APLICACIÓN DE MACHINE LEARNING PARA CAMPAÑAS DE MARKETING EN LA BANCA COMERCIAL

Valeria Triviño *

* *Universsidad Eloy Alfaro de Manabí, EL CARMEN Manabí(e-mail: vltl.1314772607@gmail.com).*

Abstract:
Los bancos usan el telemarketing para contactar directamente con los clientes potenciales para sus productos. Este canal de venta es complejo, pues requiere de grandes bases de datos de posibles prospectos y está sujeto a restricciones de tiempo y personal. Este artículo tiene tres objetivos: comparar cinco modelos de predicción basados en algoritmos de aprendizaje automático para encontrar el que ofrezca la mejor precisión predictiva; desplegar un piloto de este modelo; y recomendar una hoja de ruta para la futura arquitectura que lo soporte. Se encontró que el algoritmo seleccionado mejora considerablemente la eficacia de la identificación de clientes que aceptan el producto, que pasó de 11 implementación puede contribuir a la competitividad de estas organizaciones.

Keywords: Banca, marketing, depósitos a plazo fijo, aprendizaje automático, algoritmos de clasificación.

1. INTRODUCTION

Una parte importante del negocio bancario son las operaciones pasivas, como la apertura de cuentas dirigidas a los clientes minoristas, en sus diferentes formas, ya sean cuentas de ahorros, cuentas corrientes o cuentas a plazo fijo. Sin embargo, uno de los desafíos que enfrenta la banca comercial en este ámbito es el hecho de tener que contactar a una cantidad importante de clientes, aun con los escasos recursos, tanto de tiempo como de materiales, con los que se cuenta.

En este esfuerzo, desde hace unos años, muchos bancos han ido incorporando de manera progresiva herramientas basadas en machine learning y minería de datos con el objetivo de incrementar el nivel de éxito de sus campañas, identificando para ello los principales factores que pueden conducir a él (Dutta et al., 2020).

2. ESTADO DEL ARTE

A fin de lograr los tres objetivos planteados, se describe a continuación el estado del arte para los algoritmos de machine learning, así como para el concepto de arquitectura empresarial. Entre los principales se encuentran los siguientes:

2.1 DECISION TREE

Los árboles de decisión se han convertido en uno de los modelos más potentes y populares en la ciencia de datos, como ciencia y tecnología de exploración de grandes y complejos conjuntos de datos, donde ayuda a descubrir patrones útiles.

2.2 K-NEAREST NEIGHBORS (KNN)

Es un modelo simple y eficaz que no requiere parámetros. El proceso de clasificación de KNN consiste en realizar el cálculo de la similitud entre un objeto objetivo y los k vecinos más cercanos y similares en el conjunto de muestra de entrenamiento. (1), La distancia de similitud de KNN normalmente se mide por la distancia euclidiana

$$d(x, xi) = \sqrt{\sum_{i=1}^{n} (x - xi)^2} \tag{1}$$

Donde x es el objetivo y xi es el i-ésimo vecino similar más cercano. Luego, al estar x más cerca de sus vecinos, el destino se asignará a la clase más común entre sus k vecinos más cercanos.

2.3 REDES NEURONALES ANN (RPROP)

La ANN está inspirada biológicamente en el cerebro humano. Las neuronas están interconectadas en el cerebro humano de manera similar a como los nodos están interconectados en la red neuronal artificial. A diferencia de otros modelos adaptativos, el efecto del proceso de adaptación de RPROP no se ve empañado por la influencia imprevisible del tamaño de la derivada, sino que solo depende del comportamiento temporal de su signo.

