
MÁSTER DE MACHINE LEARNING
CON R SOFTWARE

# MODELO DE CLASIFICACIÓN PARA LA PREDICCIÓN DE LA ROTACIÓN DE CLIENTES EN UNA OPERADORA DE TELEFONÍA

AUTOR: JOSÉ LUIS LORÉN CABRERIZO

DIRECTOR: NACHO GARCÍA

FECHA: 26 DE JULIO DE 2021

ENTIDAD COLABORADORA

# RESUMEN

La rotación de clientes es uno de los grandes problemas a los que se enfrenta toda empresa y la habilidad para identificar aquellos clientes con una alta probabilidad abandonar la compañía es clave a la hora de optimizar el ROI de las actividades destinadas a retener esos clientes.

En el presente Trabajo Fin de Máster se ha procedido a la generación de modelos de predicción de clasificación de diversos tipos así, a la selección de los mejores modelos en función del AUC, así como a la explicación de la mejora que supondría para la empresa la utilización del modelo con mejor resultado.

Para el modelado se ha utilizado AutoML de h2o que genera gran cantidad de modelos de predicción de diversos tipos (GBM, Stacked Ensembled, árboles de decisión, etc.) de una manera muy sencilla y con apenas unas pocas líneas de código.

Todo el ejercicio se ha realizado en R Software y para el mismo se ha utilizado un conjunto de datos que contiene información acerca de los clientes de una operadora de telefonía.

Como se explica al final del ejercicio, con el 50% de los clientes seleccionados por el modelo, estaríamos abarcando el 80% de los clientes cuyas variables poseen valores que son compatibles con la rotación.

Palabras clave:

- Machine Learning
- Aprendizaje automático
- Churn
- Rotación de clientes
- AutoML
- R Software

# AGRADECIMIENTOS

Quiero agradecer a Rosana Ferrero y Nacho García todo el apoyo, orientaciones y sugerencias durante todos los meses del máster. Su ayuda ha sido indispensable tanto en el proceso de aprendizaje como en la elaboración del presente trabajo.

También quiero agradecer a Máxima Formación por elaborar un máster que me ha permitido adquirir unos conocimientos que desde el primer día he podido poner en práctica en mi trabajo.

No quiero acabar sin antes agradecer a mi esposa, Pilar Broch Mesado, por haber sabido respetar todos los fines de semana y festivos que tuve que dedicar al máster.

# INDICE

# INTRODUCCION

Es bien sabido, que uno de los principales problemas con los que se encuentran en la actualidad las operadoras de telefonía e internet es la alta rotación de clientes que existe así como y el continuo transvase que se produce entre unas y otras.

Todas las operadoras tienen diseñadas estrategias destinadas a mejorar la fidelidad de los clientes para disminuir su rotación y, si las estrategias de captación funcionan conforme se espera, aumentar la base de clientes y por lo tanto la cuota de mercado.

Estas estrategias están compuestas por diversas acciones como son campañas de branding pero principalmente basadas en promociones y descuentos que tienen como objetivo evitar que un cliente decida cambiar de compañía por motivos de precio.

Sin embargo, hay que tener presente que esas estrategias tienen un elevado coste entre el que cabe destacar la pérdida de ingresos causada por esas promociones y descuentos.

Por consecuencia, y para buscar la maximización del ROI, una de las actividades que se diría crítica es el dirigir todos los esfuerzos a los clientes que tienen una mayor probabilidad de cancelar el contrato.

Como se verá en el siguiente apartado, nos encontramos ante un problema en el que se deberá de clasificar clientes en dos grupos:

- Grupo A: Clientes que no se espera que vayan a dejar de serlo y
- Grupo B: Clientes que se espera que puedan dejar de serlo y que serán el foco y objetivo de las acciones comerciales y de márquetin comentadas anteriormente.

# MATERIAL Y METODOS

Como se indicó en el apartado anterior, nos encontramos ante un problema de clasificación para cuya resolución se procedió a aplicar los conocimientos en machine learning adquiridos durante el máster en cuanto a la importación, exploración y transformación de características o variables.

