Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Sistema de recomendación de productos financieros en la organización bancaria Santander

Implementación de técnicas de Aprendizaje Automático

AUTOR: DI COSTANZO GERMAN
MENTORA: MG. NATALIA SALABERRY

NOVIEMBRE 2022

Resumen

El surgimiento de nuevas tecnologías y medios digitales ha generado un gran crecimiento en los volúmenes de datos con los que cuentan las organizaciones. Poder utilizarlos representa un gran desafío para estas. Por tal motivo, es importante contar con un entorno adecuado de gestión de los datos para que puedan ser recolectados, almacenados y procesados. El rápido avance tecnológico les ha permitido a las organizaciones explotarlos mediante diferentes técnicas, permitiéndoles identificar las preferencias de sus clientes. En particular, las organizaciones financieras, como es el caso del Banco Santander, también se encuentran entusiastas con esta posibilidad.

En este contexto, el tema propuesto a desarrollar en el presente trabajo busca exponer el potencial de la utilización de grandes volúmenes de datos en una organización bancaria para el diseño eficiente de las campañas de venta. El objetivo buscado consiste en determinar cuál es el producto financiero de mayor preferencia por cliente en la organización Santander, implementando modelos de aprendizaje automático. Esto constituye un valor agregado para el diseño eficiente de una estrategia de venta en la organización.

Con el fin de cumplimentar con el objetivo planteado, en primer lugar, se aborda la gestión de grandes volúmenes de datos en organizaciones bancarias. Luego se presentan los datos a utilizar y se describen los métodos a implementar. Finalmente se aplican estos y se evalúan los resultados obtenidos para la recomendación de productos. De esta manera, el uso de un sistema de recomendación permite identificar el producto de mayor preferencia para cada cliente, permitiendo realizar campañas eficientes de venta en la organización bancaria Santander.

Palabras clave: Sistema de recomendación, Productos financieros, Organización bancaria, Aprendizaje automático.

Índice

Introducción

El fuerte desarrollo tecnológico del siglo XXI ha facilitado la creación de sitios web, aplicaciones móviles y diferentes medios de comunicación digital, que son nichos concentradores de consumidores donde se generan grandes volúmenes de datos. Estos datos son más complejos en su tratamiento ya que, además de su tamaño, poseen una estructura variada. Poder explotarlos es un tema de interés para todas las organizaciones y en particular para las del sector financiero.

En este sentido las organizaciones bancarias, como lo es el Banco Santander, también se encuentran interesadas. Esto ya que les permite realizar una estrategia de venta personalizada obteniendo mayor eficiencia de negocio. Además, pueden obtener un mayor conocimiento del cliente y personalizar la oferta de productos, lo que les permitirá establecer una relación a largo plazo entre la organización y el cliente. De esta manera, lograr la fidelización y retención de clientes, en un entorno de marcada competencia, es un punto fundamental para las organizaciones bancarias.

En este contexto, resulta interesante poder determinar cuál producto es preferido por cada cliente en particular de la organización bancaria Santander. Para eso resulta elemental utilizar los datos pertenecientes a esta que fueron publicados en el repositorio de *Kaggle*[1] y son de libre acceso, así como los métodos de aprendizaje automático que permitan analizarlos. De esta manera, el interrogante que surge para llevar adelante este trabajo es: ¿Cuál es el producto de mayor preferencia para cada cliente del Banco Santander?

Para responder el interrogante planteado, el objetivo general del trabajo de especialización consiste en identificar el producto financiero de mayor preferencia por cliente de la organización bancaria Santander, implementando modelos de aprendizaje automático. Esto constituye un valor agregado para el diseño eficiente de una estrategia de venta en la organización. De esta manera, la hipótesis principal es que el uso de un sistema de recomendación permite identificar el producto de mayor preferencia por cliente, permitiendo realizar una gestión eficiente de las campañas de venta en la organización bancaria Santander.

[1] Los datos se encuentran accesibles de manera online en el siguiente link: https://www.kaggle.com/competitions/santander-product-recommendation/data

Para resolver el objetivo planteado, el trabajo se desarrolla en tres apartados. En el primer apartado se analizarán los procesos y metodologías para la gestión de grandes volúmenes de datos en el sector bancario. En primer lugar, se desarrolla el concepto de los grandes volúmenes de datos y porque son importantes para las organizaciones. Luego se contextualizará dentro de las organizaciones bancarias. Finalmente, se determinará como una correcta gestión de datos impacta positivamente en la planificación de una estrategia de venta.

En el segundo apartado, se presentará el conjunto de datos con el cuál se va a trabajar. A continuación, se realizará un análisis exploratorio. Luego se especificarán los métodos predictivos que permiten generar el desarrollo de un sistema de recomendación de productos financieros.

En el tercer y último apartado se implementarán los modelos para la recomendación de productos. Luego se realizará la evaluación de resultados obtenidos por cada modelo. A continuación, se seleccionará el mejor. Finalmente se concluirá acerca de la importancia de contar con un sistema de recomendaciones al momento de ofrecer productos financieros.

1. Gestión de datos en organizaciones bancarias

Hacia fines de la década del 90, la expansión de las páginas web condujo al enorme crecimiento de la generación de datos y también, al desarrollo de técnicas para su análisis (Lee, 2017). Esto surge como consecuencia de que los datos son cedidos a las organizaciones por los usuarios a cambio de satisfacer una necesidad (Zuboff, 2015). A su vez les permitió obtener un mayor conocimiento de sus clientes sobre sus gustos y preferencias a la hora de elegir un determinado producto o servicio, cambiando en gran parte su modelo de negocios (Dicuonzo G., 2019). Por otra parte, el avance tecnológico, ha reducido los costos de almacenamiento de los datos lo que también ha facilitado su alta disponibilidad (Schmarzo, 2013).

Como consecuencia, surgieron nuevas fuentes de datos como ser las redes sociales, dispositivos móviles y sensores. Estas son aprovechados por las organizaciones para mejorar los procesos de interacción con sus clientes ya que le proporcionan un mayor conocimiento sobre estos. Como sostiene Schmarzo (2013), esto a su vez implicó una revolución empresarial basada en datos.

Este nuevo desafío atraviesa a muchos sectores de la industria, incluido el comercio minorista, mayorista, banca de inversiones y riesgos, entre otros. En el caso particular de las organizaciones bancarias surge el reto de trabajar con grandes volúmenes de datos de forma ágil y precisa. Para enfrentarlo, se requiere de un cambio cultural en la organización y, de ser necesario, la formación de equipos interdisciplinarios compuestos por especialistas estadísticos, matemáticos y científicos de datos que pueden combinar habilidades de análisis de datos con habilidades funcionales para crear procesos de valor (Dicuonzo G., 2019).

Es entonces como la gestión de grandes volúmenes de datos en las organizaciones bancarias toma un rol principal, ya que es un proceso que atraviesa a todas las áreas. En particular, a la hora de ofrecer un nuevo producto o servicio, resulta atractivo poder identificar a que clientes de su cartera les será relevante. Es así como surge el interés por los sistemas de recomendaciones a la hora de lanzar una nueva campaña de venta (Gigli, 2017).

El objetivo del siguiente apartado es analizar los proceso y metodologías para la gestión de grandes volúmenes de datos en el sector bancario. Para eso, en primer lugar, se abordará el concepto de grandes volúmenes de datos y su importancia para las organizaciones. Luego, se identificará como es llevada a cabo su gestión en el contexto de una organización bancaria.

Finalmente se especificará sobre la relevancia de una correcta gestión de datos para la determinación de un sistema de recomendación de productos bancarios como valor agregado para el diseño de estrategias de ventas.

1.1 Acerca de grandes volúmenes de datos

En el contexto tradicional, la estrategia del gobierno organizacional, la recopilación y el uso de datos, se han llevado a cabo por medio de prácticas establecidas y basadas en la experiencia. Pero esta forma solo ha servido para fines corporativos particulares, como el control y la contabilidad, finanzas o marketing (Porter, 1995). A las organizaciones les sigue resultando importante mantener estas formas ya que los datos son parte de una estructura cognitiva más amplia proporcionada por categorías o clasificaciones (Constantiou, 2015).

Habitualmente, la gestión y el análisis de la información eran principalmente en base a las decisiones internas. Pero esto es algo diferente cuando se refiere a grandes volúmenes de datos. Si bien hay muchos casos en los que son utilizados para ese propósito, en general, en lugar de crear informes o presentaciones que asesoren a los altos ejecutivos internos, los científicos de datos comúnmente trabajan orientados al cliente, productos y servicios (Davenport, 2014). Por lo tanto, es necesario un enfoque más amplio para realizar muestreos, analizar y actuar sobre los datos.

Surge entonces el término *Big Data*, que no solo hace referencia a grandes volúmenes de datos, también abarca a datos que tienen una mayor variedad y crecen a más velocidad (Chen H. C., 2012). Es decir, cuando nos referimos a *Big Data* hablamos de conjuntos de datos de mayor tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Resulta necesario hacer foco en las tres características mencionadas.

Por volumen se entiende la manera en que los datos son almacenados, transformados y las tecnologías que esto implica. Con el surgimiento de grandes cantidades de datos digitales en la última década, el volumen de datos comenzó a crecer de manera exponencial. Además, se suma la aparición de nuevos tipos de datos distintos a los tradicionales, agregando complejidad para su almacenamiento y procesamiento (Katal, 2013).

En cuanto a la velocidad, hace referencia a como los datos se generan de forma continua y acelerada. Por tal razón, resulta necesario almacenarlos y procesarlos apropiadamente para hacer un correcto uso de los mismos (Laborde, 2020). Es decir, a la complejidad de almacenar

y procesar grandes volúmenes de datos se incorpora el hecho de que pueden ser generados en tiempo real, o periódicamente.

Por último, el término variedad refiere a la estructura o forma que toman los datos, como mensajes, actualizaciones e imágenes publicadas en las redes sociales y lecturas de sensores. Lo mismo ocurre con los teléfonos inteligentes y otros dispositivos móviles que ahora brindan enormes flujos de datos vinculados a personas, actividades y ubicaciones (Andrew McAfee, 2012). Por lo tanto, las tecnologías de almacenamiento de datos tradicionales no son adecuadas para almacenar y procesar este tipo de datos.

De esta manera, las características mencionadas sobre *Big Data* lo convierten en un entorno de datos que conlleva un cambio cultural en toda la organización. Y es en este sentido, que se requiere una correcta gestión sobre grandes volúmenes de datos para brindar apoyo a la toma de decisiones. Este cambio también ha alcanzado a las organizaciones bancarias. Para comprender sus implicancias a continuación se desarrolla al respecto.

1.2 Gestión de grandes volúmenes de datos en organizaciones bancarias

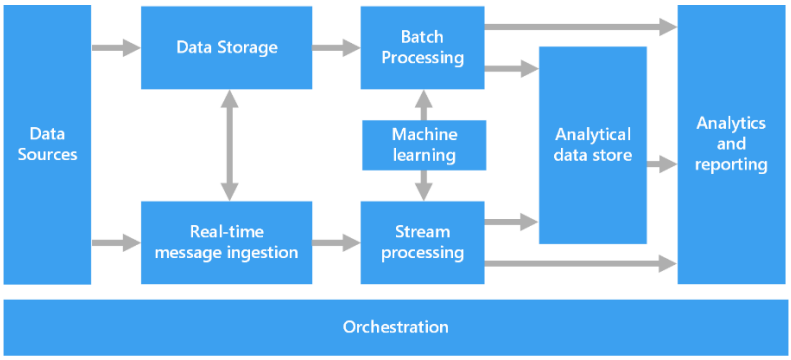
Habiendo conceptualizado sobre *Big Data* y sus características definitorias, se puso en evidencia que los grandes volúmenes de datos están ganando un amplio reconocimiento por parte de los responsables de la toma de decisiones en las organizaciones (Labrinidis, 2012). Esto tiene una importancia significativa en las organizaciones financieras, particularmente en los bancos, ya que ayuda a tomar decisiones más efectivas. Es en este sentido que los datos de carácter estructurado y no estructurado se utilizan para crear mejores estrategias y anticipar el comportamiento del cliente (Sun, 2014).

