MACHINE LEARNING

Diana Luzuriaga Calero

INDICE

I. RESUMEN

En este informe se llevará a cabo un análisis enfocado en la pérdida de clientes de una institución bancaria. Se desarrollará un modelo predictivo con el fin de identificar las variables que tienen un impacto significativo en la deserción de los consumidores. Para este propósito, se examinará una base de datos que contiene una variedad de información sobre los clientes del banco. A través de este análisis, se buscará determinar la influencia de estas variables en el objetivo, permitiendo así anticipar qué clientes tienen una mayor probabilidad de dejar de consumir en la empresa.

II. INTRODUCCIÓN

Uno de los grandes retos a los que se enfrentan los mercados es la captación y fidelización de clientes, ya que se encuentran en un momento en el que las empresas son cada vez más competitivas e incluso depredadoras en la captación de clientes.

Dentro de este contexto en el que el cliente es el centro de cualquier negocio, las organizaciones se ven con la necesidad de buscar nuevos caminos y dar un paso más en el análisis de su gestión a la hora de detectar y evitar a tiempo la fuga de clientes. Deben utilizar herramientas clave que les permitan estimar probabilidades de fuga de clientes para poder identificar el foco donde deben centrar el esfuerzo y evitar que sus clientes se vayan a la competencia. La empresa debe tener en cuenta que la inversión de captar nuevos clientes o de tratar de recuperarlos es mayor que retener a los clientes actuales.

En muchas ocasiones, la gestión diaria se ve obstaculizada por sistemas de información estáticos y complejos de analizar, que proporcionan una visión limitada de la realidad de las interacciones con los clientes. Por este motivo, se vuelve fundamental emplear análisis más sofisticados para identificar de manera precisa posibles desafíos no evidentes relacionados con la clientela.

La detección temprana de señales que apunten a una posible pérdida de clientes resulta crucial y la capacidad de reacción de las empresas es fundamental, ya que una vez que estos clientes optan por la competencia, recuperarlos se convierte en una tarea sumamente compleja y costosa desde el punto de vista económico.

Una solución es usar técnicas estadísticas combinadas con algoritmos de inteligencia artificial que permiten convertir los datos en información de valor para entender con mayor profundidad la realidad de los clientes y que permitan actuar a tiempo para evitar el abandono. Esta información es vital para la toma de decisiones tácticas y estratégicas a la hora de fidelizar la cartera de clientes, aportar ventajas competitivas respecto a la competencia y consolidar la relación cliente-servicio.

III. BASE DE DATOS

Para la realización de este análisis predictivo se utilizó una base de datos descargada de kaggle. Cuenta con 10.000 observaciones y 14 columnas incluida la del objetivo.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

La imagen anterior muestra los nombres de las columnas que se han utilizado para realizar el estudio. Cabe destacar que después del análisis se determinó que para la realización del estudio predictivo no se van a utilizar las siguientes variables, ya que no aportan información. Las variables que no se utilizarán son: RowNumber, CustomerId, Surname.

A continuación, en la siguiente imagen podemos observar que las variables y el target carecen de valores missings así como de valores nulos. Lo que si se han encontrado son valores atípicos que se verán más adelante.

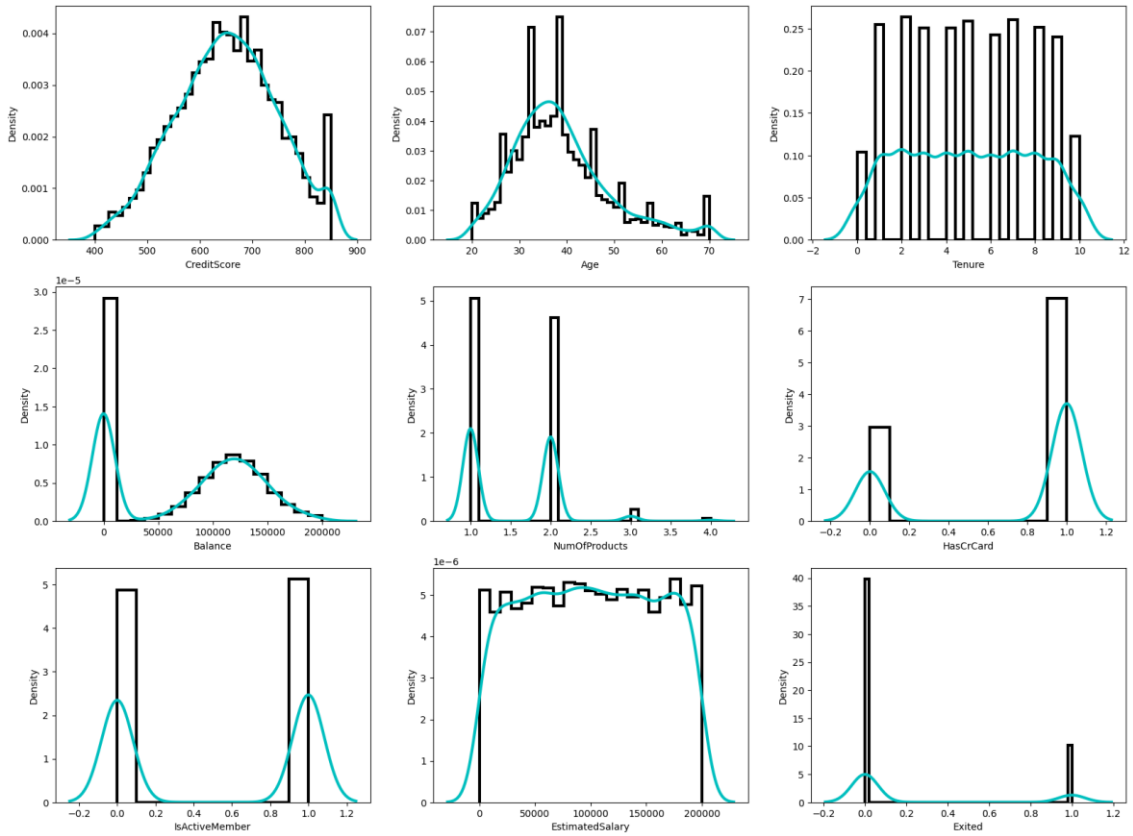
Todas las variables tienen datos numéricos que son necesarios en el análisis, excepto las columnas Geography y Gender. Debido a que en el estudio es necesario que todas las columnas sean numéricas se cambió el tipo de dato de estas columnas a numérico.

