Universidad Internacional San Isidro Labrador

Curso: Data Science

Proyecto I

Tema:

Análisis de Deserción de Clientes en una Institución Bancaria

Docente: Samuel Saldaña Valenzuela

Estudiante:

Aranza Moreira Sánchez

Noviembre, 2024

Contenido

Introducción:

Para dar inicio a la ejecución del proyecto es importante comprender que actualmente, el competitivo panorama del sector financiero, la fidelización de clientes ha adquirido un papel estratégico fundamental para los bancos y compañías financieras en general. Partiendo de lo anterior, logramos comprender como en nuestro país, las entidades financieras enfrentan el desafío de comprender los factores que incitan a sus clientes a abandonar sus servicios, esto con la intención de implementar estrategias que resulten efectivas para reducir el churn. Este fenómeno, además de representar una pérdida directa de ingresos, incrementa de manera considerable los costos relacionados con la adquisición de nuevos clientes, lo que compromete la sostenibilidad y competitividad a largo plazo de la institución.

El trabajo se estructura bajo el método CRISP-DM (Cross-Industry Standard Process for Data Mining), reconocido por su enfoque sistemático en proyectos de minería de datos. Este método permite avanzar desde la comprensión del contexto del negocio y del conjunto de datos hasta su limpieza y normalización, asegurando que la información esté en condiciones óptimas para la implementación de algoritmos de aprendizaje automático.

Además de lo anteriormente mencionado, en este capítulo se encuentra a modo de antecedentes algunos estudios que poseen una relación próxima al mismo, es decir sirven de guía para el desarrollo de la investigación, y a su vez se encuentra la justificación, la cual expresa el principal motivo de la elaboración de este trabajo, así como el principal problema; esto sin dejar de lado la presentación de los objetivos (general y específicos) siendo estos los logros esperados al termino de dicho proyecto.

Objetivos:

Objetivo General:

El objetivo general de este proyecto es:

Preparar un conjunto de datos enfocado en la deserción de clientes en una institución bancaria. Para ello, se realizará un análisis exploratorio detallado y un proceso exhaustivo de limpieza de datos…

Objetivos Específicos:

Los objetivos específicos de este proyecto son:

- Seleccionar y aplicar técnicas de pre-procesamiento de datos, se utilizarán técnicas como la imputación de valores faltantes, la codificación de variables categóricas y el escalado de variables numéricas para preparar los datos para el modelado.
- Detectar y corregir problemas de calidad que puedan comprometer la integridad y confiabilidad de los datos.
- Realizar un análisis exploratorio de datos exhaustivo, en el cual se aplicarán técnicas de visualización y estadística descriptiva para comprender la distribución de las variables, identificar relaciones.

Marco Teórico:

Se toma en cuenta que el tema abordado en general es un fenómeno crítico en las industrias de servicios financieros, donde la competencia es alta y las barreras de salida son mínimas. La deserción se refiere al acto de que los clientes dejan de utilizar los servicios de una institución y trasladan sus relaciones financieras a otra. Este comportamiento afecta negativamente a las instituciones financieras, ya que la pérdida de clientes conlleva a una reducción en los ingresos, una disminución en la lealtad y una mayor dificultad para atraer y retener clientes nuevos.

Hoy en día, los bancos enfrentan un entorno dinámico en el que los clientes demandan mejores servicios y mayor personalización, o popularmente hablando, un servicio individualizado. Por ello, los bancos han comenzado a implementar técnicas avanzadas de análisis de datos, enfocadas en identificar patrones de comportamiento y factores que inciden en la decisión de abandonar los servicios bancarios. Esto les permite optimizar tanto sus estrategias de retención como las de marketing y servicio al cliente, contribuyendo a fortalecer la lealtad.

Los datos masivos recopilados por los bancos ofrecen una oportunidad valiosa para extraer conocimientos profundos sobre el comportamiento y las expectativas de los usuarios. Esta capacidad de análisis y predicción permite a las instituciones no solo reducir el riesgo de deserción, sino también mejorar la experiencia del cliente y ajustar las estrategias de servicio y marketing.

