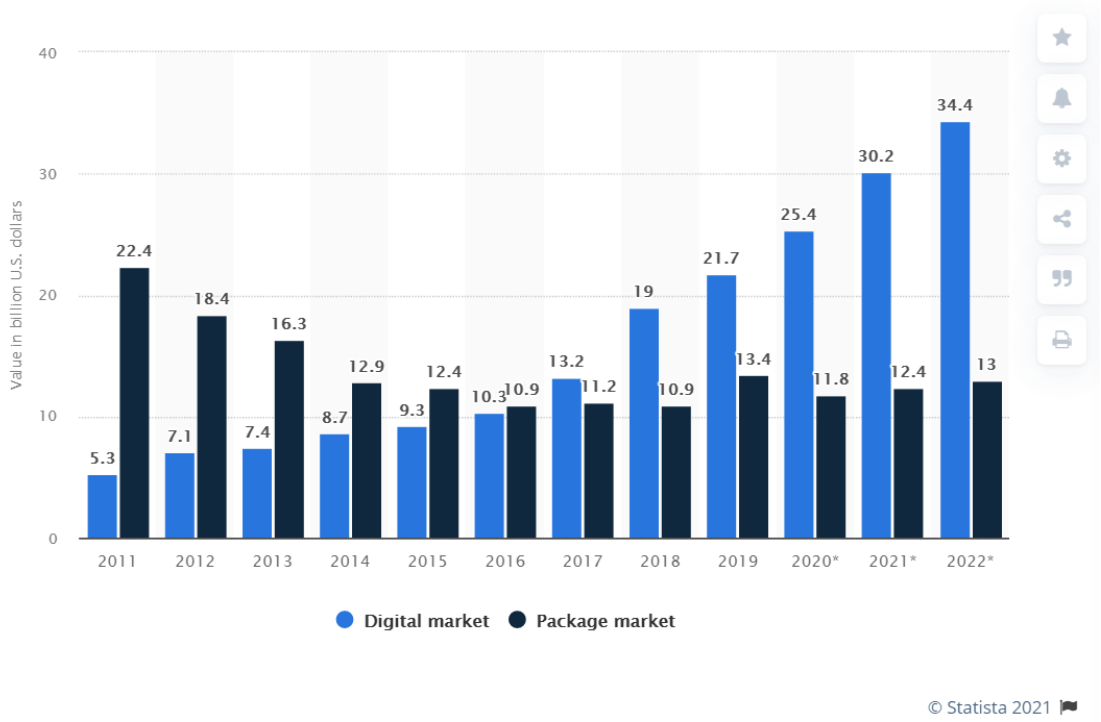
Streaming game influence

Intro

La industria de los videojuegos es el sector económico involucrado en el desarrollo, la distribución, la mercadotecnia, la venta de videojuegos y del hardware asociado. Ha sufrido un enorme desarrollo desde principios del siglo XXI debido al avance en todos los campos relacionados con la computación. Al tener las plataformas de videojuegos más capacidades a un coste menor, el interés por esta industria ha crecido desde finales de 2010 junto con otros factores externos. Este aumento del interés por los videojuegos por factores externos se debe principalmente a la aparición de plataformas donde se podían subir vídeos jugando (Gameplays) YouTube (apareció en 2005) y posteriormente plataformas especializadas en retransmisiones en directo (Streaming) que han evolucionado hasta convertirse en referentes como Twitch (nació en 2011).

En este proyecto nos centraremos en las plataformas más utilizadas. De la parte de retransmisión streaming usaremos datos de la plataforma Twitch al ser la que está más especializada en los video juegos. Por la parte de plataformas de videojuegos, la más utilizada hoy en día es la plataforma Mobile. La industria es muy opaca con relación a los datos de uso, ventas y descargas ya que ha habido un cambio de paradigma al llegar plataformas de distribución digital.



En esta gráfica podemos ver claramente este cambio. El mercado de venta de videojuegos digital va superando año a año al mercado físico desde 2017, por lo que he decidido utilizar la plataforma más usada de videojuegos de venta digital, Steam, para la elaboración del TFM.

