

The background is a blurred image of a financial market display. It features various stock indices such as OMXC 25, OMXRGI, OMXIC 8, and OMXI8, along with their respective values and percentage changes. A line chart is visible in the center, showing a sharp upward trend followed by a decline. The overall color scheme is dark blue and black with red and green highlights for price movements.

Análisis de información de Youtube para US

Cristian Hidalgo

CONTENIDO

- Contexto
- Objetivos
- Metodología
- Análisis exploratorios de datos
- Nuevas variables
- Modelado de los datos
- Conclusiones
- Recomendaciones y pasos a seguir

Contexto

- + YouTube publica diariamente una lista de los videos más populares según interacciones (vistas, compartidos, comentarios y likes), no necesariamente los más vistos del año.
- + Este conjunto de datos recopila información estructurada de videos en tendencia durante varios meses en EE. UU (nuestro caso).
- + Incluye hasta 200 videos diarios, con detalles como título, canal, fecha de publicación, etiquetas, vistas, likes, dislikes, descripción y comentarios.
- + Es una herramienta clave para analizar tendencias globales en contenido de YouTube.

Objetivos









Preguntas de investigación

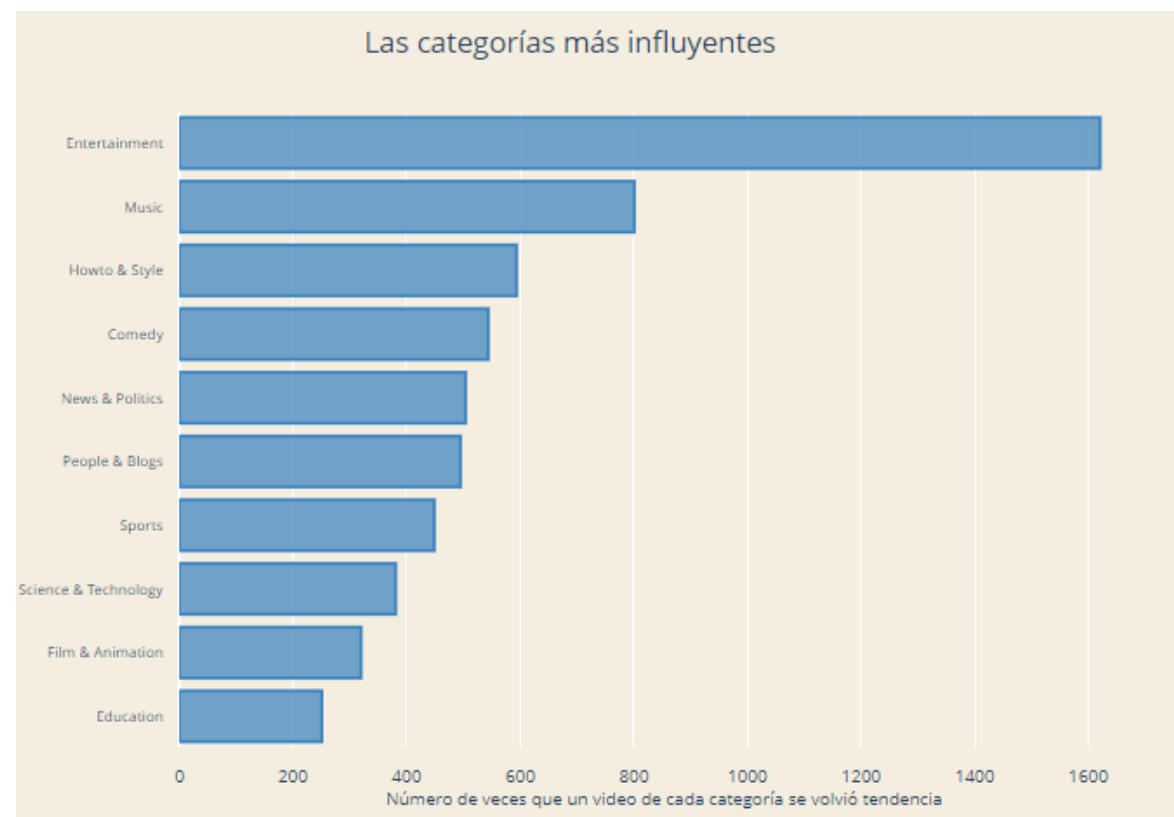
- + ¿Qué variables influyen en que un video sea tendencia?
- + ¿Se pueden definir relaciones entre el número de días en tendencia y otras variables como categorías?
- + ¿Es posible generar un modelo de predicción de likes?

