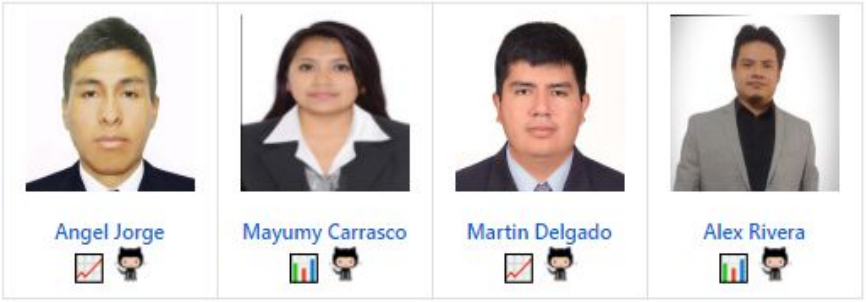
www.datascience.pe

Data Science

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

Grupo 3



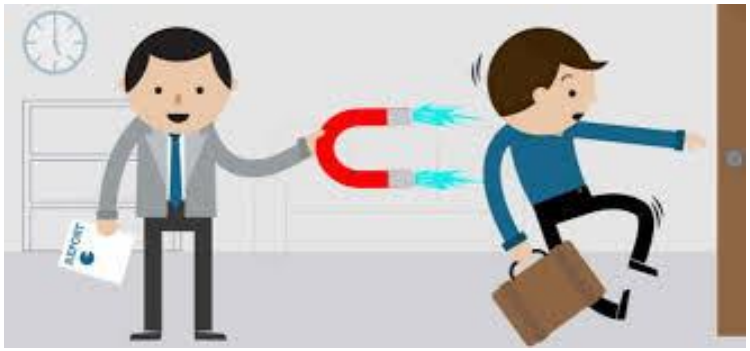
Comprensión del Negocio

www.datascience.pe

Comprensión del Negocio

Un gerente de banco se siente incómodo por que cada vez más clientes abandonan sus servicios de tarjeta de crédito.

Así que se encuentra en busca de soluciones que le permitan tomar acciones de manera proactiva frente al cliente para brindarle un mejor servicio y/u ofertas que ayuden a cambiar su decisión.



Problema

Problema

El aumento progresivo de casos de deserción de clientes del servicio de tarjeta de crédito en un banco.

Planteamiento

¿Cómo un modelo predictivo ayudará a la identificación de la fuga del cliente?

Hipotesis

Hipótesis

Un modelo de predicción nos ayudará a clasificar los tipos de clientes(fuga y permanencia) en más del 50%.

Trabajos relacionados

Artículos

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

E.A.P. DE INGENIERÍA DE SISTEMAS

Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTORES

Rosa Angela del Carmen Ordoñez Cairo

REPOSITORIO ACADÉMICO UPC

Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos

Item Type	info:eu-repo/semantics/bachelorThesis
Authors	Barrueta Meza, Renzo André; Castillo Villarreal, Edgar Jean Paul
Citation	[1] R. A. Barrueta Meza and E. J. P. Castillo Villarreal, "Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos," Universidad Peruana de Ciencias Aplicadas(UPC)., Lima, Perú, 2018. Doi: http://doi.org/ 10.19083/ tesis/626023
DOI	10.19083/tesis/626023
Publisher	Universidad Peruana de Ciencias Aplicadas (UPC)
Rights	info:eu-repo/semantics/openAccess; Attribution-NonCommercial-ShareAlike 3.0 United States
Download date	10/02/2021 00:29:43
Item License	http://creativecommons.org/licenses/by-nc-sa/3.0/us/
Link to Item	http://hdl.handle.net/10757/626023

