Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Predicción de Espectadores en un Complejo de Cines

AUTOR: ALFREDO ANTONIO JUÁREZ

MENTOR: NÉLIDA MÓNICA CANTONI RABOLINI

DICIEMBRE 2020

Resumen

Durante los últimos años la industria de exhibición cinematográfica enfrenta desafíos que afectan su rentabilidad. Desde la irrupción de nuevas tecnologías, cambios en los comportamientos de los consumidores, y achicamiento del período de exclusividad en la exhibición de títulos. Una gestión basada en datos es necesaria para tomar decisiones estratégicas, y así lograr que organizaciones de este sector se adapten a los cambios.

El principal objetivo de este trabajo es abrir la discusión sobre el uso predictivo que se le puede otorgar a los datos. Se buscarán modelos de aprendizaje automático que logren predecir la cantidad de público en un complejo de cine. Se plantea la hipótesis de que la cantidad de espectadores de las funciones depende de distintas variables, desde la fecha de exhibición o características de la película proyectada.

Se comenzará estudiando la organización elegida, planteando sus objetivos y desafíos en la gestión de datos. Se analizará una base de datos de funciones, la cual será utilizada para entrenar modelos buscando predecir la cantidad de público. Finalmente, se plantearán recomendaciones acerca del uso de metodologías ágiles para la implementación de modelos de aprendizaje automático, y se elaborarán conclusiones, explicando los resultados obtenidos y planteando interrogantes para futuros trabajos.

Palabras clave: attendance[1], espectadores, predicción, aprendizaje automático, gestión de datos.

[1] Anglicismo muy usado en la industria del cine, significa espectadores

Índice

1. Introducción

1.1 Fundamentación de la elección del tema

Durante la última década la industria de exhibición cinematográfica ha estado enfrentando nuevos desafíos relacionados a la irrupción de nuevas tecnologías, como el internet y el streaming[2]. La industria del entretenimiento es muy competitiva de por sí, ya que los individuos siempre buscan nuevas experiencias. Es por eso por lo que la innovación y herramientas de gestión son muy necesarias en esta industria. Desde la irrupción de nevos formatos como el 4D[3] y Dbox[4], la diversificación de los productos vendidos en la concesión de alimentos y bebidas, y el uso de los datos para poder aprender sobre el comportamiento de los consumidores. La gran diversidad de películas, los distintos tipos de público hacen que la decisión de la programación a exhibir sea fundamental para el negocio. Salas repletas o salas totalmente vacías son situaciones comunes para un complejo cinematográfico, por lo que un modelo que pueda predecir los espectadores de una sala es fundamental para poder tomar una decisión óptima a la hora de elegir que películas se exhiben.

Uno de los principales problemas de la industria cinematográfica es la heterogeneidad del servicio ofrecido. En la industria del retail[5] los mismos productos se ofrecen durante todo el año. La exhibición de cine por su parte depende de las películas que estén disponibles. Además, en la era digital, cada vez más los espectadores quieren ser los primeros en ver un blockbuster[6], evitar posibles spoilers[7], compartir en las redes sociales sobre las películas, es parte del comportamiento del nuevo cliente. En consecuencia, no es lo mismo una función en el primer fin de semana de estreno de una película, que pasadas varias semanas.

[2] Anglicismo. Tecnología que permite ver contenidos que se transmiten desde internet sin la necesidad de descargar previamente los datos.
[3] Sistema de proyección de películas que recrea en la sala de proyección las condiciones físicas que se ven en la pantalla, como niebla, lluvia, viento, sonidos más intensos u olores, así como vibraciones en los asientos y otros efectos.
[4] Butacas que se sincronizan con la acción en pantalla y están programadas para moverse con la película.
[5] Anglicismo. Es un sector económico que engloba a las empresas especializadas en la comercialización masiva de productos o servicios uniformes a grandes cantidades de clientes.
[6] Anglicismo, hace referencia a una película que es un éxito en taquilla.
[7] Anglicismo. Explicación de algún aspecto importante de una película, libro, etc., que a una persona que lo desconozca le puede resultar molesto.

A su vez, la oferta es limitada, ya que se cuenta con una capacidad limitada de salas y butacas, por lo que la decisión de elegir qué película exhibir en cada función plantea un problema de optimización de por sí. Además, los períodos "ventana" en los cuales los complejos tienen la exclusividad de transmitir el contenido es cada vez menor. La pregunta a plantear es, como puede un complejo de cine adaptarse a estos cambios que se van produciendo en el comportamiento de los consumidores.

1.2 Estado actual del conocimiento

La bibliografía referente a la temática descripta no es abundante, ya que se trata de un tema específico de un rubro en particular. Por su parte, gran parte de la bibliografía obtenida plantea el tema desde el punto de vista del distribuidor, buscando estimar el attendance o la recaudación total de una película. Un ejemplo es el trabajo de Hand, Chris & Judge, Guy (2017) el cual utiliza las búsquedas de Google para predecir la recaudación total de una película.

Rhee and F. Zulkernine, (2016) y Zhou, Y., Zhang, L. & Yi, Z (2019) usan redes neuronales para predecir la recaudación total por película tomando como variables las críticas, el presupuesto, el género o los actores que participan en las mismas. Otro ejemplo de redes neuronales utilizadas a la hora de predecir público es Şahin, M.; Erol, R. A (2017), el cual usa uno de estos modelos para predecir el público en eventos deportivos. Temática similar se encuentra en King, B.E., Rice, J., & Vaughan, J. (2018), quienes usan técnicas de machine learning como random forest y gradient boosting para predecir la cantidad de espectadores en partidos de hockey.

Eliashberg, J., Elberse, A., & Leenders (2006) realizan una buena descripción de la problemática que enfrenta la industria cinematográfica. Analizan el ciclo de vida de cada producto desde la producción hasta la exhibición. Analizan por ejemplo si las secuelas suelen ser más populares que las películas originales, y mencionan como las secuelas sirven a los estudios para crear franquicias. Por su parte, analizando el lado de la distribución, explican cómo los complejos cinematográficos dependen de un limitado número de blockbusters anuales para la recaudación.

El trabajo de Marshall, Dockendorff e Ibáñez (2006), utiliza un modelo Bass para modelar el ciclo de vida de una película, estimando el mercado potencial de la misma mediante un modelo econométrico. Gevaria, Wagh, y D'mello (2015) también busca estimar el attendance total de una película. Toma datos de IMDB y utiliza dos modelos,

uno de Sawhney y Eliashberg y un modelo Bayesiano jerárquico. Utiliza variables como fecha de estreno de la película, genero, y críticas.

Por su parte, el trabajo de Lim, J. (2012) tiene más similitudes tiene con este trabajo ya que se busca estimar el attendance individual de cada función. El autor realiza un estudio a partir de un modelo Bayesiano jerárquico. Este modelo presenta dificultades a la hora de introducir nuevas películas. Otro trabajo que resultó muy útil para el proyecto es el de Baranowski, Korczak & Zając. El mismo trata de hacer forecasts de corto plazo, teniendo la programación de los complejos, usando variables propias de los cines, de las regiones y de las películas. Estos dos trabajos son los que mejor se adaptan a lo que se quiere elaborar en este proyecto.

1.3 Objetivos e hipótesis

El objetivo principal del trabajo lograr predecir la cantidad de público o attendance de un complejo de cine, por función, día, semana y mes, utilizando modelos de aprendizaje automático. Contar con una predicción certera sobre esta variable es importante para el negocio ya que permite estimar los beneficios futuros y tomar decisiones estratégicas. Los objetivos específicos son los siguientes: en primer lugar, analizar los métodos por los cuales la organización elegida recopila, utiliza y analiza los datos para sugerir oportunidades de mejora. Otro objetivo es poder determinar cómo influye cada variable a la hora de explicar de qué depende la afluencia de público. Finalmente, otro objetivo de este trabajo es plantear distintos usos que se le pueden dar a un modelo de predicción de público. Se puede usar una herramienta para ayudar a decidir la programación de títulos en las salas; o para elaborar otras predicciones como ventas de alimentos y bebidas o churn rate[8] de clientes.

La principal hipótesis del trabajo es que existen distintos factores que explican la afluencia de espectadores a una función lo que permite predecir la cantidad de público del complejo. Mediante el conocimiento que se tiene sobre la industria, se plantea que estos factores son principalmente el día de la semana y la franja horaria de exhibición. Es claro que un sábado a la noche tendrá más público que un martes por la mañana. A su vez, la cantidad de semanas que una película lleva en cartel incide en el público interesado en verla. Se observa que la cantidad de espectadores disminuye progresivamente al

[8] Anglicismo utilizado en Marketing para hacer referencia a la tasa de abandono de clientes.

trascurrir las semanas. Se plantea que es posible predecir el público para un complejo de cines con la información sobre las funciones y las películas.

Cómo las películas cambian constantemente, será indispensable agruparlas según sus características y géneros. Se plantea como hipótesis que las características de las películas, como su género, distribuidora, presupuesto, país de origen, inciden en la cantidad de público que reciben. Contar con esta información para las películas aun no estrenadas será vital para poder predecir el attendance del complejo.

1.4 Contenidos y apartados

Tras este apartado introductorio, se continuará con un apartado que describa y analice a la organización elegida y su uso y gestión de datos. Se plantearán los objetivos de la organización, su modelo de negocio y el contexto en la industria de exhibición cinematográfica. Se abordará la temática gestión de datos de la organización, comenzando por describir las distintas fuentes de datos que se cuentan, el proceso de recolección, almacenamiento y uso de los datos, reflexionando sobre la transición de la empresa hacia una organización basada en datos. Finalmente se plantearán los desafíos existentes, mencionando oportunidades de mejoras en el uso de datos.

En un tercer apartado, se realizará una descripción metodológica sobre la base de datos elegida para la construcción del modelo predictivo. Se explicará como se obtuvieron los mismos, se describirá cada variable y se comentará el proceso de limpieza de la base de datos. Se realizará un análisis estadístico descriptivo mediante gráficos para explicar el razonamiento detrás de las hipótesis planteadas, es decir, mostrar como incide cada variable a la hora de explicar el público. En el último apartado se mostrará el proceso de implementación de los modelos elegidos. Estos serán comparados mediante distintos indicadores para poder elegir un modelo con mayor capacidad predictiva. Se observará como el modelo predice la cantidad de público por función, día, semana y mes; buscando así las fortalezas y debilidades del modelo.

Finalmente, se cerrará el trabajo con una conclusión que recapitule las principales afirmaciones conceptuales. Se evaluará si se respondió el interrogante inicial y si se lograron los objetivos propuestos. Se tratará de abrir una discusión acerca de las implicaciones del trabajo, incluyendo sugerencias, recomendaciones e interrogantes que planteen líneas futuras de investigación.

2. Gestión de datos en contextos organizacionales

En el siguiente apartado se describirá a la organización utilizada para la elaboración de este trabajo, describiendo su modelo de negocio y enmarcándola dentro de la industria a la cual pertenece, desarrollando brevemente su contexto histórico en Argentina. Luego, se profundizará sobre el uso y gestión de datos por parte de dicha organización, describiendo y clasificando los procesos de recolección, almacenamiento y uso de los datos. Finalmente, se mencionarán los desafíos que la organización posee en relación con el uso de los datos, describiendo las problemáticas actuales y mencionando las posibilidades de mejoras.

2.1 Descripción de la organización

2.1.1 Objetivos de la organización

La organización que se utilizará para el armado de este trabajo se trata de una cadena de complejos cinematográficos con sede en Estados Unidos y presencia en varios países de Latino América. Esta organización privada cuenta con 534 teatros de cine en todo el continente. En Estados Unidos es el tercer exhibidor más relevante en términos de Market Share[9]. En Argentina, con 21 complejos en todo el país, es el principal actor en la industria de exhibición cinematográfica, contando con aproximadamente un tercio del Market Share de espectadores.

El objetivo de la organización es proveer de un servicio de excelencia en lo que respecta a la exhibición cinematográfica, tomando como servicio la experiencia completa del cliente desde que toma decisión de comprar una entrada. La experiencia del cliente se complementa con un programa de fidelización, en el cual los socios tienen dos entradas mensuales garantizadas y suman puntos por todas sus compras, además de acceder a descuentos y beneficios exclusivos. Este es otro objetivo de la compañía, lograr fidelizar a los clientes. Por último, el objetivo principal es generar valor agregado y generar dividendos. La organización cotiza en la Bolsa de Valores de Nueva York y presenta sus resultados de EBITDA[10] a sus accionistas de forma trimestral.

[9] Anglicismo que refiere al porcentaje de participación en el mercado.
[10] Indicador financiero. Acrónimo de los términos en inglés Earnings Before Interest Taxes Depreciation and Amortization. (Beneficio antes de intereses, impuestos, depreciaciones y amortizaciones)

2.1.2 Modelo de Negocio

El modelo de negocio de la compañía se puede clasificar en dos operaciones intrínsecamente relacionadas. En primer lugar, el núcleo del negocio es la venta de entradas de cine. Con lo que respecta a la tecnología de proyección y sonido, la compañía se asegura un standard de calidad en todas sus filiales, asociando a la marca con excelencia en la exhibición. Para una mejor experiencia del cliente se ofrecen distintos formatos de exhibición, desde los tradicionales 2D y 3D, formatos con movimiento de butacas como el DBOX y el 4D, y un formato Premium con butacas más grandes y oferta gastronómica.

El otro gran aspecto del negocio es la venta de alimentos y bebidas en la concesión del cine también llamado Candy. Además de los ya conocidos pochoclos y gaseosas se ofrecen otros tipos de comidas rápidas y golosinas, merchandising asociado a las películas en cartel, y una oferta gastronómica más elaborada para las salas Premium. Por último, la tercera fuente de ingresos de la organización son los acuerdos comerciales con socios estratégicos. En estos acuerdos se otorgan descuentos 2x1 a los clientes de estas compañías.

Con lo que respecta a la estructura de costos, aproximadamente la mitad de lo recaudado en concepto de entradas debe ser pagado a las distribuidoras que comercializan las películas en exhibición. Los productos ofrecidos en la concesión de alimentos y bebidas suelen tener un costo de mercadería de aproximadamente un 25%. Por su parte se le abona un alquiler al Locador del complejo, en general las cadenas de Shoppings. Este suele rondar en un 10% de lo recaudado tanto en entradas como en productos del Candy. A esto se le suman costos laborales e impuestos que en Argentina son más altos que en otros países de la región. A su vez, parte de los ingresos se destina a inversión para garantizar que el equipamiento y las condiciones edilicias de los complejos estén en óptimas condiciones.

2.1.3 La Industria de Exhibición Cinematográfica

Analizando la industria de exhibición cinematográfica en su conjunto se observa que la cantidad de público en todo el país ronda entre los 40 y 50 millones de espectadores. Estas cifras se mantienen durante toda la última década lo que denota una industria estable y de poco crecimiento. Esto implica una entrada de cine por habitante en promedio, cuando en los años ochenta se superaban las dos entradas vendidas por habitante.

Las oscilaciones en la cantidad de público no suelen depender de los vaivenes de la economía argentina, sino del grado de atractividad de las películas en cartel. A modo de ejemplo, en el 2019 el 45% del público de cine estuvo concentrado en las 5 películas más taquilleras. Esto denota una clara estacionalidad en lo que respecta a la afluencia de público. Las vacaciones de invierno son el pico de espectadores y coinciden con los principales estrenos de películas infantiles de las grandes distribuidoras.