Entornos

Los entornos que se ha desarrollado el trabajo son:

- Plataforma Github: Repositorio donde se encuentran los notebooks de jupyterlab y algunos datos que se utilizan para el front-end -> https://github.com/JoaquinLou/Streaming_game_influence
- Jupyterlab Notebooks: Contienen la lectura, visualización, procesamiento, utilización y evaluación de los modelos de ML que se ha utilizado.
- Front-end: App con un pequeño resumen del proyecto y los modelos de regresión subida a Heroku: https://joaquinlou-tfm.herokuapp.com/

Fuentes de datos

Encontrar fuentes de datos relacionados con la industria de los videojuegos no ha sido tarea fácil. Es una industria que tiene un rápido crecimiento y los datos públicos no están disponibles debido a la alta competencia. Se querían utilizar datos actuales que estuvieran relacionados con la pandemia, para ver cuánto había afectado al sector y cómo.

En un principio se pensó en usar la API de Twitch para recoger datos de la plataforma junto con un Web scraping de la página Vgchartz:

- Twitch API: Ofrece datos muy interesantes como visualizadores, participantes en los chats de determinados vídeos, visualizaciones por categorías, … pero no es una API retroactiva (no se pueden conseguir datos del pasado), por lo que no íbamos a obtener los suficientes datos para nuestros modelos.

- Vgchartz: Ofrecía datos de las ventas de videojuegos por plataforma hasta 2018. Es una fuente de datos muy interesante para hacer Web scraping pero cambié de parecer al darme cuenta de que los datos estaban limitados y no se correspondían con la actualidad. También se empezó a hacer scraping de datos de la web obteniendo datos de ventas de años, pero los bloqueos de los servidores al scraping eran constantes y la extracción de datos era complicada.

La alternativa ha sido encontrar usuarios que hayan ido recopilando datos históricos de retransmisiones de Twitch, scrapeando webs como sullygnome y datos de Steam haciendo scraping de páginas como Steamcharts.

Estos datos han sido subidos a la plataforma Kaggle y son los que se han utilizado finalmente:

- Twitch data: Usaremos los 2 datasets disponibles:
    - Twitch_global_data.csv: Se encuentran datos de la evolución de la plataforma
    - Twitch_game_data.csv: Datos de los videojuegos/categorías más vistas
- Steam data: El dataset “SteamCharts.csv” contiene los datos de los videjuegos de Steam

Datos a trabajar

Los datos que tenemos disponibles en los datasets son:

- Twitch_global_data.csv: Datos globales de Twitch unificados por año y mes. Consta de las siguientes columnas:

COLUMN	DTYPE	DESCRIPTION
YEAR	int64	Year
MONTH	int64	Month
HOURS_WATCHED	int64	Total hours watched
AVG_VIEWERS	int64	Average viewers
PEAK_VIEWERS	int64	Peak of viewers
STREAMS	int64	Nº of streams
AVG_CHANNELS	int64	Average channels
GAMES_STREAMED	object	Nº of games streamed

- Twitch_game_data.csv: Datos con dimensionalidad por juegos de Twitch (incluyendo categorías especiales de Twitch) por mes y año:

COLUMN	DTYPE	DESCRIPTION
RANK	int64	Rank in the month 1- 200
GAME	object	Name of game or category
MONTH	int64	Month
YEAR	int64	Year Hours watched on Twitch
HOURS_WATCHED	int64	Hours watched on twitch
HOURS_STREAMED	object	Hours streamed on twitch
PEAK_VIEWERS	int64	Maximum viewers at one instant
PEAK_CHANNELS	int64	Maximum channels at one instant
STREAMERS	int64	Number of streamers who streamed the game
AVG_VIEWERS	int64	Average viewers
AVG_CHANNELS	int64	Average channels
AVG_VIEWER_RATIO	float64	Average viewer ratio

- Popularity_games_Steam_Charts.csv: Datos de los videojuegos más jugados en la plataforma Steam:

COLUMN	DTYPE	DESCRIPTION
GAMENAME	object	Game name
YEAR	int64	Year
MONTH	object	Month
AVG	float64	Average number of players at the same time
GAIN	float64	Difference in average compared to the previous month
PEAK	int64	Highest number of players at the same time

AVG_PEAK_PERC	object	Share of the average in the maximum value (avg / peak) in %

Datasets

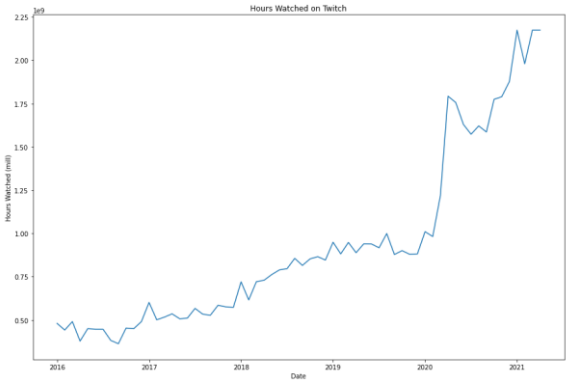
El proyecto se divide en 3 datasets para diferenciar bien qué parte se está trabajando. Cada dataset al final genera un csv con los datos limpios y procesados que se usa en el siguiente hasta la elaboración de los modelos de Machine learning en el tercero. Los 3 datasets son:

- 1-Steam data cleaning and visualization
- 2-Twitch data cleaning and analysis
- 3-Data merge & Prediction

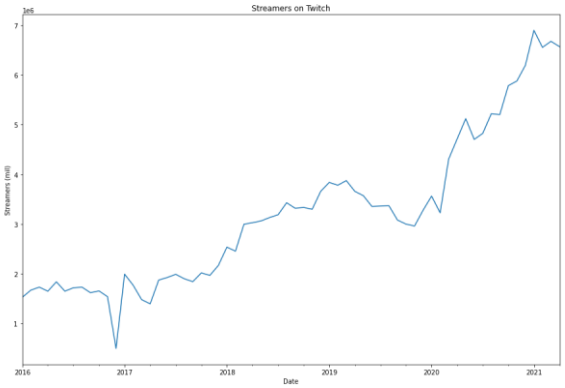
A continuación, se extraen unos insights de los datasets para entender mejor el proyecto realizado y las decisiones que se han tomado en los modelos de ML.

Twitch

Dentro de los dataset de Twitch nos encontramos la realidad de los datos de este sector. Agrupando las horas vistas por mes y año vemos como tiene un ascenso constante hasta la época covid, donde prácticamente la cantidad de horas vistas al mes se ha duplicado desde marzo 2020 a marzo 2021

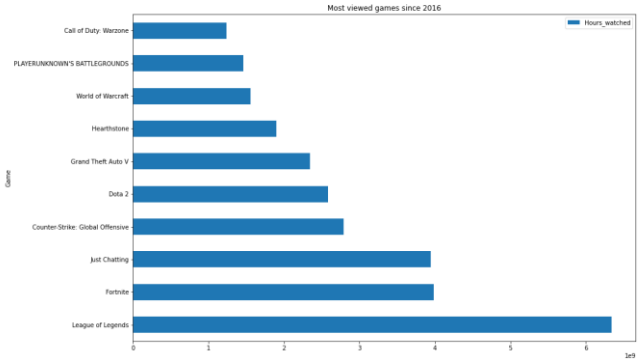


Lo mismo ha pasado para los creadores de contenido “Streamers”. Desde marzo de 2020 se ha visto un fuerte aumento en la cantidad de usuarios que se han creado una cuenta y han empezado a retransmitir contenido a través de Twitch.

