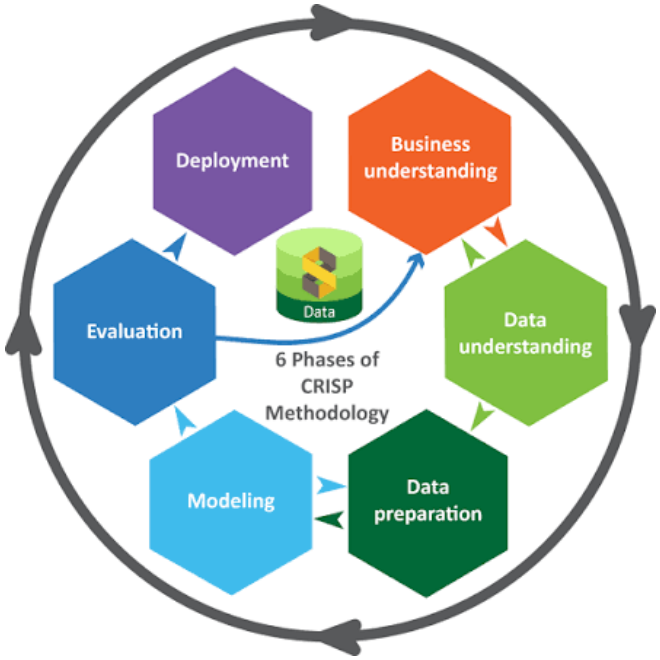


Ilustración 1 - Fases de la metodología CRISP-DM

2.11. Enfoques alternativos utilizados en forecasting

La literatura sobre forecasting incluye diferentes familias de modelos cuya aplicabilidad depende de las características del problema y de los datos disponibles. Entre ellas se encuentran los métodos estadísticos tradicionales, algoritmos de Machine Learning y modelos de Deep Learning.

Los modelos ARIMA utilizan componentes autorregresivos, diferenciación y medias móviles para representar dependencias temporales. Holt-Winters incorpora componentes de nivel, tendencia y estacionalidad, siendo especialmente relevante en series que presentan patrones estacionales.

Dentro del Machine Learning, Random Forest utiliza múltiples árboles de decisión y combina sus resultados para generar predicciones. Por otra parte, los algoritmos de Gradient Boosting construyen árboles secuencialmente con el propósito de reducir progresivamente el error, familia dentro de la cual se encuentra LightGBM.

Las redes LSTM (Long Short-Term Memory) constituyen una arquitectura de redes neuronales recurrentes diseñada para procesar información secuencial y representar dependencias temporales de largo plazo.

Cada enfoque presenta diferentes requerimientos, niveles de complejidad e interpretabilidad. Por esta razón, su consideración dentro de la revisión teórica permite contextualizar la selección metodológica realizada en la presente investigación, sin implicar que todos estos algoritmos hayan sido implementados experimentalmente.

3. Marco Metodológico

3.1. Enfoque de la investigación

La presente investigación adopta un enfoque cuantitativo, debido a que utiliza datos históricos de ventas expresados numéricamente para desarrollar y evaluar un modelo predictivo. El análisis se fundamenta en el procesamiento de observaciones históricas, la construcción de variables explicativas y la aplicación de métricas cuantitativas para determinar el desempeño del modelo.

El enfoque cuantitativo resulta adecuado para el objetivo de la investigación, debido a que permite analizar sistemáticamente el comportamiento histórico de las ventas y establecer relaciones entre variables temporales, características asociadas con tiendas y productos, y la cantidad de unidades vendidas.

Asimismo, el estudio utiliza técnicas de Ciencia de Datos y Machine Learning para transformar los registros históricos en información susceptible de ser utilizada para generar estimaciones futuras.

La variable de interés corresponde a las ventas diarias, mientras que las variables predictoras incluyen información temporal y características construidas a partir del comportamiento histórico de las ventas.

3.2. Tipo de investigación

La investigación es de tipo aplicada y predictiva.

Se considera aplicada debido a que utiliza conocimientos y herramientas de Ciencia de Datos para abordar un problema relacionado con la estimación futura de ventas empresariales. El propósito no se limita al análisis descriptivo de los registros históricos, sino que busca desarrollar una solución predictiva susceptible de apoyar procesos de planificación.

Asimismo, presenta un alcance predictivo, debido a que se construye un modelo destinado a estimar valores futuros de la variable ventas a partir de patrones identificados en observaciones históricas.

La investigación comprende además una fase descriptiva inicial, correspondiente al análisis exploratorio de los datos. Esta etapa permite caracterizar el comportamiento de las ventas antes de desarrollar el modelo predictivo.

3.3. Diseño de la investigación

La investigación presenta un diseño no experimental, longitudinal y retrospectivo.

Es no experimental debido a que las variables analizadas no son manipuladas deliberadamente. El estudio utiliza registros históricos existentes y analiza su comportamiento para desarrollar el modelo predictivo.

Se considera longitudinal debido a que las observaciones corresponden a diferentes momentos dentro de un período temporal continuo, permitiendo estudiar la evolución de las ventas a través del tiempo.

Finalmente, presenta un carácter retrospectivo, debido a que el desarrollo del modelo se fundamenta en registros históricos previamente generados.

Esta estructura resulta especialmente importante debido a la naturaleza temporal del problema. En consecuencia, el orden cronológico de las observaciones se conserva durante la preparación, entrenamiento y evaluación del modelo.

3.4. Metodología de desarrollo

Para organizar el desarrollo analítico se utiliza la metodología CRISP-DM (Cross Industry Standard Process for Data Mining).

La metodología permite estructurar el proyecto mediante seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. El material académico utilizado en la maestría presenta precisamente estas fases y describe dentro de preparación actividades de selección, limpieza y creación de variables, mientras que la fase de modelado comprende la selección y entrenamiento del algoritmo.

En esta investigación, las fases se aplican de la siguiente manera:

Fase CRISP-DM	Aplicación en la investigación
Comprensión del negocio	Definición del problema de predicción de ventas y del objetivo del modelo
Comprensión de los datos	Exploración del conjunto histórico de ventas
Preparación de los datos	Limpieza, transformación e ingeniería de características temporales
Modelado	Entrenamiento del algoritmo LightGBM
Evaluación	Validación temporal mediante MAE, RMSE, MAPE y R^2
Despliegue	Generación del pronóstico recursivo correspondiente a 2018

Tabla 1 - Aplicación de las fases CRISP-DM en la investigación

3.5. Fuente de datos

La investigación utiliza un conjunto de datos histórico correspondiente a ventas diarias registradas entre los años 2013 y 2017.

La unidad de observación está determinada por la combinación de fecha, tienda y producto, mientras que la variable sales representa la cantidad de ventas correspondiente a cada combinación.

El conjunto utilizado contiene información correspondiente a:

- 10 tiendas.
- 50 productos.

- 500 combinaciones tienda-producto.
- Registros diarios comprendidos entre 2013 y 2017.
- 913.000 observaciones históricas.

La estructura permite analizar simultáneamente la dimensión temporal de las ventas y las diferencias existentes entre tiendas y productos.

Característica	Valor
Período histórico	2013–2017
Tiendas	10
Productos	50
Combinaciones tienda-producto	500
Observaciones	913.000
Frecuencia	Diaria
Variable objetivo	Ventas

Tabla 2 - Características generales del conjunto de datos

3.6. Descripción de las variables originales

El conjunto de datos contiene las variables necesarias para identificar temporalmente cada observación, así como la tienda, el producto y la cantidad de ventas registrada.

Variable	Tipo	Descripción
date	Temporal	Fecha correspondiente a la observación
store	Numérica/categórica	Identificador de la tienda
item	Numérica/categórica	Identificador del producto
sales	Numérica	Cantidad de ventas registradas

Tabla 3 - Variables originales utilizadas en la investigación

3.7. Herramientas utilizadas

El procesamiento de los datos y el desarrollo del modelo predictivo se realizaron mediante el lenguaje de programación Python, utilizando un entorno de ejecución basado en notebooks.

Las principales librerías empleadas corresponden a herramientas para manipulación de datos, cálculo numérico, visualización, modelado y evaluación.

Herramienta	Aplicación
Python	Lenguaje principal de programación
Google Colab	Entorno utilizado para ejecutar el código
Pandas	Manipulación y transformación de datos
NumPy	Operaciones numéricas
Matplotlib	Generación de visualizaciones
LightGBM	Construcción del modelo predictivo
Scikit-learn	Cálculo de métricas de evaluación

Tabla 4 - Herramientas computacionales utilizadas

3.8. Comprensión y exploración inicial de los datos

Como parte de la fase de comprensión de los datos de CRISP-DM, se realizó una exploración inicial destinada a identificar la estructura del conjunto, período temporal, cantidad de observaciones, comportamiento de la variable objetivo y posibles problemas relacionados con la calidad de la información.

Durante esta etapa se verificaron las dimensiones del conjunto y la existencia de valores faltantes o registros duplicados. Posteriormente, se analizaron las ventas desde diferentes dimensiones temporales y comerciales.

El análisis exploratorio consideró principalmente:

- evolución histórica de las ventas;
- comportamiento anual;

- comportamiento mensual;
- comportamiento según día de la semana;
- distribución de ventas entre tiendas;
- diferencias en ventas entre productos.

Estos análisis permitieron identificar patrones relevantes antes de construir las variables utilizadas posteriormente por el modelo.

3.9. Preparación de los datos

La preparación de los datos constituye una de las etapas fundamentales del procedimiento debido a que el algoritmo requiere una estructura adecuada para representar las relaciones temporales existentes en las ventas.

Inicialmente, la variable correspondiente a la fecha fue transformada a un formato temporal. Posteriormente, los registros fueron ordenados cronológicamente considerando la fecha y las combinaciones tienda-producto.

También se realizaron verificaciones relacionadas con valores faltantes y duplicados, con el propósito de identificar posibles problemas que pudieran afectar el análisis posterior.

Una vez validada la estructura básica del conjunto, se procedió a generar características temporales derivadas de la fecha y del historial de ventas.

3.10. Ingeniería de características

La ingeniería de características se realizó con el propósito de proporcionar al algoritmo información relacionada con la estructura temporal y el comportamiento histórico de cada combinación tienda-producto.

Las características fueron construidas preservando el orden cronológico y evitando utilizar información posterior a la fecha correspondiente a cada observación.

Las variables generadas se clasificaron en tres grupos principales: características de calendario, características históricas y medias móviles exponencialmente ponderadas.

3.10.1. Características de calendario

A partir de la variable date se construyeron características destinadas a representar diferentes componentes temporales.

Entre ellas se incluyeron:

- año;
- mes;
- día;
- día de la semana;
- semana del año;
- trimestre.

Estas variables permiten representar patrones asociados con diferentes posiciones dentro del calendario y proporcionar al algoritmo información temporal adicional.

3.10.2. Variables Lag

Posteriormente se generaron valores rezagados de las ventas para cada combinación tienda-producto.

$$Lag_1, \quad Lag_7, \quad Lag_{14}, \quad Lag_{28}$$

Estos valores representan respectivamente las ventas correspondientes a 1, 7, 14 y 28 días anteriores.

La generación de los rezagos se realizó independientemente para cada combinación de tienda y producto, permitiendo preservar la estructura histórica particular de cada serie.

La utilización de shift garantiza que la observación correspondiente al día que se desea predecir no sea utilizada directamente como variable explicativa.

Variable	Información representada
Lag 1	Ventas del día anterior
Lag 7	Ventas de siete días anteriores
Lag 14	Ventas de catorce días anteriores
Lag 28	Ventas de veintiocho días anteriores

Tabla 5 - Variables rezagadas utilizadas

3.10.3. Promedios móviles

También se generaron variables correspondientes al promedio de ventas registrado durante ventanas temporales anteriores.

En el desarrollo se utilizaron:

$$RollingMean_7, \quad RollingMean_{14}, \quad RollingMean_{28}$$

Estas características representan el nivel promedio reciente de las ventas y permiten reducir parcialmente el efecto de fluctuaciones diarias.

Su cálculo se realizó para cada combinación tienda-producto utilizando exclusivamente información anterior a la observación que se desea estimar.

Variable	Ventana
Rolling Mean 7	7 días
Rolling Mean 14	14 días
Rolling Mean 28	28 días

Tabla 6 - Variables de promedio móvil

3.10.4. Variables EWMA

Las variables EWMA fueron construidas combinando cuatro factores de suavizamiento (α=0,95;0,90;0,80;0,70) con tres desplazamientos temporales (1, 7 y 14 días), generando un total de 12 características EWMA.

El propósito de estas variables consiste en representar el comportamiento histórico asignando una mayor ponderación a las observaciones recientes.

Parámetro	Valores
Alpha	0,95; 0,90; 0,80; 0,70
Lag	1; 7; 14
Total de variables	12

Tabla 7 - Configuración de las variables EWMA

3.11. Variables utilizadas por el modelo

Una vez finalizada la ingeniería de características, se construyó la matriz utilizada para entrenar el algoritmo.

De acuerdo con la implementación desarrollada, el modelo utiliza un total de 27 variables predictoras.

- Dimensiones: (500, 27)
- Número de variables: 27

Las características corresponden a información de identificación, calendario, rezagos, promedios móviles y EWMA.

Grupo	Variables	Cantidad
Identificación	store, item	2
Calendario	year, month, day, dayofweek, weekofyear, quarter	6
Lag	lag_1, lag_7, lag_14, lag_28	4

Rolling Mean	7, 14 y 28 días	3
EWMA	4 alphas × 3 lags	12
Total		**27**

Tabla 8 - Composición de las variables predictoras

3.12. Tratamiento de valores faltantes generados por rezagos

La generación de variables históricas mediante rezagos y ventanas móviles produce inevitablemente valores faltantes en las primeras observaciones de cada combinación tienda-producto, debido a que todavía no existe suficiente información histórica para calcular determinadas características.

Por ejemplo, para construir un rezago de 28 días se requiere disponer de al menos 28 observaciones anteriores.

Por esta razón, después de generar las características se realizó el tratamiento correspondiente de los registros que no disponían de la información histórica necesaria para construir la matriz final de modelado.

Este procedimiento permitió trabajar únicamente con observaciones que contaban con las características requeridas por el modelo.

3.13. División temporal de los datos

Debido a que la investigación utiliza información temporal, los datos no fueron divididos aleatoriamente entre entrenamiento y validación. Una división aleatoria podría permitir que observaciones futuras participaran en el entrenamiento de un modelo utilizado para predecir observaciones anteriores.

Para preservar el orden cronológico, se estableció una división temporal.

Los registros correspondientes al período 2013–2016 fueron utilizados para entrenar el modelo, mientras que las observaciones correspondientes a 2017 fueron destinadas a evaluar su capacidad predictiva.

Conjunto	Período	Finalidad
Entrenamiento	2013–2016	Ajuste del modelo
Validación	2017	Evaluación fuera de muestra
Pronóstico	2018	Generación de estimaciones futuras

Tabla 9 - División temporal utilizada

Esta estrategia permite evaluar el algoritmo bajo una situación similar a su utilización real: aprender utilizando información disponible hasta un determinado momento y posteriormente estimar observaciones futuras.

3.14. Justificación de la selección del modelo LightGBM

Para el desarrollo del componente predictivo se seleccionó Light Gradient Boosting Machine (LightGBM) como modelo principal, considerando las características del conjunto de datos y el alcance definido para la investigación. La base utilizada presenta una estructura tabular que combina variables temporales, identificadores de tienda y producto, rezagos (lags), promedios móviles y características EWMA, por lo que se consideró pertinente utilizar un algoritmo basado en árboles capaz de modelar relaciones no lineales entre múltiples variables predictoras.

LightGBM permite integrar dentro de un mismo modelo diferentes tipos de características derivadas del comportamiento histórico de las ventas y capturar interacciones que podrían no ser representadas adecuadamente mediante enfoques exclusivamente lineales. Adicionalmente, su implementación resulta adecuada para conjuntos de datos de gran tamaño, aspecto relevante considerando el volumen de observaciones procesadas durante el desarrollo de esta investigación.

