UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSTGRADO

SISTEMA DE MODELOS PARA LA PREDICCIÓN DE PAGO DE CUOTAS ANTICIPADAS A PRÉSTAMOS

MODELOS DE "MEJOR PRÓXIMA OFERTA" EN PRODUCTOS BANCARIOS

Trabajo final de investigación aplicada sometida a la consideración de la Comisión del Programa de Estudios de Postgrado en Estadística para optar por el grado y título de Maestría Profesional en Estadística

ADRIÁN SOTO BLANCO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2021

DEDICATORIA

A mi familia que siempre ha sido el principal apoyo a lo largo de mi vida, así como la inspiración para buscar la superación académica.

AGRADECIMIENTO

Al M.Sc Fernando Ramírez Hernández, que a lo largo del desarrollo de las investigaciones estuvo disponible como guía.

Al M.Sc. Johnny Madrigal Pana, cuyos comentarios y sugerencias dieron pie a la culminación de los estudios y siempre veló por el buen análisis y contenido del documento.

Al M.Sc Felipe Ramírez Herrera que como guía de análisis, sugerencias innovadoras y comentarios del documento fueron parte fundamental del desarrollo del mismo. Además como mentor ha tenido gran influencia en mi desarrollo profesional.

A Bach. Carla López Gamboa, quien como colega, amiga y esposa fue un apoyo indispensable, tanto como consejera en temas técnicos, como personal.

“Este trabajo final de investigación aplicada I fue aceptado por la Comisión del Programa de Estudios de Postgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística”

FERNANDO RAMIREZ HERNANDEZ (FIRMA)

Firmado digitalmente por FERNANDO RAMIREZ HERNANDEZ (FIRMA)
Fecha: 2021.09.22 14:56:44 -06'00'

M.Sc. Fernando Ramírez Hernández
Profesor Guía

JOHNNY ANGEL MADRIGAL PANA (FIRMA)

Firmado digitalmente por JOHNNY ANGEL MADRIGAL PANA (FIRMA)
Fecha: 2021.09.21 20:53:42 -06'00'

M.Sc. Johnny Madrigal Pana
Lector

FELIPE ANTONIO RAMIREZ HERRERA (FIRMA)

Firmado digitalmente por FELIPE ANTONIO RAMIREZ HERRERA (FIRMA)
Fecha: 2021.10.08 10:10:46 -06'00'

M.Sc. Felipe Ramírez Herrera
Lector

ADRIAN SOTO BLANCO (FIRMA)

Firmado digitalmente por ADRIAN SOTO BLANCO (FIRMA)
Fecha: 2021.09.21 19:38:44 -06'00'

Adrián Soto Blanco
Sustentante

“Este trabajo final de investigación aplicada II fue aceptado por la Comisión del Programa de Estudios de Postgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística”

FERNANDO RAMIREZ HERNANDEZ (FIRMA)

Firmado digitalmente por FERNANDO RAMIREZ HERNANDEZ (FIRMA)
Fecha: 2021.09.22 14:57:17 -06'00'

M.Sc. Fernando Ramírez Hernández
Profesor Guía

JOHNNY ANGEL MADRIGAL PANA (FIRMA)

Firmado digitalmente por JOHNNY ANGEL MADRIGAL PANA (FIRMA)
Fecha: 2021.09.21 20:54:06 -06'00'

M.Sc. Johnny Madrigal Pana
Lector

FELIPE ANTONIO RAMIREZ HERRERA (FIRMA)

Firmado digitalmente por FELIPE ANTONIO RAMIREZ HERRERA (FIRMA)
Fecha: 2021.10.08 10:11:26 -06'00'

M.Sc. Felipe Ramírez Herrera
Lector

ADRIAN SOTO BLANCO (FIRMA)

Firmado digitalmente por ADRIAN SOTO BLANCO (FIRMA)
Fecha: 2021.09.21 19:39:13 -06'00'

Adrián Soto Blanco
Sustentante

Tabla de Contenido

RESUMEN PRÁCTICA PROFESIONAL I

Uno de los problemas que afrontan las entidades financieras al otorgar créditos es el de la posibilidad de que el cliente cancele anticipadamente el préstamo que se le haya concedido, o que realice pagos anticipados. Por ello, como una forma de prevenir dicho problema, en este estudio se proponen diversos métodos que permiten ajustar modelos matemáticos útiles para estimar el riesgo de que se presenten esos casos. Además, se identifican aquellas variables que captan los sistemas de información de las entidades financieras y que se relacionan con los clientes a quienes se les señale esa práctica. Entre los modelos ajustados se tienen los modelos de supervivencia, para el caso de los clientes que efectúan la cancelación anticipada de sus préstamos, y los modelos de regresión logística para los que hacen pagos anticipados. Para estos últimos se obtuvieron tasas de clasificación correcta de hasta 89,4%. A la vez se identificaron grupos con un riesgo de cancelación y de pago anticipado altos, lo cual le permite a la entidad financiera saber cómo minimizar la cantidad de pagos o de cancelaciones anticipadas de las operaciones comerciales; o, en algunos escenarios, buscar cómo colocar el dinero en otras operaciones para de esta forma obtener un mejor rendimiento.

RESUMEN PRÁCTICA PROFESIONAL II

Las entidades bancarias, como cualquier negocio en la actualidad, tienen la necesidad de rentabilizar a sus clientes y una forma de lograrlo es mediante la venta cruzada de productos, la cual se puede optimizar a partir de modelos de recomendación, permitiendo a cada cliente priorizar el producto a recomendar, basado en la probabilidad de su aceptación. Este método puede aprovechar los momentos de contacto con clientes, para realizar una venta cruzada de los mismos. El estudio realizado busca establecer un sistema de recomendación de productos a clientes bancarios que permitan desarrollar un proceso de ventas eficiente en futuras implementaciones, con las que se busque mejorar la práctica de ventas cruzada en la entidad financiera que lo requiera. Para lograr esto se analizan diferentes aproximaciones, entre las que se considera una adaptación del modelo XGBoost y filtrados colaborativos basados en contenido e ítems mediante un caso de estudio en el ambiente bancario, tomando datos disponibles en línea de la página de Kaggle Inc (2021). Este caso considera productos pasivos, activos y otros servicios financieros que tradicionalmente se ofrecen en las entidades bancarias, además de una gamma de características del cliente bancario que se aprovechan para determinar similitudes entre ellos, lo que aporta a la recomendación. Se identifican tres métodos que pueden ser utilizados como sistema de recomendación, de los cuales el algoritmo propuesto en esta investigación, “Próxima mejor oferta” y “recommenderlab” de Hahsler (2021) ofrecen una solución atractiva para el caso específico de estudio.

LISTA DE CUADROS

LISTA DE GRÁFICOS

LISTA DE FIGURAS

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Adrián Soto Blanco, con cédula de identidad 4-0203-0062, en mi condición de autor del TFG titulado Sistema de modelos para la predicción de pago de cuotas anticipadas a préstamos. Modelos de "Mejor próxima oferta" en productos bancarios.

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

ADRIAN SOTO BLANCO (FIRMA)

Firmado digitalmente por ADRIAN SOTO BLANCO (FIRMA)
Fecha: 2021.09.21 19:39:48 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

1. PRACTICA PROFESIONAL I

1.1 INTRODUCCIÓN

Desde sus orígenes, los seres humanos han transado con diferentes especias, bienes o servicios para lograr los objetivos personales de cada individuo o grupo. Es cuando las entidades financieras cobran importancia en la sociedad, al ofrecer servicios financieros que permiten distribuir el dinero entre aquellos grupos que lo necesitan, y también obtenerlo de las agrupaciones que no lo tienen en uso.

El negocio bancario puede dividirse en dos principales ramas que logran que el dinero sea rentable. Se refiere esto a la captación de dinero y a su colocación. La primera se refiere a la posibilidad de guardar el dinero en la cuenta de un banco, mientras que la segunda es el caso en el que la entidad financiera ofrece dinero como préstamo a cambio de que el cliente obtenga determinado monto por concepto de interés. Ambos escenarios son importantes, pues uno se convierte en la fuente de dinero para poder colocarlo posteriormente y con ello generar rentabilidad.

Por otro lado, un banco o entidad financiera se convierte en un actor importante en la sociedad, pues, como menciona Jiménez-Sandoval (1986), esta entidad se convierte en un intermediario entre ahorrantes e inversionistas y con ello logra distribuir el dinero de forma más eficiente (p. 25).

No obstante, como todo negocio, debe entenderse que este debe ser rentable en términos monetarios, lo que se logra al colocar el dinero a una tasa de interés mayor que aquella a la que se capta. Existen múltiples factores que pueden afectar este objetivo y, por ende, la rentabilidad de una entidad financiera. Entre otros, puede considerarse el de la sincronía oportuna o coincidencia entre la colocación y la captación de dinero, pues si bien es cierto pueden parecer negocios separados realmente van de la mano. Así, una entidad financiera recibe dinero con el objetivo de obtener dividendos de él, y para ello lo utiliza en financiamiento en beneficio de personas o de empresas a una tasa de interés superior a la que les pagaría a los dueños originales del dinero.

El cliente bancario recibirá el beneficio de que se le preste el dinero que no tiene para alcanzar algunos objetivos de su vida, como tener vivienda propia o vehículo o impulsar un negocio, entre otros proyectos que los prestatarios puedan tener. Por otro lado, la entidad financiera recibe lo esperado de la colocación de este dinero pero para la entidad bancaria existen riesgos al otorgar el crédito. Entre ellos se destaca el de que los clientes ahora dueños de su dinero lo pidan de vuelta sin que el banco disponga de él. Por otro lado, existe el riesgo de que el banco no haya prestado parte de ese dinero, y que eso esté ocasionando que no se genere el retorno esperado. Como el de mayor interés, el presente estudio se enfoca en este último problema tal como aquí se enuncia.

Específicamente, lo que se persigue con la investigación es identificar principalmente el riesgo implícito en el otorgamiento de cada préstamo bancario de que sea pagado anticipadamente, o sea, que la operación sea cubierta en su totalidad antes del plazo fijado; o que se le hagan pagos parciales no considerados en el plan original. Como se desprende de lo dicho, estas posibles acciones llevan a la entidad financiera respectiva a tener dinero en sus arcas del que no se obtiene el retorno esperado.

Manola & Urosevic (2010) resaltan la importancia de ejercer el control necesario para evitar prácticas como las mencionadas, al considerar que la más reciente crisis financiera (la del 2008 en Estados Unidos de América) se debió de cierta forma a la mala gestión de los bancos al no controlar los pagos y las cancelaciones anticipadas de préstamos hipotecarios, entre otras causas (p. 43).

El problema del pago o la cancelación anticipada de un préstamo se presenta en la mayoría de las entidades financieras, y el caso en estudio no es la excepción. Por ello su estudio es de interés para toda organización que desee controlar la cantidad de pagos anticipados o la cancelación prematura de un préstamo; esto mediante el uso de algún sistema matemático que posibilite medir los riesgos implícitos en la concesión del crédito. En tal caso el banco podrá disponer de su dinero para colocarlo de la mejor forma según los diferentes productos que ofrece.

Existen diferentes metodologías en el área de la estadística que pueden contribuir a la identificación y estimación del riesgo de que se presenten pagos o cancelaciones anticipadas de los préstamos, por lo que como resultado de esta investigación se plantean diferentes posibilidades para esos fines.

Las entidades financieras almacenan millones de datos día con día que tienen que ver directamente con el comportamiento financiero de los clientes de la empresa, lo cual hace que el valor de esos datos aumente; pues, como mencionan Provost & Fawcett (2013), los volúmenes de datos existentes en la actualidad permiten que las compañías dueñas de ellos los exploten con el fin de obtener una ventaja competitiva (p. 1). Este estudio ofrece una de esas ventajas, ya que determina una metodología útil para estimar el riesgo de cancelación anticipada de préstamos o el riesgo de pagos anticipados, es decir, los pagos parciales no contemplados en el plan original, que se refieren a operaciones en las que se debe tener una consideración especial y ofrecerse una opción según sea el riesgo que amenace al préstamo; esto para que la operación siga el curso esperado y no se desgaste por medio de pagos anticipados o se cancele anticipadamente.

El presente informe se divide en seis secciones o capítulos en los que se desarrolla la investigación. En el primero de ellos se introduce el tema y el segundo contiene el marco conceptual necesario para entender lo planteado de forma adecuada; en el tercero se especifican los objetivos por alcanzar y en el cuarto se establece la metodología adoptada; mientras que en el quinto se discuten los resultados de lo investigado. En el sexto se puntualizan las conclusiones obtenidas del desarrollo de la investigación y se ofrecen algunas recomendaciones al respecto.

1.2. MARCO CONCEPTUAL

Existen diversos conceptos en torno al negocio bancario que deben tenerse claros para entender los procedimientos sugeridos en este estudio para dicha actividad, por lo que a continuación se establece un marco conceptual explicativo de dichos conceptos, con la finalidad de facilitar su aplicación.

En el negocio bancario se pueden considerar múltiples actividades mediante las cuales una entidad financiera puede obtener ganancias, como el cobro de honorarios por la gestión de servicios, las transferencias internacionales, el uso de cajeros automáticos y la intermediación financiera. Gómez-Quesada (2008) describen la intermediación financiera como la acción de captar dinero del público en general para destinarlo a inversión o a la concesión de créditos, lo cual forma parte de una actividad importante para el desarrollo de la economía (p. 2). Por otro lado, Choudhry (2012) comenta que el negocio bancario tiene dos actores principales: prestamistas y prestatarios. Los primeros tienen a su disposición su dinero para que los prestatarios lo utilicen en actividades de inversión, por otro lado una entidad financiera es la que lleva a cabo o dirige esa actividad o proceso, en la que los inversionistas depositan su dinero en las cuentas bancarias, lo que hace que los fondos de la entidad aumenten y así se pueda prestar este dinero a los prestatarios (pp. 6-7). Anteriormente se describió este procedimiento como la acción de captar dinero a determinada tasa de interés para posteriormente prestarlo a una mayor.

En lo relativo a la colocación de dinero es en lo que surge el problema de estudio, pues existen casos de préstamos que son pagados por completo de forma anticipada, así como otros a los que se les efectúan pagos parciales, lo que desgasta la operación y puede desencadenar en una cancelación anticipada. En ambos casos el problema mayor radica en que el dinero que se espera tener colocado en diferentes operaciones generando ingresos finalmente se tiene como dinero ocioso y por eso no genera la rentabilidad esperada. Esto podría no ser un problema si se considerara que ese mismo dinero se podría colocar en otros préstamos; sin embargo, como menciona Alnsour (2013), se puede considerar que obtener un cliente nuevo es costoso (p. 128), lo que sugiere que puede ser más rentable retener a un cliente que adquirir uno nuevo. Dawes & Swailes (1999); Engel *et al* (1995); Ganesh *et al* (2000), y Paulin *et al* (1998) (en Van den Poel & Lariviere, 2004) refuerzan esta afirmación al considerar que la retención de un cliente disminuye la necesidad de encontrar otro potencialmente riesgoso, y ayuda a la organización a crear relaciones más estrechas con sus actuales clientes. Por otro lado, los clientes antiguos se vuelven menos costosos debido al conocimiento que tienen del banco, lo que contribuye en el decrecimiento de los costos de servicios (p. 3). Lo anterior ayuda a determinar que existen beneficios al hacerse esfuerzos por controlar los pagos anticipados de los préstamos, así como la cancelación anticipada de ellos.

Para manejar adecuadamente la cartera de préstamos bancarios es importante que exista el menor número posible de cancelaciones o pagos anticipados, como una forma de controlar esta práctica y de procurar que no se presente en la entidad financiera. Al

respecto, Blattberg *et al* (2008) consideran que ese control de la cartera de clientes empieza por determinar cuáles clientes presentan mayor riesgo de pérdida. Esto puede determinarse utilizando modelos predictivos e información conductual del cliente, como frecuencia, montos monetarios solicitados, quejas, ofertas previas, precios pagados atcualmente, o esfuerzos previos de retención y fuentes de adquisición, entre otros (p. 615). Para este caso se considera que es riesgo de cancelación de préstamos el valor por aproximar, que se obtiene utilizando modelos predictivos y una serie de variables relacionadas con la conducta del cliente para con el banco, lo mismo que para con la operación en particular, mediante los cuales se estimará el riesgo de cancelación o riesgo de pago anticipado de los préstamos.

Por otra parte, la estimación del riesgo de pago o cancelación anticipada puede ajustarse utilizando modelos predictivos supervisados. Clarke *et al* (2009) los caracteriza como aquellos en los que se tiene una variable respuesta (p. 231), la cual -en términos del problema en estudio- puede considerarse como la acción de cancelación o pago anticipado. Por otro lado, Consalvi & Scotto di Freca (2010) consideran que los modelos conductuales pueden explicar de forma correcta las tendencias de pagos anticipados en hipotecas (p. 2), acción que se puede generalizar a otros tipos de préstamos. Este caso lleva a considerar las herramientas estadísticas como los modelos de regresión logística y de supervivencia, los que ajustan una probabilidad ante la ocurrencia de un evento, y eso puede considerarse como el riesgo de ocurrencia. Para el caso de la cancelación de préstamos se puede ajustar la probabilidad de que el préstamo sea cancelado en un tiempo dado, mientras que para determinar el riesgo de pago anticipado parcial se puede determinar la probabilidad de que ocurra en un mes o periodo dado.

En el ámbito bancario se han utilizado diversas metodologías para abordar problemas financieros clásicos. Consalvi & Scotto di Freca (2010) utilizan un análisis de supervivencia para estimar el riesgo de pagos anticipados en las hipotecas. Te-Hsin & Jian-Bang (2014) utilizaron un modelo de dos etapas para predecir el pago anticipado también en las hipotecas, en que la primera etapa es una segmentación de la cartera y la segunda un modelo predictivo ajustado a partir de modelos de supervivencia. Además, mencionan otros estudios en los que se usan árboles de clasificación, bosques aleatorios y regresiones logísticas (pp. 329, 331). Otros autores, como Blattberg *et al* (2008), sugieren la estimación de un sistema de modelos que permita determinar el valor de vida de los clientes utilizando diferentes estratégias, entre otras los modelos de supervivencia para ajustar una curva de vida de los clientes (p. 111). Todos ellos respaldan el uso de estas metodologías para abordar problemas como el que se busca solventar en este estudio, por lo que se justifica considerar estas técnicas.

Por otro lado, los modelos predictivos y de supervivencia están sujetos a un entorno de datos que los alimenta inicialmente y que podría no ser el mismo en el momento de predecir el comportamiento esperado en los préstamos, por lo cual las técnicas utilizadas deben ser validadas adecuadamente para minimizar este riesgo. Kuncheva (2004) sugiere el método "*Hold-out*", que consiste en separar de forma aleatoria parte del conjunto de datos para ajustar los modelos, utilizando una parte, y para con la otra poder calcular los

indicadores de error. También sugiere el método *Cross-validation*, o validación cruzada, que consiste en dividir el conjunto de datos en "K" partes independientes, ajustar el modelo separando en cada subconjunto K y luego estimando el error en cada subconjunto, para de esa forma determinar el error en diferentes escenarios posibles (p. 9).

Una vez determinado el riesgo de cancelación o pago anticipado de cada operación se pueden distinguir aquellos préstamos con mayor riesgo que otros, lo que ofrece un panorama de control de los préstamos, en el que se pueden aplicar estrategias de negocio distintas según el nivel de riesgo de cada una, para buscar retener aquellas operaciones que aún ofrecen rentabilidad para el negocio o para rescatar aquellas que ya no son lucrativas, o para los dos propósitos.

Algunas estrategias de rescate de operaciones que presentan alto riesgo de cancelación pueden ser el ofrecimiento de otro producto similar para financiar algún otro proyecto del cliente bajo la misma garantía del producto en riesgo. Podría considerarse también alterar las condiciones del préstamo para favorecer al cliente, siempre que no se hagan pagos o cancelaciones anticipadas durante un determinado tiempo. La aplicación de cualquier estrategia debe dejarse a discreción de la entidad financiera, de forma que no afecte negativamente su accionar, sino que más bien contribuya a su fortalecimiento.

El comportamiento de los préstamos puede ser muy variable si se considera que existen montos, plazos y tasas de interés, entre otros aspectos, que pueden influenciar de cierta forma en la decisión de cancelar el préstamo o en iniciar una serie de pagos fuera del plan original con el fin de disminuir la deuda. Una forma de categorizar los préstamos es según la garantía de pago que exija el banco para otorgar el crédito. En este estudio se consideran los préstamos hipotecarios, los prendarios y los de consumo, tal y como se clasifican seguidamente:

- Hipotecario. Se trata de aquellos préstamos cuya garantía consiste en gravar una propiedad del cliente a favor del bancario.
- Prendario. En este caso la garantía consiste en gravar un vehículo o bien móvil del cliente bancario.
- Consumo. En este caso el préstamo se otorga sin garantía de bienes directamente, aunque en ocasiones se solicita un fiador, que sería el actor que deberá hacerse cargo de la deuda en caso de incumplimiento por parte del cliente bancario.

Según datos del Banco Central de Costa Rica[1] la tasa de interés de préstamos para construcción o vivienda oscilan entre 8% y 18% anual; el de consumo y para compra de vehículos[2] varía entre 7,4% y 33,8%. Además, los plazos y montos de los préstamos para consumo y prendario pueden compararse entre sí; lo que no ocurre con los préstamos hipotecarios, pues al tenerse -en la mayoría de los casos- garantía de mayor valor al igual que deudas con plazos mayores, el monto de apertura de los préstamos suele ser mayor.

[1] Banco Central de Costa Rica, 2014, consulta realizada para cada tipo de préstamo del 01-01-2017 hasta 31-12-2017.
[2] En el sistema de consulta los préstamos para vehículos se incluyen en el rubro de consumo.

1.3. OBJETIVOS

De conformidad con lo planteado como el problema de estudio, que es la ocurrencia de pagos y cancelaciones anticipados en una entidad financiera, y que incide en el desgaste de la cartera de préstamos, se enuncian como objetivos de estudio los siguientes:

1.3.1. Objetivo general

Estimar la probabilidad de que un cliente determinado haga un pago o cancelación anticipada de un préstamo, con el propósito de darle seguimiento y estructurar estrategias que posibiliten minimizar la ocurrencia de estos hechos en la cartera de préstamos, y mantener la rentabilidad de los productos.

1.3.2. Objetivos específicos

- Estimar el posible impacto de las variables del comportamiento transaccional bancario, así como las demográficas en el fenómeno, e identificar las más importantes.
- Determinar modelos para estimar la probabilidad de que un cliente haga un pago o cancelación anticipada, y evaluar dichos modelos en términos de practicidad y eficiencia.
- Proponer una posible estratificación de la cartera de créditos según su probabilidad de cancelación o pago anticipado, o ambos.

1.4. METODOLOGÍA

En esta sección se describen los métodos y las técnicas utilizados lo mismo que algunos métodos de interpretación de datos, la fuente de ellos, las características del contenido de esos datos, el periodo de análisis y los indicadores de precisión para comparar las técnicas, entre otros aspectos.

1.4. 1. Fuente de datos

Se tuvo acceso al sistema de una entidad bancaria y se recolectaron datos de 25 meses, de setiembre del 2015 a setiembre del 2017. Estos datos incluyen la información de los préstamos activos por mes, así como su saldo pendiente, monto de apertura, plazo, fecha de apertura, moneda, entre otras características que se detallan en el anexo 1.

De setiembre del 2017 se estudiaron en total 28 901 préstamos, de los cuales la mayoría son préstamos prendarios (53%, aproximadamente); seguidos por préstamos hipotecarios o de consumo hipotecario[3], y préstamos de consumo (aproximadamente 46% y 2%, respectivamente). En estos casos se aprecia que los saldos de los préstamos hipotecarios son superiores a los de los demás tipos. Estos representan 84% de la cartera estudiada, mientras que los préstamos prendarios y de consumo acumulan entre los dos tipos 16%, aproximadamente (cuadro 1).

Cuadro 1. Promedio mensual de préstamos estudiados al corte de setiembre 2017

Tipo de préstamo	Total de operaciones	Porcentaje	Porcentaje de cartera
Consumo	523	1,8	0,7
Hipotecario	13 168	45,6	84,3
Prendario	15 210	52,6	15,0
Total general	*28 901*	*100,0*	*100,0*

Fuente: Almacén de datos de la entidad financiera. Setiembre de 2015 - setiembre de 2017.

Entre las variables consideradas existen casos de valores faltantes en variables demográficas, tales como edad, antigüedad de la relación bancaria y sexo del cliente; esto debido a errores de registro. En el caso de la edad los faltantes son apenas 0,01%, mientras que la antigüedad bancaria es de 1,24% y en sexo es de 0,02%. Para trabajar con esos datos se decidió colocar el valor cero para la edad y para la antigüedad, debido a que de esta forma se anula el efecto que esta variable tendría en el modelo al considerarse una ecuación lineal, pues el valor faltante multiplicaría por cero el parámetro de la ecuación. Para el caso del sexo se agregó la categoría "otros". Se debe aclarar que sustituir la edad y la antigüedad del cliente por cero cuando no se tiene un valor asignado

[3] Se refiere a los préstamos con una finalidad de consumo pero que dejan como garantía una hipoteca. En este estudio estos préstamos son considerados como hipotecarios.

es únicamente una medida remedial para que el modelo se pueda estimar; sin embargo, en estos casos no se tendría una interpretación en dichas variables, pues son valores imputados. Al ser pocos los casos en los que ocurre este fenómeno se espera que no haya un impacto significativo en la veracidad del análisis.

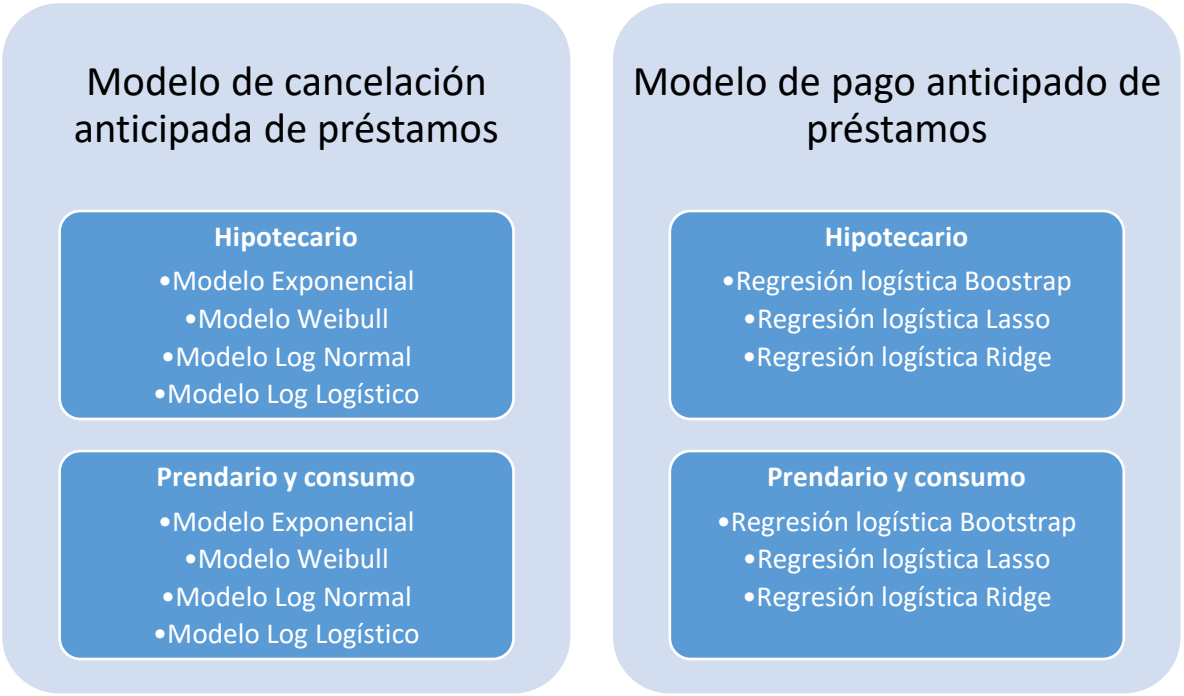
Los datos utilizados son confidenciales, por lo que no se puede dar detalle de ellos más allá de los resultados de este estudio. Se mantienen bajo confidencialidad el nombre de la entidad financiera y la naturaleza de los datos. Es decir, no se especifica si se trata de una muestra o de la población.

En este estudio se utilizarán dos estrategias para determinar la probabilidad de que un préstamo sea cancelado en los siguientes tres meses y para estimar la probabilidad de que en un préstamo se presente un pago anticipado. Estas estrategias se escogieron a partir de la necesidad de negocio de la entidad financiera, considerando que el tiempo mínimo requerido para actuar ante una operación en riesgo son tres meses, según el personal operativo de la entidad. Las variables respuesta son las que se detallan seguidamente.

- Cancelación anticipada. Se determina que un préstamo ha sido cancelado si tiene saldo pendiente igual a cero y si, además, faltan más de tres meses para su fecha de vencimiento original.
- Pago anticipado. Se determina que a un préstamo se le ha hecho un pago anticipado si en un mes dado registra un monto extra pagado superior a 25% de la cuota por pagar en dicho mes.

Para cada caso se compararon diferentes tipos de modelos. En el de cancelación anticipada se buscó el modelo de supervivencia que mejor se adaptara a los datos, mientras que en el de pagos anticipados se consideró el modelo de clasificación que ofrece la mejor predictibilidad. En cada caso se dividió el archivo de datos en dos partes. La primera estuvo compuesta por 60% de los registros totales, la cual se utiliza para ajustar los modelos, y se reserva el 40% restante para predecir el comportamiento y calcular los indicadores de precisión. En la figura 1 se resumen los modelos considerados en el análisis.

Figura 1. Resumen de modelos propuestos por tipo de análisis



1.4. 2. Modelos de cancelación anticipada de préstamos

Según los objetivos de estudio, para determinar un modelo que se adapte a la ocurrencia de cancelación de préstamos se consideraron cuatro de los principales modelos de análisis de supervivencia paramétrica con tiempo acelerado de falla. Se utilizaron estos modelos ya que al tener cada tipo de préstamo un plazo distinto se conoce que el tiempo de vigencia de la operación será una variable importante en la cancelación o no de esa operación. Los modelos basados en Rodríguez (2010, pp: 1-5) y Klein & Moeschberger (2003, pp: 38, 393 - 408) son:

- Exponencial
  - Se caracteriza por tener una distribución de riesgos constante.
  - Se relaciona con distribuciones con valores extremos en ella.
  - El modelo matemático no posee parámetros de anclaje.

$$P(T > t) = S(t|x) = e^{-t \times e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

Con p: Número de variables independientes en el modelo

- Weibull
  - La distribución de riesgos se distribuye creciente si el parámetro de anclaje es mayor que la unidad y decreciente si el parámetro es menor que uno.
  - Se relaciona con distribuciones con valores extremos en ella, al igual que con la exponencial.

$$P(T > t) = S(t|x) = e^{-t^{\alpha} \times e^{-\alpha \times (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

Con p: Número de variables independientes en el modelo

α: Parámetro de anclaje para el modelo Weibull

- Log normal
    - Supone que los errores relacionados con el modelo se distribuyen en forma normal.
    - En términos de practicidad se dificulta su programación pues se requiere identificar los cuantiles de una distribución normal estándar.

$$P(T > t) = S(t|x) = 1 - \varphi\left[\frac{\ln(t) - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{\sigma}\right]$$

Con p: Número de variables independientes en el modelo
σ: Parámetro de anclaje para el modelo Log Normal
φ: Distribución normal estándar acumulada

- Log logístico
    - Se tiende a parecer al modelo Log normal cuando el parámetro de anclaje es mayor que la unidad.
    - Supone que los tiempos de supervivencia se distribuyen en forma normal.

$$P(T > t) = S(t|x) = \frac{1}{1 + \left(t \times e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}\right)^{1/\gamma}}$$

Con p: Número de variables independientes en el modelo
Υ: Parámetro de anclaje para el modelo Log logístico

1.4. 3. Modelos de pagos anticipados de préstamos

Para esta etapa de la investigación se reestructuró el archivo de datos de forma que cada fila representa un determinado préstamo en un mes dado. La información de cada préstamo corresponde a la información del mes previo observado, pues se considera que será la verdadera observación en el momento de predecir. No obstante, la variable respuesta de pago anticipado se calcula para el mes dado. Se procede así porque en el momento de ejecutar el modelo se tiene información para un mes dado y se desea predecir el pago anticipado con un mes de antelación para poder tomar medidas que mitiguen los efectos del fenómeno estudiado.

Para este caso se utilizaron modelos predictivos que permiten la programación de una ecuación en un sistema informático simple, para que la predicción pueda ser automática mes a mes, por lo que se consideraron diferentes tipos de modelos de regresión múltiple.

Por ser la variable por predecir de carácter binario se consideran modelos de regresión logística que ofrecen una probabilidad ajustada a que ocurra el evento. En este caso se utilizó el pago anticipado de un préstamo con la codificación 1 (0, el no pago anticipado). Hosmer, Lemeshow, & Sturdivant (2013, p. 35-36) establecen el modelo de regresión logístico como un modelo lineal a partir de la transformación:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Donde: $g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

p: Número de variables independientes en el modelo.

