Video Games Sales

Eduardo Rodríguez Gil, ITC A01274913, Tecnológico de Monterrey

Este documento muestra la implementación y explicación del dataset de Video game sales.

I. INTRODUCCIÓN

Los Videojuegos son una enorme industria que ha ido creciendo a lo largo de los años. Donde año tras año su popularidad crece gracias a la gran variedad de juegos que se llegan a vender en el mercado. Todo esto es posible por la variedad de tipos de juegos que existen como los shooter, battle royal, open world, etc. Por esta razón veremos las ventas del mundo y de Norteamérica en millones, de algunas copias de los videojuegos más famosos que actualmente existen.

Figura 1. Video games.

II. DATA SET

El dataset lo obtuve de Kaggle, donde contiene un database de más de 15,000 videojuegos diferentes, donde están separados en diferentes categorías como Global_Sales, Na_Sales, etc. El dataset se redujo a 600 videojuegos aproximadamente para poder hacer uso de los datos más fácil.

	Global_Sales	NA_Sales
0	82.74	41.49
1	40.24	29.08
2	35.82	15.85
3	33.00	15.75
4	31.37	11.27

Figura 2. Data set de vgsales.

III. MODELO EN TRAIN, TEST Y VALIDATION

Para la realización del modelo separamos los datos en dos grupos. El primer grupo es el de Train model, donde se agregan varios datos para poder realizar las predicciones del modelo y la gráfica de regresión lineal (Figura 3).

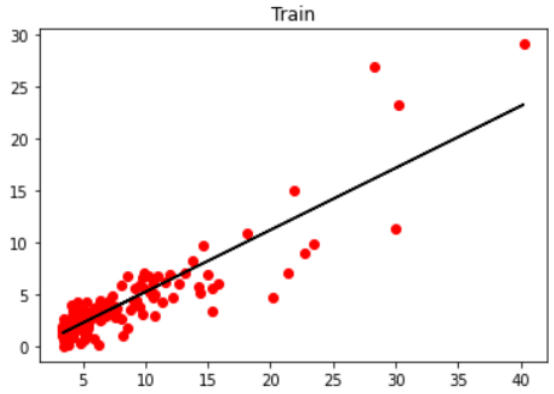

Figura 3. Train model.

Nuestro segundo modelo es el test model, donde de igual manera se agregan datos aleatorios para de igual forma poder realizar las predicciones del modelo y la gráfica de regresión lineal (Figura 4).

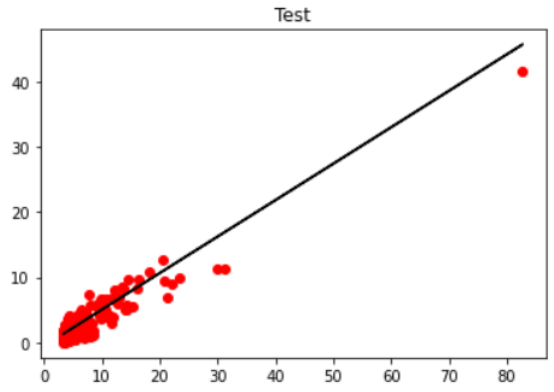

Figura 4. Test model.

Por último, nuestros últimos datos fueron los de validation que igual que los anteriores modelos tomábamos en muestra datos de manera aleatoria para la realización de la predicción y su respectiva gráfica (Figura 5).

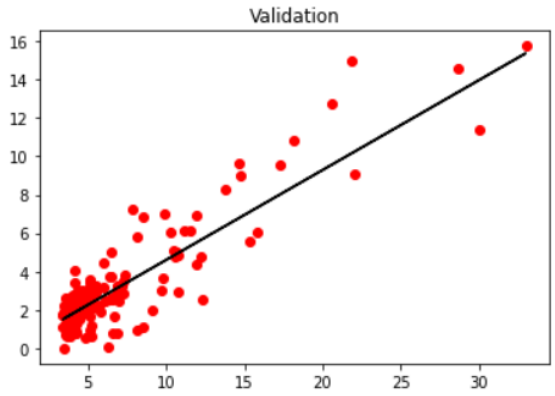

Figura 5. Validation model.

En la siguiente figura (Figura 6) podemos ver los datos que tomamos para cada una de las pruebas tanto de train, test y validation juntas en una misma gráfica.

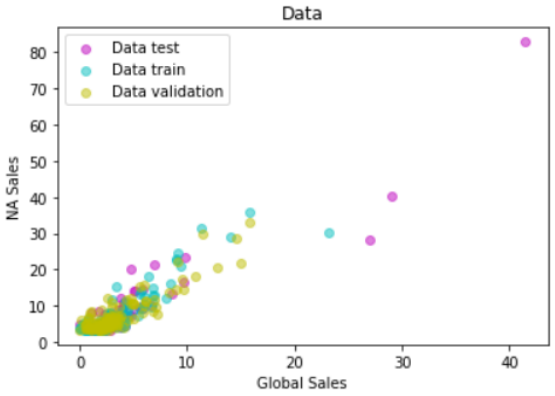

Figura 6. Data.

IV. BIAS O SESGOS

El bias con el que cuenta el modelo es bajo. Esto se debe a que los errores que llega a obtener el modelo tanto en train y en test son bajos. Diciéndonos que el porcentaje de predicción es muy alto como se puede observar en la (Figura 7), de cinco predicciones que se realizaron con datos aleatorios nos damos cuenta que nuestro modelo es bastante bueno, siendo asi 72% en train como la predicción más baja y 76% en test de las cinco veces que se realizó.

