Optimización Estratégica en Instituciones Financieras: Identificación y Predicción de Fuga de Clientes mediante Aprendizaje Automático

Eduardo David González*
Subdirector de Analítica Avanzada & ML-Ops, Grupo Regional

Este artículo aborda la problemática de la identificación y predicción de fuga de clientes en instituciones bancarias mediante la aplicación de técnicas avanzadas de ciencia de datos. Utilizando un conjunto de datos proporcionado por Kaggle, que incluye variables como score crediticio, país, género, edad, años de antigüedad, saldo, cantidad de productos, entre otros, llevamos a cabo un análisis descriptivo detallado y una limpieza exhaustiva de los datos. Posteriormente, exploramos la correlación entre las variables y aplicamos el algoritmo de clustering K-Means para identificar patrones de comportamiento y segmentar a los clientes.

Adoptando un enfoque de modelado predictivo, evaluamos varios algoritmos de clasificación utilizando la librería PyCaret y seleccionamos el modelo más efectivo. Además, optimizamos sus hiperparámetros para mejorar aún más su rendimiento. Los resultados obtenidos no solo proporcionan una visión profunda de la fuga de clientes en el contexto bancario, sino que también ofrecen aplicaciones prácticas para mejorar las estrategias de retención de clientes.

Este estudio contribuye a la comprensión de las complejidades en la retención de clientes bancarios y demuestra el potencial de la ciencia de datos en la toma de decisiones estratégicas en el sector financiero. Además, se discuten consideraciones éticas, limitaciones y se sugieren direcciones futuras para la investigación en este campo en constante evolución.

Palabras clave: ciencia de datos, aprendizaje automático, clasificación binaria, fuga
Revisión: última modificación el día 20 de noviembre de 2023

---

1. Introducción

En el dinámico escenario de la banca moderna, la retención de clientes se ha vuelto imperativa para la sostenibilidad y el crecimiento de las instituciones financieras (1). La anticipación eficaz de la fuga de clientes se presenta como un desafío crítico en la banca, y la aplicación de técnicas avanzadas de ciencia de datos proporciona una solucion viable y escalable para esta compleja problemática (2).

El análisis predictivo, respaldado por el acceso a conjuntos de datos integrales, se convierte en un aliado esencial para desentrañar los misterios de la retención de clientes en el ámbito bancario.

* M.Sc.c. en Ciencia de Datos, M.Eng. en Seguridad de la Información, M. en Finanzas, B.Sc. en Ciencias Computacionales

En este contexto, este artículo se sumerge en la aplicación de rigurosas metodologías de ciencia de datos para explorar, entender y prever la fuga de clientes con el objetivo de fortalecer las estrategias de retención (4).

2. Contextualización

La industria bancaria, inmersa en la vorágine de la transformación digital y la evolución constante de las expectativas del cliente, enfrenta desafíos cruciales para mantener y expandir su base de usuarios (1). Entre estos desafíos, la fuga de clientes emerge como un complejo fenómeno de gran importancia estratégica. La retención de clientes resulta un pilar fundamental para la estabilidad financiera y el crecimiento sostenible de las instituciones bancarias en un panorama altamente competitivo.

La pérdida de clientes puede ser causada por una variedad de factores, desde experiencias insatisfactorias hasta cambios en las circunstancias financieras del cliente. En este contexto, la capacidad de identificar tempranamente a aquellos clientes propensos a abandonar la institución se convierte en un imperativo para diseñar estrategias efectivas de retención.

Las instituciones financieras, conscientes de esta realidad, buscan incansablemente maneras de comprender y anticipar el comportamiento de sus clientes. La implementación de soluciones basadas en la ciencia de datos y el análisis predictivo se ha convertido en una herramienta invaluable en este esfuerzo, ofreciendo una visión más profunda de las interacciones cliente-banco y la capacidad de prever movimientos futuros (2).

En este escenario, nuestra investigación se inserta como una exploración detallada en el campo de la ciencia de datos aplicada a la gestión bancaria. Utilizando un conjunto de datos proporcionado por Kaggle, buscamos comprender no solo las razones detrás de la fuga de clientes, sino también para diseñar estrategias prácticas y eficaces para su retención. A través del análisis descriptivo, la identificación de patrones de comportamiento mediante técnicas de clustering y la construcción de modelos predictivos avanzados, aspiramos a comprender de mejor manera esta problemática crítica y contribuir a la evolución continua de la gestión bancaria en la era digital.

3. Datos y Metodología

Nuestra investigación se basa en un conjunto de datos proporcionado por Kaggle, que abarca una variedad de variables relacionadas con los clientes bancarios, incluyendo algunos datos demográficos tales como país, género, edad, y otros datos característicos del cliente y su relación con la institución, por ejemplo, score crediticio, años de antigüedad, saldo en débito, cantidad de productos, tenencia de tarjeta de crédito, si se considera cliente activo, y su salario estimado (3).

Para abordar nuestros objetivos de investigación, aplicamos un enfoque metodológico robusto que consta de varias fases. Iniciamos con un análisis descriptivo de los datos para comprender la

distribución y las características generales. Posteriormente, llevamos a cabo una limpieza de datos exhaustiva para abordar posibles valores atípicos, valores faltantes y otras irregularidades.

La identificación de patrones de comportamiento se realizó mediante técnicas avanzadas de análisis de correlación. Exploramos la interconexión entre las variables clave que podrían influir en la retención de clientes. Además, aplicamos el algoritmo de clustering K-Means para segmentar a los clientes en grupos homogéneos, revelando estructuras latentes en los datos.

La construcción de modelos predictivos se llevó a cabo utilizando la biblioteca PyCaret, que nos permitió evaluar y comparar diversos algoritmos de clasificación. Seleccionamos el modelo más prometedor y procedimos a la optimización de sus hiperparámetros para mejorar su rendimiento predictivo.

4. Análisis Descriptivo

La primera etapa de nuestro estudio consistió en un análisis descriptivo exhaustivo de los datos. Este proceso nos proporcionó una comprensión detallada de la distribución y las características fundamentales de las variables clave en nuestro conjunto de datos (5).

De las distribuciones de los datos, se obtuvieron las siguientes conclusiones para cada una de las variables:

1. **age**: La edad (`age`) presenta una distribución normal entre los 20 y 60 años. Figura 2.

2. **balance**: El saldo (`balance`) presenta una distribución normal entre los valores de 50,000 y 200,000. Sin embargo, cabe destacar que la mayoría de las cuentas presentan saldo 0. Figura 3.

3. **credit_score**: La puntuación de crédito (`credit_score`) presenta una distribución normal entre los valores de 500 y 800. Figura 4.

4. **estimated_salary**: La estimación del salario (`estimated_salary`) muestra una distribución uniforme. Figura 5.

5. **gender**: El género(`age`) presenta una distribución uniforme. Tabla 1.

6. **country**: La nacionalidad (`country`) francesa representa a un 50 % de los clientes, mientras que la otra mitad se distribuye uniformemente entre España y Alemania. Tabla 2.

7. **tenure**: La antiguedad de los clientes (`tenure`) presenta una distribución uniforme. Tabla 3.

8. **products_number**: El cuanto al número de productos financieros (`products_number`), naturalmente, es menor la cantidad de clientes con más productos, que aquellos con menos. Tabla 4.

9. **credit_card**: La posesión de una tarjeta de crédito (`credit_card`) corresponde a un 70 % de los clientes. Tabla 5.

