TEC | Tecnológico de Costa Rica

Escuela de Computación

Inteligencia Artificial

Profesor:

Kenneth Obando Rodriguéz

TC04 - Proyecto de Predicción de Ventas por Subcategoría

Integrantes:

Francisco Villanueva Quirós

Sebastián Cruz Guzmán

Semestre I - 2025

Fecha: 12/05/2025

Descripción General

Este proyecto corresponde al TC04, cuyo objetivo es poner en práctica el uso de herramientas computacionales para la ingesta, limpieza, transformación y exportación de datos. Siguiendo la metodología CRISP-DM para resolver un caso de minería de datos, el trabajo se dividió en varias etapas:

1. Entendimiento del negocio
2. Entendimiento de los datos
3. Preparación de los datos
4. Modelaje de los datos
5. Evaluación y conclusiones

El proyecto se enfoca en la predicción de ventas futuras para una cadena minorista global, utilizando técnicas de machine learning. El objetivo principal es desarrollar modelos predictivos capaces de pronosticar ventas por subcategoría de producto y mes para el año 2025, proporcionando información valiosa para la planificación de inventario, gestión de la cadena de suministro y estrategias comerciales.

Objetivos del negocio

El propósito de este proyecto fue implementar un sistema de predicción de ventas. Esto con el fin de mejorar lo que podría la planificación de un inventario de la tienda o sitio de ventas, además, de toda la logística para los productos de la tienda. Para esto, se buscará predecir las ventas mensuales durante el año de 2025, divididas por sub-categorías. Esto para hacer una toma de decisiones más fundamentada ante cuáles productos son los más cotizados o serán los más cotizados.

Desde la perspectiva del negocio, el proyecto será exitoso si permite estimar con suficiente precisión las ventas futuras por subcategoría. Como métrica de éxito, se espera un error absoluto medio (MAE) razonablemente bajo al predecir las ventas mensuales, junto con una visualización clara de tendencias por tipo de producto.

En el contexto del trabajo realizado, el objetivo de la minería de datos, es que, a partir de datos históricos de ventas de una empresa, en este caso (2015-2018), se puedan estimar el valor de ventas mensuales para cada sub-categoría.

Para lograr esto, se utilizaron diferentes modelos de machine learning para pronosticar las ventas por subcategoría para el año 2025. Se implementaron y compararon cuatro modelos diferentes con los datos obtenidos de la minería de datos, de esta forma se puede observar si es factible la realización de las métricas. Los modelos elegidos fueron los siguientes:

- Regresión Lineal
- Random Forest
- Gradient Boosting
- Árbol de Decisión (Decision Tree)

Estructura del Proyecto

Para organizar el trabajo de manera clara y modular, se estructuró el proyecto de la siguiente manera:

```
TC04-IA/
├── main.py                 # Archivo principal de ejecución
├── requirements.txt        # Dependencias del proyecto
├── data/                   # Datos de entrenamiento
├── plots/                  # Gráficos generados
└── src/                    # Módulos del proyecto
    ├── dataset.py          # Gestión de datos
    ├── eda.py              # Análisis exploratorio
    ├── load_data.py        # Carga de datos
    ├── modeling.py         # Implementación de modelos
    └── preprocessing.py    # Preprocesamiento de datos
```

Esta estructura permite una separación clara de responsabilidades y facilita el mantenimiento del código. El archivo principal main.py orquesta el flujo completo del proceso, mientras que los módulos específicos en la carpeta src/implementan cada una de las etapas del análisis.

Metodología

1. **Preprocesamiento de Datos**
    a. Limpieza de datos
    b. Transformación de fechas
    c. Agregación de ventas por subcategoría y mes
    d. Codificación de variables categóricas
2. **Implementación de Modelos**
    a. Se implementaron cuatro modelos diferentes
    b. Cada modelo fue entrenado con los datos históricos
    c. Se utilizó GridSearch para optimizar hiperparámetros en Random Forest, Gradient Boosting y Decision Tree
3. **Evaluación de Modelos**
    Se utilizaron dos métricas principales:
    a. MAE (Error Absoluto Medio)
    b. RMSE (Error Cuadrático Medio)

Entendimiento de los Datos

Para llevar a cabo este proyecto, se trabajó con el dataset "Superstore Sales Dataset" de Rohit Sahoo disponible en Kaggle. Este conjunto de datos contiene un total de 9,800 registros, con información de órdenes realizadas entre los años 2015-2018.