COL_N	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
DATA_TYPE	int64	object	object	int64	int64	float64	int64	int64	int64	float64	int64
MISSINGS (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
UNIQUE_VALUES	460	3	2	70	11	6382	4	2	2	9999	2
CARDIN (%)	4.6	0.03	0.02	0.7	0.11	63.82	0.04	0.02	0.02	99.99	0.02

IV. ANÁLISIS DE VARIABLES

A continuación, se llevará a cabo un análisis de las variables numéricas utilizadas en el estudio. En las imágenes que se presentan a continuación, se muestran los principales indicadores estadísticos de cada variable, así como las representaciones gráficas de sus distribuciones. Estos indicadores, junto con la visualización de la distribución de las variables, desempeñarán un papel crucial en el examen detallado de cada una de ellas y permitirán realizar transformaciones en los datos de las variables en caso de ser necesario.

	Mean	Median	Variance	Min	Max	Moda
CreditScore	651.647625	653.00	9.286445e+03	350.00	850.00	850.00
Age	38.897750	37.00	1.106484e+02	18.00	92.00	37.00
Tenure	5.003875	5.00	8.304898e+00	0.00	10.00	2.00
Balance	76102.139645	96447.52	3.901377e+09	0.00	250898.09	0.00
NumOfProducts	1.531375	1.00	3.360576e-01	1.00	4.00	1.00
HasCrCard	0.703500	1.00	2.086138e-01	0.00	1.00	1.00
IsActiveMember	0.512750	1.00	2.498687e-01	0.00	1.00	1.00
EstimatedSalary	100431.289764	100487.72	3.308426e+09	90.07	199970.74	24924.92

