EAE Business School

Trabajo Final de Máster

Memoria técnica

“MACHINE TR(AI)NING. Churn Rate Model”

David Rodríguez Alvarez
Adrià Llorens Galindo
Pau Joan Galiana Morell
Joan Castelltort Viñallonga
Aleix Masjoan Colomer

Índice

Resumen (Abstract)

Como equipo especializado en Data y Business Intelligence del ********** de Barcelona tenemos un nuevo reto: Mejorar la tasa de retención de nuestros clientes, reduciendo la tasa de abandono en un 10% a finales del 2025.

Según Fitness KPI (una de las fuentes de datos más destacadas del sector), la tasa media de retención de usuarios, a 6 meses, en los centros deportivos en España es de un 69% y, en nuestro caso es de un 60%. Esto supone que, a día de hoy, nuestro “churn rate” o tasa de abandono está 9 puntos por encima de la media. Y, teniendo en cuenta el potencial del centro (cuenta, por ejemplo, con 200 actividades dirigidas por semana, piscina, spa o servicios personalizados de entreno y nutrición) y aprovechando la fuerza de los datos, creemos que podemos mejorar este escenario actual.

Es muy importante tener en cuenta que estamos delante de un proyecto desarrollado con información 100% real, por lo que la complejidad a nivel de tratado y modelado de datos ha sido, también, real.

De este modo, teniendo claro el objetivo de negocio “Destinar recursos y esfuerzos a fidelizar clientes actuales, en lugar de captar nuevos” y gracias al uso de datos vamos a realizar y presentar una Prueba de Concepto (PoC) para reducir la tasa de abandono del ********** de Barcelona.

Una Prueba de Concepto que se ha enfocado en utilizar un modelo predictivo (algoritmo de Machine Learning) que ordenará de 1 a 0 todos los clientes del centro, por las posibilidades que tienen de darse de baja (siendo 1 lo más probable y 0 lo menos probable) y, una vez listados, cada cliente con elevado riesgo de fuga será sometido a un proceso de cribado y segmentación, que le derivará al departamento correspondiente (Atención al Cliente, Servicios, Técnico, Marketing o Durmientes) para que le propongan, o no, acciones necesarias para su retención.

Así, vamos a poder presentar y desarrollar una estrategia, juntamente con el resto de departamentos del centro, basada en datos, que nos va a permitir desarrollar acciones super personalizadas que deberían mejorar la relación con nuestros clientes (fidelizar) y, con ello, reducir la tasa de abandono del centro en un 10% a finales del 2025.

1. Introducción

1.1 Objetivo del proyecto Machine Tr(AI)ning

El objetivo del proyecto Machine Tr(AI)ning es, a través de los datos reales del ********** de Barcelona, desarrollar una Prueba de Concepto (PoC) de un "Churn Rate Model" (Modelo para calcular la tasa de abandono) para reducir la tasa de abandono en un 10% a finales del 2025.

Y, gracias a los datos, poder predecir y ordenar las personas que más posibilidades tienen de darse de baja, en función de sus características demográficas y del uso que hacen de los diferentes servicios del centro. Una vez listadas todas, desde la que más posibilidades tiene de darse de baja hasta la que menos, segmentarlas y derivarlas al departamento correspondiente, para promover diferentes acciones y estrategias que permitan fidelizarlas y retenerlas el máximo tiempo posible.

1.2 Equipo

El equipo destinado a trabajar en este proyecto se encuentra actualmente en nómina del ********** y cuenta con formación en Data, Business Intelligence y Desarrollo de Negocio. De este modo, todo el proceso de selección, extracción, modelado, procesado y lectura de datos se realizará internamente y sin apoyo externo.

Por orden alfabético, el equipo de Machine Tr(AI)ning estará formado por:

- Joan Castelltort
- Pau Galiana
- Adrià Llorens
- Aleix Masjoan
- David Rodríguez

1.3 Situación actual de la Industria del Fitness

Tradicionalmente, como en todas las industrias, el sector del fitness siempre se ha movido entre dos grandes objetivos de negocio: fidelizar y captar nuevos usuarios. Dos objetivos que permiten, por un lado, conseguir más negocio (sin incrementar notablemente los recursos) y, por otro, retener el máximo tiempo posible a cada persona abonada.

Y, entendiendo que estamos hablando de un modelo de negocio de membresía o abono, para poder elaborar la planificación a futuro de los centros deportivos, es imprescindible poder saber (o estimar con precisión) con cuántas personas abonadas contarás el próximo mes, trimestre, semestre o año. Este número de personas abonadas, multiplicado por la cuota (menos gastos fijos y variables) te da un primer valor sobre el cuál puedes verificar la salud financiera del centro, optimizar la eficiencia operativa y desarrollar planes sólidos a futuro.

De todos modos, hoy en día los hábitos de consumo de deporte y centros deportivos (y de muchos otros sectores) han cambiado completamente por culpa (o gracias) a un detonante común: La pandemia mundial del Covid-19.

Desde un punto de vista general, el comportamiento de las personas post-pandemia ha cambiado porque entendemos la vida de otro modo, las sociedades se han digitalizado, concienciado y adoptado nuevos valores y eso tiene, como siempre, aspectos positivos y negativos.

Y, si nos centramos específicamente en el sector del fitness, por un lado, la salud y el buen estado de forma física y mental se valora mucho más que antes, pero, por otro lado, hay mucha más oferta para hacer deporte. Un claro ejemplo es la expansión de gimnasios 'low cost', sin personal y abiertos 24 horas al día; la aparición de aplicaciones de suscripción que permiten acudir físicamente a diferentes centros deportivos; la explosión de nuevas disciplinas deportivas (indoor y outdoor) como CrossFit o Hyrox; la gran oferta de diferentes aplicaciones para realizar entrenos virtuales en casa sin necesidad de equipamiento (calistenia); los grupos de entreno al aire libre; nuevos conceptos como los gimnasios boutique, etc.…

A día de hoy los usuarios tienen a su disposición 16 centros deportivos diferentes al **********, sin contar el aire libre o el deporte en casa, en una superficie de menos de 2km^2 del mismo.