Este marco teórico no solo enmarca la importancia del EDA y la preparación de datos, sino que sienta las bases para garantizar modelos predictivos robustos y aplicables a problemas reales en la industria bancaria.

Además, retener a un cliente existente resulta generalmente menos costoso que captar uno nuevo. Esta regla sugiere que el 80% de los ingresos de una empresa a menudo proviene del 20% de sus clientes actuales. Por lo tanto, comprender los factores que llevan a la deserción es esencial para maximizar la rentabilidad.

Desarrollo Teórico:

- ## Fase Conceptual del Modelo CRISP-DM (

1. Comprensión del negocio

Se debe comprender que el objetivo principal de este proyecto es principalmente desarrollar un modelo que permita predecir la probabilidad de deserción de clientes (entendida como la salida o baja de un cliente en el banco).

Dicho análisis permitirá comprender de manera profunda los distintos factores que influyen en la deserción, para que así el banco pueda establecer estrategias de retención.

- ¿Qué características definen a los clientes que desertan?
- ¿Cómo se puede intervenir de manera proactiva para retenerlos?

Resultado esperado: Un modelo que prediga la probabilidad de que un cliente deserte, ayudando al banco a reducir la deserción.

2. Comprensión de los datos

Como bien se sabe, la calidad de los datos disponibles es crucial para construir modelos predictivos efectivos.

En esta ocasión el Dataset a utilizar incluye datos sobre las características demográficas y financieras de los clientes, como:

- Edad
- Antigüedad en el banco
- Saldo de cuenta
- Número de productos adquiridos
- Información sobre crédito
- Estado de deserción (variable objetivo)

Es importante mencionar que es fundamental analizar cada columna para verificar la existencia de valores “nulos, valores atípicos, o inconsistencias” en el formato. Esto facilitará la preparación de los datos y mejorará la calidad del análisis en cuestión.

Resultado esperado: Un análisis detallado de los datos y un informe que incluya patrones observados y posibles variables.

3. Preparación de los datos

En términos generales, los datos suelen recopilarse de diversas fuentes, y cada fuente puede tener un formato diferente, es por ello que la preparación de los datos garantiza que todos los datos fueron recopilados y que se utilizarán en el proyecto estén en un formato consistente y uniforme. Al dedicar tiempo a esta etapa, aumentamos la confiabilidad y la precisión de nuestros resultados.

- **Limpieza de datos**: Para asegurar la consistencia, se tratarán valores nulos o NaN, ya sea imputándolos con valores promedios o eliminándolos si no se cuenta con suficiente información.
- **Codificación de variables**: Las variables categóricas se transformarán en variables dummies (por ejemplo, sexo, estado civil).
- **Tratamiento de valores atípicos**: Valores extremos se identificarán mediante técnicas estadísticas como el análisis de cuartiles.
- **Reducción de dimensionalidad**: Para optimizar el análisis, se aplicará una reducción de dimensionalidad (como PCA) en las variables de alto número de categorías o de baja relevancia.

Por ejemplo, en un conjunto de datos sobre clientes bancarios, podríamos normalizar variables como el saldo promedio de la cuenta o las transacciones mensuales, asegurando que tengan una escala uniforme.

4. Modelado de datos

Consiste en construir modelos matemáticos o estadísticos que representen las relaciones entre las variables de nuestros datos. Estos modelos nos permiten hacer predicciones, clasificaciones, agrupamientos y otras tareas analíticas.

Con el modelado de datos podemos predecir valores futuros o eventos desconocidos. Por ejemplo, predecir si un cliente va a abandonar una empresa o el precio de una acción.

- **PCA (Análisis de Componentes Principales)**: Reducirá las dimensiones del Dataset y permitirá identificar las variables con mayor peso en la deserción.
- **K-Means (Clustering)**: Permitirá agrupar a los clientes en segmentos basados en su riesgo de deserción, lo que permitirá enfoques de retención específicos.

