Universidad Internacional San Isidro Labrador

Curso: Data Science

Proyecto 2

Tema:

Análisis de Deserción de Clientes en una Institución Bancaria

Docente: Samuel Saldaña Valenzuela

Estudiante:

Aranza Moreira Sánchez

Noviembre, 2024

Contenidos

Introduccion

Partiendo del conjunto de datos depurado en la fase anterior, se aplican métodos tanto supervisados como no supervisados con el fin de identificar características determinantes que influyen en la decisión de los clientes de abandonar el servicio, se crea un sistema predictivo que utiliza diferentes métodos de análisis de datos para predecir cuándo un cliente podría dejar de serlo. Se realizó un estudio minucioso para descubrir los motivos que llevan a los clientes a dejar de usar los servicios del banco. Utilizando PCA y K-Means, se puede dividir a los clientes de manera precisa en diferentes segmentos. Además, se utilizó el modelo supervisado XGBoost para predecir con gran precisión la probabilidad de que un cliente deje el banco.

El Análisis de Componentes Principales (PCA) fue utilizado, ya que fue una gran herramienta para poder simplificar los datos, identificando las características más relevantes y evitando problemas como la redundancia en los datos, repetición y el sobreajustar, lo que resulta en un modelo más preciso. Además, K-Means clustering agrupa observaciones parecidas para facilitar la detección de patrones ocultos en los datos. En esta etapa se eligen y ajustan los modelos para mejorar su rendimiento y garantizar que las predicciones sean lo más acertadas.

Este método crea un sistema confiable de predicción que brinda datos importantes para tomar decisiones estratégicas sobre cómo prevenir la pérdida de clientes. Este modelo ayuda al banco a retener a sus clientes actuales, reducir costos de adquisición de nuevos clientes y mejorar la fidelidad.

Objetivos

ObjetivoGeneral:

Crear e implementar un modelo predictivo utilizando técnicas de Inteligencia Artificial para prever la deserción de clientes en una entidad bancaria, aplicando enfoques tanto supervisados como no supervisados, con el propósito de generar información estratégica que permita optimizar las acciones de retención y potenciar la fidelización de usuarios en el banco.

Objetivos Específicos:

1. Analizar y segmentar los datos para identificar las características y patrones determinantes en la deserción de clientes, empleando técnicas de machine learning.
2. Crear un sistema de soporte para la toma de decisiones estratégicas, incorporando el modelo predictivo para facilitar la planificación estratégica de acciones preventivas enfocadas en la lealtad de clientes.
3. Desarrollar e implementar modelos predictivos supervisados que aseguren alta precisión y fiabilidad en la predicción de la probabilidad de abandono de clientes.

Marco Teórico

En la industria empresarial, las organizaciones se enfrentan al reto constante de gestionar y analizar datos de manera masiva pero eficiente, convirtiendo esta capacidad en una ventaja estratégica decisiva. La información ha evolucionado de ser un recurso estático a transformarse en un activo dinámico y de gran valor, esencial para la toma de decisiones basadas en datos y para la creación de ventajas competitivas. No obstante, según Mayer-Schönberger & Cukier (2013) el valor real de los datos no reside únicamente en su volumen, sino en la capacidad de analizarlos y transformarlos en información útil que facilite la toma de decisiones.

Sin duda, el ciclo de vida de los datos ofrece un marco organizado para potenciar su valor, garantizando que cada etapa, desde la recolección hasta su aplicación, se maneje de forma óptima y eficaz. Este ciclo no solo sienta las bases técnicas para el análisis de datos, sino que también sincroniza los procesos analíticos con las metas estratégicas de la organización. En este proyecto se examinará el proceso con mayor profundidad más adelante en esta investigación.

El propósito de este estudio es abordar la fase de modelado, en esta etapa crucial en el que los datos previamente procesados y analizados se transforman en herramientas predictivas que respaldan decisiones estratégicas fundamentales para entender cómo se llega a esta etapa, es esencial explorar a fondo el ciclo de vida de los datos, un proceso que abarca múltiples fases interconectadas que trabajan conjuntamente para generar valor.

Desarrollo

A continuación, se explicarán detalladamente las etapas que dan forma al ciclo de vida de un dataset y un proyecto que utiliza datos para su objetivo. Cada fase será examinada en profundidad, desencadenando tambien su importancia en el proceso de desarrollo. Este proceso abarca desde la inicial de el ciclo de vida de los datos, hasta toda la explicaciones de modelos, EDA, entre otros temas ofreciendo así una perspectiva completa y minuciosa del camino que recorre el proyecto. Comprender el también que por que es fundamental todos estos conceptos para garantizar el éxito de cada proceso de desarrollo, ademas de que cada fase conlleva decisiones cruciales que impactan de manera directa en el resultado final.

Etapas del Ciclo de Datos

El ciclo de vida de los datos es un proceso seguido que abarca distintas fases, cada una de ellas esencial esencial para su analisis, su transformación y gestion para lograr tener datos valiosos. Estas etapas están interrelacionadas y fueron aplicadas en el proceso desarrollado lo que permitió un uso eficiente y mejor estructura al momento de tomar decisiones.

Generalmente, estas estapas son las siguientes:

Obtención de Datos

La obtención de datos es la primera fase del ciclo de la data y se refiere al proceso de adquirir información valiosa de diversas fuentes como sistemas internos, redes sociales, dispositivos conectados, encuestas y más. La calidad y el volumen de los datos recolectados son esenciales para el éxito de las fases que vendrán más adelante. Poniendolo en práctica en el

análisis de abandono de clientes que es el tema del proyecto, la recopilación de datos podría abarcar detalles sobre customer services, plataformas, transacciones financieras, datos financieros de los clientes, entre otros.

Gestion y almacenamiento

Después de reunir toda la información necesaria, es imprescindible guardarla en plataformas específicas que aseguren un manejo óptimo y ágil de los datos. En esta etapa crucial, se eligen cuidadosamente bases de datos y soluciones de almacenamiento que faciliten la ordenación y resguardo de vastas cantidades de información, garantizando que esté siempre accesible y protegida de posibles amenazas.

Las alternativas para el almacenamiento abarcan una variedad de opciones, desde bases de datos estructuradas hasta las más flexibles y dinámicas de datos no estructurados. Asimismo, se destacan las potentes plataformas en la nube, que facilitan la gestión y el acceso a la información.

Limpieza y Preparación de Datos

El procesamiento de transformación, el mas indispensable que involucra normalizar la información asegurando adecuación en el analisis y aseguramiento de calidad, data cleaning y la eliminación de datos incompletos o erróneos, así como la conversión de datos no numéricos a formatos apropiados para su análisis, es decir, datos numéricos. Además, puede involucrar la creación de nuevas variables basadas en las ya existentes, lo que mejora la capacidad de análisis y proporciona una mayor profundidad a los datos (Kelleher et al., 2015).