Existen diversos métodos de estimación de cada parámetro en un modelo de regresión múltiple logística. Cada uno tiene sus ventajas y desventajas; pero siempre se debe buscar perfeccionar el método con el fin de que el resultado ofrezca la estimación más estable, lo que se puede traducir en la más predictiva, suponiendo un menor error relativo en su estimación. Para lograr esto, autores como Tibshirani (1996) proponen usar modelos de penalización que, mediante la estimación de un parámetro de penalización o encogimiento, reducen los parámetros hasta que sean los más pequeños posibles. Con esto puede mejorarse la precisión de predicción al disminuir el tamaño de los coeficientes y también se reduce su sesgo (p. 267).

Por otro lado, determinar el mejor método de estimación de una regresión logística lleva a la posibilidad de considerar un gran número de modelos. Para el caso de este estudio se consideraron dos métodos tradicionales de penalización, como son el de la regresión Lasso y el de la regresión Ridge. Adicionalmente se utilizó una regresión con parámetros resultantes de múltiples regresiones tradicionales usando Bootstrap. Al comparar estos modelos se identifica aquel que efectivamente ofrece la predicción más precisa para el problema dado.

Se consideran los siguientes modelos, con sus ventajas y desventajas:

- Regresión logística múltiple con estimación Bootstrap.
    - Cálculo rápido, lo que permite que el Bootstrap sea aplicado sin requerir mucho tiempo.
    - No requiere la estimación de parámetros adicionales.
- Regresión Lasso logística.
    - Estimación del parámetro de penalización por validación cruzada, lo que produce una estimación más lenta.
    - Elimina automáticamente parámetros que no son importantes en el modelo y que pueden haberse filtrado en la selección inicial de variables (hace una selección de variables automática).
    - Ofrece una estimación robusta
- Regresión Ridge logística.
    - Estimación del parámetro de penalización por validación cruzada, lo que produce una estimación más lenta.
    - Ofrece una estimación robusta.

1.4. 4. Medidas de precisión

Para estudiar la precisión en la predicción de los dos fenómenos planteados, se divide el archivo de datos en dos muestras aleatorias, la primera denominada tabla de aprendizaje

(60%), que se usa para ajustar una serie de modelos predictivos, mientras que la segunda muestra llamada tabla de prueba (40%), se considera para probar los modelos y estimar los siguientes indicadores de precisión:

- Linealidad de la probabilidad conforme a la tasa de ocurrencia de cada evento. Para visualizarla se dividen las probabilidades ajustadas en grupos de igual tamaño y se calcula la proporción de aquellos a los que les ocurre el evento. Se espera observar una tendencia lineal entre los grupos, en que los grupos con mayor riesgo de ser afectados tengan una tasa de ocurrencia mayor y disminuya conforme el riesgo sea menor.
- Clasificación correcta. Se consideran los registros con una probabilidad ajustada superior a determinado umbral, para clasificarlos como aquellos en los que ha ocurrido el evento. Este se optimiza para encontrar el punto de corte adecuado. En este caso se consideran los siguientes indicadores.
    - Porcentaje de clasificación correcta total.
    - Porcentaje de aciertos en cancelación o pago anticipado de préstamos.
    - Porcentaje de aciertos en no cancelación o no pago anticipado de préstamos.
- Curva ROC. Identifica el área acumulada bajo una curva de predicción correcta utilizando el modelo predictivo y un patrón aleatorio de clasificación. Intuitivamente sugiere la ganancia en la predicción por haber utilizado el modelo predictivo, en contraposición a determinar la cancelación o pago anticipado de préstamos de forma aleatoria.

Tanto el porcentaje de clasificación correcta como la curva ROC surgen de una lógica similar de predicción correcta de categoría.

Suponga que se tiene el siguiente cuadro 2:

Cuadro 2. Ejemplo para cálculo de medidas de precisión

Realidad	Predicción		Total
	No ocurrencia	Ocurrencia	
No ocurrencia	A	B	A+B
Ocurrencia	C	D	C+D
Total	A+C	B+D	A+B+C+D

Se obtienen las siguientes definiciones:

Porcentaje de clasificación correcta total:

$$\frac{A+D}{A+B+C+D}$$

Porcentaje de "Ocurrencias" correcta o sensibilidad:

$$\frac{D}{C+D}$$

Porcentaje de "No Ocurrencias" correcta o especificidad:

$$\frac{A}{A+B}$$

En el caso de la curva ROC se calcula para diferentes puntos corte, la sensibilidad y el recíproco de la especificidad (1-especificidad); se grafican los datos colocando en el eje Y la sensibilidad y en el eje X (1-especificidad), y se calcula el área acumulada entre la curva estimada y una recta identidad. Este valor cuanto más cercano a la unidad se considera mejor.

La estimación final de cada modelo se calcula con la técnica "Bootstrap", descrita por Efron (1979), en el que a partir de muestras simples aleatorias con reemplazo, seleccionadas del conjunto de datos de estimación, se ajustan los modelos y se obtienen los coeficientes, para posteriormente resumirlos en un único coeficiente centrado en el parámetro poblacional con un error estándar dado por las muestras seleccionadas.

Para el caso de la estimación del porcentaje de clasificación correcta se debe determinar un punto de corte adecuado para la probabilidad ajustada, lo que se logra al considerar todos los posibles puntos de corte en un intervalo, y se determina que el óptimo es aquel que logra una desviación estándar mínima entre la especificidad y la sensibilidad; es decir, en que la clasificación correcta de ocurrencia o no del evento es la más parecida entre sí.

En el caso de que los resultados no sean concluyentes usando la división del archivo de datos reservado para hacer una prueba, se procede a hacer una validación cruzada. Esta técnica, según Han, Kamber & Pei (2012), consiste en obtener "k" muestras independientes de un mismo archivo de datos. Luego, con la restante porción de la muestra, se ajusta el modelo y se calculan los indicadores de precisión para cada muestra. Este proceso se repite "c" veces y al final se obtiene una validación en la que se consideran múltiples escenarios, así como la estabilidad del modelo (pp: 370-371).

1.4. 5. Selección de variables

Se determinó un procedimiento para seleccionar las variables que puedan asociarse con la cancelación o pago anticipado de un préstamo. Este procedimiento se compone de los siguientes pasos:

1. Ejecutar un procedimiento de selección por pasos utilizando cada modelo considerado.
2. Calcular las medidas de asociación con los dos indicadores por predecir.
3. Asociación con tiempo de duración del préstamo.
    a. Variables métricas: coeficiente de correlación de Pearson.
    b. Variables categóricas: coeficiente Eta.

4. Asociación con cancelación o pago anticipado del préstamo.
    a. Variables métricas: coeficiente Eta.
    b. Variables categóricas: coeficiente de correlación tetracórica.
5. Confeccionar un cuadro comparativo en el que se puedan determinar las variables con mayor peso en el modelo[4].

1.4. 6. Valores extremos

Existen múltiples valores extremos identificados en las variables predictivas, los que surgen por la propia naturaleza del negocio financiero en el que se desarrolla el proyecto. Para identificarlos se aplicó para cada variable explicativa un modelo de regresión contra una constante, para el cual se calcularon las distancias de Cook usando la biblioteca del paquete en R "referenceIntervals"[5]. Posteriormente se identificó el porcentaje de individuos determinados como extremos y si era menor de 2% se buscaban los valores máximo y mínimo del archivo sin valores extremos, y este se fijó como valor máximo o mínimo para los casos considerados extremos.

1.4. 7. Variables de mayor impacto estimado en el fenómeno

Para determinar el peso relativo de cada variable considerada se parte del principio de que aquel parámetro con menor variación es el que tiene mayor impacto en el fenómeno estudiado. Para lograr esto se procedió a realizar la estimación de los coeficientes del modelo más adecuado, utilizando una estimación Bootstrap con 1 000 muestras aleatorias. Después, para determinar el parámetro ajustado se obtiene el promedio de la estimación en cada muestra y se calcula su error estándar; y luego se calcula un indicador dado por:

$$\frac{|\hat{\theta}|}{Error\ Estandar(\hat{\theta})}$$

Lo anterior permite comparar la variabilidad de los coeficientes entre sí. Seguidamente se calcula el valor absoluto del indicador anterior y se ajusta de forma que la suma de todos sea igual a 100%. A mayor valor del indicador mayor importancia en el modelo, o, lo que es lo mismo, a menor coeficiente de variación[6] mayor importancia en el modelo.

[4] Para el caso de los modelos de pagos anticipados se omite el cálculo de medidas de asociación con el tiempo de duración del préstamo.
[5] Finnegan, D. (2014)
[6] Coeficiente de Variación = $\frac{Error\ Estandar(\hat{\theta})}{|\hat{\theta}|}$

1.4. 8. Segmentación

Para contribuir con la estrategia de retención se deben clasificar los casos en los que se tenga un mayor riesgo de hacer un pago anticipado o una cancelación automática mes a mes. Para lograrlo se propone estratificar los préstamos según su probabilidad de cancelación o de pago anticipado, usando el método K-Medias. Según Friedman *et al* (2009) este algoritmo busca identificar grupos y centros de grupos en datos que no tienen agrupación, siempre que se seleccione previamente el número de categorías deseadas. El procedimiento se inicia a partir de unos centros de datos seleccionados aleatoriamente y agrupa a los sujetos según su distancia de cada grupo, para posteriormente volver a calcular los centros. Este método continúa iterativamente hasta encontrar una variancia mínima dentro de los grupos (pp. 475).

Se pretende identificar el número óptimo de agrupaciones probando entre 2 y 10 grupos; sin embargo, para poder discriminar de mejor forma se prefiere tener grupos impares que ofrezcan un punto medio entre el riesgo de cancelación o pago anticipado de los préstamos.

Se creó una agrupación usando la probabilidad de cancelación del préstamo asignada y otra usando la probabilidad de pago anticipado ajustado a cada préstamo, por lo que para calcular la suma de cuadrados de estos indicadores entre y dentro de los grupos se sugiere como criterio para identificar el número de grupos adecuado. No obstante, esta segmentación debe estar dirigida principalmente a reforzar una posible campaña en la que se usarán los datos, por lo cual el criterio de negocio se debe considerar para decidir el número de grupos, en el que no pueden ser tan pocos que no discriminen entre sí pero tampoco pueden ser demasiados de forma que se dificulte crear campañas segmentadas para cada grupo.

1. 5. RESULTADOS

En esta sección se resumen los principales hallazgos de este estudio enmarcados dentro de los objetivos propuestos. Se detallan por tipo de modelo para facilitar su análisis.

1. 5. 1. Modelos de cancelación anticipada de préstamos

En el cuadro 3 se resume la distribución porcentual promedio mensual de los préstamos cancelados anticipadamente durante el periodo de estudio. Es importante apreciar que, en promedio, 229 préstamos fueron cancelados al menos tres meses antes de la fecha de finalización, y de estos 4,2% son hipotecarios. Debe destacarse que la rentabilidad de este tipo de cartera produce que el impacto económico sea alto, pues el saldo en estos tipos de préstamo es mayor. De las cifras de este cuadro se deduce, además, que la tendencia de cancelación varía tanto por tipo de cartera como por mes. Por ejemplo, los préstamos prendarios se cancelan principalmente en enero y diciembre, mientras que los hipotecarios son cancelados en marzo y setiembre. Los de consumo tienen un patrón más difícil de detectar pero que parece aumentar levemente en octubre.

Cuadro 3. Distribución porcentual promedio por mes de los préstamos cancelados mensualmente

Mes	Consumo (%)	Hipotecario (%)	Prendario (%)	Total (%)
Enero	36,9	4,2	58,9	100,0
Febrero	53,1	3,6	43,3	100,0
Marzo	49,6	6,1	44,3	100,0
Abril	55,0	4,8	40,2	100,0
Mayo	51,7	4,2	44,1	100,0
Junio	49,1	3,3	47,6	100,0
Julio	40,1	4,6	55,3	100,0
Agosto	43,0	3,5	53,5	100,0
Setiembre	47,3	5,3	47,3	100,0
Octubre	56,7	4,0	39,3	100,0
Noviembre	40,6	3,4	56,0	100,0
Diciembre	37,9	2,9	59,2	100,0
Total general	*46,4*	*4,2*	*49,4*	*100,0*

Nota: En promedio se cancelan 229 préstamos mensualmente[7]
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

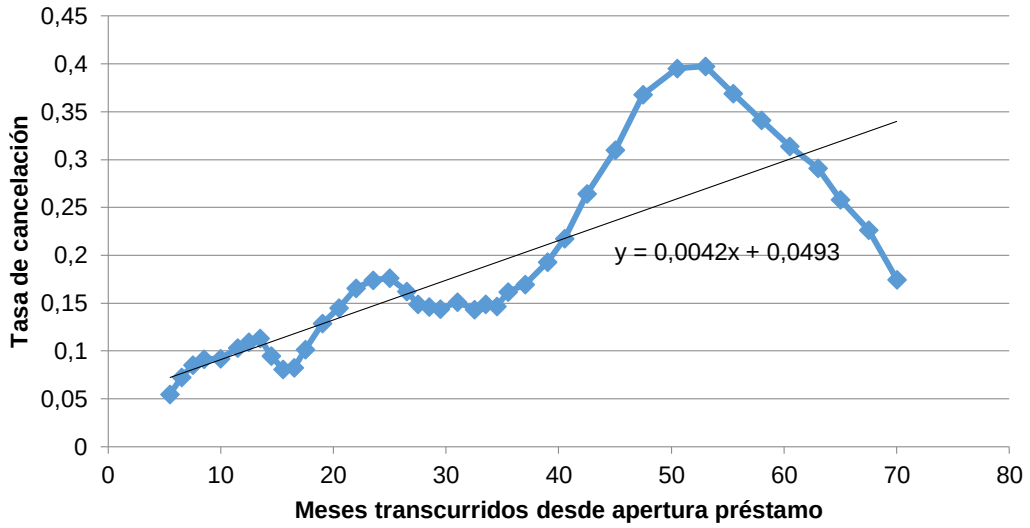
Se consideró ajustar modelos para préstamos hipotecarios, por un lado, y posteriormente préstamos de consumo y prendarios de forma conjunta. Esto debido a que los dos últimos tienen características de plazo, saldos y tasas similares, según los diferentes escenarios. Además, los estos se cancelan en una proporción similar, según el cuadro 2.

1. 5. 1. 1. Modelo de préstamos prendarios y de consumo

En el gráfico 1 se presenta la relación entre la tasa de cancelación de préstamos y el tiempo transcurrido de ellos. Puede apreciarse una curva ligeramente creciente, lo que sugiere que la probabilidad de cancelación de los préstamos se incrementa conforme aumenta su tiempo de actividad. Esta afirmación es correcta, pues la naturaleza de ellos supone un plazo establecido para cada uno. Sin embargo, al estimar un parámetro de regresión lineal para la curva se aprecia que el crecimiento en la tasa de cancelación de préstamos es 0,4 puntos porcentuales por cada mes adicional en el préstamo, y al final la tasa decae, lo cual se puede explicar por los préstamos que caen en mora. Este razonamiento no ofrece evidencia suficiente para determinar el modelo que mejor se ajusta a los datos, por lo cual se espera que los indicadores de precisión sean los que lo determinen.

[7] En promedio, por mes se cancelan 106 de consumo, 10 hipotecarios y 113 prendarios

Gráfico 1. Relación entre las tasas de cancelación de los préstamos y el tiempo transcurrido Setiembre 2015 - setiembre 2017 (suavizado, préstamos Consumo y Préndarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

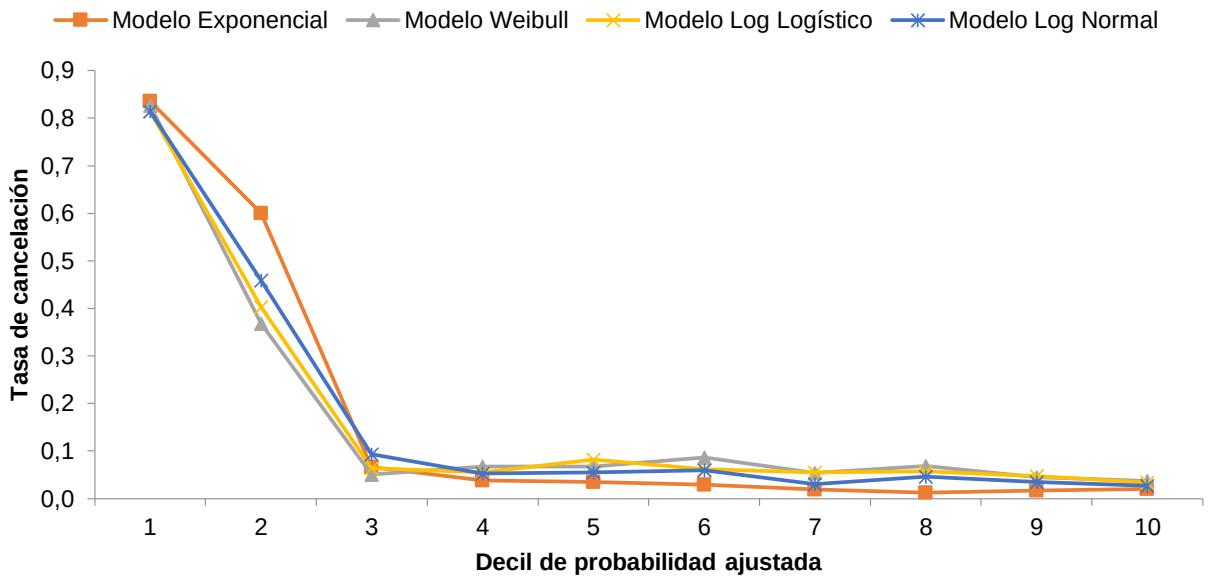
El análisis de supervivencia se realizó con 31 variables de respuesta, para las que se ejecutaron diferentes modelos con el fin de seleccionar las que puedan demostrar estar más relacionadas con el fenómeno de cancelación de préstamos, y fueron seleccionadas 17 (el detalle de la selección de cada variable se incluye en el anexo 2).

Un ejemplo de la lógica de selección se puede seguir con la variable plazo del préstamo, la que se incluyó en el método de selección automática en tres de los modelos estimados, y tiene una asociación importante[8] con la cancelación del préstamo y su antigüedad, por lo cual se selecciona para ser utilizada en el modelo. De forma recíproca ocurre con la cantidad de cuentas activas que tiene el cliente, variable que de acuerdo con el criterio utilizado no muestra asociación con su antigüedad. Además, no fue seleccionada por ningún modelo con el método de selección automática.

El desempeño de los modelos muestra que las cuatro opciones presentan una linealidad lógica, considerando que los grupos de menor probabilidad de supervivencia estimada en los próximos tres meses son los que poseen una tasa de cancelación mayor, la cual disminuye conforme aumenta la probabilidad de supervivencia estimada (gráfico 2).

[8] Para efectos de este estudio un coeficiente de asociación de 0,10 o más se consideró como una relación de importancia entre las dos características bajo análisis, dado que se está trabajando con toda la población

Gráfico 2. Tasa de cancelación por decil de probabilidad ajustada de los modelos considerados Setiembre 2015 - setiembre 2017 (préstamos Consumo y Prendario)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Por otro lado, después de obtener un punto de corte óptimo de la probabilidad ajustada por cada modelo, se determinó que los mejores indicadores de precisión son producto del modelo exponencial, considerando un porcentaje de clasificación correcta de 89,4% y utilizando un punto de corte óptimo de 0,909, lo cual indica que se consideran como préstamos con riesgo alto de cancelación aquellos con una probabilidad de supervivencia estimada inferior a dicho valor (cuadro 4).

Cuadro 4. Indicadores de precisión por modelo considerado Setiembre 2015 - setiembre 2017 (préstamos Prendario y Consumo)

Indicador de precisión	Modelo			
	Exponencial	Weibull	Log Logístico	Log Normal
Corte óptimo	0,909	0,640	0,892	0,893
Curva ROC	0,940	0,856	0,864	0,830
Porcentaje de clasificación correcta	89,40	76,00	78,60	83,00
Porcentaje de ocurrencias acertadas	89,40	78,30	78,60	83,00
Porcentaje de no ocurrencias acertadas	89,40	75,50	78,60	83,00

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

La estimación del modelo exponencial se muestra en el cuadro 5. Los coeficientes reflejan el patrón identificado en los datos para predecir el riesgo de cancelación anticipada de préstamos prendarios y de consumo.

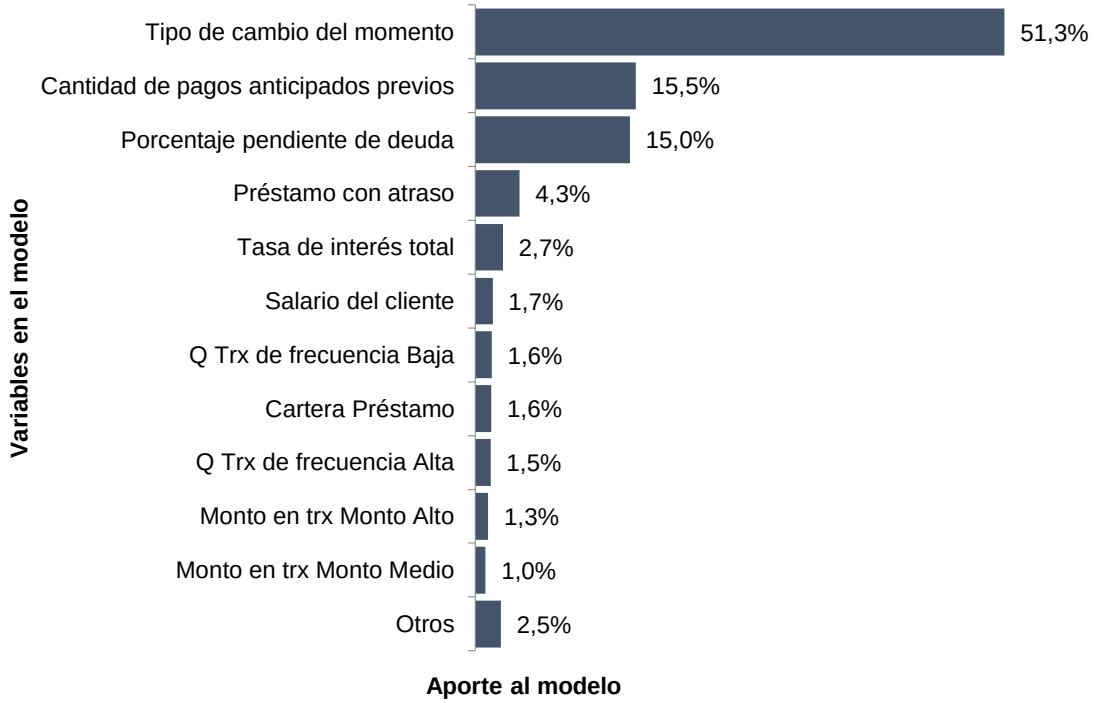
Cuadro 5. Coeficientes ajustados Modelo exponencial
Setiembre 2015 - setiembre 2017
(préstamos Prendario y Consumo)

Nombre variable	Coeficiente	Error estándar
Intercepto	-35,699557	0,380136
Cartera préstamo	-0,257267	0,071055
Edad del cliente	0,001796	0,001073
Moneda del préstamo	0,029826	0,074093
Préstamo con atraso	1,034477	0,104187
Plazo	-0,000096	0,000594
Tipo de cambio del momento	0,07706	0,000655
Tasa de interés total	-0,048717	0,007873
Porcentaje pendiente de deuda	-0,020442	0,000595
Cantidad de pagos anticipados previos	-0,122889	0,003445
Cantidad Trx de frecuencia alta	0,002332	0,000669
Cantidad Trx de frecuencia baja	0,030356	0,008147
Monto en Trx monto alto	-0,000009	0,000003
Monto en Trx monto medio	-0,000011	0,000005
Saldo en cuentas	-0,000002	0,000001
Total pagado en préstamos por el cliente	0,000017	0,000023
Salario del cliente	-0,000026	0,000007
Razón créditos a débitos cuenta	-0,013119	0,011855

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Entre las variables que más impactan el ajuste de la probabilidad estimada está el tipo de cambio del momento, lo cual tiene lógica si se considera el impacto que este puede tener en la cuota de los clientes. También inciden en eso el porcentaje de la deuda que el cliente tiene pendiente y si este ha efectuado pagos anticipados previos (gráfico 3).

Gráfico 3. Importancia de las variables en el modelo
Setiembre 2015 - setiembre 2017
(Préstamos Consumo y Prendario)



9

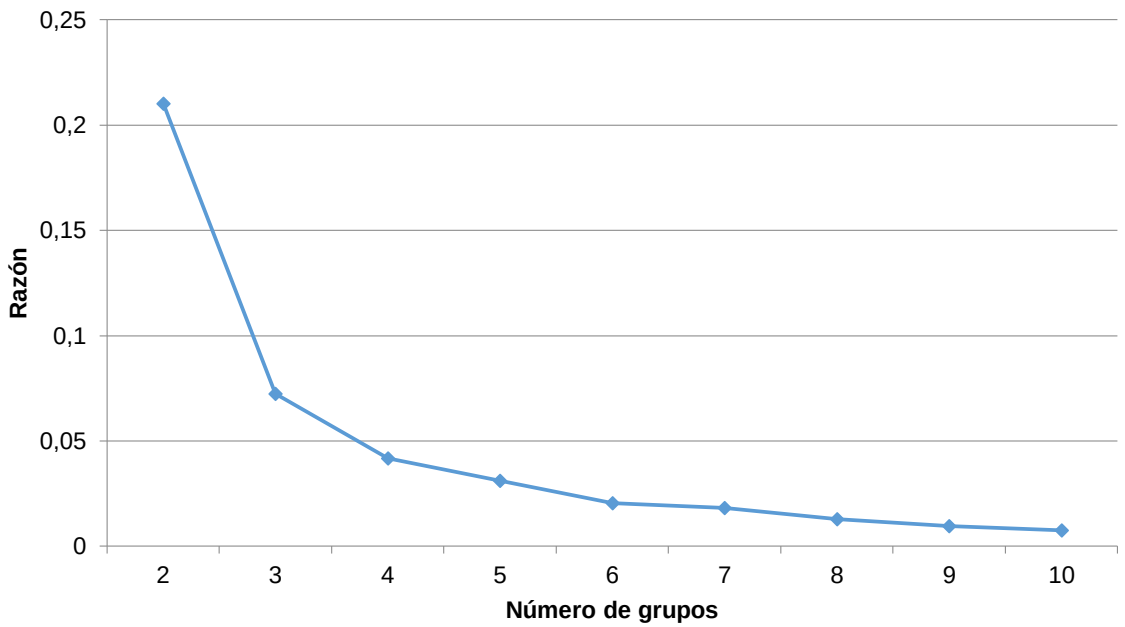
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

1. 5. 1. 2. Segmentación para préstamos de consumo y prendario

En el gráfico 4 se presenta una segmentación sugerida, la cual se fundamenta en que a partir de cinco grupos la diferencia entre la razón de la suma de cuadrados dentro de los grupos y la suma de cuadrados entre los grupos decrece levemente, por lo que la mejora en la segmentación no aumenta considerablemente. Desde un punto de vista numérico, se decidió usar cinco grupos.

[9] Como se explica en la sección 4.7.

Gráfico 4. Razón Suma de cuadrados dentro de los grupos y suma de cuadrados entre grupos según número de grupos
Setiembre 2015 - setiembre 2017
(Préstamos Prendario e Hipotecario)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Al ajustar la probabilidad para un mes determinado y asignar los grupos para cada registro se observa una segmentación que sugiere enfocarse en aproximadamente 86 operaciones, si se consideran aquellos préstamos con riesgo de cancelación medio, alto y muy alto. Esta cantidad puede ser considerable según la cantidad de personas que se dispongan para atenderlos y ofrecer una estrategia que permita retener a los clientes con sus respectivos préstamos. Por otro lado, se observa que el saldo promedio de cada uno de los cuatro préstamos con mayor riesgo de cancelación impacta de mayor forma la cartera total de préstamos que las 18 operaciones con riesgo alto, que a su vez son mayores que los 64 préstamos con riesgo medio (cuadro 6).

Cuadro 6. Agrupación propuesta para el riesgo de cancelación anticipada de préstamos Prendario y de consumo

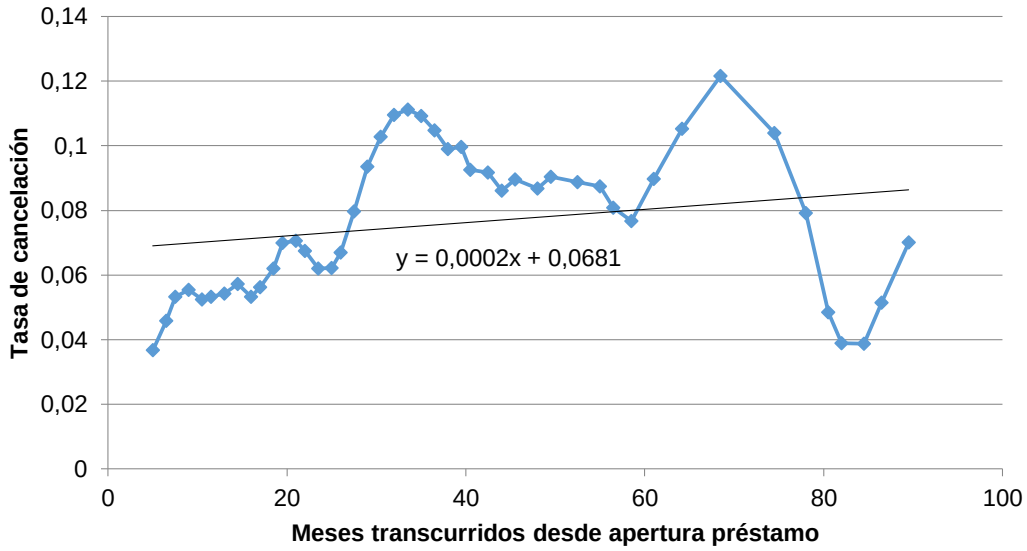
Riesgo de cancelación	Probabilidad ajustada		Cantidad de préstamos	Peso relativo (%)	*Peso cartera total (%)
	Mínimo	Máximo			
Muy Alto	0,00	0,35	4	0,03	0,0067
Alto	0,35	0,60	18	0,02	0,0058
Medio	0,60	0,80	64	0,30	0,0034
Bajo	0,80	0,95	6 899	5,60	0,0048
Muy bajo	0,95	1,00	8 748	94,05	0,0076
Total			**15 733**	**100,00**	**-**

*Se calcula como el saldo promedio de cada préstamo dividido entre el saldo total de la cartera estudiada
Fuente: Almacén de datos de la entidad financiera. Setiembre 2017.

1. 5. 1. 3. Modelo para préstamos hipotecarios

En el gráfico 5 se muestra una relación entre la tasa de cancelación de préstamos hipotecarios y el tiempo transcurrido, con una tendencia ligeramente creciente, lo que indica que al pasar el tiempo de vigencia del préstamo la tasa de cancelación crece levemente. Este comportamiento puede ser explicado por el tipo de préstamo, ya que las hipotecas tienden a ser operaciones de largo plazo, con un monto de apertura alto, lo que dificulta que los clientes tengan la capacidad de cancelarlo anticipadamente en los primeros meses. No obstante, es de esperar que al aproximarse la fecha final del préstamo se pueda cancelar, pues el saldo pendiente es menor. Esto ocurre también con los préstamos prendarios y de consumo, en los que existen casos con mucha antigüedad, y la cancelación disminuye debido a los préstamos en mora. El análisis de este gráfico no ofrece suficiente evidencia para determinar el modelo que mejor se adapta a los datos, por lo cual se consideran los indicadores de precisión para determinarlo.

Gráfico 5. Relación entre la tasa de cancelación de los préstamos y el tiempo transcurrido de los mismos
Setiembre 2015 - setiembre 2017
(suavizado, préstamos Hipotecarios)

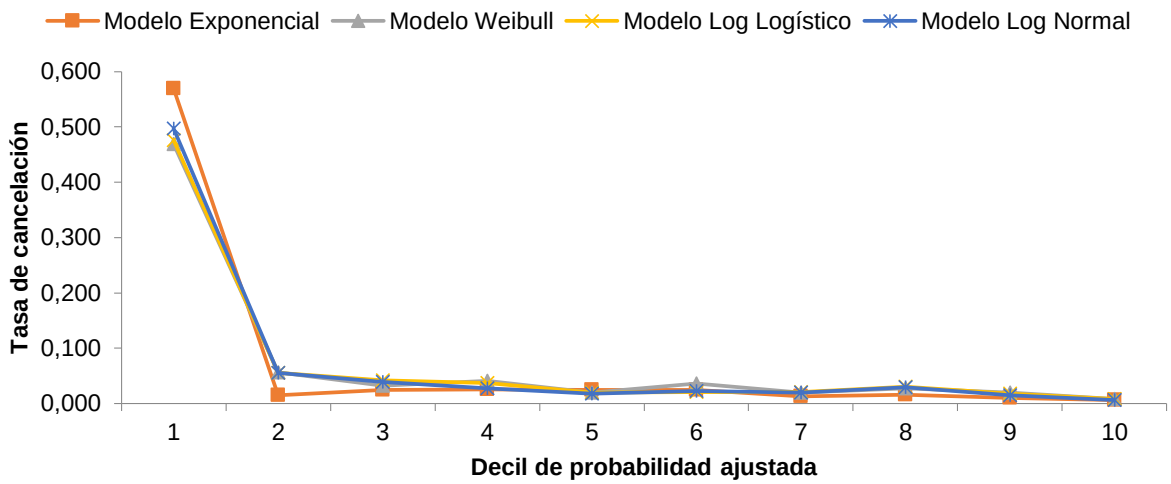


Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Para los préstamos hipotecarios se obtuvieron resultados similares a los del caso de los préstamos prendarios y de consumo, considerando el mejor modelo ajustado. Pese a eso, cambia levemente en relación con las variables seleccionadas para predecir el fenómeno. En este caso, de 32 variables se seleccionan 12, siguiendo la misma metodología descrita anteriormente y cuyo resumen se incluye en el anexo 3

El desempeño de los diferentes modelos muestra una linealidad de predicción adecuada para los cuatro y el modelo exponencial es el grupo que presenta la menor probabilidad de supervivencia (mayor de cancelación), y existe una mayor tasa de cancelación (gráfico 6).