10. **active_member**: La variable `active_member` indica que la mitad de los clientes son considerados activos, y la otra mitad inactivos. Tabla 6.

Tabla 1 Distribución de género

Género	Registros
Male	5457
Female	4543

Tabla 2 Distribución de países

País	Registros
Francia	5014
Alemania	2509
España	2477

Tabla 3 Distribución de antiguedad de los clientes (años)

Antiguedad	Registros
0	413
1	1035
3	1009
4	989
5	1012
6	967
7	1028
8	1025
9	984
10	490

Tabla 4 Distribución de cantidad de productos contratados

Productos	Registros
1	5084
2	4590
3	266
4	60

Tabla 5 Distribucion de tenencia de tarjeta de crédito (tdc)

tdc	Registros
Si	7055
No	2945

Tabla 6 **Distribución de actividad**

Activos	Registros
Si	5151
No	4849

5. Correlación y Patrones

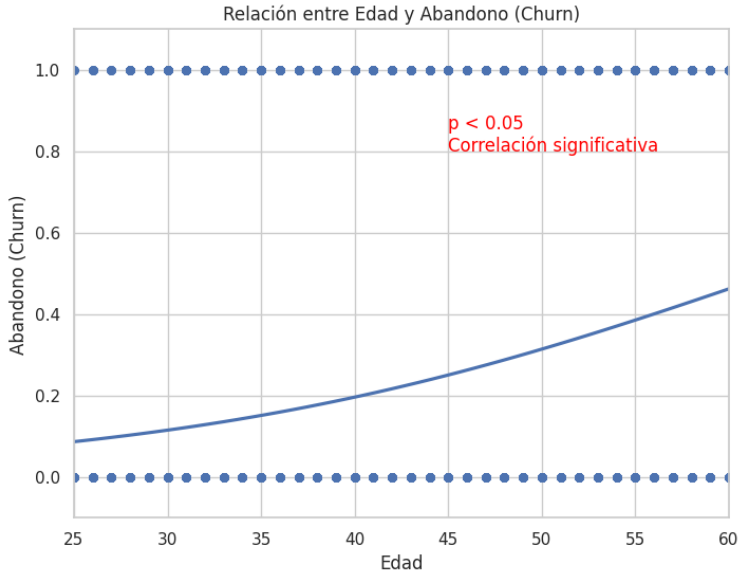
En la fase siguiente de nuestro estudio, nos adentramos en la identificación de patrones y correlaciones entre las variables clave que podrían influir en la retención de clientes bancarios. Aplicamos técnicas de análisis de correlación para evaluar las relaciones estadísticas entre diferentes atributos (6).

En nuestra investigación, nos propusimos explorar la posible relación entre la edad (*age*) y el abandono (*churn*) en el contexto de nuestro estudio. Planteamos dos hipótesis para abordar esta cuestión.

La **Hipótesis Nula (H0)** afirmaba que no existe una correlación significativa entre la edad y el abandono, mientras que la **Hipótesis Alternativa (H1)** sugería la presencia de una correlación significativa entre estas variables.

Al aplicar el coeficiente de correlación de Pearson, obtuvimos un valor de 0.2853, lo cual indica una correlación positiva moderada entre la edad y el abandono. Este resultado se ve respaldado por un *p-value* de 0.0000, que es menor que el umbral de significancia convencional de 0.05. Figura 1.

Figura 1 **Relación entre edad y abandono**



En consecuencia, rechazamos la Hipótesis Nula (H0) y concluimos que, en nuestra muestra, hay evidencia estadística que respalda la existencia de una correlación significativa entre la edad y el abandono. Este hallazgo aporta una nueva perspectiva a nuestra comprensión de los factores que podrían influir en el comportamiento de abandono en el contexto estudiado.

De la matriz de correlación anexada en la figura 6, se obtuvieron las siguientes conclusiones para cada una de las variables:

1. **customer_id**: La variable `customer_id` tiene una correlación muy cercana a cero con todas las demás variables. Esto era de esperar, ya que se trata de un identificador único para cada cliente y no debería tener una correlación sustancial con otras características.

2. **credit_score**: La puntuación de crédito (`credit_score`) muestra correlaciones débiles con la mayoría de las variables, lo que sugiere que no hay una relación lineal fuerte con otras características.

3. **age**: La edad (`age`) muestra una correlación positiva moderada con la variable `churn`, lo que sugiere que los clientes más jóvenes pueden estar más inclinados a abandonar el servicio. También muestra una correlación positiva con la variable `active_member`, lo que indica que los clientes más jóvenes pueden estar más activos en el uso de servicios bancarios.

4. **balance**: El saldo (`balance`) tiene una correlación positiva moderada con la variable `churn`, lo que sugiere que los clientes con saldos más altos pueden ser menos propensos a abandonar. Además, muestra una correlación negativa moderada con la variable `products_number`, lo que indica que los clientes con saldos más altos pueden tener menos productos financieros.

5. **products_number**: El número de productos financieros (`products_number`) muestra una correlación negativa moderada con la variable `balance`, lo que sugiere que los clientes con más productos financieros pueden tener saldos más bajos.

6. **credit_card**: La posesión de una tarjeta de crédito (`credit_card`) no muestra correlaciones fuertes con otras variables en la matriz.

7. **active_member**: La variable `active_member` tiene una correlación negativa moderada con la variable `churn`, lo que indica que los clientes inactivos pueden ser más propensos a abandonar.

8. **estimated_salary**: La estimación del salario (`estimated_salary`) no muestra correlaciones fuertes con otras variables en la matriz.

9. **churn**: La variable `churn` muestra correlaciones moderadas con las variables `age` (positiva) y `balance` (positiva), lo que sugiere que la edad y el saldo pueden influir en la probabilidad de abandono. También muestra una correlación negativa moderada con la variable `active_member`, lo que indica que los clientes inactivos pueden ser más propensos a abandonar.

6. Análisis de Clustering

En la fase de análisis de clustering, aplicamos el algoritmo K-Means para segmentar a los clientes en grupos homogéneos basados en sus características. Este enfoque nos permitió identificar patrones latentes en los datos y clasificar a los clientes en categorías que comparten características similares (7).

En la exploración de la estructura subyacente de nuestros datos, empleamos métodos claves para determinar el número óptimo de clusters en nuestro análisis de clustering. Dos enfoques destacados son el método del **Codo (Elbow)** y el método de **Silueta (Silhouette)**.

El método del **Codo** implica ajustar el modelo de clustering para diferentes cantidades de clusters y observar cómo se reduce la variabilidad intra-cluster. Al representar estos resultados gráficamente, buscamos el punto en el que la disminución de la variabilidad se asemeja a la forma de un codo. Este punto sugiere el número óptimo de clusters (8).

Por otro lado, el método de **Silueta** evalúa la cohesión y separación de los clusters asignando a cada punto un valor de silueta, que oscila entre -1 y 1. Un valor alto indica que el punto está bien asignado a su cluster y mal asignado a otros. La puntuación global más alta sugiere el número óptimo de clusters (9).

Ambos métodos ofrecen perspectivas valiosas para guiar la elección del número de clusters en nuestro análisis, equilibrando la coherencia interna de los clusters y la distinción entre ellos.

Como se puede observar en la figura 7, los resultados obtenidos de ambos métodos, coincidieron en que 4 es la cantidad óptima de clusters. Dichos clusters se agregaron al conjunto de datos como una variable nueva para ser utilizada en el modelo predictivo desarrollado. Figura 8.

