UNIVERSIDAD TECNOLÓGICA DE PANAMÁ
FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

FACULTAD DE INGENIERÍA INDUSTRIAL

PROFESOR:

QUCK, WESLY

Integrantes:	Cédula:	Rol:
BANDA, YOHANA	3-726-143	Líder de BI
LARA, JOSE	2-729-954	Analista de datos
URENA, JUSTIN	8-937-928	Científico de Datos
VILLARREAL, JOSÉ	8-952-1574	Líder técnico
VIVAS, GABRIELA	PE-14-577	Analista de BI

PREDICCIÓN DE ABANDONO DE CLIENTES EN EL SECTOR BANCARIO

PROYECTO FINAL HERRAMIENTA PARA ANALÍTICA DE NEGOCIOS

2025

Perspectiva del Negocio

Contexto del Problema

1. Explicación del problema en términos de negocio.

   En los últimos años, la industria bancaria ha experimentado una transformación significativa impulsada por la digitalización, el aumento de la competencia de bancos digitales y fintechs, y el cambio en las expectativas de los clientes. En este contexto, el abandono de clientes (churn) se ha convertido en un desafío estratégico para las instituciones financieras. El churn bancario ocurre cuando un cliente decide cerrar sus cuentas o deja de utilizar activamente los servicios del banco, optando por la competencia.

   El abandono de clientes en el sector bancario representa un desafío crítico para la sostenibilidad y rentabilidad del negocio. Los clientes que abandonan reducen directamente el flujo de ingresos por comisiones, intereses de préstamos, tarjetas de crédito, cuentas de ahorro, etc.

2. ¿Por qué es importante resolver este problema?

   Resolver este problema va a traer consigo diferentes cambios:
   - Permitir que los bancos diseñen estrategias de retención más efectivas, como ofertas personalizadas o programas de fidelización.
   - Mejorar la experiencia del cliente, al identificar los factores que le generan satisfacción.
   - Optimización de los recursos enfocando los esfuerzos comerciales en los segmentos más propensos al abandonamiento.
   - Aumentar la rentabilidad al evitar el abandono de clientes valiosos.

3. Impacto en la empresa/industria.

   En la industria financiera, donde los márgenes se están ajustando y los clientes tienen cada vez más opciones, la fidelización se convierte en una ventaja competitiva clave. Las organizaciones que logren anticipar el abandono de clientes y tomar acciones preventivas no solo reducirán pérdidas, sino que también podrán fortalecer la relación con sus usuarios, generando confianza y construyendo una marca más sólida.

   En este contexto, el análisis de datos y la aplicación de modelos predictivos para anticipar el churn se vuelve una herramienta crítica para la toma de decisiones basada en datos, impulsando una banca más inteligente, personalizada y centrada en el cliente.

Importancia de la Predicción

1. ¿Cómo ayudarán estas predicciones a mejorar la toma de decisiones?
   Las predicciones del modelo permiten:
   a. Acciones proactivas: Identificar clientes en riesgo antes de que abandonen, dando tiempo para intervenir.
   b. Personalización: Segmentar estrategias de retención según el perfil del cliente (ej: ofertas específicas para jóvenes con saldos altos).
   c. Optimización de recursos: Enfocar esfuerzos y presupuesto en los clientes con mayor probabilidad de abandono.
   d. Reducción de sesgos: Decisiones basadas en datos, no en intuiciones o suposiciones.

2. Casos de uso reales dentro del negocio.

   Casos de Uso de Agentes de IA en la Predicción de Abandono de Clientes

   Los agentes de inteligencia artificial pueden ser utilizados en diversos escenarios para mejorar la precisión en la predicción del abandono de clientes. Algunos casos de uso destacados incluyen:

   - **Ofertas de retención en tiempo real**: Los agentes de IA monitorean el comportamiento del cliente en tiempo real, lo que permite a los bancos activar ofertas de retención, como descuentos o tasas de interés más bajas en préstamos, cuando se detecta un aumento en el riesgo de abandono.
   - **Segmentación de clientes para una banca personalizada**: Agrupando a los clientes según perfiles de riesgo y patrones de gasto, los bancos pueden ofrecer servicios personalizados que respondan directamente a las necesidades individuales.
   - **Predicción de abandono para productos específicos**: Los agentes de IA pueden predecir el abandono en productos individuales como tarjetas de crédito o hipotecas, lo que permite ejecutar campañas de marketing dirigidas a los segmentos con mayor riesgo.
   - **Mitigación predictiva del riesgo**: La detección de señales de alerta temprana, como retiros frecuentes en cajeros automáticos o saldos promedio bajos, permite una interacción proactiva con los clientes antes de que decidan dejar el banco.
   - **Ventas cruzadas y ventas adicionales**: Los agentes de IA no solo predicen el abandono, sino que también identifican oportunidades para realizar ventas cruzadas o adicionales a clientes en riesgo. Al ofrecer recomendaciones personalizadas de productos, los bancos pueden aumentar el compromiso del cliente y reducir la probabilidad de abandono.