Método

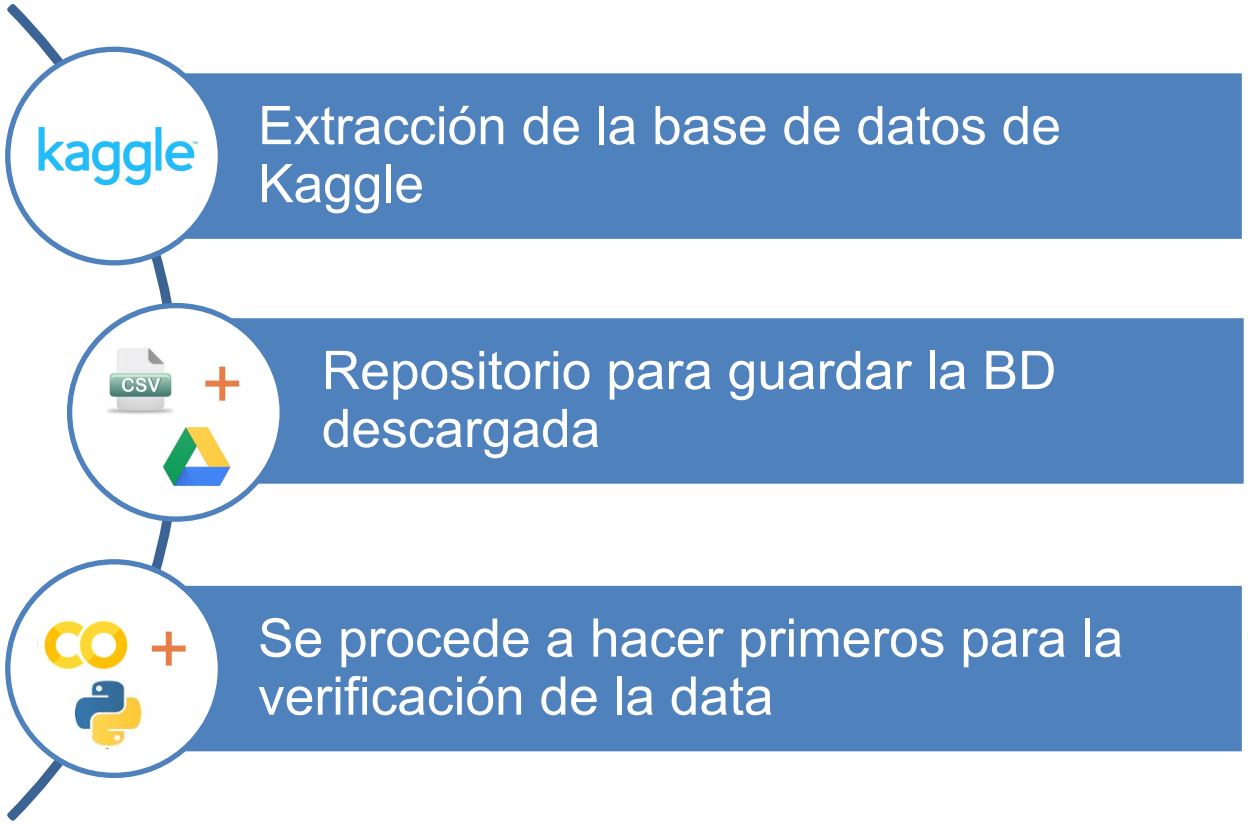
Data Science

Método

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

Experimento

Experimentación



Data Science

A. Carga de datos

CLIENTNUM	Attrition_ Flag	Customer_ Age	Gender	Dependent _count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on _book	Total_Relation ship_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_Utilizat ion_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

10127 observaciones / 23 features

B. Exploración de datos

> Se tiene:
6 variables categóricas y 15 numéricas

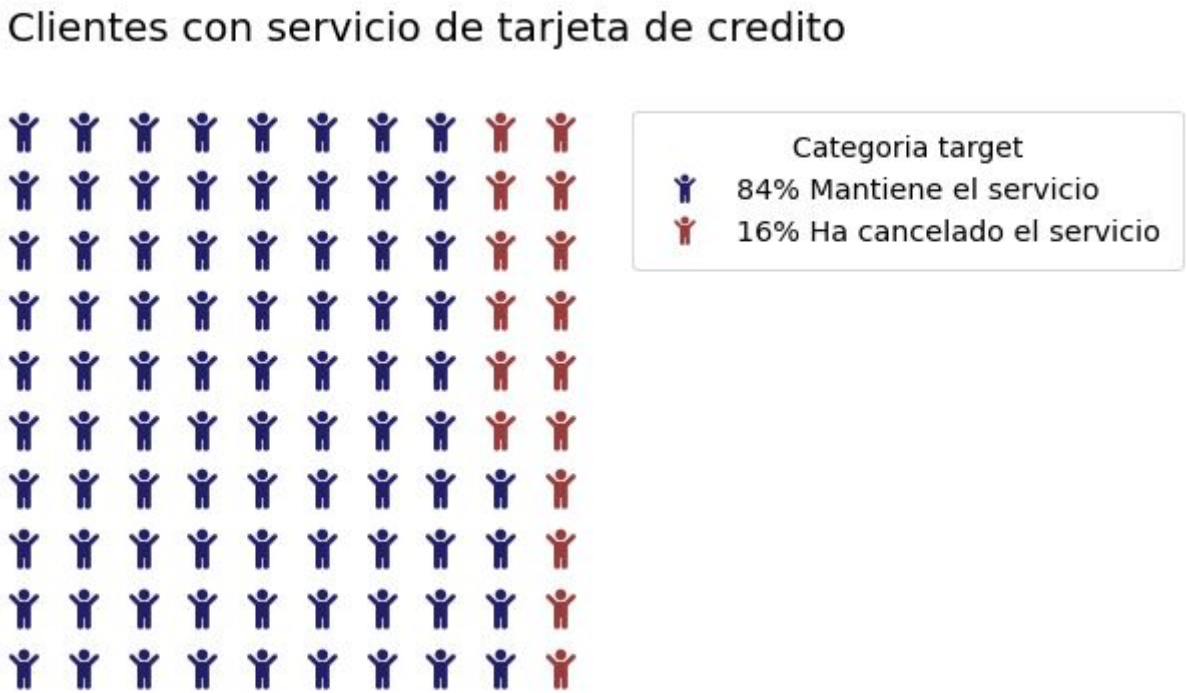
```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CLIENTNUM               10127 non-null  int64
 1   Attrition_Flag          10127 non-null  object
 2   Customer_Age            10127 non-null  int64
 3   Gender                  10127 non-null  object
 4   Dependent_count         10127 non-null  int64
 5   Education_Level         10127 non-null  object
 6   Marital_Status          10127 non-null  object
 7   Income_Category         10127 non-null  object
 8   Card_Category           10127 non-null  object
 9   Months_on_book          10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon  10127 non-null  int64
 12  Contacts_Count_12_mon   10127 non-null  int64
 13  Credit_Limit            10127 non-null  float64
 14  Total_Revolving_Bal     10127 non-null  int64
 15  Avg_Open_To_Buy         10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1    10127 non-null  float64
 17  Total_Trans_Amt         10127 non-null  int64
 18  Total_Trans_Ct          10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1     10127 non-null  float64
 20  Avg_Utilization_Ratio   10127 non-null  float64
```

> No existen datos nulos

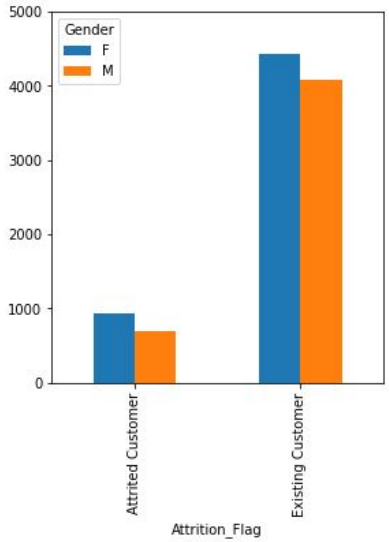
```
CLIENTNUM                 0
Attrition_Flag            0
Customer_Age              0
Gender                    0
Dependent_count           0
Education_Level           0
Marital_Status            0
Income_Category           0
Card_Category             0
Months_on_book            0
Total_Relationship_Count  0
Months_Inactive_12_mon    0
Contacts_Count_12_mon     0
Credit_Limit              0
Total_Revolving_Bal       0
Avg_Open_To_Buy           0
Total_Amt_Chng_Q4_Q1      0
Total_Trans_Amt           0
Total_Trans_Ct            0
Total_Ct_Chng_Q4_Q1       0
Avg_Utilization_Ratio     0
dtype: int64
```

