



Máster en  
Data Science & AI

MEMORIA PROYECTO CAPSTONE

**Caso easyMoney**



Autores:

**Íñigo Alegre Jiménez**  
**David Moreno Vituri**  
**Ricardo Sánchez Olivares**

Tutor:

**Matías José Hermida**

Entrega: 18/10/2024

Presentación: 28/10/2024

# Índice

<b>1. Introducción</b>	<b>3</b>
1.1. Contexto empresarial . . . . .	3
1.2. Enfoque y metodología . . . . .	4
1.3. Marco teórico . . . . .	4
1.4. Conclusiones . . . . .	5
<b>2. Objetivo</b>	<b>5</b>
<b>3. Metodología</b>	<b>5</b>
3.1. Creación de Power BI . . . . .	6
3.2. Modelos de propensión a compra . . . . .	10
3.2.1. Uso de algoritmos . . . . .	10
3.2.2. Selección de productos . . . . .	11
3.2.3. Resolución . . . . .	12
3.3. Segmentación de clientes . . . . .	19
3.3.1. Uso de KMeans . . . . .	19
3.3.2. Resolución . . . . .	20
<b>4. Resultados</b>	<b>23</b>
4.1. Modelos de propensión a compra . . . . .	23
4.2. Análisis de los clusters . . . . .	25
4.2.1. Implicaciones comerciales . . . . .	28
4.3. Campaña comercial . . . . .	28
<b>5. Conclusiones</b>	<b>30</b>
<b>6. Bibliografía</b>	<b>32</b>
<b>7. Anexo</b>	<b>33</b>

## 1. Introducción

En la actualidad, el análisis de datos se ha convertido en una herramienta fundamental para empresas que desean mantenerse competitivas y responder a las necesidades cambiantes del mercado. La creciente digitalización de los negocios y la proliferación de sistemas de almacenamiento y bases de datos han permitido que las organizaciones adopten estrategias basadas en el análisis avanzado de datos, con el fin de optimizar sus procesos comerciales, mejorar la experiencia del cliente y aumentar la rentabilidad. En este contexto, el trabajo de fin de máster que aquí se presenta aborda un caso práctico dentro del sector financiero, donde se busca aplicar técnicas de Data Science para mejorar la estrategia comercial de una empresa que comercializa productos financieros.

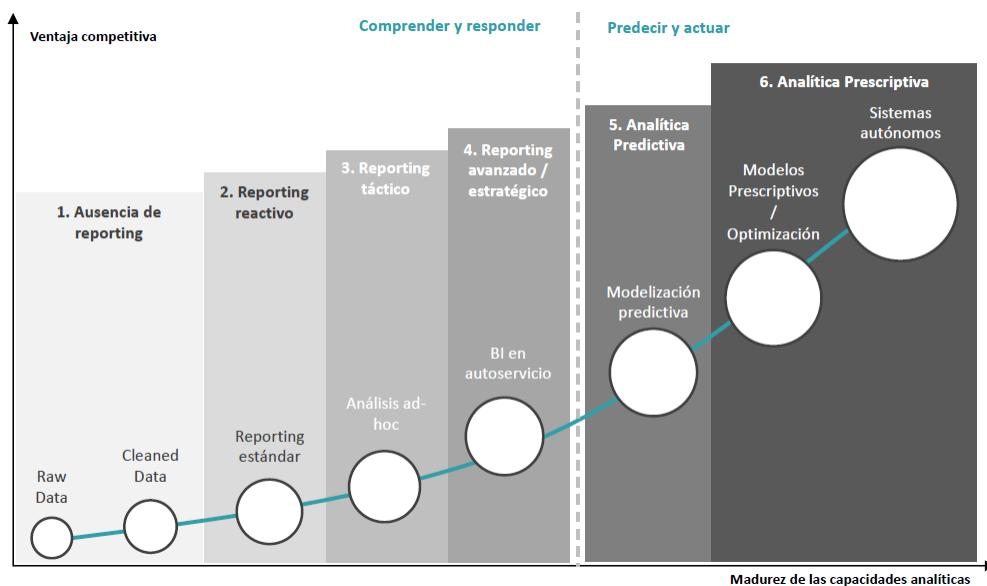


Figura 1: Modelo de madurez en Analytics, que ilustra las distintas fases por las que pasan las organizaciones en el uso de datos para obtener ventajas competitivas. El recorrido comienza con la ausencia de reporting y culmina con la analítica prescriptiva, donde los sistemas autónomos optimizan decisiones y predicciones basadas en datos. Cada fase refleja una mayor madurez analítica, pasando por reporting reactivo, táctico, estratégico y la modelización predictiva.

### 1.1. Contexto empresarial

El presente trabajo se enmarca en el rol de Data Scientist dentro de easyMoney, una joven compañía fintech que nació hace casi cuatro años con el objetivo de ofrecer productos financieros sencillos y accesibles a través de una plataforma multicanal. Fundada por Carol Denver, profesional con más de diez años de experiencia en banca de inversión, easyMoney se lanzó al mercado con su producto estrella, la cuenta hucha easyMoney, que permitía a los clientes ahorrar de forma automática mediante el redondeo de sus compras. Este primer producto fue un éxito rotundo, lo que permitió a la empresa expandir su oferta a otros productos financieros, tales como soluciones de inversión, tarjetas y servicios de financiación.

A pesar de su crecimiento inicial y la captación de numerosos clientes, easyMoney enfrenta varios retos que ponen en riesgo su continuidad. Las presiones comerciales de su

principal socio, easyBanking, y la incorporación constante de nuevos productos han desviado a la empresa de su visión original de ofrecer soluciones financieras simples. Además, el agotamiento de los fondos obtenidos en rondas de financiación y la falta de un EBITDA positivo han generado tensiones con su principal inversor, Lion Global Management, que exige mejorar la rentabilidad de la base de clientes actual antes de seguir invirtiendo. En este contexto, y con el objetivo de mejorar sus resultados financieros, easyMoney ha decidido incorporar un Data Scientist (nosotros) que ayude a optimizar la rentabilidad de sus clientes y a desarrollar nuevas estrategias comerciales basadas en el análisis de datos.

## 1.2. Enfoque y metodología

El objetivo principal de este trabajo es utilizar técnicas de Data Science para ayudar a easyMoney a mejorar su estrategia comercial y aumentar la rentabilidad de su base actual de clientes. Para lograrlo, se van a abordar cuatro tareas clave. La primera es el desarrollo de un dashboard en Power BI que permita a las áreas comerciales realizar un seguimiento detallado de las ventas a través de KPIs y análisis evolutivos. Esto facilitará la identificación de tendencias y oportunidades de mejora.

La segunda tarea consistirá en la creación de tres modelos de propensión a compra, cada uno de ellos permitirá predecir qué clientes tienen mayor probabilidad de adquirir un producto financiero específico. Además, se realizará una segmentación de clientes utilizando técnicas de clustering, lo que permitirá entender mejor los distintos grupos de clientes. De este modo se prodrán adaptar las estrategias de marketing de manera adecuada a sus características para poder aumentar así la rentabilidad. Finalmente, se estimará el impacto de una campaña comercial utilizando los resultados de los modelos anteriores para definir cuántos clientes serán impactados, qué productos alcanzarán y de qué modo a cada grupo, y estimar los ingresos potenciales de la campaña.

## 1.3. Marco teórico

Para realizar las tareas descritas anteriormente, el proyecto se fundamenta en varias disciplinas dentro del campo de la ciencia de datos, principalmente en el análisis predictivo y la segmentación de clientes.

El uso de modelos de propensión permiten a las empresas identificar de manera precisa a los clientes que tienen más probabilidades de adquirir un determinado producto. La combinación de técnicas como la regresión logística y Gradient Boosting han demostrado ser eficaces para predecir comportamientos de compra, mejorar las tasas de conversión y reducir los costes de adquisición de clientes.

En cuanto a la segmentación de clientes, técnicas de clustering como K-Means se han utilizado extensamente para agrupar a los clientes en segmentos homogéneos en base a características demográficas, transacciones y diversos comportamientos. Esta metodología ha permitido a las empresas de diversos sectores diseñar campañas de marketing personalizadas, optimizando así su impacto y retorno de inversión. Este proyecto se apoya en estos enfoques teóricos, los cuales serán desarrollados a lo largo de la memoria, con el objetivo de aplicarlos al caso particular de easyMoney.

#### 1.4. Conclusiones

Gracias a la incorporación del perfil Data Science en easyMoney, se ha implementado una analítica avanzada que ha permitido adoptar una metodología de trabajo basada en datos. Esto ha facilitado tanto la inversión en tecnología como la creación de herramientas fundamentales para áreas como marketing y ventas. Con un dashboard para la dirección de marketing y modelos de Machine Learning (XGBoost, CatBoost, LightBoost, y Regresión Logística), hemos identificado qué tipo de cliente es más propenso a adquirir productos clave como planes de pensiones, depósitos a largo plazo y tarjetas de débito. Además, mediante técnicas de clustering (K-Means), clasificamos a los clientes en siete grupos, generando buyer personas para cada cluster y ayudando a personalizar las estrategias. Tras estimar ingresos potenciales de una campaña dirigida, se prevé que estos alcancen los 6.541.000€. En los próximos pasos, estimamos que sería necesario ajustar y automatizar dashboards para otros departamentos y crear modelos de propensión a compra para los productos más relevantes para así optimizar la toma de decisiones en tiempo real y poder optimizar al máximo la rentabilidad de la base actual de clientes.

## 2. Objetivo

El objetivo principal del proyecto es proporcionar soluciones basadas en Data Science que permitan a easyMoney superar los desafíos que amenazan su continuidad. En particular, el enfoque se centra en optimizar la rentabilidad de la base actual de clientes con una campaña basada en tres acciones clave: la creación de un dashboard en Power BI para monitorear y analizar las ventas y KPIs (enfocando en el último mes y trimestre); el desarrollo de tres modelos de propensión a compra para identificar a los clientes más propensos a adquirir aquellos productos financieros que consideramos más rentables y con más potencial de afectar a la mayoría de clientes; y la segmentación de la base de clientes en grupos homogéneos, lo que permitirá personalizar las estrategias de marketing y mejorar el impacto comercial.

Por tanto, este proyecto busca desarrollar una campaña que, a la par que reduzca la inversión en captación intensiva y masiva de clientes nuevos, optimice la rentabilidad de la base actual de clientes de easyMoney, resolviendo problemas como el exceso de diversificación de productos y las crecientes demandas de los inversores en cuanto a rentabilidad. Al centrarse en estos retos, la campaña permitirá a easyMoney enfocar sus esfuerzos comerciales de manera más efectiva, mejorar la eficiencia operativa y volver a alinear su oferta con las necesidades reales de sus clientes, maximizando así el impacto en los resultados financieros.

