

# Campaña de Márqueting Directo para Banca Minorista.

## Introducción

- Una institución de banca minorista desea revisar su estrategia de márketing directo. Ésta consiste en diversas campañas dirigidas a segmentos específicos de clientes para conseguir un objetivo concreto. Dichas campañas están centralizadas en un "contact center", con el que se contacta al cliente por diferentes canales, muy a menudo el teléfono. Se dispondrá de una serie de datos relativos a las campañas de ventas, que serán analizados para incrementar la eficacia de las mismas.

## Objetivos del Proyecto

### 1. ¿ Cuales son los objetivos del negocio ?

- La tecnología de "machine learning" permite mejorar el márketing utilizando el enfoque de **maximizar el valor del tiempo de permanencia** del cliente mediante la evaluación de los datos disponibles. Específicamente, el objetivo es **aumentar la contratación de depósitos a plazo** en las campañas de márketing directo. Se podría también optimizar el número de llamadas, puesto que se obtendría la probabilidad que un cliente determinado acepte un depósito a plazo fijo. Así pues, se podrían **centrar los esfuerzos en los clientes con mayor probabilidad** de éxito.

### 2. ¿ Que decisiones o procesos específicos quereis mejorar o automatizar con ML ?

- En particular, en el presente proyecto se desea **predecir el número de éxitos o fracasos de las llamadas** de telemárketing para la venta de los depósitos a plazo. Adicionalmente se podrá **optimizar el número de llamadas** a realizar para conseguir dicho objetivo.

### 3. ¿ Se podría resolver el problema de manera no automatizada ?

- Si bien el objetivo puede plantearse mediante herramientas estadísticas, la gran cantidad de datos disponibles hace aconsejable el uso de "machine learning". Por otro lado, los algoritmos "ML" permitirían agilizar la toma de decisiones relativas a las campañas de márketing.

## Metodología Propuesta

- Dado que se desea predecir el "éxito" o "fracaso" de una venta, se escogerá un modelo de clasificación. La **"regresión logística"** no solo predice (clasifica) el resultado de la venta, sino que además nos facilita la **probabilidad de dicho éxito**. Se escoge pues dicho modelo.

- Para los modelos de clasificación pueden utilizarse dos métodos, el análisis LIFT y la gráfica ROC, utilizaremos ambos métodos.

- El análisis LIFT consiste en dividir la probabilidad de éxito predicha (la probabilidad de contratar el depósito) en deciles, esto es en 10 grupos: 0.0-0.1, 0.1-0.2, ... , 0.9-1.0.

Dichos valores formarán el eje horizontal de un gráfico LIFT. Para el eje vertical se dividirán el número de éxitos en cada grupo, por el número de clientes de dicho grupo. En los grupos con poca probabilidad deberían encontrarse pocos éxitos (pocos contratos), en los grupos con mucha probabilidad deberían contarse muchos éxitos. Lo contrario indicaría que la capacidad de predicción del método es baja. La puntuación LIFT se define como:

$$\text{lift} = (\text{probabilidad predicha}) / (\text{probabilidad media})$$

El parámetro permite centrar los esfuerzos en los grupos con un LIFT mayor, puesto que identifica los grupos con el mayor número de éxitos predichos.

- La gráfica ROC se basa en la matriz de confusión, de la que se obtienen los parámetros:

TPR = TP / P, positivos auténticos entre total positivos.

FPR = FP / N, falsos positivos entre total negativos.

En nuestro caso la matriz de confusión se definirá como:

Positivos: Se predice que realizarán el contrato.

Negativos: Se predice que no realizarán el contrato.

Falsos Positivos: Se predice que realizarán el contrato pero no aceptan.

Falsos Negativos: Se predice que realizarán el contrato pero sí lo aceptan.

Si TPR es igual a FPR estaríamos en el así llamado "clasificador aleatorio" y nuestro modelo tendría poca capacidad de predicción. Cuanto mayor sea TPR respecto a FPR tanto mejor será la predicción del modelo. Por el contrario, cuanto menor sea TPR respecto a FPR la predicción del modelo se considerará peor.

