Comparación de modelos para la predicción de pérdida de clientes en telecomunicaciones

Abel Barrios Córdova, David Márquez Cruz, Israel Martínez Jiménez

BEDU Santander, México

Fecha: 07/02/2024

Resumen— Existe en la actualidad una enorme competencia en diversos ámbitos a nivel empresarial, ya que cada vez existen más y diversas soluciones a necesidades de la vida humana en sociedad, como son las áreas de la salud, telecomunicaciones , banca, seguros, etc. Pero incluso si el mercado se vislumbra como completamente saturado y las tasas de crecimiento de los clientes son bajas, se puede optar aún por la retención y el control de rotación o pérdida de los clientes. En este sentido la Estadística y la Inteligencia Artificial, juegan un papel importante al generar herramientas de predicción que permiten identificar factores que las empresas toman como base para tomar las medidas necesarias para reducir esta pérdida. En el presente trabajo se compararon 10 modelos de predicción: Logístico (log), logístico con SMOTE (logS), Arboles de clasificación y regresión (CART, por sus siglas en inglés) (rpart), (rpart2), (rf), (bag) y (gbm), máquinas de soporte vectorial (SVM Radial), AdaBoost (ada) y Naive Bayes (bay). Se comparó su rendimiento, por medio de la medida estándar del Área bajo la curva (AUC, por sus siglas en inglés), que permite conocer la sensibilidad del modelo (positivos verdaderos) que es de mayor interés en este caso. Se encontró que los mejores modelos fueron: Bagging, Random Forest, Gradient Boosting y AdaBoost.

Palabras clave—abandono, pérdida de clientes, deserción de clientes

Abstract— Currently there is enormous competition in various areas at the business level, there are more and more solutions to the needs of human life in society, such as the areas of health, telecommunications, banking, insurance, etc. But even if the market looks to be fully saturated and customer growth rates are low, you can still choose to retain and control customer turnover or loss. In this sense, Statistics and Artificial Intelligence play an important role in generating prediction tools that allow identifying factors that companies take as a basis to take the necessary measures to reduce this loss. In the present work, 10 prediction models were compared: Logistic (log), logistic with SMOTE (logS), Classification and regression trees (CART) (rpart), (rpart2), (rf), (bag) and (gbm), vector support machines (SVM Radial), AdaBoost (ada) and Naive Bayes (bay). Its performance was compared by means of the standard measurement of the Area under the curve (AUC), which allows knowing the sensitivity of the model (true positives) that is of greatest interest in this case. The best models were found to be: Bagging, Random Forest, Gradient Boosting and AdaBoost.

Keywords — churn, customer churn, customer attrition.

INTRODUCCIÓN

La deserción de clientes aparece cuando los clientes dejan de usar los servicios o adquirir los productos de una empresa, y generalmente es un promedio mensual, trimestral o semestral. Por lo que a la empresa le apremia conocer y monitorear la tasa de abandono, para implementar estrategias de mejora y retención, como mejorar su servicio al cliente, capacitación y sensibilización de personal o estrategia de marketing. En este sentido es importante mencionar que un alto índice de deserción también puede deberse a fallas en la adquisición de nuevos clientes, se debe tener muy claro que tipo de cliente se ajusta de mejor manera a la empresa (cliente clave), ofrecer los servicios y productos adecuados con el fin de generar lealtad y fidelidad. Esto permite que incluso si existieran ofertas de menores costos por parte de la competencia, el cliente se sienta satisfecho y no opte por el cambio o abandono de nuestra empresa.

Reichheld (2001) así como Kisioglu y Topcu (2011), mencionan que captar nuevos clientes es más costoso que idear estrategias de retención, y además menciona que al aumentar las tasas de retención en un 5 %, aumentarán las ganancias en un 25 % e incluso hasta un 95 %. Sin embargo existen causas de deserción que las empresas no pueden controlar o evitar de forma directa, como lo es un desastre natural, migración o muerte; debido a ello se busca analizar factores que estén bajo la influencia de la empresa por medio de modelos estadísticos que tengan la potencia de predicción al evaluar la tendencia del riesgo de abandono, y ofreciendo un índice de

prioridades o características como factores potenciales.

Diversos trabajos se encuentran al respecto de estudios de predicción en casos de abandono en empresas de telecomunicaciones, de los que podemos mencionar algunos de ellos como: Huang et al. (2012), Yildiz y Varlı (2015), (Ahmed y Maheswari, 2017) y (Ahmad et al., 2019), entre muchos otros.