Análisis de datos

El análisis de información o exploración de data es el momento crucial en el que se desvelan patrones y conexiones relevantes, capaces de iluminar conocimientos fundamentales. En esta fase, los analistas se sumergen en un profundo examen estadístico y exploratorio en donde se descubren insignts, patrones y dinámicas ocultas en la información. Partiendo de la opinión de (Biecek & Burzykowski, 2017) La utilización de enfoques tales como el análisis descriptivo, inferencial y correlacional nos brinda una visión más rica de los datos, lo que suaviza el camino hacia la implementación de modelos predictivos de mayor complejidad.

Modelado

Trabaja con los datos procesados y analizados, listo para la creación de modelos predictivos. Es de las fases mas relevantes esta etapa transforma los datos en resultados prácticos. Durante la etapa de modelado, se desarrollan algoritmos de machine learning tanto supervisados como no supervisados que facilitan las predicciones objetivas, es decir las predicciones que buscamos para el caso de deserción de clientes. Personalmente, este es el punto destacable de la investigación.

Despliegue del Modelo

Después de que el modelo ha sido meticulosamente puesto a prueba, llega el momento emocionante de llevarlo a la acción en el entorno operativo. Esto significa fusionarlo con los sistemas empresariales existentes, lo que podría requerir la automatización de tareas o procesos fundamentados en las previsiones del modelo.

Imaginando un modelo predictivo que anticipe la deserción de clientes justo como el objetivo del proyecto, este sería un recurso valioso recurso podría ser utilizado por el banco para ser estratégicos y alcanzar retener a los clientes según preferencias, satisfacciones y necesidades de los mismos. Logrando asi una reducción significativa en la tasa de abandono.

Seguimiento

Es necesario realizar un monitoreo continuo y mantenimiento de los modelos que fueron aplicados para asegurar un rendimiento de calidad en el momento preciso y a medida que avanza el tiempo. Es indispensable tomar en cuenta que conforme va pasando el tiempo también evolucionan los patrones que fueron descubiertos y categorizados como razón de abandono de clientes o bien pueden que surgan nuevas fuentes de datos por lo tanto siempre debe de existir un monitoreo constante.

El Análisis Exploratorio de Datos (EDA)

EDA son las siglas de Exploratory Data Analysis**,** que en español se traduce como Análisis Exploratorio de Datos**.** Es una de las fases cruciales dentro del ciclo de vida de los datos en cualquier proyecto, fue en los años 70, se gestó esta revolucionaria metodología. Importante aclarar que es fundamental aplicarla antes de la etapa de modelado. Es en este proceso donde los datos comienzan a contar una historia. Lo más relevante de esta fase radica en que los patrones se revelan asi como otros insignts entonces siempre en esta etapa se va a visualizar el tema y lo que esta sucediendo con un conjunto de datos a profundidad.

Modelado Predictivo

El modelado predictivo esta basado en una variedad de enfoques, abarcando tanto el aprendizaje supervisado como el no supervisado también en este caso. Sin embargo, el aprendizaje supervisado suele ser el protagonista cuando se dispone de datos etiquetados y se conocen los resultados, asegurando así un análisis más preciso. En esta oportunidad de proyecto, XGBoost fue el modelo utilizado por su gran capacidad con altas cantidades de datos. En el modelo predictivo es donde la información procesada se transforma en herramientas predictivas que permiten tomar decisiones fundamentadas en datos.

XGBoots

A continuación, se brindara mas detalle de el modelo predictivo utilizado, de acuerdo con el centro tecnológico ESRI (s.f) XGBoost es un método de aprendizaje automático supervisado para clasificación y regresión. XGBoost es la abreviatura de las palabras inglesas "extreme gradient boosting". Este método se basa en árboles de decisión y supone una mejora sobre otros métodos, como el bosque aleatorio y refuerzo de gradientes. Funciona bien con datasets grandes y complejos al utilizar varios métodos de optimización.

Justo por lo mencionado anteriormente, fue ideal para el dataset que analizamos. Es importante resaltar que al utilizar algoritmos avanzados, los modelos predictivos permiten prever eventos futuros en este caso específico la deserción de clientes, el comportamiento de compra o incluso la probabilidad de fallos en sistemas, entre otros.

En el fragmento de código, XGBoost asume el rol protagónico como el modelo de clasificación encargado de determinar si un cliente ha decidido dar por terminado su servicio en la entidad bancaria. Se desarrolla un flujo de trabajo que comienza por normalizar las características utilizando StandardScaler, seguido por el entrenamiento del modelo XGBClassifier, donde se optimizan diversos parámetros como la cantidad de árboles (n_estimators), la tasa de aprendizaje (learning_rate) y la profundidad máxima de los árboles (max_depth). Este modelo se capacita con el conjunto de entrenamiento y se somete a una evaluación exhaustiva utilizando diversas métricas de rendimiento, tales como la precisión, el informe de clasificación, la matriz de confusión y la curva ROC. Estas métricas ofrecen una herramienta para evaluar cuán bien el modelo puede identificar a los clientes que han decidido cancelar su servicio.

Además, XGBoost proporciona una herramienta fascinante para explorar la relevancia de las características mediante su atributo feature_importances_. Esto facilita la identificación de las variables que ejercen un impacto significativo en las predicciones que se disponen, permitiéndonos entender mas que hay detrás de los resultados obtenidos.

Esto resulta valioso no solo para comprender el funcionamiento del modelo que fue empleado sino también para enriquecer el proceso de toma de decisiones fundamentadas en datos. En pocas palabras, XGBoost se utiliza para crear un modelo de clasificación potente, ágil y fácil. No solo determina si un cliente ha decidido irse, sino que también revela datos cruciales sobre los factores que influyen en esa elección.

.

Esto facilita a que las organizaciones mas grandes en este saco financieras puedan combinar diversas fuentes de datos, potenciando así la precisión de sus proyecciones. Finanlmente, es importante resaltar que hay distintos modelos que se pueden emplear, estos son sumamente efectivos y excelentes herramientas para manejar grandes cantidades de datos.

Para abordar la predicción de la deserción de clientes en esta investigación, se emplearon técnicas del modelo predictivo tales como el analissi de componentes principales, K-Means Clustering y el modelo XGBoots.

De acuerdo con IBM (s.f) El PCA resume el contenido informativo de grandes conjuntos de datos en un conjunto más pequeño de variables no correlacionadas conocidas como componentes principales. Es decir y partiendo de lp mencionado por la reconocida organización tecnológica, esta técnica busca reducir el número de variables manteniendo la mayor cantidad de información original. Esta técnica es útil cuando se trabaja con grandes cantidades de datos y muchas variables ya que ayuda a simplificar el modelo, reducir el riesgo de sobreajuste y facilitar la interpretación de los resultados.

El PCA encuentra las direcciones donde los datos cambian más y los reorganiza en esas direcciones. Este método ayuda a ver y analizar de forma más clara conjuntos de datos complicados, sobre todo cuando se trata de datos con múltiples dimensiones justo como los que se abarcan en esta investigación.