Gráfico 6. Tasa de cancelación por decil de probabilidad ajustada de los modelos consideradosSetiembre 2015 - setiembre 2017 (préstamos Hipotecarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Al considerar los indicadores de predicción, cuyos resultados se resumen en el cuadro 7, se determina que el modelo exponencial es el que se clasifica mejor, a partir de un corte de 0,957, con un porcentaje de clasificación correcto cercano a 81,7%; mientras que el siguiente modelo con mejor porcentaje de clasificación correcto es el Log normal, de 79,5%.

Cuadro 7. Indicadores de precisión por modelo considerado Setiembre 2015 - setiembre 2017 (préstamos hipotecarios)

Indicador de precisión	Modelo			
	Exponencial	Weibull	Log Logístico	Log Normal
Corte óptimo	0,957	0,924	0,960	0,958
Curva ROC	0,898	0,847	0,855	0,796
Porcentaje de clasificación correcta	81,70	74,00	78,10	79,50
Porcentaje de ocurrencias acertadas	81,80	76,70	76,40	79,60
Porcentaje de no ocurrencias acertadas	81,70	73,80	78,20	79,50

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Como producto del ajuste del modelo exponencial se obtienen los coeficientes estimados, los cuales representan el patrón de comportamiento en los datos, resultado que se visualiza en el cuadro 8.

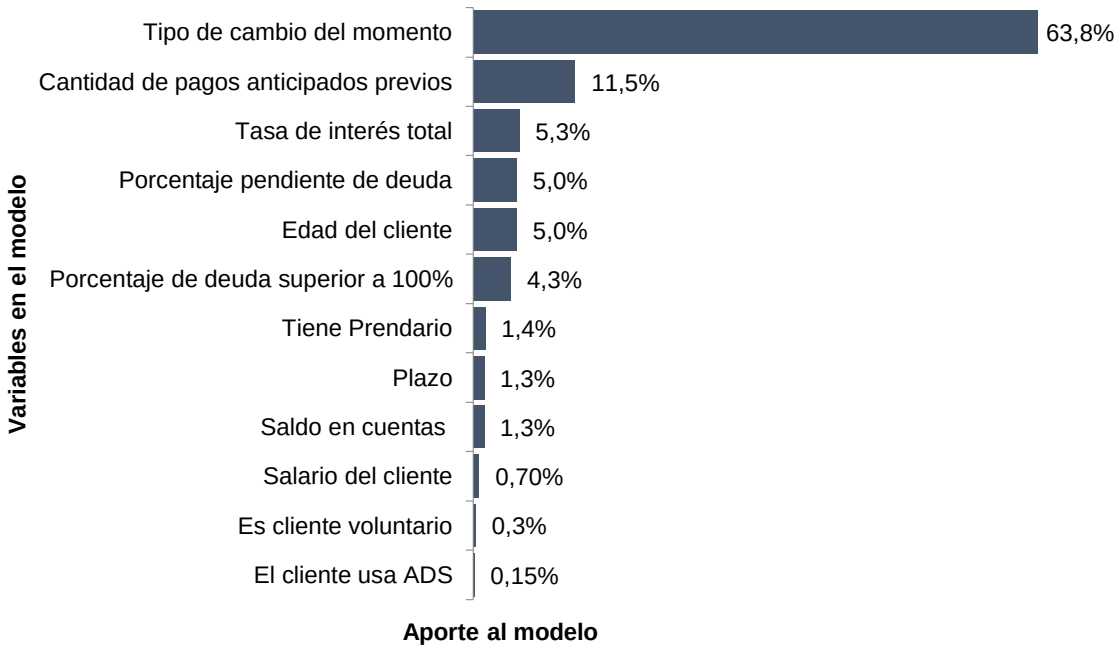
Cuadro 8. Coeficientes ajustados. Modelo exponencial
Setiembre 2015 - setiembre 2017
(préstamos hipotecarios)

Nombre variable	Coeficiente	Error estándar
Intercepto	-51,995362	0,637271
Edad del cliente	0,019278	0,002578
Plazo	-0,00082	0,000423
Tipo de cambio del momento	0,106901	0,001115
Tasa de interés total	-0,110581	0,013986
Porcentaje pendiente de deuda	-0,009207	0,001231
Porcentaje de deuda superior a 100%	0,403126	0,062206
Cantidad de pagos anticipados previos	-0,098198	0,00566
Tiene prendario	0,156058	0,073828
Saldo en cuentas	-0,000003	0,000001
Salario del cliente	-0,000022	0,000021
Es cliente voluntario	0,043716	0,088126
El cliente usa ADS	0,012778	0,057237

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Al analizar la importancia de las variables en el modelo ajustado se observa que la principal es el tipo de cambio del momento, y en segundo lugar está la cantidad de pagos anticipados que se efectuaron previamente al préstamo. Esto adquiere sentido al considerar que son operaciones que los clientes van anticipando de forma gradual debido a la alta suma del préstamo, por lo que es más difícil hacer una cancelación anticipada desde el principio, lo cual a la vez suma atractivo al modelo de predicción de pagos anticipados (gráfico 7).

Gráfico 7. Importancia de las variables en el modelo
Setiembre 2015 - setiembre 2017
(préstamos Hipotecarios)



10

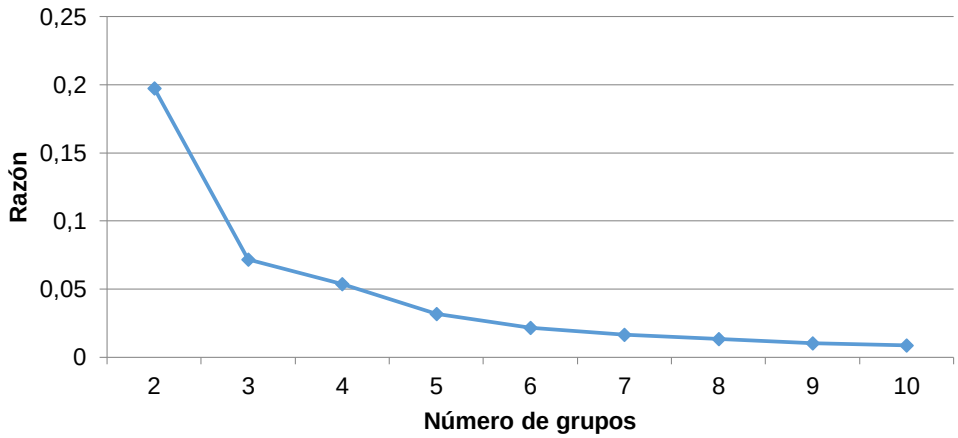
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

[10] Como se explica en la sección 4.7.

1. 5. 1. 4. Segmentación de préstamos hipotecarios

En el gráfico 8 se presenta el caso de los préstamos hipotecarios, considerando numéricamente adecuado utilizar tres o cinco grupos, pues en estos casos el cambio de la razón entre la suma de cuadrados dentro de los grupos y la suma de cuadrados entre los grupos no cambia en gran medida. Para proceder como se hizo en el caso de los préstamos prendarios y de consumo, además para disponer de mayor cantidad de agrupaciones para las cuales tomar medidas segmentadas, se determinó utilizar cinco grupos.

Gráfico 8. Razón Suma de cuadrados dentro de los grupos y suma de cuadrados entre grupos Setiembre 2015 - setiembre 2017 (préstamos Hipotecarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

La distribución establecida, según los cortes de probabilidad ajustada propuestos, da como resultado cuatro operaciones que deben atenderse de manera urgente, en que cada una representa 0,017% de la cartera total, y tienen una probabilidad de supervivencia inferior o igual a 0,35. Existen tres que surgen con un riesgo alto y 39 a los que se les debe dar cierto seguimiento o aplicar determinada estrategia, pues tienen un riesgo medio de ser cancelados anticipadamente. Estos últimos, en promedio, tienen menor saldo; sin embargo, se pueden tomar medidas menos costosas con ellos (cuadro 9).

Cuadro 9. Agrupación propuesta para el riesgo de cancelación anticipada de préstamos Hipotecarios Casos setiembre 2017

Riesgo de cancelación	Probabilidad ajustada		Cantidad de préstamos	Peso relativo (%)	*Peso cartera total (%)
	Mínimo	Máximo			
Muy alto	0,00	0,35	4	0,03	0,0174
Alto	0,35	0,65	3	0,02	0,0041
Medio	0,65	0,85	39	0,30	0,0037
Bajo	0,85	0,95	738	5,60	0,0055
Muy bajo	0,95	1,00	12 384	94,05	0,0077
Total			**13 168**	**100,00**	**-**

*Se calcula como el saldo promedio de cada préstamo dividido entre el saldo total de la cartera estudiada
Fuente: Almacén de datos de la entidad financiera. Setiembre 2017.

1. 5. 2. Modelos de pago anticipado de préstamos

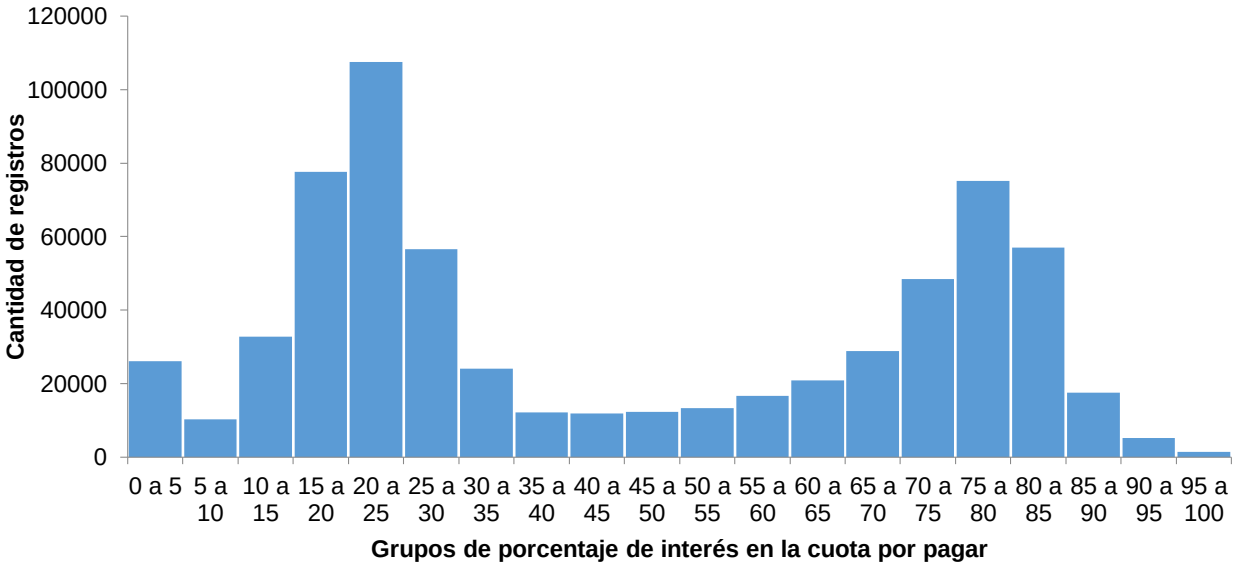
Para este análisis se trabaja con archivos de datos de mayor tamaño y cada registro es un mes de historia, pues se desea conocer el comportamiento según determinado mes. En total se trabajó con 353.801 registros, que corresponden a préstamos prendarios o de consumo y 301.540 son préstamos hipotecarios.

Se debe considerar que para el análisis de pagos anticipados la tasa de ocurrencia del evento disminuye, al pensar que no en todos los créditos se efectuarán pagos anticipados durante todos los meses, sino que solo a una porción de ellos se les hace. La tasa promedio es cercana a 3% en el caso de los préstamos prendarios y de consumo y de 2,4% en el de los préstamos hipotecarios.

Una variable recibe un tratamiento especial debido a que al analizarla muestra un comportamiento diferente para dos poblaciones: el porcentaje de interés representativo en la cuota por pagar, calculado a partir de: $100 \times \frac{Intereses\ Préstamo}{Cuota\ por\ pagar}$

En el gráfico 9 se muestra el histograma de la variable en cuestión y se observa claramente la presencia de una distribución bimodal; la primera que va desde 0% hasta 30%, aproximadamente y la segunda de 30% hasta 100%.

Gráfico 9. Histograma de porcentaje representativo de interés en la cuota por pagar
Setiembre 2015 - setiembre 2017

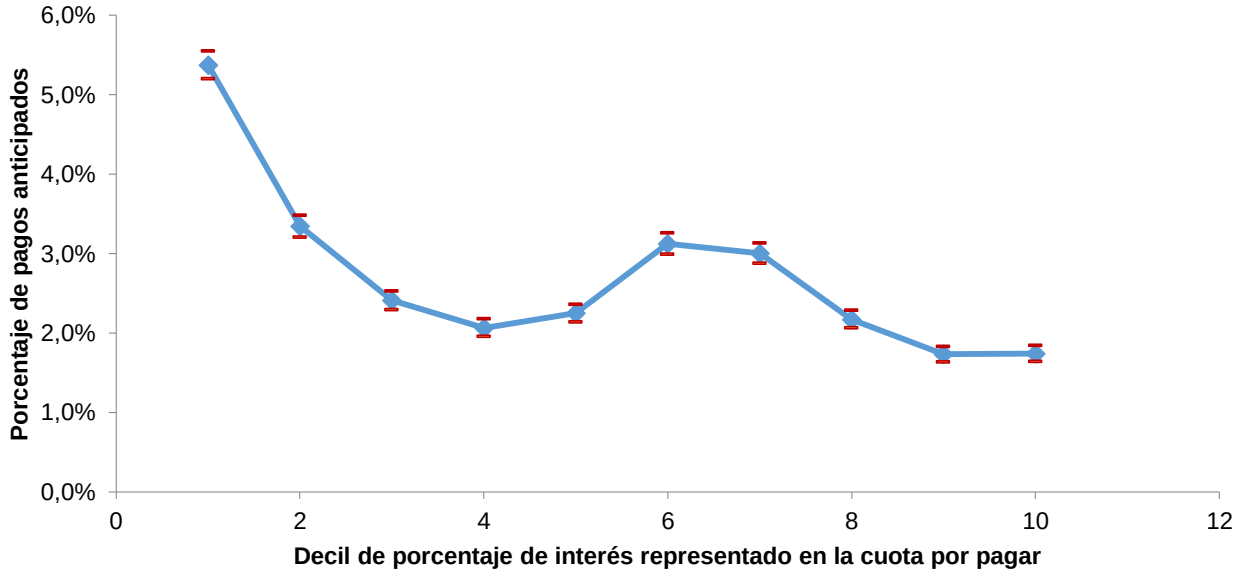


Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Otra forma de representarlo sería comparando el comportamiento específico de la variable agrupada con la variable por predecir, ajustada a partir del porcentaje de ocurrencia. Así, pueden observarse dos rectas: la primera considera los deciles del 1 al 5

mientras que la segunda considera los restantes. Debido a esta situación se consideró agregar como predictores separados aquellos casos cuyo porcentaje de interés representado en la cuota por pagar sea inferior a 30%, y en otra variable diferente, aquellos en los que sea igual o superior a este porcentaje (gráfico 10).

Gráfico 10. Porcentaje representativo de intereses en la cuota por pagar
Setiembre 2015 - setiembre 2017



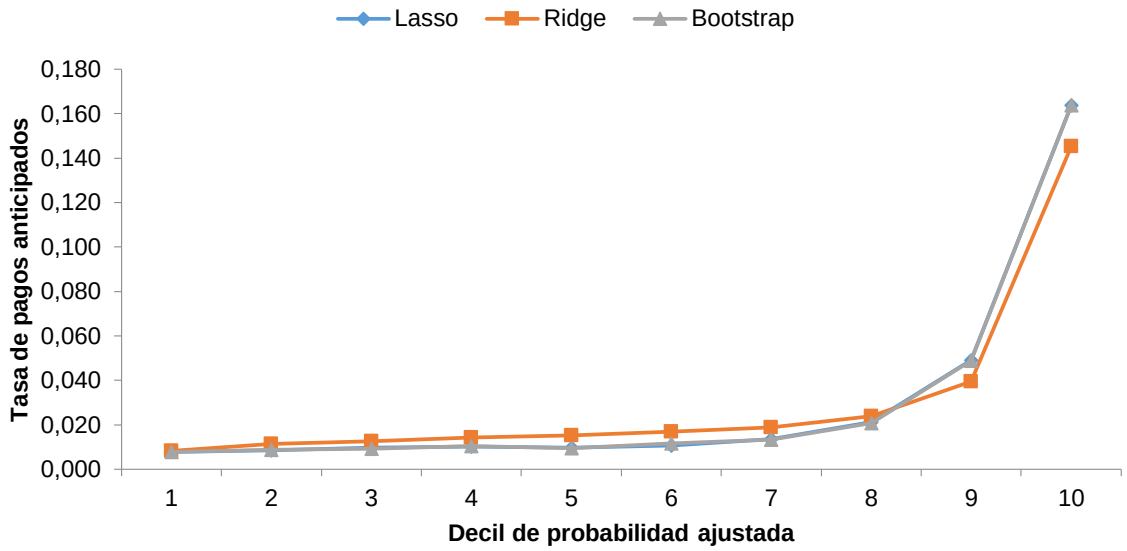
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.
* Las líneas rojas representan los intervalos de confianza estimados a partir de un 5% de significancia

1. 5. 2. 1. Modelo para préstamos prendario y de consumo

En el anexo 4 se muestran las variables que integran el modelo. Se utiliza un método similar al usado en los modelos de cancelación, con la diferencia de que el proceso automatizado por pasos se ejecuta únicamente para un modelo de regresión logística múltiple tradicional, para comparar las tres técnicas propuestas. Se parte de 37 variables y en el modelo final quedaron 26.

Al comprar los modelos ajustados se observa que en los tres casos se tiene una tendencia lineal adecuada, en la que el decil de préstamos con mayor probabilidad ajustada de sufrir un pago anticipado es el que presenta la mayor tasa de pagos. Se debe recurrir a otros métodos para determinar el mejor modelo (gráfico 11).

Gráfico 11. Tasa de pagos anticipados por decil de probabilidad ajustada de los modelos considerados
Setiembre 2015 - setiembre 2017
(préstamos Prendario y Consumo)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

En el cuadro 10 se resumen los principales indicadores de precisión, en los cuales se observa que la regresión Lasso es la que tiene mejor Curva ROC, así como el porcentaje de clasificación correcto. No obstante, la diferencia es casi nula si se compara con el modelo estimado mediante la técnica Bootstrap. Para concluir de forma adecuada, se procede a realizar una validación cruzada.

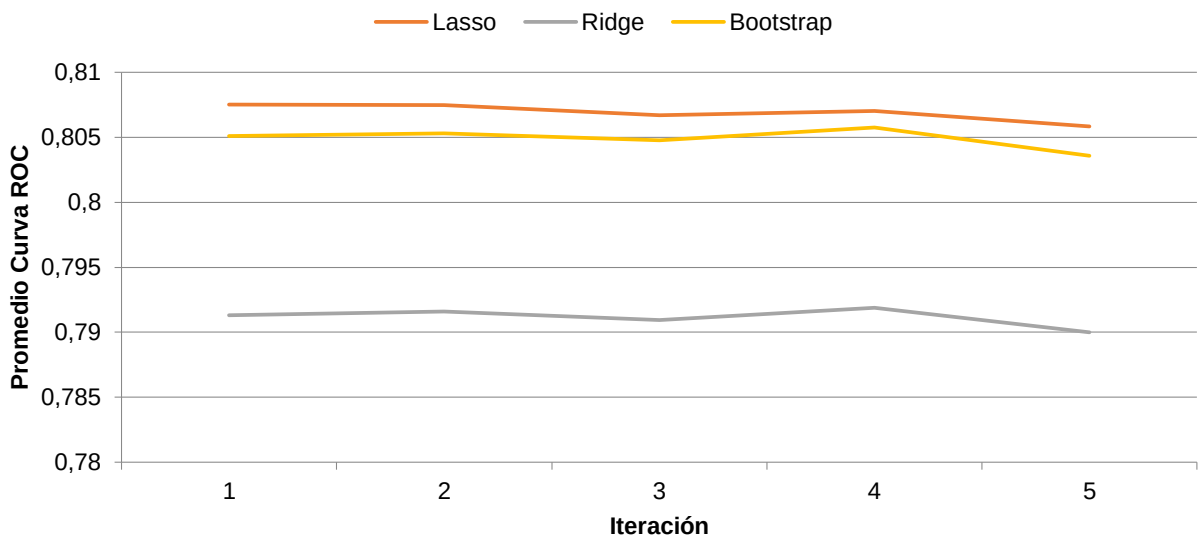
Cuadro 10. Indicadores de precisión por modelo considerado
Setiembre 2015 - setiembre 2017
(préstamos prendarios y consumo)

Indicador de precisión	Modelo		
	Lasso	Ridge	Bootstrap
Corte óptimo	0,025	0,029	0,025
Curva ROC	0,806	0,789	0,805
Desviación estándar de la curva ROC	0,016	0,015	0,016
Porcentaje de clasificación correcta	75,00	74,70	75,00
Porcentaje de ocurrencias acertadas	75,00	72,10	75,00
Porcentaje de no ocurrencias acertadas	75,00	74,80	75,00

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Al ejecutar la validación cruzada se observa que el modelo Lasso es el que ofrece la mejor precisión, aunque la diferencia con el modelo Bootstrap sigue siendo mínima, pues en promedio la diferencia es de apenas 0,2%. No obstante, al considerar que el Bootstrap es más sencillo en términos de estimación de parámetros y que las diferencias con Lasso son mínimas, se puede determinar que este es el modelo más conveniente para aproximar la probabilidad de que exista un pago anticipado en un préstamo para un mes dado, en el caso de los préstamos prendario y de consumo (gráfico 12).

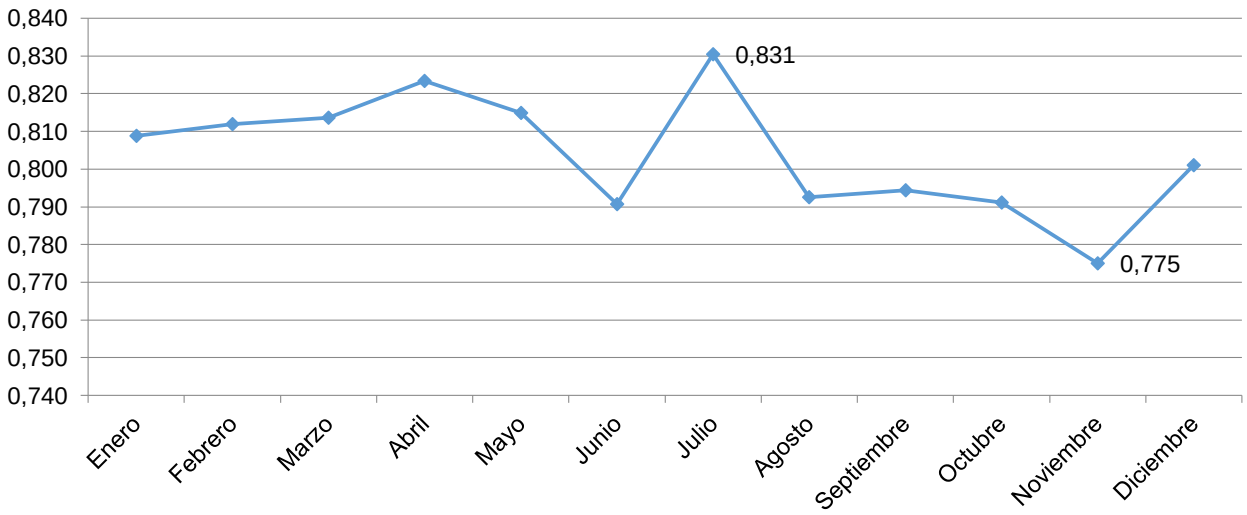
Gráfico 12. Curva ROC en validación cruzada
Setiembre 2015 - setiembre 2017
(préstamos prendario y consumo)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

El gráfico 13 demuestra la estabilidad del modelo Bootstrap en lo que respecta a la predicción mensual. Al disponer de 25 meses de historia se procede a calcular una curva ROC promedio por mes, con lo que se observa la estabilidad de predicción y se identifican los meses de mejor y peor predicción. El mes que mejor predice es julio (predice agosto) y el mes en el que la precisión de la predicción es menor es noviembre (predice diciembre). Por otro lado, se observa que en el peor de los casos se obtiene una curva ROC de 0,831 y en el mejor escenario de 0,775.

Gráfico 13. Curva ROC Promedio Mensual
Setiembre 2015 - setiembre 2017
(Modelo bootstrap, préstamos Prendario y Consumo)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Los coeficientes del modelo fueron estimados usando la técnica Bootstrap, con lo que se obtiene una ecuación que permite obtener la probabilidad de que se haga un pago anticipado a determinada operación en el mes siguiente (cuadro 11).

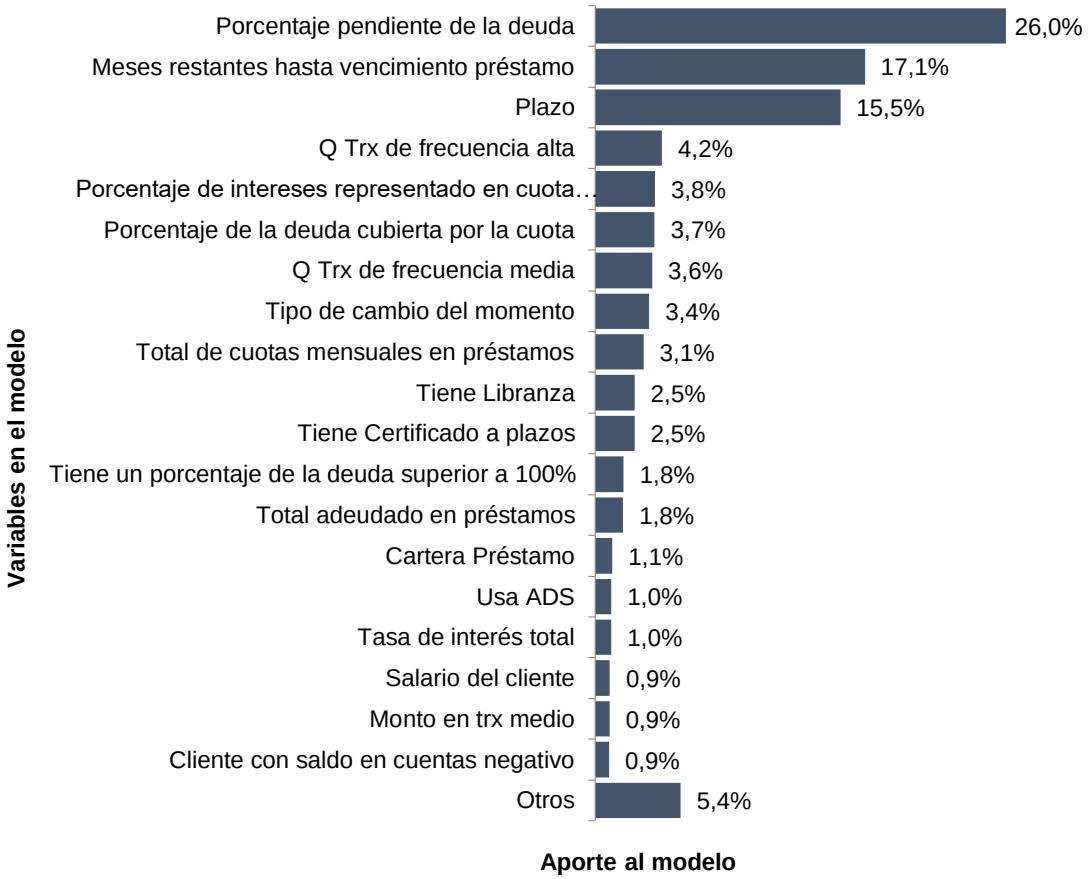
Cuadro 11. Coeficientes ajustados. Modelo exponencial
Setiembre 2015 – setiembre 2017
(préstamos Prendarios y Consumo)

Nombre variable	Coeficiente	Error estándar
(Intercept)	7,112766	0,335881
Cartera préstamo	0,26825	0,056541
Plazo	-0,058369	0,000854
Meses restantes hasta el vencimiento del préstamo	0,065939	0,000877
Tipo de cambio del momento	-0,008611	0,000575
Tasa de interés total	-0,026786	0,006053
Porcentaje pendiente de la deuda	-0,069363	0,000605
Tiene un porcentaje de la deuda superior a 100%	1.554585	0.197051
Porcentaje de la deuda cubierta por la cuota	-0.044669	0.002706
Porcentaje de intereses representado en la deuda superior 30%	0.013852	0.000835
Antigüedad	-0.00015	0.000124
Edad	0.002016	0.000871
Sexo del cliente	0.060089	0.017408
Cantidad total de cuentas	-0.029147	0.008183
Tiene certificado a plazos	0.434675	0.039895
Tiene libranza	-0.794427	0.072399
Tiene hipoteca	-0.081304	0.054397
Q Trx de frecuencia alta	-0,012118	0,000652
Q Trx de frecuencia media	0,019987	0,001265
Q Trx de frecuencia baja	-0,017266	0,00489
Monto en trx medio	0,000014	0,000003
Envía transferencias internacionales	-0,260061	0,122568
Recibe servicios regionales	-0,432284	0,116188
Cliente con saldo en cuentas negativo	-0,710452	0,18851
Total adeudado en préstamos	0,000005	0,000001
Total de cuotas mensuales en préstamos	-0,000674	0,00005
Salario del cliente	0,000021	0,000005
Usa ADS	-0,379385	0,085563
Monto en créditos a la cuenta entre monto en débitos a la cuenta	0,021153	0,009033

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Algunas de las variables más importantes del modelo son el porcentaje de deuda que tiene pendiente, para el cual cuanto mayor sea su valor, la probabilidad de efectuar un pago anticipado tiende a disminuir. También están los meses restantes que le quedan a la operación y el plazo de ella, entre otras (gráfico 14).

Gráfico 14. Importancia de las variables en el modelo
Setiembre 2015 - setiembre 2017
(préstamos Consumo y Prendario)

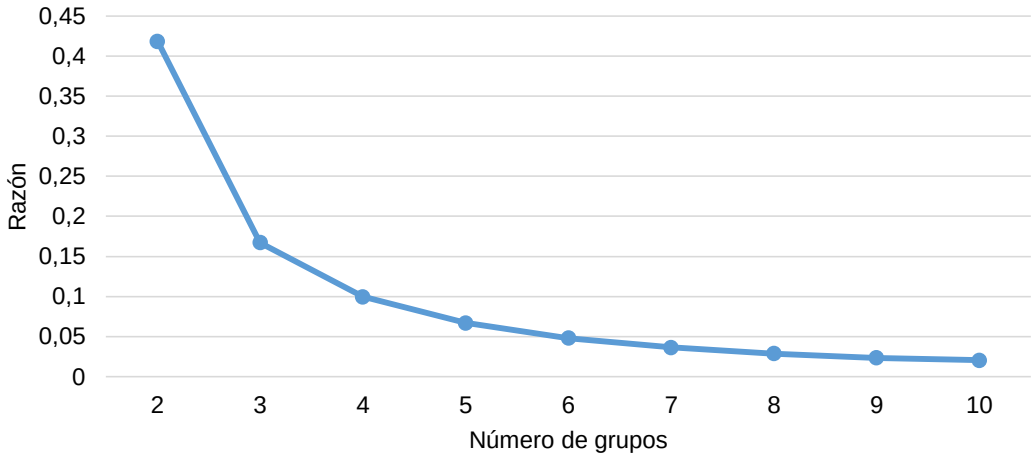


Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

1. 5. 2. 2. Segmentación de préstamos de consumo y prendario

Para determinar la cantidad adecuada de grupos para el modelo de pagos anticipados, en el caso de los préstamos prendarios y de consumo, se construyó el gráfico 15. El número óptimo de grupos en términos numéricos podría determinarse en tres o cinco grupos, y al probar ambas posibilidades se encuentra que con tres grupos se recargan en el que ajusta las probabilidades más bajas, y quedan muy pocos en las otras categorías. Con cinco grupos queda una distribución más cómoda para generar estrategias segmentadas por grupo, en términos de las necesidades del negocio.

Gráfico 15. Razón Suma de cuadrados dentro de los grupos y suma de cuadrados entre grupos Setiembre 2015 - setiembre 2017 (préstamos Prendario y Consumo)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

La distribución por grupos ofrece la posibilidad de enfocarse en 40 préstamos de forma urgente, 98 para manejarlos con prioridad alta y 199 para considerar y dar cierto seguimiento. Esto permite adecuar una estrategia de retención por parte de la entidad bancaria que busque disminuir o controlar la cantidad de pagos anticipados en esta cartera. Sin embargo, se debe considerar que los préstamos con mayor riesgo de pago anticipado son también los préstamos con menor saldo, por lo que se pueden atender estos préstamos con la misma prioridad que la de los de riesgo alto, en que, en promedio, tienen mayor saldo que los demás (cuadro 12).

