Informe modelo predictivo y evaluación

Autor: Kristhian Santiago Palomino Fajardo
Fecha: 18 de marzo de 2025

Introducción

El objetivo es desarrollar un modelo predictivo capaz de estimar el número de unidades vendidas (Units_Sold) en función de datos históricos de ventas. Con base en un conjunto de datos que incluye información sobre la fecha, tienda, categoría, unidades vendidas y precio unitario, se ha realizado un análisis previo (EDA) para comprender la estructura y comportamiento de los datos. En esta fase, se ha procedido a la construcción, tuning y evaluación de diversos modelos de regresión.

Algoritmos Utilizados

Se implementaron y evaluaron siete algoritmos distintos:

- **Regresión Lineal:**
  Modelo base que asume una relación lineal entre las variables predictoras y el objetivo.
- **Lasso:**
  Regresión lineal con regularización L1, que puede conducir a la selección de variables al forzar coeficientes a cero.
- **Ridge:**
  Regresión lineal con regularización L2, que ayuda a mitigar la multicolinealidad y a estabilizar la estimación de los coeficientes.
- **Decision Tree:**
  Un árbol de decisión que segmenta el espacio de predicción en función de umbrales en las variables predictoras.
- **Random Forest:**
  Un conjunto de árboles de decisión que mejora la capacidad de generalización mediante el promedio de múltiples árboles entrenados con sub-muestras aleatorias.
- **Gradient Boosting:**
  Modelo ensemble que construye árboles de decisión secuencialmente, corrigiendo errores del modelo anterior, lo que permite capturar relaciones no lineales complejas.
- **AdaBoost:**
  Otro método ensemble que ajusta pesos a las instancias mal predichas para focalizar el aprendizaje en aquellas, resultando en un modelo robusto ante outliers.

Proceso de Selección de Características y Tuning de Hiperparámetros

Ingeniería de Características

- Conversión de la variable Date a formato datetime y extracción de características temporales: **Mes**, **Día_de_la_semana**, **Trimestre** y **Day_of_Month**.
- Creación de la variable **Es_fin_de_semana** para distinguir entre días laborables y fines de semana.
- Se utilizó One-Hot Encoding para las variables categóricas, permitiendo representar cada categoría sin asumir un orden implícito.
- La normalización de las variables numéricas usando StandardScaler se aplicó para garantizar que todas las características se encuentren en la misma escala, facilitando el entrenamiento de modelos lineales y mejorando la convergencia de algoritmos basados en distancia..

Tuning de Hiperparámetros

Se definieron grids de hiperparámetros para cada modelo y se aplicó GridSearchCV con validación cruzada (cv=3) para optimizar los parámetros. A continuación, se resumen los mejores parámetros obtenidos:

- **Linear Regression:**
  Best Params: {} (sin hiperparámetros a ajustar).
- **Lasso:**
  Best Params: {'regressor__alpha': 0.1}.
- **Ridge:**
  Best Params: {'regressor__alpha': 1}.
- **Decision Tree:**
  Best Params: {'regressor__max_depth': 10, 'regressor__min_samples_split': 2}.
- **Random Forest:**
  Best Params: {'regressor__max_depth': 10, 'regressor__min_samples_split': 2, 'regressor__n_estimators': 50}.
- **Gradient Boosting:**
  Best Params: {'regressor__learning_rate': 0.1, 'regressor__max_depth': 3, 'regressor__n_estimators': 100}.
- **AdaBoost:**
  Best Params: {'regressor__learning_rate': 1, 'regressor__n_estimators': 100}.

Evaluación de Modelos

- **Mean Absolute Error (MAE):**
  Mide el error medio en unidades, interpretado directamente en el contexto de unidades vendidas.
- **Root Mean Squared Error:**
  Penaliza más fuertemente los errores grandes, proporcionando una medida robusta del error.
- **Coeficiente de Determinación (R^2):**
  Indica la proporción de varianza en la variable objetivo explicada por el modelo.

Los resultados obtenidos son los siguientes:

Modelo	MAE	RMSE	R^2
Linear Regression	4.90	5.77	0.80
Lasso	4.89	5.83	0.80
Ridge	4.79	5.73	0.81
Decision Tree	9.20	10.70	0.32
Random Forest	5.92	7.27	0.69
Gradient Boosting	6.65	8.23	0.60
AdaBoost	4.77	6.56	0.74

Interpretación de Resultados

- Modelos Lineales (Linear Regression, Lasso, Ridge):
  Estos modelos muestran resultados consistentes con MAE y RMSE cercanos y un R^2 alrededor de 0.80, lo que indica que explican aproximadamente el 80% de la varianza en las ventas. Entre ellos, Ridge mostró un desempeño ligeramente superior.
- Modelos Ensemble (Random Forest, Gradient Boosting, AdaBoost):
  El Random Forest y Gradient Boosting no lograron superar el desempeño de los modelos lineales en términos de R^2, posiblemente debido a la complejidad del dataset y al hecho de que cada tienda vende una única categoría, lo que reduce la variabilidad.
  AdaBoost obtuvo un MAE ligeramente menor que los modelos lineales, aunque su RMSE y R^2 son un poco inferiores, lo que sugiere que, aunque tiene buen desempeño en promedio, puede tener algunas predicciones con errores más altos.

