Universidad Torcuato Di Tella

Master in Management+Analytics

Enfoques de aprendizaje automático para tratar la fuga de suscriptores en una plataforma de streaming en el marco de una trasformación digital.

Resumen

La irrupción de las plataformas de streaming transformó muchos aspectos en la industria del entretenimiento. Las empresas líderes lograron generar valor personalizando la oferta de contenidos online de acuerdo a las preferencias específicas de cada cliente y para ello fueron cruciales los datos. En el presente trabajo se utilizan técnicas analíticas descriptivas, predictivas y prescriptivas en relación a la fuga de suscriptores de la plataforma de streaming Disney Plus. En primer término, se utilizarán técnicas de clustering para analizar las características de la oferta de contenidos disponibles en la plataforma. En segundo término, se usarán técnicas de aprendizaje automático para predecir la baja de suscriptores, utilizando múltiples variables relacionadas al producto y al contenido, para finalmente generar recomendaciones de títulos que eviten la fuga. Se indagará sobre los desafíos y limitaciones de realizar dicho trabajo en una empresa multinacional en plena transformación digital.

Alumna: Julieta Santarelli
Tutor: Gabriel Martos
Julio 2021

Universidad Torcuato Di Tella

Master in Management+Analytics

A Machine learning approach to improve subscriber's retention on a streaming platform within a digital transformation framework.

Abstract

Streaming platforms transformed many aspects of the entertainment industry. Leading companies became data-driven and reached efficiencies understanding their clients' preferences and offering personalized content. In this work, descriptive, predictive and prescriptive analytics techniques are used to deep dive towards Disney Plus's subscriber's engagement and churn behavior. First, clustering techniques will be used to analyze the characteristics of the platform's catalog. Second, machine learning techniques will be used to predict churn, working with variables related to the product characteristics and to the subscriber's content consumption. Finally, personalized content recommendations are developed seeking to avoid estimated churn. This project analyzes the challenges and limitations of developing data analysis in a multinational company living through a digital transformation.

Student: Julieta Santarelli

Tutor: Gabriel Martos

July 2021

Contenido

Introducción

Contexto: transformación en la industria del entretenimiento.

La forma de consumir entretenimiento fue cambiando en las últimas décadas a raíz de los avances tecnológicos que se fueron dando. La disrupción más fuerte se dio con la llegada de internet. Una de las industrias más afectas fue la del cine, donde tradicionalmente eran las grandes productoras las que invertían en la producción de una nueva película y se encargaban también de la gestión de la oferta al público. Las películas se lanzaban primero en los cines, luego estaban disponible para alquilar en VHS (o más tarde, DVD) y finalmente se emitían en televisión, primero por los canales de cable y luego por los de aire. Todas estas ventanas de exposición, generaban ganancias millonarias a las productoras y distribuidoras.

Con la llegada de internet, llego el *streaming*[1] que permitió el fácil acceso a películas y series en los hogares. Se generaron varias consecuencias en la industria del cine. En primer lugar, acudir a las salas de cine dejó de ser un 'acontecimiento'. Por ejemplo, en el Gráfico 1 se observan los datos de encuestas recientes respecto a la frecuencia con que una persona asiste al cine. Casi la mitad de los encuestados lo hace a lo sumo una vez al año. En segundo lugar, las ventanas de explotación tras el estreno de una película se achicaron cada vez más. Muchas ventanas desaparecieron, por ejemplo, los DVD, generando un decrecimiento importante en las ganancias de las productores y distribuidoras. Por último, se dio un cambio en la experiencia de ver una película. Tradicionalmente era una costumbre que se daba en grupo, no solo se buscaba ver determinada película, sino compartir el momento con otros, más allá de si esa película era la favorita o de una calidad extraordinaria. Con las plataformas de streaming, el contenido está a disposición del espectador en todo momento y en cualquier lugar. No hace falta coordinar con otros, con hacer clic en el salón de tu casa, se puede disfrutar de las películas favoritas. Se impuso la costumbre de ver las películas sin compañía, en cualquier tipo de pantallas, y en todo momento. El hábito del visionado individual de productos audiovisuales creció, mientras que las salas de cine han tenido que conformarse con llenar su aforo, si lo logran, durante los fines de semana. (Aresté Sancho, 2021)

[1] El streaming es un tipo de tecnología multimedia que envía contenidos de vídeo y audio a un dispositivo conectado a Internet.

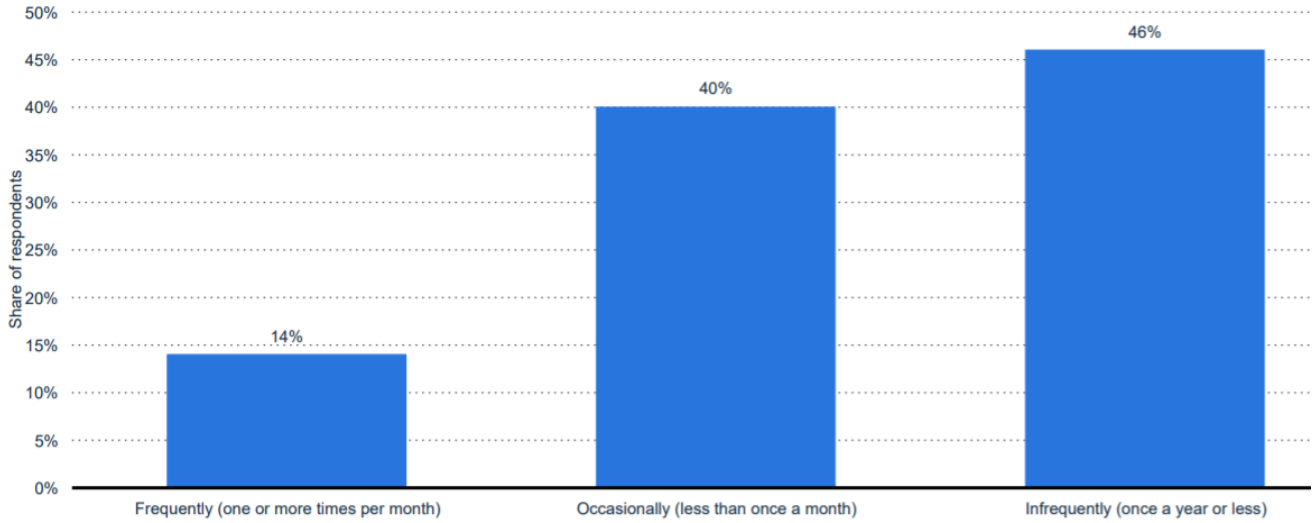

Gráfico 1: Frecuencia con la que se concurre al cine. Los datos corresponden a una encuesta elaborada en EEUU, donde participaron 2200 encuestados mayores de edad, entre el 20 y el 23 de junio de 2019. Fuentes: Morning Consult y The Hollywood Reporter.

La otra industria afectada directamente por las innovaciones tecnológicas fue la de la televisión. La llegada de internet trajo una opción no lineal para consumir el contenido, permitiendo a los espectadores seleccionar qué programas ver y cuándo verlos. Entre ellos, se encuentran los servicios VOD (video *on demand*) que transmiten contenido a través de cable, satélite e internet, pudiendo el usuario ver contenido ya emitido, grabado, o pagar por títulos exclusivos. Dentro de los VOD, están los servicios OTT (*over-the-top*) que solamente transmiten el contenido a través de internet. Generalmente se hace referencia a este último grupo al hablar de las plataformas de streaming. Es propio de estos servicios que el contenido circule libremente por internet, sin que haya un control por parte del operador.

Gran parte de la audiencia de la televisión lineal, migró a la no lineal en la última década. Primero lo hicieron las generaciones más jóvenes, y luego los de mediana y larga edad. Como se ve en el Gráfico 2, actualmente las personas dedican más tiempo a ver contenido a través de streaming que a través de la televisión lineal. No solo sucedió que el tiempo que antes una persona dedicaba a ver televisión lineal paso al consumo no lineal, si no que en momentos que antes no se podía ver televisión, por ejemplo, en un viaje al trabajo, hoy en día sí se puede usando cualquier dispositivo con conexión a internet, o previamente descargando el contenido. La portabilidad del streaming, combinada con una rica oferta de contenido y un precio de servicio accesible, generaron un crecimiento exponencial en los suscriptores de estos servicios. Como se observa en el Gráfico 3, la principal plataforma de streaming, Netflix, tuvo un crecimiento en suscripciones alrededor del mundo, del 500% entre 2013 y el 2021.

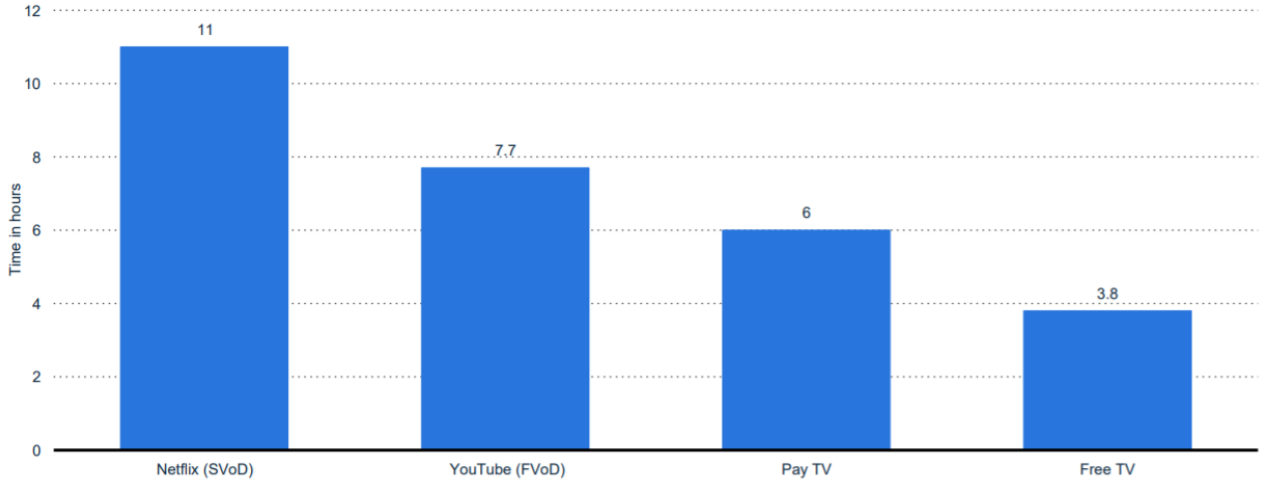

Gráfico 2: Tiempo semanal dedicado a ver contenido de video en Argentina en el cuarto trimestre de 2019, por plataforma (en horas). Fuente: Cámara Argentina de Internet y Business Bureau.

Netflix

Netflix se lanza en abril del 1998 en California, convirtiéndose años más tarde en la campeona mundial de los servicios de streaming. Sin embargo, la empresa nace como una plataforma online de renta de DVD. Con la llegada de internet, adaptan su modelo de negocio a las nuevas formas de consumir contenido. En el 2007 ofrecen un nuevo servicio donde el cliente pagaba un fee fijo, y a cambio, podía ver todo el contenido disponible en la plataforma de forma ilimitada desde su computadora a través de internet. Como se describió anteriormente, este cambio fue drástico para la industria, pero esta empresa aprovecho la oportunidad, y su modelo de negocios fue un éxito. Observando nuevamente el Gráfico 3, se ve como a partir del 2013, luego de su expansión fuera del territorio estadounidense, la empresa incrementó su cartera de clientes en 175 millones en solo 8 años.

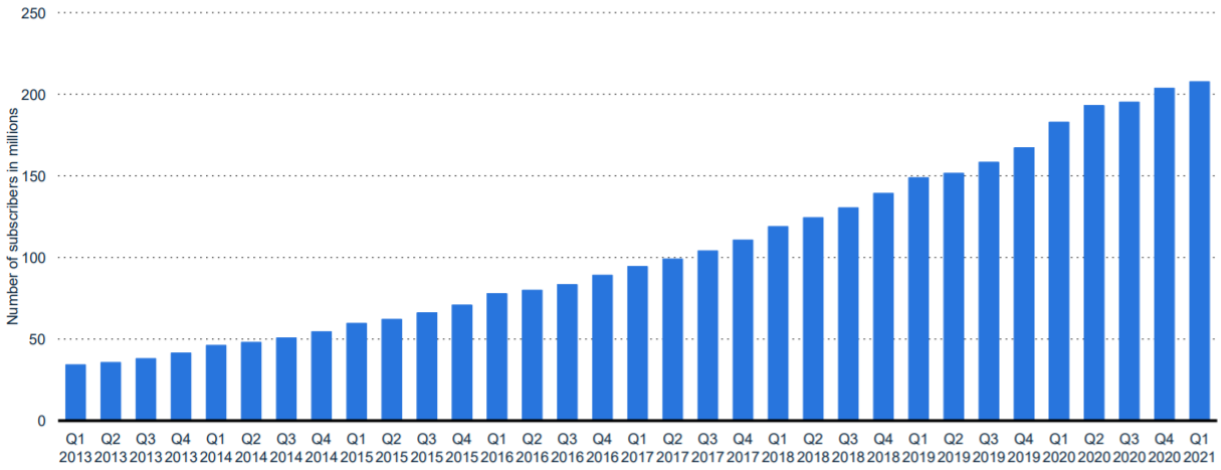

Gráfico 3: Número de suscriptores pagos de Netflix en todo el mundo desde el primer trimestre de 2013 hasta el primer trimestre de 2021 (en millones). Fuente: Netflix

Netflix buscó siempre servir a los clientes las películas que más les gustan, de modo que no tengan necesidad de buscarlas en otro sitio. Su catálogo de títulos se volvió clave. A medida que fue creciendo, se encargó de cerrar tratos con las productoras y dueñas de derechos más importantes, entre ellas The Walt Disney Company, para poder ofrecer sus títulos en su plataforma. Más tarde, decidieron lanzar sus propias producciones, lo que les permitía seguir ampliando su catálogo, sin la dependencia de las negociaciones con las productoras externas. El enriquecimiento del catálogo, fue apalancado con el desarrollo de un algoritmo de recomendación, que pone en valor todo el catálogo, incluidos los títulos más desconocidos pero susceptibles de satisfacer al suscriptor. Netflix logró así conocer a su público, sin resultar intrusivo. Solo se vende a sí misma, a unos suscriptores que ya han pagado, para lograr su satisfacción y que no se arrepienten de pagar su cuota mensual (Aresté Sancho. 2021). Para el 2020 se posicionó en el puesto 164 dentro de las 500 empresas de Estados Unidos con mayores cifras de ingresos de la lista *Fortune Global 500*[2].

The Walt Disney Company

The Walt Disney Company, de ahora en adelante Disney, es una empresa estadounidense fundada en 1923 que supo convertirse en una de las empresas de entretenimiento más poderosas e influyentes. Con casi 100 años de experiencia, aprendió a adaptarse a los cambios que se fueron dando en la industria. En su momento supo alcanzar el éxito en la transformación en la industria de la animación, incorporando

[2] https://fortune.com/global500/

la nueva tecnología CG[3], que estaba sustituyendo muy rápidamente la animación dibujada a mano. Esta adaptación la logro paulatinamente, primero con contratos de coproducción con Pixar, y luego de 10 años, en el 2006, integrando la empresa mediante una inversión de USD$7.4 billones. Durante los 15 años siguientes, Disney-Pixar produjo 16 películas las cuales generaron recaudaciones por más de USD$11 billones.

El streaming, por su parte, tuvo todo tipo de impactos en la empresa. En un principio, alrededor del 2015, se vio la oportunidad de generar beneficios extras incorporando una nueva ventana de distribución. Se cerraron contratos millonarios con servicios de SVOD Y OTT, dándoles el derecho de transmitir el contenido de la compañía por varios años. Como se muestra en el Gráfico 4, Netflix tenía derecho a transmitir películas de Disney, aproximadamente 3 años después de su lanzamiento en cine. El poder de negociación aún estaba en manos de Disney, y esto le permitía priorizar a otros negocios en el orden en la distribución de sus películas, por ejemplo, sus propios canales de televisión. También se puede observar cómo alrededor del 2016 una película seguía generando beneficios por más de tres años desde su lanzamiento en las salas de cine.

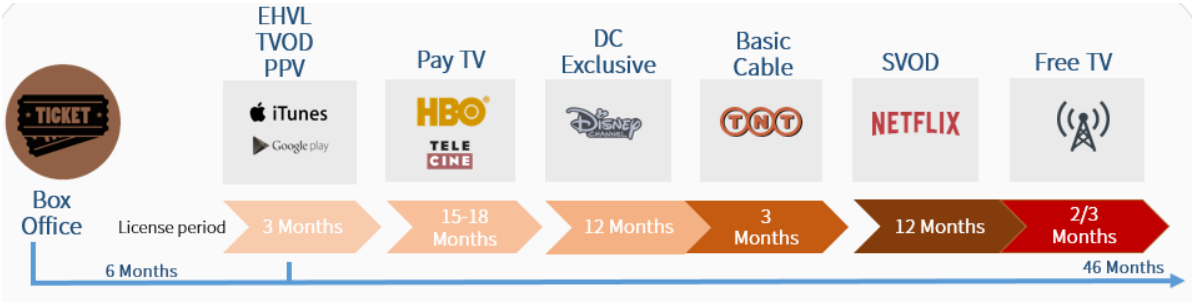

Gráfico 4: Ejemplo de un ciclo de distribución de una película producida por Disney en el 2016. La empresa firmaba contratos de exclusividad con distintos distribuidores, cuyo valor dependía de la rapidez con la que obtenían el contenido luego de su estreno en cine, y por cuanto tiempo. Primero se lanzaba en el cine. Pasado un período de 6 meses donde no estaba disponible en ningún lado, se podía acceder a verla pagando una tarifa adicional ya sea para descargarla (EHVL. Ej. ITunes), sintonizarla por streaming (TVOD. Ej. Google Play) o para verla en un canal de televisión (PPV). Luego la exclusividad de distribución la tenían los canales de Pay TV (canales Premium) por hasta un año y medio. Algunas películas, dependiendo el target de audiencia, pasaban a exclusividad de los canales Disney por 1 año. Luego llegaba a la televisión por cable básica y más tarde podía encontrarse en Netflix, por 1 año. Pasadas todas las ventanas, solían terminan en la televisión por aire.

En los años siguientes, mientras Netflix aumentaba cada vez más su poder, Disney empezó a experimentar en el mercado de streaming con su producto *Disney Life* que llego a estar disponible en tres países de Europa. También fue realizando compras

[3] Tecnología CG: Imágenes generadas por computadora, del inglés *Computer Generated Imagery* (CGI). Son el resultado de la aplicación de la infografía y más específicamente, de los gráficos 3D generados por ordenador, para su uso en diversas formas de arte, entretenimiento y medios.

estratégicas con la clara intención de competir en este nuevo mercado. Entre ellas se destaca la compra de *BAMTech* en el 2016, luego renombrada *Disney Streaming Services*, de la cual se obtuvo gran parte del *know-how* tecnológico respecto al streaming. Para el 2018 los diversos negocios de Disney se segmentaban en:

1. ***Media Networks:*** operaciones, ventas afiliadas y publicitarias de todos los canales de televisión operados por la compañía.
2. ***Parks, experiences & products:*** incluye los parques temáticos de la compañía, los resorts vacacionales, así como todos los productos comerciales de Disney.
3. ***Studio entertainement:*** producción, adquisición y distribución de películas.
4. ***Direct to consumer***: servicios de transmisión de Disney y negocios de medios en el extranjero.

Como se mencionó anteriormente, en los últimos cinco años el impacto del streaming fue feroz. Dos de los cuatros segmentos de Disney, *media networks* y *studio entertainment*, se vieron fuertemente impactados por el crecimiento de estos servicios. Por ejemplo, en el Gráfico 5 se observa cómo el número de televidentes del canal Disney Channel en Estados Unidos cayó un 82% desde el año 2014. La disminución en audiencia tiene impacto a corto y mediano plazo en el negocio de Media Networks. Los ingresos por anuncios se caen inmediatamente, ya que los clientes prefieren publicitar sus productos en lados con mayor exposición. Por otro lado, los contratos con los cable-operadores se mantienen, pero se pierde poder de negociación y si peligra el negocio del cable-operador, peligran directamente los contratos con los canales.

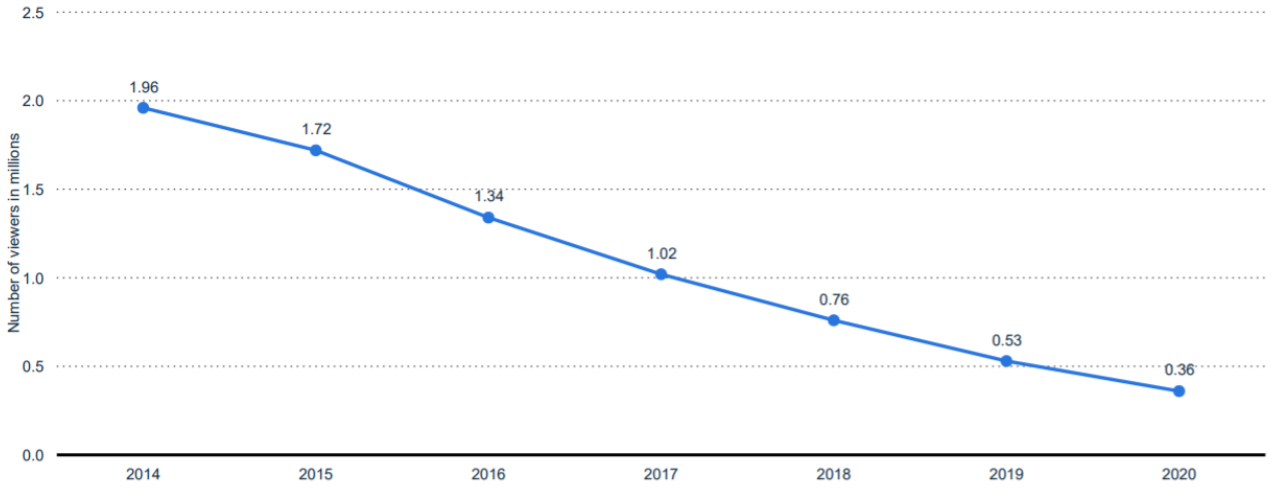

Gráfico 5: Número de espectadores de Disney Channel (canal del segmento *media networks)* en los Estados Unidos de 2014 a 2020 (en millones). La caída en los televidentes de 1.96M a 0.36M representa un decrecimiento del 82% en 6 años. Fuentes: Nielsen y Variety

Mientras estos negocios luchaban por generar nuevos ingresos, la compañía centró sus esfuerzos en sus servicios de streaming, anunciando el lanzamiento de ESPN+ en 2018, Disney+ en 2019 y Star+ en 2021. Se preparó para la batalla, haciendo foco en sus marcas, su catálogo, buscando concentrar el mejor y más diverso contenido posible. En el 2019 realizó la compra 21st Century Fox, por UDS 71 millones, que no solo le otorgó el derecho sobre el contenido de los estudios Fox, FX, National Geographic, etc.; si no que además le dio el poder mayoritario sobre Hulu, un servicio de streaming de Estados Unidos con más de 41 millones de suscriptores, y los derechos sobre Hotstar, una plataforma de streaming India la cual contaba en ese entonces con 150 millones de usuarios activos mensuales. Los lanzamientos de los distintos productos de streaming, tuvieron mucho éxito. Por ejemplo, en Estados Unidos, ofrecieron un precio combo por contratar Disney+, ESPN+ y Hulu, lo que lograba ofrecer contenido a diversos segmentos de audiencia, y ser un fuerte competidor de Netflix. En el Gráfico 6 se ve cómo Disney+ creció un 258% desde primer trimestre del 2020. Durante este período se lanzó en Europa, Asia y Latinoamérica.

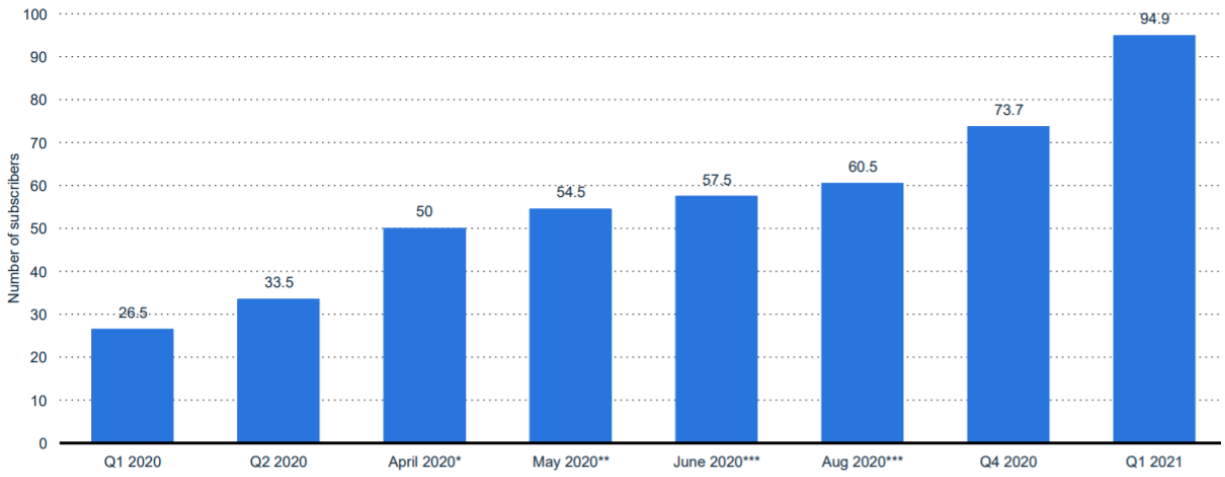

Gráfico 6: Número de suscriptores de Disney + en todo el mundo desde el primer trimestre de 2020 hasta el primer trimestre de 2021. Fuente: Disney a través de Statista.

En el 2020 la empresa anunció darles foco primordial a sus servicios de streaming para lo cual definieron una nueva estructura nombrando dos grandes divisiones. Por un lado, *Disney Media and Entertainment Distribution*, concentrando la distribución de contenido, y negocios asociados tanto para cine, televisión, música, teatro y streaming. Esta unidad además es responsable de tres grupos encargados de la producción y operatoria del contenido. Por otro lado, quedó definida la división *Disney Parks, Experiences and Products* relacionada a los parques de atracciones, hoteles, productos y experiencias. En 5 años la compañía atravesó una transformación inmensa, que culminó con esta última reestructuración, donde se decidió incluir en la misma división

a los negocios que se encuentran más en riesgo, con aquellas nuevas apuestas que prometen posicionar a Disney una vez más entre los principales competidores de la industria. En el Gráfico 7 se puede ver la estimación en números de suscritores para las plataformas de streaming mas importantes entre el 2020 y 2026. El pronóstico para Disney es prometedor, alcanzado en suscriptores a su principal rival, Netflix, en el 2026.