7. Modelado Predictivo

La etapa de modelado predictivo fue fundamental en nuestro estudio para anticipar la fuga de clientes con precisión. Utilizamos la biblioteca PyCaret, que ofrece una amplia variedad de algoritmos de clasificación y herramientas para la construcción eficiente de modelos predictivos(10).

Inicialmente, evaluamos y comparamos varios algoritmos para seleccionar el modelo base más prometedor, obteniendo los resultados mostrados en la figura 7. Se eligió el MCC como métrica de evaluación, la cual fue de 0,5265 para el modelo elegido: CatBoost.

7.1. Coeficiente de Correlación de Matthews (MCC)

El Coeficiente de Correlación de Matthews (MCC) es una métrica de evaluación ampliamente utilizada en problemas de clasificación binaria, especialmente cuando las clases están desbalanceadas, ya que ofrece una evaluación equitativa y considera la complejidad inherente a la distribución de las clases. Esta medida proporciona una evaluación robusta del rendimiento de un modelo al considerar verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Tabla 7 Comparación de Rendimiento de Modelos

Modelo	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (Sec)
catboost	0.8616	0.8571	0.4818	0.7516	0.5860	0.5077	0.5265	3.7690
gbc	0.8621	0.8618	0.4650	0.7693	0.5787	0.5023	0.5255	0.1220
rf	0.8584	0.8430	0.4432	0.7638	0.5605	0.4831	0.5088	0.1270
ada	0.8531	0.8429	0.4552	0.7238	0.5578	0.4753	0.4942	0.0490
xgboost	0.8471	0.8283	0.4812	0.6759	0.5614	0.4721	0.4825	0.0730
et	0.8501	0.8380	0.4208	0.7298	0.5333	0.4517	0.4763	0.0980
qda	0.8246	0.8055	0.3878	0.6118	0.4743	0.3752	0.3895	0.0060
dt	0.7863	0.6746	0.4860	0.4761	0.4806	0.3462	0.3465	0.0100
lda	0.8106	0.7726	0.2504	0.5837	0.3490	0.2587	0.2911	0.0090
ridge	0.8079	0.0000	0.1346	0.6354	0.2215	0.1620	0.2269	0.0080
nb	0.7873	0.7486	0.1003	0.4154	0.1606	0.0880	0.1183	0.0070
lr	0.7911	0.6813	0.0743	0.4283	0.1261	0.0704	0.1063	0.5230
knn	0.7606	0.5286	0.0806	0.2403	0.1204	0.0201	0.0243	0.1930
dummy	0.7963	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0070
svm	0.7487	0.0000	0.0540	0.0562	0.0412	-0.0152	-0.0158	0.0160

El MCC se define mediante la siguiente fórmula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Donde: - TP es el número de verdaderos positivos. - TN es el número de verdaderos negativos. - FP es el número de falsos positivos. - FN es el número de falsos negativos.

El MCC proporciona una medida equilibrada que varía entre -1 y 1. Un MCC de 1 indica una predicción perfecta, 0 indica predicciones aleatorias y -1 indica una predicción completamente incorrecta(11).

Cuando se trata de problemas de clasificación binaria con clases desbalanceadas, el MCC presenta varias ventajas:

- **Sensibilidad a Clases Desbalanceadas:** A diferencia de métricas simples como la precisión, el MCC no se ve afectado por la distribución desigual de las clases. Proporciona una evaluación equitativa del rendimiento incluso cuando una clase es significativamente más grande que la otra.
- **Considera Falsos Negativos y Falsos Positivos:** Al tener en cuenta tanto los falsos negativos como los falsos positivos, el MCC brinda una visión integral del rendimiento del modelo, lo que es crucial en situaciones donde los costos de estos errores son desiguales.
- **Robustez en Casos Desafiantes:** El MCC es particularmente robusto en casos donde las clases son desbalanceadas y las predicciones incorrectas para la clase minoritaria son críticas.

7.2. CatBoost

CatBoost es un modelo de aprendizaje automático que ha ganado popularidad en diversas aplicaciones debido a su capacidad para manejar datos categóricos de manera eficiente y su rendimiento

sobresaliente en problemas de clasificación y regresión. Fue desarrollado por Yandex, la compañía de tecnología rusa(12).

CatBoost se destaca por varias características que lo hacen atractivo en la práctica(13):

- **Manejo Eficiente de Datos Categóricos:** A diferencia de otros modelos, CatBoost puede manejar variables categóricas sin necesidad de codificación previa, simplificando el proceso de preparación de datos.
- **Regularización Incorporada:** CatBoost incorpora técnicas de regularización durante el entrenamiento, lo que ayuda a prevenir el sobreajuste y mejora la generalización del modelo.
- **Optimización de la Velocidad de Aprendizaje:** Utiliza una estrategia de optimización adaptativa de la velocidad de aprendizaje, lo que permite un entrenamiento más rápido y eficiente del modelo.
- **Tratamiento Integrado de Características Numéricas y Categóricas:** CatBoost maneja naturalmente ambos tipos de características, eliminando la necesidad de preprocesamiento adicional.

7.3. Interpretación

Podemos ver en la figura 9 la matriz de confusión correspondiente al modelo seleccionado. Asimismo, graficamos la curva ROC en la figura 10. En cuanto a la importancia de las variables, podemos ver en la figura 11 que todas las variables tienen un grado relevante de importancia, principalmente la edad y la cantidad de productos. Sin embargo, la variable con menor peso en el modelo resultó ser el cluster asociado, obtenido mediante k-means.

De la misma manera, se obtuvo la gráfica de SHAP, mostrada en la figura 12, la cual proporciona una representación visual de cómo cada característica contribuye a la predicción de un modelo(14). Cada punto en la gráfica representa una instancia de datos, y la posición horizontal muestra la contribución de una característica específica(15). Los puntos a la derecha indican una contribución positiva a la predicción. Los puntos a la izquierda indican una contribución negativa a la predicción. El color del punto indica el valor de la característica, permitiendo una fácil identificación de los patrones.

Gracias a estas gráficas, se puede determinar que, los clientes más jóvenes y con mayor cantidad de productos, tienen menor probabilidad de fuga.

8. Optimización

En la fase de optimización de hiperparámetros, buscamos mejorar el rendimiento predictivo de nuestro modelo seleccionado. Esta etapa es esencial para ajustar los parámetros del modelo de manera que se optimice su capacidad predictiva (16).

El total de iteraciones se refiere a las rondas de entrenamiento que el modelo CatBoost llevará a cabo durante el proceso de ajuste. Establecer un valor apropiado es crucial para lograr un equilibrio entre un modelo bien entrenado y evitar el sobreajuste.

CatBoost utiliza la validación cruzada estratificada por defecto. Se puede especificar el número de pliegues o particiones en las que se dividirá el conjunto de datos durante la validación cruzada. Este parámetro es fundamental para evaluar el rendimiento del modelo de manera robusta y evitar sesgos en la evaluación.

El early stopping es una técnica valiosa para evitar el sobreajuste durante el entrenamiento del modelo. Cuando se activa, CatBoost detendrá automáticamente el entrenamiento si no observa mejoras significativas en la métrica de evaluación seleccionada en un número especificado de iteraciones consecutivas. Esto ayuda a seleccionar el mejor modelo sin entrenar innecesariamente por demasiadas iteraciones.