En el planteamiento inicial se contempló la posibilidad de analizar otros enfoques de predicción. Sin embargo, el desarrollo experimental se delimitó a LightGBM con el propósito de mantener un proceso metodológico consistente y concentrar el análisis en la construcción, evaluación e interpretación de un modelo supervisado aplicado al problema de predicción de ventas. Por esta razón, los resultados obtenidos no deben interpretarse como evidencia de superioridad absoluta de LightGBM frente a otros algoritmos, sino como una evaluación de su desempeño dentro del alcance establecido para el presente estudio.

La comparación con otros algoritmos de aprendizaje automático y modelos de series temporales se mantiene como una línea de trabajo futuro, con el fin de determinar si enfoques alternativos o combinaciones de modelos pueden generar mejoras adicionales en la capacidad predictiva.

3.15. Construcción del modelo LightGBM

Una vez preparada la matriz de características y realizada la división temporal de los datos, se procedió al entrenamiento del modelo LightGBM.

La variable dependiente utilizada fue sales, mientras que las 27 características previamente construidas conformaron las variables predictoras.

El algoritmo fue entrenado utilizando exclusivamente los registros correspondientes al período 2013–2016. Posteriormente, el modelo entrenado fue aplicado al conjunto correspondiente a 2017 para obtener predicciones fuera del período utilizado durante el ajuste.

La utilización de LightGBM permitió integrar simultáneamente variables de calendario, identificadores de tienda y producto, valores rezagados, promedios móviles y características EWMA.

3.16. Configuración de hiperparámetros del modelo LightGBM

La configuración de los hiperparámetros constituye una etapa relevante en la construcción de modelos basados en gradient boosting, debido a que estos parámetros controlan aspectos relacionados con la velocidad de aprendizaje y la complejidad del modelo. En el presente estudio se utilizó una configuración experimental definida durante el desarrollo del modelo, manteniendo un valor fijo de semilla aleatoria para favorecer la reproducibilidad de los resultados.

El modelo LightGBM fue configurado con una tasa de aprendizaje (learning_rate) de 0,05, 1.000 estimadores (n_estimators), 64 hojas (num_leaves) y una semilla aleatoria (random_state) de 42. La tasa de aprendizaje de 0,05 establece una actualización gradual durante el proceso de boosting, mientras que los 1.000 estimadores determinan el número de árboles construidos secuencialmente. Por su parte, el parámetro num_leaves controla la complejidad de los árboles y su capacidad para representar relaciones no lineales presentes en los datos.

La configuración empleada no fue obtenida mediante un procedimiento exhaustivo de optimización automática, como Grid Search o optimización bayesiana. En consecuencia, los hiperparámetros utilizados deben entenderse como una configuración experimental del modelo dentro del alcance de esta investigación. Su desempeño fue posteriormente determinado sobre el conjunto temporal de validación correspondiente a 2017 mediante las métricas MAE, RMSE, MAPE y R^2.

Los resultados obtenidos con esta configuración permiten evaluar la viabilidad de LightGBM para el problema planteado; sin embargo, una búsqueda sistemática de hiperparámetros podría explorar configuraciones alternativas y determinar si existen mejoras adicionales en el desempeño predictivo.

Hiperparámetro	Valor	Función
objective	regression	Define el problema como regresión
learning_rate	0,05	Controla la contribución de cada árbol
n_estimators	1.000	Número de árboles del modelo
num_leaves	64	Controla la complejidad de los árboles
random_state	42	Favorece la reproducibilidad

Tabla 10 - Configuración de hiperparámetros de LightGBM

Nota. Los hiperparámetros corresponden a la configuración utilizada para el entrenamiento del modelo LightGBM.

3.17. Evaluación del modelo

El desempeño predictivo del modelo fue evaluado comparando las ventas reales correspondientes a 2017 con las estimaciones generadas por LightGBM para el mismo período.

Se utilizaron cuatro métricas:

- MAE;
- RMSE;

- MAPE;
- R^2.

Cada una permite analizar un componente diferente del desempeño predictivo. El MAE representa el error absoluto promedio; el RMSE incrementa la penalización de errores elevados; el MAPE expresa el error en términos porcentuales; y R^2 permite evaluar la proporción de variabilidad representada por el modelo.

Adicionalmente, el análisis no se limitó a una única métrica global. Se realizaron evaluaciones segmentadas con el propósito de analizar el comportamiento del modelo desde diferentes niveles de agregación.

Entre los análisis desarrollados se incluyeron:

- evaluación general;
- comportamiento temporal de valores reales frente a predicciones;
- análisis de residuos;
- métricas por tienda;
- métricas por producto;
- métricas por mes;
- importancia de características.

3.18. Generación del pronóstico para 2018

Una vez validado el modelo mediante las observaciones correspondientes a 2017, se procedió a generar el pronóstico para el año 2018.

Debido a que las características utilizadas por LightGBM incluyen rezagos, promedios móviles y EWMA, las variables correspondientes a fechas futuras no pueden construirse directamente para todo el horizonte, ya que las ventas reales de 2018 son desconocidas.

Por esta razón, se implementó una estrategia de pronóstico recursivo.

Para el primer día del horizonte se construyeron las variables predictoras utilizando únicamente información disponible hasta el 31 de diciembre de 2017. Posteriormente, el modelo generó la predicción correspondiente al 1 de enero de 2018.

La predicción obtenida fue incorporada al historial de la combinación tienda-producto correspondiente. Este nuevo valor permitió actualizar las características necesarias para estimar el día siguiente.

El procedimiento fue repetido sucesivamente hasta completar el horizonte correspondiente al año 2018.

3.18.1. Construcción del primer día del pronóstico

Para el 1 de enero de 2018 se generaron las características correspondientes a las 500 combinaciones tienda-producto.

Las variables Lag, Rolling Mean y EWMA se calcularon utilizando exclusivamente información histórica disponible hasta el último día de 2017.

Posteriormente, se construyó una matriz de:

- 500 observaciones × 27 variables predictoras

y se aplicó el modelo LightGBM previamente entrenado.

```
Dimensiones: (500, 27)
Número de variables: 27
```

Ilustración 2 - Construcción de la matriz predictora correspondiente al primer día del horizonte de pronóstico

Elaboración propia mediante Python.

3.18.2. Incorporación de las predicciones al historial

Una vez obtenida la predicción correspondiente a cada combinación tienda-producto, los valores generados fueron incorporados al historial utilizado para construir las características del día posterior.

Conceptualmente, el procedimiento puede representarse mediante:

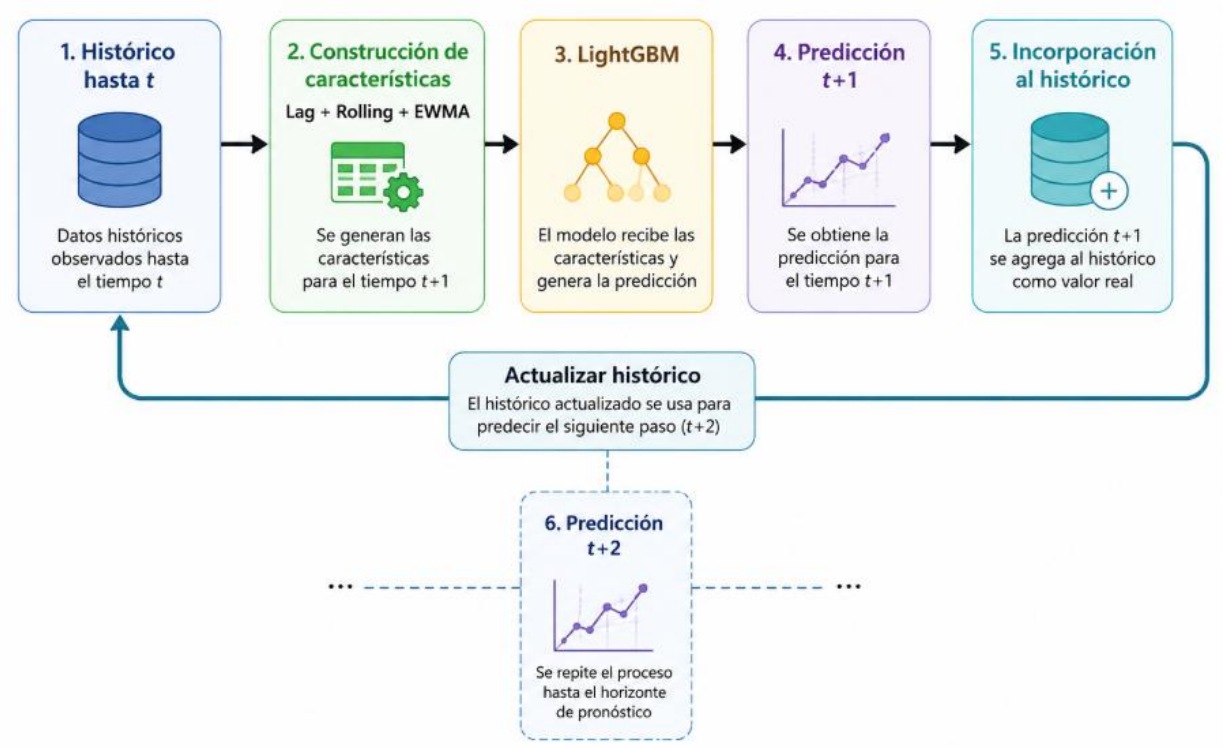


Ilustración 3 - Esquema del procedimiento de pronóstico recursivo utilizado.

Esta actualización permite continuar el proceso durante todo el horizonte futuro.

3.18.3. Validación inicial del procedimiento recursivo

Antes de automatizar la generación del pronóstico correspondiente a todo 2018, se verificó manualmente el procedimiento aplicado a los primeros días del horizonte.

Primero se generaron las características y la predicción correspondientes al 1 de enero de 2018. Posteriormente, esta predicción fue incorporada al historial y se recalcularon las características requeridas para el 2 de enero.

Esta verificación permitió comprobar que las variables históricas utilizadas para el segundo día incorporaran correctamente la información generada durante el período anterior.

Asimismo, se verificó que la matriz predictora mantuviera las mismas 27 características utilizadas durante el entrenamiento y que no existieran valores nulos antes de ejecutar las predicciones.

3.18.4. Automatización del horizonte completo

Después de validar el procedimiento correspondiente a los primeros días, se implementó un bucle encargado de repetir automáticamente el proceso para las fechas restantes de 2018.

Para cada fecha se ejecutaron las siguientes operaciones:

1. selección de las 500 combinaciones tienda-producto correspondientes al día;
2. recuperación del historial actualizado;
3. generación de variables Lag;
4. cálculo de Rolling Mean;
5. generación de características EWMA;
6. construcción de la matriz de 27 variables;
7. generación de predicciones mediante LightGBM;
8. restricción de posibles predicciones negativas a cero;
9. almacenamiento de resultados;
10. incorporación de las predicciones al historial.

El procedimiento se repitió hasta completar el último día del año.

3.19. Consolidación del pronóstico

Una vez finalizado el procedimiento recursivo, las predicciones correspondientes a todas las fechas fueron consolidadas en un único conjunto de datos.

El resultado contiene cuatro variables principales:

- fecha;
- tienda;
- producto;
- predicción de ventas.

De acuerdo con la ejecución realizada, el pronóstico final contiene:

$$365 \times 500 = 182.500$$

observaciones.

3.20. Análisis posterior del pronóstico

Finalmente, las predicciones diarias fueron agregadas desde diferentes perspectivas con el propósito de facilitar su interpretación.

Se generaron resultados correspondientes a:

- pronóstico mensual;
- pronóstico anual por tienda;
- pronóstico anual por producto;
- productos con mayor demanda estimada;
- productos con menor demanda estimada;
- volumen total pronosticado;
- promedio diario;
- mes con mayor pronóstico;
- mes con menor pronóstico;
- comparación entre ventas reales de 2017 y pronóstico de 2018.

Estos resultados permiten transformar las predicciones individuales en indicadores con mayor utilidad para la interpretación empresarial.

3.21. Procedimiento general de la investigación

Etapa	Procedimiento
1	Carga y comprensión del conjunto histórico
2	Evaluación de calidad de los datos
3	Análisis exploratorio
4	Generación de características temporales
5	Construcción de Lag, Rolling Mean y EWMA
6	División temporal 2013–2016 / 2017
7	Entrenamiento de LightGBM
8	Predicción sobre 2017
9	Evaluación mediante MAE, RMSE, MAPE y R^2
10	Validación del proceso recursivo
11	Pronóstico diario de 2018
12	Consolidación y análisis de resultados

Tabla 11 - Procedimiento metodológico de la investigación

3.22. Consideraciones metodológicas

Para garantizar la coherencia del procedimiento predictivo se consideraron diferentes precauciones relacionadas con la naturaleza temporal de los datos.

En primer lugar, se preservó el orden cronológico durante la separación de los conjuntos de entrenamiento y validación. En segundo lugar, las variables históricas fueron construidas mediante desplazamientos destinados a utilizar únicamente información disponible antes de la observación que se pretendía estimar.

Asimismo, el pronóstico futuro se desarrolló mediante un procedimiento recursivo, evitando asumir la disponibilidad de ventas reales correspondientes a 2018.

Finalmente, antes de ejecutar cada predicción se mantuvo la misma estructura de 27 variables utilizada durante el entrenamiento y se verificó la ausencia de valores faltantes en la matriz predictora.

Estas consideraciones permiten reducir riesgos de fuga de información y mantener coherencia entre las condiciones utilizadas durante el entrenamiento, la validación y la generación del pronóstico futuro.

4. Resultados y Discusión

4.1. Introducción del capítulo

El presente capítulo expone los resultados obtenidos durante la aplicación del proceso de Ciencia de Datos para el desarrollo del modelo predictivo de ventas. La presentación de los hallazgos mantiene la secuencia metodológica definida mediante CRISP-DM, comenzando con la comprensión y exploración del conjunto de datos, continuando con su preparación e ingeniería de características, y posteriormente abordando el entrenamiento, evaluación y utilización del modelo LightGBM.

El análisis considera inicialmente los registros históricos correspondientes al período 2013–2017, con el propósito de identificar patrones asociados con la evolución temporal de las ventas, diferencias entre establecimientos y comportamiento de los productos. Posteriormente se presentan los resultados obtenidos después de transformar los datos mediante variables temporales, rezagos, promedios móviles y medias móviles exponencialmente ponderadas.

La capacidad predictiva del modelo se evalúa utilizando el año 2017 como período de validación y considerando las métricas MAE, RMSE, MAPE y R^2. Adicionalmente, se analiza el comportamiento de los residuos y la correspondencia entre valores reales y estimados desde diferentes niveles de agregación.

Finalmente, se presentan los resultados del pronóstico recursivo correspondiente al año 2018, incluyendo su comportamiento mensual y su distribución por tienda y producto, así como la comparación entre las ventas observadas durante 2017 y la demanda estimada para el período siguiente.

4.2. Comprensión de los datos

4.2.1. Descripción inicial del conjunto de datos

La primera etapa del análisis consistió en verificar la estructura general del conjunto de datos utilizado en la investigación. Para ello se cargaron los archivos proporcionados por la fuente seleccionada y se analizaron sus dimensiones, variables y tipos de datos.

El conjunto destinado al entrenamiento contiene 913.000 observaciones y cuatro variables originales, correspondientes a date, store, item y sales. Los registros representan ventas diarias para diez tiendas y cincuenta productos.

La variable date permite identificar la fecha de cada observación; store representa el identificador de la tienda; item identifica el producto; y sales corresponde a la cantidad de unidades vendidas.

El análisis de las primeras observaciones permitió confirmar que cada fila representa una combinación específica de fecha, tienda y producto. Por ejemplo, para el producto 1 de la tienda 1 se dispone de una observación diaria desde el inicio del período histórico.

Característica	Resultado
Observaciones de entrenamiento	913.000
Variables originales	4
Tiendas	10
Productos	50
Período	2013–2017
Frecuencia	Diaria

Tabla 12 - Estructura inicial del conjunto de datos

La amplitud del conjunto de datos permite analizar el comportamiento histórico de 500 combinaciones tienda-producto, correspondientes a la interacción entre diez establecimientos y cincuenta productos. Esta granularidad resulta adecuada para desarrollar un modelo que no se

limite al comportamiento agregado de las ventas, sino que incorpore diferencias entre unidades comerciales y productos.