Resultado esperado: Un modelo de clasificación que permita identificar clientes con alta probabilidad de deserción y que optimice las estrategias de retención del banco.

Se puede mencionar que el modelado de datos es una herramienta poderosa que nos permite extraer conocimiento valioso de los datos y tomar decisiones más informadas.

II. Base Teórica de las Técnicas de Data Mining Aplicadas

1. Minería de reglas de asociación

Esta técnica permite identificar relaciones entre las características de los clientes que deserten y aquellas que permanezcan. Las reglas de asociación son útiles en este contexto para descubrir combinaciones de factores de riesgo, como “clientes jóvenes con bajo saldo” o “clientes con solo un producto contratado y antigüedad menor a un año”.

Adicionalmente, esta técnica puede ser útil para descubrir “combinaciones” de características que suelen estar presentes en clientes que deciden dejar el banco.

Por ejemplo, se podrían identificar reglas del tipo:

- "Si un cliente tiene un saldo bajo y un puntaje crediticio bajo, entonces es más probable que deserte".

2. Clasificación:

La clasificación se aplicará para predecir si un cliente desertará o no

Entre las técnicas comunes de clasificación para este propósito se encuentran:

- **Regresión logística**: Útil para predecir una variable binaria (deserción sí/no).
- **Árboles de decisión**: Proporcionan una representación clara de cómo diferentes factores afectan la decisión de los clientes.
- **K-Nearest Neighbors (KNN)**: Clasifica clientes en función de similitudes, agrupando clientes con características similares en el mismo grupo de riesgo.

3. Agrupación en clústeres (Clustering)

¿Cómo funciona?

La agrupación en clústeres asigna cada punto de datos a un grupo (o clúster) de modo que los puntos dentro de un mismo clúster sean lo más similares posible, mientras que los puntos de diferentes clústeres sean lo más diferentes posible.

K-Means agrupa a los clientes en clústeres según el riesgo de deserción, generando segmentos de clientes de alto y bajo riesgo. Esto permite implementar estrategias de retención de manera personalizada, maximizando la eficacia de las intervenciones.

4. Análisis de secuencias y trayectorias

Este método permite reconocer patrones y directrices en datos secuenciales, como las acciones de un usuario en un sitio web, los movimientos de un objeto en el espacio o la progresión de una enfermedad.

Aunque el análisis de secuencias y trayectorias no se aplicará en profundidad, su uso podría ser valioso si se analiza la permanencia de los clientes a lo largo del tiempo. Se podría investigar si existe una "trayectoria de deserción" común que indique los primeros pasos de un cliente en riesgo.

III. Aplicación de Data Mining en el Proyecto

1. Minería de procesos

Este análisis se centra en el ciclo de vida del usuario, permitiendo identificar en qué etapas del ciclo de relación con el banco los clientes suelen desertar. A partir de esto, se pueden definir puntos críticos para ofrecer incentivos o reforzar la retención.

2. Minería de textos

¿Cómo funciona?

El proceso de minería de textos habitualmente implica los siguientes pasos:

- **Recopilación de datos:** Se recolectan los textos de diversas fuentes.
- **Pre-procesamiento:** Se limpian los textos, se eliminan palabras vacías, se normaliza el texto y se realiza la “división”.
- **Representación vectorial:** Se convierten los textos en representaciones numéricas que pueden ser procesadas por algoritmos de aprendizaje automático.
- **Análisis:** Se aplican técnicas de clasificación, clustering, extracción de entidades y otros métodos para extraer información.
- **Visualización:** Se presentan los resultados de manera visual para facilitar la interpretación.

Si en general el banco cuenta con datos de texto, como opiniones de clientes sobre los servicios bancarios o transcripciones de interacciones de soporte al cliente, la minería de textos podría analizar los sentimientos y temas frecuentes en las quejas o sugerencias de los clientes.

