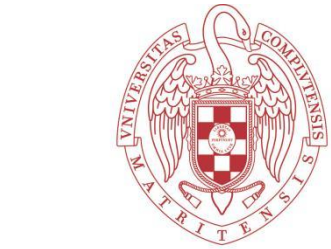


UNIVERSIDAD
COMPLUTENSE
MADRID

MASTER EN BIGDATA & DATA SCIENCE
TRABAJO FINAL DE MASTER

ANÁLISIS DEL ABANDONO DE CLIENTES EN UN E-COMMERCE MEDIANTE TÉCNICAS DE MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO

César Tulio González Franco

REPOSITORIO DEL PROYECTO:
https://github.com/haxstat/tfm-clv-ecommerce

01/15/2026

UNIVERSIDAD COMPLUTENSE MADRID

TABLA DE CONTENIDO

- **f)** CLV en riesgo por segmento
- **g)** Validación (Mann-Whitney U)

9. Propuestas de Modelo de producción

- **a)** Arquitectura Conceptual
- **b)** Flujo de datos y reentrenamiento del modelo
- **c)** Uso de CLV y P(Alive) en decisiones de negocio

10. Conclusiones

- **a)** Resultados principales de investigación
- **b)** Implicaciones resultantes para el negocio
- **c)** Limitaciones del estudio
- **d)** Líneas futuras de trabajo

11. Referencias

12. Anexos

RESUMEN

Este estudio analiza el ciclo de vida de un cliente en un entorno de e-commerce no contractual usando minería de datos y modelado probabilístico. Se utiliza el dataset público "Online Retail II UCI" que contiene transacciones entre 2009 y 201 de más de un millón de usuarios. Con estos datos se construyeron métricas RF(T)M y se segmentaron los clientes mediante K-Means en cuatro perfiles: **Premium, Potential, Occasional y Lost.** Se usaron omdelos BG/NBD y Gamma-Gamma para calcular el número de transacciones a futuro y el valor monetario por transacción obteniendo así un CLV de 6,48 millones de libras esterlinas a 12 meses. El análisis de riesgo de abandono mediante la probabilidad P(Alive) identificó que un 5,8% de los clientes está en riesgo, con diferencias estadísticamente significativas respecto a los clientes activos ($p < 0,001$, prueba Mann-Whitney U). Los resultados muestran una concentración fuerte de valor en el segmento Premium (23,7% de clientes, 74,1% del CLV total), lo que nos revela la importancia de estrategias de retención diferentes por segmentación. Al finalizar se propone un producto mínimo viable o una arquitectura conceptual para que sea implementada en un entorno de negocio.

1. INTRODUCCIÓN

1.A - OBJETIVO GENERAL

Estimar el valor económico futuro de cada cliente y medir el riesgo de abandono, usando datos transaccionales de un e-commerce, aplicando técnicas de minería de datos y modelado probabilístico y así tener un sustento para plantear diferentes estrategias de negocio a futuro.

1.B - OBJETIVOS ESPECÍFICOS

1. Analizar un Dataset con transacciones de e-commer real y entender el comportamiento de sus clientes.
2. Limpieza y transformación del Dataset para construcción de métricas RF(T)M útiles para el análisis.
3. Segmentación de los clientes con clustering no supervisado con su descripción correspondiente.
4. Usar modelos BG/NBD y Gamma-Gamma (modelos probabilísticos del ciclo del vida del cliente) y corroborar resultados.
5. Calcular el CLV (Customer Lifetime Value) y compararlo con los segmentos creados en KMeans.
6. Comparación del riesgo individual de abandono con P(Alive) y ver su correspondencia con la segmentación RFM.
7. Propuesta de arquitectura conceptual funcional para implementarla en un modelo de negocio.
8. Conclusiones enfocadas en decisiones de negocio.

1.C - PREGUNTAS DE INVESTIGACIÓN

1. ¿Qué comportamientos pueden identificarse con los datos transaccionales y métricas RF(T)M?
2. ¿Qué segmentación surge de un analisis no supervisado y como diferenciarlos en su valor y actividad?
3. ¿BG/NBD y Gamma-Gamma predicen bien el comportamiento futuro de compra y valor monetario?
4. ¿Se cumple la exigencia de los modelos? (Independencia Frequency-Monetary)
5. ¿P(Alive) aporta algo diferente frente a la segmentación descriptiva, son estos dos métodos coherentes?
6. ¿Cómo se reparte el CLV y la inactividad entre los segmentos y qué significa esto en una estrategia de negocio?
7. ¿Cómo integrar este sistema a un entorno productivo real de negocio?

2. DESCRIPCIÓN DEL CONJUNTO DE DATOS.

2.A - ORIGEN Y CARACTERÍSTICAS DEL DATASET

El dataset utilizado en este estudio es "**Online Retail II UCI**" hace parte del repositorio de **UCI Machine Learning Repository** *(UCI Machine Learning Repository, 2019)*. Este contiene un registro completo de datos de transacciones realizadas por una empres de e-commerce con sede en United Kingdom dedicada principalmente a la venta de artículos de regalo, este registro de transacciones está entre diciembre 2009 y diciembre del 2011. Este conjunto de datos tiene más de un millón de transacciones que corresponden a compras individuales realizadas por clientes con registro único, entre estos hay minoristas y mayoristas.

Desde el punto de vista analítico el dataset tiene las siguientes características:

- **Tipo:** Multivariante, temporal y secuencial
- **Dominio:** Es un negocio y un comercio electrónico
- **Periodo temporal:** Tiene un periodo temporal de 2 años consecutivos
- **Naturaleza:** Datos transaccionales reales
- **Enfoque:** Análisis de comportamiento de clientes, segmentación predicción y modelización

Las características anteriormente mencionadas hacen que el dataset sea especialmente adecuado para realizar estudios de segmentación, análisis RF(T)M y CLV (Client Lifetime Value) y análisis de CHURN.

2.B - VARIABLES DISPONIBLES

Las variables del dataset son las siguientes:

Las variables disponibles en este Dataset son las siguientes:Variable	Descripción
Invoice	Identificador único de la factura. Si comienza por "C", indica una cancelación.

Las variables disponibles en este Dataset son las siguientes:Variable	Descripción
StockCode	Código único del producto.
Description	Nombre o descripción del producto.
Quantity	Cantidad de unidades compradas en cada transacción.
InvoiceDate	Fecha y hora de emisión de la factura.
UnitPrice	Precio unitario del producto (en libras esterlinas).
CustomerID	Identificador único del cliente.
Country	País de residencia del cliente.

Estas variables permiten reconstruir el historial completo de cada cliente, siendo este su frecuencia, volumen, valor monetario y evolución temporal. Teniendo estas variables podemos generar otras derivadas de estas como lo son: Valor total de transacción, Variables RF(T)M, probabilidades P(Alive) y CLV. Estas últimas son la base del análisis de este estudio.

2.C - CALIDAD, LIMPIEZA Y LIMITACIONES DE LOS DATOS

El dataset presenta una calidad alta al tratarse de datos de transacciones reales, como estos datos son creados en un entorno real de negocio tiene problemas de registros comunes como: registros nulos, cancelaciones de pedidos, devoluciones, ajustes contables, registros duplicados y precios o cantidades no estandarizadas. Así, hay indicios de que se debe hacer un proceso de depuración de los datos previo.

Proceso de limpieza: Eliminación de registros con identificador nulo, Exclusión de facturas canceladas, Exclusión de facturas con valor monetario negativo o nulo, registros duplicados, exclusión de registros de prueba o códigos no válidos.

Haciendo este proceso, podemos tener unos datos limpios y consistente de transacciones reales que nos permita hacer un análisis de comportamiento adecuado.

Limitaciones del dataset: Las limitaciones del dataset se deben a los datos no captados por el negocio, hay una ausencia de información sociodemográfica, falta información contractual (suscripciones o contratos), valores atípicos de clientes mayoristas, y la más importante es la definición indirecta del CHURN, como es un entorno no contractual el abandono del cliente no puede observarse directamente y debe estimarse con modelos probabilísticos *(Fader, Hardie, & Shang, 2010)*.

Estas limitaciones se tuvieron en cuenta en el estudio y en la interpretación de sus resultados.

3. ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis se realizó para comprender la estructura del conjunto de datos, identificar patrones relevantes, detectar incongruencias que justificaran las decisiones

metodológicas posteriores. Esto incluye la segmentación RF(T)M la modelización probabilística.

3.A - ESTADÍSTICAS DESCRIPTIVAS

El dataset contiene 1.067.374 registros y 8 variables. Estas se describen de las siguiente manera:

Customer ID: 22,77% de valores missing.
Description: 0,41% de valores missing.

Analizando más a fondo pudimos detectar que las observaciones con **Description faltante** presentan el **Customer ID** faltante, Tienen **Price** en "0.0" y su mayoría tienen el **Country** en "United Kingdom", también incluyen con frecuencia **Quantity** negativos o igual a 0.
Quantity con valores negativos está asociada a devoluciones o a cancelaciones. Los valores inferiores o iguales a "0.0" de **Price son** ajustes contables. Los I**nvoice con letra** "C" indican cancelaciones según la documentación del Dataset. Luego de este proceso de depuración tenemos 779.415 registros válidos de ventas reales asociadas a identificadores de clientes.

El valor monetario tiene una asimetría positiva indicando una concentración significativa en un subconjunto reducido de clientes, esto sucede comúnmente en entornos de comercio minoristas.

3.B - ANÁLISIS DEL COMPORTAMIENTO DE COMPRA

El Dataset fue analizado desde tres (3) perspectivas diferentes.

Distribución geográfica: Se identificaron 43 paises distintos, siendo "United Kingdom" el que concentra el **91.9%** cantidad de clientes lo que evidencia la concentración del negocio a ese país.

Evolución Temporal: La evolución del número de facturas mes a mes y volumen total de ventas muestra un patrón estacional. Se observan picos en noviembre y diciembre coincidiendo con el **periodo navideño** y luego hay caida en un efecto post-estacional. Durante el resto del año se observa un comportamiento moderado en las compras y fluctuaciones asociadas a campañas comerciales y a factores estacionales.

Dinamica transaccional: La mayoría de los clientes realizan pocas compras, hay un conjunto de clientes que realizan muchas compras y las cantidades compradas presentan una alta dispersión. Esto incluye la presencia de transacciones con valores extremos. Debido a esto, se nota una necesidad de segmentar los perfiles de los clientes usando técnicas de segmentación.

3.C - IDENTIFICACIÓN DE PATRONES RELEVANTES

Alta concentración del valor: Una proporción de clientes reducidos generan una parte significativa de los ingresos totales, Este comportamiento se analizó con la regla de Pareto. **Evidencia de estacionalidad** debido a comportamientos en los

meses de noviembre y diciembre que se deben considerar en el modelado del comportamiento a futuro y **presencia de clientes inactivos** debio a que sus última fecha de comprar nos dice que no realizan compras por periodos de tiempos prolongados, esto nos ayuda a reforzar la idea de usar una técnica probabilística para estimar el riesgo de abandono. Debido a que **hay mucha variabilidad en la frecuencia, recencia y valor monetario** de cada cliente **este grupo no es homogéneo** y debido a esto también se asocia que **hay distribuciones que no son normales** lo cual nos da una idea de usar modelos estadísticos robustos en etapas tardías el estudio.

4. INGENIERÍA DE CARACTERÍSTICAS

4.A - CONSTRUCCIÓN DE MÉTRICAS RF(T)M

Para modelar el comportamiento de compra se construyeron metrias RFM (Recency, Frequency y Monetary), Este enfoque ha sido ampliamente utilizado en la literatura para el análisis de valor del cliente y segmentación basada en comportamiento transaccional *(Fader, Hardie, & Lee, 2005b)*. Para esto, solo se usaron datos de clientes con identificador válido de **Customer ID**, este análisis requiere estos identificadores. Además de esto usamos el día siguiente a la última transacción, eso para evaluar el comportamiento de los clientes el día siguiente al cierre del período observado.

