Aplicación de técnicas de aprendizaje supervisado para predecir el éxito en campañas de mercadeo directo bancario

Alberth Mora Brenes[1], Allison Cruz Rodríguez[1]
alberth.morabrenes@ucr.ac.cr, allison.cruzrodriguez@ucr.ac.cr

RESUMEN

El mercadeo directo es una estrategia clave para el sector bancario, ya que permite ofrecer productos o servicios personalizados a los clientes potenciales. Este trabajo tiene como objetivo desarrollar un modelo predictivo que permita identificar a los clientes potenciales para suscribirse a un depósito a largo plazo. Para ello, se comparan siete métodos de aprendizaje supervisado: árboles de decisión, bosques aleatorios, máquinas vectoriales de soporte, K-vecinos más cercanos, bagging, boosting y regresión logística. Para equilibrar los datos según la variable de respuesta, que muestra una distribución desigual entre quienes contratan un depósito y quienes no, se utiliza las técnicas de submuestreo aleatorio y NCL (Neighborhood Cleaning Rule). Además, se calibran los parámetros de los métodos de aprendizaje supervisado utilizados y se evalúan con validación cruzada. El método de árboles de decisión resulta ser el más adecuado, por su simplicidad y precisión. El modelo indica que los clientes más propensos a contratar un depósito son los ya contactados previamente, sin préstamo de vivienda y con menos de 31 o más de 60 años.

PALABRAS CLAVE: mercadotecnia, sector bancario, árboles de decisión, métodos de clasificación, indicadores de desempeño.

ABSTRACT

Direct marketing is a key strategy for the banking sector, as it allows offering personalized products or services to potential customers. This paper aims to develop a predictive model to identify potential customers to subscribe to a long-term deposit. For this purpose, seven supervised learning methods are compared: decision trees, random forests, support vector machines, K-nearest neighbors, bagging, boosting and logistic regression. To balance the data according to the response variable, which shows an unequal distribution between those who hire a depot and those who do not, random subsampling and NCL (Neighborhood Cleaning Rule) techniques are used. In addition, the parameters of the supervised learning methods used are calibrated and evaluated with cross-validation. The decision tree method proves to be the most appropriate, due to its simplicity and accuracy. The model indicates that the customers most likely to contract a deposit are those already contacted previously, without a home loan and under 31 or over 60 years old.

KEY WORDS: marketing, banking sector, decision trees, classification methods, performance indicators.

[1] Estudiantes de Estadística de la Universidad de Costa Rica

INTRODUCCIÓN

Una estrategia de comunicación que combina eficacia y rentabilidad es el mercadeo directo, que consiste en enviar mensajes personalizados y adaptados a las necesidades, preferencias e intereses de un cliente potencial (Mullin, 2002). El propósito es generar una respuesta inmediata del receptor, ya sea una compra, donación, suscripción o cualquier otra acción que aporte valor al emisor. Para lograrlo, el mercadeo directo se basa en el uso de bases de datos actualizadas y precisas (Roberts, 1992; Bose & Chen, 2009), así como en la evaluación de los resultados obtenidos. En este artículo, se exploran diversas técnicas de aprendizaje supervisado como herramienta para detectar el éxito de las campañas de mercadeo directo que promocionan la suscripción a depósitos a largo plazo según el perfil del cliente.

El mercadeo directo puede emplear diferentes estrategias para estimular la demanda de depósitos a largo plazo. Estas estrategias incluyen difundir los productos de depósito a largo plazo, destacando sus beneficios como las tasas de interés, los incentivos y otras ventajas. Según Athanassakos & Waschik (1997), el mercadeo también puede ayudar a construir una reputación de solvencia y confiabilidad del emisor, lo que puede atraer más demanda de sus productos de depósito a largo plazo. Además, el mercadeo puede educar a los consumidores sobre las características de los depósitos a largo plazo y cómo pueden alinearse con sus metas financieras globales (Athanassakos & Waschik, 1997). Sin embargo, la efectividad del mercadeo puede variar según el perfil del público, el contenido y el canal utilizado.

Debido a esta deficiencia, Alhejaily, Abdulghani y Yafooz (2022) señalan que el mercadeo directo se puede optimizar con el aprendizaje supervisado. Esta técnica permite identificar los clientes, productos y canales más adecuados para ofrecer los servicios bancarios. Además, mejora la precisión de las predicciones y con ello la rentabilidad de las campañas de mercadeo directo (Alhejaily et al., 2022). El aprendizaje supervisado también facilita la personalización de las ofertas en función de los atributos de cada cliente (Ładyżyński, Żbikowski y Gawrysiak, 2019).

Mitik, Korkmaz, Karagoz, Toroslu & Yucel (2017) predijeron el interés de los clientes en los productos bancarios y el canal de comunicación óptimo usando aprendizaje supervisado y minería de datos. Analizaron un conjunto de datos de un banco turco con información sobre las características, las transacciones y las respuestas de los clientes a las campañas de mercadeo directo. Los métodos propuestos lograron una alta precisión y sensibilidad en la predicción, y mejoraron la relación costo/beneficio de las campañas, aumentando la eficiencia del mercadeo directo utilizado.

Moro, Laureano & Cortez (2011) utilizaron minería de datos y aprendizaje supervisado para mejorar la eficiencia de las campañas de mercadeo directo de un banco portugués. Los autores utilizaron diferentes algoritmos de clasificación para predecir si un cliente se suscribiría a un depósito a largo plazo. El mejor modelo obtenido logró altos rendimientos predictivos y permitió identificar las características más relevantes que influyen en el éxito de un contacto.