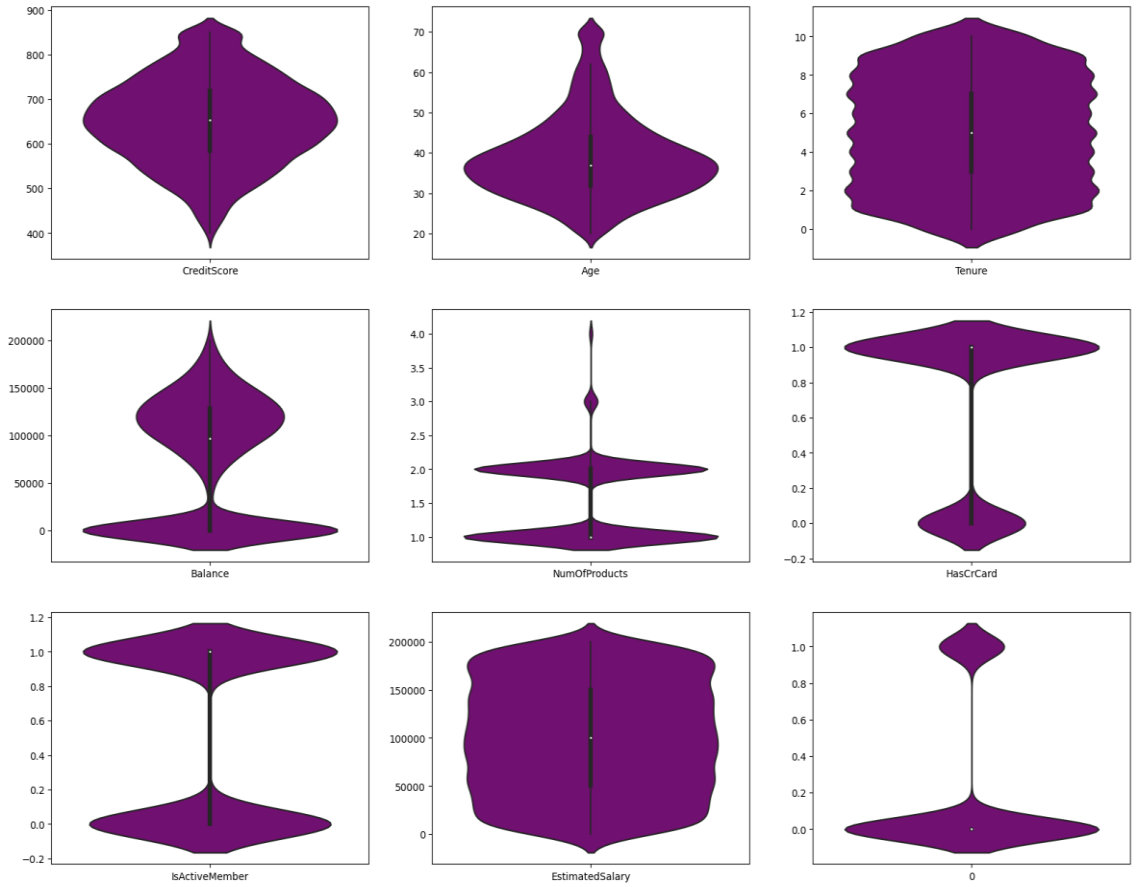


Un análisis que se puede obtener a través de estos datos es el siguiente:

- Respecto a la "Puntuación de Crédito", la media es de 651 y se observa una distribución sesgada hacia la izquierda, lo que indica la presencia de valores atípicos por debajo de 400. También es notable la concentración de puntuaciones entre 800 y 900.
- En cuanto a la "Edad" de los clientes, la media es de 39 años y se aprecia una distribución sesgada hacia la derecha. Esto sugiere que la mayoría de los clientes tienen más de 37 años y algunos incluso superan los 70 años, lo cual podría considerarse como valores atípicos.
- En la variable "Antigüedad", se puede observar que se distribuye en valores del 0 al 10, siendo el valor 2 el más recurrente. Los extremos, es decir, las antigüedades de 0 y 10 años son menos comunes.
- En cuanto al saldo o balance de la cuenta del cliente, la media es de 76.102. Los datos muestran una distribución bimodal, con una campana centrada en cero y otra que comienza alrededor de los 50.000 y se extiende hasta los 200.000. El valor central de esta última campana se sitúa entre 100.000 y 150.000. También se identifican valores atípicos por encima de 200.000.
- La variable "Número de Productos" presenta una distribución entre los valores 1 y 4, siendo los valores 1 y 2 los más comunes.
- En la variable "Tarjeta de Crédito", se observa una distribución entre los valores 0 y 1, destacando la mayor presencia del valor 1.
- La variable "Cliente Activo" muestra una distribución entre los valores 0 y 1, siendo el valor 1 el más repetido.
- Respecto a la "Estimación Salarial", los datos oscilan entre 90 y 200,000. La distribución es dispersa y la media y la mediana son prácticamente iguales, lo que sugiere una distribución casi simétrica.

- En la variable "Exited", se observa una división entre los valores 0 y 1, siendo el valor 0 el más común. Aquí se evidencia un desbalance en los datos del objetivo.

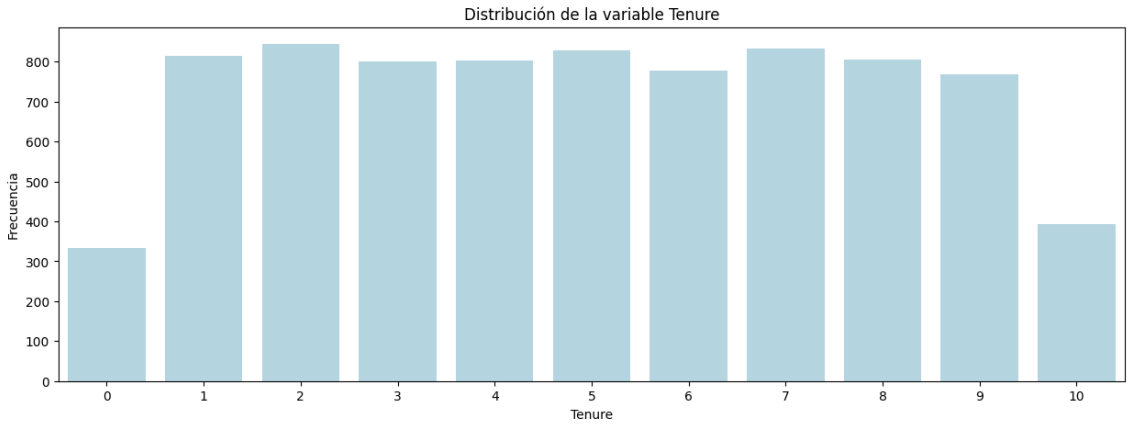
A continuación, vamos a ver representadas las variables a través de diagramas de violín.