En un análisis de deserción de clientes en un banco, PCA puede utilizarse para identificar las variables más importantes que influyen en la decisión de abandonar, eliminando datos que no son necesarios y permitiendo que el modelo se enfoque en factores más relevantes.

Esta técnica es muy útil cuando se trata de un conjunto de datos con muchas características, como transacciones, interacciones con clientes, o información demográfica, entre otros ejemplos que también se abarcan en el flujo de el código.

K-Means Clustering

Este algoritmo separa los datos en grupos según las similitudes entre sus características. El algoritmo separa los datos en k grupos, siendo k un valor que se determina antes en el proceso. La idea básica es asignar cada punto de datos al grupo cuyo centroide (promedio de las características de todos los puntos en el grupo) sea el más cercano. Se incluye inicialización, asignación, actualización y repetición en todo el proceso. Parra este caso específico de abandono de usuarios, buscando estrategias para evitar esta decisión. Esta técnica logra identificar diferentes grupos de clientes con similitud de comportamiento. Basado en una análisis previo el uso de K-Means para analizar a los clientes ayuda a identificar grupos basados en su comportamiento y además ayuda a identificar las características que podrían estar influyendo en la decisión de deserción de manera más clara. Al reconocer estos comportamientos recurrentes, el banco puede adaptar sus planes de fidelización, creando campañas a la medida de cada segmento de clientes. La segmentación ayuda a priorizar recursos, concentrándolos en grupos más propensos a abandonar, lo que optimiza las acciones de fidelización y beneficia la retención de clientes a largo plazo.

Comparacion del modelo XGBoots con otros modelos empleados

En el transcurso de mi exploración con una variedad de modelos predictivos para un conjunto de datos clasificados para poder llegar al resultado que buscaba, decidí probar con tres de los modelos más reconocidos que fueron XGBoost (modelo empleado), Random Forest y Gradient Boosting. Estos modelos fueron utilizados ya que estos modelos han sido de las mejores herramientas en el mundo del aprendizaje automático, gracias a cada una de las habilidades sobresalientes de liderar con datos que tienen.

A continuación, se relatará mi viaje personal a través de cada uno de estos métodos, desglosando las razones que me llevaron a elegir XGBoost.

Random Forest.

Random Forest es un método muy efectivo sin embargo durante mi proceso la procision con este modelo fue inferior al que tuve con XGBoost. La precisión obtenida fue de aproximadamente 84%**.** Aunque este modelo ofreció una precisión razonable, en comparación con XGBoost no fue suficiente para capturar las complejidades de las relaciones en los datos de forma tan efectiva.

De acuerdo con (Breiman, 2001) Este método es conocido por su capacidad para reducir la varianza y prevenir el sobreajuste, lo que lo hace útil cuando se tienen grandes cantidades de datos y variables. Por lo que podemos colocar a Random Forest como uno de los mejores métodos de predicción, este es otro modelo basado en árboles de decisión, pero a diferencia de XGBoost, utiliza el enfoque de bagging. Una de las ventajas que presencie de este método fue la gran capacidad que tiene para manejar datos faltantes y trabajar bien con características categóricas sin necesidad de un preprocesamiento super extenso, lo que lo hace práctico en

escenarios de datos desorganizados. Sin embargo, para mi caso específico, la precisión ligeramente superior de XGBoost hizo que me inclinara por este.

Gradient Boosting

El algoritmo de Gradient Boosting es muy bueno, sin embargo la diferencia clave entre este y XGBoots es que Gradient Boosting tiende a ser más sensible a los hiperparámetros y puede ser más propenso al sobreajuste si no se ajusta adecuadamente (Friedman, 2001).

En mi caso, el modelo de Gradient Boosting alcanzó una precisión de aproximadamente de 81%**,** lo que fue notablemente inferior al rendimiento de XGBoost y Random Forest por lo que para mi no fue la mejor opción. Aunque este modelo también es muy potente y ampliamente utilizado, sus resultados no fueron tan estables ni tan precisos en mi conjunto de datos. El modelo requirió un ajuste cuidadoso de los parámetros, pero incluso después de optimizarlo, no alcanzó el mismo nivel de rendimiento que se alcanzó con XGBoost.

Justo por el rendimiento de XGBoost fue el método electo pues destacó y le dio lugar a un codigo más preciso y sólido. En mi conjunto de datos, XGBoost se destacó al adaptarse con gran eficacia a el set de data. Logró conservar una notable precisión sin comprometer el rendimiento en las categorías menos representadas. En otra perspectiva, los otros modelos que fueron utilizados como Random Forest y Gradient Boosting lograron resultados satisfactorios no obstante su precisión no fue lo esperado. Por lo que me brindó una mayor confianza XGBoots pues logró un **88.9%** de precisión, superando a Random Forest (84%) y Gradient Boosting (81%), decidí utilizarlo para mi modelo final.

Cabe recalcar que a pesar de que el modelo XGBoost demostró ser el más idóneo en esta situación, tanto Random Forest como Gradient Boosting son herramientas increíblemente potentes. Estos modelos destacan por su impresionante rendimiento. Aunque en este análisis no lograron la misma precisión, su solidez y simplicidad de implementación los hacen elecciones sumamente valiosas, estos modelos tienen el potencial de ser valiosos en aplicaciones futuras

Desarrollo Practico

Seguidamente, se presenta la explicación del código, este código está dividido en secciones las cuales incluyen los pasos del ciclo de vida completo de los datos y su preparación, ademas PCA, K-Means y principalmente el modelo prevendivo utilizado XGBoots.

Carga y preparación inicial.

El código comienza importando datos de Google Drive. Este conjunto de datos incluye información de clientes, incluyendo si han cancelado un servicio o no (Exited). Principalmente, estamos utilizando datos organizados en filas y columnas, como en una hoja de cálculo de Excel.

Después de cargar los datos, se muestran las primeras filas para tener una idea de cómo se ven antes de procesarlos. Aquí es donde empezamos a trabajar con los datos.

```
# Proyecto 02
# Aranza Moreira

from google.colab import drive
drive.mount('/content/drive')

# dataset limpio
import pandas as pd
ruta_archivo = '/content/drive/MyDrive/Colab Notebooks/cleaned_dataset.csv'
dataset = pd.read_csv(ruta_archivo)


print("Dimensiones del dataset:", dataset.shape)
print("\nPrimeras 5 filas del dataset:")
dataset.head()
```

Tranformacion de datos categóricos

Utilizando One-Hot Encoding pasamos del dataset si variables categóricas (por ejemplo, Sexo: Hombre o Mujer), a variables numéricas utilizando justo como se aplica en el desarrollo del código.

```
# Conversión de variables categóricas a numéricas
dataset_encoded = pd.get_dummies(dataset, drop_first=True)


print("\nPrimeras 5 filas del dataset codificado:")
dataset_encoded.head()
```