Sin embargo, a la hora de seleccionar el tipo de algoritmo, y con la finalidad de aprovechar la oportunidad para adquirir nuevos conocimientos, se optó por aprender a utilizar el paquete h2o (https://rdrr.io/cran/h2o/) que se define como "*una plataforma de aprendizaje automático de código abierto que permite implementaciones paralelizadas de diferentes algoritmos supervisados y no-supervisados como Generalized Linear Models (GLM), Gradient Boosting Machines, Random Forests, Deep Neural Networks (Deep Learning), Stacked Ensembles, Generalized Additive Models (GAM), etc. así como un algoritmo de aprendizaje automático totalmente automatizado llamado H2O AutoML*".

Además de h2o, fue necesario familiarizarse con otros paquetes de R como por ejemplo:

- ggpubr, cowplot y tidyquant para la visualización de gráficos y tablas y su formato,
- recipes para el preprocesado e ingeniería de variables.

El ejercicio se realizó sobre un conjunto de datos de clientes de una operadora de telefonía que contenía una serie de variables predictoras o independientes así como una variable dependiente (Churn) que indicaba si el cliente había dejado de serlo.

En primer lugar se procedió con la importación de los datos y la transformación de las variables para posteriormente realizar un análisis exploratorio de los datos.

Para finalizar se generaron los diferentes modelos, se compararon entre sí y se procedió a elaborar un dashboard con las principales métricas de cada uno para, finalmente, generar un dashboard que comparaba el ROC, precisión vs recall, gain y lift de los 3 modelos con mejor AUC.

## Importación de datos y transformación de variables

Antes de nada, se procedió a cargar los paquetes que iban a ser necesarios a lo largo del ejercicio `tidyverse, caret,ggpubr, funModeling, tidyquant, h2o,skimr, recipes, cowplot.`

Se realizó la importación del set de datos, aprovechando para eliminar los espacios de los nombres de las variables. Asimismo, se procedió a visualizar el set de datos importado.

El set de datos contenía las siguientes variables:

- CustomerID: Identificación del cliente. Esta columna se eliminará puesto que no será de ninguna utilidad.
- Gender: Género del cliente.
- Senior_citizen: Codifica si el cliente es senior.
- Partner: Codifica si el cliente tiene pareja.
- Dependents: Codifica la existencia de dependientes del cliente.
- Tenure: Duración del contrato en meses.
- Phone_service: Codifica si el cliente tiene contratado el servicio de telefonía.
- Multiple_lines: Codifica si el cliente tiene contratadas varias líneas.
- Internet_service: Codifica si el cliente tiene contratado servicio de internet.
- Online_security: Codifica si el cliente tiene contratado el servicio de seguridad online.
- Online_backup: Codifica si el cliente tiene contratado el servicio de copia de seguridad online.
- Device_protection: Codifica si el cliente tiene contratado el servicio de protección de dispositivos.
- Tech_support: Codifica si el cliente tiene contratado el servicio de soporte técnico.
- Streaming_tv: Codifica si el cliente tiene contratado el servicio de TV a la carta.
- Streaming_movies: Codifica si el cliente tiene contratado el servicio de películas en streaming.
- Contract: Codifica el tipo de servicio que el cliente tiene contratado.
- Paperless_billing: Codifica si el cliente desea factura electrónica o no.
- Payment_method: Codifica el método de pago.
- Monthly_charges: Cargo mensual.
- Total_charges: Total de cargos aplicados al cliente.
- Churn: Codifica si el cliente canceló su contrato o si por el contrario sigue siendo cliente de la empresa.

El siguiente paso fue eliminar la columna con el ID de cliente, así como transformar numerosas columnas con variables categóricas en vector.Análisis exploratorio

Una vez realizada la importación de los datos, se procedió a visualizar un resumen de las estadísticas básicas del conjunto.

