Tesis para obtener el título de magister en economía

Sistema de recomendación de productos financieros: una aplicación empírica en banca de consumo

Daniel Ricardo Zapata Sanabria

Director: Jorge A. Gallego

Universidad del Rosario
Facultad de economía
Maestría en economía
Bogotá
2019

Sistema de recomendación de productos financieros: una aplicación empírica en banca de consumo*

Daniel Ricardo Zapata Sanabria**

16 de julio de 2019

Resumen

El objetivo de este documento es generar recomendaciones de productos de forma personalizada a los clientes y a su vez campañas comerciales enfocadas en los distintos productos de banca de consumo. Para esto se desarrolla y aplica un sistema de recomendación de productos adecuado a las necesidades prácticas de la banca de consumo. Dicho sistema se encuentra compuesto por un conjunto de modelos de propensión de productos financieros que predice para cada cliente a dos meses los productos más probables a ser adquiridos. De esta forma, el sistema permite desarrollar dos tipos de campañas comerciales: específicas para cada producto o para cada cliente. En las primeras, se busca encontrar a los clientes con mayor propensión de adquirir un producto determinado, mientras que, en las segundas, se quiere encontrar los productos más propensos a ser adquiridos. La aplicabilidad del sistema fue puesta a prueba en una entidad de banca de consumo colombiana. Como resultado, el sistema muestra un alto desempeño de predicción en todos sus productos dadas las restricciones empíricas: un AUC promedio del 87 % en la predicción sobre el mes de pruebas para los ocho productos analizados, alta exhaustividad y baja precisión, dándole importancia a la captura del total de clientes que efectivamente van a adquirir el producto ofrecido y un desempeño superior en el MAP@K comparado con el modelo base de K productos populares. Los resultados son acordes a las necesidades de la entidad desde una perspectiva de aplicación al negocio de banca de consumo, a pesar de que las métricas obtenidas no son comparadas con otros modelos más avanzados que el K populares.

Palabras clave: Personalización, modelos predictivos, Sistemas de recomendación, Banca de consumo.

Clasificación JEL: C38, C45, G21, M31.

---

*Tesis de grado para optar por el título de Magister en Economía de la Universidad del Rosario. Asesor: Jorge Gallego . El autor agradece a Daniel Arocha por sus valiosos comentarios.

**Economista y profesional en Finanzas y comercio internacional, Universidad del Rosario. email: danielr.zapata@urosario.edu.co

1. Introducción

La variedad de productos y servicios ofrecidos a los consumidores están limitados por costos operativos y logísticos tales como espacio, tiempo y transporte. Decisiones como el aumento en la capacidad instalada de comercializadoras, estudios de capacidad de pago de los clientes en entidades financieras, el número de programas diarios en una estación de radio, entre otras, implican para las empresas costos adicionales. Por tal motivo, éstas normalmente se enfocan en la venta de los productos más populares, ofreciendo sólo una variedad limitada de aquellos productos más especializados (Katsov, 2018). Los sistemas de recomendación proponen responder a esta problemática, identificando y sugiriendo los productos que son más propensos a ser comprados por un cliente dado un contexto. De forma general, un sistema de recomendación genera una lista ordenada de productos para ofrecer a cada cliente (Ricci, Rokach, y Shapira, 2011).

La historia de los sistemas de recomendación comienza a mediados de los años noventa, sin embargo su mayor uso se dio con el auge del comercio electrónico y servicios en línea. En estas plataformas los usuarios suelen tener dificultades al momento de elegir un producto, dada la gran variedad de ellos. El problema de la sobre carga de información es resuelto mediante la sugerencia de bienes que terminan limitando el universo de posibilidades. Casos de éxito como *Amazon,* (Linden, Smith, y York, 2003) actualmente calificada como la mayor compañía de comercio en línea y *Netflix* (Bennett y Lanning, 2007), plataforma que brinda el servicio de transmisión de series, documentales y películas, hacen uso continuo de estos sistemas en sus plataformas mejorando la experiencia de sus clientes, incrementando sus ventas y posicionándose dentro de su respectivo mercado.

Un reto importante y común en el sector financiero es la falta y necesidad de sistemas de asesoramiento inteligentes para la toma de decisiones (Cheng, Lu, y Sheu, 2009). Desde la perspectiva del negocio bancario, un sistema de recomendación adaptado a las necesidades y características particulares de la banca de consumo ofrece grandes beneficios al momento de diseñar campañas comerciales[1] para la venta de sus productos y servicios a los clientes. Al automatizar el diseño de estrategias específicas para cada cliente anticipándose al ofrecimiento de servicios útiles para él, se pueden mejorar los indicadores de efectividad de las campañas comerciales o brindar asesorías personalizadas a cada cliente asistiéndolos al momento de tomar la decisión de adquirir un producto (Choudhary y Goila, 2019).

A comparación de los casos exitosos anteriormente nombrados, un sistema de recomendación para el uso de banca de consumo es distinto dadas las características propias del negocio, el tipo de productos y clientes (Chirkin, a, Aleksandra. Diskin, 2018). La

[1] Las campañas comerciales son aquellas acciones generadas por el banco para profundizar o fidelizar un cliente.

base del servicio bancario involucra una relación estrecha con el cliente, que consiste en brindar asesoría en decisiones de inversión, crédito y ahorro en productos financieros. La oferta comercial de los servicios de captación y colocación monetaria puede darse por los denominados canales de venta, permitiendo llegar a más o menos clientes según sus estructuras de costos y restricciones operacionales. Entre los canales se encuentra por ejemplo, oficinas comerciales, fuerza móvil de ventas, telemercadeo y medios digitales.

El objetivo de este documento es generar recomendaciones de productos de forma personalizada a los clientes y a su vez campañas comerciales enfocadas en los distintos productos de banca de consumo. Para esto se desarrolla y aplica un sistema de recomendación de productos adecuado a las necesidades prácticas de la banca de consumo. El sistema es compuesto por un conjunto de modelos de propensión de productos financieros que predice para cada cliente a dos meses los productos más probables a ser adquiridos. De esta forma, el sistema permite desarrollar campañas comerciales de acuerdo a dos estrategias: específicas para cada producto o para cada cliente. En las primeras, se busca encontrar a los clientes con mayor propensión de adquirir un producto determinado, mientras que en las segundas, se quiere encontrar los productos más propensos a ser adquiridos para cada uno de los clientes. La aplicabilidad del modelo fue puesta a prueba en una entidad de banca de consumo colombiana. Las acciones comerciales propuestas son fácilmente ejecutables ya que siguen las recomendaciones empíricas de la aplicación de sistemas de recomendación en banca propuestas por Gorgoglione y Panniello (2011); Choudhary y Goila (2019).

Para probar el desempeño del modelo en campañas comerciales por producto, es decir, la captura de aquellos clientes que efectivamente compraron el producto de interés, se estimó la propensión de los clientes para un mes que no fue usado en la estimación y se evaluaron los resultados mediante métricas estadísticas. Para las campañas comerciales por cliente, el modelo ofrece los k productos más probables para cada cliente en el mismo mes y su resultado fue comparado con la predicción de un modelo base, K productos populares, usando el promedio de la precisión media con k productos, MAP@K [2] por sus siglas en inglés, común en la evaluación de sistemas de recomendación. Finalmente, la estrategia de modelación es cercana a la de Pryor (2016); Van de Wiele (2016) , al usar métodos de clasificación para el cálculo de las propensiones de compra. Es necesario mencionar que no se encontró un documento empírico que permitiera integrar en un único sistema ambos diseños de campañas comerciales que resultan de las propensiones de adquisición estimadas.

Como resultado se observó que el sistema de recomendación de productos financieros obtuvo un AUC promedio del 87 % en los ocho productos de banca de consumo modelados, reflejando la calidad en la predicción de adquisición de los clientes. Los modelos de propensión por producto en general muestran alta exhaustividad y baja precisión, lo

[2]Mean Average Precision at K.

que evidencia la importancia de la captura de los clientes que efectivamente compraran el producto. Respecto a las campañas por cliente, el modelo tuvo un desempeño superior al modelo base de k productos populares, permitiendo el desarrollo efectivo de este tipo de campañas. Finalmente, el modelo logra un sano balance entre un alto desempeño en la captura de la propensión de los clientes y simplicidad, dadas las restricciones tecnológicas y de información de la entidad. Los resultados son acordes a las necesidades de la entidad desde una perspectiva de aplicación al negocio de banca de consumo, a pesar que las métricas obtenidas no son comparadas con otros modelos más avanzados que el K populares.

2. Literatura de los sistemas de recomendación y aplicaciones en banca

Desde el punto de vista de cualquier proveedor de algún bien o servicio, los autores Ricci y cols. (2011) describen cuatro razones por las cuales los sistemas de recomendación son de gran interés: incrementar el número de bienes vendidos, entender mejor las necesidades del cliente, vender bienes más diversos y el aumento de la satisfacción y fidelidad del cliente. La primera razón se desarrolla en la medida en la que el sistema se adapta a las necesidades y deseos del comprador, la segunda se logra al recomendar bienes que serían difíciles de encontrar sin haber sido sugeridos, la tercera se da porque para el sistema es necesario hacer uso de las preferencias de los clientes ya sea al recolectarlas explícitamente o al predecirlas, y finalmente, el aumento tanto en la satisfacción como en la fidelidad del cliente, se produce en el momento en el que éste encuentra atractivas y relevantes las sugerencias de productos, y a su vez, les da una atención especial en tanto el sistema reconoce la antigüedad y las iteraciones pasadas del cliente.

En el sistema, los clientes o *usuarios* interactúan mediante compras, visitas o calificaciones con los productos o servicios denominados *ítems*. Características como la hora y lugar son considerados parte del *contexto*. Adicionalmente, pueden ser aplicadas *restricciones* al sistema como filtros, reglas y tipos de *ítems*. Variables descriptivas como la edad o el género para el caso de los *usuarios* y el precio para el caso de los *ítems*, pueden ser incluidas. Katsov (2018) clasifica los distintos algoritmos en seis categorías desde el punto de vista del tipo de modelo y los datos con los que es estimado, como se puede observar en la figura 1a. Las dos grandes familias son *Filtros Basados en Contenidos*(Pazzani y Billsus, 2007) y *Filtros Colaborativos* (Sarwar, Karypis, Konstan, y Riedl, 2001); los primeros se basan en información como las características de los productos, mientras que los últimos toman por ejemplo los puntajes de los productos dados por los clientes. Ambas ramas pueden usarse ya sea mediante modelos predictivos o métodos heurísticos que normal-

mente buscan la familia, vecindario de productos o clientes similares. Así mismo, existen otro tipo de soluciones que pueden combinar los métodos tradicionales para convertirlos en modelos híbridos (Modeling y Burke, 2014), adicionar información de contexto como temporadas comerciales o recomendaciones para los casos donde no existe suficiente información (Sarwar y cols., 2001). La elección del algoritmo depende del tipo y la disponibilidad de los datos que requiera, su facilidad de automatización e implementación a la tecnología disponible, el grado de personalización y el objetivo de la organización al hacer uso de estos sistemas.

Katsov (2018) realiza una segunda clasificación de los métodos de recomendación desde el punto de vista de sus usos, como se puede ver en la figura 1b. Esta categorización parte de la combinación entre el grado de personalización y el uso de información contextual. Métodos como *Filtros Colaborativos* y *Basados en Contenidos* se ubican principalmente en la intersección entre alto grado de personalización y bajo uso de contexto, dado que realizan recomendaciones, basándose en interacciones históricas como productos adquiridos, puntajes dados a los productos y búsquedas. Estas recomendaciones suelen estar en secciones como *Recomendados para usted* o *Compre de nuevo*. Al tomar en cuenta información contextual como locación, día de la semana, estaciones climáticas o fechas especiales comerciales, encontramos ejemplos como sugerencias del tipo *Restaurantes cerca a usted*. De otra parte, aproximaciones distintas que no son personalizadas, se basan en estadísticas globales y características de los productos y pueden verse en secciones como *Productos más populares* o *Nuevos lanzamientos*. Por último, las recomendaciones no personalizadas también pueden usar información contextual, por ejemplo recomendaciones como *En tendencia en su zona* o *Productos relacionados*; para mayor información consultar Katsov (2018).

2.1. Casos y retos de la aplicación de sistemas de recomendación en temas financieros

Una de las primeras revisiones de literatura del uso de sistemas de recomendación en temas financieros fue realizada por Zibriczky (2016), quien lista casos en temas como banca en línea, préstamos, seguros, finca raíz, acciones y portafolios. Varios de los siguientes trabajos se encuentran enunciados en dicha revisión de literatura. Un ejemplo de recomendaciones basadas en contexto para una aplicación móvil de un banco fue desarrollado por Vico y Huecas (2012) y su objetivo fue recomendar lugares como restaurantes o supermercados donde clientes similares hayan comprado con su tarjeta de crédito, se resalta el uso de los datos bancarios que permiten generar recomendaciones basadas en acciones reales de las personas y no calificaciones subjetivas; 100 usuarios fueron encuestados en una prueba piloto, quienes en general encontraron útil y efectivo el sistema. Adicional a lo

anterior, existen aplicaciones en banca de inversión como la realizada por Gigli, Filopanti, y Regoli (2017) quienes utilizan distintos algoritmos de recomendación para un conjunto de datos de alrededor de 200 mil inversionistas y un millón de transacciones en distintos tipos de productos de inversión para 12 meses. El sistema es construido a partir de la retroalimentación implícita (*implicit feedbacks*), que es la relación binaria de compra para cada cliente con cada uno de los productos (Hu, Koren, y Volinsky, 2008). Los autores prueban tres algoritmos: clasificación personalizada bayesiana, mínimos cuadrados alternantes y una adaptación del algoritmo *Word2Vec*. Éstos, a su vez, son comparados frente a la recomendación de productos más populares por los usuarios y el número de transacciones, empleando distintas métricas de evaluación: precisión media del usuario, ordenamiento de percentiles esperados y área bajo la curva ROC. Con el objetivo de probar que el sistema sugiera productos relevantes y específicos al usuario y expanda el portafolio de este, fueron eliminados de la base de prueba 0, 20 y 50 de los productos más populares. En general, los tres algoritmos muestran un sano desempeño, siendo la clasificación bayesiana el mejor.