Estos parámetros se utilizan en conjunto para encontrar el conjunto óptimo de hiperparámetros para el modelo CatBoost. A menudo, se realiza una búsqueda sistemática, ajustando estos parámetros y evaluando el rendimiento del modelo en un conjunto de validación.

Para este nuevo entrenamiento, se realizaron 100 iteraciones, se empleó una validación cruzada de 10 pliegues para evaluar el rendimiento de manera exhaustiva, y el entrenamiento se detuvo temprano si no se observan mejoras consistentes.

Tras la aplicación de dichos ajustes, el coeficiente de correlación de Matthews mejoró de 0,5265 a 0,5346. Esto indica que los ajustes realizados fueron efectivos en mejorar el rendimiento del modelo en la métrica MCC.

8.1. Optimización del Umbral de Probabilidad

En problemas de clasificación binaria, la optimización del umbral de probabilidad es una estrategia clave para encontrar el punto óptimo de corte que equilibre las tasas de verdaderos positivos y falsos positivos, maximizando así la métrica de interés.

El umbral de probabilidad es el valor crítico que determina cuándo una observación se clasifica como positiva o negativa. Ajustar este umbral permite explorar el trade-off entre la sensibilidad y la especificidad del modelo. La figura 13 muestra los resultados para múltiples métricas de evaluación.

El proceso implica evaluar el rendimiento del modelo en un conjunto de validación para varios valores del umbral de probabilidad. Se selecciona aquel umbral que maximiza la métrica de interés o logra el equilibrio deseado entre tasas de verdaderos y falsos positivos.

Después de optimizar el umbral correspondiente al MCC, tal como se muestra en la figura 14, se pudo incrementar el MCC de 0,5346 a 0,5682.

Este cambio implica una mejora de 8 % sobre el modelo desde el entrenamiento inicial (MCC=0.5265) hasta la optimización del umbral (MCC=0.5682)

9. Resultados y Aplicaciones Prácticas

Despues de someter al modelo a inferencia sobre datos desconocidos, se obtuvo un MCC de 0,5650, bastante similar a los resultados obtenidos en el conjunto de entrenamiento (0,5682). Los resultados de nuestro estudio ofrecen una visión profunda de los patrones de comportamiento de los clientes bancarios y proporcionan una base sólida para la toma de decisiones estratégicas. La aplicación de técnicas de ciencia de datos ha revelado correlaciones significativas entre variables clave, permitiendo la identificación temprana de clientes propensos a la fuga (5).

El análisis descriptivo detallado ha arrojado luz sobre las características demográficas y financieras que están más fuertemente asociadas con la retención de clientes. Estos hallazgos no solo contribuyen al entendimiento teórico de la fuga de clientes en el contexto bancario, sino que también ofrecen aplicaciones prácticas inmediatas.

La construcción y optimización de modelos predictivos han demostrado ser herramientas valiosas para anticipar la fuga de clientes con una precisión notable. La integración de estos modelos en los sistemas de gestión bancaria proporciona una oportunidad estratégica para implementar medidas proactivas de retención. Identificar a los clientes en riesgo y diseñar estrategias personalizadas para retenerlos puede marcar la diferencia en un entorno financiero altamente competitivo (17, 18).

10. Consideraciones Éticas y Limitaciones

La realización de investigaciones en el ámbito de la ciencia de datos y la gestión bancaria implica una serie de consideraciones éticas fundamentales. En primer lugar, es crucial garantizar la privacidad y confidencialidad de los datos de los clientes utilizados en el estudio. Todas las prácticas deben adherirse a los más altos estándares éticos, siguiendo las regulaciones y directrices pertinentes en materia de protección de datos (19).

Además, se debe tener precaución al interpretar los resultados de los modelos predictivos. Aunque estos modelos pueden proporcionar valiosas percepciones, no deben considerarse infalibles. La toma de decisiones estratégicas basadas en estos resultados debe ir acompañada de un análisis cuidadoso y, cuando sea posible, validación adicional (20).

Es importante reconocer las limitaciones inherentes a este estudio. La calidad y representatividad de los resultados dependen en gran medida de la calidad y diversidad de los datos disponibles. Además, las condiciones y variables que afectan la retención de clientes pueden ser extremadamente dinámicas, lo que podría afectar la generalización de los hallazgos a diferentes contextos temporales (21).

11. Conclusiones y Futuras Direcciones

A pesar de los valiosos hallazgos obtenidos, es importante señalar que el conjunto de datos utilizado en este estudio es relativamente básico, con un número limitado de variables. Para futuras

investigaciones, se sugiere la incorporación de conjuntos de datos más amplios y detallados. Entre las áreas de mejora sugeridas se encuentran:

- **Datos Demográficos y Financieros Detallados:** Se recomienda la inclusión de información demográfica más detallada sobre los clientes, así como datos financieros adicionales que reflejen su comportamiento profundo con la institución bancaria. Esto podría incluir detalles sobre tendencias de gastos, historial crediticio y productos financieros utilizados.
- **Interacción con Productos:** Para comprender mejor la relación entre los clientes y los productos bancarios, se deberían incorporar datos detallados sobre la interacción del cliente con diversos productos, frecuencia de uso y niveles de satisfacción.
- **Tendencias y Factores Externos:** Explorar cómo las tendencias económicas y otros factores externos podrían influir en el comportamiento de abandono, proporcionando así una visión más completa y contextualizada.

En resumen, aunque este estudio ofrece una visión inicial, se sugiere que futuras investigaciones utilicen conjuntos de datos más ricos y variados para profundizar en la comprensión del comportamiento de abandono de clientes en el ámbito bancario.