Por su parte, el público suele tratar de asistir a las películas en la primera semana de su estreno, potenciando aún más la estacionalidad y la dependencia de contar con nuevos títulos. El comportamiento del espectador ha cambiado a lo largo de las décadas. El cliente promedio dejó de ver al cine como una salida habitual, sino que está informado sobre las películas que desea ver, y decide en relación con los títulos específicos. Los cambios en los hábitos de los consumidores, producto de las nuevas tecnologías, presentan un gran desafío para la industria. El auge del streaming, y el acortamiento del período de exclusividad antes del cual las películas están disponibles por fuera del cine, hace que sea muy importante otorgar valor agregado a la experiencia del cliente.

Las cadenas de cine no buscan aumentar la cantidad de público sino potenciar la rentabilidad de los complejos, ofreciendo nuevos formatos y productos variados en el Candy. A su vez, se busca reducir la estructura de costos, mediante la digitalización de los servicios. Opciones de compra vía Internet o por máquinas de auto servicio en los complejos ayudan a reducir el costo laboral de la operación. Por otro lado, las promociones con acuerdos comerciales y los distintos programas de fidelización buscan que los clientes continúen visitando la misma cadena.

2.2 Gestión de Datos por parte de la organización

2.2.1 Descripción de los Datos

Para afrontar los desafíos descriptos en el apartado anterior, teniendo en cuenta la heterogeneidad de los títulos exhibidos y el comportamiento del cliente, es muy importante el uso del volumen de datos generado por las transacciones a la hora de tomar decisiones. La Organización en cuestión es un claro ejemplo de cómo una Empresa de un rubro tradicional, no vinculada a la era digital, puede adaptarse y hacer uso de estas nuevas herramientas. Las herramientas de Big Data permiten medir y en consecuencia conducir una empresa de forma más precisa que antes. Se pueden elaborar predicciones

que permitan tomar mejores decisiones, remplazando la intuición por estrategias basadas en datos y rigor metodológico (McAfee & Brynjolfsson 2012).

En esta organización se cuentan con datos internos tanto como externos que se usan para la elaboración de reportes y análisis que ayuden a la toma de decisiones. Cada transacción correspondiente a la compra de entradas y a los productos en la concesión de alimentos y bebidas se registra en la base de datos, con el mayor nivel de detalle posible. A su vez, se cuenta con una base de datos sobre los clientes que acceden al programa de fidelización, pudiendo identificar sus consumos mediante su ID. Una fuente externa de datos es la obtenida con la herramienta de Google Analytics, la cual permite tener una información detallada sobre el uso del sitio web de la cadena de cines. Desde cantidad de visitas, compras y tiempo de permanencia en el sitio. Por último, todos los complejos de cine deben de informar la cantidad de público diario a los organismos de control, permitiendo que los datos sobre los espectadores de competidores estén disponibles.

2.2.2 Proceso de recolección, almacenamiento y uso de los Datos

A continuación, se estudiará el proceso de recolección, almacenamiento y uso del dato para transformarlo en información. Cada transacción que se realiza en el cine ya sea mediante la web, las terminales de autoservicio, las boleterías o el Candy, queda registrada mediante la plataforma Vista[11], un software de gestión específico para complejos de cine. Cada dato cuenta con información específica de la fecha y hora de compra, la función con su respectivo horario de exhibición, película, formato e idioma; el tipo de ticket y el monto abonado. En el caso de los productos del Candy se cuenta con la fecha y hora de la transacción y el producto asociado, el monto abonado y el costo del producto. Como en general las transacciones de boletería y concesiones se hacen de forma separada, es imposible cruzar ambas. La excepción es para el grupo de clientes pertenecientes al programa de fidelización, los cuales están identificados.

Estos datos se almacenan en un servidor y se utiliza SQL[12] para generar bases de datos que podrán ser usadas a la hora de elaborar reportes. Se utiliza el software Tableau[13] para consultar estas bases de datos y generar distintos tipos de reportes de KPI[14]. Los

[11] Software de venta de entradas de cine utilizado por las principales cadenas a nivel mundial.
[12] Lenguaje de dominio específico utilizado en programación, diseñado para administrar, y recuperar información de sistemas de gestión de bases de datos relacionales.
[13] Software de visualización de datos interactivos que se enfocan en inteligencia empresarial.
[14] Acrónimo de los términos en Inglés Key Performance Indicator (Indicador clave de rendimiento)

reportes son enviados de forma mensual o semanal. También se generan tableros dinámicos de control donde cada gerente de área puede consultar para el período y agregación deseada.

Los KPI más usuales que se reportan a la gerencia y distintas áreas son los siguientes: Cantidad de público, precio promedio de entrada, ocupación de las salas, Market Share de espectadores y de recaudación, ventas promedio del Candy por espectador, horas trabajadas por los empleados sobre la cantidad de espectadores, porcentaje de compras realizadas por el sitio web, penetración del programa de fidelización sobre el total de espectadores. Esta información generada a través del uso de datos puede ser agregada por semana, mes, día de la semana, franja horaria, complejo, etc. Se utiliza para tomar iniciativas concretas como puede ser una estrategia de precios diferenciada, a través de promociones específicas. Luego se analiza el impacto de estas iniciativas en términos de Market Share de la compañía y EBITDA.

2.2.3 Transición hacia una Organización basada en Datos.

Como una empresa de rubro tradicional que comienza a usar grandes volúmenes de datos para la toma de decisiones, la aplicación de estas técnicas se da de forma paulatina y desigual en distintos sectores de la compañía. Las etapas de la aplicación de técnicas de Business Analytics se pueden clasificar en Aspiracional, Experimentado y Transformado (Lavalle et al., 2011). En el siguiente cuadro se buscará clasificar distintos aspectos de la compañía según estas etapas.

Categoría	Etapa	Detalle de la Organización
Competencia Funcional	Aspiracional	• Se usan las herramientas de BI para los procesos de management y presupuesto y para las estrategias comunicaciones de Marketing para atraer clientes. • Dado a la naturaleza de la industria no se desarrollan nuevos productos ni se cuenta con un servicio al cliente personalizado. Exceptuando los socios del programa de fidelización no se cuenta con información sobre los clientes del cine. • Tampoco se pueden usar estas herramientas para un manejo efectivo y óptimo de la fuerza de trabajo por las trabas impuestas por el Sindicato y una legislación laboral no actualizada a las nuevas formas de organización empresarial.

Desafíos del Negocio	Experimentado	• Se busca diferenciarse de los competidores mediante estrategias innovadoras, facilidad a la hora de comprar entradas y productos mediante la web o terminales automáticas. • Se pone el énfasis en una eficiencia en costos, característica de una etapa Aspiracional. Dada la naturaleza del negocio es difícil poner el foco en el crecimiento de las revenues, ya que la cantidad de espectadores suele ser estable y depender de los títulos. • Por otro lado, se hacen muchos esfuerzos para retener a los clientes mediante el programa de fidelización. Esta es una característica de la etapa Transformada. Se hace un seguimiento más exhaustivo sobre este nicho de clientes buscando otorgarles nuevos beneficios.
Manejo de los Datos	Transformado	• Se cuenta con una gran cantidad de datos sobre cada transacción de la compañía. Estos se almacenan y son de fácil acceso. • Los datos son agregados y analizados para generar información valiosa a la hora de conocer la realidad del negocio. Se arman distintos reportes de KPI que pueden ser actualizados de forma diaria, para contar con la información en tiempo real. • Esta información es fácilmente compartida con el resto de las áreas de la compañía con el objetivo de colaborar en su gestión.
Analytics en Acción	Experimentado	• Las decisiones son tomadas a partir de la información que generan los datos, analizando las distintas variables o KPI. Sin embargo, la información tiene un carácter descriptivo, no predictivo. Las estrategias se definen a partir de la realidad que muestran los datos, no sobre estimaciones a futuro usando los mismos. • Si bien la toma de decisiones a nivel gerencial está basada sobre los datos, la gestión operacional diaria de cada complejo se realiza de forma tradicional.

A partir de los descripto en la tabla se puede concluir en que la empresa está en vías de convertirse en una organización basada en datos. Las herramientas con las cuales se cuentan representan una base sólida en lo que refiere al manejo de los datos. Sin embargo, el uso de datos en lo que refiere a la toma de decisiones estratégicas, gestión y solución de desafíos presenta posibilidades de mejoras. Sobre este tema se continuará en el siguiente apartado.

2.3 Problemática de la organización y la gestión de los datos

2.3.1 El problema de la trazabilidad de los datos

Analizando el uso de los datos por parte de la organización se pueden identificar dos tipos de problemáticas. En primer lugar, no es posible determinar una trazabilidad de los datos. La Trazabilidad se define como la capacidad de retener la identidad de un producto y su origen (Khabbazi et al., 2010). Para alcanzar la trazabilidad de cada dato puntual es necesario identificar al individuo que contribuyó con el dato teniendo en cuenta las variables que identifican a dicho individuo. (Lusignan et al., 2011)

Cuando se registra una transacción no se identifica al cliente que la realiza, esto imposibilita asociar una compra de boletería con una compra en el Candy ya que estas son transacciones diferentes. En consecuencia, no se pueden establecer patrones de consumo de alimentos y bebidas en relación con los distintos tipos de películas. Tampoco es posible trazar la cantidad de visitas de cada cliente para poder determinar la frecuencia en la cual un espectador visita el cine.

La excepción a esta problemática se encuentra con los socios del programa de fidelización del cine, los cuales están registrados y cuentan con un ID que identifica cada una de sus transacciones. Si bien es posible usar este nicho de clientes como una muestra para inferir patrones de consumo, este tipo de clientes tiene un comportamiento distinto a la de los espectadores no asociados al estar fidelizados. La falta de información sobre los clientes que no se asocian al programa hace difícil estudiar su comportamiento con el objetivo de buscar herramientas para aumentar su consumo, o introducir nuevos productos en la concesión de Alimentos y Bebidas.

Incluso en el universo de los clientes fidelizados no se cuenta con variables relevantes a la hora de determinar un perfil de los clientes. Por ejemplo, no se cuenta con ninguna información sobre quienes acompañan al socio cada vez que visitan el cine, como la edad, género o relación de parentesco. Tampoco se cuenta con información geográfica sobre ningún cliente. Para el Área Metropolitana de Buenos Aires, donde existen numerosos complejos, es relevante conocer la zona de procedencia de cada cliente de un complejo, para determinar el área de influencia de este, identificando con más precisión a los cines competidores directos. A la hora de otorgar beneficios y descuentos se pone el énfasis en aquellos complejos donde un competidor está aumentando su Market Share.

2.3.2 Oportunidades de uso de los Datos

El otro gran problema que se observa en la organización es el uso que se le dan a la gran cantidad que se generan. Como se mencionó anteriormente, la información elaborada a partir de los datos suele tener un carácter descriptivo y no predictivo. Esta información funciona como termómetro sobre la salud del negocio, pero no se utiliza todo su potencial para lograr a ser una organización guiada por los datos y técnicas rigurosas de utilización de estos. Las aplicaciones más importantes que se le podrían dar a los datos son el *Targeted Marketing*[15], la publicidad online, recomendaciones de ventas cruzadas, identificación de clientes para prevenir su *attrition*[16] y maximizar el valor del cliente. (Provost & Fawcett, 2013).

En el caso concreto de la industria del cine, se pueden usar los datos para predecir la cantidad de público o las ventas en el Candy con relación a la programación futura de películas. También se pueden agrupar a los clientes según su patrón de consumo para identificar oportunidades de ventas cruzadas, poder predecir quienes podrían darse de baja del programa de fidelización, o quienes estarían dispuestos a darse de alta. Con los datos de cada cliente es posible establecer un sistema de recomendación de títulos; o incluso estrategias de captación de clientes para aquellos que vivan en zonas donde hay otro competidor fuerte. También es de gran utilidad realizar un análisis de sentimiento sobre los comentarios de los clientes en redes sociales para poder identificar oportunidades de mejora en el servicio.

Existen evidencias de como el despeño de un negocio puede mejorar sustancialmente mediante la toma de decisiones basada en datos y las tecnologías de ciencias de datos basadas en Big Data. Se estima que el modelo de toma de decisiones basadas en datos aumenta la productividad entre un 4 y 6%, mejorando también otros indicadores como la utilización de activos, rendimiento sobre el capital y valor de mercado (Brynjolfsson, Hitt & Kim, 2011). En este trabajo se buscará abordar una de estas oportunidades que la abundancia de datos genera. Se intentará poder predecir la cantidad de espectadores en un complejo determinado partiendo de los títulos en cartel y su programación.

[15] Anglicismo que hace referencia a estrategias de Marketing diferenciadas y focalizadas en grupos específicos de clientes.
[16] Anglicismo usado en Marketing, significa pérdida de clientes.

3. Descripción metodológica

En el siguiente apartado se describirá la metodología usada para recopilar la información de distintas fuentes, comentando cada una de las variables. Se proseguirá mencionando los distintos procedimientos de limpieza de errores o incongruencias de la base de datos para que pueda ser utilizada para el modelo predictivo. Finalmente se abordará un análisis estadístico descriptivo mediante gráficos que permitan explicar como incide cada variable sobre la cantidad de espectadores.

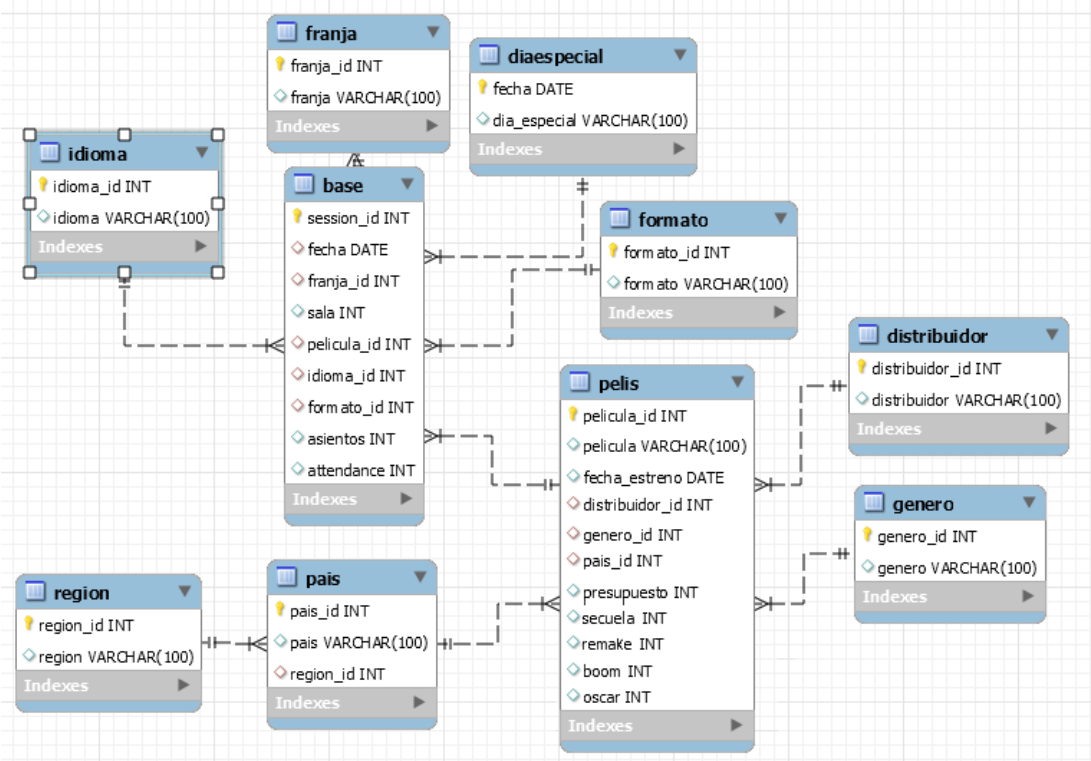
3.1 Recopilación de la información

Habiendo analizado las características de la organización, la industria en la cual pertenece y las oportunidades y desafíos que se plantean en el manejo de datos, se proseguirá por explicar la metodología aplicada a la hora de generar, procesar y analizar la información utilizada para el modelo predictivo. La recopilación de información puede clasificarse en datos internos de la empresa, generados con cada transacción en el complejo y datos externos. Estos últimos son los referentes a las variables relacionadas con las películas. En su gran mayoría fueron obtenidos del sitio Ultracine, el cual recopila información de las películas y de la industria cinematográfica en Argentina.