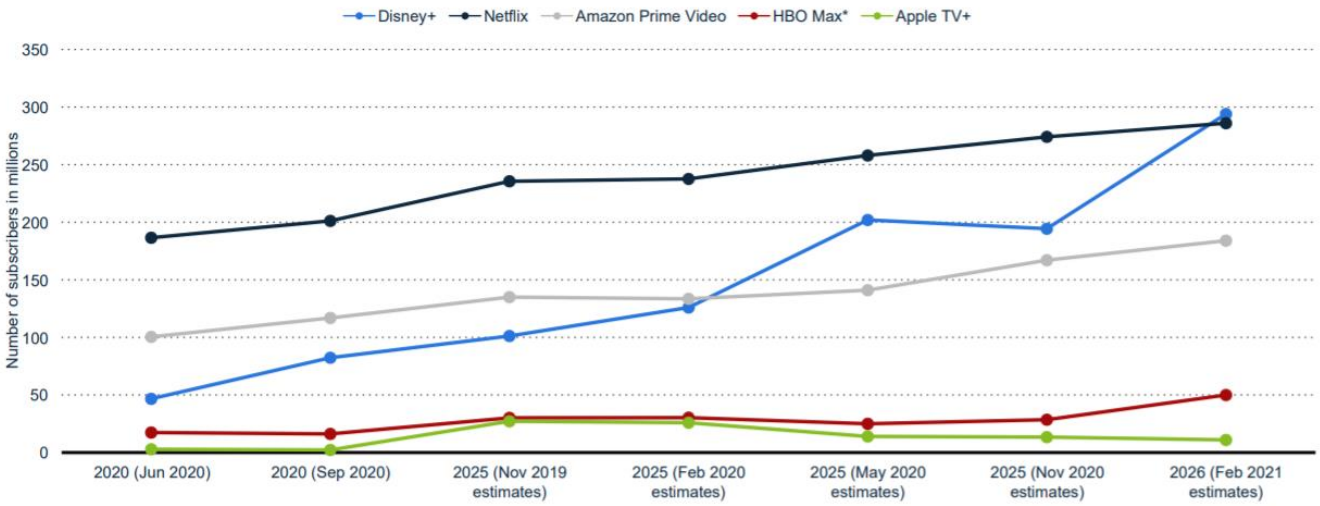

Gráfico 7: Número estimado de suscriptores de SVoD en todo el mundo en 2020 y 2026, por servicio (en millones). Fuente: Investigación de televisión digital.

Los datos en la industria del entretenimiento.

Previo a la llegada de las plataformas de streaming, los datos relacionados a la performance de los contenidos eran acotados. Estos provenían mayormente del recuento de ventas de entradas/productos, encuestas, y de muestras de consumidores generalmente sesgadas (por ejemplo, el rating, para los canales de televisión). Cuando el contenido pasa a consumirse a través de streaming, con la infraestructura adecuada, se pudo almacenar grandes *data-sets* con toda la información de los consumidores, obteniendo un *feed back* casi instantáneo. Explotando estos *data-sets* a través de la minería de datos, el aprendizaje automático y la inteligencia artificial, es posible alcanzar grandes eficiencias desde el lado de la oferta. Por ejemplo:

> Mejorar la calidad de los productos existentes.
> Optimizar los recursos y estrategias.
> Crear nueva oferta teniendo en cuenta los gustos y preferencias de los consumidores observados en los datos.
> Generar publicidad dirigida a determinados sectores para coincidir con quienes pueden estar interesados en el producto.

Desde sus comienzos Netflix hizo foco en la importancia de los datos y buscó alcanzar las eficiencias mencionadas en su modelo de negocios. Un ejemplo de esto fue el lanzamiento del concurso *NetflixPrize* que se dio entre el 2006 y el 2009, para motivar a los investigadores a mejorar la precisión del sistema de recomendación, puntualmente lograr el mejor algoritmo de *collaborative filtering* para predecir las calificaciones de los usuarios. A cambio de un premio económicamente sustancioso alcanzo mejorar su algoritmo de predicción de ratings un 10%. Asimismo, Netflix identificó la importancia de los datos en todas las áreas del negocio, por ejemplo, en la creación de nuevo contenido. Enrique Dans en su artículo "Netflix: big data, long game… y resultados" destaca como Netflix supo incorporar datos en sus áreas creativas, haciendo la fusión más impensada del uso de datos para la toma de decisiones en la dirección artística. Detalla como las producciones de la empresa se basan 70% en datos y 30% en los recursos disponibles para dicha producción. El éxito de Netflix prueba que su decisión de ser data-driven fue la acertada. Al recurrir de manera consistente al análisis de datos, un mayor porcentaje de las decisiones estarán mejor tomadas, los riesgos mejor balanceados, y los resultados serán mejores.

Por su parte, Disney, como se menciona previamente, realizó una gran inversión en contenido, apoyándose en la concentración del mismo y en la fidelidad de sus marcas como principales ventajas competitivas en el mercado de streaming. Sin embargo, en este negocio directo al consumidor no solo es fundamental tener la mayor cantidad de conocimiento respecto a los intereses de los mismos, sino que las exigencias del producto son altas. El cliente se acostumbró rápidamente a que los servicios de streaming "sepan que quiero ver", y se espera que estén a la altura de la competencia. El streaming se convirtió en un negocio donde el análisis de datos puede llevarte al éxito o dejarte fuera de juego. En este contexto Disney se ve urgida a transformarse digitalmente y convertirse en una empresa data-driven para poder competir con aquellas que sí lo son. De lo contrario, desperdiciaría grandes oportunidades generadas por la explotación de los datos.

Problema

Al decidir focalizar gran parte de los esfuerzos en el negocio de streaming, Disney no solo atraviesa un cambio de estructura, sino una transformación digital. Se puede observar como en los últimos años sucedieron dos procesos en paralelo. Por un lado, en los negocios tradicionales (cine, televisión) se mantuvo la esencia del negocio, adaptándolos para co-existir con el nuevo mercado, por ejemplo, quedándose con la exclusividad de todos sus contenidos. Por el otro se fue creando el negocio de streaming, de forma disruptiva, buscando desarrollar innovación que será la fuente de crecimiento en el futuro. Dentro de esta transformación se fue dando un "intercambio de capacidades", donde los negocios tradicionales y los nuevos comparten recursos sin cambiar la misión o la operación de ninguno. Por ejemplo, lideres con una vasta

experiencia en las estrategias tradicionales de Disney, se integraron en los equipos de streaming provenientes de las empresas nativas digitales que se fueron adquiriendo, por ejemplo, Hulu. De la misma manera, aquellos expertos digitales, se integraron con los negocios ya existentes para tener una visión holística de las estrategias futuras.

Esta transformación tiene su cuna en Estados Unidos, donde se encuentra la casa matriz y gran parte de la compañía. Se estructuraron de una manera exitosa, creando muchas divisiones en torno a los datos para abastecer a los nuevos negocios. Por ejemplo, existe un equipo dedicado al análisis de datos para la retención de clientes de las plataformas de streaming. El equipo se conforma por líderes de diversos perfiles data-driven entre ellos científicos de datos, ingenieros de datos, y otros dedicados a tareas de *business intelligence*. Todos compartiendo un mismo objetivo, mismos clientes, y contando con acceso completo a los datos y herramientas necesarias. Pero la compañía tiene un desafío adicional en esta transformación, las demás regiones. Los negocios tradicionales se podían manejar regionalmente, alineados a nivel global, pero ocupándose de gran parte de la operatoria de forma local. El negocio de streaming exige una integración diferente. La mayoría de las decisiones tienen que basarse en los datos. Así, la transformación digital tiene que llegar de forma exitosa a las regiones. El acceso a los datos y herramientas tiene que ser el mismo para cualquier equipo que trabaje con datos, sea cual sea su región.

En Latinoamérica esta transformación se continúa dando. Antes del lanzamiento de Disney+ en la región, se definió una estructura en torno a este negocio. Por un lado, los equipos relacionados a lo comercial (desarrollo de producto, *life-cycle*, marketing, contenido etc.). Por otro lado, los equipos de datos, entre ellos *data technology, data science* y *business intelligence*. Con el lanzamiento de la plataforma en Latinoamérica, en noviembre del 2020, los equipos comerciales comenzaron a necesitar cada vez más cosas de los equipos de datos, mientras que los equipos de datos todavía buscaban integrarse con la estructura de datos y herramientas utilizadas en las demás regiones. En este contexto se presenta una gran necesidad de parte del equipo de life cycle: entender los principales motivos de cancelación de suscripciones y contar con una herramienta que ayude a la retención de los mismos.

El equipo de datos dedicado a la retención de clientes en Estados Unidos, ya había trabajado en análisis descriptivos y modelos predictivos en relación a la baja de suscriptores, trabajando con la tasa de *churn*. Esta tasa mide la magnitud de suscriptores que se dan de baja Disney Plus durante un período específico. Entre las primeras observaciones, se identificó la relación entre la antigüedad de un suscriptor y dicha tasa. Como se muestra en el Gráfico 8, la tasa dentro de los suscriptores con antigüedad de uno y dos meses, M1 y M2 respectivamente, presenta una tendencia más lineal que en aquellos con mayor antigüedad. Así, los equipos norteamericanos trabajaron en modelos que predicen las bajas durante estos periodos. Pero adaptarlos a las necesidades locales era prácticamente imposible ya que no aún no se terminaban de dar las integraciones necesarias. De esta manera los resultados y descubrimientos se

compartían a Latinoamérica en presentaciones cerradas, que no permitían alteraciones para satisfacer las necesidades locales.

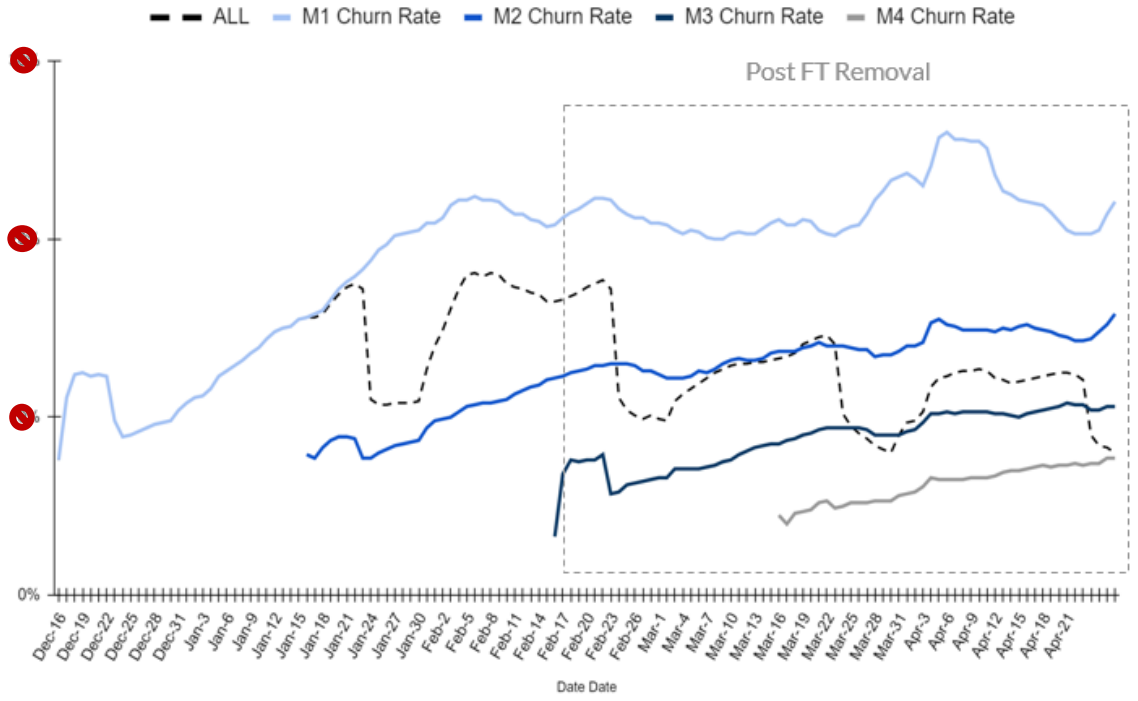

Gráfico 8: se ve el comportamiento de la tasa de churn, para los suscriptores de distinta antigüedad. Por confidencialidad se han tachado los valores de las tasas de churn, pero no hacen a la interpretación de la tendencia. Fuente: *Disney Media & Entertainment distribution.*

En un principio, los equipos locales pudieron sacar provecho de los resultados compartidos, conociendo en mayor profundidad el comportamiento de los suscriptores en sus dos primeros meses, y las variables que más influían en su baja. Pero con el crecimiento del producto, las necesidades del equipo eran mayores, y necesitaron contar con el equipo local de datos para guiar la toma de decisiones. Los problemas planteados fueron los siguientes:

1. Tener una visión completa del comportamiento de todos los suscriptores que pueden darse de baja, sin importar su antigüedad. A seis meses del lanzamiento de la plataforma, acotar los análisis a los suscriptores con uno o dos meses de antigüedad, dejaba de lado a muchos clientes, los cuales necesitaban ser observados y analizados para guiar las estrategias de del producto.
2. Predecir la baja de los suscriptores para saber hacia quienes dirigir sus estrategias de retención.
3. Poder conocer esta probabilidad con una anticipación suficiente que permita realizar las acciones de retención. Este tiempo se definió en 30 días.

4. Relacionar el consumo de contenido en los análisis y modelos de baja de suscriptores. Esto no solo les permitiría mejorar su estrategia de adquisición y lanzamiento de nuevo contenido, sino también darles mucha más información de qué busca el cliente y cómo retenerlo.

Desde el equipo de datos existían desafíos adicionales para poder ayudar a los requerimientos del área comercial.

1. Los accesos a los data-sets aun no eran completos.
2. No existía documentación respecto a los data-sets compartidos.
3. Los perfiles de permisos sobre las bases de datos eran limitadas.

Objetivo

El presente trabajo busca predecir la probabilidad de baja dentro del próximo mes, de los suscriptores de Disney Plus Latinoamérica, sin importar su antigüedad en la plataforma, e incorporando toda la información posible respecto al consumo de contenido. El objetivo es poder brindarle al negocio hallazgos y predicciones sobre la baja de suscriptores, que puedan utilizarse para mejorar la retención de clientes incorporando datos acerca del contenido en estas decisiones. Se busca aprovechar los esfuerzos de marketing ya realizados para el lanzamiento de la plataforma en 2020 y los posteriores lanzamientos de contenido de mega-producciones. Es decir, optimizar el costo ya realizado, para en un futuro poder bajar el costo de adquisición de clientes.

Para alcanzar dicho objetivo se dividirá el trabajo en tres etapas. En cada una se trabajará con distintas metodologías de minería de datos y aprendizaje automático y se obtendrán resultados que se utilizarán en las etapas siguientes. La decisión de segmentar el proyecto en tres etapas tiene varias ventajas, pero se destaca la posibilidad de reutilizar los resultados de cada una en diversos proyectos futuros.

I. Análisis y clusterización de Contenido.

En la primera etapa se trabajará con el catálogo de títulos en Disney+ Latinoamérica. Se recolectarán la mayor cantidad de características posibles para entender la oferta de contenido. El análisis descriptivo se profundizará trabajando con un modelo de *clustering*, para poder identificar grupos homogéneos dentro de la oferta de contenido. Los resultados alcanzados en esta etapa permitirán incorporar más información respecto al consumo de contenido por parte de los suscriptores en las etapas siguientes. Por ejemplo, ¿qué proporción de títulos son del género romántico? Las personas que ven contenido de este género, ¿tienen algún comportamiento específico en a la cancelación del producto? Algunas de las preguntas que se buscaran responder en esta etapa son:

- ¿Existe diversidad de contenido en el catálogo de Disney Plus?
- ¿Se cuenta con suficientes datos respecto a las características de los títulos disponibles?
- ¿Cuántos grupos homogéneos podemos identificar entre los contenidos disponibles en la plataforma?

II. Predicción de bajas.

La segunda etapa será la más importante. En la misma se trabajará con aquellos suscriptores de Disney+ que tengan la posibilidad de cancelar su cuenta dentro de los próximos 30 días. El trabajo contemplará el comportamiento del consumidor, en sus decisiones sobre el producto y en las preferencias de contenido dentro de la plataforma. Se partirá de la idea planteada en el *paper* "Behavioral attributes and financial churn prediction" de Kaya, E. sobre la mejor calidad de predicción alcanzada por los modelos de predicción de bajas que trabajan con datos que describen el comportamiento de los clientes, en contraposición a aquellos basados en características demográficas.

De esta manera se buscará identificar aquellas variables que impacten en la fuga de suscriptores. Se incluirán variables del tipo comerciales, por ejemplo "plataforma de facturación" y también variables en relación al consumo de contenido, por ejemplo "cantidad de veces que un suscriptor vio un contenido en la última semana". En un primer paso se harán análisis descriptivos para comprender el comportamiento de las variables. Luego, se trabajará en un modelo predictivo, cuyo objetivo será estimar que clientes son potenciales candidatos a abandonar la plataforma para que el equipo de lifecycle tome acciones prescriptivas que eviten la fuga de dichos usuarios. Los resultados del modelo se utilizarán en la etapa siguiente, definiendo el alcance de las acciones de retención. En esta etapa, algunas de las preguntas que se buscará responder son:

- ¿Qué atributos comerciales de un suscriptor tienen más peso a la hora de cancelar su cuenta?
- ¿De qué manera impacta el contenido consumido por un suscriptor a la hora de cancelar su cuenta?
- ¿Es necesario contar con más datos respecto a los suscriptores para entender el comportamiento de baja?

III. Recomendación de contenido

En esta última etapa se buscará brindarle una herramienta prescriptiva al equipo de *life cycle* para poder retener a los suscriptores con alta probabilidad de abandonar la plataforma. Esto se logrará generando una recomendación de contenido. Se trabajarán con los resultados tanto de la etapa I como de la etapa II. Se implementarán dos enfoques distintos de sistemas de recomendación. El primero contemplara las características del contenido, clasificando a los suscriptores en riesgo en cada uno de los

clusters de contenido identificados en la etapa I, y proponiendo los títulos más populares dentro de cada cluster. El segundo enfoque se basará en el contenido consumido por todos los suscriptores de la plataforma, para identificar estilos y preferencias de títulos para realizar una recomendación personalizada. Los resultados de cada enfoque se fusionarán con el objetivo de recomendar a cada suscriptor un contenido nuevo y atrapante, que haga cambiar su decisión de abandonar Disney Plus.

En el esquema presentado en la gráfica 9 representamos las 3 etapas analíticas (Descriptiva, Predictiva y Prescriptiva) que conforman los objetivos de esta tesis, así como también las interrelaciones entre dichas etapas.

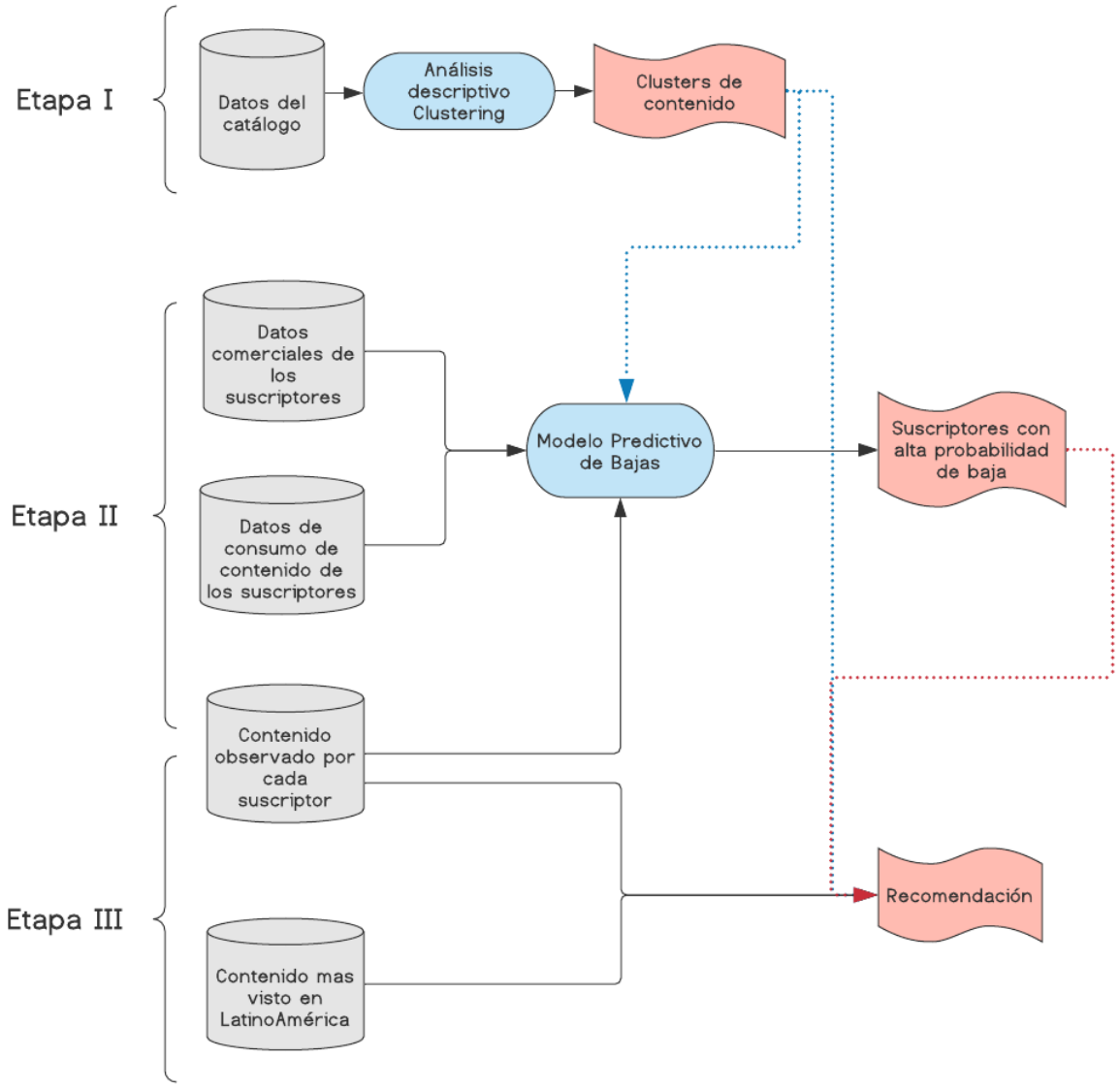

Gráfico 9: esquema de las distintas etapas del proyecto y sus procesos.

Definiciones preliminares

Datos

Se obtuvo el acceso a alrededor de 22 tablas de datos con información sobre los suscriptores y contenidos disponible en Disney+. Estas tablas son en realidad vistas generadas sobre las tablas originales, a las cuales, por un tema de permisos, no se dio el acceso directo al equipo de Latinoamérica. Entre las tablas originales y las vistas derivadas, existe un sistema de seguridad, Immuta[4], para controlar los permisos a los distintos sets de datos. Como consecuencia de esta medida tomada por seguridad, al recibir los accesos no existía documentación respecto a los datos recibidos: qué información muestra cada tabla, cuáles son sus claves primarias, relaciones existentes entre ellas, etc. Esta información era fundamental, ya sea para realizar cualquier tipo de consulta, hasta para desarrollar modelos predictivos. A la hora de trabajar con tablas de billones de datos, es fundamental este entendimiento, ya que si no se pueden generen consultas con relaciones entre tablas "*many to many*", que no solo genera resultados erróneos, sino que los tiempos de procesamiento son imposiblemente largos. La única alternativa para conocer la estructura y contenido de los datos fue explorar cada tabla, conocer las variables incluidas, qué tipo de datos era cada una, qué categorías distintas existían dentro de las variables categóricas, máximos y mínimos en el caso de las variables numéricas y las variables de fecha, conteo de nulos etc. Se identificaron las claves primarias, y las claves foráneas para entender la relación entre las tablas.

Este primer desafío que se enfrentó, es una clara consecuencia de una empresa que está dando sus primeros pasos en la democratización de los datos, entre los distintos equipos en las regiones. Fue importante que, tras terminar con la exploración, se realizó toda la documentación pertinente respecto a las tablas a las que se accedieron. Al final del documento se incluyó un apéndice donde se puede ver el detalle de las tablas que se utilizaran durante este trabajo y las variables utilizadas en cada una de ellas.

Softwares

Los datos están almacenados en Snowflake[5], una plataforma de datos construida en la nube, dentro de un ambiente perteneciente al equipo de Estados Unidos. Los accesos dados al equipo de Latinoamérica, permiten solo hacer consultas, y escribir los resultados en un espacio de almacenamiento externo. Estas acciones, son muy limitados para realizar las trasformaciones necesarias para el proyecto, pero el tamaño de los datos no permite trabajar localmente con la memoria de una computadora personal. Así, se decidió extraer los datos de interés y almacenarlos en un repositorio en la nube

[4] https://www.immuta.com/
[5] https://www.snowflake.com/

de Google Cloud Storage[6]. Dentro del ambiente de Google, se generaron todas las transformaciones entre tablas necesarias con la herramienta Big Query[7]. En este mismo entorno, se entrenaron los modelos de aprendizaje automático, a través de Big Query ML[8]. Este procedimiento que se definió entre softwares para trabajar con los datos de Disney+, fue una solución local para poder realizar tareas de data science, y business intelligence, pudiendo dar apoyo a las necesidades locales. Sin embargo, no es la manera óptima ya que se termina duplicando el almacenamiento de datos generando costos adicionales.

Códigos utilizados

Este proyecto fue desarrollado a través de los lenguajes de programación SQL y R. Los distintos códigos están disponibles en el siguiente repositorio de *Git Hub.*

https://github.com/julisantarelli/Tesis_MIM/

Confidencialidad

Para Disney es muy importante la seguridad de los datos. Por eso en este trabajo no se incluyen datos confidenciales ni que puedan resultar en una inferencia del dato real.

[6] https://cloud.google.com/storage/
[7] https://cloud.google.com/bigquery/
[8] https://cloud.google.com/bigquery-ml/

Etapa I

Análisis y clusterización de contenidos

En esta etapa se trabajará con el contenido disponible en Disney Plus para los suscriptores de Latinoamérica. Se utilizará el término título o contenido haciendo referencia a una película, un cortometraje o una serie, sin ir al detalle de capítulo o temporada. Primero se buscará consolidar la mayor cantidad de características posibles sobre los datos en una sola tabla. Luego se trabajará sobre un análisis descriptivo, que culminará con la identificación de grupos homogéneos de contenido, a través de un modelo de clustering.

Preparación de los datos

Todos los datos acerca de los títulos de Disney Plus, fueron recibidos por parte del equipo norteamericano, ya que los accesos a los mismos en *Snowflake*, todavía no están dados. Los mismos no tienen dimensiones extremadamente grandes, lo que permitió que toda su limpieza y transformación de forma local a través de R Studio, es decir, sin procesamiento en la nube. En el repositorio de código se puede encontrar el detalle del trabajo de pre-proceso de la información en el archivo: E.1 a R Contenidos. Como resultado se obtuvo una tabla final cuyas características se puede resumir en la tabla 1 a continuación:

Categoría	Descripción	Ejemplo
Relacionadas a los negocios de Disney	Estas variables describen la relación entre el título y distintos aspectos del negocio.	Marca
Relacionadas al género	Conjunto de variables binarias que valen 1 cuando el título presenta características de ese género y 0 en caso contrario.	Drama
Relacionadas a Disney Plus	Variables que describen aspectos del título dentro de la plataforma Disney +	Disponibilidad en Brasil.
Propias del título	Características puntuales de cada título	Año de lanzamiento
Categorizaciones del contenido.	Otras categorizaciones de contenido	Clase de contenido (serie, película o cortometraje)

Tabla1: Se describen las principales variables de contenido

Para el detalle completo de la tabla consultar el material *Apéndice Tabla 2.*

Análisis descriptivo

Actualmente la plataforma cuenta con un total de 1103 títulos. Como se ve en el Gráfico 10, el 64% de los títulos son películas. Como se ve en rojo, hay mucho contenido animado, que generalmente está relacionado con contenido infantil, lo que podría llegar a limitar la audiencia.

