

DATA SCIENCE I:

FUNDAMENTOS PARA LA

CIENCIA DE DATOS



PROFESOR: GERMÁN RODRIGUEZ

TUTOR: JONATAN CANCHI

COMISIÓN: 61750

ALUMNO: LUCIANO ANDRES LABUDIA

CODERHOUSE

Análisis de Datos para la Predicción de Abandono en Clientes Bancarios

Abstracto con Motivación y Audiencia

La retención de clientes es un desafío crítico para las instituciones financieras en un entorno cada vez más competitivo. Identificar los factores que contribuyen al abandono de clientes (variable: churn) permite a los bancos implementar estrategias proactivas para reducir este fenómeno, generando un impacto directo en la rentabilidad y sostenibilidad del negocio.

Este proyecto se centra en analizar un conjunto de datos de clientes bancarios, utilizando herramientas de machine learning para identificar patrones y predecir el abandono del cliente (churn). La motivación principal es proporcionar una solución basada en datos para mejorar la retención de clientes, optimizar las estrategias de marketing y tomar decisiones informadas sobre la gestión de la cartera de clientes.

Audiencias beneficiadas:

- Equipos de Marketing:** Podrán utilizar los resultados para segmentar clientes en riesgo de abandono y diseñar campañas personalizadas.
- Gerentes de Relación con Clientes:** Se beneficiarán de herramientas para priorizar esfuerzos en la retención de clientes clave para el banco.
- Analistas de Datos:** Contarán con un marco para integrar técnicas de predicción en el flujo de análisis de datos.
- Altos Directivos:** Recibirán información estratégica para tomar decisiones fundamentadas sobre políticas y recursos destinados a mejorar la experiencia del cliente.

El análisis está diseñado para ser claro, accionable y alineado con los objetivos de negocio de una institución bancaria moderna. A través de métricas precisas y visualizaciones comprensibles, este proyecto busca ser una herramienta clave para transformar datos en decisiones y mejorar el negocio.

Preguntas e Hipótesis

Antes de abordar el análisis visual de los datos, definimos una serie de preguntas clave y formulamos hipótesis que serán evaluadas a través de gráficos y estadísticas extraídas del dataset. Este enfoque permite estructurar el análisis y obtener insights claros y objetivos.

Preguntas clave:

- **¿Cuál es la distribución de clientes según su puntuación crediticia (credit_score)?**

Se busca evaluar cómo se distribuyen los clientes en función de su puntuación crediticia y analizar posibles agrupaciones o extremos que puedan influir en el abandono del cliente.

- **¿Existe una relación entre la cantidad de productos contratados (products_number) y la probabilidad de abandono (churn)?**

Se Determinara si la fidelización del cliente está relacionada con el número de productos bancarios adquiridos.

- **¿Los clientes activos (active_member) tienen menor probabilidad de abandonar el banco?**

Se tratara de Identificar si el nivel de actividad de los clientes se asocia con una mayor retención.

- **¿Cómo se distribuye el balance entre clientes que abandonaron y los que permanecieron?**

Entenderemos si el saldo promedio o el saldo nulo tienen un impacto significativo en el abandono del cliente.

- **¿Existe una relación significativa entre la edad y la puntuación crediticia?**

Exploraremos si los clientes de diferentes edades tienden a mostrar patrones similares o distintos en términos de su comportamiento financiero.

- **¿Qué patrones clave de correlación existen entre las variables?**

Evaluar las relaciones entre las diferentes variables del dataset mediante un mapa de calor (heatmap) para identificar posibles indicadores de abandono.

Hipótesis

- **H1: Los clientes con una puntuación crediticia más baja tienen mayor tendencia al abandono (churn).**

Una baja puntuación crediticia puede estar asociada a perfiles financieros menos estables, lo que podría correlacionarse con tasas de abandono más altas.

- **H2: Los clientes que utilizan más productos del banco tienden a quedarse más tiempo.**

Al adquirir múltiples productos, los clientes refuerzan su vínculo con la institución financiera, disminuyendo la probabilidad de churn.

- **H3: Los clientes activos tienen una tasa de churn más baja que los no activos.**

Razonamiento: Un cliente activo refleja mayor compromiso con el banco, lo que probablemente contribuya a una menor tasa de abandono.

- **H4: Los clientes con saldo nulo o significativamente bajo tienen mayor probabilidad de churn.**

Un balance bajo puede reflejar una relación débil con el banco, lo que podría incentivar el abandono.