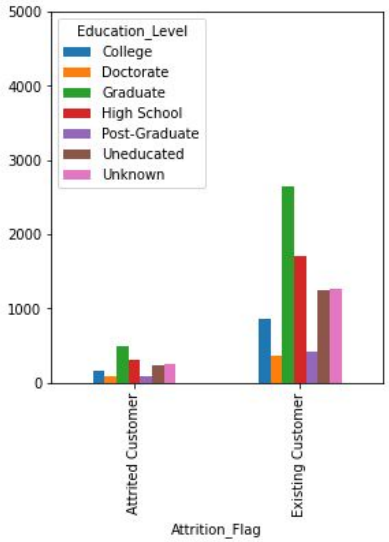
> No hay clientes duplicados

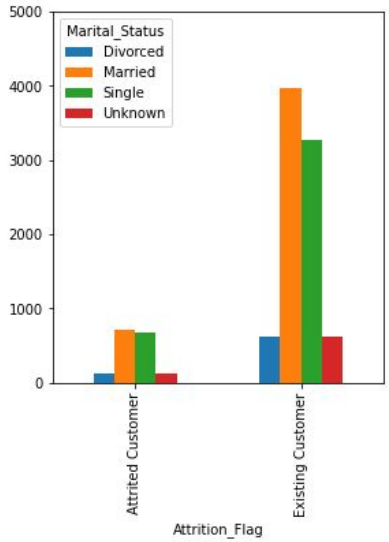
C. Elaboración de gráficos

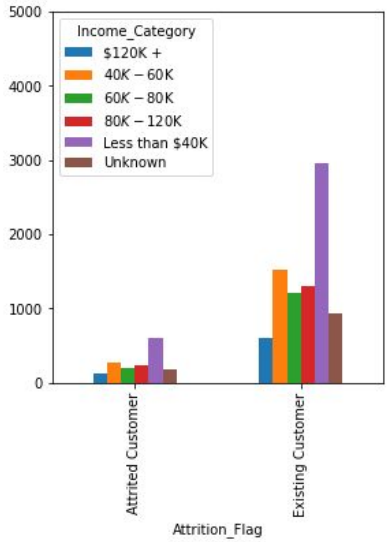


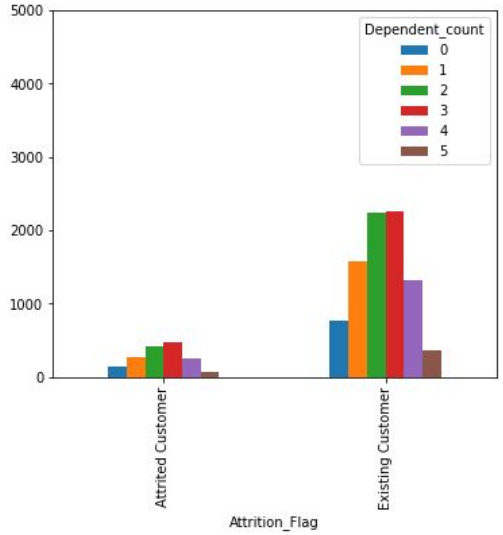
Histograma para las variables categóricas

