Aplicación de Modelos de ML para la Identificación de la Fuga de Clientes en el Servicio de Tarjeta de Créditos en un Banco.

1st Heydy Mayumy Carrasco Huaccha
Fac. Ciencias Matemáticas, Computación Cientfica
Universidad Nacional Mayor de San Marcos
Lima, Perú
heydy.carrasco.huaccha@gmail.com

2nd Alex Rivera Cruz
Fac. Ingeniería Industrial y de Sistemas, Ing. Industrial
Universidad Nacional de Ingeniería
Lima, Perú
alexriveracruz4@gmail.com

3rd Martin Adolfo Delgado Huayhua
Fac. Economia y Planificacion, Estadística Informática
Universidad Nacional Agraria la Molina
Lima, Perú
martindelhu1302@gmail.com

4th Angel Jorge Salazar
Fac. Ciencias Económicas, Economía
Universidad Nacional Mayor de San Marcos
Lima, Perú
1jssangel1@gmail.com

***Abstract*—En este artículo se presenta el análisis realizado a la data de los clientes que permanecen y abandonaron el servicio de tarjetas de crédito de un banco; esta tiene como fuente la página de Kaggle. Aquí se busca encontrar un modelo de Machine Learning que mejor identifique los potenciales clientes en fuga; para que así el banco pueda tomar acciones comerciales para retener dichos clientes.**

Todo el análisis fue realizado en Python ya que posee muchas librerias que apoyan a este fin como por ejemplo: Pandas, Numpy, Sckelearn, etc. Se realizaron análisis exploratorio de cada feature, se calcularon ciertos valores estadísticos, se trazaron gráficos de acuerdo a la naturaleza de cada variable, etc.

Para este proyecto se aplicaron los modelos de Random Forest, Regresión Logística, LGBM, Gradient Boosting Classifier (GBC) y catboost debido a que se busca encontrar patrones que apoyen a la correcta clasificación de los clientes. Luego del análisis se encontró que que el modelo catboost es el que lanza el mejor resultado con el mejor f1-score (0.91) , comparado con los otros modelos.

***Index Terms*—python, pandas, banco, fuga, retención, modelo**

I. INTRODUCCIÓN

Muchas organizaciones eventualmente enfrentan una situación en la que un cliente decide dejar la organización. Sin embargo, surgen preguntas, especialmente cuando se trata de las razones por las que el cliente decidió abandonar el producto o servicio.

Definitivamente hay razones que no son consistentes con otros clientes que han dejado una organización, pero ¿qué pasa si existe un patrón en cuanto a la razón por la que los clientes decidieron dejar la organización?

Si la empresa pudiera detectar las principales razones por las que los clientes abandonan la empresa, podría reaccionar y evitar que los clientes se vayan. Además, si la empresa puede comprender el pasado, la organización podrá evitar más pérdidas de clientes en el futuro.

Lo descrito anteriormente es lo que motivó a este proyecto, el que tiene como objetivo predecir cual es son los clientes potenciales al abandono del servicio de tarjetas de crédito a través de modelos de Machine Learning.

II. TRABAJOS RELACIONADOS

Existen varios trabajos anteriores relacionados a modelos predictivos para predecir clientes con tendencia a la deserción de los servicios de tarjetas de crédito, aquí vamos a citar 2 de ellos.

El primero tiene como título de Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine [1] en el cual usan un modelo de predicción basado en el comportamiento transaccional y datos demográficos de los clientes para la determinación de los patrones de reconocimientos con la que para ello definieron técnicas y algoritmos para validar su propuestas.

El segundo tiene como título de Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos [2], en esta se desarrolló una interfaz web como canal entre el Modelo de Análisis Predictivo propuesto y la entidad bancaria, con el fin de mostrar el resultado obtenido por el modelo indicando la exactitud, en porcentaje, de los clientes con tendencia a desertar .

III. MÉTODO

La plataforma Kaggle nos brindó la data de clientes de un banco almacenados en formato .csv y con su respectiva descripción. Esta data contiene un peso de 1.44 MB en donde se encuentra toda la información principal a utilizar.

Ahora, este conjunto de datos consta de 10,000 clientes que mencionan su edad, salario, estado marital, límite de tarjeta de crédito, categoría de tarjeta de crédito, etc. Hay casi 18 características a analizar en el presente proyecto.

A continuación, se muestra la descripción de cada variable que nos proporciona la data de los clientes:

Variable	Tipo	Descripcion
CLIENTNUM	Numérica	ID del cliente
Attrition_Flag	Categorica	1: Permanencia, 0: Fuga
Gender	Categorica	1: Male, 0: Female
Education_Level	Categorica	Graduate , High School, Unknown, Uneducated, College, Post-Graduate, Doctorate
Marital_Status	Categorica	Married, Single, Unknown, Divorced
Income_Category	Categorica	($) Menos de 40K, 40K - 60K, 80K - 120K, 60K - 80K, Unknown, 120K
Card_Category	Categorica	Blue, Silver, Gold, Platinum
Customer_Age	Numérica	Edad del cliente en años
Dependent_count	Numérica	Número de dependientes
Months_on_book	Numérica	Periodo de relación con el banco
Total_Relationship_Count	Numérica	Número total de productos en poder del cliente
Months_Inactive_12_mon	Numérica	Número de meses inactivos en los últimos 12 meses
Contacts_Count_12_mon	Numérica	Número de contactos en los últimos 12 meses
Credit_Limit	Numérica	Límite de crédito en la tarjeta de crédito
Total_Revolving_Bal	Numérica	Saldo rotatorio total en la tarjeta de crédito
Avg_Open_To_Buy	Numérica	Línea de crédito abierta para comprar (promedio de los últimos 12 meses)
Total_Amt_Chng_Q4_Q1	Numérica	Cambio en el monto de la transacción (Q4 sobre Q1)
Total_Trans_Amt	Numérica	Monto total de la transacción (últimos 12 meses)
Total_Trans_Ct	Numérica	Recuento total de transacciones (últimos 12 meses)
Total_Ct_Chng_Q4_Q1	Numérica	Cambio en el recuento de transacciones (Q4 sobre Q1)
Avg_Utilization_Ratio	Numérica	Índice de utilización promedio de la tarjeta