## 3. Metodología

En este apartado de metodología, se detallará el procedimiento seguido para llevar a cabo las tres tareas fundamentales del proyecto. En primer lugar, se abordará la creación de un dashboard en Power BI, diseñado para facilitar el seguimiento de las ventas mediante la visualización de KPIs clave y la realización de análisis evolutivos. A continuación, se describirá la creación de tres modelos de propensión a compra, cuyo objetivo es identificar a los clientes más propensos a adquirir productos financieros específicos. Por último, se explicará la segmentación de la base de clientes mediante un modelo KMeans, que permitirá agrupar a los clientes en segmentos homogéneos para optimizar las estrategias comerciales.

Cada paso del proceso incluirá detalles sobre la descarga, el tratamiento, preprocesamiento y limpieza de los datos, así como la explicación de la implementación de cada modelo, garantizando así una comprensión clara y la posible replicabilidad de las tareas descritas.

### 3.1. Creación de Power BI

A nuestra disposición tenemos 5 tablas de excel diferentes (cuyo nombre y variables podemos observar en Cuadro 1). A continuación, describimos cada una de ellas y el preprocesamiento de algunas de sus variables realizado tras la importación de las mismas a python, donde llevaremos a cabo la creación e implementación de tablas y modelos.

- Customer commercial activity

Esta tabla almacena información sobre la actividad comercial de los clientes haciendo una actualización de la misma cada mes. Incluye datos como el identificador del cliente, el canal de captación (cómo llegó el cliente a la empresa), la fecha de entrada (cuando se registró la primera interacción), el segmento comercial al que pertenece el cliente (universitario, particular o top) y si el cliente está activo en la aplicación EasyMoney.

En su preprocesamiento, hemos transformado la columna active\_customer a tipo int para un mejor manejo de los datos. También hemos cambiado el formato de la columna segment a uno más cómodo para su análisis. La columna entry\_date se ha convertido a tipo datetime, lo que nos ha permitido extraer el año y el mes correspondientes. Por último, hemos identificado que existen nulos en las columnas entry\_channel y segment, a los cuales les hemos imputado el valor de "Unknown".

- Customer products

Esta tabla registra la tenencia de productos financieros por parte de los clientes haciendo una actualización cada mes. Contiene campos como el identificador del cliente y detalles sobre diversos productos como tarjetas de crédito y débito, cuentas (easyMoney y otras), fondos de inversión, préstamos, hipotecas, y cuentas de ahorro a corto y largo plazo.

En cuanto al preprocesamiento, hemos convertido las columnas payroll y pension\_plan a tipo int, y hemos manejado los nulos asignándoles la moda, que en ambos casos es 0. A simple vista, se observa que los artículos más contratados son em\_account, debit\_card y payroll\_account, lo cual podría ser útil para futuras decisiones comerciales.

- Customer sociodemographics

Aquí se almacenan los datos sociodemográficos de los clientes. Incluye el identificador del cliente, estado civil (si está fallecido o no), género, código de región, salario, país de residencia, y edad. Esta tabla al igual que las anteriores se actualiza cada mes.

Para esta tabla, hemos decidido convertir la columna country\_ID a tipo int para optimizar el espacio de almacenamiento, a pesar de que se trata de una columna categórica. También hemos encontrado nulos en la columna region\_code, los cuales hemos atribuido a compradores que no son españoles. Para estos casos, hemos asignado un valor de -1. Observamos que

existe una gran cantidad de nulos en salary, pero hemos decidido no imputar ningún valor para preservar la distribución de los datos, dado que Power BI puede identificarlos y no los considera en sus cálculos. Finalmente, en la columna gender, hemos imputado los nulos a la moda, que resulta ser el género femenino.

- Sales

Esta tabla contiene información sobre las ventas realizadas. Incluye un identificador único para cada venta, el identificador del cliente que realizó la compra, el mes en que se llevó a cabo la venta, el identificador del producto vendido, y el margen neto generado por la venta.

El DataFrame de sales no contiene nulos, lo que es positivo para el análisis. Aunque la columna product\_ID es numérica, es importante recordar que se trata de una columna categórica. Además, hemos convertido la variable month\_sale a tipo datetime para facilitar su manipulación.

- Product description

En esta tabla se detalla la información de los productos disponibles. Contiene un identificador único del producto, una descripción del producto, y el vertical o categoría a la que pertenece (grupo). No posee nulos.

Nombre de tabla excel	Variables
customer_commercial_activity	pk.cid
	pk.partition
	entry_channel
	entry_date
	segment
	active_customer
customer_products	pk.cid
	pk.partition
	credit_card
	debit_card
	em_account_p
	em_account_pp
	em_acount
	emc_account
	funds
	loans
	long_term_deposit
	mortgage
	payroll
	payroll_account
	pension_plan
	securities
	short_term_deposit
customer_sociodemographics	pk.cid
	pk.partition
	deceased
	gender
	region_code
	salary
	country_id
sales	age
	pk.sale
	cid
	month_sale
	product_ID
product_description	net_margin
	pk.product.ID
	product_desc
	family_product

Cuadro 1: Conjunto de tablas suministradas por el director de IT que tenemos a disposición las cuales poseen 17 particiones de histórico y variables correspondientes a las mismas

Una vez preprocesada cada tabla nuestro objetivo es crear una tabla excel con la que construir un dashboard en Power BI que ayude al área comercial a explorar los datos de ventas a través del seguimiento de KPIs, evolutivos y análisis Deep Dive para profundizar en el conocimiento de cliente.

Para ello comenzamos fusionando en python la tabla sales, que contiene la información

sobre las ventas mensuales de cada producto, con la tabla customer\_sociodemographics, la cual nos proporciona datos clave sobre los clientes, como su género, localización, país, salario, edad y si están fallecidos o no.

En la segunda fusión, unimos el dataset resultante con el de customer\_commercial\_activity, que nos ofrece información adicional sobre los clientes, incluyendo si son active\_customer, el tipo de cliente según el segment, la fecha en la que se registraron (entry\_date) y el canal de entrada (entry\_channel).

Por último, realizamos la tercera fusión, combinando el dataset resultante con las columnas relevantes del dataset product\_description, que aporta más detalles sobre los productos, como su nombre y la familia a la que pertenecen. A su vez, creamos una nueva variable llamada tipo\_cliente que nos informa si el cliente es nuevo o se trata de un cliente ya existente (vemos si la variable pk\_partition es mayor o igual que la variable entry\_date. En caso de ser iguales, se trata de un cliente nuevo, y si entry\_date es menor, se trata de un cliente ya existente).

Num.col	Columna	Filas no nulas	Type
0	pk_sale	240773 non-null	int64
1	cid	240773 non-null	int64
2	month_sale	240773 non-null	datetime64[ns]
3	net_margin	240773 non-null	float64
4	pk_partition	240773 non-null	datetime64[ns]
5	country_id	240773 non-null	object
6	region_code	240773 non-null	int32
7	gender	240773 non-null	object
8	age	240773 non-null	int64
9	deceased	240773 non-null	object
10	salary	133154 non-null	float64
11	entry_date	240773 non-null	datetime64[ns]
12	entry_channel	240773 non-null	object
13	active_customer	240773 non-null	float64
14	segment	240773 non-null	object
15	product_desc	240773 non-null	object
16	family_product	240773 non-null	object
17	month_sale_year	240773 non-null	int32
18	month_sale_month	240773 non-null	int32
19	entry_date_year	240773 non-null	int32
20	entry_date_month	240773 non-null	int32
21	pk_partition_year	240773 non-null	int32
22	pk_partition_month	240773 non-null	int32
23	tipo_cliente	240773 non-null	object

Cuadro 2: Información de la tabla sales\_analysis.

De esta manera, hemos construido una tabla llamada sales\_analysis (cuyas variables e información podemos observar en el Cuadro 2), la cual nos aporta toda la información disponible tanto de la compra como del cliente que la realiza. Esto nos permitirá, en Power BI, realizar análisis detallados y segmentaciones precisas, explorando patrones de compra según el perfil del cliente, el rendimiento de los productos y su relación con factores demográficos

o comerciales. Además, podremos visualizar las tendencias de ventas a lo largo del tiempo y crear informes interactivos que faciliten la toma de decisiones basadas en datos.

### 3.2. Modelos de propensión a compra

La segunda tarea que nos encomienda Carol (la directora general) es crear modelos de propensión a compra de uno o varios productos. Este tipo de modelos es de gran utilidad ya que permite predecir la probabilidad de que un cliente adquiera un producto o servicio en el futuro. Con esta información, la empresa puede optimizar sus campañas de marketing, enfocándose en aquellos clientes que tienen una mayor probabilidad de conversión, lo que a su vez reduce los costos de adquisición y mejora la eficacia de las estrategias de venta.

#### 3.2.1. Uso de algoritmos

Los modelos de propensión a compra son herramientas analíticas que permiten predecir la probabilidad de que un cliente realice una compra en un futuro cercano. Se utilizan principalmente en marketing y ventas para identificar a los clientes con mayor predisposición a adquirir un producto o servicio, optimizando así las estrategias de segmentación, campañas publicitarias y promociones. Al tener un modelo que prediga qué clientes tienen más probabilidades de comprar, las empresas pueden enfocar sus esfuerzos en estos segmentos, logrando una mayor eficiencia en la conversión de clientes potenciales, el cual es el objetivo de este proyecto. Este tipo de modelos se basan fundamentalmente en dos tipos de técnicas de Machine Learning: el Gradient Boosting y la Regresión Logística.

- Gradient Boosting

Los algoritmos basados en Gradient Boosting son métodos potentes de aprendizaje supervisado utilizados tanto para problemas de clasificación como de regresión. Funcionan construyendo múltiples modelos simples, como árboles de decisión, de forma secuencial, y cada modelo intenta corregir los errores de los modelos anteriores. A lo largo de varias iteraciones, los modelos se 'potencian' en conjunto para mejorar el rendimiento general. Los tres algoritmos que vamos a implementar en nuestro proyecto son XGBoost, CatBoost y LightGBM.

XGBoost es uno de los algoritmos más populares basados en Gradient Boosting. Este algoritmo implementa regularización para evitar el overfitting y posee muchos hiperparámetros que pueden ser ajustados para mejorar el rendimiento del mismo en base a nuestras necesidades. Su capacidad para manejar datos desbalanceados y su soporte para diferentes tipos de tareas (clasificación, regresión) lo hacen muy adecuado para problemas complejos como la propensión a compra.

LightGBM se caracteriza por ser muy rápido y eficiente en términos de tiempo y consumo de memoria. Utiliza una técnica llamada "leaf-wise growth" que selecciona las hojas más "puras" para la división, permitiendo un aprendizaje más profundo y una mayor precisión en menos tiempo. LightGBM es especialmente útil en datasets grandes, con muchas características, donde se requiere un rendimiento computacional óptimo.

CatBoost es otro algoritmo de Gradient Boosting. Este se diferencia de los anteriores ya que está optimizado para manejar de manera eficiente variables categóricas. Tradicionalmente, las variables categóricas requieren un procesamiento de codificación antes de usarse en modelos predictivos, pero CatBoost integra este proceso directamente, evitando transformaciones que pueden degradar el rendimiento. Además, es robusto ante datasets desbalanceados y con mucho ruido, lo que lo convierte en una excelente opción para problemas reales de negocio.

