Ingeniería, investigación y tecnología
ISSN: 1405-7743
Facultad de Ingeniería, UNAM

Espinosa-Zúñiga, Javier Jesús
Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito
Ingeniería, investigación y tecnología, vol. XXI, núm. 3, 00002, 2020, Julio-Septiembre
Facultad de Ingeniería, UNAM

DOI: https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Ingeniería Investigación y Tecnología
volumen XXI (número 3), julio-septiembre 2020 1-16
ISSN 2594-0732 FI-UNAM artículo arbitrado
Información del artículo: Recibido: 23 de agosto de 2019, reevaluado: 23 y 26 de agosto de 2019, 27 de febrero de 2010, aceptado: 28 de abril de 2020

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Ingeniería
Investigación y
Tecnología

Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito

Application of Random Forest and XGBoost algorithms based on a credit card applications database

Espinosa-Zúñiga Javier Jesús
Grupo Financiero Ve por Más S.A. de C.V., México
Gerencia CRM
Correo: jjespinoza@vepormas.com
https://orcid.org/0000-0001-6828-2145

Resumen

Dentro de la gama de algoritmos de aprendizaje automático existentes destacan actualmente dos: Random Forest y XGBoost. Ambos han adquirido gran popularidad. Random Forest es un algoritmo que surgió hace casi veinte años y se utiliza ampliamente por el balance que ofrece entre complejidad y resultados. Por su parte, XGBoost es un algoritmo que ha despertado gran interés, pues aunque es relativamente reciente es considerado actualmente el estado del arte en algoritmos de aprendizaje automático por sus resultados. Uno de los sectores en los que se aplican este tipo de algoritmos es el financiero. Algunos ejemplos de su aplicación en este sector son: segmentación de clientes, detección de fraudes, pronóstico de ventas, autenticación de clientes y análisis de comportamiento de mercados, entre otros. Un área de particular interés en este sector es la identificación de clientes a quienes otorgar una tarjeta de crédito, esto es crítico para las instituciones financieras, pues una selección incorrecta de estos clientes podría derivar en un incremento de su cartera vencida. En el presente estudio se aplicaron los algoritmos Random Forest y XGBoost sobre una base de solicitudes de tarjetas de crédito (donada por un banco australiano para fines de investigación) para identificar las solicitudes con mayor probabilidad de otorgarles una tarjeta. Los modelos obtenidos se compararon estadísticamente (donde se seleccionó el modelo con el algoritmo XGBoost) y se presentaron los resultados con gráficas que permiten responder dos preguntas clave desde el enfoque de negocio: ¿Cuáles son las solicitudes a las que hay que otorgar una tarjeta? y ¿Qué resultados esperamos en caso de aplicar el modelo? La aportación más importante del presente estudio es aplicar dos algoritmos muy efectivos sobre esta base de solicitudes de tarjetas de crédito con un enfoque de negocios.
Descriptores: Aprendizaje automático, XGBoost, Random Forest, árbol de decisión, hiperparámetro.

Abstract

Two of the existing machine learning algorithms currently stand out: Random Forest and XGBoost. Both have become very popular. Random Forest is an algorithm that emerged almost twenty years ago and is widely used for the balance it offers between complexity and results. On the other hand, XGBoost is an algorithm that has aroused great interest because although it is relatively recent, it is currently considered the state of the art in machine learning algorithms for its results. One of the sectors in which this type of algorithm is applied is the financial. Some examples of its application in this sector are: customer segmentation, fraud detection, sales forecasting, customer authentication and market behavior analysis. An area of particular interest in this sector is the identification of clients to whom to grant a credit card: this is critical for financial institutions since an incorrect selection of these clients could lead to an increase in their past due portfolio. In the present study the Random Forest and XGBoost algorithms were applied on a credit card application database (donated by an Australian bank for research purposes) to identify the applications most likely to be granted a credit card. The models obtained were compared statistically (from which the model obtained with the XGBoost algorithm was selected) and the results were presented with graphs that allow answering two key questions from the business perspective: what are the requests to which a card must be awarded? and what results do we expect if the model is applied? The most important contribution of the present study is to apply two very effective algorithms on this database with a business focus.
Keywords: Machine Learning, XGBoost, Random Forest, decision tree, hyper parameter.

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Introducción

Un factor crítico para las instituciones financieras es determinar, de una base de solicitudes, a quién otorgar una tarjeta de crédito. Para ello, se apoyan cada vez más en algoritmos de aprendizaje automático, con los cuales se obtienen modelos que permiten en un momento dado tomar decisiones lo más precisas posibles en este sentido.

Sin embargo, en el sector financiero es importante que los resultados obtenidos con estos modelos se presenten con un enfoque de negocio, ya que si bien, la metodología y sustento estadístico dan soporte a los resultados, el enfoque de negocios permite aplicarlos de manera práctica. Esto es aún más importante cuando los mismos se presentan a tomadores de decisiones en el sector financiero.

En el presente estudio se determinó, de una base de solicitudes de tarjetas de crédito donada por un banco australiano para fines de investigación (Universidad de California, 2019), la probabilidad de otorgarles una tarjeta. Para ello, se aplicó un análisis exploratorio que permitió conocer las características de la base, se prepararon los datos de acuerdo con los resultados del análisis y se dividió la base en dos grupos (entrenamiento y validación) donde el primero se utilizó para entrenar los modelos y el segundo para compararlos entre sí a fin de elegir el más conveniente.

Finalmente, se muestran los resultados del modelo elegido con un enfoque de negocio mediante dos graficas: curva de ganancia acumulada y curva de respuesta acumulada.

Con esta base de solicitudes se han publicado ya múltiples estudios que utilizan algoritmos de aprendizaje automático, solo por citar algunos ejemplos: Huo *et al.* (2006) utilizaron esta base para su propuesta de algoritmo de poda de árboles, Tumer y Ghosh (2002) la emplearon para su propuesta de integración de múltiples salidas de clasificadores y Quinlan (1987), (quien de hecho fue quien donó esta base) analizó diversas técnicas para simplificar árboles de decisión sobre la misma. Sin embargo, estos estudios no analizan la base con un enfoque de negocios.

El estudio se divide en las siguientes secciones: Marco teórico, donde se comentan las características generales de los modelos Random Forest y XG Boost; Análisis exploratorio, en esta sección se estudian las características de los datos que conforman la base de solicitudes de crédito en la que se aplicarán los modelos Random Forest y XG Boost; Preparación de datos, donde se seleccionan las variables para obtener una "base limpia" para la etapa de modelado; Modelado, en esta etapa se aplican los modelos Random Forest y XG Boost sobre la "base limpia"; Análisis de resultados, aquí se obtienen las gráficas que permitirán responder las cuatro preguntas mencionadas anteriormente y finalmente las Conclusiones, que marcan los resultados obtenidos y proponen siguientes pasos a fin de dar continuidad al estudio.

Se utiliza lenguaje de programación R (lenguaje gratuito muy popular para Estadística y Ciencia de Datos) (CRAN, 2019).

Marco teórico

El aprendizaje automático es una rama de las ciencias computacionales que ha adquirido gran auge en los últimos años. Aunque existen muchas definiciones una de las más comunes es: "campo de estudio que da a las computadoras la habilidad de aprender sin ser programadas explícitamente". Abarca una amplia gama de técnicas que se clasifican en dos grandes tipos: aprendizaje supervisado (en el que se tiene una variable objetivo y se "entrena" a un programa informático con un conjunto de datos para aplicar el resultado en un nuevo conjunto de datos) y aprendizaje no supervisado (en el que no se tiene variable objetivo y el programa informático debe encontrar patrones y relaciones en un conjunto de datos) (Sandoval, 2017).

El algoritmo Random Forest (Breiman, 2001) es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado (Lizares, 2017). Cada árbol se obtiene mediante un proceso de dos etapas:

1. Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que m < M (donde M = total de predictores).
2. Cada árbol crece hasta su máxima extensión.

Cada árbol generado por el algoritmo Random Forest contiene un grupo de observaciones aleatorias (elegidas mediante *bootstrap*, que es una técnica estadística para obtener muestras de una población donde una observación se puede considerar en más de una muestra). Las observaciones no estimadas en los árboles (también conocidas como "*out of the bag*") se utilizan para validar el modelo. Las salidas de todos los árboles se combinan en una salida final Y (conocida como ensamblado) que

se obtiene mediante alguna regla (generalmente el promedio, cuando las salidas de los árboles del ensamblado son numéricas y, conteo de votos, cuando las salidas de los árboles del ensamblado son categóricas). Lo anterior se muestra gráficamente en la Figura 1.