Se procedió a hacer la lectura de los datos en la plataforma de Google Collab, con las librerías Pandas y Numpy de Python se inició una exploración de los datos para determinar si existen vacíos en la data, cuales son los máximos, promedios y mínimos de cada variable numérica, etc.

Posteriormente se realizaron gr;aficas de los datos con el uso de la librería Matplotlib y Seaborn. Los gráficos realizados fueron dependiendo del tipo de variable; para las cualitativas tenemos histogramas y pie charts. Para las variables numéricas se realizó boxplot y displots las cuales juntos con los valores estadísticos nos ayudaron a ver la distribución y asimetría.

IV. EXPERIMENTO

Se procederá a detallar cada paso realizado hasta llegar a la realización de los modelos; los tres primeros pasos pertenecen netamente al análisis de los datos.

A. Carga de datos

Una vez que la data es cargado al drive se procede a ser llamada desde Google Colab. Luego se importa las siguientes librerías: pandas, numpy, seaborn, matplotlib, etc; una vez realizado se procede a hacer la extracción de la data.
Para comprobar que se realizó satisfactoriamente, se procede a ejecutar los siguientes metodos .head() este nos muestra un vistazo de nuestra datos y .shape nos dice cuantos datos y features tiene nuestra data. Shape nos indica que se tiene 10127 observaciones y 23 features. Y head(3) nos mostrara las 1ras 3 observaciones:

CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	5	1	3	12691	777	11914	1.335	1144	42	1.625	0.061
818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	6	1	2	8256	864	7392	1.541	1291	33	3.714	0.105
713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4	1	0	3418	0	3418	2.594	1887	20	2.333	0
769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3	4	1	3313	2517	796	1.405	1171	20	2.333	0.76

Fig. 1. Dataset de los clientes del banco

B. Exploración de datos

Una vez cargado los datos kaggle nos da 2 features: ”, ” que se procede a elimnar directamente debido a que estas provienen de un analisis de naives bayes siendo features que no apoyan a nuestro modelo. Luego de eliminar se va a proceder a encontrar datos generales de la data: saber con cuantos features se queda la data, cuantas observaciones, analizar si hay datos nulos, datos duplicados, a continuacion se detalla lo encontrado.

- Se tiene 15 variables numéricas: ’Customer Age’, ’Dependent count’, ’Months onbook’, ’Total Relationship Count’, etc. Y 6 variables categoricas: ’Attrition Flag’, ’Gender’, ’Education Level’, ’Marital Status’, ’Income Category’, ’Card Category’ .
- No se tiene datos nulos
- No hay clientes duplicados

Luego de un analisis en general se procede a analizar cada variable de interés, para este analisis nos apoyaremos de graficos y estadisticos.

C. Elaboración de graficos

a) Waffle chart para la variable ’Attrition Flag’: El waffle chart elaborado muestra: El 84% de nuestros clientes conservar los servicio de la tarjeta de credito y el 16% son los clientes que fugan, el siguiente grafico muestra ese detalle:

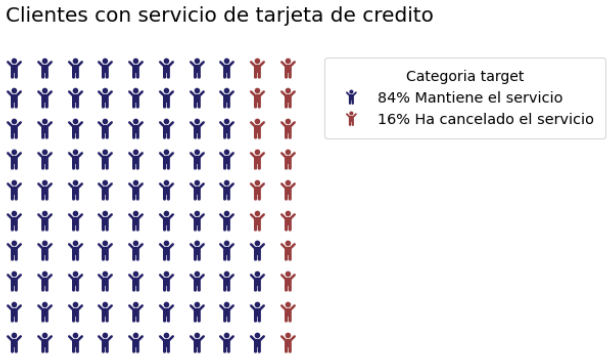
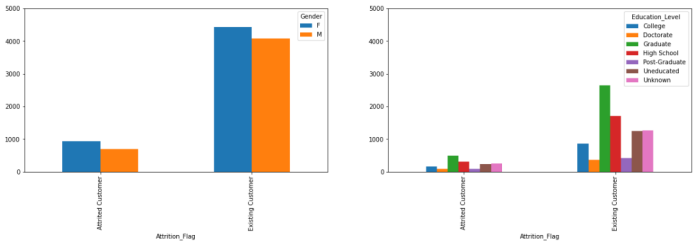


Fig. 2. Porcenataje de clientes que se van o mantienen la tarjeta de credito.

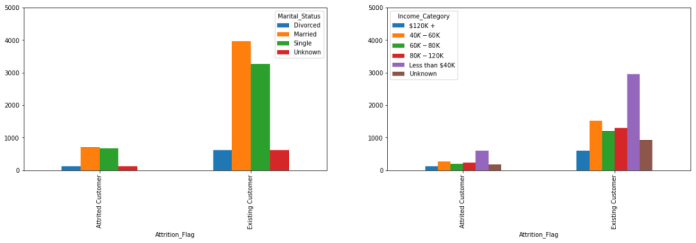
b) Histograma para las variables categoricas: Se hara un histograma para las 6 variables categoricas que se tiene.