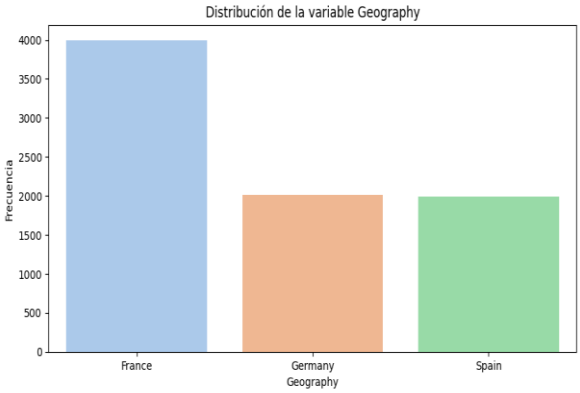
- En la "Puntuación de Crédito", notamos la presencia de valores atípicos por debajo de 400 y por encima de 800, mientras que la mayoría de los datos se distribuyen entre 600 y 700. Esto sugiere que las puntuaciones de crédito por debajo de 400 son menos comunes. La mayoría de los clientes del banco tienen puntuaciones alrededor de 650. También se observa una mayor concentración de puntuaciones de 850 en comparación con las de 800. Esto indica que hay puntuaciones que son atípicas dentro del conjunto de clientes.
- En cuanto a la "Edad", vemos que la distribución es menor por encima de los 50 años, y podrían existir valores atípicos a partir de los 70 años. También notamos un pico por debajo de los 20 años, que podrían ser considerados como valores atípicos. La mayoría de los clientes tienen más de 35 años, y el banco cuenta con un buen número de clientes mayores de 50 e incluso algunos mayores de 70.
- En la "Antigüedad", no se observa una distribución clara; los datos están dispersos entre los valores mencionados. Esto indica que no hay una distribución evidente en los valores analizados, y las puntuaciones de 0 y 10 son menos comunes.
- Con respecto al "Balance", vemos que los datos se concentran alrededor del valor 0 y la mediana, que es aproximadamente 100,000. También observamos posibles valores atípicos a partir de 200,000. Esto indica que una parte significativa de los clientes no tiene saldo en sus cuentas, mientras que la otra parte tiene más de 50,000, e incluso algunos tienen un saldo superior a 200,000.

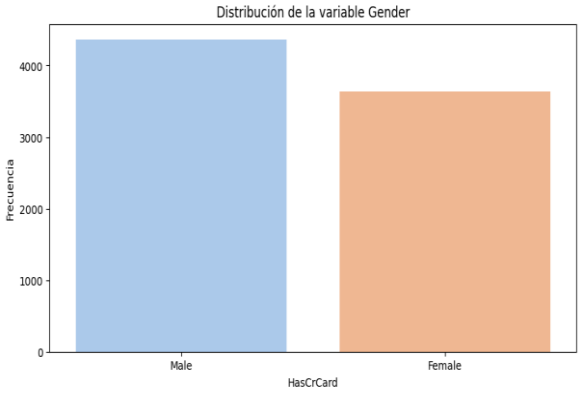
- En cuanto al "Número de Productos", notamos una mayor concentración en los valores 1 y 2, mientras que los valores 3 y 4 son menos frecuentes. Esto sugiere que la mayoría de los clientes no tienen tantos productos de la empresa, prefiriendo quedarse con 1 o 2.
- En la variable "Tarjeta de Crédito", los valores se distribuyen principalmente entre 0 y 1, con una mayor frecuencia del valor 1. Esto indica que la gran mayoría de los clientes poseen una tarjeta de crédito.
- Respecto a "Cliente Activo", la distribución se observa entre los valores 0 y 1, con una mayor repetición del valor 1. Esto sugiere que hay más clientes activos en la empresa, pero no existe una gran diferencia en la proporción de clientes activos e inactivos
- En la "Estimación Salarial", notamos posibles valores atípicos mayores a 200,000 y algunos valores iguales a 0. Esto indica que la estimación salarial de los clientes abarca un amplio rango, y no se puede determinar visualmente un rango específico en el que se mueve la variable.
- En la variable "Salida", observamos que el valor 0 es el más frecuente, lo que indica que aproximadamente el 80% de los clientes permanecen en la empresa.

En el siguiente gráfico se puede ver de forma más clara la distribución de la variable antigüedad, ya que, con los gráficos presentados antes, no se podía ver con claridad la distribución de esta.



A continuación, se va a realizar el análisis de las variables categóricas.



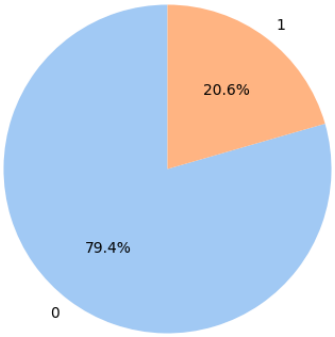


- En cuanto a la variable "Geografía", podemos observar la presencia de tres categorías de países: Francia, Alemania y España. La categoría más predominante es Francia, seguida de cerca por Alemania y España. Esto sugiere que la mayoría de los clientes provienen de Francia.