Fecha referencia=max(InvoiceDate)+1 día

```
from datetime import timedelta

fecha_referencia = df["InvoiceDate"].max() + timedelta(days=1)
print(fecha_referencia)
```

✓ 0.0s Python

```
2011-12-10 12:50:00
```

Tomando en cuenta esta fecha y su construcción se crearon las siguiente:

Recency: Número de días transcurridos desde la última compra.
Frequency: Número de facturas únicas asociadas al cliente.
Monetary: Gasto acumulado por clientes (TotalPrice)

Construcción de variables RFM

```
rfm = (
    df.groupby("Customer ID")
      .agg(
          recency=("InvoiceDate", lambda x: (fecha_referencia - x.max()).days),
          frequency=("Invoice", "nunique"),
          monetary=("TotalPrice", "sum")
      )
      .reset_index(drop=False)   # mantiene Customer ID como columna
)

rfm.head()
```

✓ 0.3s Open 'rfm' in Data Wrangler

	# Customer ID	# recency	# frequency	# monetary
0	12346.0	326	3	77352.96
1	12347.0	2	8	4921.53
2	12348.0	75	5	2019.4
3	12349.0	19	4	4428.69
4	12350.0	310	1	334.4

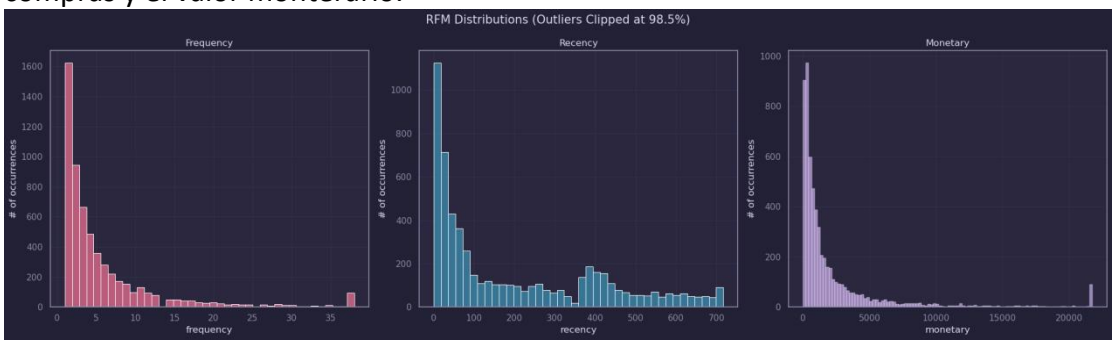
El resultado fue un dataset agregado a nivel cliente con 5.878 observaciones.

4.B - TRANSFORMACIONES DE LOS DATOS

Se evaluó la simetría de en el coeficiente de skweness:

Variable	Skewness
Recency	0.89
Frequency	12.64
Monetary	25.07

Estos valores reflejan una fuerte asimetría positiva, esto es característico en datos de transacciones en el cual un grupo de clientes pequeño concentra gran parte de las compras y el valor monterario.



Para poder estabilizar el impacto de los **valores extremos (Outliers)** se aplicaron ciertas transformaciones, esto estabiliza la varianza.

$$x\log = \log(1+x)$$

Esta transformación nos permitió reducir la asimetría significativamente, mejorar la distribución para los algoritmos que son muy sensibles a escalas, mantentar la interpretaciones económicas del estudio.

Variable (log)	Skewness
Recency_log	-0.49
Frequency_log	1.00
Monetary_log	0.27

Estandarización: Dado que vamos a usar KMeans más adelante, se ha decidido estandarizar mediante.

$$z = (x - \mu) / \sigma$$

Esto nos garantiza que ninguna de las variable sea muy dominante durante el calculo de distancia por su magnitud.

4.C - TRATAMIENTO DE OUTLIERS Y ASIMETRÍAS

Control de outliers: No se eliminaron del dataset debido a que representan un comportamiento real del negocio, esto podría representar un sesgo en la segmentación, es esperado una cola larga en la estructura de los datos debido a su naturaleza (datos de retail). Lo que sí se hizo fue aplicar mecanismos de Control para un buen uso y visualización de los datos, lo que se hizo fue un **Clipping al percentil**

98.5% en los gráficos de visualización para evitar distorción gráfica y gracias a las **transformación logarítmica** la influencia de valores extremos se ve reducida.

5. SEGMENTACIÓN DE CLIENTES

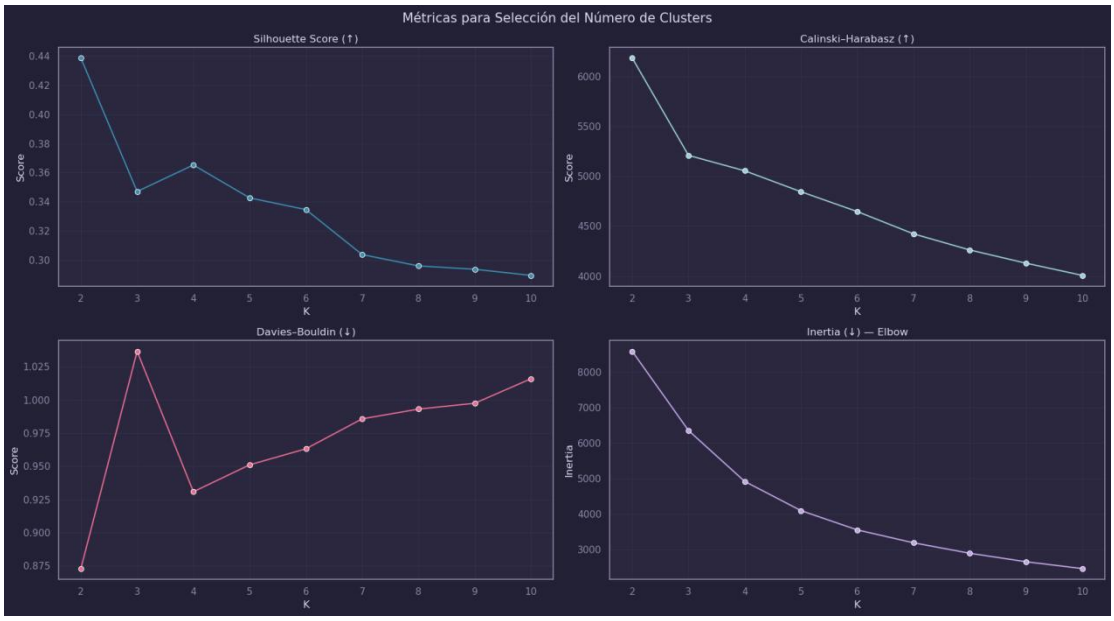
5.A - SELECCIÓN DEL NÚMERO ÓPTIMO DE CLUSTERS

Para poder determinar el número adecuado de segmentos, se han evaluado diferentes métricas internas de validación de las variables RF(T)M previamente transformadas (Transformación Logaritmica + Estandarización). Para esto se evaluaron valores **K entre 2 a 10.**

Silhouette Score: La cohesión y separación de los clusteres con este tes nos indica que la mejor definición es K = 2 (0.44)
Índice de Calinski - Harabasz: Este test mide la relación entre la dispersión interna y la dispersión entre los clusters, también se llega al **K = 2** como resultado óptimo.
Índice de Davies-Boulding: Este test evalua la similitud entre clusters, donde menos es mejor, se llega a K óptimo en **K = 2.**

Método del codo (Inercia): El método de la inercia nos da una segmentación un poco más amplia debido a que hay una reducción pronunciada de **K = 4** a partir de la cuál la mejora marginal no es mucha. Este punto de equilibrio es perfecto para nuestro estudio debido que podemos tener diferentes perfiles de clientes que sería óptimo para un entorno de negocio.

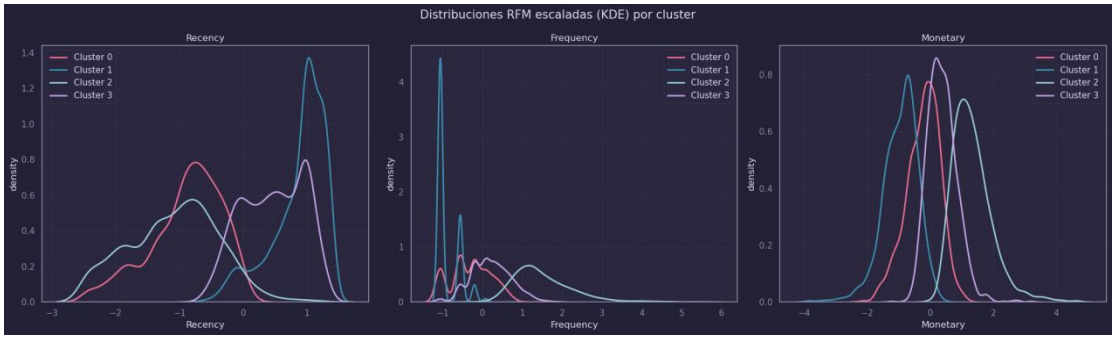
Tomando en cuenta todos estos análisis se ha optado por la métrica que nos indica la **inercia** debido a que **K = 4** nos permite una segmentación más grande entre grupos, así mismo nos permite identificar perfiles más diferenciados y ofrece un mayor rango de acción para el marketing y la retención de clientes.



5.B - CLUSTERING KMEANS SOBRE EL RFM

Para la creación del RFM se aplicó sobre las variables **previamente transformadas,** estas serían las que tienen Recency, Frequency, Monetary (Log-transformadas y estandarizadas) **.** Estas variables han sido previamente transformadas para reducir asimetrías, están estandarizadas para evitar sesgos por escala y con él modelo final **K=4**

Cluster	Recency (días)	Frequency (nº compras)	Monetary (£)
0	28.31	3.05	855.51
1	395.15	1.38	317.75
2	27.30	19.31	10,699.88
3	229.33	5.08	2,005.74



5.C - PERFILAMIENTOS DE SEGMENTOS

Las pefiles que hemos encontrado con las segmentación en clusters nos dan los siguiente datos, se le han puesto nombres a los clusters dependiendo de su comportamiento y así podemos identificar lo clientes con los qué contamos.

Segmento	Recency	Frequency	Monetary	Interpretación
2 - Premium	Baja	Alta	Muy alto	Clientes de alto valor
1 - Lost	Muy alta	Muy baja	Bajo	Clientes inactivos
0 - Occasional	Baja	Baja-media	Medio-bajo	Compradores esporádicos
3 - Potential	Media-alta	Media	Medio	Clientes con riesgo moderado

Premium: Son clientes activos, frecuentes y con un gasto elevado, es el núcleo del negocio. Son a los que se le deben prestar más atención.
Lost: Son los clientes que tienen una **Recency** Alta, no tienen buena frecuencia de compra y sus gastos son muy bajos.
Occasional: Son clientes recientes, con una baja frecuencia de compra, pero mantienen una actividad moderada.
Potencial: Estos clientes tienen un historial de compra razonable pero con menor actividad reciente, este segmento podría potenciarse con estrategias de retención.

6. MODELADO PROBABILÍSTICO BG/NBD - Gamma-Gamma

6.A - MARCO TEÓRICO: BG/NBD GAmma-GAmma

Con el fin de predecir si un cliente está activo o inactivo en un negocio tipo e-commerce, se utiliza un enfoque RFM. Los modelos BG/NBD permiten estimar el **Customer Lifetime Value** pero además se utilizan otras variables, así que agregamos

la variables **T** que representaría la antigüedad del cliente en la unidad temporal seleccionada, este es un calculo entre la primera compra del cliente y la última fecha del estudio. Asi que ahora tendríamos una matriz **RFTM** para cada cliente. Luego, el modelo **Gamma - Gamma** *(Fader, Hardie y Lee (2005))* complementa al BG/NBD porque estima la probabilidad de actividad del cliente y cuantas compras realizará.