- **Variable ’Gender’ y ’Education Level’**
  Con respecto a ’Gender’ se obtiene que el 52.9% de nuestros datos pertenecen a mujeres y se observa que el mayor porcentaje de este mantiene el servicio de su tarjeta de credito. Con respecto a ’Education level’ se

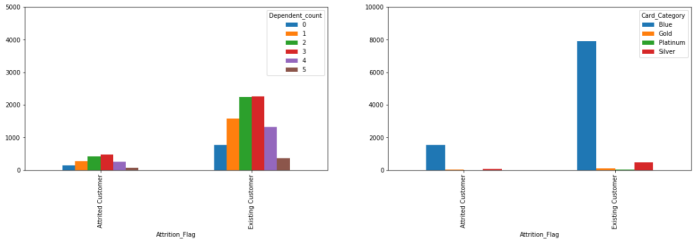
obtiene que el 40.44% de nuestros datos pertenecen a educacion superior(Graduate, Post-Graduate, Doctorate), el 29.88% pertenece a educacion basica y los demas datos pertenecen a tipo 'Uneducated' y 'Unknown'. De las cuales se observa que la mayoria de personas tienden a mantener el servicio



- **Variable 'Marital status' y 'Income category'**
  Con respecto a 'Income category' se obtiene que el 35.16% de nuestra data gana menos de 40K , seguido de 32.84% estan entre 40K y 80K. Y por otro lado la variable 'Marital status' se obtiene la gran cantidad de datos que tienen a cancelar su servicios son casados y solteros.



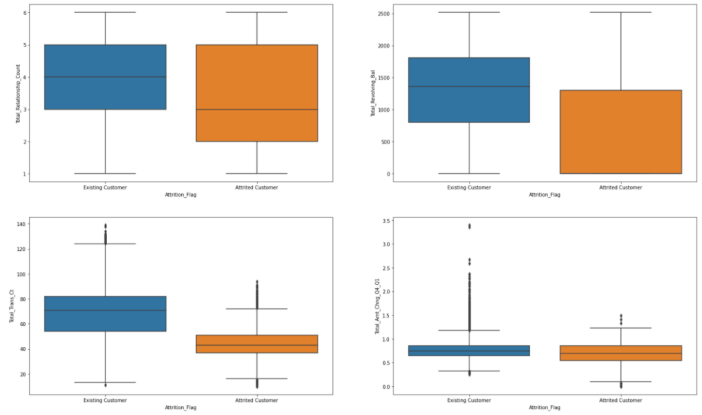
- **Variable 'Dependent count' y 'Card category'**
  Con respecto a 'Dependent count' se obtiene que los clientes tienen entre 2 o 3 personas que dependen de ellos. Y por otro lado la variable 'Card category' nos indica que el 93.2% de nuestros clientes usa la categoria "blue" seguido de silver el cual representa el 5.5%, ect.



c) Boxplot para variables numericas: Se procede a hacer un Boxplot para las variables numericas de interes.

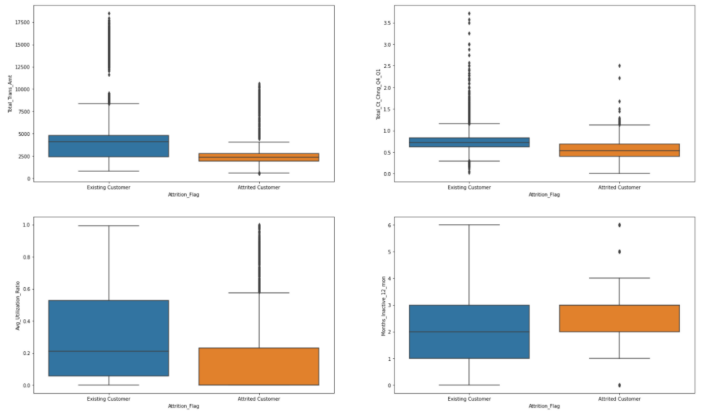
- **Variable 'Total Relationship Count', 'Total Revolving Bal', 'Total Amt Chng Q4 Q1' y 'Total Trans Amt'**

  * 'Total Relationship Count': Gráficamente no sigue una distribución normal y no posee outliers (Se deberia aplicar una transformación).
  * 'Total Revolving Bal': Se observa que no sigue una distribución normal y no posee outliers.
  * 'Total Amt Chng Q4 Q1': Gráficamente se aproxima a una normal pero posee outliers superiores e inferiores.
  * 'Total Trans Amt': No sigue una distribución normal, posee outliers superiores.

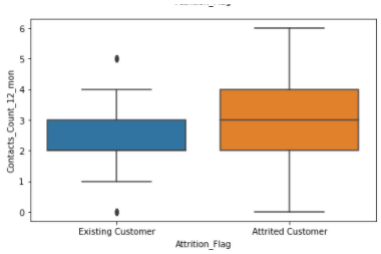


- **Variable 'Total Trans Ct', 'Total Ct Chng Q4 Q1', 'Avg Utilization Ratio' y 'Months Inactive 12 mon'**

  * 'Total Trans Ct': No sigue una distribución normal, posee outliers superiores..
  * 'Total Ct Chng Q4 Q': Sigue una distribución normal, posee outliers superiores e inferiores.
  * 'Avg Utilization Ratio': No posee una distribucion normal, no posee outliers.
  * 'Months Inactive 12 mon': Gráficamente se observa una leve aproximacion a la distribución normal , posee algunos outliers.