- Regresión Logística

La regresión logística es uno de los métodos más comunes y fáciles de interpretar para problemas de clasificación binaria, como es el caso de la propensión a compra (comprará o no). A diferencia de los modelos lineales, que predicen un valor continuo, la regresión logística estima la probabilidad de que una observación pertenezca a una clase particular, utilizando una función logística o sigmoide que limita los resultados entre 0 y 1.

La regresión logística se implementa porque es un modelo sencillo, fácil de interpretar, y es muy eficiente en situaciones donde las relaciones entre las variables no son extremadamente complejas. Además, ofrece resultados interpretables, permitiendo entender cómo cada variable afecta la probabilidad de compra. A pesar de que no es tan flexible como los algoritmos basados en Gradient Boosting, su rapidez y simplicidad lo hacen una opción muy efectiva en problemas donde se requiere un enfoque claro y directo.

En nuestro proyecto, hemos optado por un enfoque híbrido, una especie de Ensemble Learning manual que incluye tanto modelos basados en Gradient Boosting (XGBoost, CatBoost, LightGBM) como Regresión Logística, permitiéndonos obtener un balance entre poder predictivo, rendimiento, y facilidad de interpretación. Para ello hemos realizado la media aritmética de las predicciones de probabilidad de compra aportadas por cada algoritmo obteniendo así probabilidades de propensión a compra más robustas.

### 3.2.2. Selección de productos

Una vez entendemos la tarea que debemos de realizar y los algoritmos que vamos a usar, debemos escoger los productos sobre los que construir nuestra campaña, es decir, aquellos sobre los que aplicaremos el modelo de propensión a compra. Para ello, realizamos un estudio de las ventas de todos los productos con la información que disponíamos. Tras el mismo, decidimos que los productos en los que se basaría nuestra camapaña serían: tarjeta de débito (debit card), plan de pensiones (pension plan) y depósito a largo plazo (long term deposit).

En primer lugar, elegimos el producto en el que el equipo de ventas de easyMoney tiene mayor experiencia, es decir, el producto más vendido que no es una cuenta (la cuenta em es realmente el producto más vendido debido a que la mayoría de clientes la posee).

En segundo lugar, nos enfocaremos en optimizar el presupuesto apostando por el producto que ha generado más beneficios a la empresa, a pesar de no haber vendido un gran porcentaje de unidades, el plan de pensiones. Identificamos un gran potencial de crecimiento

en este ámbito, ya que la mayoría de nuestros clientes no cuenta con un plan de pensiones, el cual es un producto realmente rentable y popular en la actualidad.

Por último, seleccionaremos un producto que ofrece un buen margen medio (beneficio promedio) y en el que hemos detectado un gran potencial de crecimiento. El año pasado, en junio, la tendencia de este producto era creciente; sin embargo, en los últimos meses de este año, las ventas han experimentado una caída significativa. Por lo tanto, buscaremos revertir esta situación y aprovechar la tendencia estacional positiva que presenta este producto.

### 3.2.3. Resolución

Una vez elegidos los productos nos disponemos a resolver la tarea que se nos ha encargado. Estamos ante un problema de aprendizaje supervisado, concretamente uno de clasificación binaria en el que el objetivo es determinar si un cliente compra o no el producto.

Para llevar a cabo esta tarea, es necesario crear una tabla que nos proporcione toda la información relevante que ayude al modelo a identificar de manera correcta a aquellos clientes que son compradores y aquellos que no lo son. Comenzamos analizando la tasa de penetración de cada producto durante el último mes (ver cuadro 3)). Observamos que dichas tasas son muy bajas, lo que indica un desbalance significativo entre las clases a predecir (en este caso el número de 0, cliente que no compra, es mucho mayor que el número de 1, cliente que si compra). Para mitigar este problema, aplicaremos la técnica de undersampling en la clase mayoritaria, reduciendo su número de filas y equilibrando así el dataset, logrando un balance adecuado entre ambas clases.

Producto	Ventas último mes	Clientes sin producto último mes	Tasa de penetración
Debit card	2049	399734	0.5126 %
Pension plan	1217	425642	0.2859 %
Long term deposit	2	436866	0.0005 %

Cuadro 3: Información sobre tasa penetración de los productos protagonistas de los modelos de propensión a compra

Para la creación de esta tabla hemos creado una función en python llamada **creacion\_df**. Esta función permite generar, a partir de dos tablas clave —sales\_analysis y customer\_products— una tabla balanceada que distingue entre los clientes que han comprado un producto específico y aquellos que no lo han hecho durante un periodo de tiempo determinado.

```
creacion_df(sales_analysis, customer_products, '2019-01-01', '2019-05-01', 'debit_card').reset_index() (3.1)
```

La tabla sales\_analysis proporciona la información de las ventas, mientras que customer\_products ofrece datos adicionales sobre la tenencia de productos por parte de los clientes. A través de esta función, se filtran y combinan estos conjuntos de datos para identificar, por un lado, a los clientes que efectivamente realizaron la compra, y por otro, a un número equivalente de clientes que no la realizaron, garantizando así el equilibrio entre ambas clases.

Para cada uno de los productos, hemos utilizado distintos períodos de tiempo. Comenzamos considerando únicamente las ventas a lo largo del último año, 2019. Sin embargo, debido a que las ventas de pension\_plan y long\_term\_deposit durante este periodo no eran suficientes para construir un dataset balanceado con un número significativo de clientes, decidimos, en estos dos casos, extender el análisis para incluir las ventas de 2018.

En el caso de pension\_plan, tomamos el periodo comprendido entre el 2018-08-01 y el último mes disponible en nuestros datos, el 2019-05-01. Elegimos este periodo porque no presenta ninguna singularidad (como picos o descensos drásticos en las ventas) que pudiera influir negativamente en las predicciones del modelo.

Por otro lado, para long\_term\_deposit, consideramos todo el historial disponible, ya que de otro modo no hubiéramos obtenido una muestra significativa de datos para alimentar al modelo.

Dataset	Filas	Columnas
debit_card	22653	18
pension_plan	22884	18
long_term_deposit	8429	18

Cuadro 4: Información del tamaño de los datasets resultantes de la aplicación de la función creacion\_df

Columnas	Dtype
cid	int64
pk_partition	datetime64[ns]
country_id	object
region_code	int32
gender	object
age	int64
deceased	object
salary	float64
entry_date	datetime64[ns]
entry_channel	object
active_customer	float64
segment	object
entry_date_year	int32
entry_date_month	int32
pk_partition_year	int32
pk_partition_month	int32
tipo_cliente	object
Compra_debit_card	int64

Cuadro 5: Información sobre las columnas que tiene cada una de las tablas creadas por la función creacion\_df

Una vez generado el dataset balanceado para cada producto, procedemos a crear un DataFrame llamado info\_clientes, que consolidará toda la información de los clientes desde su primer registro en la base de datos. Este DataFrame contendrá los datos de cada cliente

para cada partición temporal registrada, proporcionando así un historial completo de cada cliente a través de todas las divisiones temporales disponibles.

Para crear este dataframe partimos de las tablas de excel que se nos proporcionaron anteriormente (cuadro 1) y a partir de las mismas extraemos variables extra las cuales pueden aportar mucha información a los modelos de propensión. Las variables nuevas que se han creado son:

- **meses\_antiguedad:** Tiempo, en meses, que un cliente ha estado asociado a la empresa.
- **total\_products:** Número total de productos que un cliente posee.
- **total\_products\_no\_account:** Número total de productos que tiene un cliente, excluyendo cuentas bancarias.
- **spent\_on\_(family product):** Dinero que el cliente gasta en cada categoría de productos.
- **max\_products:** Número máximo de productos que un cliente ha tenido a lo largo de su relación con la empresa.
- **days\_since\_last\_purchase:** Número de días que han transcurrido desde la última compra del cliente.
- **months\_since\_last\_purchase:** Número de meses que han pasado desde la última compra.
- **has\_(family\_product):** Indica si el cliente posee un producto de una familia específica.
- **has\_credit\_card\_debt:** Indica si el cliente tiene deuda en su tarjeta de crédito.
- **num\_accounts:** Número total de cuentas que posee el cliente.
- **num\_debt\_products:** Número de productos en deuda que posee el cliente.
- **(product)\_consecutive\_months:** Número de meses consecutivos que el cliente ha tenido un producto específico.

Una vez que hemos creado cada variable adicional en su tabla correspondiente, procedemos a generar la tabla info\_clientes mediante la unión de las tablas a través de diferentes operaciones de pd.merge, método de la biblioteca Pandas que permite combinar DataFrames en Python de manera similar a las uniones (joins) en bases de datos relacionales.

Finalmente, para obtener las tablas definitivas de debit\_card, pension\_plan y long\_term\_deposit, que utilizaremos para entrenar nuestros modelos de propensión, combinamos las tablas resultantes de la función creacion\_df con la tabla info\_clientes. De este modo, obtenemos tres tablas balanceadas (una para cada producto) que contienen la información completa de cada cliente.

Una vez tenemos los dataset definitivos que van a servir para resolver nuestro problema, aplicamos la pipeline común para todo problema de Machine Learning.

- Importación de librerías

El primer paso en cualquier análisis de datos es importar las librerías necesarias. En nuestro caso estas librerías son:

1. **Pandas y NumPy**: para la manipulación y análisis de datos.
  2. **Matplotlib y Seaborn**: para la visualización de datos, como gráficos y diagramas.
  3. **Scikit-learn**: para el preprocesamiento, partición de datos y evaluación de modelos.
  4. **XGBoost, LightGBM y CatBoost**: para el desarrollo de modelos de machine learning basados en gradient boosting.
- Data Understanding

El **análisis exploratorio** de datos es fundamental para comprender la estructura, las relaciones y los patrones presentes en el dataset. Se comienza examinando el tamaño del dataset, los tipos de datos, y se identifican posibles valores atípicos, datos faltantes y anomalías. Este paso nos proporciona una visión global del DataFrame y de las variables que lo componen. Decidimos imputar a los valores nulos de salary (la única variable con nulos) el valor de -1, para que los algoritmos de gradient boosting detecten este patrón de manera más sencilla.

En el **análisis univariante**, se generan visualizaciones y tablas descriptivas para analizar la relación de cada variable con la variable TARGET, en nuestro caso, la columna Compra\_(producto). Este análisis nos permite identificar qué variables tienen poder explicativo y cuáles no. Las variables que no resultan relevantes son añadidas a una lista para su posterior eliminación. Durante esta fase también realizamos algunos procesamientos, como la codificación de la columna region\_code, que contiene muchas categorías distintas, y la reducción de categorías en la variable entry\_channel. Al final de este proceso, eliminamos las variables que no aportan valor explicativo.

En el **análisis multivariante**, investigamos las relaciones entre las variables mediante el uso de herramientas como la matriz de correlación y el coeficiente de Cramér para analizar la interacción entre variables categóricas. Esto nos permite identificar posibles interdependencias entre las variables que podrían influir en la precisión del modelo. Finalmente eliminamos un de las dos variables cuya correlación de Pearson es mayor que 0.9.