4.2.2. Tipos de variables y estadísticas descriptivas

La revisión de la estructura del conjunto permitió comprobar inicialmente que la variable date se encontraba almacenada como texto, mientras que store, item y sales correspondían a variables numéricas. Por este motivo, la fecha fue posteriormente transformada a un formato temporal antes de realizar el análisis exploratorio.

Las estadísticas descriptivas de la variable ventas muestran un promedio de aproximadamente 52,25 unidades por observación, con una desviación estándar cercana a 28,80 unidades. La cantidad mínima registrada es cero, mientras que el máximo observado alcanza 231 unidades.

Los cuartiles muestran que el 25 % de las observaciones presenta ventas iguales o inferiores a 30 unidades, la mediana corresponde a 47 unidades y el 75 % de los registros se encuentra por debajo de 70 unidades.

Estadístico	Ventas
Observaciones	913.000
Media	52,25
Desviación estándar	28,80
Mínimo	0
Percentil 25	30
Mediana	47
Percentil 75	70
Máximo	231

Tabla 13 - Estadísticos descriptivos de la variable ventas

4.2.3. Calidad de los datos

Como parte del análisis de calidad se verificó la existencia de valores faltantes y registros duplicados.

Los resultados mostraron que las cuatro variables originales presentan cero valores nulos, por lo que no fue necesario implementar procedimientos de imputación sobre la información inicial.

De manera similar, la comprobación de duplicados determinó que el conjunto no contiene observaciones repetidas.

Verificación	Resultado
Valores nulos en date	0
Valores nulos en store	0
Valores nulos en item	0
Valores nulos en sales	0
Registros duplicados	0

Tabla 14 - Evaluación inicial de calidad de los datos

La ausencia de valores faltantes y registros duplicados evidencia que el conjunto presenta una estructura inicial consistente. No obstante, posteriormente se generan valores nulos de manera artificial como consecuencia de la construcción de variables rezagadas y móviles; dichos casos corresponden a una limitación matemática de las características históricas y no a problemas de calidad del dataset original.

4.2.4. Período temporal analizado

Después de transformar la variable date al formato correspondiente, se verificó el rango temporal disponible.

Los datos comienzan el 1 de enero de 2013 y finalizan el 31 de diciembre de 2017, lo que proporciona cinco años completos de información histórica.

La disponibilidad de registros continuos durante este período permite estudiar comportamientos anuales, mensuales y semanales, así como construir características históricas destinadas al entrenamiento del modelo.

Característica	Fecha
Primera observación	01/01/2013
Última observación	31/12/2017
Período total	5 años

Tabla 15 - Rango temporal del conjunto histórico

4.2.5. Evolución temporal de las ventas

El análisis agregado de las ventas diarias evidencia una estructura temporal caracterizada por tendencia creciente y patrones estacionales recurrentes durante el período 2013–2017.

La serie muestra incrementos y disminuciones que se repiten aproximadamente de manera anual. Paralelamente, el nivel general de ventas presenta un crecimiento progresivo: los máximos observados durante los últimos años superan los registrados al inicio del período.

Esta combinación de tendencia y estacionalidad confirma que las ventas no presentan un comportamiento aleatorio constante y respalda la necesidad de incorporar información temporal dentro del modelo predictivo.

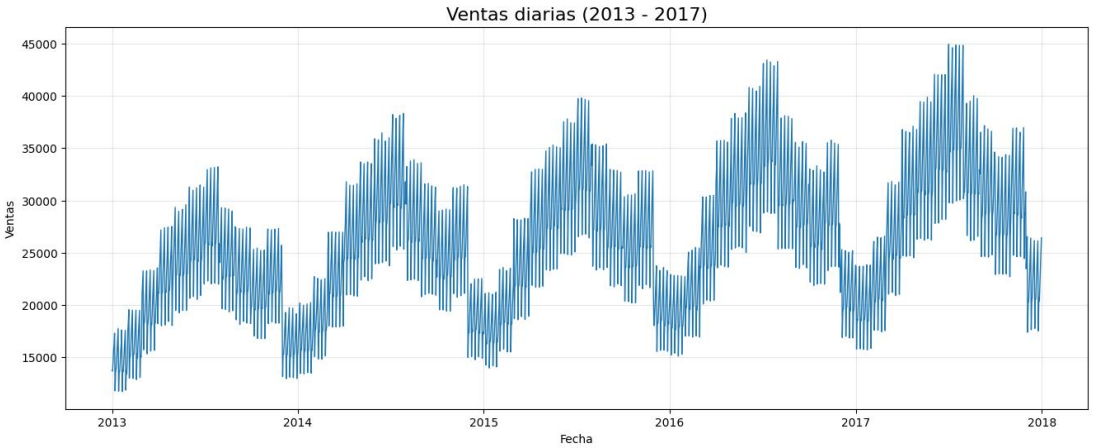


Ilustración 4 - Ventas diarias durante el período 2013–2017

Elaboración propia

Como se observa en la Ilustración 4, el volumen agregado de ventas presenta oscilaciones periódicas y un crecimiento general a lo largo del período analizado. Los niveles registrados durante 2016 y 2017 superan los observados durante los primeros años, mientras que las disminuciones identificadas en determinados períodos se repiten de forma similar entre años.

La presencia de estos patrones constituye un indicio de dependencia temporal y estacionalidad, justificando la posterior incorporación de características relacionadas con fecha, comportamiento histórico y promedios móviles.

4.2.6. Evolución anual de las ventas

La agregación de ventas por año permite observar con mayor claridad la tendencia general de crecimiento identificada en la serie diaria.

Los resultados obtenidos fueron:

Año	Ventas
2013	7.941.243
2014	9.135.482
2015	9.536.887
2016	10.357.160
2017	10.733.740

Tabla 16 - Evolución anual de las ventas

Durante los cinco años analizados se observa un incremento sostenido del volumen total de ventas. El menor valor corresponde a 2013, mientras que el máximo se registra en 2017.

Entre 2013 y 2017 las ventas pasan de 7.941.243 a 10.733.740 unidades, lo que evidencia una expansión relevante de la demanda dentro del conjunto analizado.

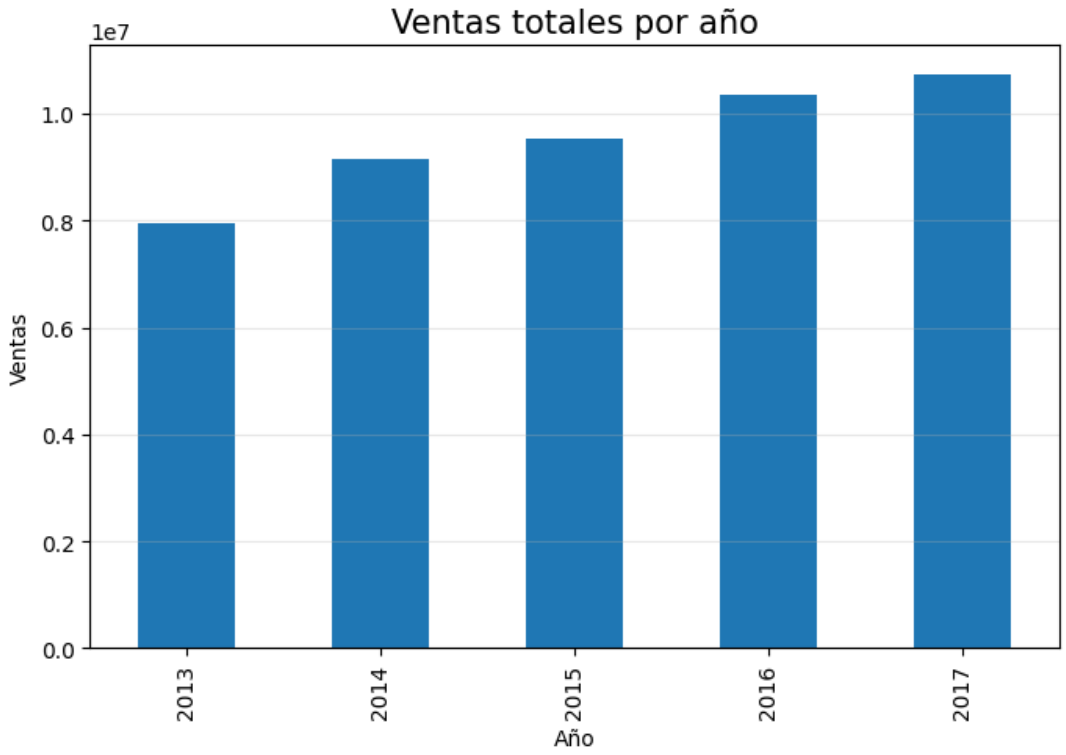


Ilustración 5 - Ventas totales por año

Elaboración propia

La Ilustración 5 confirma una tendencia creciente durante todo el período histórico. Aunque la magnitud del incremento varía entre años, no se observan disminuciones anuales en el total registrado.

Este comportamiento es relevante para el proceso predictivo debido a que indica que la estimación futura debe considerar no solamente patrones estacionales de corto plazo, sino también cambios de nivel asociados con la evolución de la demanda a lo largo del tiempo.

4.2.7. Comportamiento mensual de las ventas

Con el propósito de identificar posibles patrones estacionales, se agregaron las observaciones según el mes del año.

Los resultados muestran diferencias importantes entre los distintos meses. Julio registra el mayor volumen agregado, alcanzando aproximadamente 5,19 millones de unidades durante los cinco años analizados.

En contraste, enero y febrero presentan los menores niveles acumulados, con aproximadamente 2,75 y 2,78 millones de unidades respectivamente.

La demanda aumenta progresivamente durante los primeros meses del año, alcanza sus mayores niveles entre mayo y agosto y posteriormente presenta una reducción hacia el cierre anual.

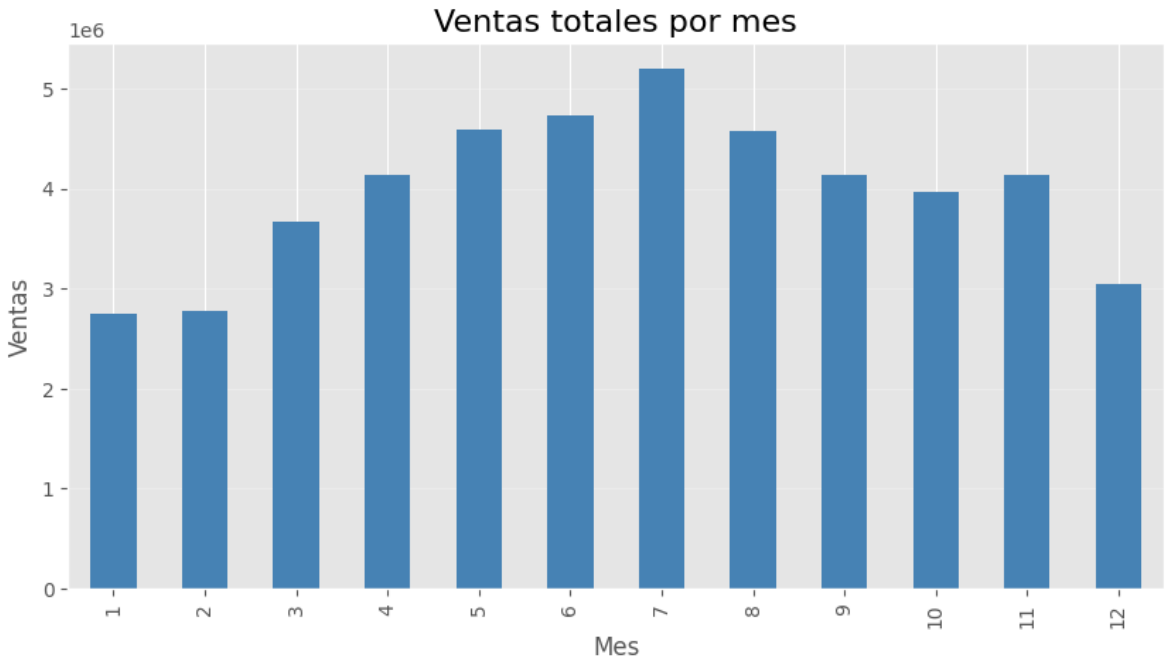


Ilustración 6 – Ventas totales según mes del año

Elaboración propia

La Ilustración 6 evidencia un patrón estacional claramente diferenciado. La concentración de mayores ventas en determinados meses y la repetición de niveles inferiores al inicio y final del año sugieren que la variable mes aporta información relevante para la predicción.

Este hallazgo sustenta la incorporación de características temporales dentro del modelo LightGBM, debido a que permite que el algoritmo represente diferencias sistemáticas entre períodos del año.

4.2.8. Comportamiento de las ventas por día de la semana

El análisis semanal permitió identificar diferencias en las ventas según el día en que se producen.

Los resultados agregados fueron:

Día	Ventas
Lunes	5.385.853
Martes	6.293.481
Miércoles	6.312.090
Jueves	6.749.880
Viernes	7.198.021
Sábado	7.655.482
Domingo	8.109.705

Tabla 17 - Comportamiento de las ventas por día de la semana

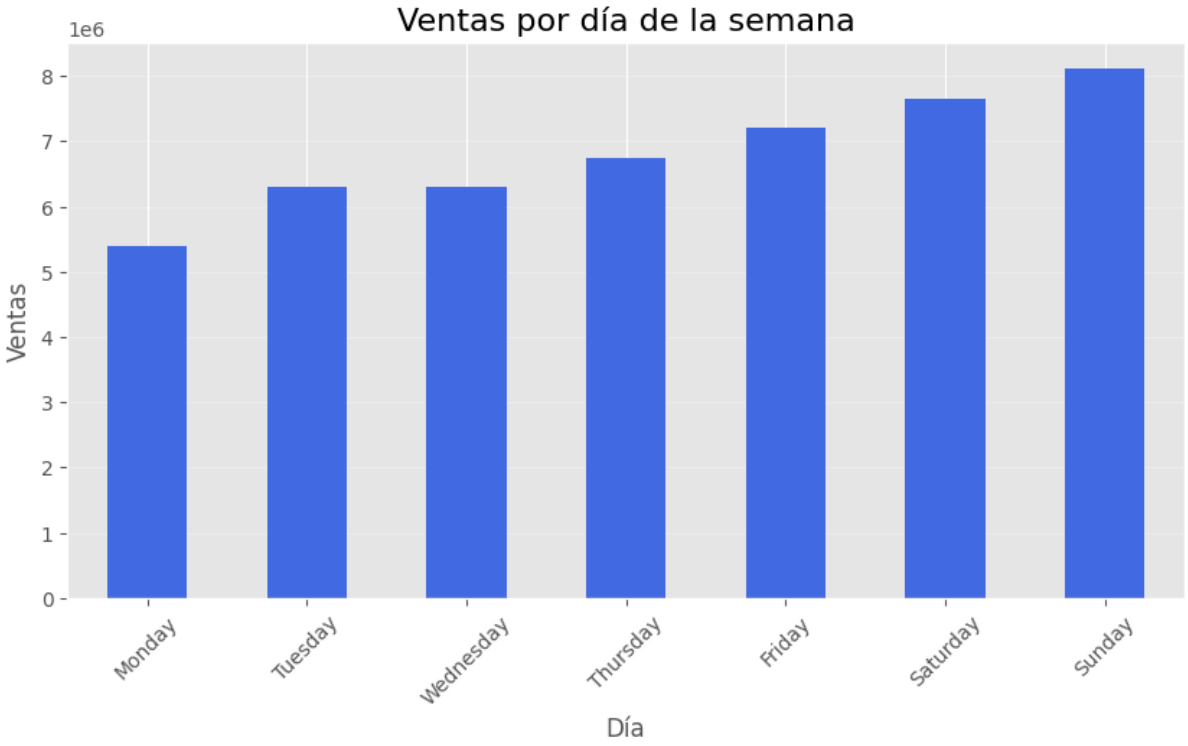


Ilustración 7 - Ventas por día de la semana

Elaboración propia

La tendencia observada sugiere la existencia de un patrón semanal en la demanda. El incremento hacia el fin de semana indica que la variable correspondiente al día de la semana puede contribuir a diferenciar períodos de menor y mayor actividad.

Este resultado respalda la utilización de dayofweek dentro de las características predictoras del modelo.

4.2.9. Análisis de ventas por tienda

El conjunto contiene diez establecimientos con volúmenes de venta diferenciados.