Separación de Características y Etiquetas

Una vez que los datos están listos, lo ideal es separar en dos segmentos el conjunto de datos, primeramente en características (X_data) y luego en etiquetas (y_data). Las características son las columnas que usaremos para predecir por ejemplo, la edad, el balance bancario, entre otros aspectos, mientras que la etiqueta es lo que queremos predecir en este caso investigar si el cliente se fue o no. Es importante recalcar que dividimos usando **train_test_split** para hacer la separación en un 70% para datos de training y un 30% para prueba, como se indica a continuación.

```
# conjuntos de entrenamiento y prueba
from sklearn.model_selection import train_test_split
X_train_data, X_test_data, y_train_data, y_test_data = train_test_split(
    X_data, y_data, test_size=0.3, random_state=42, stratify=y_data
)
```

Estandarización de los Datos

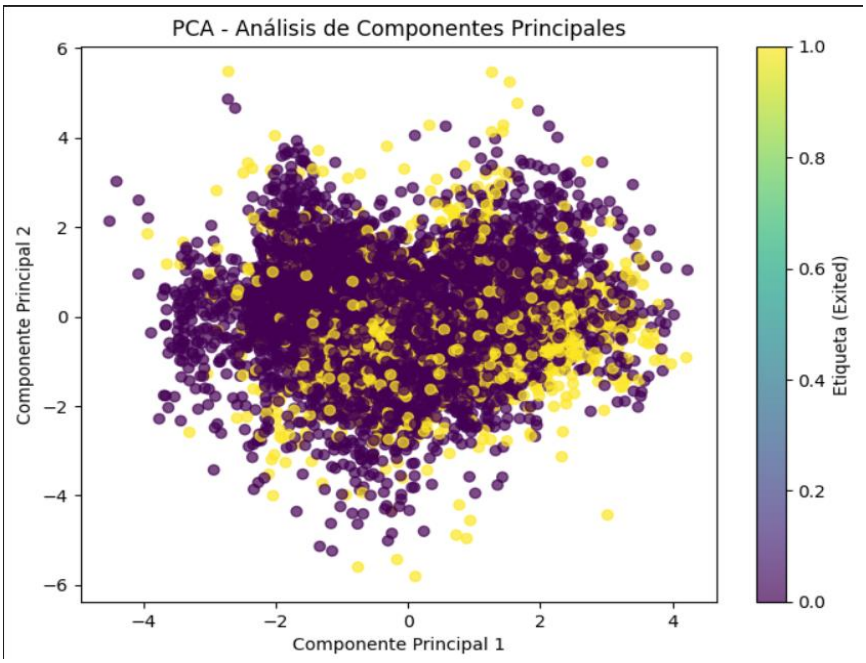
Los datos son escalados para asegurarnos de que todas las características estén en la misma escala de media 0 y desviación estándar 1. La estandarización de los datos se realiza para poner todas las características en la misma escala. Modelos como PCA o K-Means, funcionan mejor cuando los datos están estandarizados.

```
# Preprocesamiento para PCA y K-Means
from sklearn.preprocessing import StandardScaler
scaler_model = StandardScaler()
X_scaled_data = scaler_model.fit_transform(X_train_data)
```

La tecnica de modelo de analiss de componentes principales también fue utilizada para reducir la dimensionalidad de los datos. PCA lo que hizo fue tomar muchas características y las combina en un número menor de de las mismas, resaltando únicamente lo principal de ese dataset, para facilitar la visualización de análisis. En este caso especifico, reducimos a 2 componentes principales para visualizar los datos

```
from sklearn.decomposition import PCA
pca_model = PCA(n_components=2, random_state=42)
pca_results = pca_model.fit_transform(X_scaled_data)

# Gráfica de componentes principales
import matplotlib.pyplot as plt
plt.figure(figsize=(8, 6))
plt.scatter(pca_results[:, 0], pca_results[:, 1], c=y_train_data, cmap='viridis', alpha=0.7)
plt.colorbar(label='Etiqueta (Exited)')
plt.title('PCA - Análisis de Componentes Principales')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.show()
```

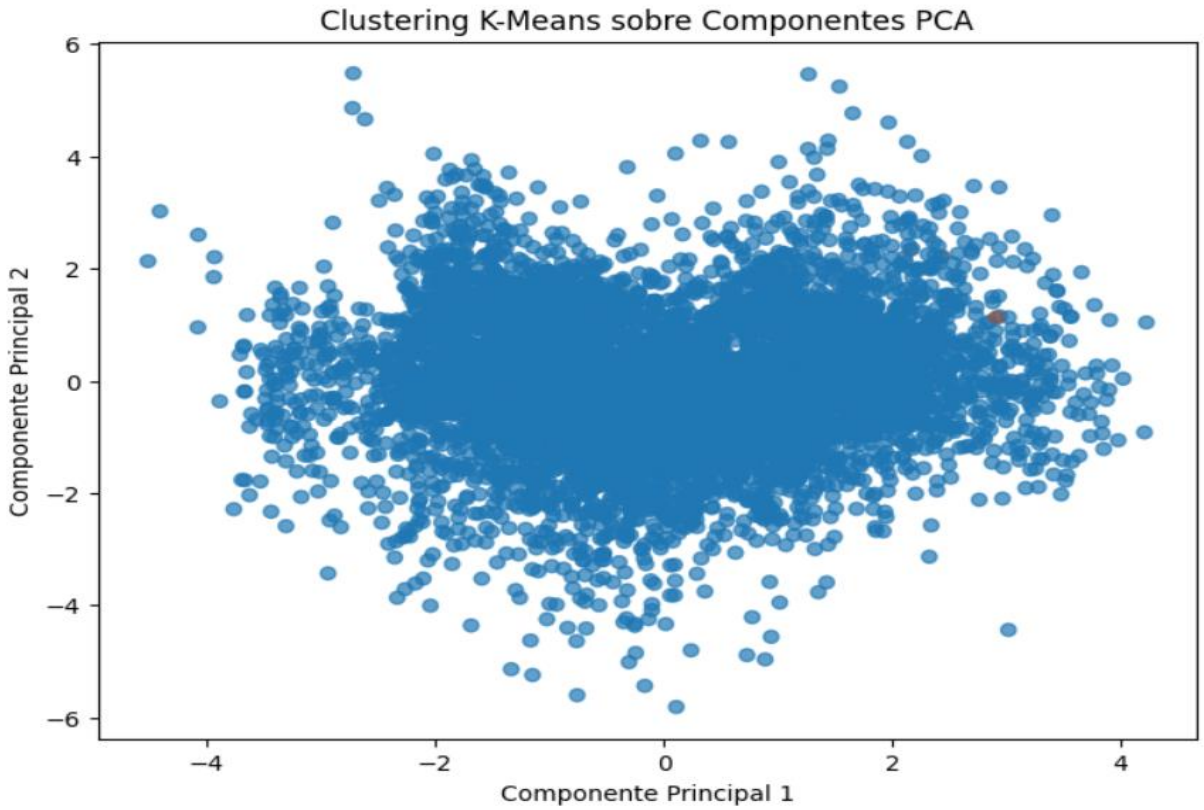
El gráfico generado por este código muestra la dispersión de los datos en las dos dimensiones principales y los puntos están coloreados según si el cliente se fue o no, como se visualiza a continuación.