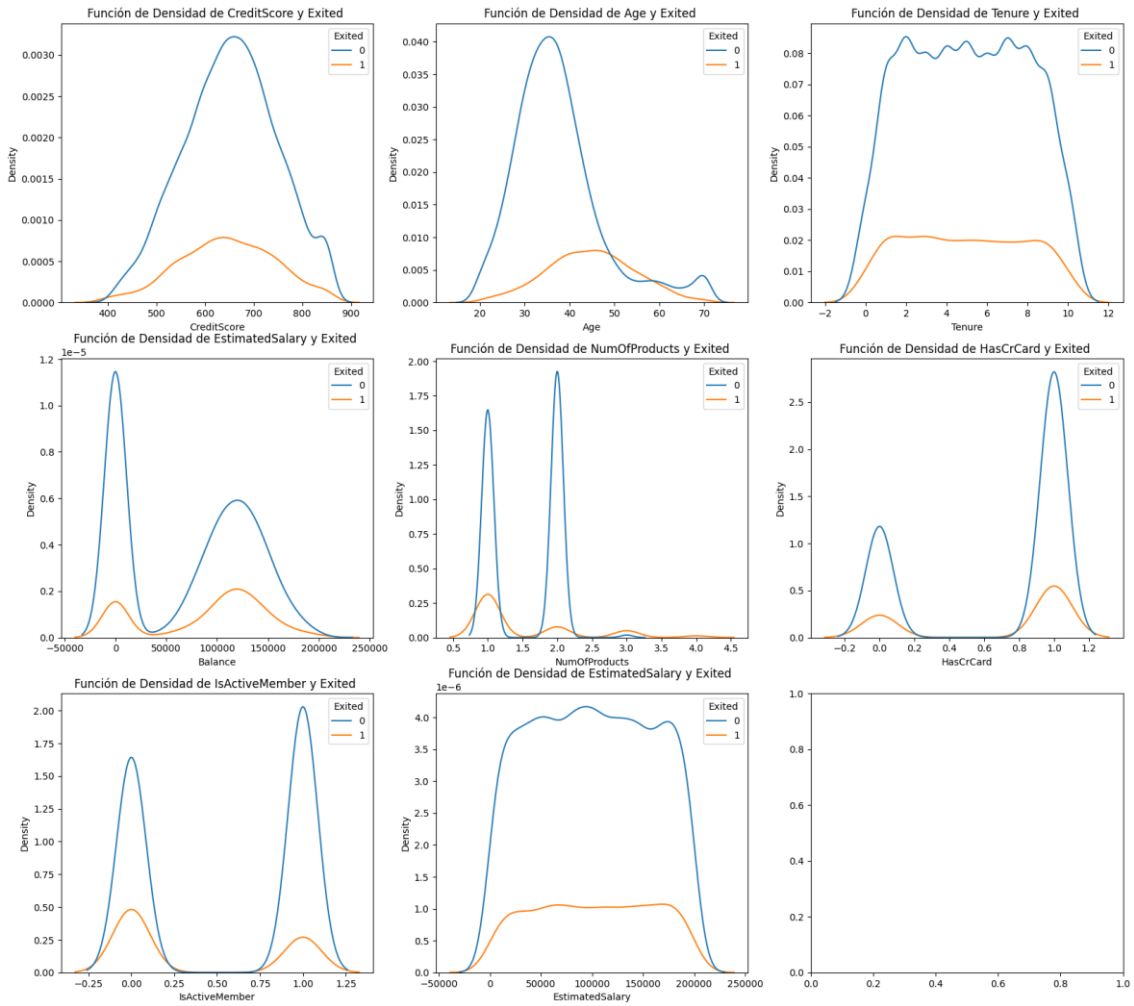
- Respecto a la variable "Género", notamos que la categoría masculina es la más representada, con alrededor de 4.000 personas. Por otro lado, la categoría femenina cuenta con aproximadamente 3.500 personas. Esto indica que la mayoría de los clientes son de género masculino.

Análisis del objetivo

La variable que estamos analizando tiene dos categorías: "fuga" representada por 1 y "no fuga" representada por 0. Como se puede apreciar en la gráfica adjunta, el 20% de los datos pertenecen a la categoría de clientes que se han dado de baja, mientras que el 80% restante corresponde a aquellos que se han mantenido. Esto confirma la observación previa sobre la variable y demuestra que los datos en esta categorización no están equilibrados.

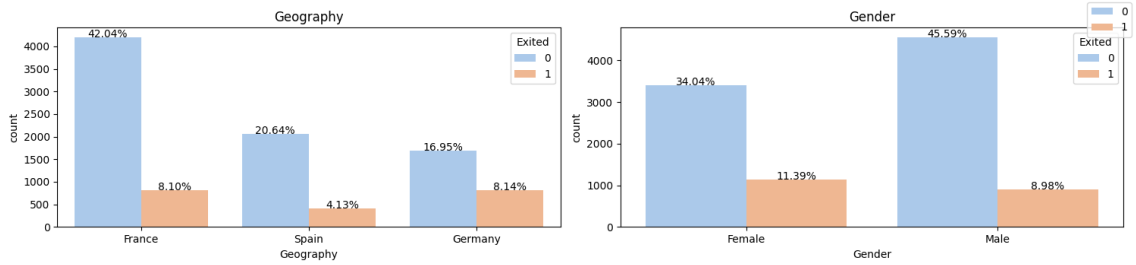


A continuación, se va a representar a través de un gráfico la relación de la variable objetivo con el resto de las variables.



