# Predicción en la deserción de clientes

Rodrigo Gomez Portillo | Roberth Hernandez
CODERHOUSE | 19/07/2022

# Contenido

Descripción del caso ..... 2
Objetivo del modelo ..... 2
Descripción de los datos ..... 2
Análisis de datos ..... 2
Algoritmo seleccionado ..... 5
Optimización ..... 6
Modelo final ..... 6
Conclusiones ..... 7

| Version del proyecto | Fecha de publicación |
|---|---|
| **Version 1** | 19/07/2022 |

## Descripción del caso

La empresa “Y” que vende su producto estrella “X” tiene una lista de los clientes que han usado la plataforma para realizar una compra de dicho producto. Algunos de sus clientes desertan de la plataforma de compras y es de interés para la empresa ser capaz de predecir la deserción de clientes de la manera más precisa posible con el fin de implementar estrategias de marketing para afianzar a dichos clientes.

La definición de la deserción se caracteriza por una columna en el dataset con un tipo de dato booleano que identifica si el cliente ha desertado en algún momento de la plataforma.

## Objetivo del modelo

El modelo debe ser capaz de identificar correctamente tanto a aquellos clientes que abandonan la plataforma como los que no lo hacen. De esta manera se espera profundizar las acciones del equipo de marketing solamente sobre aquellos usuarios que son potenciales desertores. Se espera que el modelo supere el 70% de presicion y de recall con el fin de ser lo mas efectivos con las acciones de marketing.

## Descripción de los datos

Los datos que presenta la empresa “Y” contienen una lista de 10000 clientes. De este set de datos se utilizaron 11 columnas (incluyendo la columna de la variable a predecir). Dichos datos fueron pre-procesados antes de llegar al proceso de análisis por lo que el este set de datos no presenta datos nulos o faltantes. Las columnas presentes son:

**Score**: puntaje del cliente dentro de la plataforma de ventas.
**Nationality**: nacionalidad del cliente (Alemania, Francia y España).
**Gender**: género (femenino y masculino).
**Age**: edad del cliente.
**Tenure**: permanencia del cliente en la plataforma cuantificado en un número entre el 1 y el 10.
**Balance**: balance bancario del cliente.
**Products**: cantidad de productos comprados por el cliente.
**Card**: variable booleana que define si el cliente posee tarjeta de crédito.
**Active**: variable booleana que define si el cliente figura como activo dentro de la plataforma
**Salary**: salario del cliente.
**Exited**: variable target que muestra si el cliente se ha dado de baja de la plataforma.

Del set de datos original se desestimaron las variables Row (número de fila), Id (número identificatorio del cliente), Surname (nombre del cliente) ya que no se consideraron de importancia para resolver el problema .

Fuente de datos: kaggle

## Análisis de datos

En la Figura 1 se muestra la cantidad de personas que han abandonado y permanecido en la plataforma. Como se observa la variable target con la que se trabaja presenta un desbalance.

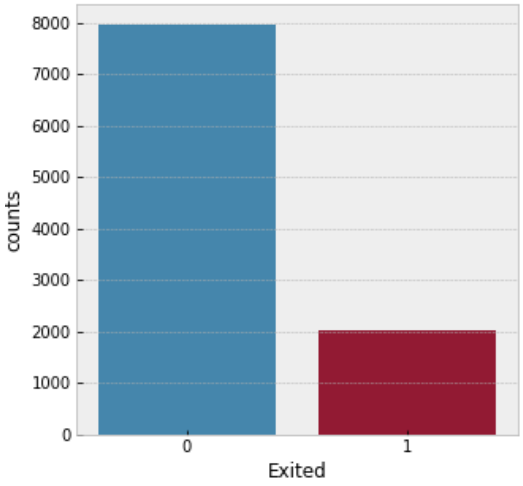


**Figura 1. Cantidad de personas que han abandonado la plataforma.**

En la Figura 2 se muestra como es la distribucion de edades de los clientes separados por si han abandonado (rojo) o no (azul) la plataforma. Como se observa las personas mayores tienen una mayor tendencia a abandonar la plataforma.