- **H5: La edad y la puntuación crediticia están relacionadas, y los clientes más jóvenes podrían mostrar un comportamiento financiero más arriesgado.**

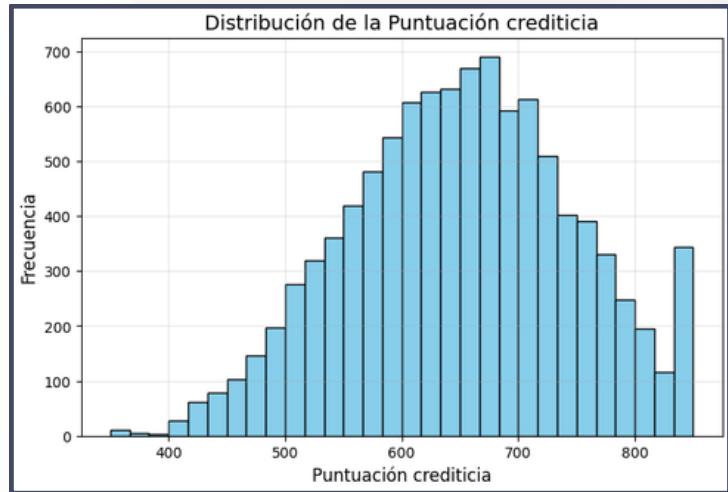
Clientes más jóvenes suelen tener menor experiencia financiera, lo que puede influir en sus patrones de puntuación crediticia.

Gráficos que abordarán las preguntas

- **Distribución del credit_score:** Para observar la dispersión de puntuaciones crediticias entre los clientes.
- **Comparación entre products_number y churn:** Para analizar la relación entre la cantidad de productos contratados y el abandono.
- **Relación entre active_member y churn:** Evaluar si ser un cliente activo tiene un impacto significativo en la retención.
- **Distribución del balance por churn:** Analizar cómo varía el saldo de los clientes entre los que abandonan y los que permanecen.
- **Relación entre edad y credit_score:** Para identificar si existen patrones comunes entre la edad y la estabilidad financiera de los clientes.
- **Heatmap de correlaciones:** Examinar las relaciones más relevantes entre todas las variables del dataset.

Visualizaciones ejecutivas que responden nuestras preguntas

Distribución de la Puntuación Crediticia (credit_score)



El análisis de la puntuación crediticia nos permite entender la distribución general de los clientes según este indicador clave, así como identificar patrones relevantes para la segmentación y la gestión de riesgos.

Distribución general

La puntuación crediticia presenta una distribución asimétrica positiva (hacia la derecha), con la mayoría de los clientes concentrados entre los rangos de 600 y 750 puntos. Este rango sugiere que los clientes tienen, en promedio, un perfil crediticio medio-alto, considerado aceptable por el banco.

Observaciones clave

Extremos de la distribución:

- Puntuaciones bajas (< 500):** Muy pocos clientes se encuentran en este rango, lo que podría indicar que el banco aplica criterios de admisión restrictivos para evitar perfiles de alto riesgo crediticio.
- Puntuaciones altas (> 800):** Existe un pequeño segmento de clientes con puntuaciones excepcionalmente altas, reflejando un historial crediticio excelente.

Pico principal alrededor de 700:

Un pico notable se observa cerca de los **700 puntos**, lo que indica que una gran parte de los clientes tiene un buen perfil crediticio, aunque sin alcanzar niveles sobresalientes.

Disminución en el rango alto:

A partir de los **750 puntos**, la frecuencia de clientes disminuye considerablemente. Esto podría deberse a las políticas de riesgo del banco o a la dificultad inherente para alcanzar puntuaciones más altas.

