Natura@economía

ISSN 2226-9479 (Versión electrónica)
Website: http://revistas.lamolina.edu.pe/index.php/neu

ARTÍCULO ORIGINAL

Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil

Using the Adaboost algorithm and logistic regression to predict customer churn in a mobile phone company

Aldo Richard Meza Rodríguez[1*]; Jorge Chue Gallardo[2]

[1] Dpto. Estadística e Informática. Facultad de Economía y Planificación. Universidad Nacional Agraria la Molina, Apartado postal 12-056 - La Molina, Lima, Perú. Email: armeza@lamolina.edu.pe; jchue@lamolina.edu.pe

Recepción: 20/06/2020; Aceptación: 15/12/2020

Resumen

El objetivo de esta investigación tiene como propósito comparar un modelo de predicción de fuga de clientes en una empresa de telefonía móvil. El modelo propuesto fue el algoritmo Adaboost, el cual se desarrolla a través de aprendizaje adaptativo. Para probar su eficiencia se comparó con la regresión logística desde la perspectiva de la minería de datos. Como la variable objetivo de respuesta era desbalanceada se utilizó procedimientos de muestreo para equilibrar los datos (sub-muestreo, sobre-muestreo y SMOTE). Las medidas de desempeño para elegir el modelo fueron la precisión, el recall (sensibilidad), el F-mesausre y el AUC (curvas ROC). La precisión, el recall y el F-mesuare arrojaron rendimientos superiores a favor del algoritmo Adaboost, también la medida principal de desempeño dio un AUC=0,93 para el Adaboost, frente a un AUC=0,86 para la regresión logística. Realizadas todas las comparaciones, la validación y las medidas de desempeño, en conclusión, el modelo óptimo para la predicción de fuga de clientes en la empresa de telefonía móvil es el algoritmo Adaboost. Finalmente, con este algoritmo se detectó que las variables más importantes para entender el patrón de fuga de los clientes fueron el tipo de reclamo, rol del cliente, comunidad (relación del cliente con otros contactos), tipo de cliente, número de reclamos, número de llamadas, nota del cliente y MOU.

Palabras clave: Algoritmo Adaboost; regresión logística; datos desbalanceados; medidas de desempeño; curva roc; validación cruzada; fuga de clientes.

Forma de citar el artículo: Meza, A.; Chue, J. 2020. Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil. Natura@economía 5(2):102-117 (2020). http://dx.doi.org/10.21704/ne.v5i2.1610

DOI: http://dx.doi.org/10.21704/ne.v5i2.1610
* Autor de correspondencia: Aldo Richard Meza Rodríguez. Email: earmeza@lamolina.edu.pe

Abstract

The objective of this research is to compare a customer churn prediction model in a mobile phone company. The proposed model is the Adaboost algorithm, which is developed through adaptive learning. To test its effectiveness, it was compared with logistic regression from the perspective of data mining. As the objective response variable was unbalanced, sampling procedures were used to balance the data (subsampling, oversampling and SMOTE). The performance measures for choosing the model were precision, recall (sensitivity), F-mesausre and AUC (ROC curves). The precision, the recall and the F-mesuare yielded superior returns in favor of the Adaboost algorithm, also the main performance measure gave an AUC = 0,93 for the Adaboost, compared to an AUC = 0,86 in the logistic regression. After carrying out all the comparisons, the validation and the performance measures, it was concluded that the optimal model for predicting customer churn in the mobile phone company is the Adaboost algorithm. Finally, with this algorithm it was detected that the most important variables to understand the pattern of customer leakage were the type of claim, role of the customer, community (customer relationship with other contacts), type of customer, number of claims, number call, customer note and MOU.

Keywords: Adaboost algorithm; Logistic regression; unbalanced data; performance measures; roc curve; cross validation; churn.

1. Introducción

La fuga de clientes (Churn en inglés), es un problema frecuente, presente en las empresas, industrias, banca, telecomunicaciones, etc. El abandono o deserción de los clientes se da en presencia de agresivas estrategias de marketing, las cuales buscan captar clientes ofreciendo nuevos y mejores productos y servicios. Según Barrientos (2012), las causas principales de la fuga de clientes en el área de telefonía están relacionadas a factores generados por el servicio de la empresa, tales como la calidad de la cobertura, señal, precio, atención al cliente, etc.

Según cifras de Osiptel (2019), entre julio del 2014 y marzo del 2019, se han realizado 16'648'547 cambios de operador, siendo Entel quien lidera la lista con 1'244'507 líneas ganadas, seguido de Claro con 617'453 clientes obtenidos. Meza (2018) describe que "Ante la necesidad de prever e identificar a los posibles clientes que fugan, los modelos de predicción son herramientas y soporte clave para identificarlos y entender el patrón y el motivo que les lleva a prescindir del servicio".

Entre las diferentes alternativas para identificar las causas y los patrones determinantes para la toma de decisión de fuga, se encuentran los modelos de predicción (Ahmad *et al.,* 2019). Diferentes investigadores en el área de la minería de datos han desarrollado constantemente nuevos modelos de clasificación y predicción, tales como el algoritmo Adaboost, la regresión logística, árboles de clasificación, entre otros. Por ejemplo, Pérez (2014) describió que el algoritmo Adaboost tiene la gran capacidad de mejorar el rendimiento y minimizar el error de predicción. Por su parte Escobar & Loas (2015) probaron una variedad de técnicas de predicción en detección temprana de fuga o deserción de estudiantes, encontrando que el algoritmo Adaboost fue el de mejor rendimiento, obteniendo la menor tasa de error. Además, al realizar un torneo en predicción de modelos de fuga de clientes, donde se evaluaron 33 modelos de predicción, la regresión logística y los árboles de clasificación tuvieron la mejor exactitud predictiva y mejor desempeño (Neslin *et al.* 2006).

Uno de los problemas al crear modelos de clasificación es la distribución o balance de los datos (Mao *et al.*, 2017). El desbalance se da cuando una de las categorías es muy inferior en comparación con la otra, causando grandes efectos negativos sobre los resultados, ocasionando sesgos en el desempeño de la clasificación (Hadad *et al.,* 2009). Ante ello, se debe buscar una técnica que equilibre los datos para luego identificar entre el algoritmo Adaboost y la regresión logística, aquel modelo que tenga la mayor precisión en la predicción, evaluando su capacidad predictiva e identificando las variables que mejor se correlacionen a la fuga de clientes.