Por lo expuesto, la relevancia del aprendizaje supervisado en el contexto del mercadeo directo bancario radica fundamentalmente en su capacidad para optimizar los recursos asignados a esta estrategia. Esta técnica permite minimizar el volumen de interacciones

necesarias para alcanzar un número equivalente de conversiones. En el caso concreto que se analiza en esta investigación, significa que se puede lograr una cantidad similar de clientes que adquieran un depósito a largo plazo con un menor número de llamadas.

El objetivo de esta investigación es encontrar la mejor regla de decisión, mediante el uso de diferentes métodos de aprendizaje supervisado, para predecir si un cliente contratará un depósito bancario a largo plazo si recibe una llamada telefónica de mercadeo directo.

METODOLOGÍA

Para efectuar el análisis planteado en este estudio, se empleó el conjunto de datos de mercadeo bancario que se encuentra en el Repositorio de Aprendizaje Automático de la Universidad de California en Irvine, una fuente abierta para la investigación. Este conjunto de datos fue introducido y examinado por Moro et al. (2011); y corresponde a una campaña de mercadeo directo de un banco portugués que promocionaba depósitos a largo plazo durante la crisis financiera global. El banco contactaba a sus clientes por teléfono o internet, ofreciéndoles intereses favorables. El conjunto de datos contiene información de 45,211 contactos realizados entre mayo de 2008 y noviembre de 2010. La Tabla 2 en Anexos muestra los 17 atributos que se registran para cada contacto con un cliente potencial.

Para abordar el desequilibrio del conjunto de datos, donde sólo el 11,7% de los clientes suscribieron el depósito, se utilizaron dos técnicas de submuestreo. La primera fue la Norma De Limpieza de Barrio (Neighborhood Cleaning Rule o NCL, por sus siglas en inglés), que reduce la clase mayoritaria mediante la combinación de dos métodos: Condensed Nearest Neighbor (CNN), que elimina las observaciones redundantes, y Edited Nearest Neighbors (ENN), que filtra las observaciones ruidosas o ambiguas (Laurikkala, 2001). La NCL se enfoca en mejorar la calidad de los ejemplos retenidos, pero no el balance de clases (Brownlee, 2021). Debido a esto, la segunda técnica usada fue el submuestreo aleatorio, que consiste en seleccionar y descartar al azar ejemplos de la clase mayoritaria del conjunto de datos de entrenamiento (Fernández, García, Galar, Prati, Krawczyk & Herrera, 2018).

La primera técnica empleada de clasificación fue la Regresión Logística, que consiste en una técnica estadística que permite examinar la relación entre una variable dependiente y varias variables independientes. Al igual que la regresión lineal, su objetivo es predecir el comportamiento de la variable dependiente; en este caso, estimando las probabilidades de un evento en función de las variables predictoras. Además, a diferencia de la regresión lineal, la regresión logística se utiliza para predecir la pertenencia a un grupo específico, utilizando una variable dependiente categórica o cualitativa. Esto implica identificar las características o factores que diferencian los grupos definidos por la variable dependiente (López-Roldán & Fachelli, 2016). Para este método se realiza una selección de variables utilizando la prueba de razón de verosimilitud y posteriormente se van eliminando variables de manera que se obtengan los mejores indicadores de desempeño.

La segunda técnica de clasificación empleada fue la de K-Vecinos Más Cercanos o K-NN, por sus siglas en inglés (K-Nearest Neighbor). La idea fundamental de este algoritmo es que un nuevo caso se clasificará en la clase que sea más común entre sus K vecinos más cercanos

(Izurieta & Moyano, 2019). Debido a la naturaleza de las variables que se incluyeron en el análisis para el cálculo del K-NN se utiliza la distancia de Gower y se calibra la cantidad de vecinos.

La tercera técnica empleada consistió en Máquinas Vectoriales de Soporte (SVM, por sus siglas en inglés). Esta técnica construye un hiperplano óptimo, de forma que el margen de separación entre las dos clases en los datos se amplíe lo máximo posible. Se llama vectores de soporte porque es un subconjunto de observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie del hiperplano. El proceso de calibración se realizó para la función de kernel.

La cuarta técnica de clasificación utilizada fue Árboles de Decisión, que proporciona una representación visual de la toma de decisiones. Este modelo se construye a partir de la descripción narrativa de un problema, especificando las variables que se evalúan, las acciones que deben tomarse y el orden en el que se tomarán las decisiones. Cada vez que se ejecuta este modelo, solo se sigue un camino, dependiendo del valor actual de la variable evaluada. Los valores que pueden tomar las variables para este tipo de modelos pueden ser discretos o continuos (Barrientos et al., 2009). Para este método se realizó la calibración del parámetro de complejidad (cp) y el parámetro de profundidad máxima (maxdepth).

La quita técnica utilizada fue Bagging, el cual implica la creación de muestras aleatorias utilizando la técnica de Bootstrap, que consiste en generar conjuntos de muestras a partir de los datos existentes, permitiendo el reemplazo. A partir de estas muestras, se construyen varios modelos. Estos modelos se comparan entre sí, y cada observación se clasifica en todos los modelos. La clasificación final de cada observación se determina por la categoría a la que fue asignada con mayor frecuencia en todos los modelos (Pang-Ning et al, 2019). En esta técnica se utilizaron los parámetros que se calibraron previamente en los árboles de decisión para posteriormente hacer una calibración de la cantidad de árboles.