Objetivos

- + Realizar un análisis descriptivo de los datos con el fin de encontrar relaciones de interés sobre los días en tendencia.
- + Desarrollar un modelo predictivo de los likes que un video puede llegar a tener.

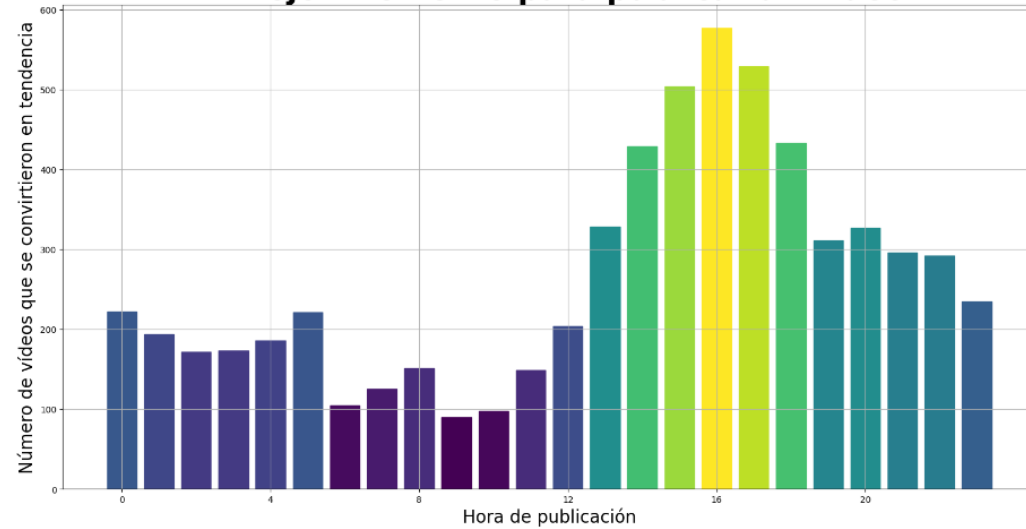
Análisis exploratorio de datos (EDA) - 1

	MickMake	#184 Making a PCB using EasyEDA. // Review	Science & Technology	2017-12-02 14:05:07
	Quad9 DNS	Quad9 How To Install with Windows	People & Blogs	2017-11-16 01:56:43
	Davie504	QUADRUPLE NECK BASS SOLO	Entertainment	2018-03-18 15:00:02
	Cleveland Browns	QB Luke Falk: Need to show you're a fearless playe...	Sports	2018-01-24 18:00:09
	Simon's Cat	Purrthday Cake (A 10th Birthday Special) - Simon's...	Pets & Animals	2018-03-04 09:00:01
	AnthonyPadilla	Puppy Bowl in Real Life (Way too many puppies)	Comedy	2018-01-29 16:56:47
	Animal Planet	Puppy Bowl Spot Center: Top 5 Plays	Entertainment	2018-02-03 23:00:04
	Hydraulic Press Channel	Punching Huge Holes Through Everything with Hydrau...	Science & Technology	2018-02-17 14:52:06

