

## **MEMORIA PROYECTO FINAL**

### **MASTER EN DATA SCIENCE & IA - BARCELONA**

Proyecto: EasyMoney

Autores: Albert Fernández, Ezequiel Velasco, Julieta Kaplan, Roger Torruella & Rory Thomson.

Fecha de entrega: 22-03-2025

Programa de Máster: Data Science & IA

Curso: 21-dsc-0924-bcn

Tutora: Raquel Revilla



## Índice

<b>Índice</b>	<b>2</b>
<b>Introducción</b>	<b>3</b>
<b>Justificación y Contextualización</b>	<b>3</b>
<b>Objetivos</b>	<b>4</b>
<b>Metodología empleada</b>	<b>5</b>
1. Recopilación y preparación de datos:	5
2. Segmentación de clientes	6
3. Modelado predictivo	7
4. Resultados y predicciones	8
5. Visualización y Seguimiento con Power BI	9
6. Implementación	9
<b>Resultados y Conclusiones</b>	<b>10</b>
1. Resultados obtenidos	10
2. Impacto en la toma de decisiones	10
3. Discusión de los resultados	10
<b>Referencias / Bibliografía</b>	<b>12</b>

## Introducción

EasyMoney es una fintech con cuatro años de trayectoria en el mercado financiero, centrada en la oferta de productos financieros accesibles y digitales. A pesar de su crecimiento y el éxito en la obtención de financiación en dos rondas de inversión, la compañía **aún no ha alcanzado rentabilidad**, lo que ha generado **presión por parte de sus inversores** para mejorar su desempeño antes de considerar nuevos aportes de capital.

En este contexto, **la empresa enfrenta varias dificultades**:

- **Estrategia comercial desalineada** con su visión original, lo que ha generado ineficiencias en la captación y fidelización de clientes.
- **Restricciones presupuestarias**, limitando las iniciativas comerciales.
- **Falta de un modelo analítico avanzado**, lo que impide tomar decisiones basadas en datos para optimizar la adquisición y retención de clientes.

El reto principal consiste en desarrollar un **modelo de predicción y segmentación de clientes** que permita a **EasyMoney mejorar su rentabilidad** y optimizar su estrategia comercial mediante la aplicación de **técnicas de Data Science y Machine Learning**.

A través de modelos de segmentación y predicción, buscamos responder preguntas clave como:

- ¿Cuáles son los clientes con mayor probabilidad de compra?
- ¿Qué características diferencian a los clientes rentables de los no rentables?
- ¿Cómo podemos personalizar la comunicación para maximizar la conversión?

La memoria se estructura de la siguiente manera:

1. Justificación y contextualización: Explicación del motivo del estudio y su importancia.
2. Objetivos: Definición de la finalidad del proyecto y los resultados esperados.
3. Metodología: Descripción del enfoque teórico-práctico aplicado y los métodos empleados.
4. Resultados y conclusiones: Presentación de hallazgos clave y evaluación del impacto del análisis de datos en la toma de decisiones comerciales.

## Justificación y Contextualización

El equipo está compuesto por profesionales de diversas disciplinas, incluyendo marketing, ingeniería, finanzas, consultoría y tecnología, lo que aporta una visión multidimensional al análisis de datos. Con experiencia en gestión de proyectos, automatización digital, business intelligence, análisis financiero y comunicación, cada integrante ha desarrollado un interés creciente en la

ciencia de datos, buscando aplicar herramientas analíticas para optimizar procesos y generar impacto en sus respectivos sectores.

Nuestra motivación para elegir esta propuesta radica en el potencial transformador de la inteligencia artificial y la analítica avanzada en el sector fintech. Nos interesa explorar cómo los datos pueden mejorar la personalización de servicios financieros y optimizar la gestión del negocio. Este proyecto representa una oportunidad para consolidar conocimientos en machine learning y análisis predictivo, alineándose con las tendencias del mercado y preparándonos para contribuir al desarrollo de soluciones basadas en datos en un entorno altamente competitivo y en expansión.

EasyMoney enfrenta desafíos en conversión de clientes, asignación eficiente de recursos y toma de decisiones basada en datos. La implementación de Data Science y Machine Learning puede aportar ventajas estratégicas clave:

- Aumentar la conversión de clientes mediante estrategias de marketing basadas en analítica avanzada.
- Optimizar la asignación de recursos, enfocando esfuerzos en clientes con mayor potencial de rentabilidad.
- Reducir la incertidumbre en la toma de decisiones con modelos predictivos que anticipen el comportamiento de los clientes.
- Mejorar la personalización de productos y campañas, segmentando eficazmente la cartera de clientes.