Cabe resaltar que la adopción de estos sistemas al contexto de productos financieros conlleva retos desde el mismo diseño. Chirkin, a, Aleksandra. Diskin (2018) contrastan las principales diferencias de los datos con los que se alimenta un sistema de recomendación en comercio electrónico y banca. La primera es la retroalimentación explicita de las preferencias de los usuarios por los productos financieros; la segunda es el cambio a través del tiempo de las características de dichos productos, que en el caso del comercio son constantes; y la tercera es el impacto de la compra para el cliente. En el caso del comercio, una mala recomendación no tiene efectos en el largo plazo, mientras que un producto financiero debe ser transparente, robusto y ajustarse a las necesidades de portafolio del cliente.

Los autores anteriormente expuestos permiten identificar las restricciones asociadas con el negocio bancario al momento de implantar un sistema de recomendación: el *manejo del riesgo* característico de la banca, la *retroalimientación implícita* asociada a la construcción de los datos necesarios para el sistema de recomendación (Hu y cols., 2008; Gigli y cols., 2017) y las restricciones asociadas a las características de los productos y clientes. Con respecto al primero, los productos crediticios implican, por su naturaleza, estudios de capacidad de pago para el adecuado manejo de riesgo, por lo que los clientes son previamente evaluados y no todos son candidatos a adquirir alguno de los productos independientemente de sus preferencias. Ésta característica es una de las principales restricciones aplicadas a un sistema de recomendación en la práctica. Por otra parte, respecto al segundo, no se cuenta con el grado de afinidad del cliente con los productos que posee o con los que no, a diferencia de las calificaciones o puntajes que se les preguntan a los usuarios en el comercio electrónico. Por ejemplo, cuando se observa que un cliente no posee un producto puede ser indicativo de falta de interés en él, simple desconocimiento

o en el caso de los productos crediticios, falta de capacidad de endeudamiento. Del mismo modo, observar la sola tenencia de un producto no permite determinar la afinidad del cliente con el mismo, por lo que suelen asumirse simplemente relaciones binarias entre los clientes respecto a la tenencia con cada uno de los productos, lo que denominamos retroalimentación implícita (Hu y cols., 2008).

Por el lado de los productos, estos están diseñados para cumplir necesidades específicas, por lo que cada uno tiene características altamente diferenciadas y en algunos casos pueden ser complementarios, como también sustitutos entre ellos. Además, algunos productos son de uso continuo, mientras que otros suelen ser adquiridos pocas veces por el cliente a lo largo de su vida. Finalmente, debido a la naturaleza de algunos productos, sus condiciones pueden variar a través del tiempo o requerir un compromiso de mediano o largo plazo y consecuentemente, la utilidad de algunos productos nos es inmediata (Chirkin, a, Aleksandra. Diskin, 2018). Para los clientes, la elección de un producto financiero no es una tarea sencilla, por lo que suelen formular expectativas más rigurosas para la adquisición de los mismos y requerir el conocimiento de un experto para una buena elección (Zibriczky, 2016). Así pues, como resultado éstos normalmente tienen pocos productos y raramente adquieren más.

2.1.1. Sistemas de recomendación de productos para la banca de consumo

Esta sección esta enfocada en explicar la línea base para el desarrollo del sistema de recomendación de productos financieros. Desde la perspectiva empírica de Choudhary y Goila (2019); Gorgoglione y Panniello (2011) se enuncian puntos estratégicos a tener en cuenta en el diseño e implementación de un sistema de recomendación en una entidad bancaria. Por otro lado, Van de Wiele (2016); Pryor (2016) permiten comprender el componente técnico requerido para elaborar los modelos de propensión como insumo para un sistema de recomendación.

Para comenzar, existen ejemplos de cómo sistemas de recomendación o a veces llamados *modelos de siguiente mejor oferta* no solo pueden incrementar la fidelidad de los clientes en banca de consumo, sino que además disminuye la fuga de los mismos. Choudhary y Goila (2019) enuncian medidas claves para asegurar la exitosa implementación de este tipo de modelos: identificar los productos, servicios y oferta, horizonte temporal a ser analizado, evaluación apropiada de los modelos, elaboración de pruebas piloto y evaluación de los impactos mediante grupos de control y tratamiento.

Por ejemplo, un caso aplicado para generan acciones personalizadas para los clientes con alta propensión de fuga es presentado por Gorgoglione y Panniello (2011). En el proceso, los autores evalúan distintas metodologías a partir de seis puntos que consideran necesarios para garantizar el éxito del sistema de recomendación en la organización.

El primer punto es el control en el proceso de decisión, que surge como la necesidad de centralizar lo que le es ofrecido a cada cliente. Segundo, la automatización en el proceso de personalización, la cual le permite a la entidad mantener costos bajos y a su vez que los procesos de manejo de relación con el cliente o *CRM* [3] por sus siglas en inglés, sean ágiles. Tercero, la efectividad en captura de los clientes, lo que quiere decir, que se maximice la probabilidad que un cliente reciba una oferta acorde a sus necesidades. Cuarto, la eficiencia en el proceso que refleja la necesidad de mantener los costos de los programas de CRM tan bajos como sean posibles. Quinto, el alcance de las acciones comerciales, el cual depende de la creatividad de las acciones a ser aplicadas a los clientes. Por último, la cobertura de los clientes, la cual consiste en aplicar una acciones personalizadas al total de clientes de la entidad.

De modo que, Gorgoglione y Panniello (2011) en el análisis encuentran que los sistemas de recomendación basados en similitud de clientes permiten cumplir cuatro de los cinco requerimientos, argumentando que dada la naturaleza de los datos con los que es estimado el modelo, las únicas acciones posibles dependerán del mismo conocimiento que ya se tenga de los clientes. En el caso aplicado, estiman un modelo de filtros colaborativos tomando solo aquellos clientes que consideran fieles para generar recomendaciones de productos a aquellos clientes con alta probabilidad de fuga. Al aplicar la recomendación resultante, encontraron que 1313 de 1609 clientes, es decir, el 81.6 %, respondieron positivamente y no cancelaron sus productos, reduciendo la tasa de fuga promedio.

Por otra parte, en términos de aplicabilidad de estas herramientas en banca de consumo, se destaca el concurso de alcance global propuesto por el Banco de Santander en la búsqueda de un sistema de recomendación de productos (Banco Santander, 2016). La razón de esta convocatoria parte de la necesidad de identificar los productos potenciales que van a adquirir los clientes en los meses siguientes, contando con la información de su portafolio y características socio demográficas. Para evaluar las diferentes propuestas recurrieron a la medida promedio de la precisión media hasta el séptimo producto recomendado (MAP@7).

En general, aunque se haya visto diversidad de técnicas en las soluciones, éstas pueden clasificarse entre aquellas que desarrollaron un modelo único para cada producto y aquellas que intentan predecir al tiempo el conjunto de productos. Al final, los modelos de propensión de adquisición para cada producto como los desarrollados por Van de Wiele (2016); Pryor (2016) ocuparon los primeros lugares. No obstante, las metodologías con base en modelos latentes y filtros colaborativos estuvieron dentro del 5 % superior. Cabe aclarar que la metodología de estimación de propensión de adquisición de Van de Wiele (2016); Pryor (2016) se encuentra basada en árboles de clasificación con potenciación de gradiente.

[3] Customer Relationship Management.

Knott y Neslin (2002) evalúan la efectividad de modelos de propensión o next producto to buy en ventas cruzadas de productos en banca de consumo. En el documento discuten el proceso de elaboración de modelos al igual que su aplicación teórica y práctica. Además, ponen a prueba el modelo en campo, como resultado muestran que el modelo incrementa los beneficios comparado con la metodología heurística. Así mismo se hace un test empírico de la metodología y como resultado encuentran que las variables de tenencias de productos son las que mejoran más la exactitud de las predicciones seguidas por las variables que representan el valor monetario del cliente ye sus respectivas variables demográficas. Los autores ponen a prueba distintos métodos como regresión logística, análisis discriminante y red neuronales encontrando que este último fue el mejor, sin embargo, no hubo gran diferencia en la exactitud de las predicciones por los métodos anteriormente nombrados. Por otro lado, usaron dos formas de muestreo una aleatoria y otra estratificada, encontrando que la primera es mejor para aumentar la exactitud de los modelos aunque resaltan que la segunda es deseable para no subestimar la predicción de productos no populares. Los modelos de propensión elaborados en este documento utilizan la matriz cliente producto y las variables demográficas, a su vez los modelos se estiman utilizando una muestra aleatoria siguiendo las recomendaciones prácticas de Knott y Neslin (2002).

3. Matriz cliente - producto y variables socio demográficas

Los datos con los que es alimentado el sistema de recomendación son provistos por una entidad bancaria colombiana que cuenta con presencia a nivel nacional. Se considera una muestra mensual de la información socio demográfica y del portafolio de productos de cada cliente perteneciente al segmento de banca de consumo. Los productos analizados son descritos en el cuadro 6. El horizonte temporal es desde junio de 2017 a diciembre de 2018. Para cada mes existen 30.000 clientes en promedio. Hay un total de 570.929 registros de clientes y 35.338 clientes en total que representan un 10 % del total de clientes de la entidad.

La información mensual del portafolio de los clientes es conocida como la matriz cliente producto; esta matriz refleja el número de unidades de cada producto que posee cada cliente. El conjunto de productos ofrecidos por el banco no cambia en el horizonte temporal analizado. Sea U el total de clientes en el segmento de banca de consumo en el mes t y D el total de productos que el banco ofrece para el segmento de consumo. R_t es entonces la matriz cliente producto para un mes t de dimensión $|U| \times |D|$ donde cada elemento $r_{ijt} \in 0 \cup \mathbb{N}^+$en el mes t. Por ejemplo $r_{ijt} = 2$ indica que el cliente i posee dos unidades del producto j en el mes t. Todos los clientes $i \in U$ poseen al menos un producto $j \in D$. Para

todos aquellos productos $j \in D$ que el cliente $i \in U$ no posee, $r_{ijt} = 0$.

$$R_t = \begin{bmatrix} r_{11t} & \dots & r_{1jt} \\ \vdots & \ddots & \vdots \\ r_{i1t} & \dots & r_{ijt} \end{bmatrix} \tag{1}$$

A partir de la evolución histórica de la matriz R_t se puede analizar la relación entre los clientes y los productos del banco a lo largo del horizonte temporal dado. La figura 2a muestra la evolución del número de unidades de cada producto y la figura 2b enseña el número de unidades de los productos para el mes de diciembre; ambas figuras permiten observar una relación entre los productos más populares y los más especializados, la demanda por productos especializados existe y crea una cola larga en el histograma. Por parte de la entidad existe un interés para que sus clientes diversifiquen sus productos con la misma. Por esta razón, los créditos como vivienda hacen parte de la estrategia comercial, aunque requieren un mayor nivel de estudio de capacidad de endeudamiento y en total existan menos que otros productos.

Las figuras 2c y 2d exponen el total de unidades de productos por cliente en los periodos analizados y en diciembre respectivamente. Dado que los clientes suelen formular expectativas más rigurosas sobre la adquisición de un producto, no solemos observarlos con bastantes unidades de productos en su portafolio financiero (Zibriczky, 2016). Este fenómeno junto con el problema de las colas largas es una de las principales razones para implementar un sistema de recomendación en la entidad bancaria.

Las variables socio demográficas de los clientes, de ahora en adelante $Demográficas_t$, son descritas en el cuadro 7 junto con las figuras de la 3a a la 3j. Éstas muestran que la mayoría de los clientes tienen una edad promedio mayor a los 50 años, se encuentran entre los estratos socio económicos dos y tres, tienen educación secundaria, son solteros, poseen una vivienda familiar, son empleados privados que viven en Bogotá y el promedio de antigüedad de su relación con el banco está entre los 200 y 300 meses. Por último, en las variables financieras como ingresos, egresos, activos y pasivos, que permiten conocer al momento de ofrecer un producto crediticio si el cliente es apto o no, existen correlaciones positivas interesantes como entre el nivel de ingresos y egresos o entre activos e ingresos.

4. Metodología

En esta sección se describe la estrategia de modelación con la que fueron construidos los modelos de propensión de adquisición de productos financieros que componen el sistema de recomendación. En adición, es descrito el modelo de k productos populares,

que será usado en la sección de resultados como modelo de referencia o ingenuo frente al anterior.

4.1. Modelos de propensión de adquisición de productos financieros

La estrategia de modelación consiste en la elaboración de un modelo base de clasificación, para cada producto, que estima la probabilidad de adquirir una unidad adicional de este en dos meses para cada cliente, condicional a su portafolio actual, características socio demográficas y demás variables útiles que se expondrán más adelante. La razón para la estimación por separado de modelos para cada producto es debido a que facilita las campañas enfocadas a estos. Además, esta estrategia es similar a las propuestas por Pryor (2016); Van de Wiele (2016) quienes obtuvieron mejores resultados de esta forma que al predecir la probabilidad de compra de múltiples productos a la vez en la competición del Banco Santander descrita en la sección 2.1.1.

Para la elaboración del modelo además de la matriz cliente - producto R_t y las variables $Demográficas_t$ expuestas en la sección 3, fueron calculadas variables adicionales a partir de la matriz R_t para cada producto, que denominaremos en conjunto como A_t. Estas variables adicionales también fueron incluidas en el modelo desarrollado por Pryor (2016) con el objetivo de mejorar el poder de predicción del modelo. La primera variable corresponde al número de meses desde la adquisición del producto [4], la segunda variable es el total de productos del portafolio en cada mes, la tercera variable es el número de veces que el cliente ha añadido cada producto al portafolio. La ultima variable es el total de transacciones como el número de veces que un cliente ha añadido o cancelado un producto. Estas variables adicionales son descritas en el cuadro 7.

Finalmente, es posible obtener los datos de entrenamiento, validación y pruebas para la estimación del modelo de propensión de adquisición de productos financieros. Para ello, como se enunció anteriormente, construiremos un modelo base para cada producto $j \in D$, siendo D el total de productos que el banco ofrece para el segmento de consumo. Sea el mes $t \in T$, donde T es el total de meses de la muestra. De esta forma podemos definir y_{ijt+2} como el evento de adquisición de producto j para el mes $t+2$ dada la información del cliente i en el mes t para el modelo base del producto j.

$$y_{ijt+2} = \begin{cases} 1 & Si\ r_{ijt+2} > r_{ijt} \\ 0 & en\ otro\ caso \end{cases} \tag{2}$$

Ahora bien podemos definir la estimación de $\hat{y}_{ijt}$ como:

$$\hat{y}_{ijt+2} \sim f(R_{it}, r_{ijt-1}, Demográficas_{it}, A_{ijt}) \tag{3}$$

[4]Debido a que el mes más antiguo de la matriz cliente – producto es junio de 2017, para algunos clientes no es posible identificar el mes en el que adquieren el producto, por lo que se usa un valor de 99.