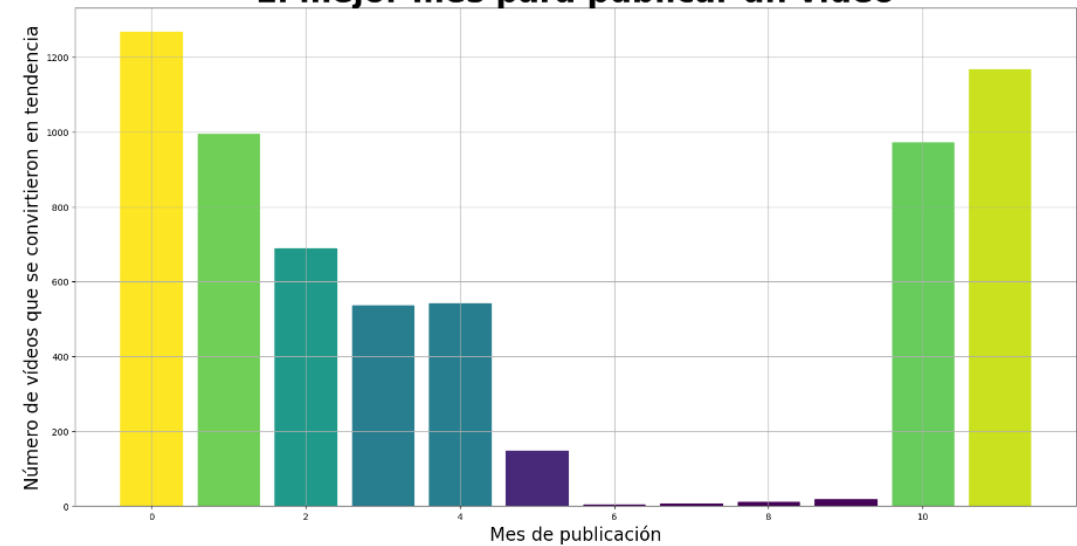


Análisis exploratorio de datos (EDA) - 2

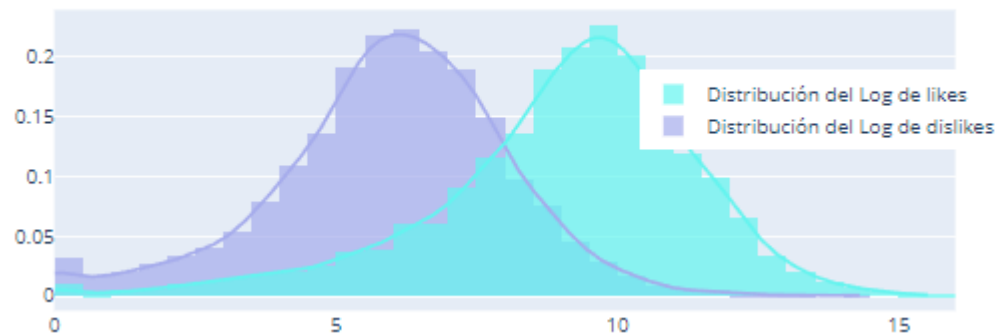