En estas estadísticas básicas, se observó lo siguiente:

- La variable dependiente Churn presentaba un cierto desbalanceo si bien éste no aparecía excesivamente preocupante.
- Las variables Gender y Partner presentaban una distribución en función de la variable dependiente muy similar.
- En la variable Total_charges había un total de 9 casos faltantes que habría que solventar.
- Las variables numéricas tienen distintos rangos, por lo que se deberán de centrar y escalar si bien esto es un proceso que el algoritmo AutoML realiza sin necesidad de intervención del usuario.
- Las variables Online_security, Online_backup, Device_protection, Tech_support, Streaming_tv and Streaming_movies presentaban una distribución muy parecida con pequeñas variaciones. Esto se tendría en cuenta más adelante.

Dado que la variable dependiente era de tipo factor, sabíamos que nos encontrábamos ante un problema de clasificación así que se procedió a estudiar su distribución, así como la del resto de variables.

Se comenzó por la visualización de la variable dependiente Churn para visualmente comprobar la distribución de la misma.

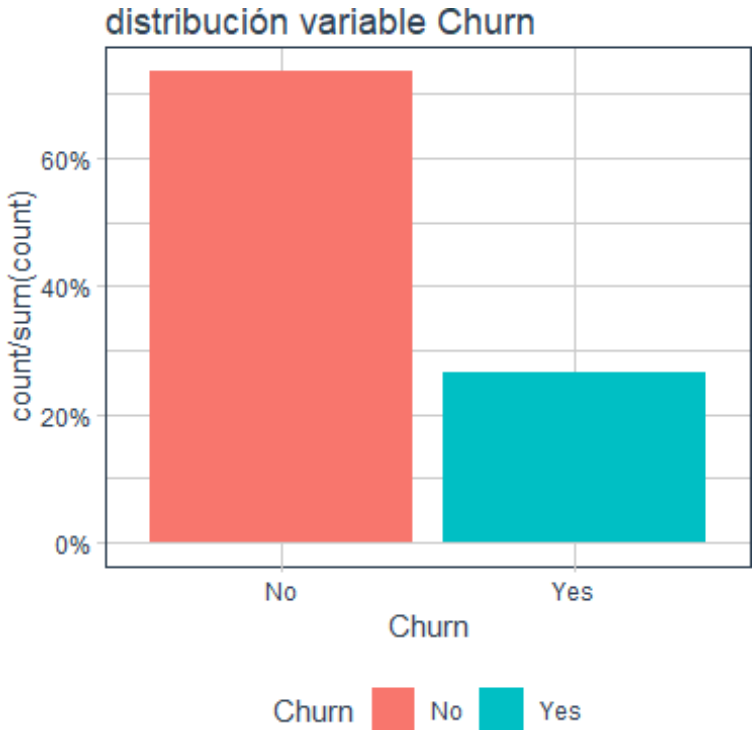


Fig 1. Proporciones de la variable Churn

En el gráfico se observó que la clase **No** era la mayoritaria (73.6%) mientras que tan sólo el 26.4% de los casos eran de la clase **Yes**.

Después se procedió a visualizar las distintas variables en función de la variable dependiente, empezando por las variables numéricas.