Sin embargo, si se demuestra que el dataset no contiene datos textuales, esta técnica no será utilizada en el análisis. Sin embargo, si se añadieran datos de encuestas o comentarios de clientes, esta práctica permitiría identificar sentimientos y razones de deserción.

3. Minería predictiva

La minería predictiva es el enfoque principal del proyecto, aquí se aplicarán modelos como retracción logística y árboles de decisión para predecir la probabilidad de deserción. Esto permitirá al banco clasificar a los clientes en grupos de “alto, medio y bajo” riesgo según corresponda.

Esta será fundamental para este proyecto, ya que se utilizarán algoritmos de clasificación (como la regresión logística o los bosques aleatorios) para desarrollar un modelo que prediga la probabilidad de deserción de cada cliente.

Este modelo permitirá al banco segmentar a los clientes en función de su probabilidad de deserción, facilitando la aplicación de estrategias de retención.

Para este, el modelo predictivo permitirá una identificación proactiva de clientes en riesgo, lo que da al banco la oportunidad de implementar acciones de fidelización específicas, como ofertas personalizadas o atención prioritaria.

Finalmente, y en consecuencia de lo anterior, se puede observar como en este proyecto, la minería predictiva será la técnica principal para crear un modelo de clasificación que identifique clientes en riesgo de desertar.

Sin embargo, la minería de procesos y la minería de textos pueden aportar un valor adicional al permitir un análisis más profundo de los patrones de comportamiento e insatisfacción que

contribuyen al abandono. Esto proporciona una visión integral que permite al banco tomar decisiones informadas y estratégicas para retener a sus clientes.

IV. Estadística Descriptiva y Análisis Exploratorio (EDA)

1. Análisis de cada técnica de estadística descriptiva

Se analizarán diferentes estadísticas para entender la estructura y consistencia de los datos:

- **Count**: Proporciona el conteo de observaciones por columna.
- **Mean (Media)**: Permite observar tendencias generales de variables numéricas.
- **Desviación estándar (std)**: Muestra la dispersión de datos, ayudando a identificar columnas con alta variabilidad.
- **Mediana**: Se usa para entender el punto medio de variables sesgadas.
- **Distribución de frecuencias**: Para variables categóricas, se identificará la frecuencia de cada valor, lo que permitirá detectar anomalías en los datos.

2. Análisis exploratorio de los datos (EDA)

- **Distribución de variables**: Se analizará la distribución de cada variable clave, como el saldo de cuenta, antigüedad, y número de productos adquiridos.
- **Correlación**: Las correlaciones entre variables como “saldo” y “deserción” ayudarán a identificar las más predictivas para la creación del modelo.
- **Transformación de datos**: Variables categóricas serán transformadas en dummies, y los valores nulos serán tratados para asegurar consistencia en el modelo.

Desarrollo Práctico:

Explicación del código:

Según el proyecto realizado, observamos que este código permite un análisis detallado en relación a un conjunto de clientes y la realidad de abandono ante compañías financieras. El proceso del código hace una visualización completa de datos, así como data cleaning, también genera consigo datos y, por ende, nos muestra como resultado el conjunto limpio.

En primera instancia el código, comienza cargando un conjunto de datos desde una URL haciendo uso de pandas y su función *–read_csv- ()*, lo que permite cargar una data frame, luego se muestran las primeras filas del conjunto de datos, seguidamente, se ejecuta una limpieza de datos de primer análisis, eliminando los valores "nulos" en cualquier columna del data frame; Esto para no generar ninguna posible afectación en el código, luego de ello, se observa como el código verifica la existencia de duplicados, para dicho resultado se utilizó el método *duplicated ()* , esta fórmula lo que realiza es la devolución de booleanos, indicando sí existe una dúplica de la fila anterior.

