**Proyecto de Predicción de Ventas por Subcategoría**

**Descripción General**

El propósito de este proyecto fue implementar un sistema de predicción de ventas. Se utilizaron diferentes modelos de machine learning para pronosticar las ventas por subcategoría para el año 2025. Se implementaron y compararon cuatro modelos diferentes:

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

**Estructura del Proyecto**

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

**Metodología**

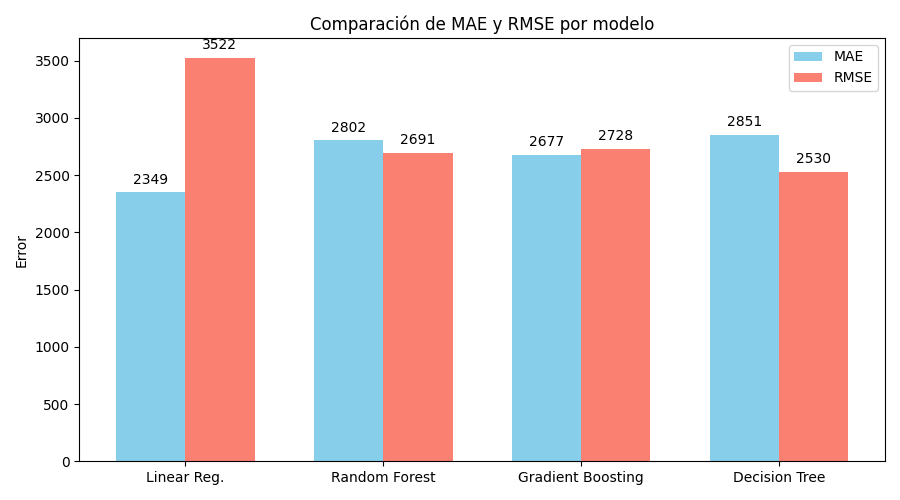
1. **Preprocesamiento de Datos**
   1. Limpieza de datos
   2. Transformación de fechas
   3. Agregación de ventas por subcategoría y mes
   4. Codificación de variables categóricas
2. **Implementación de Modelos**
   1. Se implementaron cuatro modelos diferentes
   2. Cada modelo fue entrenado con los datos históricos
   3. Se utilizó GridSearch para optimizar hiperparámetros en Random Forest, Gradient Boosting y Decision Tree
3. **Evaluación de Modelos**

Se utilizaron dos métricas principales:

* 1. MAE (Error Absoluto Medio)
  2. RMSE (Error Cuadrático Medio)

**Resultados**

Comparación de Rendimiento de Modelos



Basado en el gráfico de MAE y RMSE:

1. **Regresión Lineal**
   1. MAE: 2,349
   2. RMSE: 3,522
   3. El RMSE más alto indica que este modelo es más sensible a valores atípicos
2. **Random Forest**
   1. MAE: 2,802
   2. RMSE: 2,691
   3. Rendimiento más equilibrado entre MAE y RMSE
3. **Gradient Boosting**
   1. MAE: 2,677
   2. RMSE: 2,728
   3. El mejor rendimiento general con el MAE y RMSE más equilibrados
4. **Decision Tree**
   1. MAE: 2,851
   2. RMSE: 2,530
   3. RMSE más bajo pero MAE más alto