Habiendo recolectado los datos, se tiene una base de datos normalizada en la herramienta SQL. Se organizan los datos en distintas tablas relacionadas entre sí por foreing keys[17]. Esto se puede ver en el siguiente diagrama, el cual muestra cada tabla con sus variables, primary key[18] y foreing keys.

[17] Anglicismo usado en SQL traducido como Clave Externa. Es una columna o varias columnas, que sirven para señalar cual es la clave primaria de otra tabla.
[18] Anglicismo usado en SQL, Clave Primaria. identifica de manera única cada fila de una tabla.

Gráfico 3.1.1 Diagrama de tablas:



La tabla Base es la principal y se obtuvo con información interna del complejo en cuestión. La Primary Key corresponde al ID de la función, se cuenta también con el número de sala, la cantidad de asientos de la sala y la cantidad de público o attendance. Las siguientes foreign keys se relacionan con el resto de las tablas: fecha, franja ID, película ID, idioma ID y formato ID.

La variable fecha se usa para relacionar a la tabla Base con la tabla Dia Especial, cuya variable categoriza cada día en Nada, Feriado, Fiestas (24, 25 y 31 de diciembre y 1ro de enero) y vacaciones de invierno. La información referente a los feriados se obtiene del sitio web del gobierno nacional. Se separan los feriados de las fiestas del resto, porque el comportamiento del público es distinto. La información de las vacaciones de invierno se obtiene del calendario de la Provincia de Buenos Aires, ya que el complejo se encuentra en dicha provincia. Es importante distinguir los días de vacaciones de invierno por el fuerte impacto que tiene en la asistencia.

La variable Franja ID relaciona a la tabla Base con la tabla de Franja. La variable del mismo nombre distingue la franja horaria de la función. 1ra Matiné (11 AM a 2 PM) 2da Matiné (2 PM a 5 PM), Vermouth (5 PM a 8 PM), Noche (8 PM a 11 PM) y Trasnoche (a partir de las 11 PM). La tabla Idioma cuenta con una variable que indica el idioma de

la función, ya sea subtitulado o castellano. La tabla formato indica el formato de la función ya sea 2D, 3D, 4D (butacas con movimiento y efectos sensoriales), DBOX (Salas en las cuales algunas filas cuentan con movimiento) y Premium (Butacas más grandes y espaciadas, más variedad de oferta gastronómica). Las tres variables mencionadas en este párrafo se obtienen de la fuente interna de la organización.

La foreign key de Película ID vincula a la tabla Base con la tabla de Películas. En esta tabla Película ID es la primary key y se cuentan con variables que nombran y describen a las películas. La variable Fecha de estreno, como su nombre lo indica, muestra la fecha de estreno de la película. Distribuidor ID, Genero ID y País ID son foreign keys que vinculan a la tabla de Películas con las respectivas tablas que otorgan información de la compañía distribuidora de la película (Disney, Universal, Warner, etc); el género y el País de Origen. Las variables mencionadas hasta ahora en esta tabla fueron obtenidas del sitio de Ultracine. El país de origen también está relacionado con otra tabla que agrupa los mismos en regiones.

Por su parte, se añaden variables propias, del propio conocimiento de las películas y buscando información de estas en internet. La variable Presupuesto indica el costo de producción y distribución de la película expresado en millones de dólares. La variable secuela es binomial e indica si la película o no es una continuación de una película anterior. La variable remake, también binomial indica si la película es un remake de otra película previa. Se piensa que es importante incluir estas variables ya que son recursos usuales en la industria cinematográfica en los últimos años. Se busca asociar a las películas con tramas ya conocidas por los espectadores para aumentar las ventas. La variable binomial boom, también de creación propia indica o no si se espera que la película sea un éxito de taquilla. En la industria cinematográfica existen los denominados "Tanques", películas que atraen a una gran cantidad de público. Por último, la variable Oscar indica si la película fue nominada o no al Premio Oscar a la mejor película.

3.2 Procesamiento de la información

Habiendo indicado el proceso de obtención de la información y comentado las variables, se proseguirá a explicar la transformación de esta a una base de datos utilizable para el modelo predictivo. En primer lugar, se ejecuta una query[19] mediante la cual se genera una tabla con todas las variables previamente descriptas. En otras palabras, se

[19] Anglicismo usado en SQL, hace referencia a una consulta sobre las bases de datos.

desnormaliza la base de datos original. También, se crean nuevas variables que son resultado de un cálculo entre otras. La variable Semanas en Cartel es la diferencia medida en semanas entre la fecha de la función y la fecha de estreno de la película. La variable Funciones Diarias cuenta la cantidad de funciones de una misma película para el día de la función. La motivación detrás de la creación de la primera variable está en que una película recibe mucho más público en las primeras semanas para luego decrecer hasta que es retirada de cartel. Por su parte la cantidad de funciones diarias de cada película es una forma de mostrar la oferta, una película que es exhibida en exceso, o inferior a lo demandado, puede afectar al attendance individual de cada función. En el Anexo 7.1 se detalla un diccionario de datos de la base elegida.

A continuación, se explicará el proceso de limpieza de la base de datos. Para este proceso es fundamental contar con experiencia en el rubro, para identificar inconsistencias en los datos. Se comienza con eliminar las funciones correspondientes a las películas Nada que Perder y Nada que Perder 2. Ambas películas, de 2018 y 2019 respectivamente, son de temática religiosa y sus entradas se vendieron de una modalidad diferente a la tradicional. Se vendieron cientos de miles de entradas a una organización religiosa la cual las distribuyó a sus seguidores. Esto hace que el público en estas funciones sea mayor a lo esperado, lo que podría afectar al modelo. Como no se espera este tipo de películas en el futuro se decide eliminar estos registros.

Se encuentran inconsistencias en el género de las películas. Por ejemplo, títulos de la misma saga presentan géneros distintos. Dos películas distintas de la saga Star Wars presentaban formatos de Acción para una y Aventuras para la otra. Se decidió que el género de ambas películas debe coincidir en Aventuras. Por su parte se unifican las películas del género Thriller con las de género Suspenso. El sitio de Ultracine distingue a las películas de animación por un lado y las infantiles (sin animación) por otro. Ambos géneros se unifican en Infantiles. El género Animé, aunque su nombre hace referencia a la animación, se deja sin modificar ya que son títulos con características distintas a la del resto de las películas infantiles de animación. En 2019 la cadena Disney adquirió a la distribuidora Fox. En la información obtenida de Ultracine, las películas de Fox estrenadas después de la adquisición pasan a indicarse bajo el distribuidor de Disney. Se lograron identificar estas películas y se le modificó el distribuidor a Fox para diferenciarlas de las películas de Disney.

Por último, la variable Presupuesto, que fue generada buscando información de cada película en internet, cuenta con numerosos missing values. En la mayoría de los casos, estos están en películas desconocidas, de poca trascendencia y escaso público. Se decidió remplazar estos missing values por 0, ya que estas películas son de muy bajo presupuesto. La excepción está en algunas películas argentinas, que aun siendo populares no ha sido posible encontrar información sobre su presupuesto. Para estos casos, se buscó otras películas nacionales de similares características que si contaban con información presupuestaria. Así el presupuesto de El Robo del Siglo se completó con el de Relatos Salvajes (5 millones de dólares) y el de Mi Obra Maestra se completó con el de El Amor Menos Pensado (1.5 millones de dólares).

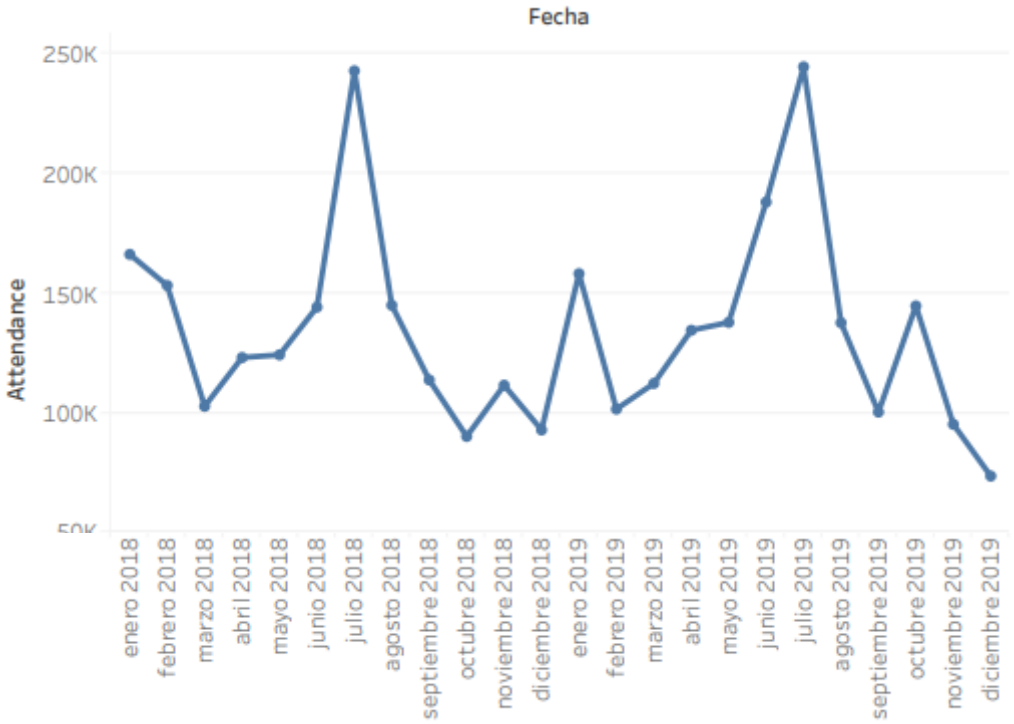
Habiendo procesado y limpiado la información, se separan las funciones de 2018 y 2019 de las funciones de enero y febrero de 2020. Así se crean dos bases de datos, una para entrenar el modelo, la otra para probar los resultados.

3.3 Análisis de la información

Una vez obtenida y procesada la base de datos que se utilizará para el modelo predictivo, se analiza la información mediante estadística descriptiva. Se toma el período 2018 a 2019 para el complejo seleccionado y se elaboran estadísticas que relacionan distintas variables con la cantidad de espectadores, variable que se querrá predecir. A través del análisis estadístico descriptivo se puede inferir como impactan distintas variables a la hora de explicar la asistencia de público.

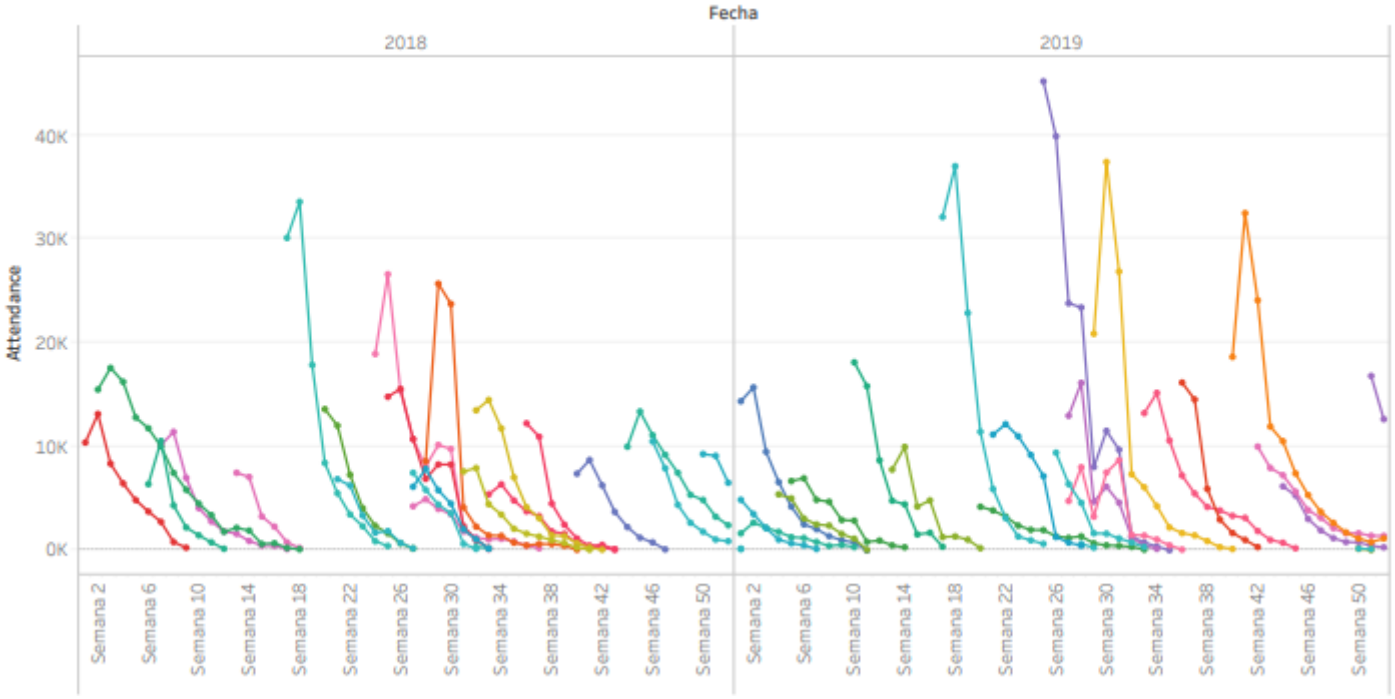
En el siguiente gráfico se observa la cantidad de público por mes para el complejo seleccionado. Es clara la estacionalidad del attendance, siendo los picos de público los meses de julio, explicado por las vacaciones de invierno. En este período las principales películas infantiles son estrenadas. Aun por debajo de julio, otros meses con una cantidad considerable de público suelen ser enero, mayo, junio y agosto. En mayo y junio suelen estrenarse películas de aventuras o super héroes, mientras que en agosto suelen estar en cartel las principales películas argentinas. Meses con muy poco publico suelen ser febrero y el período entre septiembre y diciembre. Octubre 2019 fue la excepción a esta tendencia histórica producto de la película Guasón, la cual fue un éxito en taquilla.

Gráfico 3.3.1 Attendance por mes 2018-2019



El gráfico a continuación muestra el attendance por semana de aquellas películas que han superado los 20.000 espectadores en el complejo de cine en cuestión. Cada una de estas películas es representada por una línea, y es posible identificar un patrón muy claro en el ciclo de vida de los títulos. La mayor cantidad de público se concentra en las primeras dos semanas, presentado una abrupta caída en la tercera y cuarta, para amesetarse en las semanas posteriores. El efecto es más evidente para las películas con mayor cantidad de público. En consecuencia, la cantidad de semanas que una película lleva en cartel es un factor que puede explicar el público de una función. Conociendo este fenómeno, es muy importante contar con una buena distribución de títulos por sala, para poder captar el pico de público en las primeras semanas.