El mejor momento para publicar un vídeo



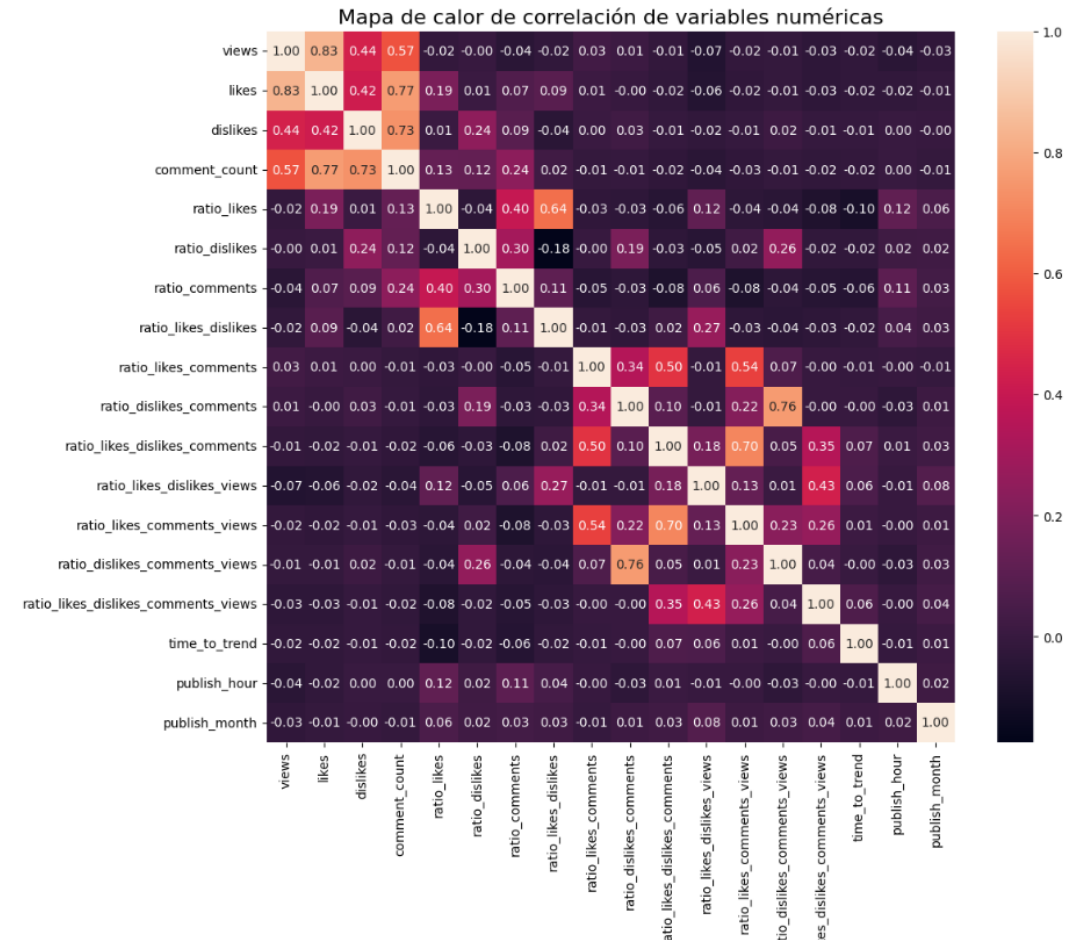
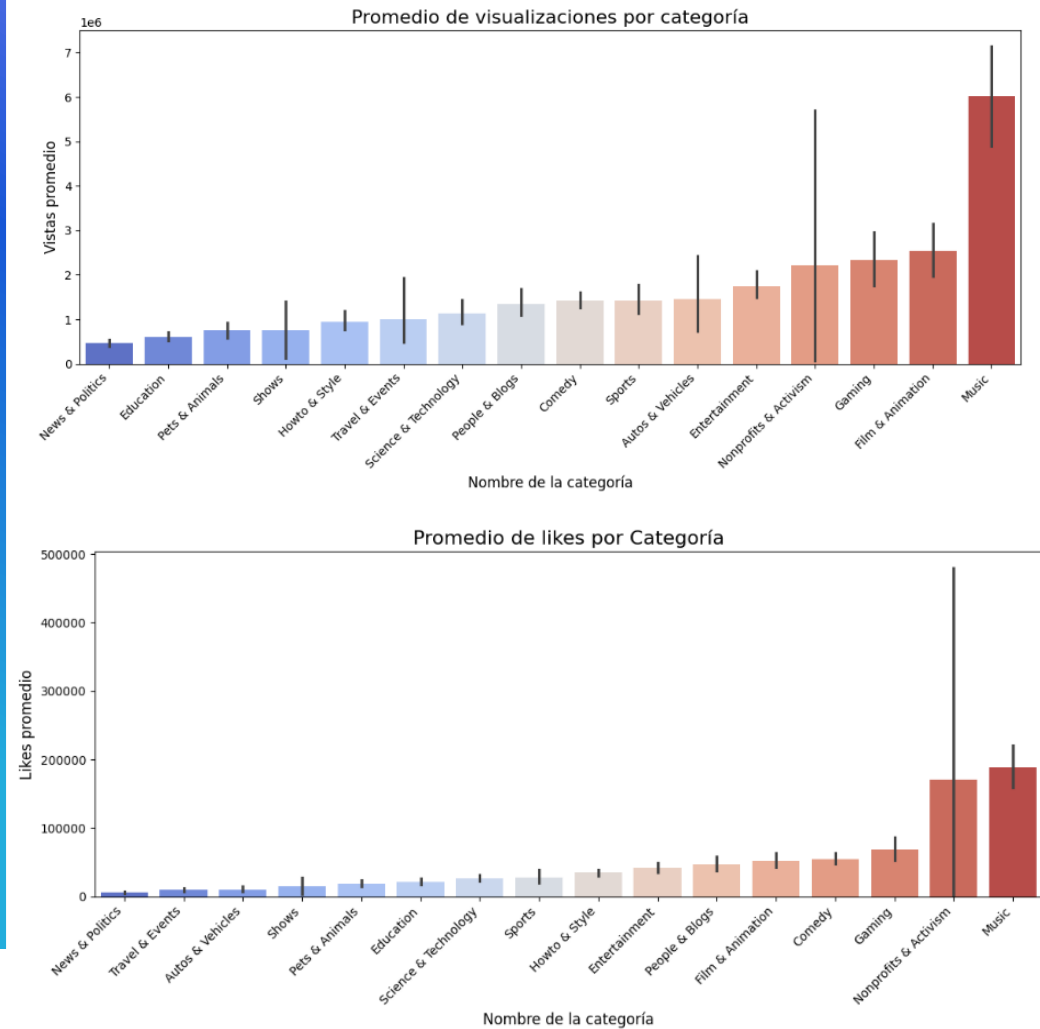
El mejor mes para publicar un vídeo



