Modelado de préstamos personales

¿PODREMOS CONVENCER A TODOS LOS CLIENTES QUE SOLICITEN UN PRÉSTAMOS EN THERA BANK?

EQUIPO

AUGUSTO BARCHI

CODERHOUSE

ÍNDICE

INTRODUCCIÓN

Thera Bank es un banco estadounidense que tiene una base de clientes en crecimiento. La mayoría de estos clientes son clientes pasivos (depositantes) con distintos tamaños de depósitos. El número de clientes que también son prestatarios (clientes de activos) es bastante pequeño, y el banco está interesado en expandir esta base rápidamente para generar más negocios de préstamos y, en el proceso, ganar más a través de los intereses de los préstamos. En particular, la gerencia quiere explorar formas de convertir a sus clientes pasivos en clientes de préstamos personales (mientras los retiene como depositantes).

OBJETIVOS

Como científico de datos en Thera Bank, debemos determinar las características mas significativas del cliente, con estudios de estadísticas para medir la precisión o el error que pueda tener el modelo y que sea alcanzable mediante iteraciones, que ayude al departamento de marketing, a identificar y predecir con una mayor probabilidad, a clientes potenciales de solicitar un préstamo, mediante un análisis de datos, que llevará un tiempo en que se realizará este proyecto, siendo este de 7 meses.

HIPÓTESIS

- ¿Qué variables son las más significativas?
- ¿La región o condado son variables significativas?
- ¿A qué segmento de clientes debería dirigirse más?
- ¿La edad tiene algún impacto en el préstamo de compra del cliente?
- ¿Las personas con menos ingresos piden préstamos?

BASE DE DATOS

La Edad de los clientes van de los 23 a los 67 años con una Experiencia de 1 a 45 años trabajando, con un Sueldo que van de los 8 a 885 k U$S, teniendo diferentes condados donde viven, también el dataset posee datos de la cantidad de Miembros de la Familia que van de 1 a 4, otro de los datos personales es la distintas Educación que poseen los clientes, Secundaria, Universitario y Master.

Ademas de ello, estos clientes poseen datos como Gasto de tarjeta de crédito que van de 0 a10 k en dólares mensuales con o sin Hipoteca que van de 0 a 700 k U$S y Tarjeta de Crédito de otro banco, entre otros datos menores.

Aquí lo que se hace es chequear que no posean datos nulos ni falta de datos, que se vieron completos en un análisis preliminar.

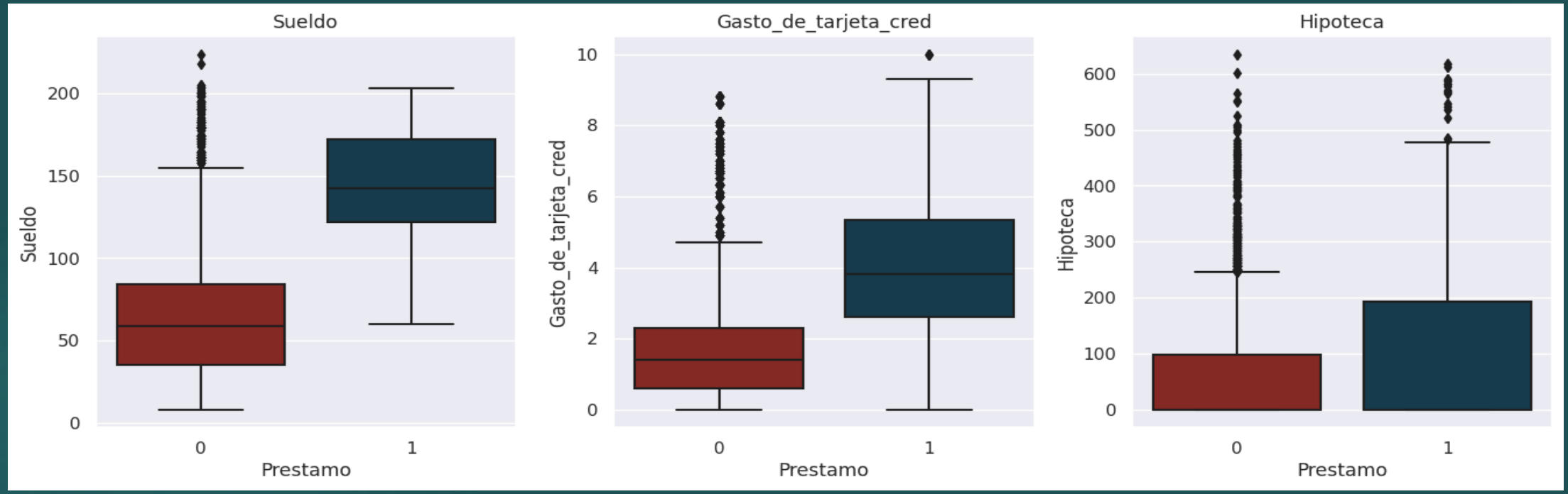
BASE DE DATOS

El dataset elegido posee 5000 Filas × 14 columnas
Edad [23-67]
Experiencia [1-45]
Sueldo [8 – 885 k]
Código Postal [varia]
Miembros de Familia [1-4]
Gasto de tarjeta de crédito [0-10 k]
Educación [1: Secundaria, 2: Universitario, 3: Master]
Hipoteca [0-700 k]
Préstamo hecho en la ultima campaña[0 o 1]
Cuenta de seguridad [0,1]
Cuenta de deposito [0,1]
Online [0,1]
Tarjeta de Crédito de otro banco [0,1]
Condado [nombres]

ESTUDIO DE LAS EDADES

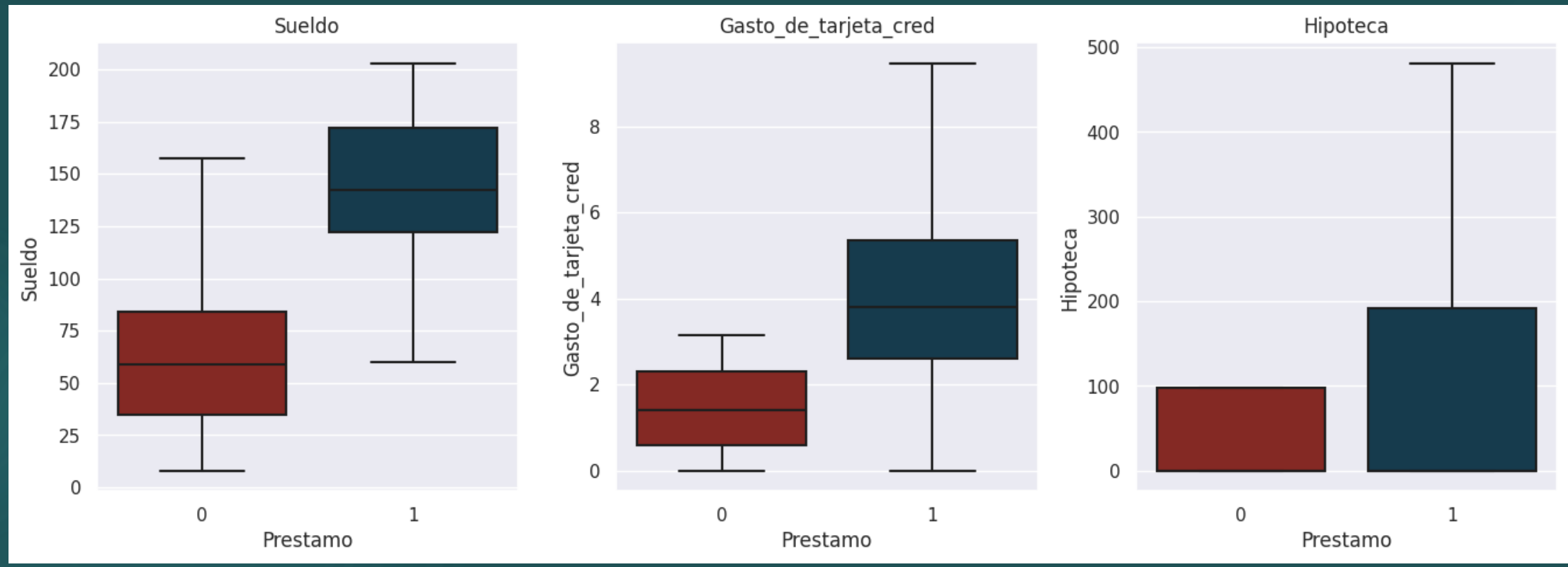
- La edad de los clientes está en el rango de 23 a 67 años, con una media y una mediana de ~45 años.
- La experiencia máxima es de 43 años. donde la media y la mediana son ~20.
- Los ingresos están en el rango de 8k a 224k USD. La media es 73k USD y la mediana es 64k USD. 224 El salario máximo debe verificarse.
- La hipoteca máxima tomada es de 635k USD. Necesito verificar esto.
- El gasto promedio en tarjeta de crédito por mes oscila entre 0 y 10.000 con una media de 2.500 USD y una mediana de 2.000 USD.

DETECCIÓN DE CLIENTES ATÍPICOS



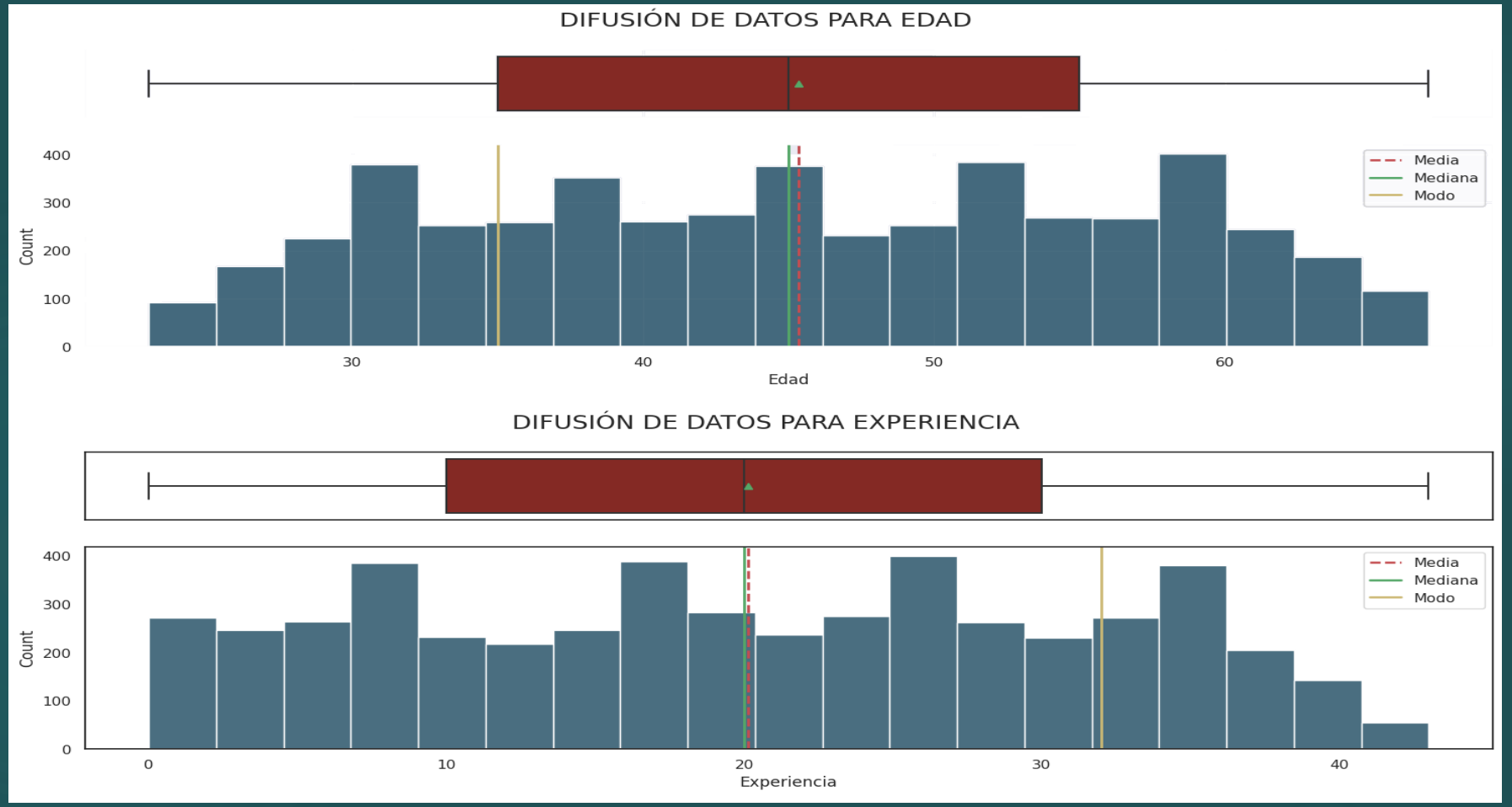
Se han identificado algunos ingresos extremadamente altos, de hasta 224.000 USD, en comparación con los ingresos promedio de personas del mismo grupo de edad y experiencia, dependiendo si solicitaron o no un préstamo alguna vez. Aunque los gastos en tarjetas de crédito e hipotecas parecen estar dentro de los límites aceptables, es importante abordar estos valores atípicos. De esta manera, se garantiza una mayor coherencia en los datos y se evita que estos valores extremos distorsionen los resultados.