Marco Teórico

Umayaparvathi y Iyakutti (2016), mencionan que en diversas investigaciones se ha empleado la tecnología de aprendizaje automático y minería de datos de forma altamente eficiente. En la figura 1 podemos observar la metodología empleada en el presente estudio, que consta de cinco fases: 1)Preprocesamiento de los registros de entrada del cliente, 2) Extracción de factores para desarrollar los modelos, 3)Construcción de modelos utilizando diferentes clasificadores y cross validation, 4) Cálculo desempeño de los modelos (sensibilidad y especificidad), así como los factores mas importantes, y 5) Políticas de retención de clientes para ejecutivos de CRM.

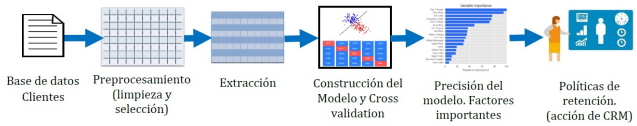

Fig. 1: Fases de un modelo de sistema de predicción de abandono. modificado de: Umayaparvathi y Iyakutti (2016)

Preprocesamiento de los registros de los clientes

Partimos de un sistema de datos, que puede ser tan extenso como sea posible y estar en diversos formatos; se realiza primeramente una exploración de los datos revisando los siguientes rubros:

- Datos faltantes. Revisar si se ignoran o imputan.
- Errores. Pueden ser valores imposibles, outlayers.
- Distribución. Transformarse, escalarse o centrarse.

Extracción de características

Realizar un análisis exploratorio de los datos para seleccionar los factores que tengan cierta evidencia de impacto sobre la decisión de abandono, esto permitirá mejorar el rendimiento predictivo de los modelos. Se pueden elegir diferentes conjuntos de factores y verificar su impacto, también se puede hacer uso de análisis estadísticos como selección de variables como stepwise, fordward o backward. De las variables predictoras podemos encontrar las siguientes:

- Demografía y datos personales (edad, sexo, entidad, etc.)
- Estadísticas de llamadas (duración, horarios, etc.)
- Facturación y pagos.
- Productos agregados (buzón, tarifa especial, etc.)
- Atención al cliente y quejas.

Construcción de modelos

Se emplearon diez modelos que se describen a continuación:

Regresión logística

Propuesta por Hosmer y Lemeshow (1989), predice el resultado de una variable categórica (un evento) en función de variables predictoras. El modelo se puede escribir como:

$$p_i = \frac{1}{1+\exp^{-(\beta_0+\beta_1 x_{1,k}+\cdots+\beta_k x_{k,i})}}$$

donde p_i es la probabilidad de que se produzca un evento, $x_1, x_2, \cdots, x_k$ son las variables independientes, $\beta_0, \beta_1, \cdots, \beta_k$ son los coeficientes de la regresión decisión.

SMOTE

Revisando los datos de este tipo se observa que son datos muy desbalanceados, ya que la proporción de abandono es muy pequeña comparada con la proporción de clientes activos, existe una proporción desequilibrada de clases. Esto se puede encontrar muy a menudo en la vida diaria, por ejemplo en estudios de enfermedad donde sólo una pequeña parte de la población está enferma, o en el caso de control de calidad de procesos al detectar productos defectuosos que casi siempre son una pequeña parte de la producción total.

Para resolver este inconveniente podemos emplear diversas técnicas como son: Submuestreo, sobremuestreo y generación de datos sintéticos, en este último existe una técnica muy utilizada para datos desequilibrados, la técnica de sobremuestreo de minorías sintéticas (SMOTE, synthetic minority oversampling technique) propuesta por Chawla et al. (2002), crea datos artificiales basados en similitudes de espacio de características y no del espacio de datos de las muestras minoritarias, empleando bootstrap y k nearest neighbor. Se utilizó esta técnica para balancear los datos de entrenamiento y prueba, para posteriormente aplicar la regresión logística de forma habitual.

Arboles de clasificación y regresión

Los Arboles de clasificación y regresión (CART, Classification And Regression Trees), son métodos que se pueden emplear para variables respuesta continuas o discretas, si es el caso continua emplean la regresión, si es el caso discreto usan clasificación. Además de que se pueden emplear con fines predictivos como en el presente caso, así como para fines explicativos de situaciones sobre decisiones. Son usados ampliamente ya que generan una representación gráfica de ramificación de tipo "árbol"que posibilita explicaciones simples para situaciones problemáticas donde el modelo lineal está muy limitado.

Existen una clasificación muy grande de estos algoritmos como son: Random Forest, Bagging, Rotation Forest, etc. Para más información sobre los modelos CART consulte Rokach y Maimon (2008). En el presente tarabajo se emplaron los algoritmos MASS none Recursive partitioning (rpart y rpart2), Random forest (rf), Bagging (treebag) y Gradient Boosting trees (gbm), implementados en la función *train* (library caret) de R Core Team (2019).