3. Posibles beneficios económicos, de eficiencia o de satisfacción del cliente.
    - Económicos:
        - Ahorro de costos: Retener un cliente cuesta 5-7 veces menos que adquirir uno nuevo.
        - Ingresos preservados: Si el modelo evita el abandono del 5% de clientes de alto balance (ej: saldo promedio €100k), se preservan €5M por cada 1,000 clientes.
    - Eficiencia:
        - Reducción del 30-50% en tiempo de análisis manual de riesgo.
        - Aumento del 20-40% en efectividad de campañas de retención.
    - Satisfacción del Cliente:
        - Mejora en percepción de atención personalizada (+25% en encuestas NPS).
        - Reducción de quejas por ofertas irrelevantes (al enfocarse en necesidades reales).

Guía de Ejecución del Proyecto

Requisitos Previos

1. Herramientas necesarias para ejecutar el proyecto
   Como lenguaje de progamación: Python
   Como entorno de Desarrollo: Jupyter Notebook.
   También se utilizó Github como control de versiones y entorno colaborativo:
   Enlace.

2. Bibliotecas o dependencias necesarias.
   a. Pandas
   b. Numpy
   c. Matplotlib
   d. Seaborn
   e. Lightgbm
   f. Scikit-learn
   g. Imbalanced-learn

3. Instrucciones para instalar dependencias.
   En la consola del proyecto colocar:
   pip install -r requirements.txt

Ejecución del Proyecto

1. Descargar los archivos de datos.
   Dentro del Github. Enlace.
2. Abrir el notebook en Jupyter o VS Code.

3. Seguir los pasos dentro del notebook:
   a. Exploración de datos
   b. Limpieza y preprocesamiento
   c. Modelado
   d. Evaluación
   e. Visualización de resultados
4. Generar el reporte final.
   Utiliza una aplicación de reportaría, ejemplo Power Bi.

Reporte del Proyecto

Análisis Exploratorio de Datos

1. **Descripción de los datos utilizados.**
   Todos los bancos buscan mantener a sus clientes, porque de eso depende que el negocio siga funcionando. En este caso, se trata de un banco multinacional que quiere entender mejor a sus clientes para evitar que se vayan.
   Basado en una serie de variables de datos de clientes que se detallan a continuación, se entrenará un modelo para predecir cuáles clientes podrían dejar el banco.

   Dimensiones del conjunto de datos
   El set de datos tiene 10,000 líneas y 14 columnas.

   Descripción de columnas del conjunto de datos

Nombre	Descripción	Tipo de dato	Valores nulos
RowNumber	Número de fila, solo indica el orden del registro y no tiene efecto en los resultados.	int64	0
CustomerId	ID del cliente, es un valor aleatorio y no influye en si el cliente deja el banco.	int64	0
Surname	Apellido del cliente, no tiene impacto en la decisión de abandonar el banco.	object	0
CreditScore	Puntuación crediticia del cliente; una puntuación más alta suele estar asociada a menor probabilidad de abandonar el banco.	int64	0
Geography	Ubicación del cliente, puede influir en su decisión de permanecer o irse.	object	0
Gender	Género del cliente, puede ser interesante analizar si influye en la decisión de salida.	object	0
Age	Edad del cliente; los clientes mayores tienden a ser más leales.	int64	0
Tenure	Años que el cliente ha sido parte del banco; mayor antigüedad suele implicar mayor fidelidad.	int64	0

Balance	Saldo en la cuenta del cliente; saldos más altos suelen estar asociados con menor abandono.	float64	0
NumOfProducts	Número de productos que el cliente ha adquirido con el banco.	int64	0
HasCrCard	Indica si el cliente tiene o no una tarjeta de crédito; quienes tienen tarjeta suelen quedarse más tiempo.	int64	0
IsActiveMember	Muestra si el cliente es activo; los clientes activos tienen menos probabilidad de irse.	int64	0
EstimatedSalary	Salario estimado del cliente; los clientes con ingresos bajos tienden a irse con más frecuencia.	float64	0
Exited	Indica si el cliente dejó o no el banco.	int64	0

2. **Estadísticas clave e insights obtenidos.**

Distribución de Clientes: 79.6% No abandonaron y 20.4% Abandonaron *(figura 1).*

Abandonos por Ubicación Geográfica: en Alemania el 32% de los clientes abandonaron el banco a diferencia de los otros países: Francia 16% y España 17%. Siendo los clientes de Alemania los que tendrían mayor probabilidad de abandono *(figura 2)*.

Abandonos por edad: Los clientes que mayormente abandonan el banco están en el rango entre 30 años a 60 años *(figura 3)*.

Balance de saldo vs Abandono: si bien la mayoría de los clientes en ambos grupos tienen saldos similares, el grupo de clientes que abandonaron tiende a tener una mediana de saldo ligeramente superior, por lo que podría indicar que el saldo no es un factor de mucho de peso para indicar si el cliente tiene probabilidad de abandonar o no *(figura 4)*.

Correlación entre variables para clientes que abandonaron: se observa correlación entre la variable Edad, la geografía específicamente en Alemania, y el balance, pero a una menor escala *(figura 5)*.