Train precision	Test precision	Validation precision
0.8116830377422517	0.8605875893203188	0.8947638261727263
0.8934909726564324	0.7817349892082548	0.7982102977098757
0.7828620079379314	0.879629943029435	0.9029446896274408
0.8980415083723439	0.7613453147837599	0.7461249584869412
0.724032890316858	0.882907823425596 5	0.845769241836191 3

Figura 7. Tabla de predicciones

V. VARIANZA

Para nuestro grada de varianza como podemos observar en la (Figura 7) para cada cambio de datos vemos que nunca son grandes los cambios. Siempre llega a cambiar la precisión un mínimo en train y test. Por lo mencionado anteriormente podemos deducir que nuestro grado de varianza es bajo.

VI. NIVEL DE AJUSTE

Volviendo a la (Figura 7) nos damos cuenta que en algunas ocasiones por un

mínimo porcentaje de error el train model es más bajo que el test model. Pero ya que los valores de nuestros datos mayormente se encuentran más óptimos podemos decir que el nivel de ajuste de nuestro modelo es óptimo.

VII. MEJORA DEL MODELO

Para poder lograr una buen desempeño del modelo fue necesario probar con diferentes tamaños del conjunto de entrenamiento, ya que en un inicio el porcentaje de predicción era menor a 50%. Empecé a ampliar el tamaño del conjunto de entrenamiento y me comencé a percatar que la predicciones que tenían eran cada vez mejores. Con esto podemos decir que entre más ampliemos el tamaño del conjunto de entrenamiento mejor va a ser nuestra predicción y por ende un mejor modelo.
También decidí agregar otra variable para las predicciones para poder tener una predicción más acertada, para lograr eso revise la correlación entre mis datos (Figura 8).

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.000000	-0.006218	-0.565873	-0.528241	-0.442389	-0.402248	-0.643929
Year	-0.006218	1.000000	-0.167779	0.175387	-0.238895	0.226980	-0.054607
NA_Sales	-0.565873	-0.167779	1.000000	0.640841	0.370595	0.461069	0.917341
EU_Sales	-0.528241	0.175387	0.640841	1.000000	0.331500	0.594572	0.847170
JP_Sales	-0.442389	-0.238895	0.370595	0.331500	1.000000	0.109502	0.547370
Other_Sales	-0.402248	0.226980	0.461069	0.594572	0.109502	1.000000	0.619082
Global_Sales	-0.643929	-0.054607	0.917341	0.847170	0.547370	0.619082	1.000000

Figura 8. Correlación entre los datos.

Podemos observar que Global_Sales y NA_Sales tienen una correlación aproximadamente de 92%, por dicha razón nuestro modelo tenía una buena predicción. Aunque de repente llegaba a tener algunas fallas. Para remediarlo agregue una segunda variable llamada EU_Sales, ya que su correlación es de aproximadamente 84%, con esto nuestras predicciones van ha estar aún más precisas como lo podemos observar en la (Figura 9).

Train precision	Test precision	Validation precision
0.81375466954581 02	0.85400273991053 68	0.80238937545892 51
0.81472905095867 76	0.86569704067217 85	0.80298108029727 58
0.80061176755968 98	0.81484515261906 34	0.87369316802107 32
0.81010751050148 46	0.86836669045703 6	0.80919271381061 59
0.85855785895458 44	0.81276106544856 44	0.81865721375034 07

Figura 9. Tabla de predicciones mejorada.

Nos damos cuenta cómo mejoraron aún más las predicciones al punto de estar muy igualadas y ninguna baja del 80% como anteriormente que se llegaban a tener predicciones de 70% y tenía una gran diferencia entre una y otra.

VIII. CONCLUSIONES

Es notorio que el modelo realiza buenas predicciones y cuenta con errores demasiado bajos, al grado de ser un modelo óptimo. Esto sucede gracias a la reducción de datos que se hizo en un inicio al dataset y a la mejora que se le fue realizando al modelo en el transcurso aumentando el tamaño del conjunto de entrenamiento. Aunque en algunos puntos todavía se podría hacer una limpieza más a fondo sobre el dataset y esperar aún mejores resultados.

IX. REFERENCIAS

[1] *Overfitting in Machine Learning - Javatpoint*. (2021). www.javatpoint.com. Recuperado 9 de septiembre de 2022, de https://www.javatpoint.com/overfitting-in-machine-learning#:%7E:text=%20What%20is%20Overfitting%3F%20%201%20Overfitting%20%26,test%2Funseen%20dataset%20and%20can%E2%80%99t%20generalize%20well.%20More%20

[2] Canadas, R. (2022, 29 mayo). *Qué es el bias en estadística y machine learning*. abdatum. Recuperado 9 de septiembre de 2022, de https://abdatum.com/tecnologia/que-es-bias

[3] Valdés, B. (2022). *Machine Learning*. Recuperado 9 de septiembre de 2022, de https://docs.google.com/document/d/1WOv6P6BzoFV0x5bN2PW_1MwP4JLdw_To/edit

[4] Smith, G. (2016, 26 octubre). *Video Game Sales*. Kaggle. Recuperado 9 de septiembre de 2022, de https://www.kaggle.com/datasets/gregorut/videogamesales