A continuación, el código utiliza el algoritmo K-Means para agrupar los datos en clusters. A pesar de que K-Means no considera las etiquetas de los clientes, podemos examinar cómo organiza los datos y encontrar posibles patrones en ellos.

Una vez que se han reunido los datos, se crea un gráfico que se asemeja al de PCA, pero en este caso los puntos se identifican por colores según su grupo o cluster. Esto nos ayuda a identificar posibles patrones como grupos de clientes con características parecidas.

El grafico resultante se visualiza de la siguiente manera:



Aplicación modelo de XGBoost

Una vez realizados los pasos previos, el código construye un modelo de aprendizaje automático usando uno de los mejore modelos estratégicos el XGBoost, que es un algoritmo muy efectivo para clasificación. Este modelo toma como entrada las características de los clientes y aprende a predecir si un cliente se va o no.

```python
from xgboost import XGBClassifier
from sklearn.pipeline import make_pipeline

xgboost_pipeline = make_pipeline(
    StandardScaler(),
    XGBClassifier(
        random_state=42,
        n_estimators=500,
        learning_rate=0.01,
        max_depth=6,
        min_child_weight=1,
        subsample=0.9,
        colsample_bytree=0.8,
        gamma=0,
        scale_pos_weight=1,
        use_label_encoder=False,
        eval_metric="logloss"
    )
)
```

Evaluación del Modelo

Una vez que el modelo fue entrenado, se realizó una evaluación usando los datos de prueba. Se calcula la precisión del modelo, identificando que tan bien predice el modelo el comportamiento de los clientes (si se fueron o no).

```python
# Entrenamiento del modelo con el conjunto de train
xgboost_pipeline.fit(X_train_data, y_train_data)

# Evaluación de moelo
train_accuracy_model = xgboost_pipeline.score(X_train_data, y_train_data)
test_accuracy_model = xgboost_pipeline.score(X_test_data, y_test_data)

print(f"Precisión en el conjunto de entrenamiento: {train_accuracy_model:.2%}")
print(f"Precisión en el conjunto de prueba: {test_accuracy_model:.2%}")
```

El modelo también genera un reporte de clasificación, que incluye métricas como la precisión, el recall y el F1-score, que ayudan a evaluar su desempeño de manera más detallada lo que nos brinda una vista mas amplia de lo que esta sucediendo.

Reporte de clasificación generado

Reporte de clasificación (conjunto de prueba):

	precision	recall	f1-score	support
0	0.88	0.97	0.92	2266
1	0.75	0.45	0.56	537
accuracy			0.87	2803
macro avg	0.82	0.71	0.74	2803
weighted avg	0.86	0.87	0.85	2803

Este reporte es esencial para esta investigación puesto a que nos revela la eficacia de nuestro modelo en la predicción del comportamiento de los clientes, en particular, su probabilidad de dejar el banco. En este recap de informacion, se nos presentan dos desenlaces posibles:

- **Clase 0 (No Exited):** Clientes leales a el banco.
- **Clase 1 (Exited):** Clientes que abandonarán el banco

Métricas clave presentes en el reporte:

Precisión: La precisión nos revela cuántas veces el modelo acertó al anticipar que un cliente dejaría el banco, confirmando así sus predicciones. Es una forma de evaluar cuán certeras son las proyecciones evaluadas.

En la categoría de clase 0 (No Exited), el modelo demostró su precisión al predecir que un cliente no se iría en un impresionante 88% de los casos.

En la Clase 1 (Abandono), el modelo mostró una muy buena precisión, acertando en el 75% de las ocasiones al anticipar que un cliente decidiría marcharse.

Si bien el modelo muestra una notable precisión, sobre todo en la identificación de los clientes que permanecen leales, aún queda margen para perfeccionar su desempeño en la detección de aquellos que deciden dejar el banco que en este caso seria mas del 70%.

Recall: El recall nos revela cuántos de los clientes que efectivamente decidieron dejar el banco fueron acertadamente detectados por nuestro modelo. En otras palabras, evalúa la habilidad del modelo para identificar a los clientes que están en peligro de irse.

En el caso de la clase 0 (No Exited), el modelo demostró su agudeza al reconocer con precisión al 97% de los clientes que decidieron permanecer fieles al banco.

En la primera clase (Exited), solo logró acertar en la identificación del 45% de los clientes que realmente decidieron dejar el banco.

Adicionalmente, el índice de recuperación para la clase 1 es notablemente inferior, lo que sugiere que el modelo se enfrenta a retos para identificar a aquellos clientes que realmente están en riesgo de marcharse. Esta es una zona donde el modelo aún tiene oportunidades para crecer y perfeccionarse.

La Métrica F1 El F1-score es una herramienta que fusiona de manera ingeniosa la precisión y el recall en una única cifra, capturando así la esencia de la efectividad de un modelo. Es valioso porque nos ofrece una perspectiva equilibrada entre la habilidad del modelo entre precisión y recall.

En la categoría 0 (No Exited), el F1-score alcanza un 0.92, lo que indica que el modelo se desempeña de manera excepcional al detectar a aquellos clientes que deciden quedarse.

En la Clase 1 (Exited), el F1-score dio 0.56, lo que se entiende como que el modelo tiene dificultades para identificar a aquellos clientes que deciden dejar el banco sin embargo si lo hace, podemos modelar aun mas el modelo.

Support; El soporte revela la cantidad de clientes que se asocian a cada categoría. Esto nos permite captar la magnitud de cada segmento de clientes y la manera en que influye en nuestras métricas.

Es importante recalcar que se puede apreciar que la cantidad de clientes que permanecen fieles supera a aquellos que deciden marcharse. La desigualdad en las clases puede mermar la habilidad del modelo para identificar con precisión a los clientes que deciden marcharse.

Categoría 0 (Clientes leales)	Categoría 1 (Clientes que se marcharon)
Un total de 2266 clientes han decidido quedarse con nosotros y no han abandonado el banco.	Un total de 537 clientes decidieron dejar el banco.

Promedios: En el informe se presentan dos variedades de promedios:

Promedio macro: Este cálculo ignora la cantidad de clientes en cada categoría. Considera únicamente cómo se desempeña el modelo en cada categoría, sin prestar atención al número de clientes que hay en cada una.

El promedio global de precisión que se sitúa en 0.82, lo que sugiere en términos generales, que el modelo se desempeña de manera bastante eficaz pero que aún queda espacio para la mejora.

El promedio general de recall se sitúa en 0.71, lo que indica que el modelo muestra una habilidad moderada para identificar a los clientes que deciden marcharse de la entidad bancaria.

El F1-score promedio se sitúa en un sólido 0.74, lo que indica que hay oportunidades para optimizar el modelo en la identificación de clientes que podrían estar dejando el banco.

Promedio ponderado: Este tipo de promedio considera la cantidad de clientes en cada categoría, y otorga más relevancia a aquellos grupos con mayor representación. Esto refleja de manera más fiel el desempeño auténtico del modelo en el conjunto de datos.