DETECCIÓN DE CLIENTES ATÍPICOS



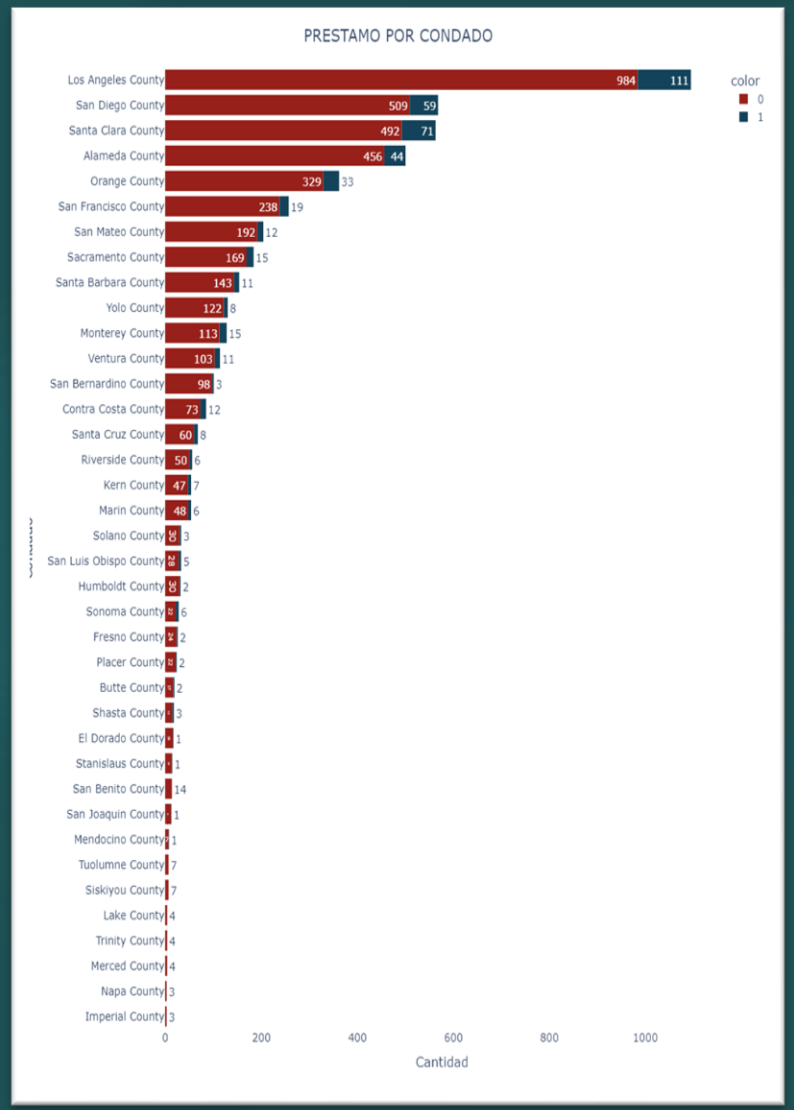
Dado que se encontraron múltiples casos de valores atípicos, solamente de los casos máximos, se los identifico y se los trató, se ha decidido reemplazarlos estos con los valores máximos, dependiendo si el individuo solicitó o no un préstamo, ya que esos clientes tienen un poder adquisitivo mas alto que el promedio, quedando de esta manera.

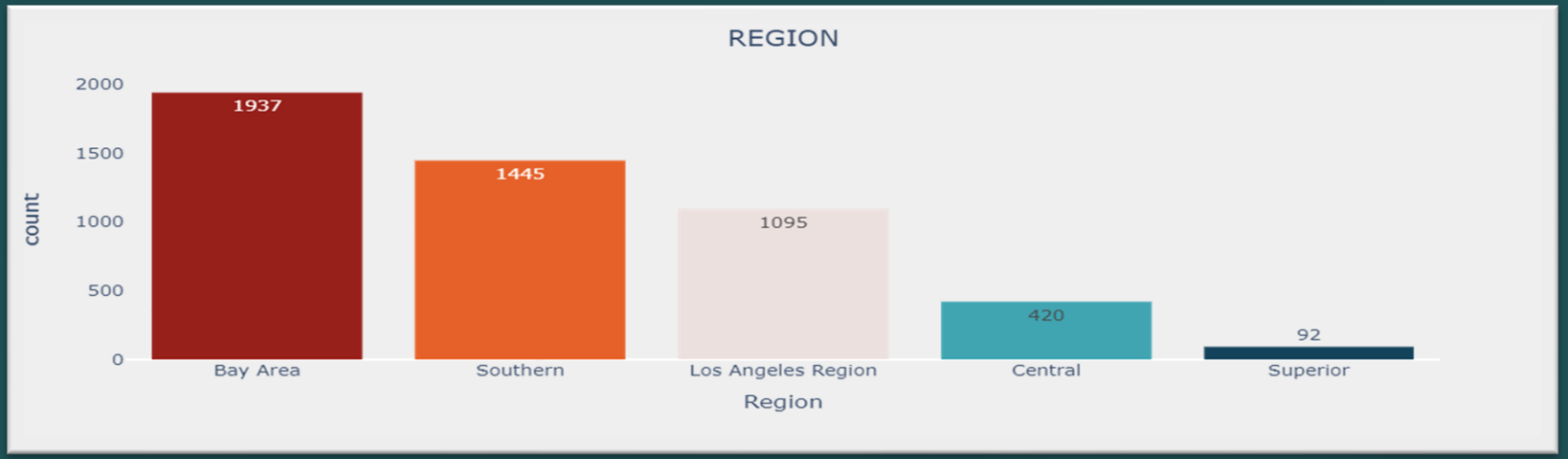
ANALIZANDO LA DIFUSIÓN DE LOS DATOS



Tanto la edad como la experiencia tienen la misma distribución con pico en 5 puntos, o sea que son fuertemente relacionales.

CANTIDAD DE PERSONAS POR REGIONES

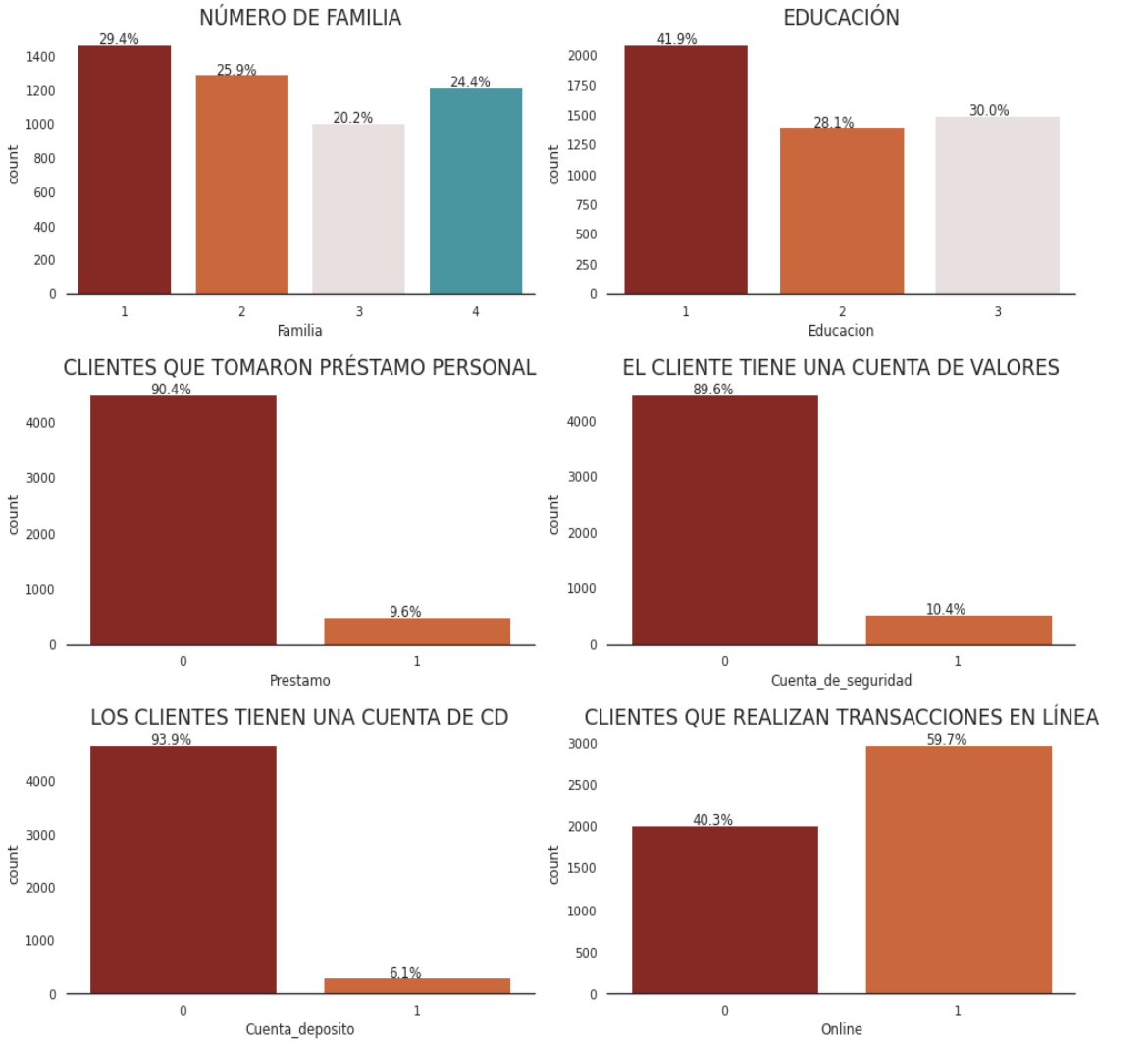


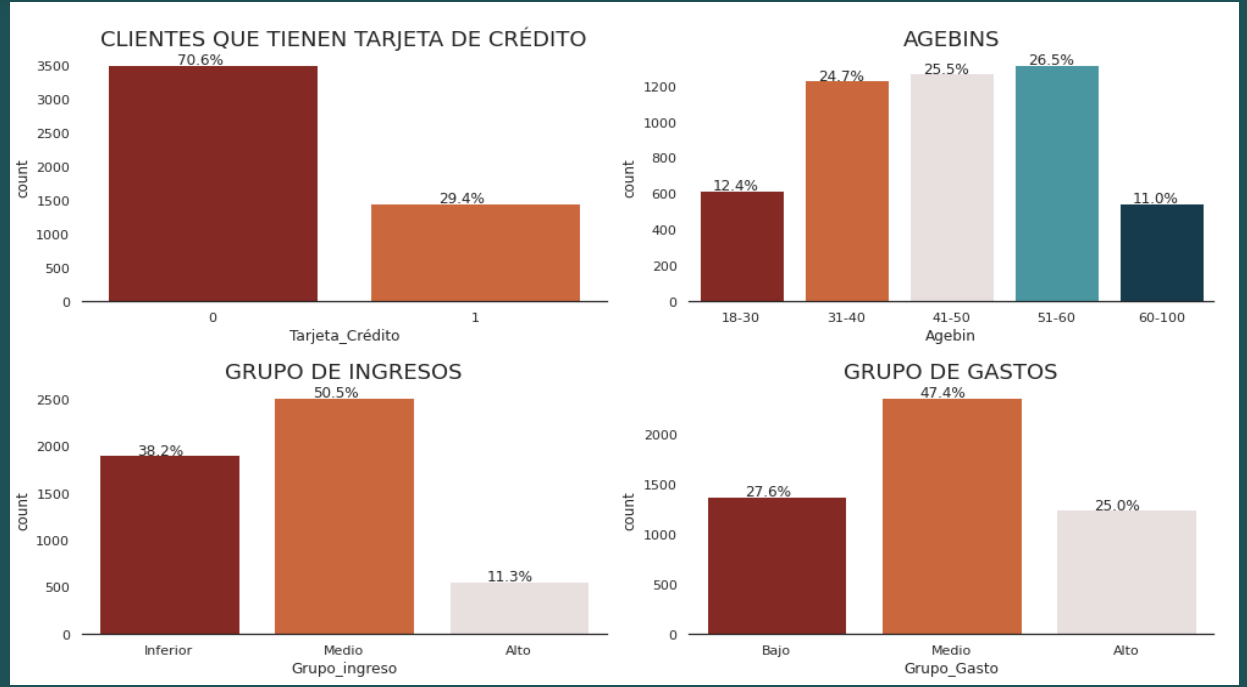


Como podemos apreciar en el condado de Los Ángeles, San Diego y Santa Clara, son los principales condados en la cual haya mayor cantidad de clientes y por ende los condados donde se podría hacer mayor énfasis, en que los clientes soliciten un préstamo. Esto se lo estudia agrupando la cantidad de clientes por región.

