

Predicción de Ventas de Videojuegos

Machine Learning Supervisado



Objetivo y Dataset

OBJETIVO

Predecir ventas globales de videojuegos (en millones)

TIPO DE PROBLEMA

REGRESIÓN
(target numérico continuo)

DATASET

16.444

juegos

10

variables utilizadas (x)

2

datasets

Hipotesis a Priori

H1

Los juegos con mejor nota de criticos venderan mas

H2

Los juegos en consolas mas vendidas venderan mas

H3

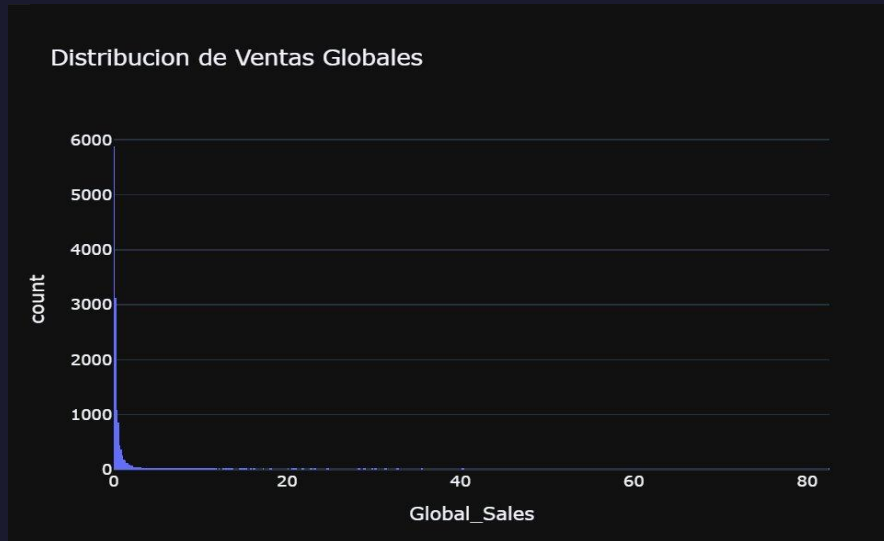
Los juegos multiplataforma venderan mas

H4

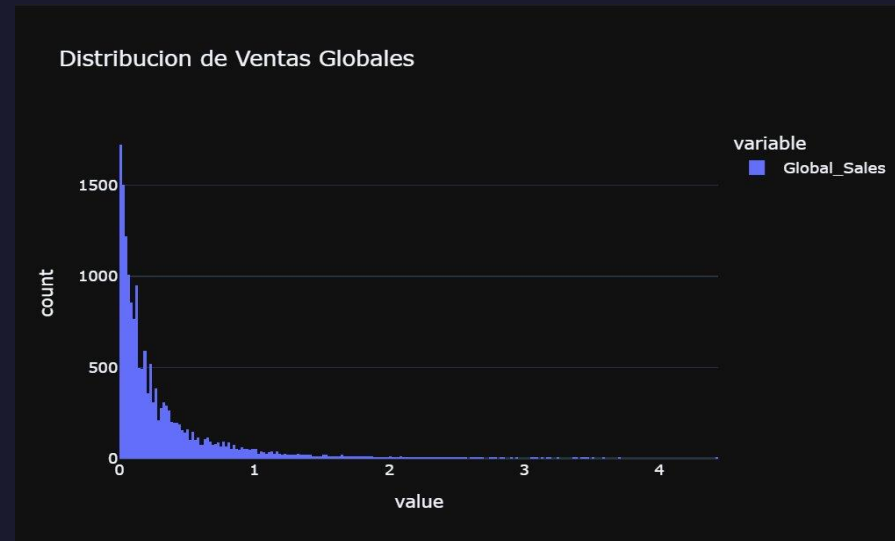
La clasificacion por edad (Rating) influira en las ventas

Validaremos estas hipotesis con el modelo OLS y los resultados finales

Variable Target: Global_Sales



ANTES (sesgado)



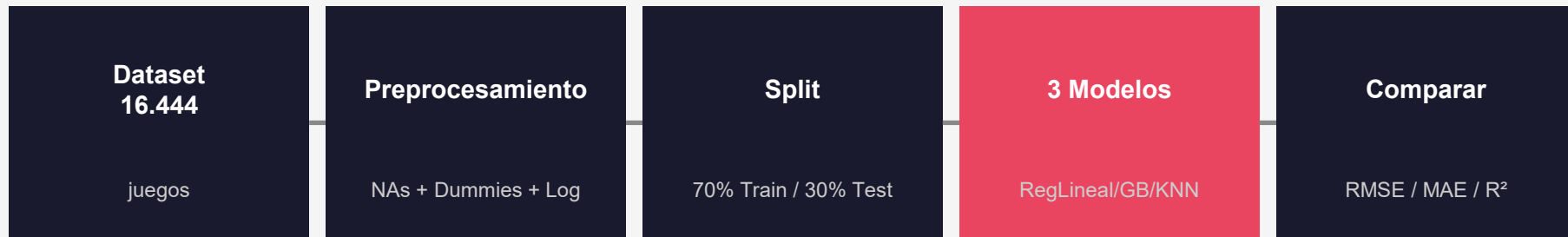
DESPUÉS (con log)

Asimetría 17,38 $\rightarrow$ 2,76 con $\log(\text{ventas}+1)$. El modelo se entrena en esa escala: sus métricas son logarítmicas

Variables Explicativas (X)