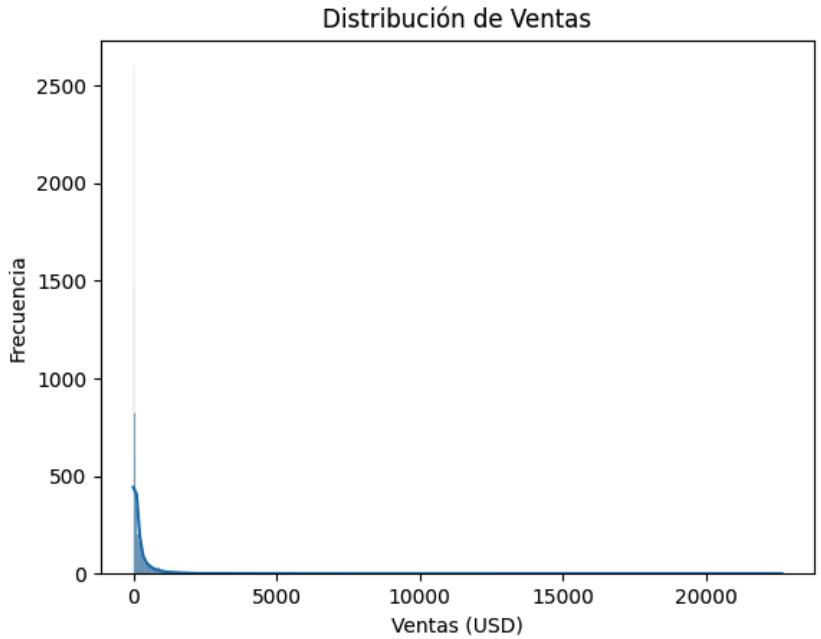
Las columnas más interesantes para nuestras predicciones incluyen:

- **Order Date, Ship Date**: Fechas de pedido y envío que permiten analizar patrones temporales
- **Product ID, Category, Sub-Category, Product Name**: Información de categorización de productos
- **Sales**: La variable objetivo que queremos predecir
- **Segment, Region, State, City**: Información geográfica y de segmentación de mercado

Para hacer dicha revisión de forma más clara, se realizaron gráficos y salidas en terminal, para comprender los datos.

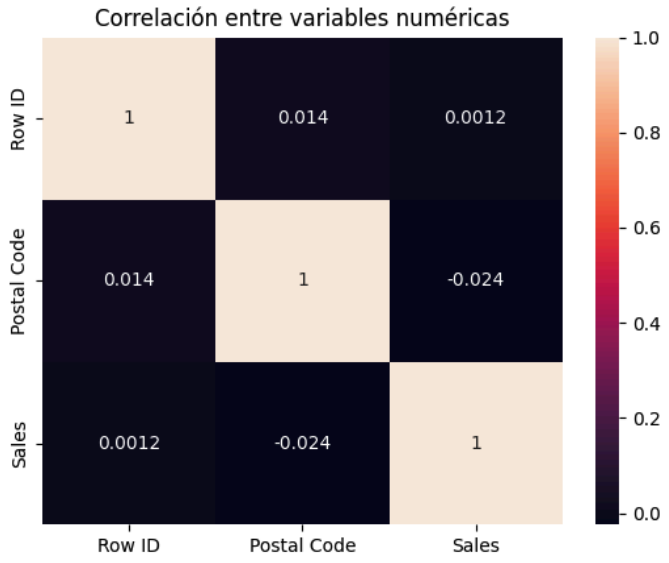
```
Descripción estadística:
            Row ID   Postal Code         Sales
count  9800.000000   9789.000000   9800.000000
mean   4900.500000  55273.322403    230.769059
std    2829.160653  32041.223413    626.651875
min       1.000000   1040.000000      0.444000
25%    2450.750000  23223.000000     17.248000
50%    4900.500000  58103.000000     54.490000
75%    7350.250000  90008.000000    210.605000
max    9800.000000  99301.000000  22638.480000

Valores nulos:
Row ID            0
Order ID          0
Order Date        0
Ship Date         0
Ship Mode         0
Customer ID       0
Customer Name     0
Segment           0
Country           0
City              0
State             0
Postal Code      11
Region            0
Product ID        0
Category          0
Sub-Category      0
Product Name      0
Sales             0
dtype: int64
```

Como se puede ver, se realiza una estadística con datos interesantes del dataset para comprender la distribución que estos siguen. Además, otra parte importante en la eliminación de columnas con valor null, y estas salidas en la terminal, nos ayudan a comprender qué columnas tienen valores null para limpiarlos.