Durante las últimas dos décadas, la industria bancaria ha experimentado cambios importantes, dando como resultado un entorno caracterizado por la intensa competencia, la globalización, la mayor movilidad y demanda de los consumidores, y la desregulación (Cohen, 2007). Pero como la mayoría de los bancos ofrecen productos y servicios financieros similares, la manera de hacer frente a la competencia es revisando sus estrategias de un producto orientado al cliente (Al-Hawari, 2006). Por lo tanto, además de incorporar nuevos clientes, el desafío consiste en retenerlos, tener una mayor fidelización y recomendación por parte de los clientes para lograr así una mayor participación de mercado.

El procesamiento y análisis de grandes volúmenes de datos ha demostrado ser una estrategia fundamental para ser implementada en la industria bancaria. Esto se debe al crecimiento de los datos que son recolectados y almacenados, siendo estos de mucho valor para las distintas áreas de la organización en sus procesos internos (Rakhman, 2019). Cabe señalar que esta gran cantidad de datos almacenados implica el uso de equipos (*Hardware)* y herramientas (*Software)* mucho más sofisticadas (Rahman, 2015).

Por lo tanto, cuando los bancos y otras empresas deciden implementar el análisis de grandes volúmenes de datos en su operación, primero deben establecer la base, es decir, el equipamiento (*Hardware*). El *Hardware* requerido consiste en instalaciones de red de información, computadoras de alto rendimiento, servidores en la nube, y almacenamiento de gran capacidad. Por otra parte, el *Software* debe mostrar características como gran capacidad para hacer minería de datos, inteligencia artificial y almacenamiento extendido (Wang, 2021). Una posible arquitectura para la implementación de estas herramientas de *Hardware* y *Software* es la que se muestra a continuación.

Figura 1: Arquitectura *Big Data*.



Fuente: Obtenido de Microsoft Docs (Big data architecture style - Azure Architecture Center)

En el esquema de la figura 1 se observa en primera instancia, las distintas velocidades por las que pueden ser generados los datos. Es decir, si existen datos generados en tiempo real deben definirse componentes de captura (*Real-Time Message Ingestion*) y procesamiento (*Stream Processing*). Lo mismo sucede con los datos que no son capturados en tiempo real y se procesan de manera periódica (*Batch Processing*).

Luego, una vez que los datos son capturados y procesados, es posible su explotación con herramientas de *Machine Learning* o con herramientas de visualización y reportería (*Analytics and Reporting)*. Además, si es necesario se puede incorporar una instancia de almacenamiento

donde los datos sufren alguna transformación para ser consumidos por estas herramientas de reportería (*Analytical Data Store*). Por último, el componente de orquestación refiere a la gobernanza y seguridad a lo largo de toda la arquitectura.

Si bien los datos son esenciales, la ventaja y la mejora en el desempeño de la obtención de información se encuentra en los modelos analíticos que permiten optimizar y predecir resultados. Esto permite que las organizaciones bancarias con la capacidad de explotar los grandes volúmenes de datos obtengan una ventaja competitiva al tener más información sobre su negocio y clientes, respecto a sus rivales (Barton, 2012). Es entonces que resulta necesario cambiar las estrategias y capacidades de captura, almacenamiento, procesamiento y análisis de datos para afrontar el desafío que implica los grandes volúmenes de datos (Bedeley, 2014).

A su vez, Dicuonzo (2019) sostiene que aplicar métodos de aprendizaje automático sobre los grandes volúmenes de datos, permite a los bancos ofrecerles a sus clientes productos personalizados que se adaptan a sus gustos y preferencias. Además, menciona que, en las organizaciones bancarias, la gestión de grandes volúmenes de datos abarca tres categorías: gestión de la relación con el cliente (CRM), gestión de riesgos (prevención y detección de fraude), y banca de inversión. También observa que los bancos que incorporan análisis de grandes volúmenes de datos están ganando un 4% de ventaja competitiva sobre otros dentro del sector bancario.

De este modo, la gestión de grandes volúmenes de datos en un contexto bancario implica cambios en todos los sentidos. Desde la conformación de equipos de trabajo con perfiles diversos, la necesidad de adaptar (y modificar de ser necesario) su arquitectura de datos hasta la incorporación de herramientas más avanzadas. Por lo cual, mediante una correcta gestión, los bancos pueden tener más información acerca de sus clientes.

1.3 Sistemas de recomendación para el diseño de una estrategia de venta

Como se mencionó en el apartado anterior, la era de los grandes volúmenes de datos presenta grandes desafíos con marcados cambios para las organizaciones bancarias. Por lo que acceder y comprender la información a través de técnicas de minería de datos (DM) se ha convertido en una tendencia emergente. Además, el contexto financiero ha sido un campo de interés para la implementación de estas técnicas durante las últimas décadas (Hassani, Huang, & Ghodsi, 2018).

Los bancos han reconocido que el conocimiento en lugar de los recursos financieros es el nuevo mayor activo (Kharote, 2014). A su vez, como se mencionó anteriormente, el desarrollo y la popularización de la banca electrónica y la banca móvil aporta al crecimiento exponencial de los datos. Estos desarrollos continuos y la disponibilidad cada vez mayor de grandes volúmenes de datos hacen que dominar las herramientas de análisis y explotación sea una de las tareas más cruciales.

En ese sentido, los servicios bancarios y financieros consideran cada vez más el campo del aprendizaje automático con el fin de aprovechar los datos a su disposición para proporcionar servicios y experiencias personalizadas a sus clientes. Dado que los bancos ofrecen una gran variedad de productos, es importante que sean ofrecidos a los clientes adecuados. Por lo que la adopción de sistemas de recomendación para ofrecer experiencias personalizadas a clientes existentes y potenciales podría tener un gran impacto en ventas de productos, que influyen directamente en la facturación y los ingresos (Oyebode, 2020).

Los sistemas de recomendación son sistemas que, mediante técnicas de aprendizaje automático, permiten obtener elementos en los que es probable que el usuario esté interesado en un contexto específico (Ricci, 2011). En general estos métodos generan un ranking de productos (música, películas, productos bancarios, libros, etc.) para cada cliente, a los fines de ofrecerle el producto más adecuado para él. Existen también, modelos que generan probabilidades y que pueden utilizarse para obtener un ranking probabilístico de compra del producto (Rezaei, 2021). En cualquier caso, es necesario entrenarlos para lo cual hay que disponer de ratings (ejemplo estrellas de *Amazon*) o relevancia; me gusta/no me gusta o compró/no compró.

Zibriczky (2016) sostiene que, desde el punto de vista comercial, un desafío común al que se enfrentan varias instituciones financieras es la falta de un sistema inteligente de apoyo a la toma de decisiones. Como las actividades de venta de productos financieros requieren conocimiento experto, los sistemas de recomendación ofrecen grandes beneficios para los servicios financieros, mejorando la eficiencia de los representantes de ventas automatizando el proceso de toma de decisiones para los clientes. Por lo tanto, se observa una importante demanda de estos sistemas de apoyo a la decisión.

En este aspecto, vale señalar la importancia de estos sistemas de recomendación a la hora de elaborar la estrategia de venta de un determinado producto. Permite ofrecerle a cada cliente,

sólo los productos relevantes o interesantes para él. A su vez, como el cliente no se ve perturbado por distintas ofertas de productos que no son de su interés, puede ser incorporado en otras campañas de venta sin tener una sobreoferta constante de productos (Deng, 2019).

Por lo cual, tomar el resultado de un sistema de recomendación como punto de partida para la campaña de venta de un producto financiero, permite ser más asertivo y eficiente a la hora de comunicarse con el cliente. Esto implica menores costos de campaña, ya que se realizan una menor cantidad de contactos, manteniendo el éxito de esta (Kosaman, 2018). A su vez, los sistemas de recomendación también pueden ser utilizados por las organizaciones financieras para mejorar el *Cross-Sell*, sugiriendo nuevos productos a un cliente que ya posee algún producto en la entidad (Schafer, 1999).

Como conclusión, en este capítulo se establece que los sistemas de recomendación son un valor agregado para el eficiente diseño de una campaña de venta a la hora de ofrecer y brindar nuevos productos en organizaciones bancarias. Para alcanzarlo se determinó que los grandes volúmenes de datos impactan en los esquemas tradicionales de las organizaciones debido a su volumen, velocidad y variedad. En este sentido, las organizaciones, enfrentan el desafío de trabajar adecuadamente con estos para lograr una gestión en términos de la relación con el cliente. De este modo, podrán implementar el sistema mencionado de manera eficiente.

Para demostrar la utilidad de este sistema, en el siguiente capítulo se comienza por presentar el conjunto de datos a utilizar, perteneciente a la organización bancaria Santander. Se desarrollará el procesamiento y análisis del conjunto de datos. Luego se llevará a cabo la descripción de los modelos predictivos seleccionados que serán utilizados en una implementación posteriormente.

2. Métodos predictivos para la recomendación de productos financieros

Como resultado del avance en la generación de grandes volúmenes de datos, surgen nuevos desafíos para las organizaciones respecto a cómo explotar estos datos. Agrawal (2011), sostiene que el análisis de datos permite, por ejemplo, obtener estadísticas generales, revelar patrones y conocimientos ocultos, descubrir grupos y relaciones. A su vez, disponer de herramientas adecuadas para el procesamiento de grandes volúmenes de datos (y en particular en tiempo real), permite la generación de distintos modelos que actúan en línea (por ejemplo, prevención de fraude, recomendaciones).

En ese sentido, las organizaciones tradicionales y de plataforma, utilizan los servicios de personalización para retener a los usuarios, lograr una mayor participación del cliente y, a la vez, obtener mayores ganancias (Kallinikos, 2019). Por lo que la personalización, es una modalidad presente en diversas áreas, por medio de la cual las organizaciones buscan estructurar la interacción con sus usuarios. Vale mencionar que la personalización también puede tener algunos efectos positivos en las organizaciones, ayudándolas en la toma de decisiones o para incorporar nuevos conocimientos sobre sus clientes (Anderson, 2006).

El objetivo del siguiente apartado consiste en identificar los métodos predictivos para la recomendación de productos financieros en la organización bancaria Santander. Para eso, se hará referencia, por un lado, al conjunto de datos con el cuál se va a trabajar y por el otro, a los algoritmos seleccionados para el desarrollo de un sistema de recomendación de productos financieros. En primera instancia se describirá el proceso de generación, obtención y procesamiento de los datos. Luego se llevará a cabo el análisis descriptivo del conjunto de datos. Finalmente se realizará la descripción conceptual de los modelos seleccionados a implementar.

2.1. Obtención y procesamiento de datos

El conjunto de datos sobre el cual se estará trabajando se obtuvo en formato CSV (*Comma-Separated Values*) mediante el lenguaje *Pyhton* en el entorno *Sypder*[2]. Este posee datos de tenencia de productos de los clientes del banco Santander para un período de un año y medio. Comienza en enero 2015 y tiene registros mensuales de productos que tiene un cliente, como ser tarjeta de crédito, caja de ahorros. A partir de estos se buscará saber que producto recomendarle a cada cliente en el último mes, junio 2016.

Los datos se encuentran disponibles en *Kaggle* y fueron publicados por la entidad Santander. *Kaggle* es una plataforma *web* que permite buscar o publicar bases de datos, explorar y construir modelos, y realizar competencias de distintos temas. En estas competencias, las empresas – como es el caso del banco Santander – publican problemas y los participantes compiten para construir el mejor algoritmo.

[2] *Spyder* (*Scientific Python Development Environment*) es un entorno de desarrollo interactivo para programación científica en el lenguaje Python.

El conjunto de datos tiene un total de 13.647.309 filas y 48 columnas. Posee 24 columnas de productos y 23 columnas con datos del cliente. A continuación, se detalla los datos contenidos en cada columna.

Tabla 1: Columnas del conjunto de datos.

