Predictor de vistas de videos de Juegos de mesa Youtube

Por Carlos Cárdenas Ruiz

***Resumen*—En este documento exploraremos hacienda el cruce de información de un sitio especializado y videos de YouTube para de esta forma poder generar un predictor de vistas utilizando la información recopilada.**

Palabras claves*—*YouTube, Juegos de mesa, vistas

I. Introducción

En la actualidad muchas personas crean canales de YouTube con la esperanza de tener ingresos extras o simplemente poder llegar a muchas personas, la idea de este artículo es poder cotejar la información de un sitio especializado (en este caso Board Game Geek) y utilizar los datos de clasificación de este sitio compararlo con las vistas de los videos de YouTube, así como la popularidad de los canales y ver encontrar una relación entre estos datos y las vistas.

II. Recolección de información y preprocesamiento

A. *Software utilizadopara extracción de datos*

Para poder extraer la información de forma directa de los sitios se utilizó como base Python debido a su versatilidad y gran cantidad de herramientas disponibles, las librerías extras utilizadas son los siguientes:

- Selenium – Esta librería nos permite tener conexión con cualquier web browser actual para de esta forma poder descargas la información de estos.
- BeautifulSoup – Nos permite tomar información en formatos HTML y XML para poder explorarlos utilizando el sistema de etiquetas que utilizan los mismos, y de esta forma poder extracción de información.
- Langdetect – Esta librería nos permite determinar el lenguaje de un texto para de esta forma poder saber el idioma del canal de videos.

B. *Datos extraidos*

La extracción de datos comenzó por buscar sitios dedicados a los juegos de mesa y de ellos extraer los videos d ellos diferentes juegos de mesa, se comenzó con la extracción de esta información, pero la identificación de a que juego pertenecía estos videos resultaba difícil y se decidió por otro enfoque.

El siguiente enfoque fue la extracción directa de la lista de juegos registrados en https://boardgamegeek.com/ de la cual se pudieron extraer la cantidad de 128,703 juegos diferentes y se extraen los siguientes datos:

- Lugar – Posición en la tabla de clasificación general del sitio
- Nombre – Nombre del juego
- Año de publicación
- Calificación comunidad
- Calificación promedio
- Cantidad de evaluadores
- URL del registro

Después de analizar la información a partir del registro 20,914 los datos de calificaciones ya no son consistentes, así como algunos datos duplicados después de eliminar registros no consistentes, se optó por una base de juegos de 10,151.

Con esta base de juegos se generan búsquedas en YouTube con la siguiente estructura:

https://www.youtube.com/results?search_query=board+game+nombre+Juego&sp=CAMSAhAB

La cual nos regresa búsquedas con respecto a los juegos aun así la búsqueda es muy general pudiendo darnos en algunos casos videos los cuales no son realmente juegos al respecto del juego en cuestión, para poder afrontar un poco el problema se agregó un factor de coincidencia el cual se explica a continuación.

- TP – corresponde al número total de palabras de las que consiste el nombre
- TO – este factor nos dice si el total de palabras de TP existen en el título del juego

De esta forma el factor de coincidencia es igual a:

Coincidencia = TO/TP en un rango de [0,1]

Un valor aceptable para para coincidencia es el siguiente:

$$\begin{cases} 1 & TP = 1 \\ TO - {}^{1}/_{TP} & TP > 1 \end{cases}$$

Figura 1 Criterio de aceptación

Al momento de escribir este texto la cantidad de juegos revisados es de 5529 y se han generado 406,148 registros de videos de los cuales se extrae la siguiente información:

- Nombre del video
- URL
- TP
- Vistas
- Fecha publicación
- Coincidencia
- URL del canal

Como se describe en la información recolectada se encuentra información que respecta al canal, ya que un canal con muchos suscriptores tiene una mayor difusión e influye en la difusión del video.