Nótese que para la estimación de $\hat{y}_{ijt+2}$ se toman las siguientes variables predictoras: R_{it} como el portafolio del cliente i en el mes t; r_{ijt-1} como el rezago de las unidades del producto j del cliente i en el mes $t-1$; A_{ijt} como el conjunto de variables adicionales propias del producto j como. Estas variables incluyen el número de meses desde la adquisición del producto hasta el mes t, el número de veces que el cliente i ha añadido el producto j hasta el mes t y las adicionales propias del cliente como el total de productos en su portafolio en el mes t y el número de veces que el cliente i ha añadido productos hasta el mes t.

En total la muestra cuenta con 19 meses desde junio de 2017 a diciembre de 2018. Debido a que el modelo predice a dos meses y que se usa un mes como rezago en la predicción, se cuenta entonces con 16 meses, o $T = 16$. De esta forma la muestra se dividió bajo cortes temporales de la siguiente forma: los datos de entrenamiento tomaron los meses de julio de 2017 a agosto de 2018, los de validación usaron el mes de septiembre 2018 y finalmente los datos de prueba se efectuaron en el mes de octubre de 2018 con el fin de predecir el mes de diciembre. Las figuras 4a a la 4h muestran el porcentaje de ocurrencia de la adquisición para cada producto, Y_t, respecto al total de clientes en cada mes.

4.2. Elección de los mejores modelos de propensión para cada producto

Para la elección del mejor modelo base o de propensión para cada producto se realizó el siguiente procedimiento: preparación de datos, competición de modelos, análisis de variables importantes, calibración de parámetros y estimación del modelo final.

Para comenzar la preparación de datos, se encontraron los registros inconsistentes o vacíos en cada variable. Para las variables financieras como ingresos mensuales, egresos mensuales, activos y pasivos, se imputó la mediana de la variable resultante del cálculo de la combinación entre departamento y estrato. También aquellos registros por encima del percentil 99 fueron imputados por el valor de percentil 99. Por su parte, en las variables categóricas con registros vacíos se imputó la categoría más recurrente para el caso en que eran escasos. En el caso contrario, cuando el número de estos registros eran comparables en tamaño con alguna otra categoría, se decidió etiquetarlos como una categoría adicional asignándoles el valor de *desconocidos*. Después, las variables categóricas fueron codificadas por valores binarios eliminando en el proceso la variable de una de las categorías. Por último, se realizó un análisis gráfico de la relación entre el evento de compra del producto j , Y_t y las variables predictivas con el fin de observar qué categorías concentraban la mayor proporción de compradores para cada producto j , o si existían diferencias de medianas para el caso de las variables financieras.

Para la competición de modelos se tomaron las métricas de *AUC*, *Precisión* y *exhaus-*

tivdad[5] y tres modelos de clasificación [6] tomando los parámetros por defecto de los paquetes estadísticos h2o de LeDell y cols. (2019) y *xgboost* de Chen y cols. (2018) en el software estadístico R (R Core Team, 2017). Los tres primeros modelos son *Logit*, *Árboles de decisión* y *Árboles de clasificación con potenciación de gradiente* usando h2o. Para el modelo de árboles de decisión y el primero de potenciación de gradiente se estimaron hasta 500 y 1000 árboles respectivamente y 10 rondas de interrupción temprana, según como mejoraba tras cada árbol la métrica de AUC en los datos de validación. De forma similar, mediante el uso del paquete *xgboost* se estimaron dos árboles de clasificación con potenciación de gradiente estimando hasta 500 árboles y 30 rondas de interrupción temprana , según como mejoraba tras cada árbol la métrica de AUC y AUCPR en los datos de validación. Justo después que fueron estimados los cinco modelos, se eligió el modelo con mayor AUC y mayor área bajo la curva resultante entre las métricas precisión y exhaustividad obtenida a partir de la variación del umbral de probabilidad.

Una vez se eligió el mejor modelo se realiza un análisis de variables importantes. En este análisis se busca entender mejor qué variables considera más relevantes el modelo y cómo los distintos valores de estas variables afectan la predicción. En el primer caso, en cada modelo podemos calcular distintas métricas. Por ejemplo en el modelo Logit es posible analizar los valores de los β resultantes. Por su parte, para los modelos basados en árboles es posible observar el Gini resultante tras cada partición, como también el resultado de la función de ganancia, la frecuencia y el número de observaciones de cada variable[7]. En el segundo caso, es posible usar los gráficos de dependencia parcial, que permiten observar el efecto marginal de una variable en la predicción de un modelo ya entrenado [8]. La relación entre $\hat{Y}_{jt}$ y alguna variable continua puede ser lineal, monótona o más compleja. También podemos comparar si cada modelo responde de forma similar o distinta. La función de dependencia parcial representa la predicción promedio cuando toda la muestra es forzada a tomar un valor único de una variable sin manipular el resto de valores de las demás variables. Sin embargo, en el proceso existen desventajas como asumir independencia entre las variables predictoras y la existencia de efectos heterogéneos ocultos al promediar las predicciones para cada valor de la variable (Molnar y cols., 2018). Para obtener los gráficos de dependencia parcial se usó el paquete estadístico Biecek (2018).

A continuación, se buscó la mejor combinación de hiper-parámetros que resulten en el modelo con la mejor predicción posible en los datos de validación, disminuyan el sobre

[5]Las métricas son definidas en la sección 8 en el apéndice

[6]La explicación teórica de estos modelos se encuentra en las secciones7.3, 7.4, 7.5 y 7.6 del apéndice.

[7]La importancia de variables solo se midió para los modelos estimados de bosques aleatorios y su metodología de medición se encuentra descrita en la sección 7.7 del apéndice.

[8]Los gráficos de dependencia parcial solo fueron calculados para los modelos de bosques aleatorios y arboles de clasificación.

ajuste, reduzcan la varianza de los errores de predicción y manejen variables predictoras correlacionadas. En el caso del modelo Logit se usaron los métodos de regularización *ridge* y *lasso*. Para los bosques aleatorios se varió la profundidad de los árboles entre 4, 5, 10 y 20 nodos y la tasa de la muestra tomada aleatoriamente por cada árbol entre 63 % y 80 %. En el caso de los árboles de clasificación con potenciación de gradiente se varió la profundidad de los árboles entre 4, 6 y 12 nodos y la tasa de aprendizaje η entre 0.1, 0.3 y 0.8 [9]. Una vez se encontró la mejor combinación de hiper-parámetros que resultó con la mejor métrica AUC en los datos de validación, se estimó un modelo final usando los hiper-parámetros encontrados (puntualmente para el caso de los métodos basados en árboles se usó la mejor iteración encontrada) y tomando los datos de validación y entrenamiento de manera unificada.

4.3. Diseño del sistema de recomendación

Para empezar, la estrategia de modelación cumple con los requerimientos necesarios para garantizar el éxito de la aplicación de un sistema de recomendación en banca de consumo enunciados por Choudhary y Goila (2019); Gorgoglione y Panniello (2011). En primer lugar, las probabilidades obtenidas permiten dos formas de dirigir campañas comerciales para el banco, lo que permite definir las acciones posibles y poseer control sobre el proceso de decisión. La campaña por producto toma los resultados del modelo base del producto analizado, los filtra y ordena por aquellos clientes con las probabilidades más altas de adquirir dicho producto. La campaña por cliente es posible al recomendar los k productos más propensos para cada cliente al ordenar sus probabilidades para cada producto. En segundo lugar, el modelo fue elaborado bajo una visión temporal mensual para ser estimado de forma automática, por lo que minimiza la intervención humana en la obtención de resultados. A su vez, es eficiente para los procesos de CRM ya que se obtiene la información más actualizada de los clientes cada mes. Tercero, las métricas de desempeño permitieron evaluar la efectividad del modelo en la captura de los clientes. Por último, se garantizó la cobertura al estimar las probabilidades de adquisición para todos los clientes activos que cuentan con al menos un producto dentro de la muestra.

Por otra parte, la elección de modelos de propensión sobre otro tipo de modelos en la literatura de sistemas recomendación, como los filtros colaborativos, radica en que permiten realizar campañas por productos, donde el orden de los clientes a los cuales se les ofrece el producto seleccionado es fundamental. La razón es que al ofrecerles primero el

[9] La tasa de aprendizaje relaciona la disyuntiva entre aprendizaje rápido versus estable. Una tasa cerca a uno permite un rápido aprendizaje pero inestable. Al contrario, una tasa cerca a cero permite un aprendizaje lento pero estable. Después de cada árbol estimado, estos son multiplicados por η para ser añadido al ensamblaje. Véase Chen y cols. (2018).

producto a los clientes más propensos, una vez ordenados, es posible llevar más productos a buen término durante el inicio de la campañas comerciales [10]. Esto permite aumentar la efectividad global de la misma, debido a que se cuenta con más tiempo en términos logísticos para el otorgamiento del producto. Por último, el ordenamiento de los clientes junto con los resultados del uplift, expuesto más adelante, permiten analizar la relación de costo y efectividad al seleccionar el número de clientes en una campaña que está sujetas a restricciones de presupuesto.

En suma, la definición de las campañas por producto como acción posible resultante del modelo, permitió dar prioridad al uso de modelos únicos para cada producto sobre los modelos de clases múltiples, que intentan predecir al tiempo el conjunto de productos para cada cliente. Finalmente, el horizonte temporal de predicción, dos meses, también fue elegido debido a que los datos necesarios para el modelo en un mes se originan en el mes siguiente y porque durante el mes intermedio en que el modelo realiza la estimación de compra en el mes siguiente. También en el mes intermedio se diseñan las campañas, permitiendo mejorar la eficiencia del proceso de estas al obtener la información más actualizada de los clientes.

4.4. K productos más populares

Siguiendo a Cremonesi, Milano, Koren, y Turrin (2015); Katsov (2018) este modelo consiste en recomendar los K productos más populares, definidos como aquellos que son adquiridos por más clientes. Puede recomendarse la lista de productos populares o aquellos que el cliente *i* no posea en su portafolio actual. La salida de este algoritmo es entonces simplemente una lista de productos recomendados ordenados por popularidad. Este modelo pertenece a la familia de modelos de baja personalización y sin contexto, el cual es finalmente usado como el modelo base o ingenuo para la comparación con el sistema de recomendación de productos financieros.

4.5. Métricas de desempeño

Las métricas elegidas para la evaluación adecuada del rendimiento de los modelos de propensión de adquisición de productos, así como el sistema de recomendación, son el uplift, exactitud, precisión, exhaustividad, AUC y el MAP@K. La primera medida, uplift, asiste en el diseño de las campañas por producto dado que permite medir un porcentaje de captura sobre un grupo de clientes seleccionados en base al ordenamiento de las probabilidades predichas. Las métricas como la exactitud, precisión y exhaustividad son medidas basadas en la captura de la respuesta de interés sobre distintos conjuntos, por lo

[10] Las campañas comerciales suelen durar entre tres y cuatro meses para la entidad bancaria.

que es necesario convertir las probabilidades predichas en resultados binarios. Para ello se usó el área bajo la curva ROC para obtener el umbral óptimo y dar una idea general del desempeño de los modelos base. Sin embargo, estas medidas no permiten medir sobre predicciones, donde es importante el orden de la recomendación, que es el propósito del diseño de campañas por persona. Para ello se tomó el promedio de la precisión media en k productos, MAP@K por sus siglas en inglés, que permite obtener una única métrica que expresa el desempeño general de todas las recomendaciones. Las métricas se encuentran descritas con más detalle en la sección 8 en el apéndice.

5. Resultados

Tomando los datos descritos en la sección 3 y formando los modelos de propensión descritos en la sección 4.1 fue posible elegir los mejores modelos base para cada producto a partir la metodología descrita en la sección 4.2. Como resultado se encontró que en 6 de los 8 productos los modelos de bosques aleatorios tuvieron las mejores métricas AUC, así como mayor área bajo la curva precisión-exhaustividad en los datos de validación como se puede observar en las figuras 7a a la 4h y 8a a la 8h. Por esta razón, fueron elegidos los modelos basados en bosques aleatorios del paquete *h20* (LeDell y cols., 2019) como ganadores para los productos de cuentas de Ahorros y Nómina, CDT, Crediservice, Tarjeta de crédito y Créditos de Libre destino. Por su parte, los modelos de árboles de clasificación usando el paquete *xgboost* (Chen y cols., 2018) y con la métrica AUC como objetivo, obtuvieron los mejores resultados en los productos de Crédito de Libranza y Vivienda.

Por otra parte, las figuras 5a a la 5h exhiben las variables más importantes para los modelos de bosques aleatorios estimados para los productos. Observamos como en general la variable de antigüedad fue la más importante, pero también otras variables como edad, ingresos mensuales, número de transacciones, activos y variables propias del producto fueron relevantes en los modelos. Al identificar que la variable antigüedad fue la más importante en todos los modelos de bosques aleatorios, se realizó la gráfica de dependencia parcial para dicha variable.

Las figuras 6a a la 6h muestran las gráficas de dependencia parcial de la antigüedad para los modelos de bosques aleatorios y árboles de clasificación con potenciador de gradiente. Se observó cómo la variable antigüedad mostró relaciones no lineales con la predicción de adquisición de los productos, lo que es propio de los modelos no lineales que se están comparando. También la relación es distinta según el modelo. En general, el modelo de bosques aleatorios exhibe una relación en forma de U, lo que sugiere que en los primeros meses que un cliente ingresa al banco al adquirir un producto, es poco probable que adquiera un segundo producto. Sin embargo, según el producto a partir de un número

de meses la probabilidad comienza a aumentar. Por su parte, los modelos de potenciación de gradiente no mostraron patrones similares en los productos e incluso exhibieron comportamientos menos suaves que el modelo de bosques aleatorios, lo que permite intuir que este último no se sobre ajustó a los datos de entrenamiento. Por otro lado, los resultados de la calibración de los hiper-parámetros de los modelos son descritos en las tablas 8 y 9 en los que resalta cómo el AUC mejoró para todos los productos en los datos de validación.