El promedio ponderado de precisión se sitúa en 0.86, lo que indica que, al tener en cuenta el tamaño de cada clase, el modelo exhibe una excelente efectividad general.

El promedio ponderado de recall se sitúa en 0.87, lo que sugiere que el modelo tiene un desempeño notable en la identificación de clientes, abarcando tanto a aquellos que deciden quedarse como a los que optan por marcharse. En líneas generales, los resultados son bastante alentadores para la organización.

El F1-score ponderado alcanza un notable 0.85, evidenciando que el modelo logra un excelente balance entre precisión y recall a lo largo de todo el conjunto de datos.

.

Conclusiones del Reporte de Clasificación

El reporte revela que el modelo se destaca con un rendimiento notable en líneas generales. Brilla con luz propia en la anticipación de aquellos clientes que deciden permanecer fieles al banco, logrando una notable precisión en sus pronósticos. Esto resulta crucial, dado que el modelo tiene la habilidad de reconocer con precisión a la gran mayoría de los clientes que continúan mostrando su lealtad.

Respecto a los clientes que deciden dejar el banco, el modelo también presenta una base sólida sobre la cual construir pero todavía hay oportunidades para perfeccionarlo. La exactitud es satisfactoria, pero hay margen para optimizar el recall, que es la habilidad de reconocer adecuadamente a los clientes que están abandonando. Este es un elemento fundamental que puede mejorarse con algunos ajustes ingeniosos, como modificar determinados parámetros o implementar técnicas para lograr un equilibrio más preciso en las predicciones.

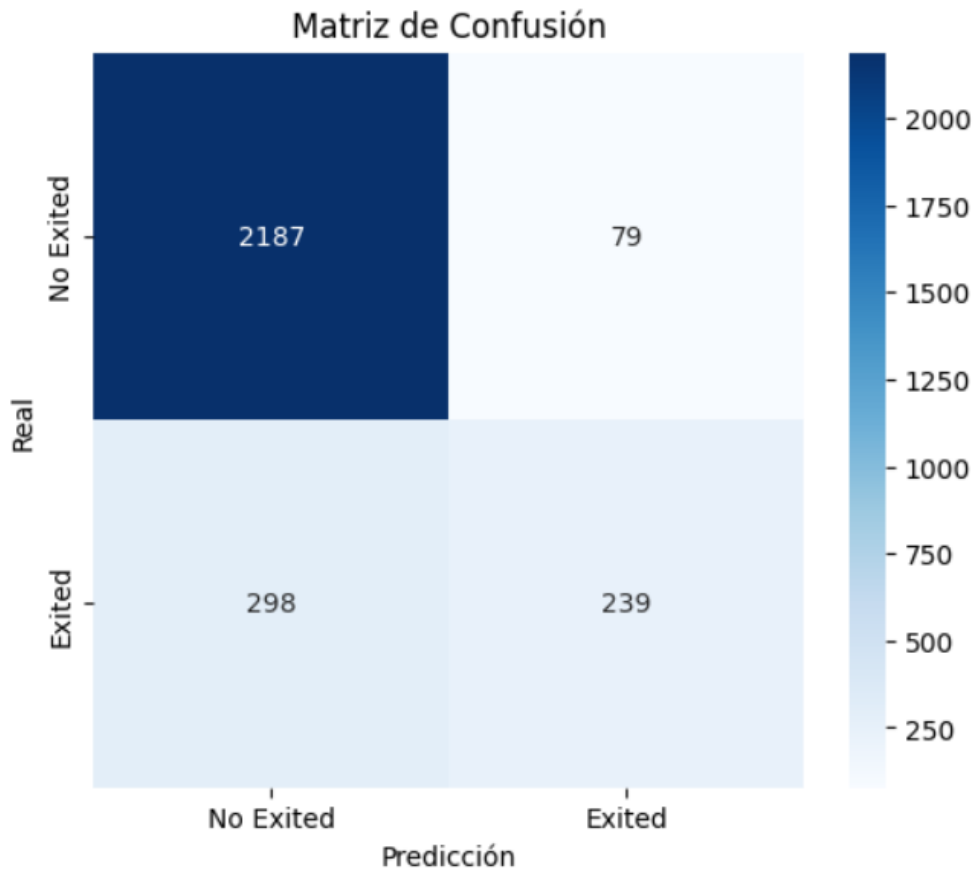
En resumen, el modelo está operando de manera eficiente y ha establecido una base robusta para anticipar la fidelidad de los clientes. Con unos pocos ajustes, se puede llevar la mejora a un nivel superior de resultado , sobre todo al detectar a los clientes que están en peligro de despedirse. Sin embargo está bastante bueno el resultado. Si se aplica lo mencionado se puede maximizar aun mas la retención de clientes.

Matriz de Confusión

Además, se genera una matriz de confusión que muestra detalladamente cuántas veces el modelo predijo correctamente y cuántas veces se equivocó es realmente muy útil esta herramienta para identificar si el modelo tiene dificultades con algún tipo de predicción.

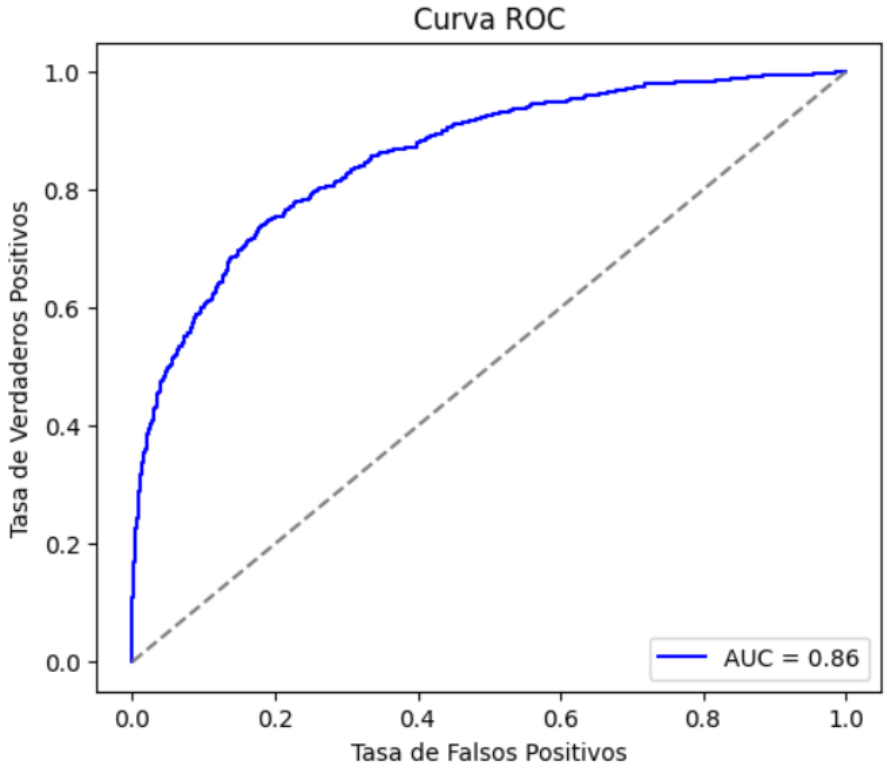
```python
# Matriz de confusión
from sklearn.metrics import confusion_matrix
conf_matrix_model = confusion_matrix(y_test_data, y_pred_model)
print("\nMatriz de confusión:")
print(conf_matrix_model)

# Visualización de la matriz de confusión
import seaborn as sns
plt.figure(figsize=(6, 5))
sns.heatmap(conf_matrix_model, annot=True, fmt='d', cmap='Blues', xticklabels=["No Exited", "Exited"], yticklabels=["No Exited", "Exited"])
plt.title("Matriz de Confusión")
plt.xlabel("Predicción")
plt.ylabel("Real")
plt.show()
```