Columna	Descripción
fecha_dato	fecha del registro
ncodpers	código del cliente
ind_empleado	índice de empleados: A activo, B ex empleado, F filial, N no empleado, P pasivo
pais_residencia	país de residencia del cliente
sexo	género del cliente
age	edad
fecha_alta	fecha de alta del cliente en la organización
ind_nuevo	índice de nuevos clientes: 1 si el cliente se registró en los últimos 6 meses.
antiguedad	antigüedad del cliente (en meses)
indrel	índice de relación: 1 (Primero/Principal), 99 (Cliente principal durante el mes pero no al final del mes)
ult_fec_cli_1t	última fecha como cliente principal (si no es a fin de mes)
indrel_1mes	tipo de cliente al comienzo del mes, 1 (primer cliente/principal), 2 (copropietario), P (potencial), 3 (antiguo principal), 4 (antiguo copropietario)
tiprel_1mes	tipo de relación del cliente al inicio del mes, A (activo), I (inactivo), P (antiguo cliente), R (Potencial), N (no se explicitó el significado de este valor por parte de la entidad)
indresi	índice de residencia (S (Sí) o N (No) si el país de residencia es el mismo que el del banco)
indext	índice de extranjería (S (Sí) o N (No) si el país de nacimiento del cliente es diferente al país del banco)
conyuemp	índice de cónyuge. 1 si el cliente es cónyuge de un empleado
canal_entrada	canal utilizado por el cliente para unirse
indfall	índice de fallecidos. N/D
tipodom	tipo de dirección. 1, dirección principal
cod_prov	código de provincia (dirección del cliente)
nomprov	nombre de la provincia
ind_actividad_cliente	índice de actividad (1, cliente activo; 0, cliente inactivo)
renta	ingreso bruto del hogar
segmento	segmentación: 01 - VIP, 02 - Particulares 03 - graduado universitario
ind_ahor_fin_ult1	caja de ahorro
ind_aval_fin_ult1	garantías
ind_cco_fin_ult1	cuentas actuales
ind_cder_fin_ult1	cuentas derivadas
ind_cno_fin_ult1	cuenta de nómina
ind_ctju_fin_ult1	cuenta junior
ind_ctma_fin_ult1	cuenta particular "más"
ind_ctop_fin_ult1	cuenta particular
ind_ctpp_fin_ult1	cuenta particular "plus"
ind_deco_fin_ult1	depósitos a corto plazo
ind_deme_fin_ult1	depósitos a mediano plazo
ind_dela_fin_ult1	depósitos a largo plazo
ind_ecue_fin_ult1	cuenta electrónica
ind_fond_fin_ult1	fondos
ind_hip_fin_ult1	hipoteca
ind_plan_fin_ult1	pensiones
ind_pres_fin_ult1	préstamos
ind_reca_fin_ult1	impuestos
ind_tjcr_fin_ult1	tarjeta de crédito
ind_valo_fin_ult1	valores
ind_viv_fin_ult1	cuenta de inicio
ind_nomina_ult1	nómina de sueldos
ind_nom_pens_ult1	pensiones
ind_recibo_ult1	débito directo

Fuente: elaboración propia

En una primera revisión del conjunto de datos se pudo detectar que el 2% (27.734) de los registros poseen valores faltantes en las columnas “pais_residencia”, “sexo”, “fecha_alta”, “ind_nuevo”, “indrel”, “indresi”, “indext”, “indfall”, “ind_actividad_cliente”. Al tener tantas columnas con valores faltantes se procede a eliminar estos registros del conjunto de datos. Además, las columnas “conyuemp” y “ult_fec_cli_1t” contienen mayoría de valores nulos (99,7% y 99,6% respectivamente) por lo que también se eliminan del conjunto.

La columna “tipodom”, no será utilizada ya que se cuenta con la variable código y nombre de la provincia como datos de ubicación del cliente. Realizadas estas transformaciones, la nueva dimensión del conjunto de datos es 13.619.575 registros con 45 columnas. Se procede con la revisión de las columnas a continuación.

A partir de la variable "fecha_dato" que contiene día, mes y año, se construye una nueva (periodo) que contiene el año y mes (por ejemplo 201501). En cuanto a la variable “ind_empleado” que posee valores "N”, "A", "B", "F", "S" se transforman a categorías de 1 a 5. Para la variable “sexo” también se pasan sus valores (“H” y “V”) a categorías (1 y 2) y los valores faltantes (70) se imputan con el valor más frecuente que es 1 (“H”).

La variable “age” presenta *outliers*, valores por debajo de 18 y también mayores a 100. Entonces, los valores menores a 18 son reemplazados por la media calculada a partir de clientes que tienen entre 18 y 30 años, y los valores mayores a 100 se reemplazan por la media calculada con clientes que tienen entre 30 y 100 años. En lo que refiere al campo “fecha_alta”, no se utilizará ya que se cuenta con la antigüedad en meses. En ese sentido, la columna “antigüedad” presenta 38 casos con valores -999999 que se reemplazan por 0.

La columna “indrel_1mes”, como indica la tabla 1, presenta valores numéricos (1 a 4) y el valor “P”. Se opta por transformar este valor a numérico, asignándole el valor 5 y se convierten todos los valores numéricos (1 a 5) en tipo categoría. En consecuencia, los valores faltantes (122.047) se imputan con la categoría más frecuente que es 1.

Lo mismo sucede con la columna “tiprel_1mes”, los valores ("A", "I", "P", "R", "N") son llevados a categorías de 1 a 5. Se obtiene entonces que la categoría 1 refiere a aquellos clientes con tipo de relación “activa” al principio de mes, la categoría 2 se utiliza para indicar el tipo de relación “inactiva”, la categoría 3 para el tipo “antiguo cliente”, el tipo de relación " potencial” es indicado con la categoría 4, y "N" (no se explicitó el significado de este valor por parte de la entidad) con la categoría 5. A los 122.047 valores faltantes se le imputa la categoría más frecuente (1).

En cuanto a los datos de lugar de residencia del cliente, las variables “indresi” e “indext”, se transforman en categorías. Se reemplazó el valor “N” (no) por la categoría 0, y el valor “S” (si) por la categoría 1. Y sobre la variable “país_residencia” se genera una nueva columna indicando si el país es España o no (1 o 0). De esta manera, la variable original ya no será utilizada.

Para la columna “canal_entrada” se imputan los valores faltantes por “KHE”, ya que resulta ser el valor más frecuente. También en base a esta se generan 4 nuevas columnas: “canal_entreada_KHE”, “canal_entrada_KAT”, “canal_entrada_KFC”, “canal_entrada_OTRO” indicando si la entrada fue por los canales principales (en el caso de los tres primeros) o por otro. De este modo, la columna “canal_entrada” no se utilizará.

En relación a la variable “indfall”, se transforma los valores ‘N’ a 0 y ‘S’ a 1. Respecto a la variable “cod_provincia” se imputan los valores faltantes (65.857) con el más frecuente (28, Madrid). Como ya que se cuenta con el código de provincia, la variable “nomprov” no será utilizada.

Las restantes variables de cliente refieren a “ind_actividad_cliente”, “renta” y “segmento”. Para la primera, no fue necesaria ninguna transformación. En cuanto a “renta” presenta 2.766.641 valores nulos. Para imputar estos valores primero se calcula la media de renta por provincia. Luego, este valor obtenido se imputa a los registros que no poseen renta, según a qué provincia pertenezcan. Y para la variable “segmento”, se generan las categorías 1 (top), 2 (particulares), 3 (universitario).

Finalizando con el procesamiento, se observan valores faltantes en las variables de productos “ind_nom_pens_ult1” e “ind_nomina_ult1” los cuales son imputados por 0. Se obtiene finalmente que el conjunto de datos quedó compuesto por 13.619.575 registros y 45 columnas siendo estas todas variables de tipo numéricas. En el próximo apartado se desarrolla el análisis exploratorio de los datos.

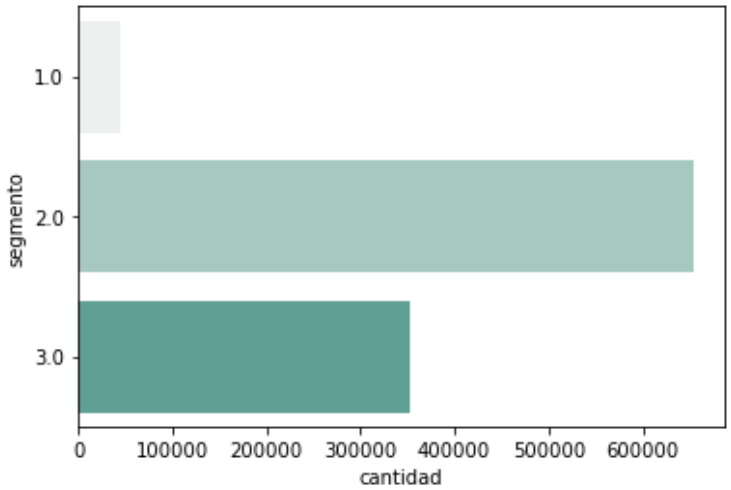
2.2. Análisis descriptivo de datos

Habiendo procesado y transformado el conjunto de datos, resulta necesario analizar las variables para comprender de mejor manera los datos con los cuales se trabajará. Al contar con datos de clientes y de productos, se visualizan variables de ambos lados que permitan obtener un panorama global. Se seleccionaron aquellos productos que más poseen los clientes, “ind_cco_fin_ult1”, “ind_ctop_fin_ult1”, “ind_recibo_ult1”. Estas variables serán contrapuestas con distintos datos de clientes.

En una primera instancia se puede observar en la figura 2 como resulta ser la distribución de los clientes por segmento. La mayoría de los clientes se encuentra en el segmento “Particulares” (2) con 654.282 clientes únicos. En segundo lugar, se encuentra el segmento “Universitarios”

(3) con 353.158, y la menor cantidad de clientes (43.838) se encuentran en el segmento “Top” (1). Esto tiene sentido ya que, para pertenecer a los segmentos prioritarios, los bancos exigen cumplir más requisitos al cliente.

Figura 2. Distribución de clientes por segmento

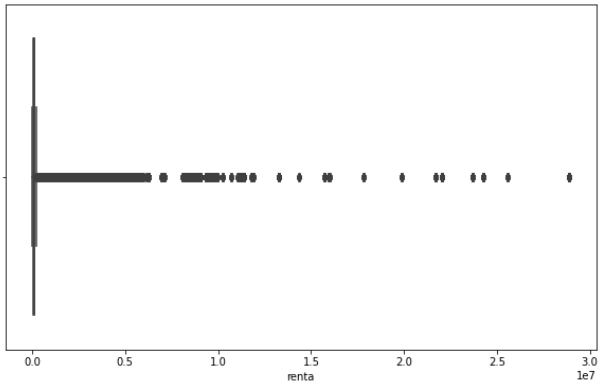


Fuente: elaboración propia.

Profundizando un poco más en los segmentos, se obtuvo que para el segmento “Top”, los productos más otorgados al cliente son "cuentas actuales", "depósitos a largo plazo" y "cuenta electrónica". Para el segmento “Particulares”, "cuentas actuales", también se ubica como el producto con más altas, seguido por "cuenta particular" y "débito directo". En cuanto al segmento “Universitario”, luego de "cuentas actuales", se encuentra "débito directo" y "cuenta de nómina".

Otra variable que puede resultar interesante al momento de ofrecerle un nuevo producto a un cliente es la renta. Como se muestra en la figura 3, esta variable presenta un rango de valores muy amplio, desde 1.202 hasta 28.894.395. Dada la dispersión de valores que posee, esta variable será estandarizada al momento de aplicar los métodos predictivos. Esto implica, para cada registro, restarle la media de la columna renta y luego dividir el valor por el desvío estándar.

