



Industria de videojuegos

¿Qué factores influyen en el éxito de un videojuego?

Autor: Carlos González



Tabla de contenido

01

Motivación

02

Objetivo

03

Metadata

04

Análisis
Exploratorio

05

Resultados

06

Insights y
Conclusiones

Motivación

En la industria de los videojuegos, comprender los factores que influyen en el éxito de un juego es de vital importancia. Se lanzan miles de títulos cada año, por lo que existe un mercado donde las preferencias de los jugadores pueden cambiar rápidamente. Según esto, la capacidad de obtener insights basados en datos es crucial para tomar decisiones informadas que maximicen las oportunidades de éxito.

El análisis de datos en la industria de los videojuegos no solo permite identificar las tendencias y patrones actuales, sino que también proporciona una base sólida para predecir el rendimiento futuro de los juegos. Estos datos se pueden utilizar para ajustar estrategias de desarrollo, marketing y ventas, asegurando que los productos se alineen con las expectativas y deseos de los jugadores.

Este análisis se centra en un dataframe que contiene información detallada sobre una colección de juegos. El dataset incluye diversas características que podrían influir en la popularidad y el rendimiento de un juego, tales como el título del juego, la fecha de lanzamiento, el equipo de desarrollo, la calificación promedio, el número de reseñas, los géneros del juego, el número de jugadas y el número de veces que un juego ha sido añadido a la lista de deseos de los usuarios.

Objetivos

Se tiene como objetivo utilizar el dataset a estudiar para realizar un análisis que permita entender mejor los factores que contribuyen al éxito de un juego. Se tiene como preguntas iniciales a responder las siguientes:

- ¿Qué géneros de juegos son los más populares?
- ¿Cómo afecta el número de jugadas al rating?
- ¿Cuál es el impacto de las reseñas en la popularidad de un juego?
- ¿Qué equipo de desarrollo tiene los juegos con mayores ratings?

Metadata

Se utiliza un dataset de la industria de videojuegos con datos de juegos populares desde 1980 a 2023, se incluyen juegos indies y de compañías grandes.

Se incluyen tanto videojuegos como DLC y versiones variadas del mismo juego.

1512

Filas

Total de filas del dataset.

14

Columnas iniciales

Columnas que contiene el dataset inicial.

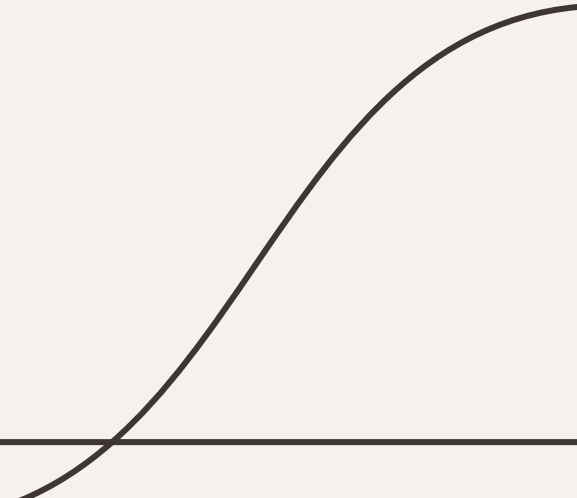
8

Columnas de interes

Columnas utilizadas para el estudio.