Figura 1: Fuente: Google Maps

En definitiva, la sensación global actual es que las ganas e iniciativa de las personas para hacer deporte están en niveles muy superiores que en pre-pandemia, pero la oferta y las posibilidades para practicarlo han crecido también al mismo ritmo. Por eso, el objetivo de

negocio de los centros deportivos se debe enfocar más hacía la fidelización que en la adquisición y, como consecuencia, se potencia la digitalización y la captación de datos e información de personas abonadas (a través de herramientas online y el uso de ERP-CRM) para poder entender mejor sus comportamientos, deseos y necesidades y, de este modo, ofrecer y mejorar todo el catálogo de servicios, ajustándolo y personalizándolo, en la medida de lo posible, a las necesidades de cada persona.

1.4 Prueba de Concepto (PoC)

El desarrollo de la Prueba de Concepto (PoC) del "Churn Rate Model" servirá para testear, como prototipo inicial, diferentes algoritmos de machine learning; validar el tratamiento de datos propuesto y las diferentes variables elegidas, además de comprobar la viabilidad del resultado final (outputs), sin destinar gran cantidad de tiempo ni recursos.

O, dicho de otro modo, esta PoC es la primera fase de exploración que, sin tomar demasiados riesgos, validará la viabilidad del proyecto, antes de comprometer el 100% de los recursos que requiere, para conseguir el objetivo de reducir la tasa de abandono en un 10% a finales de 2025.

El resultado de la misma puede ser:

- Favorable: La idea del uso del algoritmo de machine learning para reducir la tasa de abandono del centro funciona como se esperaba y los criterios técnicos elegidos cumplen con lo esperado, lo que supone que habría llegado el momento de destinar el 100% del tiempo y los recursos que necesita el proyecto Machine Tr(AI)ning.

- Parcial: Parece que la idea del uso del algoritmo de machine learning para reducir la tasa de abandono del centro funciona, pero hay que corregir aspectos técnicos o utilizar conjuntos con mayor cantidad de datos para mejorar su rendimiento.

- Fallido: El uso de un algoritmo de machine learning para reducir la tasa de abandono del centro, con el conjunto de datos actual, no funciona como se había previsto, por lo que hay que rediseñar el enfoque y realizar una nueva PoC.

Y, una vez obtenido el "output" (en forma de listado de personas con más posibilidad de darse de baja del centro), segmentar las personas y derivarlas al servicio del gimnasio que más pueda potenciar su retención, recopilando información de todo el proceso.

2. Desarrollo de la Prueba de Concepto (PoC)

2.1 Objetivos de la Prueba de Concepto (PoC)

El objetivo de esta Prueba de Concepto es, utilizando los datos actuales que dispone el centro, desarrollar un modelo predictivo que sea capaz de predecir y ordenar futuras bajas de personas abonadas, para tomar diferentes acciones sobre ellas y reducir la tasa de abandono del centro (fidelizar).

Con esta primera fase de exploración, además de conseguir el listado de personas con mayor riesgo de fuga, podremos:

- Crear nuevos segmentos de público: Clasificar a las personas abonadas, en función de diferentes criterios, y derivarlas al personal del centro que más capacidad tiene de retenerlos.

- Detectar factores de riesgo: Conocer comportamientos y usos de los servicios del centro que pueden ser indicadores de una futura baja.

- Desarrollar estrategias personalizadas: Dar apoyo 100% personalizado a cada persona del centro, entendiendo su forma de consumir el deporte en el gimnasio y potenciando sus deseos, para evitar que se den de baja.

- Mejorar la oferta de servicios: Poder ampliar o reducir el catálogo de servicios, centrándose directamente en los servicios que realmente aportan valor a los clientes, gracias a la atención personalizada brindada por los diferentes departamentos del centro y, también, a la recogida de datos de cada nueva acción y atención.

3. Solución

3.1 Descripción

El proyecto tendrá como punto de partida el listado “Anticipación de bajas” y a partir de los 50 abonados con mayor riesgo de fuga, se realizará un estudio de cada caso y haremos una segmentación (o cribado) de los diferentes perfiles, para derivarlo directamente al departamento más indicado, para que puedan intervenir y promover su continuidad en el centro.

Se eligen los primeros 50 usuarios con más riesgo de abandono como primera fase de exploración para no saturar los recursos del centro y garantizar una implantación de calidad.

La distribución se realizará según los siguientes criterios:

- Accesos
- Usos de instalaciones y servicios
- Información sociodemográfica

Estos criterios pueden ser complementarios, por lo que el responsable del cribado elegirá cuál tiene más peso en cada asignación.

3.2 Usuarios

El equipo del ****************************** que trabajará para evitar futuras bajas estará liderado por un coordinador de proyecto (perteneciente al departamento de Atención al Cliente) que distribuirá los perfiles seleccionados, en función de cada caso, a los siguientes departamentos:

- Atención al cliente, equipo que realizará mayoritariamente tareas de contacto telefónico y citas de asesoramiento con abonados que provengan de los siguientes segmentos:

    - Abonados que han realizado mayoritariamente Actividades Dirigidas: El centro cuenta con 29 actividades dirigidas diferentes, repartidas en más de 190 horarios cada semana. Se guiará al abonado para que pueda encontrar las actividades más adecuadas a sus gustos y propósitos.