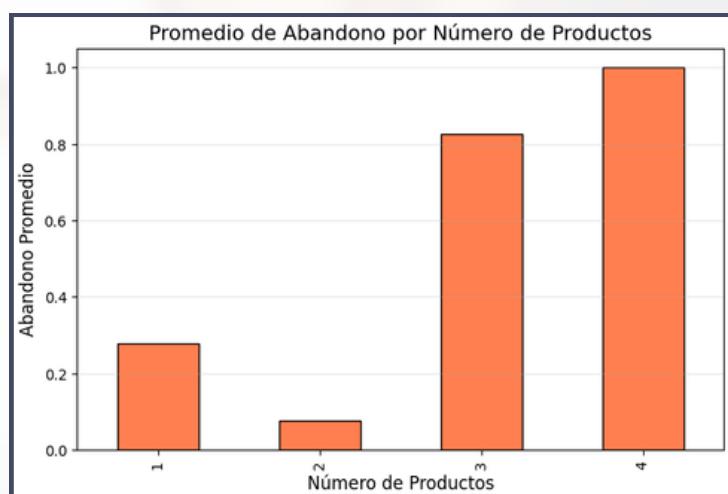
Conclusión y recomendaciones

- La mayoría de los clientes poseen una puntuación crediticia en el rango medio-alto (600-750), lo que los califica como perfiles aceptables o buenos.
- Los clientes con puntuaciones por debajo de 600 podrían ser segmentados como un grupo de riesgo para estrategias de retención o reestructuración de productos financieros.
- Los clientes con puntuaciones superiores a 750 representan un segmento de alto valor. Se recomienda considerar estrategias específicas para este grupo, como la oferta de servicios premium o personalizados, para fortalecer la relación y fidelidad hacia la institución.

Insight clave

Este análisis de distribución sirve como base para correlacionar la puntuación crediticia con otras variables, como el abandono (churn), con el fin de identificar perfiles específicos que podrían tener mayor probabilidad de abandonar el banco.

Distribución y análisis de abandono según cantidad de productos contratados



Clients con 1 producto:

- Tasa de abandono: ~28%
- Estos clientes muestran un bajo promedio de abandono, posiblemente porque mantienen una relación inicial con el banco y están satisfechos con el producto básico.

Clients con 2 productos:

- Tasa de abandono: ~8%
- Este grupo registra el abandono más bajo, indicando que los clientes con dos productos tienden a estar más comprometidos, probablemente debido a una mayor integración con los servicios del banco.

Clients con 3 productos:

- Tasa de abandono: ~82%
- Se observa un aumento significativo en el abandono. Esto podría deberse a una percepción de valor insuficiente de los productos adicionales o a una complejidad mayor en la gestión de los servicios.

Clients con 4 products:

- Tasa de abandono: 100%
- Este grupo refleja un problema crítico, ya que todos los clientes con 4 productos terminan abandonando el banco. Esto podría ser el resultado de insatisfacción generalizada, sobrecarga de servicios o falta de soporte adecuado.

Conclusiones clave:

1. **Relación no lineal:** La tasa de abandono disminuye inicialmente con el número de productos contratados, alcanzando su punto más bajo con 2 productos, pero aumenta drásticamente con 3 o más productos.
2. **Punto óptimo:** Los clientes con 2 productos representan el segmento más estable, sugiriendo un equilibrio entre la integración de servicios y la satisfacción.
3. **Segmentos críticos:** Los clientes con 3 o más productos requieren atención prioritaria, ya que la tasa de abandono en estos casos es alarmantemente alta.

Recomendaciones estratégicas:

- **Optimización de productos adicionales:** Evaluar y ajustar la propuesta de valor para clientes con más de 2 productos, asegurándose de que los beneficios justifiquen la complejidad o costos adicionales.
- **Análisis detallado:** Identificar factores específicos que afectan la experiencia de clientes con 3 o más productos, como problemas operativos o falta de soporte personalizado.
- **Estrategias preventivas:** Implementar programas de fidelización y encuestas para clientes con múltiples productos, con el objetivo de detectar posibles problemas antes de que lleven al abandono.

Relación entre Actividad de los Clientes y Tasa de Abandono



C clientes No Activos:

- Tasa de abandono: 27%
- Los clientes que no están activos tienen una probabilidad significativamente mayor de abandonar el banco.
- La falta de actividad puede estar asociada a una menor interacción, insatisfacción con los servicios ofrecidos o ausencia de incentivos claros para continuar vinculados.

C clientes Activos:

- Tasa de abandono: 14%
- Este grupo presenta una tasa de abandono considerablemente más baja.
- La actividad indica mayor compromiso y satisfacción, posiblemente debido a una participación activa en transacciones, acceso a beneficios o una mejor percepción del valor ofrecido por el banco.