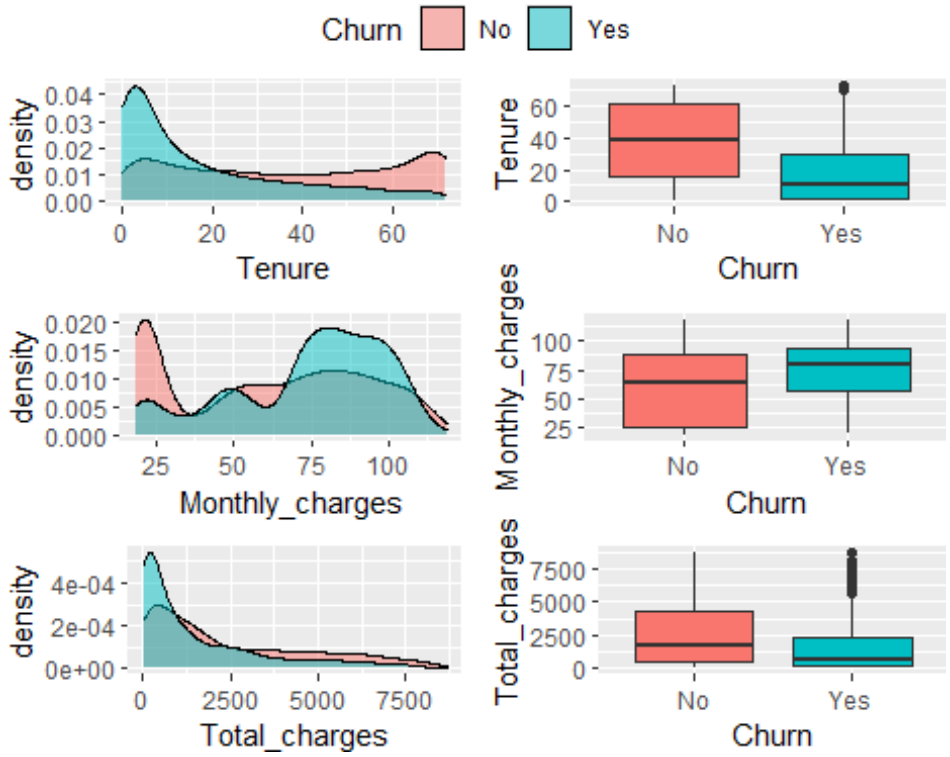


Fig 2. Visualización densidad y distribución de Tenure, Monthly_charges y Total_charges en función de la variable dependiente Churn

En las visualizaciones se pudo observar que la mayor rotación se observaba en clientes con una corta permanencia (Tenure). Consecuentemente, y como era de esperar, esta mayor rotación era mucho más pronunciada en el caso de cargos totales (Total_charges) reducidos ya que corta permanencia implica, en la mayor parte de los casos, poco gasto total.

Comprobando los datos de rotación en función de los cargos mensuales (Monthly_charges), se observa que ésta es notablemente superior cuando los cargos mensuales se encontraron entre 75 y 100 USD.

En los gráficos de densidad y cajas se pudo apreciar que las tres variables parecían adecuadas para el modelo, siendo Total_charges la única que presentaba una distribución de la densidad similar para ambos tipos de clientes (Churn y no Churn).

Asimismo, se pudo apreciar la existencia de outliers en las variables Tenure y Total_charges en el grupo de clientes que rotaron.

Se procedió a continuación a calcular las tablas de proporciones de las variables tipo factor, en función de si había habido rotación del cliente o no.

En las matrices de proporciones no se observó ninguna variable independiente con proporciones similares a la variable de respuesta aunque por el contrario sí que se detectaron varias variables con una distribución similar. Esta similitud destacaba sobre todo en el caso de clientes que cancelaron su contrato.

Estas eran las variables:

- Device_protection y Online_backup
- Online_security y Tech_support
- Streaming_tv y Streaming_movies

De estos pares de variables, se procedió a eliminar una de cada par ya que no aportaba ninguna información adicional para el modelo y por lo tanto no eran necesarias.

Asimismo, se procedió a eliminar las observaciones con datos faltantes. Se optó por la eliminación ya que estas observaciones suponían un porcentaje muy limitado del total de las observaciones (0.13%). En el caso de que hubieran supuesto un porcentaje mayor, se hubiera optado por alguna imputación (media, mediana, knn, etc.), en cuyo caso primero se hubiera dividido el conjunto de datos y después se hubiera hecho la imputación.

## Modelado

El siguiente paso, y antes del modelado, se dividió el conjunto de datos en entrenamiento y validación, en una proporción de 90% / 10%. Se comprobó el resultado de la división.

