Círculo

Descripción generada automáticamente con confianza mediaUniversidad Tecnológica de Panamá

Facultad de Ingeniería de Sistemas Computacionales

Maestría en Analítica de Datos

**Modelos Predictivos**

Juan Marcos Castillo, PhD

Por:

Aritzel Murray

**Proyecto final**

Pronóstico de Ventas mensuales en Tiendas a corto plazo para los próximos (1-2meses)

Introducción:

El pronóstico de ventas es una herramienta esencial en el sector minorista, donde la planificación precisa puede marcar la diferencia entre el éxito y el desperdicio de recursos. En este contexto, Rossmann Stores, con más de 1,110 tiendas en Europa, busca mejorar su capacidad de estimar las ventas mensuales a corto plazo, específicamente para los próximos 1-2 meses.

Factores como el número de clientes, la presencia de promociones, días festivos, y la distancia a la competencia influyen significativamente en las fluctuaciones de las ventas. Sin herramientas predictivas confiables, resulta complicado para Rossmann anticipar estas variaciones, lo que limita su capacidad para optimizar inventarios, ajustar estrategias comerciales y satisfacer la demanda de los clientes.

Este proyecto se centra en el desarrollo de un modelo predictivo basado en datos históricos de ventas y características de las tiendas. El objetivo es identificar patrones clave y proporcionar a Rossmann una solución práctica para planificar con mayor eficacia, mejorar la asignación de recursos y aumentar la competitividad en un mercado dinámico y regulado como el europeo.

Justificación

En un entorno minorista competitivo como el europeo, donde factores como días festivos, regulaciones locales y promociones afectan significativamente las ventas, la capacidad de predecir con precisión el comportamiento del mercado se convierte en una herramienta clave para la planificación estratégica.

Este análisis no solo aporta valor a Rossmann al optimizar sus operaciones, sino que también puede servir como referencia para el desarrollo de herramientas predictivas en la industria minorista en general.

Antecedentes

El sector retail se enfrenta a retos como fluctuaciones en la demanda, impacto de días festivos, competencia cercana y regulaciones locales. Estos factores afectan directamente las ventas y requieren herramientas avanzadas que permitan predecir variaciones en el comportamiento del consumidor.

En Europa, las restricciones en horarios de operación, como el cierre de tiendas los domingos, influyen en los patrones de compra y crean nichos específicos para ciertos tipos de tiendas.

El dataset Rossmann Stores Sales proporciona información detallada y estructurada, lo que lo convierte en una base ideal para el análisis predictivo.

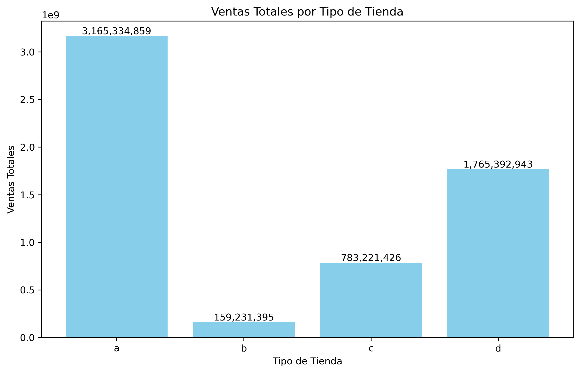
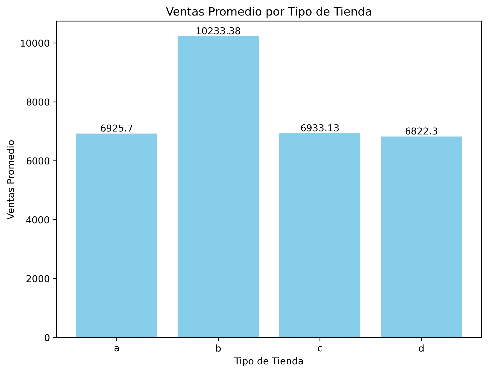
Incluye:

1. Promociones: Impacto en el volumen de ventas y el comportamiento del cliente.
2. Competencia: Distancia a la tienda competidora más cercana.
3. Estacionalidad y días festivos: Patrones de ventas influenciados por fechas especiales y regulaciones locales.