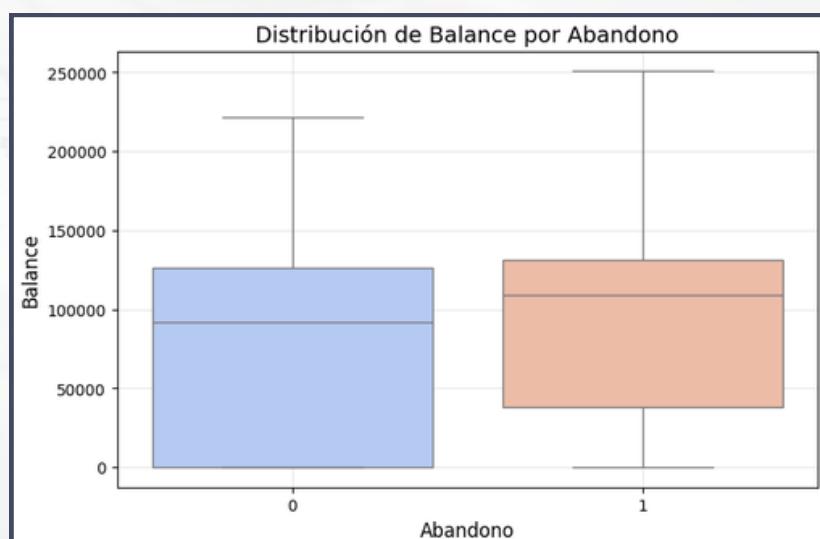
Conclusiones clave:

1. **Impacto de la actividad:** Los clientes activos tienen casi la mitad de probabilidad de abandonar en comparación con los clientes no activos. Esto refuerza la idea de que la interacción regular con el banco es un factor determinante en la retención.
2. **Segmento de riesgo:** Los clientes no activos representan un grupo de mayor vulnerabilidad que requiere atención prioritaria para evitar el abandono.

Recomendaciones estratégicas:

- **Programas de activación:** Diseñar estrategias para reactivar a los clientes inactivos, como promociones personalizadas, campañas de comunicación proactiva y ofertas exclusivas.
- **Fidelización:** Implementar programas de recompensas para clientes activos, incentivando comportamientos como el uso regular de servicios y productos financieros.
- **Análisis de causas de inactividad:** Realizar estudios para identificar barreras específicas que dificulten la participación activa, como problemas de usabilidad o falta de percepción de valor.
- **Seguimiento continuo:** Establecer indicadores de actividad para detectar señales tempranas de inactividad y anticipar acciones preventivas.

Relación entre Balance y Tasa de Abandono



Clients que No Abandonaron (Churn = 0):

- Mediana del balance: Cerca de 90,000.
- Distribución: Amplio rango de valores, desde 0 hasta más de 200,000.
- Se observa una proporción considerable de clientes con balances bajos (cerca de 0), lo que sugiere que el abandono no está únicamente relacionado con balances mínimos.

Clients que Abandonaron (Churn = 1):

- Mediana del balance: Aproximadamente 120,000, superior a la de los clientes que permanecieron.
- Distribución: Amplio rango de valores (0 a 250,000), pero con una mayor concentración de balances altos.
- Esto indica que los clientes con balances elevados tienen más probabilidades de abandonar.

Conclusiones claves:

Relación entre balance y abandono:

- La mediana del balance para los clientes que abandonan es notablemente mayor que la de los que permanecen.
- Este hallazgo sugiere que el abandono podría estar influenciado por expectativas insatisfechas de clientes con mayores balances.

Impacto estratégico:

- La pérdida de clientes con balances altos tiene un impacto directo en la rentabilidad, dado su mayor peso financiero en los ingresos del banco.

Recomendaciones estratégicas:

Estrategias de Retención para Clientes de Balances Altos:

- Beneficios exclusivos: Ofrecer tasas preferenciales, programas de lealtad, asesoramiento personalizado o servicios premium.
- Interacción personalizada: Designar gestores de cuenta para atender las necesidades específicas de estos clientes y fomentar la fidelización.

Segmentación y Campañas Dirigidas:

- Desarrollar campañas específicas para clientes con balances altos, enfocadas en destacar los beneficios únicos que ofrece el banco frente a la competencia.
- Crear programas de valor que refuerzen la percepción positiva de los servicios ofrecidos.

Ánálisis de Factores Asociados al Abandono:

- Evaluar qué otros elementos, además del balance, contribuyen al churn. Por ejemplo, nivel de satisfacción con el servicio, número de productos contratados o experiencia general del cliente.