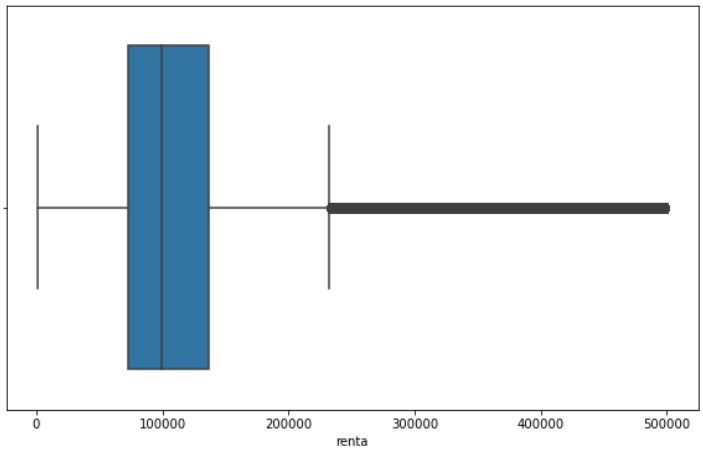
Figura 3. Distribución de la variable renta.



Fuente: elaboración propia.

Cómo la variable renta posee un rango muy amplio, se puede realizar un análisis más acotado. Obteniendo así que la concentración de valores ocurre entre 100.000 y 200.000 como muestra la figura 4. Es decir, el ingreso bruto del hogar que tienen la mayoría de los clientes del banco está comprendido en este rango de valores.

Figura 4. Concentración de valores de la variable renta.



Fuente: elaboración propia.

Con el objetivo de caracterizar a los clientes, se propone agruparlos mediante *Clustering*. Para esto se aplica el algoritmo K-medias (*K-means*), siendo este una técnica de aprendizaje automático no supervisado utilizada para identificar grupos de objetos de datos en un conjunto de datos. Es decir, se busca identificar clientes similares en cada grupo (*Cluster*), y que a la vez estos grupos sean lo más diferentes entre sí (Hartigan, 1979).

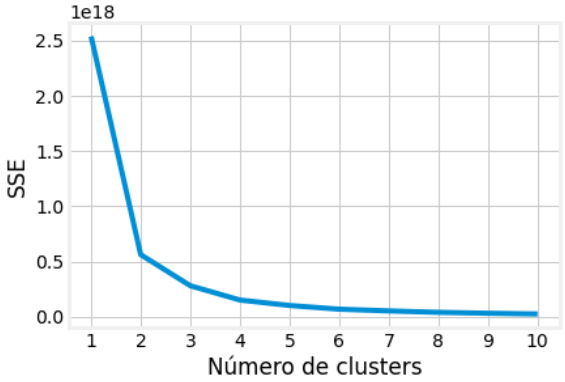
El algoritmo es iterativo y consta de pocos pasos. En primer lugar, se seleccionan los centroides para cada grupo, el centroide es el punto que representa el centro del grupo y la cantidad grupos debe ser definida con anterioridad. Luego se asigna cada observación al

centroide más cercano (mediante el cálculo de distancia), y se vuelve a calcular el centroide. Finaliza cuando la posición del centro no cambia. Existen varios cálculos de distancia, en este caso se utiliza la distancia euclidiana (Bora, 2014), donde dado dos puntos, a y b, con k grupos se calcula como $\sqrt{\sum_{j=1}^{k}(a_j - b_j)^2}$.

Para obtener el número óptimo de grupos (*Clusters*) se realizan varias ejecuciones del algoritmo. En cada ejecución, se incrementa el número de grupos (k) y se evalúa su performance, según SSE (*Sum Of The Squared Error*). SSE es la suma de las diferencias al cuadrado entre cada observación y la media de su grupo, calculada como $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Donde n es el número de observaciones, x_i es el valor de la i-ésima observación y 0 es la media de todas las observaciones (Draper & Smith, 1998).

Mediante el método del codo (Syakur, 2018) se busca un valor k óptimo en base a la performance obtenida. Es decir, la generación de grupos más allá del óptimo, no generan valor adicional. Por lo cual, en la figura 5, se identifica k = 2 cómo número óptimo de *Clusters*.

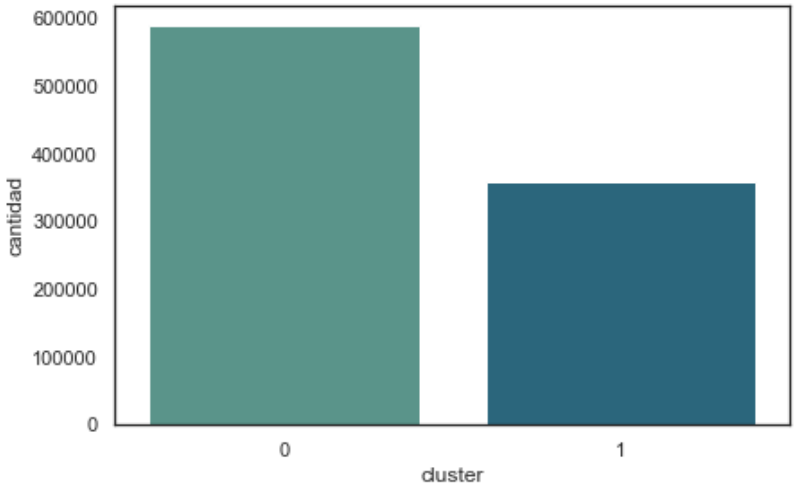
Figura 5. Número óptimo de grupos.



Fuente: elaboración propia.

Definido el número de grupos, se procede a aplicar el algoritmo de *Cluster* y se asigna cada observación a un grupo. Para ello resulta necesario estandarizar la variable renta. En la figura 6 se observa que el 62% de clientes (589.736) fueron asignados al *Cluster* 0. Mientras que 359.878 (38%) clientes pertenecen al *Cluster* 1.

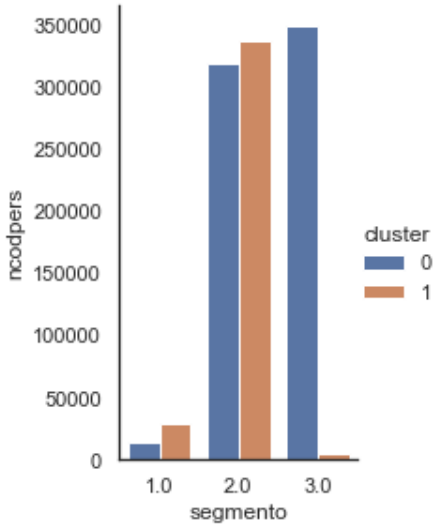
Figura 6. Asignación de clientes a cada grupo.



Fuente: elaboración propia.

Al realizar la apertura de los *Clusters* por segmento se advierte en la figura 7, que el *Cluster* 0 agrupa a clientes de los tres segmentos. En su mayoría son clientes del segmento 3 (Universitarios) y 2 (Particulares). En cambio, el *Cluster* 1 contiene muy pocos clientes del segmento Universitario y en su mayoría son clientes del segmento Particulares.

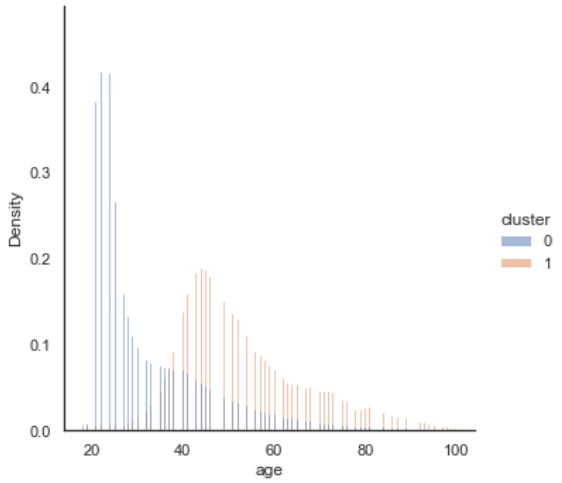
Figura 7. Segmentos según *Cluster*.



Fuente: elaboración propia.

En la figura 8, se marca la distribución por edad en cada *Cluster*. Donde los clientes pertenecientes al grupo 1, son de mayor edad y, su mayor concentración, se da entre 40 y 50 años. Respecto al grupo 0, contiene a los clientes más jóvenes donde la mayoría tiene entre 20 y 30 años.

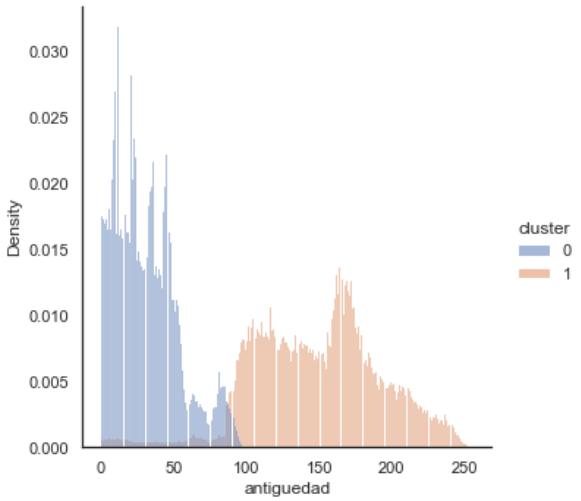
Figura 8. Distribución de edad según *Clusters*.



Fuente: elaboración propia.

Un comportamiento similar sucede con la variable "antigüedad" como se muestra en la figura 9. Para el *Cluster* 0 se observan clientes con antigüedad menor a 8 años (100 meses). Y el *Cluster* 1, si bien contiene clientes con antigüedad menor a 100 meses, la mayoría son clientes con mayor antigüedad.

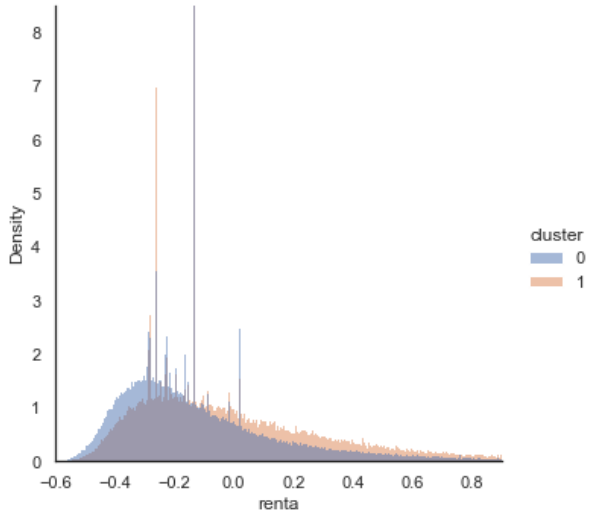
Figura 9. Distribución de antigüedad según *Clusters*.



Fuente: elaboración propia.

En cuanto a la variable renta, en la figura 10 se contempla su distribución. Donde en el *Cluster* 0 se encuentran la mayoría de los clientes, pero con una menor renta. Mientras que en el *Cluster* 1, los clientes poseen mayor renta.

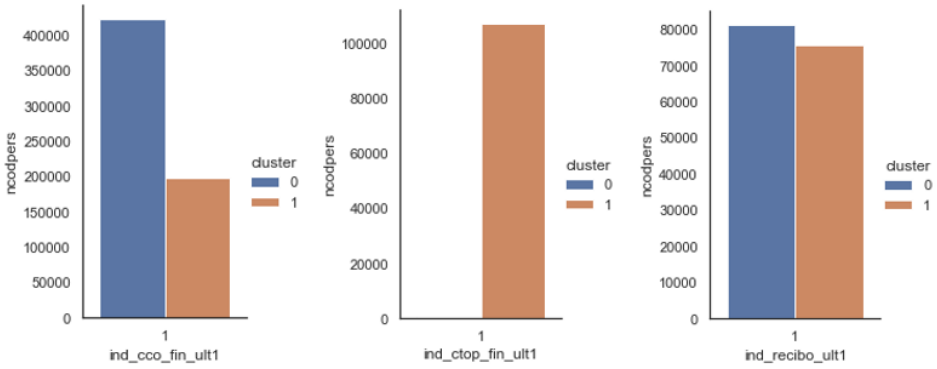
Figura 10. Distribución de renta (estandarizada) según *Clusters*.