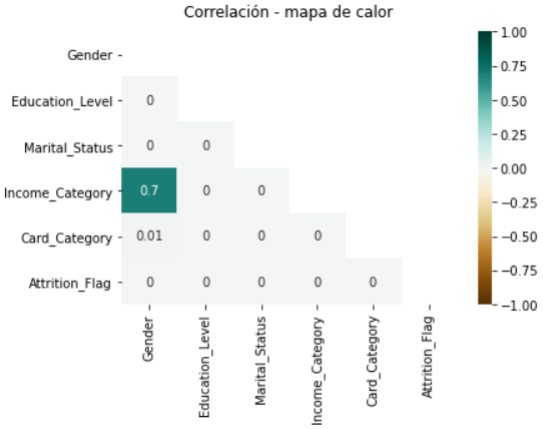


- **Variable 'Contacts Count 12 mon'**
  Gráficamente posee una aproximación a la distribución normal con algunos outliers.

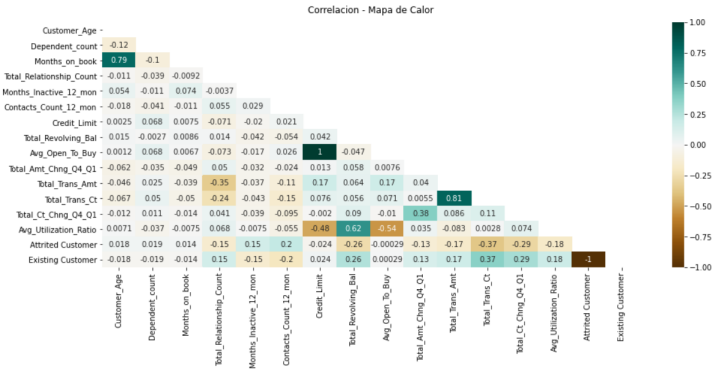


D. Análisis de Correlación

- **Variables Categoricas** Se observa que no existe alguna correlacion entre este tipo de variables y la variable objetivo.



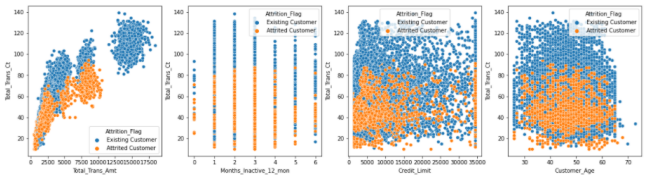
- **Variables Numericas** Dentro de las variables numericas se puede observar que tanto el limite de credito, el monto promedio para comprar, el tiempo que el cliente este en el banco, la edad y el numero de dependiente en el hogar poseen una correlacion menor a 0.1 (+,-) por lo que no seran consideradas en el modelo.



E. Resultados de analisis exploratorio

De los siguientes graficos se puede señalar que:

- Cuanto mayor sea el gasto anual, es más probable que los clientes se queden..
- Después de los 2 meses de inactividad, los clientes son más propensos a marcharse.
- Cuanto más alto sea el límite de crédito, es más probable que los clientes permanezcan.
- La distribución de la edad no importa realmente en este caso, porque los grupos se superponen en gran medida.
- Casi todos los clientes que se dan de baja han utilizado sus tarjetas menos de 100 veces.



F. Outliers

Nuestra data inicial tiene 10127 observaciones; ahora se proceder a la limpieza de Outliers de las sgtes 14 variables: 'Customer Age', 'Dependent count', 'Months on book','Total Relationship Count', 'Months Inactive 12 mon','Contacts Count 12 mon', 'Credit Limit', 'Total Revolving Bal','Avg Open To Buy', 'Total Amt Chng Q4 Q1','Total Trans Amt', 'Total Trans Ct', 'Total Ct Chng Q4 Q1', 'Avg Utilization Ratio'. Al evaluar nuestra data con el metodo z-score se encuentra 0.08% de observaciones mayores a 3 lo que nos indicaria que hay 810 observaciones que estan alejadas de la media, al ser un porcentaje bajo con respecto a nuestra data inicial se procede a eliminarlas; quedandonos entonces con 9317 datos.

G. Ingeniería de Variables

Para el presente proyecto se procedio a realizar lo siguientes cambios en las variables para que todos nuestros modelos posean el mismo pre-procesamiento:

- A todas nuestras variables categóricas se las codifico mediante la técnica de labelenconding la cual asigna a cada una de sus categorías un número empezando desde 0.

Estado Civil	Numero
Divorciado	0
Soltero	1
Casado	2

- Antes de nuestro modelo se realizó la estandarizacion de nuestras variables, ya que de esta manera se optimizara el tiempo de procesamiento en nuestros modelos a trabajar.
- Se realizo tambien el método SMOTE (Synthetic Minority Oversampling Technique) para nuestro problema de data desbalanceada. Este método sintetiza elementos para la clase minoritaria, basándose en los que ya existen. Funciona eligiendo aleatoriamente un punto k de la clase minoritaria y calculando los k vecinos más cercanos para este punto. Los puntos sintéticos generados se añaden entre el punto elegido y sus vecinos.