*Data summary*

| | |
|---|---|
| Name | train |
| Number of rows | 5850 |
| Number of columns | 20 |
| Column type frequency: | |
| factor | 17 |
| numeric | 3 |
| Group variables | None |

**Variable type: factor**

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---|---|---|---|---|---|
| Gender | 0 | 1 | FALSE | 2 | 1: 2972, 0: 2878 |
| Senior_citizen | 0 | 1 | FALSE | 2 | 0: 4896, 1: 954 |
| Partner | 0 | 1 | FALSE | 2 | No: 3023, Yes: 2827 |
| Dependents | 0 | 1 | FALSE | 2 | No: 4107, Yes: 1743 |
| Phone_service | 0 | 1 | FALSE | 2 | Yes: 5298, No: 552 |
| Multiple_lines | 0 | 1 | FALSE | 3 | No: 2824, Yes: 2474, No : 552 |
| Internet_service | 0 | 1 | FALSE | 3 | Fib: 2579, DSL: 2005, No: 1266 |
| Online_security | 0 | 1 | FALSE | 3 | No: 2899, Yes: 1685, No : 1266 |
| Online_backup | 0 | 1 | FALSE | 3 | No: 2586, Yes: 1998, No : 1266 |
| Device_protection | 0 | 1 | FALSE | 3 | No: 2568, Yes: 2016, No : 1266 |
| Tech_support | 0 | 1 | FALSE | 3 | No: 2904, Yes: 1680, No : 1266 |
| Streaming_tv | 0 | 1 | FALSE | 3 | No: 2360, Yes: 2224, No : 1266 |
| Streaming_movies | 0 | 1 | FALSE | 3 | No: 2305, Yes: 2279, No : 1266 |
| Contract | 0 | 1 | FALSE | 3 | Mon: 3221, Two: 1395, One: 1234 |
| Paperless_billing | 0 | 1 | FALSE | 2 | Yes: 3460, No: 2390 |
| Payment_method | 0 | 1 | FALSE | 4 | Ele: 1957, Mai: 1342, Ban: 1285, Cre: 1266 |
| Churn | 0 | 1 | FALSE | 2 | No: 4306, Yes: 1544 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| Tenure | 0 | 1 | 32.29 | 24.61 | 0.00 | 9.00 | 29.00 | 55.00 | 72.00 |
| Monthly_charges | 0 | 1 | 64.82 | 30.08 | 18.25 | 35.65 | 70.30 | 89.94 | 118.75 |
| Total_charges | 7 | 1 | 2283.27 | 2274.75 | 18.80 | 390.17 | 1391.65 | 3809.38 | 8684.80 |

Tabla 1. Estructura y composición del conjunto de entrenamiento

*Data summary*

| | |
|---|---|
| Name | test |
| Number of rows | 649 |
| Number of columns | 20 |

Column type frequency:

| | |
|---|---|
| factor | 17 |
| numeric | 3 |

| | |
|---|---|
| Group variables | None |

**Variable type: factor**

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---|---|---|---|---|---|
| Gender | 0 | 1 | FALSE | 2 | 0: 331, 1: 318 |
| Senior_citizen | 0 | 1 | FALSE | 2 | 0: 547, 1: 102 |
| Partner | 0 | 1 | FALSE | 2 | No: 336, Yes: 313 |
| Dependents | 0 | 1 | FALSE | 2 | No: 454, Yes: 195 |
| Phone_service | 0 | 1 | FALSE | 2 | Yes: 586, No: 63 |
| Multiple_lines | 0 | 1 | FALSE | 3 | No: 314, Yes: 272, No : 63 |
| Internet_service | 0 | 1 | FALSE | 3 | Fib: 281, DSL: 212, No: 156 |
| Online_security | 0 | 1 | FALSE | 3 | No: 309, Yes: 184, No : 156 |
| Online_backup | 0 | 1 | FALSE | 3 | No: 269, Yes: 224, No : 156 |
| Device_protection | 0 | 1 | FALSE | 3 | No: 275, Yes: 218, No : 156 |
| Tech_support | 0 | 1 | FALSE | 3 | No: 305, Yes: 188, No : 156 |
| Streaming_tv | 0 | 1 | FALSE | 3 | Yes: 264, No: 229, No : 156 |
| Streaming_movies | 0 | 1 | FALSE | 3 | No: 250, Yes: 243, No : 156 |
| Contract | 0 | 1 | FALSE | 3 | Mon: 355, Two: 170, One: 124 |
| Paperless_billing | 0 | 1 | FALSE | 2 | Yes: 376, No: 273 |
| Payment_method | 0 | 1 | FALSE | 4 | Ele: 225, Mai: 143, Ban: 141, Cre: 140 |
| Churn | 0 | 1 | FALSE | 2 | No: 478, Yes: 171 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| Tenure | 0 | 1 | 33.05 | 24.38 | 0.0 | 10.00 | 30.00 | 56.00 | 72.00 |
| Monthly_charges | 0 | 1 | 63.99 | 30.66 | 18.9 | 30.25 | 70.60 | 88.95 | 116.45 |
| Total_charges | 2 | 1 | 2279.93 | 2228.75 | 18.9 | 457.62 | 1423.05 | 3641.65 | 8476.50 |