La sexta técnica utilizada corresponde a Bosques Aleatorios, los cuales son un método de aprendizaje automático que construye múltiples árboles de decisión a partir de muestras aleatorias de observaciones seleccionadas con reemplazo (técnica de Bootstrap). Cada árbol se genera utilizando un subconjunto de variables elegidas al azar. La predicción final se obtiene a partir de las predicciones de todos los árboles; por ejemplo, por votación mayoritaria para la clasificación (Medina & Ñique, 2017). Al igual que en las técnicas de clasificación anteriores se realizó calibración: primero se estableció el tamaño de nodo terminal en 50 y se calibró el número de variables aleatorias y la cantidad de árboles.

La última técnica empleada fue Potenciación (Boosting) que funciona ajustando secuencialmente los pesos de los datos de entrenamiento en función de los errores de los modelos anteriores. Luego entrena un nuevo modelo en los datos re-ponderados. El proceso se repite, cada vez ajustando los pesos de los datos para corregir los errores de la iteración anterior. Al final, las predicciones de todos los modelos se combinan para hacer la predicción final. Se realizó calibración para el parámetro de contracción (*nu*) y la cantidad de árboles.

Se empleó la técnica de validación cruzada para evaluar los métodos de clasificación y calibración. En este enfoque, el conjunto de datos se divide en k particiones o pliegues (en este

caso, se realizaron 10 divisiones). A continuación, se lleva a cabo un proceso iterativo en el que, en cada iteración, se seleccionan k-1 particiones como conjunto de entrenamiento y la partición restante como conjunto de validación. De este modo, se crea un modelo y se calculan diversas métricas conocidas como indicadores de desempeño.

Estas métricas son empleadas tanto en el proceso de ajuste del modelo como en la validación cruzada. Permiten evaluar y comparar la calidad de un modelo de clasificación. La metodología se basa en una estructura denominada matriz de confusión. Visualmente, se representa como una "tabla de confusión" que contrasta las predicciones realizadas por el modelo de clasificación con las categorías originales de los datos en el conjunto de validación.

Los indicadores de desempeño que se tomaron en cuenta fueron el error de clasificación del modelo; el área bajo la curva (AUC, siglas en inglés) que proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles: este valor oscila entre 0 y 1; el KS (Kolmogorov-Smirnov) que corresponde a la máxima diferencia entre las distribuciones acumuladas relativas de las clases, ordenadas de menor a mayor según la probabilidad predicha por el modelo de clasificación: esta va de 0 a 100. Además, se toman en cuenta Falsos Positivos (FP), que son la proporción de casos negativos que fueron clasificados incorrectamente como positivos; y Falsos Negativos (FN), que son la proporción de casos positivos que fueron clasificados incorrectamente como negativos.

Para el análisis se utilizó el software estadístico R, versión 4.3.1 (R Core Team, 2023), y las librerías rattle (Williams, 2011), caret (Kuhn, 2008), ggplot2 (Wickham, 2016), highcharter (Kunst, 2022) y rocr (Sing, Sander, Beerenwinkel y Lengauer 2005). Para la técnica de Árboles de Decisión se utilizó la librería rpart (Therneau y Atkinson, 2022), para Bagging, Bosques Aleatorios y Boosting se utilizaron adabag (Alfaro, Gamez & García, 2013) y randomForest (Liaw y Wiener, 2002), y, por último, para Máquinas Vectoriales de Soporte se utilizó la librería e1071 (Meyer, Hornik y Leisch, 2023).

RESULTADOS

Durante el proceso de revisión de calidad de los datos, se depuraron los valores faltantes de las variables 'trabajo' y 'educacion', los cuales representaban el 0.6% y 4.10% del total de datos, respectivamente. Posteriormente, se decidió prescindir de las variables 'ucontacto' y 'resultprev'; eliminamos 'ucontacto' porque cuenta con un 28.7% de datos ausentes y no se considera relevante para el análisis. Por otro lado, 'resultprev' fue descartada debido a que contenía una alta proporción de valores faltantes, concretamente un 81.7%, lo cual limitaba su utilidad para realizar predicciones eficaces. Este estudio, además, excluye las variables relacionadas con el último contacto de la campaña de mercadeo actual ('udia', 'umes' y 'uduracion'), ya que estas variables proporcionan información extra que no debe ser usada por los modelos, pues el propósito es clasificar a los individuos antes de contactarlos.

Tras la depuración de estos datos, se implementaron los métodos de NCL y submuestreo aleatorio, dando como resultado un total de 10,042 observaciones; las cuales fueron utilizadas para todos los análisis posteriores de aprendizaje supervisado. Las características descriptivas de los datos resultantes se detallan en la Tabla 3 del Anexo. Estas muestran que aquellos que aceptaron la suscripción a un depósito a largo plazo tienden a tener un saldo medio anual más

alto. Además, los clientes que rechazaron la suscripción son más propensos a tener préstamos para vivienda y personales.