Este modelo tiene fundamentalmente 2 asunciones:
Heterogeneidad de los clientes: El valor monetario que cada cliente gasta por transacción varía, y esta variabilidad entre los clientes se modela haciendo que cad cliente tenga un nivel de gasto típico, y la población de clientes presenta una distribución de esos niveles.
Variabilidad dentro de cada cliente: Las transacciones de cada cliente también varían a su promedio personal. Esta variabilidad se modela mediante la distribución Gamma-Gamma y cada individuo tendría su propia escala.

Estos parametros de los que hablamos son los siguientes:
p: Controla la variabilidad del gasto dentro de cada cliente. Un valor alto de p indica que las transacciones de un mismo cliente son relativamente consistentes en importe.
q: Controla la forma de la distribución de promedios entre clientes.
v: Controla la escala de la distribución de promedios entre clientes. Junto con q, determina el rango y dispersión del gasto promedio en la población.

La combinación de estos 2 models permite estimar el **CLV** de forma integral, mientas el BG/NBD da el número esperado de transacciones futuras y la probabilidad de que el cliente siga activo, Gamma-Gamma da por transacción el valor esperado.

6.B - JUSTIFICACIÓN DE LA DIVISIÓN CALIBRACIÓN-HOLDOUT

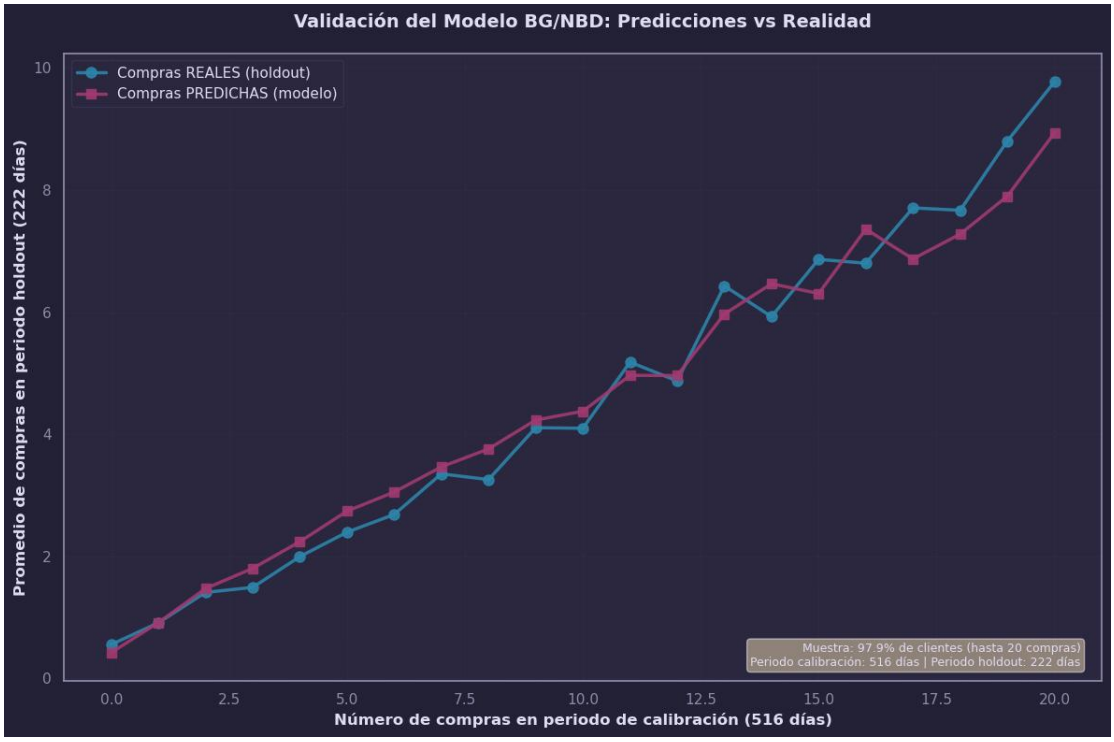
Para poder evaluar la capacidad predictiva de los estos modelos, tenemos que dividir el periodo de observaciones en 2, primero un **periodo de calibración** que corresponde a 516 días del dataset, que es el 70% de todo el periodo, con esto se ajustan los parámetros de los modelos. Luego está el **Periodo de Holdout** que serían los 222 días restantes, en total el 30% del dataset. Con estos podremos hacer las comparaciones de predicciones del modelo.

Hacemos notar que el 70/30 es una práctica estandarizada para modelos probabilísticos, esto haciendo que el periodo de calibración sea lo suficientemente largo para capturar patrones de compra de los clientes, y que el periodo de holdout sea extenso para que se puedan observar recompras y poder evaluar de forma significativa las predicciones del modelo. Es importante hacer notar que la función **calibration_and_holdout_data()** hace una construcción automática de las variables **RFTM** tanto para el periodo de calibración y de holdout.

6.C - AJUSTES DEL MODELO Y VALIDACIÓN PREDICTIVA

Para poder evaluar la capacidad predictiva de los estos modelos, se comparó las compras predichas en el periodo de holdout con las compras realmente observadas.

Para poder cuantificar la validad de las predicciones se empleó el **RMSE (Raíz del error cuadrático medio)** síendo el más óptimo para este estudio el **2.1298.**



La validación cualitativa se realizó mediante un gráfico de compras reales contra las compras predichas, y podemos observar que siguen una trayectoria similar, lo que refleja que el modelo captura adecuadamente las compras que ha hecho el cliente históricamente y su comportamiento a futuro. Las disimilaridades más pronunciadas se concentran en clientes con frecuencias de calibración extremadamente altas, dónde el muestreo es menor y por eso hay estimaciones menos estables.

La selección del **hiperparámetro L2 = 0.03** se hizo mediante una búsqueda de 20 valores entre **0.0** y **1.0** seleccionando aquel que minimiza el **RMSE** en el periodo de holdout, Se opto por **0.03** sobre **0.0** (sin regularización) por cuatro razones importantes: Prevención de sobreajuste, estabilidad de los parametros estimados, diferencia marginal en RMSE y consistencia con las recomendaciones de la literatura.

6.D - ASUNCIÓN DE INDEPENDENCIA FREQUENCY-MONETARY

El modelo Gamma-Gamma toma por dado que el valor monetario promedio de transacciones de un cliente es independiente de us frecuencia de compra. Esto es necesario ya que el modelo estima que el valor monetario no tiene correlación con frecuencia de compra. BG/NBD modela cuando y cuántas veces compra el cliente, mientras Gamma-Gamma modela cuánto gasta por transacción. Siendo cada una de su naturaleza sería arriesgado que estuvieran relacionadas porque habría un sesgo.
Para poder verificar esta asunción se puede calcular el **coeficiente de correlación de Pearson** entre la frecuencia de compra y el valor monetario promedio de cada transacción, considerando cliente que tengan más de una compra (clientes recurrentes). Sí este número es inferior a **0.3** se puede decir que hay una asociación

débil. La correlación en nuestros datos es **0.08**, así que podemos decir que el supuesto de independencia se cumple. (El detalle completo del procedimiento se presenta en el **Anexo E**.)

6.E - AJUSTE DEL MODELO GAMMA-GAMMA

El modelo se ajustó sobre todos los clientes que tiene más de una compra en el periodo de calibración sin utilizar un coeficiente de penalización debido a que la muestra de clientes recurrentes es suficientemente grande para estimar los parámetros sin necesidad de regularización adicional.
Los parámetros estimados por el modelo son:

p	2.322839969518011
q	3.4675983819029064
v	434.6884070100496

Con estos parametros combinado con la verificación de **independencia Frequency-Monetary** nos permite usar este modelo con confianza para estimar el valor monetario por transacciones de cliente.

7. ESTIMACIÓN CLV (CUSTOMER LIFETIME VALUE)

7.A - CÁLCULO DE CLV A 12 MESES

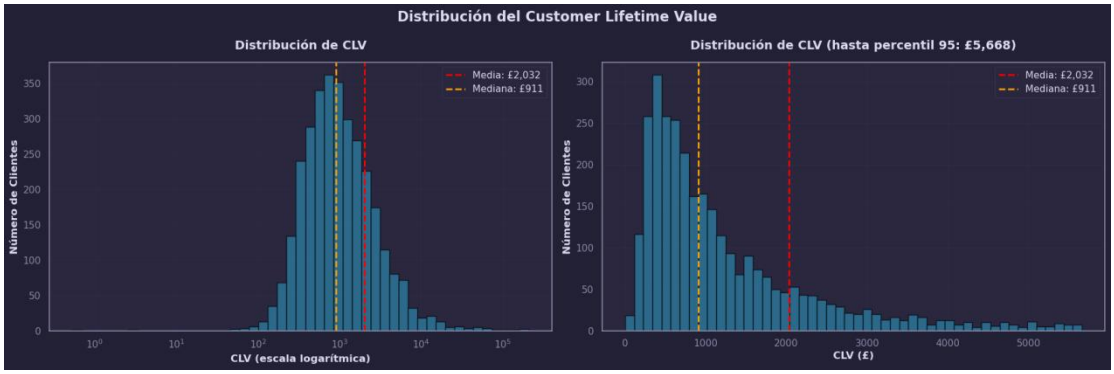
El CLV es el valor económico total que un cliente representa para una empresa en un determinado horizonte temporal. Los clientes se consideran activos ya que generan ingresos al negocio (Gupta & Lehmann, 2003). En este estudio se calcula ajustando 2 modelos probabilísticos ajustados, El BG/NBD y el Gamma-Gamma .

CLV total: £6,485,442.00
CLV promedio: £1,344.97

7.B - DISTRIBUCIÓN DE CLV POR SEGMENTOS

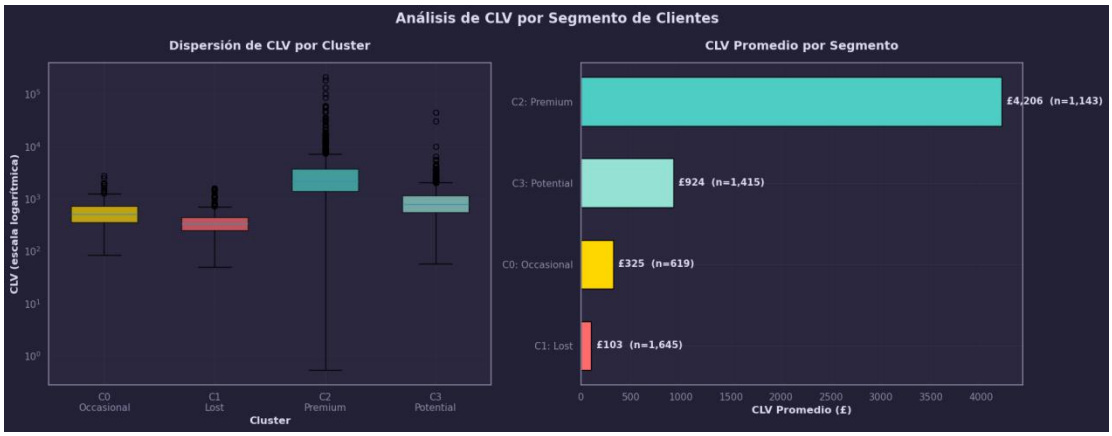
Cluster	Nº Clientes	CLV Promedio ($)	CLV Total ($)	Frecuencia Promedio	Valor Promedio ($)
0	619	325,07	201.216,04	0,94	129,21
1	1.645	103,14	169.658,65	0,31	51,09
2	1.143	4.205,99	4.807.445,14	10,27	523,15
3	1.415	923,76	1.307.122,17	2,87	400,90

El análisis de la distribución revela características consistentes con la mayoría de negocios de e-commerce: **una fuerte asimetría positiva.** La gran mayoría de los clientes tiene un CLV bajo, mientras que un grupo pequeño tienen valores superiores concentrados en la media.