Finalmente, esta investigación busca conocer con precisión a través de los modelos porqué los clientes deciden retirarse definitivamente o migrar hacia otros operadores, de tal manera que los responsables de la toma de decisiones en la empresa logren minimizar esfuerzos en la retención y focalizarse en atacar las verdaderas causas de la fuga.

2. Materiales y métodos

Materiales

La presente investigación se desarrolló con información de una empresa de telefonía móvil en el **área de** postpago (por motivos de políticas y privacidad propias de la empresa de telefonía, se mantiene en reserva el nombre) con un conjunto de datos de los registros de los últimos 6 meses del año 2017 (pudiendo adaptarse a los datos actuales), con una muestra total de 80’300 registros (Tabal 1).

Para el procesamiento de datos fue indispensable el uso del software R, en su interface Rstudio y en su versión 4.0.2, utilizando diferentes funciones, las cuales están formalmente en diferentes librerías.

Metodología

Fuga de clientes (Churn)

En el marco de las telecomunicaciones, el Churn se define como “la acción de cancelar el servicio prestado por la compañía” (Pérez, 2014), o como “la propensión de clientes a efectuar el cese de los negocios que tenga con una compañía en un periodo de tiempo determinado” (Lejenue, 2001). En esta cancelación, el cliente decide renunciar a la empresa (voluntaria) o en todo caso la empresa puede expulsarlo por algunas irregularidades o por no cumplir con las obligaciones o pagos acordados (involuntaria). Para Hu (2019), un cliente puede considerase pasivo o activo, dependiendo si la baja es directa o no, por ejemplo, para los pasivos, si se acaba el periodo de suscripción, estos no renuevan o pierden interés en el producto o servicio; los activos en cambio deciden terminar su suscripción y buscan los mecanismos formales para prescindir del servicio o darse de baja.

El modelo de regresión logística binario

La regresión logística formula un modelo para predecir una variable categórica en función a una o más variables predictoras, la cual está inmersa en el grupo de familias que usa una función de enlace llamada logit (Hosmer y Lemeshow, 2000).

El logit de las probabilidades binomiales desconocidas forman el modelo general, bajo la siguiente función lineal:

$$\text{logit}=\log\left(\frac{p}{1-p}\right)=X_i^T\beta=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p=\eta$$

La probabilidad aproximada de pertenencia a cualquiera de las dos categorías en el suceso se aproxima a través de una función logística de la siguiente manera:

$$p_i=\text{logit}^{-1}(\eta)=\frac{e^{\eta}}{1+e^{\eta}}=\frac{1}{e^{-\eta}+1}$$

Tabla 1. Descripción de las variables usadas en los modelos

Variable	Tipo de variable	Descripción de la variable
Churn	Cual. nominal	Abandono del cliente, el cual puede ser por voluntad propia, o porque la empresa decidió cancelarle el contrato (1: Fugado, 2: no fugado)
Minutos de uso (MOU)	Cuant. continua	Ratio de tiempo hablado mensual y/o datos
Días de deuda	Cuant. discreta	Nº de días promedio por mes que adeuda el cliente en los últimos 6 meses.
Reclamos	Cuant. discreta	Nº de reclamos mensual
Tipo de Reclamos	Cuant. continua	Tipo de reclamo más frecuente
Sexo	Cual. nominal	Género del cliente
Edad	Cuant. continua	Edad en años
Procedencia	Cual. nominal	Lugar de procedencia (1: Lima, 2: Provincia)
Rol	Cual. nominal	Líder, seguidor o marginal (se calcula en base a mensajes, llamadas, transferencias entre los usuarios de la misma empresa)
Comunidad	Cual. jerárquica	Representada por niveles según el porcentaje de relación del cliente con todos sus contactos, calculado en función a las redes, seguidores, llamadas, etc., con clientes de la empresa (internos), ejemplo: grado 1 (0-25%), grado 2 (26-50%), grado 3 (51-75%), grado 4 (76-100%)
Plan renuncia	Cuant. discreta	Cantidad de planes renunciados del usuario durante su periodo de vida
Antigüedad	Cuant. continua	Tiempo de afiliación del cliente
Canal	Cuant. discreta	Medio mediante el cual se vendió el plan al cliente (1: empresa, 2: vendedor individual, 3: Ejecutivo de la compañía, 4: no identificado)
Tipo cliente	Cual. jerárquica	Valoración que la empresa da al cliente, (1: bajo, 2: medio bajo, 3: medio, 4: medio alto , 5: alto)
Número mensajes	Cuant. discreta	Nº de mensajes mensual
Llamadas	Cuant. discreta	Nº de llamadas mensual
Kilovatios	Cuant. continua	Cantidad de Kilovatios de uso mensual (uso de datos, correo, internet, etc.)
Ingresos (ARPU)	Cuant. continua	Gasto promedio de usuario con el servicio
Nota de pago	Cuant. continua	Nota mensual en función a su comportamiento de pago y el tiempo que demora en pagar sus cuentas

Esta función es óptima cuando los datos son simétricos, es decir, cuando la variable dependiente categórica tiene una cantidad equilibrada de “ceros” y “unos”, lo que en muchas situaciones prácticas no se cumple, tal es el caso de esta investigación sobre la fuga de clientes, donde hay una gran diferencia entre los que fugan y los que se mantienen con el servicio.

El algoritmo Adaboost

Adaboost o Adaptative Boosting es un algoritmo de aprendizaje automático utilizado para regresión y clasificación (Freund & Schapire, 1996). Existen variadas versiones tales como Adaboost, Adaboost. M1, Adaboost.M2, entre otras (Obregón, 2016), cuya metodología consiste en entrenar en forma iterativa una serie de

clasificadores débiles tal que cada nuevo clasificador dé mayor importancia a los datos mal clasificados en los entrenamientos anteriores, para luego combinar todo el conjunto de clasificadores débiles y obtener un clasificador cuyo rendimiento sea fuerte.

Este algoritmo utiliza funciones que ponderan la importancia en relación a cada dato en el proceso del entrenamiento del clasificador, de esta manera, los datos que se han clasificado correctamente pierden peso a favor de los que fueron clasificados erróneamente, intentando conseguir que los nuevos clasificadores se enfoquen en aquellos datos clasificados erróneamente (Figura 1).

En la Figura 2 se aprecia el mecanismo del algoritmo Adaboost, cada clasificador débil asume un peso "W" mayor en diferentes iteraciones, al final en conjunto se forma el clasificador fuerte.

