MEMORIA PROYECTO:

Visualizaciones de los canales de videojuegos en la plataforma Twitch

REALIZADO POR ALVARO JAÉN ARIAS

BOOTCAMP DATA SCIENCE – THE BRIDGE

INDICE:

Proyecto EDA - Anàlisis Exploratorio de Datos

Proyecto ML: Machine Learning

INTRODUCCION:

El principal motivo de esta elección como tema de mi EDA es debido al aumento en el interés, seguimiento y uso de la plataforma Twitch desde el inicio de la pandemia por COVID-19.

Desde los 14 años llevo jugando a videojuegos con mis amigos, después de 20 años habiendo pasado por los juegos de consolas: Nintendo, Super Nintendo o Nintendo 64, como juegos de ordenador desde un jugador, multijugador o juegos online masivos, ha habido un cambio en el concepto del seguimiento de videojuegos a nivel mundial.

Hace 10 años había eventos en las ciudades dónde se jugaban los torneos e iban los jugadores más "Top" a nivel mundial que solo lo conocías si eran muy fan de esos juegos.

Desde hace pocos años este "seguimiento" lo tienes a un "click" desde tu casa. Con una variedad de juegos infinita puedes ver desde canales oficiales con presentadores a personas jugando desde su casa, siendo el mejor del mundo o tu vecino de enfrente.

El aumento de esta plataforma ha sido tan exitosa que ya es conocida mundialmente en personas de todas las edades, ya que es el "sustituto" de la televisión junto a otras plataformas de series y peliculas.

Dado que el tema a elegir era abierto, he decido hacerlo por primera vez en mi vida sobre mi hobby con el que disfruto diariamente.

HIPÓTESIS

La hipótesis principal que me llevo haciendo hace un par de años es si el concepto de las visualizaciones a mundiales de juegos en Twitch puedes ser comparable con eventos deportivos como finales de champións o finales de mundiales de futbol. La lógica me dice que no pero vamos a ver que resultado obtenemos.

Durante el comienzo de este EDA y antes de sacar los datos, se me han ocurido otras dos hipótesis secundarias acerca de esta plataforma como:

- Cuáles han sido los juegos más seguidos cada año.
- Las visualizaciones, han aumento en el tiempo o no.

Por último he sacado "hipótesis" terciarias tras obtener los datos como:

- Qué tipo de juegos son los más seguidos. Definición siglas tipo de juegos:
    - Shooter: Género de acción donde el principal objetivo es disparar y matar enemigos, generalmente con armas de fuego.
    - MOBA: Abreviatura de Multiplayer Online Battle Arena, es un subgénero de videojuegos de estrategia en tiempo real en el que dos

equipos de jugadores, normalmente 5 jugadores por equipo compiten entre sí en un campo de Batalla predeterminada.

- MMO: Los videojuegos de rol multijugador masivos en línea o MMORPG, son videojuegos de rol que permiten a miles de jugadores introducirse en un mundo virtual de forma simultánea a través de internet e interactuar entre ellos.

- Hay variación de visualizaciones según el mes del año.

PREPARACIÓN DATOS:

Para obtener los datos comencé por dos vías:

1. La primera era obtener los datos directamente desde Twitch a través de su API. Viendo la dificultad y la cantidad de horas dedicado a ello sin avanzar, investigué la segunda opción siendo la elegida finalmente.
2. Base de datos de Kaggle. Encontré varias bases de datos, que tras revisarlas enteras, vi que eran todas incompletas o no las entendía para poder analizarlas exceptuando una base de datos de 17.000 filas y con unos datos muy limpios y claros.

Esta base de datos se trata de la recopilación de datos desde el año 2016 hasta principio de 2023 de los 200 primeros canales que hay en Twitch de videojuegos seleccionados cada mes y cada año. Filtré los datos que me interesaban en función de mis hipótesis con los conocimientos aprendidos en clase hasta el día de hoy.

REPARTO DE TIEMPO DEDICADO AL PROYECTO

Aunque haya dedicado muchas horas (entre 6-8 horas) al intentar entender la API de Twitch o buscar bases de datos, el resto del trabajo se ha basado en dos bloques:

A. Tratamiento de datos: Por suerte el CSV tenía unos datos bastante limpio, por ello el orden del tratamiento ha sido:
   a. Existencia de simbolos junto a los datos que impedián su lectura. Por ello he hecho varios replaces para quitar→ “”,%,(),[],{},etc.
   b. Eliminar las columnas que no me interesaban.
   c. Al estar los datos por los 200 juegos más vistos al mes, ordenados por meses 1-12 y por años 2016-2023, tenía que agrupar por año/meses según me interesaba para la hipótesis .

B. Streamlit: Personalmente me llama la atención y me gusta el tema de "Visualizaciones". Por ello he dedicado casi la mitad del tiempo a hacer una buena presentación en Streamlit. Me he apoyado "Estructuralmente" en el EDA de mis Assistant Teachers Alejandro Cárabe Arranza y Jhon Alejandro Montecaleano Forero en las partes que me gustaban y mirando principalmente en la documentación oficial de Streamlit y Lottie para Gift e iconos.