Este es otro de los primeros gráficos realizados para visualizar la venta de productos de la tienda en el dataset. Este gráfico nos permite entender que la mayoría de las ventas que se realizan en la tienda rondan por debajo de los 3000 dólares. Lo que posteriormente, nos ayuda a entender por qué, el modelo puede fallar con las predicciones de ventas superiores a este rango.

Por último se hizo una comparación entre variables numéricas para ver si existía algún tipo de relación entre estos datos:



Como se puede observar, nos ayudó a concluir que no existen relaciones directas entre estas variables numéricas.

Preparación de los Datos

Como se mencionaba anteriormente, se hizo una limpieza de datos que primeramente eran necesarios, como las columnas y filas que tenían valores nulos, posteriormente, se eliminaron columnas que no se consideraron necesarias para las predicciones o que iban a ser utilizadas para el trabajo. Esto se realiza con la función clean_data(), que remueve las filas con datos incompletos, en este caso postal code.

Posteriormente, se hace la validación de las fechas para realizar las predicciones. Primeramente se convierte la columna Order Data a tipo datetime, así se descartan fechas inválidas. Además, se agregan columnas nuevas para el propósito de la predicción, como lo son year y month.

Por último, se hace la eliminación de columnas que no se van a utilizar. mediante la función drop_unnecesary_columns() que elimina las columnas que no se van a utilizar para fines prácticos del proyecto.

Mediante aggregate_monthly_sales(), se reestructura el dataset original para agrupar las ventas por combinación de Sub-Category, Year y Month. Esto transforma datos a nivel de transacción en un formato mucho más útil para series temporales y predicción futura.

Otra de las transformaciones de los datos, es la codificación de variables categóricas. La función encode_subcategory_column() convierte la subcategoría en una variable numérica utilizando LabelEncoder. Esto permite que los modelos puedan trabajar con ella directamente como una feature.

Por último, la función generate_2025_input() genera un DataFrame artificial con todas las combinaciones de subcategorías y meses para el año 2025. Este se utiliza como entrada para los modelos entrenados.

Modelado

En esta etapa, se implementaron y evaluaron cuatro modelos diferentes de aprendizaje automático para la predicción de ventas por subcategoría. La selección de modelos se realizó considerando diferentes enfoques, desde algoritmos simples hasta técnicas más avanzadas, con el fin de comparar su rendimiento y seleccionar el más adecuado para el caso de uso.

Para la parte del entrenamiento, se dividieron los datos en 80% de datos de entrenamiento y 20% de datos de prueba. Se utilizó una semilla aleatoria (random_ state=42). Ademas, conceptos como:

Validación cruzada:

- Se implementó validación cruzada con 5 folds para evaluar la robustez de los modelos
- Esto permitió reducir el riesgo de sobreajuste y obtener estimaciones más confiables del rendimiento

Optimización de hiperparámetros:

- Se utilizó GridSearchCV para encontrar la mejor combinación de hiperparámetros
- Para Random Forest se exploraron opciones como:

    - Número de estimadores: 50, 100
    - Profundidad máxima: 5, 10, None
    - Características máximas: "sqrt", "log2"
- Para Gradient Boosting se evaluaron:
    - Número de estimadores: 50, 100
    - Tasa de aprendizaje: 0.05, 0.1
    - Profundidad máxima: 3, 5
- Para Decision Tree se optimizaron:
    - Profundidad máxima: 3, 5, 10, None
    - Muestras mínimas para división: 2, 5, 10
    - Muestras mínimas en hoja: 1, 2, 4