## Objetivos

- Objetivo general:
  - Optimizar la estrategia comercial de **EasyMoney** mediante técnicas de **Data Science y Machine Learning**, con el fin de **aumentar la rentabilidad de su cartera de clientes** a través de una asignación más eficiente de los recursos y una mejor personalización de las ofertas comerciales
- Objetivos específicos:
  - **Realizar un análisis exploratorio de los datos (EDA)** para comprender el comportamiento de los clientes.
  - Implementar modelos de segmentación para diferenciar perfiles de clientes.
  - **Desarrollar modelos de machine learning** para predecir la propensión de los clientes a adquirir productos financieros.

- Optimizar las estrategias de conversión de clientes, diseñando **umbrales de probabilidad y escenarios comerciales** basados en análisis de ROI y valor esperado de cada cliente.
- Personalizar campañas de marketing basadas en los segmentos de clientes identificados.
- Definir métricas e indicadores clave para el seguimiento de las estrategias comerciales.
- Garantizar una implementación eficiente del modelo en un entorno de producción escalable, permitiendo su integración con sistemas existentes mediante APIs y optimizando su rendimiento a través de infraestructura en la nube.

## Metodología empleada

El marco metodológico del proyecto sigue un enfoque teórico-práctico, aplicando el método científico para validar hipótesis y contrastar resultados. Se utilizaron diversas técnicas analíticas para abordar la problemática de EasyMoney, incluyendo:

1. Recopilación y preparación de datos:
  - Carga de datos: Importación de los datasets en un entorno de análisis.
    - commercial\_activity\_df (actividades comerciales de los clientes).
    - products\_df (productos financieros contratados).
    - sociodemographic\_df (datos demográficos de los clientes).
  - Eliminación de columnas irrelevantes: Se eliminaron atributos sin valor analítico, como “Unnamed: 0”.
  - Unificación de información: Se realiza un merge de los distintos datasets para consolidarlos en un único dataset principal (df\_easy).
  - Conversión de fechas: Transformación de las columnas pk\_partition y entry\_date a formato datetime para facilitar cálculos temporales.
  - Análisis de valores nulos.
  - Estrategias de imputación:
    - Mediana para variables numéricas.
    - Moda para variables categóricas.
  - Eliminación de registros duplicados para evitar sesgos en el análisis.
  - Análisis exploratorio de datos (EDA):
    - Distribución de ingresos, edad y productos contratados.
    - Análisis de correlaciones entre variables mediante mapas de calor.
    - Identificación de patrones en la actividad comercial de los clientes.

- Feature Engineering: Se crearon nuevas variables con el objetivo de enriquecer el dataset original y mejorar la capacidad predictiva de los modelos de machine learning:
  - Cálculo de antigüedad del cliente (`days_since_entry`)
  - Variación salarial (`variacion_salarial_abs` y `salary_increase_pct`)
  - Creación de indicadores de tenencia de productos:  
Para mejorar el análisis, los productos financieros se agruparon en **tres categorías clave**:
    - Cuentas y servicios básicos (`tiene_prod_cuenta`)
    - Productos de ahorro/inversión (`tiene_prod_ahorro_inv`)
    - Productos de financiación (`tiene_prod_financiacion`)
  - Cálculo del número total de productos contratados por cliente
  - Creación de indicadores de recompra de productos (`ha_recomprado_*`)
- Separación del dataset en tres, para cada categoría de producto

## 2. Segmentación de clientes

Para optimizar las estrategias comerciales de EasyMoney, se implementó un modelo de segmentación de clientes basado en clustering no supervisado. Esto permitió identificar grupos homogéneos con características y comportamientos similares, facilitando la personalización de campañas de marketing y estrategias de fidelización.

- Selección y transformación de variables
  - Variables sociodemográficas: edad, género, tipo de cliente (universitario, particular, empresa).
  - Variables de comportamiento: antigüedad, actividad reciente, historial de compras.
  - Variables de productos financieros: número y tipo de productos contratados.
  - Variables de fidelización: recompra de productos, variación salarial y uso de servicios.
- Preprocesamiento aplicado:
  - Estandarización de variables con StandardScaler para evitar sesgos debido a diferencias en la escala de los datos.
  - Reducción de dimensionalidad para optimizar el rendimiento del modelo.
- Aplicación de K-Means Clustering:
  - Determinación del número óptimo de clusters con el método del codo y el coeficiente de silueta.
  - Entrenamiento del modelo para segmentar a los clientes en grupos homogéneos.
- Análisis de los segmentos: Caracterización según antigüedad, actividad y comportamiento de compra.