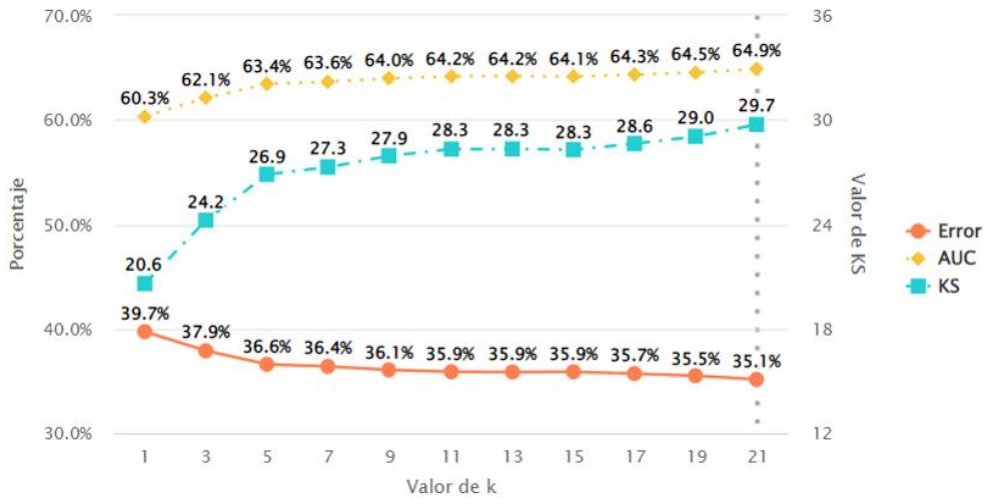
Inicialmente, para el método de Regresión Logística, se realizó el proceso de selección de variables con la prueba de razón de verosimilitud. Para esto, se compararon modelos que excluían una variable específica contra el modelo completo. Como resultado de dicha comparación, se determinó que la variable de edad no era significativa ($p > 0.80$). Por lo tanto, se decidió eliminar dicha variable del análisis.

Seguidamente, se realizó otra selección de variables, pero esta vez utilizando los indicadores de desempeño, específicamente el error de clasificación. El objetivo era identificar aquellas variables cuya eliminación contribuyera a reducir el error. Se observó que, en general, el error apenas variaba al eliminar una variable. Sin embargo, se encontró que la eliminación de las variables 'diasdespues' y 'morosidad' generaba una ligera disminución en el error de clasificación, pasando de 34.52% a 34.28%; por lo tanto, se eliminaron dichas variables.

Para analizar el método de los K-Vecinos Más Cercanos, se realizó la calibración del número de vecinos mediante el uso de validación cruzada y se evaluó el rendimiento utilizando indicadores, especialmente el AUC, KS y el error de clasificación. En este caso, se determinó que el número óptimo de vecinos para obtener los mejores resultados de los indicadores fue de 21, como se muestra en la Figura 1. Con esta configuración, se logró un AUC del 64.87%. y un KS de 29.73.

Figura 1

Calibración de la cantidad de vecinos para el método de K-NN utilizando como parámetro de decisión el AUC, KS y el error de clasificación



Posteriormente, se llevó a cabo la calibración de los distintos tipos de kernels para la técnica de Máquinas de Vectores de Soporte, con el fin de determinar cuál era el mejor. A través de la evaluación de indicadores de desempeño, se elige el kernel radial ya que ofrece los mejores resultados. Esto se debe a que presenta los mayores valores de AUC y KS (67.20% y 34.40, respectivamente), en comparación con los otros kernels. Además, se observó que tiene el menor error de clasificación, FP y FN; como se muestra en la Tabla 4 del Anexo.

Para el análisis del método de Árboles de Decisión, en primer lugar, se llevó a cabo la calibración del modelo para el parámetro de complejidad (ver Tabla 5 del anexo). En dicho proceso, se encontró que el mejor valor de cp para lograr resultados óptimos en términos de error de clasificación, AUC y KS fue de 0.002. Posteriormente, se procedió a calibrar el parámetro de profundidad, manteniendo el valor previamente seleccionado para el parámetro de complejidad.

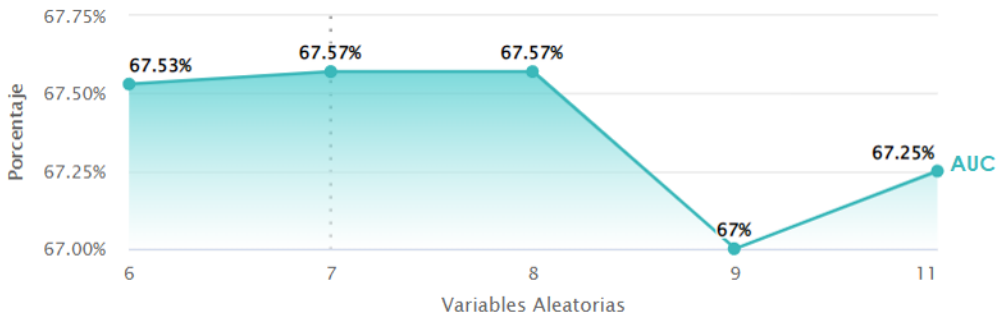
Se tomó un rango de profundidad de 1 a 15, y se observó que a partir de una profundidad de 12 se obtuvieron los mejores resultados en cuanto a los indicadores de desempeño. A medida que se aumentaba la profundidad más allá de este punto, los indicadores se estabilizaban (ver Tabla 6 del anexo). Por lo tanto, se decidió elegir una profundidad de 12 para el modelo de Árboles de Decisión.

Después, se aplicó el método de Bootstrap (Bagging). Se llevó a cabo la calibración manteniendo los parámetros encontrados en el método de Árboles de Decisión, y se procedió a calibrar el número de árboles. En este caso, se seleccionó el valor de 200, ya que se obtuvieron los mejores resultados para el error de clasificación, AUC y KS (ver Tabla 7 del anexo).

Para llevar a cabo la calibración en el método de Bosques Aleatorios, se mantuvieron 500 árboles para calibrar la cantidad de variables aleatorias utilizadas. El indicador de desempeño utilizado en este proceso de calibración fue el AUC. En la Figura 2 se puede observar que se obtiene el mejor AUC al utilizar 7 u 8 variables aleatorias. Por lo tanto, se decidió seleccionar 7 variables aleatorias para el modelo.