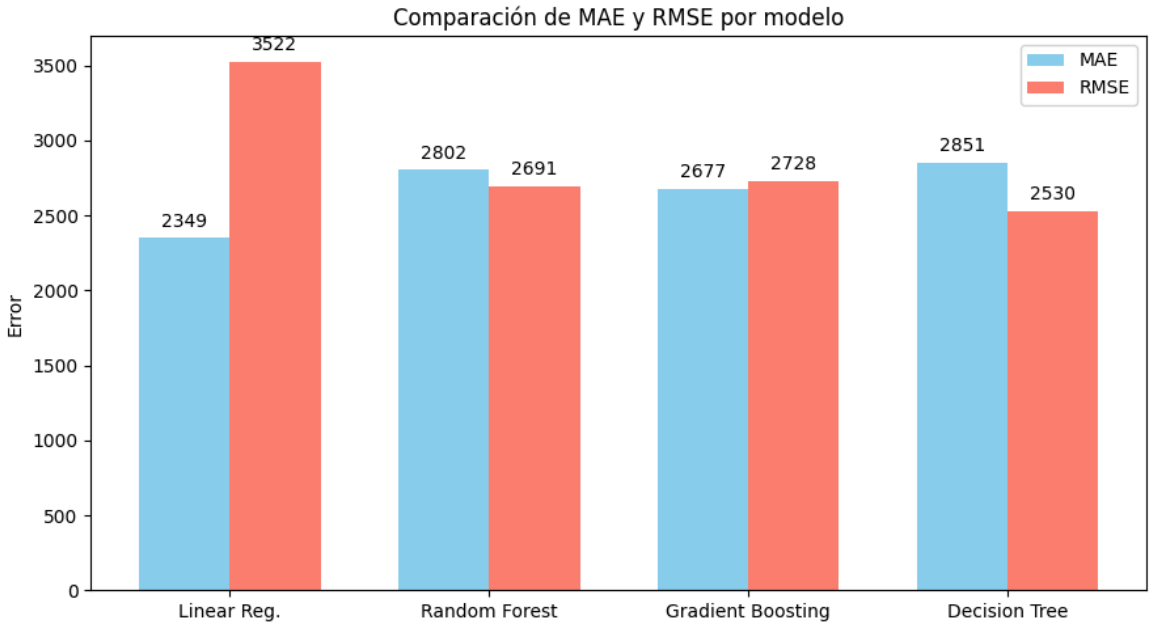
Selección de Modelos

Se seleccionaron los siguientes algoritmos de regresión:

1. **Regresión Lineal**:
    - Seleccionada como ‘baseline’ por su simplicidad e interpretabilidad
    - Útil para establecer un punto de referencia para comparar modelos más complejos
    - Permite entender relaciones lineales básicas entre las variables
2. **Random Forest**:
    - Elegido por su capacidad para capturar relaciones no lineales entre variables
    - Robusto frente a ruido en los datos y outliers
    - Adecuado para manejar patrones complejos en series temporales
    - Menos propenso al sobreajuste que un árbol de decisión individual
3. **Gradient Boosting**:
    - Seleccionado por su alta precisión y capacidad para manejar datos complejos
    - Particularmente efectivo para capturar patrones sutiles en los datos
    - Construye modelos secuencialmente, enfocándose en corregir los errores del modelo anterior
    - Conocido por su buen rendimiento en competiciones de ciencia de datos
4. **Decision Tree (Árbol de Decisión)**:
    - Elegido por su interpretabilidad y capacidad para proporcionar reglas claras de decisión
    - Permite visualizar el proceso de toma de decisiones
    - Establece un punto de comparación con modelos ensemble más complejos

Resultados

Comparación de Rendimiento de Modelos



Como se muestra en el gráfico, se utilizó el MAE que evalúa el promedio de los errores absolutos entre las predicciones y los valores reales. El RMSE, penaliza más fuertemente errores grandes, y es útil para identificar sensibilidad a valores extremos.

Basado en el gráfico de MAE y RMSE:

1. **Regresión Lineal**

   Mostró un patrón de subestimación de valores altos y una tendencia a predecir líneas rectas. Si bien su MAE fue bajo, su RMSE fue el más alto, lo que indica alta sensibilidad a outliers.

   a. MAE: 2,349
   b. RMSE: 3,522
   c. El RMSE más alto indica que este modelo es más sensible a valores atípicos

2. **Random Forest**

   Presentó mejor balance entre MAE y RMSE, y predicciones más realistas en general.

   a. MAE: 2,802

b. RMSE: 2,691
c. Rendimiento más equilibrado entre MAE y RMSE

3. **Gradient Boosting**

Presentó mejor balance entre MAE y RMSE, y predicciones más realistas en general.

a. MAE: 2,677
b. RMSE: 2,728
c. El mejor rendimiento general con el MAE y RMSE más equilibrados

4. **Decision Tree**

Aunque tuvo el RMSE más bajo, mostró un MAE elevado y una gran variabilidad en sus predicciones, indicando posible sobreajuste.