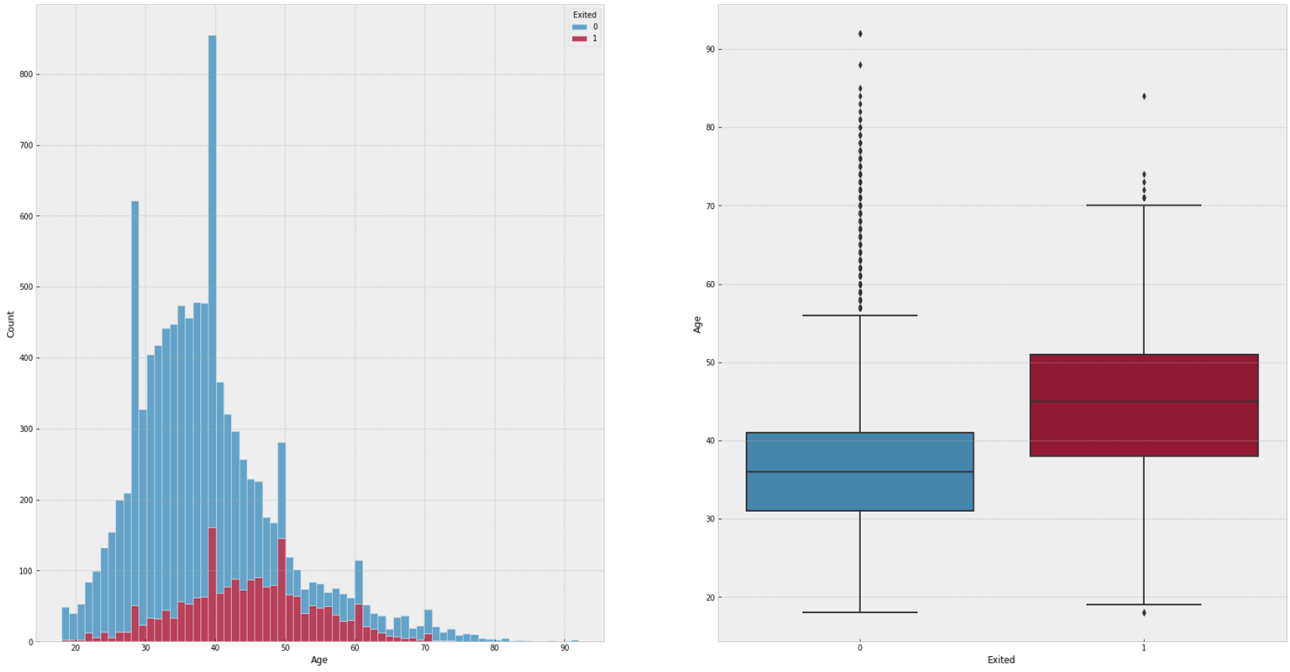


**Figura 2. (izquierda) gráfico de barass con la cantidad de personas que abandonan la plaforma en función de la edad y (derecha) un gráfico de caja que muestra la distribución de edades para aquellos que desertan y que no.**

En la Figura 3 se muestra el puntaje de cada clientes en función de la cantidad de productos que cada cliente ha comprado. Como se observa que ningun cliente ha concretado más de 4 productos y que aquellos que lo han hecho han abandonado efectivamente la plataforma. Tambien se ve que aquellos clientes que han realizado dos compras y tienen un score menos a 400 han abandonado la plataforma.

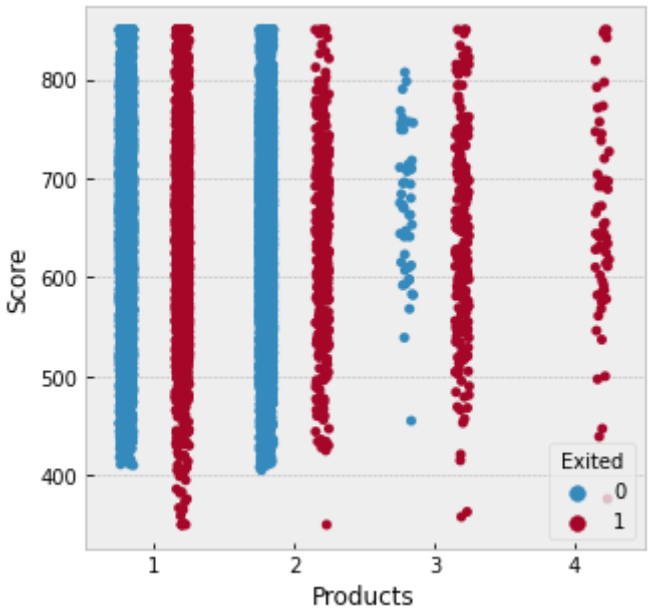


**Figura 3. Score vs. Productos**

En la Figura 4 se grafican las variables numéricas en paralelo. Como se observa hay muchos casos donde tanto el balance bancario como el salario es cero. Se asume que la empresa "Y" no cuenta con esta información sobre esos clientes. Por otro lado, se espera que haya alguna relación entre el balance bancario y el salario de los clientes, pero no es el caso.

También se puede ver que la edad vuelve a ser un factor importante a la hora de identificar a los usuarios que abandonan la plataforma.