Cuadro 12. Agrupación propuesta para el riesgo de pago anticipado de préstamos prendario y de consumo Casos setiembre 2017

Riesgo de cancelación	Probabilidad ajustada		Cantidad de préstamos	Peso relativo (%)	*Peso cartera total (%)
	Mínimo	Máximo			
Muy bajo	0,00	0,05	14 815	94,14	0,0065
Bajo	0,05	0,15	581	3,81	0,0034
Medio	0,15	0,30	199	1,22	0,0033
Alto	0,30	0,50	98	0,59	0,0032
Muy alto	0,50	1,00	40	0,25	0,0023
Total			**15 733**	**100,00**	**-**

*Se calcula como el saldo promedio de cada préstamo dividido entre el saldo total de la cartera estudiada

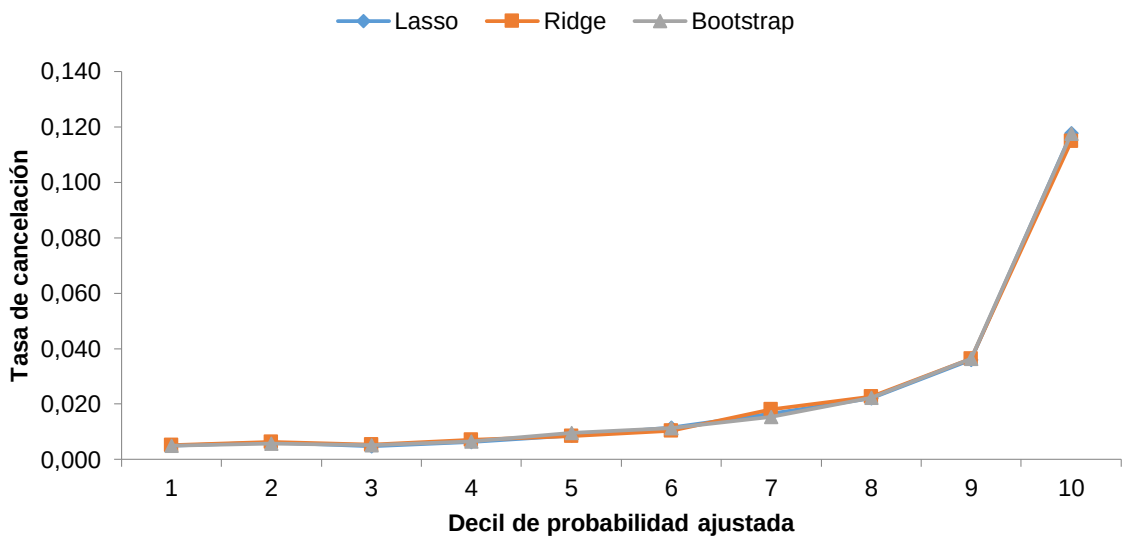
Fuente: Almacén de datos de la entidad financiera. Setiembre 2017.

1. 5. 2. 3. Modelo para préstamos hipotecarios

Se procede con la metodología propuesta para la selección de variables, empezando con 37 y terminando con 26 para ajustar los modelos finales (anexo 5).

En el gráfico 16 se puede apreciar que los tres modelos propuestos presentan una relación adecuada, considerando que los deciles de préstamos con mayor probabilidad ajustada de pago anticipado son los que, además, tienen una mayor tasa de préstamos en la que sucede el fenómeno. Por eso para determinar el mejor modelo se procede al análisis de indicadores de precisión.

Gráfico 16. Tasa de cancelación por decil de probabilidad ajustada de los modelos considerados Setiembre 2015 - setiembre2017 (préstamos Hipotecarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Similar al caso de los préstamos prendario y de consumo, los indicadores de precisión favorecen al modelo con regresión Lasso, aunque las diferencias son mínimas si se comparan con el modelo Bootstrap. El modelo de regresión Ridge parece ser el que obtiene los resultados en indicadores de precisión más bajos. Se procede a realizar una validación cruzada (cuadro 13).

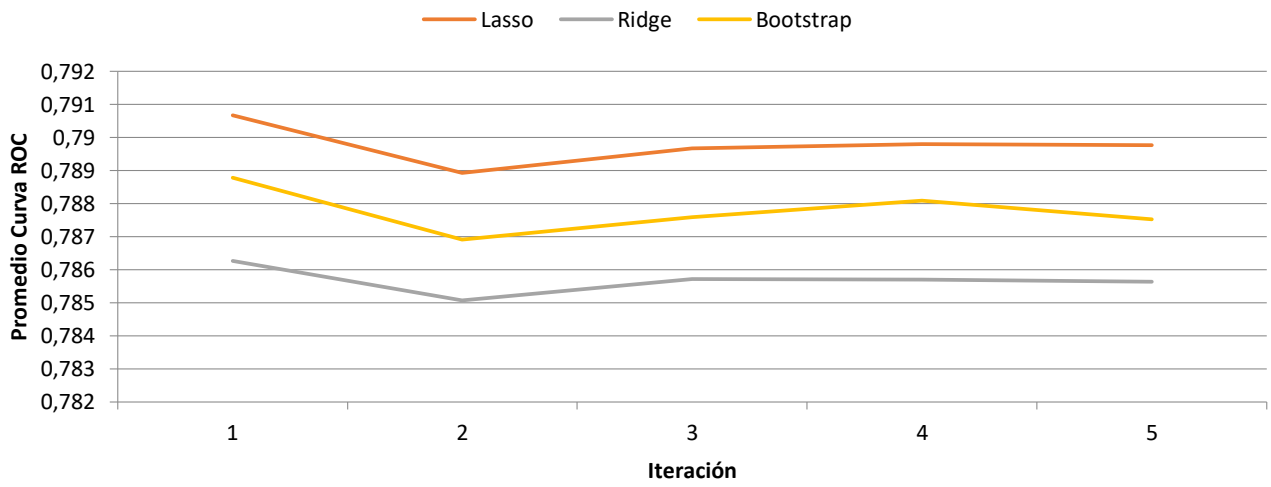
Cuadro 13. Indicadores de precisión por modelo considerado
Setiembre 2015 - setiembre 2017
préstamos hipotecarios

Indicador de precisión	Modelo		
	Lasso	Ridge	Bootstrap
Corte óptimo	0,021	0,021	0,021
Curva ROC	0,801	0,800	0,801
Desviación estándar de la curva ROC	0,009	0,010	0,009
Porcentaje de clasificación correcta	73,30	72,70	73,20
Porcentaje de ocurrencias acertadas	73,50	72,70	73,20
Porcentaje de no ocurrencias acertadas	73,30	72,70	73,20

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

El gráfico 17 muestra que no existen diferencias apreciables entre los tres modelos, por lo cual la conclusión debe ser dirigida principalmente a la conveniencia de cada uno. Se sabe, por un lado, que los modelos Ridge y Lasso se adecuan al considerarse que sus parámetros tienen un sesgo mínimo, por tratarse de modelos de penalización. Por otro lado, el modelo Lasso anula aquellos parámetros que sugieren no ser importantes, lo que lo vuelve más simple. Pese a esto, la estimación con Bootstrap parece ser un poco más estable que la de los demás, y se trata de una ecuación sencilla para la cual no se deben estimar parámetros adicionales. Por ello parece más conveniente, considerando también que se ejecuta de forma rápida, lo cual permite que su actualización posterior sea más sencilla.

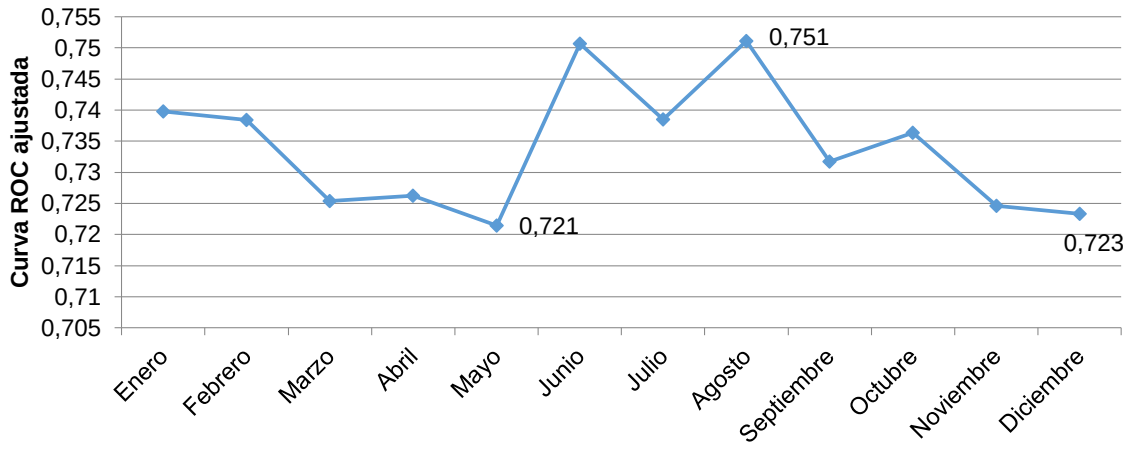
Gráfico 17. Curva ROC en validación cruzada
Setiembre 2015 - setiembre 2017
(préstamos hipotecarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

La estabilidad mensual del modelo Bootstrap parece ser adecuada. El mes de agosto (predice setiembre) es el que presenta la precisión mayor, aunque en diciembre (predice enero del siguiente año) y mayo (predice junio) disminuye. Esto es un punto importante de tomar en cuenta en el momento de diseñar una estrategia de prevención (gráfico 18).

Gráfico 18. Curva ROC Promedio Mensual
Setiembre 2015 - setiembre 2017
(Modelo Bootstrap, préstamos Hipotecarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

El modelo ajustado para préstamos hipotecarios, con los coeficientes respectivos, se presenta en el cuadro 14. Con estos valores se programa la función que permite estimar la probabilidad de pago anticipado en el siguiente mes.

Cuadro 14. Coeficientes ajustados. Modelo exponencial
Setiembre 2015 - setiembre 2017
(préstamos hipotecarios)

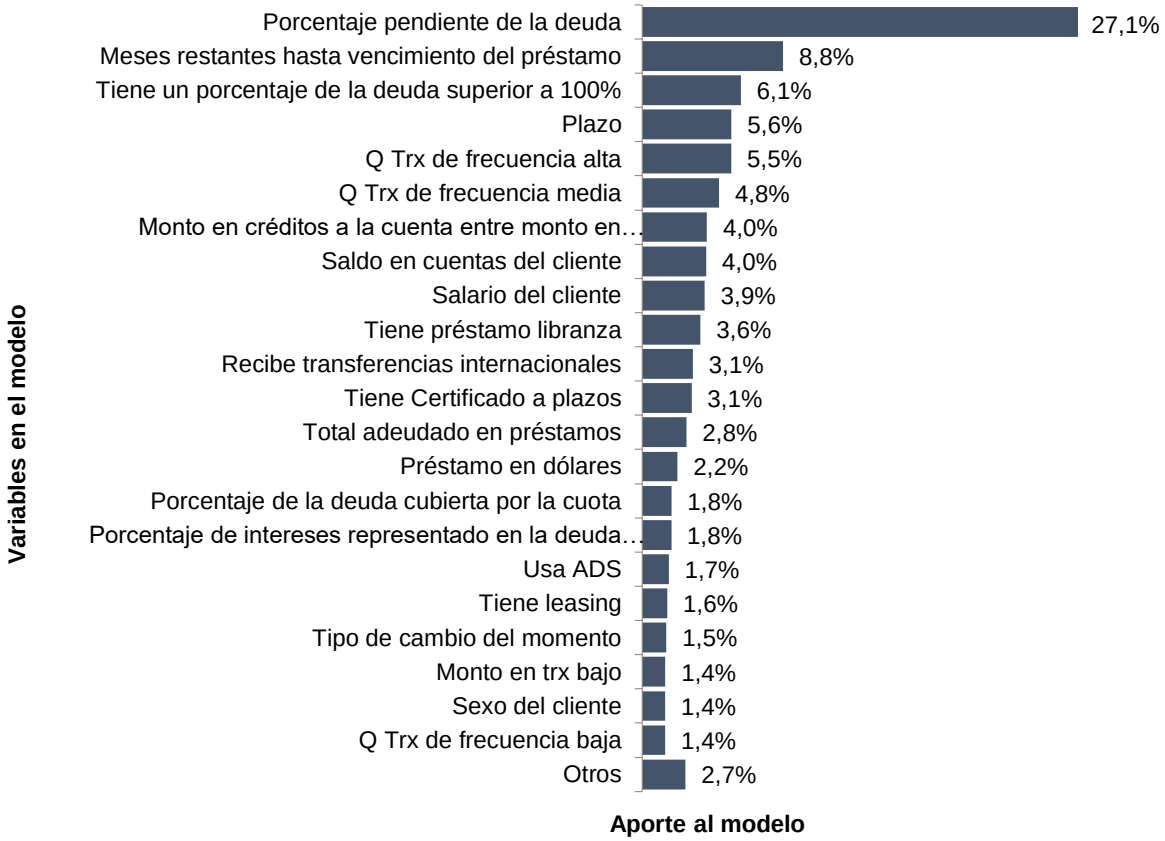
Nombre variable	Coeficiente	Error estándar
(Intercept)	1,82828000	0,37601730
Plazo	-0,00866475	0,00042878
Meses restantes hasta el vencimiento del préstamo	0,01056076	0,00033110
Préstamo en dólares	-0,28381110	0,03538650
Tipo de cambio del momento	-0,00373443	0,00067965
Tasa de interés total	0,00382890	0,00777247
Porcentaje pendiente de la deuda	-0,04293029	0,00043584
Tiene un porcentaje de la deuda superior a 100%	0,59938280	0,02688421
Porcentaje de la deuda cubierta por la cuota	-0,04971810	0,00749876
Porcentaje de intereses representado en la deuda menor 30%	-0,01462699	0,00221935
Edad	0,00142963	0,00104787
Sexo del cliente	0,10152290	0,01945704
Tiene certificado a plazos	0,42197290	0,03776950
Tiene préstamo de consumo	-0,04684980	0,03300092
Tiene préstamo libranza	-1,12969100	0,08557922
Tiene *leasing*	-0,34710260	0,06137437
Q Trx de frecuencia alta	-0,01813931	0,00090119
Q Trx de frecuencia media	0,02129293	0,00122508
Q Trx de frecuencia baja	-0,01952277	0,00374828
Monto en trx bajo	0,00006723	0,00001282
Recibe transferencias internacionales	0,49987540	0,04367604
Recibe servicios regionales	0,13564840	0,07782516
Cliente con saldo en cuentas negativo	-1,39074800	0,29009890
Saldo en cuentas del cliente	0,00001034	0,00000071
Total adeudado en préstamos	-0,00000173	0,00000017
Salario del cliente	0,00006068	0,00000429
Usa ADS	-0,82348770	0,13713400
Monto en créditos a la cuenta entre monto en débitos a la cuenta	0,13345030	0,00910182

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

El peso relativo de cada variable en el modelo muestra que el porcentaje pendiente en la deuda es el que más influye en el modelo, seguido de los meses restantes de vigencia de la operación y de aquellos casos en los que se tiene un porcentaje de la deuda adeudado

superior a 100%. Estos últimos se refieren a los préstamos en los que existió un préstamo adicional al inicial, por lo que el saldo adeudado supera el primer monto (gráfico 19).

Gráfico 19. Importancia de las variables en el modelo
Setiembre 2015 - setiembre 2017
(préstamos Hipotecarios)

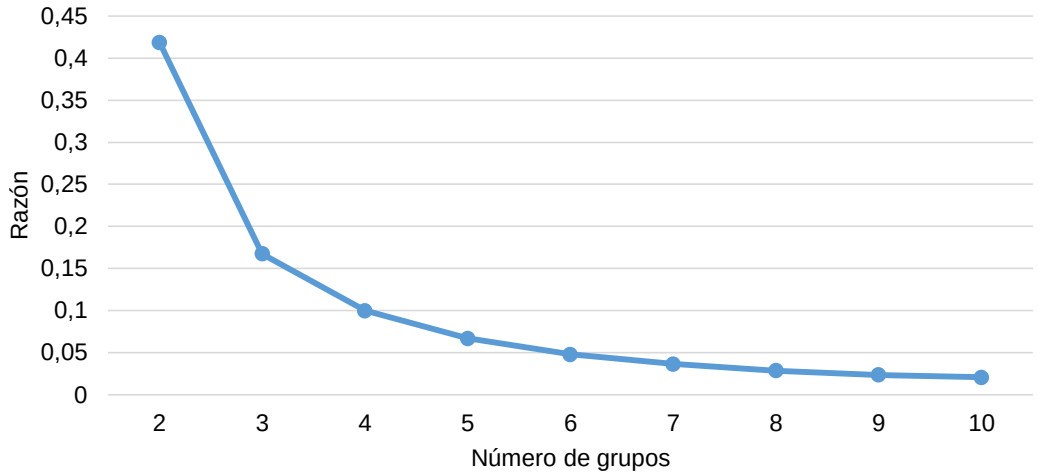


Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

1. 5. 2. 4. Segmentación para préstamos hipotecarios

El gráfico 20 muestra que el número óptimo de grupos, siguiendo el criterio de la razón de sumas de cuadrados, está entre tres y cinco agrupaciones. Se considera que cinco agrupaciones es lo más conveniente para proceder en forma análoga al caso de los modelos anteriores.

Gráfico 20. Razón Suma de cuadrados dentro de los grupos y suma de cuadrados entre grupos Setiembre 2015 - setiembre 2017 (préstamos Hipotecarios)



Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

La distribución en los cinco grupos de riesgo de préstamos, con cierta posibilidad de efectuar pagos anticipados, ofrece 90 casos que se deben atender de forma urgente; además de 304 que tienen un riesgo alto de pago anticipado. Sin embargo, el saldo de los préstamos no es tan alto. Las diferencias con los otros tipos de préstamos se deben a la cantidad de operaciones hipotecarias en la entidad bancaria, las cuales son mucho menores que las otras dos. No obstante, por la cantidad de dinero colocado en esta cartera, las cuatro operaciones con un riesgo muy alto de pago anticipado deben considerarse con cierta prioridad, pues el tipo de cartera pone en riesgo a mayor cantidad de dinero que en los otros tipos de préstamo (cuadro 15).

Cuadro 15. Agrupación propuesta para el riesgo de pago anticipado de préstamos hipotecarios Casos setiembre 2017

Riesgo de cancelación	Probabilidad ajustada		Cantidad de préstamos	Peso relativo (%)	*Peso cartera total (%)
	Mínimo	Máximo			
Muy bajo	0,000	0,015	7 587	99,01	0,0080
Bajo	0,015	0,050	4 518	0,74	0,0075
Medio	0,050	0,170	669	0,21	0,0058
Alto	0,170	0,600	304	0,01	0,0037
Muy alto	0,600	1,000	90	0,03	0,0077
Total			**13 168**	**100,00**	

*Se calcula como el saldo promedio de cada préstamo dividido entre el saldo total de la cartera estudiada
Fuente: Almacén de datos de la entidad financiera. Setiembre 2017.

1. 5. 3. Resumen de resultados

En el cuadro 15 se presenta un resumen de resultados del que se obtiene que los porcentajes de clasificación correcta superan 70% en todos los casos, que incluso es mayor que el 80% para los modelos de cancelación. Además, muestra que la distribución de préstamos por grupos de riesgo es superior en los casos de pago anticipado, lo que cobra valor al considerarse que estos son grupos de mantenimiento o prevención de los préstamos, mientras que los grupos de riesgo alto para los modelos de cancelación se

refieren a una pérdida total de la operación, por lo que estos préstamos deben trabajarse con mayor cautela.

El cuadro 16 permite observar que las variables relacionadas con el porcentaje de deuda pendiente para cada préstamo resalta en todos los modelos como una característica importante para determinar la cancelación o pago anticipado de préstamos. También lo hace el tipo de cambio en los modelos de cancelación prematura de préstamos y el plazo, así como los meses restantes del préstamo en los modelos de pago anticipado.

Cuadro 16. Resumen de hallazgos

Características de resumen	Préstamos de consumo y prendario		Préstamos hipotecarios	
	Cancelación anticipada	**Pago anticipado**	**Cancelación anticipada**	**Pago anticipado**
Modelo resultante	Modelo exponencial	Regresión logística Bootstrap	Modelo exponencial	Regresión logística Bootstrap
Porcentaje clasificación correcta	89,4	75,0	81,7	73,2
Variables más importantes	• Tipo de cambio • Cantidad de pagos anticipados previos • Porcentaje pendiente de la deuda	• Porcentaje pendiente de la deuda • Meses restantes del préstamo • Plazo	• Tipo de cambio • Cantidad de pagos anticipados previos	• Porcentaje pendiente de la deuda • Meses restantes del préstamo
Número de grupos en la segmentación	5	5	5	5
Riesgo medio	64	199	39	669
Riesgo alto	18	98	3	304
Riesgo muy alto	4	40	4	90

1. 6. CONCLUSIONES Y RECOMENDACIONES

Se estimaron dos modelos estadísticos que permiten determinar el riesgo de que las operaciones de las carteras de préstamos prendarios y de consumo, así como las de los préstamos hipotecarios, se cancelen anticipadamente, con un porcentaje de predicción correcta de 89,4% y 81,7%, respectivamente. Esto permite identificar aquellos préstamos con riesgo alto de cancelación, lo que a la vez permite identificar las operaciones en riesgo para, ya sea retenerlas, o bien, anticiparse y colocar el dinero con otros clientes. Por otro lado, se identificó que el tipo de cambio del momento tiene relación con el riesgo de que un préstamo bancario sea cancelado anticipadamente; además de la existencia de pagos anticipados previos y el porcentaje que se tiene pendiente de la deuda, estas variables representan más del 80% de impacto en el fenómeno según el indicador utilizado para medirlo. Estos resultados son importantes porque contribuyen a que se establezcan estrategias para que el dinero destinado a préstamos sea colocado de la mejor manera por parte de la entidad financiera.

El modelo de tiempo acelerado de falla exponencial resultó ser el que mejor se adapta a los datos para lograr estimar el riesgo de cancelación anticipada de los préstamos, del que se obtuvieron los mejores indicadores de precisión al compararse con los modelos Weibull, Log Logístico y Log Normal.

La cartera de préstamos se estratifica en cinco grupos, de acuerdo con el riesgo de cancelación anticipada. Esto se hace tanto para préstamos de consumo y prendario, por un lado, como para hipotecarios, por otro. De esta forma se identifican aquellos grupos con un riesgo alto de sufrir una cancelación anticipada y pueden concentrar los esfuerzos de la entidad financiera en retener los grupos de préstamos según su prioridad asignada y el valor de estos préstamos.

Determinar el riesgo de cancelación anticipada de los préstamos permite gestionar la cartera de forma que se reduzca la cantidad de dinero ocioso en la entidad financiera, y contribuye a rescatar aquellos préstamos que resultan rentables para el banco pero que pueden cancelarse debido a diferentes situaciones. De esta forma se contribuye en estabilizar la cartera de préstamos y en que esta ofrezca el retorno esperado.

En el caso del riesgo de pago anticipado de los préstamos se ajustaron dos modelos adicionales que clasifican correctamente 75% de las operaciones para el caso de los préstamos prendarios y de consumo, y 73% en el caso de los hipotecarios. Esto, en conjunto con estrategias de negocio y retención, permite minimizar la cantidad de pagos anticipados que desgastan la cartera de créditos en la entidad financiera. Además, se identificó cierta relación entre la probabilidad de efectuar un pago anticipado de los préstamos y el porcentaje de deuda pendiente de ellos, así como la relación con los meses que restan hasta su cancelación, entre los cuales representan más del 35% de impacto sobre el fenómeno estudiado.

En el caso del riesgo de que se efectúen pagos anticipados de los préstamos bancarios, no se encontraron diferencias en precisión entre los métodos de estimación de una regresión logística, considerando los métodos de penalización y Bootstrap; sin embargo, el último se prefiere debido a que con él no es necesario estimar el parámetro de penalización que requieren los métodos Lasso y Ridge.

En el estudio se identifican cinco agrupaciones adicionales mediante las cuales se categorizan los préstamos bancarios según el riesgo de pago anticipado. Esto puede desgastar progresivamente la cartera de préstamos y mediante la segmentación propuesta se pueden crear estrategias para minimizar su efecto.

Con los modelos de pago anticipado de los préstamos se aproxima el riesgo de que en una operación se dé este fenómeno, lo que lo desgasta y posteriormente podría convertirse en una cancelación del préstamo no esperada por la entidad financiera. De esta forma pueden prevenirse los pagos anticipados o gestionar el dinero de forma que lo que se anticipe en un préstamo pueda colocarse en otros para de esta forma mantener un estado óptimo de la cartera.

1. 7 REFERENCIAS BIBLIOGRÁFICAS

Alnsour, M. S. (2013). How to Retain a Bank Customer: A Qualitative Study of Jordanian Banks Relational Strategies. International Journal of Marketing Studies, 5(4), 123-131.

Banco Central de Costa Rica. (2014). http://www.bccr.fi.cr. Recuperado el 28 de Febrero de 2018, de http://www.bccr.fi.cr/indicadores_economicos_/Tasas_interes.html

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). Database Marketing: Analyzing and Managing Customers. Nueva York: Springer Science+Business Media, LLC.

Charlier, E., & van Bussel, A. (Setiembre de 2001). Prepayment Behaviour of Dutch Mortgagors: An Empiriacal Analysis. CentER Discussion Paper, 2001-64.

Choudhry, M. (2012). The Principles of Banking. Singapore: John Wiley & Sons Singapore Pte. Ltd.

Clarke, B., Fokoué, E., & Zhang, H. H. (2009). Principles and Theory for Data Mining and Machine Learning. New York: Springer-Verlag New York.

Consalvi, M., & Scotto di Freca, G. (2010). Measuring prepayment risk: an application to UniCredit Family Financing. Working Paper series(5), 1-38.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1), 1-26.

Finnegan, D. (12 de Setiembre de 2014). Package 'referenceIntervals'. Obtenido de https://cran.r-project.org/web/packages/referenceIntervals/referenceIntervals.pdf

Friedman, J., Hastie, T., & Tibshirani, R. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Nueva York: Springer-Verlag.

García-Santillán, A. (2014). Matemáticas Financieras para la toma de decisiones (Edición electrónica ed.). Málaga, España: Euromediterranean Network.

Gómez-Quesada, Y. I. (2008). El Delito de Intermediación Financiera no Autorizada y la Importancia de su Regulación en la Legislación Penal Costarricense. Tesis para optar por el grado de Licenciatura en Derecho, Universidad de Costa Rica, San José. Recuperado el 19 de Febrero de 2018, de http://iij.ucr.ac.cr/wp-content/uploads/bsk-pdf-manager/2017/07/El-delito-intermediacion-financiera-no-atorizada.pdf

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (Tercera edición ed.). Estados Unidos: Elsevier Inc.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (Tercera edición). New Jersey: John Wiley & Sons, Inc.

Jiménez-Sandoval, H. (1986). Derecho Bancario. EUNED.

Klein, J. P., & Moeschberger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data (Segunda edición ed.). Nueva York: Springer-Verlag.

Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. New Jersey: John Wiley & Sons, Inc.

Manola, A., & Urosevic, B. (2010). Option-Based Valuation of Mortgage-Backed Securities. Economic Annals, 42-66.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. Estados Unidos: O'Reilly Media, Inc.

Rodríguez, G. (2010). Parametric Survival Models. Spring, 1-14.

Te-Hsin, L., & Jian-Bang, L. (Abril - Junio de 2014). A two-stage segment and prediction model for mortgage prepayment prediction and management. International Journal of Forecasting, 30(2), 328-343.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58, 267-288.

Valcárcel-Asencios, V. (2004). Data Mining y el descubrimiento del conocimiento. Industrial Data, 83-86.

Van den Poel, D., & Lariviere, B. (2004). Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. European Journal of Operational Research, 157, 196-217. doi:10.1016/S0377-2217(03)00069-9

1. 8. ANEXOS

Anexo 1
Variables consideradas en los modelos

Variable	Tipo
Cartera de préstamo	Categórica
Antigüedad del cliente	Métrica
Edad del cliente	Métrica
Sexo del cliente	Categórica
Moneda del préstamo	Categórica
Plazo	Métrica
Meses hasta el vencimiento del préstamo	Métrica
Tipo de cambio del momento	Métrica
Tasa de interés total	Métrica
Porcentaje pendiente de deuda	Métrica
Porcentaje de intereses representado en la cuota	Métrica
Cobertura de la cuota	Métrica
Cantidad de cuentas del cliente	Métrica
Tiene CDs	Categórica
Tiene préstamo hipotecario	Categórica
Tiene préstamo libranza	Categórica
Tiene préstamo de consumo	Categórica
Tiene préstamo prendario	Categórica
Tiene *Leasing*	Categórica
Q Trx de frecuencia Alta	Métrica
Q Trx de frecuencia Media	Métrica
Q Trx de frecuencia Baja	Métrica
Monto en trx Monto Alto	Métrica
Monto en trx Monto Medio	Métrica
Monto en trx Monto Bajo	Métrica
Envía Trx Internacionales	Categórica
Recibe Trx Internacionales	Categórica
Envía servicios regionales	Categórica
Recibe servicios regionales	Categórica
Saldo en cuentas	Métrica
Cliente saldo en cuentas negativo	Categórica
Total de deuda en préstamos del cliente	Métrica
Total pagado en préstamos por el cliente	Métrica
Salario del cliente	Métrica
Es cliente voluntario	Categórica
El cliente usa ADS	Categórica
Razón créditos a débitos cuenta	Métrica

Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Variables por seleccionar para los modelos de cancelación de préstamos

Anexo 2
Comparación de variables. Modelos para préstamos consumo y prendario

Variable (*)	Tipo variable	Inclusión de variable en método por pasos				Relación con el préstamo	
		Exponencial	Weibull	LogLogistica	LogNormal (**)	Antigüedad	Cancelación
Cartera de préstamo	Categórica	X	X			0,038	0,052
Antigüedad del cliente	Métrica					0,184	0,214
Edad del cliente	Métrica					0,199	0,223
Sexo del cliente	Categórica					0,039	-0,024
Moneda del préstamo	Categórica		X	X		0,158	0,155
Préstamo con atraso	Categórica	X	X	X		0,046	-0,133
Plazo	Métrica	X	X	X		0,109	0,288
Tipo de cambio del momento	Métrica	X	X	X		-0,208	0,366
Tasa de interés total	Métrica	X	X	X		0,295	0,558
Porcentaje pendiente de deuda	Métrica	X	X	X		-0,824	0,866
Cantidad de pagos anticipados previos	Métrica	X	X	X		0,067	0,160
Cantidad de cuentas del cliente	Métrica					-0,099	0,149
Tiene Certificado a plazos (CD)	Categórica					0,020	-0,007
Tiene Libranza	Categórica					0,044	-0,025
Tiene Hipotecario	Categórica			X		0,045	0,041
Tiene Leasing	Categórica					0,009	-0,059
Q Trx de frecuencia alta	Métrica	X	X	X		-0,115	0,165
Q Trx de frecuencia media	Métrica					-0,081	0,135
Q Trx de frecuencia baja	Métrica	X	X	X		-0,024	0,132
Monto en trx Monto alto	Métrica	X	X	X		-0,035	0,120
Monto en trx Monto medio	Métrica	X	X	X		-0,037	0,125
Monto en trx Monto Bajo	Métrica					-0,086	0,155
Envía o recibe Trx internacionales	Categórica					0,011	0,001
Envía o recibe Servicios regionales	Categórica					0,011	-0,023
Saldo en cuentas	Métrica		X			-0,024	0,133
Total de deuda en préstamos del cliente	Métrica					-0,085	0,192
Total pagado en préstamos por el cliente	Métrica	X	X	X		0,049	0,208
Salario del cliente	Métrica	X	X	X		-0,041	0,112
Es cliente voluntario	Categórica					0,098	0,038
El cliente usa ADS	Categórica					0,020	-0,027
Razón Créditos a débitos cuenta	Métrica			X		-0,044	0,119

(**) El modelo Log Normal no converge para el caso que incluye todas las variables.
(*)Las variables resaltadas en negro fueron consideradas en el modelo.
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Anexo 3
Comparación de variables. Modelos para préstamos hipotecarios