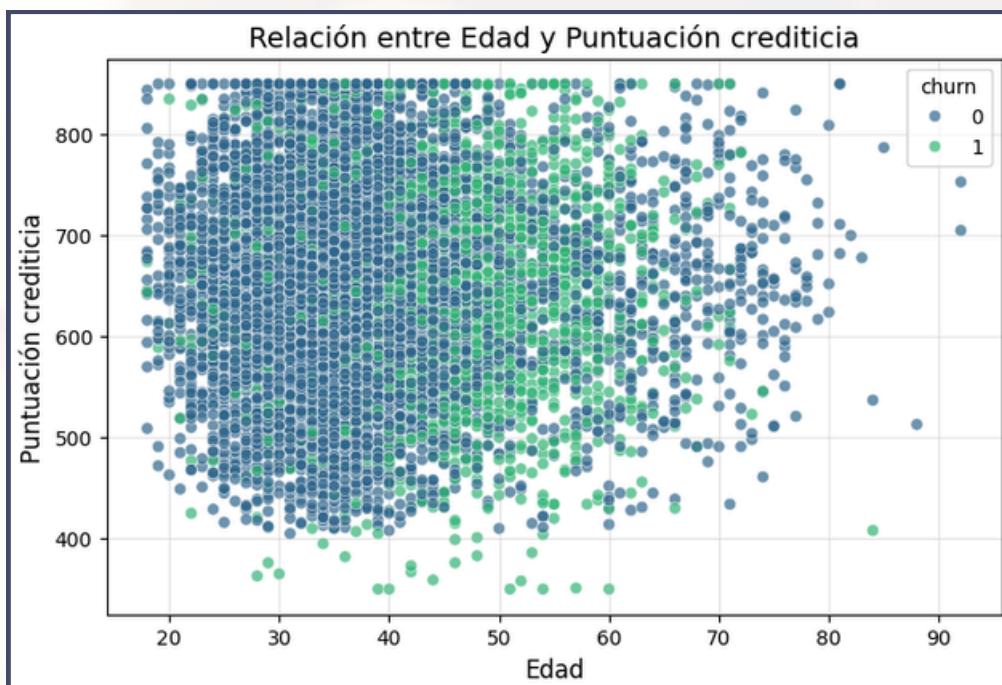
Monitoreo Proactivo:

- Implementar modelos predictivos para identificar clientes con balances altos en riesgo de abandono.
- Realizar encuestas regulares para recopilar feedback y abordar posibles puntos de insatisfacción.

Insight clave:

La mayor probabilidad de abandono entre clientes con balances elevados sugiere que este grupo tiene expectativas específicas que no están siendo satisfechas. Retener a estos clientes es crucial para la sostenibilidad financiera del banco, y requiere una combinación de personalización, beneficios exclusivos y atención prioritaria.

Relación entre Edad, Puntuación Crediticia y Abandono Bancario



Concentración general de datos:

- **Edad:** La mayoría de los clientes tienen entre 20 y 50 años.
- **Puntuación crediticia:** Se concentran entre 600 y 800 puntos, reflejando perfiles crediticios promedio o altos.

Relación con el abandono (Churn):

Clientes que abandonaron (Churn = 1):

- Predominan en edades entre 30 y 60 años.
- Se concentran en puntuaciones crediticias entre 500 y 700 puntos.
- Este patrón sugiere que los clientes de mediana edad con puntuaciones crediticias intermedias son más propensos a abandonar.

Clients en extremos de edad:

Menores de 25 años y mayores de 70 años:

- Representan una menor proporción de clientes.
- Exhiben bajos niveles de abandono, lo que podría indicar menor actividad bancaria o una relación más estable con el banco.

Puntos aislados (outliers):

- Clientes con puntuaciones crediticias muy bajas (< 400) o muy altas (> 900) no muestran un aumento significativo en la tasa de abandono.

Conclusiones clave:

Segmento de mayor riesgo:

Clientes de edad media (30-60 años) y con puntuaciones crediticias media-bajas (500-700 puntos) tienen una mayor probabilidad de abandono.

Menor abandono en extremos de edad:

Grupos de edad joven (< 25 años) o avanzada (> 70 años) presentan menor propensión al abandono, posiblemente debido a menores expectativas o necesidades financieras más específicas y estables.

Recomendaciones estratégicas:

1. Estrategias de Retención para Grupos de Riesgo:

Clientes de mediana edad con puntuaciones medias:

- Implementar programas de fidelización adaptados a sus necesidades, como productos financieros flexibles o asesoramiento personalizado.
- Enfocar campañas de reenganche y atención prioritaria para este segmento.

2. Programas de Incentivos:

- Diseñar incentivos específicos para clientes con puntuaciones crediticias entre 500 y 700, como tasas competitivas o recompensas por mantener o incrementar su actividad bancaria.

3. Segmentación basada en edad y puntuación crediticia:

- Crear segmentos específicos basados en rangos de edad y puntuación crediticia, y desarrollar ofertas diferenciadas para aumentar la retención en cada grupo.