Máquina de vectores de soporte

Los algoritmos de las máquinas de vectores de soporte fueron desarrollados por Ben-Hur et al. (2001), que sirven para clasificación o regresión. La SVM es una red estática de funciones kernel de clasificación lineal, que transforma los datos a vectores de muy alta dimensión, y que buscan un hiperplano que separe de forma óptima esos puntos proyectados. Si no se puede realizar la separación lineal, entonces encuentra el hiperplano que maximiza el margen y minimiza una función del número de clasificaciones incorrectas (término de penalización de la función). En el caso de SVM radiales, los kernel son funciones de base radial, que fueron las empleadas en este estudio.

AdaBoost

El algoritmo de aprendizaje automático Adaptive Boosting (AdaBost) fue propuesto por Freund y Schapire (1995), los algoritmos del tipo Boosting reducen la varianza y el sesgo, son clasificadores fuertes ya que pueden combinar los resultados de varios clasificadores débiles. En este sentido el algoritmo AdaBoost es uno de los Boosting más importantes hasta la fecha, pero cada vez surgen más como XGBoost, LPBoost, por mencionar algunos y que tienen resultados muy interesantes en diversas aplicaciones.

Naive Bayes

El clasificador Naive Bayes es un clasificador probabilístico débil basado en la aplicación del teorema de Bayes con fuertes supuestos de independencia (naive), de aquí su nombre. Naive Bayes clasifica por medio de modelos que asignan etiquetas a situaciones problemáticas, representadas como vectores de valores de características, y el valor de una característica particular es independiente del valor de cualquier otra característica, dada la variable de clase. Además de lo anterior para generar un clasificador de Naive Bayes se necesita una regla de decisión como por ejemplo el máximo a posteriori. Ha tenido gran uso debido a su facilidad de implementación y a su relativo éxito en diversas aplicaciones.

Cálculo desempeño de los modelos

Podemos revisar el rendimiento de los modelos por medio de dos métricas: positivos verdaderos (sensibilidad) y falsos positivos (especificidad). Note que en este caso nos interesa más la sensibilidad ya que deseamos predecir los positivos verdaderos (clientes que efectivamente abandonan), por lo que se procedió a calcular el AUC o área bajo la curva que nos da una métrica precisa al respecto.

DESARROLLO

Preprocesamiento de los registros de los clientes

Se tiene una base de datos en una tabla de Excel, que contiene 3333 registros de clientes, 20 variables predictoras (4 categóricas y 16 continuas), y una variable respuesta (churn = FALSE / TRUE). Se verificó y no se encontró ningún dato faltante ni atípico. Se observa un abandono de clientes de 14.5 % aprox.como se tenía previsto es un caso de desequilibrio, en la figura 2 se observa esta característica.

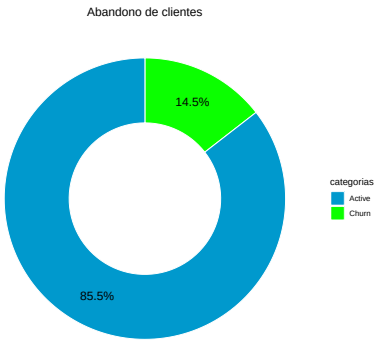

Fig. 2: Proporción de clientes activos y no activos

Extracción de características

Realizo un análisis exploratorio de las variables predictoras, se observa que *phone.number* es un valor singular para cada cliente por lo que no influye en la decisión, así que se elimina. En la parte de *state* se observa en la figura 3 que hay una deserción más alta que el promedio, por lo que no se considera, pero que sirve de referencia para darle mayor relevancia a las medidas a implementar en estos estados.

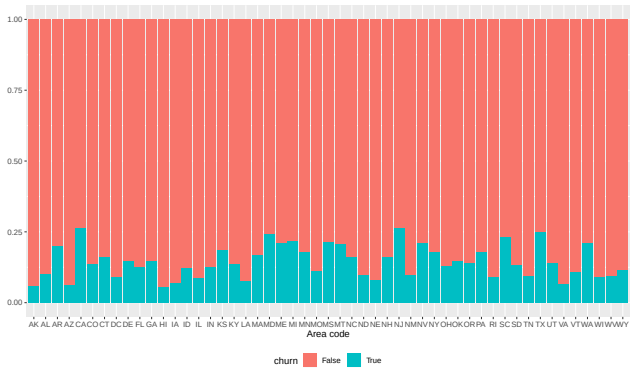

Fig. 3: Promedio de deserción por estado

Variables categóricas

Ahora se revisan las variables categóricas que son: *area.code*, *international.plan*, *voice.mail.plan* y *customer.service.call*. En la figura 4 se observa que la deserción por *area.code* es la misma, por lo que no se considera.