- Data Preparation

Una vez que el dataset ha sido limpiado y preparado, procedemos a estandarizar únicamente las variables numéricas que lo requieran. Este paso es esencial para crear un conjunto de datos de entrenamiento adecuado para el modelo CatBoost, el cual maneja bien variables categóricas pero se benefician de tener las variables numéricas estandarizadas o normalizadas. Tras este paso, aplicamos un One-Hot Encoding (OHE) a las variables categóricas que lo necesiten, preparando así los datasets para su uso en modelos como la Regresión Logística, XGBoost y LightGBM, que requieren codificación explícita de las variables categóricas.

Paralelamente, creamos el dataframe final sobre el que aplicaremos el modelo de predicción, es decir, aquel que contiene los clientes de nuestra base de datos del mes de junio que aún no poseen el producto que queremos predecir. Este dataset no incluirá la columna

objetivo (TARGET), ya que precisamente esa es la variable que deseamos predecir.

Para crear este dataframe, seleccionamos a los clientes presentes en el mes de mayo, asumiendo que la base de clientes en junio es similar a la del mes anterior. Sin embargo, eliminamos a los clientes que ya han comprado o poseen el producto, ya que consideramos que un cliente que adquiere el producto lo mantendrá y no se dará de baja, por lo que no puede volver a comprarlo. Una vez creado este nuevo dataframe, aplicamos el mismo procesamiento a las variables (estandarización de numéricas y codificación de categóricas) para que esté listo para el modelo final.

Este enfoque asegura que el dataset final esté bien preparado, siguiendo los mismos criterios de limpieza y transformación utilizados durante la fase de entrenamiento, lo que garantiza consistencia y precisión en las predicciones de nuestro modelo.

- Modeling y evaluation

En este apartado vamos a preparar el dataset para aplicar distintos modelos que nos ayuden a resolver el problema de clasificación binaria planteado y los vamos a aplicar. Para ello, comenzaremos realizando la partición de los datos en tres conjuntos: train, validation y test. Posteriormente, implementaremos los modelos y evaluaremos su rendimiento para seleccionar el que mejor se adecúe a nuestro objetivo.

Sabemos que para entrenar un modelo de forma efectiva es esencial dividir el dataset procesado en los siguientes tres conjuntos:

1. **Train:** Muestra de datos utilizada para ajustar el modelo, donde este aprende los patrones del dataset.
2. **Validation:** Muestra utilizada para proporcionar una evaluación no sesgada durante el ajuste de los hiperparámetros del modelo. Aunque, con el tiempo, la evaluación se vuelve más sesgada al incorporarse información de esta muestra en la configuración del modelo.
3. **Test:** Muestra utilizada para la evaluación final del ajuste del modelo en datos no vistos previamente, proporcionando una evaluación imparcial.

El objetivo de este proyecto es determinar cuán propenso es un cliente a comprar un producto específico en el mes siguiente. En el caso de los datasets de debit\_card y pension\_plan, cada partición está balanceada; es decir, en cada mes hay una cantidad similar de clientes que compran el producto y que no lo compran. Por ello, se decidió crear los datasets de train, validation y test basados en criterios temporales. Para debit\_card, el dataset comprende el periodo entre 2019-01-01 y 2019-05-01, mientras que para pension\_plan el rango temporal va desde el 2018-08-01 y 2019-05-01. La división de los datos sigue un esquema específico que se detalla en el cuadro 6.

	debit_card		pension_plan		long_term
	Periodo	%	Periodo	%	%
Train	2019-01-01: 2019-03-01	60.87 %	2018-08-01:2019-01-01	57.75 %	59.49 %
Validation	2019-04-01:2019-04-01	19.52 %	2019-02-01:2019-03-01	20.29 %	25.51 %
Test	2019-05-01	19.61	2019-04-01:2019-05-01	21.95 %	15.01 %

Cuadro 6: Información sobre la división de los dataset de cada producto en train, test y validacion

Por otro lado, el dataset de long\\_term\\_deposit presenta un desbalance en las diferentes particiones, ya que en los últimos meses se observa una disminución considerable en el número de compras. Para este caso, decidimos implementar un método de partición basado en Random Holdout, dividiendo los datos aleatoriamente en conjuntos de train, validation y test, con el fin de manejar mejor el desbalance en los datos y asegurar una evaluación adecuada del modelo.

Esta estrategia garantiza una correcta preparación de los datos para los modelos, considerando tanto las particularidades de cada dataset como la naturaleza del problema de clasificación binaria.

Una vez tenemos listos los datasets de train, test y validación aplicamos cada uno de los modelos descritos anteriormente, implementando para cada uno de ellos un Grid Search Cross Validation, a excepción del algoritmo de regresión logística. El Grid Search Cross Validation es una técnica utilizada para encontrar la mejor combinación de hiperparámetros de un modelo mediante una búsqueda exhaustiva. Consiste en definir un conjunto de posibles valores para cada hiperparámetro y probar todas las combinaciones posibles (Grid Search). Para evaluar cada combinación de manera más robusta, se aplica la validación cruzada, dividiendo los datos en varios subconjuntos (folds). El modelo se entrena en algunos de estos subconjuntos y se evalúa en los restantes, repitiendo el proceso varias veces. De este modo, obtenemos el conjunto de hiperparámetros que mayor rendimiento aportan al modelo.

Una vez entrenados los modelos, calculamos varias métricas para evaluar su rendimiento y compararlos de manera objetiva. Para todos los modelos, a excepción del modelo de regresión logística, utilizamos un conjunto completo de métricas. En el caso de la regresión logística, nos centramos únicamente en la curva ROC AUC y la curva precision-recall. Estas métricas nos proporcionan una visión detallada de la capacidad predictiva de cada modelo y nos permiten seleccionar el mejor enfoque para nuestro problema.

- 1. Feature Importance:** Esta métrica muestra la relevancia o el impacto que tiene cada variable en las predicciones del modelo. En los modelos como XGBoost, LightGBM o CatBoost, se puede calcular la importancia de cada característica en función de cómo contribuyen a la reducción de la pérdida del modelo durante el entrenamiento. Las características con mayor importancia indican aquellas que son más influyentes en la toma de decisiones del modelo.
- 2. Gráfica de Pérdidas (Train, Validation):** Esta gráfica muestra la evolución de la pérdida o "error" del modelo tanto en el conjunto de entrenamiento como en el de validación a lo largo de las iteraciones o épocas. Nos permite evaluar si el modelo está sufriendo de sobreajuste (overfitting), cuando el error en el conjunto de validación es mucho mayor

que en el conjunto de entrenamiento, o si está infraajustado (underfitting), cuando ambos errores son altos. Es útil para ajustar los hiperparámetros del modelo y mejorar su rendimiento.

3. **Curva ROC AUC:** La curva ROC (Receiver Operating Characteristic) muestra la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) a diferentes umbrales de clasificación. El área bajo la curva (AUC) mide el rendimiento general del modelo: cuanto mayor es el AUC, mejor es el modelo para distinguir entre clases. Un AUC cercano a 1 indica un buen rendimiento, mientras que un valor de 0.5 implica un modelo sin capacidad predictiva.
4. **Classification Report:** El classification report proporciona una descripción detallada de las métricas de precisión, recall, F1-score y soporte (support) para cada clase en un problema de clasificación.
  - **Precisión** (Precision) mide la proporción de predicciones positivas correctas frente al total de predicciones positivas realizadas por el modelo.
  - **Recall** mide la proporción de instancias verdaderamente positivas que fueron correctamente identificadas por el modelo.
  - **F1-score** es la media armónica entre precisión y recall, proporcionando una métrica equilibrada cuando ambas son importantes.
  - **Support** indica el número de ocurrencias de cada clase en los datos reales, ayudando a interpretar mejor el rendimiento del modelo en casos de clases desbalanceadas.

Cada una de estas métricas nos brinda información clave sobre el comportamiento de los modelos, permitiendo un análisis exhaustivo de su capacidad para predecir correctamente las etiquetas de los datos.

debit_card	AUC	Precision		Recall		F1-score	
		0	1	0	1	0	1
XGBoost	0.96	0.96	0.91	0.92	0.96	0.94	0.93
CatBoost	0.98	0.96	0.91	0.92	0.96	0.94	0.93
LightGBM	0.93	0.96	0.91	0.92	0.96	0.94	0.93
Log-Reg	0.97	0.97	0.86	0.86	0.96	0.91	0.91

Cuadro 7: Información sobre las métricas de evaluación de los distintos modelos para el dataframe de debit\_card

pension_plan	AUC	Precision		Recall		F1-score	
		0	1	0	1	0	1
XGBoost	1	0.99	0.99	0.99	0.99	0.99	0.99
CatBoost	1	0.99	0.99	0.99	0.96	0.94	0.93
LightGBM	1	0.99	0.99	0.99	0.99	0.99	0.99
Log-Reg	1	0.98	0.94	0.94	0.98	0.96	0.96

Cuadro 8: Información sobre las métricas de evaluación de los distintos modelos para el dataframe de pension\_plan

long\term_debit	AUC	Precision		Recall		F1-score	
		0	1	0	1	0	1
XGBoost	0.98	0.95	0.91	0.91	0.95	0.93	0.93
CatBoost	0.98	0.96	0.91	0.91	0.96	0.93	0.94
LightGBM	0.98	0.95	0.91	0.91	0.95	0.93	0.93
Log-Reg	0.96	0.92	0.88	0.88	0.93	0.90	0.91

Cuadro 9: Información sobre las métricas de evaluación de los distintos modelos para el dataframe de long\_term\_debit

Una vez que hemos verificado que nuestros modelos presentan un alto rendimiento (ver cuadros 7, 8 y 9), procedemos a calcular la probabilidad de compra que cada uno de ellos asigna a los clientes de junio, aplicando cada modelo al dataset correspondiente (con el método de predict\_proba). A continuación, calculamos la media aritmética de las cuatro probabilidades generadas por los distintos modelos, lo que nos proporciona una probabilidad final de compra para cada cliente. Este enfoque, al combinar las predicciones de múltiples modelos, resulta en estimaciones más robustas y confiables en comparación con las que ofrecería un solo modelo. Con esta información, podemos establecer un ranking de los clientes con mayor propensión a adquirir el producto en junio, lo que nos permitirá priorizar su contacto en caso de que se decida llevar a cabo una campaña de marketing.

### 3.3. Segmentación de clientes

La tercera tarea que nos asigna Carol (la directora general) es segmentar la base de clientes en 7 u 8 grupos de clientes similares para así conocer mejor a los clientes. Esta segmentación nos permitirá personalizar mejor los productos y servicios, optimizar las campañas de marketing al enfocarlas en segmentos específicos, y diseñar estrategias para retener y fidelizar a los clientes al comprender mejor sus necesidades y preferencias.

#### 3.3.1. Uso de KMeans

Para llevar a cabo esta tarea, hemos optado por utilizar el algoritmo KMeans por ser uno de los métodos más eficientes y sencillos para agrupar datos en clusters basados en similitudes. KMeans funciona asignando cada punto de datos al cluster cuyo centro o centroide esté más cercano, utilizando como métrica de proximidad la distancia euclídea. Esta métrica mide la distancia recta entre dos puntos en un espacio n-dimensional, lo que significa que los puntos dentro de un cluster estarán más cerca entre sí, minimizando la distancia a su centroide.