Variable (*)	Tipo variable	Inclusión de variable en método por pasos				Relación con el préstamo	
		Exponencial	Weibull	LogLogistica	LogNormal	Antigüedad	Cancelación
Antigüedad del cliente	Métrica		X	X	X	0,262	0,016
Edad del cliente	Métrica	X		X	X	0,364	0,015
Sexo del cliente	Categórica					0,057	0,041
Moneda del préstamo	Categórica		X			0,049	0,025
Préstamo con atraso	Categórica					0,005	0,034
Plazo	Métrica	X	X	X		-0,379	0,075
Tipo de cambio del momento	Métrica	X	X	X	X	-0,062	0,817
Tasa de interés total	Métrica	X	X	X	X	-0,257	0,005
Porcentaje pendiente de deuda	Métrica	X	X	X	X	-0,542	0,143
Porcentaje de deuda superior a 100%	Categórica	X	X	X	X	0,021	-0,075
Cantidad de pagos anticipados previos	Métrica	X	X	X	X	0,056	0,112
Cantidad de cuentas del cliente	Métrica					-0,175	0,038
Tiene Certificado a plazos (CD)	Categórica					0,030	-0,048
Tiene Préstamo Consumo	Categórica					0,034	-0,090
Tiene Libranza	Categórica					0,039	-0,033
Tiene Prendario	Categórica	X	X	X	X	0,014	-0,041
Tiene Leasing	Categórica					0,008	-0,024
Q Trx de frecuencia Alta	Métrica		X			-0,104	0,026
Q Trx de frecuencia Media	Métrica					-0,077	0,002
Q Trx de frecuencia Baja	Métrica		X			-0,001	0,019
Monto en trx Monto Alto	Métrica					-0,007	0,036
Monto en trx Monto Medio	Métrica			X	X	-0,021	0,027
Monto en trx Monto Bajo	Métrica					-0,064	0,011
Envía o recibe Trx Internacionales	Categórica					0,011	-0,017
Envía o recibe Servicios Regionales	Categórica		X			0,003	0,065
Saldo en cuentas	Métrica					-0,003	0,026
Total de deuda en préstamos del cliente	Métrica					-0,312	0,049
Total pagado en préstamos por el cliente	Métrica					-0,062	0,018
Salario del cliente	Métrica	X		X	X	-0,068	0,026
Es cliente voluntario	Categórica		X			0,117	0,042
El cliente usa ADS	Categórica	X	X	X	X	0,025	-0,082
Razón Créditos a Débitos Cuenta	Métrica					-0,072	0,017

(*)Las variables resaltadas en negro fueron consideradas en el modelo.
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Variables por seleccionar para los modelos de pago anticipado de préstamos

Anexo 4
Comparación de variables. Modelos de pagos anticipados para préstamos prendarios y consumo

Variable (*)	Tipo	Inclusión en método por pasos	Relación pago anticipado
Cartera de préstamo	Categórica	X	-0,016
Antigüedad del cliente	Métrica	X	0,005
Edad del cliente	Métrica	X	0,015
Sexo del cliente	Categórica	X	0,029
Moneda del préstamo	Categórica		0,022
Plazo	Métrica	X	0,004
Meses hasta vencimiento del préstamo	Métrica	X	0,004
Tipo de cambio del momento	Métrica	X	0,018
Tasa de interés total	Métrica	X	0,004
Porcentaje pendiente de deuda	Métrica	X	0,112
Porcentaje pendiente de deuda Mayor 100%	Categórica	X	-0,006
Porcentaje de intereses en cuota 30% o más	Métrica	X	0,003
Porcentaje de intereses en cuota menor 30%	Métrica		0,036
Cobertura de la cuota	Métrica	X	0,041
Cantidad de cuentas del cliente	Métrica		0,007
Tiene certificado a plazos	Categórica	X	0,144
Tiene libranza	Categórica	X	-0,180
Tiene hipotecario	Categórica	X	-0,022
Tiene *Leasing*	Categórica		-0,022
Q Trx de frecuencia alta	Métrica	X	0,038
Q Trx de frecuencia media	Métrica	X	0,008
Q Trx de frecuencia baja	Métrica	X	0,004
Monto en trx Monto alto	Métrica		0,014
Monto en trx Monto medio	Métrica	X	0,004
Monto en trx Monto bajo	Métrica		0,023
Envía Trx Internacionales	Categórica	X	-0,002
Recibe Trx Internacionales	Categórica		0,019
Envía servicios regionales	Categórica		0,018
Recibe servicios regionales	Categórica	X	-0,089
Saldo en cuentas	Métrica		0,017
Cliente saldo en cuentas negativo	Categórica	X	-0,170
Total de deuda en préstamos del cliente	Métrica	X	0,020
Total pagado en préstamos por el cliente	Métrica	X	0,029
Salario del cliente	Métrica	X	0,008
Es cliente voluntario	Categórica		0,026
El cliente usa ADS	Categórica	X	-0,102
Razón Créditos a débitos cuenta	Métrica	X	0,010

(*)Las variables resaltadas en negro fueron consideradas en el modelo.
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

Anexo 5
Comparación de variables. Modelos de pagos anticipados para préstamos prendarios y de consumo

Variable (*)	Tipo	Inclusión en método por pasos	Relación pago anticipado
Antigüedad del cliente	Métrica		0,011
Edad del cliente	Métrica	X	0,015
Sexo del cliente	Categórica	X	0,020
Moneda del préstamo	Categórica		-0,010
Plazo	Métrica	X	0,014
Meses hasta vencimiento del préstamo	Métrica	X	0,018
Tipo de cambio del momento	Métrica	X	0,008
Tasa de interés total	Métrica	X	0,005
Porcentaje pendiente de deuda	Métrica	X	0,169
Porcentaje pendiente de deuda Mayor 100%	Categórica	X	-0,030
Porcentaje de intereses en cuota 30% o más	Métrica		0,043
Porcentaje de intereses en cuota menor 30%	Métrica	X	0,032
Cobertura de la cuota	Métrica	X	0,016
Cantidad de cuentas del cliente	Métrica		0,014
Tiene certificado a plazos	Categórica	X	0,200
Tiene préstamo libranza	Categórica	X	-0,180
Tiene préstamo de consumo	Categórica	X	0,010
Tiene préstamo prendario	Categórica		-0,010
Tiene *Leasing*	Categórica	X	-0,030
Q Trx de frecuencia alta	Métrica	X	0,031
Q Trx de frecuencia media	Métrica	X	0,023
Q Trx de frecuencia baja	Métrica	X	0,004
Monto en trx Monto alto	Métrica		0,046
Monto en trx Monto medio	Métrica		0,014
Monto en trx Monto bajo	Métrica	X	0,008
Envía Trx Internacionales	Categórica		0,050
Recibe Trx Internacionales	Categórica	X	0,120
Envía servicios regionales	Categórica		0,080
Recibe servicios regionales	Categórica	X	0,070
Saldo en cuentas	Métrica	X	0,067
Cliente saldo en cuentas negativo	Categórica	X	-0,240
Total de deuda en préstamos del cliente	Métrica	X	0,043
Total pagado en préstamos por el cliente	Métrica		0,032
Salario del cliente	Métrica	X	0,041
Es cliente voluntario	Categórica		-0,010
El cliente usa ADS	Categórica	X	-0,170
Razón Créditos a débitos cuenta	Métrica	X	0,036

(*)Las variables resaltadas en negro fueron consideradas en el modelo.
Fuente: Almacén de datos de la entidad financiera. Setiembre 2015 - setiembre 2017.

2. PRÁCTICA PROFESIONAL II

2. 1. INTRODUCCIÓN

En la sociedad actual es importante que exista una organización capaz de canalizar el dinero de los individuos de forma óptima, en la cual los recursos se coloquen en el lugar necesario, de ahí que un banco o entidad financiera se constituye en un actor importante dentro de la colectividad, pues, como mencionan Jiménez-Sandoval (1986), esta entidad se convierte en un intermediario entre ahorrantes e inversionistas, y con ello se logra distribuir el dinero de forma más eficiente (p. 25).

Las entidades bancarias son conocidas por tener la capacidad de producir dinero a partir de la colocación de productos activos por los que cobran una tasa de interés mayor a la que pagan por los productos pasivos. López-Pascual & González (2008) mencionan lo anterior con la frase "*el sistema financiero parte de las unidades excedentarias y deficitarias de la liquidez de una economía"* (p. 2). Además, comentan que los actores con necesidades excedentarias pueden aprovechar el ahorro o la inversión y canalizar esos recursos por medio del sistema financiero para mantener seguro su dinero; o bien, rentabilizarlos. Por su parte, los intermediadores financieros se dedicarían a canalizar el dinero desde las unidades excedentarias hacia las deficitarias, mediante diferentes productos bancarios (p. 2 - 3).

Lo anterior sugiere que parte de la estrategia del negocio bancario se basa en la diversidad de productos que ofrecen, los cuales pueden resultar más convenientes para algunos clientes que para otros. Es decir, algunos podrían requerir un lugar seguro en dónde depositar su dinero mientras no lo necesitan; pero, con la intención de recuperarlo en cuanto les sea requerido. Por otro lado, muchos clientes no tienen el dinero que requieren para hacer algún tipo de inversión, por lo que obtienen una línea de crédito para cubrir ese déficit. En pocas palabras, cada cliente tiene necesidades específicas en diferentes momentos de su vida, por lo que no siempre requerirán una cuenta bancaria o un certificado a plazos. No obstante, en otros momentos podrían necesitar de una tarjeta de crédito o un seguro de vida, según el entorno que rodea a cada individuo en un momento específico.

Las entidades bancarias necesitan aprovechar las oportunidades de contacto con los clientes, sin importar el canal en el que puedan crear una ocasión de venta a partir de ese acercamiento, el cual puede ser electrónico o presencial, en sus diferentes modalidades. Abdollahpouri & Abdollahpouri (2013) resaltan en su artículo la importancia de ofrecer productos o servicios y a la vez ganar lealtad del cliente, sabiendo que ese cliente puede encontrarse a uno o dos "clicks" de distancia; además, resaltan la importancia de esta clase de estrategias en el mundo del negocio bancario (p. 1).

Los métodos de venta de productos impulsados por sistemas automáticos de recomendación actualmente son comunes en muchos ámbitos de la vida. Oyebode & Orji (2020) comentan acerca de la gran adaptación que estos sistemas han tenido en las ventas electrónicas y en los dominios de entretenimiento, tales como los de Amazon y Netflix (p. 15). Estos son ejemplos de aplicaciones en sistemas electrónicos; pero, ¿qué pasa con su aplicación en entidades bancarias, en las que lo que se busca es atrapar al cliente a partir de la venta cruzada de otros productos bancarios, pero sin llegar a molestarlo

ofreciéndole productos en los que ese cliente no está interesado? Teniendo esto como fundamento, toma fuerza la opción de utilizar los sistemas de recomendación para este tipo de ofrecimientos.

La problemática principal de la adaptación de los sistemas de recomendación de productos al mundo bancario radica en que los métodos de venta tradicionales se establecieron con el fin de cumplir con metas de colocación; es decir, que las entidades bancarias podían hacer sus esfuerzos para colocar productos según las necesidades propias de la organización, y no tomando en cuenta la necesidad del cliente que, según su etapa de vida y situación personal, podría requerir algún otro producto que no fuera el que le ofrecen, o del cual no tenga conocimiento de que existe, o de su funcionamiento.

Se debe considerar que el contacto de una organización con cada cliente es una actividad valiosa para la entidad, en la cual se tiene la oportunidad de rentabilizar más al consumidor al estimular la venta cruzada de algún producto nuevo. Esto se puede lograr si previamente se conoce, de cada uno, cuál es el producto con más probabilidad de aceptación. Jarrar & Neely (2002) hablan de la importancia y las diferencias que puede haber si se tiene un sistema que permita la venta cruzada de productos en el ámbito bancario, a favor de la rentabilidad que los clientes pueden ofrecer (p. 285).

La utilización de un sistema que permita conocer la prioridad que cada cliente le da a productos y el orden en que se le pueden ofrecer, con base en la probabilidad de que los acepte, puede contribuir a estimular la venta cruzada y a la vez optimizar los diferentes momentos de contacto de la organización con los clientes, para aprovechar mejor la oportunidad de venta. Por otro lado, el producto ofrecido podría ajustarse mejor a la necesidad del cliente según la tendencia establecida en sus datos. Como comentan Vargas-Pérez & Leiva-Olivencia (2015), los sistemas de recomendación de productos adquieren su importancia a partir de la gran variedad de opciones de productos por consumir, en que la oferta es tan grande que se requiere un sistema de recomendación que permita dirigir la venta según el perfil propio de cada cliente (p. 65).

Entre las diferentes técnicas existentes para redactar una recomendación de productos el autor Katsov (2018) sugiere los filtrados con base en contenido, así como los filtrados colaborativos, como los métodos más utilizados para crear estos sistemas de recomendación de productos. Típicamente, ambos métodos buscan determinar un vecindario para los clientes objetivo por recomendar; o bien, los productos de estos. Sin embargo, el funcionamiento del primero radica en el uso de información de contexto propia de cada cliente para identificar este vecindario; mientras que el segundo busca patrones entre los mismos productos, ya sea a partir de la tenencia de estos o de una calificación estándar de cada uno (p. 289).

Si se logra, mediante un sistema de recomendación de productos, obtener la prioridad de aceptación de cada producto, esto permitiría crear listas de clientes a los que se les puedan hacer ofertas específicas, según el producto más probable de aceptar, y enfocar los esfuerzos en la colocación de ese producto específico, y no en aquellos que según el sistema el cliente podría no tener interés. Lo anterior optimizaría los recursos destinados a la colocación de productos. Por otro lado, al priorizar la oferta utilizando el sistema disminuiría la cantidad de veces que se debe tener contacto con un cliente, pues solo se localizaría para ofrecerle aquellos productos que están dentro de su lista de prioridades. Además, cuando el cliente busque contacto por su propia voluntad con la entidad

financiera se podrá aprovechar el momento para ofrecerle algo que él tenga dentro de su lista de prioridades.

Lo anterior se propone bajo el marco del análisis de datos aplicados al negocio bancario, en el que los diferentes conceptos de "Aprendizaje de máquinas", "Minería de datos" y "*Big Data*" pueden entrar a proponer soluciones que permitan el desarrollo del negocio. Fernandez-Naveira *et al* (2018) incluso argumentan que el refinamiento de estos análisis podría aumentar las ganancias de la industria bancaria hasta en un trillón de dólares estadounidenses por año. (p. 6)

En esta investigación se propone la programación de un algoritmo de recomendación de productos bancarios utilizando como base un filtrado basado en contenido que permita, a partir de características propias de los clientes bancarios, establecer un orden de prioridad de venta de productos, determinado para cada cliente, en el que se puedan definir esfuerzos de venta específicos para cada uno. El algoritmo propuesto será puesto a prueba al contrastarlo con otros métodos, como el filtrado colaborativo y los métodos de potenciación.

Con lo anterior se espera identificar el mejor método para realizar la tarea propuesta, pero a la vez para ofrecer una alternativa que podría funcionar en los diferentes casos de estudio que puedan surgir a partir de esta investigación.

A partir del mejor modelo determinado se espera poder determinar para cada producto que entre dentro de la dinámica, definir un orden de prioridad de recomendación de cada uno, para poder ofrecerlo durante cualquier contacto del cliente con la entidad financiera. Suponga un caso donde un cliente bancario se apersona una sucursal de la entidad financiera para retirar dinero, solicitar un estado de cuenta o realizar cualquier otra gestión dentro de la agencia; se podría aprovechar este contacto del cliente, que el mismo hace de manera voluntaria para ofrecer algún producto nuevo, de la misma forma podría ocurrir cuando el cliente contacta por algún medio digital, a la entidad bancaria, siempre y cuando medie una persona preparada y con acceso al orden de prioridad de recomendación de productos del cliente, en caso que el cliente no califique o no se le pueda ofertar el producto, este podría sustituirse por el siguiente en el orden prioritario.

Siguiendo la idea del párrafo anterior, se debe aclarar, que el estudio busca determinar un orden de prioridad de recomendación para todos los productos que se consideren, lo cual difiere de obtener la probabilidad de aceptación, pues en este último caso, podrían existir clientes cuya probabilidad de aceptación de todos los productos es baja, por lo que al tener un contacto con la entidad financiera, no existiría ningún producto a recomendar, mientras que si se determina un orden de prioridad, dejando de lado la probabilidad de aceptación, siempre existirá un producto a recomendar. Una probabilidad de aceptación para cada producto puede ser una herramienta complementaria al orden de prioridad, sin embargo, esto no se considera dentro del alcance del estudio.

El presente informe se compone de seis secciones o capítulos en los que se desarrolla la investigación. En el primero de ellos se introduce el tema y el segundo contiene el marco conceptual necesario para desarrollar lo planteado como objeto de investigación. En el tercero se especifican los objetivos por alcanzar y en el cuarto se establece la metodología adoptada; mientras que en el quinto se discuten los resultados de lo investigado. Para

finalizar, en el capítulo sexto se puntualizan las conclusiones obtenidas del desarrollo de la investigación y se ofrecen algunas recomendaciones al respecto.

2. 2 . MARCO CONCEPTUAL

Existen diversos conceptos en torno al negocio bancario y a los métodos de análisis de datos que deben tenerse claros para entender los procedimientos sugeridos en este estudio, por lo que a continuación se establece un marco conceptual explicativo de dichos conceptos, con la finalidad de facilitar su comprensión.

En el negocio bancario se pueden considerar múltiples actividades mediante las cuales una entidad financiera puede obtener ganancias, como el cobro de honorarios por la gestión de servicios, el cobro de comisiones por el uso de diferentes servicios, como son las transferencias internacionales, el traslado de dinero a otros bancos, la cobertura de seguros, el uso de cajeros automáticos y la intermediación financiera. Gómez-Quesada (2008) describen la intermediación financiera como la acción de captar dinero del público en general para destinarlo a inversión o a la concesión de créditos, lo cual forma parte de una actividad importante para el desarrollo de la economía (p. 2). Por otro lado, Choudhry (2012) comenta que el negocio bancario tiene dos actores principales: prestamistas y prestatarios. Los primeros tienen a su disposición su dinero para que los prestatarios lo utilicen en actividades de inversión. Por otro lado, una entidad financiera es la que lleva a cabo o dirige esa actividad o proceso, en la que los inversionistas depositan su dinero en las cuentas bancarias, lo que hace que los fondos de la entidad aumenten y así se pueda prestar ese dinero a los prestatarios (pp. 6-7).

Existen diversos productos que pueden ofrecer las entidades financieras. Igual (2008) enuncia los productos pasivos con los que se busca captar dinero de los clientes, entre los que destacan los diferentes tipos de cuenta bancaria. Por otro lado, existen los productos activos con los que se trata de dar soluciones de financiamiento a los clientes a cambio de un pago futuro, y también existen los productos de inversión, entre los que cabe señalar los certificados de inversión a plazos. Este autor también menciona los productos de banca electrónica, que permiten a los clientes gestionar sus operaciones financieras sin necesidad de desplazarse a las oficinas bancarias, ya sea por Internet o por medio telefónico, entre otros (pp. 12-13).

En cuanto al tipo de análisis de datos que se utiliza en el estudio, Husamaldin & Saeed (2019, pp. 4 - 5) destacan cuatro diferentes tipos que se conocen en el mundo de *Big Data:*

- Descriptivos: analizan y detallan los datos en tiempo pasado y, en pocas palabras, buscan responder a la pregunta ¿qué pasó?
- Diagnósticos: buscan hacer comparaciones de datos históricos con otras covariables para responder a la pregunta ¿por qué ocurrió?
- Predictivos: utilizan los datos recolectados para identificar la probabilidad de ocurrencia en el futuro y para convertirlos en valor directamente aplicable. Responden a la pregunta ¿qué pasará?
- Prospectivos: utilizan diversas técnicas y metodologías automatizadas para hacer predicciones que ayuden a sugerir opciones de decisión.

La presente investigación podría categorizarse en el ámbito del análisis prospectivo. Siguiendo la definición anterior, se busca desarrollar un sistema que permita recomendar o prescribir un orden de prioridad de productos específicos para cada cliente bancario.

Para lograr lo anterior se tienen diversas alternativas de posibles soluciones que podrían ajustar un sistema de recomendación de productos como el que se persigue en este estudio. Galán-Nieto (2007) describe varias de estas posibilidades y destaca que este análisis se puede hacer de dos formas, principalmente, entre las que se menciona que se sugieran productos a un cliente basados en sus elecciones anteriores; o bien, hacerlo con base en las similitudes, según los puntajes que este genere a partir de los productos que ya él tiene. Lo anterior puede ser expresado por el propio cliente o este podría abstraerse de las acciones que el cliente toma y registra en los sistemas de la organización respecto de la cual presenta algún consumo (p. 2).

Cabe resaltar que el mismo autor, Galán-Nieto (2007), destaca una diferencia entre hacer una predicción que el cliente por sí mismo haría sobre determinado producto, y que se estime la prioridad de elementos que se le podrían recomendar a cada individuo. En la presente investigación se busca la segunda opción mencionada.

Por otro lado, García-Peñalvo & Gil (s.f) comentan en su estudio sobre diferentes técnicas utilizadas para crear algoritmos de recomendación, como el filtrado colaborativo, redes bayesianas, técnicas de clustering y reglas de asociación.

Formoso-López (2013, pp. 29-32) enuncia los diferentes tipos de sistema de recomendación que se resumen a continuación.

- **Basados en contenido.** Recomiendan contenido o productos según el material que anteriormente le ha gustado al individuo.
- **Filtrado colaborativo.** Basan su recomendación en la opinión de otros usuarios con patrones de comportamiento o gustos similares a los del cliente por recomendar.
- **Basados en información demográfica.** Según una agrupación demográfica, como el sexo, la edad o la población, se recomienda a partir de lo que otros usuarios del mismo grupo hayan preferido.
- **Basados en conocimiento del dominio.** En este caso un sistema le consulta las características propias al sujeto al que se le va a recomendar, a partir de las cuales se genera una recomendación. Típicamente, requiere la interacción del individuo con el sistema.
- **Basados en comunidades.** Determina las recomendaciones a partir de las preferencias identificadas en los usuarios que tienen relación social con el individuo al que se le va a recomendar. Se estudia en sistemas de redes sociales, principalmente.
- **Híbridos.** Combinan dos o más de las técnicas mencionadas anteriormente.

En este estudio se establecen sistemas de recomendación basados en un filtrado colaborativo, pues se espera tener una prioridad de recomendaciones para los clientes, según los patrones identificados en la tenencia de productos y en características del comportamiento bancario, así como en algunas demográficas de cada individuo.

Existen diferentes tipos de filtrado colaborativo que podrían resultar útiles según el escenario que se tenga. Cacheda *et al* (2011, pp. 8-15)[11] enlistan estas técnicas. Cabe destacar que algunos de estos métodos requieren que los usuarios califiquen previamente cada uno de los ítems en una escala dada por cada fenómeno estudiado. Una aproximación a esto podría considerarse una matriz de productos, en la que en cada fila se tiene a cada uno de los usuarios y en cada columna están los diferentes ítems por recomendar; y el valor 1 indica la tenencia del producto, mientras que el valor 0 indica que no lo tiene. Esto de forma similar a lo propuesto por Zapata-Sanabria (2019, pp. 10 - 11).

Los tipos de algoritmos se pueden dividir en las siguientes clasificaciones:

- **Basados en usuarios.** También se conocen como basados en vecindarios. Buscan determinar la recomendación a partir de la creación de vecindarios para cada usuario mediante el cálculo de indicadores de similitud contra el resto de usuarios. Necesitan de tres pasos para su ejecución:
    1. Calcular la similitud entre los usuarios activos y el resto de usuarios.
    2. Seleccionar un subconjunto de usuarios según su similitud con usuarios activos.
    3. Determinar la predicción utilizando la calificación del vecindario.
- **Basados en ítems.** Tiene cierta similitud con el anterior, pero en vez de identificar vecindarios de usuarios utiliza los ítems en común que los usuarios tienen para determinar la recomendación.
- **Fusión de similitudes**. Fusiona las dos ideas anteriores y utiliza los puntajes de los ítems de los usuarios activos que se parecen a los usuarios a los que se les desea hacer una recomendación, mientras que, por otro lado, utiliza los puntajes de los ítems de los usuarios activos que tienen los mismos productos que los usuarios a los que se les va a recomendar. Finalmente, utiliza una ecuación para determinar la prioridad de recomendación. Wang *et al* (2006), en Cacheda *et al* (2011), proponen otras variantes adicionales (p. 10).
- **Diagnóstico de personalidad.** Es propuesto por Pennock *et al* (2000) en Cacheda *et al* (2011). Se basa en un modelo de probabilidad en el que los usuarios califican cada ítem basados en promedios de la distribución normal. En este caso la idea es que cada usuario califique un ítem en diferentes momentos y obtenga diferentes calificaciones en cada iteración, posiblemente motivado por el contexto propio de cada individuo en el momento de calificar. Utilizando este modelo se calcula la probabilidad de que un usuario al que se le va a recomendar tenga la misma personalidad que tuvieron otros cuando se calificaron los items (p. 11).
- **Basados en regresión.** Se trata de un algoritmo basado en items con el que se obtienen sus calificaciones a partir de otros ítems similares, calculando para cada dupla un modelo de regresión entre sus calificaciones. Cuando el parámetro de la pendiente de la regresión es cercano a 1 ó a -1, esto indica que el *score* del ítem predictor muestra una buena relación entre los ítems.
- **Primera pendiente (*Slope one*).** Fue expuesto por Lemire & Maclachlan (2005) en Cacheda *et al* (2011). Se basa en predictores de la forma $f(x) = x + b$ donde la constante b se estima como la diferencia promedio entre cada ítem y el ítem por recomendar. Otra variante de este algoritmo pondera los *ítems* utilizando la cantidad de registros que han aportado una calificación a cada ítem y, además, existe otra variante adicional, donde se consideran aquellos ítems que se han

[11] Traducido por el mismo autor.

calificado de forma positiva o negativa. Cabe resaltar que para esta técnica se requiere que existan usuarios que entrenan el modelo previamente, aportando calificaciones a cada ítem (p. 11).

- **LSI/SVD.** Algoritmo propuesto por Sarwar *et al* (2000) en Cacheda *et al* (2011). Se basa en la reducción de dimensiones de la matriz de calificaciones[12], donde la nueva matriz representa, de forma latente, relaciones entre las calificaciones de los ítems, lo cual eliminaría problemas de dispersión y calificaciones ambigüas (p. 12).
- **SVD Regularizado.** Propuesto originalmente por Funk (2006) en Cacheda *et al* (2011). En este caso cada ítem es representado por un conjunto de características y cada usuario está caracterizado por una serie de aspectos que determinan sus preferencias respecto de las características de los ítems. Al final la predicción se obtiene calculando la afinidad de cada usuario con cada ítem, a partir de la asociación entre las preferencias del usuario y las características del ítem. Existen dos variantes de este algoritmo propuestas por Paterek (2007), el SVD implementado y regularizado, también conocido como RSVD2, y el NSVD2, que incorporan mejoras en cuanto a la cantidad de parámetros por estimar. Además, se encuentra el algoritmo de SVD++, propuesto por Koren (2008) en Cacheda *et al* (2011), que considera la retroalimentación implícita, modelada según los ítems que han sido calificados por los usuarios. (p. 12 - 13).
- **Basados en vecindario integrado – Modelo SVD.** Fue propuesto por Koren (2008) en Cacheda *et al* (2011). Combina el algoritmo SVD++ mencionado en el punto anterior con una aproximación basada en vecindarios o basado en ítems.
- **Suavizamiento basado en agrupaciones.** Se trata de un algoritmo similar al basado en usuarios, con algunas diferencias. Así, por ejemplo, como primer paso se deben aglomerar los usuarios en grupos separados. Esto se hace con dos objetivos, el primero de los cuales es incrementar la densidad de la matriz de puntajes, reduciendo los clientes por recomendar a un espectro dentro de un grupo y, además, mejorando la eficiencia del proceso buscando un vecindario, pero únicamente dentro del grupo de pertenencia del usuario por recomendar.

Si se desea programar algún algoritmo de los mencionados se pueden identificar algunos que se aproximan de una forma más sencilla a los del ámbito de estudio, que tienen características propias de cada cliente bancario, como son la edad y el salario recibido, entre otras, y, además, la tenencia o no de algún producto bancario. Podría dividirse el archivo de datos en dos matrices separadas, una en la que se incluye cada cliente como una fila y en las columnas los productos que este cliente tiene codificados como 1, si los tiene, y 0 si no los tiene. Esta última sería el equivalente a la matriz de puntajes que mencionan Cacheda *et al* (2011). Por otro lado, se podría confeccionar una matriz en la que cada registro sea nuevamente un cliente bancario, pero que cada columna sea una característica propia del cliente, la cual podría ser demográfica o podría ser algún reactivo de su comportamiento bancario, como por ejemplo el saldo total de todas sus cuentas bancarias.

Los casos de los filtrados colaborativos, basados en usuarios, basados en ítems y, por supuesto, la fusión de similitudes, resaltan por su simpleza en la metodología, lo que facilitaría su programación. Es algo similar a lo que podría ocurrir con el suavizamiento basado en agrupaciones, que incorpora un paso adicional al basado en usuarios; pero que

[12] Matriz de datos en los que cada usuario califica cada ítem en una escala dada.

mantiene su esencia. Por otro lado, otros algoritmos como SVD, LSI/SVD, Primera pendiente, y los basados en regresión parecen ser más difíciles de adaptar al mundo real, pues requieren que los ítems, en este caso los productos bancarios, sean calificados previamente por otros usuarios, lo que complicaría el proceso de recomendación automática que se busca, al buscar un procedimiento que se alimente de la misma base de datos de la entidad financiera, y que por sí sola devuelva una recomendación o prioridad de productos por recomendar. Herlocker *et al* (1999) comentan que en sí los filtrados colaborativos han tenido gran éxito en diferentes industrias de Internet, como son Amazon.com, CDNow.com y MovieFinder.com (p. 230).

Según Hardesty (2019), en el caso de Amazon en un principio se pensó en utilizar un filtrado colaborativo basado en usuarios; pero se terminó utilizando uno basado en ítems, ponderando la importancia de los ítems por la similitud que tienen los usuarios. Es decir, que si el producto B se relaciona con el ítem A, pero los usuarios que compraron A no se parecen al usuario por recomendar, entonces la probabilidad que le recomiende B al cliente que quiere comprar A es menor. Posteriormente se incluyeron nuevos componentes, como la selección aleatoria de clientes para recomendar a determinado cliente, con lo que se gana en procesamiento de datos, entre otros.

Un caso similar al del párrafo anterior ocurre con Netflix, como comenta Chong (2020), en el que, entre otras técnicas, utiliza un filtrado colaborativo basado en ítems. En este, en el momento en que un usuario mira una película o serie el algoritmo detecta otras películas o serie similar a partir de una matriz de similitudes entre ítems, y devuelve las que tienen mayor similitud. Adicionalmente, tiene una solución de aprendizaje de máquinas en la que se crea un puntaje utilizando un algoritmo de clasificación, y como insumos información histórica del cliente y de las páginas personalizadas que se han creado anteriormente, basadas en sus recomendaciones y en el contenido observado.

Una aproximación por considerar para el estudio es la utilización de algoritmos basados en ítems. Existe una biblioteca de R[13] llamada "recommenderlab" que, según su autor Hahsler, (2021), ofrece una infraestructura de investigación para probar y desarrollar algoritmos de recomendación, entre los que se incluyen las reglas de asociación, las cuales -según Blattberg *et al* (2008)- son reglas que se extraen a partir de productos que se obtienen en conjunto, por ejemplo en el caso de aquellos clientes que compran el producto *A,* ¿qué tan frecuente es que, además, compren el producto *B* (p. 340). La metodología anteriormente expuesta sugiere un algoritmo que podría considerarse como un filtrado colaborativo basado en ítems, que en el mundo bancario podría utilizarse para justamente promover la venta cruzada de productos y para que al obtenerse un producto bancario se revise la frecuencia con la que otros clientes que tienen el mismo producto, adicionalmente, tengan diferentes productos al obtenido, y que se cree una prioridad basada en los productos que obtengan la mayor frecuencia. La ventaja del algoritmo anterior es que se encuentra programado en una de las bibliotecas de R, y que se puede aplicar de forma directa al problema que se expone en el presente estudio. La biblioteca, además, ofrece la posibilidad de ajustar otros métodos, como son los diferentes tipos de filtrado colaborativo, por ejemplo UBCF, IBCF, SVD, SVDF, entre otros.

Por otro lado, si se considera el uso de algoritmos basados en usuarios, no se puede descartar la posibilidad de utilizar modelos de clasificación o también conocidos como

[13] R Core Team (2021)

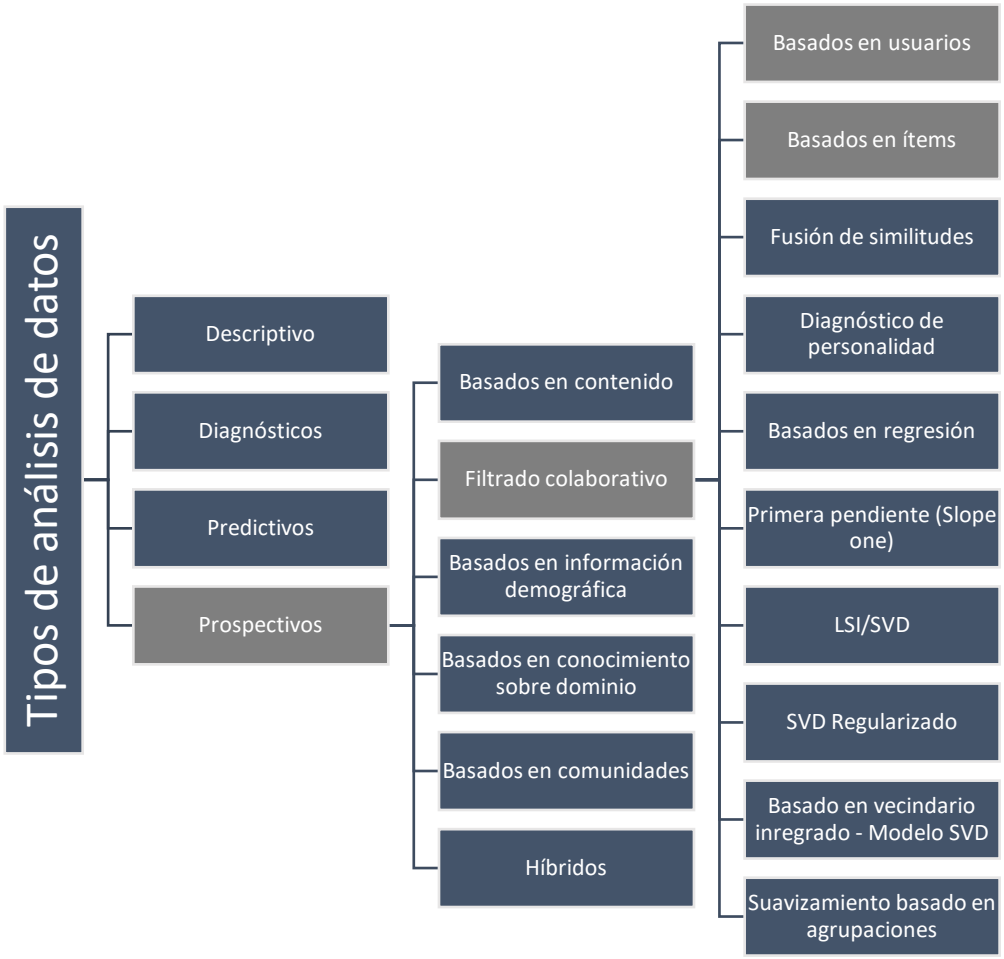
modelos predictivos supervisados. Clarke *et al* (2009) comenta que estos son aquellos en los que se tiene una variable Y por predecir, con la cual se puede buscar inferir a partir de otras dimensiones medidas. El mismo autor ejemplifica el caso de un modelo de regresión como uno de esos casos (p. 231).