Se observa que el abandono es ligeramente más alto cuando no se cuenta con el servicio de *voice.mail.plan*. En el caso de contar con *international.plan* incrementa hasta en tres veces la tasa de deserción. Y por último es interesante y lógico a la vez observar que cuando se hacen más de cuatro llamadas a la tasa de abandono en el plan internacional es 4 veces el promedio y hay un aumento significativo en el abandono a medida que las llamadas de los clientes aumentan más allá de 4 llamadas a *customer.service.call* el abandono crece significativamente, esto tiene mucho sentido.

Variables continuas

En el análisis de variables continuas podemos iniciar con una matriz de correlación entre las variables figura 5. Es evidente la correlación perfecta que se observa en cuatro pares

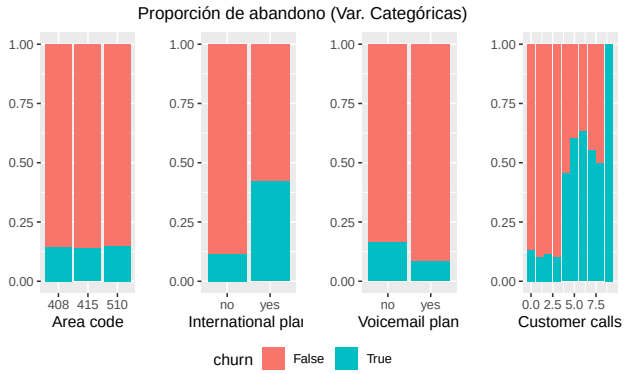

Fig. 4: Análisis de variables categóricas

de variables que son:

- total.day.minutes / total.day.charge
- total.eve.minutes / total.eve.charge
- total.night.minutes / total.night.charge
- total.intl.minutes / total.intl.charge

Por lo anterior no es necesario manejar todas las variables por lo que se elige eliminar las variables de minutos de llamadas, esto tiene sentido debido a que los minutos de llamadas están directamente relacionados con los cargos en la factura.

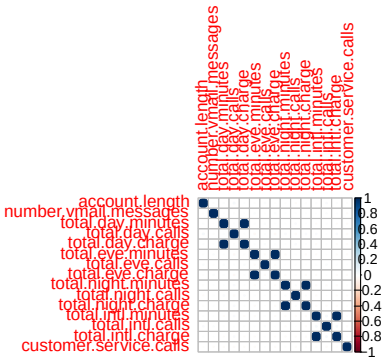

Fig. 5: Correlación de todas las variables continuas

Después de eliminar las variables reviso nuevamente la correlación y verifico la distribución de cada variable. Un comentario sería saber cuanto tiempo abarca el registro *account.length* ya que no se observa que a mayor valor de esta los demás valores aumenten, por la falta de información se considera dejar esta variable.

Construcción de modelos

Una vez seleccionado el conjunto de variables predictoras a emplear para el modelado, se procede a realizar una partición de los datos (80/20), esta partición es necesaria ya que la primer partición (80 %) es para los datos de entrenamiento de los modelos y la partición de 20 % es para probar o validar el desempeño, también se puede elegir una partición (75/25). Este procedimiento se realizó por medio del paquete *caret* de R Core Team (2019). Y se verifica que la proporción de abandono sea la misma que en los datos originales.

Es de suma importancia considerar en el aprendizaje automático el concepto de k-fold Cross-validation o validación cruzada, con esto controlamos la precisión del modelo y evitamos sobreajuste, además aseguramos que los resultados del análisis son independientes de la partición de datos de entrenamiento y validación. En este estudio usaremos un valor de k = 10, para que la estimación sea robusta.

Para todos los modelos se empleó la función *train* del paquete *caret* de R Core Team (2019), con los siguientes algoritmos:

TABLA 1: ALGORITMOS UTILIZADOS

Modelo	Algoritmo
logístico (log)	glm - binomial
logístico + SMOTE (logS)	glm - binomial
Recursive partitioning (rpart)	rpart
Recursive partitioning (rpart2)	rpart2
Random forest (rf)	rf
Bagging (bag)	(treebag)
Gradient Boosting trees (gbm)	gbm
SVM Radial (svm)	svmRadial
AdaBoost (ada)	ada
Naive Bayes (bay)	nb

Además para el modelo logS se hizo un sobremuestreo del conjunto de prueba con la función *SMOTE* del paquete *DMwR*.

