Análisis y Predicción del Puntaje de Críticos en la Industria de los Videojuegos

Raúl Quirós Morales

Abstract—Este proyecto aborda la predicción de puntajes de críticos en la industria de videojuegos a través de un enfoque metódico de análisis de datos y modelado predictivo. Se abordan desafíos como la gestión de valores nulos, la selección de modelos y la optimización de hiperparámetros. El GradientBoostingRegressor emerge como el modelo más prometedor, con resultados sólidos en validación cruzada y conjunto de prueba. Se destacan factores clave, como ventas regionales y recuentos de usuarios, en la predicción de puntajes. Se advierte sobre los límites inherentes a la precisión predictiva exacta. Las conclusiones ofrecen perspectivas valiosas para la toma de decisiones en la industria de los videojuegos.

Keywords—Videojuegos, Aprendizaje Automático, Predicción de Puntajes, Evaluación Crítica, Industria del Entretenimiento.

1 Introducción

La industria de los videojuegos ha experimentado un crecimiento significativo en las últimas décadas, convirtiéndose en un sector clave del entretenimiento global. Con el continuo surgimiento de nuevas plataformas, géneros y títulos, entender los factores que influyen en la recepción de los videojuegos se vuelve esencial. En este contexto, el presente proyecto se enfoca en la Predicción del Puntaje de Críticos y Usuarios como una herramienta analítica para comprender y anticipar la aceptación de los videojuegos en el mercado.

La evaluación de videojuegos por parte de críticos y usuarios desempeña un papel crucial en la determinación de su éxito y en la toma de decisiones por parte de los consumidores. Este análisis busca profundizando en patrones, correlaciones y características clave que contribuyan a la recepción positiva o negativa de los videojuegos. La aplicación de modelos predictivos permitirá a desarrolladores, editores y entusiastas del sector anticipar el desempeño potencial de un videojuego antes de su lanzamiento, mejorando así la toma de decisiones estratégicas en el competitivo mercado de los videojuegos.

2 Propuesta

El propósito central es desarrollar un modelo predictivo de puntajes de críticos y usuarios. Se empleará el conjunto de datos proporcionado, que contiene información sobre videojuegos, como ventas globales, distribuidor, diseñador, etc. La propuesta busca entender cómo estas variables influyen en las puntuaciones asignadas. Se pretende lograr una comprensión profunda de los factores que contribuyen a la percepción positiva o negativa de un videojuego, permitiendo la construcción de un modelo eficaz de predicción.

2 Metodología

Para alcanzar los objetivos propuestos, se ha seguido una metodología estructurada que abarca diferentes etapas clave en la construcción del modelo predictivo de puntajes de críticos y usuarios. Iniciando con el Análisis Exploratorio de Datos (EDA), se llevó a cabo una evaluación del conjunto de datos, seguida de un análisis para identificar posibles relaciones entre las diversas variables y los puntajes objetivo.

En la etapa de Preprocesamiento, se aseguró la integridad de los datos mediante la identificación y eliminación de valores nulos en el conjunto de datos. Además, las variables categóricas se transformaron utilizando técnicas de encoding para facilitar su inclusión en los modelos de aprendizaje automático.

La Selección de Métricas (2.3) involucró un análisis detallado para identificar las métricas más apropiadas que reflejaran de manera efectiva el rendimiento del modelo en la tarea de predicción. La evaluación se realizó mediante la eliminación de columnas con baja correlación en el conjunto de datos de entrenamiento. Se aplicó un criterio de selección basado en la correlación entre las métricas y los atributos objetivo (puntajes de críticos y usuarios). Se eliminaron aquellas métricas que mostraron una correlación inferior al 10% con los atributos target, resultando en la eliminación de las columnas 'Platform', 'Year_of_Release', 'Genre', y 'Rating'.