    - Abonados con accesos y usos esporádicos (sólo “picotean”) en el centro sin ningún patrón en concreto: Ofrecer un seguimiento, por parte de Atención al Cliente, para dar a conocer todas las posibilidades que tienen dentro del centro, potenciando los diferentes usos, para que encuentren los resultados que más les satisfagan.
    - Abonados que se han inscrito al centro y el porcentaje de semanas sin venir supera el 70 % de sus semanas como abonado, dentro de los primeros 9

meses: Perfiles que se inscriben, hacen uso durante las primeras semanas y luego dejan de venir hasta que se borran, normalmente entre 3er y 6º mes. Testear si, ofreciéndoles un servicio de acompañamiento en grupos pequeños vuelven a tomar la rutina deportiva y, poco a poco, ir haciendo un uso cada vez mayor del centro.

- Departamento técnico, equipo que realizará tareas de asesoramiento técnico, creación de hábitos alcanzables y podrá ofrecer gratuitamente entrenos personalizado de 30 minutos para este tipo de perfil:

    - Abonados que, por sus accesos, realizan un uso compulsivo y luego dejan de venir: Suelen ser personas a las que les cuesta crear un hábito o que se crean unas expectativas desproporcionadas, con un asesoramiento técnico personalizado preparado por profesionales se les puede reconducir la situación, hacer un plan a medio-largo plazo y ayudar a gestionar las expectativas, para que puedan realizar un uso más eficiente del centro y todos sus servicios.

    - Abonados que asisten principalmente a sala de fitness: A partir de sus métricas y datos de entrenamiento (el centro cuenta con máquinas de fuerza y de cardio digitalizadas) proponer nuevas rutinas y patrones de entreno para potenciar su deseo o necesidad.

    - Abonados que hacen uso de la piscina: Primer contacto con este segmento para saber si son usuarios a los que realmente les gusta las actividades de agua o si utilizan la piscina por falsas creencias (el entrenamiento en agua es menos lesivo o es “lo que les ha recomendado el médico”). En función de cada caso se les reconducirá a otras soluciones de actividad física, fuera del agua, o se potenciará su entreno en piscina para poder mejorar.

- Departamento de servicios, equipo que, gracias a su conocimiento de los usuarios (es muy amplio tanto personal como digitalmente), podrá ofrecer actividades a coste 0:

    - Abonados que hayan contratado servicios de nutrición: Se podrá reevaluar su situación, estudiar la evolución de sus mediciones de composición corporal con máquina de bioimpedancia *Inbody* y/o valorar cómo pueden alcanzar sus propósitos con una sesión gratuita.

    - Abonados que habían consumido servicios de entrenador personal: Se reunirán con el técnico de la persona para evaluar el caso y se hablará con el usuario para saber si el servicio ha cumplido con sus expectativas o prefiere cambiar de técnico y disponer de una sesión gratuita.

- Departamento de Marketing, equipo que trabajará la comunicación del centro a través de los diferentes canales (redes sociales, mailing o WhatsApp) para poder ofrecer y promocionar soluciones personalizadas. Es un departamento que trabajará conjuntamente con el resto de departamentos.

- Abonados que nos siguen en redes sociales: Se indagará sobre los gustos y estilo de vida de esas personas.

- Abonados que tengan su perfil público en Linkedin: Se buscará el perfil profesional por si sus horarios fueran determinantes.

- Abonados que acuden a nuestras actividades más sociales (salidas a la nieve, deportes de aventura, surfcamp, free tours, museos, etc.): Se estudiará qué tipo de usos se les pueden recomendar, según sus intereses sociales.

- Abonados detectados como durmientes: Pese a no venir al centro siguen manteniéndose como abonados y cumplen con la norma de que sus semanas sin venir superan el 50% de sus semanas como abonados, llevando más de 9 meses con nosotros. A este perfil se le dejará tranquilo sin perturbar su descanso.

En función de cada caso particular los diferentes departamentos podrán trabajar de manera individual o interactuar holísticamente entre ellos.

3.3 Caso de uso

Para cada abonado que aparezca en el listado de 50 personas con mayor riesgo de baja, exceptuando a los catalogados como “durmientes”, se elaborará el siguiente informe:

Figura 2: Ejemplo de ficha análisis cliente

4. Datos

La fuente de datos utilizada proviene de diferentes tablas que se han exportado directamente del ERP & CRM *DeporWin.*

De todos los datos existentes en el sistema se ha seleccionado un período de fechas que permita representar la realidad actual del centro. De este modo, se ha decidido trabajar con datos que van de enero de 2022 a diciembre 2023. Así evitamos sesgos de fechas anteriores, que incluían comportamientos inusuales de altas y bajas, correspondientes al período más extremo de recuperación post-covid.

A la hora de elegir las tablas, los datos seleccionados representan la idiosincrasia del club y garantizan la trazabilidad de un abonado en el centro.

En un primer lugar se ha elegido una tabla que recoge datos demográficos y datos relacionados con la membresía:

- Fecha de nacimiento
- Sexo
- Id
- Tipo de abono
- Fecha de alta
- Fecha de baja
- Valor de si la Fecha de baja es por un cambio de cuota

Además, se han exportado diferentes tablas que nos devuelven el valor de usos que ha realizado cada abonado por mes:

- Tabla con nº de accesos
- Tabla con reservas en actividades dirigidas
- Tabla con reservas en el uso de piscina
- Tabla con las compras de entreno personal
- Tabla con compras de sesiones de nutrición
- Tabla sobre si han asistido a la entrevista de bienvenida al centro

A modo de resumen de la selección recogida de datos, se presentarán los KPI's actuales de los datos que se van a utilizar.