Haciendo varios análisis de los condados, vemos que el condado Los Ángeles, es una región, por ende es unos de los candidatos en hacer mayor énfasis, ya que es uno de los que mas piden prestamos por cantidad de cliente.
Otro condado son los de la región de Bay Área, que sumados es la región mas grande de las 5, que es a la cual se debería hacer mayor énfasis que las demás, en que los clientes soliciten un préstamo.

GASTOS





- ~29,4 % de los clientes son solteros.
- ~41.9% de los clientes son estudiantes universitarios.
- ~9.6% compró un préstamo personal del banco.
- El 10,4 % de los clientes tiene cuenta de valores en el banco
- El 6 % de los clientes tiene una cuenta de CD.
- El 60% de los clientes realizan transacciones en línea.
- El 29,4% de los clientes tienen tarjetas de crédito.
- ~ 75 % de los clientes están en el rango de 31-60.
- ~ 50 % La mayoría de los clientes bancarios pertenecen al grupo de ingresos medios.
- ~48 % de los clientes tiene gasto medio.

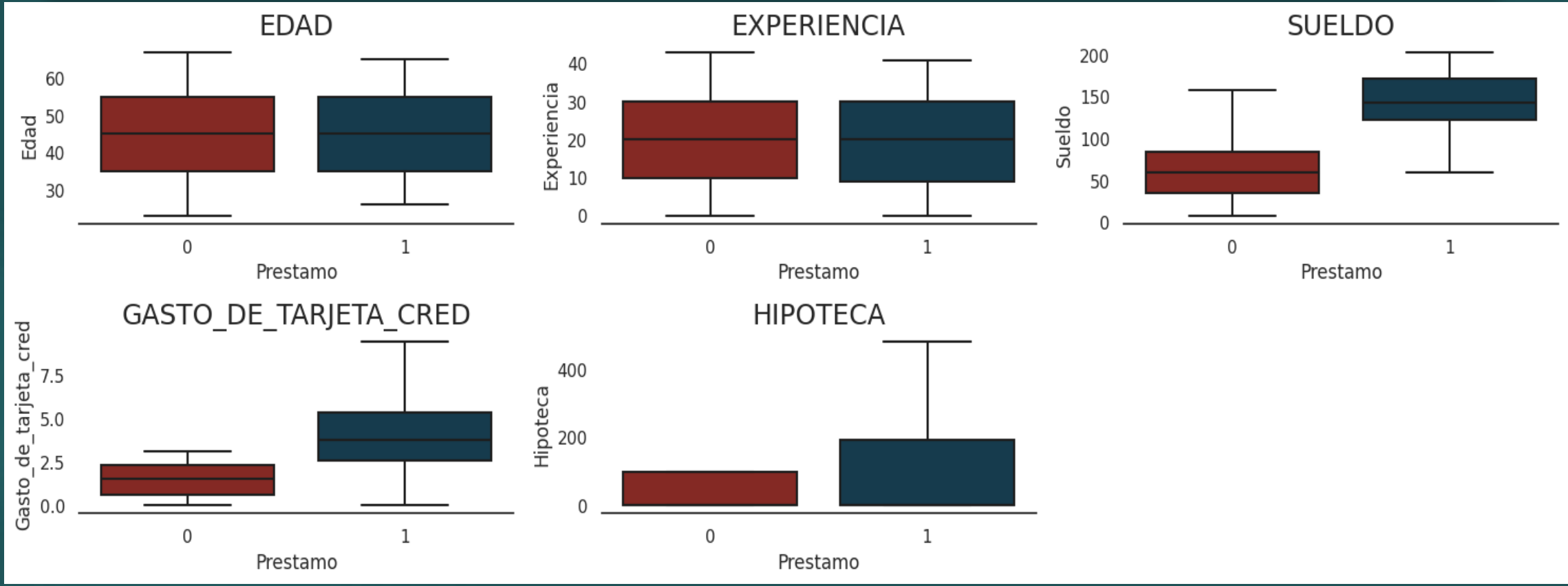
HEATMAP

Como era de esperar, la edad y la experiencia están altamente correlacionadas y uno de ellos puede descartarse. Como tuvimos que manejar 0, se descartará la experiencia.

Los ingresos y el gasto promedio en tarjeta de crédito están correlacionados positivamente.

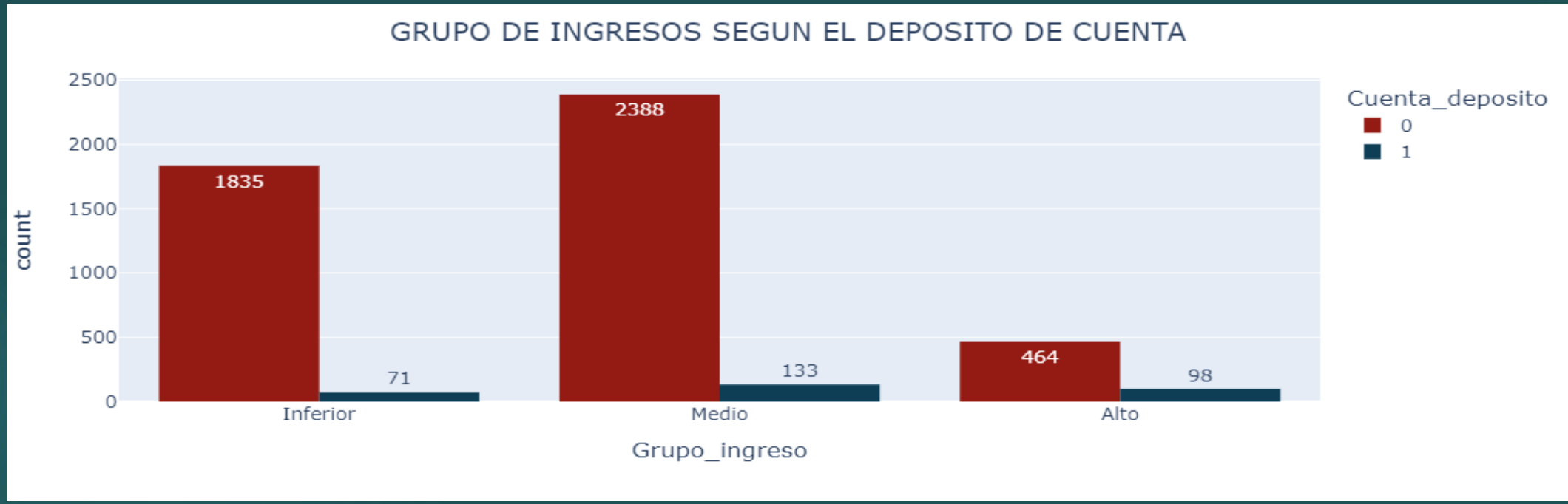
La hipoteca tiene muy poca correlación con los ingresos.

BOXPLOT DE COMPARACIÓN



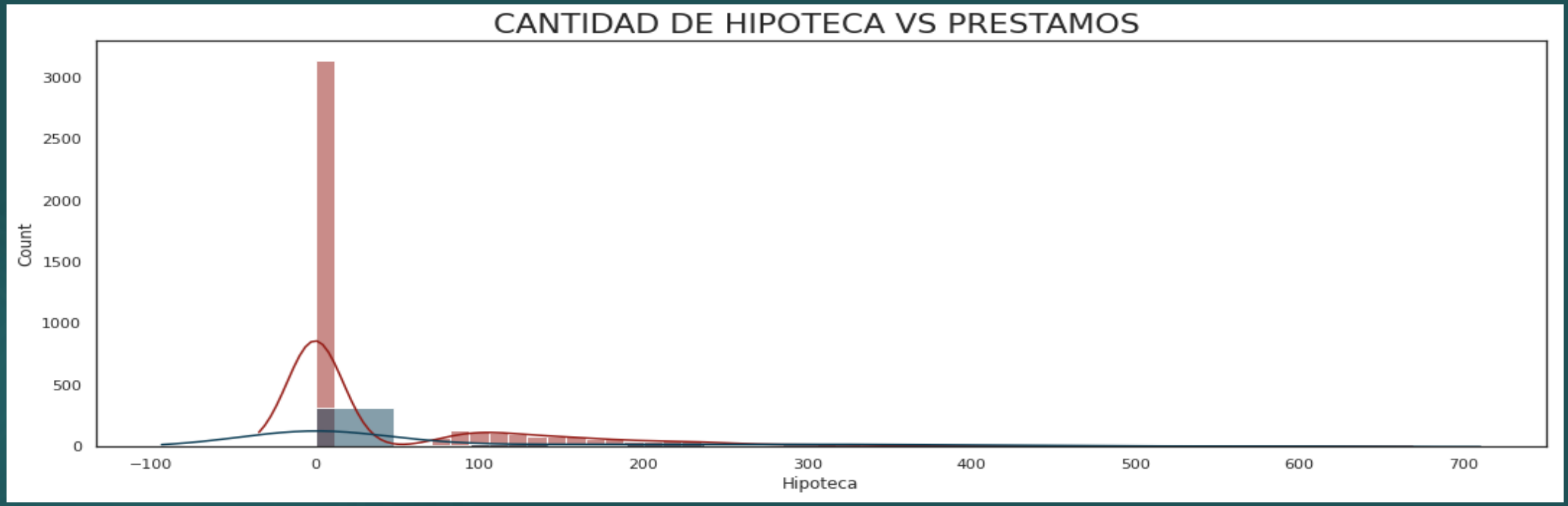
Los sueldos y los gastos de tarjetas de crédito son los valores más significativos para los prestamistas al evaluar la elegibilidad de un solicitante para recibir financiamiento. Los sueldos más altos y los gastos de tarjetas de crédito bajos indican una mayor capacidad de pago y una buena trayectoria de pago, lo que aumenta las posibilidades de obtener un préstamo y cumplir con las obligaciones financieras del mismo en el futuro.

GRUPO DE INGRESO SEGÚN LA CUENTA



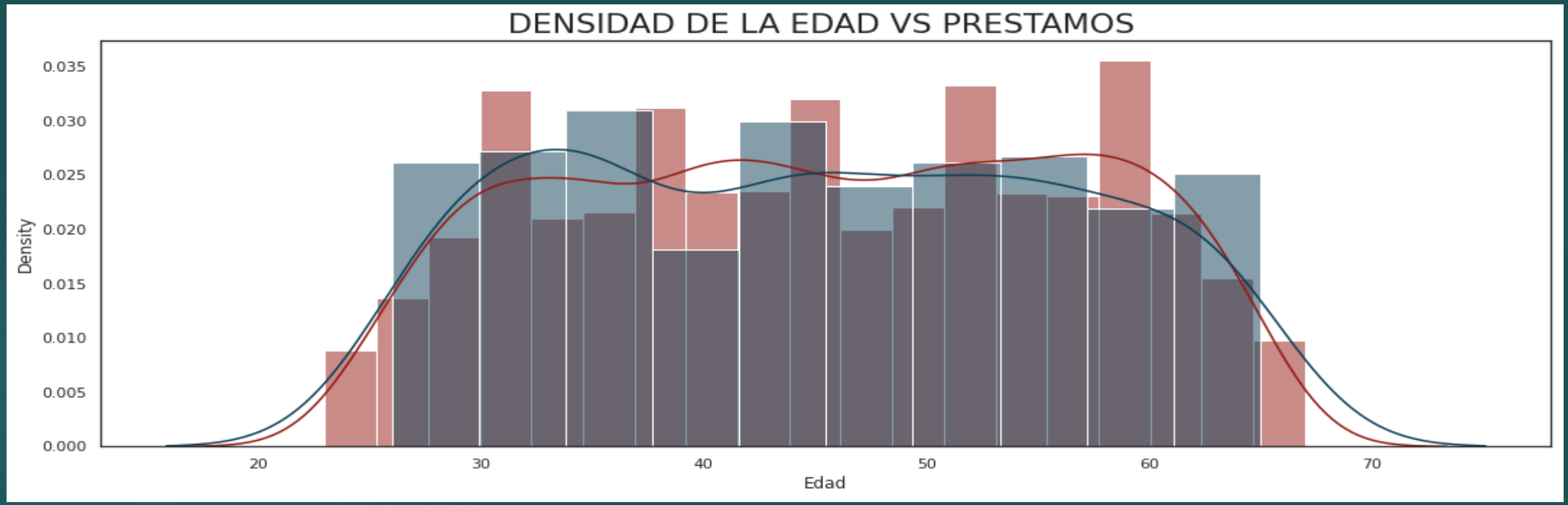
Los grupo de ingresos medios son los que tienen mas cuentas de deposito, quizás uno de los puntos débiles es que en porcentaje a la cantidad de clientes que no poseen cuenta de deposito con respecto a los que poseen una, es menor a los demás grupos de ingresos, esto debería reverse y en lo posible incrementar ese porcentaje, ya que son los clientes mas aptos para este banco.