Esta codificacion también genera una grafica para visualizar el análisis. Este gráfico muestra de forma clara los aciertos y errores del modelo, a continuación, se presenta el resultado.



Curva ROC

Este gráfica muestra como varían los verdaderos positivos frente a los falsos positivos. El propósito de esta curva muestra la relación entre la tasa de falsos positivos y la tasa de verdaderos positivos, en relación con el código este fue el resultado de grafico dado.



Importancia de las Características

El modelo también calcula la importancia de cada característica, lo que quiere decit que calcula cual es la relevancia de cada una de las características para hacer predicciones. Este paso nos ayuda a entender factores como balance o edad y verificar o entender mas a detalle cuales de las mismas son son más importantes para predecir si un cliente se irá o no de la entidad bancaria.

```
# Importancia de las características
import numpy as np
importances_model = xgboost_pipeline.named_steps['xgbclassifier'].feature_importances_
indices_model = np.argsort(importances_model)[::-1]
```

Exportación del Modelo

Como resultado final, el modelo se guardó en un archivo utilizando la librería joblib, eso para lograra que el modelo pueda ser cargado en el futuro sin problema algúno es decir sin la necesidad de volver a entrenarlo desde cero. Este proceso completa el flujo de análisis que buscamos al desarrollar el código de datos y del modelo de aprendizaje automático para predecir si un cliente las razones del cliente de irse o no.

Desde la carga de datos hasta la creación de un modelo entrenado, pasando por la exploración de los datos, la evaluación del rendimiento del modelo y la interpretación de los resultados, todo el proceso está diseñado para generar un modelo robusto y útil para hacer predicciones precisas.

```
# Exportación del modelo
import joblib

# pipeline completo
joblib.dump(xgboost_pipeline, '/content/drive/MyDrive/Colab Notebooks/pipeline_completo_XGBoost_v2.pkl')
print("Done")
```

Importancia del sistema predictivo y su función

Una vez que hemos identificado qué clientes están en riesgo de abandonar o quedarse por el código desarrollado, es crucial entender por qué la retención es tan importante. Primeramente hay que ser conscientes de que la retención de clientes no solo mejora la estabilidad financiera de la empresa, sino que también optimiza la relación con los clientes y proporciona una ventaja competitiva.

Reducción de costos: Atraer nuevos clientes generalmente requiere una inversión significativa en publicidad, marketing y otros recursos. Retener a los clientes actuales, por otro lado implica menos gasto y un retorno de inversión más alto.

Rentabilidad a largo plazo: Los clientes leales tienden a gastar más a lo largo del tiempo, o solicitando mas servicios del banco, lo cual beneficia significativamente a la entidad bancaria, ademas de la reputación que es tan importante.

Recomendaciones: Los clientes satisfechos pueden convertirse en defensores de la marca justo como lo mencioné al principio, la reputación de cualquier negocio es de suma importancia. Los clientes con tan solo recomendar los prestamos del banco, customer service, agilidad en los tramites etc, con su circulo cercano o bien por medio de redes sociales reduce la necesidad de gastar en marketing para adquirir nuevos clientes.

Mejora constante: Los clientes frecuentes proporcionan una valiosa retroalimentación que puede utilizarse para mejorar productos y servicios. Esta información ayuda a adaptar la oferta de la empresa y aumentar la satisfacción.

Estrategias para prevenir la deserción de clientes utilizando el modelo predictivo

Gracias al modelo predictivo que se uso y a el sistema que se desarrollo (explicado en paginas anteriores), el Banco puede prever qué clientes están en riesgo de abandonar y de esta manera aplicar estrategias preventivas para mejorar su experiencia y reducir la tasa de deserción. Metodos claves que se podrían implementar:

Intervenciones personalizadas: Para los clientes con alta probabilidad de abandonar, se puede estratégicamente diseñar ofertas personalizadas, como descuentos exclusivos o promociones especiales que incentivarán su permanencia.

Optimización de la experiencia del cliente: Al detectar a aquellos clientes que se sienten descontentos o que han tenido una vivencia poco satisfactoria, nos proponemos centrarnos en elevar su atención, brindando respuestas ágiles y efectivas para resolver sus inquietudes.

Encuestas y retroalimentación: Aplicar encuestas de satisfacción a los clientes en riesgo para comprender sus necesidades y preocupaciones. Esta información puede guiar mejoras específicas y hacer que los clientes se sientan valorados.

Ofrecer beneficios a clientes leales: Establecer iniciativas de lealtad que recompensen a aquellos clientes que demuestran un mayor compromiso, alentándolos a continuar eligiendo el banco una y otra vez.

Optimización de productos y servicios: Si el modelo muestra que ciertos clientes abandonan debido a fallos en el producto o servicio, se deben realizar mejoras o ajustes en lo que se ofrece asegurando que la empresa sigue siendo competitiva y atractiva.

Estrategias para reducir el porcentaje de abandono

Una vez que conocemos el porcentaje de clientes que tienen alta probabilidad de abandono, que afortunadamente en este caso en especifico es un porcentaje reducido, es posible implementar acciones estratégicas que permitan reducir dicho porcentaje.

Tales como:

Segmentación y comunicación dirigida: Con la ayuda del modelo predictivo ya desarrollado, segmentamos a los clientes en diferentes grupos según su probabilidad de permanencia. Esto nos permite comunicar de manera más eficiente ofertas personalizadas, recordatorios o promociones que mantengan el interés de los clientes.

Ajuste de precios: Para aquellos clientes que están a punto de irse debido a precios o falta de valor percibido, se pueden crear promociones especiales o descuentos para que sientan que están recibiendo más por su dinero.

Soporte proactivo: Si el modelo predice que un cliente tiene altas probabilidades de abandonar debido a problemas con el servicio al cliente, se pueden activar medidas proactivas como el seguimiento personalizado o la mejora en la rapidez de respuesta a sus consultas.

Monitoreo y actualización de estrategias: Es importante mantener actualizado el modelo predictivo y hacer ajustes en las estrategias de retención según los cambios en el comportamiento de los clientes. El monitoreo constante permite detectar cambios y adaptarse rápidamente.