12. Anexos

Anexo A: Análisis descriptivo

Figura 2 **Distribución de edades**

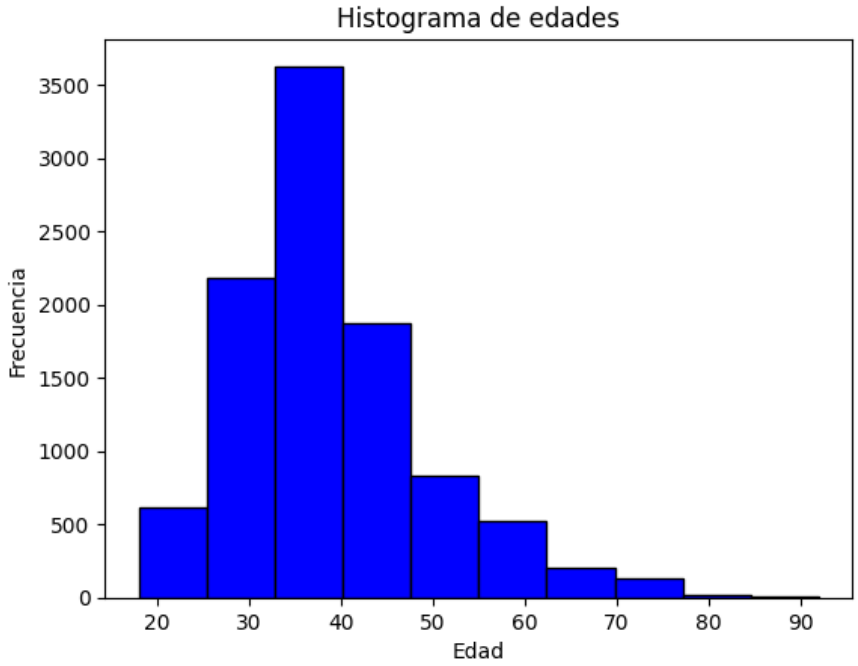


Figura 3 **Distribución de saldos en débito**

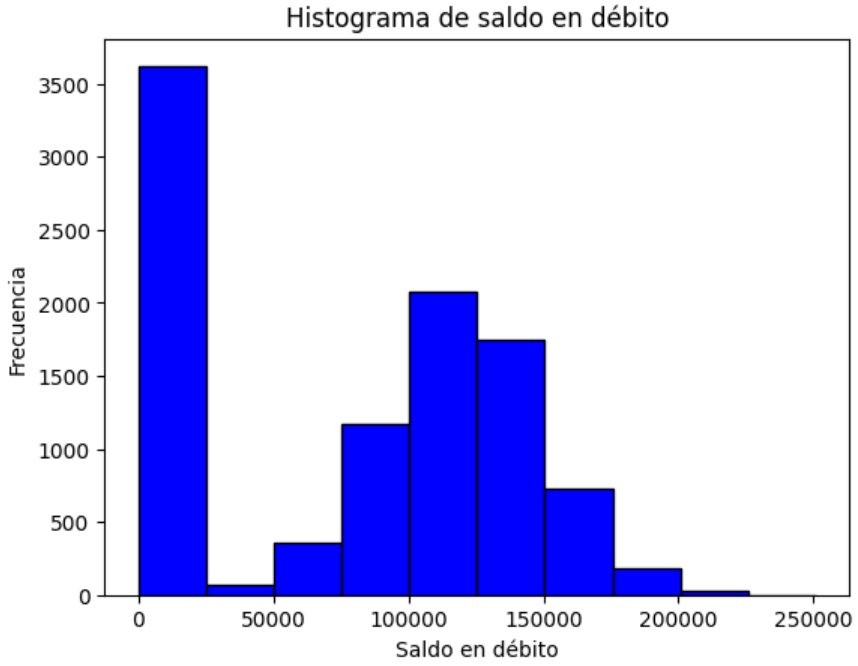


Figura 4 **Distribución de score crediticio**

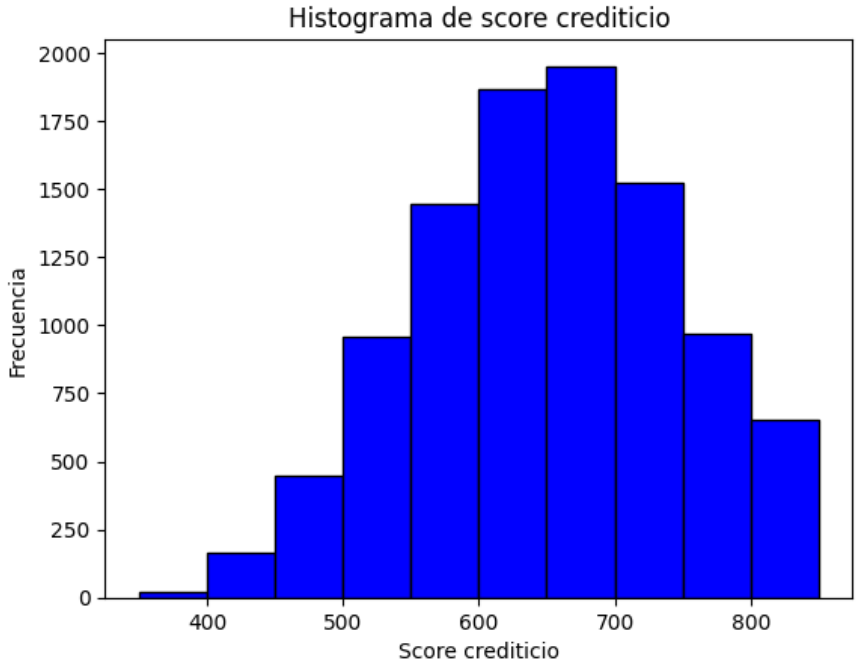
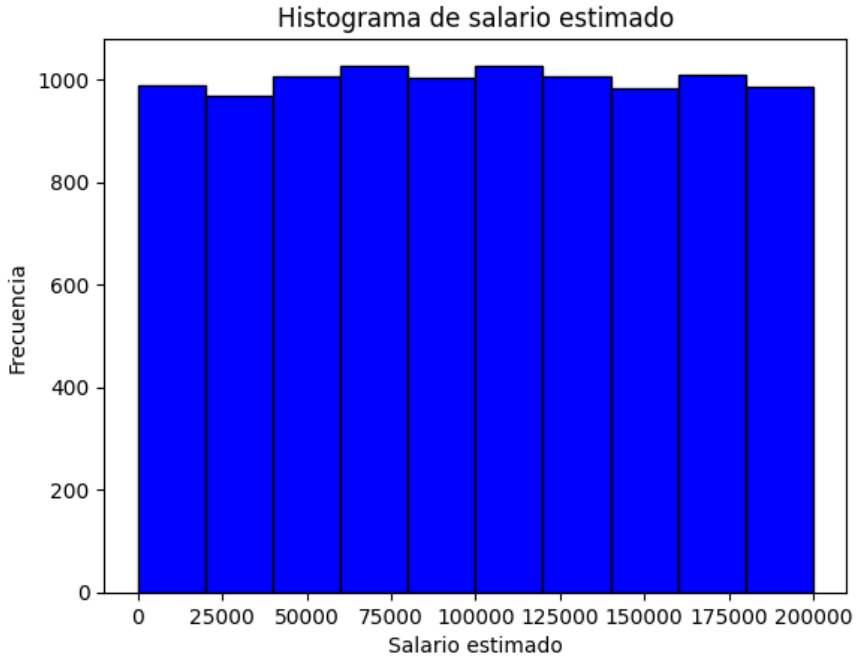
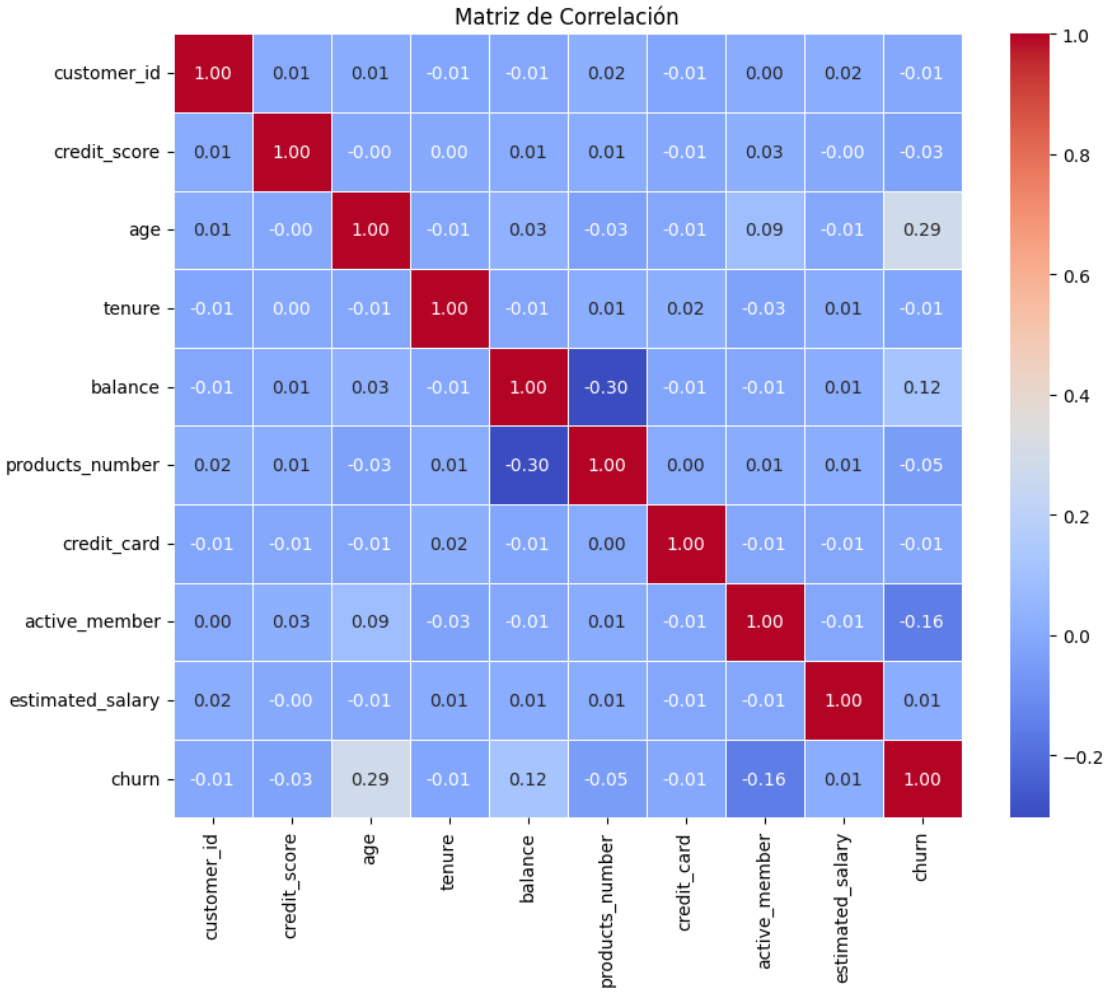