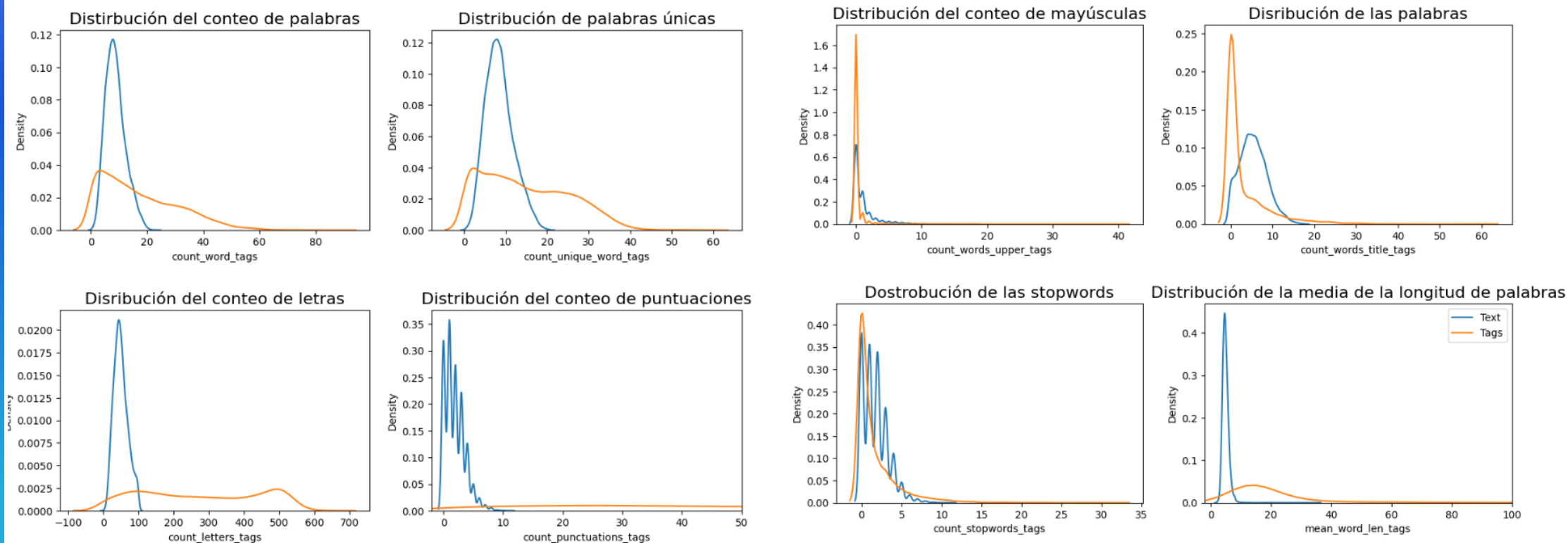
Likes vs dislikes



Análisis exploratorio de datos (EDA) - 3



Creación de nuevas variables -1



Creación de nuevas variables -2

Nube de palabras - Títulos



Nube de palabras - Descripton