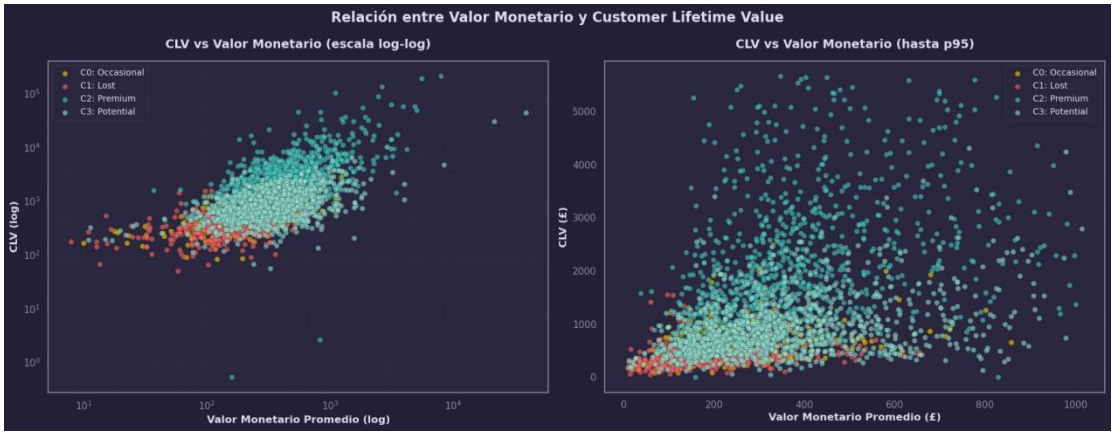


Esta asimetría refleja una diferencia entre la media y la mediana del CLV. La media es considerablemente superior por influencia de los clientes de valor superior al extremo de la derecha. La mediana vendría a ser la medida más representativa del cliente típicos. Al cruzar esta información con la segmentación anterior hecha en Kmeans podemos observar lo siguiente:

Premium: Concentra los valores más altos de CLV, tanto en media como en mediana. Estos clientes combinan alta frecuencia, baja recencia y alto valor monetario por transacción, lo que se traduce en las estimaciones de valor futuro más elevadas.
Potential: Presenta un CLV intermedio-alto. Se trata de clientes con frecuencia y valor monetario importante que representa un potencial significativo de generación de valor si se mantienen activos.
Occasional: Muestra un CLV intermedio-bajo. Son clientes con compras esporádicas y valores monetarios moderados, cuyo valor futuro depende en gran medida de si retoman la actividad de compra.
Lost: Presenta el CLV más bajo del conjunto. Su alta recencia y baja frecuencia histórica se traducen en predicciones de compras futuras muy reducidas, lo que limita su valor estimado incluso si el importe por transacción fue aceptable.



Mediante gráficos de Boxplots en la escala logarítmica podemos ver la dispersión y los valores atípicos. Se nota una variabilidad considerable dentro de cada segmento especialmente dentro del **premium, esto refuerza la idea de que cada cliente debe tener su propio CLV, no todo el segmento.**

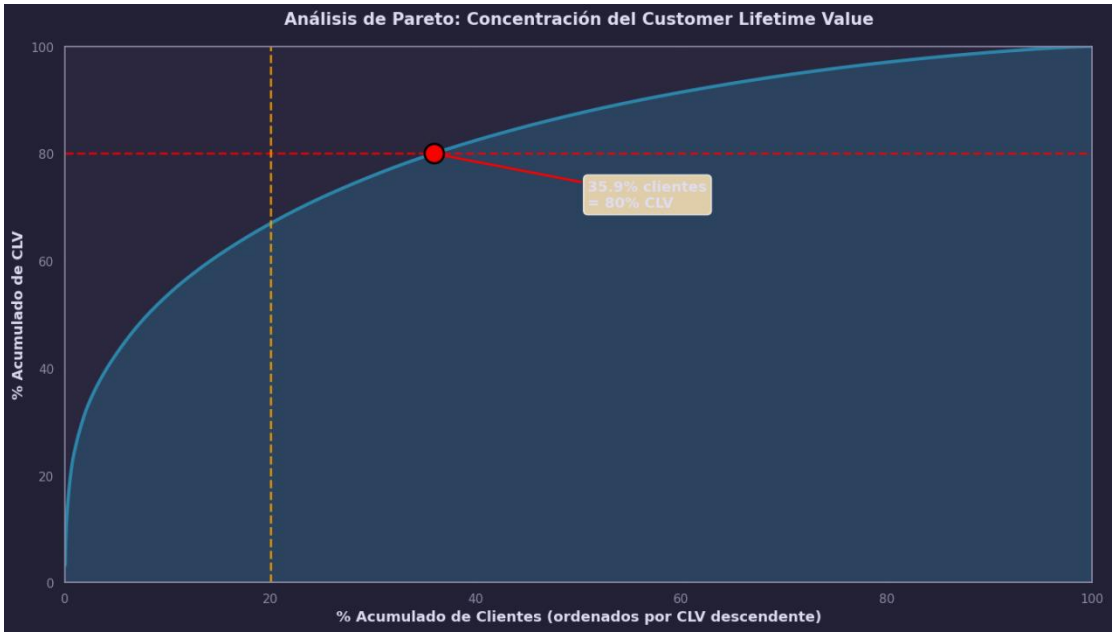


7.C - ANÁLISIS DE PARETO

Este análisis nos permite ver la concentración del valor económico en la base de los clientes. Esto quiere decir que hay una proporción reducida de clientes que tiene la mayot parte del valor económico total. Esto tiene muchas implicaciones en las estrategias de negocio ya que podría determinar que tipo de esfuerzos deben hacerse en y en qué grupo de clientes.

Este Dataset muestra que el **36%** de lo clientes concentra el **80%** del **CLV.** Esto significa que las acciones de negicio que se deben tomar después del estudio no solo deben limitarse a un solo "cluster" o segmento de clientes sino extenderse también a clientes de otrso segmentos. En otros terminos, se deben **priorizar los recursos,** la retención debe ser proporcional al valor en riesgo. Así que los segmentos **Premium y Potencial** podrian tener programas de fidelización y comuncación personalizada.

Perder un cliente en el **CLV superior** es lo mismo que perder varios clientes en el segmento **Ocassional o Lost** así que esto fundamenta la idea de un sistema de alerta temprana basado en otras métricas y sobretodo individualizar esto con **P(Alive).**



10. MODELIZACIÓN PROBABILÍSTICA DEL RIESGO DE ABANDONO Y VALIDACIÓN DE SEGMENTACIÓN

8.A - CÁLCULO DE P(ALIVE) PARA CADA CLIENTE CON BG/NBD

El modelo BG/NBD permite estimar para cada cliente la probabilidad de que permanezca activo en el momento actual: **P(Alive)**. Esta probabilidad es calculada por 3 variables, su frecuencia, su recencia y su antiguedad.

La funciona **conditional_probability_alive()** aplica una formulación bayesiana del proceso de compra y abandono. Con esto dado el comportamiento observado nos da una posibilidad posterior de que no haya abandonado. Un cliente con frecuencia alta y recencia baja tendrá un P(Alive) alto, mientras que uno con frecuencia bajay recencia alta recibirá un P(Alive) bajo. Esto se calculó para todos los 4.822 clientes del periodo de calibración y se obtuvo los siguientes descriptivos.

Estadístico	Valor
Media	0.9342
Mediana	0.9891
Desviación estándar	0.1217
Mínimo	0.0001
Máximo	1.0000

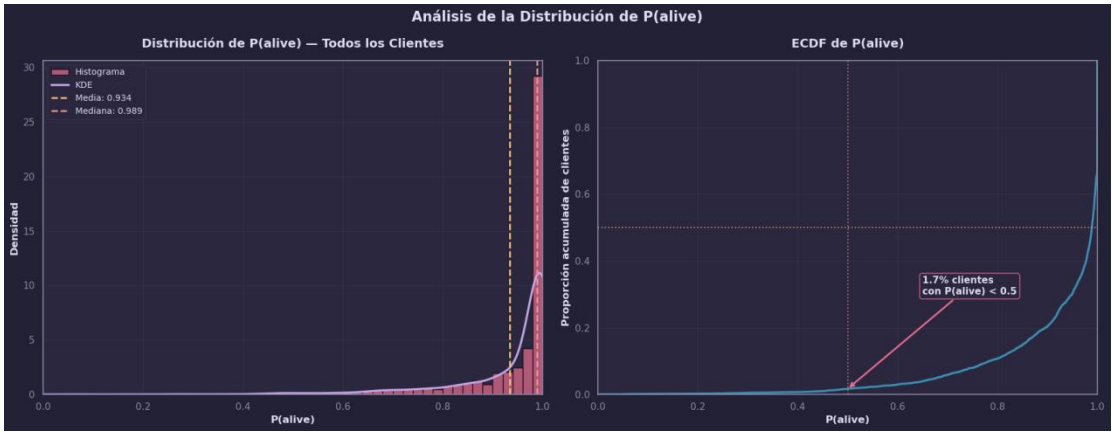
La distribución por percentiles nos revela una concentración en valores altos.

Percentil	P(alive)
P10	0.7843
P25	0.9228
P50 (mediana)	0.9891
P75	1.0000
P90	1.0000

8.B - ANÁLISIS DE LA DISTRIBUCIÓN P(ALIVE)

La distribución de P(alive) ayuda a entender la actividad de la base de clientes. Una distribución que se centra en valores entre 0 y 1 indica una separación entre clientes activos y abandonados, en este caso en riesgo.. En cambio, una distribución unimodal que se concentra en valores altos, como la que se ve en este conjunto de datos, muestra que la mayoría de los clientes sigue mostrando señales de actividad. La cola descendente hacia valores bajos representa a los clientes que están en riesgo de abandonar. Para visualizar esta distribución, se usan dos representaciones complementarias:

Histograma con estimación de densidad por kernel (KDE): Esta muestra la forma de la distribución y permite identificar de forma visual la gran concentración de clientes cerca de P(alive) = 1.0, así como la extensión de la cola izquierda. Las líneas punteadas de la media (0.934) y la mediana (0.989) evidencian la fuerte asimetría negativa. La media se ubica por debajo de la mediana, arrastrada por los clientes de la cola izquierda.



Función de distribución acumulada empírica (ECDF): Esta muestra la proporción acumulada de clientes por debajo de cada valor de P(alive). Este gráfico es especialmente útil para responder preguntas como "¿qué porcentaje de clientes tiene un P(alive) inferior a X?", lo cual es relevante para establecer un umbral de riesgo. Al no haber una bimodalidad clara, la selección del umbral no puede depender de encontrar un valle natural entre los dos modos. En su lugar, se necesita un criterio alternativo basado en consideraciones prácticas y de negocio, esto se explica adelante.

8.C - SELECCIÓN DEL UMBRAL DE RIESGO Y JUSTIFICACIÓN

Se seleccionó P(alive) < 0.7 como umbral operativo, siendo el valor más conservador (más bajo) que produce al menos un 5% de clientes clasificados como en riesgo. Este criterio se justifica por tres razones:

Relevancia práctica: Un umbral que produzca menos del 5% de riesgo no genera un volumen de clientes suficiente para justificar acciones de retención a escala.

Especificidad: Al elegir el umbral más bajo que cumple la condición del 5%, se maximiza la confianza en que los clientes clasificados como en riesgo realmente presentan un riesgo elevado de inactividad. Con P(alive) < 0.7, cada cliente clasificado tiene menos de un 70% de probabilidad de estar activo según el modelo.

Separación económica: La diferencia entre el CLV medio de activos (£1.396) y el de clientes en riesgo (£524) confirma que el umbral separa grupos con valor económico significativamente distinto.