Figura 5 **Distribución de salarios estimados**



Anexo B: Correlación

Figura 6 Matriz de correlación entre variables



Anexo C: Análisis de clustering

Figura 7 **Optimización de clusters**

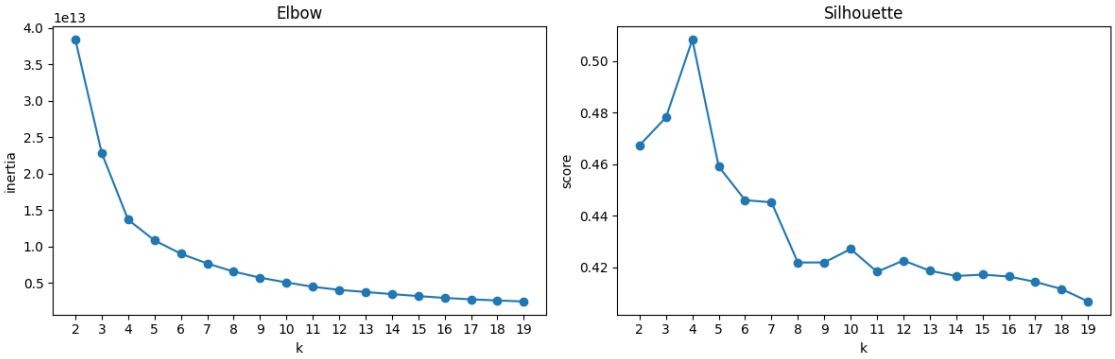
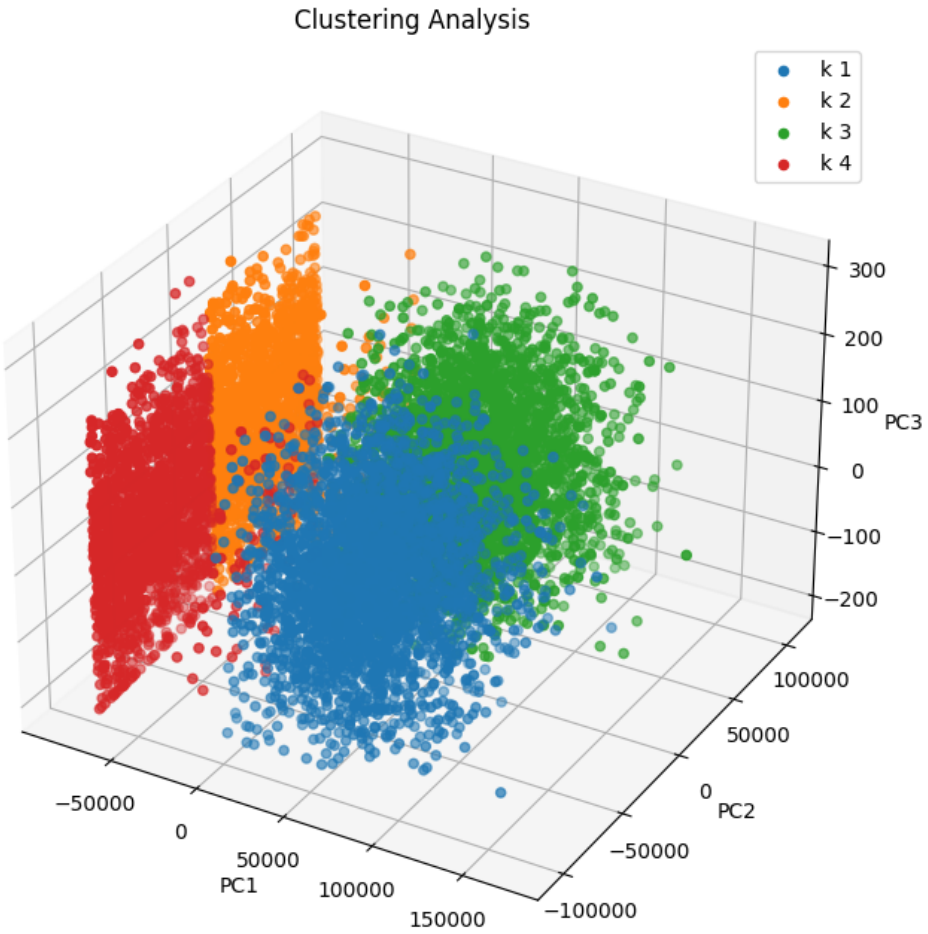