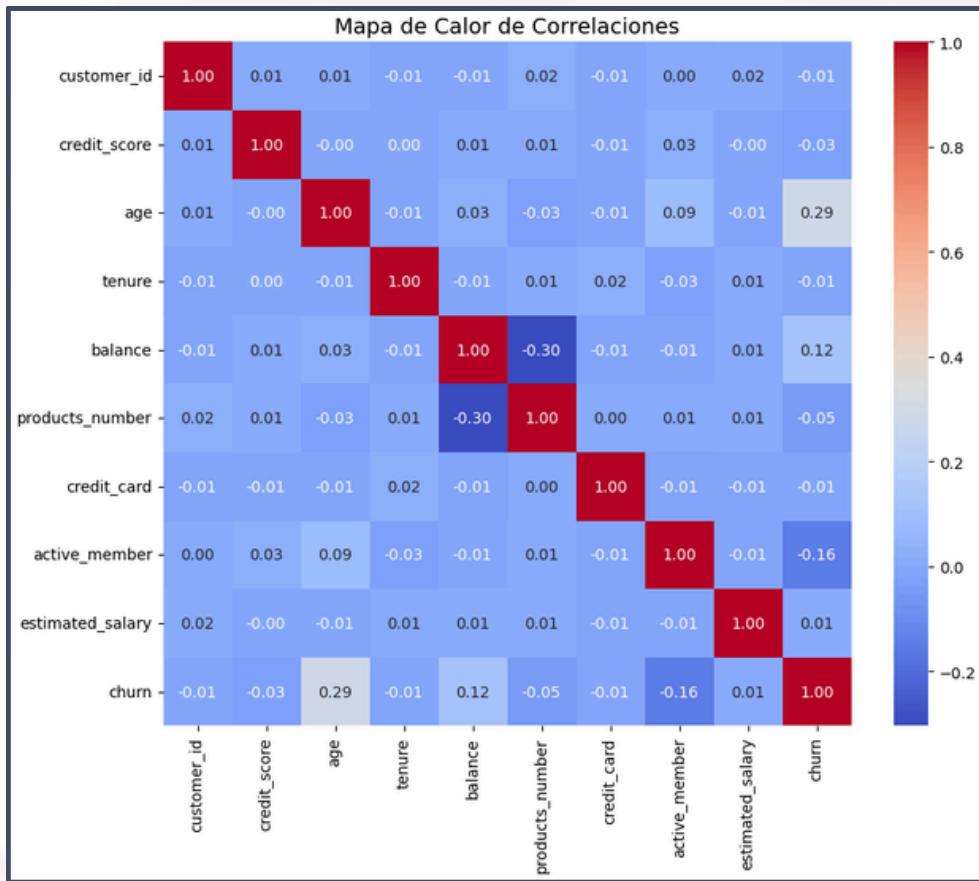
4. Monitoreo Proactivo:

- Establecer sistemas para identificar y anticipar comportamientos de abandono en clientes con edad y puntuaciones crediticias en rangos críticos.

Insight clave:

La edad y la puntuación crediticia son factores determinantes en el análisis de abandono. Identificar y atender a clientes de edad media con puntuaciones media-bajas es esencial para mejorar la retención y optimizar la oferta de productos financieros, evitando pérdidas significativas en este segmento de riesgo.

Análisis de Correlaciones: Variables Numéricas y Abandono



Principales hallazgos:

1. Edad y Abandono (Churn):

- Correlación positiva moderada (0.29):** A medida que la edad aumenta, también lo hace la probabilidad de abandono.
- Insight:** Los clientes mayores representan un grupo con mayor riesgo de churn.

2. Balance y Abandono:

- Correlación positiva débil (0.12):** Los clientes con balances más altos tienen una leve tendencia a abandonar.
- Insight:** Este comportamiento refuerza la necesidad de estrategias específicas para clientes con balances elevados.

3. Actividad del Cliente (Active Member) y Abandono:

- Correlación negativa débil (-0.16):** Los clientes activos tienen una menor propensión a abandonar en comparación con los inactivos.
- Insight:** La actividad del cliente es un factor relevante para la retención.

4. Productos Contratados y Balance:

- Correlación negativa moderada (-0.30):** A mayor cantidad de productos contratados, menor es el balance promedio del cliente.
- Insight:** Esto podría sugerir que los clientes con balances más bajos tienden a diversificar más sus productos contratados.

Conclusiones clave:

1. Edad y Balance como Predictores de Churn:

- Los clientes mayores y aquellos con balances elevados son más propensos a abandonar, lo que los convierte en segmentos clave para intervenciones específicas.

2. Actividad como Factor de Retención:

- La actividad del cliente tiene un impacto relevante en la reducción del abandono. Fomentar interacciones regulares puede ayudar a disminuir la tasa de abandono.