Figura 2

Calibración de la cantidad de variables aleatorias para el método de Bosques Aleatorios utilizando como parámetro de decisión el AUC



A continuación, se procedió a la calibración del número de árboles, manteniendo las 7 variables aleatorias seleccionadas. Como se muestra en la Tabla 8 del Anexo, el AUC y KS más alto se obtuvo cuando el número de árboles es igual a 50.

El último método que se calibró fue Boosting. Primeramente, se calibró el parámetro de contracción. Sin embargo, se observó que no había diferencia en los indicadores de desempeño para los valores del parámetro de contracción utilizados (ver Tabla 9 del anexo). Por lo tanto, se decidió utilizar un valor de 0.6 para este parámetro.

Posteriormente, se procedió a calibrar la cantidad de iteraciones, en este caso, la cantidad de árboles. Se seleccionó un valor de 150 árboles, ya que se obtuvieron los mejores resultados en términos de error de clasificación, AUC y KS (ver Tabla 10 del anexo).

Por último, una vez calibrados los métodos, se procedió a compararlos. Para esto, se realizó validación cruzada para escoger el mejor método para clasificar. En la Tabla 1 se muestra que la mejor técnica para clasificar es Boosting, ya que cuenta con los mejores resultados en los indicadores de desempeño. Tiene el error de clasificación más bajo, con un 31.94%, y el AUC y KS más altos, con un 68.09% y un 36.16, respectivamente. Sin embargo, en el porcentaje de falsos positivos, Bagging tiene los mejores resultados, con un 19.49%, y en cuanto a los falsos negativos, la regresión logística binomial presenta los mejores resultados, con un 36.19%.

A pesar de que el mejor modelo es Boosting, se decidió elegir el método de Árboles de Decisión debido a que es más sencillo. Además, al comparar los indicadores de desempeño entre los Árboles de Decisión y Boosting, se observa que son muy similares, con las únicas diferencias notables en los Falsos Positivos, de 4.04 puntos porcentuales; y en los Falsos Negativos, de 5.63 puntos porcentuales.

Tabla 1

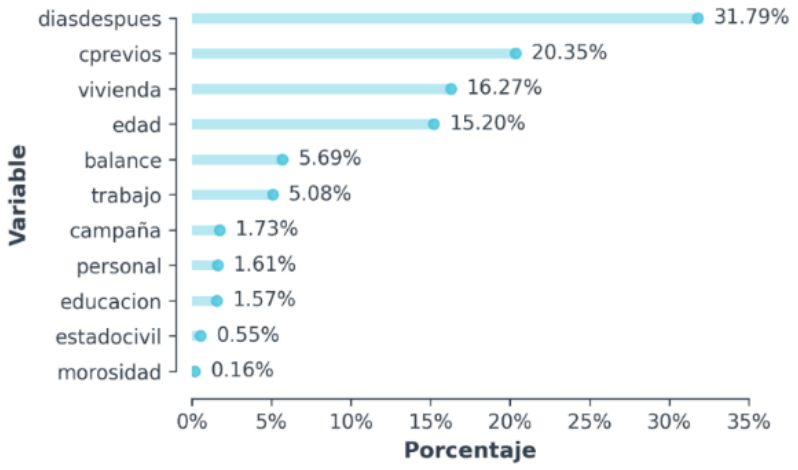
Comparación de técnicas de clasificación según indicadores de desempeño utilizando validación cruzada

Técnica	E	FP	FN	AUC	KS
Regresión Logística Binomial	36.08	35.95	**36.19**	63.93	27.86
K-vecinos más cercanos	35.14	26.16	44.11	64.87	29.73
Máquinas Vectoriales de Soporte	32.78	21.75	43.81	67.22	34.45
Árboles de decisión	32.72	20.31	**45.12**	67.3	34.56
Bagging	32.14	**19.49**	44.8	67.86	35.71
RandomForest	32.35	23.43	41.31	67.63	35.26
Boosting	**31.94**	24.35	39.49	**68.09**	**36.16**

Las variables más importantes para el ajuste del modelo de árboles de decisión se observan en la Figura 3.

Figura 3

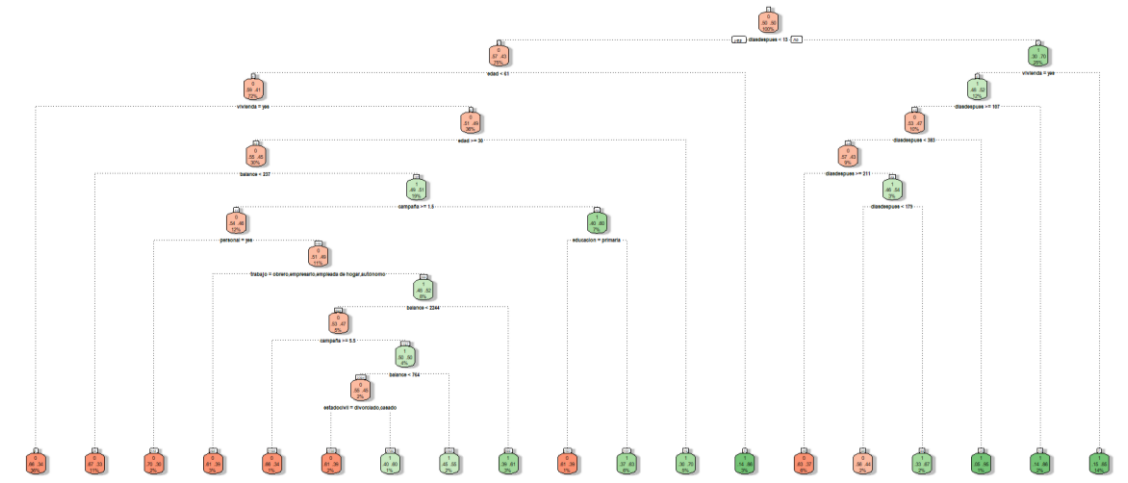
Porcentaje de relevancia de las variables utilizadas en el ajuste de modelo de Árboles de Decisión