CONCLUSIONES

Hipótesis principal:

Tras la recopilación de datos y tratamiento de los mismos, puedo comprobar que los picos de espectadores va en relación a un canal de televisión u otro tipo de canal, en el caso de Twítch, por lo que no puedo sacar el pico de espectadores a nivel mundial. He tenido que redirigir mi hipótesis y compararlo con eventos futbolísticos por canales de televisiones nacionales. Con la ayuda de las gràficas podemos comprobar que los espectadores del principal evento de los E-Sport no es comparable con la estimación de espectadores a nivel mundial ya que solo es comparable a nivel nacional. Por ejemplo con el pico de espectadores que tuvo TVE o telecincos en los mundiales de 2010 o 2022, pero muy de lejos con el pico de espectadores del principal canal Francés que casi dobla los números de visualización.

Hipótesis secundarias:

Podemos ver que la visualizaciones durante estos 7 años han ido creciendo anualmente, con una mayor subida muy notable durante los años de pandemia ya que estabamos en casa durante varios meses y por consiguiente más tiempo para ver Twitch.

En cuanto a los juegos más seguidos en estos 7 años, están muy estables estos juegos en el top 10, viendo claramente un juego ganador que es el League of Legends.

Hipótesis surgidas durante el EDA:

Por clara victoria, el tipo de juego más seguido son los tipo "Shooter", liderados por Fortnite, Counter Strike, Call of Duty y Valorant. En segundo lugar es MOBA por el liderazgo de League of Legends en el top 1 durante estos años.

Por último, las visualizaciones por mes, podemos ver que están estables exceptuando en el mes de enero que es ligeramente superior. Este hecho es debido posiblemente a mi parecer ,a que es invierno, tras navidades, aumentando la probabilidad que más

personas se queden en casa en lugar de salir y se pongan a ver los canales de sus juegos favoritos.

BIBLIOTECAS UTILIZADAS

- Pandas
- Numpy
- Streamlit
- Streamlit Lottie
- Json
- Plotly.express
- Plotly.graph
- Plotly.offline

INTRODUCCIÒN

Aprovechando el EDA realizado sobre este base de datos original con el fin de estudiar las visualizaciones en Twitch a lo largo del tiempo, he decido hacer mi proyecto de Machine Learning sobre ello. De este modo puedo tener un proyecto muy completo y sacar unas predicciones relacionadas a una idea de negocio bastante acertadas.

HIPOTÈSIS

Pongamos que somos una empresa y queremos poner publicidad en canales de videjuegos en Twitch porque pensamos que tiene un pùblico màs cercano a los productos que vendemos.

Para saber què canales de videojuegos son los mejores para nuestro propòsito, ponemos las siguientes premisas:

1- Estos canales de videojuegos deben ser los màs vists (Rank top 200 a nivel mundial en funciòn de horas vistas totales.
2- Debe seguir a un año vista en este top 200.
3- Debe crecer, y nos interesa que sea lo màximo. Cuantas màs gente estè viendo los canales de este videojuego, a màs pùblico llegarà.

PREPARACIÒN DE LOS DATOS Y DESARROLLO

Cogemos el dataset ya limpio tras el EDA y estructuramos el trabajo:

CLASIFICACIÒN

1- Creaciòn columna sintètica binaria 1/0 en funciòn de si los juegos estàn presente el mes(1-13) del año(2016-2021) y estàn tambièn presente ese mismo mes pero a un año vista. Por ejemplo, juego (x), si està presente en enero de 20216 y en enero de 2017 →(1), si no està presente, →(0). Estos datos van añadiendose a una lista vacìa y luego està lista se concatenarà como una nueva columna al dataset con el que vamos a trabajar.
2- Pretratamiento de datos:
    a. Eliminamos columnas que no nos interesan por sacar los datos a partir de otras y barajamos las filas con “Shuffle”.
    b. Separamos x_train, x_test, y_train,y_test con 0,2 para test y un Random State, luego estandarizamos los datos.

3- Creaciòn modelos de Clasificaciòn. De los siguientes modelos de clasificaciòn: (Regresiòn Logìstica, Super Vector Soporte, Arbol de decisiòn, Random Forest, Voting hard y soft, Bagging, Adaboost, Gradient Boost y XG Boost) sacamos los siguientes tres mètricas de calidad (Accuracy, Recall y Precisiòn) y los ordenamos de mayor a menor.

   Tras ver los dato que tenemos, lo que nos interesa es que nosotros como empresa de publicidad no invirtamos en canales de un videjuego que bajen en las visualizaciones, por lo que queremos asegurarnos de coger 1 en la clase binaria. Por ello escogemos el modelo con mayor Precisiòn → Gradient Boost Classifier.

REGRESION

1- Creaciòn columna sintètica: “Previsión Hours Watched a 1 año”. Del data set con la columna binaria, cogemos solo los (1). Los valores de esta columna serà el valor de Hours Watched del año siguiente. Por ejemplo, enero de 2016 pondremos el valor de enero de 2017 para ir viendo su crecimiento.
2- Pretratamiento de datos. Con una matriz de correlaciòn, eliminamos columnas que tienen colinealidad, separamos x_train, x_test, y_train,y_test con 0,2 para test y un Random State y por ùltimo hacemos una estandarizaciòn.
3- Creaciòn modelos de Regresiòn. De los siguientes modelos de regresiòn: (Regresiòn lineal, RL con Lasso, RL con Ridge, RL con Elastic, Arbol de Regresiòn, Gradient Boost Regressor y XGBoost Regressor), con el Score R2, el mejor modelo es→ XG Boost Regressor.

DATOS FINALES

Aplicar el modelo de regresiòn sobre mi muestra de datos de enero de 2023 para predecir las Horas totales visualizadas en enero de 2024 y cuales son los juegos que màs creceràn en %.