DENSIDAD DE HIPOTECA VS PRESTAMOS



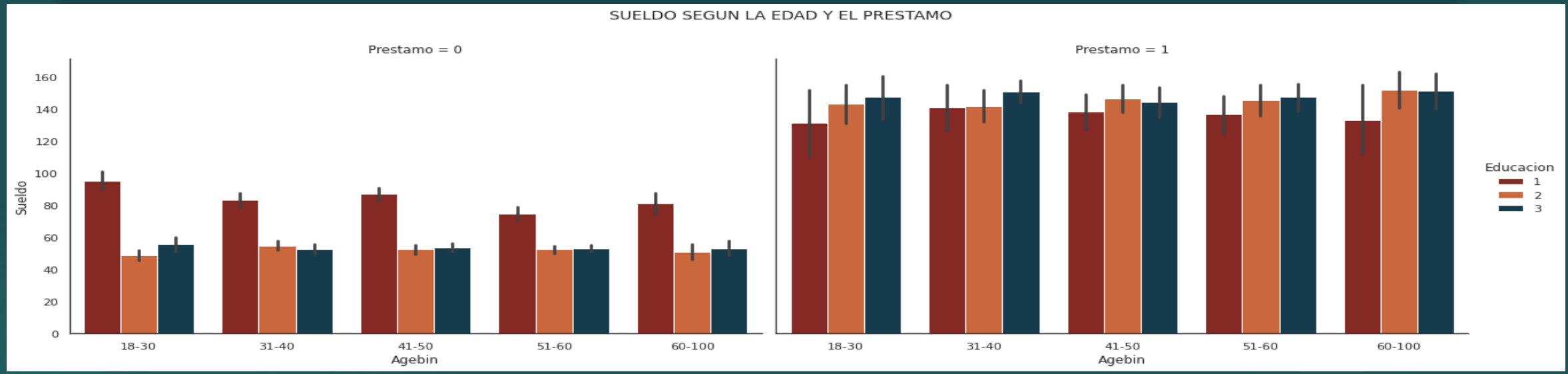
Podemos observar hay gran cantidad de clientes que no poseen prestamos activos, con respecto a la hipoteca, eso puede deberse a que no poseen un sueldo alto o que no les fue necesario solicitar un prestamos mientras poseen una hipoteca.

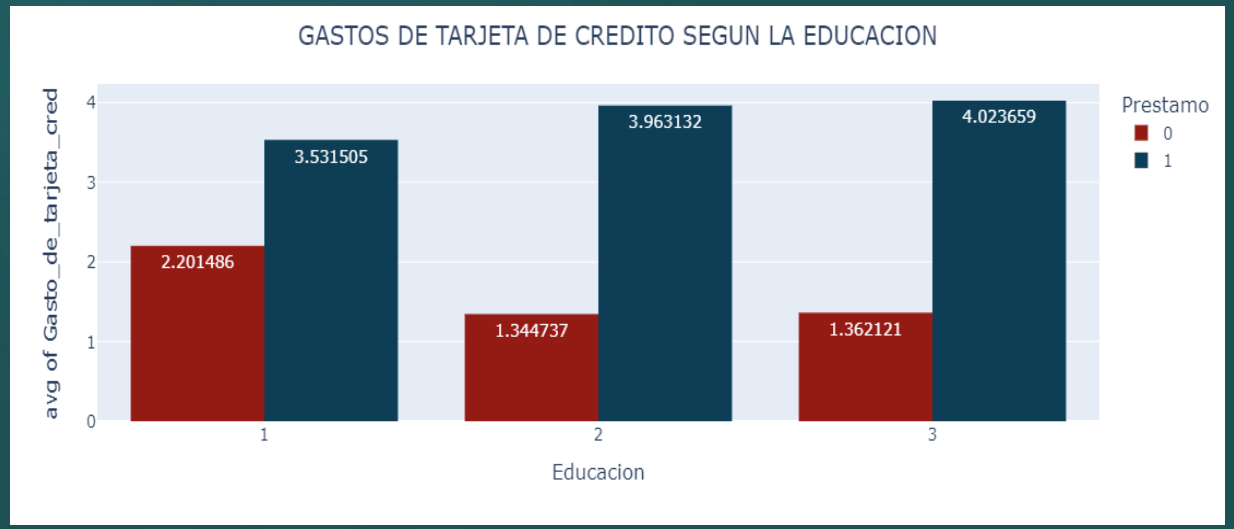
DENSIDAD DE LA EDAD VS PRESTAMOS



Podemos observar que los clientes con edades mayores a 30 y menores a 60 son los que mas poseen una densidad de prestamos activos o no, con respecto al grupo de edades fuera de ese rango.

GASTOS





Podemos observar que los clientes con mayor educación son los que mas utilizan las tarjetas de crédito y poseen prestamos con respecto al grupo de edades, aunque los que tienen una educación menor son los que poseen mejor sueldo a los que no poseen un préstamo activo.

CONCLUSIÓN DE LOS CLIENTES SEGÚN SUS CARACTERÍSTICAS DE INGRESOS

El préstamo personal variable objetivo está muy desequilibrado, donde solo el 9,6% de los clientes han optado previamente por un préstamo personal en el conjunto de datos.

- Las personas con mayores ingresos habían optado por préstamos personales antes.
- Las personas con hipotecas altas optaron por el préstamo.
- Los clientes con mayor uso de crédito promedio mensual han optado por el préstamo.
- Los clientes con mayores ingresos tenían un mayor uso promedio de tarjetas de crédito e hipotecas.

ÁRBOL DE DECISIONES

Observaciones en el EDA:

- Las personas con mayores ingresos habían optado por préstamos personales antes.
- Las personas con hipotecas altas optaron por el préstamo.
- Los clientes que hayan optado por el préstamo tendrán un uso de crédito promedio mensual más alto.
- Los clientes con Familia de 3 miembros habían tomado prestados los préstamos con el banco.
- Nivel de educación 2: Graduado y 3: Avanzado/Profesional han tomado préstamos con el banco.
- Clientes que tenían certificado de depósito con el banco habían tomado prestado préstamo previamente
- La mayoría de los clientes que tenían un préstamo personal con el banco utilizaron las instalaciones en línea.
- La mayoría de los clientes que habían tomado préstamos personales antes son de la región de Los Ángeles.
- La proporción de préstamo de endeudamiento es alta en 30 y por debajo y 60 y por encima de los clientes.
- Cuantos más ingresos obtenga, más gastará y tendrá un estilo de vida "grande que la vida".
- Segmentación de clientes para préstamo de endeudamiento basado en EDA
- Los clientes con ingresos más altos tienen hipotecas más altas y un gasto promedio mensual más alto. También tienen certificado de depósito con el banco. Son nuestros clientes de alto perfil.
- Pocos Clientes en el grupo de ingresos medios no tienen hipotecas más altas y tienen menos gasto promedio mensual con tarjeta de crédito. Son clientes de perfil promedio.
- Los clientes en el grupo de ingresos más bajos tienen menos hipotecas (hay pocos valores atípicos), menos gastos mensuales. Son nuestros clientes de bajo perfil.

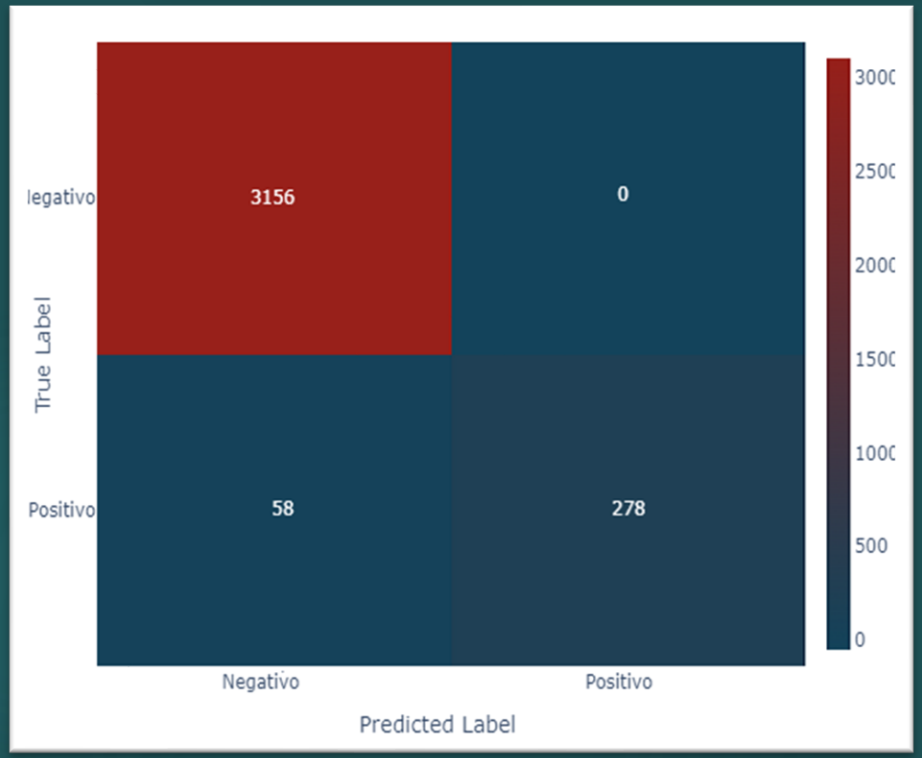
Acciones para el pre procesamiento de datos:

Muchas variables tienen valores atípicos que necesitan ser tratados. Podemos eliminar Experiencia, País, Código postal y Agebin, Grupo de ingresos, Grupo de gastos.

Que variables queremos enviar a nuestro árbol de decisión?

Edad, Experiencia, Sueldo, Familia, Gasto de tarjeta Crédito, Educación, Hipoteca, Préstamo, Cuenta de seguridad, Cuenta deposito, Online, Tarjeta de Crédito

MODELADO: ÁRBOL DE DECISIÓN




```
Train Accuracy:  0.9833906071019473
Train Recall:  0.8273809523809523
Train F1 score:  0.9055374592833876
Train Precision:  1.0
```

Perspectivas:

Verdaderos positivos:

Realidad: Un cliente quería tomar un préstamo personal.

Modelo de predicción: el cliente tomará un préstamo personal.

Resultado: El modelo es bueno.

Verdaderos negativos:

Realidad: Un cliente no quería tomar un préstamo personal.

Modelo de predicción: el cliente no tomará un préstamo personal.

Resultado: El negocio no se ve afectado.

Falsos positivos:

Realidad: Un cliente no quería tomar un préstamo personal.

Modelo de predicción: el cliente tomará un préstamo personal.

Resultado: El equipo que se dirige a los clientes potenciales desperdiciaría sus recursos en los clientes que no comprarán un préstamo personal.

Falsos negativos:

Realidad: Un cliente quería tomar un préstamo personal.

Modelo de predicción: el cliente no tomará un préstamo personal.

Resultado: el equipo de ventas extraña al cliente potencial. Esto es pérdida de oportunidad. El propósito de la campaña era dirigirse a tales clientes. Si el equipo supiera acerca de estos clientes, podrían haber ofrecido algunas buenas tasas de APR/interés.

MODELADO: ÁRBOL DE DECISIÓN

Pudimos realizar pruebas en todas las métricas, y detectamos que nuestras métricas de desempeño es mejor a las del y_train por pequeña diferencia. Podemos decir que el modelo al usar y_train está sobre ajustado o entrenado (overfitting), sirve muy bien para valores que ya entreno, pero no para valores que nunca vio.
Cuando tenemos overfitting, podemos buscar mejorar nuestra métrica de desempeño para nuestro set de validación. No hay una regla exacta, es hacer varios procesos hasta encontrar algo que sirve para nuestros datos, podríamos intentar hacerlo con algunas de las siguientes maneras:

- Itearar.
- feature engineer.
- Obtener un mayor número de datos.
- Ajustar los parámetros de nuestros modelos.
- Crear modelos más simples en caso de ser posible.
- Probar con más modelos.
- Ver si nos sirve otro algoritmo nuevo y es mejor.