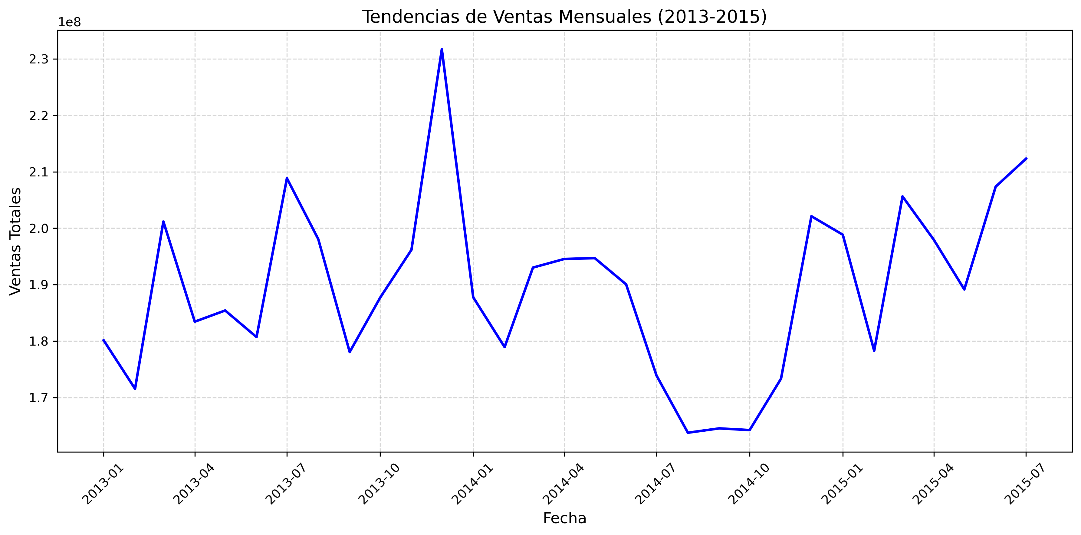
Rossmann Store opera una variedad de tipos de tiendas (a, b, c y d). cada una con características únicas en términos de rendimiento promedio.

A continuación, se presentan las visualizaciones clave obtenidas tras realizar el análisis exploratorio de los datos:

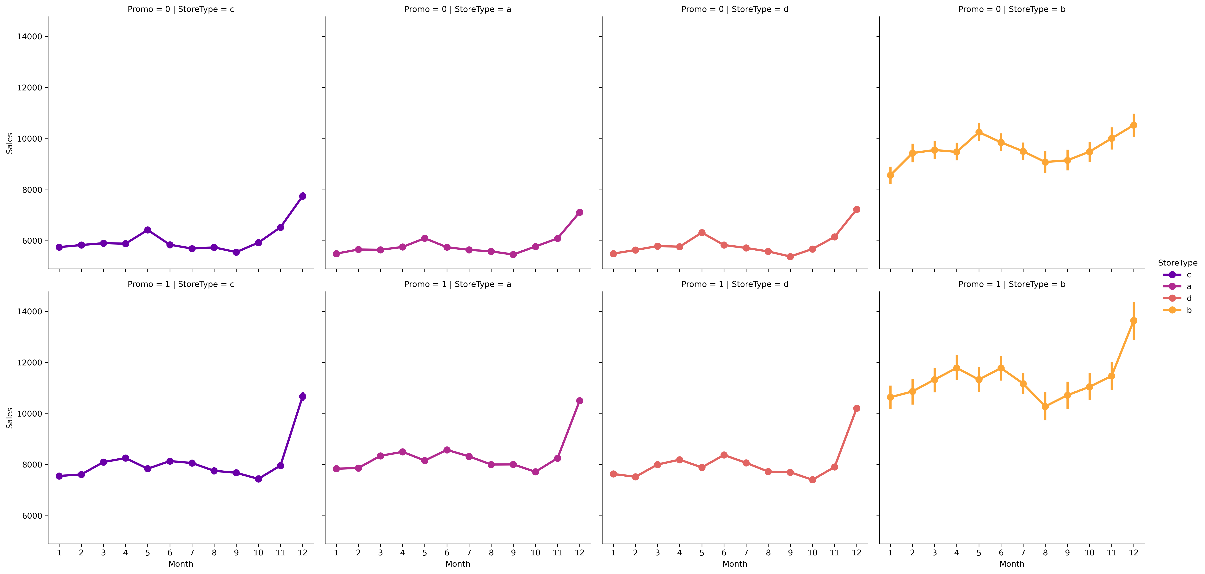
* **Ventas totales y promedios por tipo de tienda**, en donde se puede observar los tipos de tienda “a” y “d” son los más numerosos y representan más del 80% de los ingresos. Por otro lado, las tiendas tipo “b” aunque no es tan numeroso, genera mayores ventas promedio por clientes, lo que podría atribuirse a su operación continua los domingos y la facilidad de acceso para los consumidores.