Crear una experiencia única y memorable: Para evitar que los clientes abandonen por falta de conexión emocional con la marca, es necesario trabajar en la creación de una experiencia única y diferenciada que haga que los clientes no solo compren, sino que también se identifiquen con la empresa.

El uso de este tipo de tecnologías predictivas no solo mejora la retención, sino que también optimiza la toma de decisiones, lo que lleva a una gestión más eficiente de los recursos y un impacto positivo en los ingresos a largo plazo.

Este enfoque, utilizando modelos predictivos, no solo permite anticiparse a los problemas de retención, sino que también facilita la implementación de soluciones estratégicas que aumentan las probabilidades de que los clientes se queden, mejorando la competitividad y sostenibilidad de la empresa.

Importancia de la aplicación de graficas

Los gráficos cumplen con un rol imponte en el análisis, partiendo de UNIR NEWS MEXICO (2024) Los gráficos estadísticos son potentes herramientas para la visualización de datos que permiten representar de manera accesible información compleja. Consiguen presentar la información al usuario o el lector de manera clara y precisa, facilitando la comparación y la comprensión de la evolución de distintas variables. Es de suma importante entender y aplicar el rol de los graficos en el análisis de los datos, ya que ofrecen una representación visual que facilita su interpretación.

Para esta investigación especifica de deserción de clientes y modelos predictivos estos gráficos nos ayudan a analizar la manera en que los datos están distribuidos, identificar patrones de clientes al optar por el abandono, detectar posibles problemas como valores extraños, sesgos o tendencias que no son fáciles de ver a simple vista, comprender a detalle el PCA, entre otros.

Los graficos presentados conforme al avance de la investigación ayudan a visualizar los patrones y evaluar el desempeño del modelo usado en este cao XGoots. Los gráficos de PCA y K-Means presentados también nos dan una perspectiva visual de cómo los datos están distribuidos. Adicionalmente la matriz de confusión y la curva ROC fueron cruciales para

evaluar qué tan bien se logro el funcionamiento del modelo, proporcionándonos información detallada sobre las predicciones correctas o los errores cometidos en el desarrollo.

Es relevante destacar que es indispensable 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.

Una de las principales virtudes de emplear gráficos en el estudio de la deserción de clientes es su poder para desvelar de forma clara y ágil las tendencias temporales. Por lo cual se facilita la identificación de patrones estacionales y transformaciones en el comportamiento de los consumidores, lo que a su vez capacita a la empresa para tomar decisiones informadas y fundamentadas en datos, con el fin de afrontar estos desafíos actuales y futuros.

.

Conclusiones

1. Identificación de Factores Clave de Deserción: A través de la implementación del modelo predictivo, se puede identificar que ciertos factores son más determinantes en la deserción de clientes, tales como la frecuencia de uso de los servicios bancarios, el nivel de satisfacción, la calidad del servicio al cliente y el historial de interacciones con la entidad. Estos factores pueden ser utilizados para segmentar a los clientes y predecir con mayor precisión aquellos con mayor riesgo de abandonar el banco.
2. Mejora en la Toma de Decisiones Estratégicas: El modelo predictivo proporciona información valiosa para la toma de decisiones estratégicas dentro de la entidad bancaria. Al identificar a los clientes en riesgo de deserción, el banco puede aplicar medidas de retención personalizadas, lo que puede mejorar la fidelidad y reducir la tasa de abandono. Este sistema permite tomar decisiones más informadas y basadas en datos.
3. Eficiencia de los Modelos Supervisados: Los modelos supervisados aplicados, como regresión logística o máquinas de soporte vectorial (SVM), demostraron ser altamente efectivos para predecir la probabilidad de deserción. Estos modelos no solo alcanzaron una alta precisión, sino que también contribuyeron a la comprensión de las relaciones subyacentes entre las variables y la deserción, ayudando a obtener insights más profundos sobre el comportamiento de los clientes.

Recomendaciónes

1. Optimizar la Segmentación de Clientes: Utilizar los resultados de la segmentación realizada por el modelo para dirigir campañas de retención más específicas. En lugar de estrategias generales, personalizar las acciones de retención para diferentes grupos de clientes según sus características y comportamientos identificados.
2. Implementar Estrategias de Fidelización Tempranas: Basándose en las predicciones del modelo, los bancos deben implementar programas de fidelización y acciones preventivas con anticipación para los clientes con alta probabilidad de deserción. Estas acciones pueden incluir ofertas especiales, mejoras en la experiencia de usuario o incluso servicios personalizados.
3. Monitoreo Continuo de Modelos Predictivos: Los modelos predictivos deben actualizarse y monitorearse continuamente para garantizar que sigan siendo precisos a medida que los patrones de comportamiento de los clientes cambian con el tiempo. Esto puede incluir el reentrenamiento del modelo con nuevos datos y la validación periódica de las predicciones.
4. Mejorar la Calidad del Servicio al Cliente: Identificar aquellas interacciones con el cliente que afectan negativamente la probabilidad de deserción y mejorar la calidad del servicio ofrecido. Esto podría implicar capacitación adicional al personal o la implementación de canales de atención más eficientes y accesibles.

5. Fomentar Programas de Incentivos y Beneficios Exclusivos: Utilizar los resultados del modelo para ofrecer programas de incentivos dirigidos a los clientes con alta probabilidad de deserción. Estos pueden incluir beneficios exclusivos como tasas de

interés preferenciales, bonificaciones por lealtad, o acceso anticipado a productos financieros. Esto no solo aumentará la satisfacción del cliente, sino que también fortalecerá la relación con el banco, haciendo que los usuarios se sientan valorados y menos inclinados a abandonar la entidad.

Referencias bibliográficas

IBM. (s.f.). *Análisis de componentes principales*. IBM. Recuperado de https://www.ibm.com/es-es/topics/principal-component-analysis

UNIR México. (2024, enero 15). *Los gráficos estadísticos como herramienta de análisis visual*. UNIR México. Recuperado de https://mexico.unir.net/noticias/comunicacion-mercadotecnia/graficos-estadisticos/#:~:text=Los%20gr%C3%A1ficos%20estad%C3%ADsticos%20son%20potentes,la%20evoluci%C3%B3n%20de%20distintas%20variables.

Chen, T., & Guestrin, C. (2016). XGBoost: Un sistema de aumento de árboles escalable. *Actas de la 22ª Conferencia Internacional SIGKDD sobre Descubrimiento de Conocimiento y Minería de Datos*, 785–794. https://arxiv.org/abs/1603.02754

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324

Friedman, J. H. (2001). Aproximación de funciones codiciosas: Una máquina de aumento de gradiente. *Los Anales de Estadística*, 29(5), 1189–1232. https://projecteuclid.org/euclid.aos/1013203451

ESRI. (s.f.). *Cómo funciona XGBoost*. ArcGIS Pro. Recuperado de https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm

UB. (s.f.). *Análisis exploratorio de datos*. Recuperado de http://www.ub.edu/aplica_infor/spss/cap2-3.htm