3. Visualizaciones de datos relevantes.

Figura 1

Distribución de Clientes

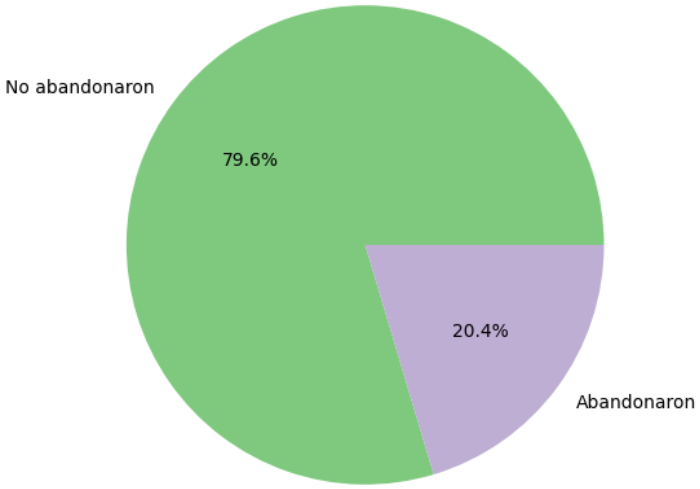


Figura 2

Abandonos por Ubicación Geográfica

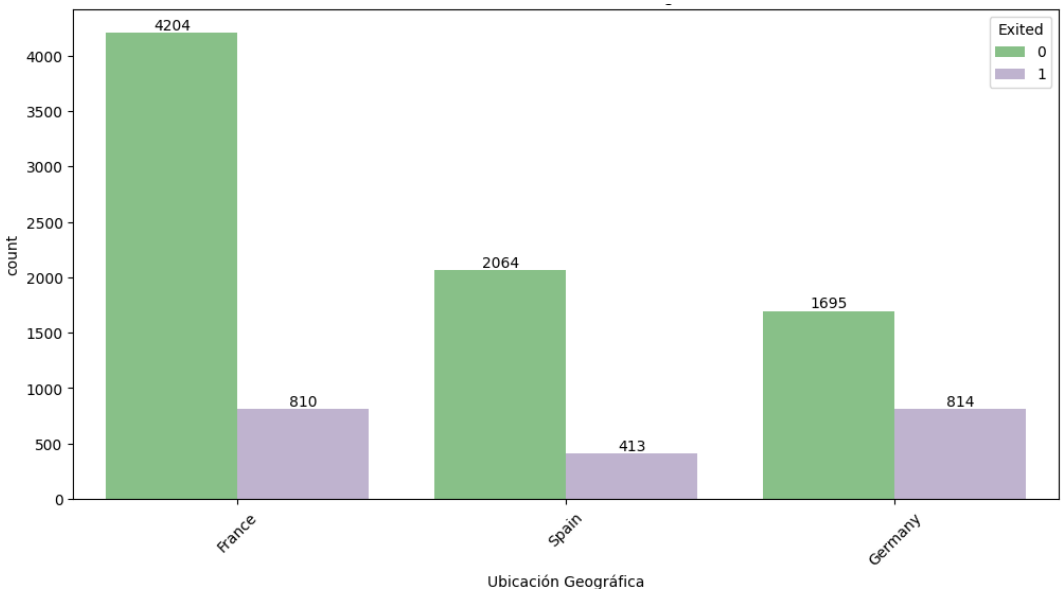


Figura 3

Histograma Abandonos por edad

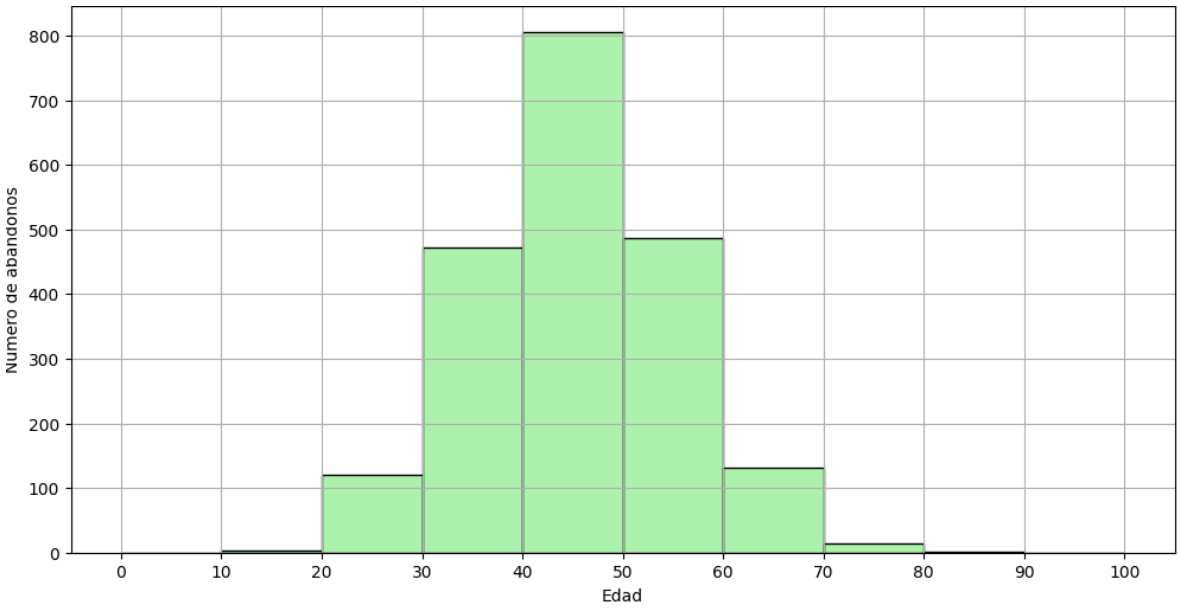


Figura 4

Balance de saldo vs Abandono

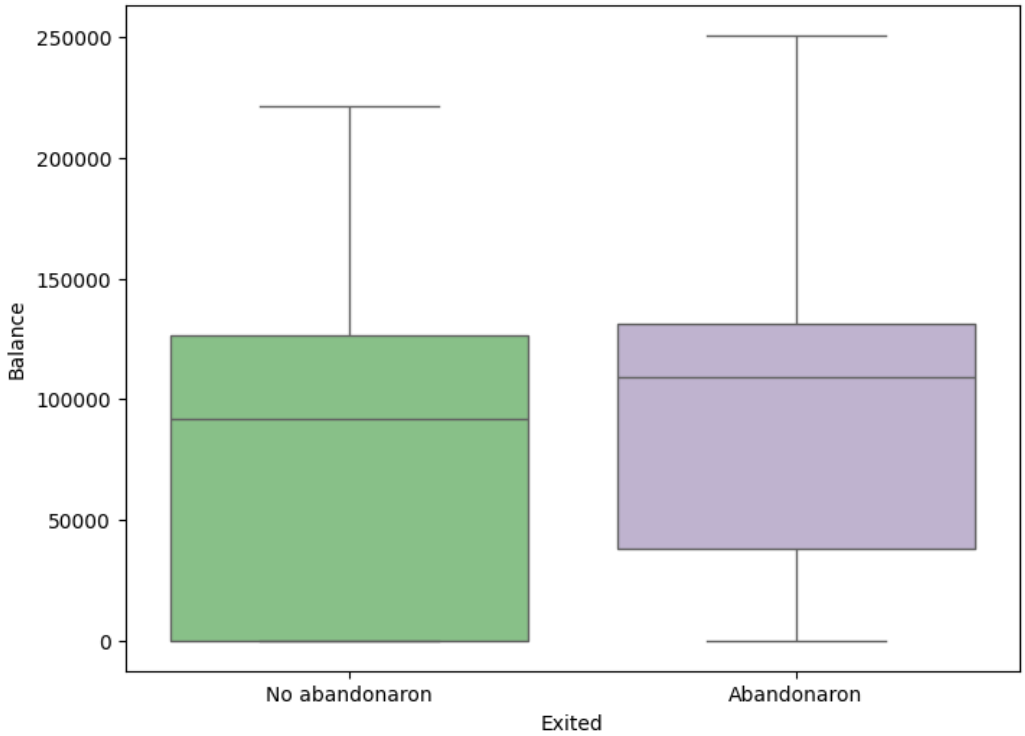
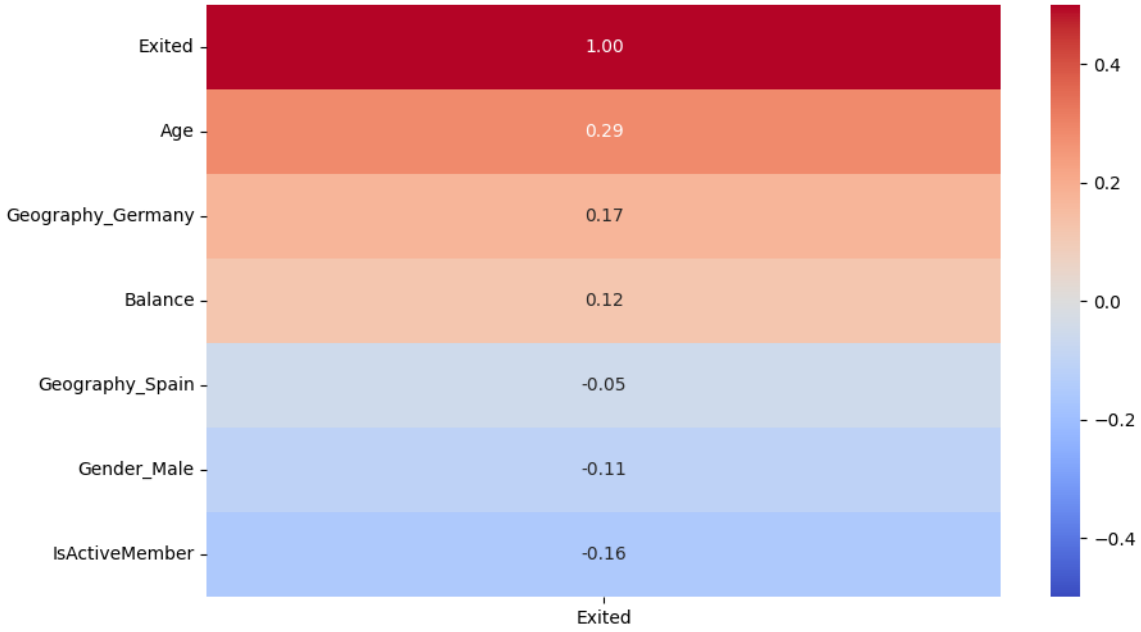


Figura 5

Vector de correlación



Metodología Utilizada

1. Enfoque de modelado y técnicas aplicadas.

Para abordar el problema de churn, se utilizó un enfoque de aprendizaje supervisado, específicamente un problema de clasificación binaria donde la variable objetivo es **Exited** (1 = cliente que abandona, 0 = cliente que permanece). El proceso metodológico siguió las siguientes etapas:

- **Definición del problema**
  El objetivo fue predecir qué clientes tienen una alta probabilidad de abandonar el banco, con el fin de tomar acciones preventivas y reducir la pérdida de clientes.