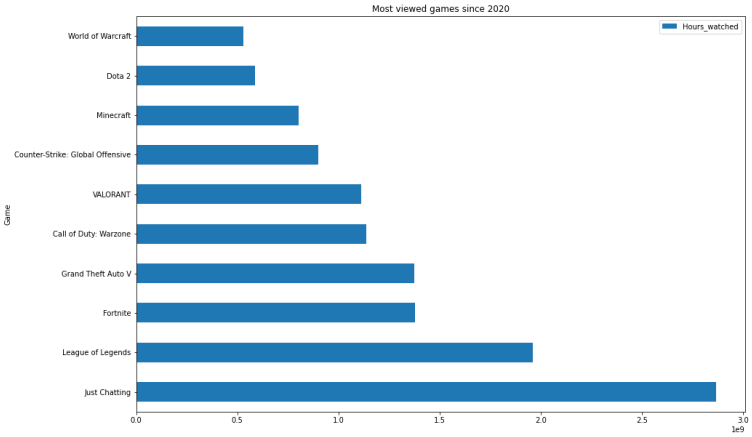


Mirando estas 2 gráficas podemos ver como el covid ha aumentado el interés en esta plataforma.

En la siguiente gráfica podemos ver los juegos más jugados desde 2016:



Y en la compararemos con la siguiente gráfica para ver cómo han cambiado las prioridades de los usuarios de Twitch durante la pandemia del covid-19:

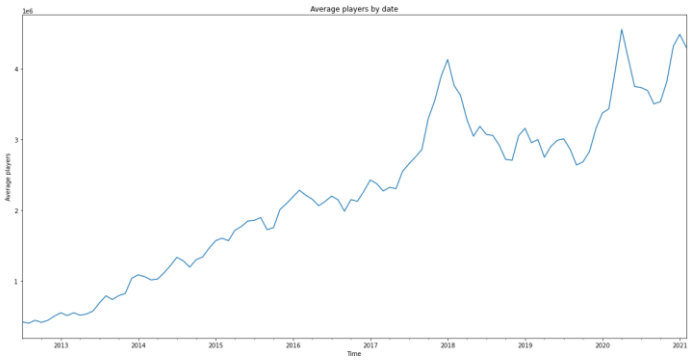


Vemos como juegos que se han lanzado durante 2020 se han colado entre los más vistos, como es el caso de Call of Duty: Warzone (salida el 10/03/2020) y Valorant (salida el 02/06/2020), pero el canal/juego más visto debido al rápido incremento del interés en los Streamers durante la pandemia ha sido Just Chatting.

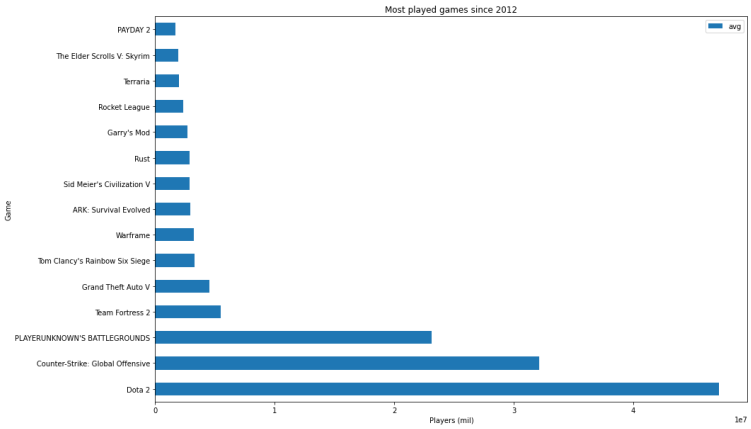
Steam

Dentro del dataset de Steam tenemos datos acerca de juegos, nº medio de jugadores e incrementos.