Cálculo desempeño de los modelos

Con la función *varImp* del paquete *caret*, se realizó un diagrama con los factores más importantes del modelo, se construyó una matriz de confusión con la función *conf_mat* del paquete *yardstick* para observar la eficiencia y sensibilidad del modelo. Por último se elaboró una gráfica *ROC* para observar la *AUC* de cada modelo y corroborar su sensibilidad, estas tres gráficas se obtuvieron con ayuda del paquete *ggplot2*.

RESULTADOS

Los resultados para cada modelo incluyen 3 secciones: A. Características importantes, B. Matriz de confusión y C. Gráfica AUC.

logístico (log)

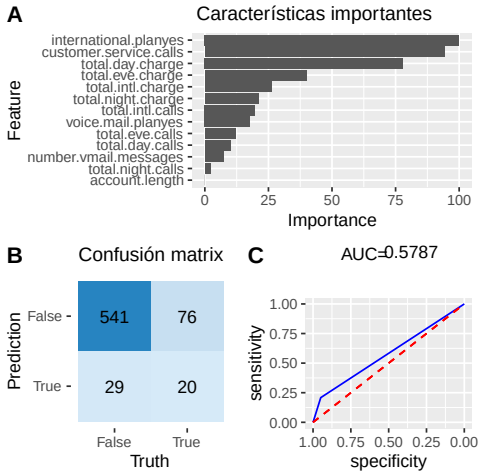

Fig. 6: Rendimiento modelo logístico

logístico + SMOTE (logS)

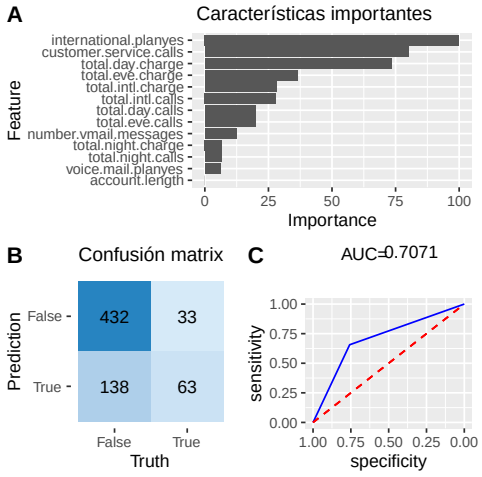

Fig. 7: Rendimiento modelo logístico + SMOTE

CART Recursive partitioning (rpart)

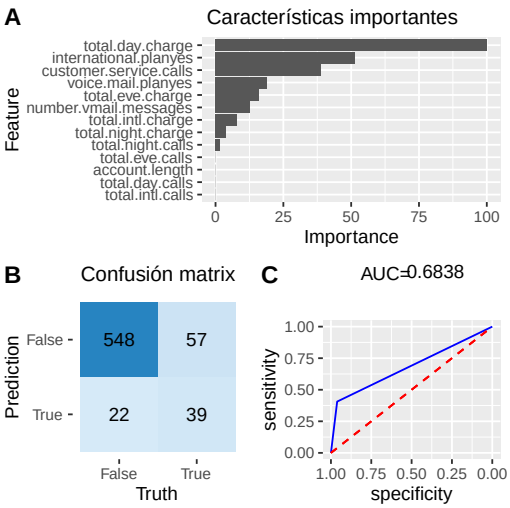

Fig. 8: Rendimiento modelo CART Recursive partitioning (rpart)

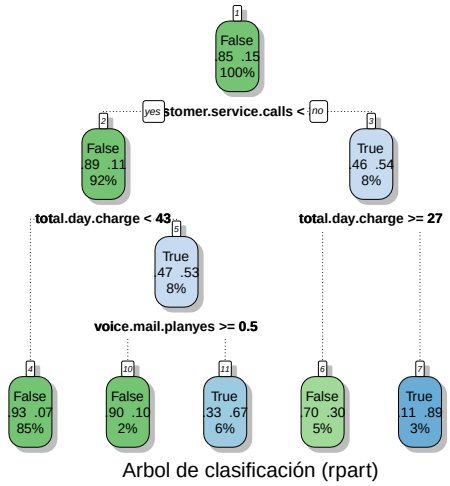

Fig. 9: Árbol del modelo CART Recursive partitioning (rpart)

CART Recursive partitioning (rpart2)