* Tendencias de ventas mensuales a lo largo de los 3 años, en donde se observa como las ventas fluctúan según el mes y año, destacándose picos importantes en meses específicos como diciembre, mientras que los meses de enero tienden a mostrar una disminución en las ventas.

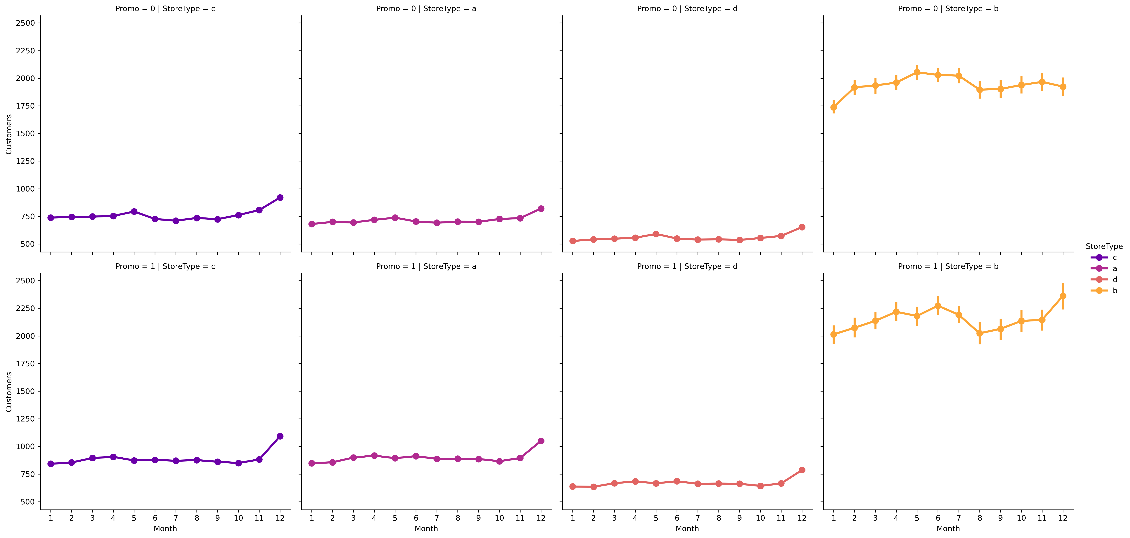


* Tendencias de Ventas Mensuales: por Promoción y Tipo de Tiendas, se presenta la evolución de las ventas mensuales para cada tipo de tienda, diferenciadas por el estado de las promociones. (Promo = 0 y Promo = 1). Donde se observa que en todos los tipos de tiendas las ventas en el mes diciembre son altas y probablemente impulsadas por la temporada navideña. Resaltando la importancia de las festividades en el comportamiento de los consumidores.

