Proyecto final

Predicción de ventas de video juegos

Por Gabriel Delgado y Felipe Barreto

Contenido

[Abstracto 1](#_Toc136843596)

[EDA 1](#_Toc136843597)

[Dataset 2](#_Toc136843598)

[Imprecisiones 2](#_Toc136843599)

[Limpieza 2](#_Toc136843600)

[Transformación de datos 2](#_Toc136843601)

[Modelo de predicción 2](#_Toc136843602)

[Conclusiones 3](#_Toc136843603)

[Bibliografía 3](#_Toc136843604)

# Abstracto

Este trabajo se realizó durante todo el semestre 2023-1 por Gabriel Delgado y Felipe Barreto, la idea de este trabajo fue responder a la pregunta “¿Pueden las características de un video juego determinar si este será un éxito o fracaso antes de su lanzamiento?” creando un modelo aprendizaje supervisado que logrará predecir las ventas de un video juego a partir de sus características. Dicho esto, el resultado no fue el mejor, opinamos que con datos más relevantes podríamos llegar a modelos más precisos, sin embargo, al trabajar todo el semestre sobre este proyecto, aprendimos mucho sobre las herramientas de inteligencia artificial, tales como las librerías de Python o temas de Deep learning como las redes convolucionales. Hasta el momento, volviendo a nuestra pregunta, el modelo aparenta mostrar que las características no tienen una correlación fuerte con el desempeño del juego en el mercado.

# EDA

Antes de empezar esta sección, queríamos mencionar que los datos que encontramos tuvieron bastantes impresiones, mayor mente dado que los data sets eran inconsistentes, por lo cual revisamos varios antes de llegar al que terminamos usando.

## Dataset

Para este trabajo se uso un data set de Kaggle (Harsh, s.f.) que usaba datos de la pagina de ranking de video juegos VgChartz.

## Imprecisiones

Lo primero que llegamos fue que, es que el numero de ventas no siempre era consistente, es decir, a veces, las ventas globales no era la suma de todas las otras ventas, incluso, en el datasets que usamos originalmente ya se había modificado en esta columna, pero esta fue con un promedio dado que la mayoría de los valores eran nulos, por lo cual no lo usamos. Con referencia al dataset que escogimos, otro problema que tuvimos fue que se agruparon varios géneros de juegos en 1 categoría: Misc (misceláneo), en esta categoría se encontraron juegos como “Rock Band 2” y “Monopoly”, dado que eran demasiados datos, dejamos esta categoría como estaba, pero creemos que con algún script que analice el significado de las palabras podrías automatizar esta tarea y crear nuevas categorías, en este caso juegos de mesa y música.

## Limpieza

Se trataron los nulos y valores a típicos, tuvimos mayor problema con estos valores, dado que, fueron muy presentes en lo que queríamos predecir, cantidad de ventas, dado que el promedio era de 0.340568 millones, mientras que Wii Sports tiene 82 millones. Por lo tanto, borramos estos valores atípicos, por lo cual el modelo será mejor prediciendo juegos con ventas cerca de 0.3 millones.

Además, también eliminamos algunas columnas que no nos aportaban información relevante, por ejemplo, en este caso solo observábamos las ventas globales, por ende, las ventas por región no nos interesaron.

Finalmente, reducimos el atributo de fecha de lanzamiento a solo el año, dado que trabajar este como un valor no numérico resulto problemático por lo que había demasiadas fechas únicas.

## Transformación de datos

Dado que la mayoría de nuestros datos no eran numéricos, usamos la técnica de one hot encoding para que estos atributos pudieran ser usados. Usando estas técnicas, logramos observar la correlación entre las ventas y el genero o las ventas y el publicador, aunque estas no fueron las más altas (en el rango de 0.03 a 0.08) se logro encontrar algo. Dicho esto, también le restamos el año de publicación del juego al año actual, para conseguir el tiempo en el que a estado en el mercado y usar este nuevo dato en el modelo.

# Modelo de predicción

Para entrenar el modelo usamos los atributos nombre, genero, publicador, año, plataforma y rank, que es el rank que le da vgchartz al juego dependiendo de sus ventas, de esto nos enteramos al entrenar varios modelos y obtener un valor R2 de 1, por lo cual nos deshicimos de este atributo.

Dicho esto, usamos la librería de Pycaret para entrenar varios modelos y comparar cual seria el mejor. Usando esta estrategia llegamos a los modelos de Bayesian Ridge, Ridge Regression, Lasso Regression y Huber Regresión, sin embargo, ningún modelo llego a una precisión alta.

# Conclusiones

Al terminar el trabajo nos sentimos algo decepcionados, dado que no logramos llegar a una respuesta definitiva a nuestra pregunta, o crear un modelo con una buena predicción. Sin embargo, aprendimos mucho de este proyecto, gracias al amplio uso de herramientas y datos que usamos, desde explorar nuevas librerías de datos, nuevos modelos o recursos en páginas como Kaggle. Dicho esto, al observar los proyectos de nuestros compañeros, se nos ocurren algunas ideas de como podría mejorarse este modelo:

1. Creemos que podríamos encontrar una correlación alta entre búsqueda en el internet sobre un juego (o genero de juego) y sus ventas, dado que juegos exitosos suelen crear comunidades o discusiones sobre este antes de su lanzamiento usando tráilers, o al ser lanzados pueden tener alto cubrimiento en redes sociales, como youtubers cubriendo el juego o streamers jugándolo en vivo
2. Dicho esto, creemos que, a esta primera parte de nuestra proposición, se le debería añadir análisis de sentimiento, dado que, aunque es verdad que un buen juego genera alta cobertura en el internet, un juego polémico también lo hace, y este sentimiento podría afectar negativamente el desempeño del juego en el mercado

Algo que no llegamos tuvimos en cuenta fue que un juego podría aparecer varias veces dado que 1 juego se puede vender en diferentes consolas, esto se puede observar en el juego “007: Quantum of Solace”, que fue lanzado en 6 plataformas diferentes. Aunque estos datos nos ayudan a ver el desempeño de juegos en diferentes consolas, tal vez unificarnos en una observación nos daría una mejor idea del desempeño del juego.

Finalmente, el uso de algún algoritmo de reducción (como LDA o PCA) podría haber sido útil, dado que teníamos demasiados atributos gracias al uso de one-hot encoding. Usar algún algoritmo de reducción nos ayudaría a reducir los atributos encontrados, reduciendo el tiempo de procesamiento y el data leakage, lo cual podría otorgarnos mejores resultados en la predicción final.

En conclusión, creemos que nuestro modelo dejo cosas por desear, pero como un primer acercamiento al mundo de la inteligencia artificial nos pareció un buen trabajo, aunque creemos que le pudimos metido más esfuerzo, podemos ver opciones para mejorar este proyecto a futuro que hacen uso de herramientas interesantes.

# Bibliografía

Harsh, S. (s.f.). *Kaggle*. Obtenido de VgChartz: https://www.kaggle.com/datasets/harshsingh2209/vgchartz-games-sales-dataset