En la fase de Selección del Modelo con Crossvalidation, se llevó a cabo un análisis exhaustivo de diferentes modelos de regresión. Se evaluaron modelos como la regresión lineal, DecisionTreeRegressor, RandomForestRegressor, XGBRegressor, GradientBoostingRegressor, y AdaBoostRegressor. Tras observar las puntuaciones de validación cruzada y las métricas de rendimiento, se seleccionó GradientBoostingRegressor (GBR) debido a su desempeño superior, con un R-squared de 0.302 en la validación cruzada.

Posteriormente, se ajustaron los hiperparámetros del modelo GBR utilizando GridSearchCV, identificando los mejores hiperparámetros. Con estos hiperparámetros optimizados, el modelo GBR fue evaluado en el conjunto de prueba, mostrando un buen rendimiento con un Mean Squared Error y un R-squared que reflejan la calidad de las predicciones.

En el Análisis Final, se evaluaron los resultados del modelo GBR en el conjunto de prueba, calculando métricas de rendimiento como el Mean Squared Error (MSE) y el R-squared (R2). Estas métricas proporcionaron información valiosa para la toma de decisiones informada en el desarrollo y lanzamiento de nuevos videojuegos en el contexto de la industria.

3 Experimentos, resultados i analisis

3.1 Preprocesamiento y selección de métricas

En la fase inicial, se llevó a cabo un análisis del conjunto de datos para comprender su estructura y desafíos potenciales. Se identificó la presencia de valores nulos en columnas clave como 'Critic_Score', 'Critic_Count', 'User_Score', 'User_Count', 'Developer' y 'Rating', lo que resultó en 7905 filas con valores nulos. En este momento se reconoció la importancia de este

problema.

En la fase de preprocesamiento, se implementaron acciones como la eliminación de la columna 'Name', la conversión de 'tbd' a NaN en 'User_Score' y la eliminación de filas con valores nulos. Este proceso resultó en 5470 filas restantes en el conjunto de entrenamiento, estableciendo una base más reducida pero segura para el análisis y modelizaciones futuras.

También se llevó a cabo la eliminación de columnas con baja correlación en el conjunto de entrenamiento. Las columnas 'Platform', 'Year_of_Release', 'Genre', y 'Rating' se identificaron y eliminaron debido a su baja correlación con las críticas. Este enfoque se basó en el análisis de la matriz de correlación, que proporcionó información valiosa sobre las relaciones entre las variables. A continuación, se presenta el gráfico de correlaciones recortado, que respalda la decisión de eliminar estas columnas, contribuyendo así a la mejora de la calidad del conjunto de datos para su posterior modelización.

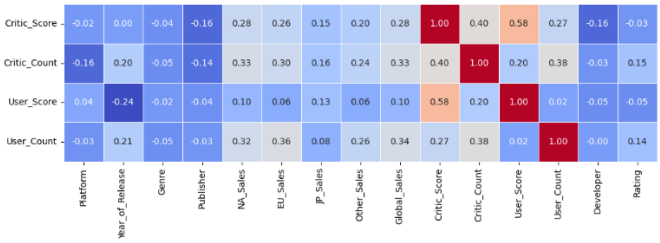

Fig. 1. Matriz de correlación (recortada).

3.2 Selección, Validación y Ajuste de Modelo de Predicciones

3.2.1 Regresión Lineal Inicial

Se inició el análisis con una regresión lineal para establecer un punto de referencia. Los resultados revelaron un Mean Squared Error (MSE) de 1.728 y un R-squared de 0.129, indicando una capacidad predictiva modesta.

3.2.2 Evaluación de Modelos Avanzados

Se exploraron varios modelos avanzados, incluyendo DecisionTreeRegressor, RandomForestRegressor, XGBRegressor, GradientBoostingRegressor (GBR), y AdaBoostRegressor. La validación cruzada proporcionó una visión general del rendimiento de cada modelo.