Datos desbalanceados

Un problema en los modelos de clasificación es el desbalance de datos, esto ocurre cuando una clase o categoría es sesgada (hacia la clase positiva o negativa), por ejemplo, si una empresa cuenta con 10,000 clientes, de los cuales solo 500 han fugado, entonces la categoría "Si fuga" tendría solo el 5%, mientras que la categoría "no fuga" tendría el 95% de los datos. El problema del desbalance es que la clase mayoritaria abruma a los modelos u algoritmos y desvían su rendimiento, los resultados pueden ser muy engañosos puesto que, las clases minoritarias tendrían un efecto mínimo en la precisión, es decir el desempeño de la clase mayoritaria superaría al pobre desempeño de la clase minoritaria (Fernández *et al.*, 2017). Dicho de otra manera, los modelos serían adecuados para predecir a los clientes que no fugan, pero tendría un desempeño bajo al identificar a los que si fugan.

Métodos para equilibrar datos desbalanceados (asimétricos)

Algunos métodos para contrarrestar el efecto de los datos desequilibrados son conocidos como "Métodos de muestreo" (Brownlee, 2015), los cuales equilibran la distribución mediante algún mecanismo propio. Kunal (2016) menciona que los métodos más recomendados para el tratamiento de conjuntos de datos desbalanceados, son el sub-muestreo, el sobre muestreo y el sobre muestreo en minorías sintéticas.

El sub-muetreo: Este método elimina los datos de la clase mayoritaria reduciendo el número de observaciones hasta equilibrar los datos, mejorando el tiempo de ejecución y almacenamiento (Haibo & Yunqian, 2013). Según Liu *et al*. (2006) el sub-muestreo produce buenos resultados, siendo un método que utiliza algoritmos sencillos y fáciles de entender, combinando resultados y entrenando secuencialmente a los clasificadores. Una posible desventaja de este método es que, al eliminar datos, puede haber cierta pérdida de información.

El sobre-muestreo: Este método replica las observaciones de la clase minoritaria, es decir, aumenta en forma aleatoria (realizando muestreo con reemplazo) las observaciones del grupo minoritario. Este proceso se repite hasta que las observaciones de la clase minoritaria y mayoritaria se equilibren. Una ventaja de este método es que no hay pérdida de información, sin embargo, al haber datos duplicados puede ocasionar un posible sobreajuste en el modelo (Kunal, 2016).

Sobre muestreo de minorías sintéticas (SMOTE): Este método utiliza inclusión de nuevas observaciones, creando arbitrariamente nuevas observaciones de la clase minoritaria a partir de los vecinos más cercanos de la clase menor. Esta técnica no

replica las observaciones, sino que erradica el desequilibrio mediante inclusión de datos artificiales, la cual se puede integrar con los algoritmos de clasificación mejorando el rendimiento de la predicción (Ijaz *et al.*, 2018).

En la Figura 2 se aprecia el mecanismo del algoritmo Adaboost, cada clasificador débil asume un peso "W" mayor en diferentes iteraciones, al final en conjunto se forma el clasificador fuerte.

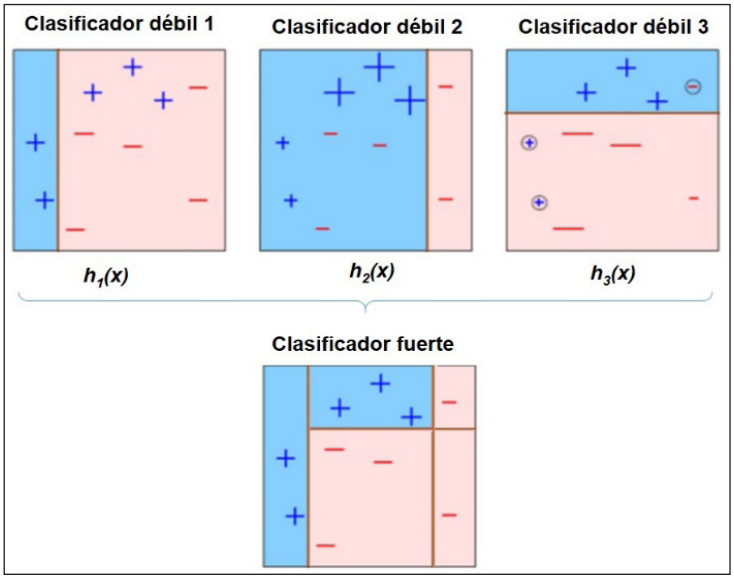

Figura 1. Formación de 3 clasificadores débiles con el algoritmo Adaboost.

Fuente: Freund & Schapire (1996)

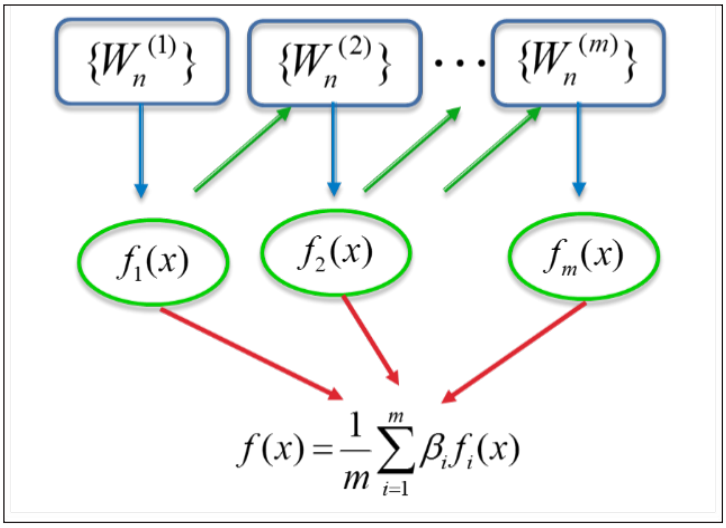

Figura 2. Mecanismo del Adaboost.

Fuente: Bhatia & Chiu (2017)

Validación cruzada

Una de las partes fundamentales y más complicadas en la creación de modelos es la evaluación de los clasificadores, puesto que muchos modelos pueden ajustarse muy bien a los datos con los cuales fueron creados, pero al intentar replicar e implementar el modelo con nuevos datos, estos muestran sobreajuste (el patrón de comportamiento no se puede generalizar) o errores en la clasificación (Obregón, 2016).