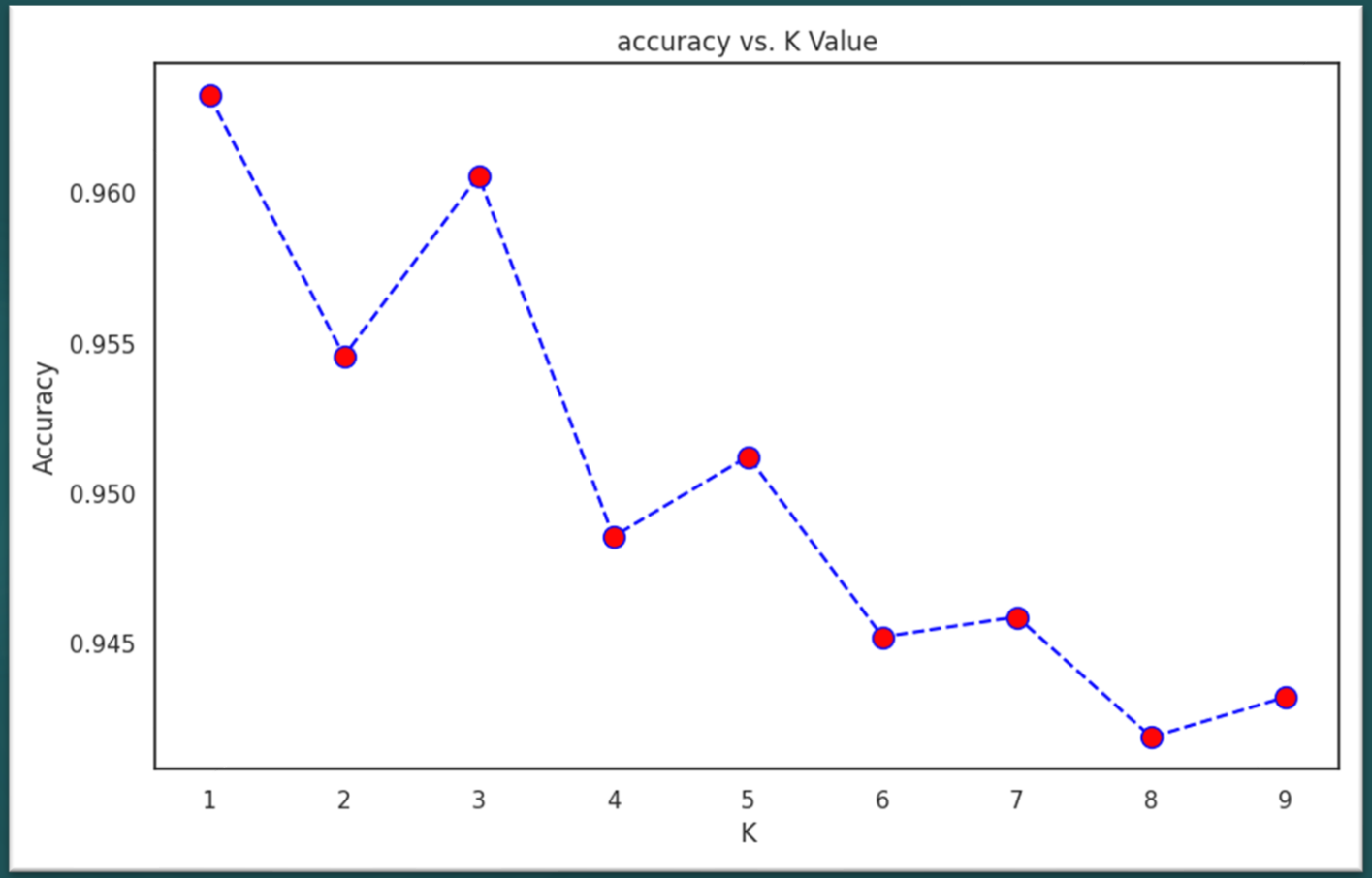
Y TEST

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1353
1	1.00	0.82	0.90	144
accuracy			0.98	1497
macro avg	0.99	0.91	0.95	1497
weighted avg	0.98	0.98	0.98	1497

Y TRAIN

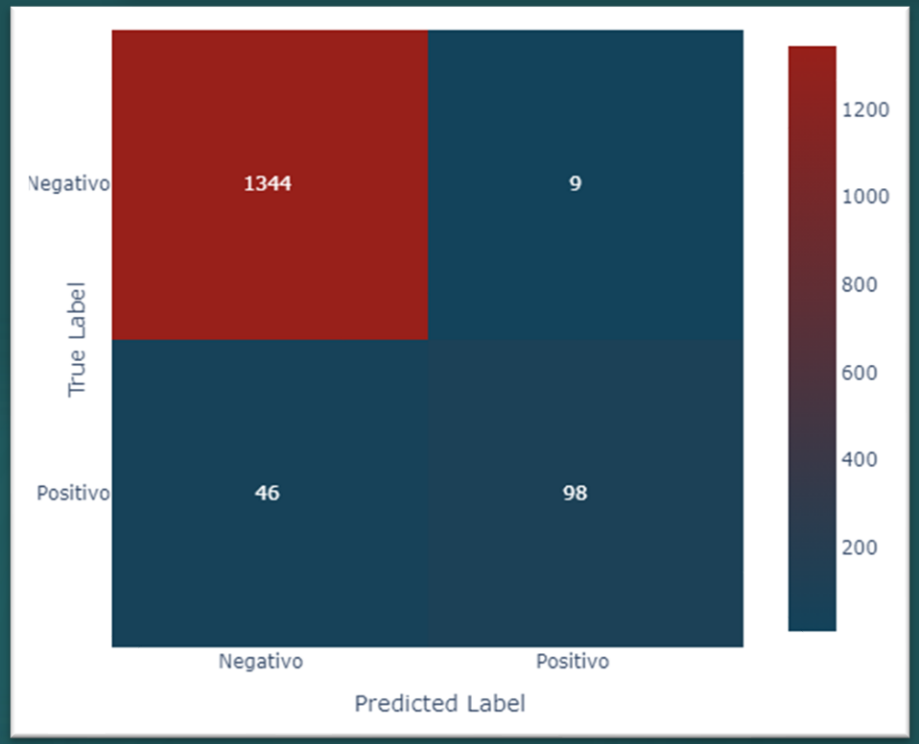
	precision	recall	f1-score	support
0	0.98	1.00	0.99	3156
1	1.00	0.83	0.91	336
accuracy			0.98	3492
macro avg	0.99	0.91	0.95	3492
weighted avg	0.98	0.98	0.98	3492

MODELOS: KNN



Se realiza un ajuste y una transformación al conjunto de entrenamiento y una transformación al conjunto de prueba. Luego se evalúa el valor óptimo de K para maximizar la precisión. Según los resultados, la mejor precisión se obtiene con K=1.

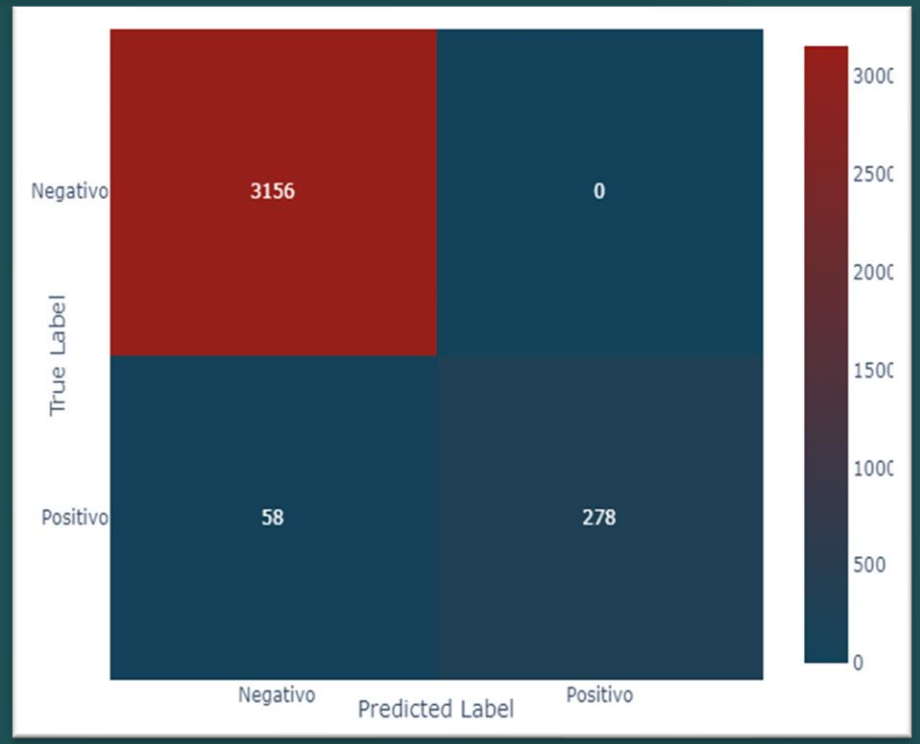
MODELOS: KNN



Y TEST

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1353
1	0.92	0.68	0.78	144
accuracy			0.96	1497
macro avg	0.94	0.84	0.88	1497
weighted avg	0.96	0.96	0.96	1497

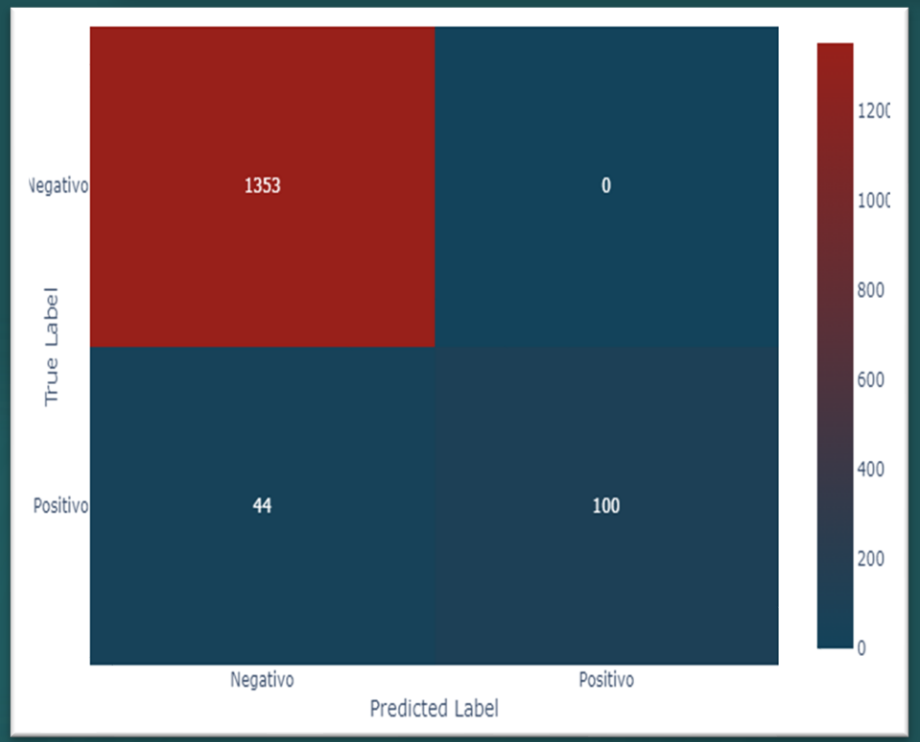
Según los resultados, el modelo KNN se ajusta mejor a los datos de entrenamiento que a los de prueba. Sin embargo, ambos tienen un buen desempeño, lo que indica que la mayoría de los clientes son elegibles para solicitar préstamos.



Y TRAIN

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3156
1	1.00	0.83	0.91	336
accuracy			0.98	3492
macro avg	0.99	0.91	0.95	3492
weighted avg	0.98	0.98	0.98	3492

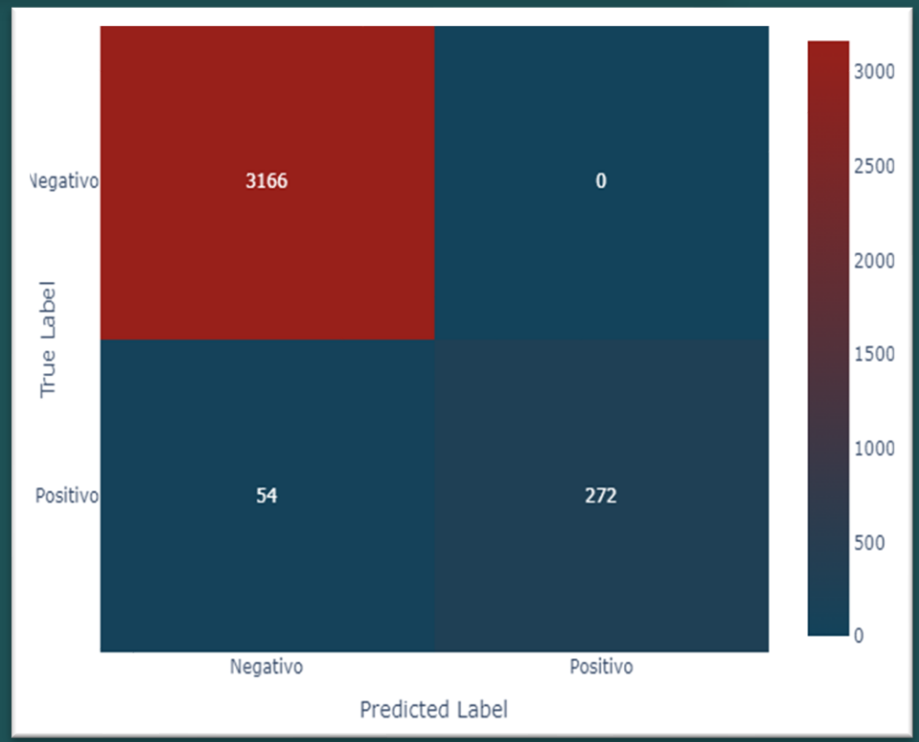
MODELOS: RANDOM FOREST



Y TEST

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1353
1	1.00	0.69	0.82	144
accuracy			0.97	1497
macro avg	0.98	0.85	0.90	1497
weighted avg	0.97	0.97	0.97	1497

Según los resultados, el modelo Random Forest se ajusta mejor a los datos de entrenamiento que a los de prueba. Sin embargo, ambos tienen un buen desempeño, lo que indica que la mayoría de los clientes son elegibles para solicitar préstamos.