Gráfico 3.3.2 Attendance por semana – principales películas (2018-2019)

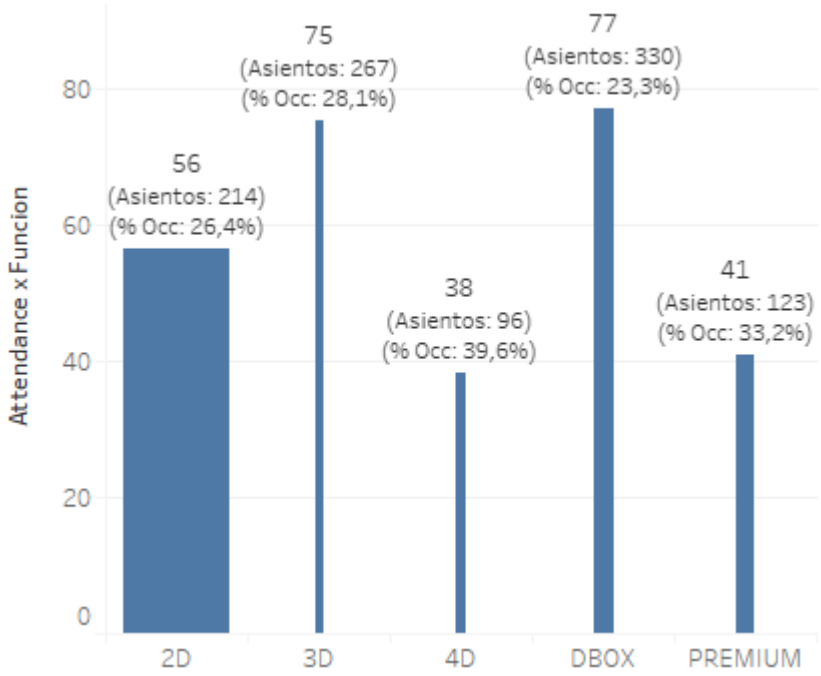


Además de la estacionalidad observada mes a mes, y la distribución de público entre semanas de estreno de cada película, es importante observar la distribución del attendance dentro de cada semana. El siguiente gráfico muestra el promedio de público por función distribuido por día de semana y franja horaria. Es evidente como la cantidad de espectadores es superior para los días de fin de semana y viernes por la noche. Dentro de cada día, la franja horaria noche es la que presenta la mayor cantidad de espectadores. La combinación de ambos factores hace que los sábados a la noche sea el período temporal con mayor cantidad de público. Es destacable también la performance de los miércoles por la noche, esto se explica por el descuento del 50% que se otorga en estos días.

Gráfico 3.3.3 Attendance promedio por función, por día de semana y franja horaria

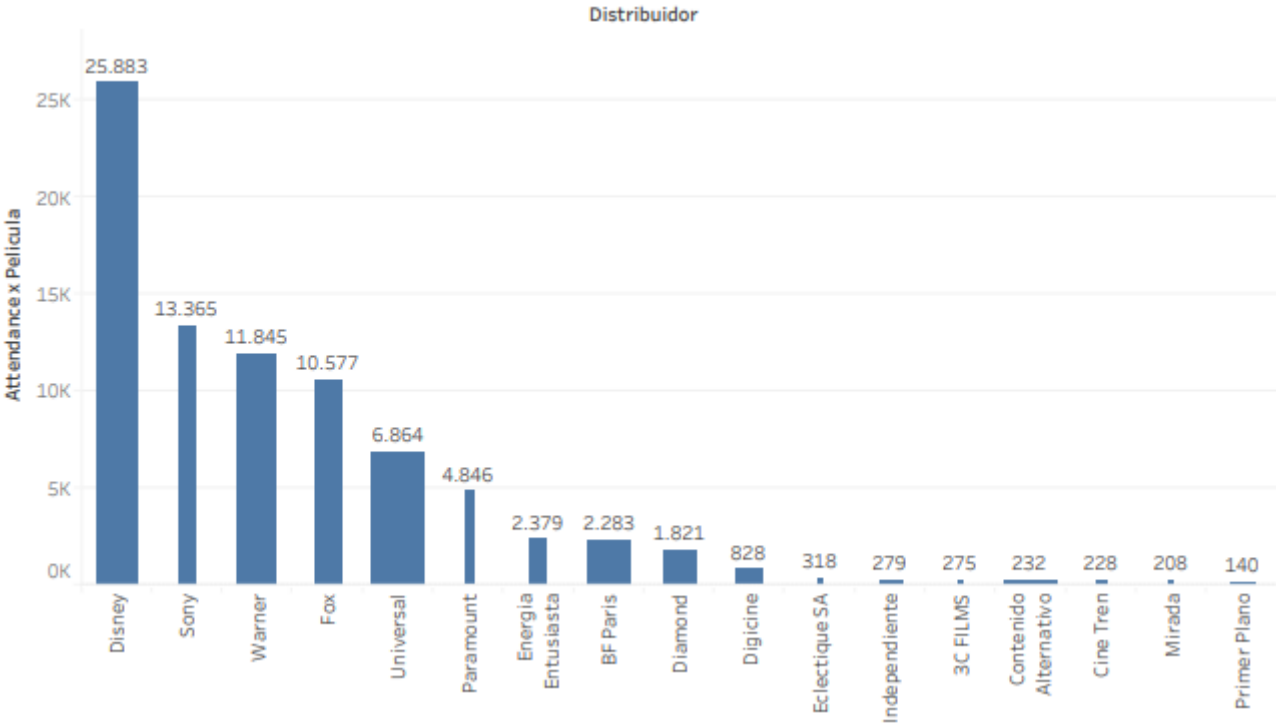
Al analizar el attendance promedio por función para cada formato se observan que los formatos 3D y DBOX presentan mayor cantidad de público. Sin embargo, hay que tener en cuenta la capacidad de la sala destinada a cada formato. Al observar la ocupación para cada formato se distingue que el 4D y el Premium tienen una mayor cantidad de público en relación con los asientos destinados a cada uno de estos.

Gráfico 3.3.4 Attendance promedio por función, asientos y ocupación por formato



Pasaremos a analizar las características de las películas que inciden en la cantidad de público. El siguiente gráfico muestra la cantidad de espectadores promedio por título. El ancho de cada columna refleja la cantidad de títulos estrenados en el período.

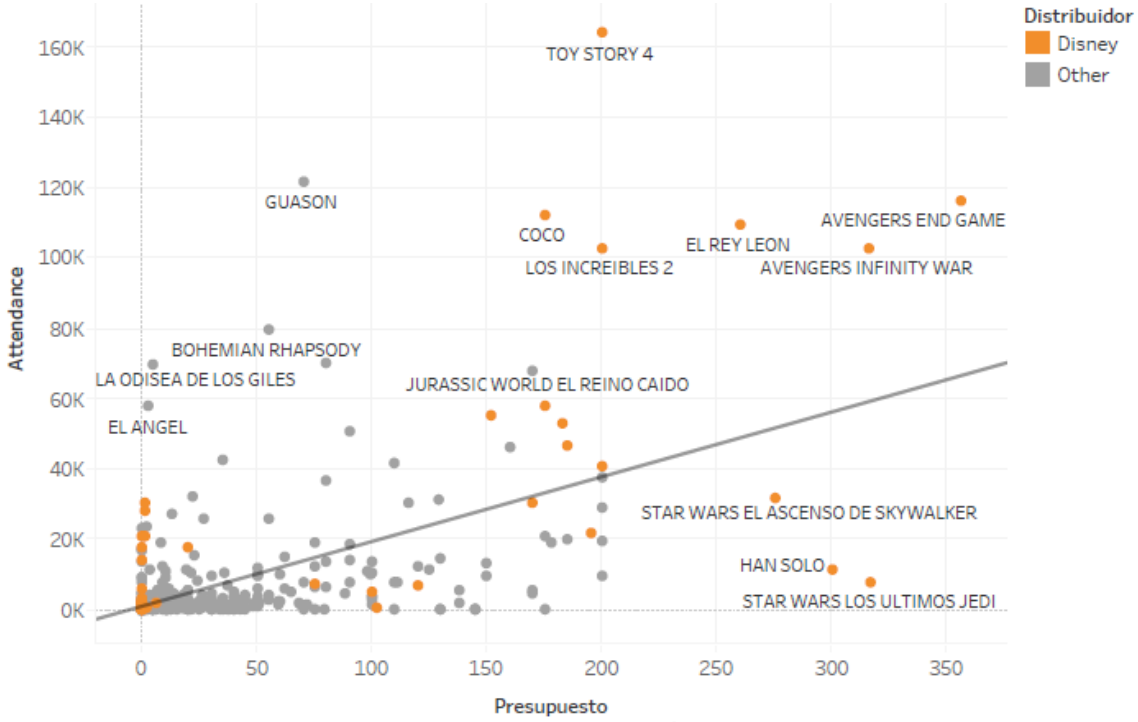
Gráfico 3.3.5 Attendance promedio por película para cada distribuidor



Es claro el dominio de Disney ya que la cantidad de espectadores por película duplica al de los siguientes competidores. Los siguientes distribuidores por cantidad de público por título también son grandes cadenas: Sony, Warner, Fox, Universal y Paramount. Otros distribuidores menos conocidos, aun estrenando gran cantidad de títulos tienen peores resultados en términos de público.

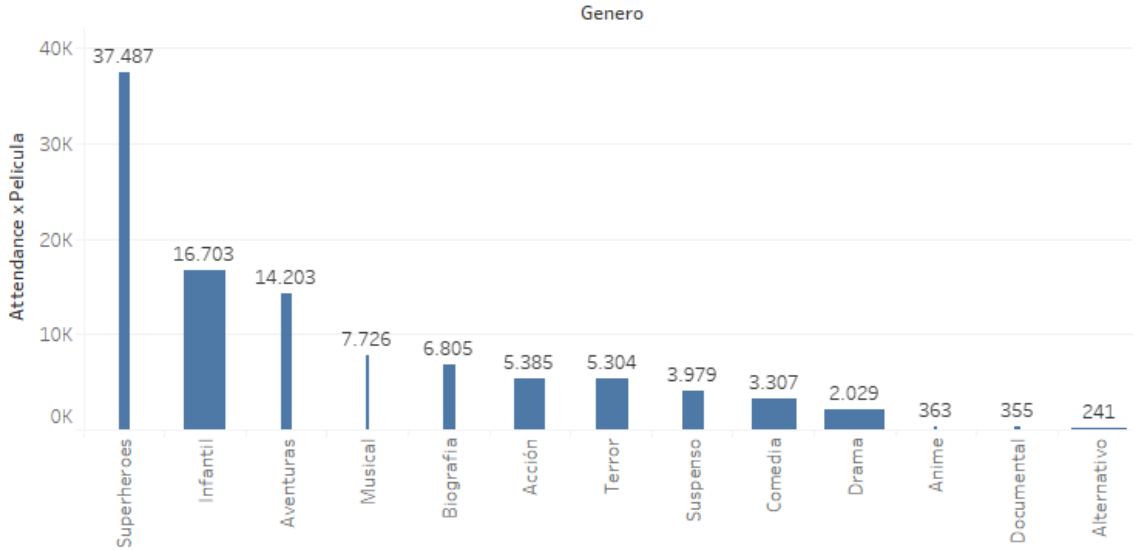
El gráfico a continuación muestra la cantidad de público en relación con el presupuesto de cada película medido en millones de dólares. Se observa cierta tendencia de que a mayor presupuesto mayor cantidad de público, aunque con claros outliers. También se diferencian con otro color los títulos de Disney, pudiéndose observar que la totalidad de las películas con un presupuesto mayor a los 200 millones de dólares corresponden a dicha cadena.

Gráfico 3.3.6 Películas por Attendance y Presupuesto



Son destacables las películas argentinas El Ángel y La Odisea de los Giles, que con un presupuesto muy limitado en comparación al resto de los éxitos extranjeros han logrado una cantidad de público muy alto. Bohemian Rhapsody y Guasón son las otras películas que con un presupuesto moderado han podido obtener una gran cantidad de espectadores. Por otro lado, se observa el claro dominio de Disney en lo que respecta a películas de muy alto presupuesto, las cuales son las dominantes en venta de entradas. La excepción se encuentra en las películas de la nueva saga de Star Wars, con un elevado presupuesto, siendo también de Disney, tuvieron un resultado muy pobre en términos de público.

Gráfico 3.3.7 Promedio de attendance por película por género



Al observar la cantidad de público promedio por película por género, el de superhéroes es el claro ganador, impulsado por los títulos de Avengers. Los géneros de Infantil y Aventuras son los que le siguen en cantidad de espectadores promedio por película. El grosor de cada columna muestra la cantidad de títulos estrenados por cada género. Se observan que géneros tradicionales como Drama o Comedia, aun con muchos estrenos presentan un bajo promedio de espectadores.

Tras haber analizado de forma descriptiva la información, es posible observar cómo se comporta la variable a predecir, el attendance en relación con las principales variables. Se reconoce una clara estacionalidad a través de los meses, y una distribución de público por día de semana. En relación con las películas se observa una preferencia por aquellas conocidas coloquialmente como “pochocleras”, películas de Disney, de genero superhéroes, infantil o aventuras y de elevado presupuesto. Si bien existen excepciones a esta regla, películas argentinas populares, o extranjeras de presupuesto moderado; es claro como la gran mayoría del público se concentra en pocos títulos.

4. Implementación

Utilizando la base de datos del apartado anterior, se proseguirá por implementar distintos modelos de aprendizaje automático con el fin de lograr predecir la cantidad de espectadores. Estos modelos serán analizados y comparados para identificar cual es el que mejor ajusta a los datos reales, pudiendo elaborar conclusiones acerca de los resultados. El mejor modelo será finalmente puesto en producción. Por último, se abordará la temática de metodologías ágiles para la implementación de modelos de aprendizaje automático, elaborando un hipotético plan de proyecto.

4.1 Entrenamiento y Prueba de modelos

Como se mostró en el apartado anterior, se cuenta con una base de datos de entrenamiento (2018 y 2019) y de prueba (Ene y Feb 2020) las cuales fueron limpiadas y analizadas. Se usará el software de Microsoft Azure[20] para entrenar distintos modelos. Se intentará predecir la variable attendance mediante el resto de las variables enunciadas en el apartado anterior. Luego, los modelos serán comparados utilizando la Raíz del Error Cuadrático Medio[21] como indicador. En el Anexo 7.2 y 7.3 se muestran los procedimientos en la herramienta Azure.

El primer modelo a entrenar será el más simple, la Regresión Lineal, el cual aproxima la relación entre una variable dependiente, las variables independientes y un término aleatorio. A través de este modelo, es posible identificar los coeficientes de la regresión, útiles para entender como impacta cada variable sobre el attendance. Si bien la regresión lineal supone independencia en cada observación, esto no siempre es así en la realidad. Por ejemplo, la existencia de un descuento del 50% en las funciones del miércoles no solo aumenta el público en dicho día, sino que disminuye el attendance en los días lunes y martes. También, si se programan varias funciones de una misma película en el mismo horario esto afecta a cada una de las funciones.