Los resultados acumulados muestran que la tienda 2 presenta el mayor nivel de ventas históricas, alcanzando 6.120.128 unidades. Le siguen principalmente las tiendas 8 y 3.

Por otra parte, la tienda 7 presenta el menor volumen acumulado, con aproximadamente 3,32 millones de unidades.

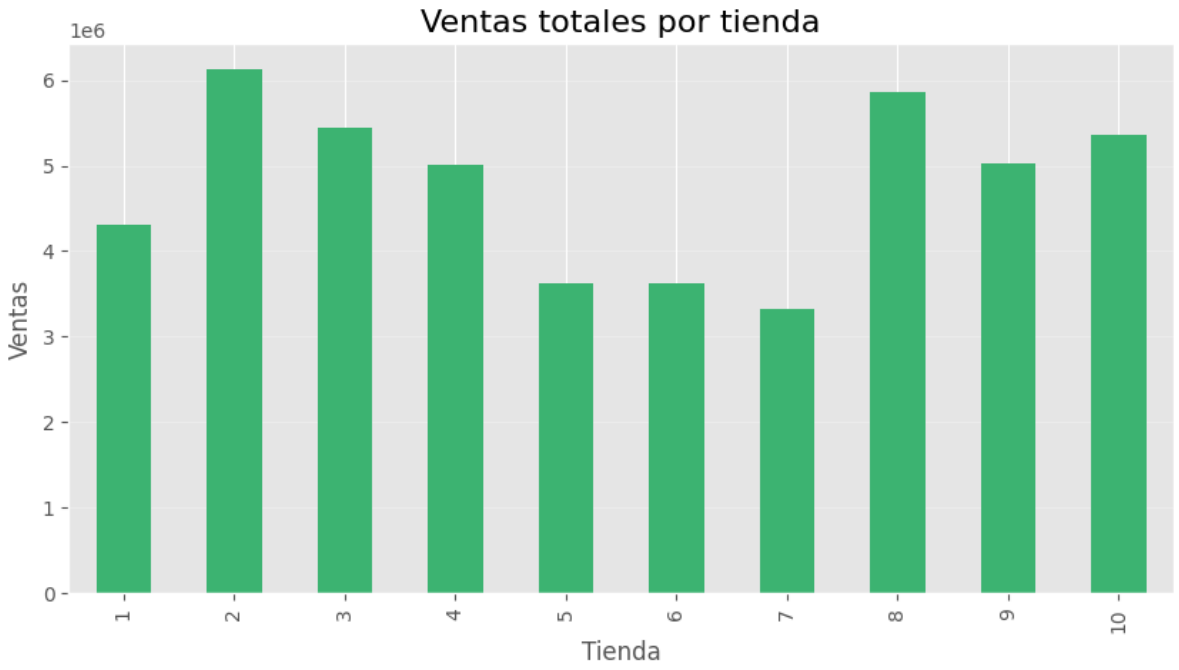


Ilustración 8 - Ventas totales por tienda

Elaboración propia

Las diferencias identificadas evidencian que el nivel de demanda no es homogéneo entre establecimientos. Por esta razón, la variable store se conserva como característica del modelo

predictivo, permitiendo que LightGBM identifique patrones específicos relacionados con cada establecimiento.

La tienda 2, que presenta el mayor nivel histórico, también constituye posteriormente uno de los establecimientos con mayor demanda pronosticada para 2018, aspecto que será analizado en la sección correspondiente al pronóstico futuro.

4.2.10. Análisis de ventas por producto

El análisis por producto evidencia una variabilidad considerable entre los cincuenta artículos incluidos en el conjunto de datos.

El producto 15 presenta el mayor volumen histórico, alcanzando aproximadamente 1.607.442 unidades, mientras que el producto 5 registra el menor nivel, con aproximadamente 335.230 unidades.

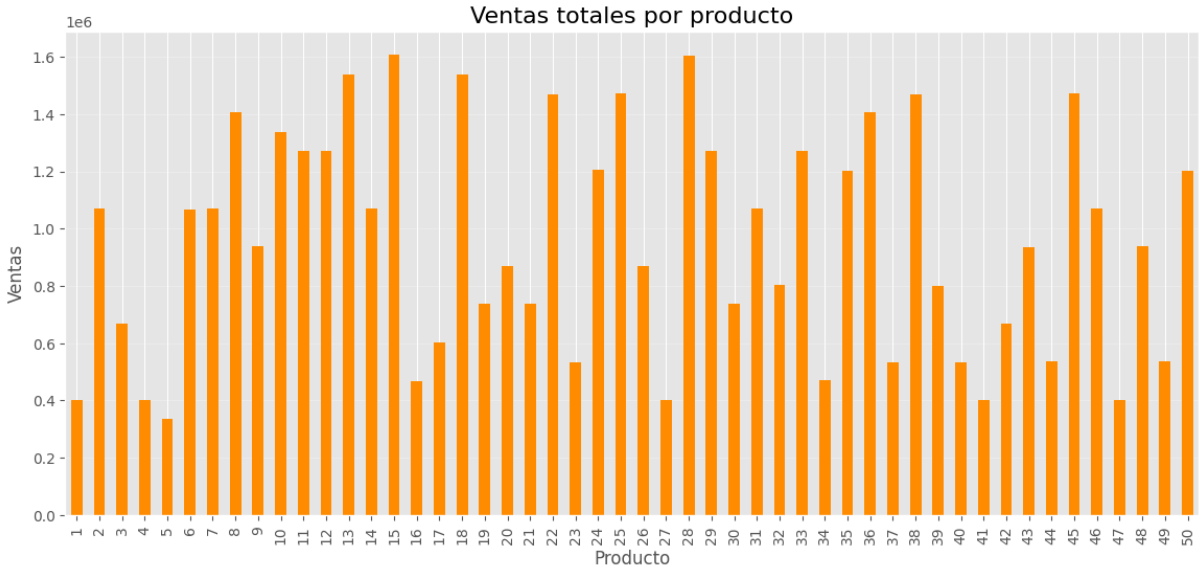


Ilustración 9 - Ventas totales por producto durante el período 2013–2017

Elaboración propia

La Ilustración 9 muestra que la demanda presenta diferencias sustanciales entre productos. Algunos artículos superan 1,5 millones de unidades durante el período histórico, mientras que otros permanecen por debajo de 500 mil.

Este comportamiento demuestra que utilizar únicamente una serie agregada de ventas ocultaría diferencias relevantes dentro de la demanda. La incorporación del identificador de producto permite que el modelo represente estas variaciones y genere estimaciones diferenciadas para cada artículo.

4.2.11. Relación entre tienda y producto

Para analizar conjuntamente ambas dimensiones se construyó una matriz de ventas por combinación tienda-producto y posteriormente se representó mediante un mapa de calor.

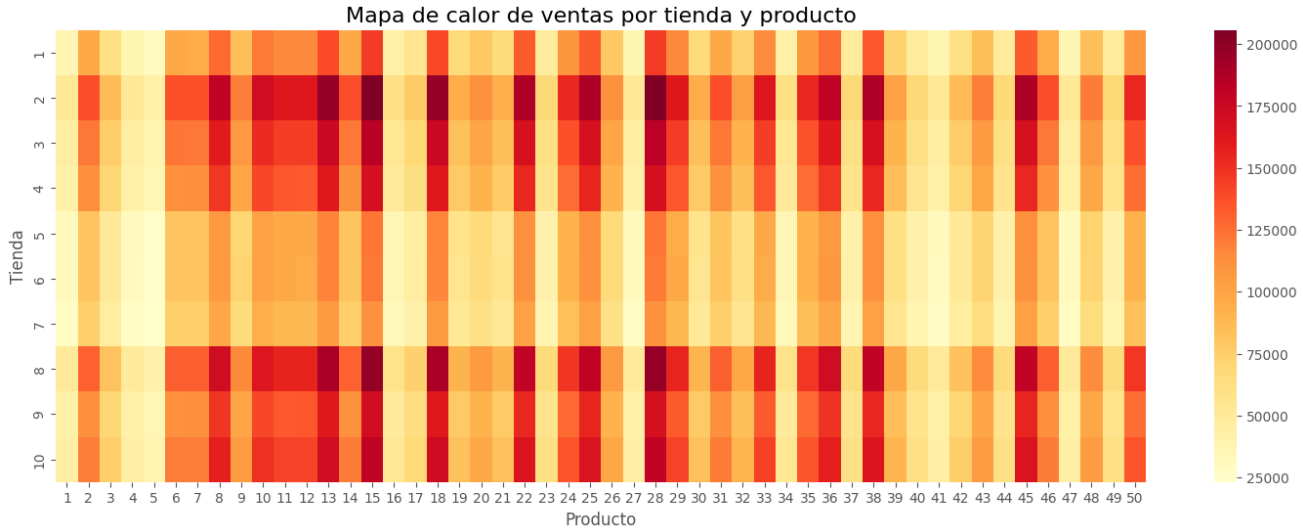


Ilustración 10 - Mapa de calor de ventas por tienda y producto

Elaboración propia

Nota. Las tonalidades de mayor intensidad representan combinaciones con mayores niveles acumulados de ventas. Elaboración propia.

El mapa de calor evidencia diferencias tanto verticales como horizontales en la intensidad de ventas. Determinados productos presentan mayores niveles de demanda de manera relativamente consistente entre tiendas, mientras que establecimientos como las tiendas 2 y 8 muestran intensidades superiores para una proporción importante de los productos.

En contraste, otras tiendas presentan volúmenes inferiores de manera sistemática.

Estos resultados confirman que la demanda depende conjuntamente del establecimiento y del producto. En consecuencia, ambas variables son incorporadas al proceso de modelado.

4.3. Preparación de los datos

Después de completar el análisis exploratorio se procedió a transformar el conjunto histórico con el propósito de construir las características utilizadas por el modelo predictivo.

La preparación consistió principalmente en la creación de variables temporales, rezagos, promedios móviles y medias móviles exponencialmente ponderadas. Estos procesos permitieron transformar la información de ventas en una estructura apropiada para un problema de regresión supervisada. Esta etapa corresponde directamente con la fase de preparación de CRISP-DM; en la tesis de ejemplo también se presenta como una sección diferenciada entre comprensión de datos y modelado.

4.3.1. Variables temporales generadas

A partir de la fecha original se generaron seis variables adicionales:

Variable	Significado
year	Año
month	Mes
day	Día del mes
dayofweek	Día de la semana
weekofyear	Semana del año
quarter	Trimestre

Tabla 18 - Variables temporales generadas

La generación de estas características permite representar diferentes componentes temporales observados durante el análisis exploratorio.

Particularmente, los resultados anteriores mostraron diferencias relevantes tanto entre meses como entre días de la semana, por lo que estas variables aportan al algoritmo información que no se encuentra explícitamente representada por los identificadores de tienda y producto.

4.3.2. Construcción de variables Lag

Con el propósito de incorporar información sobre valores históricos, se generaron cuatro características rezagadas:

- lag_1;
- lag_7;
- lag_14;
- lag_28.

Cada variable representa las ventas correspondientes a la misma combinación tienda-producto en un momento anterior.

Por ejemplo, lag_7 permite incorporar como característica la cantidad registrada siete días antes, mientras que lag_28 proporciona información correspondiente a aproximadamente cuatro semanas previas.

La generación se realizó independientemente dentro de cada combinación tienda-producto mediante desplazamiento temporal, evitando utilizar el valor correspondiente al mismo día que se pretende predecir.

Variable	Rezago
Lag 1	1 día
Lag 7	7 días
Lag 14	14 días
Lag 28	28 días

Tabla 19 - Construcción de variables Lag

4.3.3. Valores faltantes generados por los rezagos

La construcción de características históricas genera valores faltantes en las primeras observaciones de cada serie porque todavía no existe suficiente información anterior.

Los resultados obtenidos muestran:

Variable	Valores nulos
lag_1	500
lag_7	3.500
lag_14	7.000
lag_28	14.000

Tabla 20 - Valores faltantes generados por los rezagos

Estos valores son coherentes con las 500 combinaciones tienda-producto existentes.

Por ejemplo:

$$500 \times 28 = 14.000$$

Por esta razón, los 14.000 valores faltantes identificados en lag_28 no constituyen un problema de calidad del dataset original, sino una consecuencia esperada de utilizar un rezago de 28 días.

4.3.4. Construcción de Rolling Mean

Se generaron promedios móviles correspondientes a ventanas de 7, 14 y 28 días.

La utilización del desplazamiento previo al cálculo de la media permitió garantizar que cada observación utilizara únicamente información registrada antes del día correspondiente.

Los valores faltantes generados fueron:

Variable	Valores nulos
Rolling Mean 7	3.500
Rolling Mean 14	7.000
Rolling Mean 28	14.000

Tabla 21 - Construcción de Rolling Mean

Esta correspondencia nuevamente responde a la estructura de 500 series individuales.

4.3.5. Construcción de variables EWMA

Adicionalmente se construyeron características mediante medias móviles exponencialmente ponderadas.

Se utilizaron cuatro factores de suavizamiento:

$$\alpha = \{0,95;\ 0,90;\ 0,80;\ 0,70\}$$

y tres rezagos:

$$Lag = \{1;\ 7;\ 14\}$$

La combinación de estos parámetros generó:

$$4 \times 3 = 12$$

variables EWMA.

Estas características permiten asignar mayor importancia a observaciones recientes y complementar la información proveniente de los rezagos y promedios móviles.

Los resultados de la implementación confirmaron correctamente la generación de 12 variables EWMA.

Factor	Valores
Alpha	0,95; 0,90; 0,80; 0,70
Rezagos	1; 7; 14
Características generadas	12

Tabla 22 - Configuración de características EWMA

4.3.6. Eliminación de valores nulos y dimensiones finales

Una vez generadas todas las características temporales, se eliminaron las observaciones iniciales que no contaban con suficiente historial para calcular la totalidad de las variables.

Antes de este procedimiento el conjunto contenía 913.000 registros.

Posteriormente, el dataset preparado quedó conformado por:

- 899.000 observaciones y 29 columnas.

La reducción corresponde a:

$$913.000 - 899.000 = 14.000$$

observaciones, coincidiendo con el número necesario para disponer de una ventana histórica completa de 28 días en las 500 combinaciones tienda-producto.

Este resultado confirma la consistencia del proceso de preparación y permite trabajar posteriormente con una matriz sin valores nulos.

Etapa	Observaciones
Dataset original	913.000
Dataset preparado	899.000
Registros eliminados	14.000

Tabla 23 - Dimensiones antes y después de la preparación

4.4. Modelado

Una vez concluida la preparación del conjunto de datos, se procedió a construir el modelo predictivo utilizando el algoritmo Light Gradient Boosting Machine (LightGBM).

La selección del algoritmo responde a la necesidad de trabajar con un conjunto de gran volumen y múltiples características temporales. El modelo fue entrenado mediante una tarea de regresión supervisada, utilizando sales como variable objetivo y las 27 características generadas durante la etapa de preparación como variables predictoras.

A diferencia del planteamiento inicial del anteproyecto, que contemplaba la comparación de múltiples algoritmos, la implementación final se concentra en LightGBM como modelo predictivo principal. El anteproyecto originalmente proponía modelos de series temporales y Machine Learning diversos, pero esa estructura debe actualizarse en la versión definitiva para mantener correspondencia con el desarrollo experimental realmente ejecutado.

4.4.1. División temporal entre entrenamiento y validación

Con el propósito de evaluar la capacidad del modelo para generalizar sobre información futura, el conjunto de datos fue dividido respetando el orden cronológico.

El conjunto de entrenamiento contiene observaciones desde el 29 de enero de 2013 hasta el 31 de diciembre de 2016, mientras que el conjunto de validación comprende el período entre el 1 de enero y el 31 de diciembre de 2017.

Las dimensiones obtenidas fueron:

Conjunto	Observaciones	Variables
Entrenamiento	716.500	29
Validación	182.500	29

Tabla 24 - División temporal entre entrenamiento y validación

La separación temporal evita realizar una división aleatoria que pudiera introducir información futura durante el aprendizaje. Este aspecto resulta especialmente importante en problemas de forecasting, debido a que el objetivo consiste en simular un escenario en el que el modelo aprende exclusivamente con información histórica.