Una de las herramientas para detectar el sobreajuste es la técnica de validación cruzada de K iteraciones, la cual consiste en dividir las observaciones en varios subconjuntos, seleccionando aleatoriamente a uno de ellos como prueba y utilizando los demás para entrenar el modelo de clasificación, repitiendo el proceso hasta evaluar cada sub conjunto de datos y así al final se halla la media aritmética de las evaluaciones en cada iteración, obteniendo un solo resultado (Valavi *et al.,* 2019). Una de las desventajas de este método es el costo computacional, puesto que, a mayor número de iteraciones, mayor será el tiempo de procesamiento; en la práctica la elección de la cantidad de iteraciones depende del tamaño del conjunto de datos, siendo lo más frecuente usar 10 iteraciones (10-fold cross-validation).

Medidas para evaluar el rendimiento de los modelos en datos desbalanceados

La medida más utilizada para evaluar el desempeño de un modelo de clasificación es el accuracy (tasa de error), sin embargo, cuando los datos son desbalanceados, el accuracy no es una métrica apropiada, puesto que esta medida puede ser engañosa. Las medidas de desempeño más recomendadas para datos desbalanceados son Recall, Precisión, F-Measure, la ROC y AUC (Wu *et al.,* 2018).

- **Precisión:** Es una medida de la corrección o exactitud obtenida en la predicción positiva, se calcula como el porcentaje de eventos que fueron pronosticados positivos, siendo estos realmente positivas (Powers, 2008).
- **Recall (sensibilidad):** Es una medida que avalúa el porcentaje de observaciones reales que se predicen correctamente o cantidad de observaciones de clase positiva que se predicen correctamente (Chicco & Jurman, 2020).
- **F-measure (medida F):** Es la media armónica ponderada entre el Recall y la precisión (Wu *et al.,* 2018).
- **Curva ROC (**Receiver Operating Characteristic**):** Es una de las métricas principales para evaluar modelos con datos desbalanceados, la cual se forma trazando la sensibilidad (tasa de verdaderos positivos) y la especificidad (tasa de falsos positivos). Uno de los indicadores para comparar las diferentes curvas ROC es el AUC (área bajo la curva), donde el modelo que tenga mayor AUC será posiblemente el de mayor desempeño (Tharwat, 2018).

Matriz de confusión

Esta matriz valora la capacidad predictiva de un modelo y se construye cruzando los datos reales de la variable de respuesta con los datos que fueron predichos en el modelo, en la diagonal se encuentran los valores correctamente clasificados (tanto para la clase positiva como negativa) y fuera de la diagonal las clasificaciones incorrectas. En la matriz de confusión se muestran los verdaderos positivos (VP), los falsos positivos (FV), los verdaderos negativos (VN) y los falsos negativos (FV), los cuales se utilizan para hallar las medidas de desempeño (Hair *et al*., 1999).

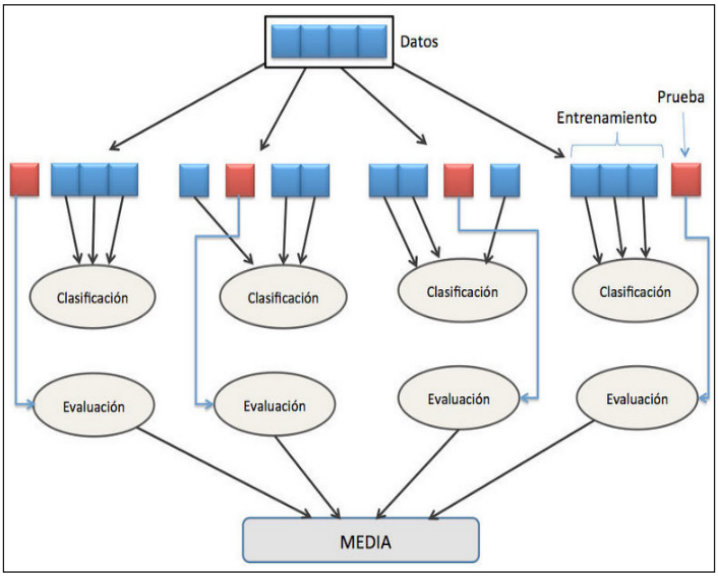

Figura 3. Validación cruzada de K iteraciones con K=4.

Fuente: Lang (2011)

3. Resultados y discusión

En la Tabla 2 muestra la distribución de la variable de respuesta, la presencia de usuarios que fugaron es solo el 10,03%, estando en un claro caso de datos desbalanceados, por cual es necesario subsanar este problema con una técnica de balanceo de datos.

Tabla 2. Distribución de la variable "Fuga" según el tipo de usuario

		Frecuencia	Porcentaje
Usuario	No fugó	72249	89,97
	Sí fugó	8051	10,03
	Total	80300	100

Como se muestra en la Figura 4, las observaciones originales muestran un desbalance en los datos, el sub-muestreo equilibró los datos, pero eliminando valores de la parte mayoritaria, el sobre-muestreo equilibró los datos a costa de la duplicidad de observaciones, mientras que el SMOTE sintetizó los datos en categorías similares.

Al aplicar los métodos de muestreo, los datos se equilibraron (Tabla 3), sin embargo, el sobre-muestreo ha aumentado la cantidad de observaciones pudiendo causar costo computacional en los modelos, por su parte el sub-muestreo ha reducido considerablemente los datos, pudiendo perder información para entender los patrones de predicción; al parecer el método Smote para el estudio de estos datos es el que mantiene el equilibro más exacto.

La Figura 5 muestra la validación cruzada para ambos modelos, con los datos de entrenamiento se realizó 10 particiones, y en cada partición se halló el AUC, se aprecia que en ambos modelos existe estabilidad, puesto que en el recorrido de los 10 *k-fold* los valores del AUC se mantienen cerca del promedio, lo cual indica que los dos modelos no estarán afectados por el sobreajuste. Por otro lado, los valores del AUC del algoritmo Adaboost se encuentran por encima de la regresión logística, dando los primeros indicios que es el modelo de mejor rendimiento.