Nucleo del proceso de adaptacion de RPROP. 1

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

$$
\begin{aligned}
&if \left(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) > 0\right) then \{ \\
&\quad \Delta_{ij}(t) = minimum\ (\Delta_{ij}(t-1) * \eta^{+}, \Delta_{\max}) \\
&\quad \Delta w_{ij}(t) = -\ sign\ \frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{ij}(t) \\
&\quad w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \\
&\} \\
&else\ if\ \left(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) < 0\right) then \{ \\
&\quad \Delta_{ij}(t) = maximum\ (\Delta_{ij}(t-1) * \eta^{-}, \Delta_{\min}) \\
&\quad w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t-1) \\
&\quad \frac{\partial E}{\partial w_{ij}}(t) = 0 \\
&\} \\
&else\ if\ \left(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) = 0\right) then \{ \\
&\quad \Delta w_{ij}(t) = -\ sign\ \frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{ij}(t) \\
&\quad w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \\
&\} \\
\}
\end{aligned}
$$

Fig. 1. núcleo del proceso de adaptación y aprendizaje de RPROP

Se supone que el operador mínimo (máximo) debe entregar el mínimo (máximo) de dos números. El operador de signo devuelve +1 si el argumento es positivo; –1 si el argumento es negativo; y 0 en caso contrario

2.4 NAIVE BAYES

Este modelo de clasificación permite estimar la probabilidad de una hipótesis a partir de la data experimental. Conforme se suma más data a la muestra, se va ajustando la probabilidad resultante.

2.5 XGBOOST

Las redes neuronales artificiales superan en su mayoría a otros marcos o algoritmos cuando predicen problemas con texto, imágenes y otros datos no estructurados. XGBoost (Extreme Gradient Boosting) es una implementación avanzada del algoritmo de refuerzo de gradiente, y ha demostrado ser un algoritmo de ML muy eficaz, ampliamente utilizado en competiciones de aprendizaje automático y hackathones. Tiene un alto poder de predicción y es casi diez veces más rápido que las otras técnicas de gradient boosting. También incluye una variedad de regularización que reduce el sobreajuste y mejora el rendimiento general, de ahí que también se conozca como técnica de refuerzo regularizado

3. METODOLOGÍA

3.1 Descripción del conjunto de datos

- Número de observaciones: 41,188.
- Número de muestras realizadas para el presente estudio:
  - · Entrenamiento: 28,831 observaciones, correspondientes al 70% del conjunto de datos.
  - · Prueba: 12,357 observaciones, correspondientes al 30% del conjunto de datos.
- Número de variables de entrada: 20.
- Número de variables de salida: 1.

Table 1. Descripción de variables

N.o	Variable	Escala	Tipo de variable	Tipo de dato
1	Edad del cliente	Cuantitativa discreta	Entrada	Cuantitativa discreta
2	Ocupación	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
3	Estado civil	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
4	Educación	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
5	Mora bancaria	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
6	Crédito hipotecario	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
7	Crédito personal	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
8	Medio de contacto	Cualitativa dicotómica nominal	Entrada	Cualitativa dicotómica nominal
9	Mes de último contacto	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
10	Día de la semana del último contacto	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
11	Duración del último contacto	Cuantitativa discreta	Entrada	Cuantitativa discreta
12	Contactos en esta campaña	Cuantitativa discreta	Entrada	Cuantitativa discreta
13	Número de días desde que fue contactado la última vez	Cuantitativa discreta	Entrada	Cuantitativa discreta
14	Número de contactos antes de esta campaña	Cuantitativa discreta	Entrada	Cuantitativa discreta
15	Resultado de campañas previas	Cualitativa politómica nominal	Entrada	Cualitativa politómica nominal
16	Tasa de variación del empleo	Cuantitativa continua	Entrada	Cuantitativa continua
17	Índice de precios al consumo	Cuantitativa continua	Entrada	Cuantitativa continua
18	Índice de confianza del consumidor	Cuantitativa continua	Entrada	Cuantitativa continua
19	Índice euribor a 3 meses	Cuantitativa continua	Entrada	Cuantitativa continua
20	Número de empleados	Cuantitativa discreta	Entrada	Cuantitativa discreta
21	Cliente suscribió depósito en cuenta	Cualitativa dicotómica nominal	Salida	Cualitativa dicotómica nominal

3.2 PREPARACIÓN DE LOS DATOS

Teniendo en cuenta que el número de clientes que aceptan campañas de telemarketing es mucho menor que aquellos que las rechazan, el dataset se encontraba desbalancead.