Regresión Lineal (LR):

Cross-Validation Scores: [0.1217, 0.0976, 0.1024, 0.1469, 0.1431]
Average R-squared: 0.1223
MSE (Critic_Score): 1.6982
R-squared (Critic_Score): 0.1446

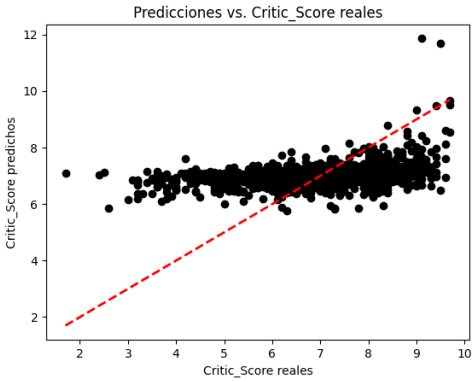

Fig. 2. Resultados de la Regresión Lineal.

DecisionTreeRegressor (DTR):

Cross-Validation Scores: [-0.4190, -0.2959, -0.5467, -0.3417, -0.2953]
Average R-squared: -0.3797
MSE (Critic_Score): 2.6456
R-squared (Critic_Score): -0.3327

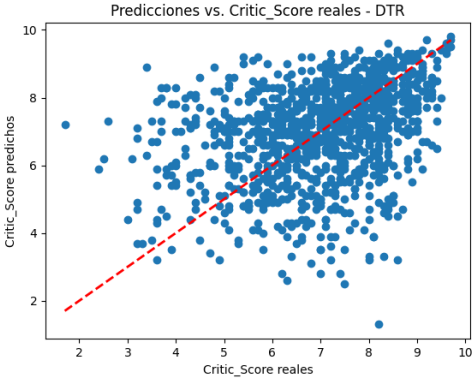

Fig. 3. Resultados al aplicar DTR.

RandomForestRegressor (RFR):

Cross-Validation Scores: [0.2376, 0.2432, 0.2444, 0.2953, 0.2949]
Average R-squared: 0.2631
MSE (Critic_Score): 1.3519
R-squared (Critic_Score): 0.3190

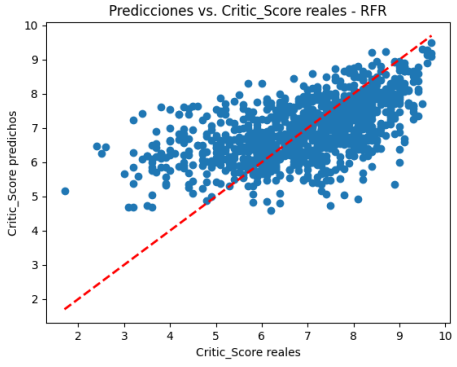

Fig. 4. Resultados al aplicar RFR.

XGBRegressor (XGBR):

Cross-Validation Scores: [0.1549, 0.1908, 0.2095, 0.2262, 0.2537]

Average R-squared: 0.2070

MSE (Critic_Score): 1.4211

R-squared (Critic_Score): 0.2842

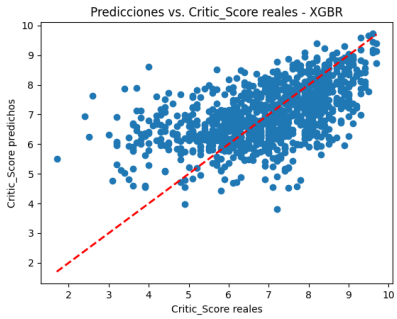

Fig. 5. Resultados al aplicar XGBR.

GradientBoostingRegressor (GBR):

Cross-Validation Scores: [0.2824, 0.2888, 0.2787, 0.3373, 0.3248]

Average R-squared: 0.3024

MSE (Critic_Score): 1.3297

R-squared (Critic_Score): 0.3302

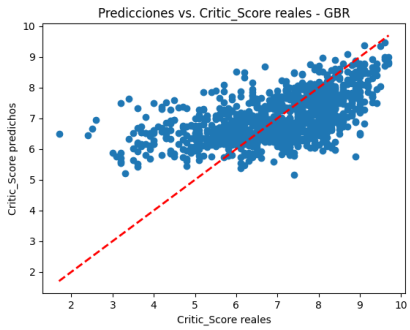

Fig. 6. Resultados al aplicar GBR.