- Decision Tree:
  El modelo de árbol de decisión mostró un desempeño significativamente inferior, con un R^2 de 0.32, lo cual indica que no es capaz de capturar adecuadamente la complejidad de los datos, probablemente por falta de robustez.

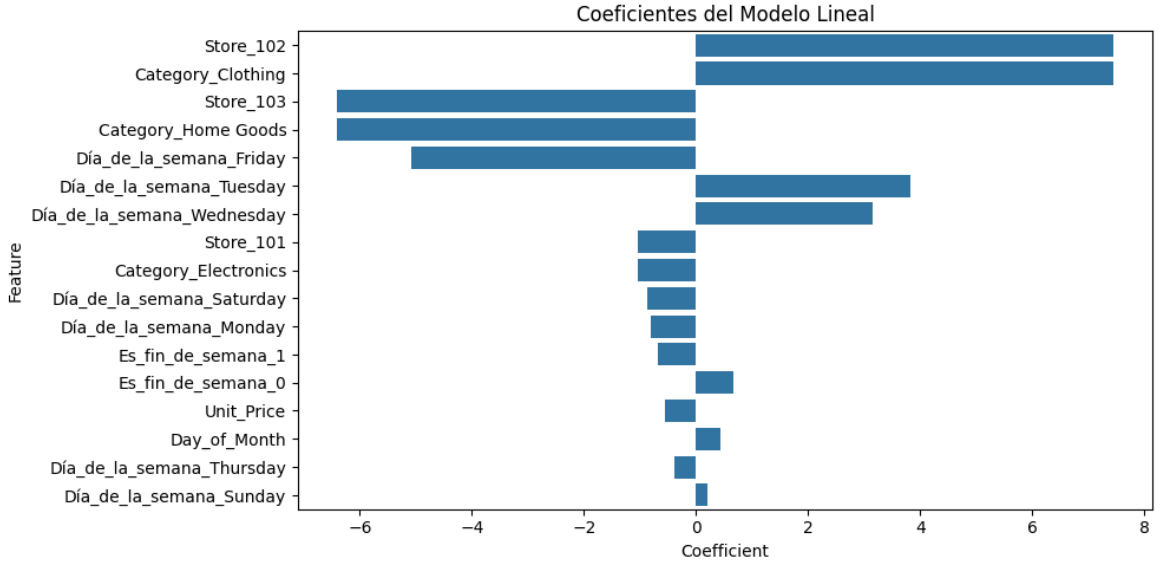
Justificación de la Elección del Modelo Final

Tras comparar los resultados, se observa que los modelos lineales (especialmente Ridge) presentan un buen equilibrio entre simplicidad, interpretabilidad y desempeño, con un R^2 del 81%. Estos modelos son especialmente adecuados cuando:

- La relación entre las variables predictoras y el objetivo es casi lineal.
- Se desea una fácil interpretación de los coeficientes, lo cual permite extraer información sobre la importancia de cada variable.

Aunque AdaBoost mostró un MAE ligeramente inferior, la estabilidad y la interpretabilidad de un modelo lineal son aspectos clave para la toma de decisiones estratégicas en este contexto. Por ello, se recomienda considerar Ridge Regression como modelo final, aunque se mantengan los otros modelos para futuras comparaciones si se dispone de más datos o se requiere explorar relaciones no lineales complejas.

Coeficientes de Modelos Lineales



Ridge Regression:
Los coeficientes sugieren que:

  - Las variables relacionadas con Store y Category tienen una alta importancia; por ejemplo, la dummy correspondiente a Store_102 y Category_Clothing

muestran coeficientes positivos, lo que indica que estas variables impulsan las ventas.

- Los coeficientes negativos asociados a Store_103 y Category_Home Goods indican una relación inversa, consistente con la diferencia en rangos de precios y comportamiento de ventas.
- Las variables derivadas de Día_de_la_semana muestran variaciones que ayudan a capturar el efecto temporal, destacando la importancia de los días centrales de la semana.

Estos hallazgos permiten entender cómo cada variable influye en la predicción y son valiosos para definir estrategias de negocio.