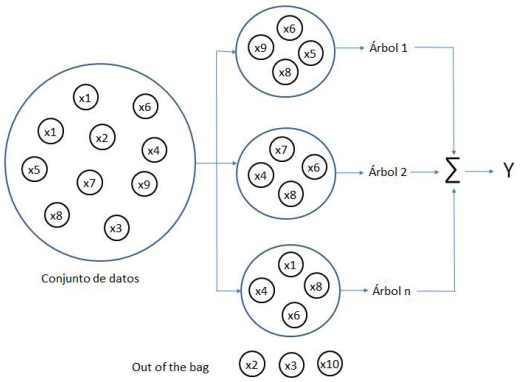

Figura 1. Algoritmo Random Forest

Las principales ventajas del algoritmo Random Forest (Cánovas *et al.*, 2017) son:

a) Pueden usarse para clasificación o predicción: En el primer caso, cada árbol "vota" por una clase y el resultado del modelo es la clase con mayor número de "votos" en todos los árboles, de forma que cada nueva observación se presenta a cada uno de los árboles y se asigna a la clase más "votada". En el segundo caso, el resultado del modelo es el promedio de las salidas de todos los árboles.
b) El modelo es más simple de entrenar en comparación con técnicas más complejas, pero con un rendimiento similar.
c) Tiene un desempeño muy eficiente y es una de las técnicas más certeras en bases de datos grandes.
d) Puede manejar cientos de predictores sin excluir ninguno y logra estimar cuáles son los predictores más importantes, es por ello que esta técnica también se utiliza para reducción de dimensionalidad.
e) Mantiene su precisión con proporciones grandes de datos perdidos.

Por otra parte, sus principales desventajas son las siguientes:

a) La visualización gráfica de los resultados puede ser difícil de interpretar.
b) Puede sobre ajustar ciertos grupos de datos en presencia de ruido.
c) Las predicciones no son de naturaleza continua y no puede predecir más allá del rango de valores del conjunto de datos usado para entrenar el modelo. En el caso de predictores categóricos con diferente número de niveles, los resultados pueden sesgarse hacia los predictores con más niveles.
d) Se tiene poco control sobre lo que hace el modelo (en cierto sentido es como una caja negra).

Las ventajas de Random Forest hacen que se convierta en una técnica ampliamente utilizada en muchos campos, por ejemplo, teledetección (para clasificación de imágenes), bancos (para detección de fraudes y clasificación de clientes para otorgamiento de crédito), medicina (para analizar historiales clínicos a fin de identificar enfermedades potenciales en los pacientes), finanzas (para pronosticar comportamientos futuros de los mercados financieros) y comercio electrónico (para pronosticar si un cliente comprará, o no, cierto producto), entre otros .

El algoritmo XG Boost (Extreme Gradient Boosting) es una técnica de aprendizaje supervisado (Chen y Guestrin, 2016) también basada en árboles de decisión y que es considerada el estado del arte en la evolución de estos algoritmos (Figura 2).

Figura 2. Evolución de los algoritmos basados en árboles de decisión

El algoritmo XG Boost tiene las siguientes características (Chen y Guestrin, 2016):

a) Consiste en un ensamblado secuencial de árboles de decisión (este ensamblado se conoce como CART, acrónimo de "Classification and Regression Trees"). Los árboles se agregan secuencialmente a fin de aprender del resultado de los árboles previos y corregir el error producido por los mismos, hasta que ya no se pueda corregir más dicho error (esto se conoce como "gradiente descendente" (Figura 3).
b) La principal diferencia entre los algoritmos XG-Boost y Random Forest es que en el primero el usuario define la extensión de los árboles mientras que en el segundo los árboles crecen hasta su máxima extensión.
c) Utiliza procesamiento en paralelo, poda de árboles, manejo de valores perdidos y regularización (optimización que penaliza la complejidad de los modelos) para evitar en lo posible sobreajuste o sesgo del modelo.

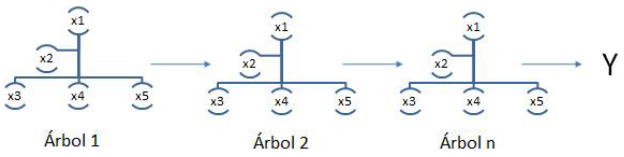

Figura 3. Algoritmo XGBoost

El algoritmo XGBoost funciona así:

a) Se obtiene un árbol inicial F_0 para predecir la variable objetivo "y", el resultado se asocia con un residual ($y - F_0$).
b) Se obtiene un nuevo árbol h_1 que ajusta al error del paso previo.
c) Los resultados de F_0 y h_1 se combinan para obtener el árbol F_1, donde el error cuadrático medio de F_1 será menor que el de F_0:

$$F_1(x) <- F_0(x) + h_1(x)$$

d) Este proceso se sigue iterativamente hasta que el error es minimizado lo más posible de la siguiente forma:

$$F_m(x) <- F_{m-1}(x) + h_m(x)$$

Las principales ventajas del algoritmo XGBoost son:

a) Puede manejar grandes bases de datos con múltiples variables.
b) Puede manejar valores perdidos.
c) Sus resultados son muy precisos.
d) Excelente velocidad de ejecución.

Por otra parte, sus principales desventajas son:

a) Puede consumir muchos recursos computacionales en grandes bases de datos, por lo que se recomienda antes de aplicar esta técnica en bases de este tipo, determinar cuáles son las variables que aportarán más información a fin de considerar solo dichas variables en la obtención del modelo.
b) Se deben ajustar correctamente los parámetros del algoritmo a fin de minimizar el error de precisión y evitar sobreajuste del modelo (lo que puede darse si se maneja un número muy grande de árboles).
c) Solo trabaja con vectores numéricos, por lo que se requieren convertir previamente los tipos de datos no numéricos a numéricos.

Las ventajas de este algoritmo hace que se aplique en campos como: identificación de huellas digitales (Luckner *et al.*, 2017), seguridad vial (Bahador *et al.*, 2020) y análisis de mercados financieros (Nobre y Ferreira, 2019), entre otros.

Análisis exploratorio

Una vez obtenida la base de solicitudes de crédito para el estudio (UCI, 2020), se realiza un análisis exploratorio de la misma. Las características principales de la base son:

1. Contiene un total de 690 registros.
2. Los nombres de las columnas se modifican de origen, a fin de proteger la confidencialidad de la misma.
3. Su layout tiene 16 columnas (15 predictores y una variable objetivo que indica si se le otorgó un crédito o no, a cada solicitud (Tabla 1).

Tabla 1. Layout original de la base de aplicaciones de crédito

Núm.	Columna	Tipo de dato	Comentario
1	A1	Categórica	Variable predictor
2	A2	Numérica	Variable predictor
3	A3	Numérica	Variable predictor
4	A4	Categórica	Variable predictor
5	A5	Categórica	Variable predictor
6	A6	Categórica	Variable predictor
7	A7	Categórica	Variable predictor
8	A8	Numérica	Variable predictor
9	A9	Categórica	Variable predictor
10	A10	Categórica	Variable predictor
11	A11	Numérica	Variable predictor
12	A12	Categórica	Variable predictor
13	A13	Categórica	Variable predictor
14	A14	Categórica	Variable predictor
15	A15	Numérica	Variable predictor
16	A16	Categórica	Variable objetivo

A fin de facilitar el manejo de las variables se cambiaron sus nombres como se indica en la Tabla 2.

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Tabla 2. Layout con nombres de variables modificadas

Núm.	Nombre original	Nombre modificado	Descripción
1	A1	Género	Género del solicitante (masculino, femenino)
2	A2	Edad	Edad del solicitante
3	A3	Saldo	Saldo del solicitante
4	A4	Estado civil	Estado civil del solicitante
5	A5	Entidad bancaria	Banco del cual el solicitante es cliente
6	A6	Educación	Nivel educativo del solicitante
7	A7	Nacionalidad	Nacionalidad del solicitante
8	A8	Años laborando	Número de años que el solicitante tiene laborando
9	A9	Bandera 1	Indica si el solicitante aplicó una solicitud anteriormente
10	A10	Bandera 2	Indica si el solicitante labora actualmente
11	A11	Calificación	Calificación crediticia del solicitante
12	A12	Bandera 3	Indica si el solicitante tiene licencia de manejo
13	A13	Bandera 4	Indica si el solicitante es ciudadano al momento de la solicitud
14	A14	Código postal	Código postal donde reside el solicitante
15	A15	Ingreso	Ingreso anual del solicitante
16	A16	Bandera5	Indica si la solicitud fue aprobada

Al obtener el análisis exploratorio sobre la base se tiene que:

- En general, las variables de la base contienen pocos nulos. Las variables con mayor volumen de nulos son: código postal (1.88 %), género (1.74 %) y edad (1.74 %), (Figura 4) donde la columna "p_na" se refiere al porcentaje de nulos por variable.
- Para la variable "Género" (Figura 5) se tienen 12 nulos, 210 registros caen en la categoría "a" y 468 registros caen en la categoría "b" (no se tienen elementos para determinar la categoría que corresponde a género masculino o femenino).