El algoritmo comienza seleccionando aleatoriamente un número predefinido de centroides. Luego, en cada iteración, asigna cada punto de datos al centroide más cercano y recalcula los centroides basados en los nuevos grupos de puntos asignados. Este proceso se repite hasta que los centroides no cambian significativamente, es decir, cuando las asignaciones de los puntos a los clusters son estables. KMeans es especialmente útil cuando se trabaja con grandes volúmenes de datos y se busca simplicidad en la segmentación.

### 3.3.2. Resolución

Para esta tarea partimos de la tabla info\_clientes que creamos con anterioridad, la cual posee la toda la información de los clientes desde su primer registro en la base de datos. Como el objetivo de la empresa es reducir la actividad intensiva de captación y aumentar la penetración de la cartera actual de productos, nos vamos a centrar en segmentar los clientes del último mes. Es por ello que de la tabla info\_clientes seleccionamos únicamente la partición del último mes (es decir, pk\_partition = '2019-05-01'). De este modo obtenemos la información de los 442995 clientes que tenemos registrados dicho mes.

Una vez tenemos la tabla que usará el modelo para realizar la segmentación, nos disponemos a aplicar la pipeline común para todo problema de Machine Learning.

- Preparación Datos

El primer paso para realizar la segmentación fue preparar y limpiar el conjunto de datos disponible a partir de la tabla info\_clientes confeccionada en el notebook CAPSTONE. Para ello comenzamos con la imputación de los nulos que los encontramos en las columnas "salary", "salary\_bracket" y "total\_spend".

En el caso de "salary" a los que tuvieran menos de 25 años y más de 65 años decidimos imputarles 0 ya que era comprensible que ya no estuvieran trabajando, por otro lado, para el resto de nulos decidimos sacar la mediana según grupos realizados a partir del segmento, la edad y el canal de entrada.

Prescindimos de la columna "salary\_bracket" al aportar la misma información de "salary".

Para "total\_spend", escogimos varios criterios de imputación:

- Ver registros nulos de clientes que nunca hayan tenido posesión de productos.
- Ver registros nulos de clientes que solo hayan tenido un producto.
- Imputando la media agrupando por segmento, edad y máximo de productos registrados en su historial.

Para el primer punto, asignamos 0 como valor. Para el segundo punto asignamos el precio del producto contratado. Para el tercer grupo realizamos la imputación planteada.

Una vez limpio de nulos, seleccionamos las columnas que más relevantes nos parezcan, siendo aquellas columnas cuyos datos incluyan variables demográficas, como edad, salario y género, así como información relativa a los diferentes productos financieros, la actividad comercial y las ventas históricas.

Por último, procedemos a realizar una transformación de variables categóricas a numéricas, a través de One Hot Encoder, y una posterior normalización de las variables numéricas (edad, salario, etc.) mediante RobustScaler, para que todas las características tuvieran el mismo peso en el algoritmo de clustering.

Esto garantizó que los datos estuvieran listos para ser procesados por el algoritmo K-Means.

Variable	Descripción
age	Edad del cliente en años
salary	Ingreso mensual estimado del cliente
total_products	Número total de productos financieros que posee el cliente
total_products_no_account	Número total de productos excluyendo cuentas bancarias
has_payment_card	Boolena: 1 si el cliente posee una tarjeta de débito o crédito, 0 si no
num_accounts	Número de cuentas bancarias que tiene el cliente
active_customer	Boolena: 1 si el cliente está activo, 0 si está inactivo
meses_antiguedad	Antigüedad del cliente en meses desde que se unió a easyMoney
total_spend	Gasto total acumulado del cliente
months_since_last_purchase	Número de meses transcurridos desde la última compra del cliente
entry_channel_KHE	Boolena: 1 si el cliente fue captado a través del canal KHE, 0 si no
entry_channel_KHQ	Boolena: 1 si el cliente fue captado a través del canal KHQ, 0 si no
entry_channel_KFC	Boolena: 1 si el cliente fue captado a través del canal KFC, 0 si no
entry_channel_KHM	Boolena: 1 si el cliente fue captado a través del canal KHM, 0 si no
entry_channel_KAT	Boolena: 1 si el cliente fue captado a través del canal KAT, 0 si no
entry_channel_KHK	Boolena: 1 si el cliente fue captado a través del canal KHK, 0 si no
entry_channel_KHN	Boolena: 1 si el cliente fue captado a través del canal KHN, 0 si no
entry_channel_RED	Boolena: 1 si el cliente fue captado a través del canal RED, 0 si no
entry_channel_KFA	Boolena: 1 si el cliente fue captado a través del canal KFA, 0 si no
entry_channel_KHD	Boolena: 1 si el cliente fue captado a través del canal KHD, 0 si no
entry_channel_Unknown	Boolena: 1 si el canal de captación es desconocido, 0 si no
entry_channel_KHL	Boolena: 1 si el cliente fue captado a través del canal KHL, 0 si no
entry_channel_KHF	Boolena: 1 si el cliente fue captado a través del canal KHF, 0 si no
entry_channel_KHO	Boolena: 1 si el cliente fue captado a través del canal KHO, 0 si no
entry_channel_Others	Boolena: 1 si el cliente fue captado a través de otros canales, 0 si no
gender_H	Boolena: 1 si el género del cliente es hombre, 0 si no
gender_V	Boolena: 1 si el género del cliente es mujer, 0 si no
segment_Particulares	Boolena: 1 si el cliente pertenece al segmento "Particulares", 0 si no
segment_Top	Boolena: 1 si el cliente pertenece al segmento "Top", 0 si no
segment_Universitario	Boolena: 1 si el cliente pertenece al segmento "Universitario", 0 si no
segment_Unknown	Boolena: 1 si el segmento del cliente es desconocido, 0 si no

Cuadro 10: Tabla con las características seleccionadas para el clustering

- Determinación del número óptimo de clústeres

Para seleccionar el número adecuado de clústeres, se utilizó la gráfica elbow curve, que evalúa la suma de las distancias cuadradas dentro de los clústeres. Se observó que el número óptimo de clústeres se encontraba en 7, ya que a partir de ese punto la mejora en la variabilidad explicada era mínima.

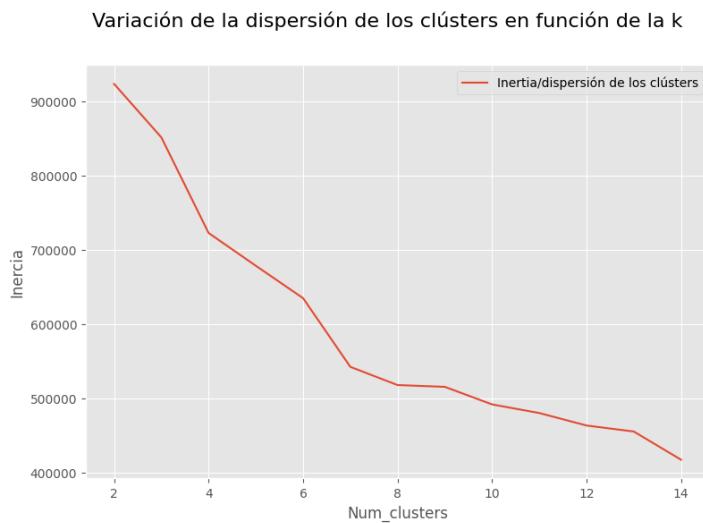


Figura 2: Gráfico Elbow Curve

- Aplicación del algoritmo K-Means

Con el número óptimo de clusters determinado, procedimos a implementar el algoritmo K-Means. Este algoritmo agrupa a los clientes basándose en las variables seleccionadas anteriormente, minimizando las distancias dentro de cada grupo y maximizando las diferencias entre los grupos. Tras entrenar el modelo, se generaron 7 clusters que capturan distintas características y comportamientos de los clientes de easyMoney. Cada cluster fue analizado para identificar patrones comunes y desarrollar perfiles que permitan una segmentación más precisa.



Figura 3: Distribución de los clústers obtenidos

## 4. Resultados

En este apartado se presentan los resultados obtenidos durante el desarrollo de los modelos de propensión a compra y segmentación de clientes. Se analizarán las variables más relevantes que han influido en las predicciones de cada uno de los modelos y los distintos clusters que se han obtenido. Finalmente, se propone una campaña comercial basada en los resultados de los modelos y alineada con los objetivos estratégicos de la empresa.

### 4.1. Modelos de propensión a compra

Como hemos explicado en la sección 3.2, hemos realizado un modelo de propensión a compra para cada uno de los productos en los que vamos a centrar nuestra campaña.

Cada uno de estos modelos asigna diferentes pesos a las variables para elaborar las probabilidades de compra para cada cliente. Para saber cuales han sido esas variables que más han aportado a cada modelo hemos hecho uso de distintas métricas.

Para los modelos de XGBoost y LightGBM, hemos observado la importancia de cada variable gracias a el método `plot_importance`. Este método permite visualizar la importancia de las características en el modelo mediante una gráfica de barras que, en el eje y muestra las variables y en el eje x, la métrica utilizada (imagen 4).

Tanto en XGBoost como LightGBM, la métrica usada es la frecuencia (Weight o F-score en XGBoost y Split en LightGBM). Esta hace referencia a el número de veces que una característica es utilizada para dividir los datos en los árboles de decisión del modelo. Cuanto más veces se usa una variable para hacer un split, mayor es su importancia.

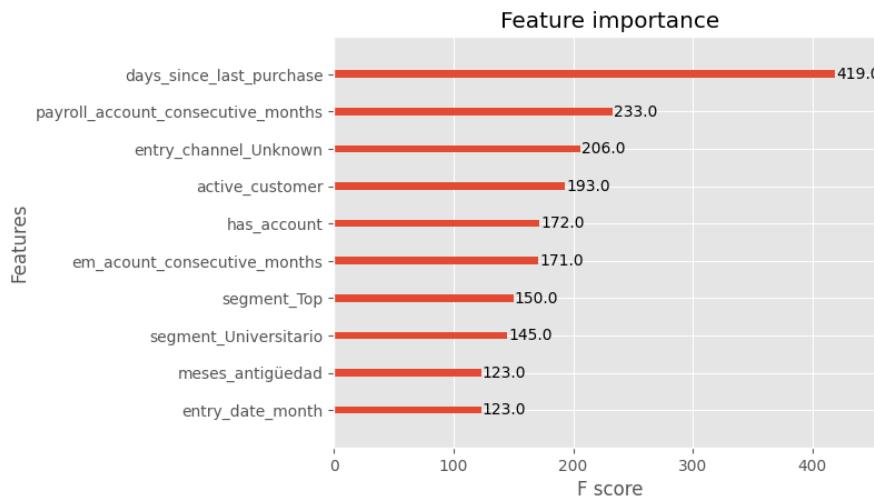


Figura 4: Gráfico resultante tras aplicar el método `plot_importance` al modelo de XGBoost entrenado en el dataframe de `debit_card`.