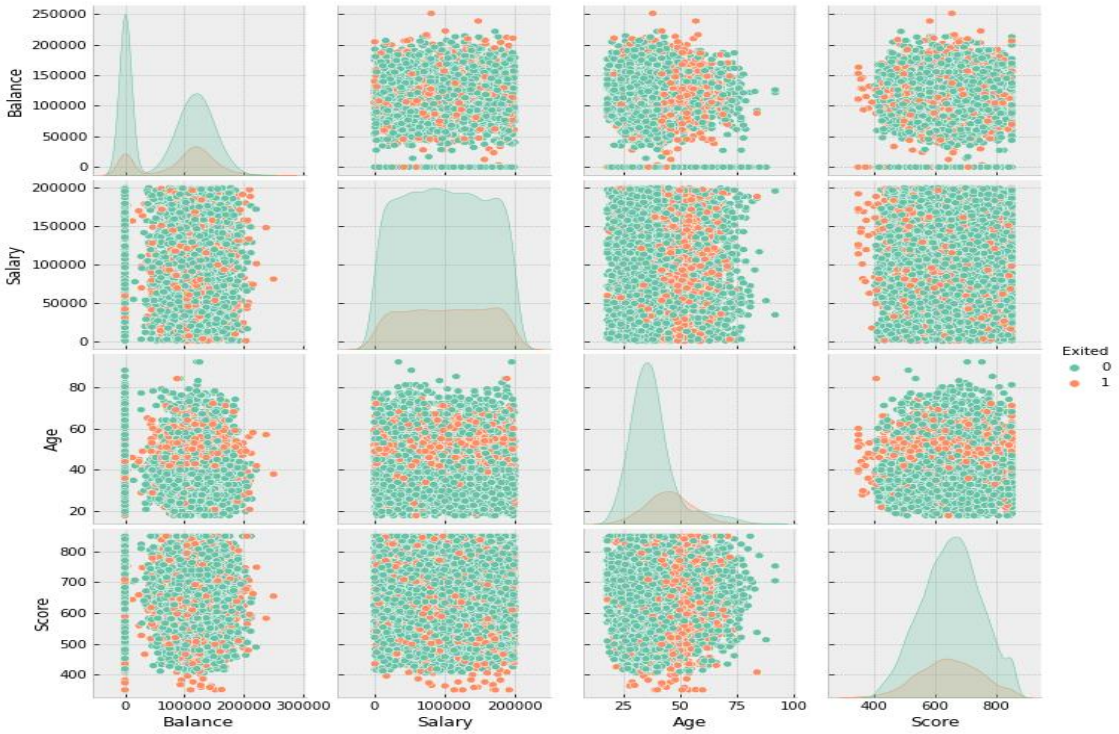


**Figura 4. Pair plot de las variables numéricas.**

También se realizó un análisis puntual sobre las variables género (gender), nacionalidad (nationality), salario (salary), balance (balance), permanencia (tenure), tarjeta (card) y activo (active). Sobre ninguna de las variables enumeradas se encontró una tendencia o una correlación que permita explicar el problema a priori.

Se realizaron análisis de correlación y componentes principales que no mostraron dependencia lineal entre las variables en estudio.

Tampoco se encontró una distribución estadística que se ajuste a ninguna de las variables.

## Algoritmo seleccionado

Se probó resolver el problema utilizando los siguientes algoritmos de clasificación:

1. Tree calssifier
2. Random Forest
3. Gradient Boosting
4. KNN
5. Logistic Regression

El que mejor se adaptó a los requerimientos, utilizando los parámetros por defecto, fue el de Gradient Boosting. En la Figura 5 se muestra la matriz de confusión resultante de la ejecución del modelo.

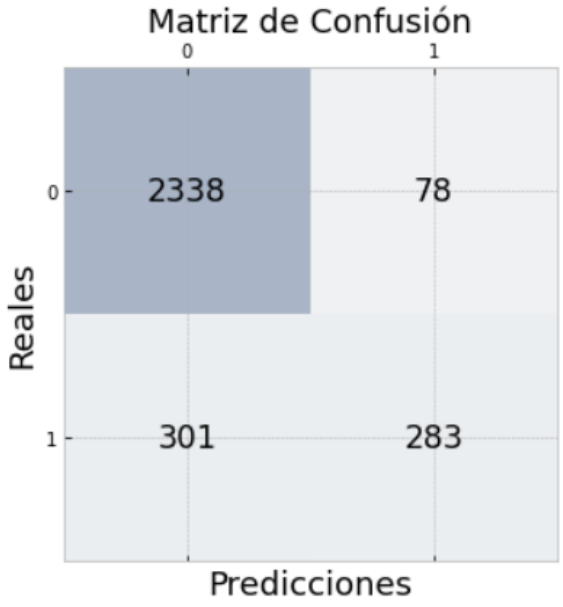


**Figura 5. Matriz de confusión de la ejecución de Gradient Boosting con los parámetros por defecto.**

En la Tabla 1 se muestra el reporte del modelo de clasificación. Como se observa se consigue una presición del 89%, un recall del 48y una exactitud total del 87%.