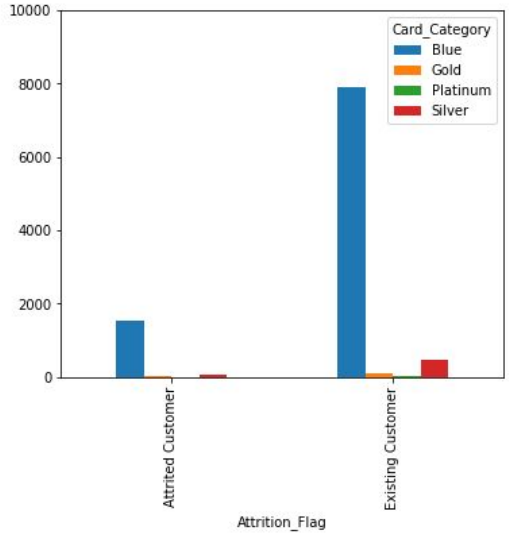




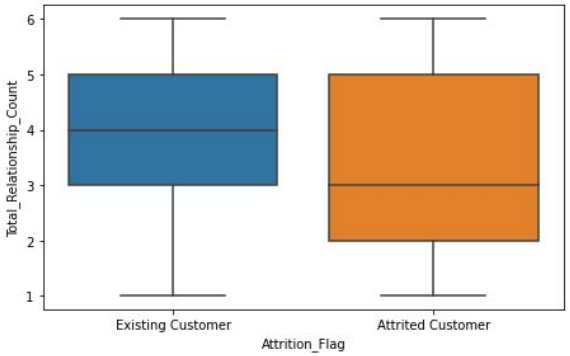


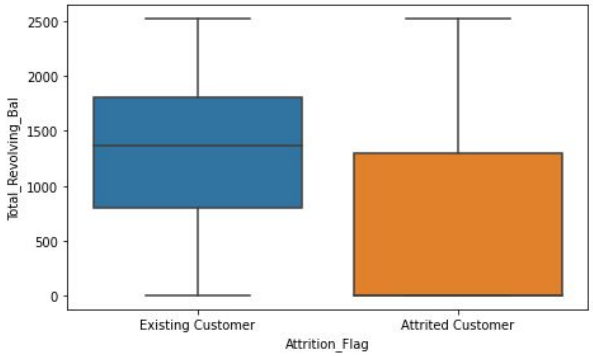


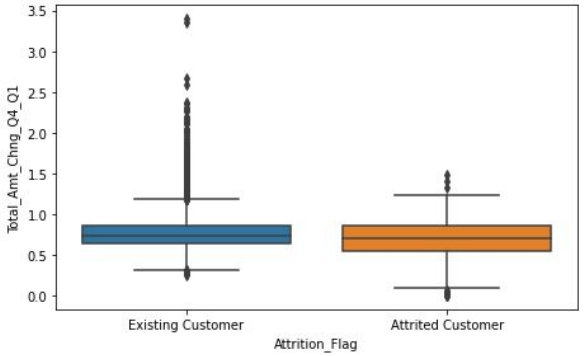


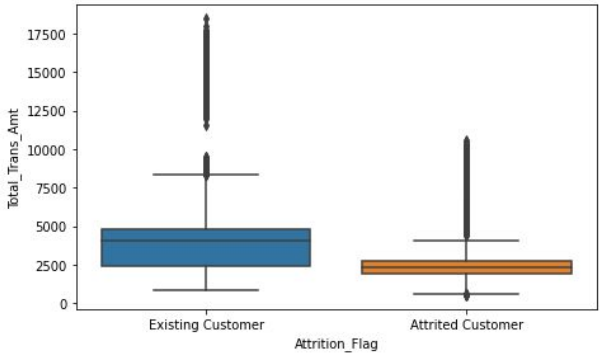


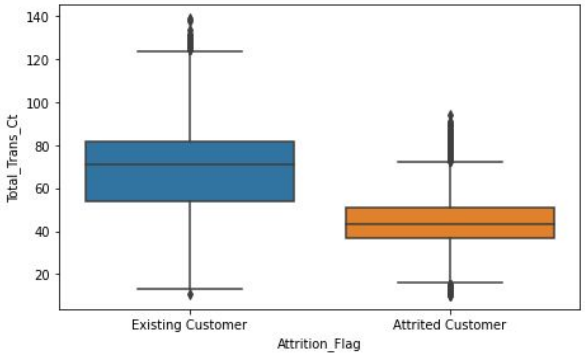
Boxplot para variables numéricas

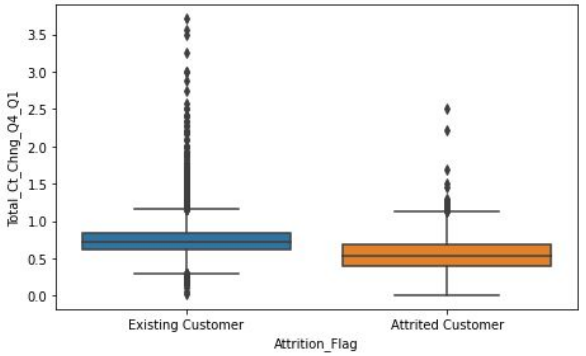


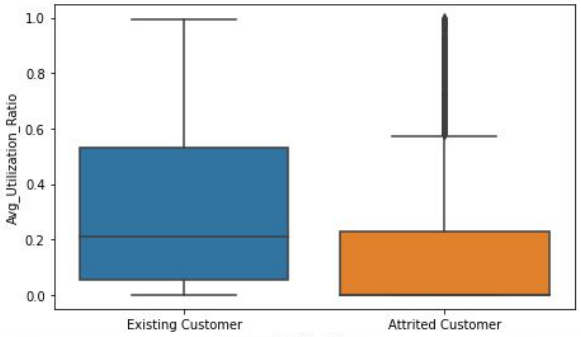


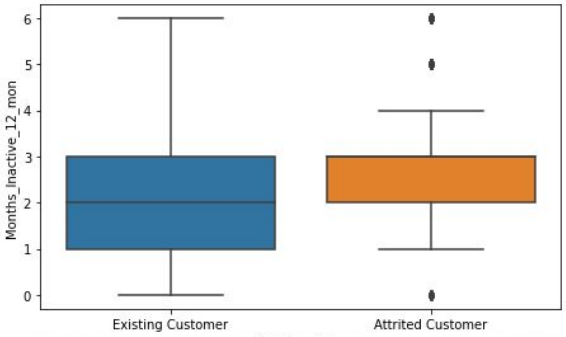


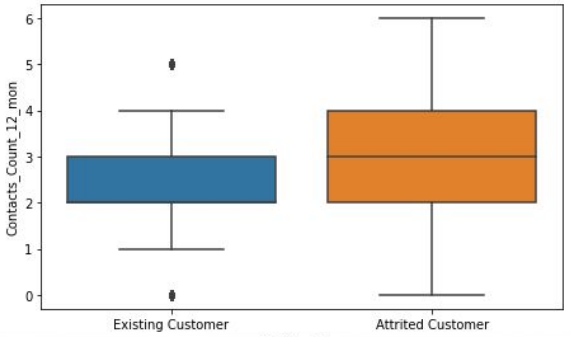








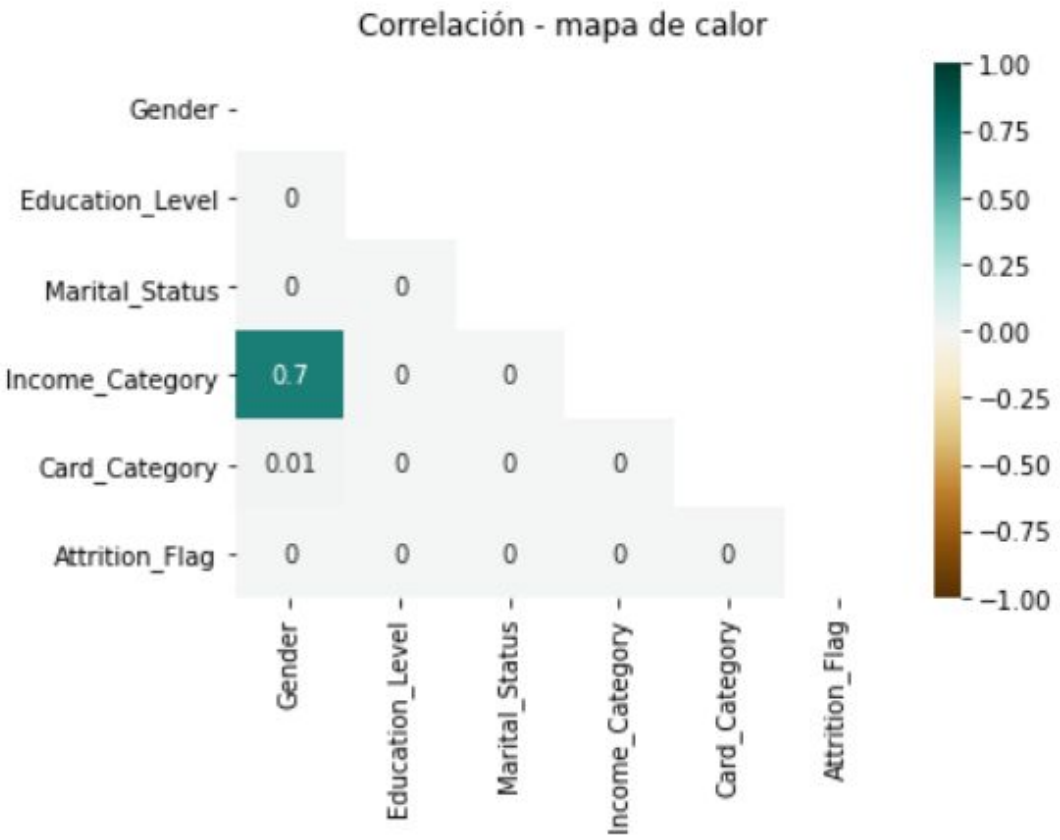