Finalmente, para poner a prueba el sistema de recomendación de producto financieros se toma la información de 29.437 clientes con corte a octubre de 2018 y se predice cuales productos son más propensos a ser adquiridos en dos meses, es decir, en diciembre de 2018. Se considera que un producto j es adquirido si $r_{ijt} < r_{ijt+2}$ [11]. El mes de diciembre de 2018 es el mes de prueba, por lo que la matriz $R_{Diciembre\ 2018}$ no es incluida al momento de estimar el modelo. Para evaluar la bondad del sistema para las campañas por producto se calcularon las métricas de exactitud, precisión, exhaustividad, AUC y Uplift en las predicciones para diciembre de 2018. Para las campañas por cliente, en las que se desea generar una recomendación de K productos a cada cliente, se comparó el sistema propuesto frente al modelo de k productos populares mediante las métricas de exactitud, precisión, exhaustividad y MAP@K.

5.1. Campañas por producto

El cuadro 1 muestra las distintas métricas para las predicciones de adquisición de cada producto durante el mes de diciembre de 2018. El producto con mejores indicadores es el Crédito de Vivienda, a excepción de la precisión y la exactitud. El modelo con mejor exactitud es Libranza y el de mejor precisión es Tarjeta de Crédito. El modelo con menor desempeño es el de Cuentas de Ahorro.

En todos los productos se observó una alta exhaustividad, pero baja precisión. Esta es una característica común en las aplicaciones de algoritmos de recomendación y búsqueda, donde existe una disyuntiva entre modelos de alta precisión y baja exhaustividad o alta exhaustividad y baja precisión (Katsov, 2018). Estos resultados no son sorpresivos, ya que hay un bajo número de clientes que no agregaron productos, por lo que hay un alto número de falsos positivos. Sin embargo, el objetivo es capturar a todos aquellos clientes que efectivamente comprarán el producto, lo que significa capturar la mayoría de positivos, por lo que se prefirió un modelo de alta exhaustividad y baja precisión. Lo anterior es debido a que en reuniones con áreas de la entidad bancaria interesadas en el modelo, se identificó que es más costoso no ofrecerle un producto a un cliente que piensa adquirirlo, que ofrecérselo a alguien que no lo desea. De manera similar Gorgoglione y Panniello (2011) desarrollan en su aplicación dos modelos de fuga de clientes y la entidad bancaria

[11] El número de unidades de cada producto para cada cliente se observa a con corte al final de cada mes.

elige el modelo con la exhaustividad más alta. La razón está basada en que el costo de un falso negativo era mucho más alto que el de un falso positivo.

El AUC o área bajo la curva ROC fue alta para todos los modelos, lo que indica que es más probable que los modelos base predigan una probabilidad alta para algún cliente que haya adquirido el producto tomado de forma aleatoria que para un cliente que no haya adquirido el producto. El AUC más alto es de 0.94 para Vivienda, a comparación de los más bajos: Tarjetas de Crédito y Crédito de Libre destino con 0.83, mientras que una predicción aleatoria se tiene un AUC de 0.5 (Google developers, 2019). Gigli y cols. (2017) logra un AUC entre 0.91 y 0.97 en el sistema de recomendación para banca de inversión.

Por último, el uplift para el producto de Crédito de Vivienda, indica que al tomar el 10 % de los clientes con las probabilidades predichas más altas y contrastando si adquirieron o no el producto en diciembre de 2018, es 8.34 veces más probable encontrar a un cliente que haya adquirido el producto comparado con un cliente tomado al azar. El Uplift más bajo en el 10 % fue de 5.85 para el producto de Tarjeta de Crédito.

5.2. Campañas por cliente

La tabla de 2 muestra el MAP@K para todos los posibles niveles de k en ambos modelos. Como fue expuesto, el MAP@K es el promedio de la precisión media calculada para cada cliente en los K productos adquiridos por el mismo. Por definición, el MAP@K depende del parámetro K y los k productos adquiridos por el cliente, por lo que no es comparable entre distintos valores de K. Cuando un cliente no agrega ningún producto, como la mayoría, su precisión es cero, afectando el promedio. La tabla 3 nos muestra el número de personas que adquirieron productos para el mes de prueba. Del total de clientes en diciembre de 2018, solo 834 clientes añadieron al menos un producto a su portafolio observado en octubre, lo que representa el 2.8 % del total de la muestra. Si la recomendación fuese perfecta para aquellos clientes que agregaron al menos un producto, el MAP@K para esta muestra sería de 0.0283. Se observó que en general el modelo de propensión de adquisición de productos supera en cada nivel de k al modelo de K productos populares y se acerca al MAP@K perfecto.

De forma similar, el modelo propuesto mostró un mejor desempeño en las métricas de exactitud, precisión y exhaustividad al recomendar distintos niveles de K productos populares. No es sorpresa que al recomendar todos los productos faltantes en el portafolio del cliente, K = 8, la exhaustividad esté cerca de 1.

5.3. Discusión de resultados

Los resultados técnicos mostrados en las tablas 1, 2 y 4 reflejan que el modelo de recomendación de productos financieros cumplió con los requerimientos de cobertura y efectividad de captura de los clientes, enunciados por Choudhary y Goila (2019); Gorgoglione y Panniello (2011) y discutidos en la sección 2.1.1 y 4.2. En primer lugar, el sistema de recomendación propuesto generó como resultado probabilidades de adquisición de los productos para todos los clientes permitiendo así ambas campañas nombradas anteriormente. En segundo lugar, en promedio los modelos por producto lograron recoger el 87 % de las adquisiciones realizadas por los clientes, promediando los AUC, lo que demostró la calidad de las predicciones para identificar productos a recomendar a cada cliente. En tercer lugar, los modelos por producto mostraron alta exhaustividad y baja precisión, lo que indicó la importancia de la captura de los clientes que efectivamente compraran el producto (casos positivos). Por último, el modelo fue superior al modelo de K productos populares al dirigir campañas por cliente.

Es preciso resaltar que el modelo no fue comparado con técnicas más avanzadas de la literatura de sistemas de recomendación más allá del modelo base K populares. Sin embargo, al comparar las métricas cualitativamente y exponiendo los resultados obtenidos con áreas de la entidad interesadas en el modelo, se consideró que los resultados son satisfactorios bajo un criterio de utilidad para la institución bancaria. Por último, también se debe tener en cuenta la metodología y enfoque empírico bajo el cual fue construido el modelo. El mismo logra un sano balance entre complejidad y practicidad. Es decir, logra un porcentaje de captura cualitativamente bueno sobre la propensión de compra de los clientes, mediante un modelo de implementación simple, automático, con mínima intervención humana y acciones plausibles dadas las restricciones tecnológicas, la calidad y la disponibilidad de los datos de la entidad.

Finalmente, la tabla 5 compara los resultados de las campañas por producto para las predicciones en diciembre. Usando el sistema de recomendación, se toma el 10 %, es decir, 2974 de los clientes más probables de adquirir cada producto. La comparación se realiza respecto a tomar el mismo número de clientes de forma aleatoria. Además, se presentan las rentabilidades esperadas para cada producto las cuales resultan de multiplicar los saldos promedio de cada producto con sus respectivas rentabilidades promedio observadas como una forma aproximada al ROA. Como mostramos en la tabla 5 con la métrica Uplift al 10 % para cada producto, el sistema de recomendación captura más clientes que de hacerse de forma aleatoria, por lo que las rentabilidades del sistema así mismo son mayores. Cabe resaltar que dado a que algunos productos presentan rentabilidades negativas, la rentabilidad al usar el modelo puede ser negativa. Al comparar las rentabilidades esperadas se obtienen los ROI para cada producto, siendo el producto de vivienda el más alto.

Cuadro 1: Métricas de evaluación por producto

Propensión de productos	Ahorros	Crediservice	Tarjeta de crédito	CDT	Libre destino	Nómina	Libranza	Vivienda
Exactitud	0.88	0.98	0.91	0.91	0.9	0.96	0.88	**0.96**
Precisión	0.05	0.04	0.04	0.03	0.02	**0.07**	0.01	0.01
Exhaustividad	0.69	0.72	0.58	0.73	0.6	0.69	0.74	**0.83**
AUC	0.84	0.89	0.83	0.89	0.83	0.87	0.84	**0.94**
Uplift al 10 %	6.54	7.6	5.84	7.31	5.98	7.3	6.79	**8.34**
Uplift al 20 %	3.77	4	3.34	3.89	3.33	4.01	3.87	**4.44**

[1] Uplift al 10 % y 20 % significa el cálculo del Uplift tomando el 10 % y 20 % de los clientes más probables de la muestra.

Cuadro 2: MAP@K para distintos niveles de K productos

MAP@K	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8
Propensión de productos	0.0162	0.0173	0.0182	0.0185	0.0187	0.0188	0.0188	0.0189
K más populares personalizado	0.0067	0.0085	0.0093	0.0099	0.0103	0.0105	0.0105	0.0105
K más populares	0.0056	0.0066	0.0090	0.0091	0.0099	0.0105	0.0108	0.0109

Cuadro 3: Conteo de adquisición de productos

Productos Agregados	Número de clientes
0	28729
1	602
2	91
3	10
4	5

Cuadro 4: Modelo de propensión de productos financieros vs. K productos populares personalizado

Métricas	Propensión de productos	K = 1 más popular	K = 2 más populares	K = 3 más populares	K = 4 más populares	K = 5 más populares	K = 6 más populares	K = 7 más populares
Exactitud	0.925	0.010	0.009	0.008	0.007	0.006	0.004	0.004
Precisión	0.0310	0.0004	0.0008	0.0010	0.0012	0.0013	0.0014	0.0014
Exhaustividad	0.668	0.235	0.445	0.551	0.653	0.724	0.766	0.771

Cuadro 5: Rentabilidad esperada del Sistema de recomendación

Producto	Rentabilidad promedio	Compradores capturados SR	compradores capturados aletoriamente	Rentabilidad esperada SR	Rentabilidad esperada aletoreamente	ROI
Ahorros	$21,402.49	157	23	$3,360,190.91	$492,257.27	583 %
Crediservice	$457,145.15	19	2	$8,685,757.85	$914,290.30	850 %
Libranza	($22,659.81)	36	6	($815,753.02)	($135,958.84)	500 %
Libredestino	$144,915.02	70	9	$10,144,051.62	$1,304,235.21	678 %
Nómina	$16,549.62	81	16	$1,340,519.21	$264,793.92	406 %
Tarjeta de crédito	$82,681.24	97	21	$8,020,080.70	$1,736,306.13	362 %
Vivienda	($914,406.42)	15	1	($13,716,096.25)	($914,406.42)	1400 %

La re

6. Conclusiones

El modelo de adquisición de productos financieros expuesto en este documento nace como respuesta a la necesidad de herramientas inteligentes que apoyen el desarrollo de campañas comerciales en la entidad financiera. El modelo, desde una perspectiva técnica, logra tener un alto desempeño de predicción en todos sus productos, dadas las restricciones mencionas: un AUC promedio del 87 % en la predicción sobre el mes de pruebas para los ocho productos analizados, alta exhaustividad y baja precisión, dándole importancia a la captura del total de clientes que efectivamente van a adquirir el producto ofrecido y un desempeño superior en el MAP@K comparado con el modelo base de K productos populares. A pesar de que el modelo no fue comparado con otras familias de modelos de la literatura, que valdrían la pena analizar en estudios posteriores, en exposiciones del mismo en la entidad y al compararlo con los procesos actuales de CRM, se identificó que el modelo propuesto tiene un gran potencial de generar un impacto positivo.

La metodología de exploración y selección de los modelos base mostró resultados adecuados. Prueba de ello fue que en versiones preliminares del sistema, en el que todos los modelos base eran estimados usando solamente los árboles de clasificación con potenciación de gradiente sin calibrar, el AUC promedio fue del 80 % y el MAP@K con k igual a 1 fue del 0.011. De esta forma la competición y calibración de parámetros de modelos mostró ser una mejor estrategia de modelación. También los bosques aleatorios mostraron ser superiores en 6 de los 8 productos, lo cual indica que para este ejercicio y dada la implementación del método de potenciador de gradiente es superior.

Si bien el modelo muestra resultados satisfactorios, es cierto que aún no ha pasado por una etapa de implementación y pruebas, en las que se deben contrastar sus predicciones sobre grupos de tratamiento y control (Chirkin, a, Aleksandra. Diskin, 2018). Esta es una siguiente etapa en el desarrollo del modelo. De la misma forma, en futuras investigaciones, es posible contrastar los resultados del modelo frente a otros algoritmos en la literatura de sistemas de recomendación en los grupos mencionados. Otro de los posibles desarrollos o mejoras del modelo se da a partir de la inclusión de un horizonte temporal más amplio, que permitiría por ejemplo, comparar las distribuciones de las probabilidades de adquisición estimadas entre varios meses. Al tener más información es posible ajustar el tamaño de los horizontes temporales o replantear los hiper-parámetros elegidos en los modelos base y contrastar los resultados respecto a otros meses de prueba con el fin de mejorar el poder predictivo. También en el proceso de selección de modelos es posible incluir otras familias de modelos de clasificación como las maquinas de soportes vectoriales y redes neuronales. Así como analizar a mayor profundidad las gráficas de dependencias parciales de más variables con el objetivo de entender mejor los perfiles de los clientes con mayor propensión de compra de cada producto.

Finalmente, se han identificado posibles usos de la herramienta en otros temas de interés para la entidad como riesgo crediticio, rentabilidad y retención de clientes. Primero, dado que el modelo identificó los productos más probables a ser adquiridos para cada cliente, este puede ser útil como insumo en el estudio de capacidad crediticia de los mismos. Al conocer la propensión de compra del cliente, se puede dar prelación al estudio de capacidad de pago para aquellos productos crediticios más propensos a ser adquiridos, lo que aportaría al flujo actual de la entidad. También el modelo también puede añadirse con indicadores ya establecidos de rentabilidad. De esta forma podría darse prioridad al ofrecimiento de productos a aquellos clientes más rentables pero a la vez más propensos. Por ultimo, como Gorgoglione y Panniello (2011) expusieron en su aplicación, el modelo puede unirse con la predicción de un modelo de propensión de fuga para generar acciones personalizadas a aquellos clientes más propensos a cancelar sus productos.

