

Proyecto de ML para aumentar la contratación de depósitos de una entidad bancaria

Introducción

Tenemos un banco que, para aumentar sus ingresos, desde el área de marketing se hacen diferentes campañas, con el fin de ofrecer diferentes productos a sus clientes. Uno de los productos de la actual campaña es la contratación de un depósito y las campañas se hacen vía telefónica, que puede ser por el canal móvil o bien fijo.

Objetivos del Proyecto

1. ¿Cuáles son los objetivos del negocio?

El objetivo es aumentar la tasa de conversión de depósitos personalizando las campañas según segmentos de clientes y priorizando las llamadas a los clientes con más probabilidad de contratación. Como objetivo opcional, se busca identificar los mejores momentos del año para ofrecer los depósitos.

2. ¿Qué decisiones o procesos específicos desea mejorar o automatizar con ML?

Para segmentar los clientes elijaremos k-means, para no encontrar segmentos de forma manual. Posteriormente, saber cuáles son los clientes con más probabilidad de éxito de contratación, permitirá que los teleoperadores tengan claro en qué momento del año y de forma automática a qué clientes debe priorizar la campaña de contratación de los depósitos.

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

Buscar el perfil de clientes se podría hacer de forma no automatizada, con estadísticos descriptivos y gráficos para visualizar clientes con características similares.

Para encontrar a los clientes con más probabilidad de contratación, se podría hacer una simplificación de todos los registros en rangos de clientes con más tasa de

conversión por alguna de las características, y por tanto también se podría encontrar un modo manual.

Por el punto de la estacionalidad, también haciendo la media de las series temporales por el grupo de clientes del punto anterior (los de mayor tasa de conversión) también podríamos encontrar una aproximación de cual momentos del año podrían parecer más buenos para llamar.

Metodología Propuesta

4. ¿Cuál es el algoritmo de Machine Learning más adecuado para resolver este problema? ¿Cómo justifica la elección de este algoritmo?

En primer lugar, utilizaremos un algoritmo no supervisado de K-means para encontrar los diferentes perfiles de clientes. Intentaremos en la medida de lo posible, no encontrar demasiados segmentos, probablemente dos sería lo ideal, para no complicar los siguientes pasos.

Esta variable de clúster, será una nueva característica de nuestro dataset. A continuación, escogeremos un algoritmo supervisado de clasificación, para determinar un modelo ajustado que nos dé un nuevo cliente con nuevas características, la probabilidad de contratación.

El objetivo opcional es identificar los mejores momentos del año para ofrecer un depósito a los clientes con más probabilidad de contratación, según el algoritmo. Para ello, se consideraría el uso de un modelo ARIMA para analizar datos temporales de nuestro dataset.

¿Qué métricas de evaluación se utilizarán para medir el rendimiento del modelo?

Para seleccionar las métricas de evaluación del modelo de clasificación del problema ML bancario, hemos de tener en cuenta ciertos factores importantes como serían:

1. Sesgo de la clase
2. Costos de errores
3. Interpretación del modelo.

En primer lugar, vemos que nuestro dataset , (sin los 512 clientes outliers), la clase de clientes que han contratado depósitos está muy equilibrada frente a la clase de clientes que no han contratado. Tenemos 5654 registros con la clase deposit=0 (no contratado) y 4996 con deposit=1 (sí contratado).

Métrica de exactitud

De momento vemos que la métrica de exactitud sería la métrica de evaluación más lógica, pues con esta métrica podemos evaluar cuántas predicciones realmente ha acertado el modelo del número total de predicciones.

Exactitud = Predicciones correctas / Total de predicciones

El problema que tenemos con la exactitud, es que no tenemos certeza del tipo de predicción que acierta más si de verdaderos positivos o de verdaderos negativos.

Eso sí, como las clases están equilibradas suele ser una métrica viable porque no se verá influenciado por el sesgo de clase. Esto significa que la precisión reflejará el rendimiento del modelo sin que haya un efecto de "inflado" en la clase mayoritaria, ya que ambas clases están igualmente representadas.

Métricas asociadas al coste de error

El siguiente punto para decidir qué métrica más a escoger para nuestro problema del banco sería el coste del error.

Para entender cuál sería la mejor métrica, viendo las consecuencias de un error, vamos a definir primero que sería para nosotros un verdadero positivo y un verdadero negativo para nuestro problema del dataset. Consideraremos un verdadero positivo, pues si el clasificador nos predice que un cliente contrataría un depósito, que coincidiría con la realidad de que realmente ese cliente lo contrató. Y obviamente un verdadero negativo, pues si el clasificador lo clasifica como que el cliente no contrataría el depósito y que realmente fue así.