Fuente: elaboración propia.

Respecto a los 3 productos que más poseen los clientes, en la figura 11 es posible observar que el producto "cuentas actuales" tiene más presencia en el *Cluster* 0. El producto "cuentas particulares" se encuentra sólo en clientes del *Cluster* 1. Mientras que para el producto "débito directo", se encuentra en ambos Clusters con cantidad de clientes similares.

Figura 11. Top 3 de productos y su distribución en cada *Cluster*.



Fuente: elaboración propia.

Por tanto, en base a los resultados obtenidos, se puede definir que el *Cluster* 0 contiene más clientes que el *Cluster* 1. Siendo estos en su mayoría del segmento Universitarios y Particulares. A su vez, en este clúster se encuentran clientes más jóvenes y con menor antigüedad en la organización. Vale mencionar que, en este grupo, no hay presencia del producto "cuentas particulares".

En contrapartida, el *Cluster* 1 concentra a clientes que en su mayoría son del segmento Particulares, y en menor medida del segmento Top, pero muy pocos del segmento Universitario. Al contener más cantidad de clientes del segmento Top que el *Cluster* 0, en este *Cluster* se observan a los clientes con una mayor renta. También, los clientes de este grupo son más antiguos y con una mayor edad.

Mediante el análisis realizado, se pudo clasificar a los clientes de la organización Santander en dos grupos con distintas características. Para esto se aplicó el algoritmo k-medias como técnica de *Clustering*. De este modo, se pudo comprender cuales serían los productos más vendidos en cada uno. Llegado a este punto, resulta necesario identificar los métodos adecuados que serán utilizados para la determinación de un sistema de recomendación que permita establecer el producto de mayor preferencia por cada cliente.

2.3. Métodos predictivos para un sistema de recomendación en organizaciones bancarias

La recomendación de productos es uno de los problemas más relevantes en el ámbito de la personalización y experiencia del cliente (Melnikov V., 2016). Brindar un entorno y experiencias personalizadas a los usuarios es un tema de interés para todas las organizaciones, incluidas las del entorno financiero. Los bancos consideran seriamente las implementaciones de *Machine Learning* con el fin de aprovechar los datos a su disposición para proporcionar servicios y experiencias personalizadas a sus clientes. Uno de los campos en los que se apoyan estas implementaciones, son los sistemas de recomendaciones (Gigli, 2017).

Para el desarrollo de estos sistemas existen algoritmos que generan un ranking de productos a recomendarle a cada cliente. En este caso lo que se obtiene no es un ranking del estilo 1, 2, 3..., sino valores que se ordenan para generar el ranking. Para este tipo de algoritmo se consideran los enfoques *Pairwise* y *Listwise*.

A la vez, existen otros tipos de algoritmos que generan como resultado una probabilidad de compra de cada producto para cada cliente. El producto más recomendable para cada cliente se obtiene ordenando estas probabilidades. Para este tipo de algoritmo se considera el enfoque *Pointwise.* A continuación, se desarrollan estos los tres enfoques propuestos.

Para explicar los tres métodos, se define a q como el cliente para el cual se tienen n productos a ser rankeados por relevancia, $D = \{d_1, \dots, d_n\}$. Los elementos de entrada de los modelos son

del estilo $x_1 = (q, d_i)$, donde para cada elemento existe un *score* de relevancia $s_i = f(x_i)$ generado por el propio modelo. La función de pérdida (por la cual se calcula el error de modelo) es el elemento distintivo de los tres enfoques (Casalegno, 2022).

Respecto al ranking mediante *Pointwise*, la función de pérdida mide directamente la distancia entre la relevancia real y_i y el s_i predicho por el modelo. Es decir, la relevancia se calcula para cada par cliente-producto. Para esto, se pueden usar algoritmos de aprendizaje automático de clasificación o regresión conocidos (Melnikov V., 2016). Como propone Cossock (2006), se puede considerar MSE (*Mean Squared Error*) como función de pérdida:

$$L(s, y) = \sum_{i=1}^{n} (s_i - y_i)^2$$

En cuanto al enfoque *Pairwise,* la pérdida total se calcula como la suma de los términos de pérdida en cada par de productos d_i, d_j para $i, j = 1 \ldots n$. Este enfoque trabaja con preferencia relativa (y no absoluta como el enfoque *Pointwise*), donde dados dos productos, el objetivo es predecir si $y_i > y_j$ o no. Es decir, cuál de los dos productos es más relevante. Por lo cual, se transforma en un problema de clasificación binaria (Casalegno, 2022).

Burges (2010) desarrolló este enfoque utilizando BCE[3] (*Binary Cross Entropy*) como función de pérdida en *RankNet*. Si bien *RankNet* es considerado una mejora respecto a los métodos *Pointwise*, los productos siguen teniendo la misma relevancia en el proceso de entramiento. En ese sentido, surge *LambdaRank* como evolución de *RankNet*, donde se utiliza el descenso del gradiente para que los productos más relevantes tengan gradientes más altos. La función de pérdida se define de la siguiente manera:

$$\lambda_j := \frac{\partial L}{\partial_{s_j}} = \frac{1}{G_{max}} \sum_{i \neq j} \sigma(s_i - s_j) \left| \boldsymbol{G_i} - \boldsymbol{G_j} \right| \left| \boldsymbol{D_i} - \boldsymbol{D_j} \right|$$

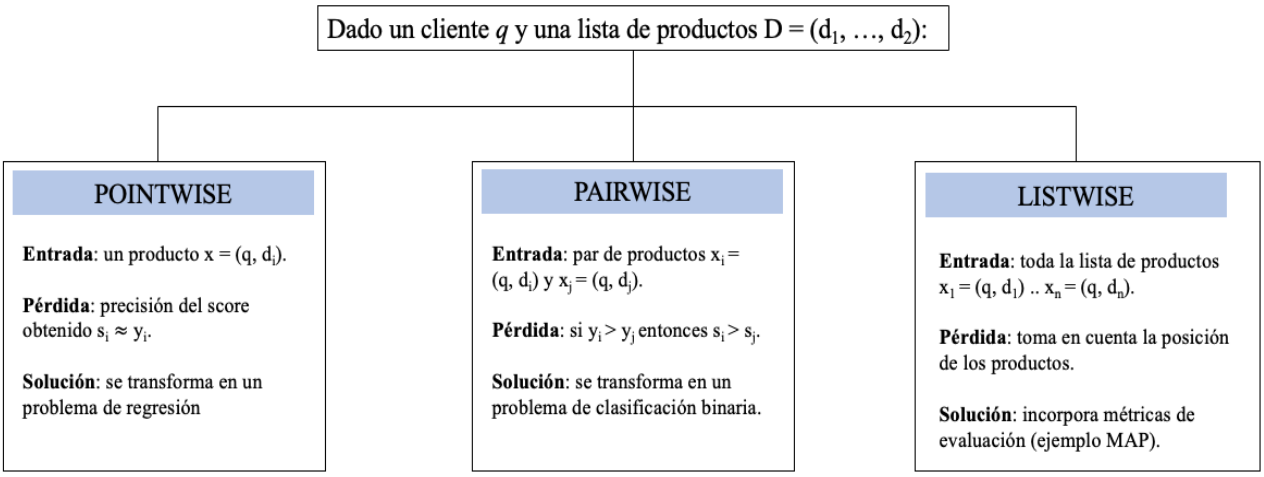
Además, Burges también desarrolla una evolución de *LambdaRank,* llamada *LambdaMart*. La cual se basa en una familia de modelos denominada MART (Árboles de Regresión Aditiva Múltiple). Esta implementación con *Boosted Trees*, generó mejores resultados que *LambdaRank.*

[3] BCE: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8943952

En los algoritmos con enfoque *Listwise,* la función de pérdida se calcula directamente con la lista completa de productos. Donde para cada cliente se intenta encontrar el orden óptimo de productos. A diferencia de los métodos *Pointwise* y *Pairwise* que transforman el problema en una regresión o clasificación binaria, los métodos *Listwise* resuelven el problema de manera más directa maximizando la métrica de evaluación.

En la figura 12, se resumen los tres enfoques desarrollados. Para el método *Pointwise* se toma sólo un producto para cada cliente y se calcula la precisión del score respecto al valor real. En el enfoque *Pairwise* se toma de a pares de producto y se busca encontrar cuál de los dos es más relevantes, transformándose en un problema de clasificación binaria. Mientras que en *Listwise* la relevancia se genera teniendo en cuenta toda la lista de productos para cada cliente, buscando ordenarla según relevancia.

Figura 12. Métodos para ranking de productos.



Fuente: elaboración propia en base a (Casalegno, 2022).

De esta manera, en este capítulo se realiza el procesamiento y transformación del conjunto de datos de la organización Santander obteniendo que el conjunto queda formado por 13.619.575 registros y 45 columnas. Luego se lleva a cabo el análisis descriptivo, en dónde se divide a los clientes del banco en dos grupos (grupo 0 y grupo 1). El grupo 0, está formado por clientes más jóvenes, de menor antigüedad y con menor renta. En este grupo se encuentran la mayoría de los clientes y, además, no hay presencia del segundo producto más vendido, “cuentas particulares”.

Respecto al grupo 1, se encuentran gran cantidad de clientes del segmento Top, por lo que el *Cluster* tiene clientes de mayor renta, y que a su vez son más antiguos en la organización. El producto "cuentas particulares", es el de más presencia en este *Cluster* (de los tres productos más vendidos en todo el conjunto). Estos *Clusters* obtenidos para cada cliente, se incorpora como una variable más del conjunto de datos, obteniendo así 13.619.575 registros y 46 columnas.

Por último, se desarrollan los tres enfoques (*Pointwise, Pairwise, Listwise*) que se utilizarán a continuación para la implementación de un sistema de recomendación de productos. Por lo que, en el siguiente capítulo, se realizará la implementación de los modelos y se evaluarán los resultados obtenidos. Finalmente, se analizará como impacta el sistema de recomendación en el diseño de una estrategia de venta.

3. Recomendación eficiente de productos financieros

La personalización es la capacidad de ofrecer contenido y servicios que se adaptan a las personas en función del conocimiento sobre sus preferencias y comportamiento (Hagen, 1999). Esta permite construir lealtad con el cliente a través de la comprensión de las necesidades de cada individuo (Riecken, 2000). Una manera de llegar a lograrlo es con el apoyo de los sistemas de recomendación. Estos producen recomendaciones que están destinadas a optimizar la experiencia del usuario (Burke, 2011).

Los sistemas de recomendación son una manera de automatizar la personalización masiva de ofertas a clientes. Esta automatización es posible mediante el uso de software, como *Python*, para entonces obtener el producto más relevante para cada cliente a través de un sistema de recomendación. De esta manera han cobrado mayor relevancia ya que para las organizaciones es muy valioso generar una relación de valor a largo plazo con los clientes (Schafer, 1999).

Frente a esta ventaja, el objetivo del presente capítulo consiste en identificar el producto financiero de mayor preferencia por cliente de la organización Santander. Para ello se implementará un sistema de recomendación mediante la utilización modelos de aprendizaje automático. Esto constituye un valor agregado para el diseño eficiente de una estrategia de venta para la organización. Para eso, se desarrollará, en primera instancia, la implementación de los modelos para la recomendación de productos. Luego se llevará a cabo el análisis de resultados

obtenidos para cada modelo y se seleccionará el que mejor performance obtenga. Por último, se mencionará la importancia de contar con un sistema de recomendaciones al momento de ofrecer productos financieros.

3.1. Implementación de algoritmos de aprendizaje automático para la recomendación de productos

Habiendo presentado los métodos a aplicar y su funcionamiento en el apartado 2.3, se realiza su aplicación. Para ello se utilizará el conjunto de datos presentado y descripto en los apartados 2.1 y 2.2. Debido a la gran cantidad de registros que posee, se requiere mucho poder de cálculo para transformarlo al formato requerido por el modelo.