Referencias

Banco Santander. (2016). *Santander bank product recommendation competition.* Descargado de https://www.kaggle.com/c/santander-product-recommendation

Bennett, J., y Lanning, S. (2007). The Netflix Prize. , 3–6.

Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, *19*(84), 1-5. Descargado de http://jmlr.org/papers/v19/18-416.html

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... Li, Y. (2018). xgboost: Extreme gradient boosting [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=xgboost (R package version 0.71.2)

Cheng, H., Lu, Y.-c., y Sheu, C. (2009). Expert Systems with Applications An ontology-based business intelligence application in a financial knowledge management system. *Expert Systems With Applications*, *36*(2), 3614–3622. Descargado de http://dx.doi.org/10.1016/j.eswa.2008.02.047 doi: 10.1016/j.eswa.2008.02.047

Chirkin, a, Aleksandra. Diskin, D. R. (2018). *A Recommender System for Private Banking.* Descargado de https://www.incubegroup.com/blog/recommender-system-for-private-banking/

Choudhary, A., y Goila, T. (2019). CAN NEXT BEST PRODUCT OFFERING HELP BANKS BOOST CUSTOMER LOYALTY ? Can Next Best Product Offering Help Banks Boost Customer Loyalty ?

Cremonesi, P., Milano, P., Koren, Y., y Turrin, R. (2015). Performance of Recommender Algorithms on Top-N Recommendation Tasks Categories and Subject Descriptors. (September). doi: 10.1145/1864708.1864721

Friedman, J., Hastie, T., y Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (n.º 10). Springer series in statistics New York.

Gigli, A., Filopanti, V. Q., y Regoli, D. (2017). Recommender Systems for Banking and Financial Services. , 5–6.

Goldbloom, Anthony. (2019). *Kaggle.* Descargado de https://www.kaggle.com/

Google developers. (2019). *Classification: ROC Curve and AUC.* Descargado de https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Gorgoglione, M., y Panniello, U. (2011). Beyond Customer Churn : Generating Personalized Actions to Retain Customers in a Retail Bank by a Recommender System Approach. , *2011*(May), 90–102. doi: 10.4236/jilsa.2011.32011

Hu, Y., Koren, Y., y Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. En *Data mining, 2008. icdm'08. eighth ieee international conference on* (pp. 263–272).

Katsov, I. (2018). *Introduction to Algoritmic Marketing*. Grid Dynamics.

Kishida, K. (2005). Property of Average Precision and its Generalization : An Examination of Evaluation Indicator for Information Retrieval Experiments Property of Average Precision and its Generalization : An Examination of Evaluation Indicator for Information Retrieval Experiments.

Knott, A., y Neslin, S. A. (2002). Next-product-to-buy models for cross-selling applications. , *16*(3), 59–75. doi: 10.1002/dir.10038

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., ... Malohlava, M. (2019). h2o: R interface for 'h2o' [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=h2o` (R package version 3.22.1.1)

Linden, G., Smith, B., y York, J. (2003). Amazon . com. (February).

Modeling, U., y Burke, R. (2014). Hybrid Recommender Systems : Survey and Experiments Hybrid Recommender Systems :. (November). doi: 10.1023/A

Molnar, C., y cols. (2018). Interpretable machine learning: A guide for making black box models explainable. *Christoph Molnar, Leanpub*.

Pazzani, M. J., y Billsus, D. (2007). Content-Based Recommendation Systems. , 325–341.

Pryor, A. (2016). *Santander Product Recommendation Competition Solution.* Descargado de `http://alanpryorjr.com/2016-12-19-Kaggle-Competition-Santander-Solution/`

R Core Team. (2017). R: A language and environment for statistical computing. Descargado de `https://www.R-project.org/`

Ricci, F., Rokach, L., y Shapira, B. (2011). *Introduction to Recommender Systems Handbook*. Springer Science + usiness Media. doi: 10.1007/978-0-387-85820-3

Rifkin, R., y Klautau, A. (2004). In defense of one-vs-all classification. *Journal of machine learning research*, *5*(Jan), 101–141.

Sarwar, B., Karypis, G., Konstan, J., y Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. En *Proceedings of the 10th international conference on world wide web* (pp. 285–295).

Van de Wiele, T. (2016). *Santander Product Recommendation.* Descargado de `https://ttvand.github.io/Second-place-in-the-Santander-product-Recommendation-Kaggle-competition/`

Vico, D. G., y Huecas, G. (2012). Generating Context-aware Recommendations using Banking Data in a Mobile Recommender System. (c), 73–78.

Zibriczky, D. (2016). Recommender systems meet finance: a literature review.

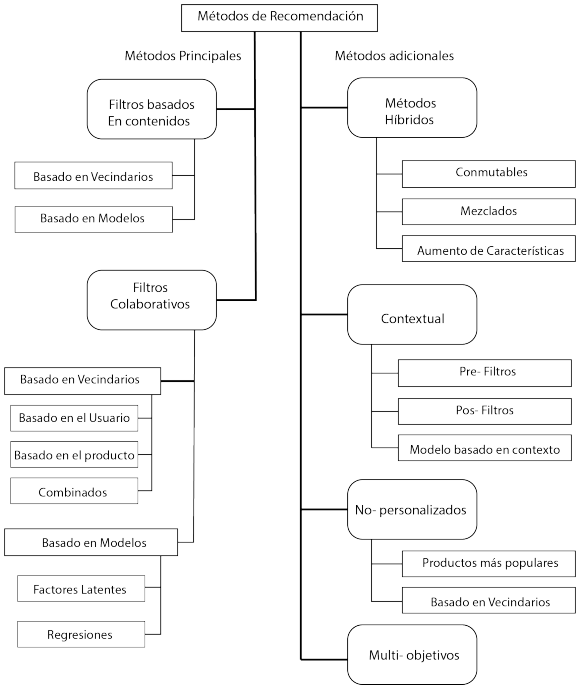
7. Apéndice

7.1. Miscelánea de gráficas

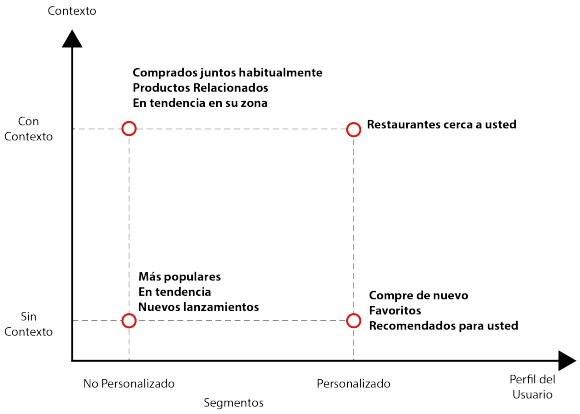
7.1.1. Clasificación de los algoritmos de recomendación

Figura 1: Diferente clasificaciones de los sistemas de recomendación, tomado de Katsov (2018).

(a) Clasificación de métodos de recomendación a partir de algoritmos.



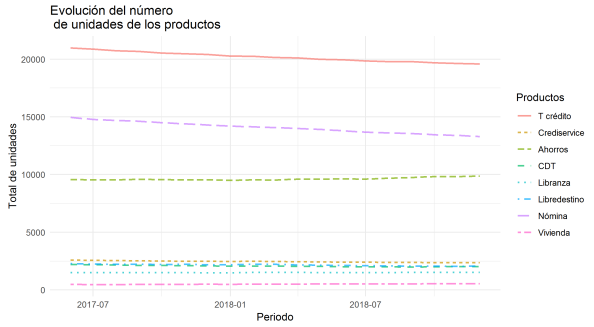
(b) Usos más recurrentes de los sistemas de recomendación.



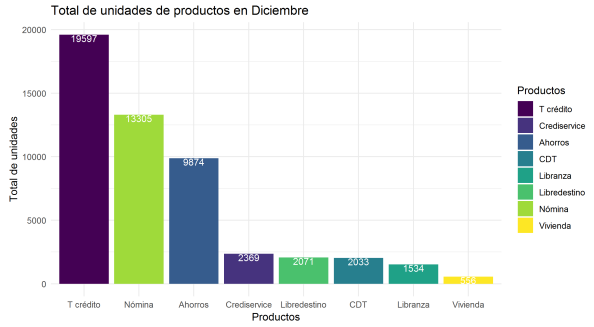
7.1.2. Clientes y productos

Figura 2: Relación entre los clientes y los productos en Banca de consumo.

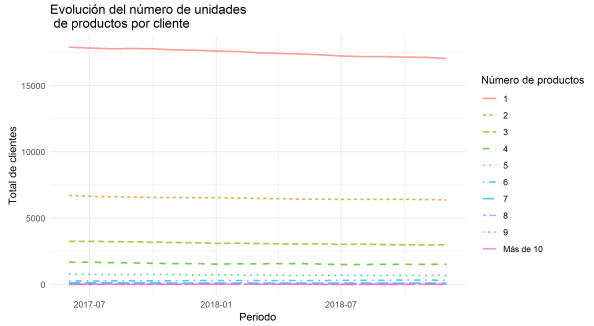
(a) Evolución del número de unidades de producto



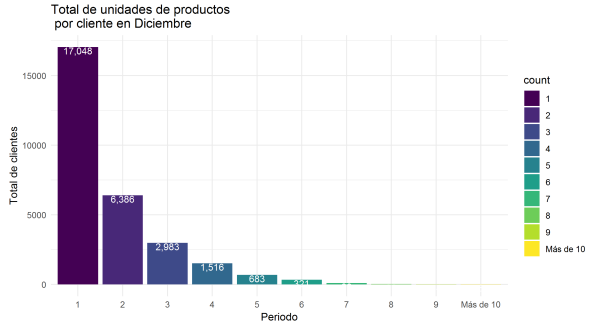
(b) Total de unidades de producto en Octubre



(c) Evolución del total de unidades de productos por cliente



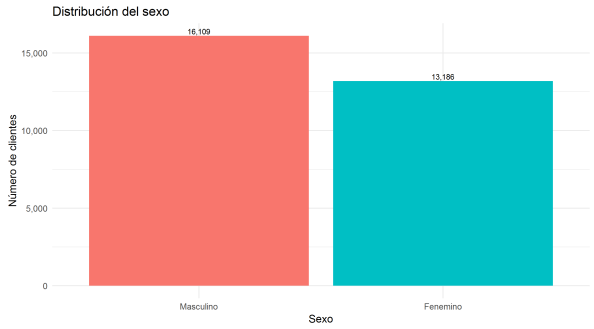
(d) Total de unidades de producto por cliente en Octubre