a. MAE: 2,851
b. RMSE: 2,530
c. RMSE más bajo pero MAE más alto

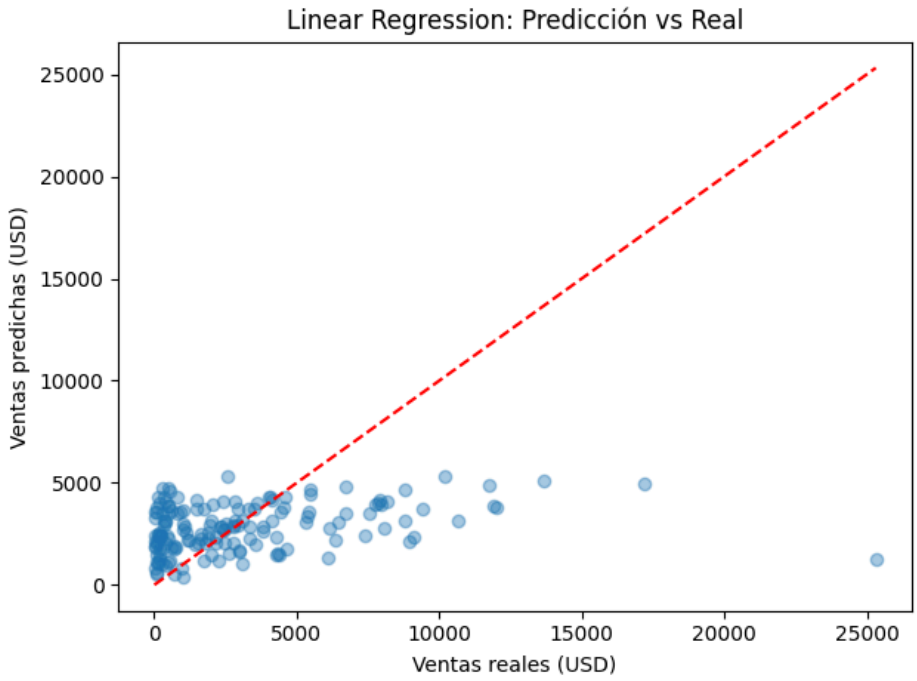
Análisis de Predicciones

- **Comportamiento de los Modelos**
  - La Regresión Lineal tiende a sobreestimar las ventas para todas las subcategorías, mostrando una tendencia lineal poco realista
  - Random Forest y Gradient Boosting muestran patrones más realistas y similares entre sí
  - El Decision Tree muestra las predicciones más variables, con cambios más abruptos entre subcategorías
- Predicciones por Subcategoría
  - Las subcategorías como "Chairs" y "Phones" muestran las predicciones más altas de ventas
  - "Envelopes" y "Fasteners" muestran las predicciones más bajas
  - Hay una notable variabilidad entre los modelos para ciertas subcategorías

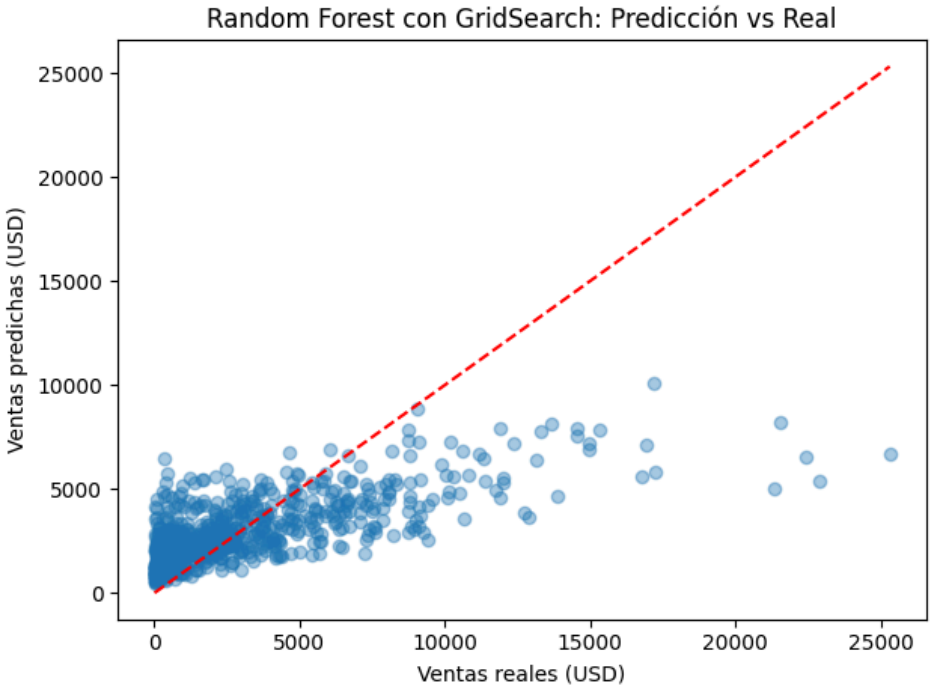
Visualización de Predicciones vs Valores Reales

Para esta parte, se realizaron diferentes gráficos con los diferentes modelos, los cuales muestran cómo se comportan las predicción y que tanto difieren de los datos reales.