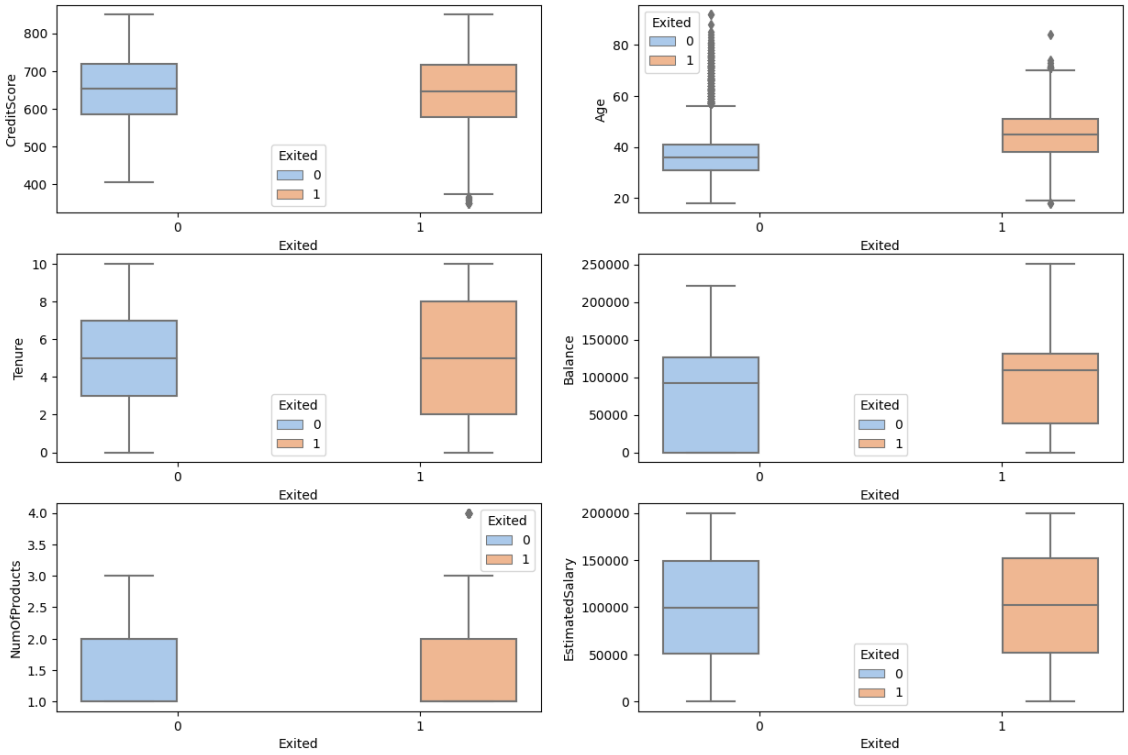
- Respecto a la "Puntuación de Crédito", se aprecia que los clientes con una puntuación inferior a 400 son quienes optan por darse de baja. La media de puntuación de los clientes que se han dado de baja es de aproximadamente 630, lo cual es menor que la media general de la variable. Visualmente, se nota que los clientes con puntuaciones superiores a la media son los que presentan mayor riesgo de darse de baja.
- En cuanto a la "Edad", la media de edad de los clientes que han decidido darse de baja es de alrededor de 45 años, lo cual es superior a la media general de la variable. Además, se observa que los clientes mayores de 65 tienden a quedarse. El enfoque debería centrarse en los clientes con edades comprendidas entre los 35 y los 55 años.
- En relación con la "Antigüedad", no se evidencia una relación clara entre esta variable y la probabilidad de darse de baja. La densidad de datos es similar en ambos casos.
- Respecto al "Balance", se observa una distribución similar entre los clientes que se han dado de baja y la variable en general. Sin embargo, es importante destacar que hay más clientes que se han dado de baja en el rango de saldos entre 100,000 y 150,000.
- En cuanto al "Número de Productos", se nota que los clientes que se han dado de baja suelen tener contratado solo un producto, seguidos por aquellos que tienen dos.
- En lo que respecta a la posesión de una "Tarjeta de Crédito", se observa que los clientes que tienen tarjeta son quienes se han dado de baja.
- Con respecto a ser un "Cliente Activo", se puede notar que aquellos que no muestran actividad en la empresa son los que deciden abandonarla.
- En cuanto a la "Estimación Salarial", no se percibe una relación clara entre esta variable y la decisión de los clientes de darse de baja.

En cuanto a las variables categóricas:



- Al analizar la variable "Geografía", se observa que hay más clientes que se han dado de baja en Alemania y Francia en comparación con España.
- En cuanto a la variable "Género", se nota que las clientes de género femenino presentan un mayor riesgo de darse de baja.

A continuación, se va a analizar la siguiente imagen en donde se puede observar la relación de las variables con el objetivo.



La distribución de la puntuación de crédito muestra que no existe una diferencia significativa entre los clientes que permanecen en el banco y los que se han dado de baja.