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
genero	0	0.00	12	1.74	0	0	character	2
edad	0	0.00	12	1.74	0	0	character	349
saldo	19	2.75	0	0.00	0	0	character	215
edo_civil	0	0.00	6	0.87	0	0	character	3
entidad_bancaria	0	0.00	6	0.87	0	0	character	3
educacion	0	0.00	9	1.30	0	0	character	14
nacionalidad	0	0.00	9	1.30	0	0	character	9
años_laborando	70	10.14	0	0.00	0	0	character	132
bandera1	0	0.00	0	0.00	0	0	character	2
bandera2	0	0.00	0	0.00	0	0	character	2
calificacion	395	57.25	0	0.00	0	0	character	23
bandera3	0	0.00	0	0.00	0	0	character	2
bandera4	0	0.00	0	0.00	0	0	character	3
cp	132	19.13	13	1.88	0	0	character	170
ingreso	295	42.75	0	0.00	0	0	character	240
bandera5	0	0.00	0	0.00	0	0	character	2

Figura 4. Porcentaje de nulos por variable

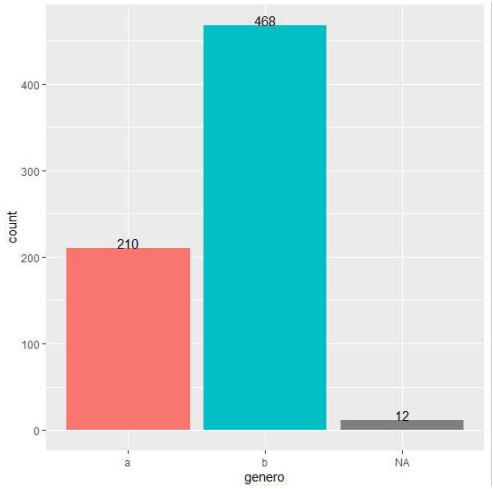

Figura 5. Género del solicitante

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

- Para la variable "Estado civil" (Figura 6) se tienen 6 nulos, 2 registros caen en la categoría "l", 519 registros caen en la categoría "u" y 163 registros caen en la categoría "y" (no se tienen elementos para determinar a qué estado civil corresponde cada categoría).
- Para la variable "Entidad bancaria" (Figura 7) se tienen 6 nulos, 519 registros caen en la categoría "g", 2 registros caen en la categoría "gg" y 163 registros caen en la categoría "p":
- Para la variable "Nivel educativo" (Figura 8) se tienen 9 nulos y 681 registros distribuidos entre 14 categorías (la categoría con mayor número de registros es la "cc" con 137 registros).
- Para la variable "Nacionalidad" (Figura 9) se tienen 9 nulos y 681 registros distribuidos entre 9 categorías (la categoría con mayor número de registros es la "v" con 399 registros).

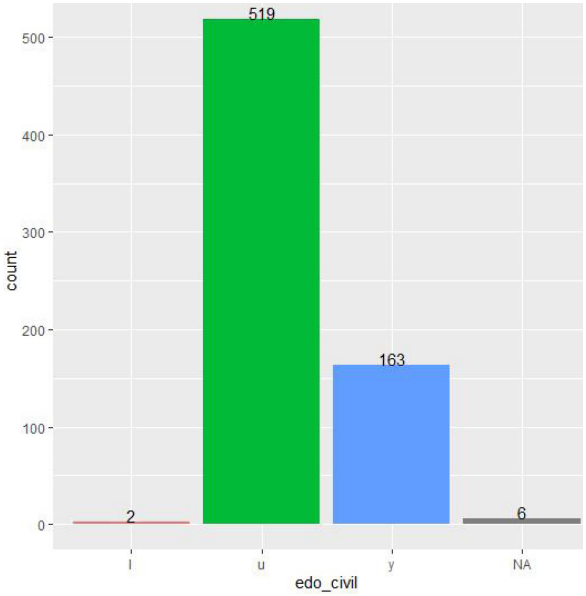

Figura 6. Estado civil del solicitante

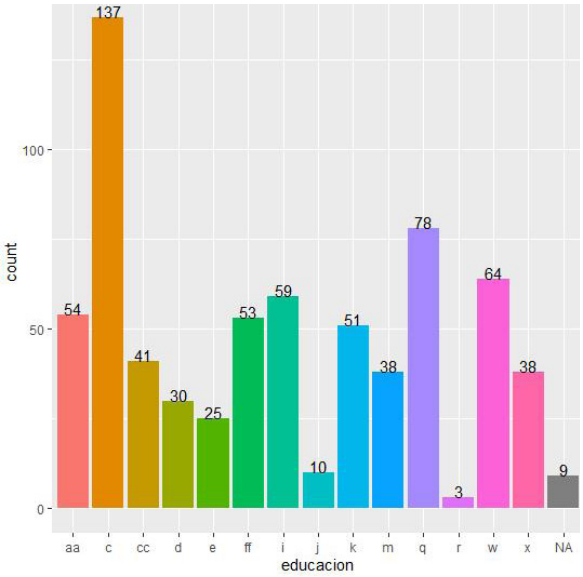

Figura 8. Nivel educativo del solicitante

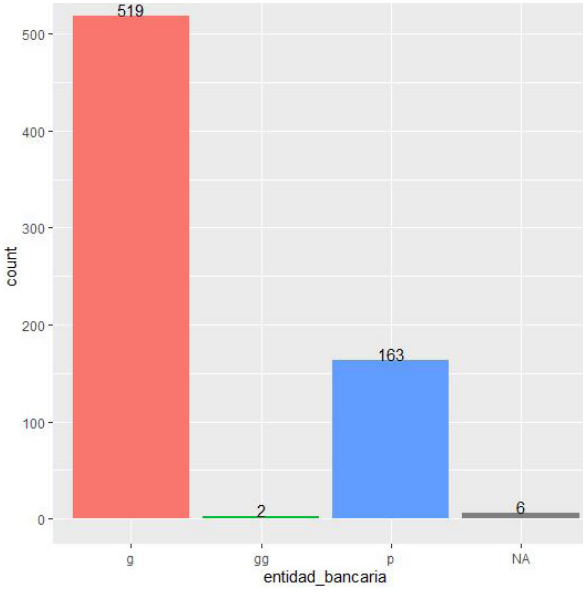

Figura 7. Entidad bancaria del solicitante

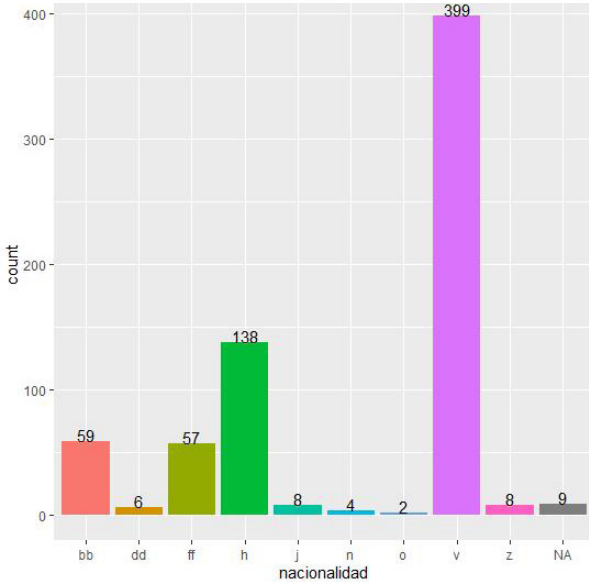

Figura 9. Nacionalidad del solicitante

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

- Para la variable "bandera1" (que indica si el solicitante ya aplicó previamente, Figura 10) no se tienen nulos, 329 registros están en la categoría "f" y 361 registros en la categoría "t". Se asume que la categoría "f" corresponde a solicitantes que no han aplicado previamente y la categoría "t" corresponde a solicitantes que ya aplicaron previamente.
- Para la variable "bandera2" (que indica si el solicitante labora actualmente, Figura 11) no se tienen registros nulos, 395 registros están en la categoría "f" y 295 registros en la categoría "t". Se asume que la categoría "f" corresponde a solicitantes que no han aplicado previamente y la categoría "t" corresponde a solicitantes que ya aplicaron previamente.
- Para la variable "bandera4" (que indica si el solicitante es ciudadano, Figura 12) no se tienen nulos, 625 registros están en la categoría "g", 8 registros están en la categoría "p" y 57 registros están en la categoría "s".
- Para la variable "bandera5" (que indica si la solicitud fue aprobada, Figura 13) no se tienen nulos, 383 registros están en la categoría "–", 307 registros es-