Tabla 3. Métodos de muestreo y procedimientos para equilibrar los datos de prueba para el modelo de regresión logística y el algoritmo Adaboost

Métodos de muestreo	No Fugó	Sí Fugó	Total	Procedimiento
Sin muestreo	50554	5656	56210	Datos originales
Sobre muestreo (down)	50554	50554	101108	Iguala la categoría alta
Submuestreo (up)	5656	5656	11312	Iguala la categoría baja
Smote	22624	16968	39592	Muestra de minoría sintética

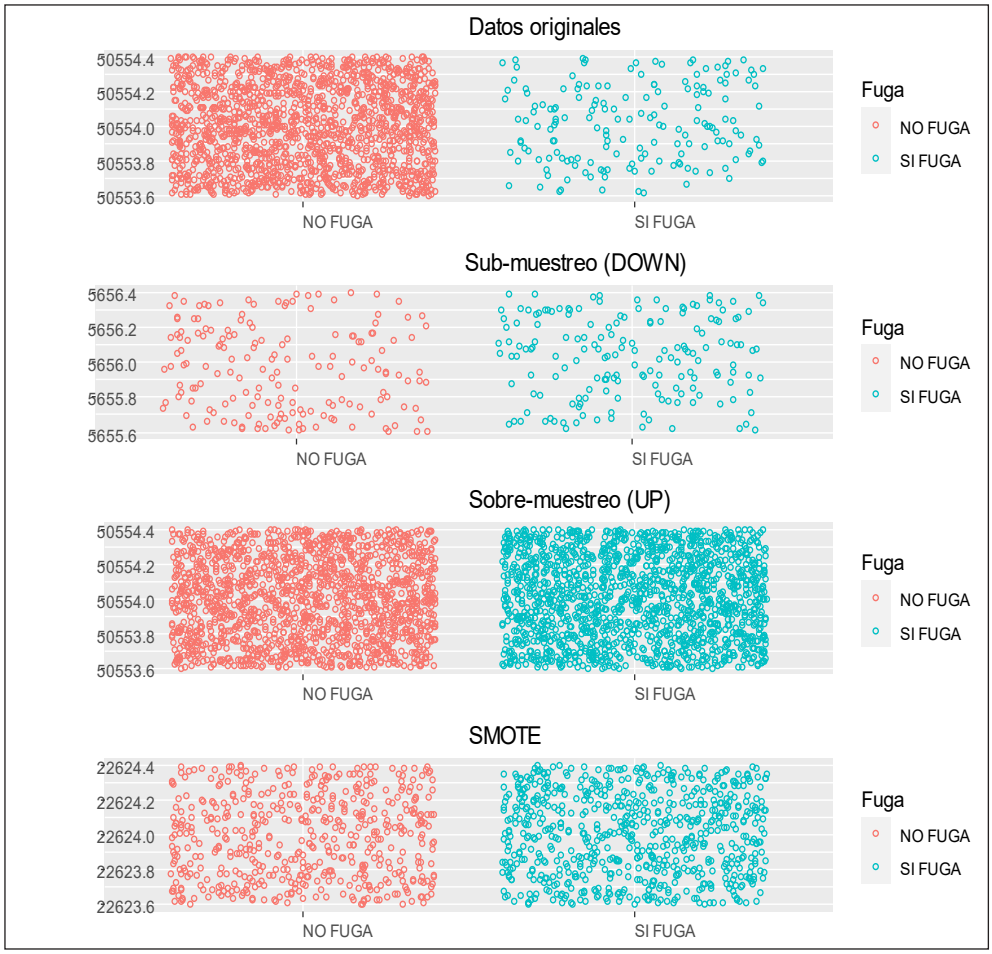

Figura 4. Distribución de la variable de respuesta “Fuga” según tipo de muestreo

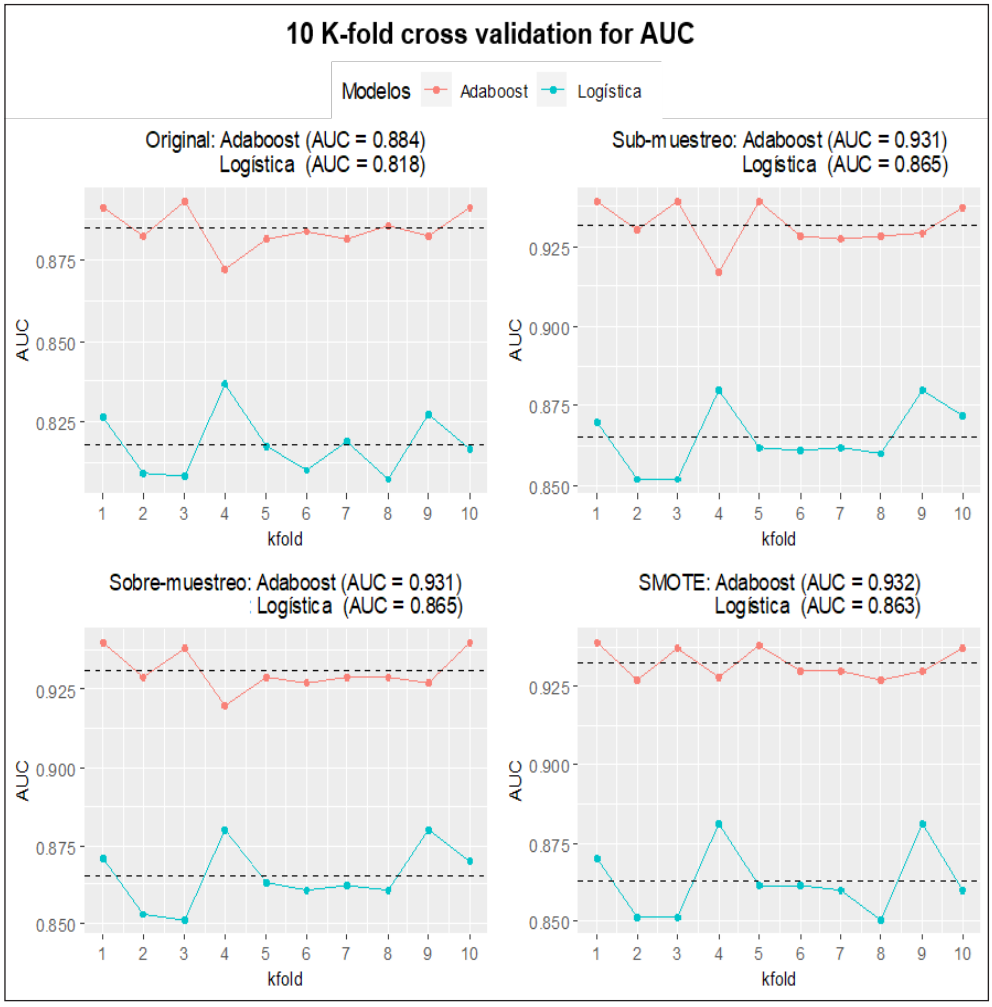

Figura 5. Validación cruzada para el AUC en los modelos Adaboost y regresión logística mediante los métodos de muestreo

Tabla 4. Comparación de las medidas de desempeño con los métodos de muestreo para los modelos de regresión logística y el algoritmo Adaboost