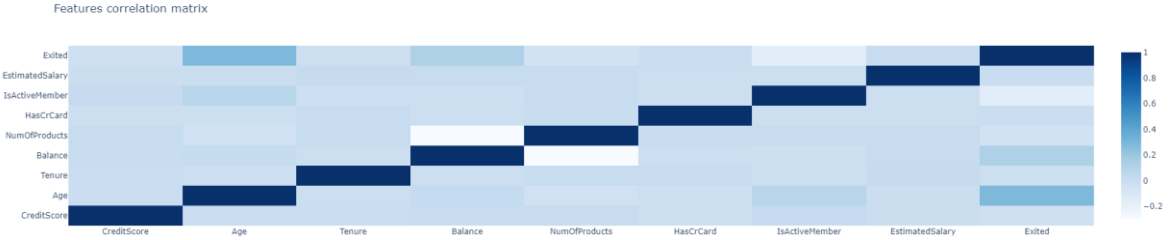
Sin embargo, se observa que los clientes de mayor edad tienen una tasa de cancelación más elevada en comparación con los más jóvenes. Esto sugiere que podría existir una preferencia de servicio diferenciada según grupos de edad. En consecuencia, el banco podría considerar la revisión de su estrategia de retención y su enfoque en diferentes segmentos de edad.

En cuanto a la antigüedad de los clientes, aquellos que se encuentran en los extremos, es decir, quienes han tenido una relación bancaria muy corta o muy larga, muestran una mayor propensión a darse de baja en comparación con aquellos con una antigüedad promedio.

De manera preocupante, el banco está perdiendo clientes que poseen saldos bancarios significativos, lo que podría tener un impacto en la disponibilidad de capital para préstamos en el futuro.

Por último, se destaca que ni el tipo de producto utilizado ni el nivel de salario tienen un efecto significativo en la probabilidad de cancelación de los clientes.

Relación lineal entre las variables y el objetivo



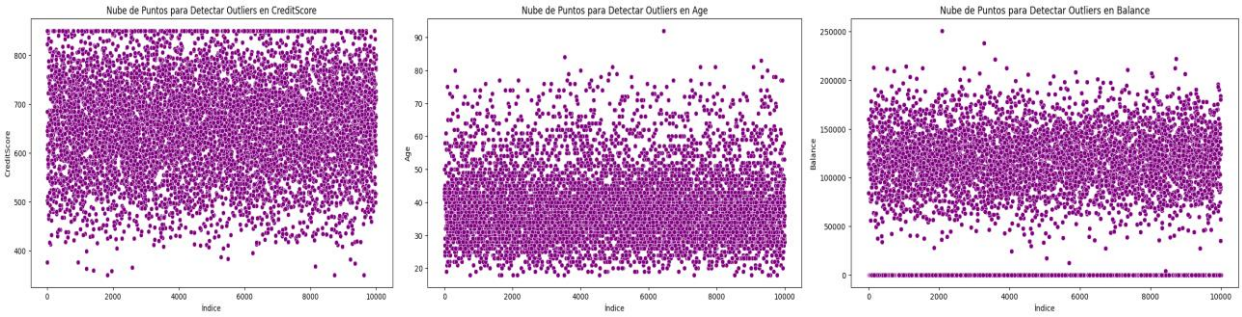
La imagen anterior nos revela que el objetivo tiene una correlación positiva con las variables edad y saldo, y una correlación negativa con el estado de miembro activo. Esto sugiere que a

medida que aumenta la edad y el saldo, hay una tendencia a la fuga de clientes. También es notable la correlación negativa entre las variables saldo y número de productos.

V. ANÁLISIS PREDICTIVO

Preprocesado de datos: Eliminación de datos atípicos en las variables

Antes de iniciar el análisis, se llevó a cabo la identificación y eliminación de valores atípicos en las variables de puntuación de crédito, edad y balance. Como se puede apreciar en la imagen siguiente, se identificaron valores que se desviaban de la tendencia general, confirmándose así los hallazgos previos en la visualización de las funciones de densidad de las variables. Estos valores atípicos serán sustituidos por la mediana con el fin de que sean considerados en el análisis.



Preprocesado de datos: Equilibrar los datos del objetivo

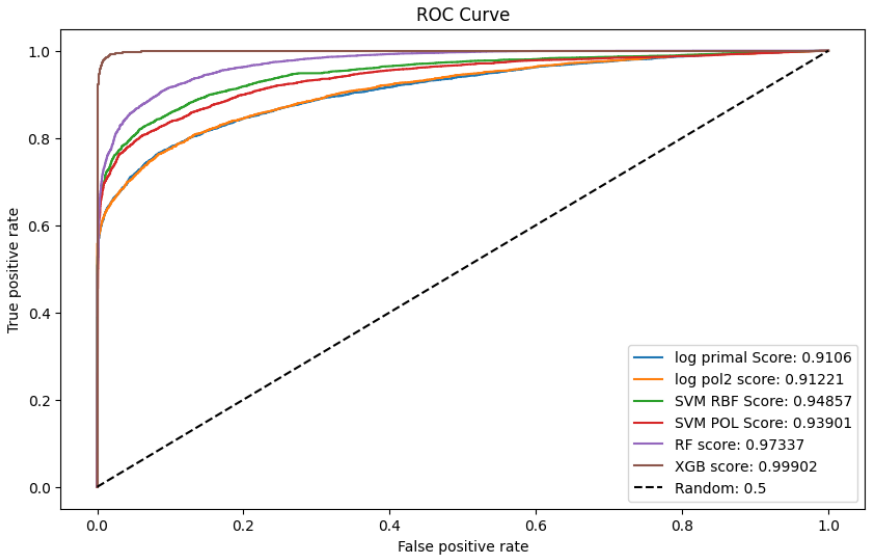
En el análisis del objetivo se observó un desbalance en los datos, con solo un 20% de casos etiquetados como 1 y un 80% como 0. Para abordar este desbalance y mejorar la precisión de la predicción, se procedió a balancear los datos. De esta forma, para la predicción de fuga de clientes se cuenta con un equilibrio de datos del 50% de casos 1 y 50% de casos 0.