Ilustración 11 - División temporal del conjunto de datos para entrenamiento y validación

La utilización de 2017 como período completo de validación permite evaluar el modelo bajo diferentes niveles de demanda y patrones estacionales a lo largo de doce meses, proporcionando una evaluación más representativa que una selección aleatoria de observaciones.

4.4.2. Selección de las variables predictoras

Para el entrenamiento se utilizaron 27 variables predictoras agrupadas en cinco categorías: identificación, variables temporales, rezagos, promedios móviles y características EWMA.

La variable objetivo correspondió a sales.

Grupo	Cantidad
Tienda y producto	2
Variables temporales	6
Variables Lag	4
Rolling Mean	3
EWMA	12
Total	**27**

Tabla 25 - Variables utilizadas para el entrenamiento de LightGBM

Esta configuración permite combinar información estructural de la observación con el comportamiento histórico reciente de cada combinación tienda-producto.

4.4.3. Configuración y entrenamiento de LightGBM

El modelo fue configurado como un problema de regresión y se entrenó utilizando los registros pertenecientes al conjunto temporal de entrenamiento.

De acuerdo con la implementación realizada, los principales parámetros utilizados fueron:

Parámetro	Valor
Objetivo	Regresión
Learning rate	0,05
Número de estimadores	1.000
Número de hojas	64
Random state	42

Tabla 26 - Configuración y entrenamiento de LightGBM

La tasa de aprendizaje controla la magnitud con la que cada árbol contribuye al modelo final, mientras que el número de estimadores determina la cantidad de árboles construidos durante el entrenamiento.

El procedimiento procesó 716.500 observaciones y 27 características predictoras, concluyendo correctamente el entrenamiento del modelo.

4.4.4. Importancia de las variables

Después del entrenamiento se analizó la importancia de las características utilizadas por LightGBM.

Los resultados muestran que las variables asociadas con el comportamiento histórico reciente presentan una participación relevante dentro del modelo.

Las características con mayor importancia fueron:

Variable	Importancia
rolling_mean_28	6.683
lag_28	4.869
day	4.252
rolling_mean_7	3.979
rolling_mean_14	2.954
dayofweek	2.871
lag_7	2.843
lag_14	2.824
weekofyear	2.820
ewm_alpha_07_lag_1	2.713
lag_1	2.555
item	2.461

Tabla 27 - Importancia de las variables

El mayor valor corresponde a rolling_mean_28, seguido por lag_28. Este resultado evidencia que la información correspondiente a períodos anteriores y al comportamiento promedio de las últimas semanas tiene una participación significativa en las decisiones realizadas por el modelo.

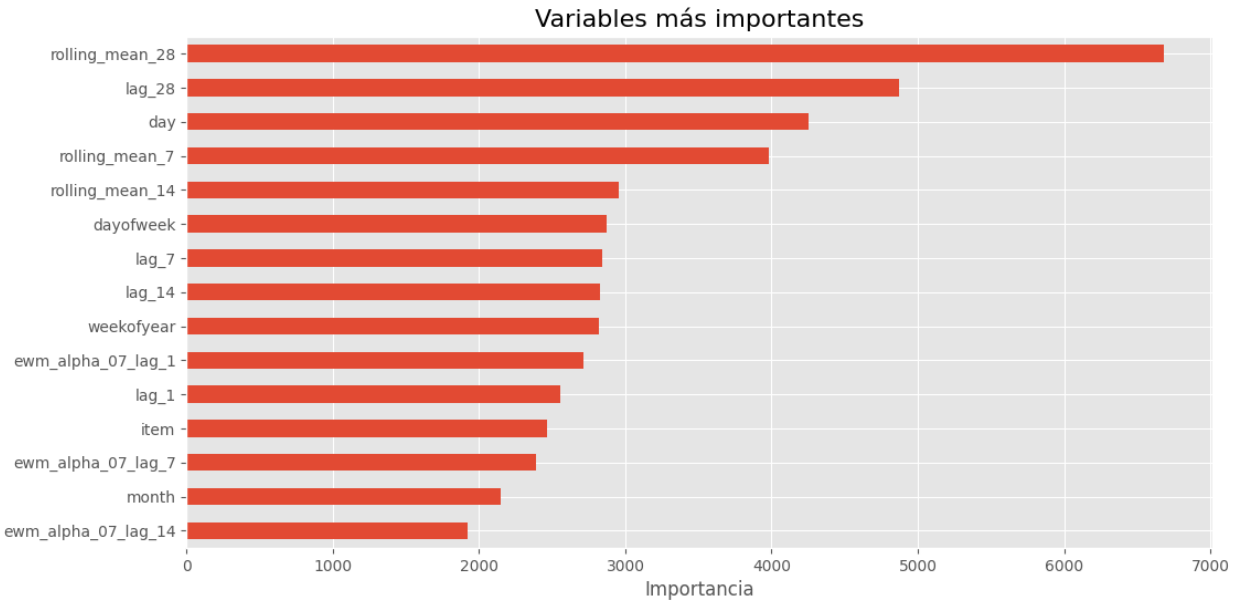


Ilustración 12 - Variables más importantes

Nota. La importancia corresponde al indicador calculado por el modelo LightGBM. Elaboración propia.

La Ilustración 12 permite observar que las variables históricas ocupan las primeras posiciones dentro del ranking. En particular, rolling_mean_28, lag_28, rolling_mean_7, rolling_mean_14, lag_7, lag_14 y lag_1 aparecen entre las características más relevantes.

Este resultado respalda la decisión metodológica de complementar las variables originales con ingeniería de características temporales, debido a que el algoritmo utiliza efectivamente la información derivada del historial para construir sus estimaciones.

También resulta relevante la presencia de variables de calendario como day, dayofweek, weekofyear y month, lo que guarda correspondencia con los patrones mensuales y semanales identificados durante el análisis exploratorio.

4.4.5. Evaluación del modelo

Una vez entrenado LightGBM, se generaron predicciones sobre las 182.500 observaciones correspondientes a 2017.

La evaluación se realizó comparando las ventas reales con los valores estimados mediante cuatro métricas: MAE, RMSE, MAPE y R^2.

Los resultados obtenidos fueron los siguientes:

Métrica	Resultado
MAE	6,0989
RMSE	7,9252
MAPE	12,58 %
R^2	0,9369

Tabla 28 - Desempeño general del modelo LightGBM

Estos indicadores muestran una capacidad predictiva favorable sobre el conjunto temporal de validación. No obstante, cada métrica representa un aspecto diferente del error, por lo que su interpretación debe realizarse de manera conjunta.

4.5. Evaluación del desempeño predictivo

4.5.1. Comparación con el modelo baseline

Con el propósito de establecer un punto de referencia para evaluar el desempeño predictivo de LightGBM, se implementó un modelo baseline mediante un pronóstico ingenuo estacional de siete días. Este enfoque utiliza como predicción las ventas registradas para la misma combinación de tienda y producto siete días antes.

El modelo baseline fue evaluado sobre las mismas 182.500 observaciones correspondientes al año 2017 utilizadas para la evaluación de LightGBM. De esta manera, ambos enfoques fueron comparados bajo las mismas condiciones mediante las métricas MAE, RMSE, MAPE y R^2.

Métrica	Baseline	LightGBM
MAE	8,8415	6,0989
RMSE	11,5967	7,9252
MAPE	17,94 %	12,58 %
R^2	0,8649	0,9369

Tabla 29 - Comparación del desempeño predictivo entre el modelo baseline y LightGBM

Los resultados evidencian que LightGBM presenta un desempeño superior al modelo baseline. El MAE disminuye de 8,8415 a 6,0989 y el RMSE de 11,5967 a 7,9252, lo que indica una menor magnitud de los errores de predicción. Asimismo, el MAPE se reduce de 17,94 % a 12,58 %, mientras que el R^2 aumenta de 0,8649 a 0,9369.

En conjunto, estos resultados demuestran que LightGBM aporta una mejora predictiva respecto de una estrategia sencilla basada únicamente en el comportamiento observado siete días antes. Por tanto, la incorporación de variables temporales, rezagos, promedios móviles y características EWMA permite capturar patrones adicionales del comportamiento de las ventas y mejorar la precisión de las estimaciones.

4.5.2. Error Absoluto Medio — MAE

El MAE obtenido fue de:

$$MAE = 6,0989$$

Este resultado indica que, en promedio, la diferencia absoluta entre la cantidad real de ventas y la estimación del modelo es de aproximadamente 6,10 unidades por observación.

Debido a que la métrica conserva la unidad original de la variable objetivo, su interpretación resulta directa.

No debe interpretarse como que todas las predicciones difieren exactamente en seis unidades, sino como la magnitud promedio del error absoluto registrado en el conjunto de validación.

4.5.3. Raíz del Error Cuadrático Medio — RMSE

El RMSE alcanzó:

$$RMSE = 7,9252$$

Su valor superior al MAE evidencia la presencia de determinados errores de mayor magnitud, debido a que el RMSE penaliza más intensamente las diferencias elevadas.

Sin embargo, la diferencia entre ambas métricas no resulta excesivamente amplia, lo que sugiere que, aunque existen observaciones con errores superiores al promedio, estos no dominan completamente el desempeño general.

4.5.4. Coeficiente de determinación — R²

El modelo obtuvo:

$$R^2 = 0,9369$$

Esto indica que aproximadamente el 93,69 % de la variabilidad observada en las ventas del conjunto de validación es representada por el modelo.

El resultado constituye un indicador favorable del ajuste obtenido sobre datos no utilizados durante el entrenamiento.

No obstante, R² no debe utilizarse de manera aislada para determinar la calidad de un modelo predictivo. Por esta razón, el resultado se analiza conjuntamente con las métricas de error y con las comparaciones gráficas posteriores.

4.5.5. Error Porcentual Absoluto Medio — MAPE

El MAPE calculado fue de:

$$MAPE = 12,58\%$$

Esto significa que, en promedio, el error absoluto representa aproximadamente el 12,58 % del valor real de ventas.

La interpretación de esta métrica requiere especial cuidado cuando las ventas reales son pequeñas, debido a que una diferencia absoluta moderada puede convertirse en un error porcentual considerable.

4.5.6. Comparación entre ventas reales y predichas

Para complementar las métricas se realizó una comparación gráfica entre valores reales y estimados.

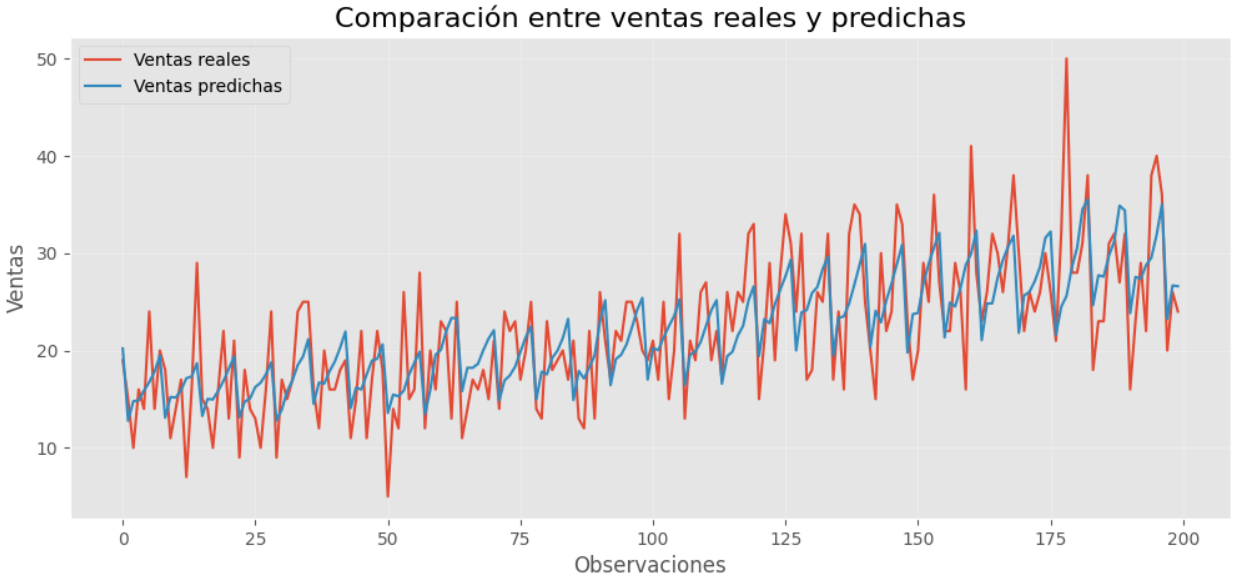


Ilustración 13 - Comparación entre ventas reales y predichas

Nota. La visualización corresponde a las primeras 200 observaciones del conjunto de validación y se utiliza únicamente como inspección gráfica del comportamiento predictivo.

La Ilustración 13 muestra que las estimaciones siguen de manera general las variaciones presentes en las observaciones reales. El modelo reproduce el comportamiento central de los datos, aunque se evidencian diferencias en determinados picos y caídas puntuales.

Las predicciones presentan una trayectoria más suavizada en comparación con los valores reales, especialmente en observaciones donde se producen cambios abruptos. Este comportamiento resulta consistente con las métricas obtenidas: el modelo presenta un ajuste general elevado, aunque mantiene errores individuales en determinadas observaciones.

4.5.7. Análisis de residuos

Los residuos fueron calculados mediante:

$$e_i = y_i - \hat{y}_i$$

donde y_i representa la venta real y $\hat{y}_i$ la predicción generada.

Se obtuvieron 182.500 residuos, con una media de:

$$\bar{e} = 0,0328$$

El valor cercano a cero indica que, en promedio, las diferencias positivas y negativas se encuentran aproximadamente equilibradas, sin evidenciar un sesgo medio considerable hacia la sobreestimación o subestimación.

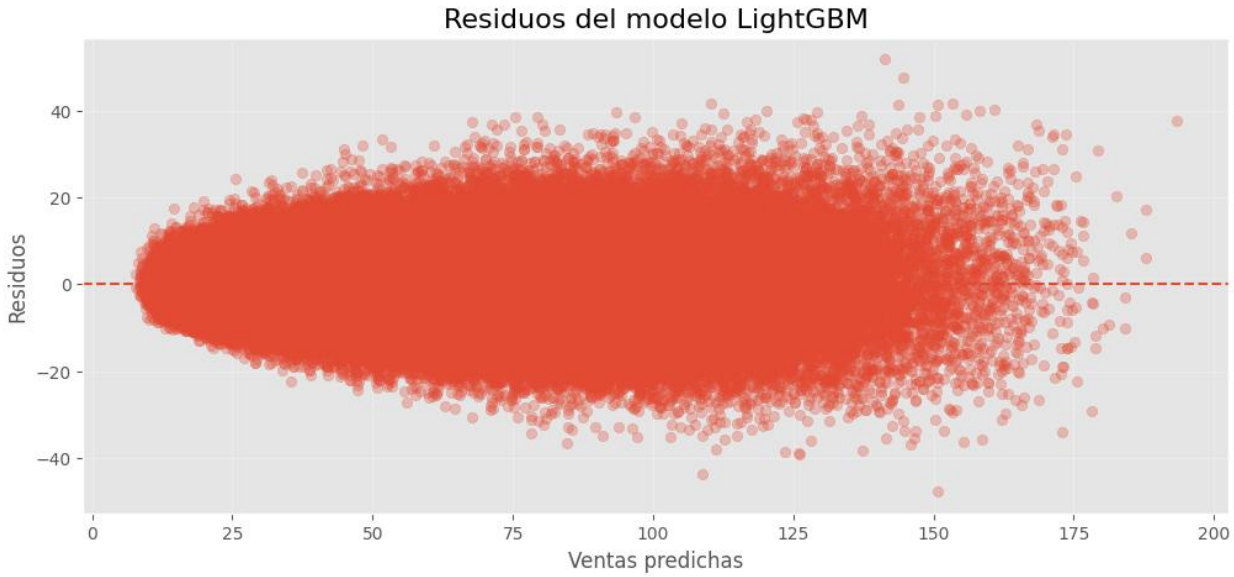


Ilustración 14 - Residuos del modelo LightGBM

La Ilustración 14 muestra que los residuos se distribuyen alrededor de cero, aunque su dispersión aumenta progresivamente conforme crece el nivel de ventas predicho.