El segundo modelo que se entrenará es el Boosted Decision Tree Regression, herramienta de machine learning[22] de tipo ensamble. Construye cada árbol usando una

[20] Servicio de computación en la nube creado por Microsoft para construir, probar, desplegar y administrar aplicaciones y servicios mediante el uso de sus centros de datos.
[21] Medida de uso frecuente de las diferencias entre los valores predichos por un modelo y los valores observados. Se calcula cada una de las diferencias y se las eleva al cuadrado. Estas se promedian y se aplica la raíz cuadrada.
[22] Anglicismo que hace referencia a Aprendizaje Automático.

función de perdida predefinida la cual mide el error y lo corrige en el paso siguiente. Los parámetros de la regresión se optimizan para encontrar el mejor modelo.

Habiendo entrenado los modelos se obtiene la base de prueba con el valor de la predicción de público para cada función, este se denomina Scored Label. El mismo debe de ser corregido para que ninguna predicción muestre espectadores negativos o superiores a la capacidad de la sala. Esta predicción de attendance, en ambos modelos, se usará para comparar resultados y obtener conclusiones sobre las predicciones obtenidas.

4.2 Análisis de Resultados y Puesta en Producción

4.2.1 Análisis de coeficientes de la regresión lineal

En el anexo 7.4 es posible observar los coeficientes de cada variable para la regresión lineal. Estos permiten obtener conclusiones que van de la mano a lo inferido cuando se elaboró un análisis estadístico descriptivo en el apartado anterior. Las variables que más inciden en el público son los sábados domingos y feriados y la franja horaria noche. Los meses de enero y julio son los que tienen un impacto positivo en el attendance. Que la película sea clasificada como boom y que sea nominada al Oscar también lo tiene. Por su parte, la cantidad de asientos en la sala donde se proyecta la película naturalmente tiene un efecto positivo en la cantidad de espectadores.

Por otro lado, las variables que inciden negativamente en el attendance son la franja horaria primera matiné, los lunes y martes y que el día no sea feriado ni vacaciones. Hay distribuidores que tienen un coeficiente negativo, como Cine Tren, Primer Plano y Digicine. Los meses que peor inciden en la cantidad de público son diciembre y marzo.

4.2.2 Comparación de modelos

A continuación, se compararán los modelos entrenados para identificar cual logra una mejor predicción. Los indicadores a utilizar son la Raíz del Error Cuadrático Medio, su relación con el promedio de espectadores por función, y el Error Cuadrático Relativo. Este último compara los desvíos al cuadrado entre la estimación y la variable real, contra los que hubieran sido obtenidos usando el promedio de la serie como la estimación. A su vez, si bien el modelo predice el público por función, se evalúa la predicción para distintos niveles de granularidad: por día, por semana y por mes. En el anexo 7.5 y 7.6 se encuentra en detalle la presentación de resultados.

Regresión Lineal

	RSME	ATTE AVG	RSME / ATTE AVG	RSE
x Funcion	45	66	68,1%	0,541
x Dia	1.113	5.280	21,1%	0,362
x Semana	3.924	33.482	11,7%	0,095
x Mes	11.221	155.774	7,2%	0,348

Boosted Decision Tree Regression

	RSME	ATTE AVG	RSME / ATTE AVG	RSE
x Funcion	38	66	58,3%	0,396
x Dia	942	5.280	17,8%	0,259
x Semana	3.185	33.482	9,5%	0,062
x Mes	1.738	155.774	1,1%	0,008

Se observa claramente cómo el modelo de Boosted Decission Tree Regression presenta mejores resultados en todos los indicadores. En consecuencia, es el modelo con mayor capacidad predictiva. Analizando los resultados por función, la raíz del error cuadrático medio es alta, representado un 58% sobre el promedio de espectadores por función. A su vez, los errores representan casi un 40% de los que se obtendrían usando el attendance promedio por sala como estimación.

Cada vez que se avanza hacia un nivel de agregación mayor, los resultados mejoran ya que los desvíos positivos y negativos se compensan. En el Anexo 7.6 se pueden observar la relación entre la predicción de attendance por día contra el attendance real. En la mayoría de los días esta es más del 85%, pero hay excepciones que perjudican los resultados en promedio. Se puede afirmar que la predicción captura bien la distribución de público que existe entre cada día. Sin embargo, en la realidad pueden existir días que rompen con la distribución tradicional, y esto no lo captura el modelo.

Cuando se observan los resultados por semana se obtienen buenos resultados: la raíz del error cuadrático medio es inferior al 10% del promedio de espectadores semanal, el error cuadrático relativo es del 6%; y viendo los resultados semana a semana, todos superan el 87% de exactitud contra el attendance real. Aunque con solo dos observaciones, los resultados son aun mejores por mes. El modelo predijo 177.243 espectadores en enero y 136.845 en febrero, cuando los números reales fueron de 174.786 y 136.762 respectivamente.

En el anexo 7.6 también se presentan la comparación entre el attendance estimado y real por película por cada mes. En ambos meses se observa que la predicción es buena para las películas más vistas, pero empeora para las menos populares. Como el público suele concentrarse en pocos títulos, una buena estimación para los principales asegura que el modelo sea bueno.

En enero, la película más vista, Frozen 2, presentó una cantidad de público inferior a la estimación. Esto se compensó con El Robo del siglo, cuyo attendance real superó el estimado. La estimación fue acertada para Jumanji, la tercera película más vista. El total de público entre estas tres películas representó un 80% del total de espectadores del mes.

Con lo que respecta a febrero, la estimación es acertada para el top 3 de películas, pero arroja muy malos resultados para Parasite, la cuarta más vista. La estimación fue la mitad del público que efectivamente fue a ver a este film. Esto se debe a que la película no es convencional, de origen surcoreano, rompió con los estándares tradicionales de la industria. Si bien el modelo tenía en cuenta la nominación al Oscar de dicho film, su región de origen, distribuidor y genero hicieron que la estimación sea baja.

4.2.3 Puesta en producción del modelo

Una vez que se han entrenado los modelos y se analizó cual otorga una mejor predicción, se procede con la puesta en producción utilizando la misma herramienta de Azure mediante la opción de crear un experimento predictivo como servicio Web. La base de prueba a cargar debe de tener excluido el label[23], en este caso el attendance, ya que es la variable a predecir. El servicio web permite cargar los datos de cada variable para una función específica y obtener la estimación de público. En el anexo 7.7 se muestra el detalle del experimento predictivo.

4.3 Metodologías ágiles para la implementación de proyectos

4.3.1 Planteo de la problemática

Habiendo analizado la base de datos y elaborado un modelo de aprendizaje automático que logre predecir el attendance a partir de las variables propuestas, se armará una estructura para la implementación de proyectos de aprendizaje automático usando metodologías ágiles. Si bien este trabajo se focaliza en una problemática concreta, la de predicción de espectadores; se tratarán distintos desafíos, que requerirían de otros modelos predictivos, para abordar el tema de implementación de proyectos desde una perspectiva global a toda la organización.

Para la compañía es de suma importancia contar con una predicción de resultados lo más acertada posible. La piedra angular para lograrlo es poder predecir la cantidad de espectadores que visitaran los complejos. Con el modelo de aprendizaje automático

[23] Anglicismo que hace referencia a Etiqueta. En este caso, la variable a predecir.

propuesto se puede estimar la cantidad de público que visitará el cine diariamente partiendo de la programación de funciones. A su vez, este modelo es útil a la hora de tomar la decisión de como programar los distintos títulos que se cuentan en distintas salas y franjas horarias. En tercer lugar, contar con una sólida predicción de espectadores es muy útil para otras áreas de la compañía, desde decisiones como campañas de marketing o contrataciones de empleados temporales dependen del attendance esperado. Dada la importancia de la información que el modelo otorga, y la interacción con las distintas áreas de la organización se propone un proyecto de aprendizaje automático. Un proyecto de tales características debe de ser integrador para afrontar los distintos desafíos que se presentan.

El modelo descripto en el trabajo debe de verse como un primer paso a la hora de abordar superar las problemáticas descriptas. Otras estimaciones que serían importantes para el proyecto son las ventas en la concesión de Alimentos y Bebidas, el impacto en la cantidad de público de un cambio de precios del propio complejo o de un competidor, y poder predecir que porción de socios del programa de fidelización se darían de baja. Si bien en este trabajo solo se plantea la estimación de público, a la hora de elaborar el proyecto se deben tener en cuenta los otros aspectos ya que estos se interrelacionan. De los títulos en cartel y la cantidad de público también dependen las ventas en Candy y el éxito del programa de fidelización. En épocas donde los títulos no son atractivos, se busca también incentivar al cliente con mayores descuentos.

4.3.2 Motivación para la implementación de Metodologías Ágiles.

En este trabajo se presenta a la fuente de datos como algo estático y las predicciones que se muestran es de un período temporal determinado. Pero en la práctica gran cantidad de datos se generan diariamente y es necesario actualizar las predicciones cada vez que la programación cambia y se tiene más información para entrenar el modelo. En consecuencia, se propone una metodología ágil que permita ser flexible para cumplir con esta necesidad de una constante actualización a partir de la nueva generación de datos. Una metodología tradicional no sería efectiva en este caso, ya que los insumos en que se cuentan, es decir los datos, cambian con frecuencia. En este proyecto es necesaria una iteración constante a la hora de la implementación, introduciendo pequeños cambios permanentemente.

Los dos métodos ágiles más populares son el Scrum y en Kanban. El Scrum es una gama de prácticas donde existen tres roles: el dueño del producto, el Scrum Master y el Equipo de desarrollo. Cada equipo encara sus tareas mediante cuatro aspectos: un "backlog[24]" del producto, un "sprint[25]" de tareas, un "incremento" (porcentaje del sprint que ha podido ser resuelto) y un "terminado" (entendimiento compartido de lo completado en el proyecto). El Backlog es una lista de Historias de Usuario, ordenadas según el valor de negocio que establece el Dueño del Producto, y que reúnen todos los requisitos y funcionalidades necesarias en el producto. En este caso, los usuarios del producto son las distintas áreas que requieren las soluciones de aprendizaje automático para realizar predicciones, por lo que las Historias de Usuario deberán capturar las necesidades de estos clientes internos. El equipo de Scrum divide su tiempo en cinco etapas: el refinamiento del Backlog del producto, el planeamiento del sprint, una reunión diaria del Scrum, revisión del sprint y análisis de este. (Sutherland & Schwaber, 2013)

En contraste con Scrum, el método Kanban es menos detallado en términos de prácticas requeridas y principios, siendo menos rígido. Las prácticas más usuales son la visualización del flujo de trabajo, limitar las tareas que están en proceso, y permitir loops para mejoras continuas en forma colaborativa. (Alqudah & Rozilawati 2017) Se sugiere tomar aspectos de ambos a la hora de diseñar el método de trabajo. Del Scrum se toma la segmentación del equipo de trabajo en roles específicos, separación de tareas y organización del tiempo. Por otro lado, es necesario adaptarse a los cambios en objetivos y permitir continuas iteraciones y cambios durante el proceso de desarrollo, características del método Kamban. (Kniberg & Skarin, 2010)

4.3.3 Plan del proyecto de implementación

A partir de las necesidades mencionadas anteriormente se plantea un proyecto de implementación. Se establecen los roles y se distinguen las etapas, así como las áreas vinculadas al proyecto. El rol de Dueño del producto, es decir, quien toma las decisiones finales se le asigna a la gerencia general de la compañía. La cantidad de tres integrantes esta explicada por el tamaño mediano de la empresa, y los límites de las funciones de cada rol en ciencia de datos deben de ser difusos para adaptarse a las necesidades de la organización.

[24] Anglicismo. Es un listado de todas las tareas por hacer durante el desarrollo de un proyecto.
[25] Anglicismo. son ciclos de ejecución cortos, con el objetivo de incrementar el valor en el producto

Se plantea un equipo del proyecto compuesto por tres miembros:

- Supervisor del proyecto: Es el nexo entre el equipo de trabajo y las distintas áreas de la compañía. Trasmite al equipo las necesidades y objetivos, comunica a las áreas los avances y limitaciones del proyecto. Dentro del equipo, toma el rol de Arquitecto de Datos y Revisa la labor del resto de los integrantes y actúa como Scrum Master. Las habilidades de este puesto son el liderazgo de equipos, estadística, finanzas y método científico.
- Analista Funcional: Se encarga de que las bases de datos se encuentren actualizadas y no presenten errores. Limpia los datos de ser necesario para que puedan ser utilizados. Es el nexo del equipo de trabajo con el área de IT en caso de que un proceso falle. Forma parte del equipo de desarrollo. Toma las tareas de un ingeniero de datos, se encarga de elaborar nuevas bases de datos y limpiarlas en caso de que sea necesario. Ayuda al Analista BI en lo que necesite para armar los modelos y se encarga del proceso técnico de ponerlos en producción. Las habilidades de este puesto son programación y arquitectura de datos.
- Analista BI: Utiliza las bases de datos armadas por el Analista Funcional. Usa los datos madre para crear nuevas variables que fueran necesarias para el armado de los modelos. Recibe las necesidades de la compañía desde el Supervisor del proyecto y elabora modelos que permitan cumplir con los objetivos propuestos. Forma parte del equipo de desarrollo y toma las tareas de un científico de datos. Analiza las variables de las bases de datos, al tener conocimiento del negocio, encuentra patrones, decide cuales son pertinentes y formula los modelos. También toma funciones de un Citizen Data Analyst[26], arma tableros de control para presentar los avances del proyecto a las distintas áreas. Las habilidades de este puesto son estadística, finanzas, método científico y visualización de datos.

Etapas del Proyecto:

El proyecto planteado se divide en etapas que distinguen los procedimientos de cada modelo descripto. Dentro de cada etapa se divide el tiempo en semanas, comenzando cada semana los jueves, día en el cual se estrenan las películas. Los jueves se establece el backlog de las tareas pendientes y se planifica el primer sprint. Cada sprint tiene una duración de una semana, y cada jueves del mes se comienza con una reunión donde se

[26] Anglicismo. El Citizen Data Analyst es una persona que crea o genera modelos que aprovechan el análisis predictivo o prescriptivo, tienen un perfil financiero y no técnico.

analizan los resultados del sprint anterior y se planifica el sprint siguiente. En el anexo 7.8 se agrega un cuadro que resume las distintas etapas del proyecto de forma simplificada. La disposición temporal de las mismas supone el éxito de cada etapa para el paso a la siguiente.

- Etapa 1: Corresponde a la planificación. El equipo se reúne con las distintas áreas de la compañía las cuales establecen sus necesidades. Se analiza la plausibilidad de cada pedido y se traza una línea de trabajo, armando el backlog para los sprints sucesivos.
- Etapa 2: Se comienza son la elaboración del modelo que busca predecir la cantidad de público. El primer paso consiste en la recolección de datos de las distintas fuentes internas y externas de la compañía, limpiar la base de datos y generar nuevas variables. Se deciden que variables se tendrán en cuenta y se arman las bases de datos, una base de entrenamiento y otra de prueba por cada complejo. En la siguiente semana se prueban los modelos, encontrando uno que logra predecir mejor los datos de la base de prueba, usando la raíz del error cuadrático medio para comparar resultados. El mejor modelo se corre para predecir el público de las semanas siguientes usando de input la programación futura.
- Etapa 3: Comenzando la nueva semana se comparan los resultados de la predicción con el attendance real de cada cine. A su vez, se usa la proyección de espectadores como input para generar nuevos modelos, uno que intente predecir el consumo de los productos en la concesión de alimentos y bebidas, y otro que analice la elasticidad precio del público. Al igual que en primer modelo, se arman las bases de datos, se deciden que variables se van a utilizar y se prueban distintos modelos. Una vez obtenido los mejores se elaboran proyecciones. Tras la primera semana, se presentan los resultados armando un dashboard[27] que compare el attendance estimado con el real. Se comparte la predicción de attendance a las distintas áreas: programación, marketing, operaciones y concesiones para ayudar a la toma de decisiones. Se analizan las variaciones obtenidas y se buscan formas de mejorar el primer modelo de predicción de público. Por ejemplo, el modelo que se elaboró en este trabajo predice el público de enero y febrero, al ser principio de año no es significativo el impacto de películas ya estrenadas. Cuando el período temporal cambie, y las películas a predecir ya hayan sido estrenadas, será necesario tener en cuenta la performance de estas para mejorar el modelo. Luego se elabora una nueva proyección.