Luego de dicho paso, se realiza *sum ()*, este método cuenta cuantos valores se encuentran duplicados, importante mencionar que el conjunto de datos también suele contar con variables categóricas como geografía y género, al menos en este caso, para que los algoritmos de Machine Learning puedan trabajar en estos, es necesario convertir dichas variables en variables numéricas por eso utilizamos *pd.get_dummies ()* que crea columnas binarias para las distintas categorías, como, por ejemplo:

- Columna de geografía

Recordamos, que el *dropfirst=true* asegura evitar la múlticolinealidad, esto lo que asegura en términos generales es que una categoría no se incluya en el modelo, y se utilice como referencia, evitando así redundancias, y que las demás categorías no se "comparen" con respecto a esta.

Seguidamente se van a identificar los valores atípicos en las variables, aquí se genera un *boxplot* para cada columna numérica como se puede visualizar en *numeric_columns* como es el caso en:

- CREDITSCORE
- AGE

El *Boxplot*, nos resulta útil para la identificación de *o*utlyers; En este caso se está utilizando *sns.boxplot ()* para dibujar los diagramas junto con el conjunto de subgraficos, luego basándonos en el rango intercuartiles hacemos la limpieza de los *outlyers.*

Después se calculan las estadísticas descriptivas de las variables haciendo uso de *describe ()*, utilizando medidas como:

- Desviación estándar
- Mínimos y máximos
- Percentiles

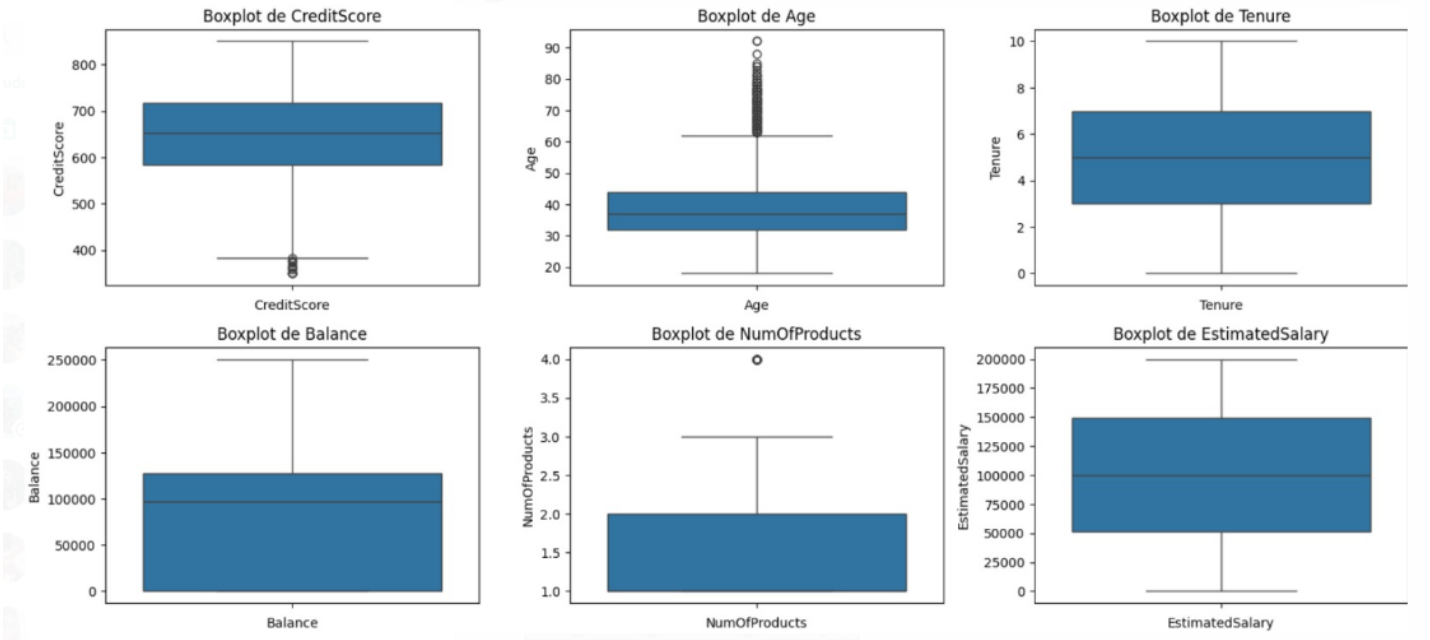
Y ya luego, se muestran las estadísticas para la realización del análisis inicial de datos.