Este patrón evidencia una mayor variabilidad de los errores en observaciones con niveles elevados de demanda. Por tanto, aunque el modelo no presenta un sesgo promedio importante, la incertidumbre de la predicción aumenta en escenarios de ventas altas.

Esta observación complementa las métricas globales y permite identificar una limitación del modelo que no sería visible utilizando únicamente MAE, RMSE o R^2.

4.5.8. Distribución de los residuos

También se analizó la distribución de los errores mediante un histograma.

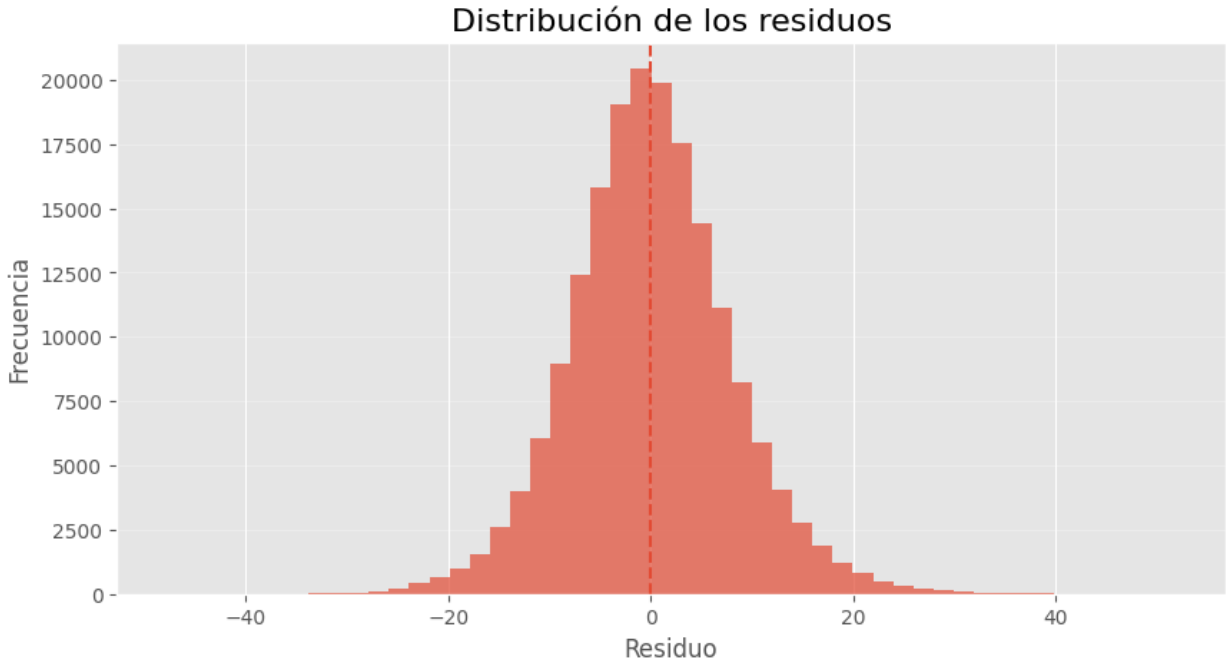


Ilustración 15 - Distribución de los residuos

La distribución muestra una concentración elevada de residuos alrededor de cero y una forma aproximadamente simétrica.

La mayor parte de las diferencias se encuentra dentro de rangos relativamente reducidos, mientras que los errores extremos presentan menor frecuencia.

La combinación entre una media residual cercana a cero y una distribución centralizada respalda la ausencia de un sesgo global marcado. Sin embargo, la presencia de colas en ambos extremos confirma que existen casos particulares donde las estimaciones se alejan considerablemente de los valores observados.

4.5.9. Evaluación agregada por tienda

Con el propósito de comprobar si el desempeño general se conserva entre establecimientos, se agregaron las ventas reales y predichas correspondientes a cada tienda durante el período de validación.

Los resultados muestran valores muy cercanos en las diez tiendas.

Tienda	Ventas reales	Ventas predichas aprox.
1	971.774	972.826
2	1.376.386	1.371.944
3	1.223.966	1.222.837
4	1.127.450	1.128.035
8	1.317.876	1.314.207
10	1.204.684	1.203.983

Tabla 30 - Evaluación agregada por tienda

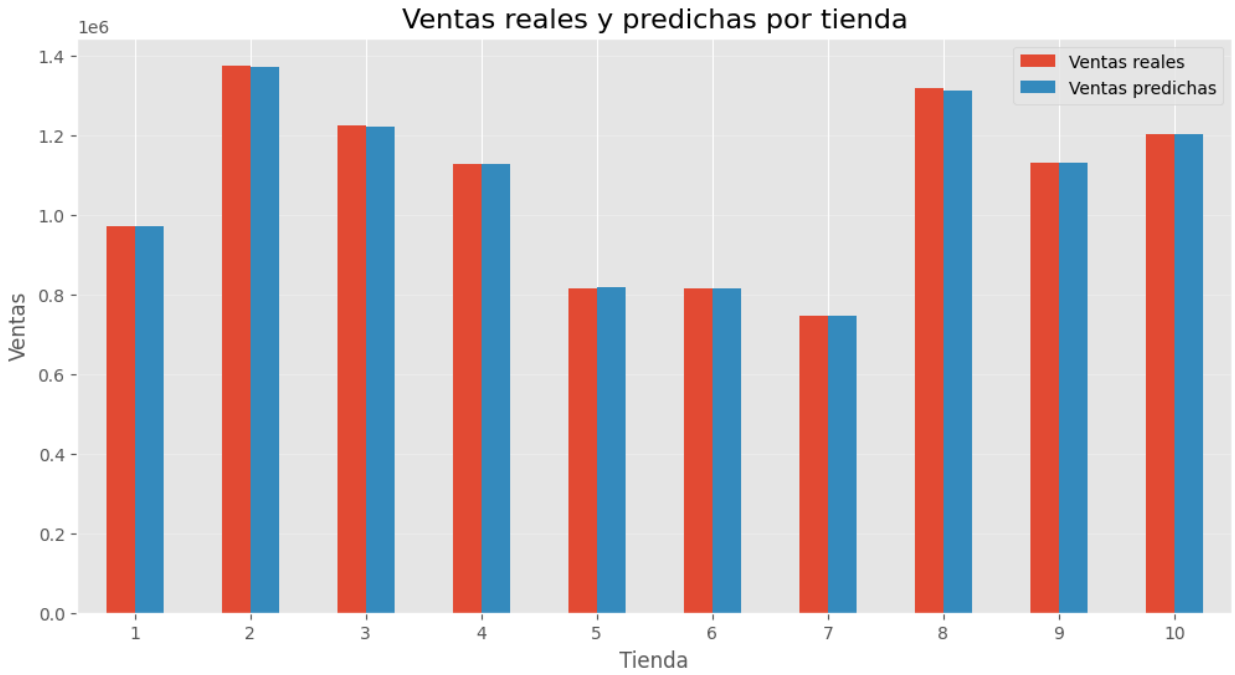


Ilustración 16 - Ventas reales y predichas por tienda

Nota. Los valores corresponden a la suma de las observaciones pertenecientes a cada tienda durante el período de validación. Elaboración propia.

La comparación evidencia una alta correspondencia entre el volumen acumulado real y predicho en los distintos establecimientos.

Las diferencias son pequeñas en relación con el nivel anual de ventas de cada tienda, lo que indica que el modelo conserva adecuadamente las diferencias estructurales de demanda entre establecimientos cuando se analiza la información de forma agregada.

Sin embargo, este resultado no significa que todas las predicciones diarias individuales presenten errores mínimos. La elevada correspondencia agregada puede estar influida por compensaciones entre sobreestimaciones y subestimaciones producidas a lo largo del año.

4.5.10. Evaluación agregada por producto

Posteriormente se realizó el mismo análisis para los cincuenta productos.

Los resultados muestran nuevamente una elevada correspondencia entre las ventas reales acumuladas y las estimaciones.

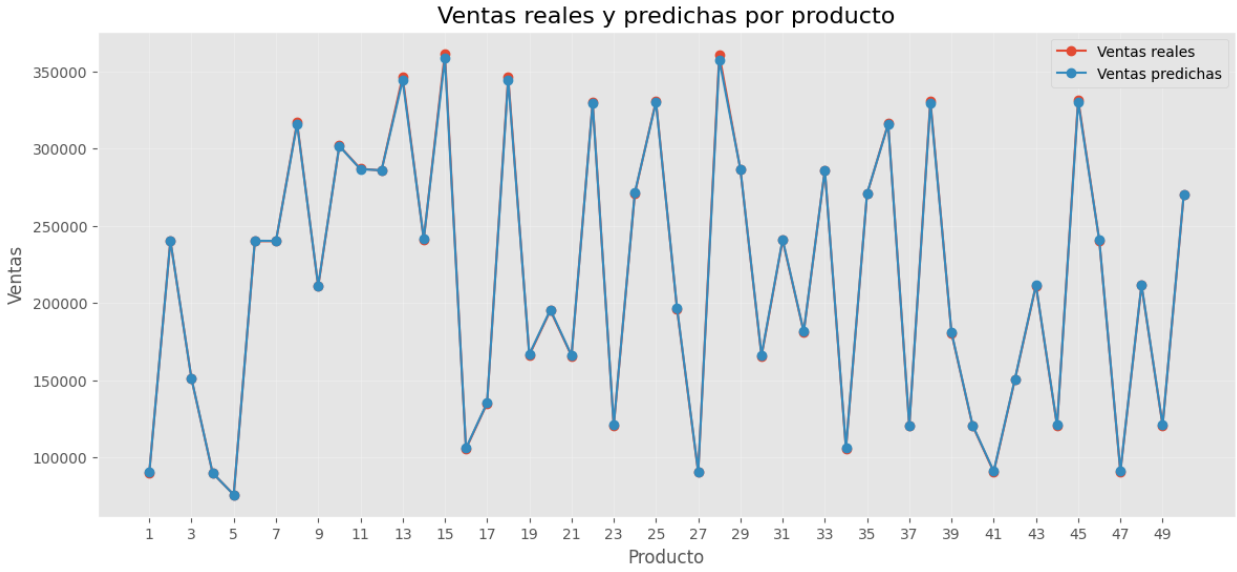


Ilustración 17 - Ventas reales y predichas por producto

Nota. Los valores corresponden a las ventas acumuladas de cada producto durante el período de validación.

Las trayectorias correspondientes a ventas reales y predichas se encuentran muy próximas para los cincuenta productos. El modelo reproduce tanto las diferencias entre artículos de alta demanda como aquellas correspondientes a productos con menor volumen de ventas.

Por ejemplo, los productos con mayores niveles acumulados mantienen posiciones similares en ambas series, mientras que los productos de menor demanda también son representados con valores próximos a los observados.

Este comportamiento evidencia que LightGBM logra mantener las diferencias relativas existentes entre productos cuando los resultados son analizados a nivel agregado.

Nuevamente, este resultado debe diferenciarse del error a nivel de observación individual, debido a que el efecto de agregación reduce parcialmente las diferencias positivas y negativas.

4.5.11. Evaluación mensual del período de validación

Para estudiar la capacidad del modelo para representar la evolución temporal anual, se agregaron las predicciones y valores reales por mes.

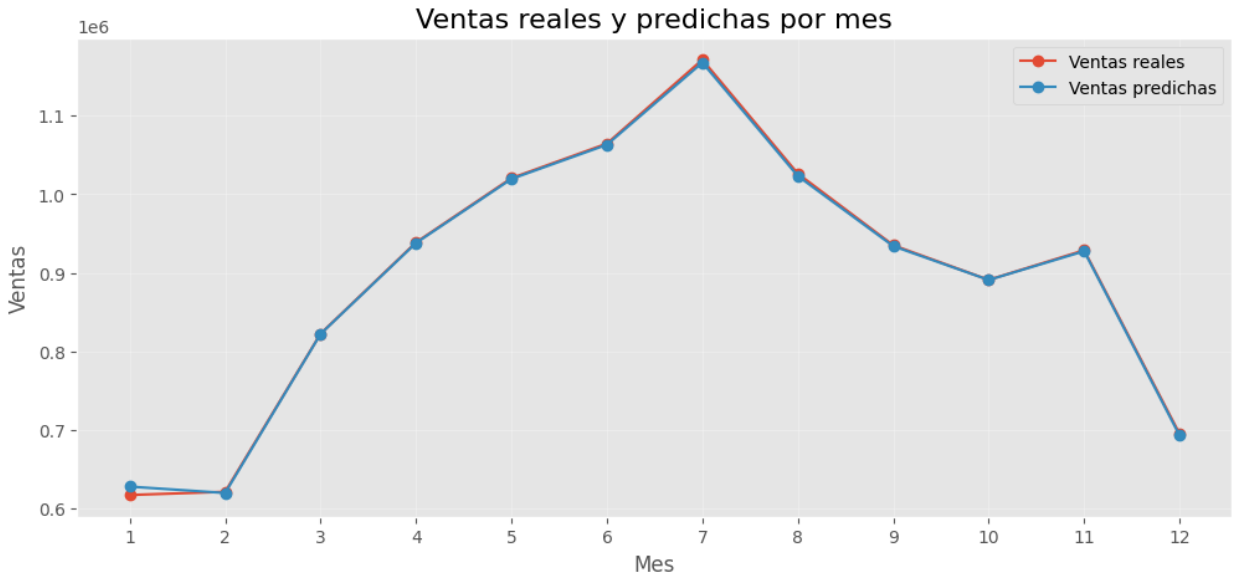


Ilustración 18 - Ventas reales y predichas por mes

La Ilustración 18 muestra una correspondencia elevada entre ambas series durante los doce meses del período de validación.

El modelo reproduce el incremento observado durante los primeros meses del año, el máximo registrado en julio y la disminución posterior hacia el cierre anual.

Las diferencias acumuladas por mes son reducidas en comparación con el volumen total registrado, lo que evidencia que LightGBM logra representar adecuadamente la evolución temporal general de la demanda cuando las predicciones se analizan de manera agregada.

El mayor error porcentual mensual se registró en enero, con aproximadamente 1,73 %, mientras que octubre presentó el menor error, cercano a 0,02 %.

En once de los doce meses el error agregado se mantuvo igual o por debajo de aproximadamente 0,32 %.

4.5.12. Síntesis del desempeño del modelo

En conjunto, los resultados obtenidos permiten afirmar que LightGBM presenta un desempeño favorable sobre el período de validación temporal.

El modelo alcanzó un MAE de 6,0989, RMSE de 7,9252, MAPE de 12,58 % y R^2 de 0,9369. Además, las comparaciones agregadas por tienda, producto y mes muestran una elevada correspondencia entre valores reales y estimados.

El análisis de residuos evidencia una media cercana a cero, aunque también muestra un incremento en la dispersión de errores conforme aumenta el nivel de ventas predicho. Esto indica que el modelo representa adecuadamente el comportamiento general de la demanda, pero presenta mayor incertidumbre en niveles elevados de ventas.

Asimismo, la importancia de variables evidencia que las características asociadas con información histórica —particularmente los promedios móviles y rezagos— tienen una participación significativa dentro del proceso predictivo.

A partir de estos resultados, el modelo fue utilizado posteriormente para generar estimaciones correspondientes al horizonte futuro de 2018.

4.6. Pronóstico recursivo de ventas para 2018

Una vez evaluado el desempeño del modelo LightGBM utilizando las observaciones correspondientes a 2017, se procedió a generar el pronóstico de ventas para el año 2018.

Debido a que el modelo utiliza variables construidas a partir de información histórica, como rezagos, promedios móviles y medias móviles exponencialmente ponderadas, no fue posible generar todas las predicciones futuras de manera independiente. Las características requeridas para determinadas fechas de 2018 dependen de ventas correspondientes a días anteriores del mismo período, cuyos valores reales no se encuentran disponibles.

Por esta razón, se implementó un procedimiento de pronóstico recursivo, mediante el cual cada predicción generada fue incorporada al historial disponible y utilizada posteriormente para calcular las características necesarias para el siguiente día.

El horizonte considerado comprende desde el 1 de enero hasta el 31 de diciembre de 2018, incluyendo las diez tiendas y los cincuenta productos disponibles en el conjunto de datos.