Dataset

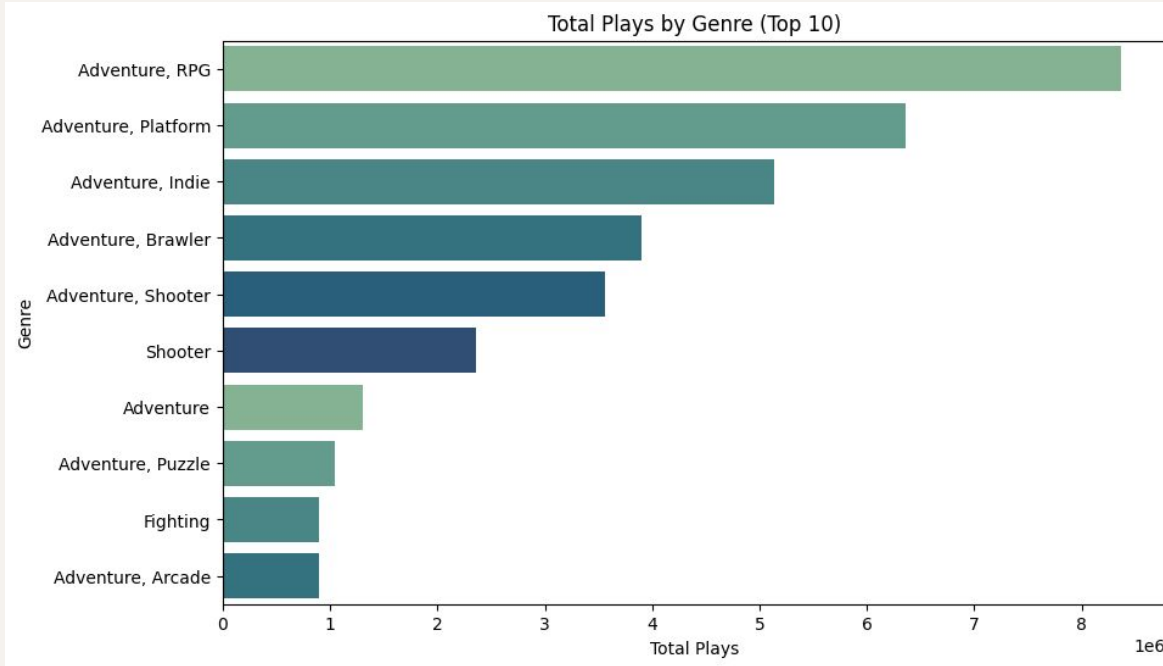
En el dataset importado se tienen las siguientes columnas:

- **Title:** Corresponde al título de cada juego, se incluyen DLC y versiones diferentes del mismo juego.
 - **Release Date:** Fecha de lanzamiento.
 - **Team:** Equipo de desarrollo.
 - **Rating:** Puntaje otorgado por los jugadores.
 - **Times Listed:** Número de usuarios que agregaron el juego a su lista.
 - **Number of Reviews:** Número de reseñas.
 - **Genres:** Generos.
 - **Summary:** Resumen de lo que trata el juego.
 - **Reviews:** Algunas reseñas dadas por los usuarios.
 - **Plays:** Número de veces que se ha jugado ese juego.
 - **Playing:** Número de jugadores activos al momento de obtener los datos.
 - **Backlogs:** Número de usuarios que tienen el juego pero aún no lo inician.
 - **Wishlist:** Número de usuarios que agregaron el juego a lista de deseados.
- 

The image features a light gray background with two thin, dark gray horizontal lines. In each of the four corners, there is a decorative wavy line that curves from the horizontal line towards the center, creating a subtle frame effect.

Análisis Exploratorio

¿Qué géneros de juegos son los más populares?



Del 'Gráfico 1' mostrado a la izquierda se puede ver que el género de mayor popularidad corresponde al de "Aventuras" (Adventure), dado que en los puestos más altos se tiene este género, acompañado de los generos 'RPG', 'Plataforma' y 'Disparos' en los primeros tres lugares de popularidad.

Gráfico 1. Top10 Género por total de veces jugados.

¿Cómo afecta el número de jugadas al rating?