Data Science

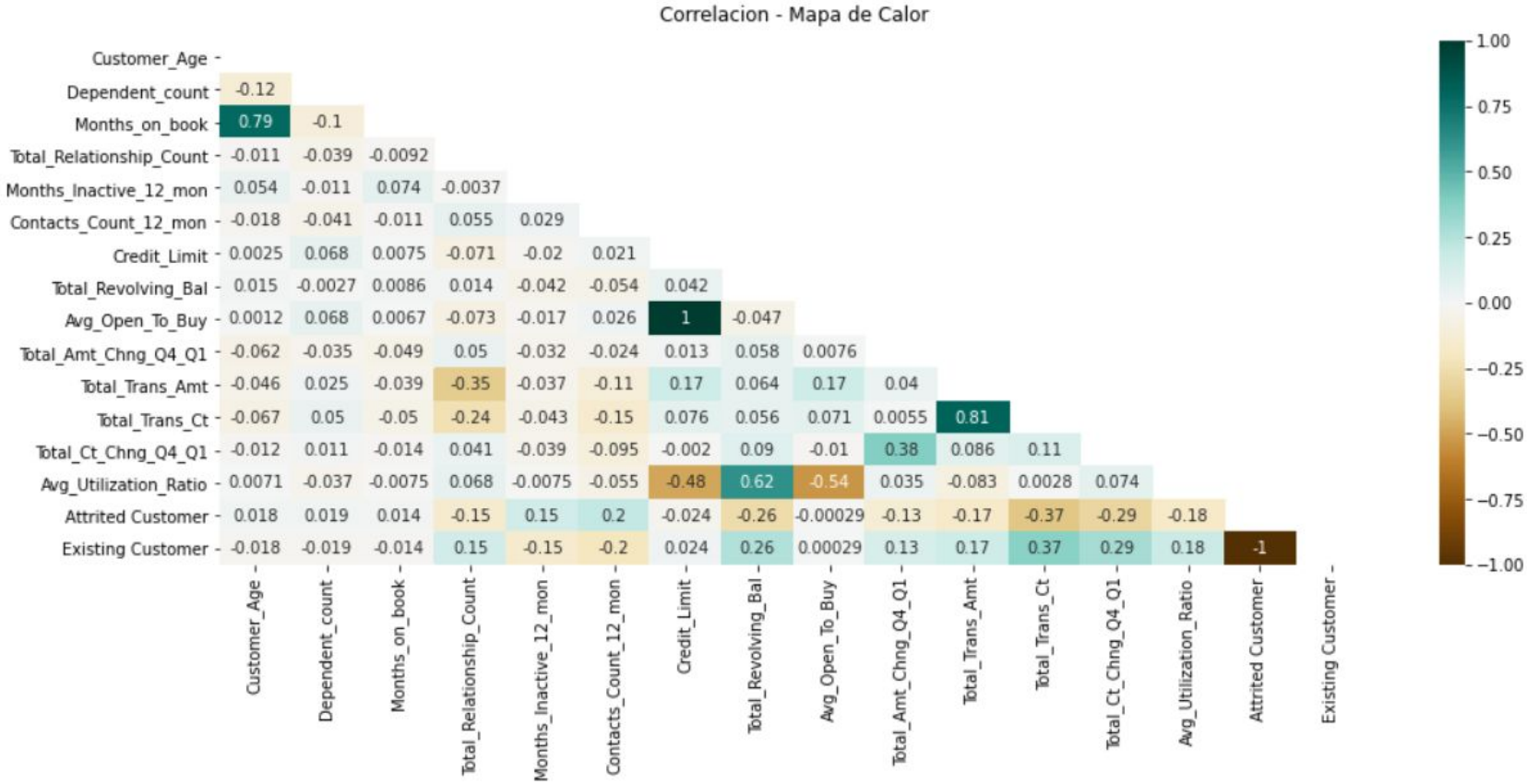
D. Análisis de Correlación

- *Variables categóricas*



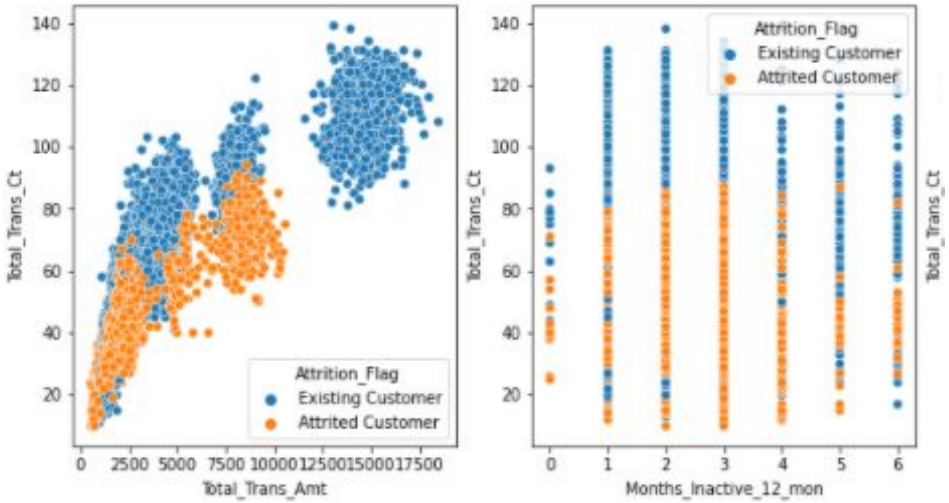
Data Science

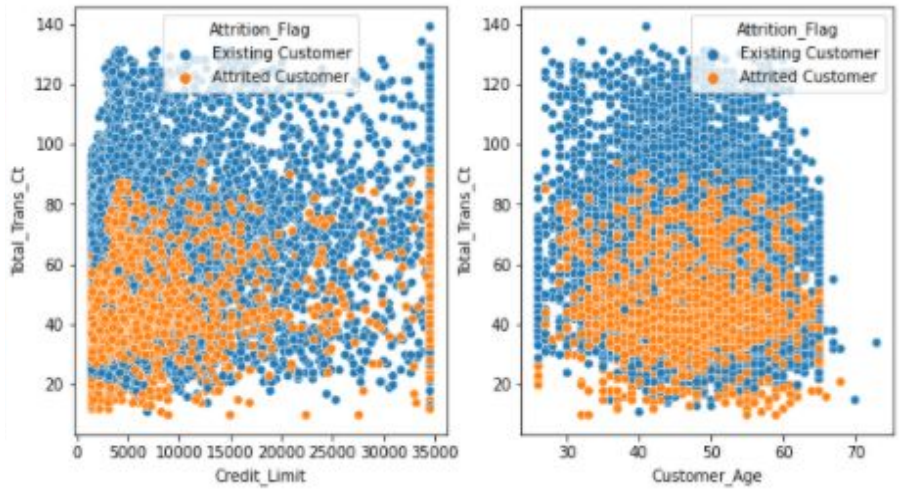
- Variables numéricas



Data Science

E. Resultado del análisis exploratorio



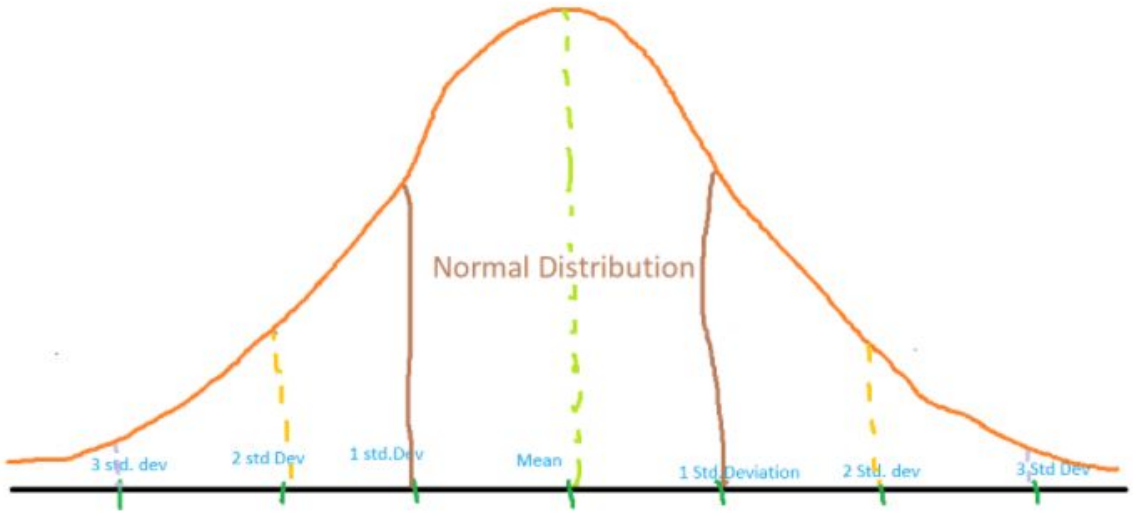


Preparación de los datos

Preparación de los datos

Tratamiento de Outliers: En nuestro caso se aplicó el Z - score. En nuestro caso era cerca del 8% por lo que se procedió a eliminarlos mediante esta técnica.

$$Z\ score = (x - mean) / std.\ deviation$$



LabelEnconding: En la codificación de nuestras variables categóricas se utilizó el método conocido como labelEnconding el cual en grandes rasgos le asigna un valor numérico a cada categoría.

Marital_Status	Encoding
Divorced	0
Married	1
Single	2
Unknown	3

Normalización de los datos: Para este caso se usó la estandarización estándar la cual sustrae la media de la observación y la divide entre la desviación estándar