Tabla 2. Estructura y composición del conjunto de validación

```
# Iniciación del clúster de h2o
h2o.init()

## Connection successful!
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         37 minutes 21 seconds
##     H2O cluster timezone:       Europe/Berlin
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.32.1.2
##     H2O cluster version age:    2 months and 24 days
##     H2O cluster name:           H2O_started_from_R_josel_dcb227
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   3.81 GB
##     H2O cluster total cores:    8
##     H2O cluster allowed cores:  8
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     H2O API Extensions:         Amazon S3, Algos, AutoML, Core V3, TargetEncoder,
Core V4
##     R Version:                  R version 4.0.5 (2021-03-31)
```

Se pudo comprobar que el set de datos de entrenamiento estaba compuesto por 5.842 mientras que el conjunto de validación estaba compuesto por 648 observaciones. Ambos sets de datos contenían 17 variables de las cuales 14 eran del tipo factor y sólo 3 numéricas, no habiendo ningún caso faltante.

En cuanto al modelado en sí, tal y como se anticipó al inicio de este apartado, se optó por la utilización del algoritmo AutoML de la librería h2o. Este algoritmo es un algoritmo de bajo mantenimientos al que le basta con disponer de las variables en el formato correcto, ncargándose del resto de las transformaciones necesarias él mismo. Otra ventaja muy importante, es que este algoritmo permite mantener los datos originales en un formato legible para cualquier ser humano.

El paquete recipes de tidymodels es un paquete que permite la creación de una secuencia con las diferentes transformaciones o preprocesado que se desean realizar para después aplicar esa secuencia a los sets de datos que se deseen. Para ello, el paquete dispone de una serie de steps que permiten desde realizar imputaciones de datos faltantes hasta centrar y escalar, pasando por diferentes transformaciones. Mas información en https://www.rdocumentation.org/packages/recipes/versions/0.1.16.

Una vez llegados a este punto, se procedió a crear una receta con el paquete recipes del framework tidymodels para eliminar las variables con una varianza igual o próxima a cero.

Y a continuación se continuó con los pasos comentados en el siguiente bloque de código:

Después se procedió a crear los modelos y a guardarlos para tenerlos disponibles a lo largo del ejercicio. Con el objeto de hacerlo reproducible, debimos de:

- Limitar el número de modelos: por motivos de coste computacional, el número de modelos se limitó a 10.
- Fijar una semilla.
- Excluir los modelos de Deep Learning ya que por el sistema de cálculo éstos no son reproducibles.
- No fijar el tiempo máximo de ejecución para que el algoritmo fuera capaz de finalizar el proceso de manera completa.

*Referencia:* *http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html?highlight=reproducibility*

Una vez generados los modelos, se procedió a comprobar las principales métricas de los 10modelos que presentaron un mejor AUC.

Fig 3. Dashboard con principales métricas para cada uno de los modelos generados en el paso anterior.

En la figura anterior se puede observar que todos los modelos presentan unos valores muy similares para todas y cada una de las diferentes métricas mostradas.