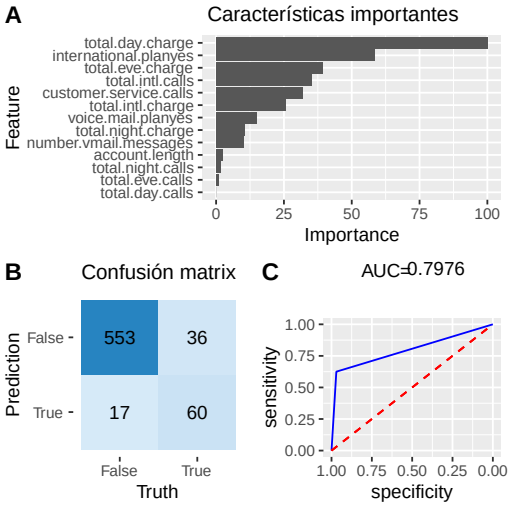

Fig. 10: Rendimiento modelo CART Recursive partitioning (rpart2)

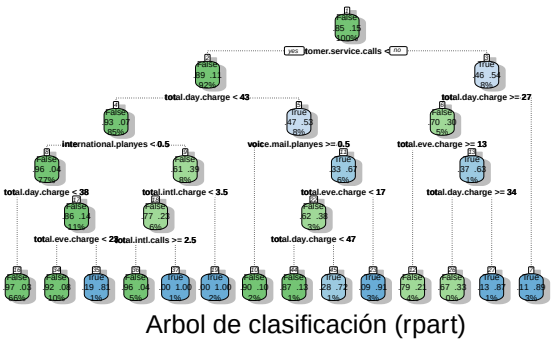

Fig. 11: Árbol del modelo CART Recursive partitioning (rpart2)

CART Random forest (rf)

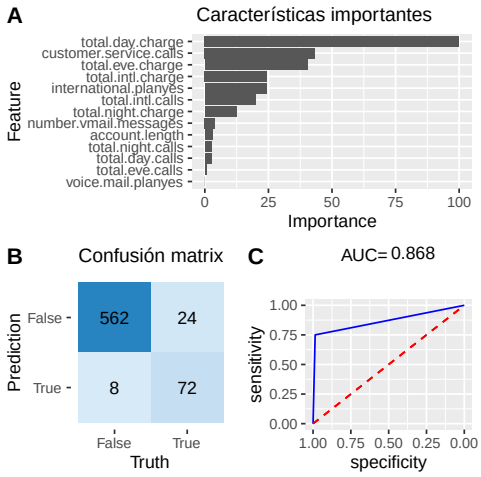

Fig. 12: Rendimiento modelo CART Random forest (rf)

CART Bagging (bag)

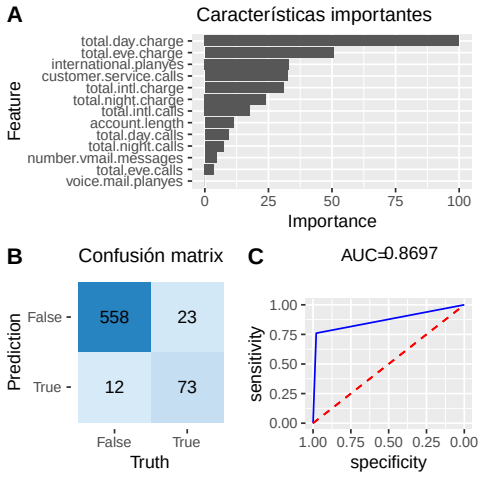

Fig. 13: Rendimiento modelo CART Bagging (bag)

CART Gradient Boosting (gbm)

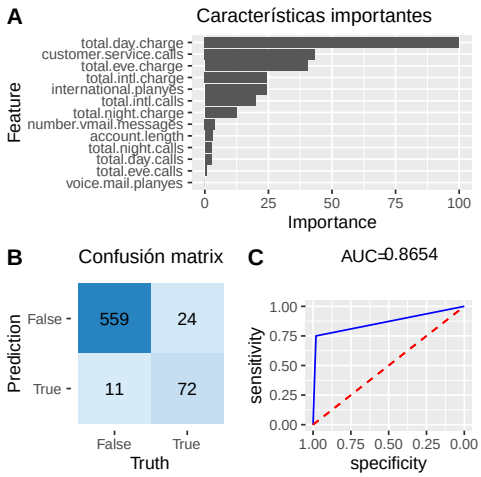

Fig. 14: Rendimiento modelo CART Gradient Boosting (gbm)

SVM Radial (svm)