Debido a que varios de los modelos de clasificación utilizados requieren operar con datos numéricos en lugar de cadenas de caracteres, también se convirtieron los atributos de tipo cualitativo dicotómico o politómico a escala numérica. Se filtraron adicionalmente en el conjunto de datos final atributos invariables de contexto social y económico.

4. PRUEBAS

A partir del conjunto de datos definidos, se implementaron los modelos de decision tree, KNN, naive Bayes, ANN (RPROP) y XGBoost. Estos modelos fueron seleccionados con el objetivo de obtener el modelo predictivo de la mayor exactitud posible. La Tabla 2 muestra la matriz de confusión de los cinco modelos desarrollados y sus respectivas gráficas en la Figura 2. Por su parte, en la Tabla 3 se aprecian las métricas de desempeño de los algoritmos evaluados. En ambos casos, la información se obtuvo luego de balancear el dataset de origen

Table 2. Matriz de confusión de los algoritmos

Algoritmo	Decision Tree	KNN	Naive Bayes	ANN (RPROP)	XGBoost
No	9950	966	9112	1804	9656
Sí	589	10,424	224	10,789	6,477

Table 3. Resultados de métricas de desempeño de los algoritmos

Algoritmo	Decision Tree	KNN	Naive Bayes	ANN (RPROP)	XGBoost
Accuracy	0.929	0.908	0.647	0.840	0.934
Recall	0.947	0.980	0.412	0.859	0.940
Precision	0.915	0.857	0.783	0.829	0.930
F-measure	0.931	0.914	0.540	0.843	0.935

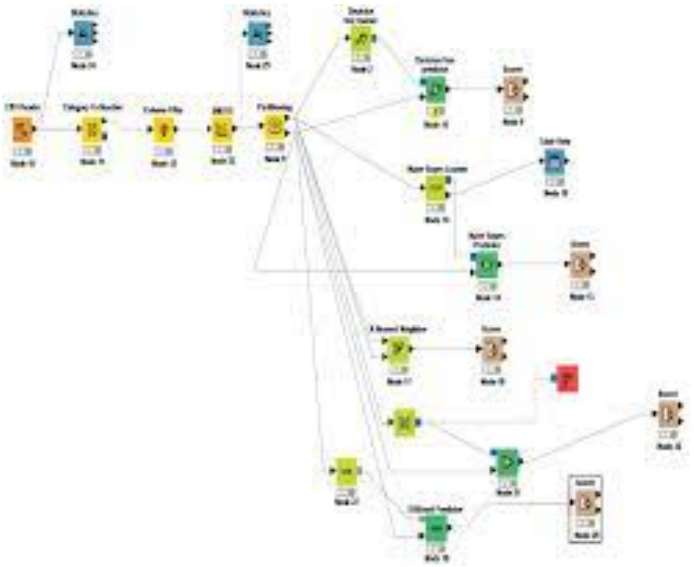

Fig. 2. Modelo del flujo de informacion analizada.

5. ANÁLISIS DE RESULTADOS

En este artículo se realizó una revisión completa de los algoritmos decision tree, KNN, naive Bayes, ANN (RPROP) y XGBoost, así como su aplicación para la predicción de decisiones en la priorización de contactos con clientes en campañas comerciales para la venta del producto de depósitos a plazo.

Para evaluar el desempeño de los cinco algoritmos en análisis, se seleccionaron las métricas de accuracy, recall, precision y la medida F1 para la clase objetivo, es decir, aquellos clientes de interés para el banco por ser potenciales ahorristas y que estarían dispuestos a aceptar la campaña. Los resultados de esta evaluación se muestran en la Tabla 3.