Procesado de datos: Resultados de los modelos en entrenamiento:

Modelos		Precisión	Recall	F1-Score	Accuracy	AUC
Regresión Logística	0 1	0.80 0.89	0.90 0.77	0.85 0.83	0.84	0.9106
Regresión Logística con extensión polinómica	0 1	0.80 0.89	0.90 0.77	0.85 0.83	0.84	0.91221
Vector Soporte con kernel rbf	0 1	0.84 0.93	0.94 0.82	0.89 0.87	0.88	0.94857
Vector Soporte con kernel polinómico	0 1	0.83 0.93	0.94 0.80	0.88 0.86	0.87	0.93901
RandomForest	0 1	0.89 0.92	0.93 0.89	0.91 0.91	0.91	0.97337
XGBClassifier	0 1	0.98 0.99	0.99 0.98	0.99 0.98	0.98	0.99902

La elección del modelo se enfocará en la métrica de calidad (accuracy) para elegir el modelo que mejor predice la probabilidad de fuga de un cliente. Se elige esta métrica dado que previamente se balancearon las categorías durante el preprocesado de los datos. De esta manera, se puede

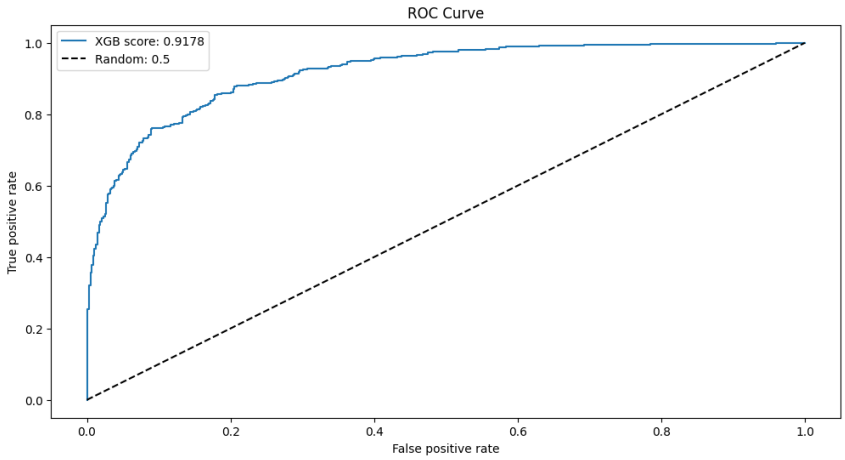
observar que los modelos XGBClassifier y RandomForest son los más destacados, con un 98% y un 91% de precisión en la predicción de la fuga de clientes, respectivamente. Optamos por el modelo RandomForest para una mejor comprensión y análisis.



La gráfica anterior nos ofrece una visión de la habilidad de los modelos para discernir entre los clientes que han abandonado y aquellos que no lo han hecho. Destacan el XGBClassifier y el RandomForest, con una puntuación de 0.99 y 0.97 respectivamente, acercándose a un puntaje perfecto de 1. Es importante notar que los otros modelos también muestran buenos resultados, sin embargo, nos inclinaremos por el RandomForest.

VI. RESULTADO DEL MODELO ELEGIDO

Modelo			Precisión	Recall	F1-Score	Accuracy	AUC
RandomForest	Train	0	0.89	0.93	0.91	0.91	0.9733
		1	0.92	0.89	0.91		
	Test	0	0.82	0.93	0.87	0.84	0.9178
		1	0.87	0.70	0.78		



El análisis de fuga de clientes está basado en la elección del modelo RandomForest. Durante el proceso de entrenamiento, este modelo demostró una precisión del 91% en el conjunto de datos de entrenamiento, lo que significa que fue capaz de acertar la condición de fuga o no fuga en un 91% de los casos dentro de este conjunto. Al evaluar el modelo con un conjunto de datos de test que no formó parte del proceso de entrenamiento, se obtuvo una precisión del 84%. Esto indica que el modelo se desempeña bien al generalizar sus predicciones a nuevos datos y no ha caído en el sobreajuste al conjunto de entrenamiento.

Adicionalmente, el AUC en el conjunto de entrenamiento fue de 0.9733, lo que sugiere que el modelo tiene una alta capacidad de discriminación entre las clases de fuga y no fuga en este conjunto. En el conjunto de prueba, el AUC fue de 0.9178, lo que confirma que el modelo también es efectivo en la discriminación de las clases en datos no vistos previamente.

Estos resultados indican que el modelo RandomForest es una elección sólida para predecir la fuga de clientes. Su buen rendimiento en el conjunto de prueba sugiere una capacidad de generalización adecuada y una alta habilidad para distinguir entre clientes que probablemente se darán de baja y aquellos que no lo harán.