4.6.1. Construcción del horizonte futuro

Inicialmente se generó una estructura temporal correspondiente a los 365 días de 2018. Para cada fecha se construyeron las combinaciones correspondientes a las diez tiendas y cincuenta productos.

Por tanto, el número esperado de observaciones futuras corresponde a:

$$365 \times 10 \times 50 = 182.500$$

Cada observación representa una combinación de:

- fecha;
- tienda;
- producto;
- ventas pronosticadas.

Adicionalmente, se generaron las mismas variables de calendario utilizadas durante el entrenamiento: año, mes, día, día de la semana, semana del año y trimestre.

Característica	Valor
Año pronosticado	2018
Días considerados	365
Tiendas	10
Productos	50
Combinaciones diarias	500
Predicciones totales	182.500

Tabla 31 - Estructura del horizonte de pronóstico

4.6.2. Generación recursiva de las predicciones

Para cada día del horizonte futuro se recuperó la información histórica disponible para cada combinación tienda-producto.

Con esta información se calcularon nuevamente las características:

- lag_1;
- lag_7;
- lag_14;
- lag_28;
- rolling_mean_7;
- rolling_mean_14;
- rolling_mean_28;
- doce características EWMA;
- variables de calendario.

Una vez construidas las 27 características requeridas por LightGBM, se obtuvo la predicción correspondiente a la fecha analizada.

Posteriormente, la estimación generada fue incorporada al historial utilizado para calcular las variables del siguiente período. Este procedimiento se repitió sucesivamente hasta completar el último día de 2018.

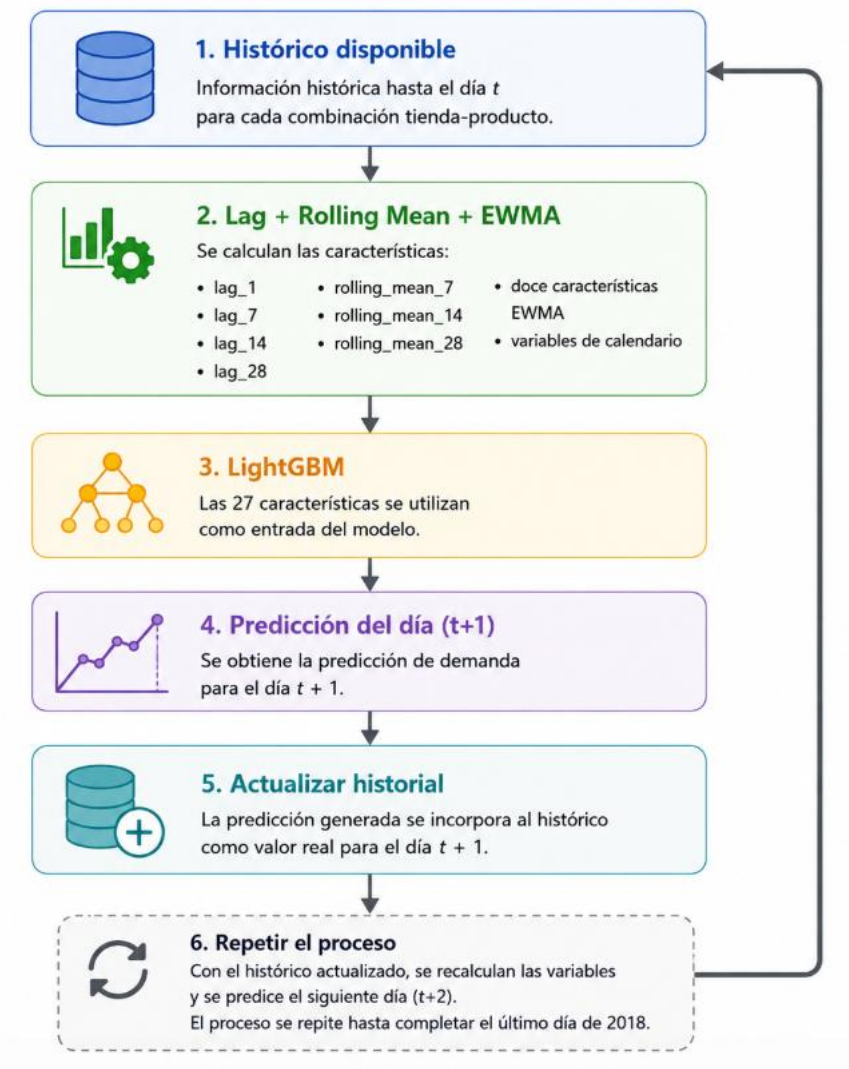


Ilustración 19 - Flujo del procedimiento de pronóstico recursivo utilizado para 2018

4.6.3. Consolidación del pronóstico

Una vez ejecutado el procedimiento para todo el horizonte, se consolidaron las predicciones en un único conjunto de resultados.

La matriz final contiene:

- 182.500 observaciones;

- cuatro variables principales;
- primera fecha: 1 de enero de 2018;
- última fecha: 31 de diciembre de 2018;
- 0 valores nulos.

```
Dimensiones: (182500, 4)
Primera fecha: 2018-01-01 00:00:00
Última fecha: 2018-12-31 00:00:00
Valores nulos: 0
```

Ilustración 20 - Verificación del conjunto consolidado de predicciones correspondientes a 2018

Elaboración propia mediante Python.

4.6.4. Volumen total de ventas pronosticado

El modelo generó un volumen anual estimado de:

11.488.304 unidades

para el año 2018.

El promedio diario agregado corresponde aproximadamente a:

31.474,81 unidades por día

Este resultado representa la suma de las estimaciones correspondientes a todas las combinaciones tienda-producto durante cada fecha.

Indicador	Resultado
Ventas totales pronosticadas	11.488.304

Promedio diario	31.474,81
Observaciones pronosticadas	182.500
Tiendas	10
Productos	50

Tabla 32 - Volumen total de ventas pronosticado

4.6.5. Pronóstico mensual

Para analizar la evolución estimada durante 2018, las predicciones individuales fueron agregadas por mes.

Los resultados muestran un comportamiento estacional similar al identificado durante el análisis histórico.

El volumen pronosticado aumenta progresivamente durante los primeros meses del año y alcanza su máximo durante julio, con:

1.242.050 unidades

Por otra parte, enero presenta el menor volumen estimado:

667.811 unidades

Posteriormente se observa una disminución gradual después de julio, acompañada de un pequeño repunte hacia noviembre.

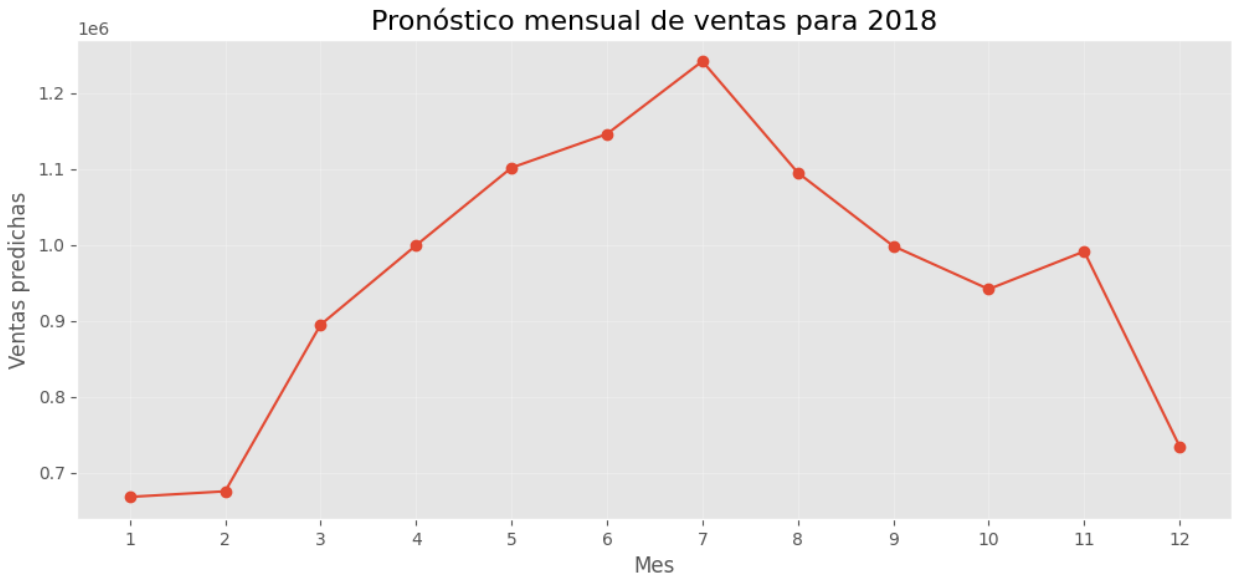


Ilustración 21 - Pronóstico mensual de ventas para el año 2018

La Ilustración 21 evidencia un comportamiento estacional definido dentro del horizonte de pronóstico. Los niveles estimados aumentan desde los primeros meses hasta alcanzar su punto máximo en julio, seguido de una reducción durante el segundo semestre.

Este patrón guarda correspondencia con el comportamiento observado durante el período histórico, donde julio también presentó uno de los mayores niveles acumulados de ventas.

La continuidad de este comportamiento dentro del pronóstico sugiere que el modelo ha incorporado información relacionada con la estacionalidad anual a través de las variables temporales y características históricas utilizadas durante el entrenamiento.

4.6.6. Pronóstico por tienda

Las predicciones fueron posteriormente agregadas según el establecimiento.

Los resultados muestran diferencias importantes en el nivel anual estimado entre las diez tiendas.

La tienda 2 presenta el mayor volumen pronosticado, alcanzando aproximadamente:

1, 46 millones de unidades

Por otra parte, la tienda 7 presenta el menor volumen estimado, con aproximadamente:

$$0,81 \text{ millones de unidades}$$

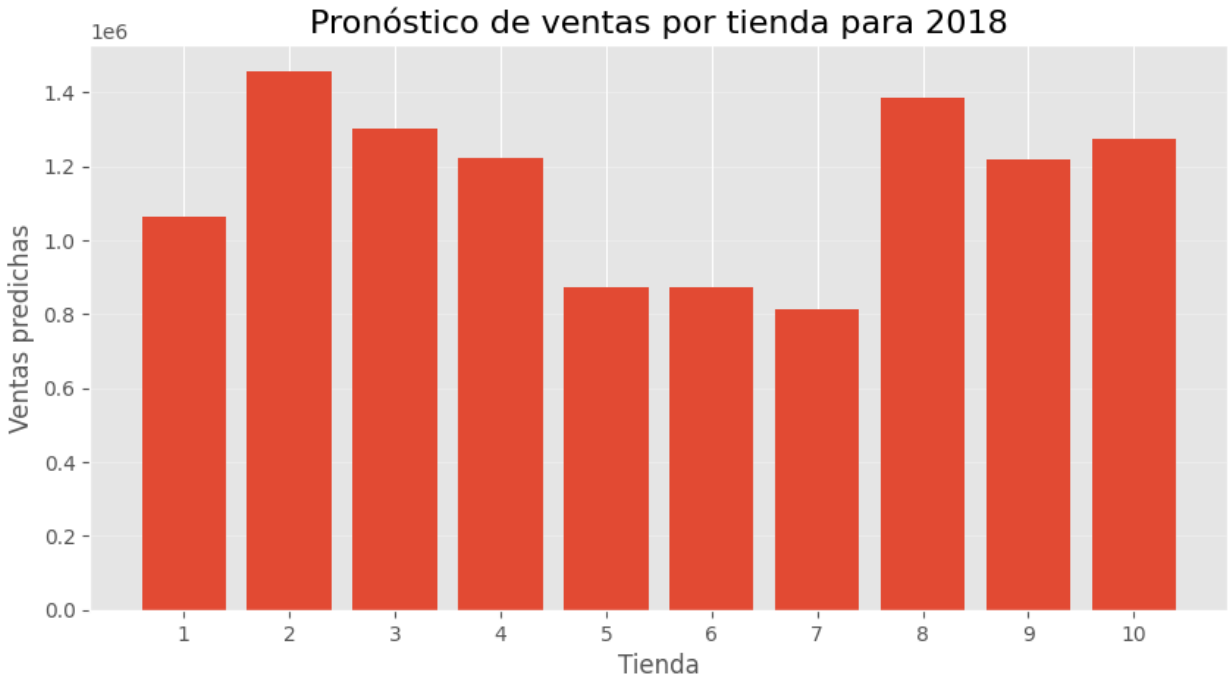


Ilustración 22 - Pronóstico de ventas por tienda 2018

La distribución estimada mantiene diferencias semejantes a las observadas durante el período histórico. En particular, la tienda 2, que presentó el mayor nivel acumulado durante los datos históricos, continúa ocupando la primera posición dentro del horizonte pronosticado.

De manera similar, los establecimientos que históricamente presentaron menores volúmenes mantienen niveles comparativamente inferiores.

Este comportamiento indica que el modelo conserva patrones estructurales asociados con cada establecimiento y no genera una distribución homogénea de la demanda entre tiendas.

Desde una perspectiva empresarial, esta información puede contribuir a priorizar recursos, inventario o capacidad operativa según el nivel de demanda esperado de cada establecimiento.

4.6.7. Pronóstico por producto

Posteriormente, las predicciones fueron agregadas por producto con el propósito de identificar aquellos artículos que concentran los mayores y menores niveles de demanda estimada.

El producto 15 registra el mayor volumen pronosticado, con aproximadamente:

$$362.016,50 \text{ unidades}$$

Los productos con mayor demanda esperada son:

- Producto 15.
- Producto 28.
- Producto 18.
- Producto 13.
- Producto 25.

	item	Ventas_predichas
14	15	362016.499847
27	28	361181.280097
17	18	352192.610343
12	13	352166.026991
24	25	343375.475549

Ilustración 23 - Productos con mayor volumen de ventas pronosticado para 2018

Elaboración propia mediante Python.

Los resultados muestran que la demanda estimada se encuentra concentrada en determinados productos. El producto 15 mantiene el liderazgo observado durante el análisis histórico y alcanza el mayor volumen estimado durante 2018.

La estabilidad relativa de los productos de mayor demanda constituye un elemento útil para procesos de planificación, debido a que permite identificar aquellos artículos que potencialmente requerirían mayores niveles de inventario o disponibilidad operativa.

4.6.8. Productos con menor demanda pronosticada

También se analizaron los productos que presentan los menores niveles de ventas esperadas.

Los cinco productos con menor demanda pronosticada corresponden a:

- Producto 5.
- Producto 4.
- Producto 1.
- Producto 41.
- Producto 47.

	item	Ventas_predichas
4	5	79180.538413
3	4	95837.383194
0	1	96137.811484
40	41	98422.479218
46	47	101438.349295

Ilustración 24 - Productos con menor volumen de ventas pronosticado para 2018

Elaboración propia mediante Python.

La identificación de productos con menor demanda estimada complementa el análisis realizado sobre los artículos de mayor volumen.

Desde una perspectiva operativa, estos resultados podrían utilizarse para revisar políticas de abastecimiento y evitar asignaciones excesivas de inventario a productos cuya demanda estimada es comparativamente reducida.

Sin embargo, debido a que el conjunto de datos no incorpora variables relacionadas con costos, niveles reales de inventario, tiempos de reposición o restricciones logísticas, estas conclusiones deben interpretarse como posibles aplicaciones del pronóstico y no como decisiones óptimas de abastecimiento determinadas directamente por el modelo.

4.6.9. Comparación entre ventas reales de 2017 y pronóstico 2018

Para contextualizar el nivel de demanda esperado, se comparó el total observado durante 2017 con el pronóstico generado para 2018.

Las ventas reales correspondientes a 2017 fueron:

$$10.733.740$$

mientras que el volumen estimado para 2018 corresponde a:

$$11.488.304$$

La diferencia absoluta es:

$$11.488.304 - 10.733.740 = 754.564$$

unidades.

En términos porcentuales:

$$\frac{11.488.304 - 10.733.740}{10.733.740} \times 100 \approx 7,03\%$$

Por tanto, el modelo proyecta un crecimiento aproximado de:

$$\boxed{7,03\%}$$

respecto al año anterior.