[27] Anglicismo. Traducido como tablero o cuadro de mandos, es un documento en el que se reflejan, mediante una representación gráfica, las principales métricas o KPI.

- Etapa 4: Iniciada una nueva semana se comparan las predicciones elaboradas con el modelo optimizado con el attendance real de cada cine. Si la medida de éxito es satisfactoria, (RSME) el modelo pasa a la etapa de producción. También se comparan las predicciones de los nuevos modelos armados con los datos reales. Las estimaciones obtenidas se comparten con las distintas áreas de la compañía. Usando las predicciones obtenidas, se comienza a armar una base de datos para el último modelo propuesto, el que busca estimar un churn rate de los clientes del programa de fidelización. Se deciden que variables de consumo de los clientes se va a utilizar y se elabora un modelo. En este caso se usa el accuracy[28] como medida de éxito ya que la variable a predecir es binaria: si un cliente se da de baja o no.
- Etapa 5: Se continúa mandando a las áreas de la compañía las distintas proyecciones. Se pasan a producción los modelos de predicción de ventas del Candy y sensibilidad precio. Se comienza a probar el modelo de predicción de churn rate. A diferencia del resto, al ser la suscripción mensual, solo será posible probar los resultados del modelo de forma mensual y no semanal como los anteriores. Una vez alcanzado un accuracy aceptable se comparte la predicción de churn rate con el área de marketing.

A continuación, se elaborará un ejemplo de cómo los distintos modelos propuestos pueden ser utilizados para resolver desafíos. Se observa que en un complejo en particular el attendance de los lunes y martes es significativamente inferior al estimado por el modelo de predicción de público. La razón es que un cine de la competencia otorgó un 50% de descuento para las entradas en esos días. Se utiliza el modelo de estimación de elasticidad precio de del público para ajustar las proyecciones a los nuevos precios de la competencia. Se usan las nuevas predicciones de público para ajustar las estimaciones de ventas de alimentos y bebidas, y finalmente se estima el impacto en el EBITDA.

El equipo de marketing recibe estos resultados y sugiere otorgar un descuento del 50% los lunes y martes al igual que la competencia. Usando el modelo de impacto de precios sobre público se calcula el impacto y se reestima la afluencia y las ventas del Candy con el nuevo escenario. Finalmente se usa toda esta información para analizar si el impacto negativo en el EBITDA es menor o mayor aplicando esta iniciativa que dejando los precios sin cambios. Sea cual sea la mejor opción se comparten nuevamente las estimaciones de público y ventas a las áreas para su propia gestión.

[28] Anglicismo. Es el grado de cercanía de las mediciones de una cantidad al valor real de esa cantidad.

5. Conclusiones

En este trabajo se abordó en profundidad la temática del uso de datos en una organización del rubro de exhibición cinematográfica. Se describió a dicha organización enunciando sus desafíos con relación a la gestión de datos. Se analizó la información proveniente de distintas fuentes con el objetivo de realizar un modelo de aprendizaje automático que logre predecir la cantidad de público de uno de los complejos de la organización. Es posible afirmar que se cumplen los objetivos planteados en la introducción. Se logró analizar a la organización sugiriendo mejoras concretas para el uso de sus datos, se logró determinar como impacta cada variable en la cantidad de público, y se elaboró un modelo predictivo con buenos resultados.

Tomando indicadores como la raíz del error cuadrático medio y el error medio relativo, se llegó a la conclusión de que el modelo de Boosted Decision Tree Regression arroja mejores resultados que la Regresión Lineal. Sin embargo, esta última es relevante para poder analizar como inciden las distintas variables en el attendance, observando los coeficientes de la regresión. Si bien ningún modelo logra predecir con exactitud la cantidad de público por función, la predicción agregada por día es aceptable, y por semana y mes muy buena. El experimento de aprendizaje automático se convierte así en una herramienta útil para la organización a la hora de tomar decisiones. Por ejemplo, se puede utilizar para comparar distintas programaciones de funciones con distintas películas para elegir la distribución de títulos que logre maximizar el público esperado.

Por su parte, partiendo de la programación futura, es posible predecir la cantidad de público, dato muy importante a la hora de estimar ingresos futuros. En el último apartado se dejó planteada una estructura para un proyecto de implementación de modelos predictivos que tomen al modelo desarrollado en este trabajo como punto de partida. La organización cuenta con una gran cantidad de datos que pueden ser utilizados para predecir distintas variables pertinentes a la hora de tomar decisiones estratégicas.

Una crítica pertinente que se le puede hacer al modelo aquí desarrollado es que requiere la programación futura funciones a exhibir. Esta sólo está disponible para un determinado horizonte temporal, no mayor al de un mes. Aunque se conocen las fechas de estreno de las películas, la distribución de títulos por función no se hace con tanta antelación. En consecuencia, este modelo podrá ser utilizado para predecir el público de un período temporal acotado.

Si se quisiera lograr una predicción de público para un período temporal más largo, será necesario poder predecir la distribución de títulos por sala utilizando un nuevo modelo. Conociendo las fechas de estreno de las películas, se podría predecir cuantas semanas durarán en cartel, y cuantas funciones recibirán por semana, usando como datos las películas ya estrenadas. Otra opción sería elaborar un modelo más simple, que tome como datos las fechas de estreno de las películas, pero no su programación; con el objetivo de predecir los espectadores diarios y no por función. Una vez que se cuente con una estimación de público acertada para un período de un año, esta podrá ser usada a la hora de elaborar el presupuesto anual de la compañía, actualizando la predicción cada vez que se elabore un nuevo forecast.

A su vez, al contar con la información del público futuro de cada complejo, será posible estimar otras variables como venta en concesiones de alimentos y bebidas y altas y bajas en el programa de fidelización. Estas herramientas podrán ser utilizadas para tomar decisiones estratégicas con el fin de aumentar los beneficios y otorgar una mejor experiencia al cliente.

A modo de conclusión, se afirma que la elaboración de este trabajo fue relevante como punto de partida para el uso predictivo de los datos por parte de la organización. Analizando al detalle las distintas fuentes de datos que se poseen, y el uso que actualmente se le dan, se concluye que la empresa está a medio camino en lo que refiere a la transformación hacia una organización basada en datos. Se elaboró un primer modelo que arroja resultados aceptables que pueden ser utilizados en la práctica para tomar decisiones, conociendo por adelantado la afluencia de espectadores. A su vez, se deja planteado distintas sugerencias y pasos a seguir para continuar mejorando las herramientas de predicción del público y otras variables relevantes para el negocio.

6. Referencias Bibliográficas

- Marshall, Pablo & Dockendorff, Monika & Ibáñez, Soledad. (2013). A forecasting system for movie attendance. Journal of Business Research. 66. 1800-1806. 10.1016/j.jbusres.2013.01.013.
- Eliashberg, J., Elberse, A., & Leenders, M. A. A. M. (2006). The motion picture industry: Critical issues in practice, current research and new research directions. Marketing Science, 25(6), 638–661.
- Lim, J. (2012). Forecasting movie attendance of individual movie showings: a hierarchical Bayes approach.
- Gevaria, K., Wagh, R., & D'mello, L.R. (2015). Movie Attendance Prediction. International Journal of Computer Applications, 130, 14-17.
- Baranowski, Paweł & Korczak, Karol & Zając, Jarosław. (2020). Forecasting Cinema Attendance at the Movie Show Level: Evidence from Poland. Business Systems Research Journal. 11. 2020. 10.2478/bsrj-2020-0006.
- Hand, Chris & Judge, Guy. (2012). Searching for the picture: forecasting UK cinema admissions using Google Trends data. Applied Economics Letters. 19. 1051-1055. 10.1080/13504851.2011.613744.
- Şahin, M.; Erol, R. A (2017) Comparative Study of Neural Networks and ANFIS for Forecasting Attendance Rate of Soccer Games. Math. Comput. Appl, 22, 43.
- King, B.E., Rice, J., & Vaughan, J. (2018). Using Machine Learning to Predict National Hockey League Average Home Game Attendance. The Journal of Prediction Markets, 12, 85-98.
- T. G. Rhee and F. Zulkernine, (2016) "Predicting Movie Box Office Profitability: A Neural Network Approach," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, pp. 665-670, doi: 10.1109/ICMLA.2016.0117.
- Zhou, Y., Zhang, L. & Yi, Z. Predicting movie box-office revenues using deep neural networks. Neural Comput & Applic 31, 1855–1865 (2019).
- Silver, Jon & Mcdonnell, John. (2007). Are movie theaters doomed? Do exhibitors see the big picture as theaters lose their competitive advantage?. Business Horizons. 50. 491-501. 10.1016/j.bushor.2007.07.004.
- Gonzales Leandro (2015) Pp. 76-88, Exhibición y consumo de cine en la Argentina (1980-2013) La reconfiguración del espectáculo cinematográfico en

cifras, Estudios de Comunicación y Política Número 36/mayo-octubre 2015, ISSN 2007-5758

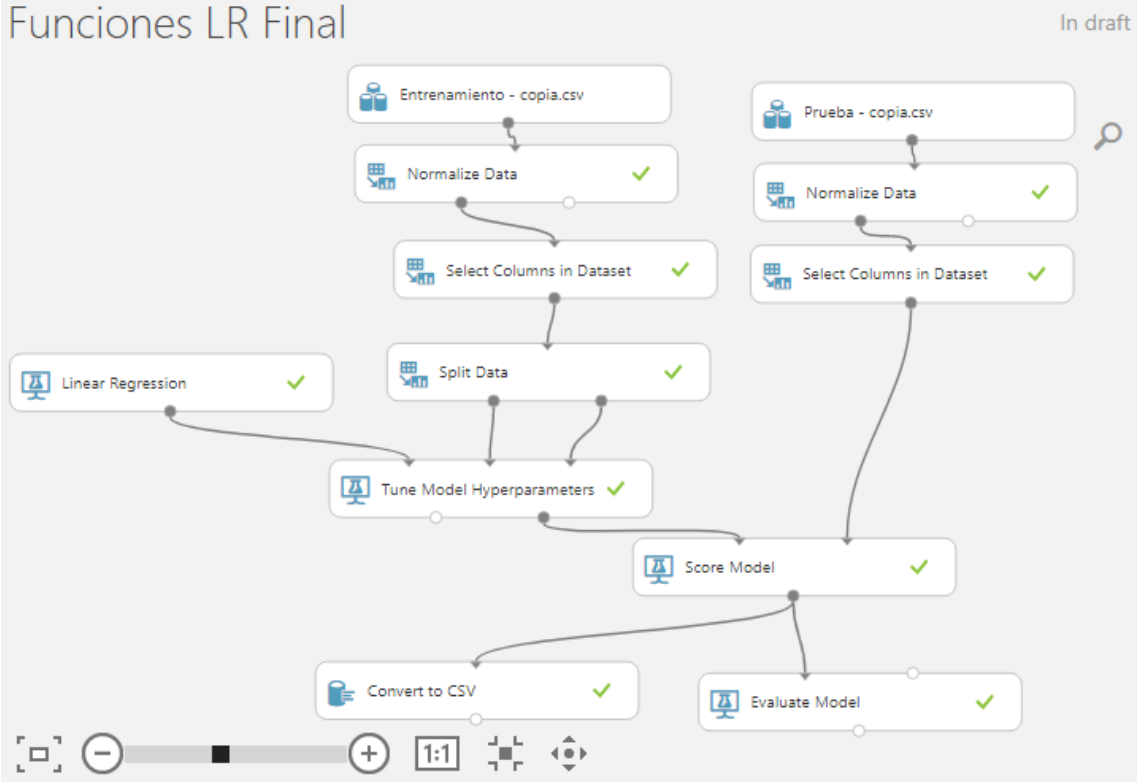
- Provost, Foster & Fawcett, Tom. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. Big Data. 1. 10.1089/big.2013.1508.
- McAfee, Andrew & Brynjolfsson, Erik (2012) Big Data: The Management Revolution, Harvard Business Review
- Lavalle, Steve & Lesser, Eric & Shockley, Rebecca & Hopkins, Michael & Kruschwitz, Nina. (2011). Big Data, Analytics and the Path From Insights to Value. MIT Sloan Management Review. 52. 21-32.
- Kiviat, Barbara (2019) The art of deciding with data: evidence from how employers translate credit reports into hiring decisions, Socio-Economic Review, Volume 17, Issue 2, Pages 283–309
- De Lusignan, Simon & Liaw, Siaw-Teng & Krause, Paul & Curcin, Vasa & Vicente, M & Michalakidis, Georgios & Agreus, Lars & Leysen, Peter & Shaw, Nicola & Mendis, K. (2011). Key Concepts to Assess the Readiness of Data for International Research: Data Quality, Lineage and Provenance, Extraction and Processing Errors, Traceability, and Curation. Contribution of the IMIA Primary Health Care Informatics Working Group. Yearbook of medical informatics. 6. 112-20. 10.1055/s-0038-1638748.
- Brynjolfsson, Erik & Hitt, Lorin & Kim, Heekyung. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?. SSRN Electronic Journal. 1. 10.2139/ssrn.1819486
- Khabbazi, M. R., Yusof Ismail, M., Ismail, N., & Mousavi, S. (2010). Modeling of traceability information system for material flow control data. Australian Journal of Basic and Applied Sciences, 4(2), 208–216.
- Fiallos Ordoñez, Angel. (2019). Técnicas para la gestión efectiva de Proyectos de Data Science: Un Enfoque Multidisciplinario. Congreso Internacional de dirección de proyectos.
- Kniberg Henrik & Skarin Mattias, (2010). Kanban and Scrum: making the most of both. InfoQ.
- Alqudah, Mashal & Razali, Rozilawati. (2017). A comparison of scrum and Kanban for identifying their selection factors.
- J. Sutherland and K. Schwaber, (2013) "The SCRUM guide. The definitive guide to SCRUM: The rules of the game," SCRUM. orgOctober