Asimismo, el árbol de decisión resultante (Figura 4) muestra que la variable más importante para predecir la suscripción al depósito es el número de días transcurridos desde el último contacto (diasdespues). Según la primera regla de decisión, las personas que han sido contactadas hace más de 13 días tienen más posibilidades de aceptar la suscripción. Esto indica que el contacto previo genera interés en el producto. Por otro lado, los nodos terminales con más observaciones son el primero, el segundo y el último, que abarcan el 38%, el 11% y el 14% del total de observaciones, respectivamente. En estos nodos, se aprecia que las personas con préstamo de vivienda ('vivienda') suelen rechazar el depósito. Estos hallazgos son consistentes con las correlaciones entre estas variables y la respuesta (0.10 para 'diasdespues' y -0.21 para 'vivienda') (Figura 5 del anexo).

Figura 4

Representación gráfica del modelo resultante de Árboles de Decisión



La Figura 6 en Anexos ilustra la relación entre el número de contactos previos al cliente antes de la campaña actual y el porcentaje de suscripciones obtenidas. Se observa una tendencia positiva: a más contactos previos, más porcentaje de suscripciones; con un incremento de 29.24 puntos porcentuales al pasar de 0 a 3 o más contactos.

La edad y el saldo medio anual no parecen tener mucha influencia en la decisión de suscribirse al depósito a largo plazo, ya que su correlación con la respuesta es muy baja (cerca de 0) (Figura 5 de anexos). Esto se refleja también en el árbol de decisión (ver Figura 4), donde las ramas que se derivan de estas variables no muestran una clara distinción entre los grupos de clientes. Sin embargo, a las personas con edad menor a 31 o más de 60 años se les suele clasificar como que sí se suscribirán al depósito.

CONCLUSIONES

El objetivo general de este análisis fue encontrar la mejor regla de decisión para predecir si un cliente contrataría un depósito bancario a largo plazo si recibe una llamada de mercadeo directo. La mejor regla la proporcionó el modelo de Árboles de Decisión, que se seleccionó por su sencillez. Aunque este modelo no tuvo el mejor desempeño en términos de error, AUC y KS, la diferencia con los otros métodos no fue relevante como para elegir un modelo más complejo.

Las cuatro variables más relevantes del modelo resultante para identificar a los clientes potenciales que se interesan por suscribirse al depósito son: 'diasdespues', 'cprevios', 'vivienda' y 'edad'. Estas variables permiten distinguir entre los clientes que aceptarán o rechazarán la oferta. Según el análisis, los clientes más propensos a suscribir un depósito son aquellos que ya han sido contactados anteriormente, no tienen un préstamo de vivienda y tienen menos de 31 o más de 60 años. Las demás variables no presentaron reglas de decisión que ayudaran a identificar claramente a los clientes.

No obstante, hay que tener en cuenta que el modelo elegido es un clasificador regular de clientes que se suscriben o no al depósito a largo plazo. El modelo obtuvo una tasa alta de falsos negativos (casi 50%) (ver Tabla 1), lo que significa que el banco podría perder oportunidades de ofrecer el depósito a clientes interesados, y tendría que hacer más llamadas para alcanzar su objetivo. Este problema persiste, aunque se usen otros métodos de clasificación, ya que las variables en estudio no parecen ser suficientes para distinguir entre los clientes que se suscriben o no.

A pesar de lo anterior, este estudio contribuye a la investigación previa sobre esta base de datos (Moro et al., 2011; Parlar & Acaravci, 2017; Mitik et al, 2017) al eliminar las variables que se suponían más relevantes para predecir si un cliente se suscribiría o no al depósito a largo plazo; ya que son variables que se obtienen después del último contacto con el cliente. En cambio, este estudio enfatiza las variables anteriores a ese contacto para desarrollar un modelo que identifique a los clientes más interesados de manera que reduzca las cantidades necesarias de esos últimos contactos.

En este estudio se reconocen diversas limitaciones. La primera es que los datos se recolectaron de una campaña de mercadeo directo de un banco portugués durante la Gran Recesión del 2008 y la crisis económica de Portugal (Pettinger, 2017); lo que puede reducir la generalización de los resultados a otros contextos o países. La segunda es que la calidad de los datos no permitió incluir dos variables que podrían tener un impacto en los modelos de clasificación: el resultado de la campaña anterior (resultprev) y el medio de contacto (ucontacto). La tercera es que los datos no incorporan otras variables que pueden influir en la decisión de los clientes, como el contenido del mensaje, el tipo de relación con el banco (cliente nuevo o recurrente) o la satisfacción previa con el banco.

Para futuras investigaciones, se sugiere incluir en el análisis el efecto de diferentes estrategias de segmentación de mercado en las campañas de mercadeo directo en el sector bancario, ya que la calidad y la eficacia de estas campañas dependen en gran medida de la segmentación del público objetivo (Martin, 2011; Dutta, Bhattacharya & Guin, 2015). También se aconseja obtener más datos sobre la relación del cliente con el banco, para evaluar si se pueden desarrollar modelos predictivos de alta calidad sin usar información posterior al contacto; con el objetivo de reducir el número de contactos requeridos para lograr una campaña de mercadeo directo exitosa.