Con estos datos, se procedió a crear un dashboard para evaluar el rendimiento de los 3 modelos con mejor AUC que incluía las siguientes métricas:

- ROC
- Precision vs Recall
- Gain
- Lift

La elaboración de ese dashboard se realizó en diversos pasos, como se puede observar a continuación:

En primer lugar, se extrajeron los 3 mejores modelos (mejor AUC) y guardaremos cada unoen una variable.

Para posteriormente proceder a extraer las métricas de rendimiento de cada uno de los modelos a visualizar.

A continuación, se generaron los gráficos ROC y Precision vs Recall y se guardaron en dos variables para usarlas posteriormente a la hora de graficar.

Para poder generar los gráficos de Gain y Lift se tuvo que generar primero las tablas que contenían los datos necesarios.

Una vez generados todos los gráficos, se pudo proceder a combinar los 4 en uno para obtener el dashboard que nos iba a ayudar con la interpretación de los resultados así como con una posible explicación a personas no versadas en aprendizaje automático y sus conceptos.

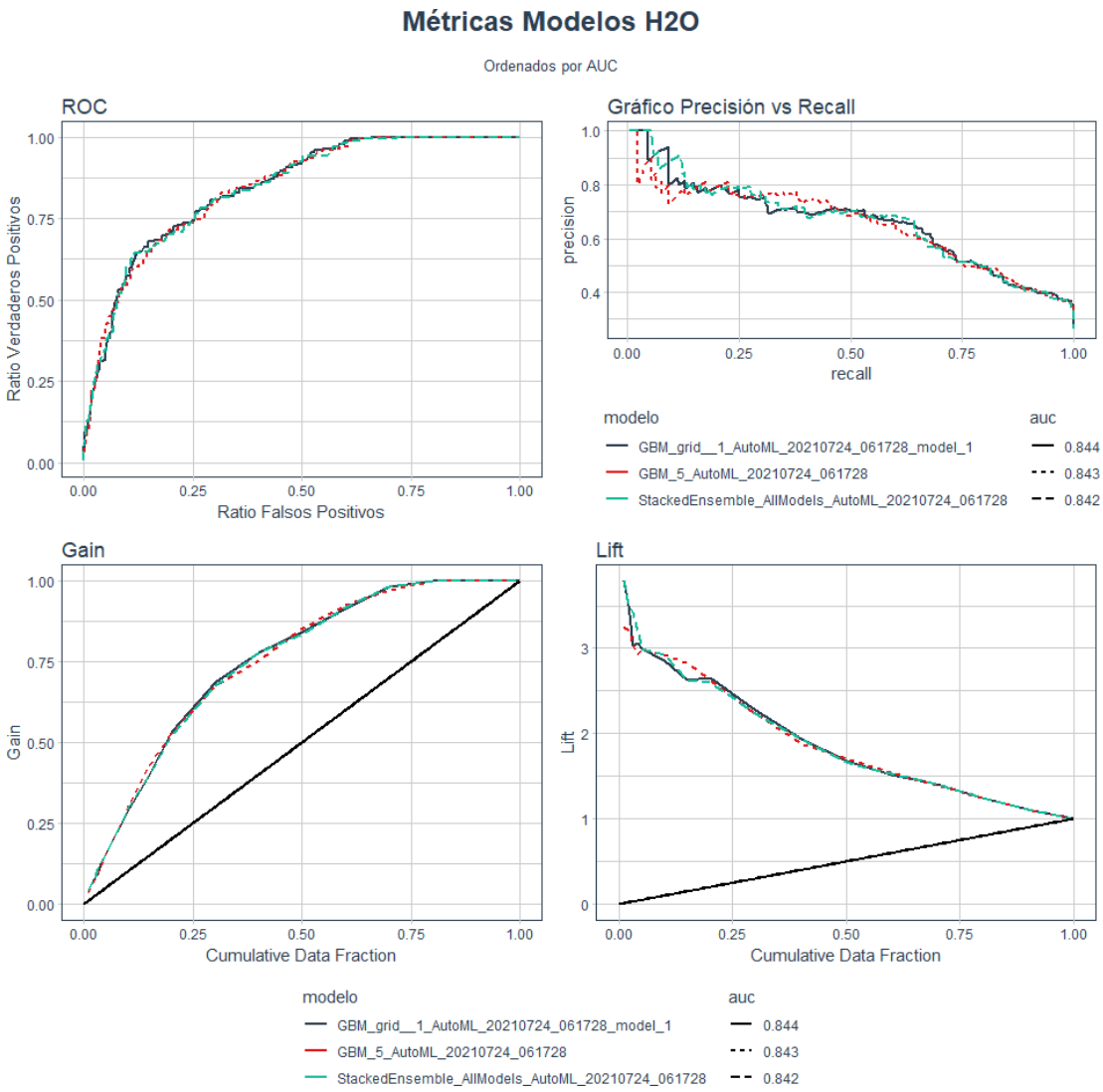


Fig 3. Dashboard mostradno ROC, Precisión vs Recall, Gain y Lift de los 3 modelos con mejor AUC

# RESULTADOS

Como ya se podía esperar de los resultados del gráfico de principales métricas por modelo, las diferencias entre los 3 mejores modelos generados son casi nulas y los 3 modelos presentan valores muy similares en cuanto al ROC, precisión vs recall, gain y lift.

No obstante, en un caso real no basta con todo lo presentado anteriormente ya que falta la parte más importante y que tiene que ver con la explicación que se da a ejecutivos y otros miembros de la empresa de los resultados obtenidos.

Considerando que las mencionadas personas no tienen por qué saber de algoritmos, estadística se debe de hacer un esfuerzo adicional en explicar las razones por las que, por ejemplo en este caso, se debería de utilizar un modelo para seleccionar los clientes a los que se deberían de dirigir las acciones comerciales para obtener un mayor retorno de la inversión.

En este caso, podríamos justificar la utilización de un modelo a la hora de seleccionar los clientes objeto de las posibles acciones de marketing simplemente fijándonos en los dos gráficos inferiores.