Métricas	Sin muestreo			Sub-muestreo			Sobre-muestreo			SMOTE		
	Logística		Adaboost	Logística		Adaboost	Logística			Logística		Ada- boost
Precisión	0,607	<	0,754	0,282	<	0,365	0,281	<	0,755	0,314	<	0,494
Recall	0,208	<	0,407	0,821	<	0,904	0,822	>	0,409	0,742	<	0,751
F - measure	0,309	<	0,529	0,419	<	0,52	0,419	<	0,53	0,441	<	0,596

Al realizar la comparación de las diferentes métricas de medidas para ambos modelos (Tabla 4), la precisión del algoritmo Adaboost fue superior a la regresión logística en todos los métodos de muestreo, teniendo una mejor capacidad para detectar correctamente a los clientes que fugan, similares hallazgos se encuentran en el Recall (sensibilidad) y el F-measure. Con estas tres medidas el desempeño del algoritmo Adaboost tendrá un mejor rendimiento y mayor capacidad para clasificar correctamente a los clientes.

Al comparar las curvas ROC para cada tipo de muestreo (Figura 6), claramente se observa que el algoritmo Adaboost está por encima de la regresión logística, teniendo aproximadamente 93% de AUC, mientras que la logística 87% aproximadamente; confirmando así que para este estudio cualquiera de los métodos de muestreo aplicados (sobre muestro, sub muestreo o SMOTE) al algoritmo Adaboost va a permitir estimar con mayor precisión y rendimiento la predicción de los clientes que fugan en la empresa.

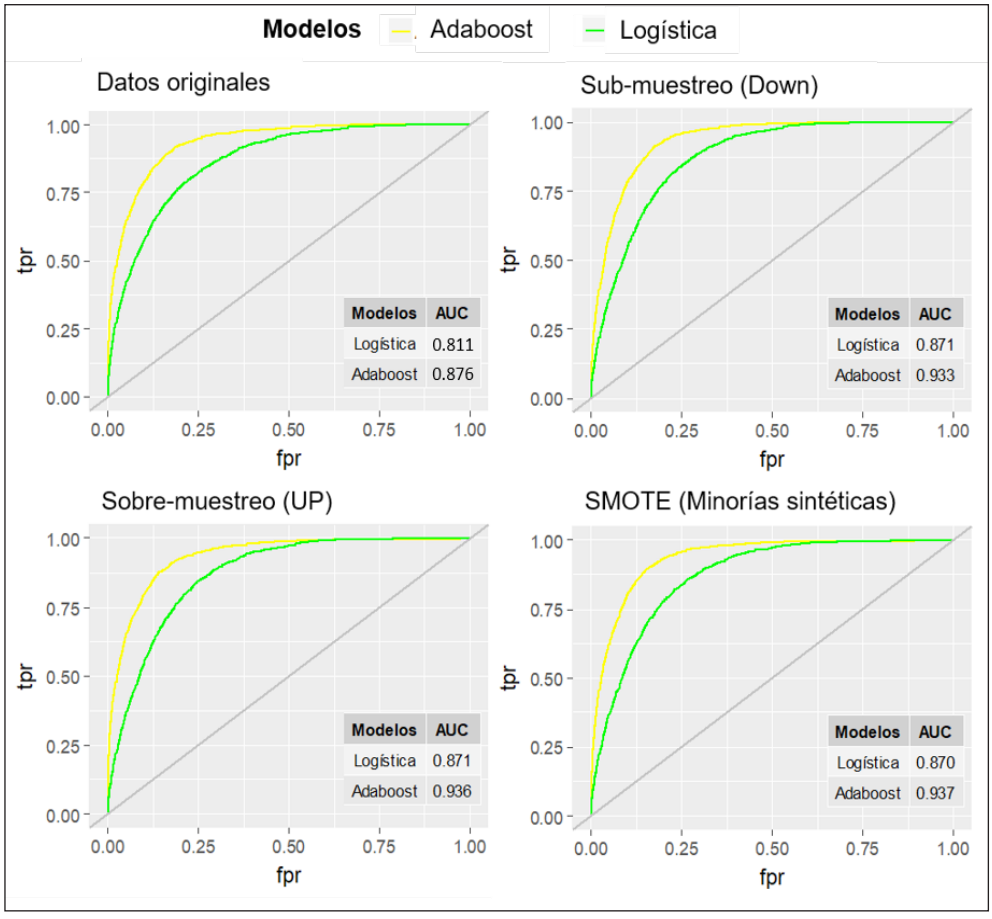

Figura 6. Comparación del AUC en los modelos de regresión logística y el algoritmo Adaboost mediante los métodos de muestreo

Para efectos de clasificación o regresión, los algoritmos seleccionan internamente (mediante árboles de clasificación) las variables más importantes para crear los modelos. En el algoritmo Adaboost, las variables más importantes para identificar si un posible cliente fugará (Figura 7) son el tipo de reclamo, generalmente por problemas de facturación (Figura 8), el ROL del cliente (clientes seguidores que no realizan llamadas, mensajes o transferencias), la Comunidad (en la Figura 8, los clientes que fugan tiene nivel de comunidad Grado 1, es decir casi no tiene contacto con otros usuarios de la misma empresa), número de reclamos (en la Figura 9 se observa que los clientes que fugan tiene en promedio 2.24 reclamos mensuales en los últimos 6 meses), días de deuda y tipo de cliente (valoración que le da la empresa), nota de pago (los que fugan tiene valoración inferior en comparación a los que no fugan).

4. Conclusiones

Al comparar los modelos de clasificación se encontró que el algoritmo Adaboost es el de mejor rendimiento en comparación a la regresión logística, esto fue corroborado con las diferentes métricas, las cuales dieron evidencias del mejor desempeño a favor del algoritmo Adaboost (aunque con un costo computacional más alto). Si bien es cierto, en este estudio el algoritmo Adaboost tuvo mejor desempeño, esto no implica que en todos los estudios de fuga de clientes en telefonía móvil este algoritmo será el más óptimo (tener en cuenta diferentes factores para seleccionar un algoritmo, tales como el tamaño de base de datos, la cantidad de variables, el tipo de variables, el balanceo de datos, las variables con exceso de outliers, entre otros). A través del algoritmo Adaboost se identificó que los clientes deciden fugar