Nube de palabras - Tags



Modelado de datos

Información: XGBoost

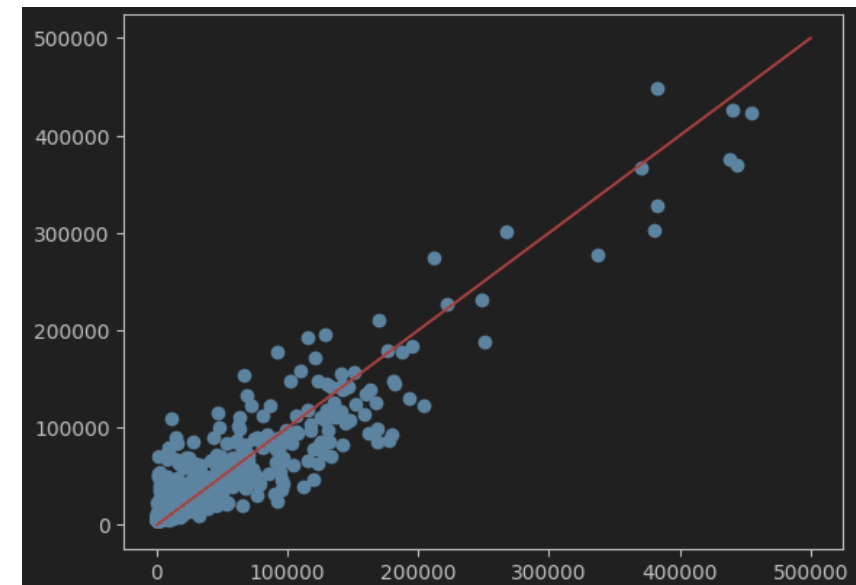
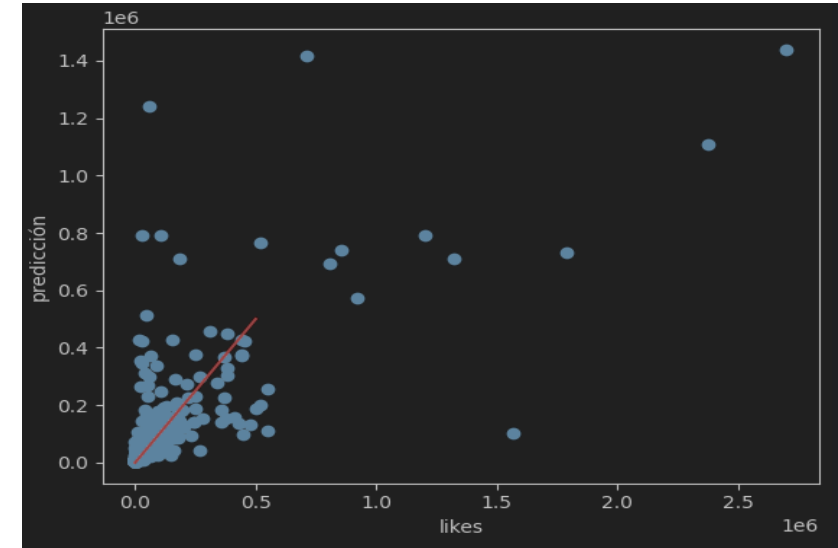
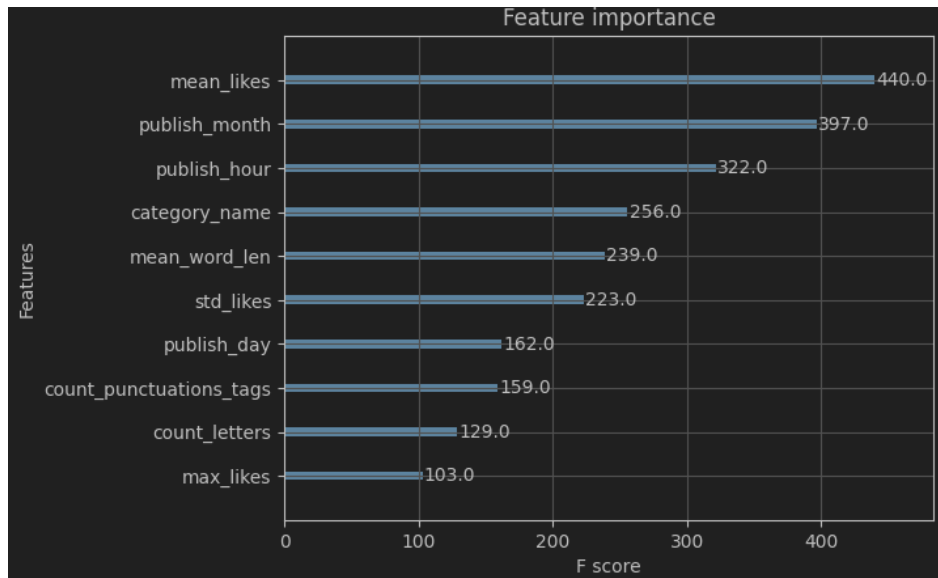
Train: 70% - **Test:** 15% - **Val:** 15%