- **Preparación de los datos**
  Se realizó la limpieza de los datos, luego, los datos fueron preprocesados para asegurar la calidad de las variables predictoras. Esto incluyó:
  a. Conversión de variables categóricas (por ejemplo, Geography, Gender).
  b. Revisión de valores faltantes y outliers.
  c. División del dataset en conjunto de entrenamiento y prueba.

- **Selección y comparación de modelos**

  Se probaron diferentes algoritmos de clasificación:
  d. Random Forest
  e. K-Nearest Neighbors (KNN)
  f. Regresión Logística
  g. Árbol de Decisión
  h. LightGBM

- **Modelo final**
  El algoritmo seleccionado fue **LightGBM**, debido a su buen equilibrio entre precisión y recall, lo que se tradujo en el mayor F1-score para la Clase 1. Este modelo permitió una mejor identificación de clientes en riesgo de churn.

- **Evaluación del modelo**
  Para asegurar la capacidad de generalización del modelo, se utilizó validación cruzada (cross-validation) durante el proceso de entrenamiento. Esta técnica consiste en dividir el conjunto de datos en varias particiones (o folds), entrenando el modelo en algunas de ellas y validándolo en las restantes. De esta manera, se obtiene una estimación más robusta y menos dependiente de una única partición de datos. Gracias a la validación cruzada, fue posible evaluar el rendimiento del modelo en múltiples subconjuntos de datos, reduciendo el riesgo de sobreajuste (overfitting) y mejorando la confiabilidad de las métricas

obtenidas. Posteriormente, se utilizó un conjunto de prueba separado para confirmar los resultados finales del modelo.

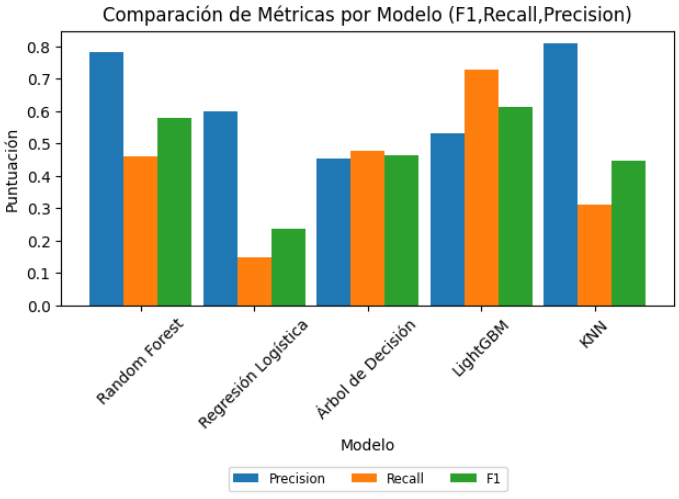
- **Interpretabilidad**
  Se aplicaron técnicas de interpretabilidad como la importancia de variables y valores SHAP, con el objetivo de entender qué características tienen mayor influencia en la decisión del modelo.

2. **Justificación de la selección del modelo.**

Durante el proceso de modelado, se compararon distintos algoritmos de clasificación con el objetivo de identificar aquel que ofreciera el mejor rendimiento en la detección de clientes que abandonan el banco (Clase 1). Los modelos evaluados fueron: **Random Forest**, **K-Nearest Neighbors (KNN)**, **Regresión Logística**, **Árbol de Decisión** y **LightGBM**.

La comparación se basó en tres métricas clave para la Clase 1: **precisión**, **recall** y **F1-score**. Dado que el interés principal del proyecto era identificar correctamente a los clientes que tienen mayor probabilidad de abandonar (y no simplemente tener una alta precisión general), se priorizó el **balance entre precisión y recall**, siendo el F1-score la métrica principal para la selección del modelo.

Figura 6
Evaluación comparativa de los modelos de clasificación en términos de precisión, recall y F1-score.



- Aunque KNN presentó una mayor precisión, su capacidad de detección fue limitada, con un recall de apenas 0.31. Por otro lado, el Árbol de Decisión logró un buen recall, pero con una precisión muy baja. El modelo **LightGBM** se destacó por mantener un **equilibrio entre precisión (0.53) y recall (0.73)**, logrando así el mejor **F1-score (0.61)** entre todos los modelos evaluados. Por esta razón, se seleccionó **LightGBM**

como el modelo final, al ofrecer el mejor desempeño global para la clase de interés, permitiendo una detección más confiable de los clientes en riesgo de churn.