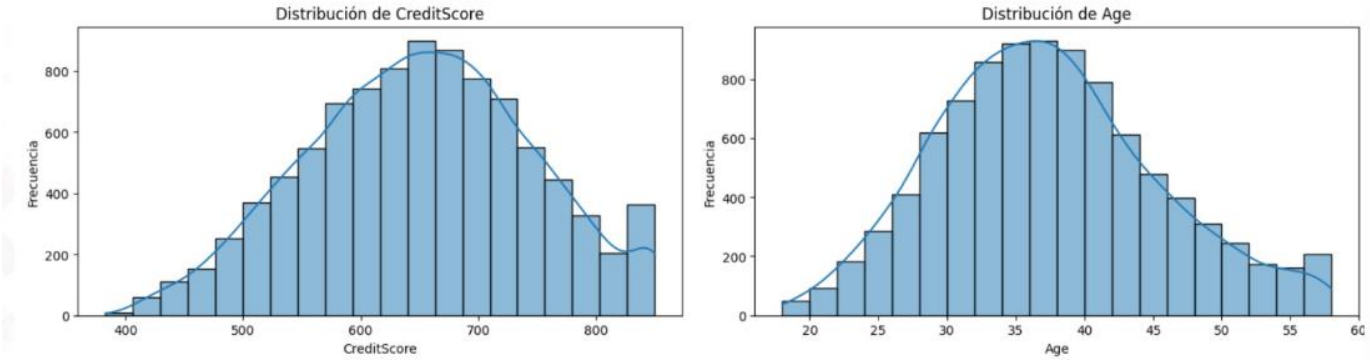
Seguidamente, se generan histogramas para cada una de las columnas numéricas de *data_cleaned* (*con sns.histplot()*)

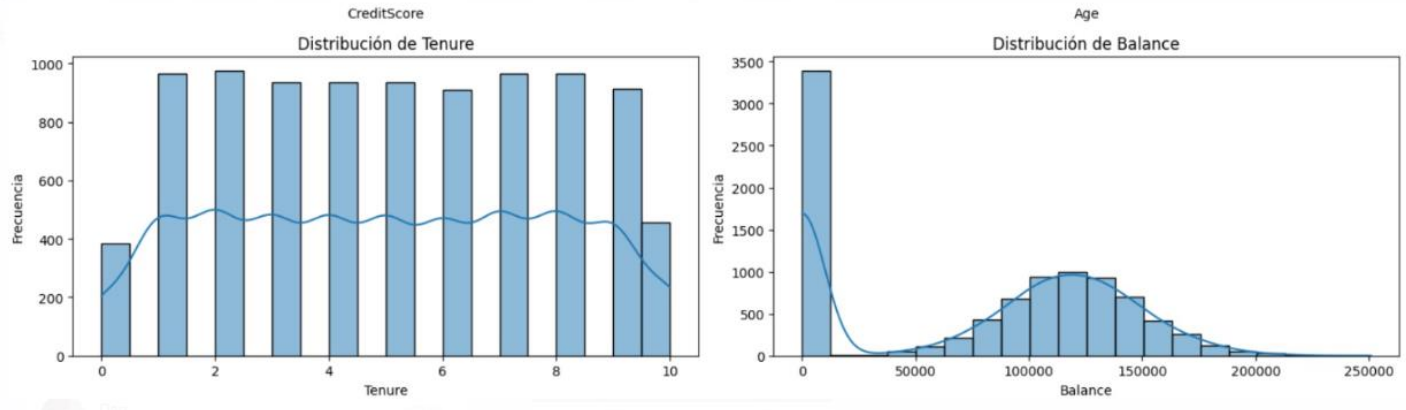
Los histogramas ayudan a visualizar la distribución de los datos. También se incluye una curva de densidad (con el parámetro *kde=True*), que muestra la distribución estimada de los datos.