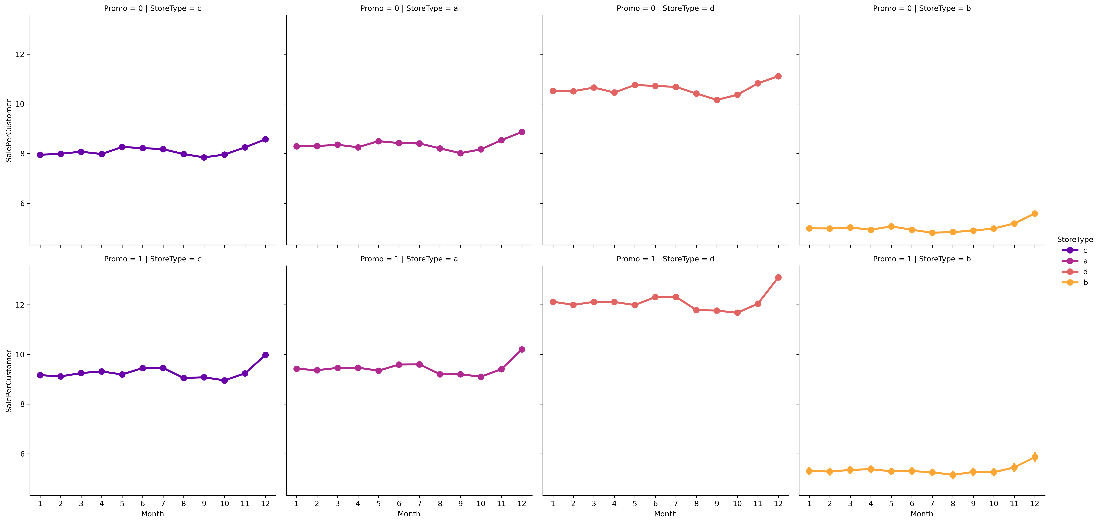


En las tiendas tipo b sobresalen con las ventas significativamente mas altas. Incluso en periodos sin promociones. Esto está relacionado con características únicas de estas tiendas como su horario de operación. El resto muestran patrones de ventas similares.

* Tendencias de Ventas Mensuales: Análisis por Cantidad Clientes, Promoción y Tipo de Tiendas, en continuación con la evolución de las ventas mensuales se observa patrones clave en los hábitos de consumo y como las promociones afectan el comportamiento de compra, para las tiendas b destacan por mantener niveles consistentes de ventas altas por clientes y para el resto de las tiendas muestran comportamientos más uniformes, pero incrementan sus ventas por clientes durante periodos promocionales.



* Tendencias de Ventas Mensuales: Análisis ventas por clientes, Promoción y Tipo de Tiendas, se observa como las ventas promedias por cliente varían mensualmente según el tipo de tiendas y estados promocionales. Y se evidencia que las tiendas tipo b, a pesar de que generan altos niveles de ventas promedios, tienden a registrar valores más bajos en las ventas por cliente a comparación con otros tipos de tiendas. Esto sugiere que los clientes en las tiendas tipo b suelen adquirir artículos de menor valor unitario pero en mayor cantidades, lo que contrasta con los otros tipos de tiendas donde los clientes tienden a realizar compras de mayor valor promedio.



Estos hallazgos resaltan la importancia de variables claves como customers y promo y el tipo de tiendas para entender las dinámicas de negocios, proporcionando insights valiosos para la toma de decisiones estratégicas.

Definición del problema

Actualmente, Rossmann Stores opera en una red de más de 1,100 tiendas, incluidas en el dataset utilizado para este proyecto. La empresa enfrenta el desafío de prever las ventas mensuales con precisión.

Carece de una herramientas predictivas que le permitan anticipar la demanda en el corto plazo (1-2 meses), dificultando la planificación de inventarios, asignación de personal y diseño de estrategias promocionales.

Las ventas están influenciadas por diversos factores, entre ellos:

* Promociones activas: Impacto directo en el volumen de ventas y atracción de clientes.
* Estacionalidad: Fluctuaciones relacionadas con periodos clave, como festividades y patrones de consumo.
* Comportamiento de los clientes: Variaciones en el número de clientes según el tipo de tienda y el día de la semana.

Este proyecto busca desarrollar un modelo de pronóstico que permita:

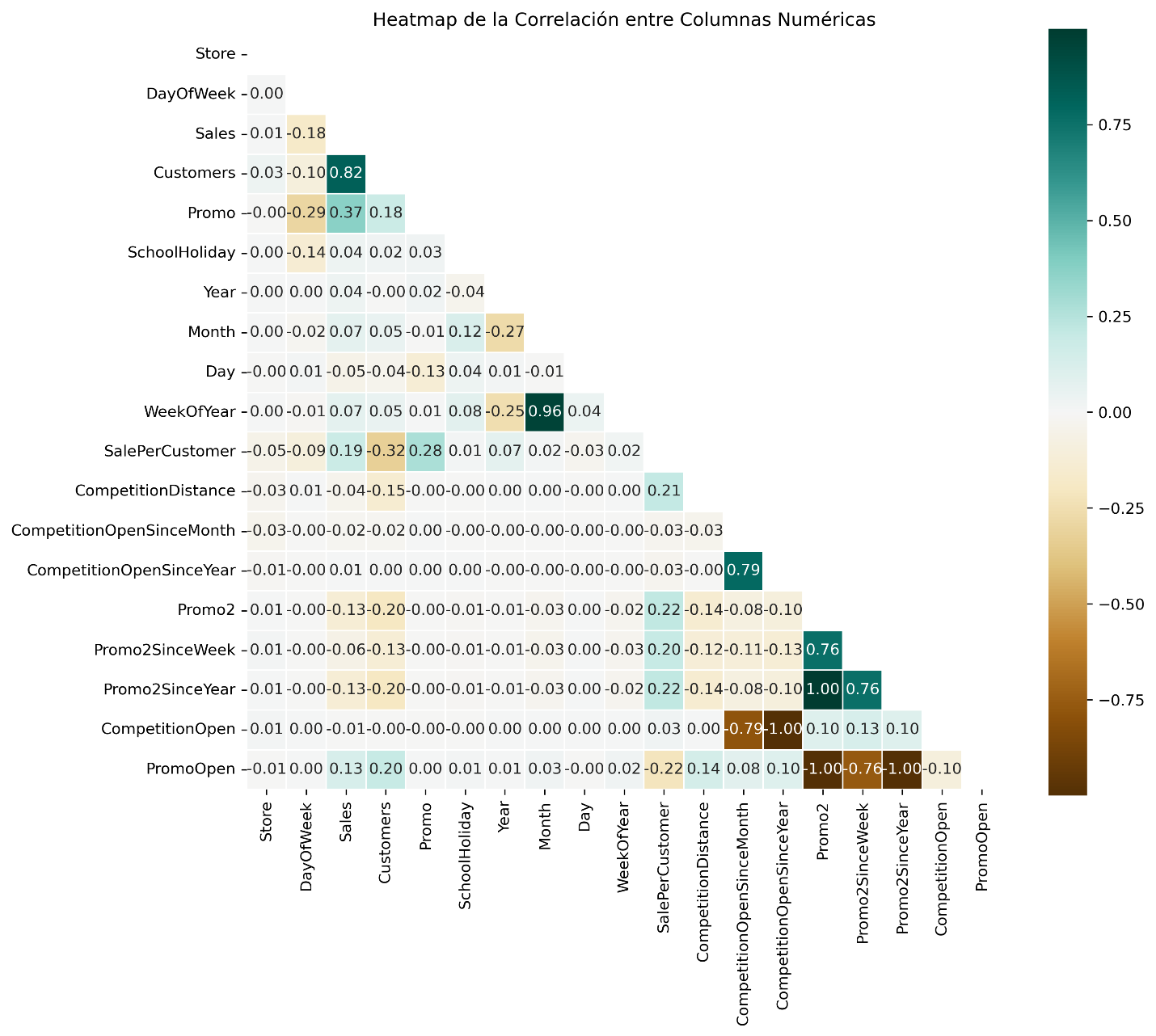
* Estimar las ventas mensuales en el corto plazo con datos de 1,115 tiendas.
* Identificar los factores clave que influyen en las ventas, como las promociones y el comportamiento de los clientes.

Este análisis busca llenar el vacío actual al proporcionar una herramienta práctica y accionable que permita a Rossmann Stores adelantarse a las variaciones de la demanda y mantener su ventaja competitiva en el mercado.

Avance del Análisis Predictivo

Antes de profundizar en los modelos predictivos, se realizó un análisis de correlación para identificar las variables más influyentes en las ventas (Sales). Este análisis permitió priorizar las variables clave y entender mejor las relaciones entre los datos, estableciendo una base sólida para la construcción de los modelos.

Variables como Customers y Promo mostraron correlaciones significativas con las ventas, destacando su relevancia en la predicción del comportamiento del mercado.



Dado que las variables con mayor correlación con la variable objetivo (Sales) son Customers y Promo, se tomó la decisión de utilizar modelos de regresión para explorar y cuantificar su impacto. Esta correlación significativa refuerza la relevancia de estas variables como predictores claves, justificando el uso de:

* Regresión Lineal Simple: Para evaluar la relación directa entre Customers y Sales.
* Regresión Múltiple: Para capturar el efecto combinado de Customers y Promo sobre las ventas.

Resultados:

Regresión Lineal Simple:

* R²: 0.68 – El modelo explica el 68% de la variabilidad en las ventas.
* RMSE: 1,758.59 – Error promedio en las predicciones de ventas.
* MAE: 1,271.24 – Desviación promedio en las estimaciones.

Interpretación:

Aunque este modelo es útil para observar la relación entre los clientes y las ventas, los resultados indican que otros factores también influyen significativamente en las ventas, sugiriendo la necesidad de un modelo más complejo.