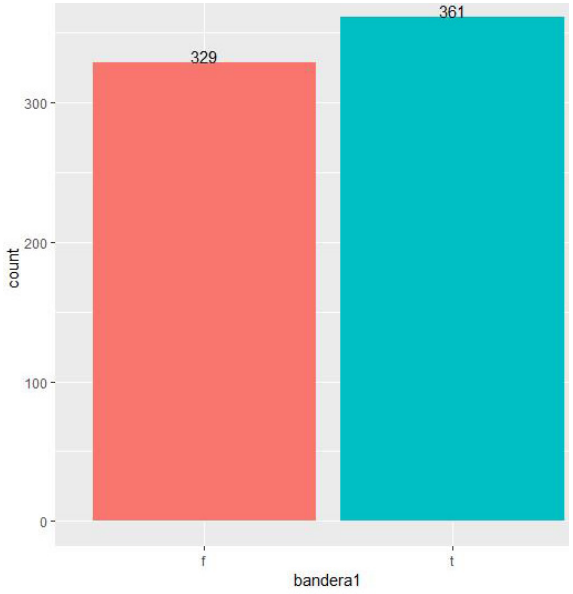

Figura 10. ¿El solicitante ya aplicó previamente?

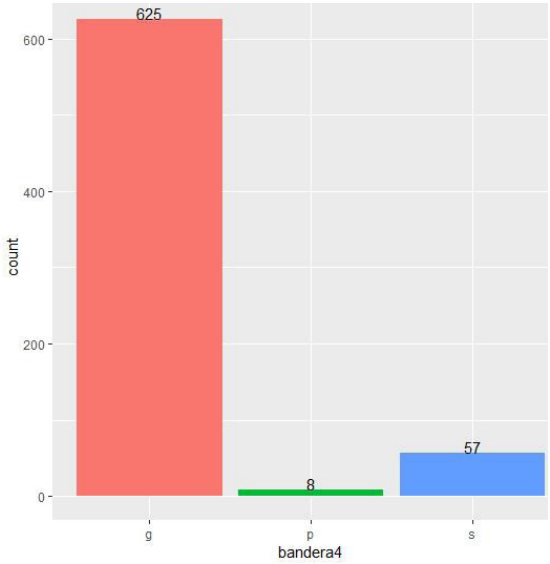

Figura 12. ¿El solicitante es ciudadano al momento de la solicitud?

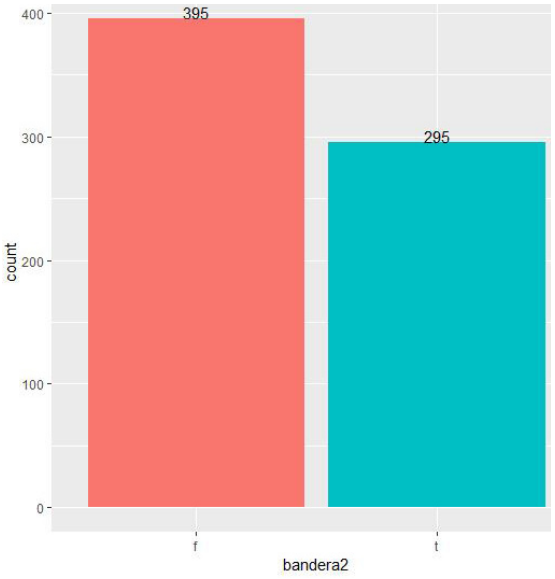

Figura 11. ¿El solicitante labora actualmente?

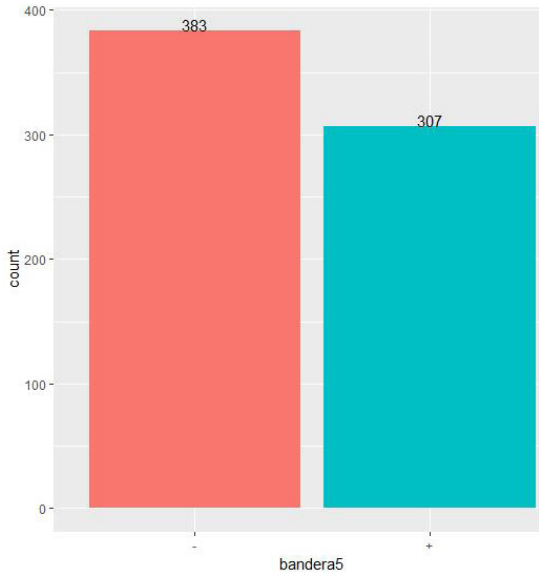

Figura 13. ¿La solicitud fue aprobada?

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

tán en la categoría "+". Se asume que la categoría "–"corresponde a solicitudes rechazadas y la categoría "+" corresponde a solicitudes aprobadas.
- Para la variable "código postal" se tienen un total de 169 códigos postales donde residen los solicitantes. La Tabla 3 muestra los diez códigos postales con mayor volumen de solicitantes, notar que el mayor volumen corresponde a solicitantes con código postal 0, también se incluye el total de 13 solicitantes sin código postal.

Tabla 3. Códigos postales con mayor volumen de solicitantes

Núm.	Código postal	Total
1	0	132
2	120	35
3	200	35
4	160	34
5	100	30
6	80	30
7	280	22
8	180	18
9	140	16
10	240	14
11	NA	13

- Para la variable "Edad" (Figura 14) la mediana está en 28.46 años, el mínimo en 13.75 años, el máximo en 80.25 años y se tiene regular proporción de valores extremos (no se descartaron los 12 nulos en la variable). El histograma muestra que las edades se cargan ligeramente a la izquierda, por lo que en su mayoría se trata de solicitantes jóvenes.
- Para la variable "Saldo" (Figura 15) la mediana está en $2.75 dólares australianos (se supone este tipo de moneda por el origen de la base de datos), el mínimo es $0.00 (es decir, hay solicitantes sin saldo), el máximo es $28.00 dólares australianos y se tiene regular proporción de valores extremos. El histograma muestra una alta concentración de solicitantes con saldos bajos.
- Para la variable "Años laborando" (Figura 16) la mediana está en 1 año, el mínimo en 0 años (es decir, hay solicitantes que son clientes nuevos del banco en cuestión, ya que ni siquiera tienen un año de antigüedad con el mismo), el máximo es 28.5 años y se tienen regular proporción de valores extremos. El histograma muestra una alta concentración de solicitantes con pocos años laborando en su empleo actual.
- Para la variable "Calificación" (Figura 17) la mediana y el mínimo están en 0 puntos, el máximo es 67 puntos y se tienen alta proporción de valores extremos. El histograma muestra una alta concentración de solicitantes en calificaciones bajas.
- Para la variable "Ingreso" (Figura 18) la mediana está en $5 dólares australianos, el mínimo en $0 dólares australianos (es decir, hay solicitantes que reportaron no tener ingresos, o bien, podría tratarse de un error de calidad de datos, ya que difícilmente se le daría crédito a un solicitantes sin ingresos), el máximo en $100,000.00 dólares australianos y se tiene alta proporción de valores extremos. El histograma muestra que la mayoría de los solicitantes son de ingresos bajos.