También al tener ya una gran cantidad de registros también se tiene una relación de aparición en lo que respecta a los canales presentando una cantidad de 56,303 canales diferentes, también podemos ver la aparición que hay de videos de cada uno de estos canales, así como videos de diferentes juegos ya que pueden existir canales con muchas apariciones, pero ser de únicamente un juego, esto nos quiere decir que este canal es solo sobre un juego en específico como por ejemplo ajedrez y para este estudio solo nos interesan canales que hablen sobre muchos juegos de mesa diferentes.

Con esta información y utilizara como discriminador aceptar videos de canales los cuales por lo menos tengan 50 apariciones y al menos cuenten con 30 juegos diferentes de esta forma podremos disminuir la totalidad de registros y asegurar que estos son sobre juegos de mesa.

Después del filtrado anterior la cantidad de registros de canales analizados al final es de 1,279 canales y esto nos permitió disminuir la cantidad de videos analizados bajo al número de 87,287.

III. Conjunto de información completo

Con la recaudación de información y discriminación de videos por fin tenemos un conjunto de datos el cual contara con la información necesaria y poder proceder a la selección de características para poder entrenar el modelo.

Nuestro set completo de datos contara con las siguientes características.

Nombre Campo	Tipo
juegoNombre	Cualitativo
juegoURL	Cualitativo
juegoAno	Numérico
juegoPalabras	Numérico
juegoLugar	Numérico
juegoRating	Numérico
juegoAvRating	Numérico
juegoVoters	Numérico
videoNombre	Cualitativo
videoMeses	Numérico
videoURL	Cualitativo
canalMeses	Numérico
canalNombre	Cualitativo
canalURL	Cualitativo
canalvistas	Numérico
canalSuscriptores	Numérico
canalIdioma	Cualitativo

Tabla 1 Datos videos completos

A. *Variable objetivo*

La variable objetivo para este caso es las vistas de cada video ya que tratamos de predecir las vistas que nos daría un video para de esta forma poder elegir los videos que nos generen más vistas y por lo tanto tener más ingresos de estos.

B. *Selección de caracteristicas y tratamiento*

Inicialmente se eliminarán características que de antemano no aportan información al modelo por sus características como, por ejemplo, todas las que tienen que ver con URL y nombres específicos como de canales.

Para discriminar las variables que posteriormente, utilizaremos para variables numéricas el análisis de correlación el cual nos arroja los siguientes datos.

Variable	Correlación
videoVistas	1.000000
canalVistas	0.678406
canalSuscriptores	0.562101
videoMeses	0.033045
canalMeses	0.027702
juegoVoters	0.020725
juegoRating	-0.004481
juegoAvRating	-0.022976
juegoAno	-0.026716

Tabla 2 Correlación variable objetivo videoVistas

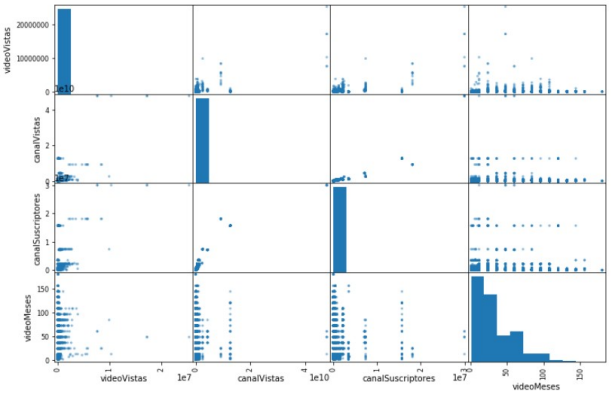


Figure 2 Matriz de dispersion

Como podemos ver en lo que respecta a las variables numéricas solo CanalVistas y canalSuscriptores cuentan con un factor de correlación lo suficientemente alto para ser tomadas en cuenta en la predicción, como recordaremos la correlación se encuentra en un rango de [-1,1] y se calcula con la siguiente formula:

$$corr\,(x,y) = \frac{cov\,(x,y)}{SD\,(x) * SD\,(y)}$$

Figura 3 Formula correlación

A continuación, veamos el mapa de calor de variables en pares para verificar la independencia de estas.