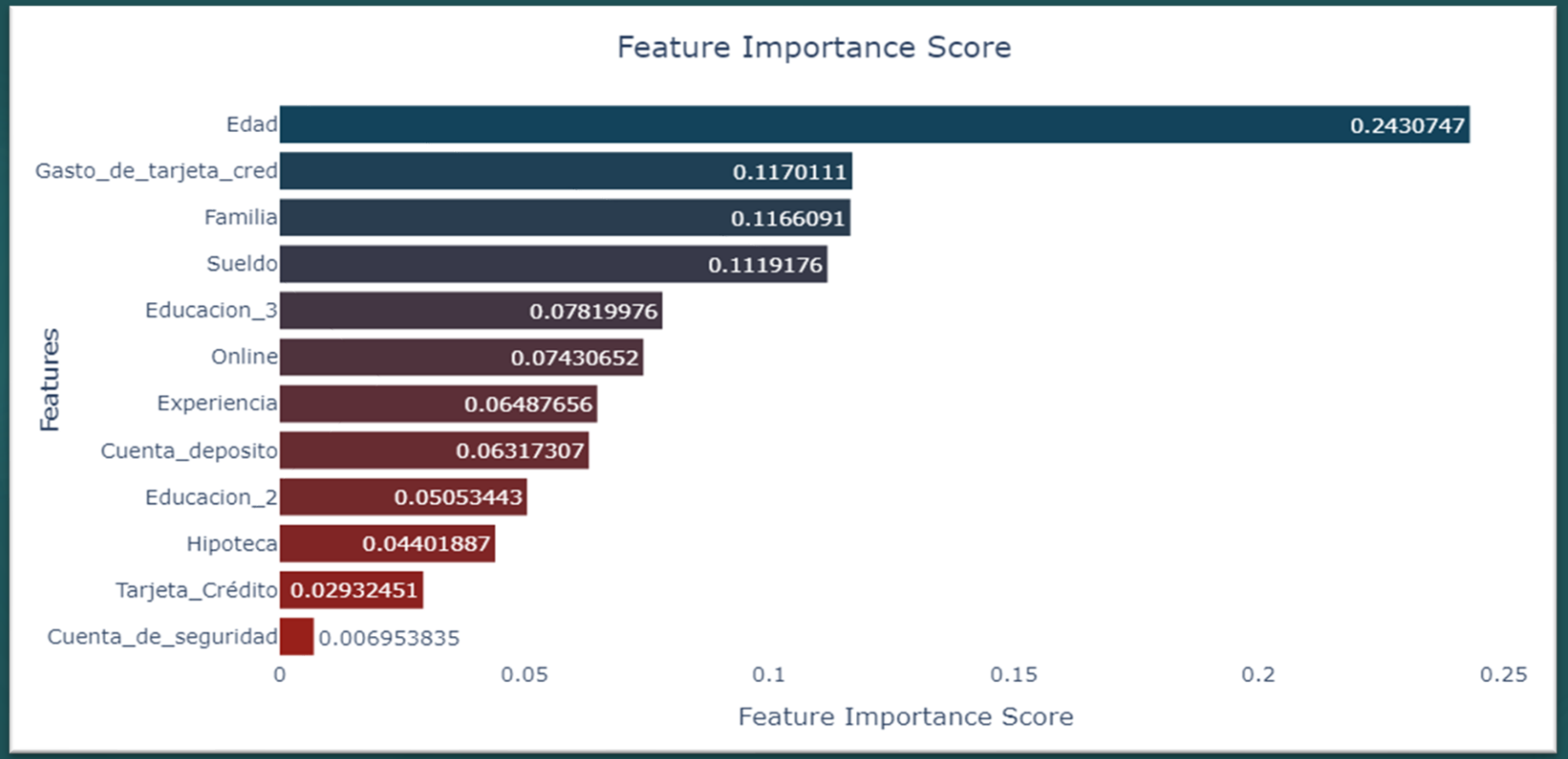


Y TRAIN

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3166
1	1.00	0.83	0.91	326
accuracy			0.98	3492
macro avg	0.99	0.92	0.95	3492
weighted avg	0.98	0.98	0.98	3492

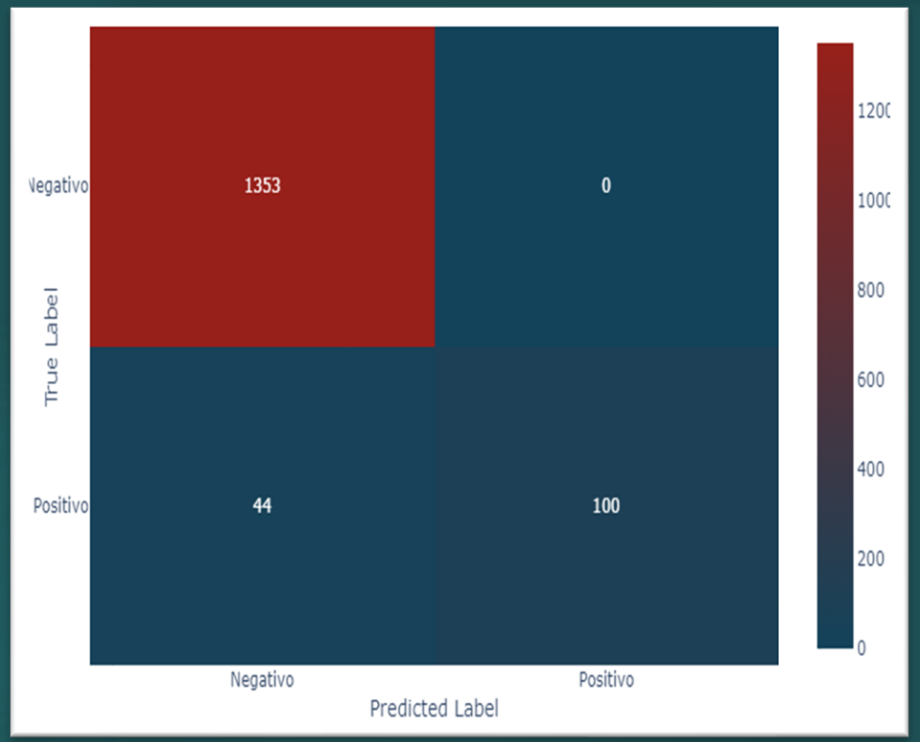
MODELOS: FEATURE IMPORTANCE

Viene de un calculo que involucra la impureza en los nodos. NO es un porcentaje NI una probabilidad. Nos da una idea de cuales variables son importantes para el modelo. Feature importance igual a 0 significa que feature no fue usado por el modelo para predecir.



Se puede observar que la edad es el factor más influyente, seguido por los tres siguientes: los gastos de tarjeta de crédito, el tamaño del grupo familiar y el sueldo. Sin embargo, aunque estos últimos también tienen importancia, no alcanzan el nivel del primero.

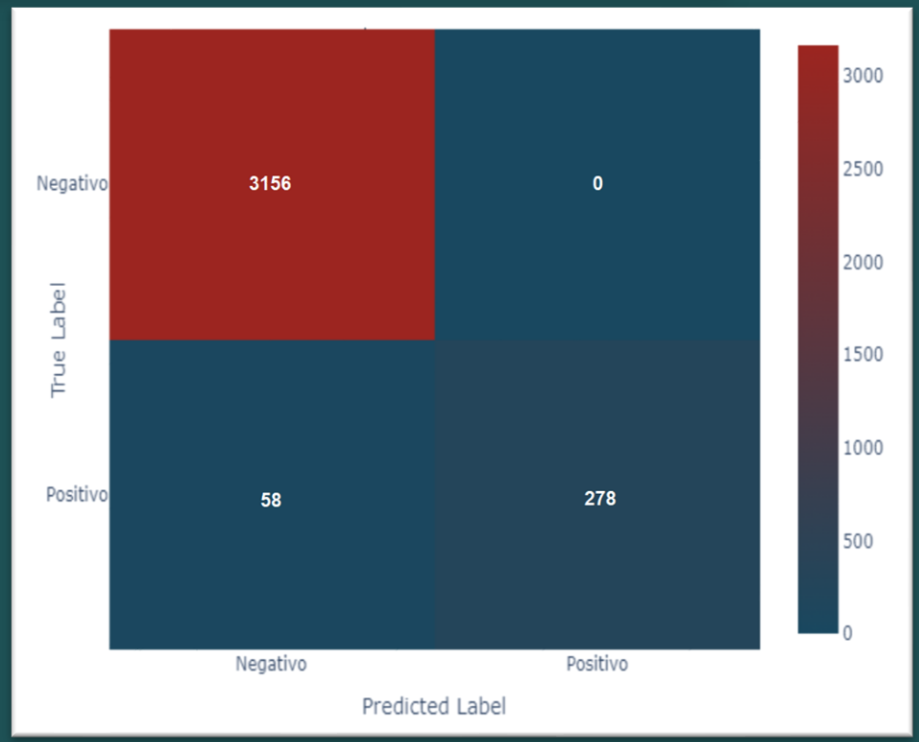
MODELOS: SVM



Y TEST

	precision	recall	f1-score	support
0	0.97	1.00	0.99	1353
1	0.97	0.74	0.84	144
accuracy			0.97	1497
macro avg	0.97	0.87	0.91	1497
weighted avg	0.97	0.97	0.97	1497

Según los resultados, el modelo SVM se ajusta mejor a los datos de entrenamiento que a los de prueba. Sin embargo, ambos tienen un buen desempeño, lo que indica que la mayoría de los clientes son elegibles para solicitar préstamos.



Y TRAIN

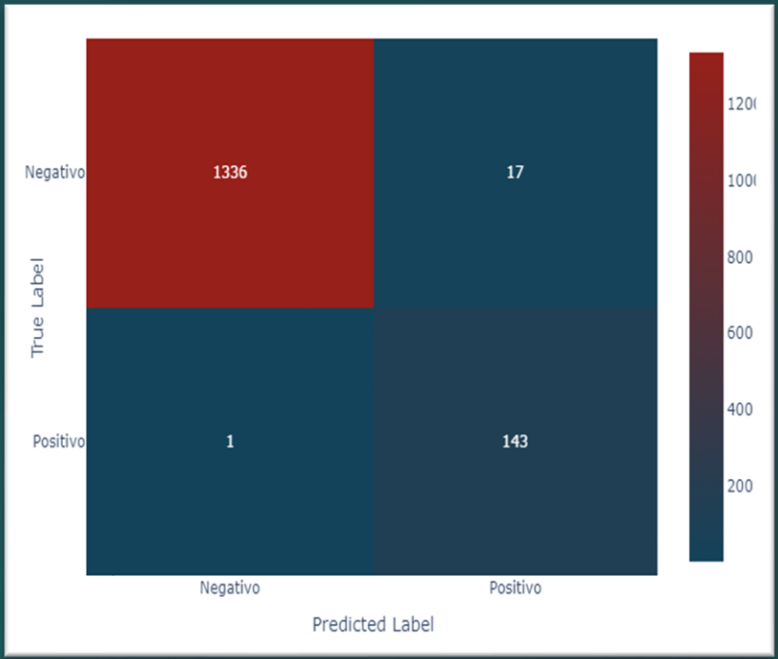
	precision	recall	f1-score	support
0	0.98	1.00	0.99	3156
1	1.00	0.83	0.91	336
accuracy			0.98	3492
macro avg	0.99	0.91	0.95	3492
weighted avg	0.98	0.98	0.98	3492

CONCLUSIÓN MODELOS

En resumen, tanto el Random Forest como el SVM se destacan en términos de rendimiento en este conjunto de datos. El Random Forest muestra una alta precisión y recall para ambas clases, lo que indica que es capaz de clasificar correctamente la mayoría de las instancias de ambas clases. Por otro lado, el SVM también logra una alta precisión en la clasificación de la clase 0 y un recall decente para la clase 1. Esto sugiere que el SVM puede ser efectivo en identificar correctamente las instancias de la clase mayoritaria mientras mantiene un buen equilibrio en la clasificación de ambas clases.

En contraste, el KNN y la Regresión Logística tienen un rendimiento ligeramente inferior. El KNN muestra una precisión y recall más bajos para la clase 1, lo que implica que puede tener dificultades para identificar correctamente las instancias de esta clase. Por su parte, la Regresión Logística tiene una precisión relativamente baja para la clase 1, lo que sugiere que puede haber problemas en la clasificación de esta clase en particular.