El formato requerido implica tener al cliente-periodo (y todas sus variables) repetido para cada producto, indicando en otra columna si tenía el producto en ese mes o no (valor 1 o 0 respectivamente). Por lo tanto, se seleccionan 3 períodos (agosto, septiembre 2015 y enero 2016) para utilizar como muestra más acotada. Como se observa en la tabla 2, en estos períodos seleccionados, los clientes y productos (en cantidades) muestran un comportamiento normal respecto al resto de períodos.

Tabla 2. Selección de períodos para desarrollo.

periodo	cantidad de productos	cantidad de clientes
201501	1.102.494	618.504
201502	1.106.946	621.454
201503	1.114.056	624.118
201505	1.116.969	626.075
201504	1.121.059	628.360
201506	1.130.732	630.249
201507	1.147.960	829.817
201508	1.151.798	843.201
201509	1.173.705	865.440
201510	1.198.063	892.251
201511	1.207.748	906.109
201512	1.220.647	912.021
201601	1.206.284	916.269
201602	1.222.968	920.904
201603	1.230.966	925.076
201604	1.233.896	928.274
201605	1.240.538	931.453

Fuente: elaboración propia.

A partir de la muestra seleccionada, se crean tres subconjuntos: uno de entrenamiento, otro de testeo y un tercero de validación. El conjunto de datos de entrenamiento es el que se utiliza para entrenar el modelo. Mientras que el conjunto de validación también se utiliza en el proceso de entrenamiento, pero con el objetivo de ajustar los hiper-parámetros del modelo. Y el conjunto

de testeo se utiliza para evaluar la performance del modelo, aplicado a un conjunto de datos nunca visto.

De esta manera se obtiene, para un cliente-período, el listado de los 24 productos indicando cuál posee en este. En la tabla 3 se ejemplifica el formato necesario para un cliente (cliente con ID 15889) en el periodo de entrenamiento agosto 2015. Vale mencionar que las variables restantes de ese período se mantienen fijas en estos 24 registros.

Tabla 3. Ejemplo transformación al formato necesario.

ncodpers	periodo	producto	compro
15889	201508	ind_ahor_fin_ult1	0.0
15889	201508	ind_aval_fin_ult1	0.0
15889	201508	ind_cco_fin_ult1	1.0
15889	201508	ind_cder_fin_ult1	0.0
15889	201508	ind_cno_fin_ult1	0.0
15889	201508	ind_ctju_fin_ult1	0.0
15889	201508	ind_ctma_fin_ult1	0.0
15889	201508	ind_ctop_fin_ult1	0.0
15889	201508	ind_ctpp_fin_ult1	1.0
15889	201508	ind_deco_fin_ult1	0.0
15889	201508	ind_deme_fin_ult1	0.0
15889	201508	ind_dela_fin_ult1	0.0
15889	201508	ind_ecue_fin_ult1	0.0
15889	201508	ind_fond_fin_ult1	0.0
15889	201508	ind_hip_fin_ult1	0.0
15889	201508	ind_plan_fin_ult1	0.0
15889	201508	ind_pres_fin_ult1	0.0
15889	201508	ind_reca_fin_ult1	0.0
15889	201508	ind_tjcr_fin_ult1	0.0
15889	201508	ind_valo_fin_ult1	1.0
15889	201508	ind_viv_fin_ult1	0.0
15889	201508	ind_nomina_ult1	0.0
15889	201508	ind_nom_pens_ult1	0.0
15889	201508	ind_recibo_ult1	0.0

Fuente: elaboración propia.

Es necesario indicarle al modelo los productos y su relevancia (si tiene el producto o no) que corresponden al cliente para cada período, por eso surge el concepto de grupo. Es decir, es una manera de identificar la situación del cliente en los distintos períodos. El identificador de grupo se define como código de cliente + últimos tres dígitos del periodo. En la tabla 4 se ejemplifica para el cliente con ID 15889, en el período de entrenamiento agosto 2015, su código de grupo es 15889508 donde 15889 corresponde al id cliente, 5 al año y 08 al mes.

Tabla 4. Se incorpora el “groupid”.

ncodpers	periodo	groupid	producto	compro
15889	201508	15889508	ind_aval_fin_ult1	0.0
15889	201508	15889508	ind_deco_fin_ult1	0.0
15889	201508	15889508	ind_ctpp_fin_ult1	1.0
15889	201508	15889508	ind_dela_fin_ult1	0.0
15889	201508	15889508	ind_tjcr_fin_ult1	0.0
15889	201508	15889508	ind_deme_fin_ult1	0.0
15889	201508	15889508	ind_ctma_fin_ult1	0.0
15889	201508	15889508	ind_valo_fin_ult1	1.0
15889	201508	15889508	ind_pres_fin_ult1	0.0
15889	201508	15889508	ind_plan_fin_ult1	0.0
15889	201508	15889508	ind_ecue_fin_ult1	0.0
15889	201508	15889508	ind_cco_fin_ult1	1.0
15889	201508	15889508	ind_nomina_ult1	0.0
15889	201508	15889508	ind_recibo_ult1	0.0
15889	201508	15889508	ind_ctop_fin_ult1	0.0
15889	201508	15889508	ind_viv_fin_ult1	0.0
15889	201508	15889508	ind_hip_fin_ult1	0.0
15889	201508	15889508	ind_nom_pens_ult1	0.0
15889	201508	15889508	ind_fond_fin_ult1	0.0
15889	201508	15889508	ind_cder_fin_ult1	0.0
15889	201508	15889508	ind_reca_fin_ult1	0.0
15889	201508	15889508	ind_ahor_fin_ult1	0.0
15889	201508	15889508	ind_ctju_fin_ult1	0.0
15889	201508	15889508	ind_cno_fin_ult1	0.0

Fuente: elaboración propia.

Como se observa en la tabla 4, los productos están identificados con un nombre. Este formato de campo no es soportado por el modelo, por lo que se generan 24 columnas (una por cada producto) indicando con 1 o 0 a cuál pertenece. Por otra parte, la columna "compro", que indica si tiene el producto o no, se separa del conjunto de datos ya que es el valor por predecir.

Antes de comenzar con el entrenamiento de los distintos modelos se procede a buscar los mejores parámetros. Para ello, se utiliza el algoritmo *XGBoost*[4]. A continuación, se explica su funcionamiento.

XGBoost es un algoritmo predictivo supervisado que utiliza el principio de *Boosting* (Friedman, 2001). La noción detrás de este es generar múltiples modelos de predicción, y que cada uno de estos tome los resultados del modelo anterior, para generar un modelo con mejor poder predictivo y mayor estabilidad en sus resultados. Para conseguirlo, emplea el algoritmo de optimización *Gradient Descent* (descenso de gradiente) (Chen, 2017).

Durante el entrenamiento con *XGBoost*, los parámetros de cada modelo son ajustados iterativamente tratando de encontrar el mínimo de una función objetivo y cada modelo es comparado con el anterior. Si un nuevo modelo tiene mejores resultados, entonces se toma este

[4] Documentación *XGBoost*: https://xgboost.readthedocs.io/en/stable/index.html

como base para realizar nuevas modificaciones. Si tiene peores resultados, se regresa al mejor modelo anterior y se modifica.

Este proceso se repite hasta llegar a un punto en el que la diferencia entre modelos consecutivos es insignificante. Esto indica que se ha encontrado el mejor modelo posible. Si no, sucede cuando se llega al número de iteraciones máximas definida por el usuario. *XGBoost* usa como modelos árboles de decisión de diferentes tipos, que pueden ser usados para tareas de clasificación y de regresión (Chen G. , 2016).

Para realizar la optimización de parámetros y entrenamiento, los conjuntos de entrenamiento, testeo y validación son transformados a matriz con la función *DMatrix*[5]. Esta función genera una estructura de datos interna utilizada por *XGBoost*, que optimiza el uso de la memoria y la velocidad de entrenamiento. También, es necesario indicar el tamaño de los grupos en la matriz de entrada del modelo. En este caso el tamaño del grupo es para todos 24 (productos).

Se excluyen de las variables predictoras: "ncodpers", "periodo", y "groupid" (ya que la matriz tiene asociados todos los tamaños de los grupos). El identificador de grupo será necesario nuevamente en la etapa de evaluación del modelo. La optimización de parámetros es realizada mediante *RandomizedSearchCV*[6]. Este método recibe una serie de valores de parámetros y realiza varias ejecuciones encontrando la mejor combinación.

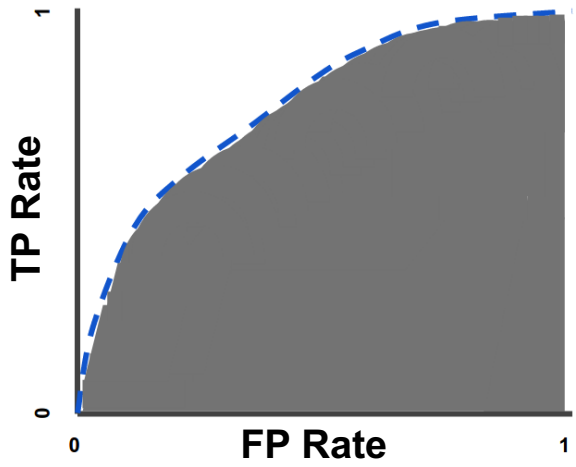
Una vez que se obtiene la mejor combinación de parámetros para cada modelo, se ejecuta el entrenamiento para los tres enfoques propuestos. La diferencia entre cada enfoque está dada por la función objetivo y métrica de evaluación de entrenamiento. Para el método *Pointwise* se declara como función objetivo "*binary:logistic*" y métrica de evaluación AUC. Mientras que para la prueba *Pairwise* se utiliza la función objetivo "*rank:pairwise*" y para *Listwise* "*rank:map*". El entrenamiento de estos dos métodos es evaluado según *Mean Average Precision* (MAP). A continuación, se explican los métodos de evaluación y se analizan los resultados obtenidos.

3.2. Análisis y evaluación de resultados

[5] Documentación de la función *DMatrix*: https://xgboost.readthedocs.io/en/stable/python/python_api.html
[6] Documentación *RandomizedSearchCV*: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

El entrenamiento del modelo binario (*Pointwise*) fue evaluado mediante AUC. Esta métrica, representa el área bajo la curva ROC (*Receiver Operating Characteristic*), siendo una medida de rendimiento para los problemas de clasificación. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Es decir, cuánto es capaz el modelo de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir 0 clases como 0 y 1 clases como 1. Como se observa en la figura 13, la curva ROC se traza con TPR (*True Positive Rate*) contra FPR (*False Positive Rate*), donde TPR está en el eje y y FPR está en el eje x (Wu, 2005).

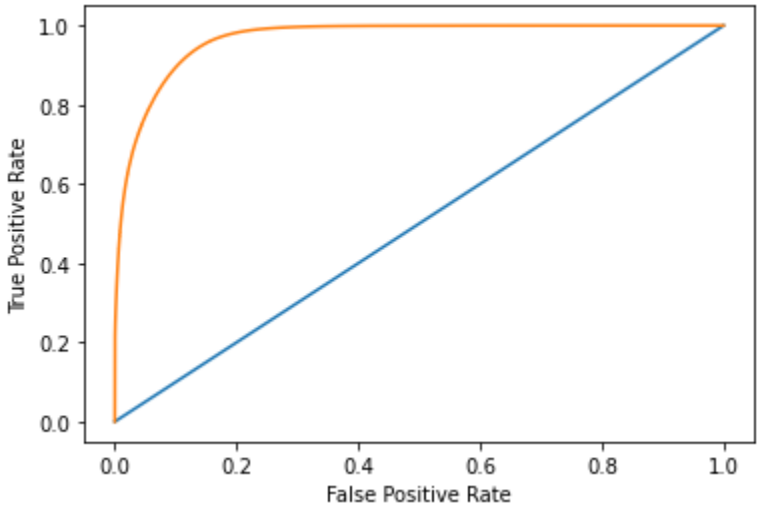
Figura 13. Área bajo la curva ROC.