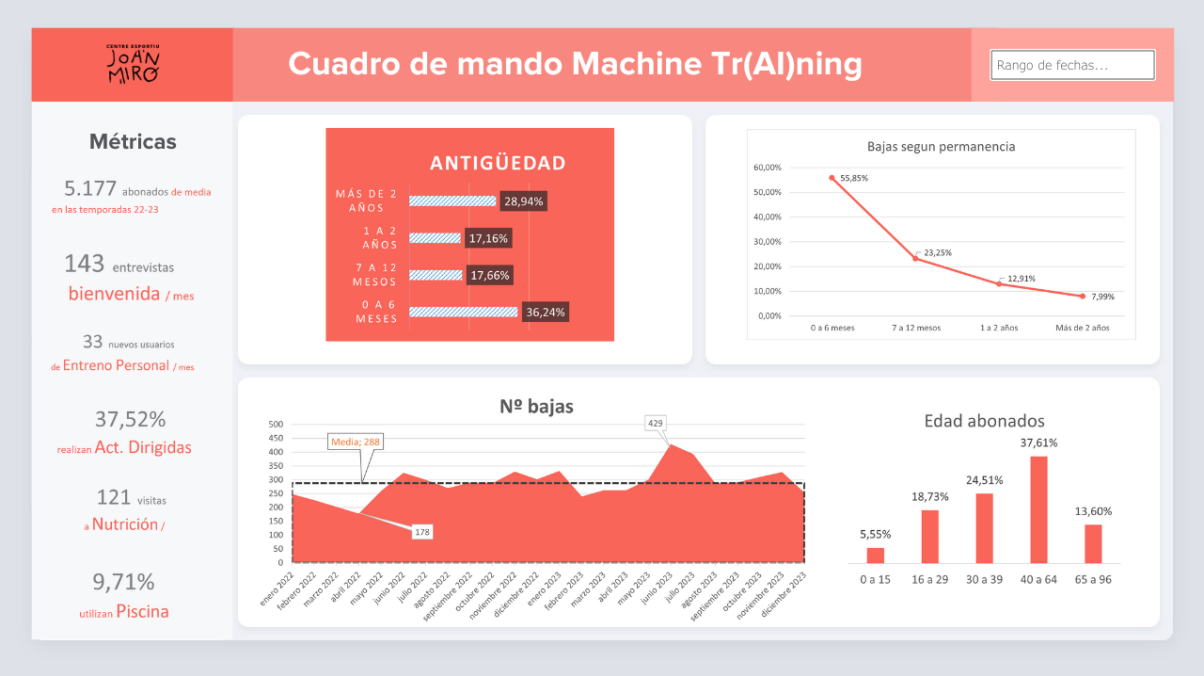

Figura 3: Dashboard análisis extracción de datos

4.1 Normalización de datos

Después de la exportación de datos se realizará una limpieza, para evitar valores faltantes, outliers y normalizar las variables numéricas.

En la interpretación de datos se excluirán los que, por su naturaleza, puedan alterar los parámetros del modelo como, por ejemplo, cuotas que, por tratarse de un centro municipal, no tengan un rendimiento económico (pertenecen a proyectos sociales o de colaboración con otros entes públicos).

5. Modelo predictivo

Un modelo es una herramienta matemática o computacional que utiliza datos históricos o actuales para predecir eventos o resultados futuros.

En este caso de estudio, se utilizan los datos de los diferentes usuarios del gimnasio para encontrar patrones en estos que ayuden a detectar los clientes con mayor riesgo de darse de baja.

5.1 Modelos

Una vez los datos han sido limpiados y se han ordenado y estructurado de forma adecuada para ser tratados, se procesan con diferentes modelos con el fin de encontrar el que presente una menor imprecisión a la hora de detectar correctamente los clientes que se dan de baja y los que no.

Para la PoC se han empleado los siguientes modelos:

- Regresión Logística
- Random Forest Classifier
- Gradient Boosting Classifier

Se contempla la posibilidad de utilizar redes neuronales como modelo, pero la idea queda descartada al presentar una exigencia computacional muy elevada en relación al resultado esperado con este problema de clasificación.

5.2 Métrica

La métrica es una función que se utiliza para evaluar el rendimiento de un modelo.

El caso de este estudio se trata de un problema de clasificación con 2 clases a distinguir - baja y no baja (clasificación binaria).

Para problemas de estas características, el AUC (Area under the curve) es una métrica que se ciñe muy bien, y en este caso AUC-ROC (Receiver Operating Characteristic) al proporcionar una medida general de su capacidad para separar ambas clases.

5.3 Validación cruzada

Para determinar qué modelo es más adecuado para el conjunto de datos empleados se realiza validación cruzada con todos ellos.

Se debe tener en cuenta cuando se opera con series temporales que hay que respetar la cronología de los datos* y evitar utilizar registros futuros para entrenar el modelo, evitando así la fuga de datos (data leakage).

**En este estudio se trabaja con múltiples series temporales (1 por cada usuario) y estos usuarios se consideran independientes entre sí. Este factor permite que se puedan emplear datos futuros de los usuarios para entrenar el modelo, siempre y cuando los usuarios seleccionados para testear el modelo (conjunto de test) no estén presentes en dicho conjunto de entrenamiento (o los datos presentes sean siempre anteriores a los datos utilizados en el conjunto de test).*

Proceso de selección de validación cruzada:

Una vez delimitado el perímetro del problema, se realiza la prueba de validación cruzada, la cual pasó por varias hipótesis, todas ellas llevadas a cabo:

- Hipótesis primera - Variante de Blocked Cross Validation:

  Función que separa todos los registros en diferentes bloques categorizados por meses en orden cronológico. Cada bloque de registros es separado en 5 splits con estratificación en el output (para garantizar la correcta distribución de las bajas en cada subconjunto) y se itera sobre los 5-folds para la validación. Para el cálculo del resultado se realiza el promedio de todas las iteraciones.