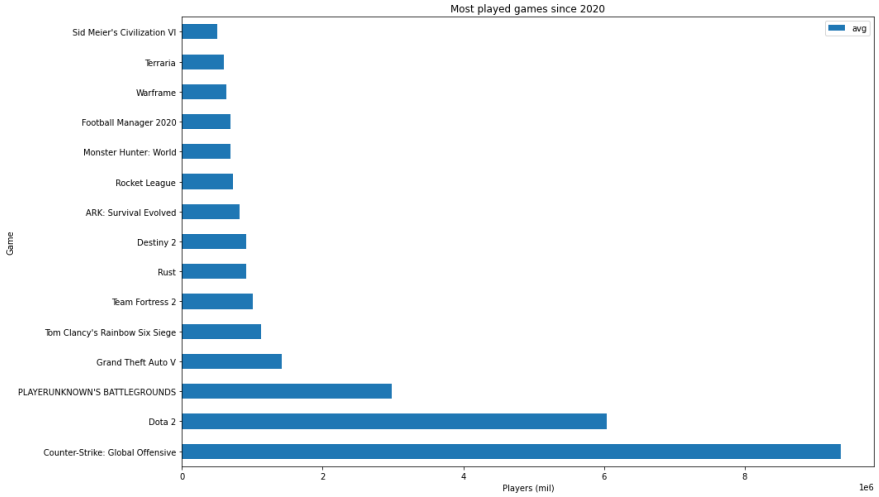
La principal métrica que nos vamos a fijar es si la situación del covid ha afectado al número de jugadores en los diferentes juegos y cuáles son los juegos más jugados después de la pandemia.



En esta gráfica tenemos la evolución del número de jugadores en la mayor plataforma digital de videojuegos. Vemos como en marzo de 2020 hay un incremento del número de jugadores probablemente debido a los confinamientos y al incremento del número de horas que la gente pasaba en casa. En la gráfica vemos que cuando las medidas de restricción empezaron a ser más leves, sobre todo en verano, el número de jugadores disminuyó, pero cuando las medidas volvieron a imponerse el número de jugadores volvió a subir.



En la gráfica anterior de videojuegos más jugados desde 2016 vemos como predominan, principalmente 2, Counter Strike: Global Offensive y Dota 2. Mientras que en la siguiente gráfica vemos como el podio se ha turnado desde al año 2020:



Viendo los datos de estas gráficas me di cuenta de que se iba a tener que dividir nuestros modelos en 2 partes para que se ajustaran mejor los datos, ya que antes de la pandemia existía una realidad completamente diferente a la que existe en la actualidad, sobre todo si nos centramos en las métricas de número de jugadores y consumidores de contenido.

Preparación del dataset para train/test

Se han tenido que limpiar varias columnas y organizar los datos para poder unificar los diferentes datasets.

- Limpieza de texto de columnas
- Transformación del tipo de columnas
- Creación de una columna fecha para graficar mejor

Después de analizar datos recogidos, se ha segmentado los datos en dos partes, tanto los datos de Steam como de Twitch debido al cambio en la tendencia de uso de las plataformas derivado por el covid-19. Por ello es que el proyecto presenta dos modelos de predicción:

- Un modelo de predicción con los datos totales desde 2016 a 2021
- Modelo de predicción con los datos desde febrero 2020 debido Aplicación de diferentes modelos de ML

A la hora de elegir las variables para la regresión nos han aparecido diferentes situaciones. Existe una gran colinealidad entre algunas variables, se ha intentado esquivar eligiendo las variables más equilibradas entre correlación y uso práctico.

Las variables que hemos escogido son:

- Avg_viewers: Se trata de una variable que recoge el número medio de usuarios que están viendo un determinado canal/juego en Twitch.
- Avg_players: Recoge el número medio de jugadores por mes en Steam

Evaluación de los modelos

Se ha recopilado información de los principales modelos de regresión. Como se trataba de predecir el número de jugadores en función del número de visualizadores de la plataforma Steam, se ha decidido aplicar los siguientes modelos de regresión:

- OLS: Ordinary least squares
- Bayesian linear regression
- ElasticNet
- KNNregressor
- Decision Tree Regressor

A la hora de determinar qué modelo era el que mejor se ajustaba a los datos y predecía mejor los resultados, se han evaluado con las principales métricas para detectar el mínimo error:

- MSE: Mean squared error (error cuadrático medio)
- RMSE: Root Mean squared error. La hemos utilizado al final para decidir el número de kvecinos era el óptimo para nuestro algoritmo
- MAE: Mean absolute error (error absoluto medio)
- MAPE: Mean absolute porcentaje error (porcentaje del error absoluto medio)
- R2: R-cuadrado (porcentaje de la variación en la variable de respuesta)
- EV: Explained variance regression score

Una vez visto que los algoritmos que menor error daban eran los KNRegression, se ha visualizado la curva para determinar por dónde estaría el mínimo y hemos usado la función GridSearchCV para encontrar el parámetro k-vecinos óptimo.