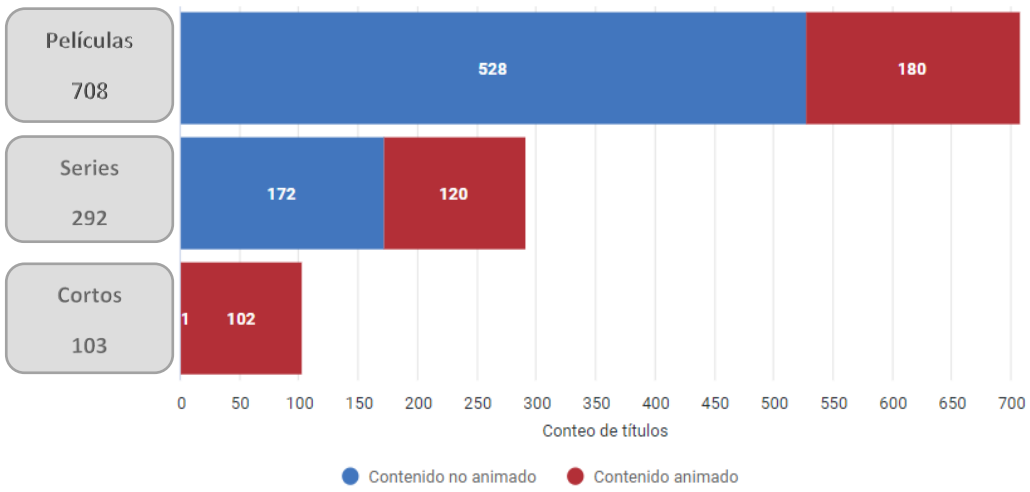

Gráfico 10: Distribución del total de títulos por tipo de contenido y la proporción de contenido animado y no animado dentro de cada clase.

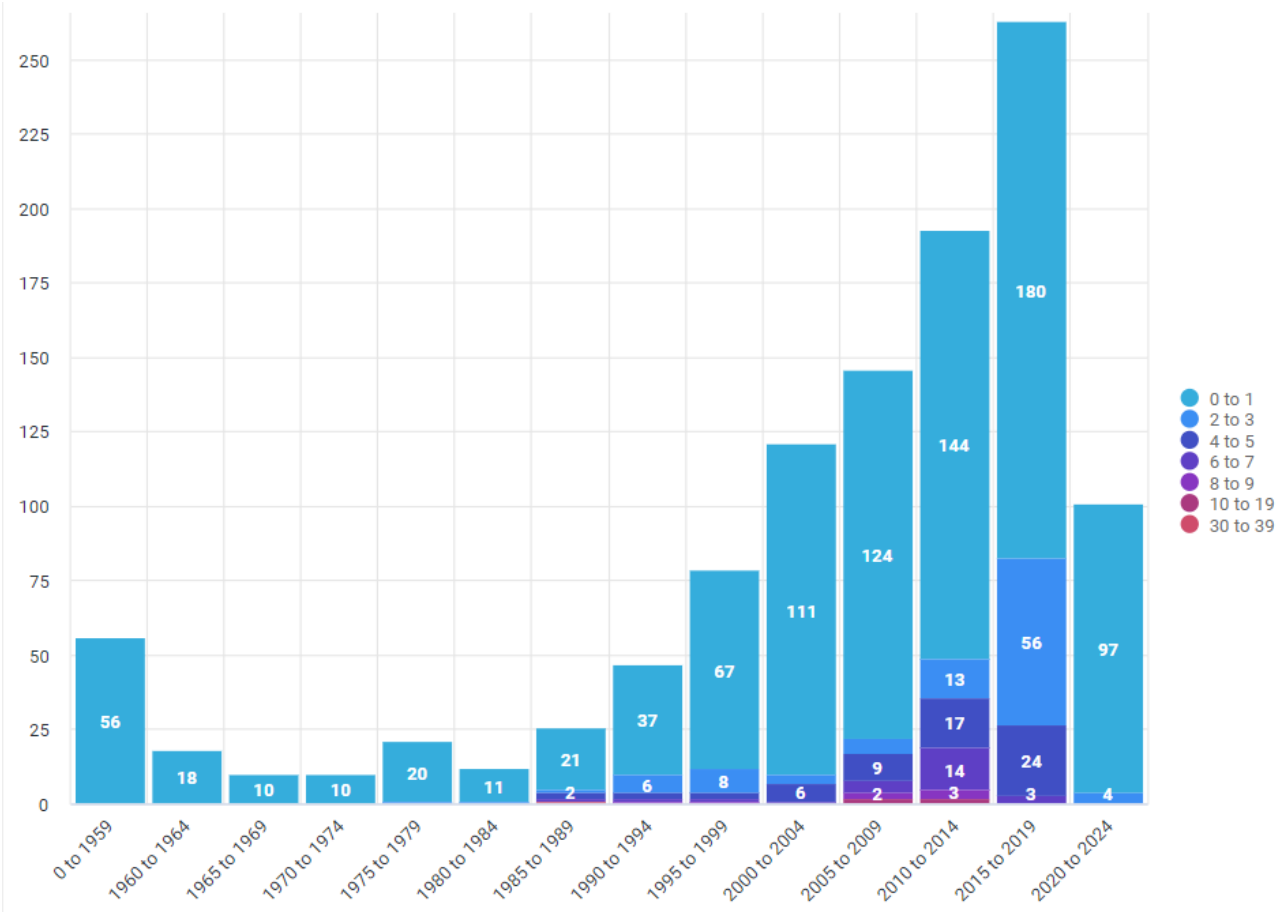

Gráfico 11: Distribución de títulos según su año de lanzamiento. A su vez, se descompone la cantidad estrenada en cada año, por la duración, en años, de cada contenido,

Observando en el Gráfico 11 la distribución de los títulos en función de su año de lanzamiento, se puede identificar que hay diversidad en este aspecto. Si bien la mayoría se lanzó a partir del año 2000, existe un 25% de títulos más antiguos. Esta característica del catálogo puede apuntar a un contenido más nostálgico, pudiendo conquistar audiencia apelando a los sentimientos de los suscriptores. Además, podemos identificar que las series, que cuentan con más de 1 temporada, son las más recientes, concentrándose a partir del 2010, coincidiendo con la época en la que arrancan las plataformas de streaming.

Se analiza a través del Gráfico 12, la cantidad de producciones originales pensadas directamente para Disney+. Las mismas representan solamente un 8% del catálogo, y son en su mayoría series. Este punto es importante ya que las producciones originales, son generalmente utilizadas para promocionar la plataforma, y el hecho de que sean tan pocas puede jugar en contra, ya que una vez que se logró traer al cliente al, este puede verse decepcionado con el resto del contenido que fue creado con otro

objetivo. Por ejemplo, es probable que las demás películas de *Marvel* producidas para generar record de ventas en cine, ya hayan sido vistas por muchos de los usuarios que se suscribieron atraídos por la serie de *WandaVision.*

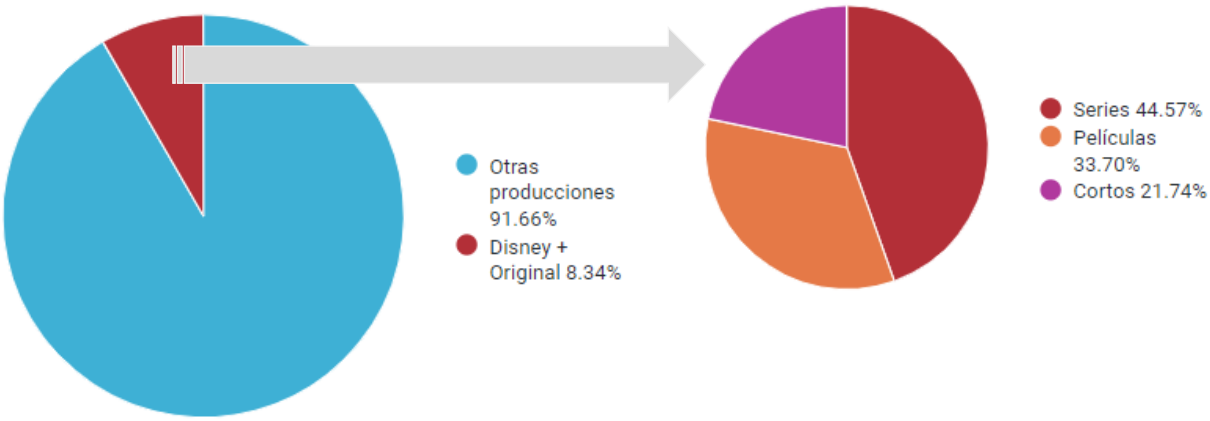

Gráfico 12: A la izquierda se observa la proporción del total de títulos que son originales de Disney Plus y la derecha el desglose de los mismos entre tipo de contenidos.

A continuación, se analizan los títulos en función a los negocios de Disney. Como se describió antes, en los últimos años se llevaron a cabo adquisiciones millonarias para ofrecer una mayor diversidad de contenido, pudiendo alcanzar una audiencia más diversa. En el Gráfico 13 se observa, que, de todos modos, el catálogo de Disney+ ofrece el 63% de títulos de la marca Disney. También se puede ver cómo las marcas adquiridas más recientemente, tienen menos contenido animado que las adquiridas hace más tiempo. Esto se condice con las trasformaciones en las estrategias que se comentaron en la introducción, tratando de no limitarse a contenido exclusivo infantil. Por ejemplo, en los títulos de Star y National Geographic, ambos provenientes de la adquisición de Fox, se observa que no hay contenido animando alguno.

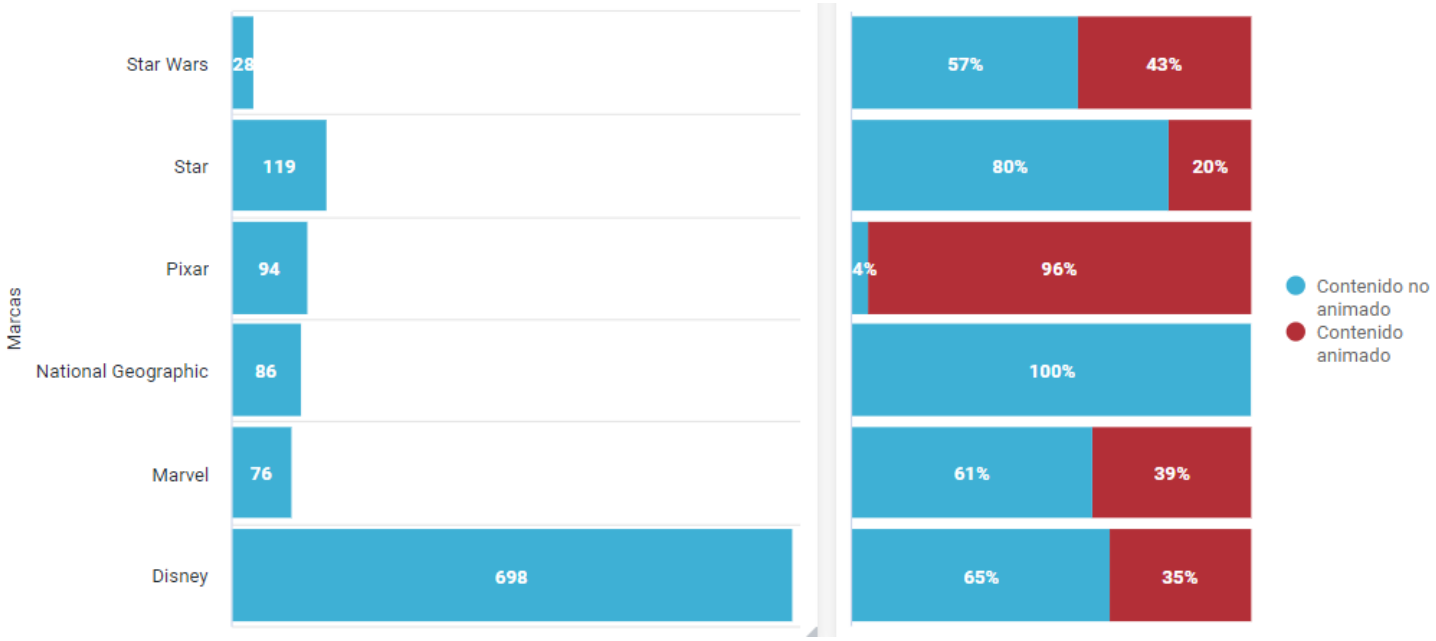

Gráfico 13: A la izquierda se observa la distribución de títulos en función de las marcas de la compañía, y a la derecha la proporción de contenido animado dentro cada una.

Por último analizaremos la variedad de títulos en base a los géneros. Un título puede tener mas de un género. Se observa en el Gráfico 14, que muchos de los géneros disponibles son acordes a una audiencia familiar, por ejemplo los géneros familia, comedia, animacion, fantasia, niños, adolescencia, etc. Para un detalle adicional analizaremos los 10 géneros mas frecuentes dentro de las marcas a traves del Gráfico 15. Existe una similitud en los géneros de las marcas Marvel y Star Wars. También son muy similares los géneros dentro de los títuos Disney, Pixar y Star, mientras que los contenidos de National Geographic son los mas diferentes dentro del catálogo.

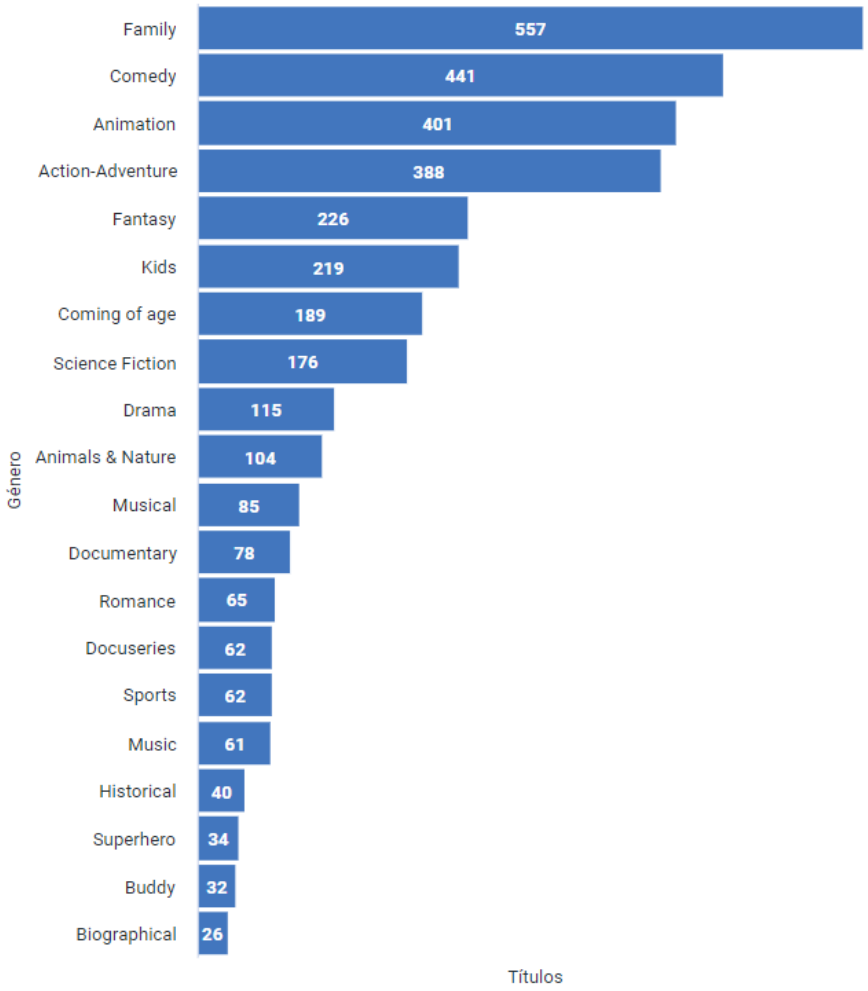

Gráfico 14: Cantidad de títulos disponibles en Disney+ por géneros.

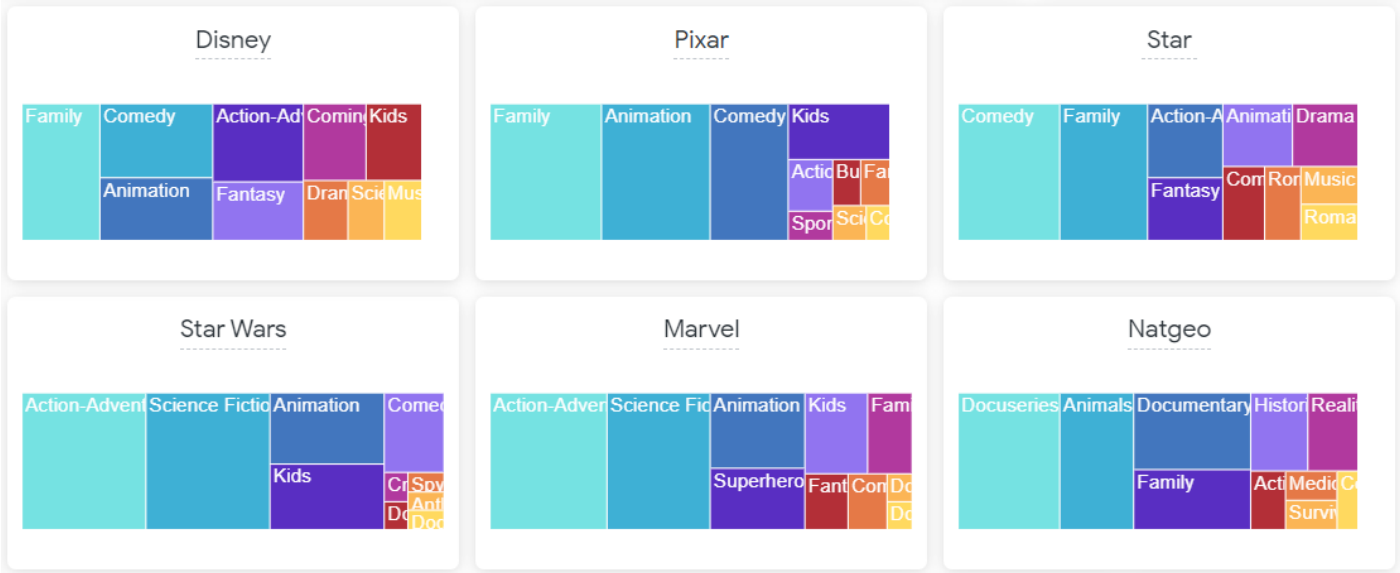

Gráfico 15: Se muestran los 10 géneros mas frecuentes por marca.

En este punto es posible responder a una de las preguntas realizadas en los objetivos, ¿Existe diversidad de contenido en el catálogo de Disney Plus? A juzgar por el análisis descriptivo, no se considera que haya mucha diversidad. Si bien se identificaron algunas diferencias importantes, como un 25% de los títulos lanzados previos al año 2000 y una marcada diferencia en los géneros del contenido de Star Wars, Marvel y National Geographic; hay una vasta mayoría de títulos Disney. Históricamente esta marca tuvo un contenido muy definido, dirigido a un público más infantil, familiar y apto para todo público. Como consecuencia, el catálogo termina siendo bastante homogéneo. Se entiende que esto es parte de la estrategia de la compañía, ya que en las demás regiones existen otras plataformas de streaming lanzadas con posterioridad a Disney+, con contenido completamente diferente, por ejemplo, ESPN+ con deportes o Hulu con contenido adulto. En comparación con las demás plataformas del mercado, la oferta de Disney+ es muy acotada. Esto puede entenderse como parte de la estrategia de segmentar contenido en distintas plataformas y ofrecerlas más tarde en un paquete promocional, como se realiza en Estados Unidos. En el Gráfico 16 se ve la cantidad el contenido que ofrece cada servicio de streaming en diciembre del 2020 dentro de Estados Unidos, Disney+ tiene una oferta escasa en comparación con los principales competidores, Amazon Prime y Netflix. De todas maneras, al combinarse con la oferta de contenido de Hulu, propiedad también de la misma compañía, la oferta es más balanceada.

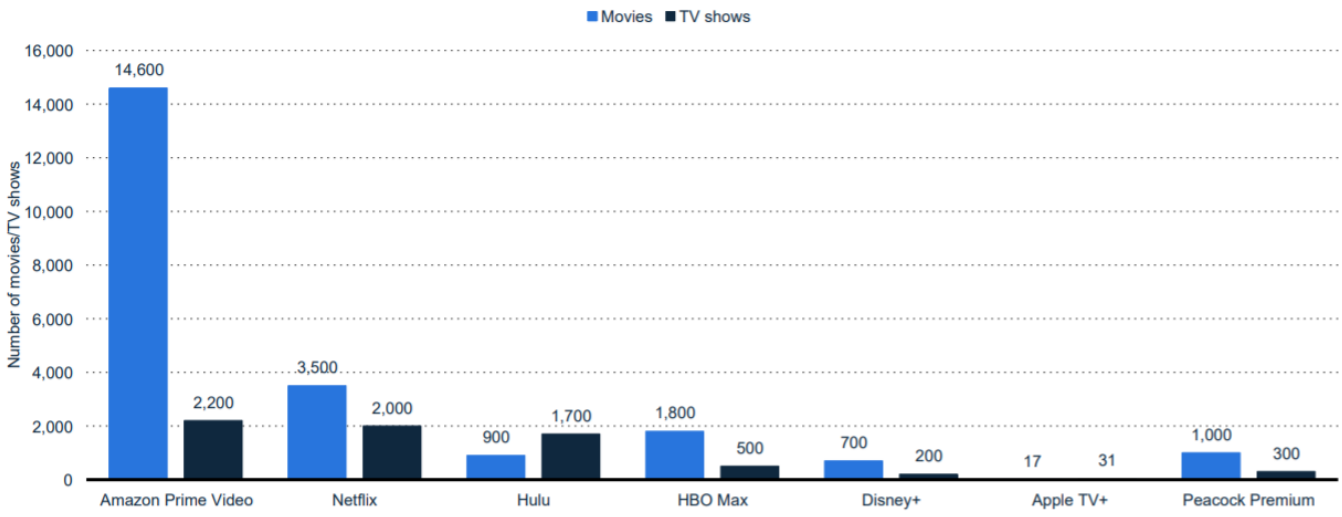

Gráfico 16: Contenido disponible en las principales plataformas SVoD en los Estados Unidos a diciembre de 2020, por servicio. Fuente: Reelgood

Modelo de Clustering K-Medias

Actualmente la única forma de segmentar los títulos, es clasificándolos por los atributos de los mismos, por ejemplo, series vs. películas, o títulos de Disney vs. títulos de Star Wars. El objetivo de esta etapa es identificar grupos de contenido homogéneos, en base a varios atributos, y sin ser ninguno excluyente. Esta clasificación tiene más sentido en el mundo real. Por lo general, los suscriptores suelen tener afinidad por un estilo de contenido que puede encontrarse en distintas marcas, productoras, formatos, etc. Para ello, se usará el algoritmo K-Medias, perteneciente al terreno del aprendizaje automático no supervisado que permitirá encontrar cierto número clusters. Estos clusters serán un conjunto de títulos de Disney+ tal que:

- Los títulos dentro del mismo cluster sean lo más similares entre sí.
- Los títulos de distintos clusters sean lo más distinto posibles.
- Cada título quede asignado a uno y solo un cluster.

En el Anexo de modelos punto 1 se detalla todo lo relacionado al algoritmo utilizado y la implementación del mismo. Cabe destacar que, atendiendo a las prácticas habituales, mediante la técnica de validación cruzada, se determinó que el parámetro K a utilizar es 10. Es decir, la división optima entre títulos, se alcanza generando 10 clusters.

Resultados

A continuación, mediante las tablas 2 y 3, se describen las principales características de los clusters resultantes. En los casos en que no existía una homogeneidad en el cluster respecto a una variable, se dejó en la tabla el campo vacío.

Cluster	Títulos	Género 1	Género 2	Género 3	Contenido Animado	Tipo	Marcas	Lanza-miento
1	55	Acción y aventura	Ciencia Ficción		No	Series y Películas	Marvel, Star Wars	Desde 2005
2	25	Acción y aventura	Ciencia Ficción		Si	Series y Películas	Marvel, Star Wars	Desde 2005
3	32	Comedia	Familia	Docu-series	No	Series y Películas	Natgeo, Disney, Star	Desde 1990
4	193	Documental	Animales y Naturaleza	Musical	No	Series y Películas	Natgeo, Disney, Star	Desde 2005

5	96	Animación	Animación	Comedia	Si	Películas y Cortos	Pixar	
6	104	Familia	Animación			Películas y Cortos	Disney	1920-1980
7	92	Niños				Series	Disney	
8	192	Familia	Acción y aventura	Comedia	No	Películas	Disney	Desde 1975
9	152	Comedia	Adolescencia		No	Series y Películas	Disney	
10	162	Animación	Animación	Acción y aventura	Si		Disney	

Tabla 2: Clusters resultantes del modelo K-medias. Descripción de las principales características.

En base a las características de cada grupo, se propone una etiqueta para cada uno, asumiendo una audiencia posiblemente interesada en el contenido.

Cluster	Etiqueta	Descripción
1	Acción y Universos ATP	El contenido tiene personajes no animados que encaran aventuras en universos ficticios, como lo es Marvel o Star Wars. No es contenido exclusivamente infantil como la mayoría del catálogo, sino que puede atraer a todo tipo de edades (ATP). Ejemplo: Black Panther.
2	Acción Infantil	El contenido tiene personajes 100% animados que encaran aventuras con mucha acción. Por su formato y tramas más aniñadas se lo considera apropiado para niños. Ejemplo: The Super Hero Squad.
3	Comedias en Familia	Contenido no animado entretenido para todas las edades, lanzados a partir del 1990, ideal para ver en familia. Ejemplo: Doctor Doolittle.
4	Entretenimiento con aprendizaje	En su mayoría son títulos no animados, relacionados a distintas áreas de aprendizaje. Por ejemplo: geografía, musicales, documentales, detrás de escenas, etc. Ejemplo: Bios
5	Animación ATP	Se incluyen la gran mayoría de películas Pixar, donde las historias animadas suelen ser apropiadas para todas las edades. En este grupo también se incluyen todos los cortometrajes de este estilo. Ejemplo: Monsters University.
6	Nostalgia	Contenido más antiguo. Acá se encuentran las historias y cuentos cortos más famosos de la marca Disney, con los personajes más legendarios. Ejemplo: Bambi.
7	Niños	Series pensadas para los más pequeños, proveniente en su mayoría de los canales Disney Junior y Natgeo Kids. Ejemplo: Morko & Mali.
8	Dramas Adolescentes	Se incluyen películas no animadas, lanzadas en los últimos 40 años. Pueden ser ideales para ver en familia, incluyendo en sus géneros comedia, drama y acción y fantasía. Ejemplo: Annie.