AdaBoostRegressor (ABR):

Cross-Validation Scores: [0.1769, 0.1511, 0.1036, 0.0770, 0.1213]

Average R-squared: 0.1260

MSE (Critic_Score): 1.5972

R-squared (Critic_Score): 0.1955

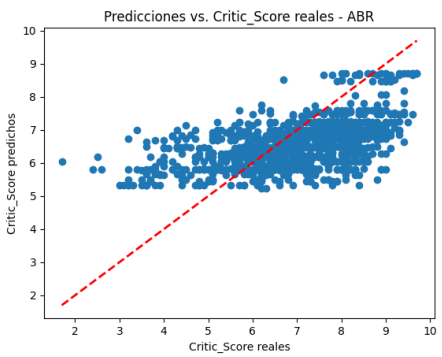

Fig. 7. Resultados al aplicar ABR.

3.2.3 Selección del Modelo

En base a los resultados, el modelo GradientBoostingRegressor (GBR) se seleccionó debido a que es el modelo con mejores resultados (pese a no ser excelentes), con el R-squared promedio más alto en la validación cruzada (0.3024). Además, tiene un MSE (Critic_Score) relativamente bajo (1.3297) y un R-squared (Critic_Score) superior (0.3302) en comparación con los otros modelos, respaldando su elección como el modelo preferido para la predicción de puntajes de críticos.

3.2.4 Ajuste de Hiperparámetros

Se procedió al ajuste fino de los hiperparámetros del modelo GBR utilizando GridSearchCV. Los mejores hiperparámetros identificados tras una larga ejecución fueron:

- Learning Rate: 0.1
- Max Depth: 4
- Min Samples Leaf: 4
- Min Samples Split: 2
- N Estimators: 100
- Subsample: 0.8

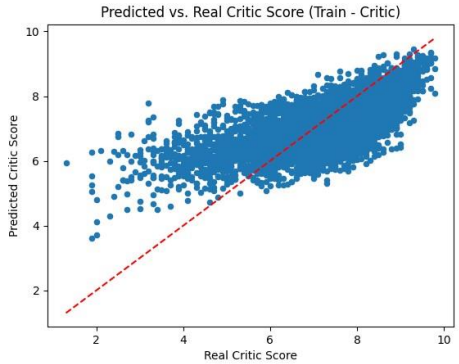

Fig. 7. Resultados al aplicar GBR con los hiperparámetros ajustados.

3.2.5 Evaluación Final del Modelo

Con los hiperparámetros óptimos, el modelo GBR fue evaluado en el conjunto de prueba, mostrando un MSE de 1.311 y un R-squared de 0.340. Estos resultados confirman la mejora sustancial en la capacidad predictiva del modelo después del ajuste.

5 CONCLUSIONES

En la culminación de este estudio, se extraen conclusiones significativas sobre el rendimiento y la utilidad del modelo GradientBoostingRegressor (GBR) en la predicción de puntajes de críticos para videojuegos, así como sobre la relevancia de los atributos seleccionados.

Precisión del Modelo: El modelo GBR exhibe una capacidad relativamente sólida para predecir los puntajes de críticos, respaldada por un R-squared promedio de 0.3024 en la validación cruzada y un Mean Squared Error (MSE) de 1.3297 en el conjunto de prueba. Estas métricas sugieren una buena capaci-

dad explicativa y una precisión aceptable en las predicciones, aunque no exenta de margen de mejora.

Margen de Error: Es importante destacar que, a pesar de su rendimiento prometedor, el modelo aún puede tener limitaciones y errores inherentes en la predicción de puntajes de críticos. Factores externos no considerados y la naturaleza dinámica del mercado de videojuegos pueden contribuir a variaciones imprevistas en las predicciones.

Atributos Relevantes: El análisis de correlación y la selección de atributos revelaron que ciertos atributos, como 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales' y 'User_Count', influyen significativamente en la predicción de puntajes de críticos. Estos atributos han demostrado tener una correlación sustancial con los puntajes objetivo y, por lo tanto, desempeñan un papel crucial en la determinación de la calidad percibida de un videojuego.