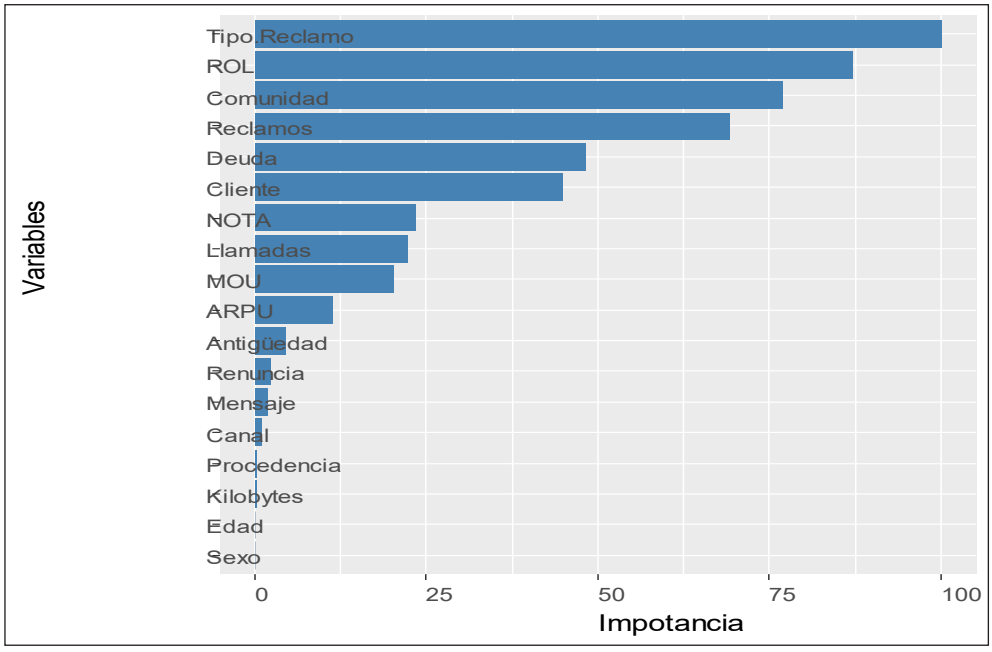

Figura 7. Distribución según Importancia de variables en el algoritmo Adaboost

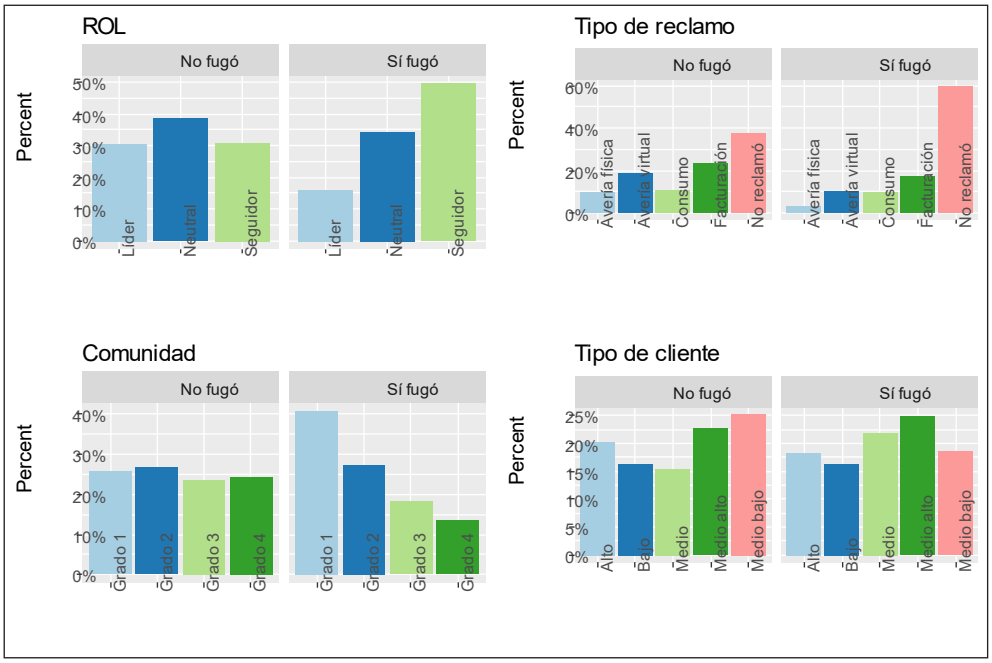

Figura 8. Distribución de las variables más importantes para la predicción de los clientes en el algoritmo Adaboost (variables cualitativas)

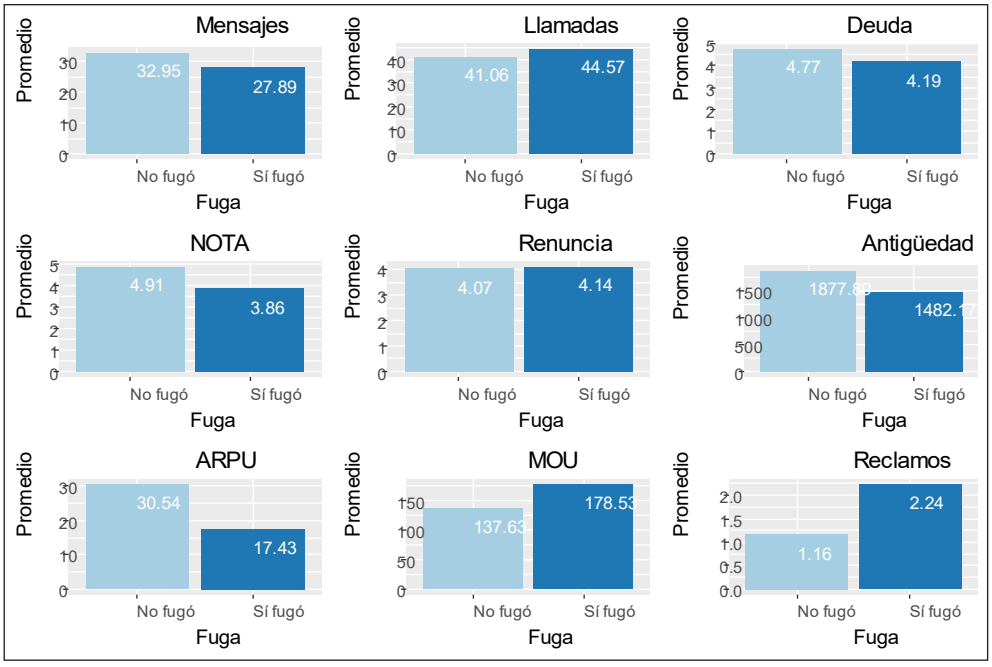

Figura 9. Distribución de las variables más importantes para la predicción de los clientes en el algoritmo Adaboost (variables cuantitativas)