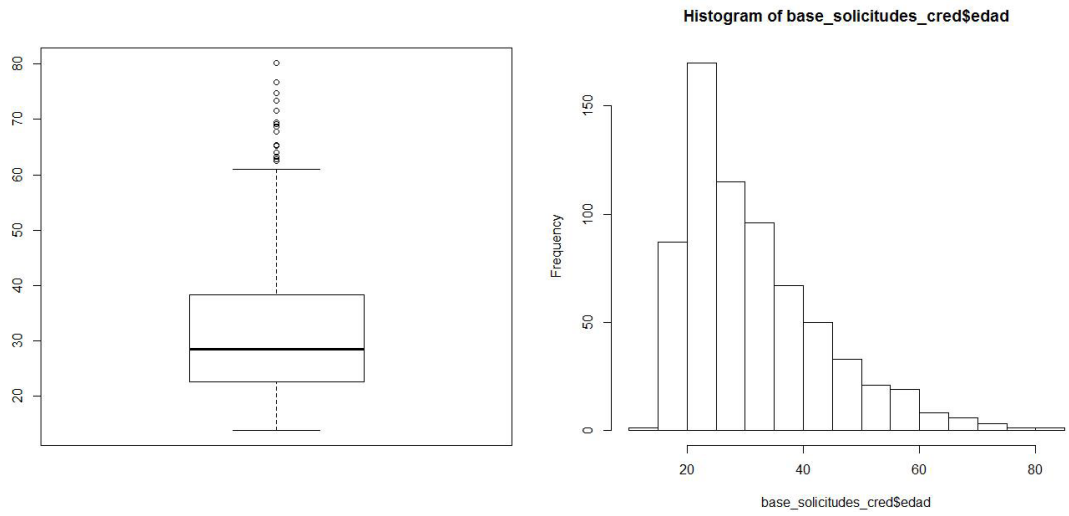

Figura 14. Edad de solicitantes

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

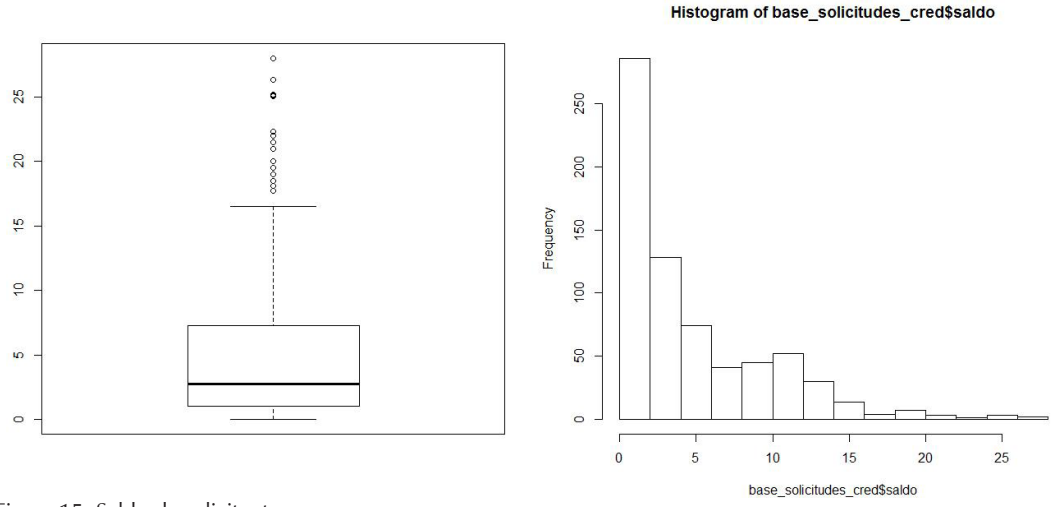

Figura 15. Saldo de solicitantes

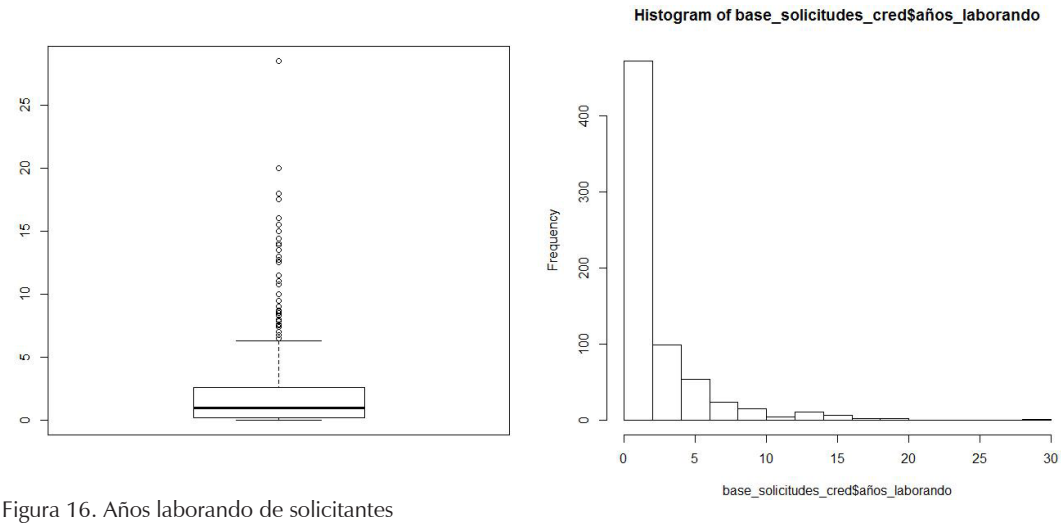

Figura 16. Años laborando de solicitantes

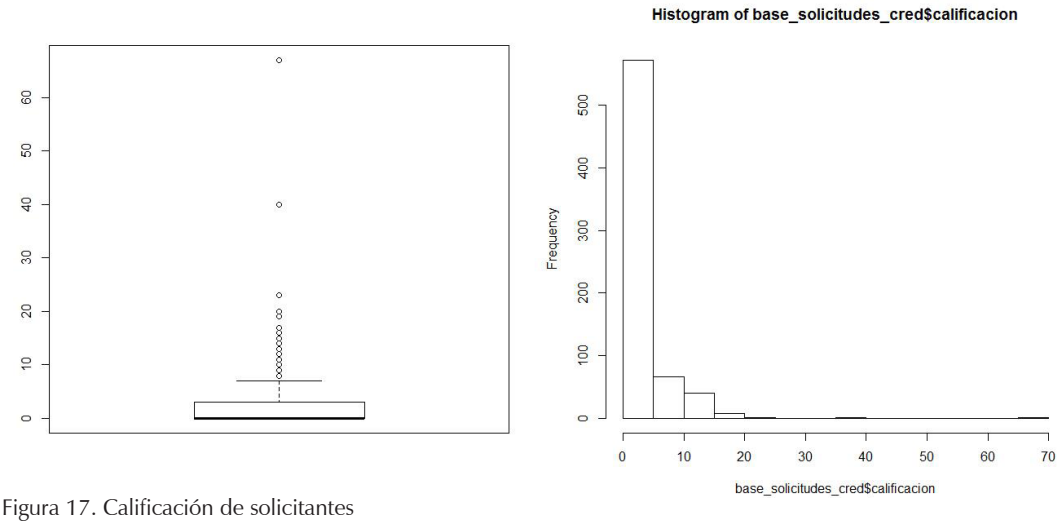

Figura 17. Calificación de solicitantes

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

La Tabla 4 muestra los estadísticos básicos para las variables numéricas en la base.
Finalmente, la Figura 19 muestra la matriz de correlación entre las variables numéricas, en la misma se observa que existe cierta correlación entre la edad y los años laborales del solicitante, así como entre el saldo y los años laborales del solicitante. Asimismo, se observa

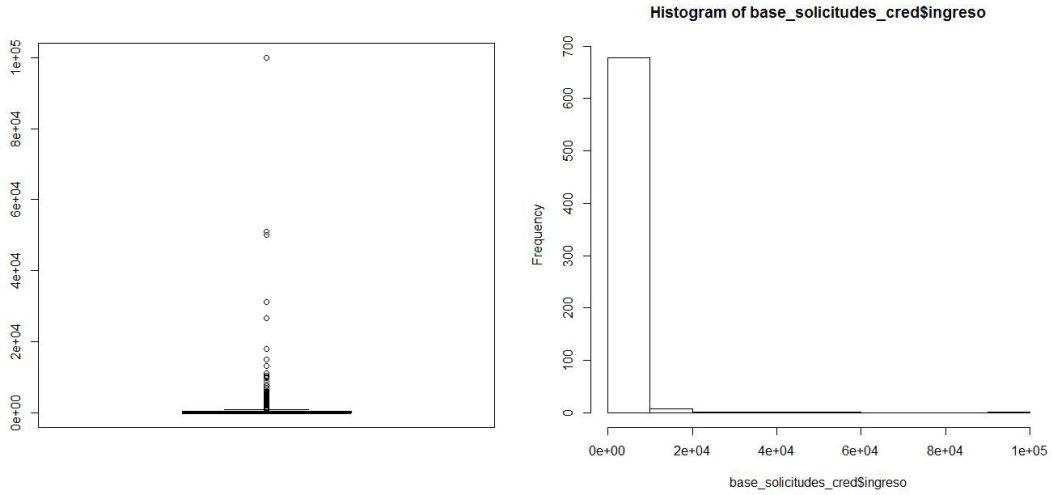

Figura 18. Ingreso de solicitantes

Tabla 4. Estadísticos básicos de variables numéricas

Variable	Mínimo	Cuartil 1	Media	Mediana	Cuartil 3	Máximo
Edad	13.75	22.60	31.57	28.46	38.23	80.25
Saldo	0.000	1.000	4.759	2.750	7.207	28.000
Años laborando	0.000	0.165	2.223	1.000	2.625	28.500
Calificación	0.0	0.0	0.0	2.4	3.0	67.0
Ingreso	0.0	0.0	1,017.4	5.0	395.5	100,000.0