Umbral	En Riesgo (%)	Activos (%)	CLV Medio en Riesgo (£)	CLV Medio Activo (£)
0,1	0,2%	99,8%	97	1.347
0,2	0,3%	99,7%	231	1.348
0,3	0,5%	99,5%	413	1.350
0,4	0,7%	99,3%	342	1.352
0,5	1,7%	98,3%	677	1.357
0,6	3,0%	97,0%	555	1.369
0,7	5,8%	94,2%	524	1.396
0,8	10,7%	89,3%	537	1.442
0,9	20,3%	79,7%	626	1.529

Este enfoque es superior a las definiciones ad hoc del abandono (como "sin compras en 6 meses") porque tiene en cuenta la frecuencia histórica del cliente, su antigüedad en el sistema y se fundamenta en un modelo estadístico validado con datos reales.

8.D - CLASIFICACIÓN: CLIENTES ACTIVOS/ CLIENTES EN RIESGO

Con el umbral seleccionado podemos clasificar los clientes en 2 estados.

Estado	Nº Clientes	Porcentaje
Activo (P(alive) ≥ 0,7)	4.540	94,2%
En riesgo (P(alive) < 0,7)	282	5,8%

Así podemos capturar 2 comportamientos de compras diferentes, los clientes activos presenta una recencia menor, una frecuencia mayor y un valor monetario superior al cliente en riesgo. Se evidencia en que el CLV de los clientes activos es superior al de los clientes en Riesgo.

Es importante recalcar que esta clasificación **no pretende sustituir la segmentación RFM ni redefine el concepto de "LOST". La etiqueta "En riesgo" es un estado dinámico de actividad estimado por un modelo probabilístico**, mientras que los segmentos RFM describen perfiles estructurales históricos.

8.E - CRUCE CON LAS SEGMENTACIÓN RFM: COHERENCIA ENTRE ENFOQUE DESCRIPTIVO Y PROBABILÍSTICO.

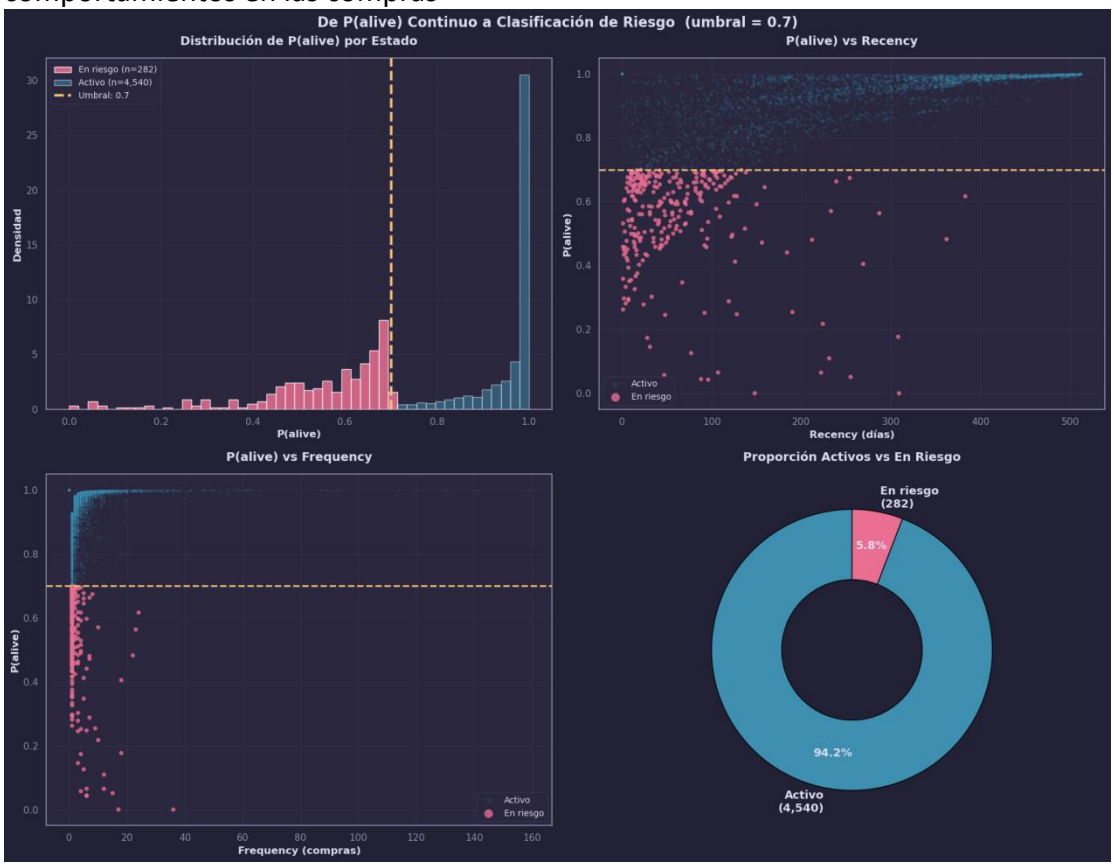
El cruce que hacemos entre un métedo descriptivo y otro probabilístico es para saber si ambos enfoques producen resultados coherentes. Este análisis muestra los siguientes datos.

Segmento	Clientes en riesgo	Total	Tasa de riesgo
Potential	106	1.415	7,5%
Lost	121	1.645	7,4%
Occasional	37	619	6,0%
Premium	18	1.143	1,6%

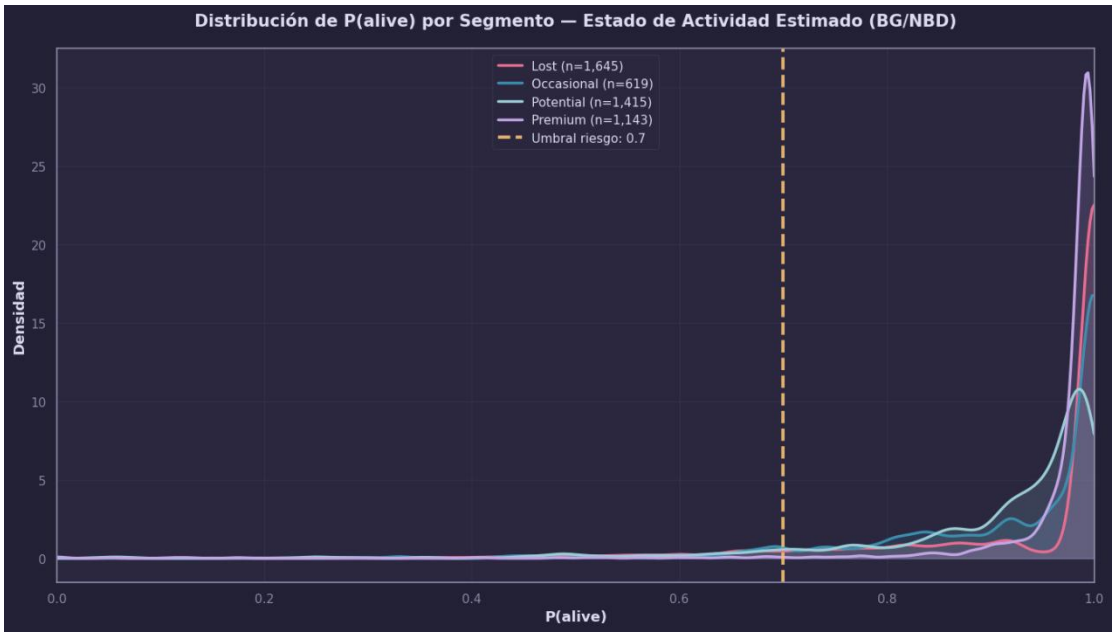
Premium tiene la tasa de riesgo más baja **1.6%**, hay **18 clientes que podrías estar en riesgo**. **Lost** presenta una tasa de 7.4%, este segmento está clasificado como inactivo así que coincide. **Potential muesta una tasa mucho más alta 7,5%** este segmento tiene unas ganancias potenciales mayores debido a que estos clientes tienen un valor individual más alto. **Occasional** tiene un nivel intermedio del **6,0%,** es un perfil de compras esporádicas.

Este análisis no redefine que es un segmento, lo que hace es responder a una pregunta diferente a la que hace el clustering, mientas Kmeans responed a ¿Qué tipo de cliente es? Tomando en cuenta el perfil histórico, con P(Alive) estamos preguntando ¿Está activo ahora? Basándonos en un modelo probabilístico y esto nos

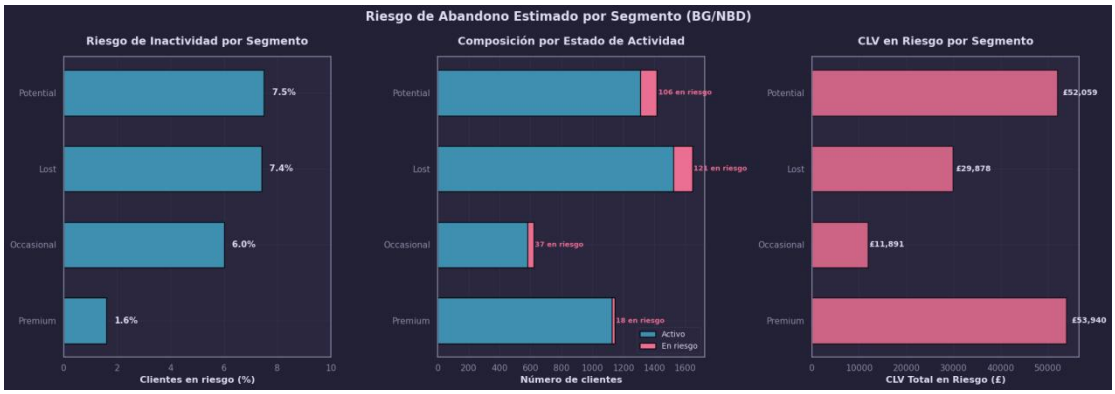
ayuda a identificar que hasta en un segmento excelente hay deterioros de comportamientos en las compras



La distribución del segmento **Premium** está muy concentrada cerca de 1.0, mientras que el segmento **Lost** presenta una dispersión mayor hacia valores bajos. **Potential y Occasional** muestran distribuciones intermedias con colas de riesgo que se pueden notar.



8.F - CLV EN RIESGO POR SEGMENTO



¿Cuánto valor económico está en riesgo? Esto lo podemos ver en los gráficos. Este análisis nos revela un hallazgo sumamente importante para un negocio. **18 clientes premium en riesgo concentran £53.940** del valor, mucho más que 121 clientes de Lost juntos £29.878. Así qe esto tiene implicaciones en la toma de decisiones de negocio donde se tiene que priorizar una atención individualizada e inmediata a cada uno de estos clientes. Esto también pasa con **106** clientes de **Potential** donde el poder económico es de **£52.059,** no es el mismo CLV pero si están con poder económico más elevado. Los demás sectores podrían tratarse con campañas de reactivación automatizadas.

8.G - VALIDACIÓN MANN WHITNEY U

Variable	U	p-valor	Resultado
P(alive): activos > en riesgo	1.280.280	$1{,}85 \times 10^{-182}$	Diferencia altamente significativa
CLV: activos > en riesgo	725.116	$6{,}66 \times 10^{-5}$	Diferencia altamente significativa
Frequency: activos > en riesgo	622.688	0,785	No significativa

Este test no confirma que P(Alive) > 0.7 nos produce grupos estadísticamente diferentes. Este test es para variables con distribución asimétrica. Los clientes activos presentan un CLV significativamente superior, lo que confirma su relevancia económica. Se nota la importancia adicional del modelo BG/NBD debido a que la frecuencia no discrimina entre grupos. La segmentación es estadísticamente robusta y funciona a nivel de empresa.