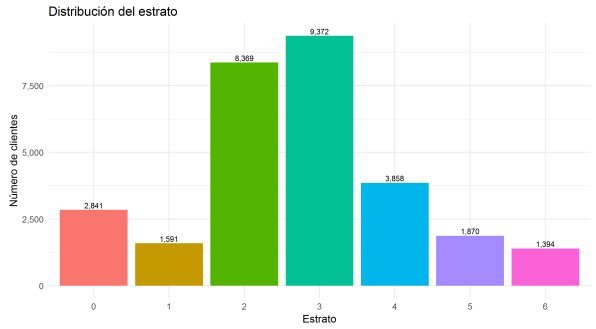
7.1.3. Variables demográficas

Figura 3: Composición socio demográfica de los clientes en Banca de consumo

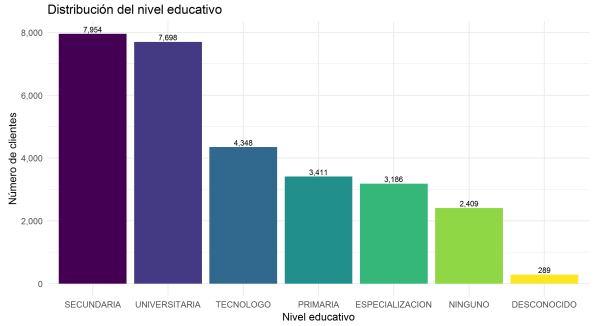
(a) Sexo



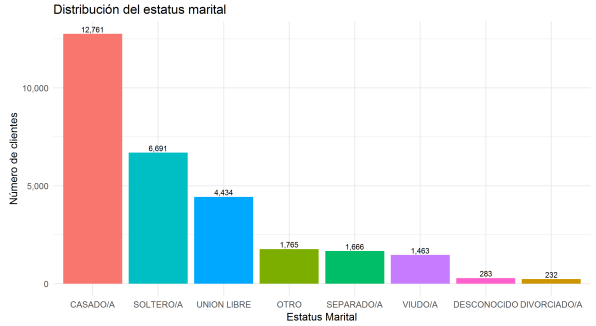
(b) Estrato socio económico



(c) Educación

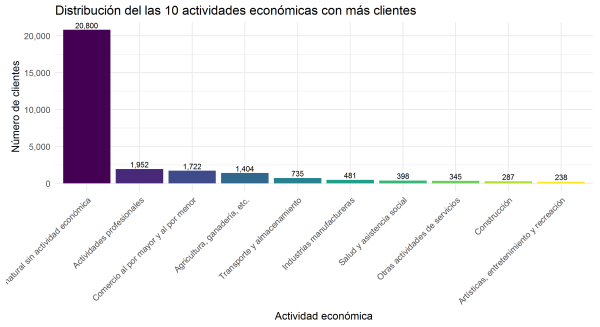


(d) Estatus marital

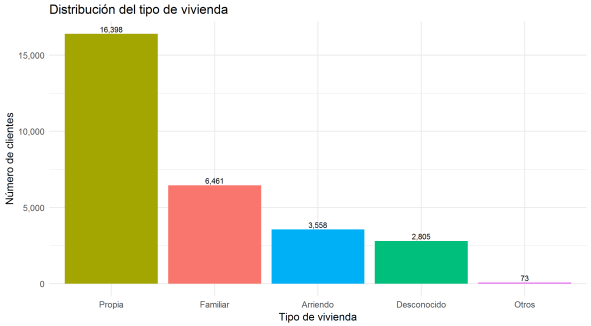


(e) Ocupación

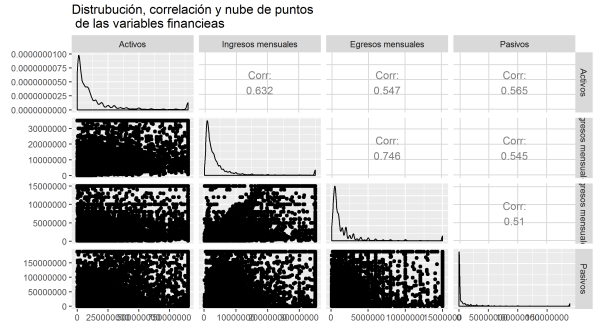
(f) Actividad económica



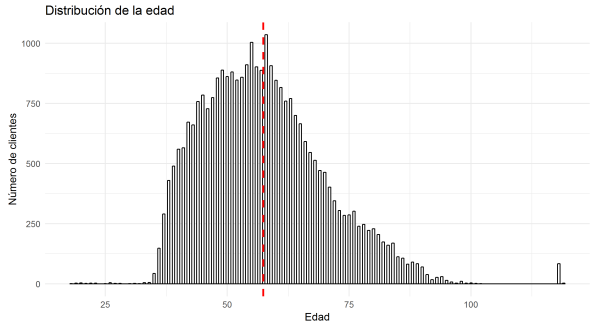
(g) Vivienda



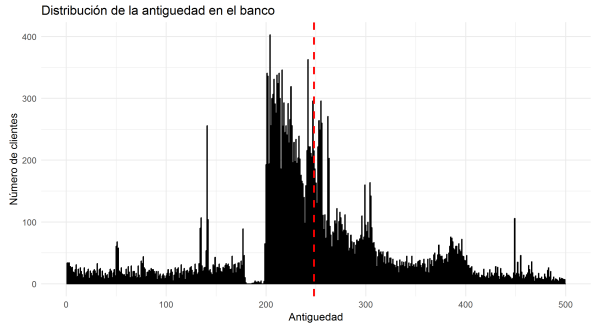
(h) Variables financieras



(i) Edad



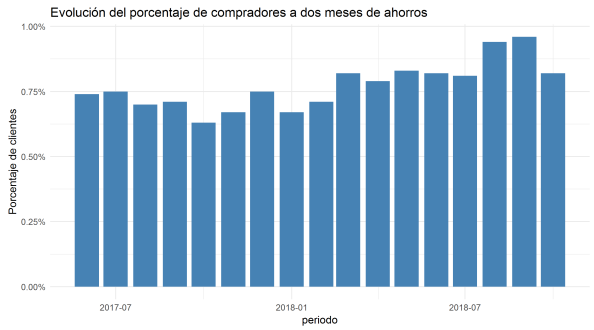
(j) Antigüedad



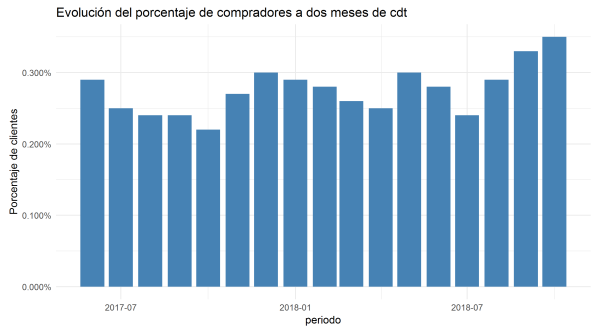
7.1.4. Ocurrencia de variable objetivo

Figura 4: Evolución de la variable objetivo por producto

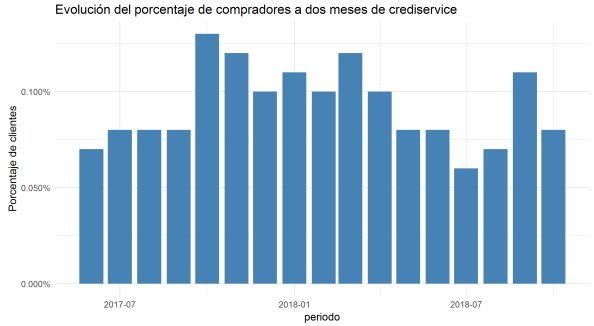
(a) Ahorros



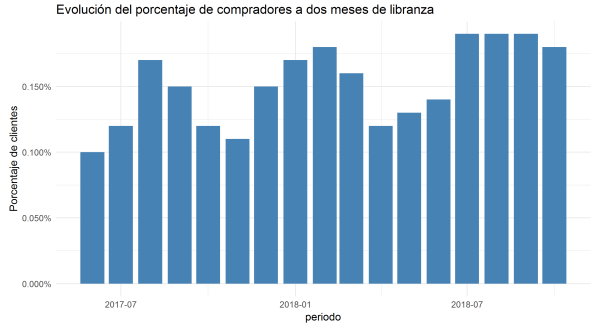
(b) CDT



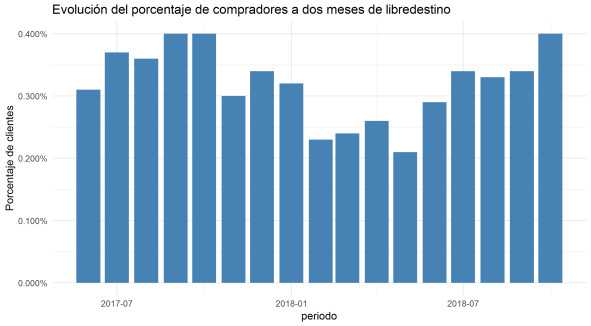
(c) Crediservice



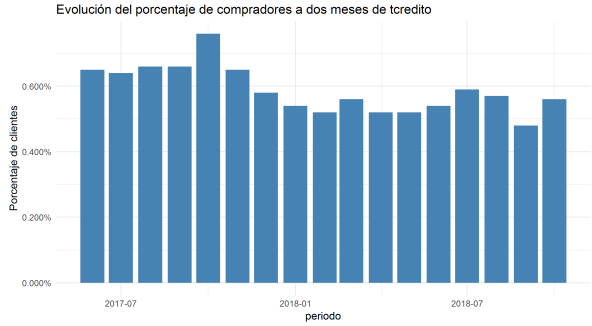
(d) Libranza



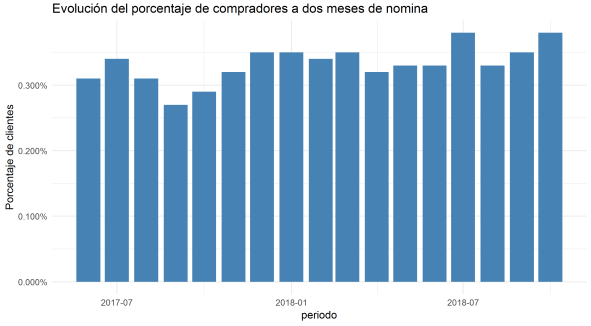
(e) Libre destino



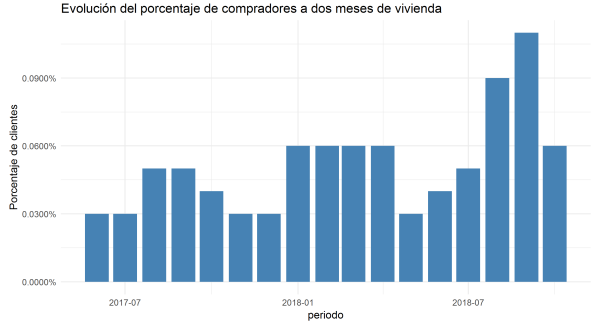
(f) Tarjeta de crédito



(g) Nómina



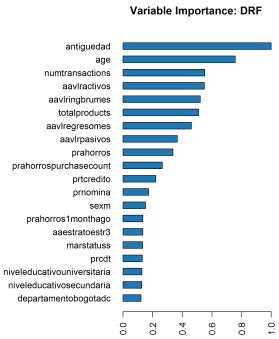
(h) Vivienda



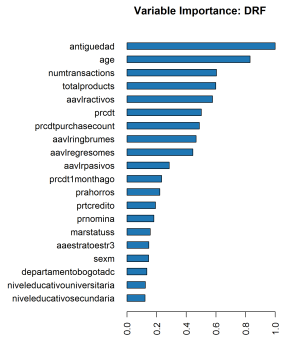
7.1.5. Variables importantes por producto

Figura 5: Variables importantes por producto

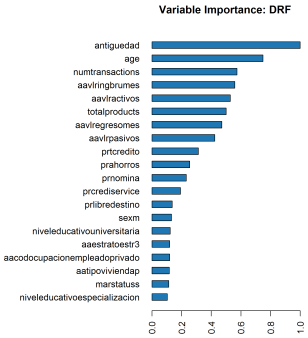
(a) Ahorros



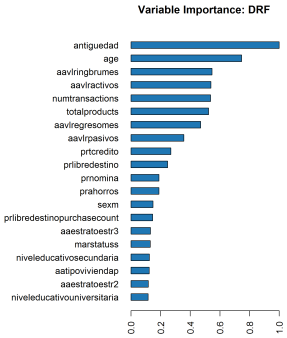
(b) CDT



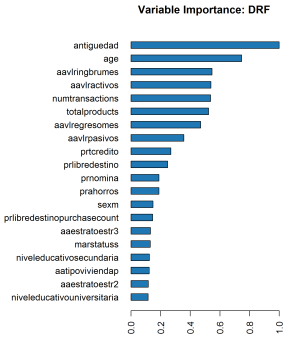
(c) Crediservice



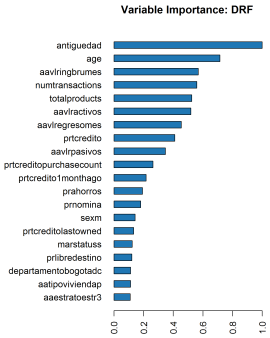
(d) Libranza



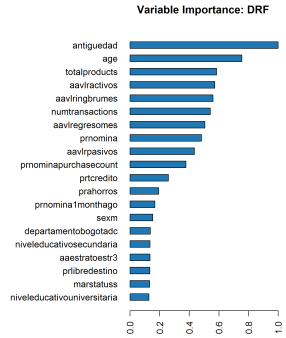
(e) Libre destino



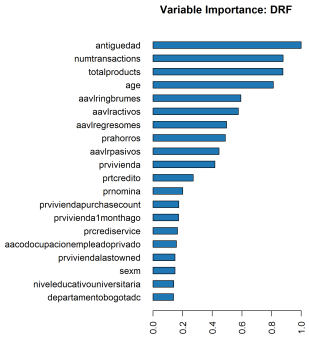
(f) Tarjeta de crédito



(g) Nómina



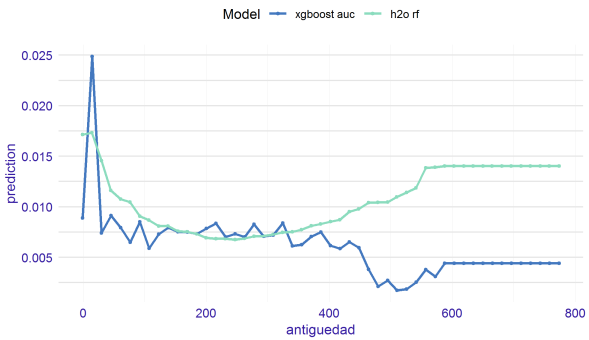
(h) Vivienda



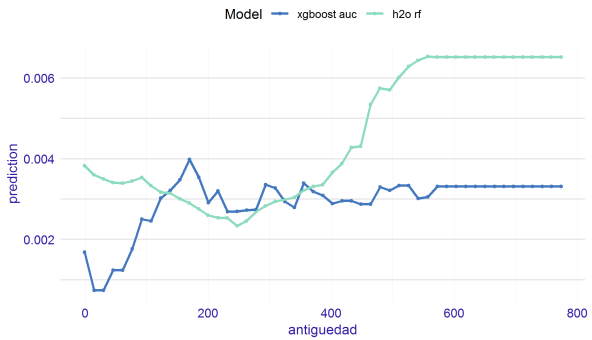
7.1.6. Gráficas de dependencias parciales de la antigüedad

Figura 6: Modelos bosques aleatorios y árboles de clasificación con potenciación de gradiente