BIBLIOGRAFÍA

Alfaro, E., Gamez, M., & Garcia, N. (2013). adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software, 54(2), 1-35. https://doi.org/10.18637/jss.v054.i02

Alhejaily, B. A., Abdulghani, R. M., & Yafooz, W. M. (2022, November). *Machine Learning and Data Mining Use Cases in the Development of Marketing Strategies*. In Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022 (pp. 581-591). Singapore: Springer Nature Singapore.

Athanassakos, G., & Waschik, R. (1997). The demand for long-term deposits of a financial intermediary: Theory and evidence. *Journal of Economics and Business*, 49(2), 127-147. https://doi.org/10.1016/S0148-6195(96)00078-1

Barrientos, R.E., Cruz, N., Acosta, H.G., Rabatte, I., Gogeascoechea, M., Pavón, P. & Blázquez, S. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Rev Med UV*, 9(2), 19-24. http://www.soporte.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, 195(1), 1-16. https://doi.org/10.1016/j.ejor.2008.04.006

Brownlee, J. (2021). *Undersampling Algorithms for Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/#:~:text=The%20simplest%20undersampling%20technique%20involves,referred%20to%20as%20random%20undersampling.

Dutta, S., Bhattacharya, S., & Guin, K. K. (2015). *Data mining in market segmentation: a literature review and suggestions*. In Proceedings of Fourth International Conference on Soft Computing for Problem Solving: SocProS 2014, Volume 1 (pp. 87-98). Springer India.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10, pp. 978-3). Cham: Springer.

Izurieta, G., & Moyano, R. (2019). *Predicción de clientes potenciales utilizando el algoritmo k vecino más cercano en el área de negocios de la COAC "Riobamba" ltda*. Universidad Nacional De Chimborazo, Riobamba, Ecuador. http://dspace.unach.edu.ec/bitstream/51000/6043/1/UNACH-EC-ING-SIT-COMP-2019-0010.pdf

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1–26. https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

Kunst, J. (2022). *_highcharter: A Wrapper for the 'Highcharts' Library_*. R package version 0.9.4, https://CRAN.R-project.org/package=highcharter.

Ładyżyński, P., Żbikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, 134, 28-35.

Laurikkala, J. (2001). *Improving identification of difficult small classes by balancing class distribution*. In Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8 (pp. 63-66). Springer Berlin Heidelberg.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*. 2(3), 18--22.

López, P., & Fachelli, S. (2016). *Metodología de la investigación social científica*. Universidad Autónoma de Barcelona. https://ddd.uab.cat/pub/caplli/2016/163570/metinvsoccua_a2016_cap3-10.pdf

Martin, G. (2011). The importance of marketing segmentation. *American journal of business education*, 4(6), 15-18. https://doi.org/10.19030/ajbe.v4i6.4359

Medina, R., & Ñique, C. (2017). *Bosques aleatorios cómo extensión de los árboles de clasificación con los programas R y Python*. Interfases. N. 10 https://dialnet.unirioja.es/descarga/articulo/6230447.pdf

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*, TU Wien_. R package version 1.7-13, https://CRAN.R-project.org/package=e1071

Mitik, M., Korkmaz, O., Karagoz, P., Toroslu, I. H., & Yucel, F. (2017). Data mining approach for direct marketing of banking products with profit/cost analysis. *The Review of Socionetwork Strategies*, 11, 17-31.

Moro, S., Laureano, R., & Cortez, P. (2011). *Using data mining for bank direct marketing: An application of the crisp-dm methodology*. https://hdl.handle.net/1822/14838

Mullin, R. (2002). *Direct marketing: a step-by-step guide to effective planning and targeting*. Kogan Page Publishers.

Pang-Ning, T., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining (2nd ed.).* Pearson. https://www.pearson.com/store/p/introduction-to-data-mining/P100001265344/9780133128901

Parlar, T., & Acaravci, S. K. (2017). Using data mining techniques for detecting the important features of the bank direct marketing data. *International journal of economics and financial issues*, 7(2), 692-696.

Pettinger, T. (2017). *Portugal Economic Crisis.* Economics Help. https://www.economicshelp.org/blog/6423/economics/portugal-economic-crisis/

Roberts, M. L. (1992). Expanding the role of the direct marketing database. *Journal of Direct Marketing*, 6(2), 51-60. https://doi.org/10.1002/dir.4000060208

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). *ROCR: visualizing classifier performance in R*. _Bioinformatics_, *21*(20), 7881. http://rocr.bioinf.mpi-sb.mpg.de

Therneau, T., & Atkinson, B. (2022). *_rpart: Recursive Partitioning and Regression Trees_*. R package version 4.1.19, https://CRAN.R-project.org/package=rpart

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Williams, G. J. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery* (Use R!). Springer.

ANEXOS

Tabla 2

Definición de las variables del conjunto de datos de mercadeo directo bancario

ID	Variable	Tipo	Definición
Variables sociodemográficas del cliente			
1	edad	Numérica	Edad en años cumplidos
2	trabajo	Categórica	Ocupación ("administración", "desconocido", "desempleado", "directivo", "empleada de hogar", "empresario", "estudiante", "obrero", "autónomo", "jubilado", "técnico", "servicios")
3	estadocivil	Categórica	Estado civil ("casado", "divorciado o viudo", "soltero")
4	educacion	Categórica	Nivel educativo ("desconocido", "secundario", "primario", "terciario")
5	morosidad	Binaria	Si la persona tiene crédito en mora
6	balance	Numérica	Saldo medio anual, en euros (numérico)
7	vivienda	Binaria	Si la persona tiene préstamo para vivienda
8	personal	Binaria	Si la persona tiene préstamo personal
Variables relacionadas con el último contacto en la campaña actual			
9	ucontacto	Categórica	Medio de contacto ("desconocido", "teléfono", "móvil")
10	udia	Numérica	Día en que se realizó el último contacto
11	umes	Categórica	Mes en que se realizó el último contacto ("enero", "febrero", …, "diciembre")
12	uduracion	Numérica	Duración de la llamada del último contacto
Otros atributos			
13	campaña	Numérica	Número de contactos realizados al mismo cliente previamente y durante la campaña actual
14	diasdespues	Numérica	Número de días transcurridos desde que el cliente fue contactado por última vez en una campaña anterior (-1 significa que el cliente no fue contactado previamente)
15	cprevios	Numérica	Número de contactos realizados al mismo cliente antes de la campaña actual