9	Problema de la Adolescencia	Contenido no animado, que cuentan historias de jóvenes atravesando problemas cotidianos de la adolescencia. Ejemplo: BIA.
10	Nostalgia Moderna	Se incluyen todos los títulos basados en las historias tradicionales de Disney, por lo general re-hechos en producciones modernas. También las películas más taquilleras de Disney para la familia, lanzadas en los últimos años. Ejemplo: Frozen.

Tabla 3: Etiquetas y descripciones propuestas sobre cada cluster en relación a una audiencia asumida.

Oferta y demanda

Luego del análisis realizado en la oferta de contenido, vale la pena analizar la relación de estos grupos con la demanda. En el gráfico 17 se combina la distribución de la oferta de títulos entre los clusters, y la cantidad de 'complete streams' que ha tenido cada uno (obtenidos del historial de contenido visto por los suscriptores de Latinoamérica en cierta fecha, pasados 6 meses del lanzamiento de la plataforma).

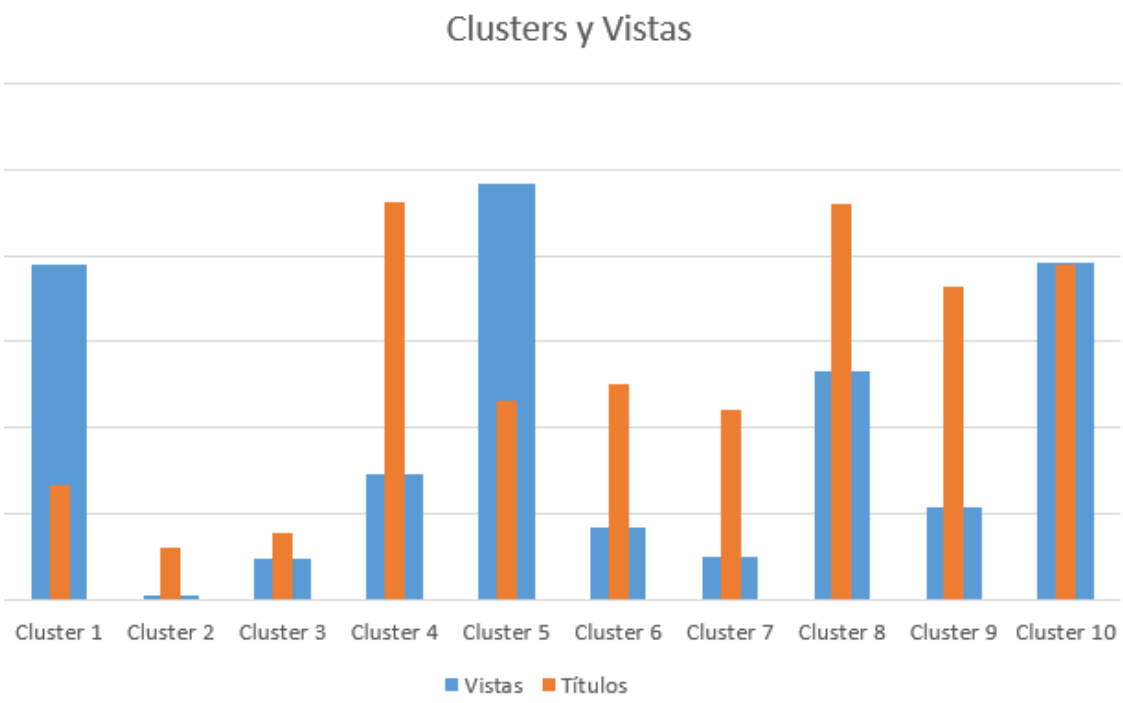

Gráfico 17: se presentan sobre a escala del eje Y derecho, la distribución de vistas recibidas por los distintos clusters, en una diferente escala sobre el eje Y izquierdo, la distribución de títulos entre los clusters. Las escalas han sido removidas por un tema de confidencialidad, pero no hacen a la interpretación del balance entre la distribución de oferta y la distribución de demanda.

En base a los clusters identificados, no es proporcional la demanda con la cantidad de títulos ofrecidos de cada grupo. Es posible identificar ciertos grupos de contenido que tienen muchos títulos para ofrecer, y no son elegidos por los suscriptores.

Por ejemplo, en el caso del cluster 4, “Entretenimiento con aprendizaje”, no existen muchas visualizaciones en comparación con el tamaño de la oferta. Sería interesante preguntarse si es por falta de marketing, o por falta de interés en la audiencia. Al haber asumido un posible segmento para este cluster, se podrían pensar acciones dirigidas al mismo, para conocer el motivo de esta poca demanda. Este análisis también permite entender dónde está el mayor interés de los suscriptores para las próximas decisiones de adquisición de contenido, viendo el gráfico, valdría la pena incorporar contenido del tipo del cluster 1 y 5. Además, no sería importante en estos casos realizar muchos esfuerzos de marketing ya que la audiencia ya se encuentra en estos grupos, “preparados” para mayor contenido.

Conclusiones

El trabajo realizado en esta sección alcanzó excelentes resultados. A través del análisis descriptivo se logró entender la oferta de contenido que ofrece el producto Disney+ y comprender que la homogeneidad de contenidos observada, está alineada con la estrategia de la compañía en relación a sus productos de streaming. También se observó cómo los títulos de aquellas marcas adquiridas en los últimos años, permitieron expandir la audiencia, sin perder la esencia de la marca. A través de los títulos de Marvel, Star Wars, Natego y Fox (Star), se despertó el interés sobre un segmento jóvenes-adultos; que antes no se interesaba por las historias Disney. Sin embargo, se logró mantener cierta línea, es decir, el 100% del contenido permite verse en familia, es decir, en un grupo de personas de distintas edades, contando historias atrapantes que apelan a los sentimientos para ser memorables.

Identificar clusters de contenido similar, permite un enfoque adicional a la hora de trabajar con contenido. Más adelante en este trabajo se utilizarán los clusters para generar variables de consumo de los suscriptores (etapa II) y para realizar las recomendaciones de contenido (etapa III). El breve análisis de oferta y demanda planteado en los resultados da pie a múltiples y diversos análisis que exceden el foco de este trabajo. Lo se quiso mostrar es la potencialidad de usos de la categorización generada. Desde segmentar campañas de marketing o tomar decisiones sobre nuevas adquisiciones, buscando que el negocio empalme la oferta con la demanda y logre optimizar todos sus recursos.

Respeto a la implementación práctica de esta sección, dado el inmenso interés por parte de los equipos de producto de Disney+ respecto al seguimiento y conocimiento del catálogo, todo el análisis descriptivo y el detalle de los clusters, se expuso en reportes automatizados y dinámicos a través de la herramienta Looker[9]. Los equipos recibieron acceso a los mismos, y los convirtieron en una herramienta más a la hora de trabajar con el contenido. El catálogo va cambiando con el paso del tiempo,

[9] https://looker.com/

sumando nuevo contenido, y descartando otro tanto. Respecto a los clusters, definidos en base al contenido actual, se decidió mantener los mismos, siempre y cuando las variables descriptivas se mantengan. Al agregarse un nuevo título a la plataforma, se definirá qué cluster tiene características similares al mismo, y se lo asignara manualmente.

Etapa II: Predicción de bajas

En esta segunda etapa se trabajará a nivel de suscriptor, y se buscará predecir aquellos que se darán de baja de Disney+ de forma voluntaria en los próximos 30 días. Las variables que se usaran se pueden agrupar en atributos comerciales y atributos de contenido. Se incluirán los resultados de la etapa I, en la creación de variables de contenido. Se usarán muchos términos propios del negocio y su estructura de datos, para lo cual se incluyó una sección de definiciones al final del documento. En determinadas ocasiones se incluirán las definiciones en esta sección, considerándolo necesario para una mejor compresión. Se usarán los términos en su idioma original, inglés, pero se podrá buscar la definición respectiva en español en la sección mencionada.

Datos

Para esta sección se extrajeron datos de 6 tablas distintas, detalladas en el anexo (desde Tabla 2 hasta Tabla 7). La extracción y transformación de los mismos se realizó con los códigos guardados en el repositorio de Git Hub con los prefijos E.2.ET. y E.2.FE.

El alcance de los datos está definido por las cuentas de Disney Plus que, 30 días previos a la fecha de entrenamiento, cumplan con las siguientes características:

- Pertenezcan a Latinoamérica.
- Sean *primary account*.
- Su estado de suscripción sea *paid* o *paid_grace*
- Hayan contratado el servicio directamente a través de Disney, no mediante socios.
- Tengan que renovar el pago de su suscripción dentro de los próximos 30 días.

Variable dependiente

La variable a predecir CHURN es una variable dummy que vale 1 cuando un suscriptor:

1. Pasa de estado *paid* a estado *churned* o *paid cancel* durante los últimos 30 días previos a la fecha de entrenamiento.
2. Es *primary account*.
3. Su estado previo no es *paid_hold*

La variable vale 0 en caso contrario.

Primary Account

Un suscriptor puede tener más de una suscripción. Se identifica como *primary account* a la suscripción con mejor estado. Por ejemplo, Juan tiene una suscripción de Disney+ mensual, y decide cancelarla para suscribirse a un plan anual. A la fecha, la cuenta de Juan tendrá dos

suscripciones, una dada de baja y otra activa. Se tomará esta última suscripción como *primary account*, y será la que se tome definir el estado de Juan. En la muestra de entrenamiento definimos que siempre se seleccione la *primary account* del suscriptor, para evitar incluir estados erróneos de las cuentas por suscripciones secundarias.

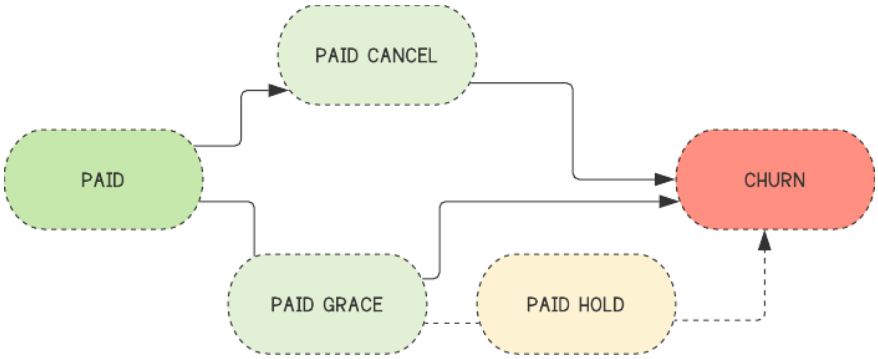

Gráfico 18: Simplificación del flujo de estados de una cuenta de Disney Plus.

Para la definición de cada estado consultar la sección Definiciones de Estado. En el gráfico 18 se ve hace una simplificación en la transición de estados que puede atravesar una cuenta. Cuando no es posible llevar a cabo los cobros de la suscripción, la cuenta en estado *Paid* pasa automáticamente a estado *Paid Grace*. En este plazo al usuario se le notifique la imposibilidad de cobro para que rectifique el método de pago. El mismo mantiene todos sus accesos al contenido de la plataforma. Si la imposibilidad de cobro persiste, pasado cierto plazo, la cuenta pasa al estado *Paid hold*, donde ya no tiene acceso al contenido. Finalmente, pasado ciertos días en estado de *Paid hold*, la cuenta queda automáticamente dada de baja, con estado *churned*.

Baja voluntaria:

Se considera que los casos de churn incluidos en la muestra son voluntarios. Se busca trabajar únicamente con las cancelaciones originadas por la disconformidad del producto y no en problemas de cobro. Para esto, excluimos los casos de churn, cuyo estado previo es *Paid hold*.

Balance de la muestra en función de la variable dependiente: Se cuenta con una muestra de entrenamiento desbalanceada, con solamente el 6% de suscriptores con CHURN = 1.

Variables independientes

Para definir las variables independientes se seguirá la siguiente estrategia:

1. Se incorporarán en los datos de entrenamiento todas las variables que se intuya relevantes respecto al consumo de contenido o a las características comerciales de un

suscriptor. Estas variables vendrán de tablas existentes, o se generarán en la combinación de las mismas, como se detallará en la sección de ingeniería de atributos.

2. Una vez entrenado el modelo, se analizará qué variables fueron importantes en el modelo para clasificar a un suscriptor. Las variables no utilizadas se descartarán, ya que de incorporarlas incrementarán sin sentido el esfuerzo de procesamiento.

Las variables independientes se pueden como se muestran en la tabla 4:

Categoría	Descripción	Tablas de Origen
Características de la suscripción	Atributos de producto y comerciales	2. Suscriptores
Características de uso del producto	Atributos que muestran el total de veces que un suscriptor uso el producto con determinadas acciones. Cada uno se repite para las ventanas de tiempo: últimos 7 días, últimos 28 días y tiempo transcurrido desde el inicio de suscripción.	3. Dispositivos 5. Consumo de producto
Características de consumo de contenido	Atributos que muestran el consumo de un suscriptor respecto a contenidos específicos o categorías de contenido. También se incluyen diversos atributos en relación al primer título consumido.	4. Marcas 1. Contenidos 7. Total Contenido Visto 8. Ranking de Contenido Etapa 1. Clusters
Características del primer contenido consumido	Diversos atributos en relación al primer título consumido al crearse la suscripción.	4. Marcas 1. Contenidos 6. Primer contenido 8. Ranking de Contenido Etapa 1. Clusters

Tabla 4: grupos de variables independientes en los datos de entrenamiento

Ingeniería de atributos

Combinaremos variables para generar variables más complejas, que muestran los grados de interacción entre ellas. Para mayor detalle ver la sección de Transformación de Variables en el Apéndice.

Características de la suscripción:

Cálculo de períodos: Con las variables del tipo fecha, se generarán distintos cálculos para quedarnos con la información de tiempo transcurrido entre distintos sucesos. Luego, se descartarán las variables del tipo fecha utilizadas.

Características de uso del producto:

Promedios diarios: Se crearán nuevas variables que calculen el promedio diario de *streams* y *stream-time* para poder comparar la intensidad de *streams* a lo largo de la suscripción, entre cuentas disímiles en cuanto a antigüedad.

Ratio días de streams: Se calculará qué porcentaje del total de días de la suscripción, la cuenta generó un *stream.*

Variación en períodos: Se calculará la variación en las variables distintas variables de uso, en el último mes sobre el total de la suscripción. También, la variación de las mismas en la última semana sobre el último mes. Se espera que aquellas cuentas con alta probabilidad de baja, muestren variaciones negativas en los últimos períodos.

Ratio stream vs. login: Se considera importante las veces que un suscriptor entra a la plataforma (hace un login) pero no ve ningún contenido (stream). Se podría entender que no encontró nada interesante y decidió buscar en otra parte. Para captar esta acción, calcularemos las ratios entre streams y logins para la última semana y el último mes.

Variación en la ratio stream vs. login: Se calculará la variación de dicha ratio del último mes respecto al periodo total de la suscripción, y de la última semana respecto al último mes.

Todas estas variables que trabajan con ventanas de tiempo son muy útiles en los casos que la antigüedad del suscriptor es mayor a un mes, para poder captar variaciones significativas en última semana, último mes y desde el inicio. Sabemos que muchas de las cuentas que se dan de baja lo hacen dentro del mes de suscripción. Esto podría quitar valor a estas variables a la hora de predecir una baja.

Características de consumo de contenido:

Tiempo destinado a cada formato de contenido: Se calculará el tiempo en horas que cada cuenta dedicó a sintonizar contenido de distintos formatos (película, serie, cortos, promocionales, etc.) Además, se calculará el porcentaje de tiempo destinado a cada uno, y se creará una variable que muestre que formato es el más consumido por la cuenta (en base a los porcentajes calculados).

Cluster de contenido: Se analizará el contenido consumido por cada cuenta, para clasificar a cada una en uno de los distintos clusters identificado en la Etapa 1. Una cuenta pertenecerá al cluster de contenido del cual haya visto la máxima cantidad de títulos.

Variables de conteo de 'complete streams' sobre distintos tipos de contenido: Se generarán variables que sumen la cantidad de 'complete streams' generados por cada cuenta, sobre grupos de contenidos definidos en base a distintas características. Estas características se obtienen de la tabla de Contenido utilizada en la Etapa 1.

Variables de diversidad de contenido: Se calculará el porcentaje de títulos vistos de distintas categorías de un atributo del contenido, para definir variables relacionadas a la diversidad del perfil respecto a dichos atributos. Se definirán cuotas de porcentajes para clasificar a una cuenta como 'mono', 'bi', o 'multi', respecto a ciertos atributos.

Variable de episodios por serie: Se sumará la cantidad de episodios vistos por serie, por cuenta, para alguno de los títulos más relevantes.

Variables Ranking: Se generarán variables que indiquen la cantidad de títulos vistos dentro de un *top 5* y *top 10* de contenidos más vistos en la plataforma.

Características del primer contenido consumido

Descriptivas del primer título: Se generarán variables categóricas para indicar si el primer contenido consumido por el suscriptor fue de ciertas características: a qué cluster pertenece, formato de contenido, marca, período de lanzamiento, etc.

Flags: Se generarán variables binarias para indicar si el primer contenido consumido por el suscriptor fue de ciertas características: pago, producción original para Disney+, animado, género y franquicia.

Algoritmos de clasificación: XGBoost y DART

En esta etapa se trabajará con algoritmos de aprendizaje supervisado para el desarrollo de un modelo de clasificación, donde se usarán los datos de entrenamiento para predecir la variable objetivo, CHURN. Para un mayor detalle sobre los modelos y su implementación ver el Anexo de modelos punto 2.

El algoritmo XGBoost fue elegido para esta etapa ya que es uno de los más potentes y flexibles en aprendizaje automático para predecir una variable dependiente. Permite trabajar con variables mixtas, no requiere mucho pre-proceso en los datos cuando se implementa vía árboles y puede lidiar tranquilamente con variables altamente correlacionadas. Como desventaja, no es muy explicativo respecto a la relación entre las variables independiente y la dependiente, sino que, funciona más bien como una 'caja negra'.

XGBoost pertenece a la categoría de modelos de ensamble Boosting. Los modelos de ensamble combinan predicciones de modelos más pequeños. En este caso se ensambla una secuencia de modelos de árboles simples (bastante débiles por sí mismos) y se sigue una estrategia de re-ponderaciones para obtener un estimador complejo. También se entrenará el modelo utilizando una variante de XGBoost, el algoritmo DART, que utiliza técnicas de descarte para potenciar el ensamblado de árboles. El mismo logra descartar árboles en el entrenamiento para evitar el sobreajuste y prevenir la construcción de árboles triviales. (Rashmi, Gilad-Bachrach). Luego de analizar la performance de ambos algoritmos, seleccionaremos aquel que reporte mejores resultados.

Hiperparametros

El algoritmo al ser tan complejo, permite ajustes a través de múltiples hiperparametros. En la tabla 5 se detallan aquellos más relevantes. Los valores utilizados en el modelo para los hiperparametros fueron seleccionados utilizando técnicas de validación cruzada.

Parámetro	Descripción
MIN_TREE_CHILD_WEIGHT	Peso mínimo necesario en un hijo para generar un corte. Rango: [0; ∞]
MAX_TREE_DEPTH	Profundidad máxima de un árbol. Rango: $[0,\infty]$
ALPHA	La cantidad de regularización L1 aplicada. Rango: [0; 1]
LAMBDA	La cantidad de regularización L2 aplicada. Rango: [0; 1]
ETA	Proporción que aprende de cada árbol. Rango: [0; 1]
nrounds	Cantidad de árboles a construir. Rango: (0;∞]
sample_type	Exclusivo DART. Tipo de muestreo del algoritmo. 'Uniform' descarta arboles al azar mientras que 'weighted' descarta los árboles en proporción al peso.
normalize_type	Exclusivo DART. Tipo de normalización del algoritmo. Con 'tree' cada nuevo árbol tiene el mismo peso que cada uno de los arboles descartados mientras que con 'forest' cada nuevo árbol tiene el mismo peso que la sumatoria de todos los arboles descartados.

Tabla 5: hiperparametros más relevantes del algoritmo XGBoost

Análisis de la performance del modelo

A continuación, se describen una batería de métricas que luego utilizaremos para contrastar las diferencias entre los modelos predictivos propuestos.

Precision: Mide qué proporción de los registros estimados como positivos, realmente lo fueron. A mayor precisión, menor errores de falsos positivos.

$$Precision = TP / (TP + FP)$$

Recall: Mide qué proporción de las observaciones que resultaron positivas, se estimaron como positivas.

$$Recall = TP / (TP + FN)$$

F1-score: Como ambas métricas tienen una relación inversa, es decir, a mayor Precision menos Recall y viceversa; también usaremos F1-score que pondera de manera conjunta precisión y Recall, y permite mejorar ambas tasas a la vez.

$$F_1 = \left(\frac{recall^{-1} + precision^{-1}}{2} \right)^{-1} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Estas métricas se basan en las clases definidas, basadas en la probabilidad estimada. Para clasificar a una observación como positivo o negativo, se considera cierta probabilidad de corte, conocida como *threshold*, límite mínimo de probabilidad para lo cual una observación se clasifica como positiva. Trabajaremos también con las siguientes métricas de performance, que permitirán evaluar la calidad de la probabilidad estimada, independientemente del punto de corte.

AUC ROC: Trabaja sobre la clasificación producida por la probabilidad. Una buena estimación de probabilidad debe clasificar en primer lugar a toda la minoría observaciones de clase y luego las de la clase mayoritaria. Para esto se utilizará la métrica AUC que mide el grado de separación de las clases.

La curva AUC relaciona la tasa de verdaderos positivos con la de falsos positivos. Puede tomar valores entre 0 y 1, siendo:

- 1: separación perfecta entre clases.
- 0.5: separación entre clases al azar
- Menor a 0.5: separación entre clases con performance peor que el azar.

Logg-loss: mide el rendimiento del modelo de clasificación, comparando la probabilidad estimada entre 0 y 1 para cada observación y la real. La métrica es un promedio negativo del logaritmo de probabilidades predichas versus las reales, para cada instancia.

$$L_{\log}(y, p) = -\log Pr(y|p) = -(y \log(p) + (1 - y) \log(1 - p))$$

A menor log-loss mejor es la precisión del modelo. Buscaremos un menor log-loss que aquel que se obtiene a través del azar, según el balanceo de nuestros datos:

Balanceo (1 vs 0)	*Log Loss por azar*
35% 65%	*0.6474*

Importancia de las variables en el modelo resultante

Para hacer la selección de variables, trabajaremos con distintas importancias definidas por el modelo.

Importancia por peso/frecuencia: el porcentaje que representa el número relativo de veces que una variable se usa en la construcción de un árbol. Es importante que las variables binarias no tendrán mucha importancia en este caso, ya que no se pueden utilizar más de una vez para delimitar regiones de valores.

Importancia ganancia: muestra la ganancia promedio para mejorar la predicción, que alcanza una variable en todos los cortes en los que se usa.

Importancia cobertura: Mide la cantidad relativa de observaciones afectadas por una variable.

Implementación

Undersampling y calibración de la probabilidad estimada

Los datos de entrenamiento están muy desbalanceados. Solamente el 6% de ellos son de la clase que se busca predecir. Esto puede influenciar en el aprendizaje del algoritmo, ya que, por lo general en estos casos, es incapaz de generalizar el comportamiento de la minoría y, por lo tanto, la precisión predictiva del algoritmo funciona mal.

A través de la estrategia de undersampling se toma una submuestra de la clase mayoritaria en el conjunto de entrenamiento antes de entrenar el clasificador (Akbani, Kwek, Japkowicz. 2004). La suposición detrás de esta estrategia es que en la clase mayoritaria hay muchas observaciones redundantes y la eliminación aleatoria de algunas de ellos no cambia la estimación de la clase distribución. Sin embargo, esto genera que trabajemos con una distribución distinta en entrenamiento y testeo, violando una de las normas básicas en el aprendizaje automático de que los datos de entrenamiento y testeo se extraen de la misma distribución subyacente.

Utilizaremos el método analítico propuesto por Dal Pozzolo, Caelen, Johnson y Gianluca Bontempi en su paper 'Calibrating Probability with Undersampling for Unbalanced Classification' en el 2015 para corregir el sesgo introducido artificialmente por el undersampling realizado para balancear los datos de entrenamiento. Al utilizar técnicas de rebalanceo de la base de datos, será necesario reajustar las probabilidades de fuga estimadas para cada suscriptor. A este respecto, llamando ρs a la probabilidad de fuga predicha de un modelo correspondiente a una cuenta particular, la probabilidad corregida ρ' se computa como:

$$p' = \frac{\beta p_s}{\beta p_s - p_s + 1}$$

Donde β es aproximadamente la ratio entre casos positivos y negativos en la muestra re balanceada. En este trabajo, la muestra estará artificialmente balanceada para obtener una relación 35% - 65%, por tanto $\beta = 0.53$. De esta forma, y a modo de ejemplificar lo anterior, para una probabilidad de fuga estimada con los datos re balanceados de $\rho s = 0.8$, se corresponde una probabilidad calibrada de:

$$\rho' = 0.53 * 0.80 / (0.53 * 0.80 - 0.80 + 1) = 0.6829$$

Cabe destacar que esta calibración en la probabilidad, no afectara a la calidad en la clasificación producida por la probabilidad (medida por AUC), pero sí impactara sobre las tazas de Precision y Recall. A costas de esta alteración, introduciremos una mayor varianza en las estimaciones, ya que al fin y al cabo trabajaremos con datos de entrenamiento balanceados artificialmente.

Análisis descriptivo

En esta sección se hará un análisis descriptivo de los datos de entrenamiento obtenidos luego de la ingeniería de atributos y balanceo. Se observa el comportamiento de los suscriptores dados de baja en función de distintas variables. Interesan aquellos casos donde la proporción entre 'churn' y no 'churn' rompe con la distribución de la muestra de 35% 65%, lo que indicaría que tal variable podría ser importante a la hora de predecir una fuga. Se indica la proporción de observaciones "churn" con el color violeta, y aquellas "no churn" con celeste. Para los datos de entrenamiento se utilizan aproximadamente 900 mil suscriptores. Se espera que este análisis pueda anticiparnos los resultados presentados por el modelo predictivo.

Lo primero a analizar es la relación entre la fuga de suscriptores y la antigüedad de las mismas. Anteriormente se habló de la mayor propensión a cancelar las cuentas durante los dos primeros meses. En el Gráfico 19, podemos constatar que efectivamente es mayor. Es interesante ver que si bien a mayor antigüedad, menor es la fuga, hay muchas suscripciones canceladas luego de los 60 días. Las mismas representan el 67% del total. Esto sustenta la preocupación de la gente de *life cycle*, que necesitaban que las herramientas de análisis y predicción de bajas no se limite a los dos primeros meses de antigüedad.