Variable	Tipo	Origen
Critic_Score	Numérica	Dataset principal
Critic_Count	Numérica	Dataset principal
User_Score	Numérica	Dataset principal
User_Count	Numérica	Dataset principal
Year_of_Release	Numérica	Dataset principal
Genre	Categórica (11 dummies)	Dataset principal
Rating	Categórica (8 dummies)	Dataset principal
Console_Units_M	Numérica	Enriquecimiento
Num_Platforms	Numérica	Creada por código
Sin_Consola	Binaria (0/1)	Creada por código

Flujo de Machine Learning



Modelos utilizados:

1. **Regresión Lineal:** modelo base, relaciones lineales
2. **Gradient Boosting:** ensemble de árboles, captura no linealidades
3. **KNN:** busca los K juegos más parecidos y promedia sus ventas

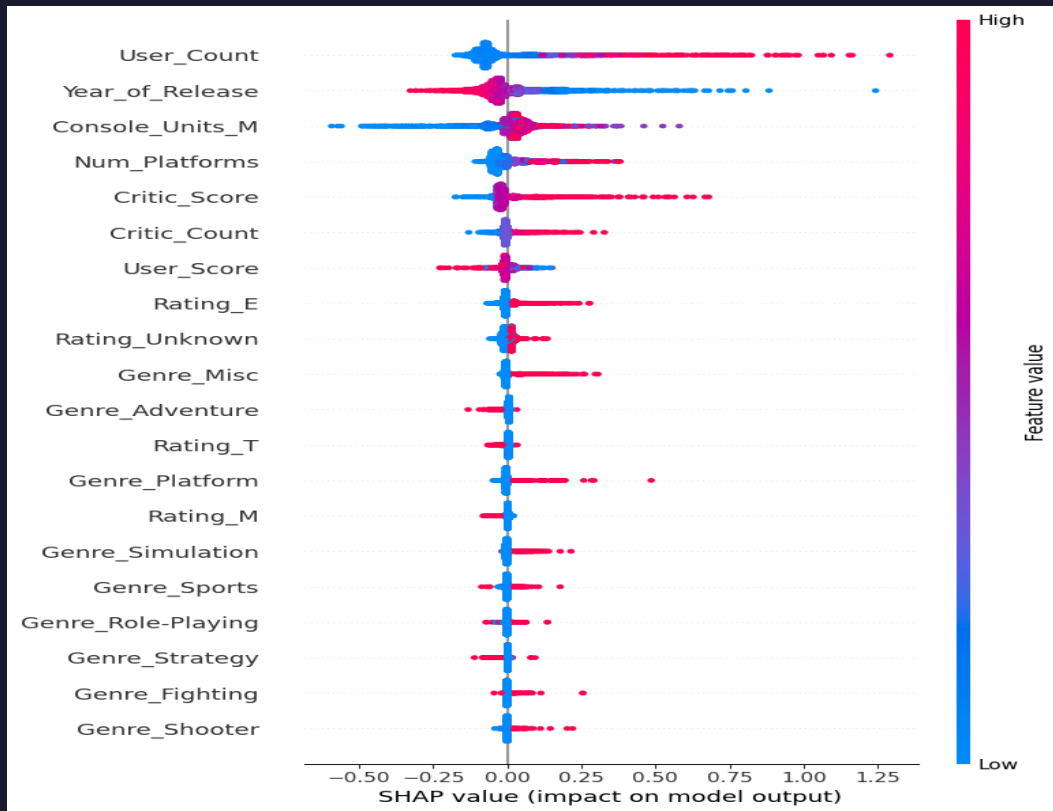
Resultados

Modelo	RMSE (log)	MAE (log)	R ² (log)
Baseline (media)	0,40	0,27	0,00
Regresión Lineal	0,34	0,22	0,29
KNN (K=11)	0,33	0,20	0,35
Gradient Boosting	0,29	0,18	0,50

GANADOR: Gradient Boosting

$R^2 = 0,50$ sobre $\log(\text{ventas}+1)$ · En millones: $R^2 = 0,23$ · RMSE 1,60 M · MAE 0,38 M

Interpretabilidad: SHAP Values



Variables más importantes:

1. User_Count
2. Year_of_Release
3. Console_Units_M
4. Num_Platforms
5. Critic_Score
6. Critic_Count

Rojo = valor alto
Azul = valor bajo
Derecha = aumenta ventas

Validación de Hipótesis

SÍ

Mejor nota de críticos → más ventas

Significativo con errores robustos HC3 y 5° en SHAP. Ojo: falta en el 51,5% de las filas

SÍ

Consola más vendida → más ventas

Significativo con HC3 y 3° en SHAP. Sin_Consola también significativa y negativa

SÍ

Multiplataforma → más ventas

Significativo con HC3 y 4° en SHAP. Pero es información posterior al lanzamiento

NO

Rating influye en ventas

Sin evidencia con estos datos: ninguna categoría es significativa, y el 41% de la columna es Unknown

Limitaciones y Utilidad

LIMITACIONES

- Critic_Score y Num_Platforms son datos posteriores al lanzamiento
- Console_Units_M es el total histórico de la consola
- Dataset cortado en dic-2016: las ventas de 2016 están sin acumular
- $R^2 = 0,50$ en log · 0,23 sobre ventas en millones

UTILIDAD PRÁCTICA

- Tras lanzar demo: predecir ventas con las primeras críticas
- Evitar fracasos: identificar combinaciones que venden poco
- Entender qué variables influyen en las ventas

Conclusión Principal

Gradient Boosting: $R^2 = 0,50$ sobre $\log(\text{ventas}+1)$, 0,23 sobre ventas en millones

Las variables en las que más se apoya el modelo son:

User_Count, Year_of_Release, Console_Units_M, Num_Platforms

Tras lanzar una demo, usar las primeras críticas para estimar ventas finales y decidir inversión en marketing