¿Qué sería más perjudicial a nivel de costos de error para una campaña de márketing de contratación de depósitos, que el modelo nos dé un falso positivo (el modelo nos predice que el cliente contrataría el depósito cuando realmente no fue así), o bien que nos dé un falso negativo (el modelo predice que el cliente no contrataría cuando realmente sí contrató)?

Tenemos un ejemplo muy claro para ver el coste de error en un problema de clasificación de alguien que tendrá una enfermedad mortal o no la tendrá. Un falso negativo, sería más grave que un falso positivo, pues esa persona pensaría que está sana, cuando realmente va a enfermar.

Vamos a ver para el caso de los depósitos si un falso negativo o un falso positivo tienen el mismo costo de error para el banco. El objetivo del modelo de clasificación es que los teleoperadores tengan una lista de clientes que si les llaman exista más probabilidad de que contraten un depósito. Así pues, entendemos, que, si de la lista de candidatos hubiera muchos falsos positivos, es decir muchos clientes que en realidad no contratarían el depósito, sería ineficiente. Pero, por otro lado, si el modelo tuviera muchos falsos negativos, y por lo tanto estos clientes no se les pasara a los teleoperadores para ser candidatos a llamarles, también estaríamos perdiendo clientes con alto potencial de contratación. Como conclusión, en este caso, un falso positivo y un falso negativo tienen la misma importancia, es decir el mismo coste de error para el banco.

Por este motivo podemos optar por calcular cualquiera de las dos métricas de evaluación que plantearemos, **la precisión** o bien **Recall**. Una se calcula con Falsos positivos y la otra con Falsos Negativos. Ambas obviamente dan el mejor rendimiento, idílicamente 1, cuando los Falsos Negativos u Positivos fueran cero.

Precisión

$$\text{Precision} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}}$$

Recall

$$\text{Recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

Pero en el caso que el modelo generara más Falsos positivos o bien más Falsos negativos de lo normal, entonces la métrica más acertada sería la **F1 score**, ya que haríamos que tanto Falsos Positivos como Falsos Negativos tuvieran el mismo peso en el rendimiento del modelo que es lo que queremos, ya que ambos parámetros tenían la misma importancia.

$$\text{F1 score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

Cómo conclusión para nuestro problema de clasificación de contratación de depósitos, podríamos **emplear la métrica de la exactitud**, pues tenemos un dataset sin sesgo de clase. También podríamos complementar la evaluación con la **métrica de precisión o la métrica de recall** pues ambas tienen la misma importancia en cuanto al coste de error. Pero en caso que el modelo generara más falsos positivos o bien más falsos negativos, **podríamos optar por el F1 score**, que nos quitaría este efecto de desproporción.

Datos Disponibles

5. *¿Qué datos están disponibles para abordar este problema?*

Es el dataset que tenemos colgado en esta tarea que se llama bank_dataset.csv

Es posible que quizás hagan falta algunos datos más que se puedan buscar en la web. <https://archive.ics.uci.edu/dataset/222/bank+marketing>

Métrica de Éxito

6. *¿Cuál es la métrica de éxito para este proyecto?*

La métrica de éxito será el aumento de contratación de depósitos, gracias a unas campañas más personalizadas en diferentes segmentos de clientes y en momentos del año más adecuados para los clientes que tienen más probabilidad de contratación.

Responsabilidades Éticas y Sociales

7. *¿Qué responsabilidades éticas y sociales es importante tener en cuenta?*

Como hemos visto la ética en la implementación de algoritmos de Machine learning de forma simplificada podríamos tener difusos pilares donde los más destacables serían: **Transparencia, Justicia, Privacidad, Responsabilidad e Impacto social.**

Transparencia

Todos los datos que se utilizarán para el modelo, el cliente debería haber aceptado en algún momento que el banco los podrá utilizar por objetivos comerciales. Si no se hubiera hecho esta comunicación, habría falta de transparencia al cliente.

Justicia

En ningún caso, el algoritmo, debería presuponer que hay ciertas profesiones, y ciertos estados civiles que sean más favorables para ofrecer un depósito a los clientes. Para evitar este sesgo, se evitarán estas asunciones, y se buscará que las conclusiones las saquen los propios algoritmos con los datos.

Privacidad

El cliente debe dar siempre su consentimiento para que la empresa recoja y almacene sus datos privados. Además, se debe proporcionar una manera clara y fácil para que el cliente pueda darse de baja y eliminar sus datos cuando lo desee.

Responsabilidad

Si un cliente recibe masas llamadas no deseadas por un error del algoritmo, el banco sería responsable (lo pondría por escrito) y el cliente podría reclamar daños con total transparencia.

Impacto social

Hay que tener en cuenta, que este algoritmo, debe ser beneficioso para el banco y por los clientes, que podrán recibir una atención más personalizada, así mismo como que el algoritmo nunca podrá excluir a nadie ni por nivel étnico ni por nivel cultural.