Figura 8 **Clusters**



Anexo D: Modelado predictivo

Figura 9 **Matriz de confusión**

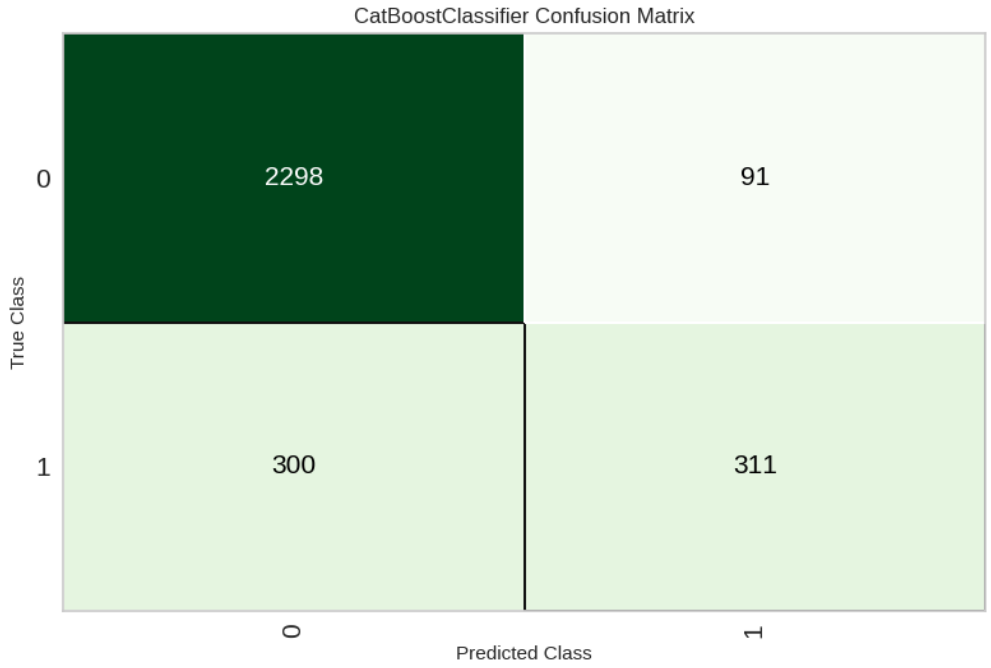
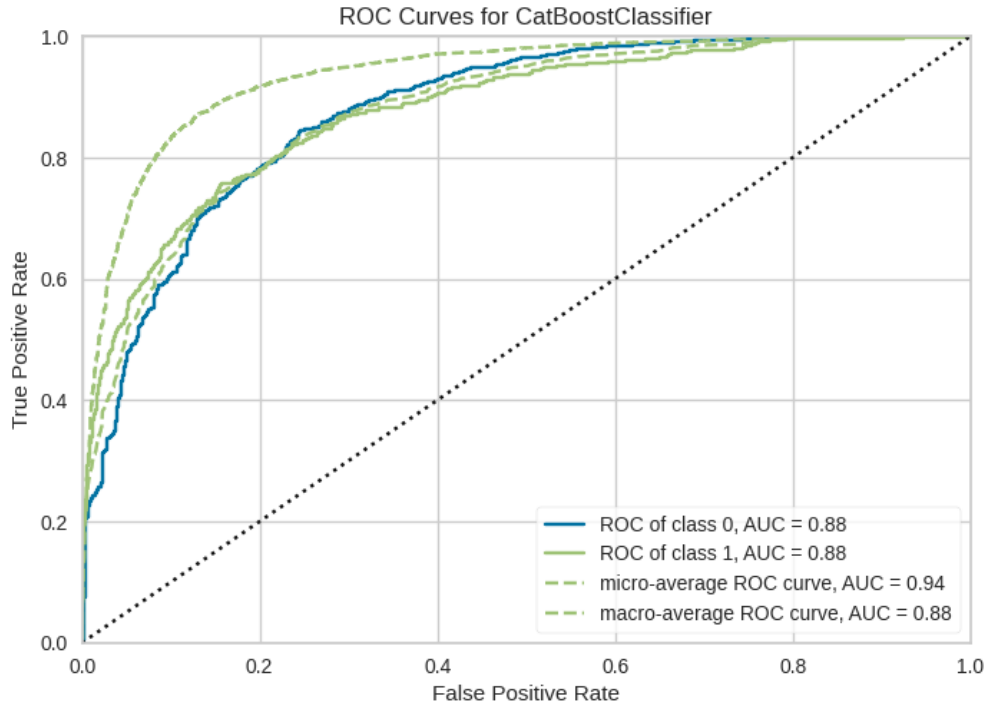


Figura 10 **ROC**



Anexo E: Selección de características

Figura 11 **Importancia de características**

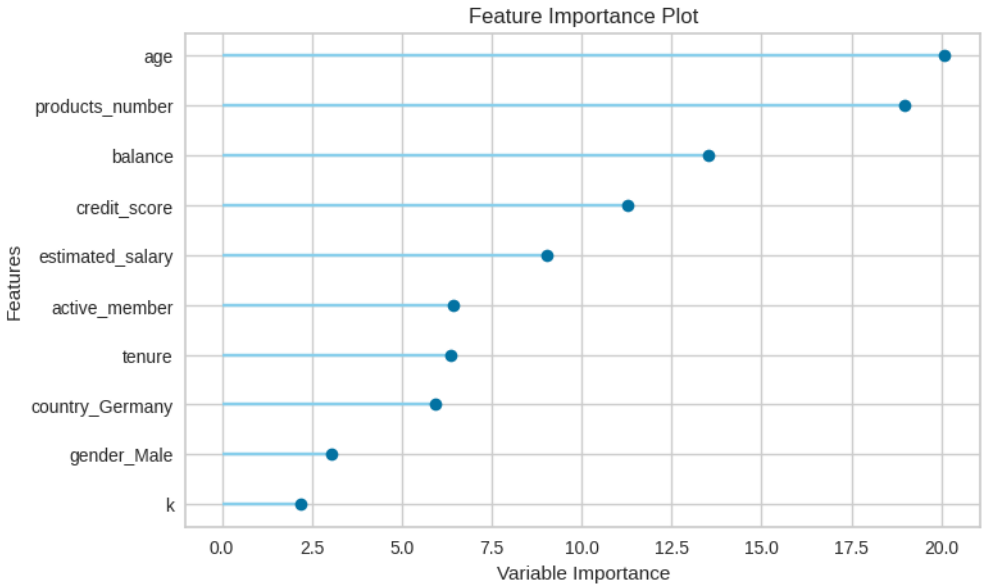
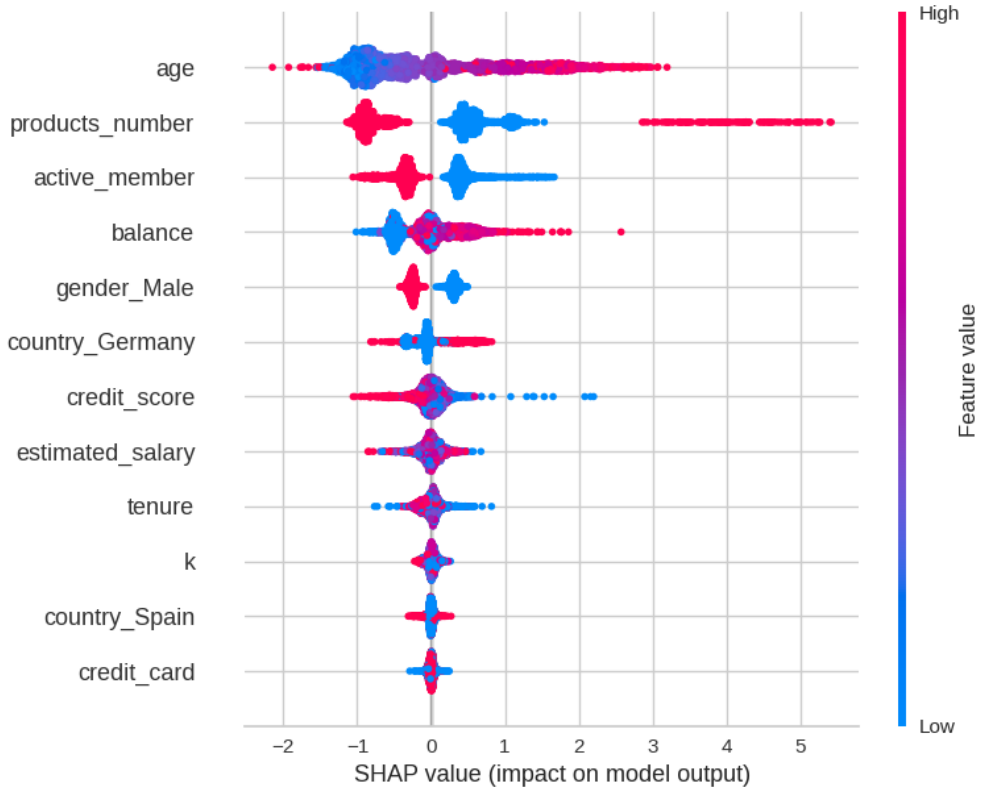