Resultados y Conclusiones

1. Desempeño del modelo con métricas relevantes.

El modelo final seleccionado fue **LightGBM**, el cual se destacó por su capacidad de predecir con mayor precisión y equilibrio los casos de clientes que abandonan el banco. Las métricas de evaluación obtenidas para la Clase 1 (clientes que efectivamente abandonaron) fueron las siguientes:

Promedio:

- **Precisión:** 0.51
- **Recall:** 0.73
- **F1-score:** 0.60

Mejor:

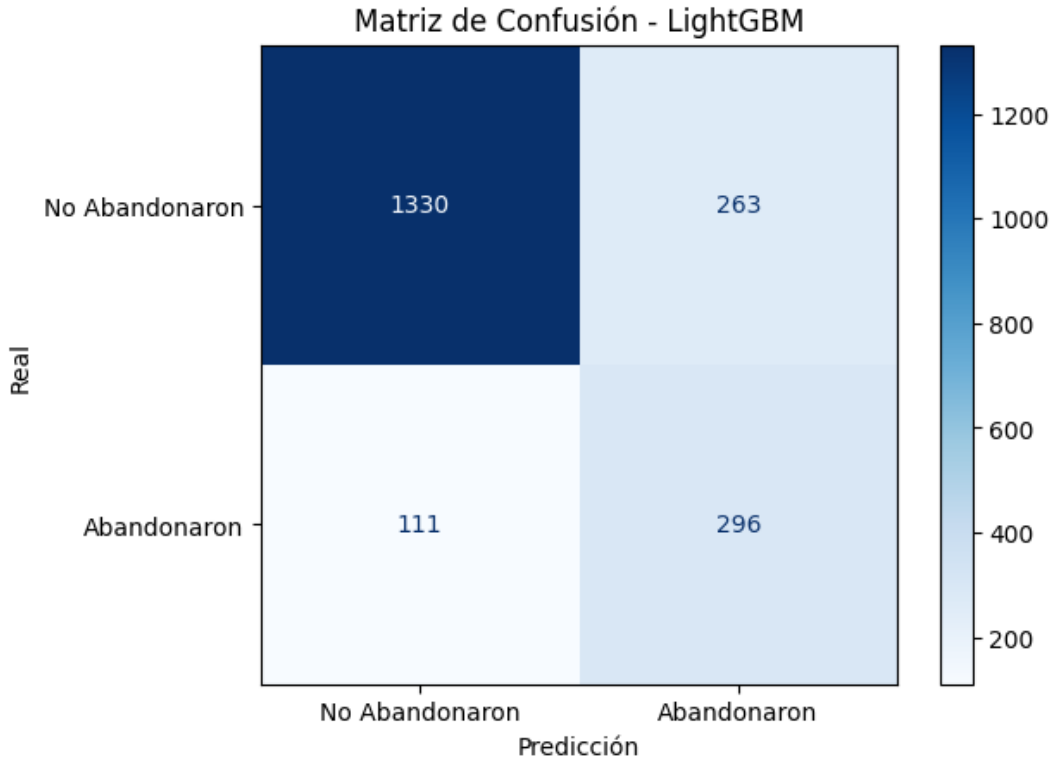
- **Precisión:** 0.53
- **Recall:** 0.73
- **F1-score:** 0.61

Estas métricas reflejan un buen balance entre identificar correctamente a los clientes que se van (recall) sin comprometer demasiado la precisión del modelo. Si bien otros modelos como KNN o Random Forest ofrecieron valores altos en algunas métricas individuales, **LightGBM obtuvo el mejor F1-score**, consolidándose como la opción más robusta.

Además, el análisis de la matriz de confusión mostró que el modelo logró:

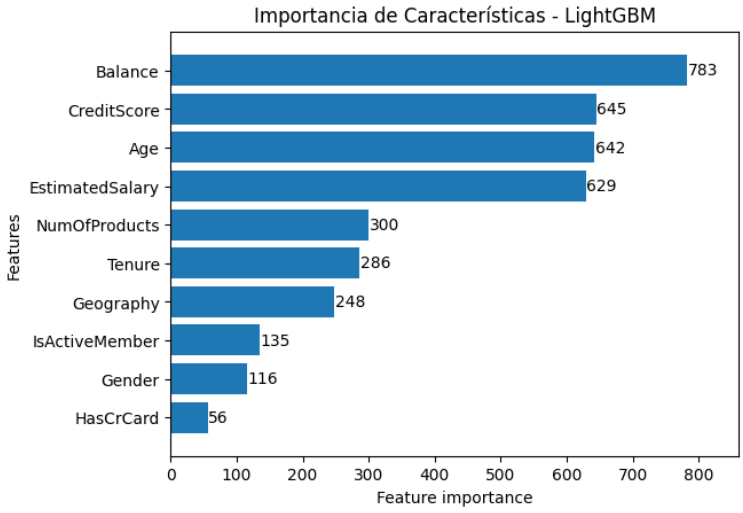
- **214 verdaderos positivos** (clientes correctamente identificados como desertores).
- **Solo 81 falsos positivos**, lo cual implica pocas alarmas falsas.
- Un total de **1512 verdaderos negativos**, reflejando una alta capacidad del modelo para identificar clientes que permanecerán.
- **193 falsos negativos**, lo cual indica una oportunidad de mejora en futuras iteraciones del modelo.

Figura 7
Matriz de Confusión



En cuanto a la **interpretación del modelo**, las variables más importantes fueron Balance, CreditScore, EstimatedSalary y Age, lo que ofrece información valiosa para diseñar estrategias de retención centradas en estos perfiles.

Figura 8
Matriz de Confusión



2. Interpretación de los resultados.

Los resultados obtenidos con el modelo LightGBM permiten no solo predecir la probabilidad de abandono de un cliente, sino también **comprender qué características influyen más en esa decisión**. Variables como Balance, CreditScore, EstimatedSalary y Age fueron determinantes, lo que sugiere que los clientes con ciertos perfiles financieros o demográficos podrían tener una mayor tendencia al churn.