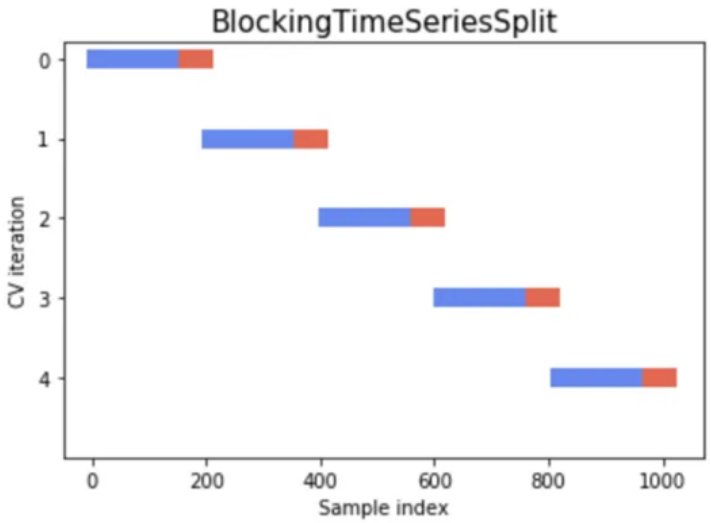

Figura 4: Ejemplo Blocking Time Series Split

  Primera hipótesis descartada al realizar muchas particiones, entrenando el modelo con conjuntos de datos pequeños, siendo poco representativos del conjunto global de los datos.

- Hipótesis segunda - Forward Chaining Cross Validation:

  Función que agrupa todos los registros en diferentes bloques categorizados por meses en orden cronológico. Se emplea un conjunto para el entrenamiento y el conjunto inmediatamente posterior para el testeo. Se itera sobre el total de meses para que el conjunto de entrenamiento vaya incrementando su volumen con cada iteración, agregando los datos anteriormente usados para testear como nuevos datos de entrenamiento.

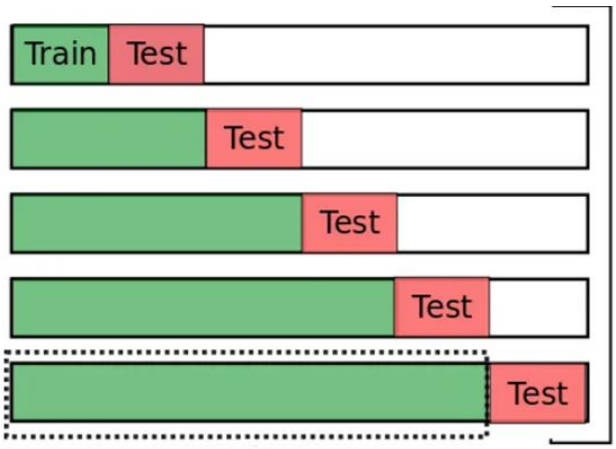

Figura 5: Ejemplo Forward Chaining Time Series Split

  Segunda hipótesis descartada al presentar una demanda computacional bastante elevada y la partición de los datos de entrenamiento en conjuntos muy pequeños (especialmente en las primeras etapas de la iteración), siendo poco representativos del conjunto total de datos.

- Hipótesis final - Group KFolds con estratificación:

  Función que separa el dataset completo en múltiples splits agrupados por ID de clientes y con estratificación, asegurando que tengamos una división uniforme de clientes en cada split (puede conllevar pequeñas diferencias en la cantidad de registros que tiene cada subconjunto teniendo en cuenta que no todos los usuarios tienen el mismo número de registros) y que cada subconjunto tenga una distribución parecida de bajas.

  La opción de emplear esta solución con Leave One Out (resultado más representativo del rendimiento real del modelo) fue descartada dada la extremadamente alta demanda computacional.

  Esta función es considerada la más adecuada dada la estructura de los datos, en que empleamos gran parte de los datos para entrenar el modelo (muy representativo del dataset) y validamos sobre un subconjunto nuevo, sin posibilidad de que en el entrenamiento haya podido observar patrones de los datos a testear.

Al realizar la validación cruzada de la hipótesis final y definitiva, se itera adicionalmente sobre los datos de entrenamiento, en que el dataset empleado para entrenar y testear el modelo va descartando columnas en cada iteración (realizando todas las combinaciones posibles entre las columnas a descartar) con el fin de verificar si hay datos que no aportan valor a la predicción.

5.4 Resultado

Una vez finalizado el proceso de entrenamiento y validación cruzada se obtiene el siguiente resultado:

Modelos	AUC_ROC
Logistic Regression	0,83
Gradient Boosting Classifier	0,86
Random Forest Classifier	**0,87**

Tabla 1: Resultados AUC_ROC de la validación cruzada

El modelo que presenta un mayor rendimiento es el **Random Forest Classifier** con un valor de **AUC_ROC = 0,87.**

Los parámetros para dicho modelo son: Profundidad - 14 / Estimadores - 300

El mejor resultado se ha logrado usando el dataset entero a excepción de las variables referentes al mes de entrada en el gimnasio.

5.5 Análisis del resultado

El resultado obtenido para esta primera prueba de concepto es bastante alentador al obtener un valor de AUC bastante próximo a 1.0, sin embargo, hay otras perspectivas para analizar el resultado del mismo.

En este caso, se han evaluado los errores de predicción que genera el modelo, que pueden ser los dos siguientes escenarios:

- Falsos positivos - El modelo indica erróneamente que el cliente se dará de baja
- Falsos negativos - El modelo indica erróneamente que el cliente no se dará de baja

Se obtiene la siguiente tabla estimativa de las predicciones realizadas:

Predicciones	Valores reales: Baja	Valores reales: No Baja
Baja	3297	13985
No Baja	2030	96298

Tabla 2: Tabla de confusión

A partir de estos datos se pueden obtener dos reflexiones:

- El modelo presenta una mejora significativa de la eficiencia respecto al azar a la hora de detectar los casos que se dan de baja, dado que evaluando el 15% de los usuarios que presentan una mayor probabilidad de darse de baja se estará encontrando más del 60% de los usuarios que realmente se dan de baja.

- El 87% de los errores de predicción que genera el modelo son falsos positivos (es decir, clientes que no se van a dar de baja que el modelo interpreta que sí lo van a hacer).
  Considerando que el resultado del modelo se empleará como una lista de usuarios con mayor probabilidad de darse de baja para a través de diferentes acciones promulgar su retención, los errores por falsos positivos son considerados menos críticos que el 13% restante, usuarios que se dan de baja y no han sido detectados.