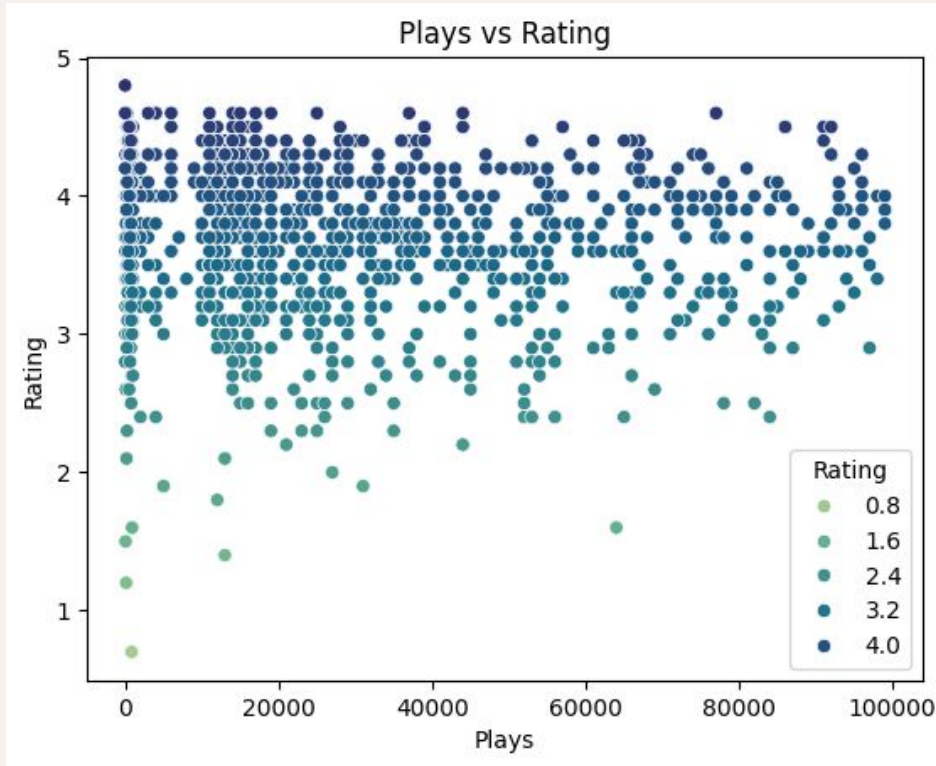


Gráfico 2. Plays vs Rating

Del 'Gráfico 2' mostrado a la izquierda se puede ver que los juegos con menor cantidad de usuarios muestran calificaciones más variadas, del 1 al 5, mientras que los con más usuarios tienen las calificaciones más concentradas en el rango de 3 a 5. Además, se observa la tendencia de que los juegos más populares tienden a tener calificaciones más altas, indicando que los juegos más populares son mejor recibidos por los usuarios.

¿Cual es el impacto de las reseñas en la popularidad de un juego?

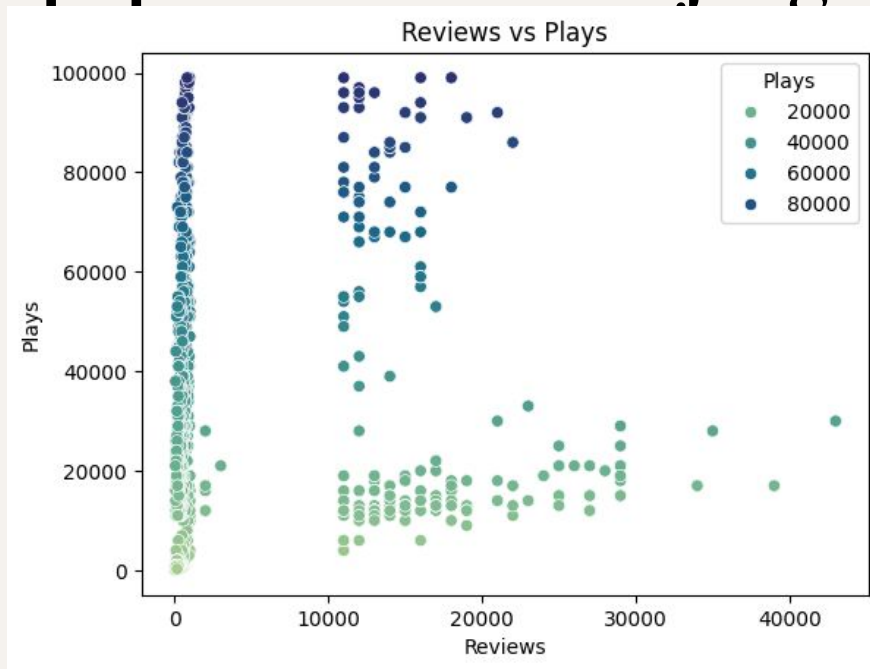


Gráfico 3. Reviews vs Plays

Del 'Gráfico 3' mostrado a la izquierda se muestra que la mayoría de los juegos tienen menos de 1000 reseñas, y estos juegos pueden tener desde pocas hasta cerca de 100000 jugadas, indicando que el número de reseñas no siempre refleja la popularidad en términos de jugadas. Los juegos con más de 1000 reseñas están más dispersos y no presentan una relación clara entre el número de reseñas y las jugadas; algunos tienen muchas jugadas y otros, menos.