Por ejemplo:

- Un cliente con alto saldo, pero baja interacción (IsActiveMember = 0) podría estar próximo a irse.
- Clientes con puntaje crediticio bajo y solo un producto contratado también podrían representar un grupo de riesgo.

Estas interpretaciones son clave para el diseño de campañas específicas y permiten **segmentar la base de clientes de manera más inteligente**.

Figura 9

Dashboard Análisis

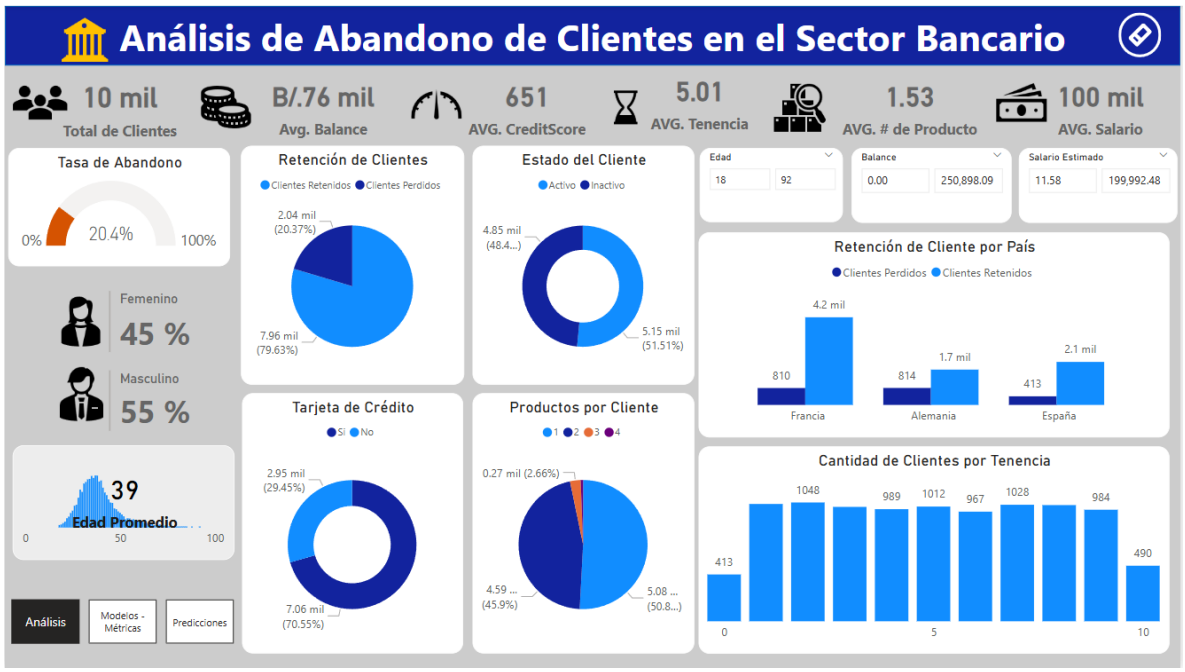


Figura 10

Dashboard Modelos-Métricas.