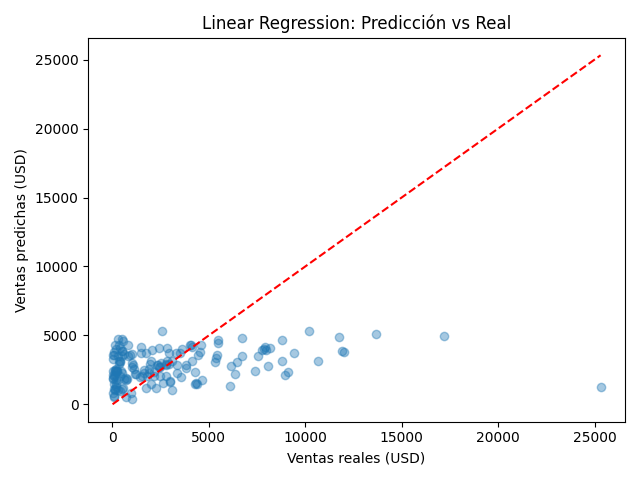
**Análisis de Predicciones**

* **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

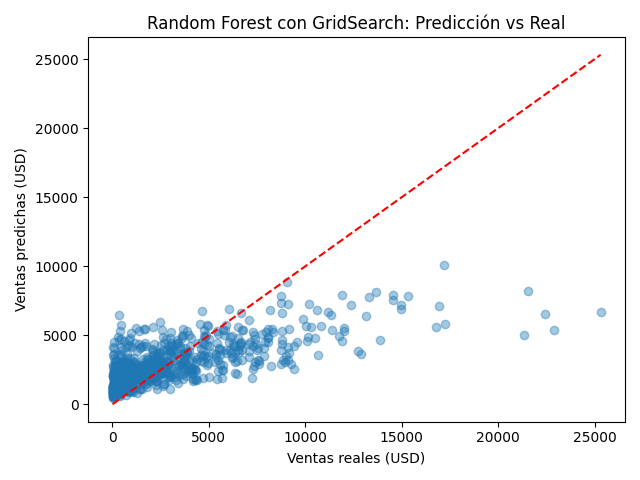
**Visualización de Predicciones vs Valores Reales**

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

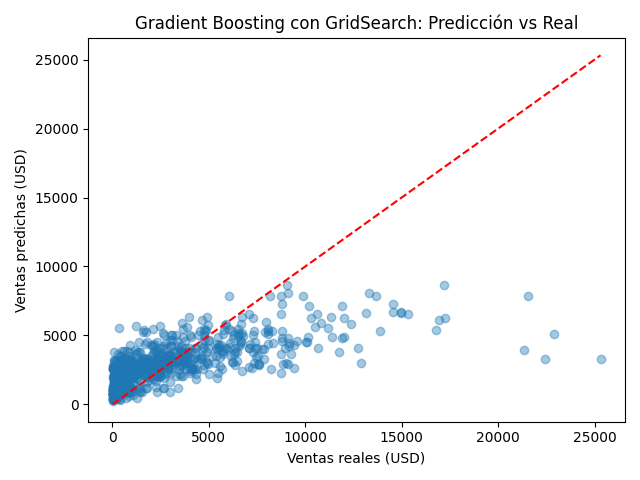
1. Regresión Lineal



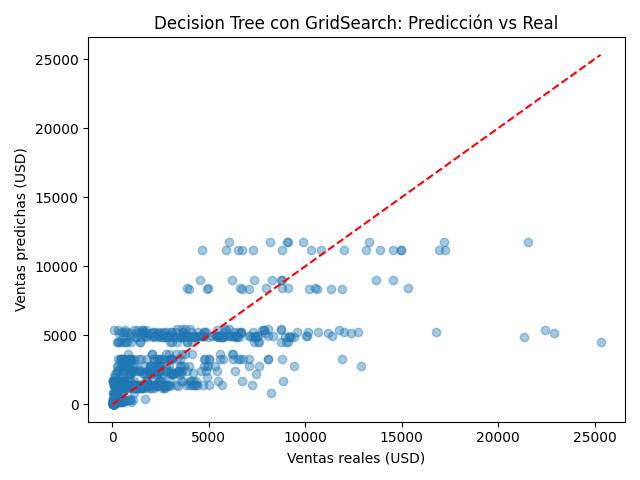
1. Random Forest



1. Gradient Boosting



1. 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 esta 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

**Conclusiones**

1. Mejor Modelo: El Gradient Boosting muestra el mejor rendimiento general, con el mejor balance entre MAE y RMSE.
2. Consideraciones de Uso:
   1. Para predicciones conservadoras: Usar Random Forest
   2. Para predicciones más precisas en el rango medio: Usar Gradient Boosting
   3. Para estimaciones rápidas: La Regresión Lineal puede servir como baseline
3. Limitaciones:
   1. Todos los modelos muestran dificultades con valores extremos
   2. Hay cierta variabilidad en las predicciones entre modelos
   3. Los patrones estacionales podrían no estar completamente capturados

**Recomendaciones**

1. Considerar la implementación de un modelo ensemble que combine las predicciones
2. Incorporar más variables predictoras si están disponibles
3. Realizar un análisis más detallado de los patrones estacionales
4. Implementar un sistema de monitoreo para validar las predicciones con datos reales futuros