**Tabla 1. Parámetros resultantes del modelo.**

| | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.89 | 0.97 | 0.93 |
| 1 | 0.78 | 0.48 | 0.60 |
| macro avg | 0.83 | 0.73 | 0.76 |
| weighted avg | 0.87 | 0.87 | 0.86 |

La elección del algoritmo de clasificación se realizó en base a plantear un modelo económico en el cual se asumió que el producto “X” toma valores de venta entre 500 y 20000 dólares y que el costo de publicidad representa el 10%, 30% y 50 % del valor del producto. En base a esto se calculó la diferencias entre el modelo y los ingresos si no existiera tal modelo que ayude a identificar a los clientes que abandonan. En base a esto Gradient Boosting es el que mejores ingresos generó.

## Optimización

Con el fin de encontrar el mejor set de parámetros del modelo, se realizó un análisis utilizando GridSearchCV variando entre los siguientes parámetros del modelo:

1. learning_rate = entre 0,1 y 1,5
2. n_estimators = 100 y 200
3. subsample = de 0,5, 0,7 y 1
4. max_depth = de 2 a 6 con con incrementos de 1

A continuación se enumeran los distintos parametros utilizados en la ejecucion de GridSearchCV:

1. random_state = 42
2. scoring = accuracy, recall, balanced_accuracy y average_precision
3. cv = 5

## Modelo final

En la Figura 6 se muestra la matriz de confusión con los parametros optimizados. Como se ve no hay un cambio apreciable en los resultados.

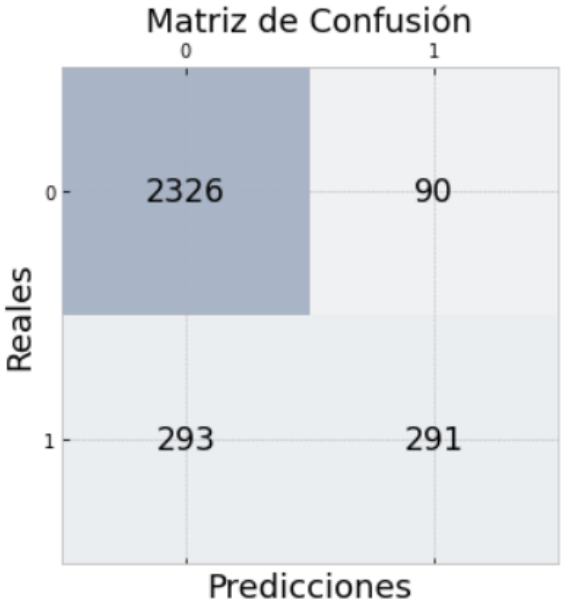


**Figura 6.Matriz de confusión de la ejecución de Gradient Boosting con los parámetros opmizados.**

En la Tabla 2 se muestran los datos resultantes del modelo.

**Tabla 2. Parámetros resultantes del modelo opmimizado.**

| | Precision | Recall | F1-score |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 0 | 0.89 | 0.96 | 0.92 |
| 1 | 0.76 | 0.50 | 0.60 |
| macro avg | 0.83 | 0.73 | 0.76 |
| weighted avg | 0.86 | 0.87 | 0.86 |

Finalmente los parametros utilizados en el modelo son:

1. learning_rate = 0,1
2. n_estimators = 200
3. subsample = 0,5
4. max_depth = 3

Teniendo este modelo, se realizó el mismo análisis con la matriz de confusión, pero con datos de entrenamiento. Con esto se logró ver un comportamiento similar al que tienen los datos de testing, lo que permite concluir que no hay sobreajuste por parte del modelo.

## Conclusiones

No se logró llegar al objetivo del modelo (precisión y recall mayor al 70%). Aún así es preferible hacer uso del mismo, ya que mostró ser ventajoso respecto a no tenerlo y generar acciones de marketing sobre toda la población de clientes.

En base a un modelo económico asumiendo valores para el producto “X” y la proporción de costo que se gasta en publicidad se determinó que Gradient Boosting es el mejor algoritmo de clasificación para resolver este problema.

La optimización no tuvo diferencias con respecto al modelo original.