- Dividiremos los datos en dos grupos, datos de entrenamiento y datos de prueba, en una relación de 70% - 30%. Entrenaremos el modelo con el 70% de los datos y lo evaluaremos con los datos restantes. Naturalmente que para el entrenamiento y las pruebas se conocerá si el cliente ha aceptado el depósito a plazo. Podremos evaluar el modelo mediante las métricas LIFT y ROC obtenidas en la fase de pruebas.

- Por último, dado que en los negocios todo error tendrá un costo, utilizaremos la validación cruzada. Dividiremos los datos de prueba y error en **cinco** grupos. **Cuatro** de dichos grupos se usarán para el entrenamiento y el restante se utilizará como prueba. Se realizarán **tantas pruebas como grupos**. El **grupo de prueba se irá alternando** en cada serie de pruebas, hasta que cada grupo se haya usado como grupo de prueba. Los grupos restantes serán usados como datos de entrenamiento. Este método nos permitirá una **evaluación más robusta** de los errores. Esto es en cada una de las **cinco** pruebas se pueden obtener valores distintos para LIFT y ROC, de modo que se tendrá una idea más ajustada del rendimiento del modelo.

## Datos Disponibles

- Se dispone de una serie estadística de datos de una longitud de 11.162 registros de ventas anteriores de depósitos a plazo.

- Las primeras 4 columnas contienen datos demográficos, la edad, el empleo, el estado civil y el nivel educativo del cliente.

- Las siguientes 4 columnas muestran datos bancarios del cliente. A saber, si está en números rojos, el saldo medio en su cuenta, si posee una hipoteca y si posee un crédito personal.

- Los siguientes 7 campos se refieren a las propias campañas de telemarketing, el tipo de contacto, día de la semana y mes del contacto, la duración de la llamada en segundos y otras relacionadas.

- Las dos últimas se refieren al éxito o fracaso de la llamada. La primera, "outcome", se refiere a si el cliente contrató un depósito a plazo en anteriores campañas. La segunda es nuestro objetivo, nuestra predicción, esto es, si el cliente ha contratado un depósito a plazo fijo en la presente campaña.

## Métrica de éxito del proyecto

- El proyecto trata de optimizar el número de clientes que contratan un depósito a plazo fijo durante las sucesivas campañas de marketing. Durante las mismas se realizan llamadas a los clientes para ofrecerles dicho contrato. Así pues dos métricas posibles podrían ser el porcentaje del número éxitos respecto del número de llamadas, o bien, el porcentaje del número de éxitos respecto el número de clientes llamados. Dado que un cliente podría hacer/recibir un número arbitrario de llamadas escogemos la segunda. Esto es "número de contratos" / "número de clientes llamados" x 100.

## Responsabilidades Éticas y Sociales

De cara a los principios éticos aplicados a la IA, se definen cuatro roles: Gobiernos, empresas, desarrolladores y usuarios.

En calidad de **desarrolladores** tendremos en cuenta la privacidad, seguridad, equidad y transparencia durante el diseño e implementación del proyecto.

En cuanto a la **privacidad**, en los datos a considerar aparecen los nombres, emails y direcciones de los usuarios que serán tratados como privados.

En cuanto a la **seguridad**, el proyecto no trata con materias relacionadas con la seguridad de personas o equipos, a diferencia de aplicaciones como la conducción autónoma o aplicaciones industriales.

En cuanto a la **equidad**, aparecen en los datos la edad, el empleo, el estado civil y el nivel educativo. Deberán por tanto evitarse los sesgos relacionados con tales datos.

En cuanto a la **transparencia**, se utiliza como modelo la regresión lógica, siendo éste un modelo ampliamente utilizado. Se comprobará su capacidad de predicción mediante las métricas ROC y LIFT.

Las **empresas**, por su parte, recopilarán sus datos de forma transparente, mentedrán un registro del método seguido y ofrecerán información accesible sobre las predicciones del modelo. Realizarán, además, estudios de impacto ético.