Regresión Múltiple:

* R²: 0.73 – El modelo explica el 72.8% de la variabilidad en las ventas, superando al modelo lineal simple.
* RMSE: 1620.14 – Mejora en la precisión del modelo, con un error reducido.
* MAE: 1155.38 – Predicciones más cercanas al valor real.

Interpretación:

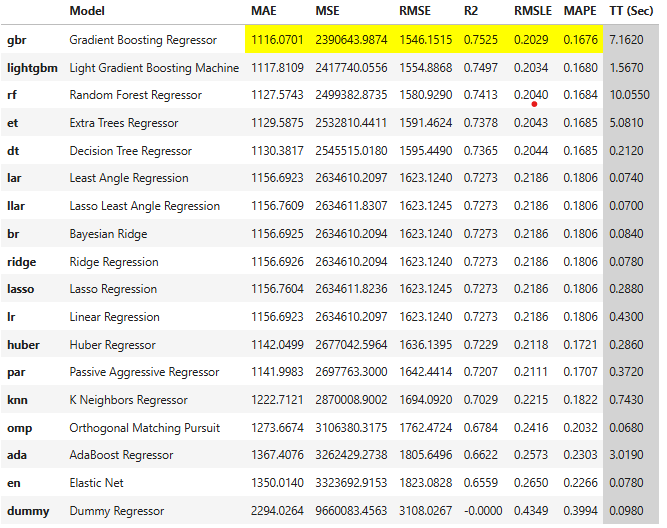
La incorporación de Promo como variable adicional muestra cómo las promociones influyen directamente en las ventas. Este modelo captura de manera más efectiva las dinámicas del negocio, destacando la importancia de considerar múltiples variables.

Dada la precisión relativamente baja obtenida con los modelos de regresión lineal y múltiple, se decidió utilizar la librería pycaret como una herramienta eficiente para la selección de modelos. PyCaret permite probar múltiples algoritmos de machine learning de manera rápida, evaluando métricas clave como R², RMSE y MAE en pocos pasos.

Esta herramienta fue fundamental para realizar un análisis comparativo entre varios modelos estándar, como regresiones lineales y múltiples, Random Forest y Gradient Boosting. PyCaret proporcionó una visión inicial de cuáles algoritmos tenían mayor potencial para capturar la complejidad de las ventas en las tiendas Rossmann.

Resultados de librería Pycaret, utilizando las variables independientes Customers y Promo para predecir la variable objetivo Sales.

Donde se puede observar Gradient Boosting Regressor con un R2 en 0.7525



Tras identificar los mejores algoritmos en términos de precisión y estabilidad, se decidió implementar **XGBRegressor** debido a su capacidad para manejar relaciones no lineales, mejorar el ajuste de los datos y trabajar eficientemente con variables adicionales relevantes como: DayOfWeek, StoreType, Assortment, StateHoliday, CompetitionDistance.