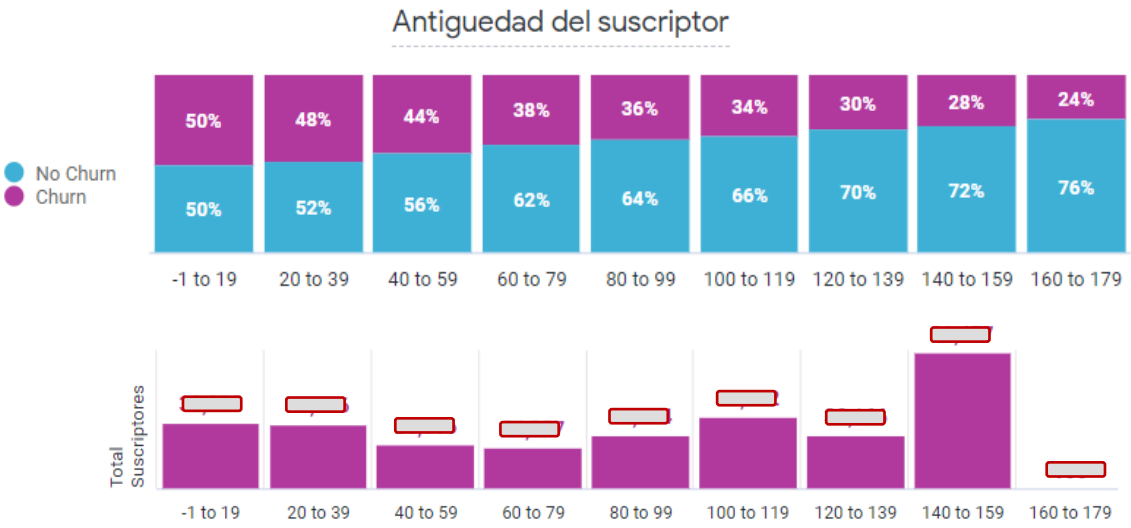

Gráfico 19: Suscriptores dados de baja (violeta) en función de la antigüedad de su suscripción. Los valores del total de suscriptores por antigüedad han sido tachados debido a su confidencialidad, pero no hacen a la interpretación de la distribución de los mismos entre los meses de antigüedad.

En relación a las variables de producto/comerciales, se observa en el Gráfico 19 la influencia de la prueba gratuita en la propensión a la fuga. Se puede interpretar que comercialmente es ventajoso ofrecer esta prueba, interpretando que los suscriptores están más contentos con el producto y por eso tienen menos tendencia a cancelar la cuenta. También hay un mayor comportamiento de fuga en aquellas cuentas que son facturadas a través de plataformas tercerizadas. Se puede ver en el Gráfico 20, cómo los casos que utilizan plataformas de facturación directas (internas de la empresa) tienen menor proporción de cancelaciones, que aquellos que utilizan plataformas intermedias, por ejemplo, Hulu, Roku o Amazon. Esto es un punto a investigar, ya que puede ser que existan fallas en las plataformas que deterioren la experiencia de usuario, o tiendan a tener problemas con el procesamiento de pagos. Si bien se excluyó del análisis los churn involuntarios, siendo los mismos ocasionados por el cese del pago, un primer intento de cobro fallido, dispara un correo electrónico hacia el suscriptor para que regularice su situación, y esto puede hacer que se replantee si quiere seguir o no pagando por la plataforma.

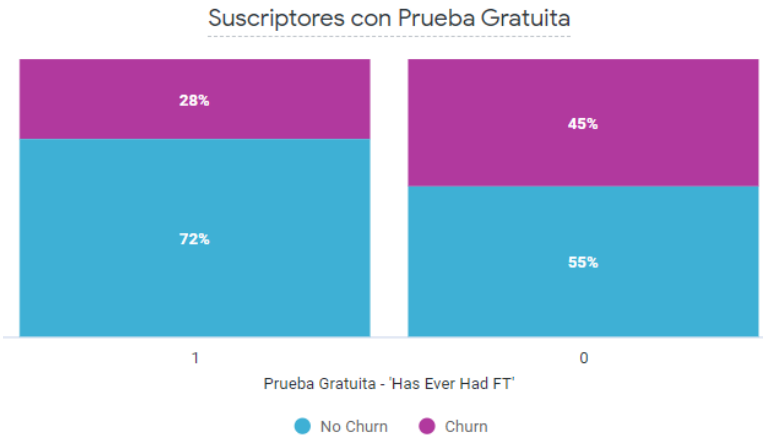

Gráfico 19: Suscriptores dados de baja (violeta) en función de si la cuenta conto con una prueba gratuita del producto (1) o no (0).

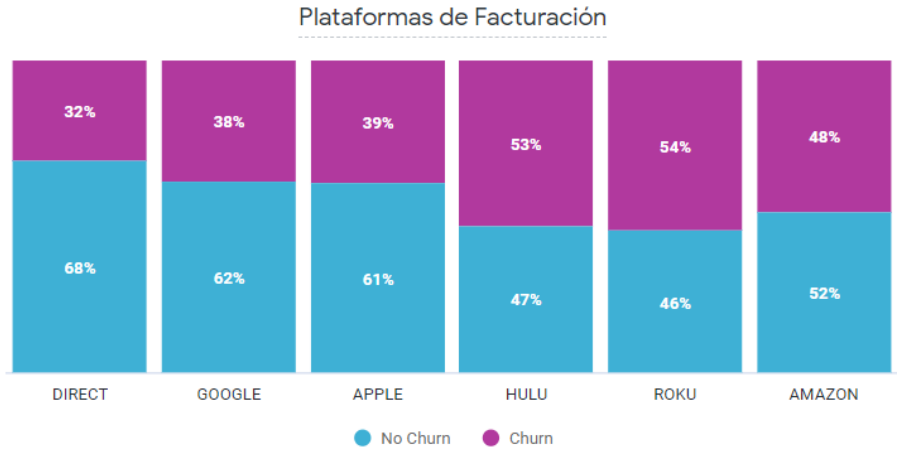

Gráfico 20: Suscriptores dados de baja (violeta) en función de las plataformas mediante las que se realizan los cobros de las suscripciones.

A continuación, se detallan algunos aspectos a destacar en relación a las variables de consumo de contenido. A través del Gráfico 21 se puede interpretar que aquellas cuentas que consumen una mayor cantidad de contenido lanzado en los últimos dos años, tienen menor propensión a la baja. Esto se puede relacionar con las producciones originales de Disney+ anteriormente mencionadas, las cuales están diseñadas pensando en un producto de streaming y fueron las más promocionadas al lanzar la plataforma.

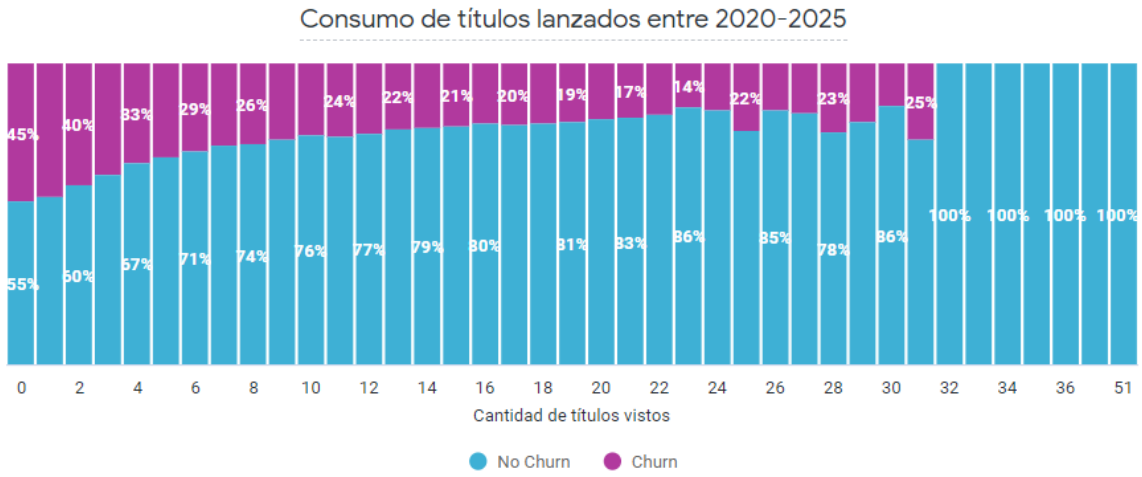

Gráfico 21: proporción de cuentas dadas de baja en función a la cantidad de títulos vistos lanzados a partir del 2020.

Por último, analizamos cómo se comporta la variable creada con los resultados de la etapa 1. En el Gráfico 22 se ve que no hay una marcada diferencia de proporción de bajas entre los distintos clusters. Cabe destacar que la mayor cantidad de suscriptores fueron clasificados en base a su consumo, en los clusters 1, 5, 8 y 10. Es interesante analizar la distribución de los suscriptores entre los clusters de contenido, ya que podría suceder que, por falta de marketing, exploración del usuario, o falla en el algoritmo de recomendación, muchos usuarios no sepan de la existencia de todo el contenido disponible, y consuman los títulos más famosos. Por ejemplo, en el cluster 1 se encuentran los títulos de Star Wars y Marvel, que fueron apalancados por el lanzamiento de nuevo contenido de estos universos como la serie WandaVision y Mandalorian. En el cluster 5 se encuentran en su mayoría las películas de Pixar, que también fueron apalancadas por el lanzamiento de nuevos títulos por ejemplo Soul y Onward, así también como por muchísimos cortometrajes creados para Disney +. En relación a la cancelación de cuentas, a juzgar por el análisis descriptivo no se espera que esta variable tengan mucha importancia en el modelo predictivo. Sin embargo, podría darse que su combinación con otras variables sí lo tenga, tema que se analizará al final de la sección.

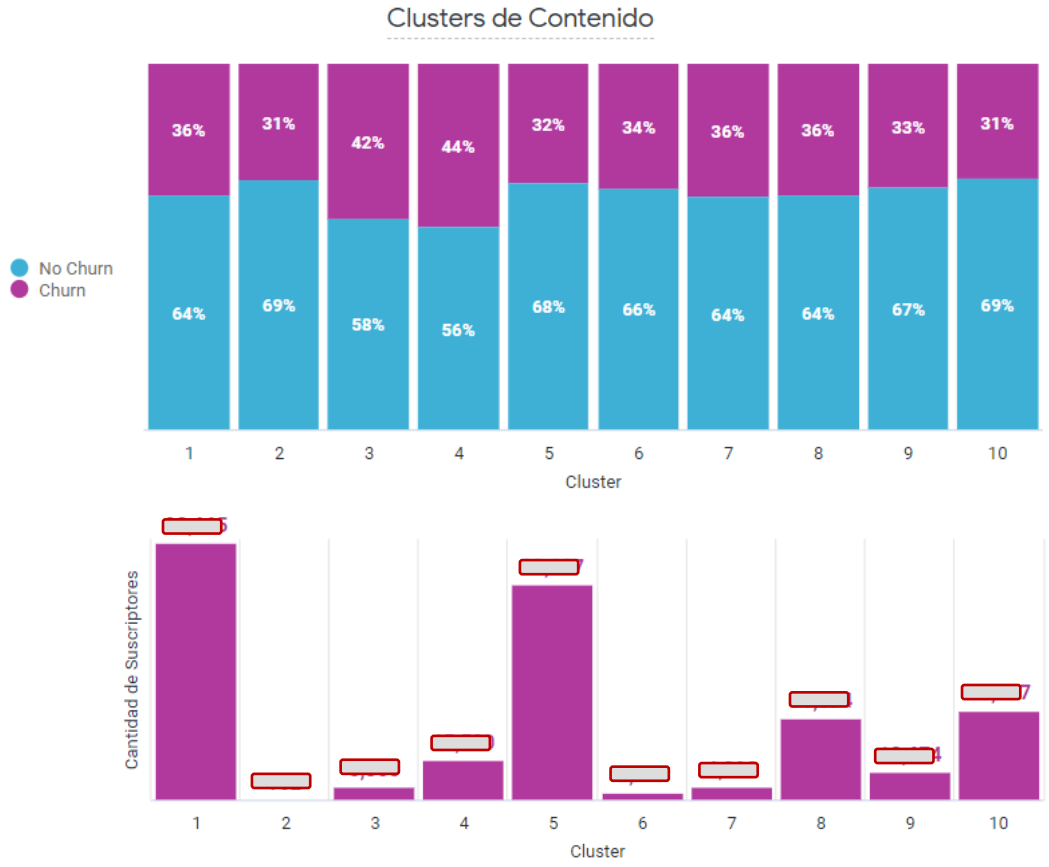

Gráfico 22: Proporción de fugas entre los grupos de suscriptores en función de los clusters de contenido definidos en la etapa 1. Abajo se puede ver la desbalanceada distribución de cuentas entre los grupos. Los valores del total de suscriptores por cluster han sido tachados debido a su confidencialidad, pero no hacen a la interpretación de la distribución de los mismos entre los clusters.

Entrenamiento y resultados

En línea con la metodología entrenamiento-evaluación-testeo, detallada en el Anexo de modelos punto 2, definiremos qué algoritmo a utilizar (XGBoost o DART) y los mejores valores para los hiperparametros en base a su performance en evaluación y testeo, buscando maximizar la AUC ROC y minimizar el Logg-loss. Una vez seleccionado el modelo, analizaremos las variables importantes para concluir la selección de variables, descartando aquellas no utilizadas. Por último, se calibrarán las probabilidades estimadas, y se analizarán los resultados obtenidos para distintas probabilidades de corte (threshold).

Resultados en Test

Métricas	XGBoost	DART
AUC	0.6849	0.7125
Log loss	0.6067	0.5840

Probabilidad de corte y Propuesta de escenarios

En esta etapa calibraremos las probabilidades para evitar el sesgo del sub-sampling. La calibración no afecta las métricas de AUC y log loss, ya que las mismas se relacionan con la dispersión de la probabilidad. (Dal Pozzolo, Caelen, Bontempi. 2015) A continuación, mediante la tabla 6, analizamos los resultados en testeo respecto a las métricas relacionadas a la matriz de confusión, para distintas probabilidades de corte. |

Threshold	*TP*	*FP*	*TN*	*FN*	*PRECISION*	*RECALL*	*F1*
0.25	1,563	1,537	3,412	1,082	0.50	0.59	0.54
0.27	1,411	1,290	3,659	1,234	0.52	0.53	0.53
0.29	1,252	1,063	3,886	1,393	0.54	0.47	0.50
0.31	1,101	873	4,076	1,544	0.56	0.42	0.48
0.33	929	686	4,263	1,716	0.58	0.35	0.44
0.35	755	527	4,422	1,890	0.59	0.29	0.38
0.37	47	13	4,936	2,598	0.78	0.02	0.03
0.39	477	298	4,651	2,168	0.62	0.18	0.28
0.41	357	217	4,732	2,288	0.62	0.13	0.22
0.43	266	152	4,797	2,379	0.64	0.10	0.17

Tabla 6: Se muestra la cantidad de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) que se obtuvieron en los resultados de testeo, dentro de un rango posible de probabilidades de corte. Para cada una, además se muestra las métricas de Precision, Recall y F1 obtenidas.

En base a métricas de Precision, Recall y F1, se proponen dos tasas de corte dependiendo de posibles escenarios en relación a las acciones a tomar con los resultados del modelo. Cuando la acción de retención es costosa, se va a querer gastar el esfuerzo en un suscriptor que seguramente termine en fuga, y no desperdiciar en un suscriptor que no corre peligro de baja. Llevado al extremo, si se ofrecen suscripciones gratuitas por cierto tiempo, a los usuarios en riesgo, si se da la promoción por error a un suscriptor "sano", no solo se pierde el costo de la acción, sino también el costo de oportunidad del ingreso que hubiera generado el cliente en ese momento, ya que en realidad no planeaba darse de baja. En otro contexto, si la acción de retención es económica, por ejemplo, la aparición de un *pop up* dentro de la plataforma, se va a intentar no invadir a aquellos que no corren riesgos de baja, pero en el caso de incluir en la acción a una cuenta que no se termina fugando, las consecuencias no son tan graves. De esta manera, se proponen dos tasas de corte, para cada uno de estos escenarios, el escenario 1, la acción costosa, y el escenario 2, la acción económica.

Para el escenario 1, la acción costosa, se buscará aumentar el Precision, sin que los verdaderos positivos sean extremadamente pocos. Seleccionamos el threshold de 0.35 que nos permitirá identificar 1 de cada 3 personas que se darán de baja en la plataforma, sin embargo, el costo de la acción será en vano solo en el 40% de los casos, ya que los falsos positivos son menos al tener un mayor threshold.

- ✓ Precision: 0.59
- ✓ Recall: 0.29
- ✓ F1-Score: 0.38

		Predicho	
		1	0
Actual	1	755	1890
	0	527	4422

		Predicho	
		1	0
Actual	1	29%	71%
	0	11%	89%

Para el escenario 2, la acción económica, se buscará aumentar el Recall, intentando buscar que los falsos positivos no sean extremadamente altos, ya que, si no, daría igual hacer la acción para todos los suscriptores. Seleccionamos el threshold de 0.27 con la que se podrá identificar 1 de cada 2 personas que se darán de baja, sin embargo, casi la mitad de la acción será en vano, ya que se incluirán muchos falsos positivos con un threshold tan bajo.

- ✓ Precision: 0.52
- ✓ Recall: 0.53
- ✓ F1-Score: 0.53

		Predicho	
		1	0
Real	1	1,411	1,234
	0	1,290	3,659

		Predicho	
		1	0
Real	1	53%	47%
	0	26%	74%

Selección de variables

Como parte del entrenamiento del modelo, se busca identificar qué variables fueron utilizadas para estimar la probabilidad de fuga y cuáles no. De las 230 variables independientes, descartaremos 80, ya que las mismas no se utilizaron por el algoritmo, es decir, no se consideraron relevantes. Como se mencionó, existen distintos tipos de

importancia respecto a una variable dentro del algoritmo de Boosting. A continuación, se analizan las 8 variables más relevantes según cada tipo de importancia. Acompañaremos el análisis con gráficos que describen las predicciones realizadas sobre una muestra de suscriptores en meses posteriores al desarrollo del modelo. En este caso se usará la probabilidad de corte del escenario 1, es decir, no habrá tantas predicciones de fuga, sino más bien predicciones más precisas. La idea es poder entender la dirección en los valores de las variables, respecto a la fuga de suscriptores.

Importancia por ganancia

Atributo	Ganancia
SUBSCRIPTION_PERIOD_NUMBER	901
TOTAL_STREAM_TIME_MS_L28	347
HAS_EVER_HAD_FT	273
NUM_STREAMING_PROFILES_L28	200
ACCOUNT_TOTAL_STREAM_DAYS_L28	199
TOTAL_STREAM_TIME_MS_L7	174
DAYS_SINCE_LAST_STREAM	123
FREE_TRIAL_LENGTH	100

Tabla 7: Se muestra en orden descendiente las 8 variables con mayor importancia por ganancia en el modelo entrenado

La ganancia es la importancia más relevante desde el punto de vista predictivo. Es comprensible viendo la tabla 7, que la variable que mayor ganancia alcanzo fue 'SUBSCRIPTION_PERIOD_NUMBER'. Como ya se discutió, este punto fue la principal razón de limitar la población según antigüedad en el desarrollo de los primeros modelos de predicción de fuga. Resultaron muy importantes las variables que miden el uso dado a la plataforma dentro de los últimos 28 días, como se ve por ejemplo en el gráfico 23, la probabilidad de baja fue más alta a menor uso de la plataforma durante el último mes. El hecho de que estas variables son nulas cuando los suscriptores no cuentan con más de 28 días de antigüedad, permitió combinar el riesgo en el primer mes, con la propensión de baja en base al tiempo de uso en los demás meses.

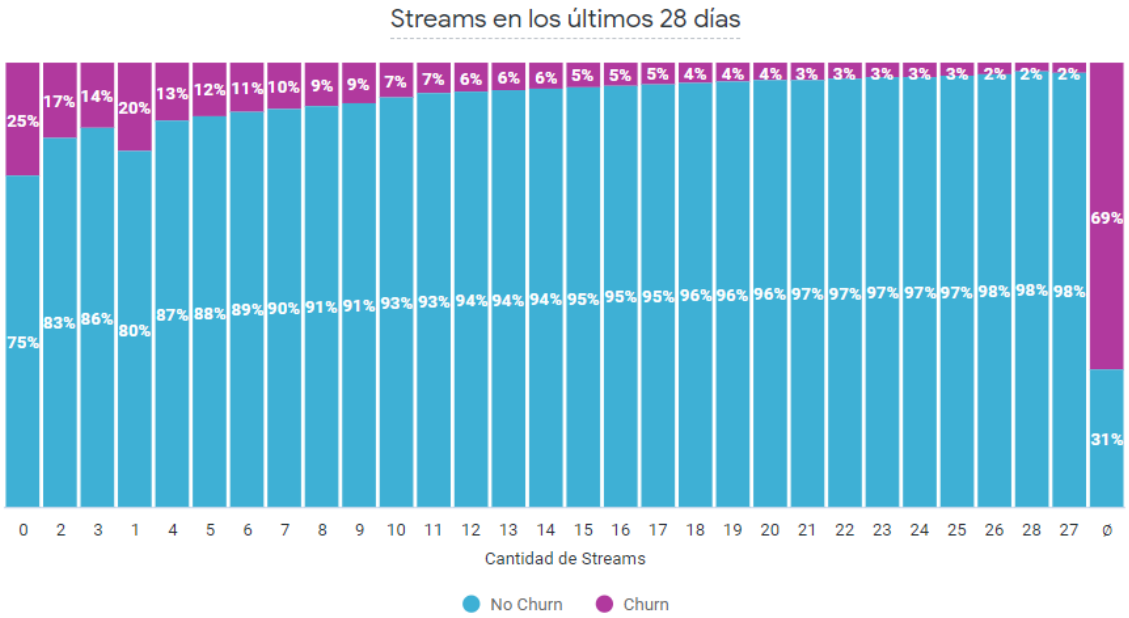

Gráfico 23: Proporción de suscriptores que se estima su baja en el próximo mes, y aquellos que no según la cantidad total de streams realizados en la plataforma en los últimos 28 días.

Respecto a las variables relacionadas al producto, como se había anticipado en el análisis descriptivo, resultó de mucho poder predictivo las variables respecto a la prueba gratuita. Se observa en el Gráfico 24 que las cuentas que tuvieron la oportunidad de probar el producto antes de comprarlo, tienen menos probabilidades de baja, que aquellas que no la tuvieron. Es interesante que estos suscriptores, ya tuvieron la oportunidad de darse de baja de la prueba gratuita finalizada la misma, y eligieron comprar el producto. Este compromiso se refleja con una menor propensión a la fuga. Si bien los casos de baja con prueba gratuita son los menos, resulta muy útil observar que las probabilidades bajan en las pruebas dadas de 7 días. Esto permitiría limitar siempre las pruebas a una semana para evitar perder ingresos regalando días de acceso a la plataforma en cuentas que más tenderán a cancelar su suscripción.

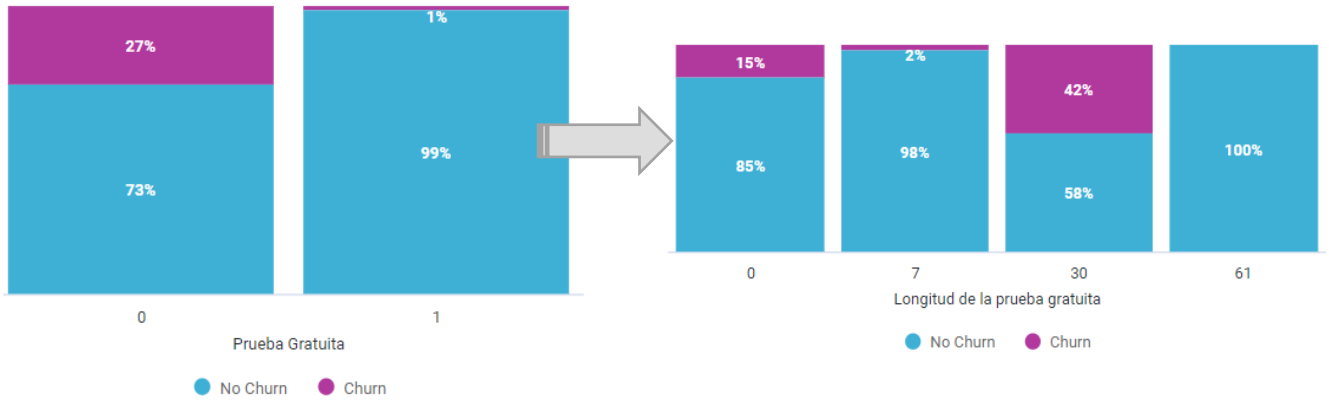

Gráfico 24: Proporción de cuentas con mayor probabilidad de churn dentro del proximo mes según si existio o no una prueba gratutia y sus longitudes, en los casos que sí hubo.

Importancia por cobertura

Atributo	Cobertura
SERIES_FULL_TITLE	372,606
centroid_6	241,455
WATCH_TOP5_MOVIES	197,495
period1965_1970	189,949
centroid_7	155,916
SUBSCRIPTION_PERIOD_NUMBER	112,199
SUBSCRIPTION_SOURCE_SALES_PLATFORM	111,385
SHARE_PROMOTIONAL_WATCH_LENGTH	108,275

Tabla 8: Se muestra en orden descendiente las 8 variables con mayor importancia por cobertura en el modelo entrenado

Respecto a la cobertura resultó muy importante al observar la tabla 8, la variable que muestra el título del primer contenido visto. En el caso de las series, como se observa en el gráfico 25, aquellas series más exitosas, como lo son 'Wanda Visión' o 'The Falcon & The Winter Soldier', tienden atraer a un público con más tendencia a darse de baja en el corto plazo. Sería interesante hacer un detalle más profundo sobre los suscriptores que vieron como primer contenido "The Falcon and the Winter Soldier", ya que los mismos presentan una probabilidad de baja mucho más alta. Podría explicarse una vez más en la antigüedad de las suscripciones, ya que dicha serie fue lanzada en los últimos meses, y se generó mucha publicidad sobre la misma. Otro motivo posible que también ya se discutió, es el tema de los contenidos nuevos originales de Disney+. Como se dijo, los mismos generan muchos nuevos suscriptores, con importantes campañas de marketing, pero es posible que los nuevos suscriptores interesados en estos universos ya hayan consumido (en los últimos años) las demás películas o series del mismo universo, Marvel o Star Wars.

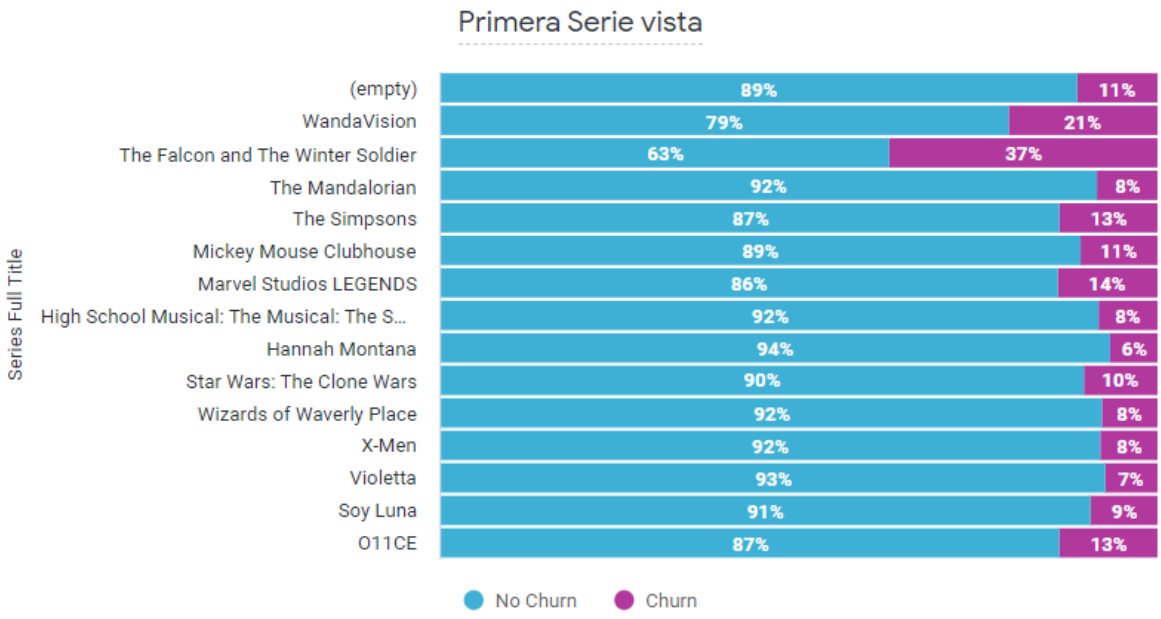

Gráfico 25: Proporción de suscriptores que se estima su baja en el próximo mes, y aquellos que no según el primer contenido consumido por la cuenta, cuando se trata de una serie.