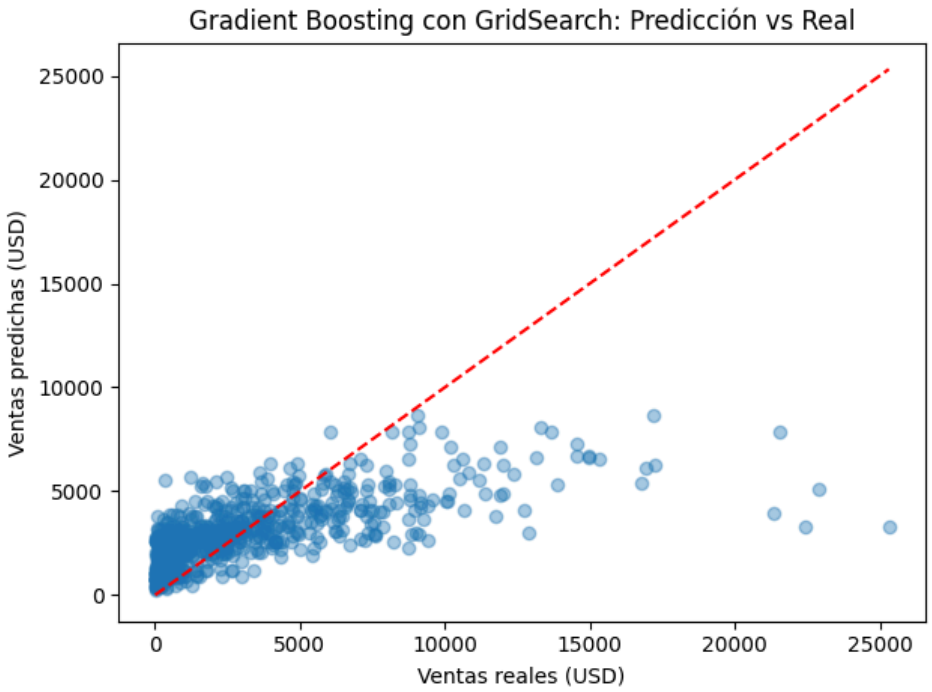
a. Regresión Lineal



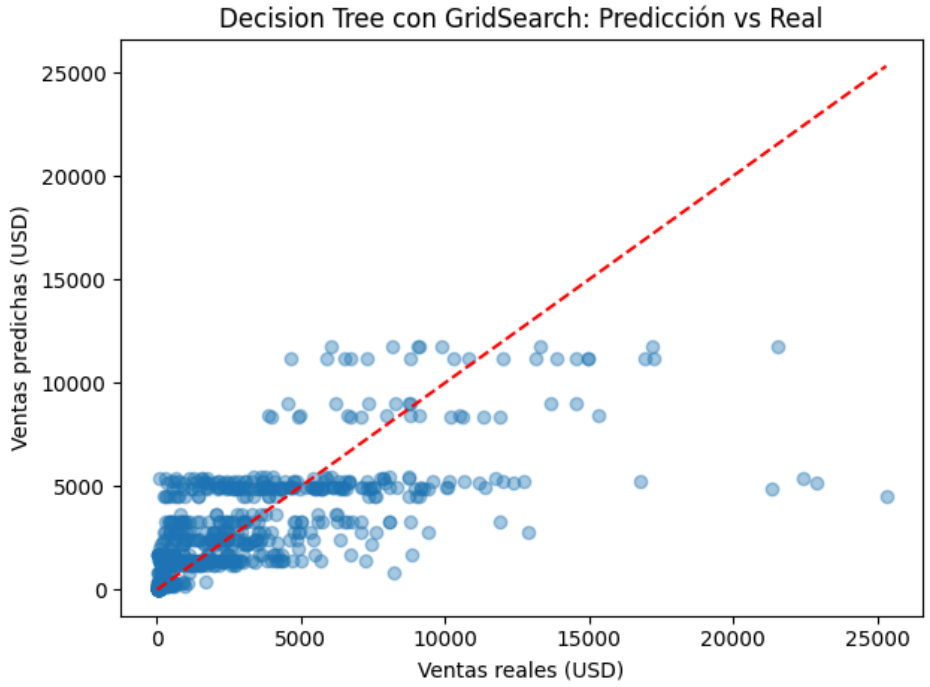
b. Random Forest



c. Gradient Boosting



d. Decision Tree



Explicación de los gráficos:

- Los gráficos muestran en el eje “x” las ventas y en el eje “y” las ventas predichas.