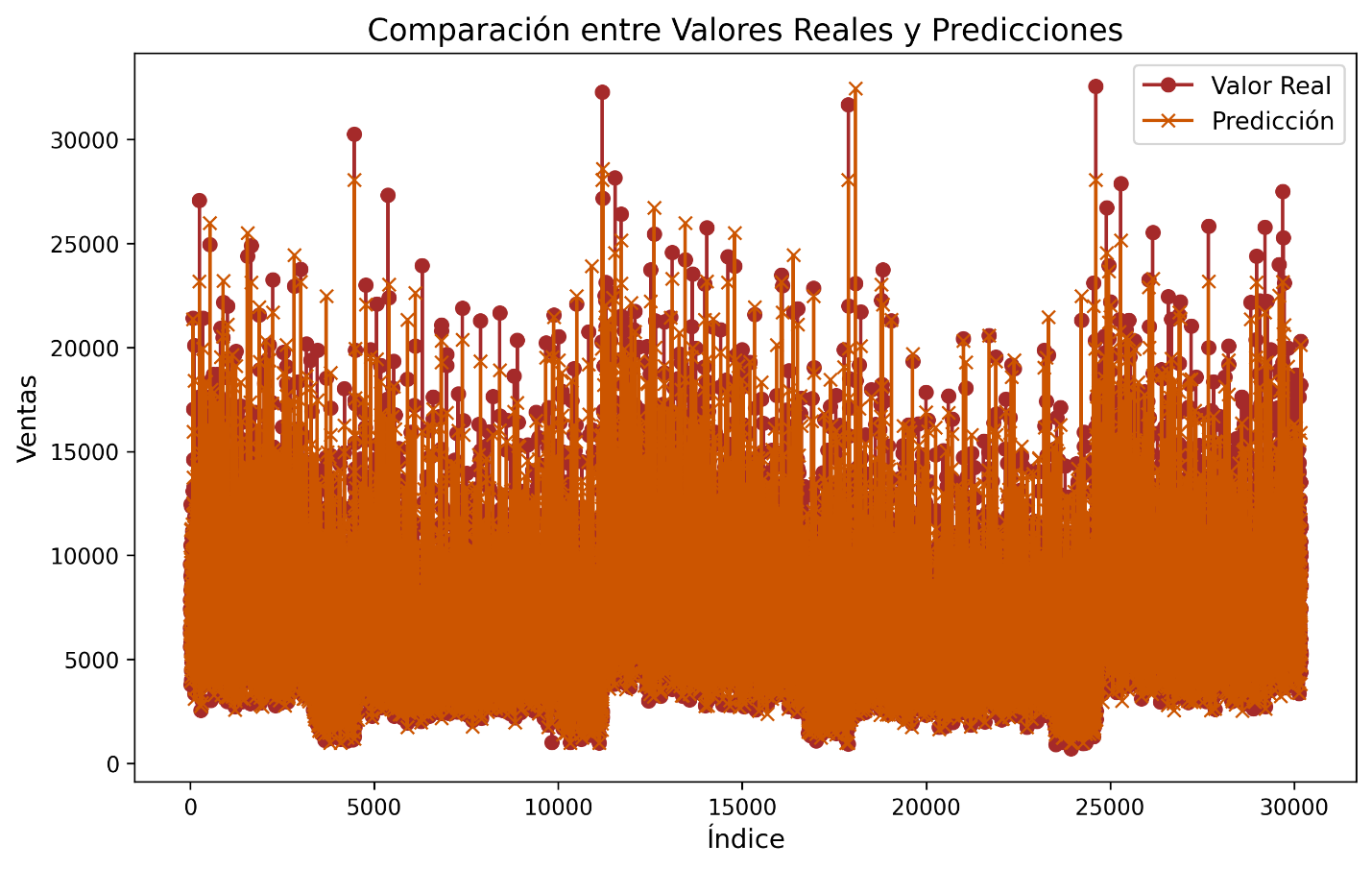
Para maximizar el rendimiento del modelo **XGBRegressor**, se realizó un ajuste de hiperparámetros utilizando **GridSearchCV**. Esta técnica permitió explorar de manera sistemática diferentes combinaciones de valores para parámetros clave, como el número de estimadores (n\_estimators), la tasa de aprendizaje (learning\_rate), y la profundidad máxima de los árboles (max\_depth).

Hiperparámetros óptimos seleccionados:

* {'learning\_rate': 0.2, 'max\_depth': 8, 'n\_estimators': 200}

Con los hiperparámetros optimizados, el modelo fue validado en conjuntos de prueba con datos no entrenados, utilizando el mes de julio 2015, confirmando su capacidad de generalización para nuevos escenarios.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Conjunto | R2 | RMSE | MAE | MAPE |
| Prueba | 0.9546 | 661.92 | 455.33 | 6.77% |
| Entrenamiento | 0.9580 | 637.37 | 441.77 | 6.60% |
| Julio 2015 | 0.9374 | 734.70 | 527.63 | 7.40% |



**Línea de Tiempo de la Investigación**

|  |  |
| --- | --- |
| **22 - 23 marzo** | búsqueda y revisión del dataset a utilizar. |
| **25 - 26 marzo** | Exploración del dataset, limpieza de datos y manejo de valores nulos. |
| **28 - 30 marzo** | Análisis descriptivo: visualización de tendencias, estacionalidad y factores que afectan las ventas. |
| **1 - 2 Abril** | Desarrollo del análisis predictivo utilizando regresión lineal/múltiple. Y gráficos de visualización. |
| **3 - 5 Abril** | Desarrollo análisis predictivo utilizando librería pycaret, para mejorar el modelo. Se utiliza XBGRegressor como algoritmo por la capacidad de predecir mejor. |
| **6 - 9 Abril** | Realiza Validaciones del modelo con datos que no se entrenó para ver su forma de generalizar. Y se crea los entregables. |