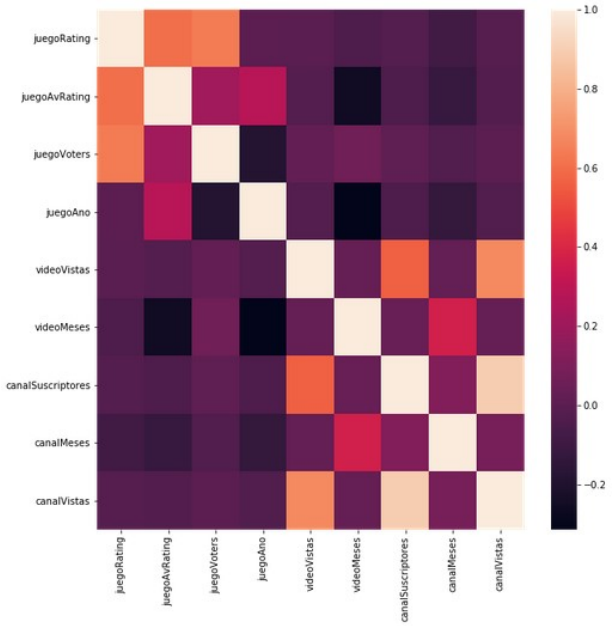


Figura 4 Mapa de calor

En este análisis Podemos ver que solo las variables juegoRating y juegoAvRating tienen una dependencia lineal pero no realmente fuerte, así que podríamos utilizarla.

Las variables categóricas son solo dos JuegoNombre y canalIdioma para poder utilizar es estas variables se crearon variables dummies, las cuales nos genera una matriz esparza de tipo binario donde tendrá 1 si se tiene la característica de la variable, en este caso si pertenece al juego o al idioma al que hace referencia, aún así estas variables no se utilizaran debió al bajo impacto que se tiene en la predicción.

Finalmente, antes de ingresar los datos numéricos a nuestro modelo se optó por estandarizar los datos para evitar un efecto adverso por datos extremos

C. Modelo utilizado

Inicialmente utilizaremos una regresión lineal simple, después regularización con Ridge, arboles de regresión y finalmente debido al riesgo latente de los árboles de regresión al sobre entrenamiento (ovefitting) para de esta forma verificar los resultados de la predicción de cada modelo utilizando la métrica siguiente.

El análisis de F-Anova nos muestra que las variables cualitativas tienen una influencia demasiado baja en el resultado, por este motivo las variables del nombre del juego y el idioma del canal serán descartados ya que el aporte que generan no es lo suficiente para ser tomadas en cuenta, con esto nuestro análisis será totalmente sobre variables numéricas.

Antes de proceder con el análisis de cada modelo, se mostrará una gráfica de dispersión la cual se encuentra en la figura 5, en el que se encuentran las dos variables con mayor poder de predicción, esto con la finalidad de mostrar como se agrupan los datos y dar una intuición al lector del porque, unos modelos dan un mejor resultado que otros.

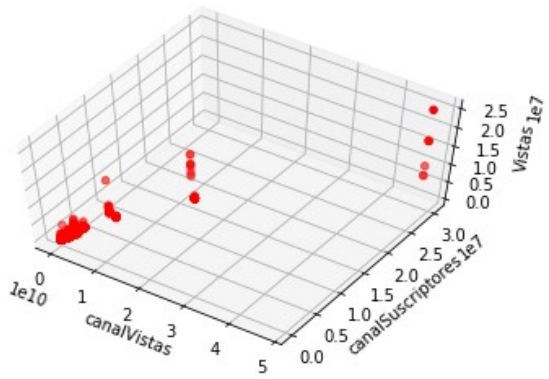


Figura 5 gráfica de dispersión general

Para hacer una calificación de los modelos se utilizará una la siguiente formula que nos retorna el coeficiente de predicción R^2 :

$$R^2 = 1-\frac{u}{v}$$

Donde:

$$u = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$v = \sum_{i=1}^{n}(y_i-\bar{y})^2$$