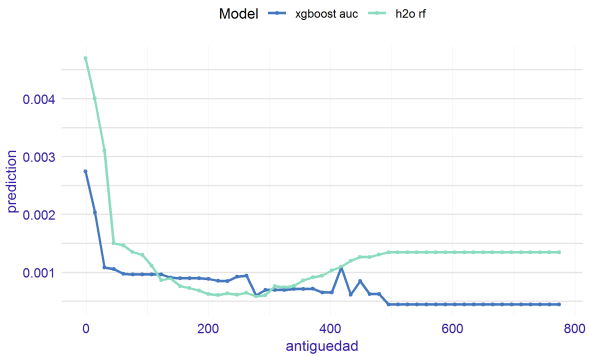
(a) Ahorros



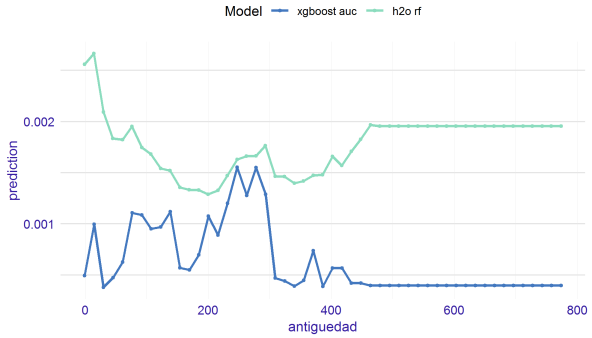
(b) CDT



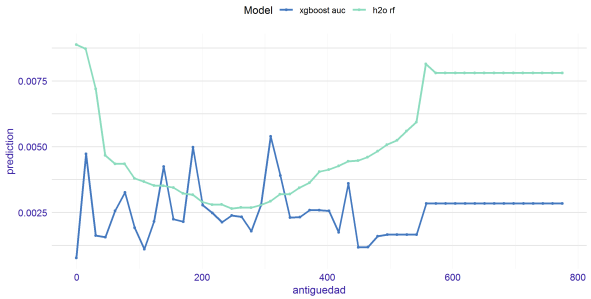
(c) Crediservice



(d) Libranza

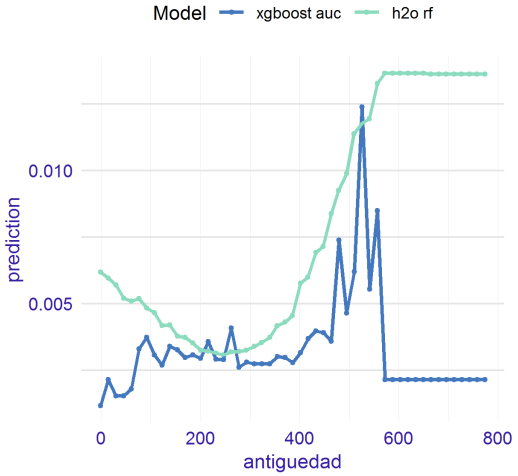


(e) Libre destino

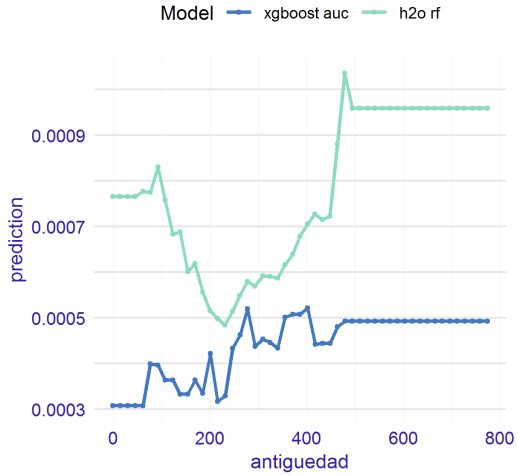


(f) Tarjeta de crédito

(g) Nómina



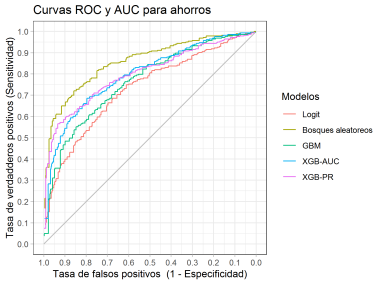
(h) Vivienda



7.1.7. Competición de modelos usando AUC

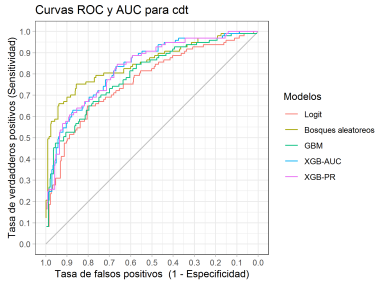
Figura 7: Competición de modelos usando AUC[12]

(a) Ahorros



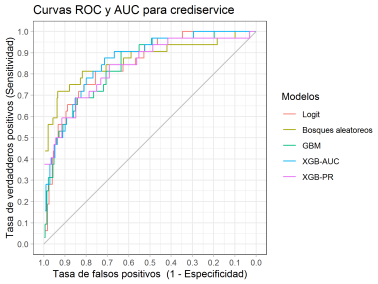
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.73	0.86	0.77	0.8	0.8

(b) CDT



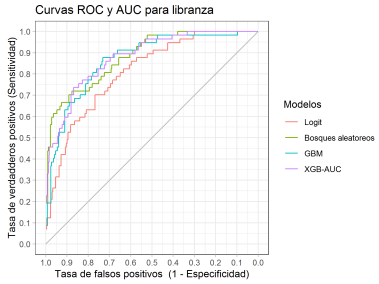
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.77	0.86	0.8	0.84	0.84

(c) Crediservice



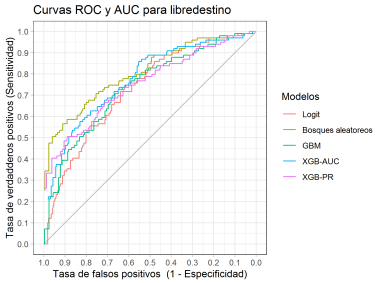
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.85	0.87	0.85	0.86	0.84

(d) Libranza



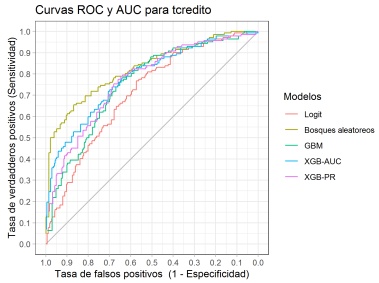
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.81	0.89	0.87	0.88	0.56

(e) Libre destino



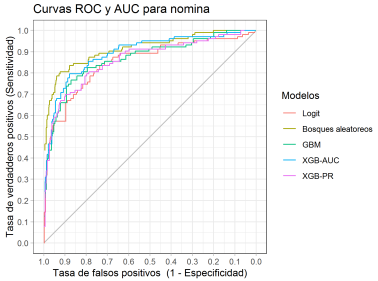
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.73	0.81	0.74	0.78	0.76

(f) Tarjeta de crédito



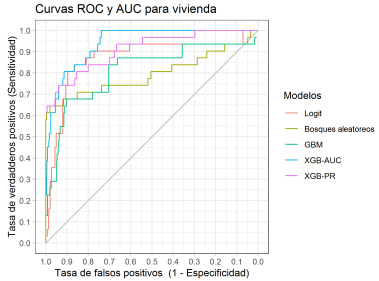
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.72	0.83	0.76	0.79	0.77

(g) Nómina



	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.86	0.92	0.87	0.9	0.87

(h) Vivienda



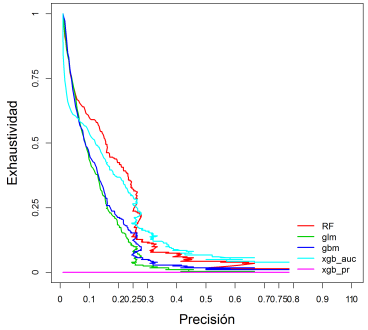
	Logit	**Bosques aleatoreos**	GBM	XGB – AUC	XGB – PR
AUC	0.88	0.8	0.82	0.94	0.91

[12] gbm, xgb_{auc} y xgb_{pr}: árboles de clasificación con potenciador de gradiente usando h2o y xgboost respectivamente.

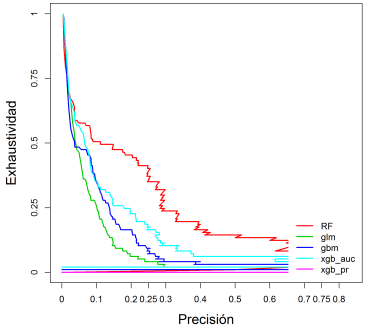
7.1.8. Competición de modelos usándola curva Precisión-Exhaustividad

Figura 8: Competición de modelos usando la curva precisión-exhaustividad[13]

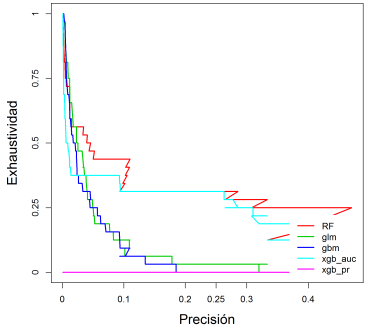
(a) Ahorros



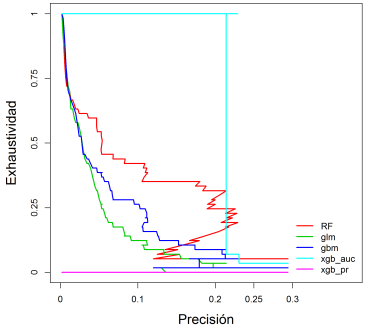
(b) CDT



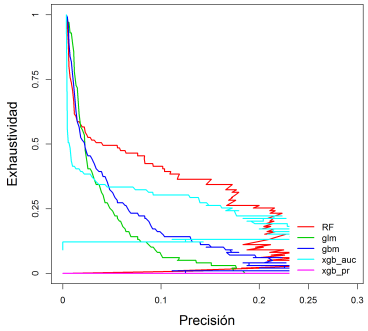
(c) Crediservice



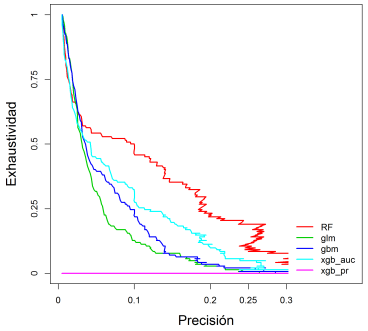
(d) Libranza



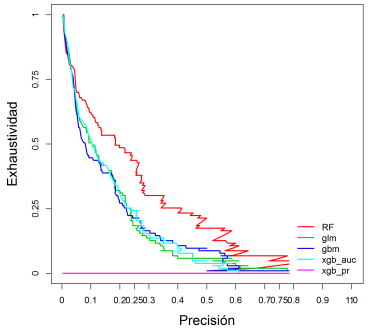
(e) Libre destino



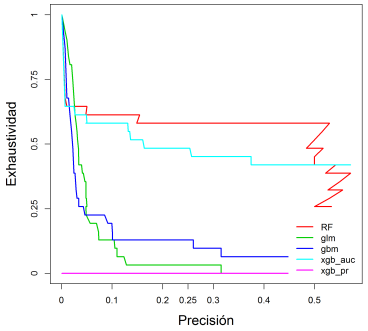
(f) Tarjeta de crédito



(g) Nómina



(h) Vivienda



[13] *RF*: bosques aleatoros, *glm*: logit, *gbm*, xgb_{auc} y xgb_{pr}: árboles de clasificación con potenciador de gradiente usando h2o y xgboost respectivamente.

7.2. Tablas

Cuadro 6: Descripción de productos

Productos	Descripción
Ahorros	Esta cuenta es un depósito a la vista, en la que el cliente tiene disponibilidad inmediata de sus recursos y además genera cierta rentabilidad en un periodo de tiempo según el monto ahorrado y el tipo de cuenta que elija el cliente.
CDT	Es un título valor, negociable, transferible por endoso, constituido a nombre de una o varias personas naturales o jurídicas, a una tasa de interés periódica dentro de un pazo pactado.
Crediservice	Es un crédito rotativo y revolvente. Rotativo, debido a que disminuye con su utilización y aumenta según los pagos mensuales o extraordinario que realiza el cliente. Revolvente porque el saldo a capital es diferido cada mes a 48 meses.
Libranza	Es un crédito de libre inversión dirigido a empleados, pensionados, y FFMM, que se caracteriza por brindar a nuestros clientes una alternativa de financiación con la comodidad del pago de sus cuotas a través del descuento de nómina.
Libre destino	Línea de crédito de cuota y tasa fija, dirigida a personas naturales como empleados, pensionados e independientes que se encuentren entre los 18 y 70 años de edad cuyos ingresos sean superiores a 1.5 SMMLV.
Tarjeta de crédito	Es un medio de pago que permite a los clientes financiar sus necesidades a través de un cupo de crédito rotativo. La tarjeta posee aceptación nacional e internacional y respaldo de la franquicia.
Nómina	Es un producto dirigido a empresas y empleados, basados en un modelo de soluciones integrales de ahorro, rentabilidad y seguridad por medio del cual la empresa podrá gestionar el pago de nómina y los empleados podrán manejar su dinero.
Vivienda	Es una línea de financiación para la compra de vivienda, que tiene como respaldo la hipoteca sobre el inmueble a financiar. El mercado objetivo son personas naturales en edades de 18 a 70 años con experiencia crediticia y con ingresos superiores a 1 SMMLV.

Cuadro 7: Variables del modelo de propensión de adquisición de productos financieros

Variables	Descripción
Tenecia de productos	Número de unidades de cada producto
Sexo	Sexo masculino femenino del cliente
Nivel educativo	Niveles de educación del cliente; ninguno, primaria, segndaria, tecnologo, universitaria, especialización y otros
Tipo de vivienda	Tipo de vivienda del clinete; familiar, propia, en arriendo, desconocido y otros
Estrato	Estrato socioeconómico del cliente de 0 a 6
Estatus marital	Estado marital del clinete; Soltero, casado, unión libre, separado, viudo, divorsiado y otros
Edad	Edad en años del cliente
Antigüedad	Antigüedad en el banco medida en meses
Activos financieros	Activo financieros reportados por el cliente
Pasivos Fiancieros	Pasivos financieros reportados por el cliente
Ingresos mensuales	Ingresos mensaulaes reportados por el cliente
Egresos mensuales	Egresos mensuales reportados por el cliente
Rezagos de tenecia del producto	Tenencia del producto analisado uno y dos meses anteriores al mes base de predicción
Antigüedad del producto	Meses de antigüedad que el cliente posee el producto a predecir
Total de productos	Total de productos que posee el cliente en el mes base de predicción
Número de veces que se ha añadido el producto	Número de veces que el cliente ha añadido el producto analizado historicamente
Número de veces que se a añadido los productos	Número de veces que el cliente ha añadido productos históricamente

7.2.1. Resultados de calibración de hiper-parámetros

Cuadro 8: Calibración de hiper-parámetros del modelo de bosques aleatorios

Productos	Número de árboles	Profundidad máxima	Minimas hojas	Máximas hojas	Tasa de muestra	AUC validación
Ahorros	235	20	2484	4177	0.8	0.87
Crediservice	500	20	164	999	0.8	0.88
Tarjeta de crédito	145	20	1769	3240	0.63	0.83
CDT	232	20	972	1965	0.63	0.85
Libre destino	197	20	1206	2240	0.63	0.81
Nómina	246	20	941	2485	0.8	0.92

Cuadro 9: Calibración de hiper-parámetros del modelo de árboles de clasificación con potenciador de gradiente

Productos	AUC Validación	Profundidad máxima	Eta	Gamma	Interrupción temprana	Mejor iteración	Número de variables
Libranza	0.88	6	0.3	0	30	139	20
Vivienda	0.94	6	0.3	0	30	73	22

7.3. Logit

La regresión logística es usada para problemas de clasificación binaria donde la variable a predecir o respuesta tiene dos niveles. El objetivo es modelar la probabilidad que una observación corresponda a alguna de las dos categorías $Pr(y = 1|x)$. La estimación de $\hat{y}$es igual a :

$$\hat{y} = Pr(y = 1|x) = \frac{e^{x^T\beta+\beta_0}}{1 + e^{x^T\beta+\beta_0}} \tag{4}$$

β y β_0 son estimados entonces maximizando :

$$max_{\beta,\beta_0}\frac{1}{N}\sum_{1=1}^{N}(y_i(x_i^T\beta+\beta_0) - log(1+e^{x^T\beta+\beta_0})) - \lambda(\alpha||\beta||_1 + \frac{1}{2}(1-\alpha)||\beta||_2^2) \tag{5}$$

7.4. Árboles de clasificación

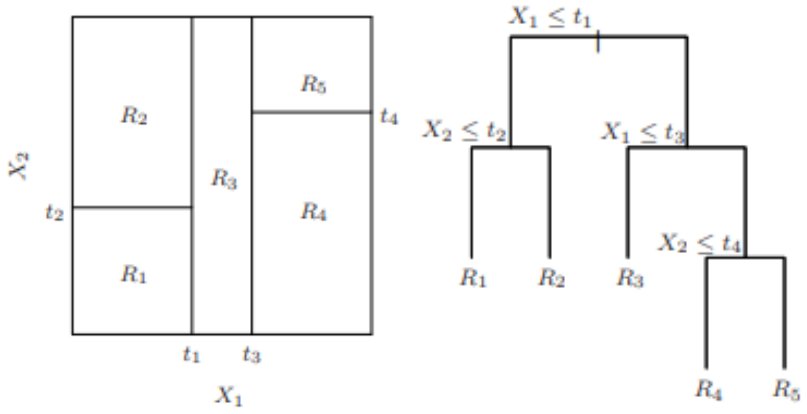
Un árbol de clasificación es un tipo de algoritmo supervisado [14] que particiona el espacio de predictores para generar una predicción de manera constante en cada parte resultante. En cada partición se elige la variable y corte dado el mejor ajuste posible. Luego una o ambas de las regiones resultantes son particionadas de nuevo en dos regiones más, este proceso continua hasta que alguna regla de cierre sea aplicada. Por ejemplo en la figura 8a la primera partición es en $X_1 = t_1$. Luego la región $X_1 \leq t_1$ es separada en $X_2 = t_2$. El resultado de este proceso resulta en cinco regiones $R_1, \ldots R_5$. El resultante árbol de clasificación 8b predice $\hat{Y}$en la región R_m como la clase de mayor proporción de Y en esta región (Friedman, Hastie, y Tibshirani, 2001).

