# Introducción a la Analítica Empresarial - Proyecto Final

**Prof. Roberto Villar**

## Parte 1: Introducción

### 1.1 Contexto del problema

Un gerente de un banco está preocupado al tener el problema de que cada vez más clientes abandonan sus servicios de tarjetas de crédito. Necesita de alguien que pueda predecir quién se verá afectado para poder acudir de manera proactiva al cliente, brindarle mejores servicios y cambiar sus decisiones en la dirección opuesta.

Nos parece importante notar que a medida que los bancos compiten para obtener una ventaja competitiva, la necesidad de gestionar grandes cantidades de datos y análisis se hace más relevante. La ciencia de datos y el big data han cambiado la forma en que funcionaban los bancos tradicionales en el pasado y ha sido de gran ayuda para la toma de decisiones. A través de distintas herramientas de datos, los bancos pueden obtener una mayor visibilidad del comportamiento de los clientes y evaluar la probabilidad de riesgo.

### 1.2 El objetivo

Este proyecto se lleva a cabo en una secuencia de pasos, el primero de los cuales consiste en un análisis exploratorio, donde el objetivo es conocer el comportamiento de las variables y analizar atributos que indiquen una fuerte relación con la cancelación de los clientes del servicio de tarjeta de crédito. Al final del proyecto, luego de completar todos los pasos, se desarrollará un modelo de aprendizaje automático, capaz de predecir, en base a los datos de un sistema, si un cliente dejará o no el servicio de tarjeta de crédito.

### 1.3 Conjunto de datos

Este conjunto de datos consta de 10.000 clientes que mencionan su edad, salario, estado civil, límite de tarjeta de crédito, categoría de tarjeta de crédito, etc.

Solo tenemos un 16,07% de clientes que han cancelado. Por lo tanto, es un poco difícil entrenar nuestro modelo para predecir la rotación de clientes.

**Parte 2: Desarrollo**

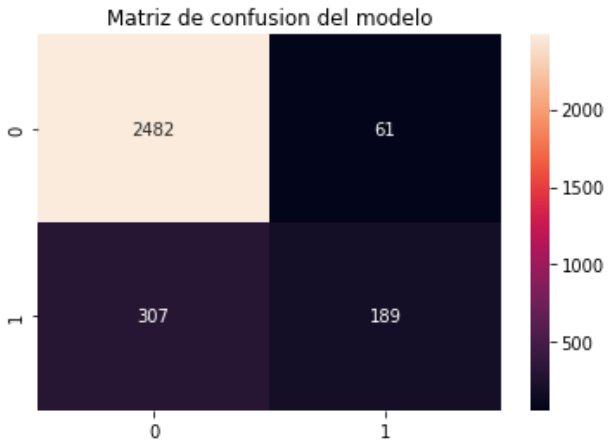
El modelo desarrollado para analizar los datos es **regresión logística**.

En primer lugar, estudiamos el tipo de valor que contiene cada columna, y eliminamos las columnas que no aportan valor al modelo. También verificamos los valores nulos del dataset, y al no tener ninguno no fue necesaria la limpieza de datos.

En segundo lugar, decidimos aplicar técnicas de "feature engineering" a las columnas categóricas, para poder trabajar con ellas adecuadamente más adelante.

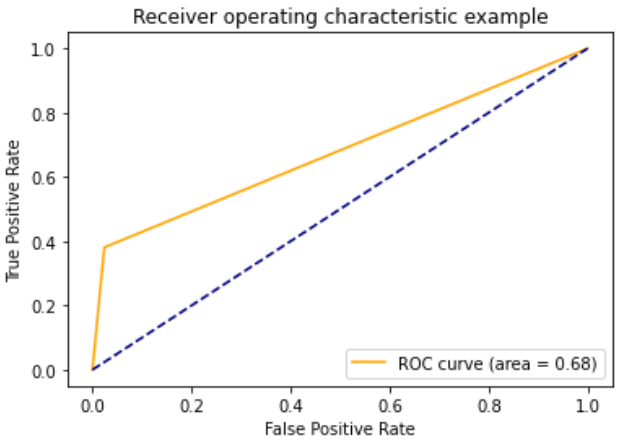
En tercer lugar, estudiamos la correlación entre las variables mediante un mapa de calor, y seleccionamos las variables con una correlación absoluta mayor a 0.01, para realizar el entrenamiento del modelo.