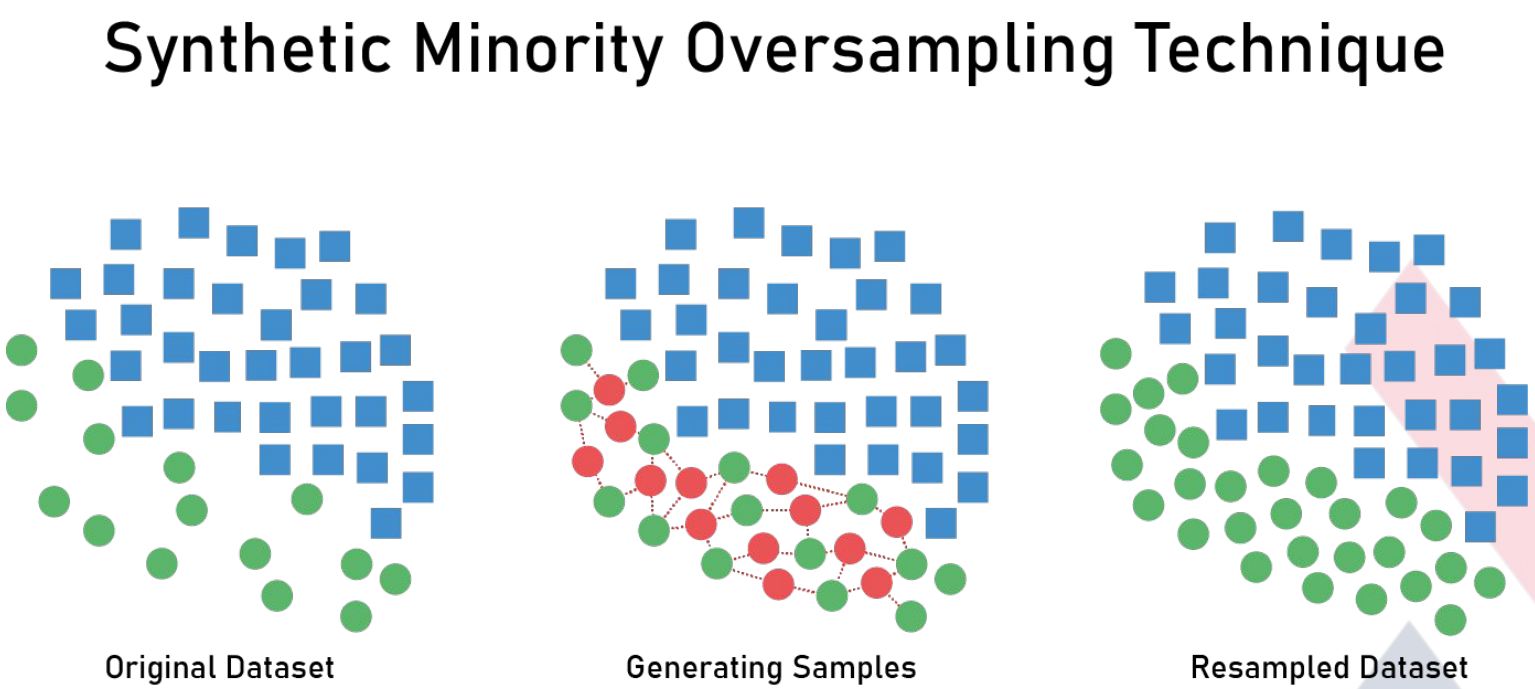
	Gender	Education_Level	Marital_Status	Income_Category	Card_Category	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Total_Revolving_Bal
0	1.079693	-0.592235	-0.637965	-1.243382	-0.249152	-0.550958	-1.394369	-0.408686	0.117766
1	1.079693	1.040910	2.071824	-1.907141	-0.249152	0.739782	0.753847	-0.408686	0.387645
2	-0.926189	-0.592235	-0.637965	0.747896	-0.249152	0.739782	-0.320261	-0.408686	-0.577784
3	-0.926189	-0.592235	-0.637965	1.411655	-0.249152	1.385153	-1.394369	-0.408686	0.007362
4	1.079693	-1.136617	-1.992860	-0.579623	-0.249152	0.739782	-0.320261	-2.261594	0.796142
...	...	...	...	...	...	...	...	...	...
9312	-0.926189	-0.592235	0.716929	0.747896	-0.249152	0.094412	-1.394369	1.444222	-0.668561
9313	1.079693	1.585292	-1.992860	-1.243382	-0.249152	0.094412	-0.320261	0.517768	1.269656
9314	-0.926189	-0.047853	-0.637965	0.747896	-0.249152	0.739782	0.753847	1.444222	-1.411953
9315	1.079693	-0.592235	2.071824	-1.243382	-0.249152	0.094412	0.753847	0.517768	-1.411953
9316	-0.926189	-0.592235	-0.637965	0.747896	4.226461	1.385153	-0.320261	1.444222	0.993644

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

Técnicas de Balanceo de Datos: En nuestro caso aproximadamente el 16% de nuestra base era considerado como un cliente que 'fuga' por lo que se utilizó el Método SMOTE para generar datos sintéticos de la clase minoritaria.