16	resultprev	Categórica	Resultado de la campaña de mercadeo anterior ("desconocido", "otro", "fracaso", "éxito")
Variable de respuesta			
17	y	Binaria	Si el cliente se suscribió a un depósito a largo plazo

Tabla 3

Estadísticas descriptivas de las variables numéricas por suscripción aceptada o rechazada

Variables	**Suscripción a depósito a largo plazo**			
	Aceptada		**Rechazada**	
	Media	**Desviación estándar**	**Media**	**Desviación estándar**
edad	41.48	13.31	40.80	9.99
balance *(Saldo medio anual)*	718.00	3483.19	387.00	2990.60
campaña *(Número de contactos totales al mismo cliente)*	2.15	1.93	2.00	3.14
diasdespues (Días transcurridos desde el último contacto)	68.49	119.00	36.57	97.36
cprevios *(Número de contactos realizados previamente)*	1.18	2.58	0.49	1.78
morosidad *(Si el cliente tiene crédito en mora)*	0.01	0.10	0.02	0.14
vivienda *(Si el cliente tiene préstamo para vivienda)*	0.37	0.48	0.59	0.49
personal *(Si el cliente tiene préstamo personal)*	0.09	0.29	0.18	0.38

Tabla 4

Indicadores de desempeño según kernel para la técnica de SVM

Kernel	**Error**	**FP**	**FN**	**AUC**	**KS**
Lineal	35.00	34.80	35.20	65.00	30.00
Polinomial	33.90	20.30	47.50	66.10	32.20
Radial	32.80	21.70	43.80	67.20	34.40
Sigmoidal	40.00	34.50	45.50	60.00	20.00

Tabla 5

Calibración del parámetro de complejidad (cp) para la técnica de Árboles de Decisión

Tabla 6

Calibración del parámetro de profundidad para la técnica de Árboles de Decisión

Profundidad	**Error**	**FP**	**FN**	**AUC**	**KS**
1	39.9	15.2	64.6	60.1	20.2

3	37.1	5.4	68.9	62.8	25.7
6	33.1	19.4	46.9	66.8	33.7
9	32.7	19.9	45.5	67.3	34.6
12	32.7	20.3	**45.1**	**67.3**	**34.6**
15	32.7	20.3	45.1	67.3	34.6

Tabla 7

Calibración del número de árboles para la técnica de Bagging

Árboles	Error	AUC	KS
5	32.6	67.4	34.9
50	32.5	67.5	35
100	32.3	67.7	35.5
200	**32.1**	**67.9**	**35.7**

Tabla 8

Calibración del número de árboles para la técnica de Árboles Aleatorios.

Árboles	Error	AUC	KS
50	32.4	**67.6**	**35.3**
100	32.7	67.3	34.6
200	32.5	67.5	34.9
300	32.6	67.4	34.8
500	32.4	67.6	35.1
700	32.4	67.6	35.2

Tabla 9

Calibración del parámetro de contracción para la técnica de Boosting.

nu	Error	AUC	KS
0.1	32.4	67.6	35.2
0.3	32.4	67.6	35.2
0.6	32.4	67.6	35.2
0.9	32.4	67.6	35.2

Tabla 10

Calibración del número de árboles para la técnica de Boosting.

Árboles	Error	AUC	KS
30	32.6	67.42	34.83
90	32.13	67.88	35.74
150	**31.48**	**68.53**	**37.06**

300	32.01	68.01	36.02
400	31.93	68.08	36.16

Figura 5

Correlaciones entre las variables empleadas en el ajuste del modelo de Árboles de Decisión

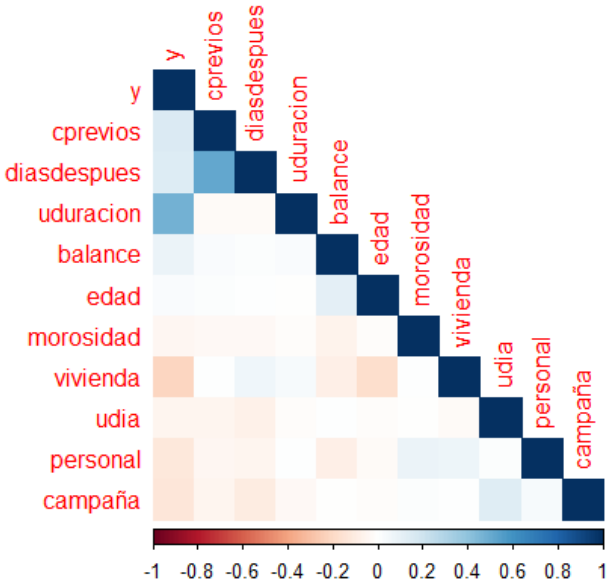


Figura 6

Porcentaje de suscripciones según el número de contactos realizados al mismo cliente antes de la campaña de mercadeo actual