7. Anexos

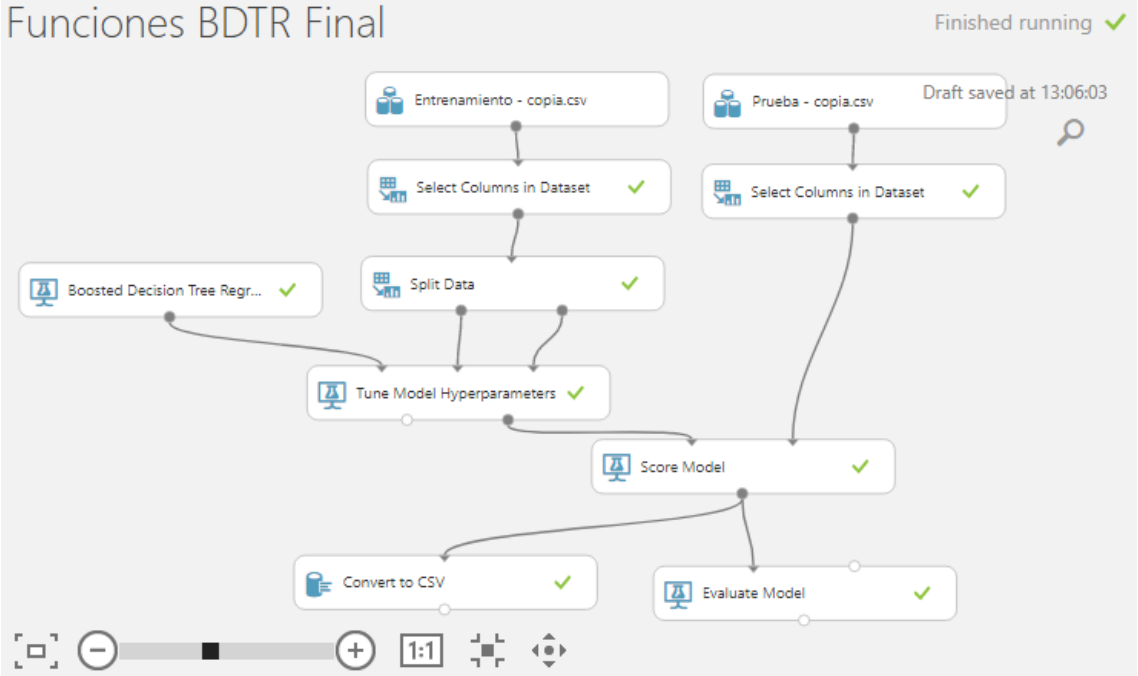
7.1 Diccionario de Datos

Variable	Tipo	Descripción
Session_ID	INTEGER (ID)	ID de cada función.
Fecha	DATE	Fecha de la función.
Dia	INTEGER	Dia de la función.
Dia Especial	STRING	Clasificación del día de la función, entre Nada, Feriado, Vacaciones y Fiestas.
Mes	STRING	Mes de la función.
Weekday	STRING	Día de la semana de la función.
Franja	STRING	Franja Horaria de la función, clasificadas en 1ra Matiné (11 AM - 2 PM), 2da Matiné (2 PM - 5 PM), Vermouth (5 PM - 8 PM), Noche (8 PM - 11 PM), Trasnoche (11 PM - Cierre).
Pelicula	STRING	Nombre de la película a exhibir
Secuela	BOOLEAN	Indica si la película es una secuela de otra anterior.
Remake	BOOLEAN	Indica si la película es un remake de otra anterior.
Fecha_Estreno	DATE	Fecha de estreno de la película.
Semanas en Cartel	INTEGER	Cantidad de semanas que transcurrieron entre la fecha de estreno de la película y la función.
Funciones por dia	INTEGER	Cuenta la cantidad de funciones de una misma película en el día.
Distribuidor	STRING	Nombre del distribuidor de le película.
Genero	STRING	Género de la película.
Region	STRING	Región de origen de la película.
Presupuesto	INTEGER	Presupuesto de la película en millones de USD.
Boom	BOOLEAN	Indica si se espera que la película sea un éxito en taquilla.
Oscar	BOOLEAN	Indica si la película fue nominada a un Premio Oscar a la mejor película.
Idioma	STRING	Idioma de la función (Castellano o Subtitulado).
Formato	STRING	Formato de la función (2D, 3D, 4D, DBOX o PREMIUM).
Sala	INTEGER	Número de Sala.
Asientos	INTEGER	Cantidad de asientos de la sala.
Attendance	INTEGER (Label)	Cantidad de público asistente a la función.

7.2 Regresión Lineal en Microsoft Azure.



7.3 Boosted Decision Tree Regression en Microsoft Azure.



7.4 Coeficientes de la Regresión Lineal

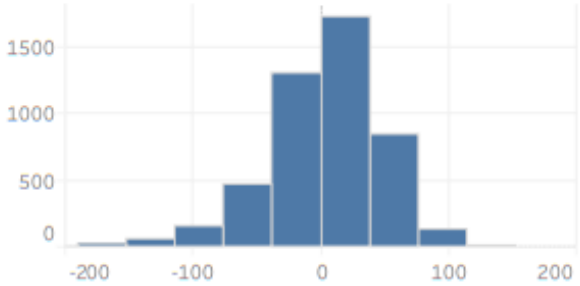
Batch Linear Regressor				
Settings				
Setting	Value			
Bias	True			
Regularization	0.001			
Allow Unknown Levels	True			
Random Number Seed				
Feature Weights				
Feature	Weight		Feature	Weight
Weekday_Saturday_2	0.538871		Genero_Alternativo_1	0.0723872
Franja_1ra Matine_0	-0.534697		Secuela	0.0669979
Weekday_Sunday_3	0.521611		Distribuidor_3C FILMS_0	-0.0667875
Franja_Noche_2	0.513458		Mes_Febrero_4	0.0656966
Dia Especial_Nada_2	-0.499602		Dia Especial_Fiestas_1	0.0630463
Dia Especial_Feriado_0	0.414167		Franja_2da Matine_1	-0.0584446
Weekday_Tuesday_5	-0.379467		Weekday_Friday_0	-0.0556739
Weekday_Monday_1	-0.351743		Formato_4D_2	-0.0554929
Distribuidor_Cine Tren_2	-0.289964		Formato_PREMIUM_4	-0.0553346
Genero_Musical_9	0.287244		Mes_Junio_6	0.055126
Mes_Enero_3	0.243708		Formato_DBOX_3	0.054475
Mes_Marzo_7	-0.205978		Oscar	0.0541238
Asientos	0.190173		Region_Argentina_0	0.0508047
Mes_Julio_5	0.182169		Genero_Comedia_5	0.0506466
Weekday_Thursday_4	-0.177905		Semanas en Cartel	-0.0480034
Mes_Diciembre_2	-0.169666		Region_Europa+Australia_2	0.0469369
Boom	0.167043		Mes_Octubre_10	-0.041978
Distribuidor_Contenido Alternativo_3	0.163379		Distribuidor_Warner_17	0.0411152
Genero_Acci&#65533;n_0	-0.158996		Genero_Biografia_4	0.0387714
Region_USA_4	0.157385		Genero_Infantil_8	-0.0383429
Distribuidor_U.I.P_15	0.149821		Genero_Documental_6	0.0300699
Distribuidor_BF Paris_1	0.145422		Genero_Anime_2	0.0292342
Region_ASIA_1	-0.14069		Distribuidor_Universal_16	0.0289326
Formato_2D_0	0.140594		Genero_Drama_7	-0.0260549
Distribuidor_Primer Plano_13	-0.136585		Genero_Suspenso_11	0.0256326
Distribuidor_Digicine_5	-0.124198		Remake	-0.0253828
Idioma_Castellano_0	0.123277		Idioma_Subtitulado_1	-0.0243874
Dia Especial_Vacaciones_3	0.121278		Distribuidor_Disney_6	0.0234468
Distribuidor_Energia Entusiasta_8	0.118765		Distribuidor_Diamond_4	-0.0160514
Mes_Noviembre_9	-0.108241		Region_LATAM_3	-0.0155475
Distribuidor_Paramount_12	-0.107187		Formato_3D_1	0.014648
Genero_Superheroes_10	-0.105589		Mes_Mayo_8	-0.0119714
Genero_Aventuras_3	-0.102886		Distribuidor_Sony_14	-0.010615
Franja_Vermouth_4	0.0990338		Mes_Septiembre_11	0.00867894
Bias	0.0988893		Mes_Agosto_1	-0.00653414
Funciones por dia	0.0943677		Distribuidor_Independiente_10	0.00324584
Distribuidor_Eclectique SA_7	0.0927297		Genero_Terror_12	-0.00322833
Presupuesto	0.0917487		Weekday_Wednesday_6	0.00319671
Mes_Abril_0	0.0878803		Dia	-0.00244342
Distribuidor_Fox_9	0.0846676		Distribuidor_Mirada_11	-0.00124657
Franja_Trasnoche_3	0.0795399			

7.5 Visualización de resultados: Regresión Lineal

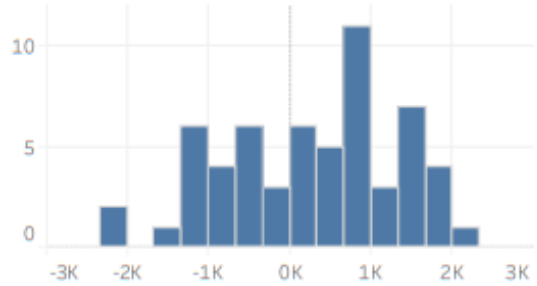
Resumen de Resultados - Linear Regression

	RSME	ATTE AVG	RSME / ATTE AVG	RSE
x Funcion	45	66	68,1%	0,541
x Dia	1.113	5.280	21,1%	0,362
x Semana	3.924	33.482	11,7%	0,095
x Mes	11.221	155.774	7,2%	0,348

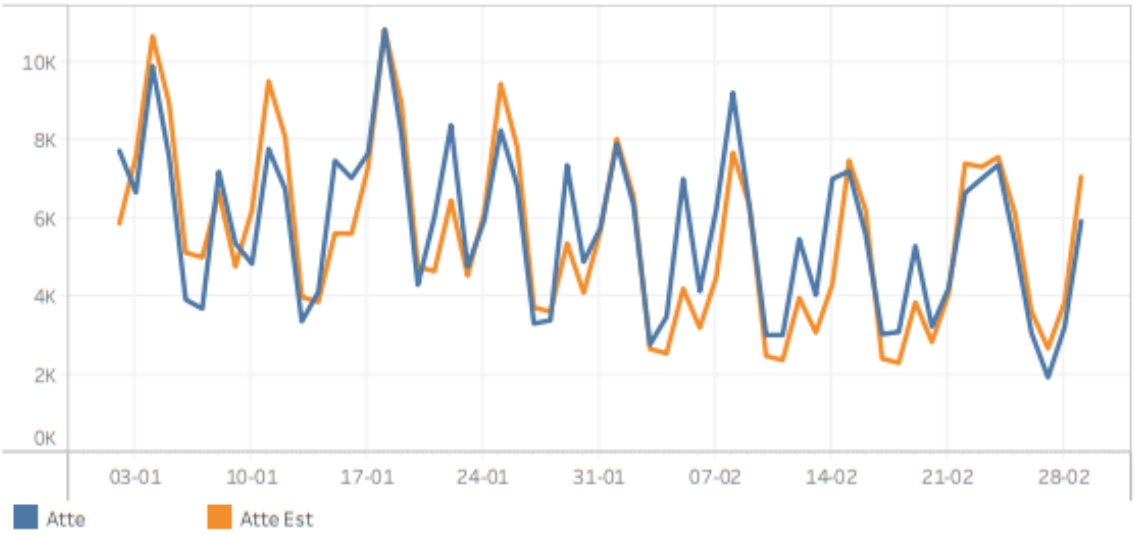
Histogramas de Desvios x Funcion



Histogramas de Desvios x Día



Comparación Attendance vs Attendance Estimada x Día



Attendance vs Attendance Estimada x Semana

Semana	Atte	Atte Est	Ratio
2	43.109	49.830	87%
3	36.637	41.982	87%
4	48.438	48.525	100%
5	36.784	40.572	91%
6	35.232	33.723	96%
7	34.504	30.473	88%
8	32.562	29.601	91%
9	34.072	38.890	88%

Detalles Enero - Linear Regression

Dia1	Weekday1	Atte	Atte Est	Ratio
2	Thursday	7.143	5.862	82%
3	Friday	6.152	7.580	81%
4	Saturday	9.144	10.663	86%
5	Sunday	7.002	8.947	78%
6	Monday	3.617	5.116	71%
7	Tuesday	3.403	4.997	68%
8	Wednesday	6.648	6.665	100%
9	Thursday	4.949	4.758	96%
10	Friday	4.466	6.203	72%
11	Saturday	7.187	9.508	76%
12	Sunday	6.230	8.072	77%
13	Monday	3.101	3.982	78%
14	Tuesday	3.798	3.851	99%
15	Wednesday	6.906	5.608	81%
16	Thursday	6.497	5.599	86%
17	Friday	7.077	7.299	97%
18	Saturday	10.008	10.828	92%
19	Sunday	7.568	8.954	85%
20	Monday	3.967	4.751	83%
21	Tuesday	5.573	4.641	83%
22	Wednesday	7.748	6.454	83%
23	Thursday	4.395	4.528	97%
24	Friday	5.487	6.131	90%
25	Saturday	7.623	9.442	81%
26	Sunday	6.304	7.802	81%
27	Monday	3.049	3.710	82%
28	Tuesday	3.128	3.606	87%
29	Wednesday	6.798	5.353	79%
30	Thursday	4.512	4.087	91%
31	Friday	5.306	5.658	94%
Total		**174.786**	**190.653**	**92%**

Pelicula1	Atte	Atte Est	Ratio
FROZEN 2	65.324	85.463	76%
EL ROBO DEL SIGLO	51.291	41.885	82%
JUMANJI EL SIGUIENTE NIVEL	22.860	21.627	95%
STAR WARS EL ASCENSO DE SKYWALKER	11.164	9.495	85%
DOLITTLE	3.053	4.518	68%
PARASITE	2.844	2.551	90%
GUASON	2.832	2.555	90%
JOJO RABBIT	2.481	3.080	81%
NUEVA YORK SIN SALIDA	1.993	2.285	87%
ENTRE NAVAJAS Y SECRETOS	1.649	2.093	79%
1917	1.602	1.668	96%
ELO DE RICHARD JEWELL	1.528	2.134	72%
LA HORA DE TU MUERTE	1.423	1.385	97%
ESPIAS A ESCONDIDAS	1.236	1.242	99%
MUJERCITAS	721	527	73%
LA POSESION DE MARY	621	1.830	34%
EL ARO	577	664	87%
MALEFICA DUENA DEL MAL	416	413	99%
CATS	302	1.362	22%
LO MEJOR ESTA POR VENIR	245	555	44%
LA MUERTE NO EXISTE Y EL AMOR TAMPOCO	172	573	30%
BACURAU	122	394	31%
JUGANDO CON FUEGO	85	168	51%
EL RECORDADOR	81	35	44%
FROZEN 2 OPEN CAPTION	69	954	7%
SOMOS CALENTURA	50	742	7%
RUMBO AL MAR	45	457	10%
Grand Total	**174.786**	**190.653**	**92%**

Detalles Febrero - Linear Regression

Comparación x Día

Dia1	Weekday1	Atte	Atte Est	Ratio
1	Saturday	7.305	8.030	91%
2	Sunday	5.868	6.569	89%
3	Monday	2.556	2.650	96%
4	Tuesday	3.212	2.534	79%
5	Wednesday	6.473	4.194	65%
6	Thursday	3.816	3.190	84%
7	Friday	5.761	4.493	78%
8	Saturday	8.518	7.679	90%
9	Sunday	5.801	6.323	92%
10	Monday	2.778	2.469	89%
11	Tuesday	2.779	2.363	85%
12	Wednesday	5.051	3.957	78%
13	Thursday	3.726	3.070	82%
14	Friday	6.478	4.307	66%
15	Saturday	6.657	7.474	89%
16	Sunday	5.167	6.214	83%
17	Monday	2.798	2.402	86%
18	Tuesday	2.850	2.292	80%
19	Wednesday	4.886	3.843	79%
20	Thursday	2.978	2.825	95%
21	Friday	3.880	4.048	96%
22	Saturday	6.142	7.398	83%
23	Sunday	6.479	7.312	89%
24	Monday	6.804	7.566	90%
25	Tuesday	4.958	6.126	81%
26	Wednesday	2.831	3.615	78%
27	Thursday	1.777	2.664	67%
28	Friday	2.961	3.845	77%
29	Saturday	5.472	7.058	78%
Total		**136.762**	**136.509**	**100%**