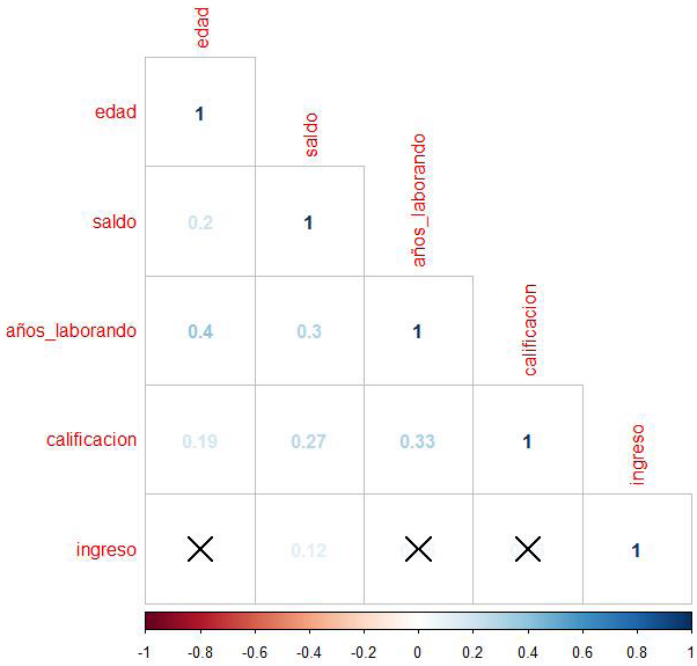

Figura 19. Matriz de correlación de variables numéricas

que no hay correlación entre el ingreso del solicitante y el resto de variables numéricas.

Preparación de datos

De acuerdo con los resultados del análisis exploratorio y a las pruebas realizadas para generar las bases de entrenamiento y validación se realizarán las acciones indicadas en la Tabla 5 como paso previo a la etapa de modelado.

Una vez aplicadas dichas acciones la base queda en 677 registros y se llamará en lo sucesivo "base limpia".

Posteriormente, se normalizan las variables numéricas: edad, saldo, años laborando, calificación e ingreso. La Figura 20 muestra la distribución de las variables sin normalizar (se observa que ninguna se parece a una distribución normal) y en la Figura 21 se muestra un fragmento de las variables normalizadas.

Finalmente, se divide la base limpia en dos partes para la etapa de modelado:

1. *Base de entrenamiento*: Con esta base se "entrenarán" ambos modelos (Random Forest y XGBoost) y tendrá 70 % del total de registros elegidos aleatoriamente (473 registros).
2. *Base de validación*: Con esta base se validarán los resultados de ambos modelos y contendrá 30 % restante del total de registros (204 registros).

Modelado

Se aplicaron los algoritmos Random Forest y XGBoost sobre la base de entrenamiento a fin de "entrenar" a los correspondientes modelos, con los parámetros default de cada algoritmo en R.

Posteriormente, se aplicaron los modelos "entrenados" sobre la base de validación. La Figura 22 muestra las curvas ROC obtenidas, que permiten comparar ambos algoritmos en términos de error tipo I (falsos positivos, eje X) y clasificaciones correctas (verdaderos positivos, eje Y) (Fawcett, 2005). Se observa que la curva

Tabla 5. Acciones previas a la etapa de modelado

Variable	Acción
Género	Se imputará con la moda (valor "b") los 12 registros con valor faltante
Estado civil	Se imputará con la moda (valor "u") los 6 registros con valor faltante y se descartan los 2 registros con valor "l"(a fin de evitar que queden concentrados en la base de entrenamiento o en la de validación)
Entidad bancaria	Se imputará con la moda (valor "g") los 6 registros con valor faltante
Nivel educativo	Se descartarán los 9 registros con valor faltante
Nacionalidad	Se imputará con la moda (valor "v") los 9 registros con valor faltante y se descartan los 2 registros con valor "o" (a fin de evitar que queden concentrados en la base de entrenamiento o en la de validación)
¿Ya aplicó previamente a una solicitud?	Se considerará sin cambio alguno
¿El solicitante labora actualmente?	Se considerará sin cambio alguno
¿El solicitante es ciudadano?	Se descartan los 8 registros con valor "p" (a fin de evitar que queden concentrados en la base de entrenamiento o en la de validación)
¿La solicitud de crédito fue aprobada?	Se reemplazarán los "+" por 1 y los "–" por 0
Código postal	Se descartará la variable por el alto volumen de registros asignados a código postal 0
Edad	Se imputará con la mediana (28.46 años) los 12 registros con valor faltante
Saldo	Se considerará sin cambio alguno
Años laborando	Se considerará sin cambio alguno
Calificación	Se considerará sin cambio alguno
Ingreso	Se considerará sin cambio alguno

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

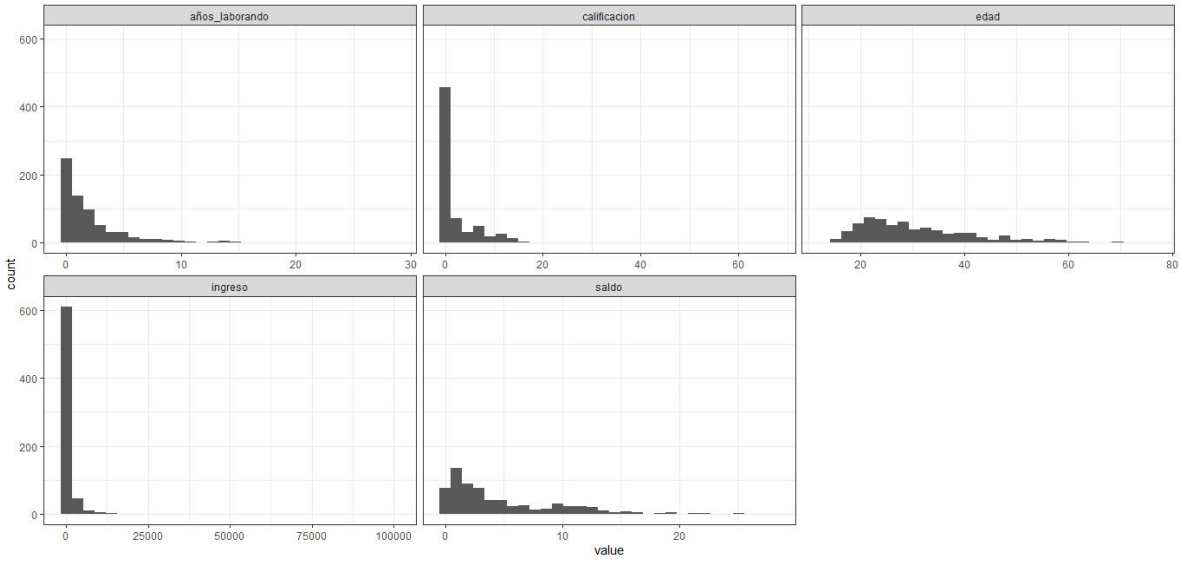

Figura 20. Distribución de variables numéricas sin normalizar

	edad	saldo	años_laborando	calificacion	ingreso
1	0.27111111	0.000000000	0.043859649	0.01492537	0.00000
2	0.71301587	0.159285714	0.106666667	0.08955224	0.00560
3	0.17063492	0.017857143	0.052631579	0.00000000	0.00824
4	0.22349206	0.055000000	0.131578947	0.07462687	0.00003
5	0.10190476	0.200892857	0.060000000	0.00000000	0.00000
6	0.29095238	0.142857143	0.087719298	0.00000000	0.00000
7	0.30825397	0.037142857	0.228070175	0.00000000	0.31285