¿Qué equipo de desarrollo tiene los juegos con mayores rating?

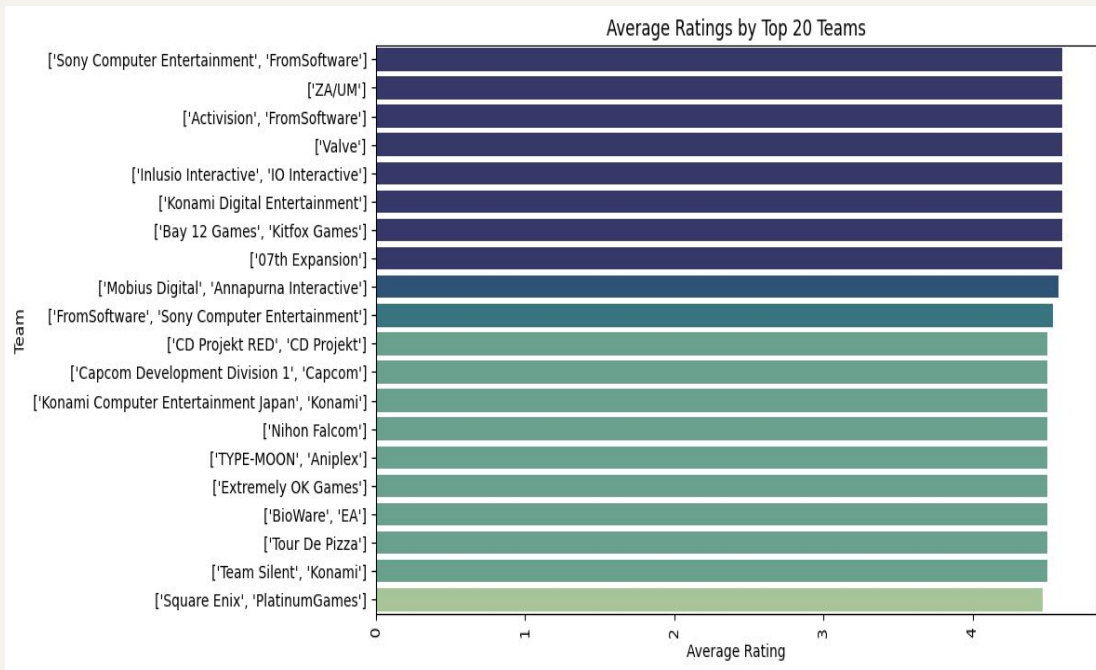


Gráfico 4. Average Rating by Top 20 Teams.

Del 'Gráfico 4' mostrado a la izquierda se puede ver que los equipos con las calificaciones promedio más altas tienen calificaciones cercanas a 4.5 o superiores. Esto indica que estos equipos consistentemente producen juegos de alta calidad según las calificaciones de los usuarios.

Equipos como "Sony Computer Entertainment" y "FromSoftware" aparecen múltiples veces, lo que sugiere una reputación consistente de alta calidad en sus juegos.

Algoritmos Machine Learning

Se utilizó Lazy Predict para ver los mejores modelos para el estudio. Dentro de estos se eligen los modelos:

- **Random Forest Regressor:** Funciona creando múltiples árboles de decisión en el proceso de entrenamiento y luego promedia sus predicciones para obtener un resultado más preciso. Cada árbol se entrena con un subconjunto diferente de los datos y características, lo que ayuda a reducir el sobreajuste y mejora la generalización del modelo.
- **Light Gradient Boosting Machine Regressor:** LGBM utiliza una técnica llamada "histograma" para reducir el tiempo de entrenamiento y manejar grandes volúmenes de datos con mayor eficiencia. Es conocido por su alta velocidad, eficiencia en memoria, y su capacidad para manejar datos desbalanceados.