5.6 Código

El código puede encontrarse en el siguiente directorio de GitHub:

https://github.com/JoanCastelltort/Chrun-Model.git

6. Consumo del modelo

Existen varios conceptos asociados y relacionados a lo que se refiere al *consumo del modelo* como tal. Algunos de estos tipos de consumo son:

- *Business*
  El consumo de los resultados por parte de las personas encargadas, como analistas, responsables de negocio, estrategia, marketing, etc.

- *Operativo*
  El consumo de herramientas, horas, personas y recursos que la compañía destina para llevar a cabo el proyecto.

- *Computacional*
  El consumo de poder de computación o "fuerza bruta" que necesita el modelo para ejecutarse: on premise, hybrid, cloud, clústers, arquitectura e infraestructura, etc.

6.1 Consumo *Business*

A nivel de negocio el modelo predictivo se traducirá en unos resultados que nos permitirán extraer valor empresarial.

La forma de consumo más inmediata, será una lista con nombres y apellidos de los 50 socios del gimnasio con más probabilidad de darse de baja, ordenada de mayor a menor probabilidad. Esta lista se obtendrá cada mes de manera cíclica y periódica, y será la hoja de ruta principal para conseguir el objetivo primordial del proyecto: reducir la tasa de abandono.

A partir de esta hoja de ruta, de manera automática y personalizada para cada cliente, se generará una ficha o informe en formato dashboard con sus principales datos: accesos, usos, actividades, etc.

Mediante un cribado, se derivará cada caso particular al departamento correspondiente en función de la ficha de características de ese cliente. Los departamentos involucrados y encargados para estas funciones, van a ser; Atención al cliente, departamento Técnico, departamento de Servicios y departamento de Marketing.

Una segunda forma de consumo, alternativa, no tan directa y algo más general, será un dashboard con los principales KPIs y datos relevantes a nivel total del gimnasio.

Estos datos de manera bien estructurada y ordenada, nos proporcionarán insights y nos permitirán analizar y conocer profundamente los patrones de comportamiento del cliente. Con la misma periodicidad que la lista anterior, este dashboard también se actualizará el día 1 de cada mes, dando al departamento de Dirección y gerencia un margen de tiempo suficiente para tomar decisiones.

Esencialmente, se tratará de un contenido más bien económico para aquellas áreas relacionadas directamente con el negocio y con el fin de obtener un valor empresarial de los resultados.

6.2 Consumo *Operativo*

A nivel operativo, como mínimo hará falta un científico de datos encargado de desarrollar el modelo predictivo y de evaluarlo con tal de obtener el más eficiente, y así contribuir a obtener el mejor de los resultados. Además, y por el momento, también se encargará de todas aquellas tareas relacionadas con la extracción y limpieza de datos (ingeniería de datos) y la interpretación de resultados (análisis de datos), a la vez que asegura que su trabajo se alinea a la visión de negocio.

Dada la necesidad de reentrenar periódicamente el modelo (como mínimo 1 vez al mes para cumplir las necesidades de negocio) y de monitorizar constantemente sus resultados, el mismo científico de datos deberá estar contratado en formato permanente o semi permanente.

En un futuro cercano se plantea ir escalando en tamaño e ir recopilando datos y más datos, con el fin de alimentar el modelo predictivo y aumentar su eficiencia y precisión.

Esta escalabilidad podría provocar la necesidad de incorporar a un Ingeniero o Arquitecto de datos que dé soporte en estas tareas. Este profesional podría aportar conocimiento en plataformas híbridas o cloud, poder computacional, desarrollo de pipelines y automatización de procesos.

6.3 Consumo *Computacional*

Para la PoC que nos ocupa, el modelo predictivo no ha requerido de utilizar servicios de pago. Para el caso de uso se ha utilizado la plataforma de código Google Collab y Google Drive, ambos de servicio gratuito hasta cierta computación. La computación que actualmente necesita el modelo, con el tamaño de los datos que le hemos introducido, hace que su ejecución sea posible y viable, pero costosa de tiempo.

Cabe añadir también, que los datos que se suben a Collab, en el preciso instante en que se suben dejan de ser privados, por lo que hay que ir con cuidado con la ley de protección de datos.

A corto y medio plazo no se prevé un aumento de clientes ni de datos tan considerable como para tener que escalar el pipeline y migrar a plataformas cloud. Aun así, en un caso hipotético, podría plantearse como una opción viable.

Para garantizar una gran escalabilidad a futuro, evitar problemas de protección de datos, ganar computación y mucho tiempo de espera, lo más recomendable es utilizar los recursos Cloud. Actualmente el mercado está tirando hacia allí, y existen ya tarifas muy buenas y competentes en el mercado. Las principales son AWS (Amazon), GCP (Google) y Azure

(Microsoft). Además, estas mismas plataformas cuentan con múltiples servicios, hasta el punto que te permiten integrar todo el ciclo de tus datos en ella. Ofrecen servicios de computación, de bases de datos relacionales y no relacionales, desarrollo de pipelines, servicios de análisis de datos, de machine learning, etc.

A continuación detallamos el coste aproximado de lo que podría ser una hipotética pipeline 100% Cloud. Este coste total se basa en unas hipótesis, estados y suposiciones estándares para el caso de uso actual. El coste del mismo puede variar en función del servicio Cloud elegido, por eso se muestra en un formato de horquilla.