Lo anterior cuando los modelos de recomendación se pueden operar como un algoritmo basado en usuarios, en que se pueden considerar características del usuario como variables que permitan predecir una variable *Y*, que indicaría la tenencia o no de un producto específico. El problema se complica al considerarse que no se desea predecir únicamente un producto, sino que son muchos de ellos, por lo que se debería considerar un modelo por cada producto y posteriormente estandarizar las probabilidades ajustadas de cada uno de ellos, para que puedan ser comparables entre sí y no existan efectos de desbalance en sus probabilidades.

No obstante, aun considerando la posibilidad de incorporar modelos de aprendizaje supervisado en los modelos de recomendación se abre la opción de utilizar diferentes técnicas que podrían aportar en la precisión en el momento de recomendar un producto. Una de las técnicas por considerar es la del XGBoost que describen Chen & Guestrin (2016). Se trata de un algoritmo de potenciación de árboles de regresión que ha tenido un gran éxito en competencias como el "Netflix prize" y el "KDD Cup 2015", el cual ofrece ventajas como la eficiencia en su programación, lo que lo hace una técnica fácil de aplicar y escalable a diversos escenarios, que parte del método de potenciación del gradiente expuesto por Friedman (2001) en Chen & Guestrin (2016), que hace modificaciones que permiten que el algoritmo sea más escalable. Los detalles matemáticos de la técnica se encuentran en el artículo mencionado de los autores (p. 1).

Las técnicas descritas anteriormente se resumen en la figura 2, de las cuales se resaltan en gris aquellas que se espera considerar dentro del marco de la investigación actual con base en las necesidades planteadas y en lo explicado anteriormente. Estas técnicas se pueden suponer filtrados colaborativos basados en usuarios cuando se consideren variables propias del usuario para hacer la recomendación, o también análisis basados en ítems, como es el caso del algoritmo de "recommenderlab" de Hahsler, (2021).

Figura 2. Tipos de análisis de algoritmos relacionados con el análisis de datos de un sistema de recomendación de productos



Con fundamento en los párrafos anteriores se ha determinado un conjunto de técnicas que podrían aplicarse al sistema de recomendación de productos, que se pueden comparar entre sí para determinar el método que mejor se adapte al caso de estudio, y esto expone la necesidad de contar con medidas que permitan determinar la mejor técnica.

Los autores Bernardes *et al* (2015) recomiendan el uso de algunos indicadores utilizados tradicionalmente en los modelos supervisados, como son el área bajo la curva ROC o AUC, que parte de la matriz de confusión, compuesta por el porcentaje de aciertos positivos y los aciertos negativos. Además, proponen otras medidas también utilizadas en el mundo de la predicción de variables continuas, como son el error cuadrático medio o RMSE, por sus siglas en inglés, y el error medio absoluto o MAE, también por sus siglas en inglés (p. 28).

Por otro lado, Zapata-Sanabria (2019) también utilizan el área bajo la curva ROC en sus análisis, pero también incorporan la medida de "MAP@K", también llamada "Promedio de la precisión media". Esta calcula la precisión de cada producto recomendado al cliente y posteriormente el promedio de aciertos en cada uno de ellos, para posteriormente promediarlos a todos en un único valor. Además, sugiere el uso del "Uplift", el cual se

describe como una razón del porcentaje de clientes que adquieren determinado producto ofrecido entre el total de clientes que adquieren el mismo (pp. 52 - 54).

Las medidas anteriores pueden aplicarse al contexto de estudio; no obstante, al buscar determinar un orden de prioridad entre los productos por recomendar no se debe ver el problema únicamente a partir de la aceptación o no de un determinado producto, pues se busca que el sistema sugiera un orden y no una recomendación por sí solo, por lo que medidas como el área bajo la curva ROC, "Uplift", "MAP@K", RMSE y MAE se consideran aplicables cuando se determine un punto corte para el orden de prioridad, por ejemplo que se determine recomendar el producto cuando la prioridad se encuentre entre las primeras tres opciones, lo que convertiría el problema en binario. Otra opción en el caso del área bajo la curva ROC, es considerar el orden de priorización de cada producto como una especie de probabilidad inversa, y esto utilizarse de forma equivalente pero inversa a una probabilidad de aceptación.

Se destaca que la eficiencia de estos algoritmos, depende de la disponibilidad y calidad de los datos que se toman para confeccionar las recomendaciones a los clientes, típicamente las entidades bancarias cuentan con un sistema que recolecta información transaccional y demográfica de los mismos, lo cual puede ser utilizado para ajustar modelos de recomendación, entre mayor diversidad y calidad en los datos, se espera obtener un mejor sistema de recomendación. Específicamente en este caso, los datos utilizados no fueron recopilados con el propósito de realizar este tipo de análisis, lo que constituye una limitación importante del estudio.

Los datos utilizados en el estudio surgen de una competencia propuesta en el sitio de Kaggle Inc. (2021) en que el Banco de Santander provee datos. Según el sitio web de la organización Banco Santander, S.A., (2021), se trata de una institución financiera nacida al norte de España a mediados del siglo XIX, que se ha mantenido sólida hasta la actualidad y que ha logrado crecer en España y en los continentes americano y europeo, y que ofrece diversos productos financieros a sus clientes.

2. 3. OBJETIVOS

Con base en lo indicado el problema de estudio se define como la necesidad de establecer un sistema que determine un orden de prioridad para determinar la próxima mejor oferta de productos para cada cliente en un ambiente bancario, y como objetivos de estudio los siguientes:

2. 3. 1. Objetivo general

Establecer un sistema de recomendación de productos a clientes bancarios que permita desarrollar un proceso de ventas eficiente en futuras implementaciones, con las que se busque mejorar la práctica de venta cruzada de productos en la entidad bancaria que lo implemente.

2. 3. 2. Objetivos específicos:

1. Identificar técnicas de recomendación de productos que puedan ser utilizadas para la oferta de productos en el ámbito bancario.

2. Programar un algoritmo de recomendación de productos basado en características propias de los clientes.
3. Determinar el modelo más preciso y eficiente que pueda aplicarse para recomendar productos en el ámbito bancario.

2. 4. METODOLOGÍA

Este capítulo detalla los aspectos metodológicos considerados al momento de realizar el estudio, así como las fuentes de información consultadas, las técnicas estadísticas, los métodos de calibración y evaluación de las técnicas que permiten desarrollar los resultados obtenidos durante la investigación.

2. 4. 1. Fuente de datos

A partir de datos identificados en Kaggle Inc, (2021), que ofrece una plataforma de datos de diversas índoles para realizar análisis estadísticos, se obtienen los datos del Banco de Santander en España[14], en el que en el año 2016 se propuso un concurso para escoger el mejor producto que aceptaría cada cliente bancario el próximo mes, esto habiendo puesto a disposición los datos necesarios para hacerlo.

Se cuenta con un archivo de un total de 13.647.309 de registros, en el que se consideran 956.645 clientes bancarios con una historia de 17 meses. No obstante, este archivo contiene datos de clientes declarados fallecidos, así como con otros que se etiquetan como inactivos. Además, se considerarán únicamente dos periodos del estudio, los clientes que se encontraban activos en mayo del 2015 y en el 2016. Esto para poder hacer una comparación de los mismos clientes, lo que reduce el archivo de datos a un total de 286.081 clientes diferentes por considerar para hacerles una recomendación de productos.

Se busca que estos datos puedan servir para simular un caso de estudio, donde se calibren y prueben los diferentes métodos propuestos para generar un orden de prioridad de productos que permita satisfacer el objetivo general.

2. 4.1.1 Variables de estudio

Al proponer un sistema de recomendación de productos se considera como variable respuesta la posesión de los productos que se deseen ofrecer, para lo cual se consideran: cuentas bancarias, certificados a plazos, cuentas electrónicas, pensiones, hipotecas, préstamos, pago de impuestos y tarjetas de crédito.

Es importante destacar que la representación computacional de estos productos estará dada por una tabla de datos que se denominará “Matriz de productos”, para la cual su primera columna tendrá el identificador de cada cliente, mientras que las demás serán los nombres de los posibles productos por recomendar. Finalmente, cada celda de estas columnas de productos contiene un código “1” cuando el cliente dado tiene el producto respectivo en la columna en el momento del análisis y un “0”, como por ejemplo en el caso del cuadro 17.

[14] En la página de Kaggle Inc (2021) se aclara que los datos no necesariamente corresponden a los de clientes reales del Banco de Santander.

Cuadro 17. Ejemplo de matriz de productos

ID Cliente	Producto 1	…	Producto P
1	1	…	0
…	…	…	
N	P_1	…	P_n

P_n puede tomar los valores de 1 ó 0 en el caso de que el cliente tenga o no el producto, respectivamente.

Como posibles variables de contexto que alimenten el algoritmo programado, así como para las demás técnicas aplicadas se consideran las características del cuadro 18.

Estas variables poseen valores nulos que provienen desde la fuente de datos. Para cada caso se analizará la mejor forma de imputar o eliminar los registros, según corresponda.

Cuadro 18. Descripción de variables de contexto

Variable	Rango de variación
Sexo del cliente	Codificado como 1: Mujer, 0: Hombre.
Edad del cliente	Entre 2 y 164 años.
Antigüedad cliente	Entre 0 y 256 meses.
Condición empleado bancario	Activo, exempleado, filial, no empleado y pasivo.
Residente del país	1: El cliente vive en España, 0: No vive en España.
Extranjero	1: El cliente es extranjero, 0: El cliente nació en España.
Cliente nuevo	1: Ingresó en los últimos 6 meses, 0 Cliente con más de 6 meses de antigüedad.
Provincia de residencia	Provincia de España en la que vive el cliente.
Ingreso del hogar	Entre 1.203 y 28.894.396 (No se indica la moneda)
Segmento del cliente	VIP, Individual o graduado universitario.
Esposa/Esposo empleado bancario	Indica si el cliente es esposo o esposa de un empleado del Banco

Fuente: Kaggle Inc (2021)

Las variables anteriores se representan mediante una matriz que se denominará matriz de características, que es representada por una tabla en la que la primera columna contiene el identificador de cada cliente bancario, mientras que en las restantes columnas se tienen cada una de las características del cuadro 18 para el correspondiente cliente y en el momento dado.

2. 4.2. Técnicas por considerar

Se propone el uso de tres diferentes enfoques para identificar el mejor algoritmo. Para los datos dados el primero considera la técnica "XGBoost" explicada en el marco conceptual, debido a que ha demostrado su efectividad en diferentes concursos de sistemas de recomendación, como los mencionados por Chen & Guestrin (2016). Su desarrollo se basa en un filtrado colaborativo basado en usuarios en que se utilizan variables de contexto del cliente bancario, tales como las mencionadas en la sección 4.1.1. También se considerará el uso de la biblioteca de R, "recommenderlab", lo que vendría a ser similar a un filtrado colaborativo basado en ítems, utilizando la matriz de productos como los ítems por recomendar.

Por último se propone la programación de un algoritmo de filtrado colaborativo basado en ítems, cuyo detalle se podrá consultar en la sección 2.4.3. Con este método se espera tener un procedimiento automatizado que permita tener un orden de prioridad de productos por recomendar para cada cliente, a partir de un perfil específico determinado por sus particularidades en la matriz de características.

2. 4. 3. Algoritmo de recomendación "Próxima mejor oferta"

El algoritmo de recomendación propuesto se basa en el principio de recomendación de persona a persona, como los descritos por King, Lyu & Ma (2010), quienes ejemplifican esta práctica como la utilizada en sitios de internet como Facebook o MySpace, al recomendar a amigos, basados en los amigos de los amigos de cada individuo; o en el ejemplo de Google de recomendar contenido a partir de palabras claves que se escriben en una barra de búsqueda, entre otros sistemas (p. 1356).

Supóngase que cada cliente bancario pudiera recibir una recomendación de sus amigos o parientes más cercanos para saber cuál es el producto bancario que mejor se adapta a sus posibilidades. Este escenario es poco probable de realizar, pues no se pueden conocer los parientes o amigos de cada uno de los clientes de una organización. Sin embargo, se pueden considerar aquellos clientes que tienen alguna similitud con el cliente al que se desea recomendar un producto, con base en otras características, como las que se podrían identificar en la matriz de características expuesta en la sección 2.4.1.1.

Para poder medir la similitud entre los clientes una fórmula de fácil implementación es la de la distancia euclidea (fórmula 1), como la expuesta por Hernández-Rodríguez (2013). El autor resalta que este método tiene como limitaciones que las variables dependen de las unidades de medida y que, además, se les da mayor peso a las variables con mayor magnitud, por lo que, de previo a su uso, debe hacerse una estandarización, y sugiere el uso de la transformación a una distribución normal estándar con media cero y varianza igual a la unidad, lo que se logra mediante la fórmula 2, la cual podría ser cambiada de sentido, al obtener su inverso, el cual se puede interpretar como un indicador de semejanza entre los clientes (pp. 228 - 229).

$$d(I_i, I_j) = \sqrt{(X_{1i} - X_{1j})^2 + \cdots + (X_{ki} - X_{kj})^2} = \frac{1}{s(I_i, I_j)} \qquad \text{Fórmula 1}$$

Donde

- I_{ij}: $Representa\ al\ individuo\ i\ o\ j$
- $X_{k,ij}, ij$: $La\ característica\ k\ del\ individuo\ i\ o\ j$
- $d(I_i, I_j)$: Representa la distancia entre el individuo i y j.
- $s(I_i, I_j)$: Representa la semejanza entre el individuo i y j.

$$Z_i = \frac{X_i - \bar{X}}{S_i} \qquad \text{Fórmula 2}$$

Donde

- $\bar{X}$ *representa la media de la variable* X_i
- S_i *la desviación estándar de la variable* X_i

Por otro lado, también se puede considerar una estandarización de rangos, utilizando la fórmula 3.

$$C_i = \frac{X_i - \min(X)}{\max(X) - min(x)} \qquad \text{Fórmula 3}$$

Donde

- min(X) *representa el valor mínimo de la variable* X_i
- max(X) *representa el valor máximo de la variable* X_i

2. 4. 3. 1. Lógica algorítmica

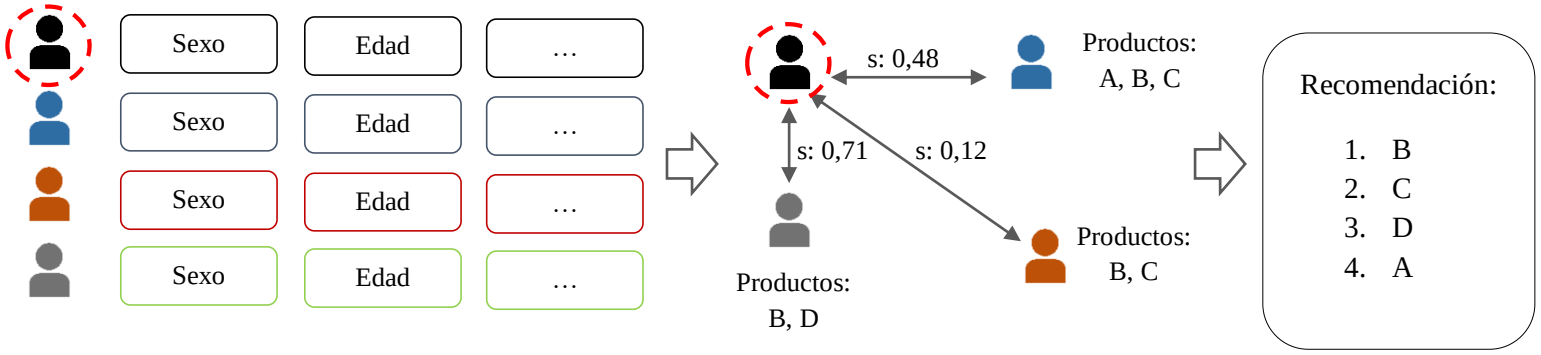
Siguiendo la lógica planteada anteriormente se programó una función en el lenguaje de programación *R* (R Core Team, 2021), que de ahora en adelante se denominará modelo de "Próxima mejor oferta", el cual sigue cuatro sencillos pasos para cada uno de los clientes a fin de generar un orden de recomendación de productos.

1. Determina los K clientes más parecidos al cliente objetivo por recomendar. Esto con base en una distancia euclídea, utilizando como insumo las variables de la matriz de características.
2. Para cada cliente por recomendar se hace un conteo de cada uno de los productos que tienen los K clientes obtenidos en el paso 1.
3. La frecuencia de productos de cada uno de los K clientes con mayor semejanza al cliente objetivo se pondera por el indicador de semejanza entre ellos, es decir, se le da mayor importancia a aquellos productos que tiene el cliente con mayor semejanza.
4. Se ordenan los productos con base en la frecuencia ponderada obtenida del paso anterior y se asigna un orden de prioridad basado en eso.
5. Se descartan los productos que el cliente objetivo ya tenga asignados para evitar recomendaciones innecesarias, cuando esto se aplique.

La figura 3 representa este proceso. Cuando se desea identificar una recomendación para el cliente resaltado con el círculo rojo se toman diferentes variables propias de cada uno de los otros clientes de la tabla que más se parecen a este, y posteriormente, con base en los productos que estos tengan, se hace la recomendación de los productos. Como primera parte se cuenta la cantidad de veces que cada producto aparece. En este caso el producto B lo tienen los tres clientes similares, por lo que se prioriza en el primer lugar, seguido de C, que lo tienen dos de los tres clientes; para finalizar en los casos de D y A, que únicamente uno de los clientes similares los tienen, por lo que el papel del indicador de semejanza toma más valor para este caso, en el que el cliente verde tiene un mayor indicador (0,71), por lo que sus productos adquieren mayor relevancia en el momento de

proponer un producto que los del cliente azul no tienen, pues según esa lógica el siguiente producto en recomendar sería el D, y por último el A.

Figura 3. Representación del sistema de recomendación



Supóngase que se tienen los datos del cuadro 19.

Cuadro 19. Ejemplo de datos

	Matriz Características estandarizadas de 0 a 10		Matriz Productos		
Cliente	**Edad**	**Salario**	**Producto A**	**Producto B**	**Producto C**
Carlos	0,0	0,0	0	0	1
Daniela	1,8	10,0	1	1	0
José	10,0	4,7	1	1	1
Ana	4,4	6,7	0	1	0

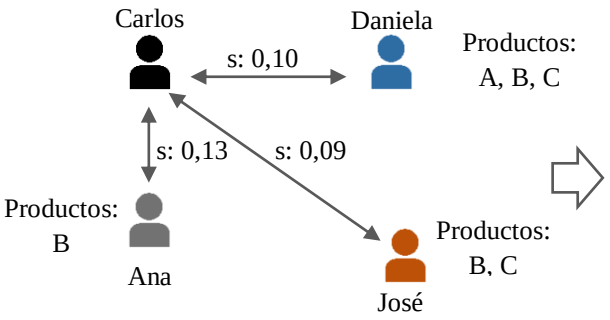
Se puede calcular la distancia entre cada cliente siguiendo la fórmula 1, lo que deja como resultado la matriz de similitudes del cuadro 20:

Cuadro 20. Matriz de similitudes entre clientes

	Carlos	Daniela	José	Ana
Carlos	0,00	0,10	0,09	0,13
Daniela	0,10	0,00	0,10	0,23
José	0,09	0,10	0,00	0,17
Ana	0,13	0,23	0,17	0,00

Si se busca determinar una recomendación para Carlos se tiene que la similitud de este cliente con Ana es de 0,13, con Daniela de 0,10 y con José de 0,09. Si se sustituyen estos valores en la matriz de productos, para los casos en que los clientes similares a Carlos tengan efectivamente el producto, y posteriormente se suman estos valores, da como resultado un puntaje ponderado de la tenencia de los productos por la similitud entre los clientes similares y Carlos, por lo que la recomendación sería, en primer lugar, el producto B, seguido del A y, como Carlos ya tiene el producto C, a este último no se le recomienda (figura 4).

Figura 4. Ejemplo de lógica algorítmica



Matriz Productos Ponderada

	Producto A	Producto B	Producto C
Daniela	0,1	0,1	0
José	0,09	0,09	0,09
Ana	0	0,13	0
Total	0,19	0,32	0,09

La lógica de este proceso es que para cada cliente al que se le va a recomendar se pueda identificar un subconjunto de clientes similares, que podrían tener necesidades similares según el comportamiento de los datos de la matriz de características; por lo que se pueden aislar y en conjunto ofrecerles una probabilidad conjunta de tenencia de un producto, pues al ser un subconjunto de clientes las probabilidades sumadas de cada producto terminarán siendo la unidad en cada caso de recomendación, por lo que estas pueden ser comparadas entre sí, y si bien puede existir un desbalance entre los productos, este sería causado por el perfil propio del cliente por recomendar, lo que coadyuvaría en lograr la priorización de recomendación de productos requerida.

2. 4. 3. 2. Integración de análisis de componentes principales

Una mejora que de forma opcional podría producir el algoritmo es incorporar un análisis de componentes principales a la matriz de características, de previo al proceso de recomendación. Hernández-Rodríguez (2013) comentan que esta técnica típicamente se utiliza para reducir la cantidad de variables por considerar. Además, permite la creación de componentes, explicados por las variables de entrada de datos en forma de una combinación lineal, con la característica de que estos no estarán correlacionados entre sí (p. 93).

La opción de obtener componentes compuestos por las mismas variables de la matriz de correlaciones y de que estos no se encuentren correlacionados entre sí tiene dos ventajas. La primera es que podría mejorar levemente la eficiencia computacional del sistema, pues al considerar menor cantidad de variables el procesamiento también disminuye y, por otro lado, al combinar las variables en constructos contribuye a evitar posibles relaciones directas entre las características propias y los productos por recomendar.

Para explicar lo anterior se considera el ejemplo del producto de una cuenta bancaria. Si entre las variables de la matriz de características se encontraran algunas que pudieran estar directamente relacionadas con este producto, como por ejemplo el saldo que el cliente mantiene en sus cuentas, podría ocurrir que aquellos clientes que tienen saldo se parezcan a sus semejantes que también lo tienen; por lo tanto, al recomendar un producto este podría ser el mismo producto de cuenta bancaria, por lo que el sistema podría descartarlo pues lo considera una recomendación innecesaria. A su vez, aquellos clientes que no tengan saldo en su cuenta se parecerían principalmente a aquellos que tampoco lo tienen, por lo que no tienen el producto, lo que disminuye su prioridad en el momento de recomendarlo. Al incorporar un análisis de componentes principales al proceso se podría

considerar el saldo como una porción de un constructo, por lo que no sería lo único de tomar en cuenta en el momento de hacer la recomendación.

A pesar de que el análisis de componentes principales ofrece ventajas en el momento de proponer el sistema de recomendación, cabe destacar que, como con otras técnicas estadísticas, se debe calibrar el número de componentes por considerar. Hernández-Rodríguez (2013) plantean una serie de posibilidades, entre las que destacan definir una porción de varianza explicada, es decir, tomar el número de componentes que expliquen un X porcentaje de la variabilidad. Esto supone la necesidad de definir un umbral por considerar por el investigador. Por otro lado, se plantea tomar aquellos componentes cuyos valores característicos sean mayores a la unidad, lo que supondría un método sencillo y fácil de aplicar, entre otros métodos expuestos en el documento del autor (pp. 102 - 103).

Para la programación del algoritmo se definirá un parámetro dentro de la función en el que por defecto tendrá un valor nulo. Esto indica que no se debe hacer el análisis de componentes principales de previo a la ejecución de la recomendación; sin embargo, si el parámetro se define como 0 esto indica que se tomarán los componentes cuyos valores característicos sean mayores que la unidad, mientras que cualquier otro valor mayor que cero indica el número de componentes por considerar.

2. 4. 3. 3. Calibración de parámetros

Para esta investigación se plantea trabajar con los datos mencionados en la sección 4.1, en la que para el escenario establecido se deben calibrar los parámetros del número de individuos por seleccionar para cada cliente objetivo que se va a recomendar; o K Vecinos, así como la opción de formular la recomendación utilizando un análisis de componentes principales previamente o no.

Para realizar lo anterior se utilizará el indicador sugerido en la fórmula 4 de la sección 4.6, en la cual, para diferentes opciones de K Vecinos similares al cliente objetivo, se tratará de minimizar este indicador. De la misma forma se decidirá si se utiliza un análisis de componentes principales o no previamente a la recomendación final.

2. 4. 4. Modelos de clasificación

Los modelos de clasificación ofrecen virtudes de análisis como las mencionadas anteriormente por Clarke *et al* (2009); sin embargo, para el escenario actual también suponen una limitación, y es que para su funcionamiento se requiere una variable por predecir Y, cuando en el escenario actual realmente se tienen diferentes opciones de productos, para las que se pretende ofrecer un orden de prioridad por recomendar. Si bien estos modelos permiten que la variable por predecir tenga múltiples categorías estas deben ser mutuamente excluyentes, según lo sugieren Hosmer, Lemeshow & Sturdivant (2013), quienes comentan sobre la aplicación de regresión logística, en los casos en los que las variables por predecir son nominales u ordinales. Para el primer caso sugieren que la lógica sea trabajar con estas variables de forma similar a una variable *"dummie"* (p. 35 - 36), por lo que el uso de esta técnica no se adapta tal cual al escenario de estudio.

A pesar de lo anterior se plantea utilizar cada producto por sí solo, como una variable por predecir, por lo que un modelo de clasificación podría ofrecer la probabilidad de

adquisición de este y aprovechar esta información para establecer un sistema de recomendación.

Lo anterior supone una limitación, y esta es que la proporción de productos en la muestra de estudio podría no estar balanceada o estar con un desbalace diferente por ítem por recomendar. Esto ocasiona que las probabilidades de adquisición de cada producto no sean comparables entre sí, por lo que, si se quiere recomendar un producto u otro, estas probabilidades ajustadas deben ser estandarizadas previamente. En este caso se sugiere el uso de una estandarización a una distribución normal estándar, como la descrita en la fórmula 2, a partir de la cual se establecen todas las probabilidades transformadas en una normal con media igual a cero y desviación estandar igual a la unidad.

Una vez determinada la estandarización de la probabilidad de tenencia de cada producto, se espera que estos productos puedan ser comparables entre sí para de esa forma poder ordenarlos por el indicador resultante, lo que ofrecería un orden de prioridad de recomendación como el propuesto para este análisis. En este caso se plantea el uso de un método de clasificación como el XGBoost, descrito por Chen & Guestrin (2016), por ser una técnica ya utilizada en el mundo de la recomendación de productos.

La calibración de este modelo se ejecuta de forma similar a la del algoritmo de recomendación propuesto en esta investigación, en la que se busca minimizar el indicador propuesto. Sin embargo, cabe resaltar que realmente se plantea crear un modelo de XGBoost por producto por recomendar, lo cual vuelve complicada la posibilidad de calibración para cada uno de ellos, por lo que los parámetros se determinan iguales para cada producto, pero buscando minimizar el indicador de precisión propuesto.

2. 4. 5. Reglas de asociación y otros métodos de recomendación en R

Entre los algoritmos documentados en la sección II resalta el del filtrado colaborativo basado en ítems, para el cual, a partir de una puntuación de cada ítem, o de forma más enfocada en el caso de estudio, la matriz de productos, con la tenencia de cada ítem, se puede obtener una recomendación de productos basada en reglas de asociación como las que ofrece la biblioteca de R, “recommenderlab” (Hahsler, 2021). El uso de esta herramienta permite hacer una recomendación de productos de forma sencilla y rápida, pues el algoritmo ofrece pocos parámetros para calibrar y la única variable estricta que se debe seleccionar es la del método de recomendación, para el caso de una matriz de productos binaria como la que se considera en el caso de estudio. Este parámetro puede tomar los valores, entre los que destacan los diferentes métodos anteriormente mencionados de filtrado colaborativo y de reglas de asociación. Este parámetro será seleccionado de forma que el indicador propuesto en la fórmula 4 de la sección 4.6 sea mínimo, de forma similar a los otros métodos. Los posibles valores que puede tomar el parámetro que se va a calibrar se muestran en el cuadro 21. Cabe resaltar que aquellos métodos que requieran puntuación en los ítems o que no puedan ser estimados, debido a que se requiere información adicional, se descartan del análisis. Además, el caso del método “RERECOMEND” basa su recomendación en aquellos productos que el cliente podría volver a obtener; sin embargo, esto se sale de los alcances de este proyecto.

Cuadro 21. Posibles valores del parámetro "method"

Valor en parámetro "method"	Modelo considerado	Requiere puntuación ítems	Considerado en estudio
UBCF	Filtrado colaborativo basado en usuarios	X	
IBCF	Filtrado colaborativo basado en ítems		X
SVD	Filtrado colaborativo SVD	X	
SVDF	Filtrado colaborativo SVDF	X	
ALS	Mínimos cuadrados alternados		
LIBMF	Matriz de factorización con LIBMF	X	
AR	Reglas de asociación		X
POPULAR	Ítems populares		X
RANDOM	Selección aleatoria de ítems para comparación		X
RERECOMMEND	Ítems re recomendados		
HYBRID	Recomendadores híbridos		

Fuente: Hahsler (2021).

Al final, con la recomendación dada por el sistema de "recommenderlab" se establecerá un orden de prioridad de productos por recomendar para que el mismo sistema sea comparable con el algoritmo propuesto en la sección 4.3.

2. 4. 6. Medidas de precisión

Dado que uno de los objetivos específicos de este estudio es determinar el mejor modelo para el caso expuesto, seguidamente se proponen algunas técnicas con ese propósito.

Utilizando el archivo encontrado en Kaggle Inc (2021) y aprovechando que la información se encuentra de forma histórica para cada cliente con al menos 17 meses para cada uno de ellos, se plantea observar a cada cliente en dos tiempos diferentes, con una diferencia de 12 meses entre uno y otro. Esto para poder ejecutar los procedimientos en el primer momento y posteriormente revisar la tenencia de la proyección de productos recomendados por el sistema para cada cliente en el segundo momento. De esta forma se espera poder evaluar el rendimiento de las diferentes técnicas para comparar su eficiencia en una simulación dentro de la misma base de datos, en la que se espera que aquellos productos que el cliente obtuvo en el segundo momento mencionado hayan tenido una mayor prioridad en el sistema de recomendación. Cabe resaltar que para ambos tiempos de estudio se considerará a los mismos individuos.

En la sección 2.2 se expusieron diferentes métodos para medir la precisión en los sistemas de recomendación, y entre ellos destaca el uso del área bajo la curva ROC. Esta identifica el área acumulada bajo la curva de predicción correcta de un modelo predictivo, y un patrón aleatorio para diferentes puntos de corte de la probabilidad o puntaje ajustados, e intuitivamente sugiere la diferencia entre la predicción utilizando el modelo y lo que se obtendría al no utilizar ninguno. Para el caso en estudio deberá calcularse una curva ROC para cada ítem utilizando el orden de prioridad asignado por cada una de las técnicas anteriormente mencionadas. Finalmente, se obtendrá un indicador promedio a partir del resultado de cada producto por recomendar. Esta técnica se asemeja a la también mencionada anteriormente *MAP@K,* pero en esta, en vez de utilizarse la precisión de aceptación de cada producto, se utiliza el valor del área bajo la curva ROC.

Por otro lado, se sugiere el cálculo de un indicador llamado "Razón de adquirencia promedio" (RAP), basado en el orden de prioridad dado promedio, para aquellos productos que el cliente efectivamente obtiene, dividido entre el promedio de los que no obtiene. De esta forma un indicador cercano a cero indica que el promedio de la prioridad

asignada a los productos que el cliente efectivamente obtuvo fue menor que el de los productos que el cliente no adquirió (fórmula 4).