H. Modelado

Elegimos nuestras variables predictoras denotadas por ”X” y nuestra variable a predecir denotada por ”y” que vienen a ser las siguientes variables que fueron creadas en el anterior paso, finalmente nuestros dataframes tal como se muestra en el siguiente gráfico, queda conforme para aplicar los distintos modelos:

X

	Gender	Education_Level	Marital_Status	Income_Category	Card_Category	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon
0	1.0	2.0	1.0	1.0	0.0	3.0	1.0	2.0
1	1.0	5.0	3.0	0.0	0.0	5.0	3.0	2.0
2	0.0	2.0	1.0	4.0	0.0	5.0	2.0	2.0
3	0.0	2.0	1.0	5.0	0.0	6.0	1.0	2.0
4	1.0	1.0	0.0	2.0	0.0	5.0	2.0	0.0
...	...	...	...	...	...	...	...	...
9312	0.0	2.0	2.0	4.0	0.0	4.0	1.0	4.0
9313	1.0	6.0	0.0	1.0	0.0	4.0	2.0	3.0
9314	0.0	3.0	1.0	4.0	0.0	5.0	3.0	4.0
9315	1.0	2.0	3.0	1.0	0.0	4.0	3.0	3.0
9316	0.0	2.0	1.0	4.0	3.0	6.0	2.0	4.0

9317 rows × 14 columns

Fig. 3. Variables predictoras.

Total_Revolving_Bal	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
1247.0	1.376	1088.0	24.0	0.846	0.311
1467.0	0.831	1201.0	42.0	0.680	0.217
680.0	1.190	1570.0	29.0	0.611	0.279
1157.0	0.966	1207.0	21.0	0.909	0.080
1800.0	0.906	1178.0	27.0	0.929	0.086
...	...	...	...	...	...
606.0	0.570	14596.0	120.0	0.791	0.164
2186.0	0.804	8764.0	69.0	0.683	0.511
0.0	0.819	10291.0	60.0	0.818	0.000
0.0	0.535	8395.0	62.0	0.722	0.000
1961.0	0.703	10294.0	61.0	0.649	0.189

Fig. 4. Las demás variables predictoras.

```
y

0       1
1       1
2       1
3       1
4       1
       ..
9312    1
9313    0
9314    0
9315    0
9316    0
Name: Attrition_Flag
```

Fig. 5. Variable a predecir ”y”.

Luego se procede a importar la librería sklearn el cual nos permitirá el uso de algoritmos tales como el random forest, regresion logística y gradient boosting y sus respectivas métricas, además se importan librerias como lightgbm y catboost, para el uso de sus respectivos algoritmos, que se mencionan a continuación LGBM y catBoost y sus respectivas métricas. Para esta investigación se de analizará 5 modelos(Random forest, regresión logística, LGBM, gradient boosting y catboost) y poder compararlos y elegir el mejor modelo tomando en cuenta la métrica del f1-score y que dicho modelo tenga una mayor precision en cuanto a la predicción del ”Attrited Customer”, a continuación se procede a mostrar cada modelo desarrollado y al final la comparación de los modelos:

a) Modelo Random Forest: Al desarrollar el modelo de random forest, podemos obtener estos 3 figuras importantes que nos ayudaran a entender el modelo y de que forma se trabajó y que resultados nos arrojan.

- **Matriz de confusión**
  Esta herramienta nos permite visualizar el desempeño de nuestro algoritmo de Random forest. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. De 434 datos reales de fuga el algoritmo predijo que 39 eran de permanencia y de 2362 datos reales de permanencia predijo que 81 era de fuga.

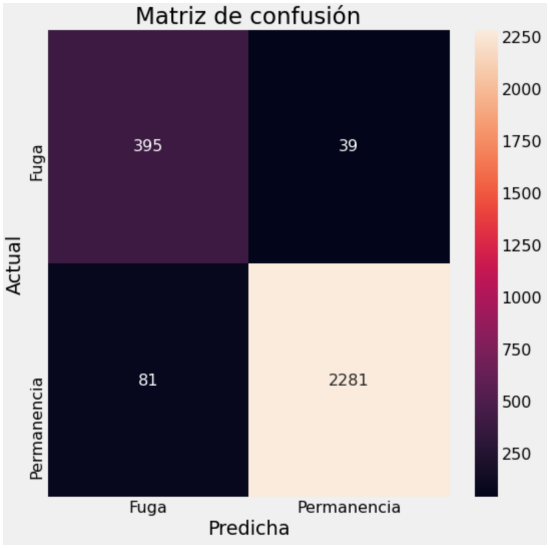


Fig. 6. Matriz de confusión.

- **Métricas**
  En el siguiente gráfico podemos visualizar las distintas métricas con las cuales podemos tomar la decisión de nuestro modelo, pero en esta investigación nos enfocaresmos en visualizar y obtener el mayor valor de la métrica f1-score por el lado de ”Attrited Customer”(fuga). Además nos enfocamos en esta métrica(f1-score) porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad. Con Random forest obtenemos un f1-score de 0.87.

	precision	recall	f1-score	support
0	0.84	0.91	0.87	434
1	0.98	0.97	0.98	2362
accuracy			0.96	2796
macro avg	0.91	0.94	0.92	2796
weighted avg	0.96	0.96	0.96	2796

Fig. 7. Métricas del modelo.

- **Curva ROC**
  La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad. Además este gráfico es la representación de la razón o proporción de verdaderos positivos frente a la razón o proporción de falsos positivos. La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Con Random forest obtenemos un AUC de 0.98.