Figura 21. Variables normalizadas (fragmento)

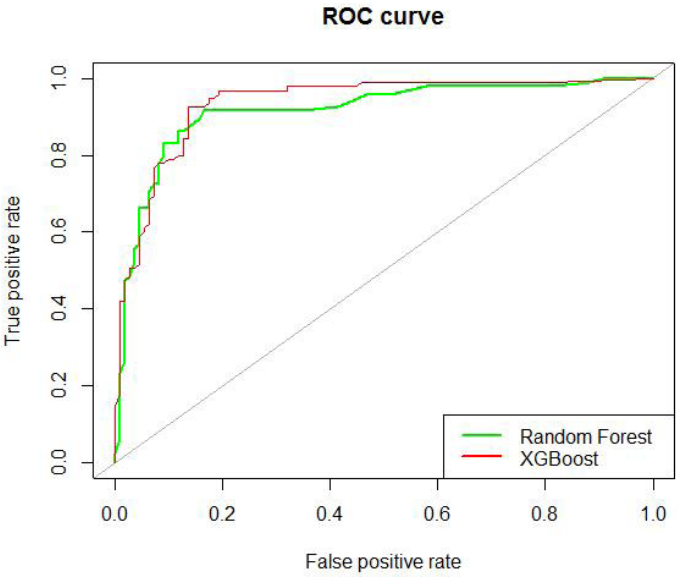

Figura 22. Curvas ROC (base de validación)

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

del modelo XGBoost tiene una tasa de verdaderos positivos ligeramente mejor al modelo Random Forest.

La Figura 23 muestra el AUC (área bajo la curva, métrica que mide la precisión de un modelo donde entre más cercano se encuentre a uno se considera que el modelo es más preciso) de los modelos Random Forest y XGBoost (en ambos casos sobre la base de validación) donde se observa que el modelo XGBoost tiene un AUC mayor al modelo Random Forest.

La Figura 24 muestra la importancia de predictores (es decir, cuáles son los predictores que más pesan para decidir si a una solicitud se le dará tarjeta de crédito en este caso) de acuerdo con el modelo Random Forest, donde se observa que "bandera1" (indica si el solicitante ya aplicó previamente para una tarjeta de crédito) es el predictor que pesa más.

La Figura 25 muestra la importancia de predictores de acuerdo con el modelo XGBoost, donde se observa que "bandera1" es también el predictor que más pesa para determinar si se da una tarjeta de crédito a una solicitud.

Análisis de resultados

Las curvas ROC y la métrica AUC obtenidas en la sección anterior dan el sustento técnico para comparar los modelos obtenidos (en este caso se elige el modelo XGBoost dado que su AUC es mayor), sin embargo, presentar solo estos resultados a una audiencia de negocios puede no ser suficiente. Uno de los mayores retos en minería de datos es presentar los resultados de un modelo a audiencias no técnicas en términos de negocio.

En el caso particular de la base analizada surgen dos preguntas de interés desde el enfoque de negocios: ¿Cuáles son las solicitudes a las que hay que otorgar una tarjeta? y ¿Qué resultados esperamos en caso de aplicar el modelo?

La Figura 26 responde la primera pregunta mediante la curva de ganancia acumulada de la base limpia ya considerando las probabilidades calculadas por el modelo XGBoost. Algunas características de esta gráfica son:

a) El eje X contiene deciles de clientes donde el decil 1 corresponde a las solicitudes con mayor probabilidad de que se les otorgue una tarjeta de crédito mientras que el eje Y contiene la probabilidad acumulada de otorgar una tarjeta de crédito por decil.
b) En el decil 5 se alcanza la máxima probabilidad acumulada lo que en términos de negocio significa que en lugar de gestionar la base completa de solicitudes el banco australiano podría enfocarse solo a 50 % de las solicitudes con mayor probabilidad de otorgarles una tarjeta de crédito, lo que redunda en una reducción de costos y esfuerzos. La probabilidad acumulada en el decil 5 para la base de validación es de 92 % (en la base de entrenamiento es de 98 %).
c) La curva continua azul (correspondiente al modelo obtenido en la base de validación) está cercana a la curva punteada del mismo color (correspondiente al modelo óptimo sobre dicha base), lo que indica que el modelo XGBoost obtenido ofrece resultados aceptables.

La Figura 27 responde la segunda pregunta mediante una curva de respuesta acumulada, donde se indica que el porcentaje de respuesta (solicitudes a las que se otorga una tarjeta de crédito en este caso) en la base de validación al considerar los 5 deciles con mayor probabilidad, de acuerdo con el modelo XGBoost, el porcentaje de respuesta esperada es de 85.3 %. De acuerdo con la gráfica en caso de no aplicar el modelo (líneas punteadas) el porcentaje de respuesta esperado es poco menor a 50 %.

```
> auc_rf <- auc(test$bandera5, pred_rf)
Setting levels: control = No, case = Si
Setting direction: controls < cases
> auc_rf
Area under the curve: 0.9138
>
> auc_xgb <- auc(test$bandera5, pred_xgb)
Setting levels: control = No, case = Si
Setting direction: controls < cases
> auc_xgb
Area under the curve: 0.9348
```

Figura 23. AUC de modelo Random Forest

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

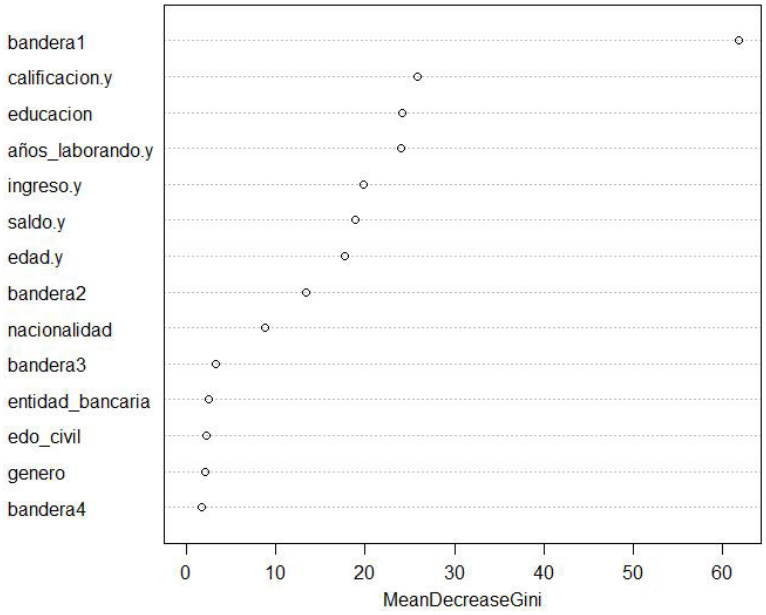

Figura 24. Importancia de predictores (modelo Random Forest)

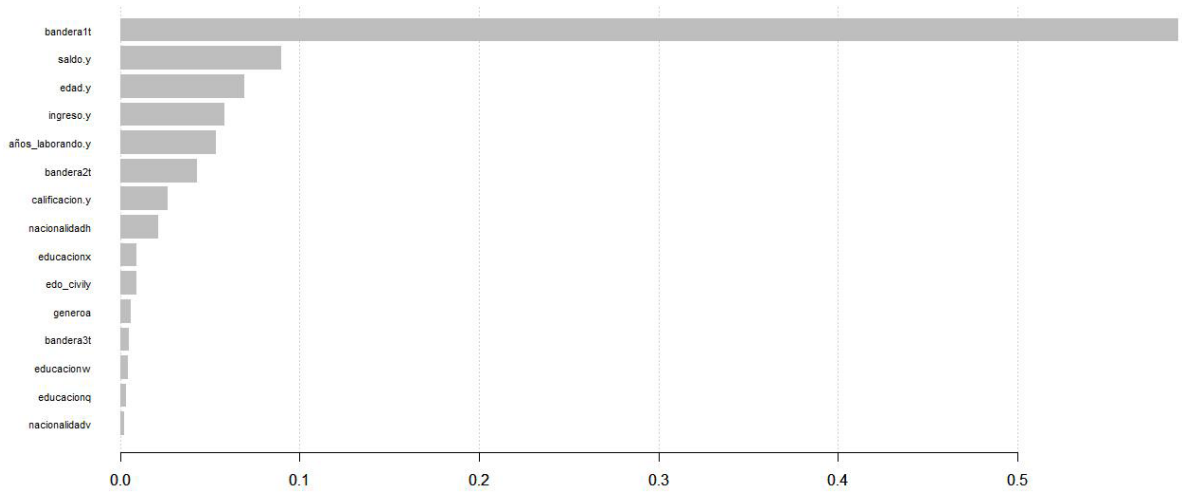

Figura 25. Importancia de predictores (modelo XGBoost)