- Definición grupos de clientes:
  - Clientes activos recientes.
  - Clientes inactivos con baja contratación.
  - Clientes inactivos y recientes.
  - Clientes inactivos, antiguos y fidelizados.
  - Clientes activos, antiguos y fidelizados.
  - Clientes premium y fidelizados.
- Estrategias recomendadas:
  - Reactivación dirigida para clientes antiguos fidelizados que no realizan compras mediante campañas de recompra.
  - Fidelización temprana en los clientes recientes para evitar que pasen a inactivos.
  - Segmentación diferenciada en clientes particulares y universitarios para optimizar las conversiones.

### 3. Modelado predictivo

El modelado predictivo fue una de las fases clave del proyecto, con el objetivo de estimar la probabilidad de que un cliente adquiera productos financieros en función de sus características y comportamiento histórico. Para lograrlo, se utilizaron varios algoritmos de machine learning y se comparó su rendimiento.

Preparación de los datos para el modelado:

- División del dataset: Se separaron los datos en conjunto de entrenamiento (70%) y conjunto de prueba (30%) para evaluar el rendimiento del modelo.
- Preprocesamiento:
  - Normalización con **StandardScaler** para asegurar que todas las variables numéricas tengan la misma escala.
  - Codificación **OneHotEncoder** para convertir variables categóricas en variables numéricas y **balanceo de clases** para mitigar el sesgo en datos desbalanceados.
- Selección de características: Se utilizó **Feature Importance (RandomForest)** para identificar las variables más relevantes.
- Ingeniería de características aplicada en cada dataset:  
 Se entrenaron modelos predictivos para tres categorías de productos financieros:
  - Cuentas y servicios básicos (*tiene\_prod\_cuenta*)
  - Productos de ahorro e inversión (*tiene\_prod\_ahorro\_inv*)
  - Productos de financiación (*tiene\_prod\_financiacion*)
- Competición de modelos:
  - **CatBoostClassifier**.

- **XGBClassifier.**
- **GradientBoostingClassifier.**
- **LogisticRegression.**
- **SGDClassifier.**
- **Perceptron.**
- Selección del modelo ganador: Tras evaluar todas las métricas, CatBoostClassifier fue seleccionado como el modelo más eficiente debido a:
  - Su alta precisión y capacidad de generalización en comparación con otros modelos.
  - Su capacidad de manejar relaciones no lineales y detectar patrones complejos en los datos.
  - **Mayor estabilidad** frente a outliers y datos faltantes
- Evaluación del modelo:
  - Accuracy: Medida general de precisión, indicando el porcentaje de predicciones correctas.
  - Matriz de confusión: Analiza los aciertos y errores en la clasificación, diferenciando Verdaderos Positivos, Falsos Positivos, Falsos Negativos y Verdaderos Negativos.
  - AUC-ROC: Indicador clave del rendimiento del modelo en términos de diferenciación entre clases positivas y negativas.
- Para entender las decisiones del modelo y mejorar su interpretabilidad, se utilizaron **valores SHAP**.
  - Se generaron gráficos de importancia de características.
  - Se analizaron las variables con mayor impacto en la probabilidad de compra.
  - Se identificaron patrones clave en el comportamiento de los clientes.
- Se integró el modelo en el dashboard de Power BI, lo que permitió visualizar la probabilidad de compra de cada cliente y mejorar las estrategias comerciales.

#### 4. Resultados y predicciones

Se realizó un análisis cuyo objetivo fue optimizar la selección de clientes a impactar con productos financieros, maximizando la ganancia esperada según distintos niveles de riesgo. Se nos indica que existe presupuesto para impactar a 10.000 clientes y, por tanto, el objetivo será encontrar esos 10.000 con mayor propensión de compra, y con mayor retorno de la inversión.

Se definieron tres escenarios (Conservador, Moderado y Agresivo) para ajustar el umbral mínimo de probabilidad de compra y encontrar la mejor distribución de clientes.

Esto permite enfocar los esfuerzos en aquellos con mayor potencial de conversión, equilibrando rentabilidad y riesgo. Logrando una estrategia más eficiente y alineada con los objetivos comerciales.