1) Regresión Lineal

Como primer modelo se utilizo una regresión lineal simple como podemos observar en la gráfica de dispersión de la figura 5, la disposición de los datos no se puede expresar perfectamente como una línea, el coeficiente de predicción R^2 para este caso es 0.3722, y predicciones un valor de R^2 de 0.4922, lo cual como se menciono anteriormente es debido a la representación de los datos en la gráfica de dispersión.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$$

Figura 6 Ecuaciones regresión lineal y función de costo

2) Regresión Ridge

Al igual que en la que con la regresión lineal simple, la distribución de los datos no puede ser representada correctamente por una línea y aún peor al distribuirse los

datos mayormente a la parte superior del eje Z, podemos darnos cuenta de que al regularizar la regresión los resultados empeoran a medida que se aumenta λ, se comenzó dándole valores a λ desde 1 pero solo al ser 0 se igualaba a regresión lineal claro sabemos que esto es equivalen debido a la formulación de la función de costo de la regresión.

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda\sum_{j=0}^{p}\beta_j^2$$

Figura 7 función de costo regularización

El valor del coeficiente de predicción R^2 no difiere mucho del presentado en la regresión lineal, claro esto depende en gran medida de λ la cual no asigna la penalización de la información.

3) Arbol de regresión

Como pudimos ver en la figura 5 la disposición de los datos con relación a las 2 variables con mayor poder de predicción hace posible que un árbol de regresión tenga un poder de predicción arriba del 0.91 dentro del conjunto de prueba, ya que al tener la posibilidad de segmentar de manera binaria los datos con base en las diferentes variables permite una predicción bastante alta, los parámetros utilizados son los siguientes:

- Función de costo: MSE
- La profundidad del árbol: 10

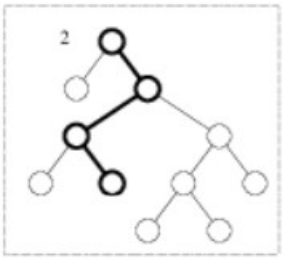


Figura 8 Arbol de regresión ejemplo

Se elaboraron varias pruebas y la profundidad de 10 ofrecía un valor con poco Bias y Variación aceptable con valores en el entrenamiento de 0.9226 y en el set de prueba de 0.8410 para el coeficiente de predicción R^2 lo cual nos muestra que existe algo de sobre entrenamiento.

4) Random Forest

Debido a que algunas veces el uso de Arboles de regresión pueden tender al sobre entrenamiento, se utilizaran los Random Forest para tratar de mejorar el desempeño mostrado por los arboles de regresión, como sabemos los arboles de regresión crean diferente arboles usando diferentes conjuntos datos del set de entrenamiento permitiendo duplicidad de datos, así como diferentes combinaciones de las características, con esto se podrán generar diferentes arboles los cuales al predecir el resultado se tomara en cuenta un promedio de los mimos, y con esto evitar sobre entrenamiento en del modelo.

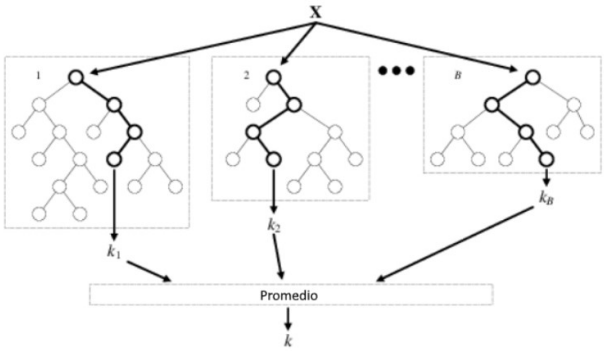


Figura 9 Random Forest ejemplo

Los parámetros utilizados para este análisis son los siguientes:

- Profundidad: 10
- Estimadores: 4

Los resultados fueron de 0.9105 para el entrenamiento y de 0.8469 para la prueba.