Después, generamos gráficos de barras para variables geográficas, aquí se seleccionan las columnas que contienen “geography_” en su nombre, lo que sugiere que son variables denominadas categóricas, y posterior a ello, muestra la frecuencia de cada categoría según corresponda.

Para ejemplificar con mayor exactitud y amplitud lo desarrollado anteriormente, se generan los siguientes gráficos, a continuación, una muestra de ellos:


Boxplot de CreditScore
CreditScore
Boxplot de Age
Age
Boxplot de Tenure
Tenure
Boxplot de Balance
Balance
Boxplot de NumOfProducts
NumOfProducts
Boxplot de EstimatedSalary
EstimatedSalary


Distribución de CreditScore
Frecuencia
CreditScore
Distribución de Age
Frecuencia
Age


CreditScore
Age
Distribución de Tenure
Frecuencia
Tenure
Distribución de Balance
Frecuencia
Balance

Roles:

Los gráficos y histogramas son importantes para entender y analizar los datos, ya que ofrecen una representación visual que facilita su interpretación. Estos gráficos nos ayudan a analizar la manera en que los datos están distribuidos, identificar patrones y detectar posibles problemas como valores extraños, sesgos o tendencias que no son fáciles de ver a simple vista. Un histograma es una forma de ver cómo se distribuyen los datos de una variable continua, mostrando con barras la frecuencia de valores en distintos rangos.

Este gráfico ayuda a identificar si los datos tienen algún tipo de sesgo, si hay más de una moda presente o si la distribución es normal. Esto puede ser útil al momento de decidir qué métodos estadísticos o de modelado utilizar. Los diagramas de caja, también conocidos como boxplots, son útiles para encontrar valores atípicos y ver cómo se distribuyen los datos. Muestran de forma sencilla los diferentes cuartiles y los valores extremos.

Los gráficos de dispersión facilitan la comprensión de la relación entre diferentes variables, lo que es fundamental para la creación de modelos predictivos y la toma de decisiones fundamentadas. Es importante mostrar los datos de forma clara para que los resultados sean comprendidos por personas que no tienen un conocimiento técnico, como gerentes o partes interesadas.

Es entonces, que, en resumen, dichos gráficos e histogramas son herramientas fundamentales para analizar datos y tomar decisiones informadas, desarrollar modelos precisos y encontrar soluciones basadas en datos.

Una vez concluido lo anterior y para ir finiquitando, se continua con el gráfico de dispersión entre “edad y balance”, este es simple, en general muestra la relación entre -Age y Balance- y como estas están relacionadas con la variable *Exited* que indica si un cliente / usuario abandonó el servicio.

Gráfico de relación

Tabla de comparación y resultados según el gráfico:

Se logra comprobar según lo visto la gran difusión independientemente de las edades y guiándonos bajo los siguientes colores:

AZUL	NARANJA
Menor edad	Mayor edad
Balances bajos	Balances altos

Finalmente, se guarda el data frame limpio con la siguiente formula de código:

```
data_cleaned.to_csv('/content/cleaned_dataset.csv', index=False)
```

Pasado lo anterior, el código lo que hace es que se guardará limpio de *outliers* en un archivo CSV.

El proceso de análisis exploratorio de datos (EDA) y limpieza de datos es fundamental para cualquier proyecto de análisis de datos o machine learning. Antes de comenzar a modelar o extraer conclusiones de un conjunto de datos, es esencial comprender su estructura, distribuciones y posibles problemas, como los valores atípicos.

La visualización de datos mediante gráficos como los *boxplots* y los histogramas permite detectar patrones, anomalías y outliers que pueden afectar la precisión de los modelos predictivos.