Consideraciones

- **Cross Validation:** Se utilizó el método de validación cruzada “K-Fold Cross Validation”, método que divide el conjunto de datos en K subconjuntos y realiza K iteraciones, entrenando el modelo en $K-1$ subconjuntos y validándolo en el conjunto restante. Se utilizó un $K=5$.
- **Optimización:** Se realizó el proceso de optimización de los modelos mediante el ajuste de hiperparámetros para mejorar el rendimiento de los modelos utilizado. Se hizo uso del método “**Randomized Search**”, este busca de manera aleatoria en un espacio predefinido de hiperparámetros para encontrar la combinación que produce el mejor rendimiento del modelo.

Resultados

El análisis muestra que los modelos **RandomForest** y **LGBM** presentan buenos resultados según LazyPredict. RandomForest tiene un mejor rendimiento con un RMSE=0.36 y $R^2=0.54$, en comparación con LGBM, que tiene RMSE=0.38 y $R^2=0.48$.

Al aplicar K-Fold cross-validation, ambos modelos presentan un ligero descenso en el rendimiento, lo que indica un posible sobreajuste en los modelos base. Finalmente, tras optimizar con Randomized Search, RandomForest mantiene su rendimiento con RMSE=0.36 y $R^2=0.54$, mientras que LGBM mejora significativamente con un RMSE=0.14 y $R^2=0.50$.

| | Simple | | K-Fold | | Randomized Search | |
|--------------|--------|------|--------|------|-------------------|------|
| Modelo | R2 | RMSE | R2 | RMSE | R2 | RMSE |
| RandomForest | 0.54 | 0.38 | 0.48 | 0.38 | 0.54 | 0.36 |
| LGBM | 0.48 | 0.38 | 0.42 | 0.40 | 0.50 | 0.14 |

Insights

Insights:

- * Los juegos con mayor cantidad de veces jugados tienden a ser mejor calificados.
- * La calidad de los juegos es percibida de manera diversa, especialmente entre aquellos con menos jugadas.
- * La cantidad de reseñas no siempre refleja la popularidad del juego, lo que indica que esta es influenciada por otros factores.
- * Algunos desarrolladores tienen una reputación constante de alta calidad, tanto en estudios grandes como independientes.

Conclusiones

En conclusión, se tiene una relación entre la calidad percibida por los usuarios y la popularidad de los juegos. Los juegos más jugados generalmente reciben calificaciones más altas, lo que indica que los jugadores suelen dedicar más tiempo a los juegos que consideran de alta calidad. Pese a esto, es importante destacar que la percepción de la calidad entre los usuarios varía significativamente. Esta diversidad es particularmente evidente en los juegos con menos jugadas, donde se pueden tener calificaciones más variables debido a los gustos y experiencias de los jugadores.

Por último, aunque la popularidad y la calidad están relacionadas, los gustos son subjetivos, lo que puede llegar a afectar la popularidad de un videojuego según como lo aprecie el usuario. Los desarrolladores que entregan regularmente juegos de alta calidad pueden lograr mayores ventas al sacar nuevos títulos, debido a que los usuarios confían en que estos serán de la calidad esperada. Esto demuestra la importancia de comprender las diversas expectativas de los jugadores y mantener altos estándares de calidad para lograr el éxito a largo plazo en el sector de los videojuegos.

Conclusiones

Se observa que RandomForest tuvo un rendimiento estable a lo largo de los análisis, con un RMSE de 0.36 y un R^2 de 0.54 tanto en los modelos base como después de la optimización con Randomized Search. La validación cruzada mostró una pequeña reducción en su rendimiento, lo que sugiere que el modelo es relativamente robusto y generaliza bien. Mientras que en el caso de LGBM se observa que aunque comenzó con un rendimiento ligeramente inferior, experimentó una mejora significativa después de la optimización, reduciendo su RMSE de 0.38 a 0.14. Esto indica que el modelo de LGBM es altamente sensible a la optimización de hiper parámetros y puede llegar a ser muy preciso. Sin embargo, su R^2 de 0.50, aunque cercano al de RandomForest, no logró superar su capacidad explicativa.

Por lo tanto, después de realizado el análisis, es posible concluir que para el estudio el mejor modelo corresponde a **RandomForestRegressor**, esto según los resultados de R^2 y RMSE en todas las formas aplicadas al modelo.