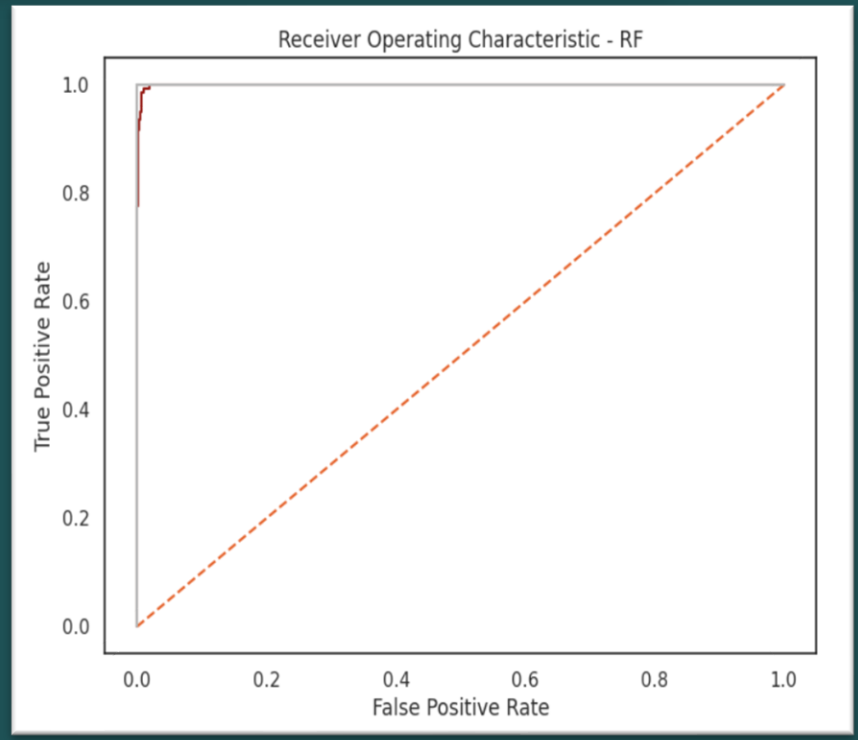
MEJORA DE MODELOS: RANDOM FOREST



Y TEST

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1353
1	1.00	0.80	0.89	144
accuracy			0.98	1497
macro avg	0.99	0.90	0.94	1497
weighted avg	0.98	0.98	0.98	1497

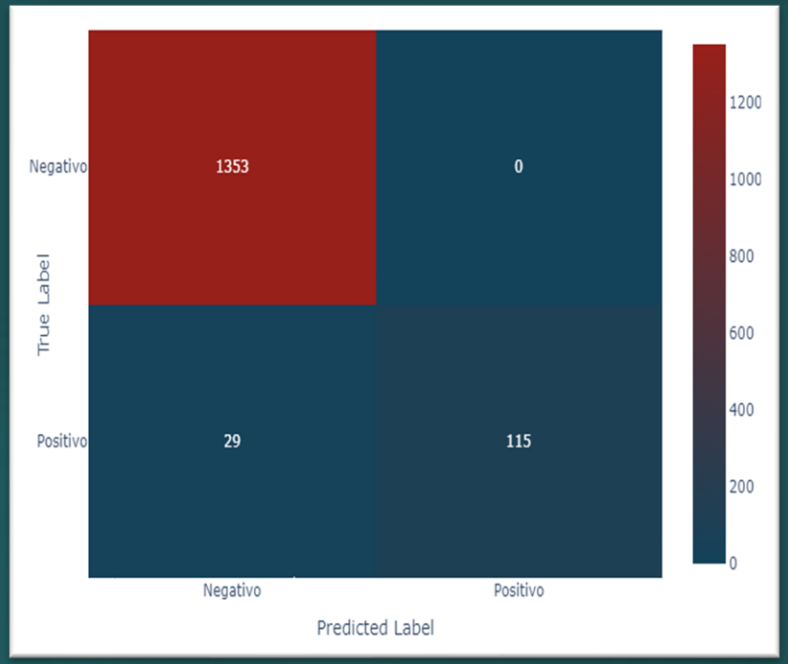
El Random Forest muestra un rendimiento excepcional en este conjunto de datos, con una precisión del 100% para la clase 0 y una precisión del 89% para la clase 1, un recall del 99% para ambas clases, y un f1-score promedio del 97%. Además, obtiene un valor muy alto de roc_auc_score de 0.9993, lo que indica una capacidad sobresaliente para distinguir entre las clases y un rendimiento destacado en la tarea de clasificación. En resumen, el Random Forest demuestra ser un modelo altamente preciso y confiable en la clasificación de este conjunto de datos.



ROC - Random Forest

```
roc_auc_score for Random Forest:  0.9993224932249323
```

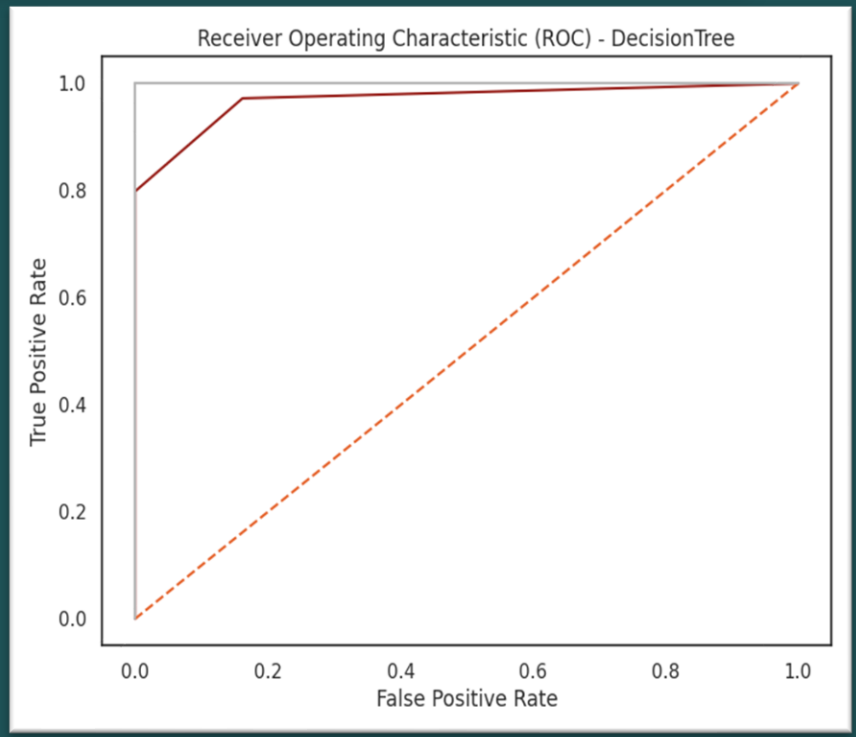
MEJORA DE MODELOS: DECISION TREE CLASSIFIER



Y TEST

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1353
1	1.00	0.80	0.89	144
accuracy			0.98	1497
macro avg	0.99	0.90	0.94	1497
weighted avg	0.98	0.98	0.98	1497

El Random Forest muestra un rendimiento excepcional en este conjunto de datos, con una precisión del 100% para la clase 0 y una precisión del 89% para la clase 1, un recall del 99% para ambas clases, y un f1-score promedio del 97%. Además, obtiene un valor muy alto de roc_auc_score de 0.9993, lo que indica una capacidad sobresaliente para distinguir entre las clases y un rendimiento destacado en la tarea de clasificación. En resumen, el Random Forest demuestra ser un modelo altamente preciso y confiable en la clasificación de este conjunto de datos.



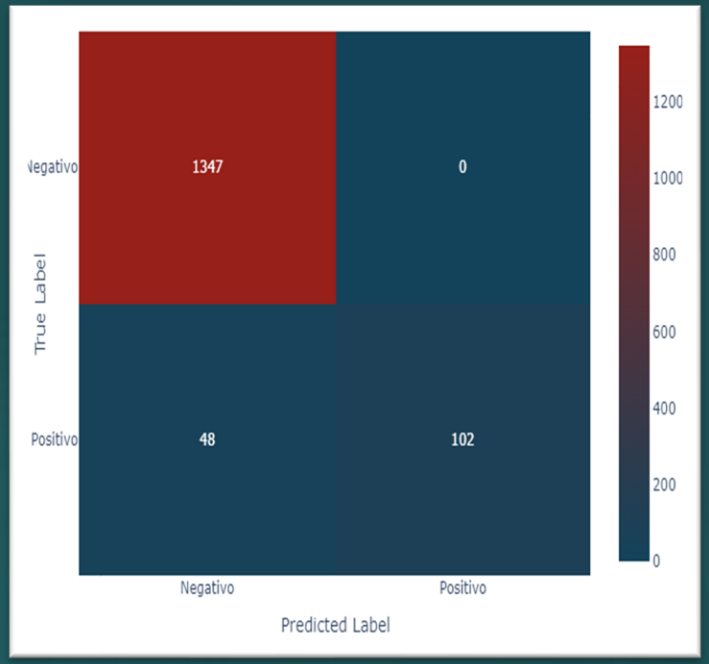
ROC - DecisionTree

```
roc_auc_score for DecisionTree:  0.9698124538063563
```

CONCLUSIÓN MEJORA DE MODELOS

En comparación con la mejora del modelo Decision Tree Classifier, el modelo Random Forest sigue siendo la opción preferida debido a su mayor rendimiento general en términos de precisión, recall y f1-score. El Random Forest logra una precisión del 100% para la clase 0 y una precisión del 89% para la clase 1, mientras que el Decision Tree Classifier obtiene una precisión del 98% y del 100% respectivamente. Además, el Random Forest tiene un recall del 99% para ambas clases, superando el recall del 80% obtenido por el Decision Tree Classifier para la clase 1. En términos de f1-score, el Random Forest alcanza un promedio del 97% en comparación con el 89% del Decision Tree Classifier. Estas métricas más altas en el Random Forest indican una mejor capacidad para clasificar correctamente las instancias de ambas clases. Por lo tanto, el Random Forest sigue siendo el modelo preferido debido a su rendimiento general superior en la clasificación de este conjunto de datos.

ANALISIS PCA



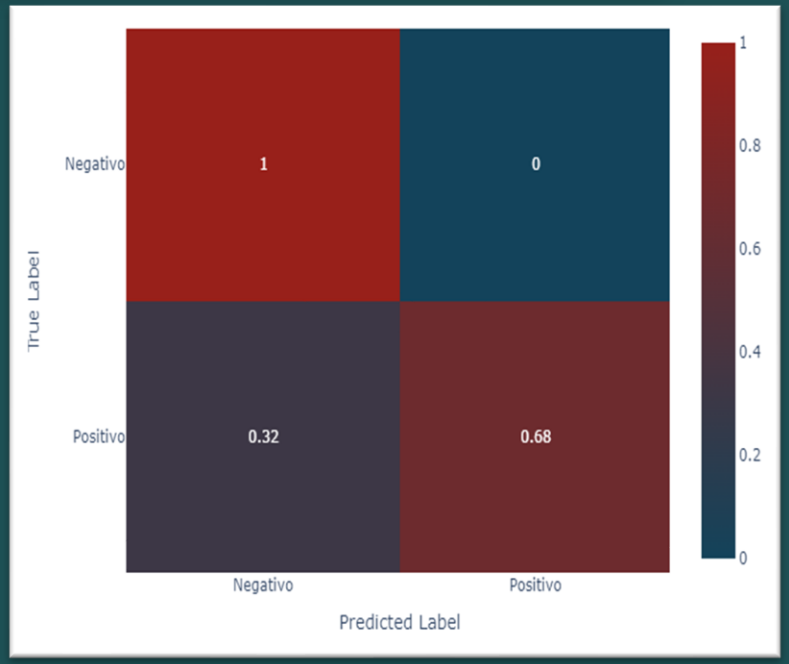
Y TEST

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1347
1	1.00	0.68	0.81	150
accuracy			0.97	1497
macro avg	0.98	0.84	0.90	1497
weighted avg	0.97	0.97	0.97	1497

Principalmente no hubo cambios en la performance del modelo, ya que se encuentra con una precisión del modelo para la clase 0 es alta, con un 96%, lo que indica que el modelo ha identificado correctamente la mayoría de los casos de la clase 0. La recuperación (recall) del modelo para la clase 1 es un poco baja, con un 68%, lo que indica que el modelo ha perdido algunos casos de la clase 1.

El puntaje F1 del modelo para la clase 1 es de 0.81, lo que indica un equilibrio razonable entre la precisión y la recuperación. En general, el modelo tiene una precisión promedio del 97%, lo que significa que ha clasificado correctamente la gran mayoría de los casos.