Por otro lado, en CatBoost, la importancia de las variables se mide utilizando lo que se conoce como Ganancia Total (Total Gain). Este método evalúa la importancia de una característica al observar la ganancia en la función de pérdida cuando esa característica se usa para dividir los datos en los árboles de decisión.

En el modelo de Regresión Logarítmica la importancia de las variables se mide a través de los coeficientes del modelo. Estos coeficientes indican la magnitud y dirección del impacto de cada característica sobre la variable objetivo de modo que cuanto mayor sea el valor absoluto del coeficiente, mayor es la influencia de la variable en la predicción. Los coeficientes positivos aumentan la probabilidad de la clase positiva, y los negativos la disminuyen.

	Variable
XGBoost	days_since_last_purchase
	payroll_account_consecutive_months
	entry_channel_Unknown
LightGBM	days_since_last_purchase
	meses_antiguedad
	em_acount_consecutive_months
Catboost	days_since_last_purchase
	active_customer
	segment
Log-Reg	days_since_last_purchase
	entry_channel_Unknown
	active_customer

Cuadro 11: Variables más relevantes por algoritmo para el modelo de propensión a compra de debit\_card

	Variable
XGBoost	num_accounts
	payroll_account_consecutive_months
	days_since_last_purchase
LightGBM	days_since_last_purchase
	age
	em_acount_consecutive_months
Catboost	num_accounts
	days_since_last_purchase
	payroll_account_consecutive_months
Log-Reg	num_accounts
	em_acount_consecutive_months
	has_account

Cuadro 12: Variables más relevantes por algoritmo para el modelo de propensión a compra de pension\_plan

long_term_deposit	Variable
XGBoost	days_since_last_purchase
	age
	meses_antiguedad
LightGBM	days_since_last_purchase
	age
	meses_antiguedad
Catboost	active_customer
	days_since_last_purchase
	age
Log-Reg	active_customer
	entry_channel_KHQ
	entry_channel_KHL

Cuadro 13: Variables más relevantes por algoritmo para el modelo de propensión a compra de long-term\_deposit.

Una vez identificadas las variables más significativas para cada modelo y producto, calculamos la probabilidad de compra de cada cliente para cada producto mediante la media aritmética de las probabilidades proporcionadas por cada modelo, como se detalla en la sección 3.2.3. A continuación, elaboramos un listado de los top clientes con mayor probabilidad de compra para cada producto y analizamos las características que los distingúan. Este análisis nos permite identificar el perfil de los clientes más propensos a adquirir los productos, facilitando así la segmentación de estrategias de venta y publicidad (gráficas disponibles en el Anexo 7).

## 4.2. Análisis de los clusters

La implementación del algoritmo K-Means permitió segmentar a los clientes en distintos clusters, cada uno con características diferenciadas en cuanto a comportamiento y atributos financieros. En este apartado, se realizará un análisis detallado de los grupos obtenidos, describiendo las principales características que definen a los clientes en cada cluster. Este análisis permitirá comprender mejor los perfiles de los clientes, facilitando la identificación de oportunidades estratégicas y de personalización de ofertas, así como el potencial de crecimiento de cada segmento.

### ■ Clúster 1: Mujeres jóvenes con productos limitados

Este grupo está compuesto exclusivamente por mujeres menores de 35 años, quienes poseen un único producto financiero, predominantemente asociado a cuentas bancarias. Estas clientas no tienen tarjetas de pago y su interacción con la entidad se concentra en la posesión de una sola cuenta bancaria. El canal de entrada más común para este grupo ha sido KHE, con una antigüedad que varía entre 15 y 50 meses. Aunque su gasto total acumulado asciende a 70K, la mayoría de estas clientas han dejado de ser activas, con un 75 % de inactividad. Su salario se encuentra en el rango de 60K a 90K, lo que refleja un poder adquisitivo moderado-alto, aunque su categoría dentro de la segmentación corresponde mayoritariamente al segmento Universitario. Entre los productos más contratados en este grupo destacan debit\_card y payroll\_account.

A pesar de la inactividad de una gran parte del grupo, el nivel salarial relativamente alto sugiere que podría haber oportunidades de mejorar la relación con estos clientes mediante la oferta de productos adicionales o servicios más personalizados.

#### ■ Clúster 2: Mujeres con comportamiento financiero variado

El clúster 2 está compuesto por mujeres entre los 30 y 45 años, quienes poseen uno o ningún producto financiero, y en todos los casos estos productos están asociados a cuentas bancarias. Estas clientas no poseen tarjetas de pago y presentan una heterogeneidad significativa en su comportamiento financiero, tanto en términos de antigüedad como de gasto. Algunas de ellas registran un gasto total de 70K, mientras que otras no han efectuado transacciones relevantes. Su relación con la entidad ha sido captada a través del canal KFC, y su actividad está dividida, con una proporción 50/50 entre clientas activas e inactivas. Las clientas activas tienen un salario que oscila entre 75K y 90K, mientras que las inactivas no registran ingresos o transacciones recientes. El principal producto adquirido por este grupo es la emc\_account, y el segmento predominante es el de Particulares.

El comportamiento de este grupo es variado, lo que dificulta su análisis, pero la existencia de un número significativo de clientas activas con niveles de ingresos altos sugiere que puede haber margen para ampliar su relación con la entidad.

#### ■ Clúster 3: Hombres jóvenes con productos básicos

Este grupo está compuesto exclusivamente por hombres menores de 25 años que poseen un único producto financiero, generalmente asociado a cuentas bancarias. Ninguno de ellos posee tarjetas de pago, y la mayoría de estos clientes presentan una actividad baja, con más del 70 % de inactivos. La antigüedad de estos clientes en la entidad es reducida, entre 5 y 10 meses, y en su mayoría no registran ingresos salariales, lo que limita sus interacciones financieras con la entidad. A pesar de esto, algunos de estos clientes han alcanzado gastos acumulados entre 50K y 75K, lo que indica cierta actividad, aunque intermitente. Estos clientes fueron captados principalmente a través del canal KHO, y pertenecen en su totalidad al segmento Universitario. Entre los productos más contratados se encuentran la payroll\_account y la debit\_card.

El bajo nivel de ingresos y la inactividad de la mayoría de estos clientes limitan considerablemente su potencial de crecimiento, aunque aquellos que han mostrado actividad pueden ser objeto de ofertas para fortalecer su relación con la entidad.

#### ■ Clúster 4: Adultos con productos diversificados

Este clúster agrupa a tanto hombres como mujeres, con una proporción de 60 % de hombres y 40 % de mujeres, cuyas edades oscilan entre 30 y 45 años. Estos clientes tienen una relación más diversificada con la entidad, ya que poseen al menos dos productos financieros, entre los que se incluyen tanto cuentas bancarias como productos no relacionados con cuentas. A diferencia de otros clústeres, estos clientes ya poseen tarjetas de pago y han sido captados principalmente a través del canal KFC. Aunque su antigüedad es variada, su actividad se mantiene elevada, con la mayoría de ellos siendo clientes activos y con un nivel salarial que oscila entre 75K y 100K. Sin embargo, no han registrado gastos recientes, con 17 meses desde la última compra. La debit\_card es el producto más contratado por este grupo, y el segmento

predominante es el de Particulares.

Este grupo tiene un alto potencial de crecimiento, dado que ya tienen una relación diversificada con la entidad y podrían incrementar su uso de productos adicionales si se les ofrece una estrategia de fidelización adecuada.

- **Clúster 5: Mujeres jóvenes con baja actividad financiera**

Este clúster está compuesto por hombres menores de 25 años, quienes poseen un único producto financiero, principalmente asociado a cuentas bancarias. No tienen tarjetas de pago y su actividad financiera es baja, con la mayoría de ellos inactivos (70%). Aunque su antigüedad varía, pocos registran ingresos salariales, lo que limita su capacidad de interacción con la entidad. A pesar de esto, algunos han registrado gastos acumulados de entre 60K y 80K. El canal de entrada predominante para este grupo es KHO, y pertenecen principalmente al segmento Universitario. Los productos más contratados por estos clientes son la payroll\_account y la debit\_card.

Este grupo presenta un potencial limitado debido a la baja actividad y falta de ingresos en la mayoría de los clientes. Aunque algunos han registrado actividad, es necesario evaluar las posibilidades de reactivación para este perfil de cliente.

- **Clúster 6: Hombres jóvenes con productos limitados**

Este grupo está compuesto exclusivamente por hombres menores de 30 años, quienes poseen un único producto financiero, generalmente cuentas bancarias. No tienen tarjetas de pago y la mayoría presentan niveles bajos de actividad, con un 75 % de ellos inactivos. Su antigüedad en la entidad varía entre 20 y 45 meses, y aunque algunos han alcanzado un gasto total de 70K, su comportamiento financiero es intermitente. Este grupo fue captado principalmente a través del canal KHE, y su salario presenta una variabilidad considerable entre los 60K – 90K. Pertenece en su mayoría al segmento Universitario y entre los productos más adquiridos destacan la debit\_card y la payroll\_account.

Este grupo tiene un potencial moderado, con algunos clientes que muestran niveles de gasto significativos. Sin embargo, la inactividad de la mayoría limita las oportunidades de crecimiento.

- **Clúster 0: Hombres con productos limitados y actividad variada**

Este clúster agrupa a hombres de entre 35 y 45 años que poseen uno o ningún producto financiero. Estos clientes no tienen tarjetas de pago y su comportamiento financiero es muy heterogéneo. La mayoría de estos clientes fueron captados a través del canal KFC, y aunque su antigüedad varía, tiende a ser menor a los 10 meses. El comportamiento financiero de estos clientes es mixto, con algunos alcanzando gastos de hasta 70K, mientras que otros no han realizado transacciones recientes. Sus ingresos salariales oscilan entre 65K y 95K, y presentan un perfil de actividad intermitente, con un 50 % de clientes activos y el otro 50 % inactivos. El producto más comúnmente adquirido por este grupo es la emc\_account, y pertenecen al segmento de Particulares.

El potencial de crecimiento para este grupo es moderado, con algunos clientes mostrando comportamientos de gasto recurrentes que podrían incrementarse si se fortalecen las estrategias de vinculación.

#### 4.2.1. Implicaciones comerciales

La segmentación de clientes realizada mediante K-Means permite a easyMoney identificar de manera clara las diferencias entre los distintos grupos de clientes y, por lo tanto, personalizar las campañas de marketing de acuerdo con las necesidades y características de cada segmento tal como abordaremos en la siguiente sección 4.3. Para cada cluster hemos definido un buyer persona lo que permite identificar y comprender a los diferentes tipos de clientes que conforman su público objetivo. Esto ayudará a el equipo de marketing de easyMoney a tomar decisiones más informadas y centradas en el cliente, lo que incrementa la efectividad de las campañas y las interacciones comerciales (buyer persona disponibles en el Anexo 7).