11. PROPUESTAS DE MODELO DE PRODUCCIÓN

Como pureba de concepto, se desarrolló un dashboard interactivo con la ayuda de Streamlit que permite explorar resutaldos del modelo, esto incluye segmentación, CLV y clasificación de riesgo. Puede consultarse en:
https://tfm-clv-ecommerce.streamlit.app

9.A - ARQUITECTURA CONCEPTUAL

La propuesta de producción modela el ciclo de vida de un cliente y se basa en una arquitectura de 3 capas:

Ingesta y almacenamiento: las transacciones de almacenan en una base de datos, esta capa aplica la limpieza y se usa para alimentar los modelos

Motor de modelización: Un proceso llamado batch construye las métricas del RFM, ejecuta los modelos BG/NBD y Gamma-Gamma para los cálculos ya creados en este proyecto para generar los resultados y los almacena en una tabla de salida

Visualización y acción: Los resultados se generan en un dashboard interactivo donde se puede analizar el CLV en riesgo por segmento, prioriza campañas y simular el impacto económico, el usuario no necesita conocimientos técnicos para ver los resultados del modelo.

9.B - FLUJO DE DATOS Y REENTRENAMIENTO DEL MODELO

El sistema operacionalmente tiene ciclos:
Actualización de datos: incorpora nueva transacciones del e-commerce
Scoring: recalculo del P(Alive) y CLV con los modelos
Re-entrenamiento periódico: Ajustas los modelos y compara con el modelo anterior
Validación y monitorización: Se hace un seguimiento de RMSE, distribución del P(Alive) y estabiliza los parámetros para evitar degradación del sistema.

9.C - USO DE CLV Y P(ALIVE) EN DECISIONES DE NEGOCIO

La mezcla del CLV y el P(Alive) Permite tomar decisiones basadas en el valor esperado en el futuro. Esto permite hacer una **priorización de retención de clientes**, tener **alertas tempranas** antes de que un cliente cruce el umbral de riesgo, una **segmentación dinámica** donde se puede combinar segmentos, una **presupuestación basada en valor**, lo cual estima el impacto económico del abandono, además se puede hacer una **evaluación de campañas** debido a que se miden los cambios de P(Alive) antes y después de cualquier acción comercial. El dashboard desarrollado es una prueba funcional de la arquitectura del concepto, esto demuestra viabilidad técnica y una aplicación directa en un entorno empresarial.

12. CONCLUSIONES

10.A - RESULTADOS PRINCIPALES

Este estudio analizó el ciclo de vida de un cliente en un entorno de e-commerce no contractual usando segmentación descriptiva y modelización probabilística con el objetivo de calcular el valor económico de un cliente así como su riesgo de abandono. La segmentación del RFM mediante KMeans identificó 4 perfiles diferente: Premium, Potencial, Occasional y Lost. Se descubrió una fuerte concentración de valor monetario en Premium, representando el 23,7% de toda la base, donde se concentra el 74,1% del CLV total de la empresa. Esta tiene un promedio significativamente mayor a los otros segmentos. El modelo BG/NBD ajustado con **L2 = 0,03** mostró una capacidad predictiva adecuada con un **RMSE = 2,13 compras en validación holdout.** Esto indica una tasa baja de abandono coherente con la distribución observada en P(Alive) cuya media fue de **0.93.** Paso siguiente el modelo Gamma-Gamma permitió estimar el gasto esperado por transacción. La integración de ambos modelos nos dio una proyección estiamada de £6,48 millones a 12 meses. El análisis de riesgo basado en P(Alive) ,según criterios decritos por este estudio, es de **0,70** así podiendo identificar un **5.8% de clientes en riesgo. Las pruebas estadísticas confirman significativas diferencias entre clientes activos y clientes en riesgo de actividad como en CLV.El cruce entre la segmentación descriptiva y la clasificación probabilística reveló** un hallazgo clave: El valor económico no depende solamente del número de clientes, sino de su valor individual. **Este resultado refuerza la necesidad tomar acciones de retención en función del valor esperado, no solo del volumen.**

10.B - IMPLICACIONES PARA EL NEGOCIO

El estudio demuestra que pequeños grupos de clientes pueden concentrar un alto valor en proporción relevante del valor en riesgo, lo que justifica intervenciones diferenciadas según el segmento (perfil económico). Esto justifica que P(Alive) exista y pueda usarse como una alerta temprana para detectar patrones en declive en relación comercial antes de que el abandono de un cliente sea irreversible. Estas observaciones nos da a entender que una estrategia uniforme de retención para todos los segmentos sería ineficiente, y no solo eso, podría ser fatal para el bienestar del negocio. Abordar el problema de esta manera generas diferentes estrategias para cada segmento donde se encuentre un cliente específico y poder tomar diferentes estratégias dependeiendo de su perfíl.

10.C - LIMITACIONES

Este dataset tiene limitaciones al no tener variables demográficas o comportamientos adicionales, esto no permite generalizar resultados. Además los modelos probabilísticos asumen estacionariedad temporal y ciertos supuestos estructurales que no pueden cumplirse en todos los contextos. Dentro de todos el más importante es la selección del umbral de riesgo, este sa basó en critarios pragmáticos, si validación directa frente a abandono observado, dado el carácter no contractual del negocio.

10.D - LÍNEAS FUTURAS DE TRABAJO

Las futuras versiones, o extensiones de este estudio tendrían la posibilidad de incorporar variables individuales o técnicas de deep learning para capturar dinámicas temporales más complejas. Dentro de las posibilidades están implementar una validación temporal con ventanas deslizantes, optimización del umbral de riesgo mediante funciones de costo en función del retorno de inversión. Así mismo podría pensarse en un pipeline mas robustos para contextos empresariales más completos como multicanal y multimercado.

13. BIBLIOGRAFÍA

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005a). “Counting your customers” the easy way: An alternative to the Pareto/NBD model. *Marketing Science, 24*(2), 275–284. https://doi.org/10.1287/mksc.1040.0098

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Fader, P. S., Hardie, B. G. S., & Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science, 29*(6), 1086–1108. https://doi.org/10.1287/mksc.1100.0582

Gupta, S., & Lehmann, D. R. (2003). Customers as assets. *Journal of Interactive Marketing, 17*(1), 9–24. https://doi.org/10.1002/dir.10045

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research, 9*(2), 139–155. https://doi.org/10.1177/1094670506293810

UCI Machine Learning Repository. (2019). Online retail II dataset. University of California, Irvine. http://archive.ics.uci.edu/ml/datasets/Online+Retail+II

14. ANEXOS

ANEXO A – Proceso de limpieza y construcción del dataset

```
df = df.query("TotalPrice > 0")
df = df[df["Customer ID"].notna()]
df = df[~df["Invoice"].astype("string").str.startswith("C", na=False)]
df = df[~df["StockCode"].astype("string").str.contains("TEST",
na=False)]
df = df.drop_duplicates()
assert (df["InvoiceDate"] > datetime(2009,1,1)).all()
assert (df["InvoiceDate"] < datetime(2012,1,1)).all()
```

```
<class 'pandas.core.frame.DataFrame'> Index: 779415 entries, 0 to
1067370 Data columns (total 9 columns): # Column Non-Null Count Dtype -
-- ------ -------------- ----- 0 Invoice 779415 non-null object 1
StockCode 779415 non-null object 2 Description 779415 non-null object 3
Quantity 779415 non-null int64 4 InvoiceDate 779415 non-null
datetime64[ns] 5 Price 779415 non-null float64 6 Customer ID 779415
non-null float64 7 Country 779415 non-null object 8 TotalPrice 779415
non-null float64 dtypes: datetime64[ns](1), float64(3), int64(1),
object(4) memory usage: 59.5+ MB
```

ANEXO B – Construcción de métricas RFM

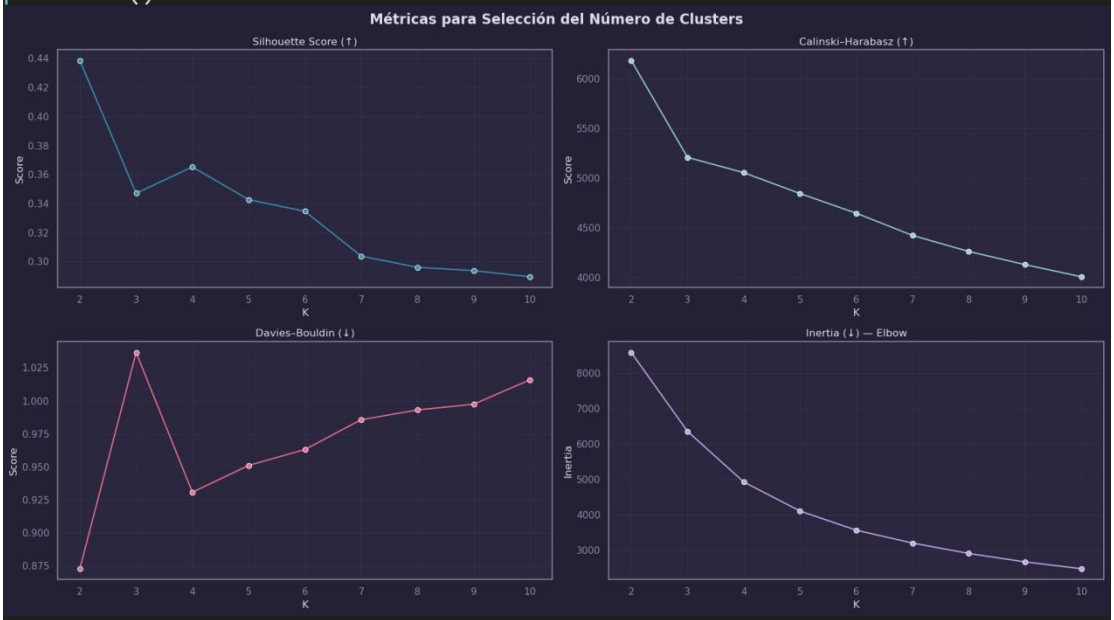
```
rfm = (
    df.groupby("Customer ID")
      .agg(
          recency=("InvoiceDate", lambda x: (fecha_referencia -
x.max()).days),
          frequency=("Invoice", "nunique"),
          monetary=("TotalPrice", "sum")
      )
      .reset_index(drop=False)   # mantiene Customer ID como columna
)
rfm.head()
```

	# Customer ID	# recency	# frequency	# monetary
0	12346.0	326	3	77352.96
1	12347.0	2	8	4921.53
2	12348.0	75	5	2019.4
3	12349.0	19	4	4428.69
4	12350.0	310	1	334.4

ANEXO C – Selección de K y validación de clustering

```
fig, axes = plt.subplots(2, 2, figsize=(18, 10))
# 1) Silhouette (↑ mejor)
sns.lineplot(x=list(k_values), y=silhouette_scores, marker="o",
color=palette[1], ax=axes[0, 0])
axes[0, 0].set_title("Silhouette Score (↑)")
axes[0, 0].set_xlabel("K")
axes[0, 0].set_ylabel("Score")
axes[0, 0].set_xticks(list(k_values))
# 2) Calinski-Harabasz (↑ mejor)
```

```
sns.lineplot(x=list(k_values), y=ch_scores, marker="o",
color=palette[2], ax=axes[0, 1])
axes[0, 1].set_title("Calinski-Harabasz (↑)")
axes[0, 1].set_xlabel("K")
axes[0, 1].set_ylabel("Score")
axes[0, 1].set_xticks(list(k_values))
# 3) Davies-Bouldin (↓ mejor)
sns.lineplot(x=list(k_values), y=db_scores, marker="o",
color=palette[0], ax=axes[1, 0])
axes[1, 0].set_title("Davies-Bouldin (↓)")
axes[1, 0].set_xlabel("K")
axes[1, 0].set_ylabel("Score")
axes[1, 0].set_xticks(list(k_values))
# 4) Inertia (↓ mejor) — Elbow
sns.lineplot(x=list(k_values), y=inertias, marker="o", color=palette[3],
ax=axes[1, 1])
axes[1, 1].set_title("Inertia (↓) — Elbow")
axes[1, 1].set_xlabel("K")
axes[1, 1].set_ylabel("Inertia")
axes[1, 1].set_xticks(list(k_values))
plt.suptitle("Métricas para Selección del Número de Clusters",
fontsize=16, fontweight='bold')
plt.tight_layout()
plt.savefig('_img/KMeans_All_Metrics.png', dpi=300, bbox_inches='tight')
plt.show()
```




```
kmeans = KMeans(n_clusters=4, random_state=42, n_init=10)
rfm["cluster"] = kmeans.fit_predict(rfm[sc_cols])
# Reportar métricas finales del modelo elegido (K=4)
print(f"Silhouette Score (K=4):    {silhouette_score(rfm[sc_cols],
rfm['cluster']):.4f}")
print(f"Calinski-Harabasz
(K=4):   {calinski_harabasz_score(rfm[sc_cols], rfm['cluster']):.4f}")
```

```
print(f"Davies-Bouldin (K=4):      {davies_bouldin_score(rfm[sc_cols], rfm['cluster']):.4f}")
```

```
Silhouette Score (K=4): 0.3652
Calinski-Harabasz (K=4): 5054.2432
Davies-Bouldin (K=4): 0.9307
```