Channel Name: Mean, std, median encoding

R2: 0,54 - **MAE:** 33206

R2 (whit out outliers): 0,85

MAE (whit out outliers): 12864



Conclusiones

- + La categoría, la hora de publicación, el mes, son variables que influyen a la hora de saber si un video será tendencia. Enero y febrero y entre las 2 y las 6 pm son los picos de los videos que se han vuelto virales.
- + Todas las categorías, excepto *nonprofits & activism* tienden a ser más homogéneas en su comportamiento.
- + Existe una alta relación entre *likes, dislikes, comment count* y *views*. Es decir que se podría estimar una a partir de las otras si fuera el interés.
- + Es posible predecir los likes y variables como canal, mes, hora, categoría, promedio de palabras en el título, son importantes para estimar esta variable. Siendo el promedio histórico de cada canal la de mayor relevancia.
- + Los valores atípicos pueden distorsionar la efectividad del modelo.

Recomendaciones

- + Es recomendable realizar una segmentación a nivel de canal (o video) con el fin de detectar aquellos atípicos y su agrupación, esto con fines de entendimiento, campañas, etc., ya que asignar una nueva observación sin histórico resulta complejo sin las variables que componen la segmentación.
- + Se puede realizar más de una segmentación con fines diferentes: Tipos de videos en cuanto a variables numéricas (vistas, likes, dislikes, etc.) y otra respecto a variables de texto (tags, títulos, descripciones).