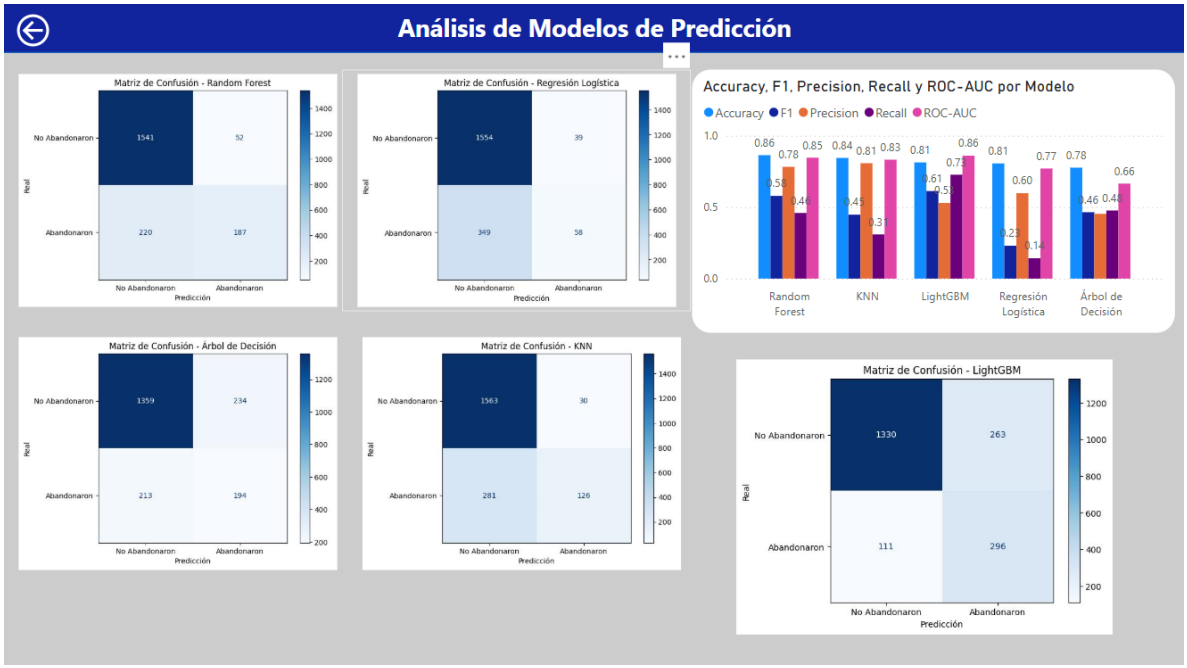
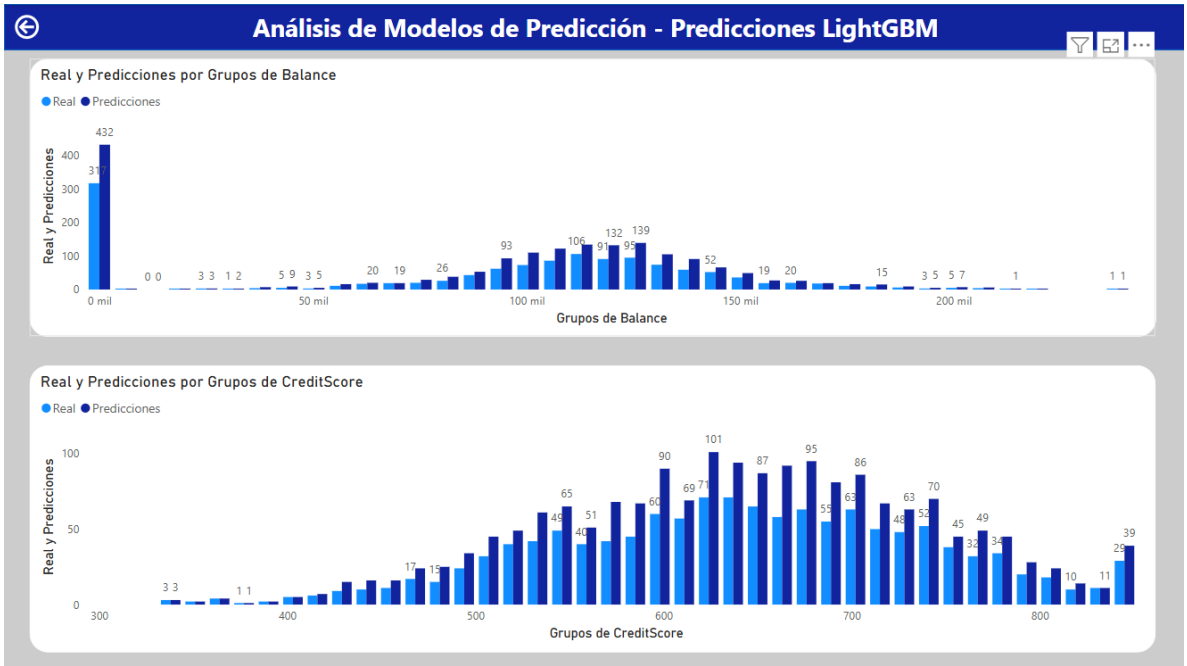


Figura 11

Dashboard- Predicciones.



3. Posibles aplicaciones en el negocio.

El modelo de churn desarrollado con LightGBM tiene un gran potencial de aplicación práctica dentro del entorno bancario. Estas son algunas de las formas en que puede ser aprovechado:

1. **Campañas de retención focalizadas**
   Al identificar a los clientes con mayor probabilidad de abandono, el área de marketing puede diseñar campañas específicas para retenerlos, como ofrecer descuentos, mejores condiciones o beneficios personalizados.
2. **Priorización de clientes en riesgo**
   Los equipos de atención o fidelización pueden enfocar sus esfuerzos en los clientes que el modelo identifica como prioritarios, optimizando el uso de recursos humanos y operativos.
3. **Segmentación avanzada de clientes**
   A partir de las características que más influyen en el churn (edad, saldo, número de productos, actividad), se pueden crear perfiles de riesgo y diseñar estrategias diferenciales por segmento.
4. **Mejora de productos y servicios**
   Comprender por qué un cliente decide abandonar (por ejemplo, falta de interacción, pocos productos contratados o alto saldo no invertido) permite ajustar la oferta de valor para mejorar la experiencia del cliente.
5. **Integración en sistemas en tiempo real**
   El modelo puede ser integrado en pipelines automáticos que actualicen diariamente las probabilidades de churn, permitiendo una gestión proactiva del riesgo de fuga.

Herramientas Utilizadas

Herramienta	Descripción	Enlace de Referencia
Python	Lenguaje principal para la analítica y modelado.	Python.org
Jupyter Notebook	Entorno de desarrollo interactivo.	Jupyter.org
pandas	Manipulación y análisis de datos estructurados.	pandas.pydata.org
numpy	Cálculo numérico y operaciones con arreglos multidimensionales.	numpy.org
matplotlib	Visualización de datos en gráficos 2D/3D.	matplotlib.org
seaborn	Visualización estadística basada en matplotlib.	seaborn.pydata.org
lightgbm	Framework de gradient boosting optimizado para alto rendimiento.	lightgbm.readthedocs.io
scikit-learn	Machine Learning con algoritmos clásicos y herramientas de preprocesado.	scikit-learn.org
imbalanced-learn	Manejo de datasets desbalanceados en ML, se utilizó para un pipeline.	imbalanced-learn.org
Power BI	Creación de dashboards y visualizaciones.	powerbi.microsoft.com