5) Comparación de resultados

En la siguiente tabla se encuentran los resultados del coeficiente de predicción R^2 para los diferentes métodos usados en la predicción, estos resultados son los que se calcularon después del entrenamiento con relación al conjunto de entrenamiento y prueba.

Modelo	Entrenamiento	Prueba
Regresión lineal	0.3722	0.4922
Ridge	0.3722	0.4922
Árbol regresión	0.9226	0.8410
RF	0.9105	0.8469

Tabla 3 comparación de resultados modelos

Como podemos ver finalmente RF mejor en un poco el resultado obtenido por el Árbol de regresión aún que la diferencia es mínima, por ultimo se hizo el análisis de Validación cruza que se presenta a continuación.

6) Validación cruzada

La validación cruzada nos permite verificar que las predicciones basadas en el entrenamiento y su posterior verificación contra el conjunto de prueba no se deben solamente a una selección fortuita de la división seleccionada, de esta forma se seleccionan diferentes particiones y se busca encontrar las diferentes calificaciones para los mismos y con esto poder generar una media los datos para este análisis.

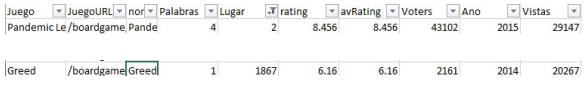
Los resultados son los siguientes y se encuentran en la tabla siguiente:

Modelo	Promedio Validación cruzada
RL	0.0834
Ridge	0.0834
AR	0.5001
RF	0.4931

Tabla 4 Validación cruzada

IV. Conclusiones

El análisis sobre los datos recopilados nos llevó a la conclusión que en realidad las vistas de un video en particular no depende en gran medida del juego al que haga referencia, lo más importante tiene que ver con el tamaño del canal que quiere decir esto, aunque existan juegos que en las bases de datos de Board Game Geek no cuenten con una calificación similar, por ejemplo uno con una calificación muy alta y otro con una calificación muy baja las vistas serán similares por lo cual lo realmente importante es el tamaño del canal que se traduce en la cantidad de suscriptores y vistas globales.



Juego	JuegoURL	nor	Palabras	Lugar	rating	avRating	Voters	Ano	Vistas
Pandemic Le	/boardgame	Pande	4	2	8.456	8.456	43102	2015	29147
Greed	/boardgame	Greed	1	1867	6.16	6.16	2161	2014	20267

Cada modelo mostrado anterior mente nos permite ver que dependiendo del problema nos darán mejor o peor resultado y también depende mucho los hiper parámetros utilizados, esto es más que nada una invitación a la prueba de muchos más modelos y técnicas que permitan mejorar las predicciones planteadas a cada problema.

V. Futuros trabajos

Para futuros trabajos el uso de series de tiempo en relación con las vistas de videos específicos estratificados con respecto al tamaño de los diferentes canales podría darnos una mejor idea en cuestión de la predicción de vistas generadas por un video a corto, mediano y largo plazo.

VI. Referencias

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* . Secaucus: Springer-Verlag New York.

Board Game geek. (2021). *Board Game geek*. Obtenido de https://boardgamegeek.com/

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn & TensorFlow.* Sebastopol: O'REILLY.

ICHI. (s.f.). *ICHI.PRO*. Obtenido de https://ichi.pro/es/por-que-random-forest-es-el-mejor-152703913514341

Selenium. (2021). *Selenium*. Obtenido de https://www.selenium.dev/

YouTube. (2021). *YouTube*. Obtenido de http://www.youtube.com