Es interesante que, en la importancia de cobertura, se destaquen dos variables derivadas de los clusters de contenido generados en la etapa 1. Al analizar las variables, en el Gráfico 26 se observa que aquellos suscriptores que no han visto contenidos de estos clusters 6 y 7 tienen más alta probabilidad de baja que aquellos que sí lo hicieron. Para una mejor comprensión hace falta entender la distribución de suscriptores entre los clusters, según los contenidos más consumidos. En el gráfico 27, se puede ver que son muy pocos los suscriptores que consumen la mayor parte de títulos de los clusters 6 y 7. Si bien esto quita relevancia a las variables que cuentan la cantidad de títulos observados de los clusters 6 y 7, es muy interesante plantear que podría necesitarse más promoción o recomendación de este contenido, ya que cuando se consume, genera un mayor compromiso de los clientes con la plataforma.

Gráfico 26: Se segmenta a los suscriptores entre los que han visto al menos un contenido de los clusters 6 y 7 y aquellos que no lo han hecho y se muestra la proporción de fugas estimadas entre cada segmento.

Gráfico 27: Cantidad de suscriptores asignados a cada cluster según el contenido que más consumieron.

Por último, destacamos la importancia que tuvo la variable WATCH_TOP5_MOVIES que muestra cuántas películas vieron los suscriptores de aquellas en el ranking de las 5 más vistas. Se ve en el Gráfico 28 de forma muy clara que cuantas más películas dentro del ranking se consumieron, menores son las probabilidades de fuga. Es una información simple pero potente, ya que como se hace más adelante en la etapa 3, recomendar las películas más vistas por todos los suscriptores de la plataforma, puede impactar fuertemente sobre la retención de los usuarios.

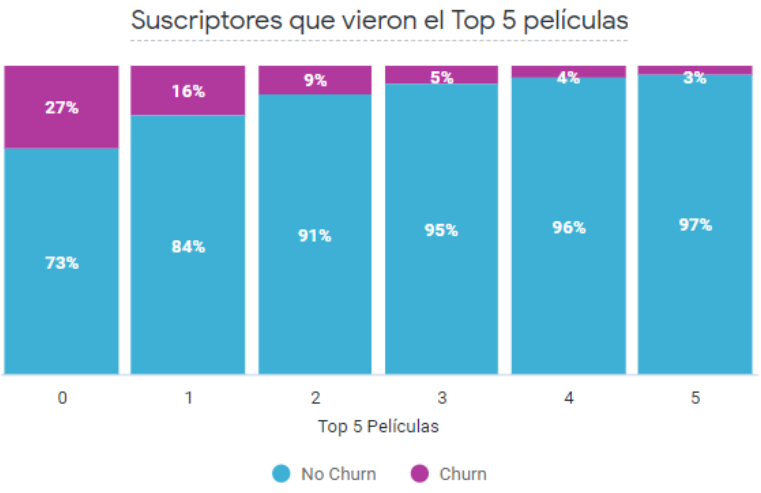

Gráfico 28: Se segmenta a los suscriptores por la cantidad de películas que vieron dentro de las que se encuentran en el ranking top 5 de las más vistas; y se muestra la relación entre suscriptores con probabilidad de baja y aquellos sin para cada grupo.

Importancia por peso/frecuencia

Atributo	Peso/Frecuencia
ACCOUNT_HOME_COUNTRY	230
PRODUCT_NAME	136
DAY_SINCE_PREVIOUS_DIFFERENT_STATE	110
SUBSCRIPTION_PERIOD_NUMBER	82
BILLING_PLATFORM	78
VAR_NUM_STREAMING_DEVICES_ITD_L28	70
PREVIOUS_DIFFERENT_STATE	58
EPISODES_SIMPSONS	58

Tabla 9: Se muestra en orden descendiente las 8 variables con mayor importancia por peso en el modelo entrenado

En relación a las variables relevantes en función de la frecuencia de uso para hacer cortes, se ve en la tabla 9 que esta importancia está relacionada a la cantidad de categorías que las mismas tienen. Esto quita peso al análisis, ya que se excluyen todas las variables binarias, o de dos categorías, que son muchas. De todas formas, vale la pena apreciar la importancia de las mismas en su relación con los suscriptores en fuga. Como se esperaba al realizar el análisis descriptivo, la variable BILLING_PLATFORM tuvo un rol importante como así también SUBSCRIPTION_PERIOD_NUMBER. Es llamativo la relevancia que tuvieron las variables relacionadas al estado. Vale la pena aclarar que el estado previo, es aquel que tuvo la cuenta antes de estar en "Paid", estado que le permitió estar dentro de la población alcanzada. Al ver el gráfico 29, es llamativo la cantidad de fuga estimada en los suscriptores cuyo estado previo al Paid, fue Churned. Es decir, ya se habían dado de baja una vez y tienen muchas posibilidades de volver a hacerlo. Observando el gráfico 30, son más las chances de cancelar una cuenta, a menor tiempo transcurrido desde el último cambio de estado.

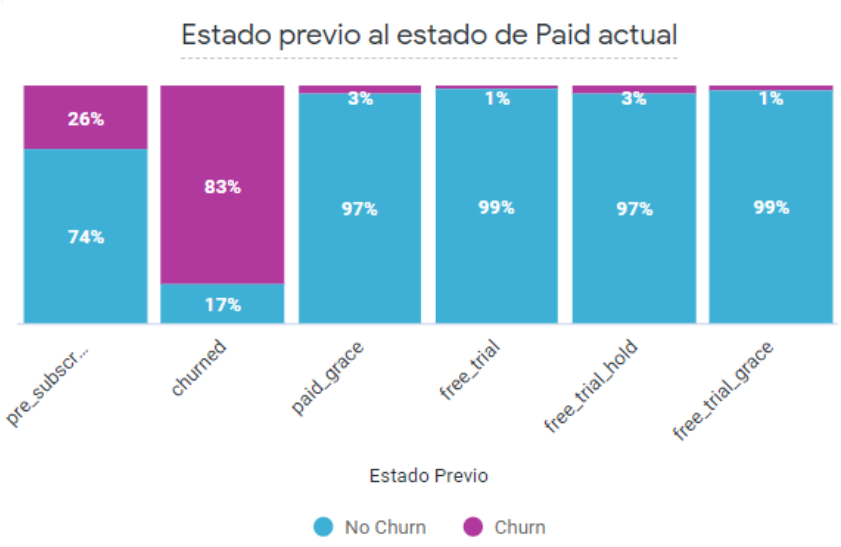

Gráfico 29: Proporción de suscriptores estimados como bajo, en función al estado previo de las cuentas.

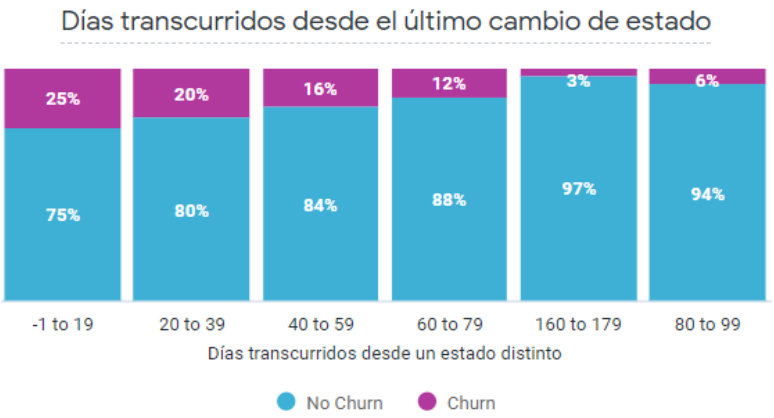

Gráfico 30: Proporción de suscriptores estimados como bajo, en función a la cantidad de días transcurridos desde el último cambio de estado.

Conclusiones

El trabajo realizado en esta sección fue muy rico en cuanto a la integración de una gran cantidad de atributos de consumo de los suscriptores y el comportamiento de fuga. Esto permitió generar análisis muy interesantes para el negocio, que mostraron direcciones posibles a tomar para aprovechar oportunidades en pos de la calidad del producto y la retención de clientes. Esta era una de las necesidades principales del equipo de life cycle, y se lograron excelentes resultados, agregando mucha nueva información acerca consumidor. Además, dichas variables, se tornaron muy importantes a la hora de predecir una fuga, logrando mejorar la calidad de la predicción.

Por el lado de las variables de producto, también se dieron hallazgos muy enriquecedores para el negocio, lo discutido acerca de la plataforma de facturación y los días de prueba gratuita, son ejemplos de espacios en donde se puede trabajar para incrementar los suscriptores o retener los mismos.

Desde el punto de vista del diseño de una estrategia efectiva de retención de clientes, el modelo aporta una visión más general respecto de las variables que más influyen a la hora de generarse una cancelación. El alto alcance de la población incluida en el modelo puede afectar, la performance del mismo, ya que es más difícil la identificación de patrones fuertes que predigan una baja, existiendo tanta mescla de suscriptores. Esto se ve reflejado en el AUC ROC del modelo de 0.71, que fue menor al esperado. En el apartado de próximos pasos se detallan las acciones consideradas para mejorar este aspecto.

Etapa III: Recomendación de contenido

En esta última sección se lleva adelante el análisis prescriptivo, que se apoya en los inputs generados en las atapas anteriores: Las predicciones de los modelos discutidos en la Etapa II (Predictiva) y el análisis de la estructuración de la oferta de contenidos online que surgen de la Etapa I (Descriptiva). El objetivo es entregarle al negocio una posible herramienta para retener a los suscriptores. Esto se generará a través de una recomendación de contenido personalizada para cada cuenta con propensión a la baja.

Se usarán dos enfoques distintos para hacer una recomendación final. Por un lado, se realizará una recomendación basada en los títulos más destacados dentro del cluster de contenido al que pertenece cada cuenta, según lo trabajado en la etapa I y en la etapa II. En un segundo enfoque se implementará un sistema de recomendación basado en el comportamiento del suscriptor dentro de la plataforma. Se buscará identificar con alta confianza los títulos que debería ver un suscriptor, en base a aquellos suscriptores que suelen ver contenido similar. Se combinarán ambas recomendaciones para lograr retener al suscriptor o bien por un contenido con características homogéneas al que suele consumir, o por un contenido que los sorprenda en base a lo consumido por perfiles similares.

Recomendación según características del contenido

Para realizar la primera parte de la recomendación basada en los clusters de contenido, se tendrán en cuenta dos hallazgos realizados en las secciones previas.

1. La mayor cantidad de suscriptores tienden a ver la mayor cantidad de títulos de 4 clusters, de los 10 identificados. Esto presenta ventajas y desventajas. Lo positivo se da por el lado de la recomendación, ya que, al no haber mucha diversidad en el consumo de los suscriptores, hace más simple una recomendación de títulos que seguramente atraiga al usuario. El comportamiento en masa hace más fácil proponer algo atractivo para cada uno. La desventaja, es la cantidad de catálogo no consumido que se “desperdicia” en el producto. Buscaremos enmendar este aspecto proponiendo en ciertos casos, contenidos de clusters distintos.
2. Los títulos más vistos por todos los usuarios de la plataforma, generan una mayor retención en los suscriptores, como se observó en el gráfico 27. Esta información se usará para definir el contenido a recomendar, ya que el objetivo es la retención.

Implementación

Para realizar la recomendación se llevarán a cabo los siguientes pasos. Los mismos fueron simplificados mediante un ejemplo en el gráfico 31. El detalle de código puede verse en E.3. o BQ RecomendacionCluster

1. Se partirá de una cuenta que se estima que se dará de baja
2. Se buscará el contenido histórico que vio en la plataforma
3. Se buscará el cluster de contenido al cual pertenece definido por la variable creada en la Etapa 2.
4. Paralelamente se agruparán todos los contenidos que pertenecen al cluster en cuestión, y se los ordenara de manera descendente en base a la cantidad de cuentas en Latinoamérica que han visto a cada uno.
5. Se descartarán los contenidos del cluster, que la cuenta en cuestión, ya haya visto.
6. Se seleccionará los 3 mejores contenidos del cluster, que aún no ha visto la cuenta, para realizar la recomendación.
7. En el caso de que el subscriptor ya haya visto todos los contenidos del cluster, se recomendaran los mejores 3, del segundo cluster que mayor contenido haya consumido. Así se seguirá hasta completar la recomendación de 3 títulos.

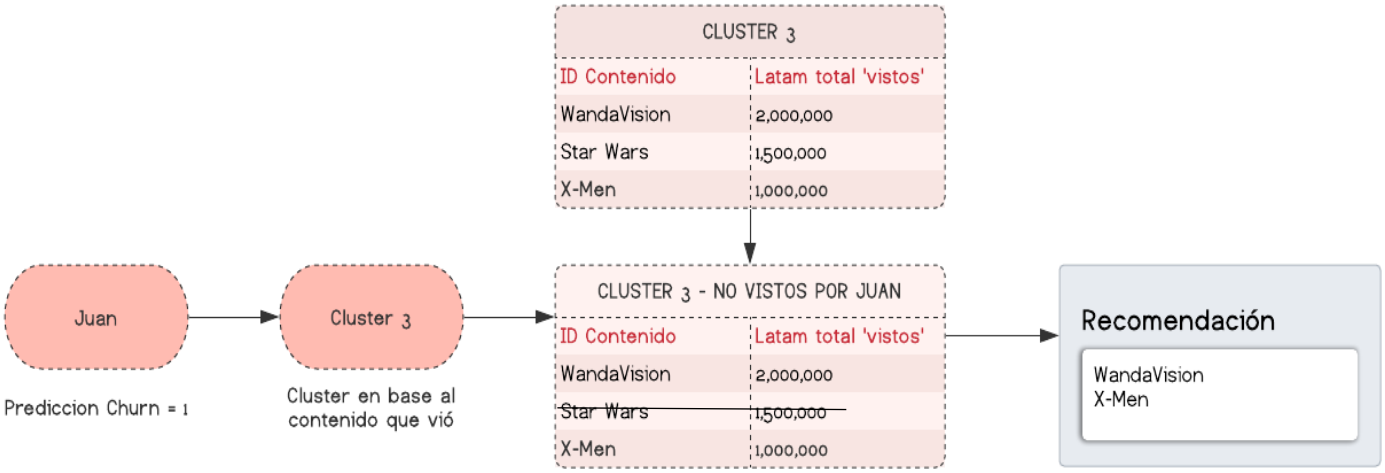

Gráfico 31: Simplificación del esquema de pasos llevados a cabo para la recomendación por ranking del cluster. A modo de ejemplo la recomendación es solo de 2 títulos.

Recomendación según el comportamiento de los suscriptores.

En este segundo enfoque se buscará una recomendación de contenido, trabajando con la similitud de títulos vistos entre los suscriptores de Disney Plus. La ventaja de este enfoque es la posibilidad de capturar los intereses específicos de un suscriptor y recomendarle contenido que lo sorprenda, considerando que perfiles

similares descubrieron contenido que se estima que el suscriptor también estará interesado. Se busca que la recomendación contemple la diversidad de contenido en la plataforma, apuntando a una recomendación personalizada en base al gusto de los usuarios, en contraposición a la recomendación en base a las características del contenido realizada en el enfoque anterior.

Se utilizará un sistema de recomendación de filtrado colaborativo *(Collaborative Filtering)* que analiza las relaciones entre los suscriptores y las interdependencias sobre el contenido, para identificar nuevas asociaciones entre un suscriptor y un contenido. El objetivo es realizar predicciones sobre los intereses de un usuario mediante la recopilación de preferencias sobre contenido consumido por muchos usuarios.

Para llevar a cabo las recomendaciones, el algoritmo utiliza información de *feedback* donde se identifica la posición de un suscriptor frente a un título. Se inferirá las preferencias de un usuario en base a *feedback* implícito generado por su comportamiento, a contraposición del *feedback* explicito[10] donde los usuarios declaran una apreciación del producto consumido. Cuando un suscriptor consuma un contenido, generando un *complete stream*, se contará el mismo para usar de input como feedback. Se usará el trabajo de Yifan Hu, Yehuda Koren, Chris Volinsky realizado en su paper 'Collaborative Filtering for Implicit Feedback'.

Las ventajas de este sistema es que la recolección de datos para construirlo es sencilla, ya que necesita solamente el dato de consumo de los suscriptores sobre los títulos. La desventaja es que requiere trabajo para limpiar el "ruido" generado por asumir que, si se vio un título, significa que a uno le gusta ese título. A esto se le llama la confianza de la suposición. El sistema de recomendación se basará en que a mayor cantidad de complete streams sobre un título, mayor es la apreciación del suscriptor. En este caso, al trabajar en recomendaciones a nivel cuenta (donde cada una puede tener más de un perfil) el hecho de que se sintonicen varias veces un contenido, puede explicarse por el hecho de que una misma persona haya repetido su consumido, o por que varias personas de esa cuenta hayan decidió ver el mismo título. En ambos casos, considerar que existe una mayor apreciación por la repetición del consumo del título en la cuenta, tiene sentido para que la recomendación sea de buena calidad.

Uno de los desafíos que presenta el sistema es el tratamiento de los títulos no consumidos, que a priori representan una falta de interés, pero también pueden significar desconocimiento por parte del usuario. Frente a este desconocimiento, se trabajará con el algoritmo ATS[11] (*Alternating Least Squares*) para "aprender" sobre los vacíos, es decir sobre los títulos no consumidos.

El sistema trabajara con la técnica de *matrix factorization*, que buscará generar una representación de la matriz de feedback original (con dimensiones muy altas) en un matriz de

[10] El feedback explícito son datos en los que existe algún tipo de calificación generado por el usuario sobre algún producto, contemplando cierta escala.

[11] Proceso de optimización iterativo en el que, en cada iteración, se busca ir acercándose cada vez más a una representación factorizada de la matriz de feedback original.

dimensiones más pequeñas cuyo producto tienda a la matriz original. Como se observa en el grafico 32, el algoritmo partirá de la matriz R, con dimensiones U (total de suscriptores) por V (total de títulos), y buscará aprender dos matrices más pequeñas de dimensiones:

1. Matriz U: Total de suscriptores por distintos "gustos/estilos" identificados.
2. Matriz V: Total de "gustos/estilos" identificados por el total de títulos.

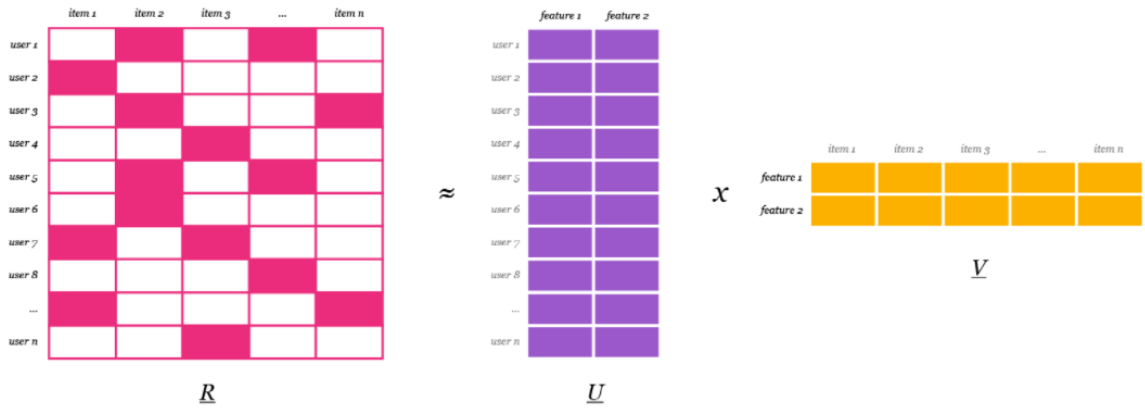

Gráfico 32: Representación gráfica de la trasformación que se realiza en el sistema de recomendación de filtrado colaborativo para obtener vectores que sinteticen el tipo de contenido que más sintoniza cada suscriptor

Se reducirá cada suscriptor a un vector que represente su "gusto/estilo" de contenido, y se usará el mismo para generar una recomendación. Seleccionaremos como recomendación, el listado de 3 contenidos con mayor confianza para cada cuenta que estimamos se dará de baja.

Como recomendación final, se combinarán ambos enfoques. Al equipo de lifecycle se le entregará una recomendación personalizada para cada cuenta clasificada como futura baja. La misma, conformada por 6 títulos, resultado de la sumatoria de los títulos recomendados en cada uno de los enfoques.

Próximos pasos

El proyecto se verá mejorado a través de la incorporación de nuevos datos. En relación a las características del contenido se incorporarán en el corto plazo datos provenientes de la plataforma IMDB[12]. Se trabajará en la vinculación de cada ID interno de Disney, con el ID respectivo de IMDB. De esta manera se accederá a una mayor cantidad de variables descriptivas para cada título entre ellas: idioma, elenco, origen, director/a, nominaciones, certificado, locación del set, especificaciones técnicas. Se espera mejorar aún más las agrupaciones propuestas por el modelo de clustering, y generar mayor cantidad de variables de consumo de contenido en la etapa de ingeniería de atributos del modelo trabajado en la etapa 2.

En relación al modelo de predicción de bajas, también se obtendrá acceso a tablas de datos que permitirán incorporar más información acerca de los suscriptores. Por un lado, se accederá a una tabla con variables demográficas de los mismos. Estos datos son el resultado de un modelo de aprendizaje automático que infiere según características de cada suscriptor un segmento excluyente para cada uno. Así se clasifican como:

a) Casas de familia con hijos jóvenes
b) Casas de familia con hijos pequeños
c) Adultos sin hijos – intereses femeninos
d) Adultos sin hijos – intereses femeninos

Por otro lado, se accederá a datos más detallados respecto a la facturación del producto y también a datos acerca de la navegación de los usuarios en la plataforma. Se espera que, con estas incorporaciones, se identifiquen variables de mucha importancia respecto a la predicción de bajas, que logren mejorar la performance obtenida.

Propuesta de implementación

Se propone como próximos pasos la implementación de un A/B *testing* para poder monitorear el impacto de una acción de retención de clientes sobre los suscriptores estimados de baja. Este tipo de test permite optimizar el gasto en marketing minimizando el costo de decisiones en campañas de marketing que no alcanzan el éxito buscado, permitiendo aprender rápidamente cual es la mejor acción que permite alcanzar los objetivos. En este caso el objetivo será minimizar las bajas de suscriptores, y para ellos se medirá la siguiente ratio:

[12] https://www.imdb.com/

Suscriptores que se estima que cancelaran su cuenta

Suscriptores que efectivamente cancelaron

El experimento se basará en la hipótesis de que los suscriptores se dan de baja principalmente por no encontrar contenido de interés.

Para realizar el test se tomará como población a todos los suscriptores de Disney Plus que hayan contratado el servicio de manera directa, su suscripción venza dentro de los próximos 30 días y que al menos una vez hayan sintonizado contenido desde un dispositivo móvil. Se partirá la población en partes iguales, de forma aleatoria, para obtener una población de control y una de testeo. Esta última será impactada con notificaciones 'push'[13] móviles. Estas notificaciones requieren que el suscriptor tenga descargada la aplicación de Disney Plus en su teléfono móvil. Se propone que la notificación incluya un contenido aleatorio dentro de la recomendación realizada en la etapa 3, y se proponga con un texto en el que sea observe un esfuerzo por entender los gustos del suscriptor en base al contenido ya consumido. Se muestra un ejemplo de la misma en el grafico 33.

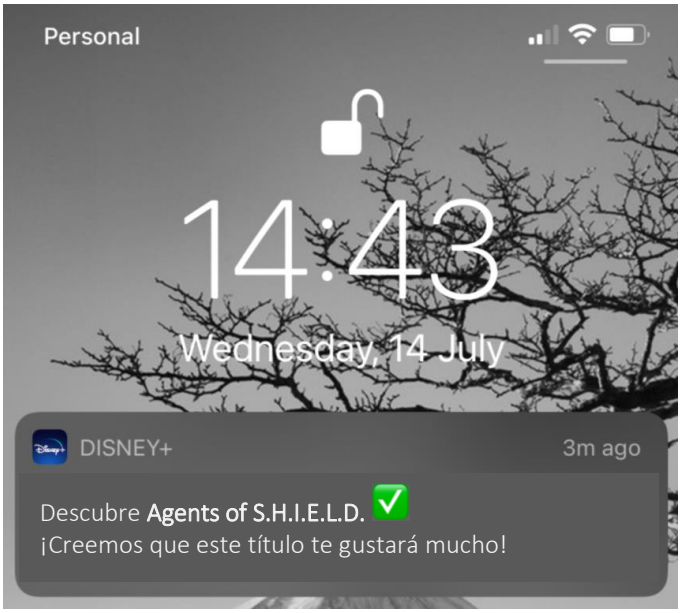

Grafico 33: Ejemplo de campaña propuesta para el grupo de testeo del experimento A/B testing para mejorar la tasa de retención de suscriptores.

Se propone que la duración del experimento sea de 30 días, y se evaluará si la misma tuvo éxito, buscando una diferencia significativa entre la ratio en test y en control.

[13] Notificación push: mensaje emergente corto en una aplicación móvil o web.

Discusión final

En el trabajo realizado se expusieron las problemáticas reales que se dan en una empresa multinacional que atraviesa una transformación digital. Disney Latam se encuentra inmersa además en un proceso de readecuación de su inteligencia de negocios al contexto particular del mercado Latino. En este sentido, resulta fundamental promover acciones analíticas como las que se describen en este trabajo, que tienen como fin último brindar acciones prescriptivas concretas que le permitan a la empresa solucionar problemas específicos, como la fuga clientes, entre otros.

Los objetivos de este trabajo giraron en torno a 3 cuestiones: (1) Describir adecuadamente la oferta de contenidos de Disney Plus Latam, (2) Identificar a los clientes en fuga de la plataforma y (3) Diseñar mecanismos de retención de clientes. Los puntos (1) y (2) se abordaron utilizando modelos de aprendizaje no supervisado y supervisado respectivamente, que luego fueron integrados en una etapa prescriptiva (punto 3) de manera adecuada. Segmentar el proyecto en etapas permitió que los resultados de cada una puedan usarse por muchos grupos de interés, y disparar nuevos proyectos mediante adaptaciones según nuevos casos de uso. Se consideró clave poner en disposición de todo el negocio los hallazgos que se fueron dando a través de dashbaords, que permitió que los interesados, con sus conocimientos del negocio, profundicen según su interés, las variables, su interacción, y resultados dados. Existía una necesidad crítica de conocer más al consumidor del producto para agregar valor a la toma de decisiones, y todo lo entregado a través del proyecto fue muy apreciado.