Período	Ventas
2017 – Real	10.733.740
2018 – Pronóstico	11.488.304
Diferencia	+754.564
Variación	+7,03 %

Tabla 33 - Comparación entre ventas reales de 2017 y pronóstico para 2018

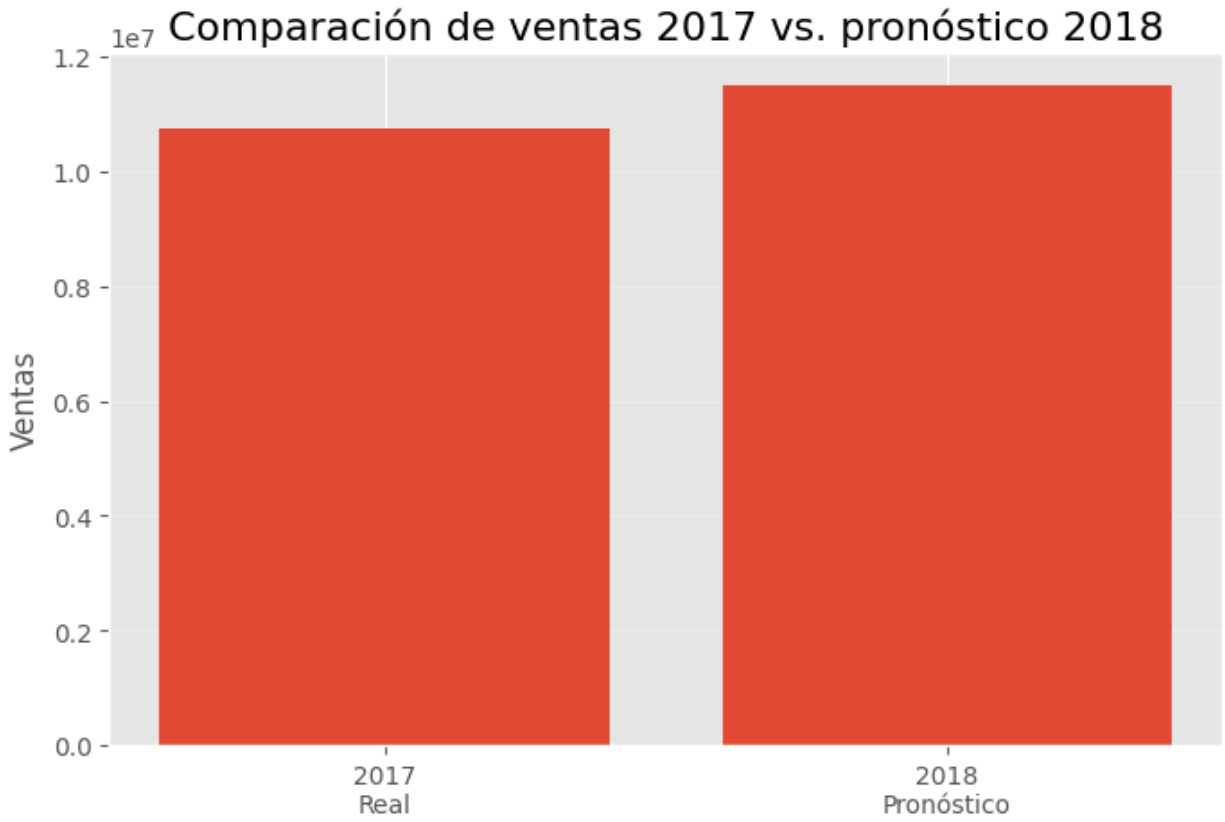


Ilustración 25 - 2017 Real vs. 2018 Pronóstico

El pronóstico indica un incremento del 7,03 % respecto al volumen registrado durante 2017. Esta evolución resulta consistente con la tendencia creciente identificada durante los cinco años históricos utilizados para desarrollar el modelo.

No obstante, la cifra correspondiente a 2018 representa una estimación generada por el modelo y no un valor observado. Adicionalmente, el procedimiento recursivo implica que las predicciones

de períodos anteriores pasan a formar parte de las características utilizadas posteriormente, por lo que la incertidumbre puede acumularse conforme aumenta el horizonte.

Por esta razón, el crecimiento estimado debe interpretarse como un escenario predictivo derivado de los patrones históricos disponibles, y no como una garantía de incremento real de las ventas.

4.6.10. Principales resultados del pronóstico

Para sintetizar los hallazgos obtenidos, se presenta la siguiente tabla:

Indicador	Resultado
Predicciones generadas	182.500
Ventas pronosticadas	11.488.304
Promedio diario	31.474,81
Mes de mayor demanda	Julio
Ventas julio	1.242.050
Mes de menor demanda	Enero
Ventas enero	667.811
Tienda con mayor demanda	Tienda 2
Tienda con menor demanda	Tienda 7
Producto con mayor demanda	Producto 15
Variación vs. 2017	+7,03 %

Tabla 34 - Resumen de resultados del pronóstico 2018

4.7. Discusión de los resultados

Los resultados obtenidos permiten analizar la utilidad de combinar técnicas de Machine Learning con ingeniería de características temporales para desarrollar modelos de forecasting sobre grandes conjuntos de datos de ventas.

El análisis exploratorio permitió identificar inicialmente que la demanda presenta una estructura temporal definida. Durante el período 2013–2017 se observó un crecimiento sostenido del volumen anual de ventas, acompañado de diferencias mensuales y semanales. Asimismo, se identificaron variaciones importantes entre tiendas y productos.

Estas características justifican el uso de un modelo capaz de integrar diferentes tipos de información. En particular, LightGBM permite considerar simultáneamente identificadores de tienda y producto, componentes de calendario y características generadas a partir de valores históricos.

La importancia de variables mostró que rolling_mean_28 y lag_28 se encuentran entre las características más utilizadas por el modelo. Este resultado evidencia la relevancia de incorporar información histórica acumulada y valores rezagados para representar el comportamiento de la demanda.

También aparecen entre las variables relevantes características como dayofweek, weekofyear, month, diferentes promedios móviles y otros rezagos. Este comportamiento mantiene coherencia con los patrones temporales identificados previamente durante el análisis exploratorio.

En términos predictivos, LightGBM alcanzó un R^2 de 0,9369, MAE de 6,0989, RMSE de 7,9252 y MAPE de 12,58 % sobre el conjunto temporal de validación correspondiente a 2017.

Estos resultados muestran que el modelo logra representar una proporción elevada de la variabilidad presente en las ventas y mantener un nivel de error moderado en relación con los valores observados.

El análisis de residuos aporta una perspectiva adicional. Aunque el promedio de los errores se encuentra cercano a cero, la dispersión aumenta conforme crece el volumen de ventas estimado. Por tanto, la capacidad predictiva no es completamente homogénea y existe mayor incertidumbre en observaciones asociadas con niveles elevados de demanda.

Por otra parte, las comparaciones agregadas por tienda, producto y mes muestran una elevada correspondencia entre ventas reales y estimadas. Sin embargo, esta precisión agregada debe interpretarse con precaución, debido a que errores positivos y negativos pueden compensarse durante la suma.

Este fenómeno explica también por qué el MAPE calculado a nivel individual alcanza 12,58 %, mientras que los errores obtenidos después de agregar las ventas por mes son considerablemente inferiores.

Finalmente, el procedimiento recursivo genera una estimación total de 11.488.304 unidades para 2018, equivalente a un incremento estimado del 7,03 % respecto a 2017.

La evolución mensual pronosticada conserva la estacionalidad identificada durante el período histórico, alcanzando su mayor volumen en julio. De manera similar, la estructura relativa por tienda y producto mantiene patrones observados previamente, destacándose la tienda 2 y el producto 15 como las unidades con mayor demanda estimada.

4.8. Aplicación empresarial de los resultados

Los resultados obtenidos pueden constituir un insumo analítico para apoyar diferentes procesos de planificación empresarial.

En primer lugar, el pronóstico de ventas permite anticipar posibles variaciones en los niveles de demanda y proporcionar una referencia cuantitativa para la planificación de inventarios.

La identificación de productos con mayores ventas estimadas puede contribuir a priorizar aquellos artículos cuya disponibilidad resulte más crítica, mientras que los productos de menor demanda pueden ser analizados con el propósito de reducir riesgos asociados con exceso de inventario.

De manera similar, el pronóstico por tienda permite reconocer establecimientos que potencialmente requerirán mayores niveles de abastecimiento o capacidad operativa.

La información mensual permite además identificar períodos donde la demanda esperada es superior, como julio, facilitando la planificación anticipada de recursos.

No obstante, la aplicación directa del pronóstico en decisiones empresariales debería complementarse con información adicional que no se encuentra disponible en el conjunto utilizado, como:

- inventario existente;
- costos de almacenamiento;

- tiempos de reposición;
- promociones;
- precios;
- restricciones logísticas;
- eventos especiales;
- condiciones macroeconómicas.

Por tanto, el modelo debe considerarse como una herramienta de apoyo a la decisión, y no como un mecanismo autónomo para determinar políticas operativas.

4.9. Limitaciones de los resultados

A pesar del desempeño obtenido, la investigación presenta diferentes limitaciones que deben considerarse durante la interpretación de los resultados.

En primer lugar, el conjunto de datos contiene exclusivamente información relacionada con fecha, tienda, producto y ventas. Por esta razón, el modelo no incorpora variables externas que podrían influir sobre la demanda, como precios, promociones, campañas comerciales, disponibilidad de inventario, feriados, competencia o condiciones macroeconómicas.

En segundo lugar, el pronóstico correspondiente a 2018 utiliza una estrategia recursiva. En este procedimiento, las estimaciones realizadas durante los primeros períodos son incorporadas posteriormente como parte del historial utilizado para generar nuevas predicciones. En consecuencia, los errores pueden propagarse y acumularse conforme aumenta el horizonte temporal.

Otra limitación se relaciona con la interpretación de la importancia de variables. La relevancia asignada por LightGBM permite identificar las características utilizadas con mayor intensidad durante la construcción del modelo, pero no establece relaciones causales entre estas variables y las ventas.

Asimismo, los resultados corresponden específicamente al conjunto de datos analizado. Por tanto, no puede asumirse que las métricas obtenidas se reproducirán automáticamente en organizaciones pertenecientes a otros sectores o con estructuras de demanda diferentes.

Finalmente, la investigación concentra la experimentación principal en LightGBM. Otros métodos de forecasting pueden presentar resultados diferentes según las características de los datos, por lo que futuras investigaciones podrían ampliar la comparación mediante modelos estadísticos, otros algoritmos de Machine Learning o arquitecturas de Deep Learning.

5. Conclusiones y Recomendaciones

5.1. Conclusiones

La presente investigación permitió desarrollar y evaluar un modelo predictivo de ventas empresariales basado en LightGBM e ingeniería de características temporales, utilizando información histórica correspondiente al período 2013–2017. A partir de los resultados obtenidos se establecen las siguientes conclusiones:

1. El comportamiento histórico de las ventas presenta patrones temporales relevantes para la predicción. El análisis exploratorio de 913.000 registros permitió identificar una tendencia general de crecimiento y diferencias sistemáticas según el mes y el día de la semana. Las ventas aumentaron de 7.941.243 unidades en 2013 a 10.733.740 unidades en 2017, confirmando que la dimensión temporal constituye un componente relevante para explicar el comportamiento de la demanda.

2. La ingeniería de características permitió representar adecuadamente la dinámica histórica de las ventas. A partir de las variables originales se construyeron características de calendario, valores rezagados (lags), promedios móviles y medias móviles exponencialmente ponderadas (EWMA), obteniéndose un conjunto final de 27 variables predictoras. Estas características permitieron incorporar información sobre el comportamiento reciente y los patrones temporales de cada combinación tienda-producto.

3. LightGBM presentó una elevada capacidad predictiva sobre datos no utilizados durante el entrenamiento. La separación temporal empleó el período 2013–2016 para entrenamiento y reservó 2017 para validación. Sobre las 182.500 observaciones de este último

período, el modelo obtuvo un MAE de 6,0989, un RMSE de 7,9252, un MAPE de 12,58 % y un R^2 de 0,9369. En conjunto, estos resultados evidencian una adecuada capacidad para reproducir el comportamiento observado de las ventas.

4. La comparación con un modelo baseline confirmó el valor predictivo adicional aportado por LightGBM. El pronóstico ingenuo estacional de siete días utilizado como línea base obtuvo un MAE de 8,8415, un RMSE de 11,5967, un MAPE de 17,94 % y un R^2 de 0,8649. LightGBM presentó menores errores y un mayor coeficiente de determinación, demostrando que la incorporación conjunta de características temporales, rezagos y estadísticas móviles mejora la predicción respecto de una estrategia sencilla basada únicamente en las ventas observadas siete días antes.

5. El modelo permitió extender el análisis hacia un escenario de pronóstico empresarial para 2018. Mediante una estrategia recursiva se generaron estimaciones para las 500 combinaciones tienda-producto durante los 365 días del horizonte futuro. El pronóstico obtenido permitió analizar la evolución esperada de la demanda a nivel temporal, por tienda y por producto, proporcionando información potencialmente útil para procesos de planificación comercial, gestión de inventarios y asignación de recursos.

6. El alcance de los resultados debe interpretarse considerando las limitaciones del estudio. El modelo fue construido exclusivamente con información histórica de fecha, tienda, producto y ventas, sin incorporar variables externas como precios, promociones, inventario, feriados o condiciones económicas. Asimismo, la experimentación principal se delimitó a LightGBM y la configuración de hiperparámetros utilizada no provino de una búsqueda automática exhaustiva. Por tanto, los resultados demuestran la viabilidad del enfoque desarrollado, pero no implican superioridad absoluta frente a otros algoritmos predictivos.

5.2. Recomendaciones

A partir de los resultados obtenidos y de las limitaciones identificadas durante el desarrollo de la investigación, se establecen las siguientes recomendaciones:

1. Incorporar variables externas relacionadas con la demanda. Futuras aplicaciones del modelo podrían incluir información sobre precios, promociones, disponibilidad de inventario,

feriados, campañas comerciales y otras variables del entorno empresarial. Su incorporación permitiría evaluar si estos factores contribuyen a mejorar la capacidad predictiva del modelo.

2. Comparar LightGBM con modelos predictivos alternativos. Se recomienda ampliar la experimentación mediante algoritmos como XGBoost, Random Forest, modelos estadísticos de series temporales o arquitecturas de Deep Learning. Esta comparación permitiría determinar si existen enfoques que mejoren el desempeño obtenido por LightGBM o si la combinación de diferentes modelos puede aportar ventajas adicionales.

3. Realizar una optimización sistemática de hiperparámetros. Debido a que la configuración utilizada en esta investigación no fue obtenida mediante una búsqueda automática exhaustiva, futuros estudios podrían aplicar técnicas como Grid Search, Random Search u optimización bayesiana, utilizando esquemas de validación que respeten la estructura temporal de los datos.

4. Implementar un proceso de actualización y monitoreo del modelo. En una aplicación empresarial, se recomienda incorporar periódicamente nuevas observaciones de ventas y evaluar la evolución de métricas como MAE, RMSE, MAPE y R^2. Esto permitiría identificar cambios en los patrones de demanda y determinar cuándo resulta necesario reentrenar o ajustar el modelo.

5. Utilizar los pronósticos como herramienta de apoyo para la planificación empresarial. Las estimaciones generadas pueden complementar procesos relacionados con planificación de inventarios, compras, distribución y asignación de recursos. Sin embargo, se recomienda utilizar los pronósticos como apoyo cuantitativo a la toma de decisiones y no como sustituto absoluto del criterio empresarial, especialmente ante eventos extraordinarios no representados en los datos históricos.

6. Evaluar estrategias alternativas para horizontes de pronóstico extensos. Debido a que el procedimiento recursivo utiliza predicciones anteriores para generar estimaciones posteriores, los errores pueden acumularse conforme aumenta el horizonte temporal. Se recomienda explorar estrategias de pronóstico directo o modelos especializados en horizontes múltiples cuando se requieran proyecciones de largo plazo.

6. Referencias

Capa Santos, H. (2017). Series temporales: La ciencia y el arte de la modelación y los pronósticos. Escuela Politécnica Nacional.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann.

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.

Kaggle. (n.d.). Store Item Demand Forecasting Challenge [Data set].

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29–40).

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Brownlee, J. (2018). Deep learning for time series forecasting: Predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.