Fuente: Obtenido de *Google Developers* (Classification: ROC Curve and AUC).

Al aplicar el modelo se obtuvo AUC de 0.9665, lo cual significa que el modelo clasifica correctamente los productos relevantes y no relevantes para cada clientes. Para los conjuntos de test y validación, el AUC obtenido fue de 0.9669 y 0.9652 respectivamente. En la figura 14 se grafica la curva ROC obtenida en entrenamiento.

Figura 14. Curva ROC entrenamiento.



Fuente: elaboración propia.

En la figura 14, se traza sobre el eje x la tasa de falsos positivos y en el eje y, la tasa de verdaderos positivos para las predicciones generadas por el modelo. La curva representa que la capacidad de separabilidad es buena, ya que es cercana a 1. Mientras que la diagonal en azul corresponde al caso donde el modelo no tiene capacidad para separar los productos relevantes de los no relevantes. Esta se generó prediciendo todos valores 0.

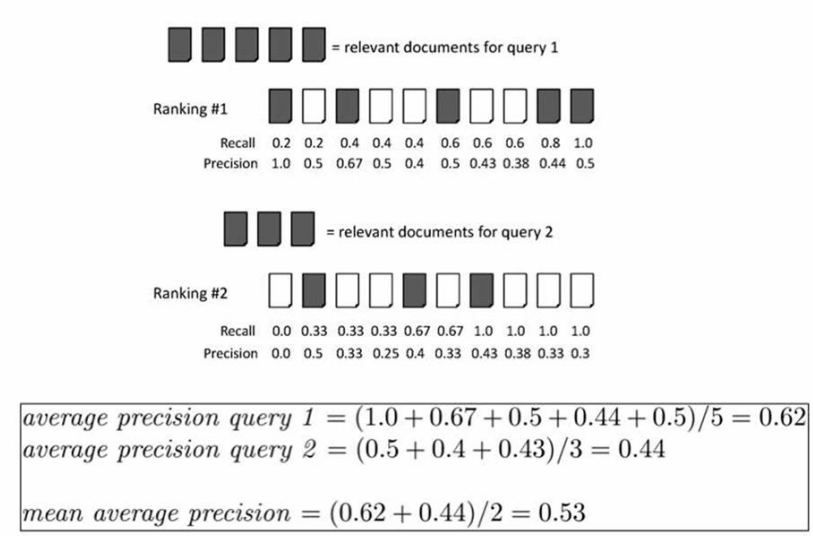
Sin embargo, para poder establecer una comparación con los dos enfoques restantes, es necesario evaluar el modelo *Pointwise* con otra métrica. Esto se debe a que los métodos *Pairwise* y *Listwise* devuelven como resultado un ranking. Al evaluar modelos de ranking, no es posible usar la precisión de una predicción para problemas de clasificación o regresión. Esto se debe a que las métricas como, precisión (*Accuracy*), *Root Mean Squared Error*, *Mean Absolute Error*, *Precision*, *Recall* y F1 *Score*, no son suficientes porque solo calculan las brechas entre el valor real versus la puntuación prevista.

En problemas de ranking, lo importante es el orden de los productos más que el valor obtenido. Una manera de evaluar este tipo de problemas es mediante MAP (*Mean Average Precision*). Donde el AP (*Average Precision*) es una medida que indica que tan bien quedan los productos relevantes en la posición más alta para cada cliente (Liu P., 2017). EL AP se calcula para cada cliente-periodo (por “groupid”) y el MAP es el promedio de todos los AP.

En la figura 15 se ejemplifica el cálculo de esta métrica. Donde para el cliente 1 (*Query* 1) se tienen 5 productos relevantes sobre un total de 10 productos. El AP se calcula como la suma de la precisión obtenida para cada producto relevante y se divide por la cantidad de productos

relevantes. Lo mismo sucede para el cliente 2 (*Query* 2), para luego obtener el MAP promediando los dos AP obtenidos.

Figura 15. Calculo MAP.



Fuente: Obtenido de Victor Lavrenko (Text Technologies - Evaluation, 2009) .

Además, con el objetivo de tener un valor base con el cual comparar la métrica de los tres modelos, se desarrolla un modelo *Baseline* (modelo base). Los modelos a desarrollar deberían tener una performance superior a este para ser considerados como buenos. El modelo *Baseline* es una regresión logística[7] cuya implementación se realizó sin ajuste de parámetros. La performance obtenida fue MAP: 0.69 en entrenamiento y testeo, 0.70 en validación.

Por tanto, es posible comparar los tres métodos entre sí y también respecto al modelo *Baseline*, con la métrica MAP. Como se observa en la tabla 5, la performance es muy pareja en los tres modelos para los tres conjuntos de datos, pero todos superan ampliamente el modelo base. A fin de seleccionar un modelo, se destaca al enfoque *Pairwise* como el que mejor resultados obtuvo.

Tabla 5. Comparación performance modelos.

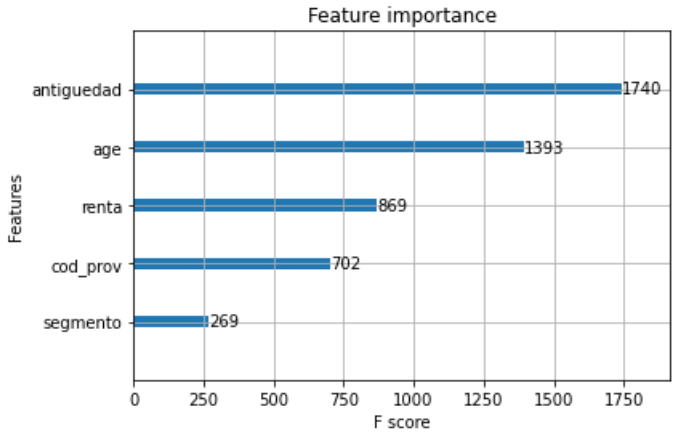
	BASELINE	POINTWISE	PAIRWISE	LISTWISE
ENTRENAMIENTO	0.6953	0.8662	0.8684	0.8666
TESTEO	0.6987	0.8673	0.8696	0.8613
VALIDACION	0.7039	0.8642	0.8673	0.8657

[7] Algoritmo regresión logística: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Fuente: elaboración propia.

Por último, es posible obtener que variables del conjunto de datos fueron más importantes en el entrenamiento del modelo. En la figura 16 se observan las 5 más importantes (antigüedad, edad, renta, provincia y segmento). La importancia de las variables se obtiene mediante la función *plot_importance ()*[8], propia de *XGBoost*, la cual calcula la importancia en base a los árboles utilizados durante el proceso de entrenamiento.

Figura 16. Importancia de variables en el modelo *Pairwise*.



Fuente: elaboración propia.

De esta manera, es posible concluir que, de los tres métodos propuestos, *Pairwise* es el que mejor se adapta al problema a resolver, obteniéndose un valor MAP de 0.8686 para el conjunto de datos de validación y 0.8694 para testeo. Sin embargo, es mínima la diferencia con el resto de los enfoques, por lo cual se considera que los tres métodos resuelven el problema de manera correcta. A continuación, se desarrolla sobre la importancia de utilizar este modelo en la campaña de venta de los productos.

3.3. Relevancia de un sistema de recomendación para el diseño de una estrategia de venta

Mediante los sistemas de recomendación, es posible ofrecer productos de manera individual o, en el caso contrario, no ofrecerlos cuando no son de interés para el cliente (Leimstoll., 2007). Entonces, según la situación, es posible identificar el producto más recomendable para cada

[8] Función plot_importance() : https://xgboost.readthedocs.io/en/stable/python/python_api.html

cliente o identificar para que clientes un producto es el más recomendable. En base a los resultados obtenidos, se ejemplifican ambas situaciones.

Mediante el uso del modelo *Pairwise*, ya que obtuvo una mejor performance, es posible identificar para cada cliente el producto de mayor preferencia (de los que no posee). En la tabla 6 se observa la lista de productos que el cliente 15889 no posee, ordenados de mayor a menor según su ranking. De esta manera el producto a ofrecerle a este cliente es débito directo ya ranquea en primer lugar. Esta predicción se realizó sobre el período mayo 2016, para conocer cuál producto es de mayor preferencia en el próximo mes.

Tabla 6. Ranking de productos para el cliente 15889.

ncodpers	producto	ranking
15889	ind_recibo_ult1	1,628585
15889	ind_ecue_fin_ult1	1,276466
15889	ind_dela_fin_ult1	0,732734
15889	ind_fond_fin_ult1	0,659103
15889	ind_cno_fin_ult1	0,480488
15889	ind_nomina_ult1	0,412761
15889	ind_nom_pens_ult1	0,366905
15889	ind_ctop_fin_ult1	0,341933
15889	ind_reca_fin_ult1	0,244216
15889	ind_plan_fin_ult1	-0,452718
15889	ind_hip_fin_ult1	-2,541362
15889	ind_viv_fin_ult1	-2,601382
15889	ind_deme_fin_ult1	-3,392520
15889	ind_pres_fin_ult1	-4,351931
15889	ind_deco_fin_ult1	-4,825956
15889	ind_ctma_fin_ult1	-5,089623
15889	ind_cder_fin_ult1	-5,198501
15889	ind_ahor_fin_ult1	-5,918215
15889	ind_aval_fin_ult1	-6,659149
15889	ind_ctju_fin_ult1	-7,714739

Fuente: elaboración propia.

Como se mencionó anteriormente, la columna ranking indica la salida predicha por el modelo y no un ranking del estilo: 1, 2, 3. Es decir, este valor es el *score* que genera el modelo para cada cliente-producto, y se utiliza para ordenar la lista de productos. Otro ejemplo, es el que se observa en la tabla 7, donde para el cliente 1464736 el producto de mayor preferencia para el período junio 2016 es depósito a corto plazo.

Tabla 7. Ranking de productos para el cliente 1464736.

ncodpers	producto	ranking
1464736	ind_deco_fin_ult1	3,148101
1464736	ind_cco_fin_ult1	2,875912
1464736	ind_dela_fin_ult1	2,627871
1464736	ind_cno_fin_ult1	-0,502170
1464736	ind_nom_pens_ult1	-1,177413
1464736	ind_recibo_ult1	-1,237943
1464736	ind_ecue_fin_ult1	-1,484821
1464736	ind_fond_fin_ult1	-1,733419
1464736	ind_valo_fin_ult1	-2,303842
1464736	ind_reca_fin_ult1	-2,605222
1464736	ind_nomina_ult1	-3,084516
1464736	ind_tjcr_fin_ult1	-3,401702
1464736	ind_plan_fin_ult1	-5,053013
1464736	ind_deme_fin_ult1	-5,129472
1464736	ind_ctpp_fin_ult1	-5,274130
1464736	ind_cder_fin_ult1	-5,933240
1464736	ind_ctop_fin_ult1	-6,764782
1464736	ind_hip_fin_ult1	-7,023962
1464736	ind_ctju_fin_ult1	-7,174019
1464736	ind_viv_fin_ult1	-7,240596
1464736	ind_pres_fin_ult1	-7,750008
1464736	ind_aval_fin_ult1	-7,821479
1464736	ind_ahor_fin_ult1	-7,949812

Fuente: elaboración propia.

Por último, puede ser interesante identificar, a partir de un producto en particular, para que clientes este es considerado el más relevante. Por ejemplo, si fuera necesario realizar una campaña de venta para el período junio 2016 del ítem préstamos ("ind_pres_fin_ult1"), en la tabla 8 se observan 10 clientes para los cuales este producto es el más recomendable. Es decir, estos clientes tienen al producto préstamos en la primera posición del ranking. Además, se obtiene que este producto es el más relevante para el 0.12 % de los clientes (871).