3. Relación entre Productos Contratados y Balance:

- Aunque no afecta directamente al abandono, esta relación puede ser un indicador para identificar perfiles de clientes y ajustar estrategias de oferta.

Recomendaciones Estratégicas:

1. Segmentación y Programas Personalizados:

Clientes mayores y con balances altos:

- Ofrecer beneficios exclusivos, como tasas preferenciales o servicios premium.
- Diseñar programas de fidelización para fortalecer la relación con estos clientes.

2. Incentivar la Actividad del Cliente:

- Implementar campañas de reactivación para clientes inactivos, destacando beneficios de uso de productos o servicios adicionales.

3. Optimización de la Oferta de Productos:

- Analizar la rentabilidad de clientes con múltiples productos y diseñar estrategias para equilibrar la relación entre productos contratados y satisfacción del cliente.

4. Modelo Predictivo:

- Utilizar estas correlaciones para desarrollar un modelo de Machine Learning que permita identificar con mayor precisión los segmentos de alto riesgo de churn y anticiparse al abandono.

Insight clave:

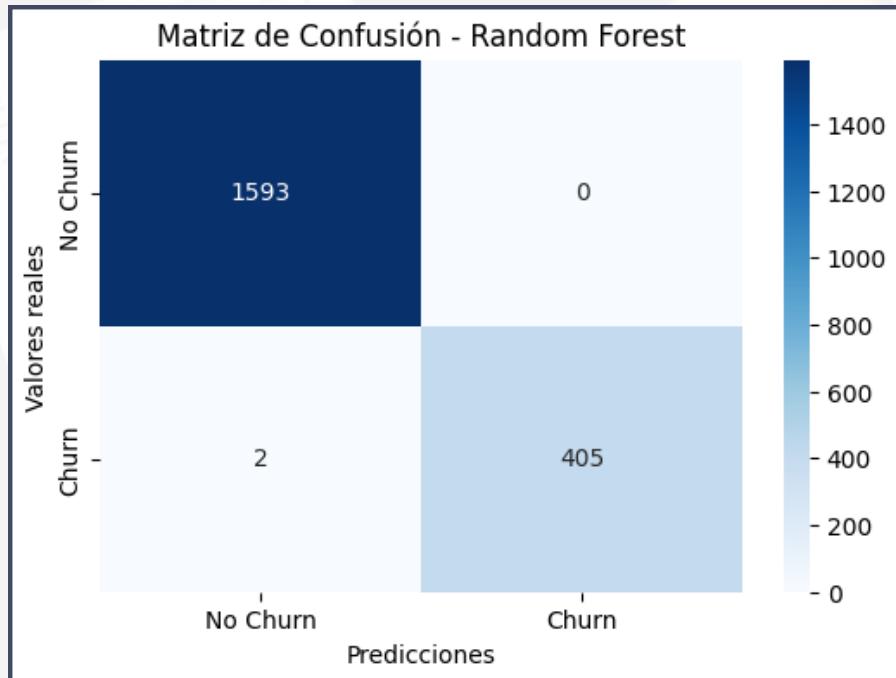
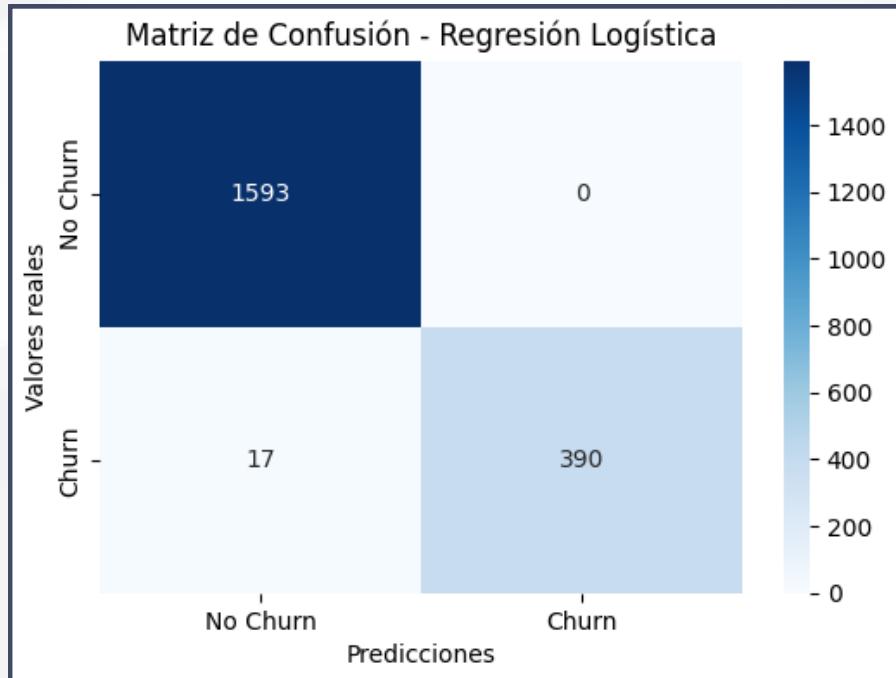
Los clientes mayores, con balances altos y baja actividad son los más propensos a abandonar. Fomentar su participación y ofrecerles beneficios personalizados es fundamental para mejorar la retención y maximizar el valor de la relación con el cliente.