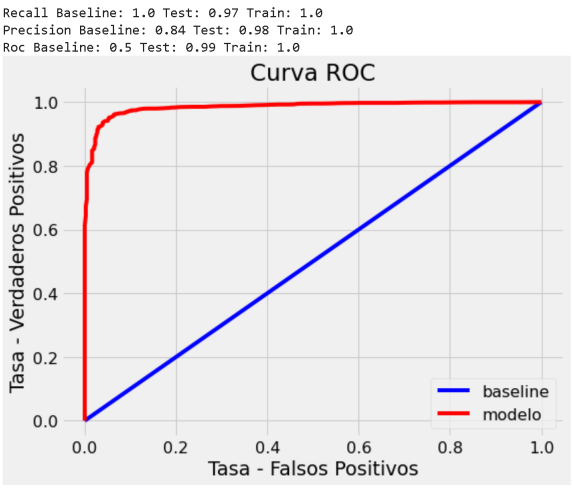


Fig. 8. Curva ROC.

b) Modelo Regresión Logística: Al desarrollar el modelo de Regresión Logística, podemos obtener estos 3 figuras importantes que nos ayudaran a entender el modelo y de que forma se trabajó y que resultados nos arrojan.

- **Matriz de confusión**
  Esta herramienta nos permite visualizar el desempeño de nuestro algoritmo de Regresión Logística. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. De 434 datos reales de fuga el algoritmo predijo que 66 eran de permanencia y de 2362 datos reales de permanencia predijo que 386 era de fuga.

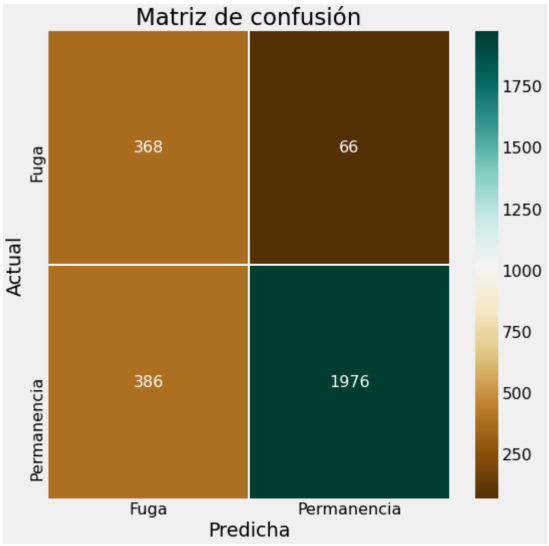


Fig. 9. Matriz de confusión.

- **Métricas**
  En el siguiente gráfico podemos visualizar las distintas métricas con las cuales podemos tomar la decisión de nuestro modelo, pero en esta investigación nos enfocaresmos en visualizar y obtener el mayor valor de la métrica f1-score por el lado de ”Attrited Customer”(fuga). Además nos enfocamos en esta métrica(f1-score) porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad. Con Regresión Logística obtenemos un f1-score de 0.62.

	precision	recall	f1-score	support
0	0.49	0.85	0.62	434
1	0.97	0.84	0.90	2362
accuracy			0.84	2796
macro avg	0.73	0.84	0.76	2796
weighted avg	0.89	0.84	0.85	2796

Fig. 10. Métricas del modelo.

- **Curva ROC**
  La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad. Además este gráfico es la representación de la razón o proporción de verdaderos positivos frente a la razón o proporción de falsos positivos. La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Con Regresión Logística obtenemos un AUC de 0.93.

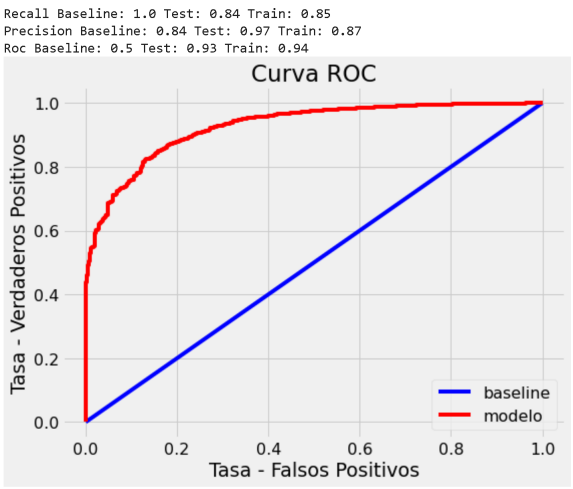


Fig. 11. Curva ROC.

c) Modelo LGBM: Al desarrollar el modelo de LGBM, podemos obtener estos 3 figuras importantes que nos ayudaran a entender el modelo y de que forma se trabajó y que resultados nos arrojan.

- **Matriz de confusión**
  Esta herramienta nos permite visualizar el desempeño de nuestro algoritmo de LGBM. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. De 434 datos reales de fuga el algoritmo predijo que 41 eran de permanencia y de 2362 datos reales de permanencia predijo que 48 era de fuga.

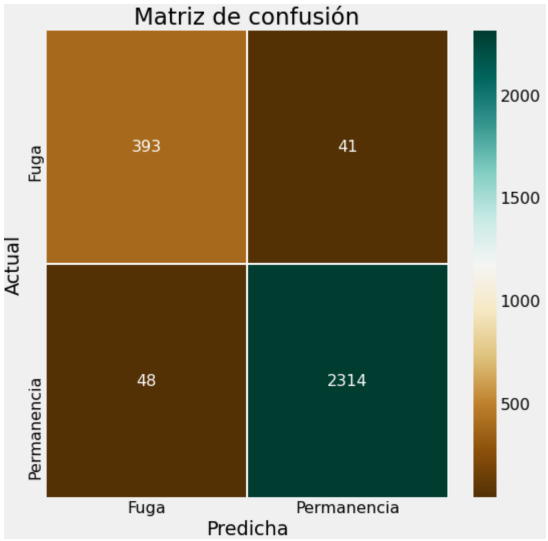


Fig. 12. Matriz de confusión.