Escenario	Cuenta	Ahorro	Financiación
<b>Escenario Conservador</b>	96,48 %	95,10 %	95,01 %
<b>Escenario Moderado</b>	96,82 %	90,02 %	90,00 %
<b>Escenario Agresivo</b>	97,53 %	85,00 %	85,00 %



## 5. Visualización y Seguimiento con Power BI

Se creó un dashboard en Power BI con:

- Análisis general: Resumen de características principales de clientes y productos.
- Segmentación de clientes: Visualización de clusters.
- Análisis de modelos predictivos: Comparación de métricas.
- Implementación: Impacto de estrategias aplicadas.
- Seguimiento: definición de KPIs para medir el éxito de la implementación y propuestas de mejora para futuro.

Este enfoque metodológico permitió estructurar el análisis de datos de manera robusta, asegurando la calidad de los datos, la selección del mejor modelo predictivo y la implementación de estrategias comerciales basadas en información objetiva.

## 6. Implementación

Para una implementación eficiente en producción, se recomienda el despliegue del modelo en un entorno cloud escalable como AWS SageMaker o Google Vertex AI, garantizando actualizaciones automáticas y procesamiento en tiempo real. Además, se sugiere la exposición del modelo a través de una API en FastAPI o Flask, permitiendo una integración fluida con los sistemas existentes y dashboards interactivos. Es crucial implementar un monitoreo continuo del desempeño del modelo para detectar y mitigar el *data drift*, causado por cambios en la distribución de los datos a lo largo del tiempo. En caso de degradación en las métricas de precisión, se debe realizar un reentrenamiento del modelo con datos actualizados para garantizar su efectividad y capacidad de generalización.

## Resultados y Conclusiones

### 1. Resultados obtenidos

- Se lograron identificar 6 segmentos de clientes con características distintas, lo que permitió personalizar estrategias comerciales de manera más efectiva y orientar las campañas de marketing a públicos más específicos.
- El modelo de CatBoostClassifier alcanzó valores de AUC-ROC por encima del 91%, lo que indica una alta capacidad de predicción en la propensión de los clientes a adquirir productos financieros.
- La segmentación permitió diferenciar a los clientes en función del comportamiento de consumo y frecuencia de actividad, permitiendo estrategias diferenciadas para cada grupo.
- Se validó la importancia de la variable antigüedad del cliente como un factor clave en la fidelización y rentabilidad de los clientes.

### 2. Impacto en la toma de decisiones

- La implementación del dashboard en Power BI permitió a la empresa visualizar en tiempo real los datos y métricas clave, facilitando el análisis y la toma de decisiones basadas en datos.
- Se identificaron clientes con **alta propensión de compra**, permitiendo al equipo de marketing **priorizar esfuerzos en clientes con mayor potencial de conversión**.
- Se optimizó el **envío de correos electrónicos segmentados**, reduciendo costos operativos y aumentando la **eficiencia de las campañas publicitarias**.
- Se identificó la necesidad de ofrecer productos financieros específicos a segmentos particulares, permitiendo diseñar nuevas estrategias de fidelización y retención de clientes.
- La empresa ahora contaría con un sistema automatizado de análisis de datos, reduciendo el tiempo de procesamiento de informes y mejorando la eficiencia operativa.

### 3. Discusión de los resultados

- Los hallazgos confirmaron que la segmentación adecuada y la personalización de campañas comerciales pueden mejorar significativamente la rentabilidad.
- La precisión del modelo de machine learning indica que es posible prever el comportamiento de los clientes con un margen de error bajo, lo que permite mejorar las estrategias comerciales de la empresa.
- La visualización de datos en Power BI demostró ser una herramienta clave para el análisis interactivo y la rápida toma de decisiones.

#### 4. Conclusión

Este proyecto permitió establecer un **marco sólido de análisis de datos y toma de decisiones basada en información cuantificable**, logrando mejoras significativas en la **rentabilidad y efectividad comercial** de EasyMoney.

Gracias a la combinación de **modelos de machine learning, segmentación avanzada y visualización de datos interactiva**, se sentaron las bases para una evolución continua en la aplicación de **Data Science dentro de la empresa**, permitiendo una gestión comercial **más eficiente y orientada a resultados**.

- Optimización comercial basada en datos
- Mayor precisión en la predicción de clientes con alta propensión de compra
- Mayor retorno de inversión en campañas comerciales
- Toma de decisiones más rápida e informada

#### Impacto final:

EasyMoney ahora dispone de una estrategia de negocio **basada en inteligencia artificial y analítica avanzada**, garantizando un enfoque **más rentable, eficiente y escalable** en un mercado altamente competitivo.

## Referencias / Bibliografía

Las referencias utilizadas en esta memoria se presentan en formato APA:

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Artículos y documentación de Scikit-learn y Power BI.
- Bases de Datos internos de EasyMoney.