De estos gráficos podemos deducir, por ejemplo, que seleccionando el 25% de los clientes de manera aleatoria (como se estaba haciendo hasta el momento), alcanzaríamos alrededor de un 25% de clientes con alta probabilidad de cancelar el contrato.

Sin embargo, usando el modelo seleccionado, el porcentaje que podríamos alcanzar rondaría el 60%. Del mismo mod, y siguiendo el eje de abscisas, seleccionado únicamente el 50% de los clientes usando el modelo, el porcentaje superaría el 80%.

El gráfico de Lift también se podría utilizar para explicar que el beneficio de usar el modelo multiplica el retorno, en el ejemplo del 25% de los clientes, por 2.3.

# DISCUSION

A lo largo del presente documento se ha dado solución a un problema de predicción binaria que se da en todas las empresas como es la previsión de la rotación de un cliente basándose en sus datos históricos.

A la hora de enfrentarse a un problema como el que nos ocupa, se puede optar por diferentes caminos si bien un algoritmo que requiere de poco mantenimiento y cuyo requerimientos iniciales son pocos como AutoML de h2o puede ser una opción muy interesante para muchas empresas.

En el ejemplo que se ha trabajado, se ha podido comprobar como en unas pocas líneas de código, AutoML ha sido capaz de comparar entre diferentes modelos usando diversos algoritmos para ofrecer un resultado más que aceptable.

Si bien está fuera del alcance del presente TRABAJO DE FIN DE MÁSTER, se podría mejorar la predictibilidad del modelo final mediante ingeniería de variables y/o la modificación de parámetros.

# REFERENCIAS/BIBLIOGRAFÍA


- Material del Master en Machine Learning con R Software
- h2o: https://rdrr.io/cran/h2o/
- Ayuda AutoML de H2o: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html
- Recipes: https://www.rdocumentation.org/packages/recipes/versions/0.1.16
- Tidyquant: https://www.rdocumentation.org/packages/tidyquant/versions/1.0.3
- Ggpubr: https://www.rdocumentation.org/packages/ggpubr/versions/0.4.0
- Cowplot: https://www.rdocumentation.org/packages/cowplot/versions/1.1.1