https://doi.org/10.22201/fi.25940732e.2020.21.3.022

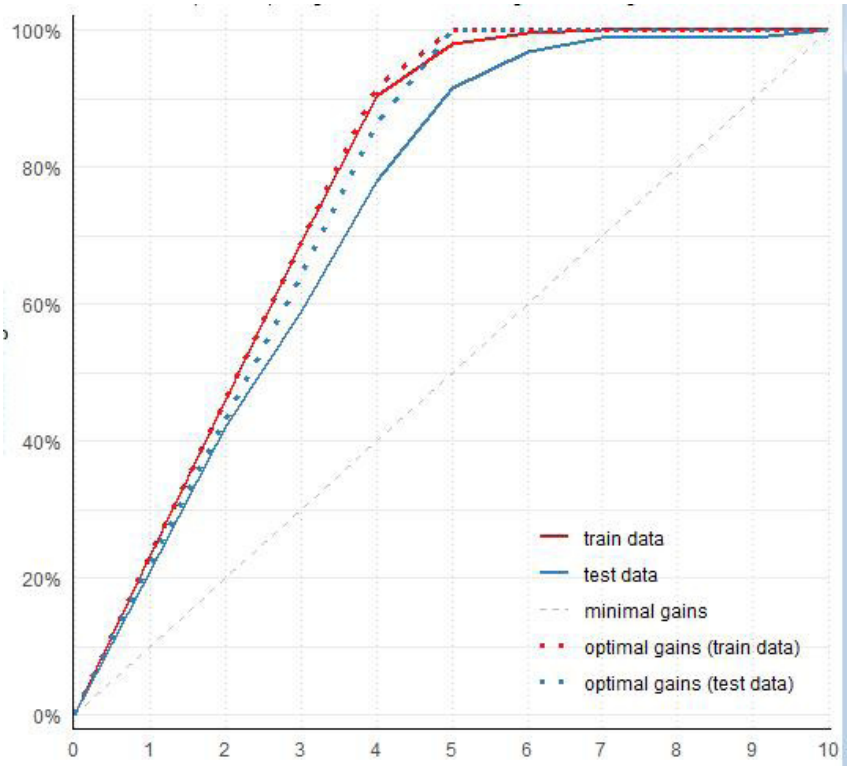

Figura 26. Curva de ganancia acumulada (modelo XGBoost)

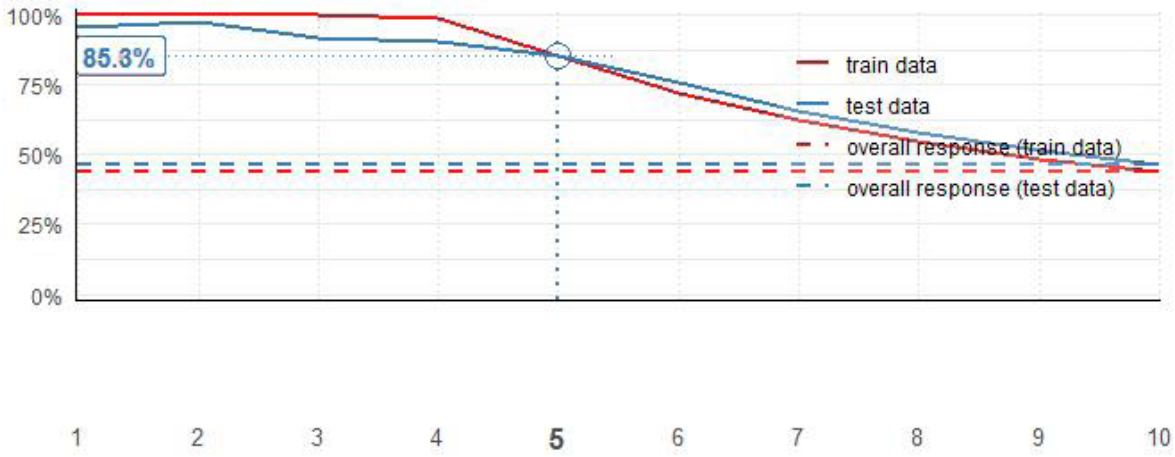

Figura 27. Curva de respuesta acumulada (modelo XGBoost)

Conclusiones

De acuerdo con el estudio realizado se tiene que:

1. El modelo XGBoost fue más preciso que el modelo Random Forest en esta base de solicitudes (esto coincide con la literatura, donde se señala que los resultados del algoritmo XGBoost generalmente superan a otros algoritmos).
2. Los modelos coinciden en determinar que el predictor más importante es que el solicitante haya solicitado previamente una tarjeta de crédito.
3. Las curvas de ganancia acumulada y respuesta acumulada dan un enfoque práctico a los resultados del modelo elegido (XGBoost en este caso), el cual es importante en el ámbito del que proviene la base (sector financiero).

Se proponen los siguientes pasos a fin de dar continuad al presente estudio:

- Replicar el estudio con otros modelos sobre la misma base de datos. Otros modelos que se puedan aplicar a fin de comparar su precisión con los resultados obtenidos en el presente estudio son: análisis discriminante, logit multinomial, redes neuronales, algoritmos genéticos y SVM,
- Considerar otras métricas de comparación adicionalmente a la curva ROC y métrica AUC, por ejemplo: coeficientes Kappa o Alfa.
- Replicar el análisis con una función de entrenamiento personalizada para XGBoost, este algoritmo permite utilizar funciones de entrenamiento creadas por el usuario, por lo que sería interesante replicar el análisis con una función de entrenamiento que el modelo no incluya por default.

Agradecimientos

El autor agradece a Grupo Financiero Ve por Más S.A. de C.V. su apoyo para realizar el presente artículo.

Referencias

Bahador, A., Movahedi A., Taghipour, H., Derrible, S. & Mohammadian, A. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136. https://doi.org/10.1016/j.aap.2019.105405.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32 https://doi.org/10.1023/A:1010933404324.

Cánovas, F., Alonso, F., Gomariz, F. & Oñate, F. (2017). Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. *Computers & Geosciences*, 103, 1-11. https://doi.org/10.1016/j.cageo.2017.02.012.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data 785-794. KDD '16. https://doi.org/10.1145/2939672.2939785.

CRAN. (2019). The comprehensive R archive network. Recuperado el 12 de junio de 2019 de The Comprehensive R Archive Network. https://cran.r-project.org/

Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

Huo, X., Bum, S., Tsui, L. & Wang S. (2006). FBP: A Frontier-Based tree-pruning algorithm. *INFORMS Journal of Computing*, 18 (4), 407-530. https://doi.org/10.1287/ijoc.1050.0133.

Lizares, M. (2017). Universidad Nacional Mayor de San Marcos. Recuperado el 10 de marzo de 2020 de http://cybertesis.unmsm.edu.pe/bitstream/handle/cybertesis/7122/Lizares_cm.pdf?sequence=1&isAllowed=y

Luckner, M., Topolski, B. & Mazurek, M. (2017). Application of XGBoost algorithm in fingerprinting localisation task. 16th IFIP TC8 International Conference, CISIM 2017 661-671. Bialystok, Poland: CISIM. https://doi.org/10.1007/978-3-319-59105-6_57.

Nobre, J. & Ferreira, R. (2019). Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. *Expert Systems with Applications*, 125, 181-194. https://doi.org/10.1016/j.eswa.2019.01.083.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 27 (3), 221-234. https://doi.org/10.1016/S0020-7373(87)80053-6.

Sandoval, L. L. (2017). Machine Learning algorythms for analysis and data prediction. 2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII), 1-5. Managua, Nicaragua. IEEE. http://doi.org/10.1109/CONCAPAN.2017.8278511

Tumer & Ghosh. (2002). Robust combining of disparate classifiers through order statistics. *Pattern Anal Appl*, 5, 189-200. https://doi.org/10.1007/s100440200017.

UCI. (2020). UCI Machine learning repository. Recuperado el 21 de mayo de 2019 de Center for Machine Learning and Intelligent Systems. http://archive.ics.uci.edu/ml/datasets/Credit+Approval

Universidad de California, I. (2019). Recuperado el 21 de mayo de 2019 de https://archive.ics.uci.edu/ml/datasets/Credit+Approval