ANEXO D – Ajuste de BG/NBD

```
df_rftm_cal = calibration_and_holdout_data(
    transactions=df,
    customer_id_col="Customer ID",
    datetime_col="InvoiceDate",
    monetary_value_col="TotalPrice",
    calibration_period_end=end_date_cal,
    observation_period_end=end_date_obs
)
# Verificaciones
print(f"Shape: {df_rftm_cal.shape}")
print(f"Missing values: {df_rftm_cal.isna().sum().sum()}")
print(f"Columnas: {list(df_rftm_cal.columns)}")
print()
df_rftm_cal.describe()
```

	# frequency_cal	# recency_cal	# T_cal	# monetary_value_cal	# frequency_holdout	# monetary_value_holdout	# duration_holdout
Missing:	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Distinct:	7 (88%)	7 (88%)	8 (100%)	7 (88%)	7 (88%)	7 (88%)	3 (38%)
	Min 0.0 Max 4822.0	Min 0.0 Max 4822.0	Min 3.0 Max 4822.0	Min 0.0 Max 38662.955	Min 0.0 Max 4822.0	Min 0.0 Max 4822.0	Min 0.0 Max 4822.0
count	4822.0	4822.0	4822.0	4822.0	4822.0	4822.0	4822.0
mean	3.503940273745334	174.777478224803	335.00580671920363	275.66503154257305	1.896930734135213 5	18.64532276438867	222.0
std	7.409908673767798	173.59757328831802	147.24589088591273	749.6574081798872	3.918951093821607	70.99705006729192	0.0
min	0.0	0.0	3.0	0.0	0.0	0.0	222.0
25%	0.0	0.0	209.0	0.0	0.0	0.0	222.0
50%	1.0	136.0	368.0	187.41	1.0	5.780626569082166	222.0
75%	4.0	336.0	462.0	358.270625	2.0	19.320731557377048	222.0
max	159.0	513.0	516.0	38662.955	95.0	3096.0	222.0

```
l2_grid = [0.0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09,
           0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]

scores = []
for l2 in l2_grid:
    model = BetaGeoFitter(penalizer_coef=l2)
    model.fit(
        frequency=df_rftm_cal["frequency_cal"],
        recency=df_rftm_cal["recency_cal"],
        T=df_rftm_cal["T_cal"]
    )
    y_hat = model.predict(
        df_rftm_cal["duration_holdout"],
        df_rftm_cal["frequency_cal"],
        df_rftm_cal["recency_cal"],
        df_rftm_cal["T_cal"]
    )
    eval_df = (
        df_rftm_cal
        .reset_index()
        .assign(pred_frequency=np.asarray(y_hat))
```

```
        .dropna(subset=["frequency_holdout", "pred_frequency"])
    )
    rmse = np.sqrt(
        mean_squared_error(
            eval_df["frequency_holdout"],
            eval_df["pred_frequency"]
        )
    )
    scores.append({"L2 coefs": l2, "rmse_score": rmse})
resl = pd.DataFrame(scores)
# Mostrar resultados
best = resl.loc[resl["rmse_score"].idxmin()]
print(f"Menor RMSE: L2={best['L2 coefs']} con
RMSE={best['rmse_score']:.4f}")
print(f"Se selecciona L2=0.03 por criterio de regularización (ver
justificación abajo)")
best_l2 = 0.03
resl
```

	# L2 coefs	# rmse_score
	Missing: 0 (0%) Distinct: 20 (100%)	Missing: 0 (0%) Distinct: 20 (100%)
	Min 0.0 Max 1.0	Min 2.102409825... Max 2.252289124...
0	0.0	2.102409825265597
1	0.01	2.2522891240923166
2	0.02	2.162114887403784
3	0.03	2.1298455202205298
4	0.04	2.126464576674574
5	0.05	2.1264843983478
6	0.06	2.1237025871540993
7	0.07	2.122471735188781
8	0.08	2.1210482473790315
9	0.09	2.1198515817512837
10	0.1	2.1182859133613303
11	0.2	2.1108729932818098
12	0.3	2.104315686906474
13	0.4	2.1058066474591226
14	0.5	2.1056393896753893
15	0.6	2.105708360939373
16	0.7	2.109307447661588
17	0.8	2.1198579354465816
18	0.9	2.1194168952770087
19	1.0	2.119717315538586

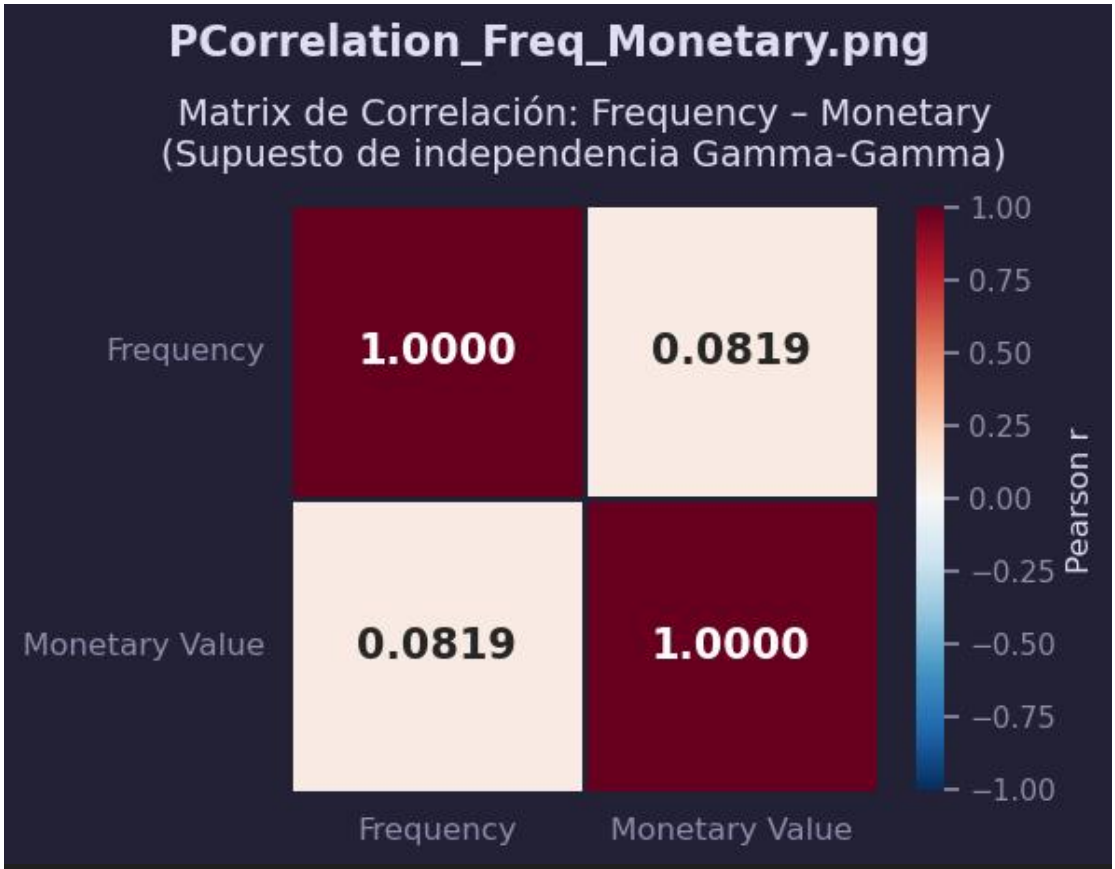
```
BetaGeo = BetaGeoFitter(penalizer_coef=best_l2)
BetaGeo.fit(df_rftm_cal['frequency_cal'],
        df_rftm_cal['recency_cal'],
        df_rftm_cal['T_cal'])
BetaGeo.summary
```

	# coef	# se(coef)	# lower 95% bound	# upper 95% bound
r	0.6411980265489421	0.016875393624842486	0.6081222550442508	0.6742737980536334
alpha	62.87271528308277	2.15577569485963	58.6473949211579	67.09803564500764
a	0.024304084664904788	0.004306530807443778	0.015863284282314982	0.03274488504749459
b	0.3215131684166498	0.04098854701780905	0.2411756162617440 5	0.4018507205715555

```
params = BetaGeo.params_
for param_name, param_value in params.items():
    print(f"  {param_name:10s} = {param_value:12.6f}")
print()
r = 0.641198
alpha = 62.872715
a = 0.024304
b = 0.321513
```

ANEXO E – Ajuste Gamma-Gamma

```
# Correlation Matrix: Frequency vs Monetary (independencia requerida
por Gamma-Gamma)
corr_matrix = returning_customers_cal[['frequency_cal', mv_col]].corr()
fig, ax = plt.subplots(figsize=(6, 5))
sns.heatmap(
    corr_matrix,
    annot=True,
    fmt=".4f",
    cmap="RdBu_r",
    center=0,
    vmin=-1, vmax=1,
    square=True,
    linewidths=1,
    linecolor="#232136",
    annot_kws={"size": 16, "weight": "bold"},
    cbar_kws={"shrink": 0.8, "label": "Pearson r"},
    ax=ax
)
ax.set_xticklabels(["Frequency", "Monetary Value"], fontsize=12)
ax.set_yticklabels(["Frequency", "Monetary Value"], fontsize=12,
rotation=0)
ax.set_title("Matrix de Correlación: Frequency – Monetary\n(Supuesto de
independencia Gamma-Gamma)",
             fontsize=14, pad=15)
plt.tight_layout()
plt.suptitle('PCorrelation_Freq_Monetary.png', fontsize=16,
fontweight='bold')
plt.savefig("_img/Correlation_Freq_Monetary.png", dpi=150,
bbox_inches="tight")
plt.show()
print(f"\nPearson r = {corr_matrix.iloc[0,1]:.4f}")
if abs(corr_matrix.iloc[0,1]) < 0.3:
    print("✓ Correlación Baja – El supuesto de independencia se cumple
razonablemente.")
else:
    print("⚠ Moderada/Alta Correlación – Los resultados de Gamma-Gamma
deben interpretarse con precaución.")
```




```
# PASO 2: Entrenar Gamma-Gamma
print("PASO 2: Entrenando modelo Gamma-Gamma...")
ggf = GammaGammaFitter(penalizer_coef=0.0)
ggf.fit(
    returning_customers_cal['frequency_cal'],
    returning_customers_cal[mv_col]
)
ggf.params_
```