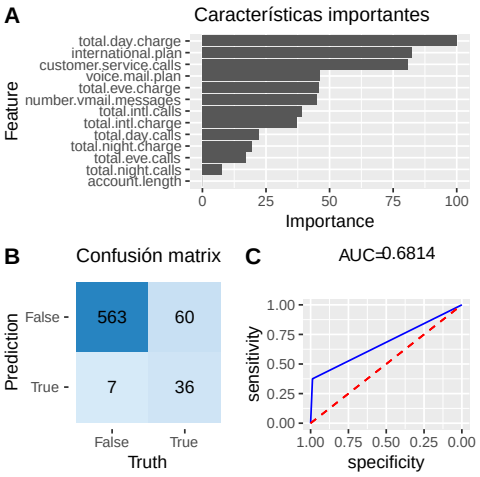

Fig. 15: Rendimiento modelo SVM Radial (svm)

AdaBoost (ada)

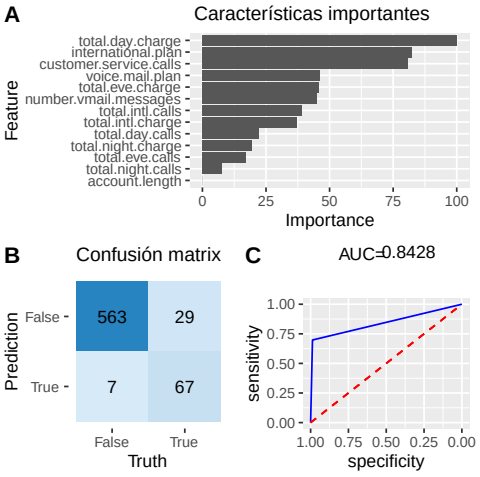

Fig. 16: Rendimiento modelo AdaBoost (ada)

Naive Bayes (bay)

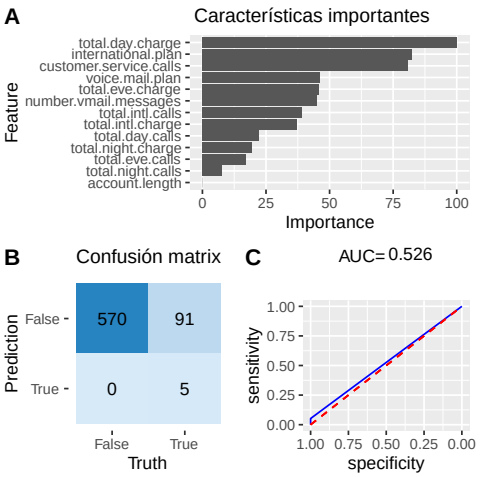

Fig. 17: Rendimiento modelo Naive Bayes (bay)

Selección del mejor modelo

De acuerdo a la métrica elegida que fué el área bajo la curva AUC de cada modelo se elige el que tenga mayor valor, ya que eso indica que tiene mayor sensibilidad en la predicción. En la figura 18 que se encuentra en el apéndice se exhiben las gráficas ROC mostrando los valores AUC para cada modelo, de lo cual seleccionamos a los que tuvieron mayor valor.

TABLA 2: MODELOS SELECCIONADOS

Modelo	AUC
Bagging (bag)	0.8697
Random forest (rf)	0.868
Gradient Boosting trees (gbm)	0.8654
AdaBoost (ada)	0.8428

CONCLUSIONES

Tomando como sustento la métrica usada en el presente trabajo podemos concluir que el mejor modelo predictivo considerando el conjunto de variables predictoras seleccionado y las condiciones mencionadas es el modelo Bagging con una sensibilidad de casi un 87 % seguido del modelo Random Forest con una sensibilidad de 86,8 %, pero se pueden considerar los modelos Gradient Boosting y AdaBoost que lograron un 86,5 y 84,3 % respectivamente. Lo anterior apoyado con los resultados de diversos trabajos mencionados en el presente trabajo, que muestran resultados muy coincidentes respecto a los mejores modelos de predicción.

De acuerdo a las características importantes que arroja cada uno de los métodos se puede idear una estrategia de marketing junto a los ejecutivos de CRM para poder tener un mayor impacto en la retención de los clientes considerando las siguientes acciones:

1. Realizar una encuesta a clientes cuyos pagos y/o minutos de llamadas están por encima del promedio para identificar posibles causas de deserción y tomar las medidas necesarias.

2. Verifique que los clientes con más de dos llamadas a servicio a clientes sea contactado por un grupo con la suficiente experiencia técnica y de marketing para solucionar cualquier situación que lo afecte de forma directa. Además tome acciones de capacitación o implementación técnica en el departamento correspondiente.

3. Contacte a los clientes con plan internacional para identificar debilidades o intenciones de búsqueda de nuevos servicios u oportunidades, puede elaborar un programa de lealtad o de tarifas preferenciales de acuerdo al consumo o atención Premium a este tipo de clientes.