Conclusión Final

1. Solución del Problema:

El objetivo principal del proyecto fue identificar y predecir los clientes con alta probabilidad de abandono (churn) utilizando técnicas de Machine Learning. A través de un análisis exhaustivo del dataset y de la selección de modelos adecuados, se logró implementar un sistema de predicción robusto que permite:

- Identificar clientes en riesgo de churn con alta precisión.
- Priorizar estrategias de retención dirigidas a segmentos específicos basados en características relevantes como edad, balance y actividad del cliente.



2. Justificación del Modelo para Producción:

Después de la evaluación de múltiples algoritmos, se determinó que **Random Forest** es el modelo más adecuado para llevar a producción debido a su rendimiento superior:

Desempeño de Random Forest:

- **Accuracy:** 99.9%. Clasifica correctamente casi todas las observaciones del dataset.
- **Precision:** 1.0. Garantiza que todas las predicciones de abandono son precisas, sin falsos positivos.
- **Recall:** 99.51%. Detecta casi todos los clientes que abandonaron, minimizando falsos negativos.
- **F1-Score:** 99.75%. Logra un balance casi perfecto entre precisión y recall.
- **Matriz de Confusión:** Solo 2 clientes fueron clasificados incorrectamente como no propensos al abandono, lo que refuerza su confiabilidad.

Importancia de las Variables:

- La variable **age_churn_interaction** fue la más relevante, destacando la interacción entre la edad y la probabilidad de abandono.
- Variables como **age**, **products_number**, y **active_member** también jugaron un rol clave en el modelo, proporcionando insights accionables para estrategias de retención.

Ventajas del Modelo Random Forest:

- Capacidad para manejar datos no lineales y relaciones complejas.
- Robustez frente a variables menos relevantes, como **estimated_salary** y **country**, que tienen un impacto reducido en la predicción.

Regresión Logística como Alternativa: Aunque la regresión logística también mostró un desempeño destacado (accuracy del 99.15%), Random Forest supera este modelo al manejar mejor los falsos negativos y capturar patrones no lineales más complejos.

3. Recomendaciones para Implementación en Producción:

Integración del Modelo:

- Utilizar el modelo de **Random Forest** para analizar y monitorear clientes en tiempo real.
- Implementar un sistema de alertas que identifique clientes con alta probabilidad de abandono para acciones proactivas.

Optimización de Estrategias:

- Segmentar campañas de retención basadas en las características clave identificadas (edad, balance, número de productos contratados).
- Fomentar la actividad de los clientes menos comprometidos con programas de reactivación.

3. Recomendaciones para Implementación en Producción:

Integración del Modelo:

- Utilizar el modelo de **Random Forest** para analizar y monitorear clientes en tiempo real.
- Implementar un sistema de alertas que identifique clientes con alta probabilidad de churn para acciones proactivas.

Optimización de Estrategias:

- Segmentar campañas de retención basadas en las características clave identificadas (edad, balance, número de productos contratados).
- Fomentar la actividad de los clientes menos comprometidos con programas de reactivación.

Mantenimiento del Modelo:

- Actualizar y reentrenar el modelo periódicamente con nuevos datos para mantener su efectividad.
- Evaluar y ajustar las variables incluidas para simplificar el modelo sin perder precisión.

4. Conclusión sobre la Producción:

El modelo de **Random Forest** es totalmente apto para producción. Su rendimiento superior en términos de precisión, recall y capacidad predictiva lo convierten en una herramienta confiable para abordar el problema del abandono de clientes.

Su implementación permitirá al banco anticiparse al abandono de clientes, optimizar recursos y diseñar estrategias de fidelización efectivas, mejorando así la sostenibilidad del negocio a largo plazo.