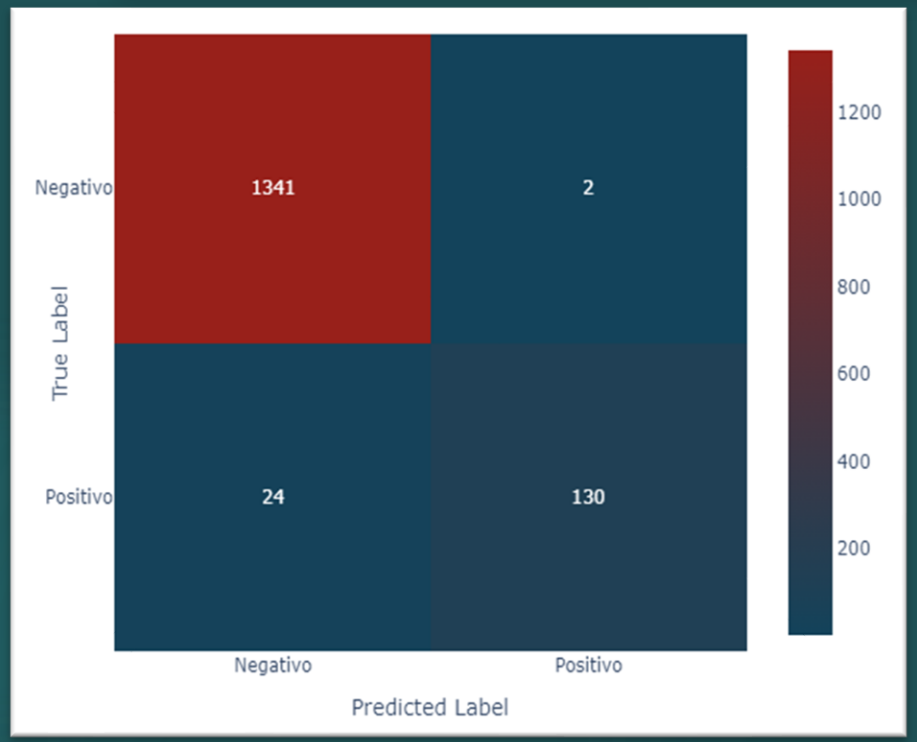
El puntaje F1 promedio ponderado es de 0.97, lo que sugiere que el modelo tiene un buen rendimiento general en términos de clasificación. En conclusión, aunque hay margen de mejora en la recuperación de la clase 1, el modelo parece ser bastante efectivo en la clasificación de la mayoría de los casos. El informe sugiere que el modelo tiene más dificultades para clasificar correctamente las muestras de la clase 1 que las de la clase 0.



AVG SCORE

Average score: 0.97

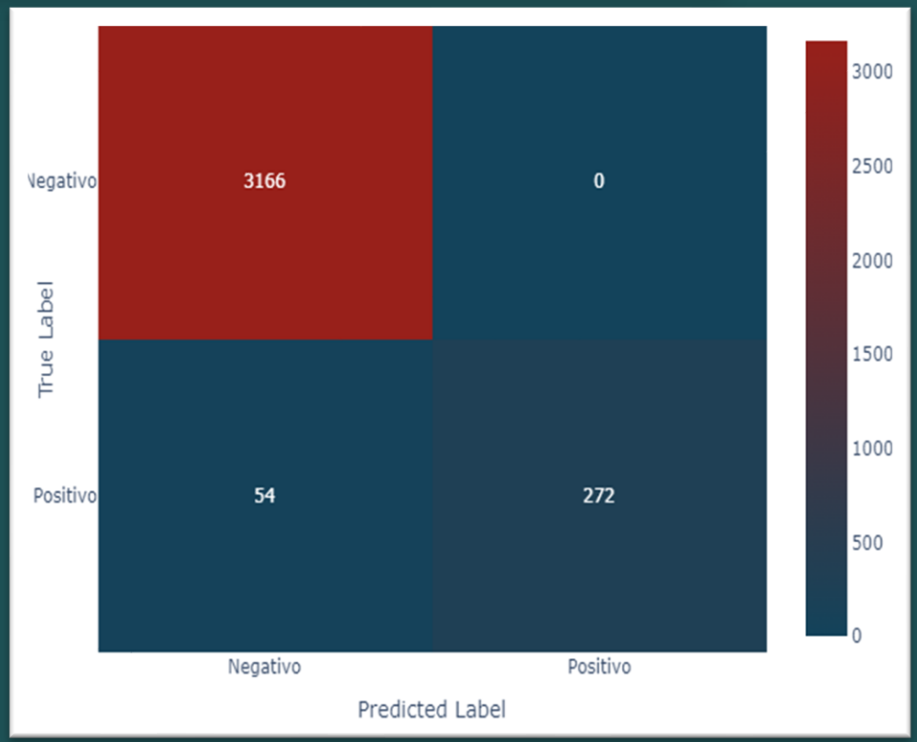
BOOSTING MODELS: XGBOOST



Y TEST

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1343
1	0.98	0.84	0.91	154
accuracy			0.98	1497
macro avg	0.98	0.92	0.95	1497
weighted avg	0.98	0.98	0.98	1497

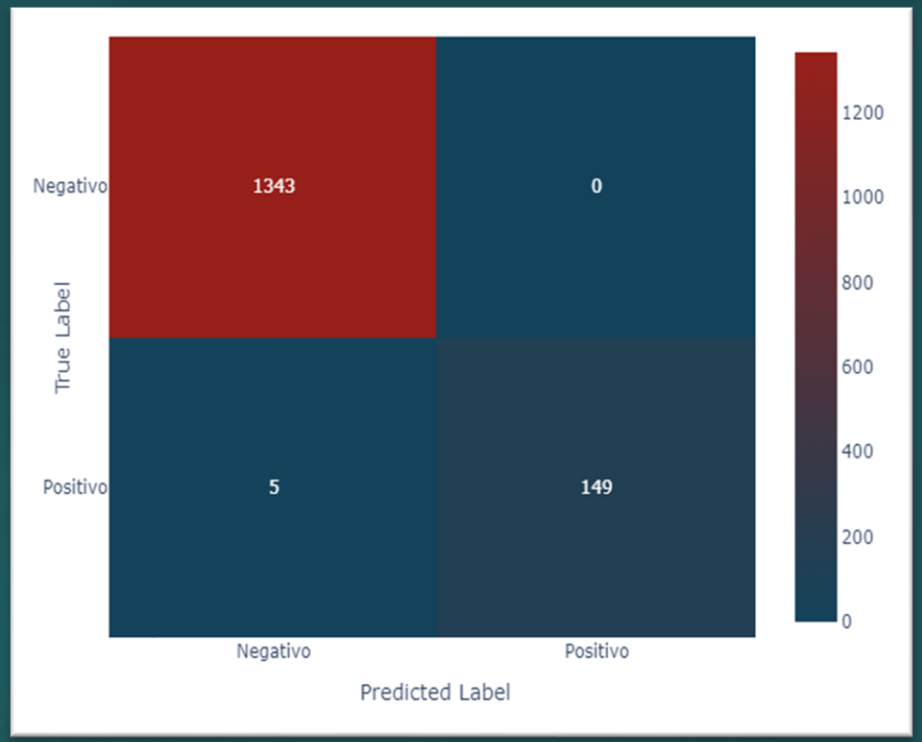
Según los resultados, el modelo XGBOOST se ajusta mejor a los datos de prueba que a los de entrenamiento. Sin embargo, ambos tienen un buen desempeño, lo que indica que la mayoría de los clientes son elegibles para solicitar préstamos.



Y TRAIN

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3166
1	1.00	0.83	0.91	326
accuracy			0.98	3492
macro avg	0.99	0.92	0.95	3492
weighted avg	0.98	0.98	0.98	3492

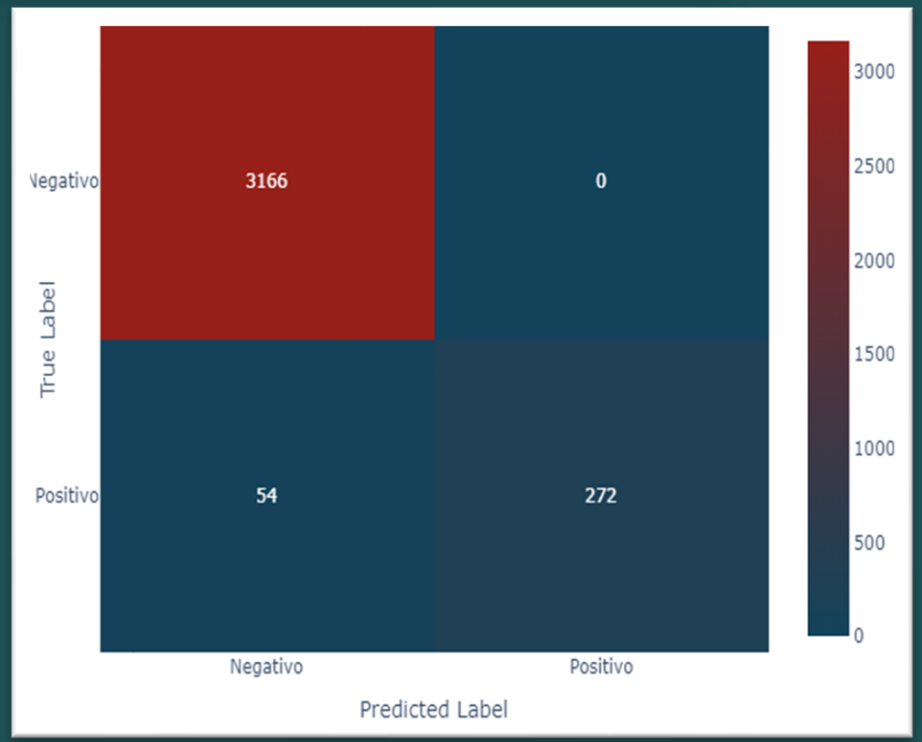
BOOSTING MODELS: LIGHTGBM



Y TEST

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1343
1	1.00	0.97	0.98	154
accuracy			1.00	1497
macro avg	1.00	0.98	0.99	1497
weighted avg	1.00	1.00	1.00	1497

Según los resultados, el modelo LIGHTGBM se ajusta mejor a los datos de prueba que a los de entrenamiento. Sin embargo, ambos tienen un buen desempeño, lo que indica que la mayoría de los clientes son elegibles para solicitar préstamos.



Y TRAIN

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3166
1	1.00	0.83	0.91	326
accuracy			0.98	3492
macro avg	0.99	0.92	0.95	3492
weighted avg	0.98	0.98	0.98	3492

CONCLUSIÓN BOOSTING MODELS

Ambos modelos, XGBoost y LightGBM, muestran un rendimiento muy sólido en el conjunto de datos proporcionado. Sin embargo, LightGBM parece tener un rendimiento ligeramente mejor en términos de métricas de precisión, recall y F1-score para la clase 1.

En comparación con XGBoost, LightGBM logra un recall del 97% para la clase 1 en lugar del 84% de XGBoost, lo que indica una mejor capacidad para identificar correctamente los casos positivos. Además, el F1-score para la clase 1 también es mayor en LightGBM (0.98) en comparación con XGBoost (0.91), lo que implica un equilibrio más óptimo entre precisión y recall.

Sin embargo, ambos modelos son altamente precisos en la clasificación de la clase 0, con una puntuación perfecta en el caso de LightGBM. Por lo tanto, si la prioridad principal es la identificación precisa de la clase minoritaria y se busca un modelo con un rendimiento general ligeramente mejor, LightGBM podría considerarse como la mejor opción en este caso.

CONCLUSIÓN FINAL

- Analizamos los datos de la campaña de Préstamos personales usando EDA y usando diferentes modelos como Clasificador de árboles de decisión para generar una probabilidad de que el Cliente solicite un Préstamo.
- Los árboles de decisión no requieren mucha preparación de datos o manejo de valores atípicos como la regresión logística. Son fáciles de entender. El árbol de decisión puede sobre ajustarse fácilmente, por lo que debemos tener cuidado al usar el árbol de decisión.
- Coeficiente de Ingresos, Graduados y Educación Avanzada, Cuenta de deposito, Edad, son positivos, es decir, un aumento de una unidad en estos conducirá a un aumento en las posibilidades de que una persona tome un préstamo.
- El árbol de decisión puede sobre ajustarse fácilmente. Requieren menos pre procesamiento de datos en comparación con la regresión logística y son fáciles de entender.
- Los ingresos, los clientes con título de posgrado, los clientes que tienen 3 miembros en la familia son algunas de las variables más importantes para predecir si los clientes comprarán un préstamo personal.
- Los clientes que tienen ingresos superiores a 98k dólares, educación de nivel avanzado/graduado, una familia de más de 2, estos clientes tienen mayores posibilidades de tomar préstamos personales.
- Entonces para esta campaña podemos tener diferentes perfiles para los clientes:
- Clientes de alto perfil: ingresos más altos, educación de nivel avanzado/graduado, 3/4 miembros de la familia, gastos elevados.
- Perfil promedio: - Grupo de ingresos medios, educación de nivel de posgrado. 3 a 4 miembros de la familia, gasto medio Perfil bajo: grupo de bajos ingresos, estudiantes universitarios, 3 a 4 miembros de la familia, gastos bajos.
- El gasto promedio del cliente y las hipotecas también se pueden considerar basados en EDA y regresión logística. Estos parámetros también juegan un papel en la probabilidad de comprar un préstamo.
- Primero, podemos dirigirnos a clientes de alto perfil, brindándoles un administrador de relaciones personal que pueda abordar sus inquietudes y perseguirlos para comprar un préstamo del banco con tasas de interés completas.
- La precalificación para el préstamo también puede atraer a más clientes.
- Nuestro segundo objetivo serían los clientes de perfil medio.
- El modelo no puede identificar bien si hay algunos casos excepcionales cuando el cliente de bajo perfil está listo para comprar un préstamo personal.