Comparación x Película

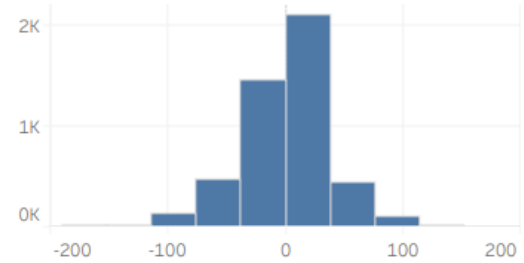
Pelicula1	Atte	Atte Est	Ratio
EL ROBO DEL SIGLO	36.488	36.375	100%
SONIC	15.199	15.474	98%
1917	12.380	12.068	97%
PARASITE	11.727	6.661	57%
ESPIAS A ESCONDIDAS	10.894	6.801	62%
AVES DE PRESA	10.809	14.476	75%
FROZEN 2	9.788	12.465	79%
BAD BOYS PARA SIEMPRE	6.837	7.622	90%
MUJERCITAS	6.271	4.383	70%
JUMANJI EL SIGUIENTE NIVEL	5.082	1.975	39%
LA MALDICION RENACE	2.378	3.808	62%
DOLITTLE	1.600	791	49%
EL ESCANDALO	1.565	2.898	54%
EL HOMBRE INVISIBLE	1.329	1.721	77%
AMENAZA EN LO PROFUNDO	1.034	1.814	57%
GRETEL Y HANSEL	803	893	90%
JUDY	645	1.822	35%
HABIA UNA VEZ... EN HOLLYWOOD	398	306	77%
EL LLAMADO SALVAJE	395	1.328	30%
BUSCANDO JUSTICIA	342	552	62%
STAR WARS EL ASCENSO DE SKYWALKER	328	216	66%
GUASON	238	208	88%
RUMBO AL MAR	101	1.168	9%
LOVE LIVE FEST 2020 9TH ANNIVERSARY	77	29	38%
EL PRINCIPE	38	421	9%
RESPIRA	16	234	7%
Grand Total	**136.762**	**136.509**	**100%**

7.6 Visualización de resultados: Boosted Decision Tree Regression

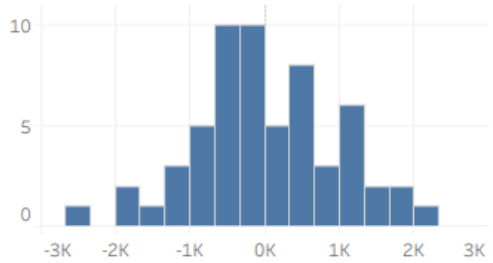
Resumen de Resultados - Boosted Decision Tree Regression

	RSME	ATTE AVG	RSME / ATTE AVG	RSE
x Funcion	38	66	58,3%	0,396
x Dia	942	5.280	17,8%	0,259
x Semana	3.185	33.482	9,5%	0,062
x Mes	1.738	155.774	1,1%	0,008

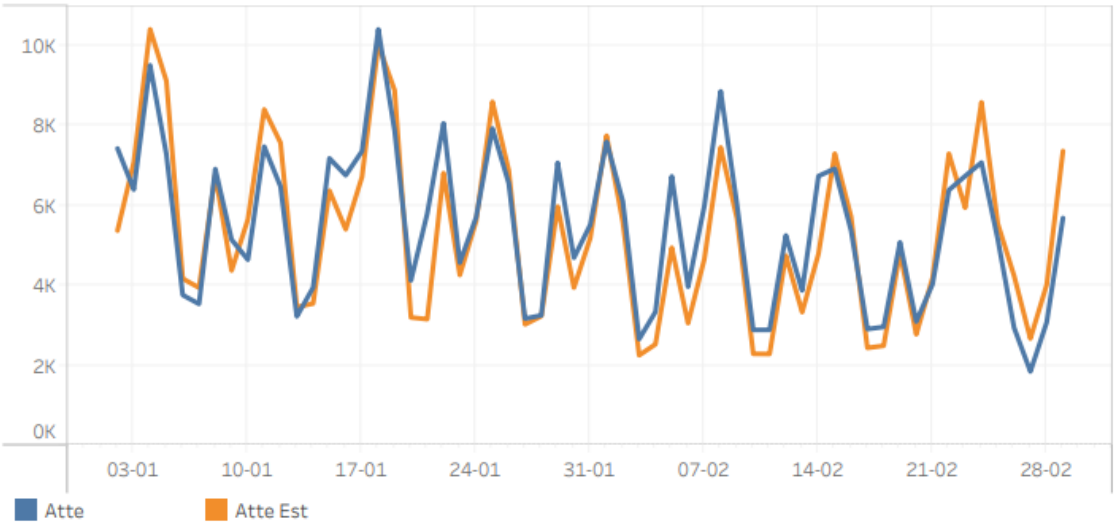
Histogramas de Desvios x Funcion



Histogramas de Desvios x Día



Comparación Attendance vs Attendance Estimada x Día



Attendance vs Attendance Estimada x Semana

Semana	Atte	Atte Est	Ratio
2	43.109	46.878	92%
3	36.637	39.406	93%
4	48.438	44.221	91%
5	36.784	37.602	98%
6	35.232	32.210	91%
7	34.504	30.189	87%
8	32.562	30.918	95%
9	34.072	38.603	88%

Detalles Enero - Boosted Decision Tree Regression

Comparación x Día

Dia1	Weekday1	Atte	Atte Est	Ratio
2	Thursday	7.143	5.374	75%
3	Friday	6.152	7.085	87%
4	Saturday	9.144	10.413	88%
5	Sunday	7.002	9.120	77%
6	Monday	3.617	4.173	87%
7	Tuesday	3.403	3.947	86%
8	Wednesday	6.648	6.768	98%
9	Thursday	4.949	4.376	88%
10	Friday	4.466	5.658	79%
11	Saturday	7.187	8.408	85%
12	Sunday	6.230	7.574	82%
13	Monday	3.101	3.462	90%
14	Tuesday	3.798	3.550	93%
15	Wednesday	6.906	6.378	92%
16	Thursday	6.497	5.407	83%
17	Friday	7.077	6.720	95%
18	Saturday	10.008	10.032	100%
19	Sunday	7.568	8.886	85%
20	Monday	3.967	3.202	81%
21	Tuesday	5.573	3.159	57%
22	Wednesday	7.748	6.815	88%
23	Thursday	4.395	4.263	97%
24	Friday	5.487	5.614	98%
25	Saturday	7.623	8.601	89%
26	Sunday	6.304	6.883	92%
27	Monday	3.049	3.028	99%
28	Tuesday	3.128	3.235	97%
29	Wednesday	6.798	5.978	88%
30	Thursday	4.512	3.948	87%
31	Friday	5.306	5.188	98%
Total		**174.786**	**177.243**	**99%**

Comparación x Película

Pelicula1	Atte	Atte Est	Ratio
FROZEN 2	65.324	71.885	91%
EL ROBO DEL SIGLO	51.291	44.048	86%
JUMANJI EL SIGUIENTE NIVEL	22.860	23.438	98%
STAR WARS EL ASCENSO DE SKYWALKER	11.164	8.981	80%
DOLITTLE	3.053	5.390	57%
PARASITE	2.844	1.623	57%
GUASON	2.832	3.014	94%
JOJO RABBIT	2.481	2.186	88%
NUEVA YORK SIN SALIDA	1.993	2.801	71%
ENTRE NAVAJAS Y SECRETOS	1.649	2.263	73%
1917	1.602	1.507	94%
ELO DE RICHARD JEWELL	1.528	2.109	72%
LA HORA DE TU MUERTE	1.423	1.088	76%
ESPIAS A ESCONDIDAS	1.236	1.331	93%
MUJERCITAS	721	406	56%
LA POSESION DE MARY	621	778	80%
EL ARO	577	605	95%
MALEFICA DUENA DEL MAL	416	492	84%
CATS	302	1.158	26%
LO MEJOR ESTA POR VENIR	245	323	76%
LA MUERTE NO EXISTE Y EL AMOR TAMPOCO	172	343	50%
BACURAU	122	245	50%
JUGANDO CON FUEGO	85	232	37%
EL RECORDADOR	81	32	40%
FROZEN 2 OPEN CAPTION	69	555	12%
SOMOS CALENTURA	50	277	18%
RUMBO AL MAR	45	133	34%
Grand Total	**174.786**	**177.243**	**99%**

Detalles Febrero - Boosted Decision Tree Regression

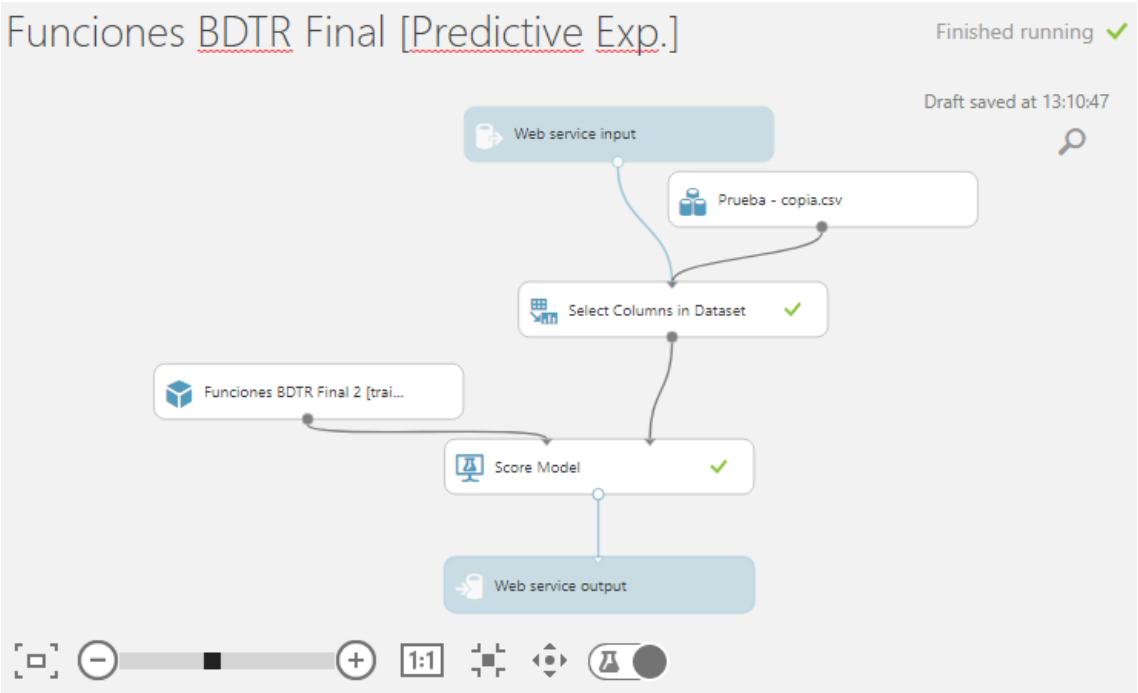
Comparación x Día

Dia1	Weekday1	Atte	Atte Est	Ratio
1	Saturday	7.305	7.751	94%
2	Sunday	5.868	5.589	95%
3	Monday	2.556	2.255	88%
4	Tuesday	3.212	2.531	79%
5	Wednesday	6.473	4.949	76%
6	Thursday	3.816	3.058	80%
7	Friday	5.761	4.664	81%
8	Saturday	8.518	7.458	88%
9	Sunday	5.801	5.680	98%
10	Monday	2.778	2.292	82%
11	Tuesday	2.779	2.287	82%
12	Wednesday	5.051	4.751	94%
13	Thursday	3.726	3.334	89%
14	Friday	6.478	4.794	74%
15	Saturday	6.657	7.302	91%
16	Sunday	5.167	5.723	90%
17	Monday	2.798	2.440	87%
18	Tuesday	2.850	2.493	87%
19	Wednesday	4.886	4.832	99%
20	Thursday	2.978	2.779	93%
21	Friday	3.880	4.196	92%
22	Saturday	6.142	7.301	84%
23	Sunday	6.479	5.946	92%
24	Monday	6.804	8.587	79%
25	Tuesday	4.958	5.550	89%
26	Wednesday	2.831	4.244	67%
27	Thursday	1.777	2.672	67%
28	Friday	2.961	4.020	74%
29	Saturday	5.472	7.368	74%
Total		136.762	136.845	100%

Comparación x Película

Pelicula1	Atte	Atte Est	Ratio
EL ROBO DEL SIGLO	36.488	38.659	94%
SONIC	15.199	14.499	95%
1917	12.380	11.361	92%
PARASITE	11.727	5.739	49%
ESPIAS A ESCONDIDAS	10.894	7.740	71%
AVES DE PRESA	10.809	14.667	74%
FROZEN 2	9.788	12.192	80%
BAD BOYS PARA SIEMPRE	6.837	7.434	92%
MUJERCITAS	6.271	3.652	58%
JUMANJI EL SIGUIENTE NIVEL	5.082	4.702	93%
LA MALDICION RENACE	2.378	3.799	63%
DOLITTLE	1.600	1.235	77%
EL ESCANDALO	1.565	2.384	66%
EL HOMBRE INVISIBLE	1.329	1.742	76%
AMENAZA EN LO PROFUNDO	1.034	1.730	60%
GRETEL Y HANSEL	803	738	92%
JUDY	645	1.397	46%
HABIA UNA VEZ... EN HOLLYWOOD	398	503	79%
EL LLAMADO SALVAJE	395	1.164	34%
BUSCANDO JUSTICIA	342	399	86%
STAR WARS EL ASCENSO DE SKYWALKER	328	221	67%
GUASON	238	311	76%
RUMBO AL MAR	101	311	33%
LOVE LIVE FEST 2020 9TH ANNIVERSARY	77	18	24%
EL PRINCIPE	38	154	25%
RESPIRA	16	96	17%
Grand Total	136.762	136.845	100%

7.7 Experimento Predictivo de Boosted Decision Tree Regression en Microsoft Azure. (Puesta el Producción)



Anexo 7.8: Plan de Proyecto de implementación

Objetivo	**Etapa 1**	**Etapa 2**		**Etapa 3**		**Etapa 4**		**Etapa 5**			
	Sem 1	**Sem 2**	**Sem 3**	**Sem 4**	**Sem 5**	**Sem 6**	**Sem 7**	**Sem 8**	**Sem 9**	**Sem 10**	**Sem 11**
Estimación de público	Planifi-cación	Armado de la base de datos	Elaboración del modelo	Prueba del modelo	Presentar resultados y forecasts, Optimizar el modelo	Presentar resultados y forecasts, Probar el modelo	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción
Estimación de Ventas Candy	Planifi-cación			Armado de la base de datos	Elaboración del modelo	Prueba del modelo	Presentar resultados y forecasts, Optimizar el modelo	Presentar resultados y forecasts, Probar el modelo	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción
Sansibilidad Precio	Planifi-cación			Armado de la base de datos	Elaboración del modelo	Prueba del modelo	Presentar resultados y forecasts, Optimizar el modelo	Presentar resultados y forecasts, Probar el modelo	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción	Presentar resultados y forecasts, Modelo en producción
Churn Rate Clientes Fidelización	Planifi-cación					Armado de la base de datos	Elaboración del modelo	Prueba del modelo	Presentar resultados y forecasts, Optimizar el modelo	Presentar resultados y forecasts, Probar el modelo	Presentar resultados y forecasts, Modelo en producción