APÉNDICES

En la figura 18 se observan los valores de AUC de todos los modelos presentados en el presente trabajo. Por su extensión la gráfica se muestra en la siguiente página.

REFERENCIAS

Ahmad, A. K., Jafar, A., y Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1):28.

Ahmed, A. A. y Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*, 18(3):215 – 220.

Ben-Hur, A., Horn, D., Siegelmann, H. T., y Vapnik, V. (2001). Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Freund, Y. y Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

Hosmer, D. W. y Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons.

Huang, B., Kechadi, M. T., y Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414 – 1425.

Kisioglu, P. y Topcu, I. (2011). Applying bayesian belief network approach to customer churn analysis: A case study on the telecom industry of turkey. *Expert Syst. Appl.*, 38:7151–7157.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reichheld, F. (2001). Prescription for cutting costs. *Bain & Company, Inc.*, pages 1–3.

Rokach, L. y Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications*, volume 69. World scientific.

Umayaparvathi, V. y Iyakutti, K. (2016). A survey on customer churn prediction in telecom industry: Datasets, methods and metrics. *International Research Journal of Engineering and Technology (IRJET)*, 4(4):1065–1070.

Yildiz, M. y Varlı, S. (2015). Customer churn prediction in telecommunication. *2015 23rd Signal Processing and Communications Applications Conference, SIU 2015 - Proceedings*, pages 256–259.

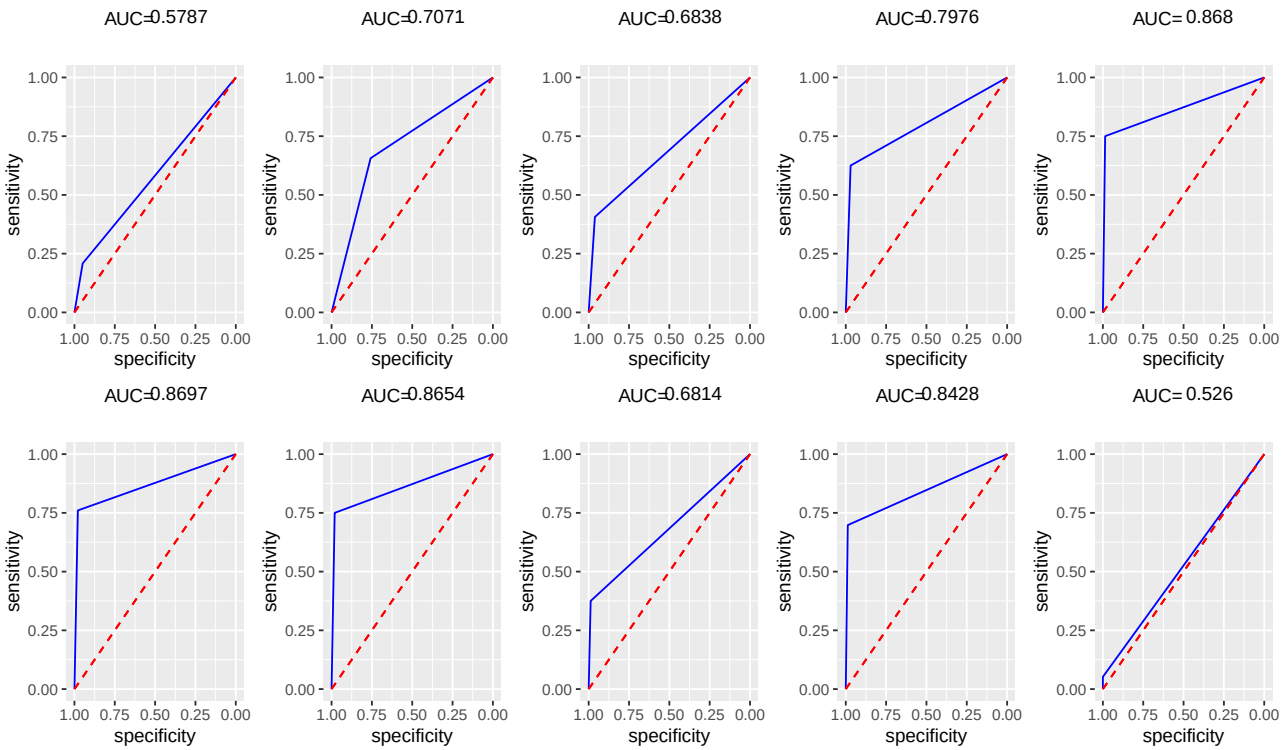

Fig. 18: Valores AUC de todos los modelos