Entre los aspectos más destacables de los hallazgos obtenidos en la etapa I cabe mencionar la alineación en la diversidad del catálogo con las estrategias de productos de streaming de la compañía. Se planteó una valiosa categorización que permite al negocio conocer su oferta y una posible audiencia detrás cada grupo. Disparar preguntas del tipo ¿El consumidor sabe de la existencia de este estilo de contenido? ¿Se necesita publicidad, o directamente no hay mucho interés en este grupo homogéneo de títulos? Como empresa multinacional, es lógico que catálogo tienda a ser global, pero analizar los gustos y segmentos de interés propios de Latinoamérica puede llevar a optimizar los recursos generando importantes eficiencias.

Respecto de la etapa II se identificó la gran población de bajas que se dan pasados los 60 días de suscripción, hecho que se ignoraba por parte del negocio. Analizando las variables más determinantes se observó que la variación en el último mes en la intensidad de uso de la plataforma es una clara alerta de anticipación a la baja. Combinando esto con lo advertido sobre las pruebas gratuitas, se intuye una posible estrategia de retención ofreciendo no más de 7 días gratuitos a esas cuentas que no muestran movimiento en el último mes. Se descubrieron hechos puntuales que también son de interés, como la convocatoria que generan las producciones originales, y la tendencia de alguno de estos títulos a conquistar suscriptores que rápidamente se ven dan de baja.

En relación a la última etapa, la etapa III, el modelo prescriptivo logró incorporar gran parte de los hallazgos mencionados a través de un proceso sencillo de

recomendación de contenido. La recomendación se hizo a nivel de suscripción, pero es posible re-utilizar los resultados buscando aquellas recomendaciones más frecuentes para realizar acciones por segmentos o países, controlando así el gasto de la acción y la complejidad de implementación.

Este trabajo alcanzó excelentes resultados dentro del contexto mencionado. Previo al mismo, existía mucha necesidad del negocio por conocer a sus clientes. Lanzada la plataforma, no había hallazgos locales tangibles o análisis profundos realizados que puedan ayudarlos. Este proyecto logro varios entregables al negocio, que no solo fueron útil de forma inmediata, sino que dieron visibilidad de hacia dónde dirigir los próximos estudios de datos.

Definiciones

Account / cuenta: Un registro de account_id único que contiene propiedades relacionadas con la plataforma y el usuario. Usuario único a nivel de Disney+.
Subscriber / suscriptor: Una cuenta con una suscripción que puede estar en período de prueba o de pago. Para los suscriptores anónimos, esta identidad única será el subscription_id hasta que se conozca la cuenta.
Subscription / suscripción: Un suscriptor pude tener más de una suscripción. Las mismas se identifican por el suscription_id. Una suscripción tiene determinadas características frente a una cuenta, la plataforma, la habilitación para consumir contenido, y demás.
Profile / perfiles: Dentro de la plataforma el suscriptor puede generar perfiles especiales para ver contenido que permiten una experiencia más personalizada del contenido.
Entitled /habilitación: Un suscriptor está habilitado cuando tiene acceso a los contenidos de la plataforma.
Primary Account /cuenta primaria: Un suscriptor puede tener más de una suscripción. Se identifica como cuenta primaria a la suscripción con mejor estado respecto al pago.

Definiciones de estados: Posibles estadíos de una suscripción en relación a su situación de facturación y cobro.

- Free_trial: Periodo de prueba del producto en donde la suscripción tiene acceso al contenido.
- Free_trial_cancel: La Suscripción se encuentra en un período de prueba gratuito, pero el cliente ha elegido "cancelar" (optar por no renovar) su Suscripción.
- Free_trial_grace: La suscripción acabo su período de prueba gratuito y el sistema ha intentado procesar un pago, pero el pago no se ha realizado correctamente. Este estado representa un breve período de tiempo en el que la prueba gratuita se extiende de manera efectiva mientras el sistema de procesamiento de pagos intenta solucionar el problema. Se continúa teniendo acceso al contenido.
- Free_trial_hold: Periodo al terminar los días prueba de gratuita en gracia, donde se continúan los intentos fallidos de cobro. La suscripción ya no está habilitada para consumir contenido
- Free_trial_abandon: Cuando un suscriptor probó el período de prueba gratuito, pero no se convirtió en un cliente pago. La suscripción ya no está habilitada para consumir contenido
- Paid: Cuando una suscripción cumple con un período regular de pago y renovación. El mismo está habilitado a consumir contenido y tiene una fecha definida de renovación. Intitulado.
- Paid_cancel: La Suscripción se encuentra en un período de pago regular, pero el cliente ha elegido "cancelar" (optar por no renovar) su Suscripción. Tendrá

acceso al contenido durante el resto de su período de suscripción que ya fue pagado.

- Paid_grace: La suscripción ha pasado la fecha de renovación de su suscripción y el sistema ha intentado procesar un pago, pero el pago no se ha realizado correctamente. Este estado representa un corto período de tiempo en el que el período de suscripción se extiende de manera efectiva mientras el sistema de procesamiento de pagos intenta remediar el problema. Se continúa teniendo acceso al contenido.
- Paid_hold: Periodo al terminar los días de Paid_grace, donde se continúan los intentos fallidos de cobro. La suscripción ya no está habilitada para consumir contenido
- Churned: Cuando un suscriptor ha decidido dejar de ser cliente. Ya no tiene acceso al contenido.

Definiciones de consumo de contenido:

- Watch / visualización: Una colección de reproducciones de video en la misma plataforma, en el mismo dispositivo y mismo perfil, con un tiempo de inactividad mínimo entre reproducciones.
- Minplayhead / mínima reproducción: Tiempo mínimo alcanzado de reproducción durante una visualización.
- Maxplayhead / máxima reproducción: Tiempo máximo alcanzado de reproducción durante una visualización.
- Watchlength / largo de la reproducción: Duración de la visualización.
- Runtime / largo del contenido: Duración total del contenido
- Stream: Una visualización que muestra la intención del usuario de ver contenidos específicos. Es una visualización que dura más de 10 segundos.
- Completed Stream: Un stream donde se completa el contenido. Se alcanzan ciertas condiciones respecto al largo del stream y la duración del contenido.
- First Stream: La primera transmisión válida por cuenta.

Apéndice

Detalles de tablas utilizadas y variables

Tabla 1: Contenidos

Categoría	Variable	Tipo	Descripción
Categorizaciones del contenido.	CONTENT_CLASS	STRING	Tipo de contenido que se clasifica el título, entre serie, película o cortometraje.
	IS_ANIMATED	STRING	Vale 1 si el título es animado, y 0 en caso contrario.
Propias del título	idmix	STRING	ID del título
	titlemix	STRING	Título
	RELEASE_YEAR_MIN	INTEGER	Primero año en el que se lanzó el título.
	MAX_YEAR	INTEGER	Ultimo año en el que se lanzó el título.
	period	INTEGER	Cantidad de años durante los cuales se lanzaron unidades del título.
Relacionadas a Disney Plus	Brazil	STRING	Vale 1 si el título es está disponible en el territorio 'Brazil', y 0 en caso contrario.
	Caribbean	STRING	Vale 1 si el título es está disponible en el territorio 'Caribbean', y 0 en caso contrario.
	LATAM_HISP	STRING	Vale 1 si el título es está disponible en el territorio 'LATAM_HISP', y 0 en caso contrario.
	latam_available	INTEGER	Vale 1 si el título es está disponible en el territorio 'latam_available', y 0 en caso contrario.
	IS_DISNEY_PLUS_ORIGINAL	BOOLEAN	Línea de negocio a la cual se relaciona el título
	IS_PAY_1	STRING	Vale 1 si el título requiere un pago adicional en la plataforma para consumirse, y 0 en caso contrario.
Relacionadas a los negocios de Disney	BRANDS_DETAILED	STRING	Detalle de marca a la cual se relaciona el título
	BRANDS	STRING	Marca a la cual se relaciona el título
	FRANCHISES	STRING	Franquicia a la cual se relaciona el título
	BUSINESS_UNIT	STRING	Línea de negocio a la cual se relaciona el título
	PRODUCTION_COMPANY	STRING	Compañía productora del título
Relacionadas al género	Animation	BOOLEAN	Vale 1 si el título es del género 'Animation', y 0 en caso contrario.
	Family	BOOLEAN	Vale 1 si el título es del género 'Family', y 0 en caso contrario.
	Comedy	BOOLEAN	Vale 1 si el título es del género 'Comedy', y 0 en caso contrario.
	Action_Adventure	BOOLEAN	Vale 1 si el título es del género 'Action_Adventure', y 0 en caso contrario.
	Fantasy	BOOLEAN	Vale 1 si el título es del género 'Fantasy', y 0 en caso contrario.
	Drama	BOOLEAN	Vale 1 si el título es del género 'Drama', y 0 en caso contrario.
	Historical	BOOLEAN	Vale 1 si el título es del género 'Historical', y 0 en caso contrario.
	Dance	BOOLEAN	Vale 1 si el título es del género 'Dance', y 0 en caso contrario.

	Documentary	BOOLEAN	Vale 1 si el título es del género 'Documentary', y 0 en caso contrario.
	Thriller	BOOLEAN	Vale 1 si el título es del género 'Thriller', y 0 en caso contrario.
	Coming_of_age	BOOLEAN	Vale 1 si el título es del género 'Coming_of_age', y 0 en caso contrario.
	Kids	BOOLEAN	Vale 1 si el título es del género 'Kids', y 0 en caso contrario.
	Science_Fiction	BOOLEAN	Vale 1 si el título es del género 'Science_Fiction', y 0 en caso contrario.
	Superhero	BOOLEAN	Vale 1 si el título es del género 'Superhero', y 0 en caso contrario.
	Docuseries	BOOLEAN	Vale 1 si el título es del género 'Docuseries', y 0 en caso contrario.
	Musical	BOOLEAN	Vale 1 si el título es del género 'Musical', y 0 en caso contrario.
	Buddy	BOOLEAN	Vale 1 si el título es del género 'Buddy', y 0 en caso contrario.
	Police_Cop	BOOLEAN	Vale 1 si el título es del género 'Police_Cop', y 0 en caso contrario.
	Romance	BOOLEAN	Vale 1 si el título es del género 'Romance', y 0 en caso contrario.
	Animals___Nature	BOOLEAN	Vale 1 si el título es del género 'Animals___Nature', y 0 en caso contrario.
	Crime	BOOLEAN	Vale 1 si el título es del género 'Crime', y 0 en caso contrario.
	Soap_Opera___Melodrama	BOOLEAN	Vale 1 si el título es del género 'Soap_Opera___Melodrama', y 0 en caso contrario.
	Romantic_Comedy	BOOLEAN	Vale 1 si el título es del género 'Romantic_Comedy', y 0 en caso contrario.
	Horror	BOOLEAN	Vale 1 si el título es del género 'Horror', y 0 en caso contrario.
	Music	BOOLEAN	Vale 1 si el título es del género 'Music', y 0 en caso contrario.
	Sports	BOOLEAN	Vale 1 si el título es del género 'Sports', y 0 en caso contrario.
	Reality	BOOLEAN	Vale 1 si el título es del género 'Reality', y 0 en caso contrario.
	Survival	BOOLEAN	Vale 1 si el título es del género 'Survival', y 0 en caso contrario.
	Biographical	BOOLEAN	Vale 1 si el título es del género 'Biographical', y 0 en caso contrario.
	Spy_Espionage	BOOLEAN	Vale 1 si el título es del género 'Spy_Espionage', y 0 en caso contrario.
	Medical	BOOLEAN	Vale 1 si el título es del género 'Medical', y 0 en caso contrario.
	Game_Show___Competition	BOOLEAN	Vale 1 si el título es del género 'Game_Show___Competition', y 0 en caso contrario.
	Mystery	BOOLEAN	Vale 1 si el título es del género 'Mystery', y 0 en caso contrario.
	Anthology	BOOLEAN	Vale 1 si el título es del género 'Anthology', y 0 en caso contrario.
	Lifestyle	BOOLEAN	Vale 1 si el título es del género 'Lifestyle', y 0 en caso contrario.

	Parody	BOOLEAN	Vale 1 si el título es del género 'Parody', y 0 en caso contrario.
	Variety	BOOLEAN	Vale 1 si el título es del género 'Variety', y 0 en caso contrario.
	Disaster	BOOLEAN	Vale 1 si el título es del género 'Disaster', y 0 en caso contrario.
	Procedural	BOOLEAN	Vale 1 si el título es del género 'Procedural', y 0 en caso contrario.
	Western	BOOLEAN	Vale 1 si el título es del género 'Western', y 0 en caso contrario.
	Anime	BOOLEAN	Vale 1 si el título es del género 'Anime', y 0 en caso contrario.
	Concert_Film	BOOLEAN	Vale 1 si el título es del género 'Concert_Film', y 0 en caso contrario.
	Film_Noir	BOOLEAN	Vale 1 si el título es del género 'Film_Noir', y 0 en caso contrario.
	Talk_Show	BOOLEAN	Vale 1 si el título es del género 'Talk_Show', y 0 en caso contrario.

Tabla 2: Suscriptores

- DS: Fecha de la observación.
- ACCOUNT_ID: Cuenta asociada a la suscripción.
- SUBSCRIPTION_STATE: Estado de la suscripción relacionado a su pago.
- PREVIOUS_SUBSCRIPTION_STATE: Estado de la suscripción el día anterior.
- PRODUCT_NAME: Nombre del producto de Disney +, incluyendo país y plataforma de facturación.
- FREE_TRIAL_LENGTH: número de días de prueba gratuita
- HAS_EVER_HAD_FT: Indicador binario que indica 1 cuando la suscripción tuvo acceso a una prueba gratuita.
- START_DATE: Fecha del comienzo de la suscripción.
- END_DATE: Fecha del final de la suscripción.
- RENEWAL_TIMESTAMP_EST: Fecha de renovación estimada de la suscripción.
- SUBSCRIPTION_START_COHORT_DATE: Fecha del primer registro del plan de suscripción
- SUBSCRIPTION_RENEWAL_COHORT_DATE: Fecha esperada de renovación del plan de la suscripción.
- SUBSCRIPTION_PARTNERSHIP: Socio comercial de la suscripción
- SUBSCRIPTION_LENGTH: Duración del plan de suscripción o frecuencia de facturación
- SUBSCRIPTION_TYPE: Indica si el producto es una suscripción independiente o en paquete
- BILLING_PLATFORM: Plataforma desde la que se factura el plan de suscripción
- ACCOUNT_HOME_COUNTRY: País donde se generó la suscripción.

- PRE_LAUNCH_COHORT: Indicador binario que Indica si la suscripción se generó en el periodo de pre venta.
- IS_ENTITLED_ACCOUNT: Indicador binario que indica si la suscripción tiene acceso a ver el contenido de la plataforma.
- SUBSCRIPTION_PERIOD_NUMBER: Período de suscripción actual o veces en que se renovó el plan de suscripción.
- EXPECTED_RENEWAL_DATE: Fecha esperada de renovación del plan de la suscripción.
- IS_RENEWAL_EVENT: Indicador binario que indica 1 si en la fecha de la observación, la suscripción debe renovarse.
- IS_FREE_TRIAL_CONVERSION_EVENT: Indicador binario que indica 1 si el plan de suscripción se convirtió en pago después del período de prueba gratuita.
- PREVIOUS_DIFFERENT_STATE: Estado de la suscripción previo, distinto al actual.
- PREVIOUS_DIFFERENT_STATE_DS: Ultima fecha en la que el estado de suscripción fue distinto al actual.
- ACTIVATION_DATE: Fecha en que se creó la cuenta.
- PURCHASE_DEVICE_TYPE: Tipo de dispositivo desde el que se compró la suscripción
- SUBSCRIPTION_MODEL: Indica si la suscripción se compró al por mayor o al por menor
- SUBSCRIPTION_SOURCE_SALES_PLATFORM
- CAMPAIGN_CODE: código de campaña correspondiente a la suscripción
- VOUCHER_CODE: código de voucher correspondiente a la suscripción
- SUBSCRIPTION_SOURCE: descripción de los estados
- PRODUCT_SKU_PLATFORM: plataforma relacionada con el producto SKU.
- SUBSCRIPTION_REPORTING_TYPE: Categorización de productos de suscripción utilizada para informes

Tabla 3: Dispositivos

- ACCOUNT_ID: Cuenta asociada a la suscripción.
- DEVICE_COUNT: recuento de distintos dispositivos utilizados por una cuenta para transmitir Disney Plus.

Tabla 4: Marcas

- ACCOUNT_ID: Cuenta asociada a la suscripción.
- HAS_ANY_KIDS_PROFILE_STREAMING: Indicador binario que indica 1 si se ha creado un perfil de niños en la cuenta.
- BRAND_SEGMENT: Marca asignada en función de su consumo de contenido.

Tabla 5: Consumo de producto

- ACCOUNT_ID: Identificador único para todas las cuentas
- TOTAL_LOGIN_DAYS_L7: Total de días en los que la cuenta inició sesión al menos una vez al día. (7 días antes de la DS)
- TOTAL_LOGIN_DAYS_L28: Total de días en los que la cuenta inició sesión al menos una vez al día. (28 días antes de la DS)

- TOTAL_LOGIN_DAYS_ITD: Total de días en los que la cuenta inició sesión al menos una vez al día. (Desde la creación de la cuenta)
- TOTAL_STREAMS_L7: Reproducción total de 'streams', reproducciones de videos que duran más de 10 segundos, 7 días antes de la DS.
- TOTAL_STREAMS_L28: Reproducción total de 'streams', reproducciones de videos que duran más de 10 segundos, 28 días antes de la DS.
- TOTAL_STREAMS_ITD: Reproducción total de 'streams', reproducciones de videos que duran más de 10 segundos, desde la creación de la cuenta
  NUM_STREAMING_PROFILES_L7: Recuento de perfiles que tienen al menos una transmisión en el período de tiempo determinado (7 días antes de DS).
- NUM_STREAMING_PROFILES_L28: Recuento de perfiles que tienen al menos una transmisión en el período de tiempo determinado (28 días antes de DS).
- NUM_STREAMING_PROFILES_ITD: Recuento de perfiles que tienen al menos una transmisión en el período de tiempo determinado (Desde la creación de la cuenta).
- NUM_STREAMING_DEVICES_L7: Recuento de todos los dispositivos que tienen al menos una transmisión en el período de tiempo dado (7 días antes de DS)
- NUM_STREAMING_DEVICES_L28: Recuento de todos los dispositivos que tienen al menos una transmisión en el período de tiempo dado (28 días antes de DS)
- NUM_STREAMING_DEVICES_ITD: Recuento de todos los dispositivos que tienen al menos una transmisión en el período de tiempo dado (Desde la creación de la cuenta)
- ACCOUNT_TOTAL_STREAM_DAYS_L7: Cantidad de días únicos que tienen al menos una transmisión desde la cuenta en el período de tiempo dado (7 días antes de DS).
- ACCOUNT_TOTAL_STREAM_DAYS_L28: Cantidad de días únicos que tienen al menos una transmisión desde la cuenta en el período de tiempo dado (28 días antes de DS).
- ACCOUNT_TOTAL_STREAM_DAYS_ITD: Cantidad de días únicos que tienen al menos una transmisión desde la cuenta en el período de tiempo dado (Desde la creación de la cuenta).
- TOTAL_STREAM_TIME_MS_L7: Tiempo total de transmisión en milisegundos (7 días antes de DS)
- TOTAL_STREAM_TIME_MS_L28: Tiempo total de transmisión en milisegundos (28 días antes de DS)
- TOTAL_STREAM_TIME_MS_ITD: Tiempo total de transmisión en milisegundos (Desde la creación de la cuenta)
- TOTAL_STREAM_TIME_WEB_MS_ITD: Tiempo total de transmisión en milisegundos en el dispositivo web (Desde la creación de la cuenta)
- TOTAL_STREAM_TIME_MOBILE_MS_ITD: Tiempo total de transmisión en milisegundos en el dispositivo móvil (Desde la creación de la cuenta)
- TOTAL_STREAM_TIME_CONNECTED_TV_MS_ITD: Tiempo total de transmisión en milisegundos en una televisión conectada a internet (Desde la creación de la cuenta)
- TOTAL_STREAM_TIME_UNKNOWN_MS_ITD: Tiempo total de transmisión en milisegundos en un dispositivo desconocido (Desde la creación de la cuenta)
- DAYS_SINCE_LAST_STREAM: Días transcurridos desde la última transmisión de contenido.

Tabla 6: Primer Contenido

- ACCOUNT_ID: Cuenta asociada a la suscripción.
- FIRST_STREAM_PROGRAM_ID: Identificador único para la unidad de contenido transmitida por primera vez por la cuenta.
- FIRST_STREAM_PROGRAM_TITLE: Título de la unidad de contenido transmitida por primera vez por la cuenta
- FIRST_STREAM_DATE_TIME_EST: Fecha en la que la cuenta transmitió el primer contenido
- first_PROGRAM_FULL_TITLE: variable que describe el título completo del primer contenido consumido por la cuenta.
- SERIES_FULL_TITLE: variable que describe el título de la serie a la que pertenece el primero contenido consumido por la cuenta.

Tabla 7: Total Contenido Visto

- ACCOUNT_ID: Identificador único para todas las cuentas
- PROFILE_ID: Identificador único para todos los perfiles
- TOTAL_WATCH_LENGTH_MS: Duración de la transmisión en milisegundos
- PROGRAM_TYPE: Clasificación de contenido en cuanto a su duración.
- PARTNER_SERIES_ID: Identificador único de la serie.
- PROGRAM_ID: identificador único para cada unidad de contenido
- EPISODE_SERIES_SEQUENCE_NUMBER: Número de episodios desde el inicio de la serie, más allá de la temporada.

Consideraciones:

- Solamente se tienen en cuenta los 'complete streams'.
- Se tienen en cuenta todos los títulos vistos por la cuenta, sin importar los distintos perfiles creados dentro de la misma.
- No se tienen en cuenta títulos vistos por las cuentas que no están disponibles en Latinoamérica. Esto puede darse cuando una cuenta de un país latinoamericano, fue dada de alta en otra región, pudiendo acceder a los títulos de la misma.
- Cada título visto por el suscriptor cuenta como 1.
- En el caso de las series, se considera que un suscriptor vio una serie, cuando completo más del 50% de la primera temporada.

Tabla 8: Ranking de Contenido

- CONTENT_UNIT_ID: Id único del contenido
- CONTENT_TITLE: Título del contenido

- ACCOUNT_WATCHES_LATAM: Número total de cuentas que realizaron un 'complete stream' del contenido, en los últimos 28 días en Latinoamérica

Transformación de variables

Cálculo de períodos

Con las variables del tipo fecha, generaremos distintos cálculos para quedarnos con la información de tiempo transcurrido entre distintos sucesos.

- DAY_TO_RENEWAL: días que faltan para renovar la suscripción.
- DAY_SINCE_PREVIOUS_DIFFERENT_STATE: días que transcurrieron desde la última vez que se modificó el estado de la cuenta.
- AGE_IN_DAYS: días que transcurrieron desde que se creó la cuenta.

Descartaremos las variables del tipo fecha utilizadas.

Promedios diarios

Se crearán nuevas variables que calculan el promedio diario de *streams* y *stream-time* para poder comparar la intensidad de *streams* a lo largo de la suscripción, entre cuentas disimiles en cuanto a antigüedad.

- TOTAL_STREAMS_AVG = TOTAL_STREAMS_ITD / Age_in_days
- TOTAL_STREAM_TIME_MS_AVG = TOTAL_STREAMS_ITD / AGE_IN_DAYS

Ratio de días de streams

Se calculará qué porcentaje del total de días de la suscripción, la cuenta generó un *stream.*

- STREAM_DAYS_OVER_AGE = ACCOUNT_TOTAL_STREAM_DAYS_ITD / AGE_IN_DAYS

Variación del último mes respecto al período total de la suscripción

Para el caso de las variables que cuentan el total de dispositivos y perfiles activos en un periodo de tiempo, se calculara su variación respecto a la variable que cuenta lo mismo, pero a lo largo de toda la suscripción.

- VAR_NUM_STREAMING_PROFILES_ITD_L28 = (NUM_STREAMING_PROFILES_L28 / NUM_STREAMING_PROFILES_ITD)-1
- VAR_NUM_STREAMING_DEVICES_ITD_L28 = (NUM_STREAMING_DEVICES_L28/NUM_STREAMING_DEVICES_ITD)-1

Para el caso de las variables que cuentan logins y streams, que su conteo se ve incrementado día a día, haremos un tratamiento especial para analizar la variación durante el último mes. Compararemos la variable del último mes, contra un cálculo que represente el valor promedio de un mes, tomando los valores históricos de la cuenta.

- VAR_TOTAL_STREAMS_ITD_L28 = TOTAL_STREAMS_L28/ ((TOTAL_STREAMS_ITD / Age_in_days)*28) -1
- VAR_TOTAL_STREAM_TIME_MS_ITD_L28 TOTAL_STREAM_TIME_MS_L28/ (TOTAL_STREAM_TIME_MS_ITD/ AGE_IN_DAYS)*28) -1
- VAR_ACCOUNT_TOTAL_STREAM_DAYS_ITD_L28 = ACCOUNT_TOTAL_STREAM_DAYS_L28/ ((ACCOUNT_TOTAL_STREAM_DAYS_ITD/ AGE_IN_DAYS)*28) -1

Variación de la última semana respecto al último mes

Realizamos los mismos calculo que para la variación mensual.

- VAR_NUM_STREAMING_PROFILES_ITD_L7 = NUM_STREAMING_PROFILES_L7 / NUM_STREAMING_PROFILES_ITD)-1
- VAR_NUM_STREAMING_DEVICES_ITD_L7 = NUM_STREAMING_DEVICES_L7 / NUM_STREAMING_DEVICES_ITD)-1
- VAR_TOTAL_STREAMS_L28_L7 = TOTAL_STREAMS_L7 / (TOTAL_STREAMS_L28/4))-1
- VAR_ACCOUNT_TOTAL_STREAM_DAYS_L28_L7 = ACCOUNT_TOTAL_STREAM_DAYS_L7 / (ACCOUNT_TOTAL_STREAM_DAYS_L28/4))-1
- VAR_TOTAL_STREAM_TIME_MS_L28_L7 = TOTAL_STREAM_TIME_MS_L7 / (TOTAL_STREAM_TIME_MS_L28/4))-1

Ratio stream vs. login

Se considera importante las veces que un suscriptor entra a la plataforma (hace un login) pero no ve ningún contenido (stream). Se podría entender que no encontró nada interesante en la plataforma y decidió buscar en otra parte. Para captar esta acción, calcularemos las ratios entre streams y logins para la última semana y el último mes.

- STREAMS_VS_LOGINS_DAYS_L7 = ACCOUNT_TOTAL_STREAM_DAYS_L7 / TOTAL_LOGIN_DAYS_L7
- STREAMS_VS_LOGINS_DAYS_L28 = ACCOUNT_TOTAL_STREAM_DAYS_L28 / TOTAL_LOGIN_DAYS_L28,2

Variación en la ratio stream vs. login

Se calculará la variación de dicha ratio del último mes respecto al periodo total de la suscripción, y de la última semana respecto al último mes.