	# 0
p	2.322839969518011
q	3.4675983819029064
v	434.6884070100496

ANEXO F – Cálculo CLV

```
# PASO 3: Calcular CLV
print("PASO 3: Calculando CLV...")
time_horizon = 12  # meses
monthly_discount_rate = 0.01  # 1%

returning_customers_cal['clv'] = ggf.customer_lifetime_value(
    BetaGeo,
    returning_customers_cal['frequency_cal'],
    returning_customers_cal['recency_cal'],
    returning_customers_cal['T_cal'],
    returning_customers_cal[mv_col],
```

```
    time=time_horizon,
    discount_rate=monthly_discount_rate
)
# Asignar CLV a todos
df_rftm_cal['clv'] = 0.0
df_rftm_cal.loc[df_rftm_cal['frequency_cal'] > 0, 'clv'] =
returning_customers_cal['clv'].values
print(f"  CLV total: £{df_rftm_cal['clv'].sum():,.2f}")
print(f"  CLV promedio: £{df_rftm_cal['clv'].mean():,.2f}")
print()
```

```
PASO 3: Calculando CLV…
CLV total: £6,485,442.00
CLV promedio: £1,344.97
```

```
cluster_analysis = df_rftm_cal.groupby('cluster').agg({
    'clv': ['mean', 'median', 'sum', 'count'],
    'frequency_cal': 'mean',
    'recency_cal': 'mean',
    'T_cal': 'mean',
    mv_col: 'mean',
}).round(2)

cluster_analysis.columns = ['_'.join(col) for col in
cluster_analysis.columns]
cluster_analysis = cluster_analysis.rename(columns={
    'clv_mean': 'CLV_promedio',
    'clv_median': 'CLV_mediana',
    'clv_sum': 'CLV_total',
    'clv_count': 'N_clientes',
    'frequency_cal_mean': 'Freq_promedio',
    'recency_cal_mean': 'Recency_promedio',
    'T_cal_mean': 'T_promedio',
    f'{mv_col}_mean': 'Valor_promedio'
})
cluster_analysis = cluster_analysis.sort_index()
print(cluster_analysis[[
    'N_clientes', 'CLV_promedio', 'CLV_total',
    'Freq_promedio', 'Valor_promedio'
]].to_string())
print()
# Etiquetar clusters usando los nombres dinámicos
cluster_labels = {k: f"Cluster {k}: {v}" for k, v in
cluster_names.items()}
```

cluster	N_clientes	CLV_promedio	CLV_total	Freq_promedio	Valor_promedio
0	619	325.07	201216.04	0.94	129.21
1	1645	103.14	169658.65	0.31	51.09
2	1143	4205.99	4807445.14	10.27	523.15
3	1415	923.76	1307122.17	2.87	400.90

```
total_clv = df_rftm_cal['clv'].sum()
total_clientes = len(df_rftm_cal)
for cluster_id in sorted(cluster_names.keys()):
    if cluster_id in cluster_analysis.index:
        row = cluster_analysis.loc[cluster_id]
        print(f"  {cluster_labels[cluster_id]}")
        print(f"  N Clientes: {int(row['N_clientes']):,}
({int(row['N_clientes'])/total_clientes*100:.1f}%)")
        print(f"  CLV promedio: £{row['CLV_promedio']:,.2f}")
        print(f"  CLV total: £{row['CLV_total']:,.2f}")
        print()
# Concentración
print("=" * 80)
print("CONCENTRACIÓN DE CLV:")
print("-" * 80)
print()
for cluster_id in sorted(cluster_names.keys()):
    if cluster_id in cluster_analysis.index:
        pct_clientes = cluster_analysis.loc[cluster_id, 'N_clientes'] /
total_clientes * 100
        pct_clv = cluster_analysis.loc[cluster_id, 'CLV_total'] /
total_clv * 100
        ratio = pct_clv / pct_clientes if pct_clientes > 0 else 0

        print(f"  {cluster_labels[cluster_id]}: {pct_clientes:5.1f}%
clientes -> {pct_clv:5.1f}% CLV (ratio: {ratio:.2f}x)")
print()
```

```
  Cluster 0: Occasional
  N Clientes: 619 (12.8%)
  CLV promedio: £325.07
  CLV total: £201,216.04

  Cluster 1: Lost
  N Clientes: 1,645 (34.1%)
  CLV promedio: £103.14
  CLV total: £169,658.65

  Cluster 2: Premium
  N Clientes: 1,143 (23.7%)
  CLV promedio: £4,205.99
  CLV total: £4,807,445.14

  Cluster 3: Potential
  N Clientes: 1,415 (29.3%)
  CLV promedio: £923.76
  CLV total: £1,307,122.17

================================================================================
CONCENTRACIÓN DE CLV:
--------------------------------------------------------------------------------

  Cluster 0: Occasional:  12.8% clientes ->   3.1% CLV (ratio: 0.24x)
  Cluster 1: Lost:  34.1% clientes ->   2.6% CLV (ratio: 0.08x)
  Cluster 2: Premium:  23.7% clientes ->  74.1% CLV (ratio: 3.13x)
  Cluster 3: Potential:  29.3% clientes ->  20.2% CLV (ratio: 0.69x)
```

ANEXO G – Validación estadística man-whitney u

```
# ═══════════════════════════════════════════════════════════════
# VALIDACIÓN ESTADÍSTICA
# ═══════════════════════════════════════════════════════════════
from scipy.stats import mannwhitneyu
print("RESUMEN COMPARATIVO: ACTIVOS vs CHURNED")
print("=" * 90)
summary = df_rftm_cal.groupby('churn').agg({
    'frequency_cal':      ['mean', 'median'],
    'recency_cal':        ['mean', 'median'],
    'T_cal':              ['mean', 'median'],
    mv_col:               ['mean', 'median'],
    'p_alive':            ['mean', 'std', 'min', 'max'],
    'clv':                ['mean', 'median', 'sum'],
    'predicted_purchases': ['mean']
}).round(2)
summary.columns = [f'{c[0]}_{c[1]}' for c in summary.columns]
summary.index = ['Activo (0)', 'Churned (1)']
print(summary.T.to_string())
print()
# ── Test Mann-Whitney U: P(alive) ──
active_pa  = df_rftm_cal[df_rftm_cal['churn'] == 0]['p_alive']
churned_pa = df_rftm_cal[df_rftm_cal['churn'] == 1]['p_alive']
stat1, p1 = mannwhitneyu(active_pa, churned_pa, alternative='greater')
print(f"Test Mann-Whitney U — P(alive) activos > churned:")
print(f"  U = {stat1:,.0f},  p-valor = {p1:.2e}")
if p1 < 0.001:
    print("  → Diferencia altamente significativa (p < 0.001)")
print()
# ── Test Mann-Whitney U: CLV ──
active_clv  = df_rftm_cal[df_rftm_cal['churn'] == 0]['clv']
churned_clv = df_rftm_cal[df_rftm_cal['churn'] == 1]['clv']
stat2, p2 = mannwhitneyu(active_clv, churned_clv, alternative='greater')
print(f"Test Mann-Whitney U — CLV activos > churned:")
print(f"  U = {stat2:,.0f},  p-valor = {p2:.2e}")
if p2 < 0.001:
    print("  → Diferencia altamente significativa (p < 0.001)")
print()
# ── Test Mann-Whitney U: Frequency ──
active_freq  = df_rftm_cal[df_rftm_cal['churn'] == 0]['frequency_cal']
churned_freq = df_rftm_cal[df_rftm_cal['churn'] == 1]['frequency_cal']
stat3, p3 = mannwhitneyu(active_freq, churned_freq,
alternative='greater')
print(f"Test Mann-Whitney U — Frequency activos > churned:")
print(f"  U = {stat3:,.0f},  p-valor = {p3:.2e}")
if p3 < 0.001:
    print("  → Diferencia altamente significativa (p < 0.001)")
```

```
RESUMEN COMPARATIVO: ACTIVOS vs CHURNED
================================================================================
                          Activo (0)  Churned (1)
frequency_cal_mean              3.58         2.35
frequency_cal_median            1.00         1.00
recency_cal_mean              181.81        61.57
recency_cal_median            154.00        42.50
T_cal_mean                    333.25       363.23
T_cal_median                  368.00       384.00
monetary_value_cal_mean       270.74       354.96
monetary_value_cal_median     184.93       212.30
p_alive_mean                    0.96         0.55
p_alive_std                     0.07         0.15
p_alive_min                     0.70         0.00
p_alive_max                     1.00         0.70
clv_mean                     1395.96       524.00
clv_median                    530.41       291.88
clv_sum                   6337673.90    147768.09
predicted_purchases_mean        2.14         0.74

Test Mann-Whitney U — P(alive) activos > churned:
  U = 1,280,280,  p-valor = 1.85e-182
  → Diferencia altamente significativa (p < 0.001)

Test Mann-Whitney U — CLV activos > churned:
...
  → Diferencia altamente significativa (p < 0.001)

Test Mann-Whitney U — Frequency activos > churned:
  U = 622,688,  p-valor = 7.85e-01
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

ANEXO H - Clasificación de Riesgo y Umbral de Churn

```
# ═══════════════════════════════════════════════════════════════════════
# ANÁLISIS CRUZADO: RIESGO DE INACTIVIDAD × SEGMENTO
# ═══════════════════════════════════════════════════════════════════════

# Tabla cruzada
print("TABLA CRUZADA: ESTADO DE ACTIVIDAD (BG/NBD) × SEGMENTO (RFM)")
print("=" * 80)
ct = pd.crosstab(
    df_rftm_cal['segment'],
    df_rftm_cal['churn'].map({0: 'Activo', 1: 'En riesgo'}),
    margins=True
)
print(ct)
print()
# Tasa de riesgo por segmento
print("TASA DE RIESGO DE INACTIVIDAD POR SEGMENTO:")
print("-" * 55)
churn_by_segment = df_rftm_cal.groupby('segment')['churn'].agg(['sum',
'count', 'mean'])
churn_by_segment.columns = ['N_riesgo', 'N_total', 'Tasa_riesgo']
churn_by_segment['Tasa_riesgo_pct'] = (churn_by_segment['Tasa_riesgo']
* 100).round(1)
churn_by_segment = churn_by_segment.sort_values('Tasa_riesgo',
ascending=False)
for seg, row in churn_by_segment.iterrows():
    print(f"  {seg:15s}: {row['Tasa_riesgo_pct']:5.1f}% en
riesgo  ({int(row['N_riesgo']):,} / {int(row['N_total']):,})")
print()
# CLV en riesgo por segmento
```

```
print("CLV EN RIESGO POR SEGMENTO (CLV de clientes con P(alive) < umbral):")
print("-" * 65)
clv_at_risk = df_rftm_cal[df_rftm_cal['churn'] == 1].groupby('segment')['clv'].agg(['sum', 'mean', 'count'])
clv_at_risk.columns = ['CLV_total_riesgo', 'CLV_medio_riesgo', 'N_riesgo']
for seg, row in clv_at_risk.iterrows():
    print(f"  {seg:15s}: £{row['CLV_total_riesgo']:>12,.2f} total | £{row['CLV_medio_riesgo']:>10,.2f} medio | n={int(row['N_riesgo']):,}")
print()
```

```
TABLA CRUZADA: ESTADO DE ACTIVIDAD (BG/NBD) × SEGMENTO (RFM)
================================================================================
churn       Activo  En riesgo   All
segment
Lost          1524        121  1645
Occasional     582         37   619
Potential     1309        106  1415
Premium       1125         18  1143
All           4540        282  4822

TASA DE RIESGO DE INACTIVIDAD POR SEGMENTO:
------------------------------------------------------
  Potential      :   7.5% en riesgo  (106 / 1,415)
  Lost           :   7.4% en riesgo  (121 / 1,645)
  Occasional     :   6.0% en riesgo  (37 / 619)
  Premium        :   1.6% en riesgo  (18 / 1,143)

CLV EN RIESGO POR SEGMENTO (CLV de clientes con P(alive) < umbral):
-----------------------------------------------------------------
  Lost           : £   29,878.21 total | £    246.93 medio | n=121
  Occasional     : £   11,891.12 total | £    321.38 medio | n=37
  Potential      : £   52,058.62 total | £    491.12 medio | n=106
  Premium        : £   53,940.14 total | £  2,996.67 medio | n=18
```