Sin embargo, decidimos primero entrenar el modelo con todas las variables obtenidas luego del procesamiento, para luego compararlo con el modelo que utiliza únicamente las variables que cumplen con el mínimo de correlación establecido. Luego de haber entrenado el modelo y obtenido las predicciones, medimos su precisión a través del "Accuracy Score", que dio como resultado un 87,89%. Además, creamos la matriz de confusión, de la que sacamos las siguientes conclusiones:



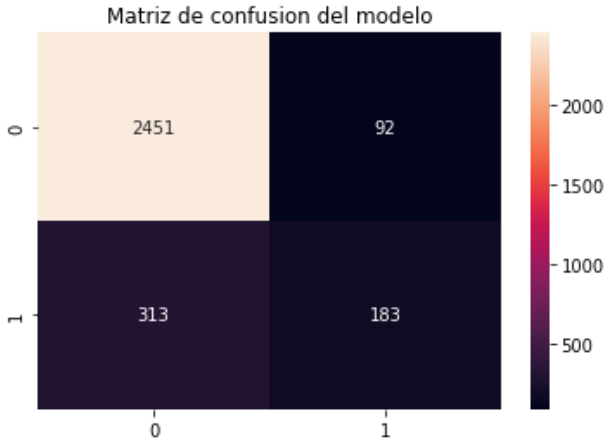

- 2482 verdaderos negativos: clientes que se quedaron en el banco y el modelo hizo la predicción correcta.
- 307 falsos negativos: clientes que abandonaron el banco pero el modelo predijo que se quedaban en el banco.
- 61 falsos positivos: clientes que se quedaron en el banco y el modelo predijo que lo abandonaron.
- 189 verdaderos positivos: clientes que abandonaron el banco y el modelo lo predijo de esa manera.

Por último, graficamos la curva ROC (Receiver Operating Characteristic), que muestra el rendimiento del modelo en todos los umbrales de clasificación. Esta curva representa dos parámetros: la tasa de verdaderos positivos y la tasa de falsos positivos. El área debajo de la curva proporciona una medida agregada del rendimiento en todos los umbrales de clasificación posibles. Una forma de interpretarla es la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio.




En este caso, el valor de esta área es 0,68. Este resultado es debido a que los datos se encuentran desbalanceados, es decir, tenemos más datos que corresponden a la clase de los clientes que se quedan en el banco (más de 8000 datos) y tenemos menos datos que corresponden a la clase que abandona el banco (menos de 2000 datos). Esto puede causar cierta "confusión" en el modelo, porque tiende a sesgarse en las predicciones, dando como resultado más predicciones donde el cliente se queda que en las que abandona.

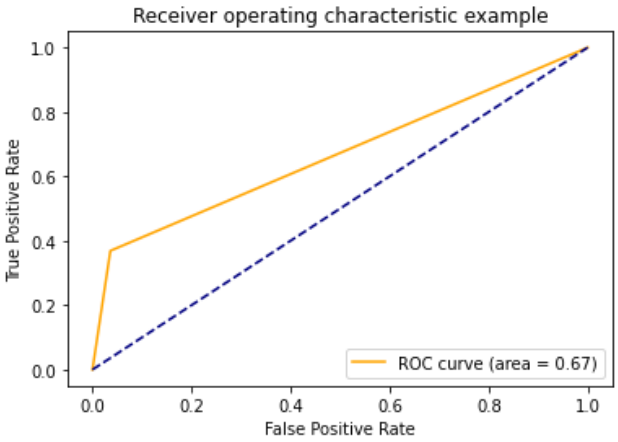
Repetimos todo el proceso con las variables que tenian una correlación mayor a 0,01. Este modelo resultó en un “Accuracy Score” de 86,67%, menor al anterior, y en la siguiente matriz de confusión:




- 2451 verdaderos negativos: clientes que se quedaron en el banco y el modelo hizo la predicción correcta.

- 313 falsos negativos: clientes que abandonaron el banco pero el modelo predijo que se quedaban en el banco.
- 92 falsos positivos: clientes que se quedaron en el banco y el modelo predijo que lo abandonaron.
- 183 verdaderos positivos: clientes que abandonaron el banco y el modelo lo predijo de esa manera.

En cuanto al área debajo de la curva ROC, también disminuyó a un 0,67, lo que significa que el modelo bajó en su desempeño, dando como resultado un incremento en los falsos positivos y falsos negativos.




**Parte 3: Conclusión**

Finalmente, al comparar el modelo con ambos grupos, se puede ver como el dataset completo sin filtrar presenta una mayor precisión, lo que evidentemente es beneficioso para predecir la cantidad de clientes que abandonan los servicios del banco.