- VAR_TOTAL_STREAMS_VS_LOGINS_DAYS_ITD_L28 = ((ACCOUNT_TOTAL_STREAM_DAYS_L28/TOTAL_LOGIN_DAYS_L28)/ (ACCOUNT_TOTAL_STREAM_DAYS_ITD/TOTAL_LOGIN_DAYS_ITD))-1
- VAR_TOTAL_STREAMS_VS_LOGINS_DAYS_ITD_L7 = ((ACCOUNT_TOTAL_STREAM_DAYS_L7/TOTAL_LOGIN_DAYS_L7)/ (ACCOUNT_TOTAL_STREAM_DAYS_L28/TOTAL_LOGIN_DAYS_L28))-1

Cantidad de tiempo visto por tipo de contenido.

Se calculará el tiempo en horas que cada cuenta dedico a sintonizar contenido de distintas características. Se generan las siguientes variables:

- PROMOTIONAL_WATCH_LENGTH_HS
- SHORTFORM_WATCH_LENGTH_HS
- MOVIE_WATCH_LENGTH_HS
- SERIE_WATCH_LENGTH_HS
- SUPPLEMENT_WATCH_LENGTH_HS

Proporción de tiempo visto por tipo de contenido.

Se calculará el porcentaje de tiempo, sobre el total, que destino cada cuenta a cada tipo de contenido. Se genera las siguientes variables:

- SHARE_PROMOTIONAL_WATCH_LENGTH_MS
- SHARE_SHORTFORM_WATCH_LENGTH_MS
- SHARE_MOVIE_WATCH_LENGTH_MS
- SHARE_SERIE_WATCH_LENGTH_MS
- SHARE_SUPPLEMENT_WATCH_LENGTH_MS

Tipo de contenido más visto: CONTENT_TYPE_MOST_SEEN

Generaremos una nueva variable que asigne a cada cuenta cuál es el tipo de contenido más consumido. Para ello se seleccionará el máximo de las variables de SHARE generados en el paso anterior, para cada cuenta.

Cluster de contenido: CENTROID_ID

Analizaremos el contenido consumido por cada cuenta, para clasificar a cada una en uno de los distintos clusters identificado en la etapa 1. Una cuenta pertenecerá al cluster de contenido, del cual haya visto la máxima cantidad de títulos. Algunas consideraciones:

- El título del 'first stream' tiene mayor peso, contando como 5.
- En los casos que hay un empate entre clusters, dado porque ambos alcanzan la cantidad máxima de títulos vistos por la cuenta. Se asignará algún cluster que haya empatado, de forma aleatoria.
- En aquellos casos que la cuenta no haya completado ningún *stream*, no tendrá un cluster asignado y la variable será nula.

Variables de conteo de 'complete streams' sobre distintos tipos de contenido.

Se generarán variables que sumen la cantidad de 'complete streams' generados por la cuenta, sobre grupos de contenidos definidos en base a distintas características. Estas características se obtienen de la tabla de contenido utilizada en la etapa 1. Las variables resultantes son:

- TOTAL_SEEN: cantidad de complete streams por cuenta

- IS_ANIMATED: cantidad de complete streams por cuenta del contenido animado
- IS_DISNEY_PLUS_ORIGINAL cantidad de complete streams por cuenta del contenido original creada para Disney Plus
- IS_PAY_1: cantidad de complete streams por cuenta del contenido pago
- Cantidad de complete streams por cuenta sobre distintos tipos de contenido:
    - Movie
    - episode
    - short_form
- Cantidad de complete streams por cuenta sobre los títulos definidos dentro de cada cluster:
    - centroid_1
    - centroid_2
    - centroid_3
    - centroid_4
    - centroid_5
    - centroid_6
    - centroid_7
    - centroid_8
    - centroid_9
    - centroid_10
- Cantidad de complete streams por cuenta sobre los títulos de cada marca:
    - BRANDS_Marvel
    - BRANDS_StarWars
    - BRANDS_Star
    - BRANDS_Disney
    - BRANDS_Natgeo
    - BRANDS_Pixar
- Cantidad de complete streams por cuenta sobre los títulos de cada franquisia:
    - franchise_WDSLA
    - franchise_DC.
    - franchise_DCOM_DC
    - franchise_Villains
    - franchise_DJ
    - franchise_Cars
    - franchise_ToyStory
    - franchise_DisneyXD
    - franchise_MickeyFriends
- Cantidad de complete streams por cuenta sobre los títulos de cada compañía productora:
    - Prod_Com_Walt_Disney_Pictures
    - Prod_Com_Disney_Channel
    - Prod_Com_Pixar
    - Prod_Com_Walt_Disney_Productions

    - Prod_Com_20Th_Century_Fox
    - Prod_Com_Marvel
    - Prod_Com_Lucasfilm
    - Prod_Com_Walt_Disney_Television_Animation
    - Prod_Com_Walt_Disney_Animation_Studios Walt_Disney_Animation
    - Prod_Com_Walt_Disney_Television
    - Prod_Com_Disney_XD
    - Prod_Com_Disney_Junior
    - Prod_Com_Disneynature
    - Prod_Com_ABC_Family
    - Prod_Com_National_Geographic
    - Prod_Com_DisneyToon_Studios
    - Prod_Com_Disneyplus_Studios
- Cantidad de complete streams por cuenta sobre los títulos de cada unidad de negocios:
    - BU_STUDIO_WDSLA
    - BU_DCWW_DISNEY_CHANNEL
    - BU_STUDIO_ANIMATED
    - BU_NATIONAL_GEOGRAPHIC
    - BU_STUDIO_FOX
    - BU_PIXAR
    - BU_DISNEYPLUS
    - BU_MARVEL_STUDIOS
    - BU_DCWW_DISNEY_JUNIOR
    - BU_DCWW_INTERNATIONAL
    - BU_DCWW_DISNEY_XD
    - BU_MARVEL_ENTERTAINMENT
    - BU_STUDIO_OTHER
    - BU_STUDIO_DISNEYNATURE
    - BU_STUDIO_BLUE_SKY
- Cantidad de complete streams por cuenta sobre los títulos de cada género:
    - Comedy
    - Animation
    - Action_Adventure
    - Fantasy
    - Kids
    - Coming_of_age
    - Science_Fiction
    - Drama
    - Animals___Nature
    - Musical
    - Documentary
    - Romance
    - Docuseries

    - Sports
    - Music
- Cantidad de complete streams por cuenta sobre los títulos lanzados dentro de cada período:
    - period0_1960
    - period1960_1965
    - period1965_1970
    - period1975_1980
    - period1980_1985
    - period1990_1995
    - period2000_2005
    - period2005_2010
    - period2010_2015
    - period2020_2025
- Cantidad de complete streams por cuenta, sobre los X títulos de X tipo de contenido, más vistos en Disney Plus.
    - WATCH_TOP5_SERIES
    - WATCH_TOP10_SERIES
    - WATCH_TOP5_MOVIES
    - WATCH_TOP10_MOVIES
    - WATCH_TOP5_SHORTS
    - WATCH_TOP10_SHORTS

Variables de diversidad de contenido

Se calculará el porcentaje de título vistos de distintas categorías de un atributo del contenido, para definir variables relacionadas a la diversidad del perfil respecto a dichos atributos.

Cuando una de las categorías tenga más de cierto porcentaje de 'streams' respecto al total, se considerará a esa cuenta como "mono". Así se construyen las variables:

- MONO_TYPE: Vale 1 cuando la cuenta ve más del 70% de un solo tipo de contenido (entre series, películas y cortometrajes). 0 en caso contrario.
- MONO_CENTROID: Vale 1 cuando la cuenta ve más del 50% del contenido, de un solo cluster. 0 en caso contrario.
- MONO_BRAND: Vale 1 cuando la cuenta ve más del 50% del contenido, de un solo marca. 0 en caso contrario.

Cuando dos categorías tengan más de cierto porcentaje de 'streams' respecto al total, se considerará a esa cuenta como "BI". Así se construyen las variables:

- BI_TYPE: Vale 1 cuando la cuenta ve más 25% de una clase de contenido y más del 25% de otra clase. 0 en caso contrario.
- BI_CENTROID: Vale 1 cuando la cuenta ve más 30% de un cluster de contenido y más del 30% de otro cluster. 0 en caso contrario.

- BI_BRAND: Vale 1 cuando la cuenta ve más 30% de títulos de una marca y más del 30% de títulos de otra. 0 en caso contrario.

Se considerará a una cuenta como "MULTI" cuando los porcentajes de streams sobre el total, para distintas categorías sean superiores a cierto porcentaje. Así se construyen las variables:

- MULTI_CENTROID: Vale 1 cuando la cuenta vio más del 10% del contenido en un cluster, para más de 5 clusters distintos. 0 en caso contrario.
- MULTI_BRAND: Vale 1 cuando la cuenta vio más del 10% del contenido de una marca, para más de 5 marcas distintas. 0 en caso contrario.

Variable de episodios por serie:

Se sumará la cantidad de episodios vistos por serie, por cuenta, para alguno de los títulos más relevantes. Se construyen las variables:

- EPISODES_SIMPSONS
- EPISODES_WANDAVISION
- EPISODES_MANDALORIAN
- EPISODES_JESSE
- EPISODES_MMCLUBHOUSE

Descriptivas del primer título: Se generarán variables categóricas para indicar si el primer contenido consumido por el suscriptor fue de ciertas características:

- CENTROID_ID: a qué cluster de contenido pertenece el primer título consumido por la cuenta
- PROGRAMTYPE: qué tipo de contenido pertenece el primer título consumido por la cuenta
- BRANDS: de qué marca es el primer título consumido por la cuenta
- BRANDS_DETAILED: de qué detalle de marca es el primer título consumido por la cuenta
- first_period: en qué period de tiempo fue lanzado el primer título consumido por la cuenta.

Flags: Se generarán *dummies* para indicar si el primer contenido consumido por el suscriptor fue de ciertas características:

- IS_PAY_1: indica 1 si el primer título consumido por la cuenta es de pago adicional, 0 caso contrario.
- IS_DISNEY_PLUS_ORIGINAL: indica 1 si el primer título consumido por la cuenta es original de Disney Plus, 0 caso contrario.
- IS_ANIMATED: indica 1 si el primer título consumido por la cuenta es animado, 0 caso contrario.

A continuación, el listado *dummies* que se generarán cuando el primer título consumido es de cierto género:

- Family

- Comedy
- Animation
- Action_Adventure
- Fantasy
- Kids
- Coming_of_age
- Science_Fiction
- Drama
- Animals___Nature
- Musical
- Documentary
- Romance
- Docuseries
- Sports
- Music

A continuación, el listado de dummies que se generarán cuando el primer título consumido es de franquicia:

- first_franchise_WDSLA
- first_franchise_DC
- first_franchise_DCOM_DC
- first_franchise_Villains
- first_franchise_DJ
- first_franchise_WDAS
- first_franchise_Cars
- first_franchise_ToyStory
- first_franchise_DisneyXD
- first_franchise_MickeyFriends

Anexo

Detalle de modelos y su implementación

1) Modelo K-medias

El modelo K-Medias es un modelo de aprendizaje automático no supervisado ideal para usar en etapas de análisis descriptivos. El modelo permite encontrar subgrupos de observaciones dentro de los datos tal que:

- Observaciones dentro del mismo cluster sean lo más similares entre sí.
- Observaciones de distintos clusters sean lo más distintas posibles.

Al utilizar el modelo k-medias, se debe indicar el parámetro K, que representará el número de clusters a armar. Cada cluster estará definido por un punto centroide, que será igual al promedio de todos los puntos en el conjunto, e ira cambiando en cada iteración del algoritmo. Se buscará asignar cada punto en uno de los K clusters, minimizando la sumatoria al cuadrado de la distancia entre cada punto del cluster y su centroide, para cada uno de los clusters. (Gareth, Witten, Hastie, Tibshirani. 2013)

Esta optimización es NP-difícil, por lo cual se utiliza una aproximación a través del algoritmo Lloyd, donde se parte de una asignación aleatoria de cada punto a un cluster, luego se asigna cada punto al centroide más cercano, se actualizan los puntos centroides y se continúa iterando hasta converger.

La primera asignación aleatoria de los puntos a los distintos clusters tiene algunas desventajas. En primer lugar, si la primera “semilla” (es decir, asignación al azar) es buena, el tiempo de convergencia puede ser corto, de lo contrario si esta asignación es mala, el algoritmo puede tardar mucho tiempo en converger. En segundo lugar, la optimización puede quedar atrapada en un mínimo local, y no encontrar una buena solución. Como consecuencia lógica, la asignación final de cada punto a un cluster, se verá completamente afectada por la asignación inicial, pudiendo armarse buenos o malos clusters, dependiendo del azar.

Para soslayar este problema se trabajará con el algoritmo K-Medias++, que implementa una técnica diferente de iniciación. Solamente el primer punto se asigna al azar. Para elegir el siguiente, se realiza un muestreo aleatorio de los puntos, eligiendo aquel que tenga mayor probabilidad de estar lo más lejos posible del primer punto seleccionado. De esta manera, no solo se logra una convergencia más rápida, sino que la calidad de los clusters finales es mejor. (Arthur, Vassilvitskii. 2007)

Para analizar la performance se computará el valor mínimo alcanzado de la distancia media al cuadrado y el índice de Davies-Bouldin, un esquema de evaluación interna, donde la validación de qué tan bien se ha realizado la agrupación se realiza utilizando

cantidades y características inherentes al conjunto de dato. A menor valor del índice, mejor será la calidad de los clusters.

Implementación del modelo: K-medias

Para entrenar el modelo dejaremos de lado algunas variables de nuestro dataset.

- "Idmix" y "titlemix": estas variables son identificativas de cada registro, es decir, son variables categóricas de n categorías, siendo n igual al total de registros. No agregan valor como atributo de agrupación.
- "PROGRAMTYPE" y "CONTENT_CLASS": no buscamos que los títulos se traten distinto por ser una película, serie o cortometraje; queremos que se agrupen por sus demás características.
- "Brazil", "Caribbean" y "LATAM_HISP": no buscamos que los títulos se traten distinto dependiendo su disponibilidad en cada región; queremos que se agrupen por sus demás características.

Las variables numéricas serán normalizadas, para dejar de lado cualquier número extremo. En los casos que la variable contenga nulos, se reemplazaran por el valor promedio de la variable en cuestión. (Zheng, Casari. 2016)

Si bien K-medias trabaja solo con variables numéricas, también incluiremos variables categóricas. Realizaremos one hot encoding, para generar una variable dummy por cada categoría que valga 1 en el caso que se corresponda con esa categoría y 0 en caso contrario. En el caso de que la variable categórica contenga nulos, se adiciona una columna dummy adicional, como si se tratase de una categoría más. (Zheng, Casari. 2016)

Definiendo el número de K.

Para definir el número de clusters final, entrenaremos 15 modelos que generaran de 5 a 20 clusters. Trabajaremos con las dos métricas de performance para seleccionar el mejor número de cluster. Graficaremos tanto el comportamiento de la distancia media al cuadrado y el índice de Davies-Bouldin, en función de la cantidad clusters generados en cada modelo. Para mas detalle ver E.1_b_BQML_Kmedias.

Una manera de elegir K, es analizando gráficamente el valor de la distancia media al cuadrado, alcanzado el modelo con k clusters. El valor óptimo de K suele darse en aquel punto donde la función grafica un codo. En el Gráfico 1 se trazaron dos líneas rojas estimando la línea recta entre los puntos para buscar el cambio de pendiente y en él, el "codo". Este se encontró en k = 10. (Gareth, Witten, Hastie, Tibshirani. 2013)

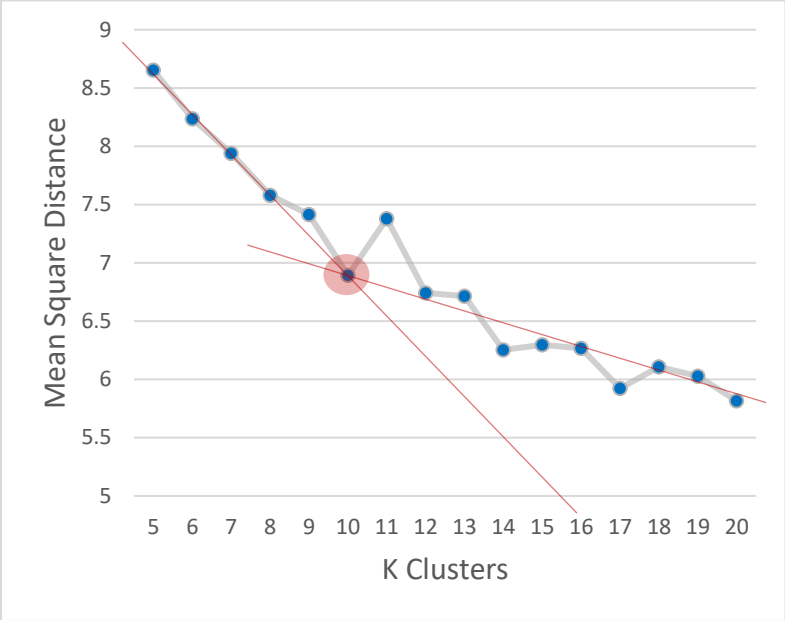

Gráfico 1: Valor la distancia media al cuadrado para cada modelo de K clusters.

Otra manera, es aquel K en donde el índice de Davies-Bouldin alcanza su menor valor. Esto también se da en k = 10, observando en el Gráfico 2 que en este K se alcanza el valor mínimo de la función, coincidiendo con el análisis del "codo" en la función de la distancia media al cuadrado.

Gráfico 2: Valor del índice de Davies-Bouldin para cada modelo de K clusters.

Concluimos por ambos métodos que el número óptimo de K para maximizar las diferencias entre títulos es 10, donde se alcanza la siguiente performance:

- Davies–Bouldin índex: 1.7946
- Mean squared distance: 6.8929

2) Descripción del modelo: XGBoost y DART

El algoritmo XGBoost es uno de los modelos más potentes y flexibles en ML para predecir una variable dependiente. Pertenece a la categoría de modelos de ensamble Boosting. Los modelos de ensamble combinan predicciones de modelos más pequeños. En este caso se ensamblan una secuencia de modelos de árboles simples (bastante débiles por si mismos).

Arboles: los pequeños modelos a ensamblar.

Los árboles son modelos de partición recursiva, que pueden utilizarse tanto para regresión como para clasificación. Como es imposible considerar todas las posibles particiones del espacio multidimensional, se toma una approach top-down, greedy, conocido como *recursive binary splitting.* En este caso buscamos predecir una variable categórica, binaria. Así, cuando una observación pertenece a una región, se le asigna la clase mayoritaria a cuál las observaciones de esa región pertenecen. Se buscará minimizar una medida de pureza (índice de gini o entropía), que suelen medir la variación total a lo largo de todas las K clases, captando que tan homogénea es la caja que analiza. A menor valor de la medida, más puro es el nodo. (Tan, Steinbach, Kumar. 2005) (Gareth, JWitten, Hastie, Tibshirani. 2013)

Boosting

El nombre del modelo XGBoost viene de "Extreme Gradient Boosting". En Boosting los árboles son construidos de manera secuencial usando información de árboles previos. Cada árbol se *fitea* modificando cosas de árboles anteriores. Secuencia del algoritmo:

1. Entreno un árbol de decisión.
2. Veo en qué observaciones de entrenamiento predijo mal.
3. Construyo un segundo árbol de decisión que se enfoque en aquellas observaciones que el primero predijo mal.
4. Tomo como mi predicción final alguna combinación de las predicciones de cada árbol.

XGBoost es un algoritmo de aprendizaje lento. Cada árbol suele ser pequeño, con pocos cortes. Cada uno predice un poco, por si solos no son muy potentes. Al evitar que un solo árbol prediga todo, se evita el sobreajuste.

Hiperparametros

El algoritmo al ser tan complejo, permite ajustes a través de múltiples hiperparametros. En la tabla 1 se detallan aquellos más relevantes. Los valores utilizados en el modelo para los hiperparametros fueron seleccionados utilizando técnicas de validación cruzada.

Parámetro	Descripción	Rango	Valor por default	Sensibilidad	
				ALTA	MEDIA
MIN_TREE_CHILD_WEIGHT	Peso mínimo necesaria en un hijo para generar un corte.	[0; ∞]	1	X	
MIN_SPLIT_LOSS (GAMMA)	Mínima reducción del error necesaria en una hoja para generar una nueva partición.	[0; ∞]	0		X
MAX_TREE_DEPTH	Profundidad máxima de un árbol.	$[0,\infty]$	6	X	
L1_REG (ALPHA)	La cantidad de regularización L1 aplicada.	[0; 1]	0	X	
L2_REG (LAMBDA)	La cantidad de regularización L2 aplicada.	[0; 1]	1	X	
LEARN_RATE (ETA)	Proporción que aprende de cada árbol.	[0; 1]	0.3	X	
MAX_ITERATIONS (nrounds)	Cantidad de árboles a construir.	(0;∞]	20	X	

Tabla 1: hiperparametros del algoritmo XGBoost

En XGBoost hay dos formas de controlar el overfitting. Por un lado, controlando directamente la complejidad del modelo. Para ello podemos utilizar los hiperparametros pintados de celeste. Otra alternativa es agregar mayor aleatoriedad a los datos, para sumar "ruido". Eso se puede lograr al utilizar en conjunto los hiperparametros pintados de verde; o a través de los siguientes hiperparametros que nos permitirán trabajar con muestras del total de los datos, para entrenar cada árbol. [14]

Parámetro	Descripción	Cuándo ocurre	Valor por default	Sensibilidad	
				ALTA	MEDIA
COLSAMPLE _BYTREE	Proporción de submuestra de columnas al construir cada árbol.	En cada árbol	1		X
COLSAMPLE _BYLEVEL	Proporción de submuestras de columnas para cada nivel. Las columnas se submuestrean del conjunto de columnas elegido para el árbol actual (COLSAMPLE_BYTREE)	En cada nuevo nivel de profundidad	1		X
COLSAMPLE _BYNODE	Proporción de submuestra de columnas para cada nodo (división). Las columnas se submuestrean del conjunto de columnas elegido para el nivel actual. (COLSAMPLE_BYLEVEL)	Cuando se evalúa una nueva división	1		
SUBSAMPLE	Proporción de los datos de entrenamiento para entrenar cada árbol.	En cada árbol	1		X

Tabla 2: hiperparametros del algoritmo XGBoost que logran contrarrestar el sobreajuste.

[14] https://xgboost.readthedocs.io/)

DART

XGBoost combina principalmente una gran cantidad de árboles con una pequeña tasa de aprendizaje. Así, los árboles que se agregan en las primeras iteraciones, son significativos y los árboles agregados más tarde no son tan importantes. Vinayak y Gilad-Bachrach propusieron un nuevo método para agregar técnicas de dropout de la comunidad de redes neuronales, para potenciar el ensamblado de árboles y en algunas situaciones reportaron mejores resultados. Este algoritmo se llama DART Booster. El mismo logra descartar árboles en el entrenamiento para evitar el sobreajuste y prevenir la construcción de árboles triviales (para corregir errores triviales).

Parámetros:

DART utiliza los mismos hiperparametros que XGboost, y agrega los detallados en la tabla 3:

Parámetro	**Descripción**	**Valor por default**
sample_type	Tipo de muestreo del algoritmo. 'Uniform' descarta arboles al azar mientras que 'weighted' descarta los árboles en proporción al peso.	Uniform
normalize_type	Tipo de normalización del algoritmo. 'tree' cada nuevo árbol tiene el mismo peso que cada uno de los arboles descartados mientras que 'forest' cada nuevo árbol tiene el mismo peso que la sumatoria de todos los arboles descartados.	Tree
rate_drop	Tasa de descarte. Rango: [0.0, 1.0]	0
skip_drop	Probabilidad de evitar el descarte. En este caso se armarían los mismos árboles que XGbosot. Rango: [0.0, 1.0]	0

Tabla 3: hiperparametros específicos del algoritmo Dart.

Sistema de validación: Training-Validation-Testing Set

Dado que los modelos a entrenar tienen varios hiperparametros, se trabajará separá la totalidad de los datos en:

- Entrenamiento: El conjunto de datos usado para entrenar el modelo. El modelo ve y aprende de estos datos.
- Validación: La muestra de datos utilizada para evaluar el modelo, que se ajusta al conjunto de datos de entrenamiento, mientras se ajustan los hiperparametros. La evaluación se vuelve más sesgada a medida que la habilidad del conjunto de datos de validación se incorpora a la configuración del modelo.

- Testeo: La muestra de datos utilizada para proporcionar una evaluación imparcial del modelo final que se ajusta al conjunto de datos de entrenamiento.

Dado que nuestros datos llegan al millón de registros, separaremos para validación un 2.5% y para testeo otro 2.5%. Los datos se dividirán al azar y la división será determinística, es decir que mientras los datos de entrenamiento sean los mismos, las diferentes ejecuciones de entrenamiento producirán los mismos resultados de división.

Bibliografía

- Aresté Sancho, J. M. (2021). “La Guerra Del Streaming. El Ascenso De Netflix. EDICIONES RIALP S.A
- Neira, E. (2020). “Streaming Wars: La nueva televisión”. Libros Cúpula
- ALCACER, J., COLLIS, D.j., Furey, M. (2010). “Walt Disney Company y Pixar Inc.: ¿Adquirir o no adquirir?”. Harvard Business School.
- Roels, G., Carrick, a. m. 2018. “Reshaping Disney’s Strategy for the Digital Age”. INSEAD
- Kaya, Erdem, et al. Dec. 2018. “Behavioral Attributes and Financial Churn Prediction.” EPJ Data Science, vol. 7, no. 1, © 2018 The Authors
- Abrardi, L., Cambini, C., & Rondi, L. (2019). “The economics of Artificial Intelligence: A survey”. Robert Schuman Centre for Advanced Studies Research.
- Dans, E. (2019) “Netflix: big data, long game… y resultados”. Extraído de https://www.enriquedans.com/2020/01/netflix-big-data-long-game-y-resultados.html
- Gilbert, C., Eyring, M., Foster, R. N. (2012). “Two Routes to Resilience”. Extraído de https://hbr.org/2012/12/two-routes-to-resilience
- Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). “An introduction to statistical learning”. Springer
- Bramer, M. (2007). “Principles of data mining”. Springer
- Tan P.N., Steinbach M., Kumar, V. (2005). “Introduction to Data Mining”. Pearson
- Alpaydin, E. (2014). “Introduction to machine learning”. MIT press.
- Friedman, J., Trevor H., Tibshirani, R. (2001). “The elements of statistical learning”. Springer.
- Zheng, A., Casari, A. (2016). “Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists”. O'Reilly Media.
- Arthur, D., Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding". Extraído de http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf
- Rashmi, K. V., Gilad-Bachrach, R. “DART: Dropouts meet Multiple Additive Regression Trees”. Extraído de http://proceedings.mlr.press/v38/korlakaivinayak15.pdf
- Dal Pozzolo, C., Bontempi, J. Bontempi, G. (2015). “Calibrating Probability with Undersampling for Unbalanced Classification”. Extraído de https://www3.nd.edu/~dial/publications/dalpozzolo2015calibrating.pdf
- Hu, Y., Koren, Y., Volinsky, C. (2008) “Eighth IEEE International Conference on Data Mining”. Extraído de http://yifanhu.net/PUB/cf.pdf