Servicio Cloud	Coste mensual
Almacenamiento de datos *Almacenamiento de **100 GB** de datos de clientes y registros de gimnasio*	€3.72 a €9.30
Procesamiento ETL (Transformación de datos) *Procesamiento diario de datos (ETL) durante **2 horas al día** para limpiar y transformar datos de clientes, registros de acceso, y datos de IoT.*	€27.90 a €46.50
Entrenamiento del modelo *Entrenamiento del modelo diariamente durante **2 horas**, usando una instancia estándar para el modelo predictivo de churn*	€46.50 a €69.75
Evaluación y validación del modelo *Evaluación diaria del rendimiento del modelo, usando instancias más pequeñas durante **2 horas** al día.*	€37.20 a €55.80
Monitoreo y almacenamiento de logs *Monitoreo continuo de logs y métricas durante todo el mes, para detectar anomalías en el rendimiento del modelo o infraestructura.*	€9.30 a €18.60
Predicción mensual *Generación de **1 predicción mensual** usando un endpoint dedicado para procesar las predicciones de churn del gimnasio.*	€1 a €5
Total	**€126 a €205**

Tabla 3: Resumen de costes del consumo del modelo

Y a continuación detallamos la canalización actual y la que podría llegar a ser la futura:

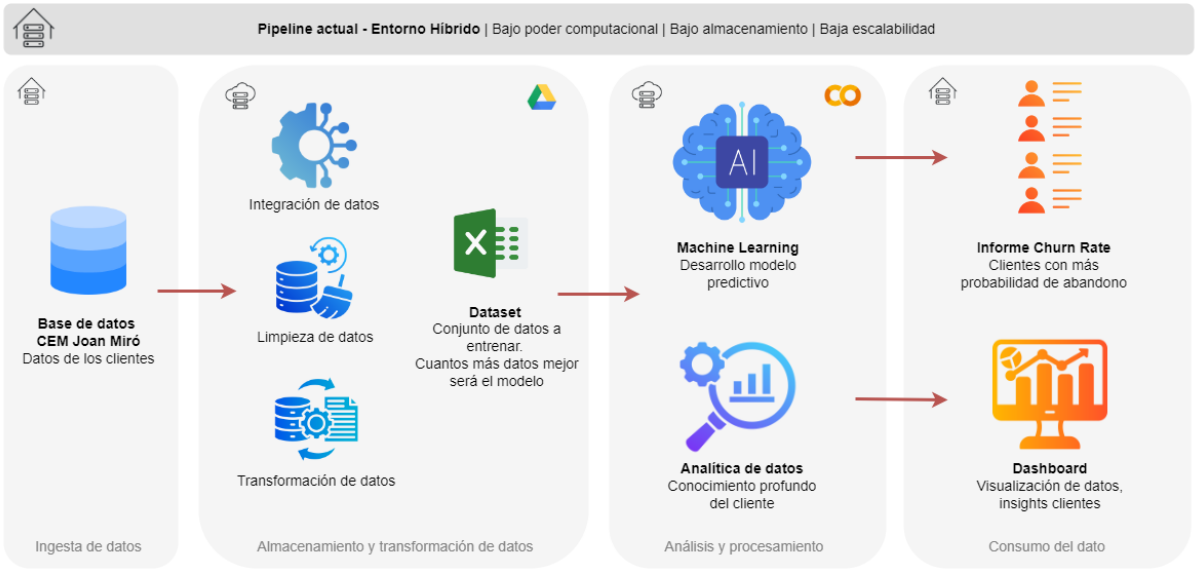

Figura 6: Pipeline Actual

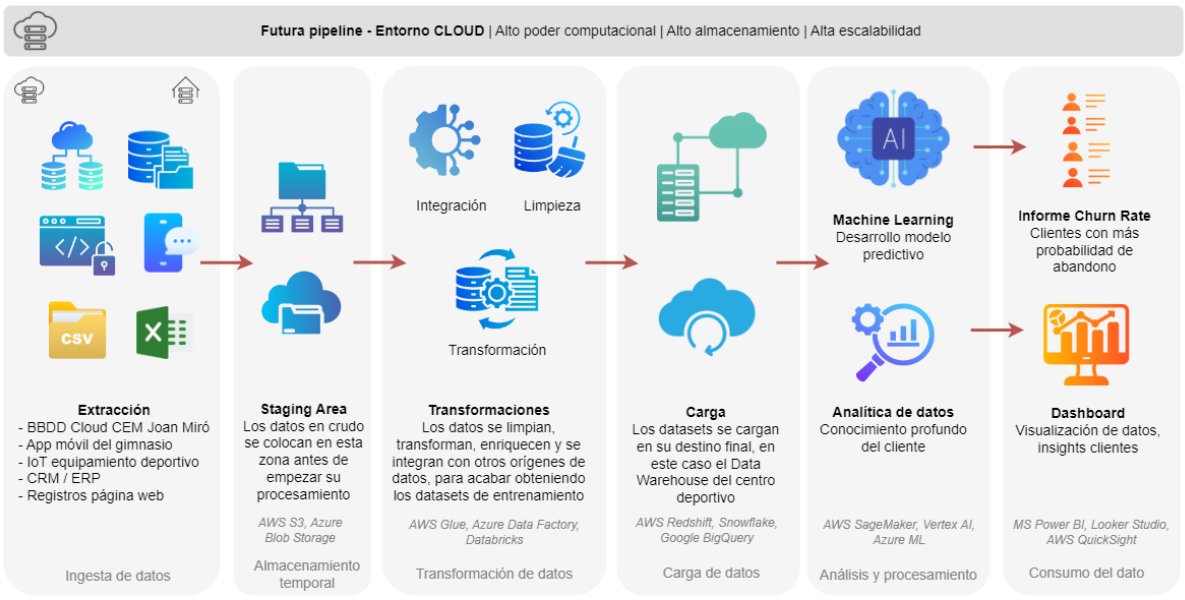

Figura 7: Pipeline futura con escalabilidad del proyecto

7. Siguientes pasos y líneas futuras de trabajo

Siguiendo con la mejora continua del proyecto a futuro, se plantean una serie de acciones estrategias que hemos considerado dividir en tres áreas clave: Calidad de los datos, optimización del modelo y automatización e implementación. Estos próximos pasos son cruciales para alcanzar el objetivo de mejorar la tasa de retención de abonados del **********.