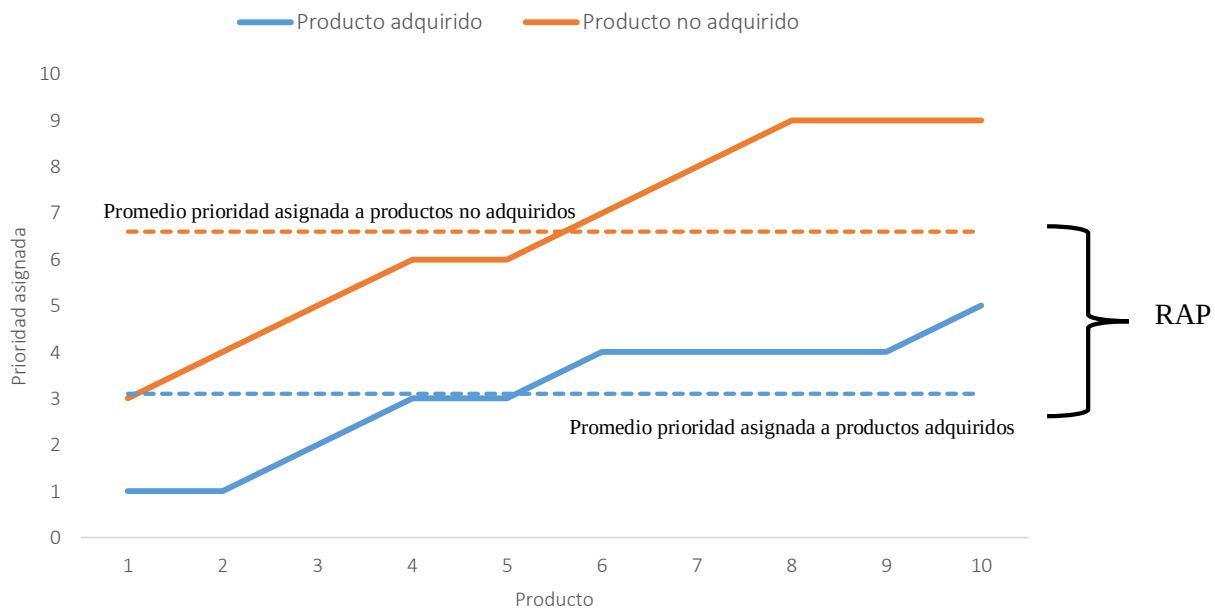
$$RAP = \frac{\overline{PA}}{\overline{PNA}} \qquad \text{Fórmula 4}$$

Donde

- $\overline{PA}$: denota la prioridad promedio de los productos que el cliente adquirió en el futuro.
- $\overline{PNA}$: denota la prioridad promedio de los productos que el cliente no adquirió en el futuro.

Con esto se busca obtener un indicador que logre representar la separación entre la prioridad asignada a los productos que efectivamente se colocaron entre los clientes, y aquellos en los que esto no se logró, como se representa en el gráfico 21:

Gráfico 21. Representación Razón de Adquirencia Promedio



Para el proceso de calibración considera la tabla de datos de entrenamiento únicamente, para dejar la tabla de resultados, dada por el segundo momento de observación para la comparación final de las técnicas, por lo que para calibrar los parámetros, se eliminan aleatoriamente el 10% de los productos bancarios de los clientes, lo que se utiliza para simular los métodos con diferente combinación de parámetros y se selecciona aquella que minimice el indicador RAP descrito anteriormente.

2. 5. RESULTADOS

En esta sección se resumen los principales hallazgos de este estudio enmarcados dentro de los objetivos propuestos. Se detallan por tipo de modelo para facilitar su análisis.

2. 5. 1. Descripción general de datos

Ya sea para ser utilizadas en el algoritmo de "Próxima mejor oferta", o para el modelo de clasificación, las variables propias del cliente, que se almacenan en la también conocida "matriz de características", son necesarias para poder determinar un perfil del cliente a partir del cual se espera hacer una recomendación. Estas variables incluyen información sobre el sexo del cliente, la edad en el momento de capturar la información, los meses desde que ingresó como cliente a la entidad financiera, la condición de empleo, la residencia en el país de España; si nació en un país diferente de España y si es un cliente con menos de seis meses de antigüedad; la provincia de España en la que vive, en caso de ser residente del país, el ingreso familiar reportado y un segmento socioeconómico de ese ingreso. En el cuadro 6 se resumen los principales indicadores relacionados con las variables continuas de la matriz de características, que incluye las variables de edad, antigüedad e ingreso.

Estos datos se obtienen del archivo de Kaggle Inc (2021) y reflejan toda la información que se tiene para los registros de los clientes, por lo que los datos, como la edad mínima de dos años y máxima de 115, se mantienen por simular ser parte de los datos reales de una entidad financiera. En el caso de la antigüedad existen clientes que fueron nuevos ingresos durante el primer momento de observación, motivo por el cual aparecen con cero meses de antigüedad. Por último, se encuentran clientes sin ingreso reportado, casos en los cuales se les anota el ingreso de cero, lo cual se hace para que en el momento de calcular la distancia euclídea estos clientes tiendan a parecerse entre sí y no a aquellos que sí reportan ingreso; caso contrario ocurre con los clientes con ingresos altos, como cuando se reporta "casi 29 millones", pues no se tiene evidencia de que estos valores sean erróneos, por lo que para que el cliente tenga una prioridad de recomendación se mantienen en ese valor (cuadro 22).

Cuadro 22. Indicadores descriptivos de variables continúas en la matriz de características. Periodo 2015

Indicador	Variable		
	Edad (años)	Antigüedad (meses)	Ingreso
Mínimo	2,0	0,0	0,0
Promedio	43,9	96,1	123.008,0
Mediana	44,0	93,0	97.990,0
Máximo	115,0	246,0	28.894.396,0
Desviación estándar	15,6	68,5	243.679,6
Coeficiente variación	35,6%	71,2%	198,1%

Fuente: Kaggle Inc, 2021

Por otro lado, el cuadro 23 muestra los datos para las variables de carácter nominal, en las que se reporta el dato en valor absoluto, por ejemplo, en porcentaje. Para el caso de los clientes que no reportan ingreso, residentes o extranjeros, se considera el total de clientes que cumplen con esta característica, por lo que se reporta el total que lo cumple, así como el porcentaje del ingreso.

En ese caso se tienen variables como "Indicador de empleado bancario" o "Residente", que tienen una predominancia en una categoría, la cual agrupa 99,9% y 99,4% de los registros, respectivamente. La inclusión de estas variables varía de forma muy leve respecto de los resultados generales medidos a partir de los indicadores de RAP y del

promedio de área bajo las curvas ROC, pues únicamente afecta a 400 y a 1.722 de los 286.081 registros totales. Sin embargo, sí se consideran los resultados específicos de priorización de productos para estos clientes. La inclusión de estas variables podría hacer una diferencia en el momento de determinar el producto indicado por ofrecer cuando se realice un contacto con el cliente. Este estudio busca replicar un posible escenario real, por lo que se considera que lo correcto es dejar la variable, para que la recomendación a los empleados bancarios o "No residentes" sea lo más personalizada posible. (Cuadro 23).

Cuadro 23. Frecuencia de categorías de variables nominales en matriz de características. Periodo 2015

Variable	Categoría	Cantidad	Porcentaje
Total		**286.081**	**100,0**
Indicador de empleado bancario	Activo	136	0,0
	Ex empleado	141	0,0
	Filial	122	0,0
	No empleado	285.681	99,9
	Pasivo	1	0,0
Sexo	Hombre	167.155	58,4
	Mujer	118.923	41,6
	No Reportado	3	0,0
Segmento	Top	29.705	10,4
	Particulares	201.530	70,4
	Universitario	54.839	19,2
	Sin Segmento	7	0,0
Sin ingreso reportado		42.895	15,0
Residente		284.359	99,4
Extranjero		11.286	3,9

Fuente: Kaggle Inc, 2021

Para el caso de las provincias de España, en donde habitan los clientes bancarios, esta distribución se muestra en el anexo 1, debido a la gran cantidad de categorías que contiene esta variable.

Por otro lado se tiene el cuadro 24, donde se muestra una comparación de la tenencia de productos en año 2015 y 2016, se observa que la tenencia de productos de forma general aumentó levemente entre los dos periodos de análisis, sin embargo puntualizando en la tenencia de cada uno de ellos, muchos productos disminuyeron, como es el caso de los depósitos de corto y mediano plazo, así como la cuenta "Más particular", que disminuyeron 76%, 29% y 17%, respectivamente. Por otro lado, los productos que más fueron adquiridos por los clientes de un periodo a otro fueron el pago de impuestos (18%) y las cuentas electrónicas (13%). Se destaca, además, que 206.475 clientes tienen cuentas corrientes, que es el producto que más clientes tienen, seguido de "cuenta particular" (con 61.725) y cuentas de planilla con 59.972 clientes que poseen este producto.

Cuadro 24. Distribución porcentual de productos en los dos periodos de análisis

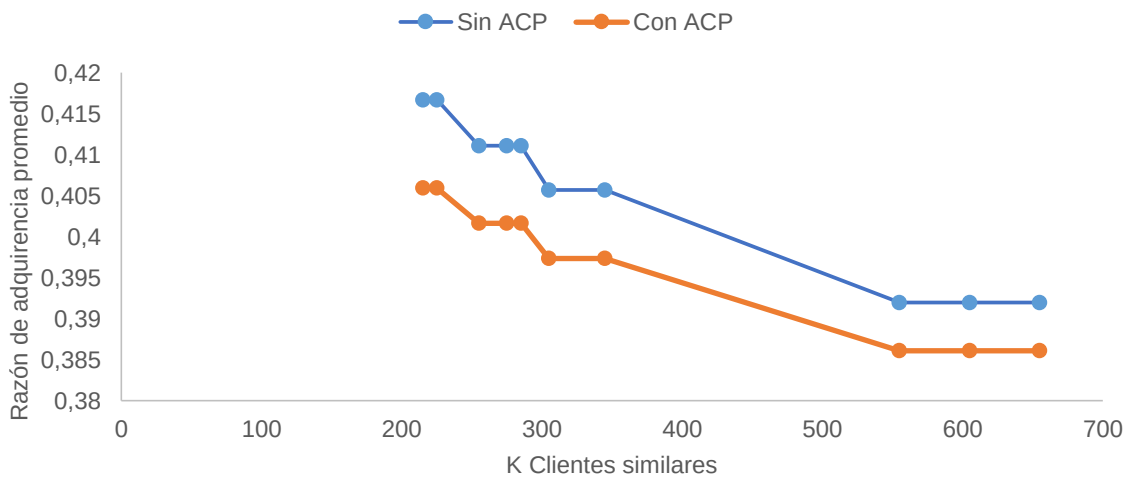
Producto	Periodo				Diferencia	% Diferencia
	2015		2016			
	Total de clientes con el producto	% Clientes con el producto	Total de clientes con el producto	% Clientes con el producto		
Total Clientes	**286.081**	**100,0**	**286.081**	**100,0**		
Total Productos	**555.478**	**-**	**558.250**	**-**	**2.772**	**0,5**
Cuenta corriente	206.475	72,2	204.105	71,3	-2.370	-1,1
Cuenta derivada	272	0,1	273	0,1	1	0,4
Cuenta de planilla	59.972	21,0	63.234	22,1	3.262	5,4
Cuenta junior	5.636	2,0	5.228	1,8	-408	-7,2
Cuenta "Más particular"	4.636	1,6	3.834	1,3	-802	-17,3
Cuenta Particular	61.725	21,6	59.495	20,8	-2.230	-3,6
Cuenta Particular Plus	26.990	9,4	25.415	8,9	-1.575	-5,8
Depósitos a corto plazo	1.304	0,5	315	0,1	-989	-75,8
Depósitos a mediano plazo	1.422	0,5	1.015	0,4	-407	-28,6
Depósitos a largo plazo	32.630	11,4	27.922	9,8	-4.708	-14,4
Cuenta electrónica	55.406	19,4	62.431	21,8	7.025	12,7
Hipoteca	4.699	1,6	4.498	1,6	-201	-4,3
Plan de Pensiones	7.161	2,5	7.167	2,5	6	0,1
Préstamos	1.643	0,6	1.613	0,6	-30	-1,8
Pago de impuestos	31.726	11,1	37.447	13,1	5.721	18,0
Tarjeta de crédito	34.453	12,0	33.712	11,8	-741	-2,2
Seguridades	19.328	6,8	20.546	7,2	1.218	6,3

Fuente: Kaggle Inc, 2021

2. 5. 2. Algoritmo de recomendación "Próxima mejor oferta"

Utilizando el método descrito en la sección 2.4.3.3 se calibró el parámetro de la cantidad de clientes similares que se seleccionan para la recomendación de productos utilizando el algoritmo de "Próxima mejor oferta", descrito en la sección 4.3. Además, se realizó el ejercicio considerando un análisis de componentes principales (ACP) y sin este, para lo cual se identificaron los resultados del gráfico 22. Ahí se muestra que existe una ligera mejoría al utilizar el análisis de componentes principales para ejecutar el algoritmo de recomendación; además, a partir de los 555 clientes similares o vecinos deja de percibirse una mejoría en la razón de adquirencia promedio, por lo que se define este valor como el parámetro utilizado para ejecutar la recomendación.

Gráfico 22. Razón de adquirencia promedio por cada K clientes similares considerados



Fuente: Kaggle Inc, 2021

Con la configuración deseada de k clientes similares igual a 555 y utilizando un análisis de componentes principales previamente.

2. 5. 3. Modelos de clasificación

De forma similar a la calibración realizada en la sección anterior, se seleccionaron diferentes valores para los parámetros por calibrar del algoritmo XGBoost, a partir de lo cual se obtienen aquellos que minimizan el indicador de la razón de adquirencia promedio, y se determinó que la mejor combinación para ejecutar la recomendación de productos es utilizar "max.depth" de 6, "eta" 0,1, "nrounds" igual a 55 (cuadro 25).

Cuadro 25. Combinación de parámetros XGBoost que ofrecen mejores resultados (10 combinaciones con RAP óptimo)

Parámetro			RAP
max.depth	eta	nrounds	
6	**0,1**	**55**	**0,645**
6	0,1	75	0,647
6	0,1	35	0,647
6	0,5	15	0,652
6	0,1	105	0,653
6	0,5	25	0,662
6	0,5	35	0,667
10	0,1	35	0,668
10	0,1	55	0,671
6	1	15	0,673

Fuente: Kaggle Inc, 2021

2. 5. 4. Sistema de "recommenderlab"

Para este análisis de se calibra el parámetro "method" del paquete de R, "recommenderlab" (Hahsler, 2021), para el cual el método que ofrece mejores resultados, según la Razón de adquirencia promedio, es el de "POPULAR". Cabe destacar que los demás métodos mencionados en la sección 4.5 no se pudo estimarlos, pues esos métodos requieren datos adicionales que no se encuentran disponibles, por ejemplo la calificación

de cada cliente de los productos, para el caso de las reglas de asociación, estas no generaron la reglas con la suficiente importancia, para ser considerado dentro del estudio (cuadro 26).

Cuadro 26. Calibración del parámetro "method

Valor en parámetro "method"	RAP
POPULAR	**0,330**
IBCF	0,881
RANDOM	1,013

Fuente: Kaggle Inc, 2021

2. 5. 5. Comparación de técnicas

Al analizar los resultados de los métodos de calibración se pudo observar que todos los métodos probados, logran valores de Razón de adquirencia promedio menores que la unidad, lo que quiere decir que la prioridad asignada a los productos eliminados de los clientes, es menor que las que el cliente efectivamente no tiene, por otro lado los algoritmos de "recommenderlab", así como el de próxima mejor oferta ofrecen indicadores de RAP superiores al del XGBoost; no obstante, esta calibración se hace utilizando los mismos datos de entrenamiento, con la diferencia de que se elimina aleatoriamente 10% de los productos, lo cual difiere del método de evaluación propuesto, pues en este último se espera observar los productos que a lo largo del tiempo los clientes efectivamente adquirieron. Con esto se espera que aquellos productos que se adquirieron tengan un orden de prioridad asignado por cada técnica menor que los que no se adquirieron, para lo que se utilizan los indicadores de la RAP, así como las curvas ROC promedio por producto.

Los resultados del cuadro 11 ofrecen el promedio de las curvas ROC obtenidas para cada producto por cada una de las diferentes tres metodologías utilizadas, así como el detalle del indicador en cada producto. Cabe resaltar que la metodología que ofrece el mejor resultado es el de "Próxima mejor oferta", seguido muy de cerca por el "recommenderlab", esto basado en el promedio de las curvas ROC, no obstante la diferencia entre ambas técnicas es mínima, por lo que no se puede concluir que alguna sea mejor que la otra.

Por otro lado y tomando como referencia el cuadro 27, este ofrece el coeficiente de variación de las curvas ROC en los productos, como indicador de la variabilidad en la precisión de cada producto, resalta que este es menor en el modelo de "Próxima mejor oferta" que el de los otros métodos y al comparar los productos donde ofrece un mayor indicador, sobresale que en once de los diecisiete productos considerados, es decir en un 65%, el modelo de "Próxima mejor oferta" supera al de "recommenderlab". Aun así las curvas ROC para todos los modelos resultan ser bajas, pues están bajo el umbral del 50%, lo que impide concluir cual es el mejor modelo, por lo que considerar otros indicadores como el de la "Razón de adquirencia promedio", puede ser de utilidad para llegar a una conclusión sobre este tema.

Cuadro 27. Comparación de curvas ROC entre métodos de recomendación de productos

Producto	Técnicas consideradas		
	Próxima Mejor Oferta	XGBoost	Recommenderlab
Promedio	**44,2%**	**37,6%**	**42,2%**
Coeficiente Variación	**25,4%**	**51,3%**	**38,1%**
Cuentas corrientes	50,3%	45,3%	50,0%
Cuenta derivada	48,9%	38,7%	50,0%
Cuenta de Planilla	54,5%	38,0%	55,0%
Cuenta junior	57,0%	0,0%	83,6%
Cuenta "Más particular"	37,8%	6,7%	57,7%
Cuenta "Particular"	52,5%	21,9%	53,9%
Cuenta "Particular Plus"	46,4%	23,0%	34,2%
Depósitos a corto plazo	6,6%	90,0%	19,9%
Depósitos a mediano plazo	46,1%	44,0%	28,3%
Depósitos a largo plazo	47,0%	38,3%	43,8%
Cuenta electrónica	51,7%	40,8%	51,6%
Hipoteca	39,7%	42,9%	20,2%
Plan de pensiones	37,9%	39,0%	29,1%
Préstamos	46,8%	52,4%	31,6%
Pago de impuestos	44,3%	39,7%	40,3%
Tarjeta de crédito	42,8%	37,8%	37,8%
Seguridades	40,2%	40,0%	31,0%

Fuente: Kaggle Inc, 2021

Como segundo indicador se obtiene la Razón de adquirencia promedio, en la que se observa que el modelo "recommenderlab" ofrece una ligera ventaja sobre el de "Próxima mejor oferta" de aproximadamente 3,8% de mejoría, lo cual aún se considera poco. No obstante la diferencia de los métodos del algoritmo de "recommenderlab" y "Próxima mejor oferta", contra el del XGBoost, si parece ser importante, lo que permite definir que los primeros dos métodos resultan ser los que mejor proporcionan un orden de prioridad para la recomendación de productos (Cuadro 28).

Cuadro 28. Comparación de Razón de Adquirencia Promedio entre métodos de recomendación de productos

Método	RAP
Recommenderlab	0,346
Próxima Mejor Oferta	0,359
XGBoost	1,248

Fuente: Kaggle Inc, 2021

La figura 5 muestra las principales ventajas y desventajas de cada técnica. Cabe resaltar que para esto se considera que los modelos que demostraron tener una precisión, considerando el orden de prioridad establecido para la recomendación de productos en el caso de estudio, fueron el de "Próxima mejor oferta" y el "recommenderlab", con una ligera diferencia entre ambos. Por otro lado, el algoritmo de "XGBoost", si bien ha tenido gran aceptación científica en otros casos de estudio, se ha propuesto con el fin de realizar predicciones a categorías mutuamente excluyentes o binarias, según fuera el caso, por lo que la adaptación realizada para generar un orden de prioridad, después de ajustar un modelo predictivo para cada producto de manera independiente, resulta en una metodología que no ofrece lo que se busca en este análisis. Por otro lado, una de las

desventajas de los métodos utilizados en el algoritmo de "recommenderlab", es el que estos no consideran la matriz de características entre sus parámetros, por lo que en una eventual mejora de la conformación de este insumo, no aportaría a una mejora en la precisión del modelo, cosa que si podría ocurrir con los otros métodos.

Figura 5. Ventajas y desventajas según los resultados y el análisis de cada método

Próxima mejor oferta

Ventajas

- Fácil programación.
- Metodología sencilla.
- Precisión y estabilidad en el caso de estudio.

Desventajas

- Requiere mejorar eficiencia computacional.

XGBoost

Ventajas

- Aceptación científica como modelo predictivo.

Desventajas

- No fue hecho para el caso de estudio, por lo que su adaptación influye en la pérdida de precisión.

Recommenderlab

Ventajas

- Fácil uso.
- Pocos parámetros a optimizar.
- Precisión en el caso de estudio.

Desventajas

- Basa su recomendación únicamente en la matriz de productos y no considera la de características.

La forma operativa en la que el resultado de estos métodos debe interpretarse es específica para cada cliente al que se le haga una recomendación. Si se toma como ejemplo el caso del cliente 1000027, este cliente tiene las características del cuadro 29. Se resalta que este es un hombre de 44 años de edad, que no es empleado del banco pero que tiene una antigüedad de 43 meses de relación bancaria, que es residente en España y nació en ese mismo país; además, vive en Madrid y reporta un ingreso de 133.379, aproximadamente, y, por último, pertenece al segmento "Particulares".

Cuadro 29. Características del cliente 1000027

Indicador	Valor
Id Cliente	1000027
Indicador empleado bancario	N
Sexo	V
Edad	44
Indicador cliente nuevo	0
Antigüedad en meses	43
Residente	S
Extranjero	N
Provincia	Madrid
Ingreso reportado	133.379
Indicador de no Ingreso	0
Segmento	Particulares

Fuente: Kaggle Inc, 2021

Este cliente, a lo largo de los 12 meses transcurridos entre el primer y el segundo momento de observación, obtuvo como producto nuevo una cuenta corriente.

Por otro lado, los resultados de priorización de los métodos de estudio se muestran en el cuadro 30, en el que se representan con prioridad "0" aquellos productos que el cliente ya tenía en el momento de hacerse la recomendación, que son la cuenta planilla, la cuenta electrónica, el pago de impuestos y la tarjeta de crédito. Se observa, además, que tanto el

método de “Próxima mejor oferta” como el de “recommenderlab” priorizan el producto obtenido en el primer lugar, mientras que XGBoost lo coloca en la prioridad número 7.

Cuadro 30. Priorización de productos por método de recomendación para el caso del cliente 1000027

Producto	Método de recomendación		
	Próxima Mejor Oferta	XGBoost	"recommenderlab"
Cuenta Corriente	1	7	1
Cuenta Derivada	10	17	17
Cuenta de planilla	0	0	0
Cuenta Junior	17	3	7
Cuenta "Más particular"	17	2	9
Cuenta "Particular"	2	12	2
Cuenta "Particular plus"	4	11	4
Depósitos a corto plazo	11	1	12
Depósitos a mediano plazo	9	10	11
Depósitos a largo plazo	7	4	3
Cuenta electrónica	0	0	0
Hipoteca	6	9	8
Plan pensiones	5	6	6
Préstamo	8	5	10
Pago de impuestos	0	0	0
Tarjeta de crédito	0	0	0
Seguridades	3	8	5

Fuente: Kaggle Inc, 2021

Tomando como ejemplo el orden de prioridad dado por el método de “Próxima mejor oferta” observado en el cuadro 30, el banco del caso de estudio, podría tomar acciones para colocar productos que el cliente 1000027 no tenga. Por ejemplo, suponga que este requiere hacer el pago correspondiente a su tarjeta de crédito, por lo que va a una sucursal bancaria a hacerlo y al llegar, la persona que lo atiende ingresa su número de identificación en el sistema de gestión de clientes, el cual muestra los primeros tres productos a recomendar. En este caso los productos serían una cuenta corriente, una cuenta “Particular” y un producto de seguridades crediticias, por lo que con esta información el agente bancario que lo atiende podría preguntar por la posibilidad de adquirir alguno de estos productos y en caso que el cliente acepte, se gestiona su apertura en el mismo momento.

El escenario anterior, puede replicarse para cualquier contacto del cliente, con la entidad bancaria, donde medie de antemano un colaborador bancario que podría actuar como vendedor del producto, como por ejemplo al llamar al centro de llamadas del banco o al escribir por medio de algún chat de servicio digital, entre otros. Esto sugiere una serie de retos, como capacitar a los colaboradores en todos los posibles productos a ofertar, o tener una estrategia que permita comunicar al cliente con el encargado de producto más apto para realizar la venta una vez que se haya tenido contacto con el cliente. Por otro lado también se debe integrar la información resultado del sistema de recomendación con los sistemas de gestión de clientes bancarios, que permitan saber el orden de prioridad de productos para el cliente en cualquier canal de contacto con la entidad financiera, pero la solución de estos sistemas en conjunto con el sistema de recomendación, podría resultar en una oportunidad de rentabilizar a los clientes bancarios.

2. 6. CONCLUSIONES Y RECOMENDACIONES

Las entidades financieras son instituciones en las cuales la riqueza de los datos generados por la actividad de sus usuarios permite desarrollar diferentes análisis que pueden fundamentar la creación de estrategias de negocio para fomentar la venta cruzada de productos, lo que podría contribuir a incrementar la rentabilidad de los clientes.

Para fomentar la venta cruzada de productos se pueden considerar diversos métodos y sistemas de recomendación establecidos, o bien, crearse híbridos entre ellos que se adapten de la mejor forma a las necesidades del negocio y que permitan optimizar los recursos y el contacto con los clientes.

En esa línea, en el presente estudio se identificaron diferentes métodos de recomendación de productos bancarios que se pueden considerar en el momento de crearse un sistema con el fin de obtener un orden de prioridad de recomendación específica para cada cliente, que pueda ser utilizado en el momento del contacto entre la entidad bancaria y el cliente al que se desea recomendar productos.

El uso del algoritmo "Próxima mejor oferta" puede resultar en una técnica con virtudes que bien puede ser utilizada en el ámbito bancario, pues mediante ella pueden ordenarse los productos según su probabilidad de aceptación por parte de cada cliente, lo que permite que en el momento de tenerse contacto con el cliente se sepa qué ofrecer, y que en el caso de que el cliente no acepte o no califique para esta recomendación se tenga una variedad de opciones de productos que, de igual forma, se podrían ofrecer en posteriores contactos con el cliente.

Si bien el sistema aquí propuesto permite darle prioridad al ofrecimiento de productos, esto no debe confundirse con la probabilidad de aceptación; es decir, que un cliente podría mostrar una baja probabilidad de aceptación de equis producto, pero que ese siga siendo el producto más necesario de recomendar, ya que simplemente los otros productos son todavía menos probables de aceptación. Esta premisa sugiere que la combinación del algoritmo "Próxima mejor oferta" con otros modelos de aceptación podría crear un sistema óptimo para la recomendación y aceptación de productos, en que se descarten aquellos con una baja probabilidad de aceptación, en los casos o canales de contacto en los que sea más difícil colocarlos.

El algoritmo de XGBoost permite generar un orden de prioridad de aceptación de productos para cada cliente, al realizarle algunas modificaciones en la forma en que usualmente se utiliza o creando modelos específicos para cada producto, y posteriormente estandarizando sus resultados para que puedan ser comparables. Si bien en el caso de estudio no resultó ser el modelo con la mayor precisión sí podría serlo para otros casos de estudio; o bien, en el caso de que se utilice para determinar la aceptación o no de un producto específico, dejando de lado el orden de prioridad del ofrecimiento.

El sistema propuesto por Hahsler (2021), es decir, el "recommenderlab", resultó ser un sistema que se adapta bien al caso de estudio pues su facilidad de uso y practicidad lo convierten en una buena opción como método de recomendación.

Se puede afirmar entonces que para el caso de estudio los métodos más precisos para generar un orden de prioridad de productos son el de "Próxima mejor oferta" y el

"recommenderlab". La diferencia entre los indicadores RAP y el promedio de curvas ROC es poca, por lo que ambos métodos se pueden considerar precisos.

Todo lo anterior deja como motivo de reflexión la importancia de considerar otros aspectos en el momento de decidir sobre un método de recomendación de productos, más allá de su precisión, tales como la practicidad de su uso, el consumo computacional de recursos, la adaptación con otros sistemas de ventas, la interacción con el cliente, y su entendimiento a nivel gerencial de una organización bancaria, aspectos que para el caso de estudio no fueron considerados.

Para el buen funcionamiento de un sistema de recomendaciones, se requiere como material fundamental, datos precisos y diversos que compongan la matriz de características, esto condiciona el análisis a la disponibilidad de estos datos en los sistemas de la entidad financiera que lo maneje. Lo recomendable es confeccionar bases de datos con las características que se consideren adecuadas para realizar este tipo de análisis y así mejorar sus resultados.

En el caso del algoritmo de "Próxima mejor oferta", destaca que el sistema requiere de información de entorno del cliente para determinar sus vecinos, por lo que robustecer la variedad de características del cliente que alimenten el sistema, podría mejorar sus resultados.

2. 7. REFERENCIAS BIBLIOGRÁFICAS

Abdollahpouri, H., & Abdollahpouri, A. (Mayo de 2013). An approach for personalization of banking services in multi-channel environment using memory-based collaborative filtering. 208-213. doi:10.1109/IKT.2013.6620066

Banco Santander S.A. (2021). https://www.santander.com/. Recuperado el 28 de Abril de 2021, de https://www.santander.com/es/sobre-nosotros/nuestra-historia

Bernardes, D., Diaby, M., Fournier, R., Fogelman-Soulié, F., & Viennet, E. (Mayo de 2015). A Social Formalism and Survey for Recommender Systems. ACM SIGKDD Explorations Newsletter, 16(2), 20 - 37. doi:https://doi.org/10.1145/2783702.2783705

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). Database Marketing: Analyzing and Managing Customers. Nueva York: Springer Science+Business Media, LLC.

Cacheda, F., Carneiro, V., Fernández, D., & Formoso, V. (Febrero de 2011). Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems. ACM Transactions on the Web, 5(1). doi:10.1145/1921591.1921593

Chen, T., & Guestrin, C. (Agosto de 2016). XGBoost: A Scalable Tree Boosting System. KDD'16, 13-17.

Chong, D. (30 de Abril de 2020). Towards data science. Recuperado el 21 de Abril de 2021, de towardsdatascience.com: https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48

Choudhry, M. (2012). The Principles of Banking. Singapore: John Wiley & Sons Singapore Pte. Ltd.

Clarke, B., Fokoué, E., & Zhang, H. H. (2009). Principles and Theory for Data Mining and Machine Learning. New York: Springer-Verlag New York.

Fernandez-Naveira, C., Jacob, I., Rifai, K., Simon, P., & Windhagen, E. (19 de Setiembre de 2018). Smarter analytics for banks. Global Banking. Recuperado el 12 de Abril de 2021, de https://www.mckinsey.com/industries/financial-services/our-insights/smarter-analytics-for-banks

Formoso-López, V. (2013). Técnicas eficientes para la recomendación de productos basadas en filtrado colaborativo. Tesis doctoral, Universidad da Coruña, Departamento de Tecnoloxías da Información e as Comunicacións, La Coruña.

Galán-Nieto, S. M. (2007). Filtrado Colaborativo y Sistemas de Recomendación. Madrid: Universidad Carlos III de Madrid.

García-Peñalvo, F. J., & Gil, A. B. (s.f.). Personalización de Sistemas de Recomendación. Departamento de Informática y Automática – Universidad de Salamanca.

Gómez-Quesada, Y. I. (2008). El Delito de Intermediación Financiera No Autorizada y la Importancia de su Regulación en la Legislación Penal Costarricense. Tesis para optar por el grado de Licenciatura en Derecho, Universidad de Costa Rica, San José. Recuperado el 19 de Febrero de 2018, de http://iij.ucr.ac.cr/wp-content/uploads/bsk-pdf-manager/2017/07/El-delito-intermediacion-financiera-no-atorizada.pdf

Hahsler, M. (2021). recommenderlab: Lab for Developing and Testing Recommender. 0.2-7. Obtenido de https://CRAN.R-project.org/package=recommenderlab

Hardesty, L. (22 de Noviembre de 2019). The history of Amazon's recommendation algorithm. Recuperado el 21 de Abril de 2021, de amazon.science: https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (Agosto de 1999). An Algorithmic Framework for Performing Collaborative Filtering. SIGIR'99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrival, 230-237. doi:https://doi.org/10.1145/312624.312682

Hernández-Rodríguez, Ó. (2013). Temas de análisis estadístico multivariante. San José: Editorial Universidad de Costa Rica.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (Tercera edición ed.). New Jersey: John Wiley & Sons, Inc.

Husamaldin, L., & Saeed, N. H. (2019). Big Data Analytics Correlation Taxonomy. Information, 1-12. doi:10.3390/info11010017

Igual, D. (2008). Conocer los productos y servicios bancarios: Productos de tesorería, de inversión, de financiación, leasing, factoring, renting, tarjetas. Barcelona: Profit.

Jarrar, Y. F., & Neely, A. (Febrero de 2002). Cross-selling in the financial sector: Customer profitability is key. Journal of Targeting Measurement and Analysis for Marketing, 10(3), 282–296. doi:10.1057/palgrave.jt.5740053

Jiménez-Sandoval, H. (1986). Derecho Bancario. EUNED.

Kaggle Inc. (25 de 03 de 2021). kaggle.com. Obtenido de https://www.kaggle.com/c/santander-product-recommendation

Katsov, I. (2018). Introduction to Algorithmic Marketing. Artificial Intelligence for Marketing Operations. Grid Dynamics.