Identificación de Posibles Mejoras en la Precisión del Modelo

Mejoras en la precisión del modelo:

- **Ampliación del Historial de Datos:**
  Actualmente el modelo se ha desarrollado a partir de datos correspondientes a un único mes. Recopilar y analizar datos a lo largo de varios meses o años permitirá capturar patrones estacionales, tendencias de largo plazo y variaciones puntuales (por ejemplo, cambios en el comportamiento del consumidor durante períodos promocionales o festivos). Esto, a su vez, facilitará el entrenamiento de modelos más robustos y precisos.
- **Incorporación de Variables Externas:**
  La precisión del modelo podría mejorar significativamente si se incluyen variables contextuales, como promociones, días festivos, eventos locales o indicadores macroeconómicos. Estas variables adicionales ayudarían a explicar variaciones en las ventas que el modelo actual no captura, y permitirían que el modelo se ajuste a cambios en el entorno comercial.

Recomendaciones Prácticas para el Equipo de Ventas:

- **Enfoque Diferenciado por Categoría:**
  El análisis de coeficientes ha evidenciado que cada tienda se especializa en una única categoría, lo que implica que la estrategia de ventas debe adaptarse a las características propias de cada segmento:
  - Para Electronics (Tienda 101), dado su alto precio unitario y el considerable impacto en ingresos, se recomienda asegurar un stock suficiente y ejecutar promociones premium que refuercen la imagen de alta calidad y exclusividad.

    - Para Clothing (Tienda 102), con un elevado volumen de ventas pero menores ingresos totales, es conveniente implementar estrategias de cross-selling y bundle, además de optimizar promociones que incentiven el aumento del ticket promedio.
    - Para Home Goods (Tienda 103), la baja contribución a los ingresos sugiere la necesidad de potenciar esta categoría mediante campañas de marketing focalizadas o diversificación de la oferta, especialmente si se identifican nichos de mercado poco atendidos.

- **Optimización de Estrategias Diarias:**
  Dado que los días centrales de la semana (martes y miércoles) presentan mayores promedios de ventas, se pueden diseñar campañas específicas para esos días, así como estrategias para potenciar ventas en días con menor actividad (como viernes), ya sea mediante ofertas especiales o mejoras en la experiencia de compra.
- **Monitoreo y Retroalimentación Continua:**
  Se sugiere implementar un sistema de seguimiento del desempeño del modelo en producción y combinarlo con la retroalimentación del equipo de ventas. Esto permitirá ajustar rápidamente las estrategias y mejorar continuamente el modelo en función de nuevos datos y cambios en el comportamiento del mercado.

Limitaciones del Análisis Actual:

- **Datos Limitados en el Tiempo:**
  El análisis se realizó con datos de un único mes, lo que puede limitar la detección de patrones estacionales y tendencias a largo plazo. La variabilidad en las ventas podría estar influenciada por factores que no se reflejan en un corto período.
- **Falta de Variables Externas:**
  La ausencia de información sobre promociones, eventos locales, condiciones económicas o cambios en el comportamiento del consumidor restringe la capacidad del modelo para capturar todas las fuentes de variabilidad en las ventas.
- **Relación One-to-One entre Tienda y Categoría:**
  La asignación única de categorías a cada tienda simplifica el análisis, pero también puede limitar la generalización del modelo si en el futuro se incorporan tiendas que manejen múltiples categorías o se modifique la estrategia comercial.

Sugerencias para Futuros Trabajos:

- **Ampliar el Historial de Datos:**
  Recopilar y analizar datos de períodos más largos permitirá identificar patrones estacionales y tendencias más robustas. Esto facilitará la actualización del modelo

con datos históricos ampliados y ayudará a capturar cambios en el comportamiento de los consumidores.

- **Incorporar Variables Externas y Contextuales:**
  Considerar la integración de variables que reflejen promociones, eventos especiales, festividades y datos macroeconómicos. Esto puede enriquecer el análisis y mejorar la capacidad predictiva del modelo.
- **Implementación de un Sistema de Monitoreo en Producción:**
  Desarrollar un dashboard o sistema de monitoreo que permita evaluar en tiempo real el desempeño del modelo y facilitar la retroalimentación para ajustes y mejoras continuas.
- **Validación Cruzada y Evaluación Periódica:**
  Establecer procedimientos regulares de validación y actualización del modelo para asegurar que este se mantenga robusto y relevante frente a cambios en el entorno del mercado.

Conclusiones

Los modelos lineales, en particular Ridge Regression, han demostrado un buen desempeño, explicando aproximadamente el 81% de la varianza en Units_Sold.

La capacidad de interpretar los coeficientes es un valor añadido que facilita la toma de decisiones estratégicas, permitiendo identificar qué variables (tienda, categoría, día de la semana) impactan de forma positiva o negativa en las ventas.

Aunque se exploraron modelos ensemble, la simplicidad y robustez de los modelos lineales hacen que sean una opción sólida para este conjunto de datos, especialmente dado que cada tienda se especializa en una única categoría.