o migrar a otros operadores dependiendo de los diferentes tipos de reclamos, el tipo de cliente, la cantidad de reclamos, la cantidad de mensajes y llamadas que realizan, el grado de actividad, el contacto con otros usuarios, entre otras variables; con lo cual los responsables de toma de decisiones de la empresa podrán focalizarse en estas causas y minimizar esfuerzos en la retención de los clientes. Para estudios posteriores se recomienda comparar otros modelos y/o algoritmos de aprendizaje con el algoritmo Adaboost para evaluar el desempeño de los indicadores en la clasificación e identificación de fuga de clientes, además utilizar otros puntos de corte o umbrales de clasificación para ajustar al máximo los modelos, por último, complementar el análisis de desempeño con otras curvas tales como las PPROC (área bajo la curva entre la Precisión y el Recall), puesto que puede ser una medida implícitamente más informativa.

5. Literatura citada

Ahmad, A.; Jafar, A.; Aljoumaa, K. 2019. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data 6(28). doi: 10.1186/s40537-019-0191-6

Bhatia, A.; Chiu, Y. 2017. Machine Learning with R Cookbook. 2da Edición. Editorial Packt Publishing, Birmingham B3 2PB, UK. 274 p.

Barrientos, F. 2012. Diseño e implementación de una metodología de predicción de fuga de clientes en una compañía de telecomunicaciones. Memoria para optar al título de ingeniero civil industrial. Departamento de Ingeniería Industrial. Universidad de Chile. Disponible en http://repositorio.uchile.cl/handle/2250/104421

Brownlee, J. 2015. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. Machine Learning Process. Disponible en: http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Chicco, D.; Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1): 6. doi: 10.1186/s12864-019-6413-7

Escobar, C.; Lolas, F. 2015. Desarrollo de un sistema prototipo para la detección temprana de la deserción escolar en escuelas públicas chilenas. Memoria de Título, Universidad Adolfo Ibáñez, Santiago de Chile.

Fernández, A.; Río, S.; Chawla, N.; Herrera, F. 2017. An insight into imbalanced Big Data classification: outcomes and challenges. Complex & Intelligent Systems 3: 105-120. doi: 10.1007/s40747-017-0037-9

Freund, Y.; Schapire, R. 1996. Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference. Murray Hill, NJ 07974-0636.

Hadad, A.; Evin, D.; Drozdowicz, B. 2009. Modelo para el tratamiento de datos desbalanceados basado en redes neuronales autoorganizadas. XVII Congreso Argentino de Bioingeniería, Rosario, Santa Fe.

Haibo, H.; Yunqian, M. 2013. Imbalanced Learning: Foundations, Algorithms, and Applications. 1era Edición. Editoria John Wiley & Sons, Hoboken, New Jersey. 86 p.

Hair, J.; Anderson R.; Tatham R.; Black W. 1999. Análisis Multivariante. 5ta edición. Editorial Prentice Hall Iberia , Madrid. 195 p.

Ijaz, M.; Alfian, G.; Syafrudin, M.; Rhee, J. 2018. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest. Applied Sciences 8(8): 1325. doi: 10.3390/app8081325

Hosmer, D.; Lemeshow, S. 2000. Applied Logistic Regression. 2da Edición. Editorial Wiley. ISBN 0-471-35632-8. 88-102 pp.

Hu, H. 2019. Research on Customer Churn Prediction Using Logistic Regression Model. Advances in Intelligent Systems and Computing 885: 344-350. doi: 10.1007/978-3-030-02804-6_46

Kunal, J. 2016. Practical Guide to deal with Imbalanced Classification Problems in R. Analytics Vidhya. Learn Everything About Analytics. Disponible en: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Lang, J. 2011. Predictors tutorial, Bioinformatic Department Projects. Disponible en: https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada

Lejenue, M. 2001. Measuring the impact of data mining on Churn Management. Research Internet, 11(5): 374-384. doi: 08/10662240110410183

Liu, X.; Wu, J.; Zhou, Z. 2006. Exploratory Under-Sampling for Class-Imbalance Learning. 965-969. 10.1109/ICDM.2006.68

Mao, W.; Wang, J.; Xue, Z. 2017. An ELM-based model with sparse-weighting strategy for sequential data imbalance problem. Int. J. Mach. Learn. & Cyber. 8:1333–1345. doi: 10.1007/s13042-016-0509-z

Meza, A. 2018. Predicción de fuga de clientes en una empresa de telefonía utilizando el algoritmo Adaboost desbalanceado y la regresión logística asimétrica. Tesis para optar el grado de Magister. Universidad Agraria la Molina. Disponible en: http://repositorio.lamolina.edu.pe/handle/UNALM/3245

Neslin, S.; Gupta, S.; Kamakura, W.; Lu, J.; Mason, C. 2006. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research American Marketing Association ISSN 43(2): 204-211. doi: 10.1509/jmkr.43.2.204

Osiptel. 2019. PERÚ: Portabilidad móvil se mantiene arriba de las 800,000 portaciones por cuarto mes consecutivo. Disponible en: https://www.osiptel.gob.pe/noticia/np-portabilidad-movil-mantiene-arriba-800000-portaciones-cuarto-mes

Obregón, S. 2016. Desarrollo de una Herramienta de Diagnóstico de Fallos en Motores de Inducción Mediante la técnica Adaboost. Trabajo fin de Máster para obtener el título de Ingeniero Industrial. Universidad de Valladolid. Disponible en: http://uvadoc.uva.es/handle/10324/18912

Pérez, P. 2014. Modelo de predicción de fuga de clientes de telefonía móvil post pago. Memoria para Optar al Título de Ingeniero Civil Industrial. Departamento de Ingeniería Industrial. Universidad de Chile. Disponible en: Disponible en http://repositorio.uchile.cl/handle/2250/115942

Powers, D. 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach. Learn. Technol 2.

Tharwat, A. 2018. Classification assessment methods. Applied Computing and Informatics. doi: 10.1016/j.aci.2018.08.003

Valavi, R.; Elith, J.; Lahoz-Monfort, J.; Guillera-Arroita, G. 2018. blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution 10(2): 225-232. doi: 10.1111/2041-210X.13107

Wu, Z.; Lin, W.; Ji, Y. 2018. An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics. in IEEE Access 6: 8394-8402. doi: 10.1109/ACCESS.2018.2807121