- **Métricas**
  En el siguiente gráfico podemos visualizar las distintas métricas con las cuales podemos tomar la decisión de nuestro modelo, pero en esta investigación nos enfocaremos en visualizar y obtener el mayor valor de la métrica f1-score por el lado de "Attrited Customer"(fuga). Además nos enfocamos en esta métrica(f1-score) porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad. Con LGBM obtenemos un f1-score de 0.90.

	precision	recall	f1-score	support
0	0.89	0.91	0.90	434
1	0.98	0.98	0.98	2362
accuracy			0.97	2796
macro avg	0.94	0.94	0.94	2796
weighted avg	0.97	0.97	0.97	2796

Fig. 13. Métricas del modelo.

- **Curva ROC**
  La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad. Además este gráfico es la representación de la razón o proporción de verdaderos positivos frente a la razón o proporción de falsos positivos. La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Con LGBM obtenemos un AUC de 0.99

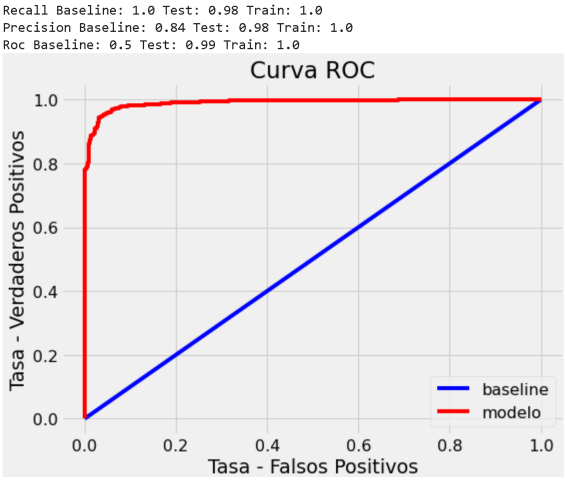


Fig. 14. Curva ROC.

d) Modelo Gradient Boosting: Al desarrollar el modelo de Gradient Boosting, podemos obtener estos 3 figuras importantes que nos ayudaran a entender el modelo y de que forma se trabajó y que resultados nos arrojan.

- **Matriz de confusión'**
  Esta herramienta nos permite visualizar el desempeño de nuestro algoritmo de Gradient Boosting. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. De 434 datos reales de fuga el algoritmo

predijo que 40 eran de permanencia y de 2362 datos reales de permanencia predijo que 84 era de fuga.

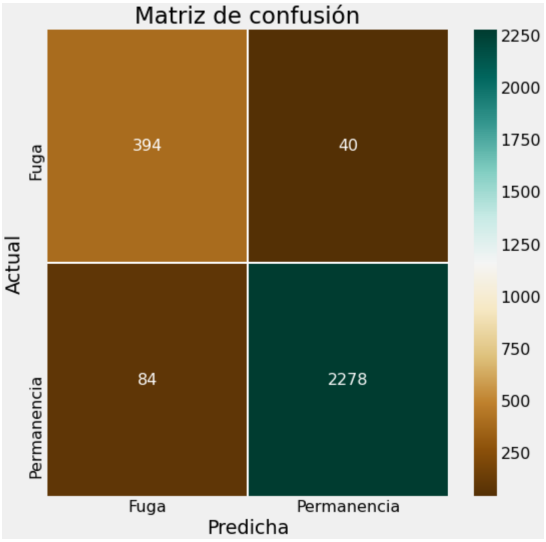


Fig. 15. Matriz de confusión.

- **Métricas**
  En el siguiente gráfico podemos visualizar las distintas métricas con las cuales podemos tomar la decisión de nuestro modelo, pero en esta investigación nos enfocaresmos en visualizar y obtener el mayor valor de la métrica f1-score por el lado de "Attrited Customer"(fuga). Además nos enfocamos en esta métrica(f1-score) porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad. Con Gradient Boosting obtenemos un f1-score de 0.86.

	precision	recall	f1-score	support
0	0.82	0.91	0.86	434
1	0.98	0.96	0.97	2362
accuracy			0.96	2796
macro avg	0.90	0.94	0.92	2796
weighted avg	0.96	0.96	0.96	2796

Fig. 16. Métricas del modelo.

- **Curva ROC**
  La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad. Además este gráfico es la representación de la razón o proporción de verdaderos positivos frente a la razón o proporción de falsos positivos. La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Con Gradient Boosting obtenemos un AUC de 0.99.

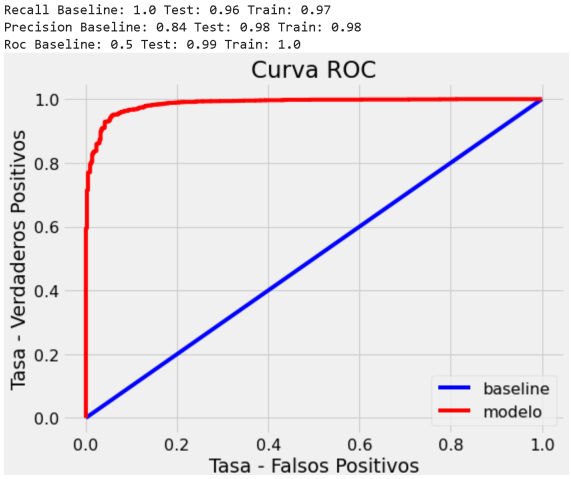


Fig. 17. Curva ROC.

e) Modelo Catboost: Al desarrollar el modelo de Catboost, podemos obtener estos 3 figuras importantes que nos ayudaran a entender el modelo y de que forma se trabajó y que resultados nos arrojan.