Figura 12 **SHAP**



Anexo F: Optimización

Figura 13 Umbral de probabilidad

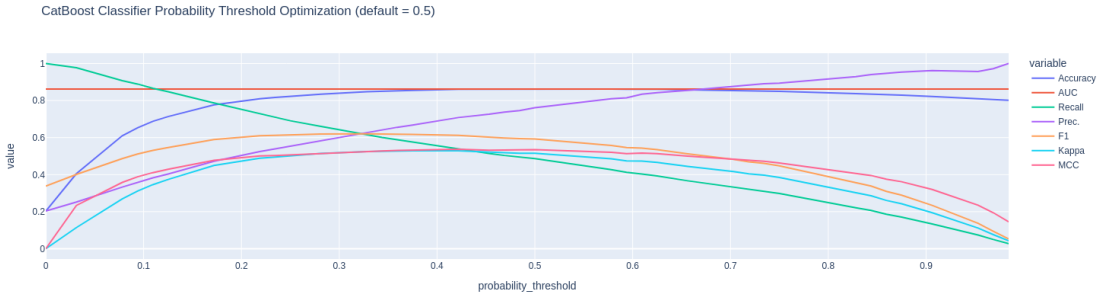
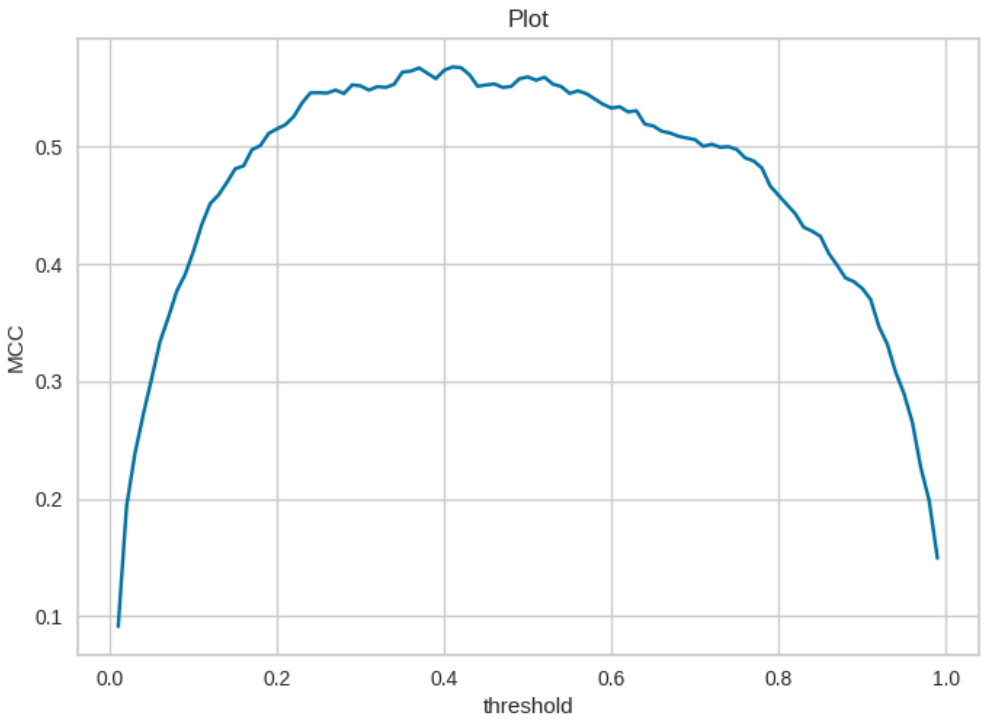


Figura 14 Optimización de MCC



Referencias

[1] Smith, J., *Strategies for Customer Retention in Banking*, *Journal of Banking and Finance*, 2020.

[2] Jones, A., *Predictive Analytics in the Banking Industry*, *International Journal of Finance and Economics*, 2019.

[3] kaggle.com, *Bank Customer Churn Prediction*, *Kaggle*, `https://www.kaggle.com/janiobachmann/bank-customer-churn-prediction`, Consultado en Noviembre 2023.

[4] Doe, M., *Advancements in Data-Driven Customer Retention*, *Journal of Data Science Applications*, 2021.

[5] Chen, S., *Patterns in Bank Customer Data*, *Data Science Journal*,

[6] Wang, Y., *Correlation Analysis in Customer Retention Studies*, *International Journal of Data Science*,

[7] Chen, Z., *Comparative Analysis of Clustering Methods in Customer Segmentation*, *Journal of Business Analytics*,

[8] Ketchen, D. J., Shook, C. L. (1996). *The application of cluster analysis in strategic management research: An analysis and critique. Strategic Management Journal*, 17(6), 441-458.

[9] Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics*, 20, 53-65.

[10] Krasny, F., *PyCaret: An Open Source, Low-Code Machine Learning Library in Python*, *Journal of Open Source Software*,

[11] Chicco, D., Jurman, G. (2017). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics*, 18(Suppl 5), 1-13.

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems*, 31.

[13] Dorogush, A. V., Ershov, V., Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.*

[14] Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems*, 30.

[15] SHAP Developers. (2023). *SHAP: SHapley Additive exPlanations. https://github.com/slundberg/shap.*

[16] Bergstra, J., *Random Search for Hyper-Parameter Optimization*, *Journal of Machine Learning Research*,

[17] Lee, K., *Predictive Models for Customer Churn in Banking*, *Journal of Predictive Analytics*,

[18] Gupta, R., *Strategic Customer Retention in the Banking Sector*, *International Journal of Bank Marketing*,

[19] Johnson, M., *Ethical Implications of Data Privacy in Predictive Analytics*, *Journal of Business Ethics*,

[20] García, A., *Challenges and Opportunities in Interpretable Machine Learning*, *Nature Machine Intelligence*,

[21] Wang, L., *Data Quality Issues in Predictive Modeling*, *Journal of Data Science*,

[22] García, R., *Descriptive Analysis Techniques in Data Science*, *Journal of Data Exploration*,

[23] Wong, L., *Data Visualization for Descriptive Analysis*, *Journal of Visualization*,

[24] Chen, Z., *Pattern Recognition Techniques in Customer Segmentation*, *Journal of Business Analytics*,

[25] Gupta, R., *Patterns in Predictive Models for Customer Churn*, *Expert Systems with Applications*,

[26] Jain, A., *Cluster Validity Measures for Customer Segmentation*, *Pattern Recognition*,

[27] Wang, L., *Applications of Clustering in Customer Relationship Management*, *Expert Systems with Applications*,

[28] Hastie, T., *Elements of Statistical Learning*, *Springer*,