King, I., Lyu, M. R., & Ma, H. (2010). Introduction to social recommendation. (págs. 1355 - 1356). Raleigh, Carolina del Norte: WWW 2010. doi:10.1145/1772690.1772927

López-Pascual, J., & González, A. S. (2008). Gestión Bancaria. Factores claves en un entorno competitivo (Tercera ed.). Madrid: McGraw-Hill.

Oyebode, O., & Orji, R. (Enero de 2020). A hybrid recommender system for product sales in a banking environment. Journal of Banking and Financial Technology(4), 15-25. doi:10.1007/s42786-019-00014-w

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Viena, Austria. Obtenido de https://www.R-project.org/

Vargas-Pérez, P., & Leiva-Olivencia, J. L. (2015). Prototipo de sistema de recomendación grupal en un destino turístico. ROTUR, Revista de Ocio y Turismo, 9, 62-81. Obtenido de http://www.rotur.es

Zapata-Sanabria, D. R. (2019). Sistema de recomendación de productos financieros: una aplicación empírica en banca de consumo. Tesis magister en economía, Universidad del Rosario, Facultad de economía, Bogotá.

2. 8. ANEXOS

Anexo 1.

Anexo 1
Distribución porcentual de clientes por provincia de España.
Parte 1

Provincia	Cantidad	Porcentaje
Total	**284.361**	**100,0**
Alava	899	0,3
Albacete	2.004	0,7
Alicante	5.568	2,0
Almeria	1.289	0,5
Asturias	4.463	1,6
Avila	912	0,3
Badajoz	2.155	0,8
Balears, Illes	2.486	0,9
Barcelona	22.456	7,9
Bizkaia	4.293	1,5
Burgos	1.946	0,7
Caceres	1.678	0,6
Cadiz	4.423	1,6
Cantabria	3.559	1,3
Castellon	1.757	0,6
Ceuta	163	0,1
Ciudad Real	2.203	0,8
Cordoba	2.155	0,8
Coruña, A	7.596	2,7
Cuenca	993	0,3
Gipuzkoa	1.724	0,6
Girona	1.366	0,5
Granada	3.039	1,1
Guadalajara	1.451	0,5
Huelva	1.578	0,6
Huesca	693	0,2
Jaen	1.320	0,5
Leon	1.832	0,6
Lerida	1.049	0,4
Lugo	1.753	0,6
Madrid	123.442	43,4
Malaga	5.962	2,1
Melilla	186	0,1
Murcia	4.518	1,6
Navarra	1.737	0,6
Ourense	1.553	0,5
Palencia	1.011	0,4
Palmas, Las	4.196	1,5
Pontevedra	4.948	1,7
Rioja, La	1.742	0,6
Salamanca	3.288	1,2
Santa Cruz De Tenerife	1.734	0,6
Segovia	1.046	0,4
Sevilla	10.878	3,8
Soria	364	0,1
Tarragona	1.728	0,6
Teruel	333	0,1
Toledo	3.438	1,2
Valencia	11.987	4,2
Valladolid	4.887	1,7
Zamora	1.059	0,4
Zaragoza	5.521	1,9
No Reporta	1.720	0,6

Fuente: Kaggle Inc, 2021

Anexo 2. Código utilizado durante el desarrollo del análisis.

```
#Exploración y depuración de datos.
library(readr)
library(dplyr)
library(tidyr)
setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")

db <- readRDS(file = "train.rds")

db %>% filter(ncodpers == 1050611) %>% View()

length(unique(db$ncodpers))
length(unique(db$fecha_dato))
length(unique(db$ncodpers)) * length(unique(db$fecha_dato))
nrow(db)
length(unique(db$ncodpers[db$fecha_dato == max(db$fecha_dato)]))
length(unique(db$ncodpers))
min(db$fecha_dato)
max(db$fecha_dato)

table(db$nomprov)
table(db$conyuemp)

summary(db)
summary(as.numeric(db$age))

summary(db$tipodom)
table(db$indfall)
table(db$nomprov)

db %>%
  filter(indfall == "") %>%
  View()

db <- db %>%
  select(-tipodom, -ult_fec_cli_1t) %>%
  filter(ind_actividad_cliente == 1 & tiprel_1mes == 'A' & indfall == "N")

summary(db)
db %>%
  filter(is.na(ind_nomina_ult1) == TRUE) %>%
  View()

db <- db %>%
  mutate(ind_nomina_ult1   =   if_else(is.na(ind_nomina_ult1)   ==   TRUE,   0,
as.double(ind_nomina_ult1)),
         ind_nom_pens_ult1  =  if_else(is.na(ind_nom_pens_ult1)  ==  TRUE,  0,
as.double(ind_nom_pens_ult1))
         )

summary(db)

db %>%
  filter(is.na(cod_prov) == TRUE) %>%
  View()

db <- db %>%
  mutate(nomprov = if_else(is.na(cod_prov) == TRUE, "NoReporta", nomprov) #Se
usa cod_prov porque es mas sencillo el codigo, pero al final uso nombre
  ) %>%
  select(-cod_prov)

table(db$age)
db <- db %>%
  mutate(age = as.numeric(age),
```

```
         antiguedad   =   if_else(as.numeric(antiguedad)   <   0   ,   0   ,
as.numeric(antiguedad))
  )

db %>% filter(is.na(renta) == TRUE) %>%
  select(ncodpers) %>% #View()
  unique() %>%
  count()


db  %>%
  select(ncodpers) %>%
  unique() %>%
  count()


db %>% filter(ncodpers == 1051174) %>% View()

db %>%
  mutate(SinIngreso = if_else(is.na(renta) == TRUE, 1, 0)
         ) %>%
  group_by(segmento) %>%
  summarise(PSinIng = mean(SinIngreso),
            Ingreso = mean(renta,na.rm = TRUE)
            )

de filtro.
db <- db %>%
  mutate(SinIngreso = if_else(is.na(renta) == TRUE, 1, 0),
          ) %>%
  mutate(renta = if_else(is.na(renta) == TRUE, 0, renta),
         segmento = if_else(segmento == "", "SinSegmento", segmento),
         conyuemp = if_else(conyuemp == "", "N", conyuemp)

        ) %>%
  select(-ind_actividad_cliente, -tiprel_1mes, -indfall, -pais_residencia, -
canal_entrada, -fecha_alta)


summary(db)
table(db$conyuemp)


table(db$fecha_dato)


##Dejamos unicamente las fechas de interes 2015-05-28 y 2016-05-28
db <- db %>%
  filter(fecha_dato %in% c("2015-05-28", "2016-05-28"))

db %>% select(ncodpers) %>% unique %>% count() #378367 personas distintas

#Personas que aparecen en ambos meses:
personas <- db %>%
  group_by(ncodpers) %>%
  summarise(Q = n()) %>%
  filter(Q == 2) %>%
  select(-Q)

personas <- personas$ncodpers
length(personas) #Quedan 286081 que se encuentran en ambos periodos


db <- db %>%
  filter(ncodpers %in% personas) %>% #Dejamos solo las que aparecen en ambas
fechas
  arrange(ncodpers, fecha_dato)
rm(personas)
```

```
A2015 <- db %>%
  filter(fecha_dato == '2015-05-28')

A2016 <- db %>%
  filter(fecha_dato == '2016-05-28')

rm(db)

#Separamos los archivos en productos y similitudes para cada anno
A2015.Prod <- A2015 %>%
  select(ncodpers,  ind_cco_fin_ult1,
         ind_cder_fin_ult1,         ind_cno_fin_ult1,         ind_ctju_fin_ult1,
ind_ctma_fin_ult1,
         ind_ctop_fin_ult1,        ind_ctpp_fin_ult1,        ind_deco_fin_ult1,
ind_deme_fin_ult1,
         ind_dela_fin_ult1,         ind_ecue_fin_ult1,         ind_hip_fin_ult1,
ind_plan_fin_ult1,
         ind_pres_fin_ult1,        ind_reca_fin_ult1,        ind_tjcr_fin_ult1,
ind_valo_fin_ult1
)

A2016.Prod <- A2016 %>%
  select(ncodpers,  ind_cco_fin_ult1,
          ind_cder_fin_ult1,         ind_cno_fin_ult1,         ind_ctju_fin_ult1,
ind_ctma_fin_ult1,
          ind_ctop_fin_ult1,        ind_ctpp_fin_ult1,        ind_deco_fin_ult1,
ind_deme_fin_ult1,
          ind_dela_fin_ult1,         ind_ecue_fin_ult1,         ind_hip_fin_ult1,
ind_plan_fin_ult1,
          ind_pres_fin_ult1,        ind_reca_fin_ult1,        ind_tjcr_fin_ult1,
ind_valo_fin_ult1)

A2015.Sim <- A2015 %>%
  select(ncodpers,  ind_empleado,  sexo,  age,  ind_nuevo,  antiguedad,  indrel,
indrel_1mes,
         indresi, indext, conyuemp, nomprov, renta, SinIngreso, segmento,
)

A2016.Sim <- A2015 %>%
  select(ncodpers,  ind_empleado,  sexo,  age,  ind_nuevo,  antiguedad,  indrel,
indrel_1mes,
         indresi, indext, conyuemp, nomprov, renta, SinIngreso, segmento,
  )

##Escribimos los archivos en RDS
saveRDS(A2015.Prod,file = "DatosAntes201505PROD")
saveRDS(A2016.Prod,file = "DatosAntes201605PROD")

saveRDS(A2015.Sim,file = "DatosAntes201505SIM")
saveRDS(A2016.Sim,file = "DatosAntes201605SIM")

###### Creacion tabla de calibración #######
library(dplyr)
library(tidyr)
library(tictoc)
library(openxlsx)
setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")
source("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Algoritmos/Re
comendacion10.R")
datos.sim <- readRDS(file = "DatosAntes201505SIM")
datos.prod <- readRDS(file = "DatosAntes201505PROD")

###Reconfiguracion de campos a numerico
datos.sim$ELIMINAR <- 1
datos.sim.mod <- as.data.frame(model.matrix(ELIMINAR ~., data = datos.sim[,-1]))
datos.sim.mod %>% head(n = 100) %>% View
```

```
datos.sim.mod$ncodpers <- datos.sim$ncodpers

datos.sim.mod <- datos.sim.mod %>%
  select(ncodpers,ind_empleadoB:segmentoSinSegmento)

datos.sim <- datos.sim %>%
  select(-ELIMINAR)

set.seed(1989)
Porcentaje.eliminacion.productos <- 0.9
prod.antes <- datos.prod
tic()
for(i in 1:nrow(datos.prod)){
  for(j in 2:ncol(datos.prod)){
    if(prod.antes[i,j]  == 1){
      prod.antes[i,j]    <-    rbinom(n    =    1,    size    =    1,    prob    =
Porcentaje.eliminacion.productos)
    }

    if((i*j %% round((0.05*nrow(datos.prod)*ncol(datos.prod)))) == 0){
      print(paste("Avance:",round(100*i*(j-
1)/(nrow(datos.prod)*ncol(datos.prod)),digits = 1)))

    }

  }
}
toc()

sum(prod.antes[,-1])
sum(datos.prod[,-1])
sum(prod.antes[,-1])/sum(datos.prod[,-1]) #Se eliminaron el 10% de los datos de
forma aleatoria

saveRDS(object = prod.antes, file = "Sin10PorcientoProductos")

########   Calibración del K optimo para algoritmo de próxima mejor oferta
###############

source("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Algoritmos/Se
leccionKNBO.R")
#Con muestra de 2500 registros
prod.antes <- readRDS(file = "Sin10PorcientoProductos")
k.optimo <- seleccionKNBO(datos.sim = datos.sim.mod,
              productos.antes = prod.antes,
              productos.despues = datos.prod,
              vector.k = c(5,15,25,55,105,215, 305, 445, 555, 885),
              muestra = 2500)

#Con muestra de 25000 registros
tic()
k.optimo <- seleccionKNBO(datos.sim = datos.sim.mod,
                          productos.antes = prod.antes,
                          productos.despues = datos.prod,
                          vector.k  =  c(215,225,255,275,285,305,345,555,  605,
655)
                          )
                          #muestra = 40000)
toc()
k.optimo
write.xlsx(x = k.optimo$indicas,file = "CalibracionSinPCA.xslx")

gc()
###Con PCA
tic()
k.optimo <- seleccionKNBO(datos.sim = datos.sim.mod,
                          productos.antes = prod.antes,
                          productos.despues = datos.prod,
```

```
                          vector.k   =   c(215,225,255,275,285,305,345,555,605,
655),
                          factores = 0
                          #muestra = 25000
)
toc()
k.optimo
write.xlsx(x = k.optimo$indicas,file = "CalibracionConPCA.xslx")

pruebaKNBO(datos.sim = datos.sim,
           productos.antes = productos.antes,
           productos.despues = productos.despues,
           kvecinos = 55,factores = 0
           )
 ######### Recomendaciones   ##################
##Caso sin ACP
tic()
rec <- recomend(datos = datos.sim.mod, productos = datos.prod, kvecinos = 555)
toc()
saveRDS(object = rec, file = "RecomendacionNBOSinPCA")

##Caso con ACP
tic()
rec <- recomend(datos = datos.sim.mod, productos = datos.prod, kvecinos =
555,factores = 0)
toc()
saveRDS(object = rec, file = "RecomendacionNBOConPCA")

########### Calibración para XGBoost ###############

xg.func <- function(max.depth, eta = 1, nthread = 2, nrounds = 55){

probs <- matrix(data = NA, nrow = nrow(datos.prod), ncol = (ncol(datos.prod)-
1))

for(i in 2:ncol(datos.prod)){
  print(paste("Vamos por columna",i, "de ", ncol(datos.prod)))
  datos.prod.i <- datos.prod[,c(1,i)]
  datos.sim.var <- datos.sim %>%
    left_join(datos.prod.i, by = names(datos.prod.i)[1])

  Nombre.y <- names(datos.prod.i)[2]
  names(datos.sim.var)[which(names(datos.sim.var) %in% names(datos.prod.i)[2])]
<- "y"

  SP <- sparse.model.matrix(y ~ ind_empleado + sexo + age + ind_nuevo +
antiguedad + indrel + indrel_1mes + indresi + indext + conyuemp + nomprov +
renta + SinIngreso + segmento,
                            data = datos.sim.var)
  mod <- suppressMessages( xgboost(data = SP,
                 label = datos.sim.var$y,
                 max.depth = max.depth,
                 eta = eta,
                 nthread = nthread,
                 nrounds = nrounds,
                 objective = "binary:logistic",
                 verbose = 0, eval_metric = 'logloss'
  ))
  probs[,i-1] <- predict(mod, SP)


}

##Estandarizacion de probabilidades

probs.est <- (scale(probs))
summary(probs.est)
summary(probs)
```

```
probs.est <- cbind.data.frame(datos.prod[,1],probs.est)
names(probs.est) <- names(datos.prod)

datos.prod.t <- datos.prod %>%
  pivot_longer(!ncodpers, names_to = "Productos", values_to = "Tenencia")

recomendacion <- probs.est %>%
  pivot_longer(!ncodpers, names_to = "Productos", values_to = "Score") %>%
  left_join(datos.prod.t, by =c("ncodpers", "Productos")) %>%
  mutate(Score = if_else(Tenencia == 1, -9999, Score)) %>%
  arrange(ncodpers, desc(Score)) %>%
  group_by(ncodpers) %>%
  mutate(Prioridad = row_number()) %>%
  mutate(Score = if_else(Tenencia == 1, 9999, Score)) %>%
  mutate(Prioridad = (1-Tenencia)*Prioridad,
         MaxScore = min(Score)
         ) %>%
  ungroup() %>%
  mutate(Prioridad = if_else(Score == MaxScore, 24, as.double(Prioridad))) %>%
  select(-Score, -Tenencia, -MaxScore) %>%
  pivot_wider(names_from = Productos, values_from = Prioridad)

#Reordenamos para que queden en el mismo orden que en el archivo original.
recomendacion<-recomendacion[,names(datos.prod)]


##Prueba de eficiencia comparable con las del sistema de recomendacion
productos.despues <- readRDS(file = "DatosAntes201505PROD")

return(Eficiencia_NBO(rec = recomendacion,productos.despues = readRDS(file =
"DatosAntes201505PROD")))

}


#setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")
source("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Algoritmos/Ef
iciencia_NBO.R")
datos.sim <- readRDS(file = "DatosAntes201505SIM")
datos.prod <- readRDS(file = "Sin10PorcientoProductos")
setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")
source("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data/Codigos/
RecomendacionXGBOOSTPruebaFunction.r")

set.seed(1989)
max.depth. <- c(6,10,15)
eta. <- c(0.1,0.5,1)
nrounds. <- c(15, 25, 35, 55, 75, 105)
#nrounds. <- c(15,16)
q = 1
mat <- matrix(data = NA, nrow = 55, ncol = 4)
for(md in 1:length(max.depth.)){
  for(e in 1:length(eta.)){
    for(nr in 1:length(nrounds.)){
      print("############################################################")
      print(paste("Vamos por ",q, "de", 55, round(100*(q/55),2),"%"))
      print("############################################################")
      a <- suppressWarnings(xg.func(max.depth = max.depth.[md], eta = eta.[e],
nrounds = nrounds.[nr]))
      mat[q,1] <- max.depth.[md]
      mat[q,2] <- eta.[e]
      mat[q,3] <- nrounds.[nr]
      a <- a$indicador
      mat[q,4] <- a$Promedio
      q <- q+1
    }
  }
```

```
}

mat2 <- as.data.frame(mat)
names(mat2) <- c("max.depth","eta","nrounds","RAP")

write.xlsx(x = mat2, file = "CalibracionXGBoost.xlsx")
####### Recomendación de XGBoost #########
setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")
datos.sim <- readRDS(file = "DatosAntes201505SIM")
datos.prod <- readRDS(file = "DatosAntes201505PROD")


###Para la primer variable

probs <- matrix(data = NA, nrow = nrow(datos.prod), ncol = (ncol(datos.prod)-
1))

for(i in 2:ncol(datos.prod)){
  print(paste("Vamos por ",i))
  datos.prod.i <- datos.prod[,c(1,i)]
  datos.sim.var <- datos.sim %>%
    left_join(datos.prod.i, by = names(datos.prod.i)[1])

  Nombre.y <- names(datos.prod.i)[2]
  names(datos.sim.var)[which(names(datos.sim.var) %in% names(datos.prod.i)[2])]
<- "y"

  SP <- sparse.model.matrix(y ~ ind_empleado + sexo + age + ind_nuevo +
antiguedad + indresi + indext + nomprov + renta + SinIngreso + segmento,
                            data = datos.sim.var)
  mod <- xgboost(data = SP,
                 label = datos.sim.var$y,
                 max.depth = 55,
                 eta = 0.1,
                 nthread = 6,
                 nrounds = 15,
                 objective = "binary:logistic"
  )
  probs[,i-1] <- predict(mod, SP)


}


##Estandarizacion de probabilidades

probs.est <- (scale(probs))
summary(probs.est)
summary(probs)

probs.est[is.na(probs.est)] <- (min(probs.est,na.rm = TRUE) - 0.1)
probs.est <- cbind.data.frame(datos.prod[,1],probs.est)
names(probs.est) <- names(datos.prod)

datos.prod.t <- datos.prod %>%
  pivot_longer(!ncodpers, names_to = "Productos", values_to = "Tenencia")

df <- probs.est %>%
  pivot_longer(!ncodpers, names_to = "Productos", values_to = "Score") %>%
  left_join(datos.prod.t, by =c("ncodpers", "Productos")) %>%
  mutate(Score = if_else(Tenencia == 1, -9999, Score)) %>%
  arrange(ncodpers, desc(Score)) %>%
  group_by(ncodpers) %>%
  mutate(Prioridad = row_number()) %>%
  mutate(Score = if_else(Tenencia == 1, 9999, Score)) %>%
  mutate(Prioridad = (1-Tenencia)*Prioridad,
         MaxScore = min(Score)
         ) %>%
```

```
  ungroup() %>%
  mutate(Prioridad = if_else(Score == MaxScore, 24, as.double(Prioridad))) %>%
  select(-Score, -Tenencia, -MaxScore) %>%
  pivot_wider(names_from = Productos, values_from = Prioridad)

#Reordenamos para que queden en el mismo orden que en el archivo original.
df<-df[,names(datos.prod)]

saveRDS(object = df, file = "RecomendacionXGBOOST")

######## Calibración recomenderlab ######
set.seed(1989)
datos.sim <- readRDS(file = "DatosAntes201505SIM")
datos.prod <- readRDS(file = "Sin10PorcientoProductos")

ids <- datos.prod$ncodpers

matriz <- datos.prod %>%
  select(-ncodpers) %>%
  as.matrix() %>%
  as("binaryRatingMatrix")

recommenderRegistry$get_entry_names()
# Se cambia a mano el parámetro "method" para "POPULAR", "IBCF" y "RANDOM"
Recomendacion <- Recommender(data = matriz, method = "POPULAR")

rec <- predict(object = Recomendacion, matriz, type = 'topNList', n = 17)
rec.original <- rec
#rec <- rec.original
rec <- as(rec, "list")

prods <- colnames(datos.prod)[-1]
rec.mat <- matrix(data = NA, nrow = length(rec), ncol = length(prods))

q.prods <- ncol(datos.prod)-1
for(i in 1:length(rec)){
  cl <- rec[[i]]
  p <- rep(q.prods,q.prods)
  if(length(cl) > 0){
    for(j in 1:length(prods)){
      for(e in 1:length(cl)){
        if(prods[j] == cl[e]){
          p[j] <- e
        }
      }
    }
  }
  rec.mat[i,] <- p
}

  rec.mat <- (1-datos.prod[,-1])*rec.mat

  #rec.mat <- as.data.frame(rec.mat)
  names(rec.mat) <- prods
  rec.mat$ncodpers <- datos.prod$ncodpers
  rec.mat<-rec.mat[,names(datos.prod)]

  recomendacion <- rec.mat %>%
    pivot_longer(!ncodpers, names_to = "Productos", values_to = "Prioridad") %>%
    arrange(ncodpers, Prioridad) %>%
    filter(Prioridad != 0) %>%
    group_by(ncodpers) %>%
    mutate(Prioridad2 = row_number()) %>%
    mutate(MaxScore = max(Prioridad)) %>%
    ungroup() %>%
    mutate(Prioridad     =     if_else(Prioridad     ==     MaxScore,     24,
as.double(Prioridad2))) %>%
    select(-Prioridad2, -MaxScore) %>%
```

```
    pivot_wider(names_from = Productos, values_from = Prioridad)

  recomendacion[is.na(recomendacion)] <- 0
  recomendacion<-recomendacion[,names(datos.prod)]
  recomendacion <- as.data.frame(recomendacion)
  Eficiencia_NBO(rec      =      as.data.frame(recomendacion),productos.despues
=readRDS(file = "DatosAntes201505PROD"))

######## Recomendación con recommenderlab ########
datos.prod <- readRDS(file = "DatosAntes201505PROD")

ids <- datos.prod$ncodpers

matriz <- datos.prod %>%
  select(-ncodpers) %>%
  as.matrix() %>%
  as("binaryRatingMatrix")

recommenderRegistry$get_entry_names()
metodos. <- c("IBCF", "POPULAR", "RANDOM", "RERECOMMEND", "UBCF") #Se elimina
el hibrido pues da un error de que no base recommender specified
#Se elimina ALS porque no corre
#Se elimina AR por no encontrar reglas que cumplan con support o confianza
Recomendacion <- Recommender(data = matriz, method = "POPULAR")

rec <- predict(object = Recomendacion, matriz, type = 'topNList', n = 17)
rec.original <- rec
#rec <- rec.original
rec <- as(rec, "list")

prods <- colnames(datos.prod)[-1]
rec.mat <- matrix(data = NA, nrow = length(rec), ncol = length(prods))

q.prods <- ncol(datos.prod)-1
for(i in 1:length(rec)){
  cl <- rec[[i]]
  p <- rep(q.prods,q.prods)
  if(length(cl) > 0){
    for(j in 1:length(prods)){
      for(e in 1:length(cl)){
        if(prods[j] == cl[e]){
          p[j] <- e
        }
      }
    }
  }
  rec.mat[i,] <- p
}

rec.mat <- (1-datos.prod[,-1])*rec.mat

names(rec.mat) <- prods
rec.mat$ncodpers <- datos.prod$ncodpers
rec.mat<-rec.mat[,names(datos.prod)]
recomendacion <- rec.mat %>%
  pivot_longer(!ncodpers, names_to = "Productos", values_to = "Prioridad") %>%
  arrange(ncodpers, Prioridad) %>%
  filter(Prioridad != 0) %>%
  group_by(ncodpers) %>%
  mutate(Prioridad2 = row_number()) %>%
  mutate(MaxScore = max(Prioridad)) %>%
  ungroup() %>%
  mutate(Prioridad = if_else(Prioridad == MaxScore, 24, as.double(Prioridad2)))
%>%
  select(-Prioridad2, -MaxScore) %>%
  pivot_wider(names_from = Productos, values_from = Prioridad)

recomendacion[is.na(recomendacion)] <- 0
```

```
recomendacion<-recomendacion[,names(datos.prod)]
recomendacion <- as.data.frame(recomendacion)

saveRDS(object = recomendacion, file = "RecomendacionRECOMMENDERLAB")

####### Funciones de evaluación ########
####### Selección KNBO #######
pruebaKNBO<-function(datos.sim,productos.antes,productos.despues,ponderado    =
TRUE, reescalar ="rangos", ordinal = TRUE ,kvecinos, anulacion = TRUE, factores
= NULL)
 {
   rec<-recomend(datos  =  datos.sim,productos  =  productos.antes,ponderado  =
ponderado,reescalar = reescalar, ordinal = ordinal,kvecinos = kvecinos,anulacion
= anulacion,transpuesta = TRUE, factores = factores)
   rec <- rec[rec$Prioridad != 0,]
   ###BD despues transpuesta
   identificador <- names(productos.despues)[1]
   df<-gather(data  =  productos.despues,key  =  Producto,value  =  Tenencia,  -
identificador)

   df[,1] <- as.character(df[,1])
   ##Union de BD
   bd.f<-inner_join(df,rec,by = c(names(df)[1], names(df)[2]))
   mindicador<-aggregate(x = bd.f[,4],by = list(bd.f[,3]),FUN = mean)
   indicador<-mindicador[2,2]/mindicador[1,2]

   return(list(indicador = indicador, tabla = mindicador))
 }


seleccionKNBO<-function(datos.sim,productos.antes,productos.despues,ponderado
= TRUE,ordinal = TRUE, reescalar = "rangos",anulacion = TRUE, kvecinos,vector.k,
tolerancia = 0.01,muestra = NULL, factores = NULL){
  vector.k <- sort(vector.k)
  indicas <- rep(NA,length(vector.k))
  if(is.null(muestra) == FALSE){
    idm<-sample(x = datos.sim[,1],size = muestra)
    datos.sim <- datos.sim[datos.sim[,1] %in% idm,]
    productos.antes <- productos.antes[productos.antes[,1] %in% idm,]
    productos.despues <- productos.despues[productos.despues[,1] %in% idm,]
  }
    for(i in seq(vector.k)){
        if(i == 1){
          print(paste("Vamos por el k:",vector.k[i], "inicio a las",Sys.time()))

          prueba      <-           suppressWarnings(pruebaKNBO(datos.sim      =
datos.sim,productos.antes     =     productos.antes,     productos.despues     =
productos.despues,ponderado  =  ponderado,  ordinal  =  ordinal  ,reescalar  =
reescalar,anulacion = anulacion,kvecinos = vector.k[i], factores = factores))
          tabla<-prueba$tabla
          menor <- prueba$indicador
          k<-vector.k[i]
          print(paste("En la iteracion: ",i,"con K:",vector.k[i], " el indicador
es:",round(prueba$indicador,4)))
        } else {
            print(paste("Vamos    por    el    k:",vector.k[i],    "inicio    a
las",Sys.time()))
            prueba     <-          suppressWarnings(pruebaKNBO(datos.sim     =
datos.sim,productos.antes     =     productos.antes,     productos.despues     =
productos.despues,ponderado  =  ponderado,  ordinal  =  ordinal  ,reescalar  =
reescalar,anulacion = anulacion,kvecinos = vector.k[i], factores = factores))
            print(paste("En  la  iteracion:  ",i,"  con  K:  ",vector.k[i],  " el
indicador    es:    ",round(prueba$indicador,4),"    La    diferencia    es    de:
",round(((100*abs(prueba$indicador - menor)/menor)),2),"%" ,sep = ""))
            if((menor    -    prueba$indicador)    >    0       &    (abs(menor    -
prueba$indicador)/menor)  >= tolerancia){

              tabla<-prueba$tabla
```

```
              menor <- prueba$indicador
              k<-vector.k[i]
            }
        }
      indicas[i] <- menor
    }
  df <- cbind.data.frame(vector.k, indicas)
  invisible(gc(reset = TRUE))
  return(list(k = k, indicador.menor = menor,tabla = tabla, indicas = indicas))
}


####### Eficiencia NBO #######
suppressWarnings(suppressMessages(library(dplyr)))
suppressWarnings(suppressMessages(library(tidyr)))

Eficiencia_NBO<-function(rec, productos.despues)
{
  rec[,1] <- as.character(rec[,1])
  productos.despues[,1] <- as.character(productos.despues[,1])
  identificador <- names(rec)[1]
  rec <- rec %>%
    pivot_longer(!all_of(identificador), names_to = "Productos", values_to =
"Prioridad") %>%
    filter(Prioridad != 0)

  ###BD despues transpuesta
  identificador <- names(productos.despues)[1]
  df <- productos.despues %>%
    pivot_longer(!all_of(identificador), names_to = "Productos", values_to =
"Prioridad")

  #df<-gather(data = productos.despues,key = Producto,value = Tenencia, -
identificador)

 # df[,1] <- as.character(df[,1])
  ##Union de BD
  bd.f<-inner_join(df,rec,by = c(names(df)[1], names(df)[2])) %>%
    group_by(Prioridad.x) %>%
    summarise(Promedio = mean(Prioridad.y))
  #mindicador<-aggregate(x = bd.f[,4],by = list(bd.f[,3]),FUN = mean)
  indicador<-bd.f[2,2]/bd.f[1,2]

  return(list(indicador = indicador, tabla = bd.f))
}

####### Curvas ROC #######
library(tidyr)
library(dplyr)
library(pROC)
setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")

rec.NBO <-readRDS(file = "RecomendacionNBOConPCA") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)


Real.prod <-readRDS(file = "DatosAntes201605PROD") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

Original.prod <-readRDS(file = "DatosAntes201505PROD") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

recom <- rec.NBO
#Para primer producto
```

```
rocas <- function(recom, orig, real){
  productos <- colnames(recom)[-1]
  rocs <- rep(NA, length(productos))
  for(i in 1:length(productos)){
    prod <- productos[i]
    print(paste("Vamos por el producto ",prod," " ,
round(100*i/length(productos),1),"%"))
    orig <- Original.prod[,c("ncodpers",prod)]
    clientes.sin.prod <- orig[orig[,prod] == 0,1]
    rec <- recom[,c("ncodpers",prod)] %>%
      filter(ncodpers %in% clientes.sin.prod) %>%
      as.data.frame()
    real <- Real.prod[,c("ncodpers",prod)] %>%
      filter(ncodpers %in% clientes.sin.prod)
    if(sum(real[,prod]) > 0){
      c.roc <- roc(response = real[,prod], predictor =
as.vector(rec[,prod]),direction = "<")
      rocs[i] <- c.roc$auc
    }
  }
  df <- cbind.data.frame(productos, rocs)
  return(list(df = df, promedio = mean(rocs,na.rm = TRUE)))
}


#####  Para NBO  #####
rec.NBO <-readRDS(file = "RecomendacionNBOConPCA") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

nbo <- rocas(recom = rec.NBO, orig = Original.prod, real = Real.prod)
nbo
#####  Para XGBOOST  #####
rec.xg <-readRDS(file = "RecomendacionXGBOOST") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

xgb <- rocas(recom = rec.xg, orig = Original.prod, real = Real.prod)
xgb
#####  Para Recommenderlab  #####
rec.recom <-readRDS(file = "RecomendacionRECOMMENDERLAB") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

recomenderlab <- rocas(recom = rec.recom, orig = Original.prod, real = Real.prod)
recomenderlab

####### Razón de adquirencia promedio #######
library(tidyr)
library(dplyr)
setwd("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Data")

source("C:/Users/Adrian/OneDrive/UCR/Maestría/2021/PPII/Analisis/Algoritmos/Ef
iciencia_NBO.R")

#####  Para NBO  #####
rec.NBO <-readRDS(file = "RecomendacionNBOConPCA") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

nbo <- Eficiencia_NBO(rec = rec.NBO,productos.despues = readRDS(file =
"DatosAntes201605PROD"))
nbo
#####  Para XGBOOST  #####
rec.xg <-readRDS(file = "RecomendacionXGBOOST") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers) %>% as.data.frame()
```

```
xgb <- Eficiencia_NBO(rec = rec.xg,productos.despues = readRDS(file =
"DatosAntes201605PROD"))
xgb
#####  Para Recommenderlab  #####
rec.recom <-readRDS(file = "RecomendacionRECOMMENDERLAB") %>%
  mutate(ncodpers = as.character(ncodpers)) %>%
  arrange(ncodpers)

recomenderlab <- Eficiencia_NBO(rec = rec.recom,productos.despues = readRDS(file
= "DatosAntes201605PROD"))
recomenderlab
```