Tabla 8. 10 clientes para los cuales el producto préstamos es el más recomendable.

ncodpers	producto	ranking
734059	ind_pres_fin_ult1	4,006289
738160	ind_pres_fin_ult1	3,869064
607773	ind_pres_fin_ult1	3,831758
738255	ind_pres_fin_ult1	3,805022
749638	ind_pres_fin_ult1	3,776161
729689	ind_pres_fin_ult1	3,659668
754447	ind_pres_fin_ult1	3,633823
579644	ind_pres_fin_ult1	3,633432
676059	ind_pres_fin_ult1	3,616075
723598	ind_pres_fin_ult1	3,576689

Fuente: elaboración propia.

El mismo análisis puede realizarse sobre otro producto, por ejemplo, "cuenta particular plus". En la tabla 9 se observan 10 clientes para los cuales este producto es el más preferido. El producto cuenta particular plus, encabeza la lista de preferencias de 14.328 (2.05%) clientes. El valor de la columna ranking, no implica que para el cliente 873060 este producto sea más

recomendable que para el cliente 872764, sino que este valor fue el que obtuvo el producto en la lista, para cada cliente.

Tabla 9. 10 clientes para los cuales el producto más recomendable es “cuenta particular plus”.

ncodpers	producto	ranking
873060	ind_ctpp_fin_ult1	3,786446
872764	ind_ctpp_fin_ult1	3,772290
882828	ind_ctpp_fin_ult1	3,728951
884959	ind_ctpp_fin_ult1	3,716773
875536	ind_ctpp_fin_ult1	3,707994
761646	ind_ctpp_fin_ult1	3,699388
856681	ind_ctpp_fin_ult1	3,679216
881317	ind_ctpp_fin_ult1	3,675840
875546	ind_ctpp_fin_ult1	3,673536
866157	ind_ctpp_fin_ult1	3,661264

Fuente: elaboración propia.

Por otro lado, el producto que más clientes poseen en el primer puesto de su lista es impuestos (“ind_reca_fin_ult1”). A la vez, existen 4 productos que no son considerados de mayor preferencia por ningún cliente para el periodo junio 2016: caja de ahorro (“ind_ahor_fin_ult1”), garantías (“ind_aval_fin_ult1”), cuentas derivadas (“ind_cder_fin_ult1”) y depósitos a mediano plazo (“ind_deme_fin_ult1”). Es decir, estos cuatro productos no están en el primer puesto de la lista para ningún cliente.

De esta forma, se logra determinar cómo mediante el uso de un sistema de recomendación, es posible realizar una oferta de productos más eficientes a los clientes. Esto supone, por un lado, menores costos de comunicación, ya que se pueden realizar una menor cantidad de contactos, pero más asertivos. Y, por otro lado, el cliente no se ve perturbado con ofrecimientos de productos que no son de su interés.

Conclusión

A lo largo del presente trabajo se determinó como la personalización y relación con el cliente puede ser lograda a través de una adecuada gestión y análisis de datos. Para esto es fundamental contar con un entorno adecuado para el almacenamiento y explotación de los grandes volúmenes de datos. También, se mostró como, mediante el desarrollo de un sistema de recomendación, es posible identificar el producto financiero de mayor preferencia para cada cliente de la organización bancaria Santander. Mediante este, se logró argumentar que la implementación de estos sistemas genera una mayor eficiencia en las campañas de venta de los productos, ya que permite realizar una menor cantidad de contactos, pero con mayor

efectividad. A la vez, el cliente no se ve afectado por comunicaciones de productos que no son de su interés.

En el primer capítulo se analizaron los procesos y metodologías para la gestión de grandes volúmenes de datos en el sector bancario. En primer lugar, se abordó el concepto de grandes volúmenes de datos, el cual se puede definir en base a volumen, variedad y velocidad. Para las organizaciones tradicionales, como es el caso del banco Santander, este nuevo contexto requiere de un cambio en todas sus áreas y procesos internos de negocio. A su vez, este nuevo contexto se caracteriza por una mayor demanda y movilidad de los clientes, desregulación, y una marcada competencia. Por estas razones, la fidelización del cliente resulta fundamental, y una manera de lograrlo es a través de la personalización mediante la aplicación de sistemas de recomendación.

En el segundo capítulo se logró identificar dos grupos de clientes mediante técnicas de *Clustering*. Para ello, en primer lugar, se presentó el conjunto de datos utilizado, al cual se le realizó las transformaciones necesarias para su explotación y luego se llevó a cabo el análisis exploratorio, que permitió identificar dos grupos principales de clientes. Un grupo contiene a la mayoría de los clientes de la organización, pertenecientes al segmento Universitarios y Particulares, de menor antigüedad en la entidad y más jóvenes. El otro, concentra a clientes de más antigüedad y con mayor renta, ya que aquí se encuentran la mayoría de los clientes del segmento Top. Finalmente, se realizó la selección y descripción de los tres métodos para llevar a cabo la aplicación de recomendación de productos: *Pointwise, Pairwise y Listwise*.

Por último, en el tercer capítulo, se aplicaron los métodos propuestos a través del algoritmo *XGBoost*. A partir de la performance obtenida mediante el coeficiente MAP, se concluyó que *Pairwise* es el mejor para la determinación del producto de mayor preferencia por cliente de la organización Santander. Sin embargo, no se detectó una diferencia considerable respecto a los otros dos métodos. Los resultados alcanzados, permitieron aproximar a concluir que un sistema de recomendación permite generar una estrategia de venta de los productos financieros de manera más eficiente.

En cuanto al alcance del trabajo se realizó tomando únicamente los datos proporcionados por la organización Santander en la plataforma *Kaggle* y no se incorporaron fuentes externas. En ese sentido, incorporar otros tipos de datos como por ejemplo los provenientes de redes sociales de los clientes del banco, podría mejorar el rendimiento del modelo. De manera similar

sucedería con los datos provenientes de campañas de ventas realizadas anteriormente, los cuales darían la opción de, por ejemplo, no volver a ofrecerle un producto a un cliente que fue contactado recientemente.

Como futuras líneas de trabajo, puede resultar de interés aplicar otros algoritmos y métodos para la recomendación de productos. Por ejemplo, redes neuronales o algoritmos de filtrado colaborativo. Este último tipo, permitiría generar recomendaciones por similitud entre clientes, entre productos, o ambos. Así mismo, también podría ser interesante ampliar la problemática de negocio, y conocer como atraer nuevos clientes en base a qué primer producto recomendarles.

Referencias bibliográficas

Agrawal. (2011). Challenges and Opportunities with Big Data. *Purdue University*.

Al-Hawari. (2006). The effect of automated service quality on bank financial performance and the mediating role of customer retention. *Journal of Financial Services Marketing*.

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more.* Hachette Books.

Andrew McAfee, E. B. (2012). Big Data: The Management Revolution. *Harvard Business Review*.

Barton, D. C. (2012). Making Advanced Analytics Work for You. *Harvard Business Review*.

Bedeley, R. (2014). Big Data Opportunities and challenges: The Case Of Banking Industry. *SAIS 2014 Proceedings.*

Bora, G. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab. *International Journal of Computer Science and Information Technologies*.

Burges C. (2010). From RankNet to LambdaRank to LambdaMART: An Overview. *Microsoft Research Technical Report MSR-TR-2010-82*.

Burke, F. G. (2011). Recommender Systems: An Overview. *AI Magazine*.

Casalegno, R. (2022). *Learning to Rank: A Complete Guide to Ranking using Machine Learning*. Obtenido de https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4

Chen, G. (2016). XGBoost: A Scalable Tree Boosting System. *Association for Computing Machinery*.

Chen, H. (2017). *xgboost: eXtreme Gradient Boosting.* Obtenido de https://cran.microsoft.com/snapshot/2017-12-11/web/packages/xgboost/vignettes/xgboost.pdf

Chen, H. C. (2012). *Business intelligence and analytics: From big data to big impact. MIS Quarterly.* Obtenido de https://is.wcu.edu/wchung/rsch/cbia/doc/BIA_Rsch_MISQ12.pdf

Cohen, D. G. (2007). Customer retention by banks in New Zealand. *Banks and Bank Systems*.

Constantiou, I. D. (2015). New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology*.

Cossock D., Z. T. (2006). *Subset Ranking Using Regression.* Obtenido de https://link.springer.com/chapter/10.1007/11776420_44

Davenport. (2014). *Big data at work: dispelling the myths, uncovering the opportunities.* Harvard Business Review Press.

Deng, S. C. (2019). *Recommender system for marketing optimization.* Obtenido de https://link.springer.com/article/10.1007/s11280-019-00738-1

Dicuonzo G., G. G. (2019). Risk Management 4.0: The Role of Big Data Analytics in the Bank Sector. *International Journal of Economics and Financial Issues*.

Draper, N., & Smith, H. (1998). *Applied Regression Analysis.* (3, Ed.) John Wiley.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Institute of Mathematical Statistics*.

Gigli, L. R. (2017). Recommender Systems for Banking and Financial Services. *RecSys Posters*.

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

Hagen, P. (1999). Smart personalization. *Cambridge, MA: Forrester Research, Inc.*

Hartigan, W. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Royal Statistical Society*.

Hassani, H., Huang, X., & Ghodsi, M. (2018). Big Data and Causality. *Annals of Data Science*.

Kallinikos, J. (2019). Recommender Systems. *Oxford University Press*.

Katal, W. G. (2013). Big Data: Issues, Challenges, Tools and Good Practices. *Sixth International Conference on Contemporary Computing (IC3)*.

Kharote, K. (2014). Data mining model for money laundering detection in financial domain. *International Journal of Computer Applications*.

Kosaman, H. S. (2018). Recommendation System to Optimize Email Marketing Campaign Using Apriori Algorithm Case Study: Webeli.Com. *International Journal of Engineering and Technology*.

Laborde, R. (2020). *The Three V's of Big Data: Volume, Velocity, and Variety*. Obtenido de Oracle Health Sciences Blog: https://blogs.oracle.com/health-sciences/post/the-three-vx27s-of-big-data-volume-velocity-and-variety

Labrinidis, J. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*.

Lavrenko, V. (2009). *Text Technologies - Evaluation*. Obtenido de The University of Edinburgh: http://www.inf.ed.ac.uk/teaching/courses/tts/pdf/rf-1x2.pdf

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*.

Leimstoll. (2007). Collaborative Recommender Systems for Online Shops. *Association for Information Systems*.

Liu P., C. M. (2017). Performance Evaluation of Recommender Systems. *Int J Performability Eng*.

Melnikov V., G. P. (2016). Pairwise versus Pointwise Ranking: A Case Study. *Schedae Informaticae*.

Microsoft. (s.f.). Obtenido de Big data architecture style - Azure Architecture Center: https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data

Oyebode, O. (2020). A hybrid recommender system for product sales in a banking environment. *Journal of Banking and Financial Technology*.

Porter. (1995). Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. *Princeton University Press*.

Rahman, N. &. (2015). Big data business intelligence in bank risk analysis. *International Journal of Business Intelligence Research (IJBIR) 6(2)*.

Rakhman, R. W. (2019). Big data analytics implementation in banking industry – Case study cross selling activity in Indonesia's Commercial bank. *International Journal of Scientific & Technology Research*.

Rezaei, M. R. (2021). Amazon Product Recommender System. *University of Toronto* .

Ricci, R. S. (2011). *Introduction to recommender systems handbook.* Springer.

Riecken, D. (2000). *Personalized views of personalization*. Obtenido de Communications of the ACM : https://dl.acm.org/doi/fullHtml/10.1145/345124.345133

Schafer, K. R. (1999). Recommender Systems in E-Commerce. *University of Minnesota*.

Schmarzo, B. (2013). *Big Data: Understanding How Data Powers Big Business.* Wiley.

Sun, N. M. (2014). iCARE: A framework for big data-based banking customer analytics. *IBM Journal of Research and Development*.

Syakur. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*.

Wang, Y. X. (2021). Can fintech improve the efficiency of commercial banks? An analysis based on big data. *Research in International Business and Finance*.

Wu, F. (2005). Ascored AUC Metric for Classifier Evaluation and Selection. *Second workshop on ROC analysis in ML*.

Zibriczky, D. (2016). Recommender Systems meet Finance: A literature review. *Proc. 2nd Int. Workshop Personalization*.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*.