7.1 Mejora de la calidad de los datos

Revisión y limpieza de datos: corregir registros erróneos y establecer protocolos claros para la recogida y actualización de información. Se deberá asignar a una persona que revise el CRM para asegurar que la información recogida del abonado, se ajuste a la situación real.

Automatización del control de calidad: Implementar controles automáticos para detectar y corregir errores en tiempo real, evitando que las inconsistencias afecten a futuros análisis. Con scripts de validación en los procesos de carga de datos, podríamos corregirlo.

7.2 Incremento de volumen de datos y variables

Recopilación de más datos: Es necesario aumentar el volumen de datos para una mejor capacidad predictiva del modelo. Esto implica una recopilación de información más detallada de los usuarios como podrían ser las preferencias, uso de servicios específicos que ofrece el centro o objetivos personales.

7.3 Optimización y ajuste del modelo predictivo

Reentrenamiento del modelo: Mejorados los datos, se debe re entrenar el modelo utilizando la nueva información recogida y validada. Se evaluarán nuevamente los modelos utilizados ajustando los parámetros según las nuevas métricas.

7.4 Mejoras tecnológicas

Automatización de procesos: Implementar pipelines automáticos para la extracción, transformación y carga de datos, así como para el reentrenamiento y monitoreo del modelo predictivo. Esto nos reduciría la carga operativa garantizando que el sistema se actualice de manera más eficiente y periódica.

7.5 Monitoreo y análisis de impacto

Análisis continuo: Realizar un seguimiento del impacto del modelo en la retención de clientes, comparando las predicciones con la realidad. Permitiendo así, ajustar futuras estrategias.

Dashboards de Análisis: Creación de dashboards interactivos, para que los equipos de Marketing y Comercial, puedan interpretar de manera clara los resultados del modelo, ayudando así a la toma de decisiones a tiempo real.

Encuestas a los abonados que se dan de baja: se implementará un sistema de encuestas dirigido a los abonados que decidan darse de baja. Este paso, permitirá obtener información de primera mano sobre los motivos reales de la baja, proporcionando información importante para desarrollar estrategias de retención más efectivas en un futuro.

7.6 Implementación de estrategias personalizadas de fidelización

Con el modelo mejorado, los equipos de marketing y Comercial, deberán implementar estrategias personalizadas para retener a los abonados identificados con alto riesgo de baja. Esto incluirá tanto campañas de marketing específicas, ofertar en los servicios personalizados o contacto directo con los mismos.

Estos próximos pasos a seguir, han sido pensados en base a las distintas dificultades que se han encontrado a lo largo del proyecto. Con ellos, se pretende mejorar la retención de abonados y el rendimiento del negocio a medio y largo plazo.

8. Conclusión

La conclusión del proyecto “Machine Tr(AI)ning. Churn Rate Model”, habla sobre los logros y desafíos encontrados durante su desarrollo. Valoramos en primer lugar, **de manera favorable**, esta primera prueba de concepto realizada, logrando desarrollar un modelo predictivo capaz de anticipar aquellos abonados del ********** que tienen mayor probabilidad de darse de baja. Este primer paso, nos permite generar un listado mensual de clientes con riesgo de abandono, lo que facilitará a los equipos del centro, la implementación de estrategias personalizadas para mejorar la retención de los usuarios.

Desde el punto de vista de la **viabilidad**, la implementación progresiva del proyecto **es factible** tanto a nivel económico, operativo y computacional. La infraestructura actual del centro, junto con las herramientas tecnológicas utilizadas, permiten que la adopción del modelo sea gradual sin generar una carga excesiva tanto en términos financieros como operativos. Esto nos asegura que el modelo pueda ser integrado de forma eficiente en los procesos del centro, adaptándose a sus capacidades sin afectar a su rendimiento.

Sin embargo, a pesar de los logros conseguidos, se han identificado áreas clave donde el modelo **tiene potencial de mejora**. Durante su desarrollo, surgieron varios inconvenientes relacionados con la calidad de los datos, tales como bajas mal registradas, altas duplicadas, y otros errores similares que afectaron a las predicciones. Además, la limitación de volumen de datos y la falta de variables adicionales con valor para la predicción, como el uso de servicios específicos del centro (piscina, clases guiadas, etc.), también influyeron en el rendimiento del modelo. Ampliar el volumen de datos disponibles, será crucial para mejorar la precisión del modelo y obtener resultados más concisos. Aumentar el número de registros y la calidad de los datos, permitirá generar predicciones más confiables para ayudar a la toma de decisiones.

Por último, a pesar de los retos actuales, consideramos que es **factible alcanzar el objetivo final del proyecto**, que consiste en reducir la tasa de abandono del centro en un 10% a finales de 2025. Con las mejoras esperadas en la calidad de los datos, incorporación de variables adicionales, automatización de procesos y ajuste del modelo, el proyecto tiene un potencial significativo de lograr este objetivo fijado. A medida que se avance en la implementación y se realicen las correcciones necesarias, esperamos que el modelo se convierta además de una herramienta predictiva precisa, en un factor clave para la estrategia de retención de abonados del centro.

9. Bibliografía

Referencias de mercado:

1. Fitness KPI. (https://new.fitness-kpi.com/center)

2. T. Innova. (https://t-innova.na4.teamsupport.com/knowledgeBase)

Referencias para el desarrollo del modelo:

1. Brownie, J. (2020-07-21). How to Fix k-Fold Cross-Validation for Imbalanced Classification. *Machine Learning Mastery*. https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/

2. Cochrane, C. (2018-05-19). Time Series Nested Cross-Validation. *Medium.* https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

3. Muskin, I. Time-series grouped cross-validation. *Data Science.* https://datascience.stackexchange.com/questions/77684/time-series-grouped-cross-validation

4. Shrivastava, S (2020, 14 de Enero). Cross Validation in Time Series. *Medium.* https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4#:~:text=Cross%20Validation%20 on%20 Time%20 Series,for%20the%20forecasted%20data%20points