- Cada punto azul representa el total de ventas de una subcategoría por mes.
- La línea roja es la representación de que tan alejado de la perfección está la predicción realizada por el modelo.

Los gráficos de dispersión muestran que:

- Todos los modelos tienen dificultades para predecir valores extremos
- Gradient Boosting y Random Forest muestran la mejor distribución de predicciones
- La Regresión Lineal tiende a subestimar valores altos
- El Decision Tree muestra patrones más discretos en sus predicciones

Predicciones Finales

Como se puede observar con la salida se presentan los resultados de las ventas por mes y por sub-categoría dadas por los diferentes modelos:

Predicciones Regresión lineal de ventas por subcategoría y mes en 2025:

	Sub-Category	Year	Month	Predicted Sales
0	0	2025	1	4394.865221
1	0	2025	2	4692.217215
2	0	2025	3	4989.569208
3	0	2025	4	5286.921202
4	0	2025	5	5584.273195
5	0	2025	6	5881.625188
6	0	2025	7	6178.977182
7	0	2025	8	6476.329175
8	0	2025	9	6773.681169
9	0	2025	10	7071.033162
10	0	2025	11	7368.385156
11	0	2025	12	7665.737149

Predicciones Random Forest de ventas por subcategoría y mes en 2025:

	Sub-Category	Year	Month	Predicted Sales
0	0	2025	1	2483.303405
1	0	2025	2	2191.623651
2	0	2025	3	3132.526095
3	0	2025	4	2803.521438
4	0	2025	5	2948.871757
5	0	2025	6	3041.585530
6	0	2025	7	2953.935072
7	0	2025	8	3511.441741
8	0	2025	9	5298.021408
9	0	2025	10	4722.524618
10	0	2025	11	5474.461221
11	0	2025	12	4875.169296

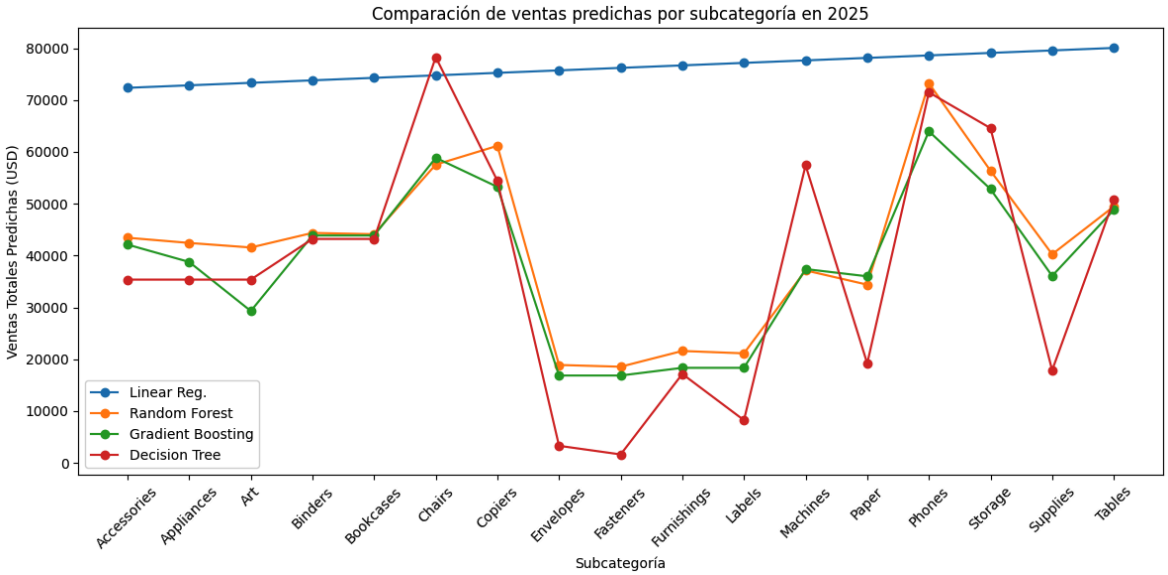
Predicciones Gradient Boosting de ventas por subcategoría y mes en 2025:

	Sub-Category	Year	Month	Predicted Sales
0	0	2025	1	2575.422299
1	0	2025	2	2427.462665
2	0	2025	3	2914.830934
3	0	2025	4	2837.673189
4	0	2025	5	2837.673189
5	0	2025	6	2837.673189
6	0	2025	7	2885.484238
7	0	2025	8	2911.130883
8	0	2025	9	4598.599914
9	0	2025	10	4598.599914
10	0	2025	11	5343.305916
11	0	2025	12	5343.305916

Predicciones Decision Tree de ventas por subcategoría y mes en 2025:

	Sub-Category	Year	Month	Predicted Sales
0	0	2025	1	2780.389943
1	0	2025	2	2780.389943
2	0	2025	3	2780.389943
3	0	2025	4	2780.389943
4	0	2025	5	2780.389943
5	0	2025	6	2780.389943
6	0	2025	7	2780.389943
7	0	2025	8	2780.389943
8	0	2025	9	3278.648708
9	0	2025	10	3278.648708
10	0	2025	11	3278.648708
11	0	2025	12	3278.648708

Con estos datos, se presenta el siguiente gráfico que muestra las ventas por categoría en el predichas para el año 2025:



Como se puede observar y se venía comentando anteriormente, el modelo de Regresión Lineal mostró una línea más constante, sin capturar la variabilidad real, esto podría indicar que le falta mayor entrenamiento con los datos. Por su contraparte, el modelo Decision Tree, presenta cambios bruscos y picos enunciados, esto podría ser un indicador que el modelo está haciendo un sobre ajuste con los datos. Sin embargo, podemos ver que los modelos de Random Forest y Gradient Boosting, se comportaron de una buena forma, lo que indica que podrían ser buenos candidatos para ser los elegidos para la solución final.

Conclusiones

1. **Selección del mejor modelo**: El Gradient Boosting demostró ser la opción más equilibrada para este caso de uso, ofreciendo el mejor balance entre precisión (medida por MAE) y consistencia (medida por RMSE).

2. **Consideraciones prácticas para el uso de modelos**: La evaluación comparativa permite recomendar diferentes modelos según las necesidades específicas:
    - Para predicciones conservadoras donde la subestimación sería más costosa que la sobreestimación: El Random Forest ofrece mayor robustez.
    - Para predicciones más precisas en el rango medio, donde se concentra la mayoría de las ventas: El Gradient Boosting proporciona el mejor rendimiento.

    - Para estimaciones rápidas o cuando la interpretabilidad es prioritaria: La Regresión Lineal puede servir como baseline, aunque con las limitaciones mencionadas.
    - Para casos donde se necesitan reglas explícitas de decisión: El Decision Tree ofrece mayor transparencia en el proceso predictivo.
3. **Patrones de ventas identificados**: El análisis por subcategoría reveló que productos como "Chairs" y "Phones" consistentemente muestran las predicciones más altas de ventas a lo largo de 2025, lo que sugiere áreas de enfoque para la gestión de inventario y estrategias promocionales.

4. **Limitaciones detectadas**: Es importante reconocer las limitaciones del enfoque actual:

- Todos los modelos mostraron dificultades para predecir valores extremos, particularmente ventas inusualmente altas.
- Existe cierta variabilidad en las predicciones entre modelos, lo que sugiere incertidumbre inherente en algunas subcategorías.
- La cantidad limitada de datos históricos (4 años) podría no ser suficiente para capturar ciclos económicos más largos o cambios estructurales en el mercado.

Recomendaciones

1. Implementar un modelo ensemble que combine las predicciones de Gradient Boosting y Random Forest para obtener mayor robustez. Este enfoque podría mitigar las debilidades individuales de cada modelo y mejorar la precisión general.
2. Incorporar variables predictoras adicionales como datos de promociones, eventos especiales, factores económicos o tendencias de mercado que podrían influir en las ventas.
3. Considerar la ampliación del horizonte predictivo más allá de un año para planificación estratégica a largo plazo.
4. Desarrollar un sistema de monitoreo continuo que compare automáticamente las predicciones con los datos reales a medida que se generan, permitiendo ajustes rápidos a los modelos.

Referencias

Sahoo, R, (2020). *Super Store Sales DataSet.* Recuperado de: https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting

Scikit-learn, (s.f). *Scikit-Learn Machine Learning.* Recuperado de: https://scikit-learn.org/stable/index.html