- **Matriz de confusión'**
  Esta herramienta nos permite visualizar el desempeño de nuestro algoritmo de Catboost. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. De 434 datos reales de fuga el algoritmo predijo que 38 eran de permanencia y de 2362 datos reales de permanencia predijo que 45 era de fuga.

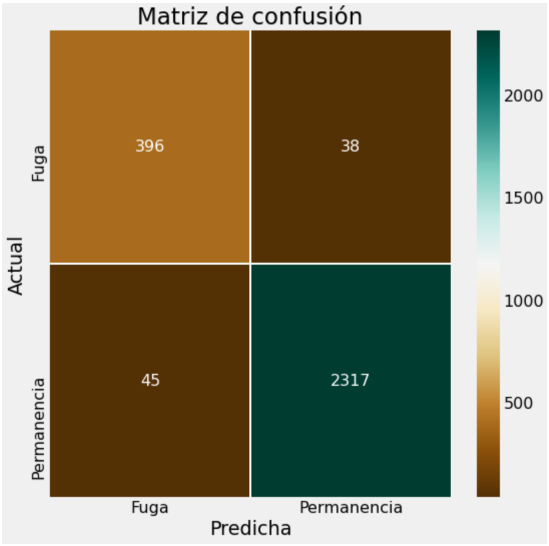


Fig. 18. Matriz de confusión.

- **Métricas**
  En el siguiente gráfico podemos visualizar las distintas métricas con las cuales podemos tomar la decisión de nuestro modelo, pero en esta investigación nos enfo-

caresmos en visualizar y obtener el mayor valor de la métrica f1-score por el lado de "Attrited Customer"(fuga). Además nos enfocamos en esta métrica(f1-score) porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad. Con Catboost obtenemos un f1-score de 0.91.

	precision	recall	f1-score	support
0	0.90	0.91	0.91	434
1	0.98	0.98	0.98	2362
accuracy			0.97	2796
macro avg	0.94	0.95	0.94	2796
weighted avg	0.97	0.97	0.97	2796

Fig. 19. Métricas del modelo.

- **Curva ROC**
  La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad. Además este gráfico es la representación de la razón o proporción de verdaderos positivos frente a la razón o proporción de falsos positivos. La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Con Catboost obtenemos un AUC de 0.99.

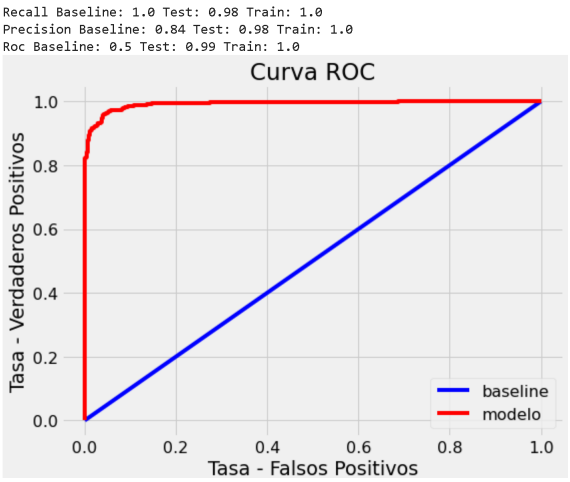


Fig. 20. Curva ROC.

f) Comparación de modelos: Finalmente comparamos los 5 modelos desarrollados en la investigación y elegimos al mejor modelo.

- **Comparación de métricas**
  Al visualizar las 3 metricas dentro de nuestra siguiente figura podemos concluir que el mejor modelo es el Catboost y si nos enfocamos en la métrica f1-score también concluimos que catboost es el mejor modelo para el desarrollo de nuestra predicción de la fuga y permanencia del servicio de tarjetas de crédito para un determinado banco.

Modelo	precision	accuracy	f1-score
Catboost	0.90	0.97	0.91
LGBM	0.89	0.97	0.90
Random Forest	0.83	0.96	0.87
GradientBoosting	0.82	0.96	0.86
Regresion Logística	0.49	0.84	0.62

Fig. 21. Comparación de métricas.

V. CONCLUSIONES Y RECOMENDACIONES

Luego de realizar nuestro análisis y modelos podemos afirmar lo siguiente:

- Cuando un cliente tiene un menor uso de su tarjeta y a la vez se le ha ofrecido una línea de crédito menor es mucho más probable que abandone el servicio de tarjeta de crédito.
- Nuestro mejor modelo para la clasificación de la permanencia o fuga de un cliente para el servicio de tarjeta de crédito, según la métrica f1-score será el modelo catboost el cual nos arroja un valor de 0.91, para la categoría de interés el cual es, 'Attried Customer'(fuga).
- Sería recomendable tener el comportamiento histórico de los clientes en el banco, dentro de nuestra base de datos, para un mejor análisis y poder clasificar no solo en permanencia y fuga sino también poder clasificar con clientes que se encuentren propensos a la fuga y de esta forma tener una mejor alternativa para la toma de decisiones.

REFERENCES

[1] R. Ordoñez, M. Pastor, Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine, Tesis Ing. de Sistemas, Universidad Nacional Mayor de San Marcos, Lima, Perú, 2016

[2] R. Barrueta, E. Castillo, Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos, Tesis Ing. de Sistemas, Universidad Peruana de Ciencias Aplicadas, Lima, Perú, 2018
.