Al dividir los datos en dos dataset diferentes para ajustarlos mejor a la nueva realidad del covid-19, se ha tenido que repetir el proceso 2 veces, pero casualmente el modelo que menor error nos devolvía seguía siendo el KNNRegressor.

Frontend

En el front-end se ha decidido por usar la librería Streamlit y subirla a un servidor de Heroku. Se ha intentado no saturarla con datos para ir dirigida a un público más amplio y con un conocimiento analítico básico. El problema se ha ido describiendo en una línea temporal para entender las realidades de las dos plataformas.

Para montar el front con algunos datos básicos, se necesitaban tener los datos por separado para que los pueda leer la app directamente desde el repositorio y es por ese motivo por el que aparecen algunos archivos csv dentro del repositorio. Se han añadido principalmente gráficas con evolutivos de las 2 plataformas para ver visualmente el impacto que ha tenido el covid.

Para crear las gráficas se ha usado la librería altair ya que tenía un apartado front interactivo con el cual se pueden ampliar y mover las gráficas dentro de la aplicación. También en este apartado gráfico se han añadido un par de visualizaciones de Tableau con disgregaciones de datos por juego ya que se dio también la materia en el máster.

En el apartado de Machine learning se ha optado por un selector que carga por detrás con la librería Joblib los modelos de regresión hechos en el proyecto. Cuando seleccionas el modelo en función del rango de datos (datos totales o datos después del covid-19), aparece un selector para introducir la cifra de espectadores esperado y un botón.

Cuando se presiona el botón el algoritmo calcula el número de jugadores esperado en función de los algoritmos que hemos extraído de los notebooks y devuelve un resultado, que se trata del valor predicho por nuestro algoritmo.

Conclusiones

Tras trabajar en 2 notebooks con dos fuentes de datos distintas, hemos conseguido unir datos de dos plataformas del mismo sector y que tengan sentido. Se ha podido demostrar la correlación entre visualizadores de la plataforma de Twitch y su repercusión en el número de jugadores en Steam.

Ha quedado demostrado que el impacto del covid-19 en este sector ha sido importante, pero de una forma positiva. Los confinamientos y por lo tanto la necesidad de entretenimiento en nuestras casas ha hecho que estas dos empresas se beneficien en mayor y menor medida.

A raíz de este impacto tan fuerte por factores externos al sector, se ha decidido desarrollar 2 modelos de regresión para poder predecir en diferentes ámbitos el número de jugadores que va a tener un determinad videojuego en función del número de visualizadores en Steam. Unos ejemplos de estas aplicaciones serían:

- Promoción de un nuevo videojuego: Cuando una empresa quiere lanzar un videojuego no promociona a través de diferentes canales. Como ha quedado demostrado Twitch engloba a un perfil de usuarios muy específico y que está en pleno crecimiento con millones de horas visualizadas por mes. Saber una cifra aproximada del retorno de la inversión (ROI) puede determinar las colaboraciones con Streamers de Twitch.
- Promoción de eventos dentro de videojuegos: Existe una tendencia reciente de crear eventos dentro de videojuegos con fines específicos. Fornite los popularizó con conciertos dentro de su juego y se ha extendido en otros formatos.

Próximos pasos

Habría que seguir recopilando datos porque tenemos varios factores que afectan al proyecto para completarlo:

- Falta transparencia por parte de las empresas del sector. Se entiende que, el sector de los videojuegos y el streaming, es un sector muy competido y con un amplio crecimiento como hemos visto en los datos, pero la facilitación de más datos para explotar mejoraría la búsqueda de nuevas oportunidades y nichos en el mercado.
- Impacto de factores externos: La aparición del covid-19 ha cambiado las costumbres de los consumidores. La tendencia es muy alcista desde febrero de 2020 pero habrá que esperar a tener más datos durante varios años para ver si se mantiene.