### 4.3. Campaña comercial

En este punto, una vez aplicados los modelos de propensión a compra para los 3 productos seleccionados y teniendo a nuestros clientes clasificados en 7 grupos diferentes, hemos procedido a desarrollar las bases de una campaña de penetración de mercado con los clientes ya existentes estimando cuánto dinero podríamos ganar y aportando herramientas al equipo de marketing y ventas (como buyer persona que muestren un perfil de cada uno de los 7 grupos de clientes, así como el número de clientes de cada grupo a los que impactar con cada producto) para poder optimizar al máximo los recursos de Easy Money y tratar de revertir la situación actual de la empresa.

Así pues, para escoger primeramente los grupos a los que impactar con cada producto, hemos analizado los perfiles de los clientes más propensos a adquirir cada uno de los 3 productos (tarjeta de débito, plan de pensiones y depósito a largo plazo) y los hemos comparado con los 7 grupos de clientes clasificados. .

Para ello, hemos tenido en cuenta sobretodo las variables que más han aportado a cada modelo, según hemos destacado en la sección 4.1, pero también otras variables como por ejemplo si el cliente es considerado como 'active' o no (ya que el hecho de que fuera activo era un punto en común en el cien por cien de los clientes con mayor propensión a compra) o aspectos demográficos como la edad, el sexo o el salario.

Una vez relacionados los perfiles más propensos a la compra con los perfiles de cada cluster, hemos tratado de extraer un número aproximado de clientes a los que impactar dentro de cada cluster para cada uno de los productos. Para ello hemos cogido el total de clientes dentro de cada cluster, le hemos extraído aquellos que ya tenían el producto en cuestión y, solo en aquellos casos en los que la similitud entre el perfil del cluster y el perfil de propensión a compra no eran perfectos, hemos limitado por aquellos clientes considerados como activos (desechando así los no activos).

De este modo, el número de clientes dentro de cada cluster a los que impactar con cada producto quedaría del siguiente modo:

Product	Cluster	Nº de clientes/ Cluster
debit_card	2	53.100
	4	44.400
	0	70.800
pension_plan	2	25.400
	4	27.900
	0	32.700
long_term_deposit	2	51.200
	4	39.000
	0	32.100

Cuadro 14: Clusters a los que impactar con cada producto así como volumen de clientes a los que impactar dentro de cada cluster

A continuación, teniendo los productos a promocionar, el público al que impactaríamos (escogido por compartir ciertas características que lo hacen más proclive a adquirir estos productos) y el precio medio de cada uno de ellos (lo podemos ver en la siguiente tabla, en la columna de “average margin”), podemos estimar los posibles ingresos que conseguiríamos con esta campaña de penetración de mercado.

AVERAGE MARGIN	PRODUCT_DESC	FAMILY_PRODUCT
60,00	debit_card	payment_card
60,02	credit_card	payment_card
69,93	payroll	account
69,97	payroll_account	account
69,97	em_account	account
70,01	emc_account	account
1.486,43	short_term_deposit	investment
1.496,92	long_term_deposit	investment
1.497,83	securities	investment
1.499,05	funds	investment
2.183,92	loans	loan
2.532,51	mortgage	loan
5.976,44	pension_plan	pension_plan

Cuadro 15: Información precios medios por producto en dólares.

Asumiendo que la tasa de cierre media (o tasa de conversión) podría rondar en torno a un 3 % en debit card, un 1 % en pension plan y un 0,6 % en long term deposit (teniendo en cuenta que se trata de clientes ya existentes del banco y que hemos escogido los grupos más proclives a adquirir estos productos), hemos calculado los posibles ingresos que nos podría traer esta campaña:

$$\text{debit card} \rightarrow 303.000 \$ \\ (168.300 * 0,03) * 60$$

$$\text{pension plan} \rightarrow 5.140.000 \$ \\ (86.000 * 0,01) * 5976,44$$

**long term deposit → 1.098.000 \$**  
 $(122.300 * 0,006) * 1496,92$

En **Total**, podríamos conseguir alrededor de **6.541.000 \$** extra con esta campaña.

## 5. Conclusiones

Gracias a la nueva incorporación del perfil Data Science en la empresa easyMoney, se ha conseguido empezar a implementar una analítica más avanzada, adoptando así una metodología de trabajo data-drive y, de este modo, dando pie a emprender un camino hacia la inversión en tecnología, aspecto que tantos problemas y fricciones estaba empezando a generar en todas las áreas de la empresa.

Para empezar, se ha provisto a Dan (director de marketing) con un Dashboard con el que pueda mostrar de manera visual los principales KPIs a la dirección, así como poder analizar él mismo en su día a día las fortalezas, debilidades y puntos de mejora de su departamento.

Por otra parte, dejando a un lado el avance tecnológico de la empresa, otro de los puntos a abordar por easyMoney es el de obtener rentabilidad de la base actual de clientes, punto que su principal inversor Lion Global Management ha exigido a la dirección.

Para ello, y gracias al poder de los modelos de Machine Learning supervisados de clasificación empleados (XGBoost, CatBoost, LightBoost y Regresión Logística), hemos conseguido destacar qué tipo de cliente es más propenso a comprar ciertos productos, facilitando enormemente el trabajo en los departamentos de marketing y ventas. Teniendo estos datos a su alcance, los departamentos de Marketing y Ventas podrán enfocar su esfuerzo y presupuesto únicamente en aquellos clientes con mayor propensión a adquirir estos productos. En este caso, hemos aplicado los modelos a los 3 productos que podrían traer más beneficios a la empresa (pension\_plan, long\_term\_deposit, debit\_card), pero en un futuro podríamos aplicar modelos de propensión a compra al producto que necesitemos.

Además de destacar qué tipo de cliente es más propenso a adquirir uno de estos 3 productos, hemos empleado técnicas de clustering (o agrupamiento) utilizando el modelo K-Means, que nos ha permitido clasificar a toda nuestra base de clientes en 7 grupos, dotando así al departamento de marketing de una herramienta fundamental para el estudio del tipo de cliente que tenemos en la actualidad. Además, nos hemos tomado la libertad de extraer un buyer persona por cada uno de estos clusters (buyer persona disponibles en el Anexo 7) con el fin de facilitar la comprensión de los mismos y 'humanizar' los datos.

Finalmente, una vez clasificados los clientes en 7 grupos y destacados los tipos de perfil más propensos a adquirir cada uno de los 3 productos que habíamos comentado, hemos realizado un caso de uso que nos ha permitido estimar, en base a una campaña X, cuánto dinero podemos ganar. Así pues, hemos definido qué productos ofrecer, a cuántos clientes tenemos que impactar y finalmente hemos destacado unos posibles ingresos. La cantidad que podríamos obtener con una hipotética campaña es de alrededor de 6.541.000€.

Con todos estos datos, el departamento de Marketing va a tener las premisas y herramientas suficientes como para poder plantear una campaña exitosa de penetración de nuestra cartera actual de clientes.

Los siguientes pasos ya irían más enfocados en acabar de ajustar los Dashboards, automatizar la recogida y procesamiento de datos y adaptar nuevos Dashboards para otros departamentos (a parte del de Marketing y la directiva capitaneada por Carol).

Así pues, necesitaríamos sentarnos con Dan y Carol para ver de qué manera podemos pivotar ciertos aspectos del Dashboard para cubrir al 100 % sus necesidades. Posteriormente tendríamos que encontrar un método para automatizar la recogida y procesamiento de datos con el fin de que podamos tener un Dashboard en tiempo real con los KPIs más importantes para dirección y para marketing. Después de ello, sería interesante copiar este proceso para los diferentes departamentos que lo vayan necesitando para afrontar su día a día de la forma más eficiente. Y paralelamente a todo ello, tendremos que estar preparados por si Carol nos necesitase para generar modelos predictivos para el negocio.

## 6. Bibliografía

### Referencias

CatBoost Team (2024). *CatBoost Documentation*. Accessed: 2024-10-21.

**URL:** <https://catboost.ai/en/docs/>

LightGBM Developers (2024). *LightGBM Documentation: LGBMClassifier*. Accessed: 2024-10-21.

**URL:** <https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html>

Pandas Development Team (2024). *pandas - Python Data Analysis Library*. Accessed: 2024-10-21.

**URL:** <https://pandas.pydata.org/>

Scikit-learn Developers (2024a). *LogisticRegression: scikit-learn 1.5 documentation*. Accessed: 2024-10-21.

**URL:** [https://scikit-learn.org/1.5/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html)

Scikit-learn Developers (2024b). *scikit-learn: Machine Learning in Python*. Accessed: 2024-10-21.

**URL:** <https://scikit-learn.org/stable/>

## 7. Anexo

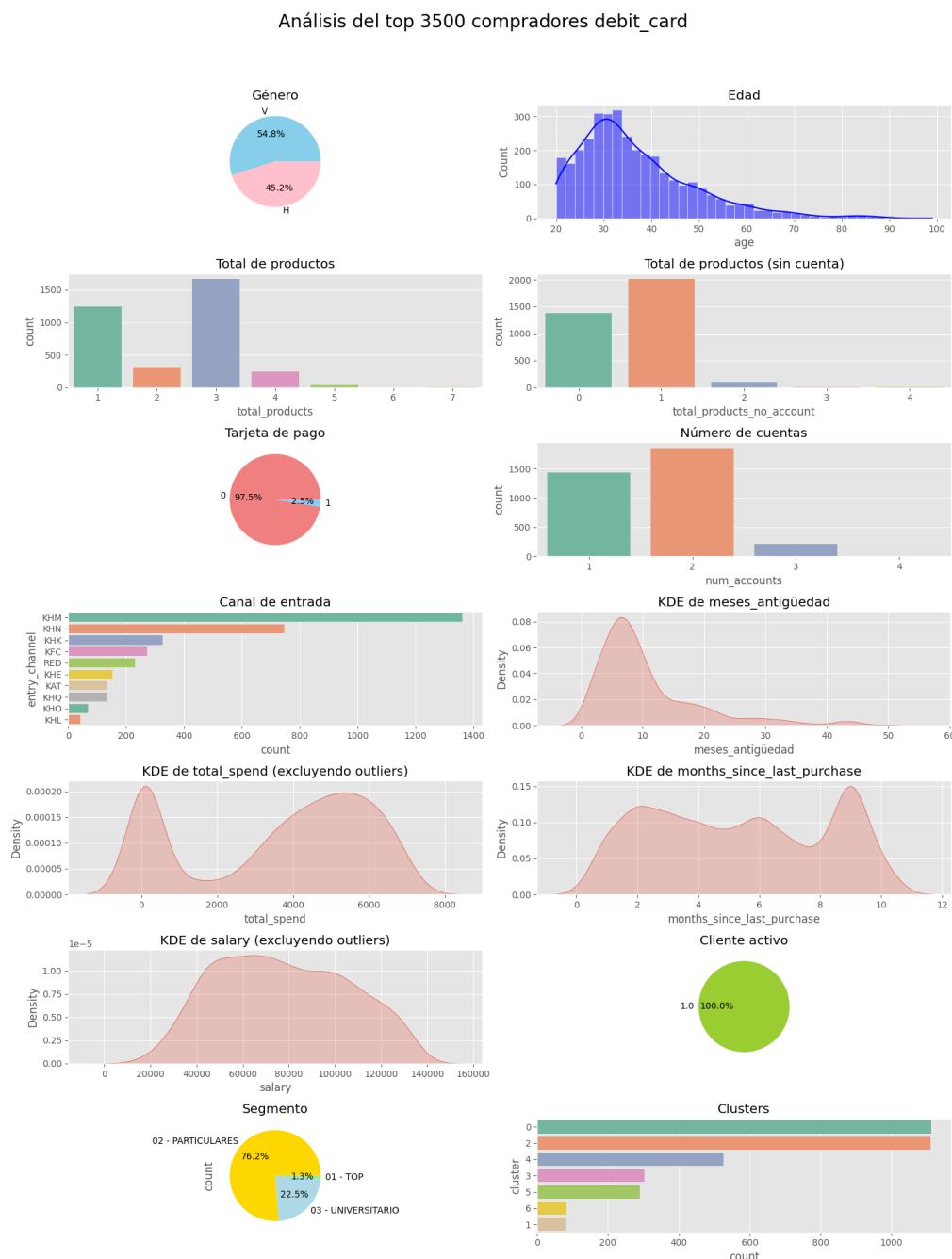


Figura 5: Análisis del top 3500 clientes más propensos a comprar debit\_card.