Para la métrica accuracy, el algoritmo XGBoost es el que muestra un mayor valor con 0,934, seguido de decision tree con 0,929; luego KNN con 0,908; ANN (RPROP) con 0,840; y, finalmente, naive Bayes con 0,647. Para la métrica recall, el algoritmo KNN obtiene el mayor valor con 0,980, seguido por decision tree con 0,947; luego ANN (RPROP) con 0,859; y, finalmente, naive Bayes con 0,412. En caso de la métrica precision, el algoritmo XGBoost obtiene 0,930, seguido de decision tree con 0,915; luego KNN con 0,857; luego ANN (RPROP) con 0,829; y, finalmente, naive Bayes con 0,783.

6. FUTUROS TRABAJOS

Uno de los objetivos de los sistemas de información es su alineamiento con las estrategias de negocios, y la arquitectura empresarial es la disciplina que permite a la organización que sus recursos de tecnologías de la información respondan adecuadamente a las fuerzas disruptivas de su entorno.(Seleeme, Fakieh, 2020).

En este sentido, consideramos que un análisis bajo el enfoque de la arquitectura empresarial es necesario para identificar los siguientes pasos en la evolución de soluciones de campañas de marketing en instituciones financieras, para extender el modelo más allá de los modelos algorítmicos de machine learning, a fin de integrar este componente con las demás aplicaciones y procesos de la institución que se relacionan con la comercialización de productos bancarios.

La capa de infraestructura, si bien las instituciones financieras usualmente han favorecido la utilización de recursos on-premise, las opciones en nube permiten acceder a recursos para atender integralmente un proceso de campañas. La habilitación de elementos de integración a las aplicaciones core financieras permitiría que una capa de campañas que resida en la nube pueda acceder a la información para identificar eventos que gatillen las ofertas a los clientes.

La utilización de recursos como data lakes de marketing permitiría consolidar masivamente la información de manera que pueda ser aprovechada en múltiples usos posteriores, sin necesidad de una transformación previa

7. CONCLUSIONES

La clasificación de datos mediante técnicas de machine learning se puede utilizar para mejorar la eficacia en la toma de decisiones de los responsables de las áreas comerciales en las entidades financieras, según las variables seleccionadas y sus ponderaciones Dado que el conjunto de datos de entrada es desbalanceado, por el menor ratio de

clientes que acepta la campaña comercial respecto a los que la rechazan, fue necesario realizar un balanceo de los datos utilizando la técnica SMOTE (Martínez Heras, 2020).

8. REFERENCIA

Reyes, G. T. R., Coral, X. A. G., y Marchinares, A. E. H. (2022). Aplicación de machine learning para campañas de marketing en la banca comercial. Interfases, 016, 187-200. https://doi.org/10.26439/interfases2022.n016.5953

Arango Serna, M. D., Londoño Salazar, J. E., Zapata Cortés, J. A. (2010). Arquitectura empresarial: una visión general. Revista Ingenierías Universidad de Medellín, 9(16), 101-111.

Asha, R. B., Kumar, K. R. (2021). Credit card fraud detection using artificial neural network. Global Transitions Proceedings, 2(1), 35-41. https://doi.org/10.1016/j.gltp.2021.01.006

Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. En Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics (vol. 1, pp. 403-412). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20473-1

Data Science Team. (2019). XGBoost. ¿Qué es? https://Datascience.Eu/Es/Programacion/Xgboost-4/

Dutta, S., Bose, P., Goyal, V., Bandyopadhyay, S. K. (2021). Applying convolutional-GRU for term deposit likelihood prediction. International Journal of Engineering and Management Research, 11(3), 265-272. https://doi.org/10.20944/preprints202007.0101.v1

Goethals, F. G., Snoeck, M., Lemahieu, W., Vandenbulcke, J. (2006). Management and enterprise architecture click: The FAD(E)E framework. Information Systems Frontiers, 8(2), 67-79. https://doi.org/10.1007/s10796-006-7971-1

Martínez Heras, J. (2020, 9 de octubre). Precision, recall, F1, accuracy en clasificación. IArtificial.net. https://www.iartificial.net/precision-recall-f1-accuracy-en- clasificacion/