Además, la limpieza de estos datos, eliminando o corrigiendo los valores atípicos, mejora la calidad de las predicciones y asegura que los resultados obtenidos sean más robustos y representativos

Una de las principales razones por las que la limpieza de datos es crucial es que los outliers pueden distorsionar significativamente las estadísticas descriptivas y los resultados de los modelos. Por ejemplo, en el análisis de variables como "edad" o "puntaje crediticio", un valor extremadamente alto o bajo puede sesgar las medias, desviaciones estándar y otros parámetros importantes.

Al eliminar estos valores extremos, los análisis se vuelven más fiables y precisos. Asimismo, las distribuciones de las variables se normalizan, lo que facilita el uso de métodos estadísticos o algoritmos de machine learning que requieren datos sin distorsiones.

Además, la transformación de variables categóricas en variables numéricas mediante técnicas como *pd.get_dummies ()* es esencial para que los algoritmos de machine learning puedan procesar los datos de manera efectiva. Este tipo de pre-procesamiento no solo facilita la construcción de modelos, sino que también permite realizar análisis más complejos y obtener insights más profundos sobre el comportamiento de los datos.

Por consiguiente, en resumen, un adecuado análisis y limpieza de datos es un paso indispensable para asegurar que cualquier modelo o análisis posterior se base en información sólida y precisa, evitando interpretaciones erróneas y maximizando el valor extraído de los datos.

Conclusiones y recomendaciones

Conclusiones:

1. La transformación de las variables categóricas en el conjunto de datos fue esencial para preparar la información que se utilizará en el modelo. Este proceso asegura que los datos estén en su mejor forma posible, lo que facilita la aplicación de algoritmos de aprendizaje automático, los cuales requieren datos claros, coherentes y estandarizados para mejorar la precisión en los resultados del proyecto.

2. La depuración y normalización de los datos fueron esenciales para asegurar la calidad del conjunto de datos, eliminando errores e inconsistencias. Esto no solo favorece la creación de modelos predictivos más confiables, sino que también asegura que las decisiones basadas en dichos modelos sean más precisas. Un conjunto de datos bien preparado permite realizar análisis más exactos y mantener la fiabilidad del modelo a largo plazo, minimizando el riesgo de sesgos y errores en los resultados.

3. El análisis exploratorio de datos es clave para entender los factores que influyen en la deserción de clientes, al identificar patrones y relaciones importantes entre las variables. Esta fase no solo mejora la exactitud del modelo al centrarse en los aspectos fundamentales que afectan el comportamiento de los clientes, sino que también optimiza la capacidad del modelo para predecir la pérdida de clientes de manera más eficiente. En resumen, el EDA se presenta como una herramienta vital para construir modelos sólidos y confiables, capaces de generar resultados prácticos y útiles para tomar decisiones estratégicas.

Recomendaciones:

1. Crear incentivos para clientes con alto riesgo de abandonar: Para disminuir la tasa de deserción en los grupos de clientes más propensos a irse, se sugiere diseñar incentivos personalizados y específicos.

- Estos incentivos podrían incluir beneficios como descuentos, recompensas por fidelidad, programas de puntos o acceso a servicios adicionales sin costo durante un periodo limitado. La efectividad de estos incentivos debe ser monitoreada de manera continua, ajustando las estrategias según los resultados obtenidos.

2. Realizar actualizaciones frecuentes del modelo predictivo: Para garantizar que el modelo predictivo siga siendo preciso y relevante, es aconsejable actualizarlo regularmente con datos nuevos. Esto permitirá que el modelo se ajuste a los cambios en los comportamientos de los clientes y a los nuevos riesgos que puedan surgir, mejorando su capacidad para predecir la deserción con mayor precisión. Las actualizaciones constantes también aseguran que el modelo siga siendo competitivo frente a las dinámicas cambiantes del mercado financiero y las expectativas de los clientes.