Figura 6: Análisis del top 2000 clientes más propensos a comprar pension\_plan.

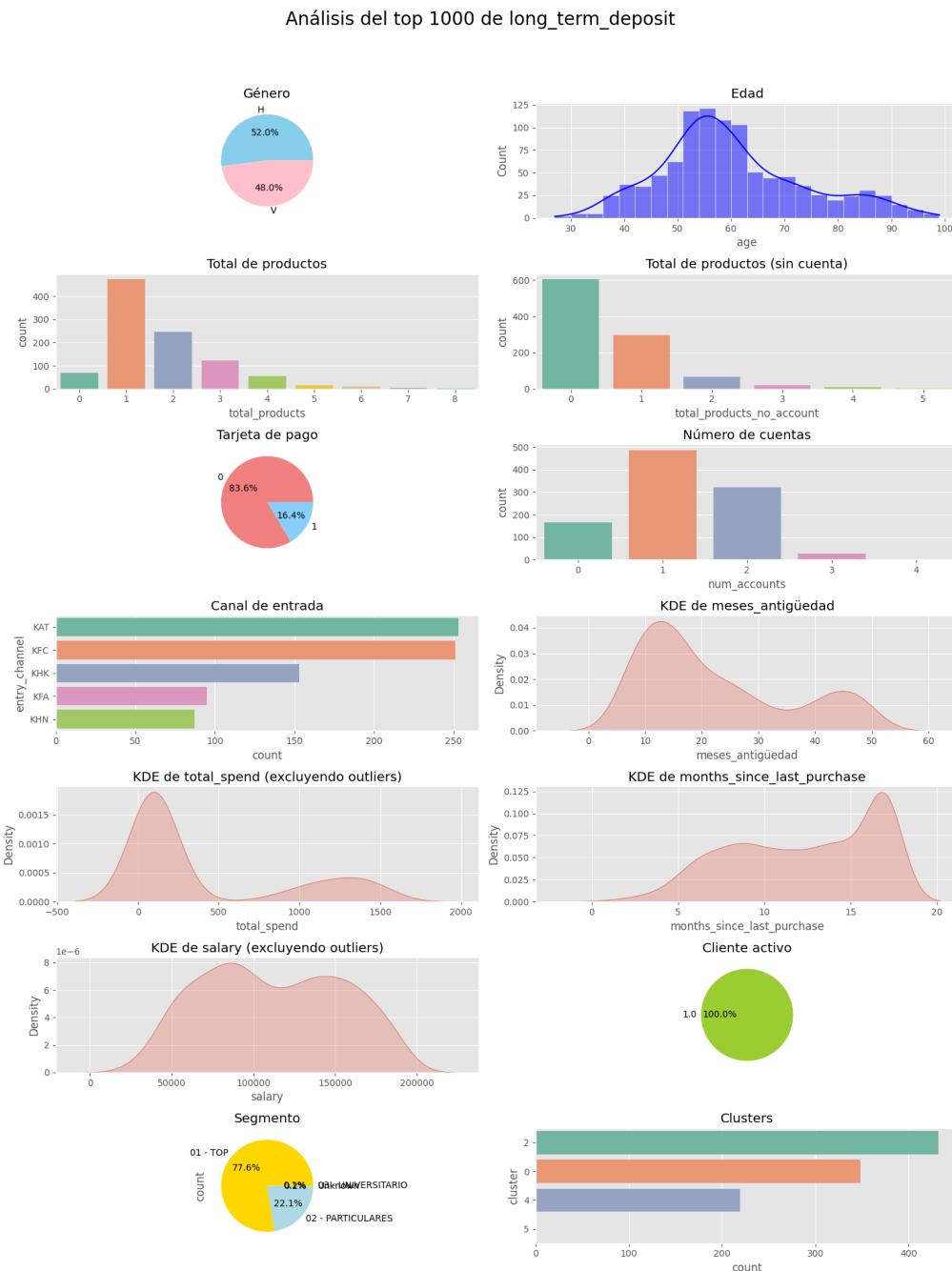


Figura 7: Análisis del top 1000 clientes más propensos a comprar long\_term\_deposit.

### Buyer Persona

## Cluster 1:

### Buyer Persona:



- **Nombre:** Lucía, 24 años
- **Ocupación:** Consultora en una empresa tecnológica
- **Ingresos:** 75K anuales
- **Objetivos:** Quiere comenzar a invertir en su futuro, pero aún no utiliza muchos productos bancarios.
- **Comportamientos:** Lucía apenas interactúa con su banco, lo ve más como una herramienta de fondo que no ocupa mucho en su vida diaria. Tiene su cuenta para gestionar su salario, pero no se ha preocupado por añadir más productos financieros. Hace tiempo que no revisa su app bancaria, ya que su rutina es sencilla: gestiona sus pagos y gastos de forma automática y sin demasiada complicación.
- **Necesidades:** Busca simplicidad en la gestión de su cuenta principal y quizás beneficios de tarjetas de débito o crédito si decide activarlas.

## Cluster 2:

### Buyer Persona:



- **Nombre:** Laura, 42 años
- **Ocupación:** Directora de marketing en una empresa de publicidad
- **Ingresos:** 100K anuales
- **Objetivos:** Mantener una relación básica con el banco, sin buscar productos adicionales complejos.
- **Comportamientos:** Laura tiene una relación funcional con su banco. No le dedica mucho tiempo a gestionar sus productos, porque no siente la necesidad. Su

prioridad es mantener sus cuentas en orden, y rara vez necesita asesoramiento. Solo revisa su cuenta cuando es absolutamente necesario, pero tiene una visión muy clara de que su banco es una herramienta práctica para sus finanzas diarias.

- **Necesidades:** Laura busca un servicio bancario que le permita gestionar sus cuentas de forma sencilla, con productos como cuentas a largo plazo (emc\_account) y posiblemente un plan de pensiones a medida.

## Cluster 3:

### Buyer Persona:



- **Nombre:** David, 21 años
- **Ocupación:** Estudiante universitario y consultor en una de las Big 4
- **Ingresos:** 50K anuales
- **Objetivos:** Manejar sus ingresos de prácticas y posibles futuros ahorros con una gestión sencilla.
- **Comportamientos:** David es de esos clientes que abre una cuenta porque lo necesita para cobrar su beca o gestionar sus primeros ingresos, pero no interactúa mucho más allá de lo esencial. Su relación con el banco es bastante esporádica, suele dejar que las cosas fluyan sin preocuparse demasiado. Sus compras son pocas y siempre ha preferido mantener un perfil financiero bajo por el momento.
- **Necesidades:** David necesita herramientas básicas como una cuenta nómina (payroll\_account) y una tarjeta de débito para sus compras diarias.

## Cluster 4:

### Buyer Persona:



- **Nombre:** Sergio, 35 años
- **Ocupación:** Gerente de ventas en una empresa de telecomunicaciones
- **Ingresos:** 100K anuales
- **Objetivos:** Mantener una relación activa con el banco mientras gestiona sus varias cuentas y productos financieros.
- **Comportamientos:** Sergio es activo con su banco. Consulta su cuenta regularmente, no solo para verificar su saldo, sino también para asegurarse de que sus productos financieros están funcionando a su favor. Le gusta estar al tanto de sus movimientos y sentirse en control de su dinero. Si bien tiene todo automatizado, no duda en revisar detalles cuando es necesario, buscando siempre la mejor opción para sus ahorros e inversiones.
- **Necesidades:** Sergio necesita una gestión integral de sus productos financieros, con prioridad en la comodidad y flexibilidad de sus pagos y ahorros.

## Cluster 5:

### Buyer Persona:



- **Nombre:** Paula, 21 años
- **Ocupación:** Estudiante de último año de carrera
- **Ingresos:** 80K anuales (subvenciones y apoyo familiar)
- **Objetivos:** Administrar sus finanzas como estudiante, aunque es poco activa en la adquisición de productos.
- **Comportamientos:** Paula es bastante relajada con sus finanzas. Sabe que tiene una cuenta bancaria, pero no la revisa con frecuencia. Al ser joven y estar enfocada

en sus estudios, no tiene demasiadas preocupaciones financieras por el momento. Para ella, el banco es solo un lugar donde su dinero está, y por ahora, no tiene prisa en cambiar eso.

- **Necesidades:** Paula requiere soluciones bancarias básicas como cuentas nómina y tarjetas de débito, sin demasiadas complicaciones adicionales.

## Cluster 6:

### Buyer Persona:



- **Nombre:** Javier, 25 años
- **Ocupación:** Ingeniero industrial
- **Ingresos:** 85K anuales
- **Objetivos:** Mantener una relación limitada con el banco mientras se enfoca en su carrera.
- **Comportamientos:** Javier no le da demasiada importancia a su banco. Para él, tener una cuenta es más una formalidad que una herramienta que usa activamente. Las pocas veces que interactúa con el banco es para revisar su saldo o hacer una transacción puntual. No siente que su relación con la entidad sea algo que requiera demasiada atención.
- **Necesidades:** Javier puede beneficiarse de productos como una tarjeta de débito o una cuenta nómina para gestionar mejor sus finanzas diarias.

## Cluster 0:

### Buyer Persona:



- **Nombre:** Manuel, 40 años
- **Ocupación:** Consultor independiente

- **Ingresos:** 90K anuales
- **Objetivos:** Administrar sus cuentas personales y profesionales de manera eficiente.
- **Comportamientos:** Manuel tiene una relación estable y equilibrada con su banco. Revisa sus cuentas de vez en cuando, especialmente cuando tiene que hacer una transferencia o comprobar alguna cuestión relacionada con su negocio. No es alguien que busque nuevos productos constantemente, pero aprecia tener sus cuentas en orden y estar seguro de que todo funciona como debería.
- **Necesidades:** Manuel requiere productos básicos como cuentas EMC y, ocasionalmente, un plan de pensiones o una cuenta nómina.