[14]Los algoritmos supervisados son aquellos que estiman una función que relaciona variables explicativas con una variable de respuesta

Figura 9: Tomado de Friedman y cols. (2001)

(a) Particiones de un espacio vectorial de dos dimensiones

(b) Arbol de decisión



7.5. Bosques aleatorios

Los bosques aleatorios estan compuestos por conjunto de arboles de clasificación donde cada uno de ellos son denominados predictores débiles, peor que en conjunto logran predicciones bastante acertadas. Cada uno de estos árboles es construido a partir de la elección aleatorias tanto de variables predictoras como de muestra. A medida que aumentan los árboles se reduce la varianza.La predicción para cada observación es compuesta por el promedio o voto de cada árbol estimado. El algoritmo es entonces (Friedman y cols., 2001):

1. De $b = 1$ a B, donde B es el tamaño de bosque.
   (a) Tomar una muestra aleatoria Z^* de tamaño N de la muestra de entrenamiento.
   (b) Crear un conjunto de arboles de forma aleatoria, repitiendo recursivamente en cada nodo de cada árbol los siguientes pasos, hasta que el tamaño mínimo del nodo sea alcanzado n_{min}. :
      i. Seleccionar aleatoriamente m variables del total p
      ii. Elegir la mejor combinación variable/partición entre las m variables.
      iii. Dividir el nodo.
2. Obtener el conjunto de arboles $\{T_b\}_1^B$

La predicción es :

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

7.6. Árboles de clasificación con potenciación de gradiente

El método de árboles de clasificación con potenciación de gradiente elegido fue el denominado *Xgboost* (Cheng y cols., 2009). Esta versión del algoritmo está diseñada para mejorar la velocidad y rendimiento de este tipo de modelos. Actualmente es uno de los algoritmos más populares del aprendizaje automático y de las competiciones del portal Kaggle[15] debido a su fácil uso, eficiencia, precisión y factibilidad (Chen y cols., 2018). Para explicar este algoritmo, primero empezaremos por los conceptos de árbol de regresión, *Boosting* y *Gradient boosting*.

La potenciación (*Boosting*) es una técnica de ensamblaje donde los nuevos modelos, árboles de clasificación en este caso, son añadidos de forma secuencial para corregir los errores de los modelos previos hasta que no sea posible hacer futuras mejoras. Una forma de medir la ganancia al añadir nuevos modelos es mediante la potenciación de gradiente (*Gradient boosting*) que utiliza el algoritmo de gradiente descendiente para minimizar alguna función de perdida.

Los árboles de decisión pueden ser escritos de forma matemática como:

$$f(x) = w_{q(x)} \tag{6}$$

Donde $q(x)$ es una función que direcciona cada punto de x a un nodo terminal con un valor $w_{q(x)}$. Definamos también el conjunto de puntos asignados a cada hoja como:

$$I_j = \{i|q(x_i) = j\} \tag{7}$$

Supongamos que han sido estimados K árboles de decisión, el modelo es entonces un conjunto de árboles ensamblados de la forma:

$$\sum_{k=1}^{K} f_k \tag{8}$$

Las predicciones son estimadas entonces de forma iterativa como:

[15]Kaggle es un popular portal de ciencia de datos donde son propuestos y resueltos problemas analíticos por una comunidad de 536,000 miembros activos (Goldbloom, Anthony, 2019).

$$
\begin{aligned}
&\hat{y}_i^{(0)} = 0 \\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\ldots \\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
\tag{9}
$$

Para entrenar el modelo, es necesario optimizar una función de perdida, la más indicada para problemas de la clasificación es la función de perdida logarítmica (*Logloss*):

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i - log(p_i) + (1-y_i)log(1-p_i)) \tag{10}$$

La reguralización también es importante en el modelo, ya que controla la complejidad del modelo para evitar el sobre ajuste (*overfitting*)[16]

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{11}$$

Donde T es el número de nodos terminales u hojas y w_j^2 es el valor predicho en la hoja j. La función objetivo es entonces la unión de la función de perdida, que controla el poder predictivo, y la de regularización, que controla la simplicidad del modelo:

$$Obj = L + \Omega \tag{12}$$

De forma iterativa a partir de la ecuación 7.6, podemos definir la función objetivo como:

$$Obj^{(t)} = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \tag{13}$$

El método de árboles de clasificación potencializados utiliza el algoritmo de gradiente descendiente, el cual en cada iteración es re-estimada y mejorada $\hat{y}$ sobre la dirección del gradiente para minimizar la función objetivo. Para ello consideramos el gradiente de primer y segundo orden : $\partial_{\hat{y}_i^{(t)}} Obj^{(t)}$, $\partial^2_{\hat{y}_i^{(t)}} Obj^{(t)}$. La función objetivo es aproximada mediante una aproximación de taylor de segundo orden donde:

[16]El sobre ajuste o overfitting es una característica que debe ser evitada y ocurre cuando un modelo se ajusta de forma muy precisa a los datos de entrenamiento, pero tiene un pésimo ajuste en datos de prueba o fuera de muestra

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{14}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{15}$$

El objetivo es encontrar f_t, el árbol de decisión en la iteración t, que optimice la ecuación:

$$Obj^{(t)} \simeq \sum_{i=1}^{N} [L(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \sum_{i=1}^{t} \Omega(f_i) \simeq \sum_{i=1}^{N} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i))] + \sum_{i=1}^{t} \Omega(f_i) \tag{16}$$

La forma de encontrar el árbol f_t que mejore la predicción de $\hat{y}$ sobre el gradiente de la iteración t con lleva a dos preguntas esenciales: ¿Cómo encontrar la mejor estructura para dicho árbol?¿Cómo asignar la mejor predicción? tomando las ecuaciones 6 , 7 y 11, podemos reescribir la ecuación 16 en términos de la suma de las predicciones de cada hoja del árbol dado que cada punto asignado a una misma hoja comparte la misma predicción.

$$Obj^{(t)} \simeq \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{17}$$

A partir de la ecuación podemos estimar las mejores predicciones w_j de cada hoja dado que es un problema cuadrático. El peso óptimo w_j^* depende entonces de la primera y segunda derivada de la función de pérdida y del parámetro de regularización λ, además podemos obtener el valor de la función objetivo:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{18}$$

$$Obj^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{19}$$

En cada partición del árbol queremos encontrar el mejor punto de partición que optimice la función objetivo ya sea en el nodo raíz o en un nodo intermedio. A partir de la ecuación 7 y 19 definamos entonces I_L y I_R como el conjunto de puntos asignados a nuevas hojas provenientes del nodo con el conjunto I, podemos calcular la función de ganancia para una partición del conjunto I como:

$$Ganancia = -\frac{1}{2} [\sum_{j=1}^{T} \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \sum_{j=1}^{T} \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \sum_{j=1}^{T} \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}] - \gamma \tag{20}$$

De esta forma se encuentran los mejores puntos de partición de forma recursiva hasta que cada árbol llega a la máxima profundidad permitida, este es un parámetro pre-definido, después los nodos del árbol son recortados tomando las ganancias negativas de cada nodo desde los nodos terminales al nodo raíz.

7.7. Calculo de variables importantes

En el modelo de bosques aleatorios las variables importantes son calculadas a partir de sus influencias relativas que depende de si la variable fue elegida en el proceso de partición de nodos y cuando del error cuadrado mejoró tras la partición. el error cuadrado es el resultado entonces de (Rifkin y Klautau, 2004):

$$Error\ cuadrado = MSE * N = VAR * N \tag{21}$$

donde la varianza es igual a:

$$VAR = \frac{1}{N}\sum_{i=0}^{N}(yi - \hat{y})^2 \tag{22}$$

Entonces el error cuadrado es:

$$Error\ cuadrado = VAR * N = (\frac{1}{N}\sum_{1=0}^{N} y_i^2 - N\hat{y}^2) * N \tag{23}$$

8. Métricas de desempeño

8.0.1. Uplift

Esta métrica es usual para medir la efectividad de campañas de mercadeo o algoritmos de selección y es la diferencia entre la tasa de conversión entre el grupo de tratamiento y el grupo de control. Sin embargo, para medir la capacidad de captura del modelo en el mes de prueba, realizaremos un ajuste a la métrica. Primero, los clientes son ordenados de mayor a menor respecto las probabilidades de adquisición predichas. El grupo de tratamiento es entonces el primer x % de los clientes con mayor probabilidad. Por ejemplo, el primer 10 % de los clientes con mayor probabilidad. El grupo de control, ahora bien, es el total de la base. Esta métrica, de manera aplicada, permite seleccionar para un mes predicho, la cantidad de registros a ser candidatos a alguna acción comercial dada la efectividad medida en datos de prueba.

$$Tasa\ de\ conversión = \frac{Clientes\ queadquirieron\ el\ producto}{Total\ de\ clientes} \tag{24}$$

$$Uplift = \frac{Tasa\ de\ conversión\ x\%más\ probable}{Tasa\ de\ conversión\ total\ de\ la\ muestra} \tag{25}$$

8.0.2. Área bajo la curva ROC

La curva ROC [17] en inglés, es una gráfica que muestra el desempeño de un modelo de clasificación binaria respecto a todos los posibles umbrales que conviertan las probabilidades predichas en un resultado binario. Para ello, es graficada la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos en distintos umbrales:

$$Tasa\ de\ verdaderos\ positivos = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos} \tag{26}$$

$$Tasa\ de\ falsos\ positivos = \frac{Falsos\ positivos}{Falsos\ positivos + Verdaderos\ negativos} \tag{27}$$

El área bajo la curva resultante, AUC [18] por sus siglas en inglés, es un índice que permite medir de forma agregada el desempeño del modelo sobre todos los posibles umbrales de clasificación. EL AUC va de cero a uno, siendo cero un modelo con predicciones totalmente erróneas y lo contrario si es iguala uno. El AUC es deseable por dos razones: no cambia según la escala ya que mide que tan bien las predicciones son ordenadas, sin importar sus valores y es invariante respecto al umbral, ya que mide la calidad de las predicciones sin importar que umbral es elegido (Google developers, 2019).

8.0.3. Precisión, exhaustividad y exactitud

En campos como reconocimiento de patrones, búsqueda y recuperación de información y en general para problemas de clasificación binaria las métricas de precisión y exhaustividad son muy populares ya que miden la relevancia de las predicciones propuestas. La precisión es la proporción de verdaderos positivos predichos (el número de casos correctamente predichos de la clase positiva) respecto al total de positivos predichos por el modelo. La exhaustividad mide la proporción de verdaderos positivos respecto al total de casos positivos de la muestra.

Por un lado, la exhaustividad permite comparar cuantos verdaderos positivos predichos pertenecen al total de positivos sin importar el tamaño de positivos predichos, por lo que es posible obtener la máxima exhaustividad, 1, prediciendo la adquisición de todos los

[17] Receiver Operating characteristic Curve.
[18] Area Under the Curve.

productos. Por su parte la precisión habla sobre la densidad de verdaderos positivos sobre los valores positivos predichos, sin contemplar el total de casos positivos de la muestra (Katsov, 2018). La exactitud es la proporción de resultados verdaderos respecto al total de casos examinados, en otras palabras, es la relación entre el total de verdaderos positivos y verdaderos negativos respecto al total de casos positivos y negativos. Es una medida general para entender el poder de predicción del modelo.

$$Precisión = \frac{Verdaderos\ positivos}{Positivos\ predichos} \tag{28}$$

$$Exhaustividad = \frac{Verdaderos\ positivos}{Total\ de\ casos\ positivos} \tag{29}$$

$$Exactitud = \frac{Verdaderos\ positivos + Verdaderos\ falsos}{Total\ de\ casos\ positivos\ y\ negativos} \tag{30}$$

8.0.4. Promedio de la precisión media con k productos

Como su nombre lo indica el MAP@k calcula la precisión promedio para cada k productos adquiridos que se encuentren en la recomendación de cada cliente para luego promediar todos los valores obtenidos (Katsov, 2018; Kishida, 2005). Sea el número de recomendaciones Q, el número de productos adquiridos en la recomendación q igual a R_q y la precisión en el K-ésimo producto adquirido es P_{qk} entonces:

$$MAP@K = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{R_q} \sum_{k=1}^{R_q} P_{qk} \tag{31}$$