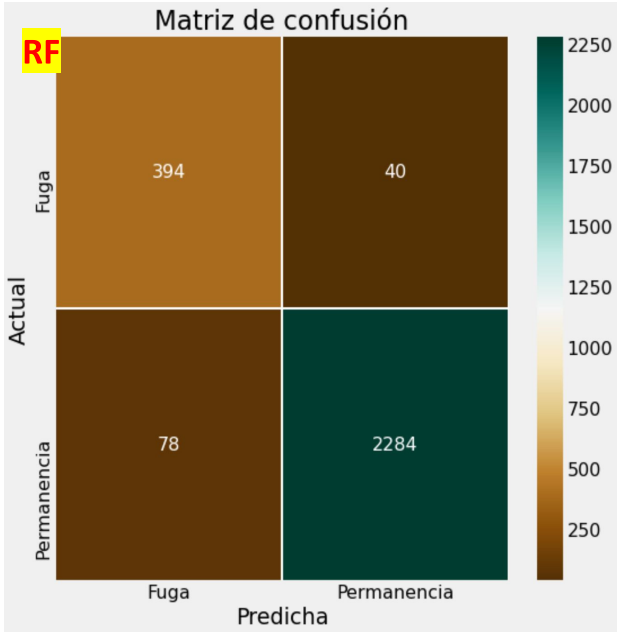
Modelo

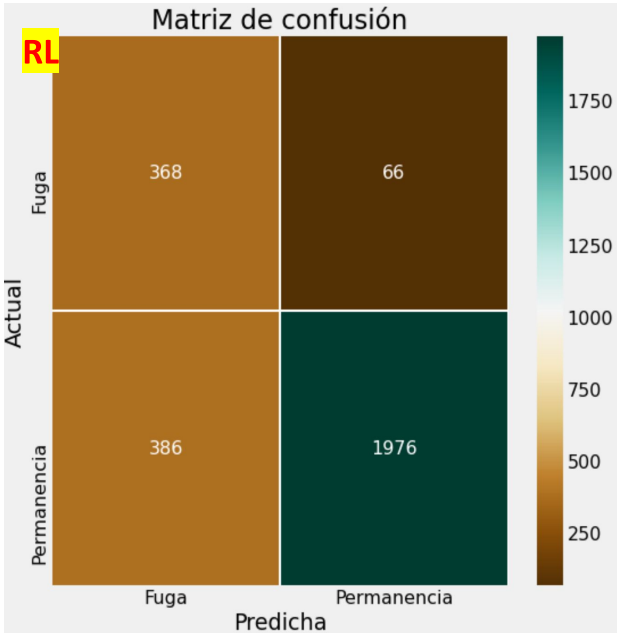
Modelamiento

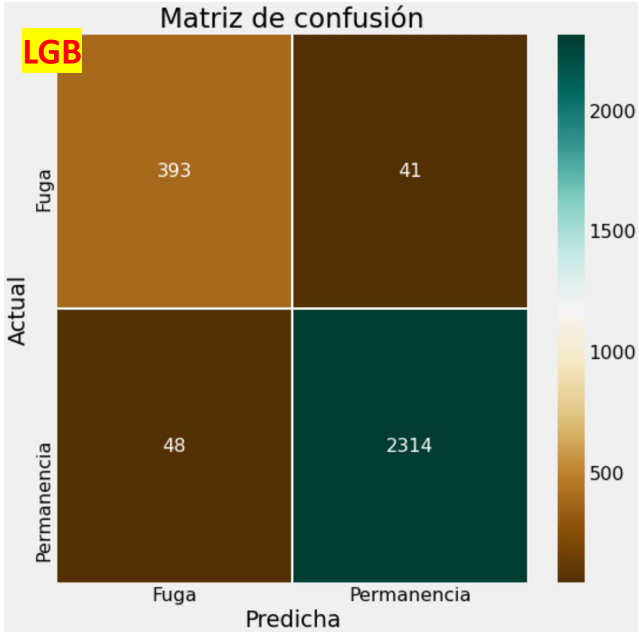
Algoritmos a desarrollar:

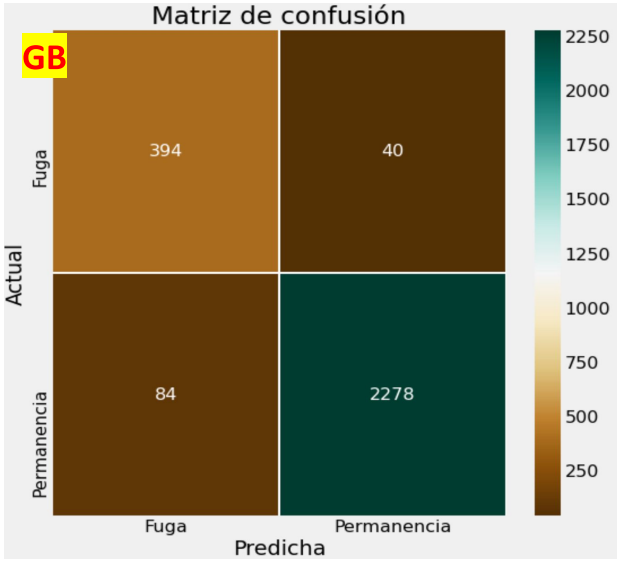
1. Random Forest(RF)
2. Regresión Logística(RL)
3. Light GBM(LGB)
4. Gradiente Boosting(GB)
5. Catboost(CB)

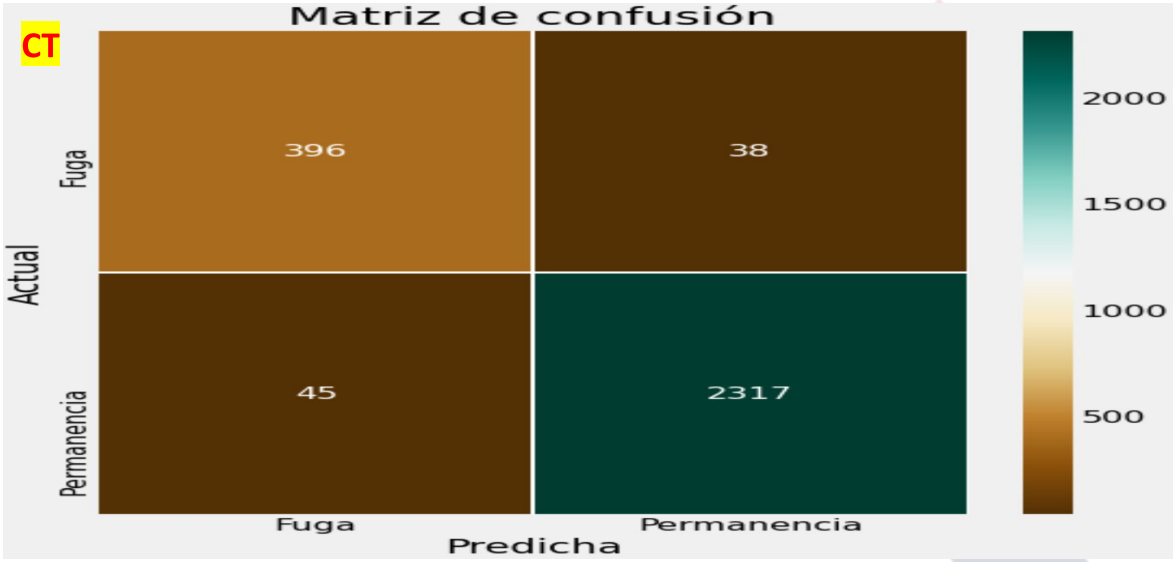
Comparación de matriz de confusión











Matriz de confusión y curva roc

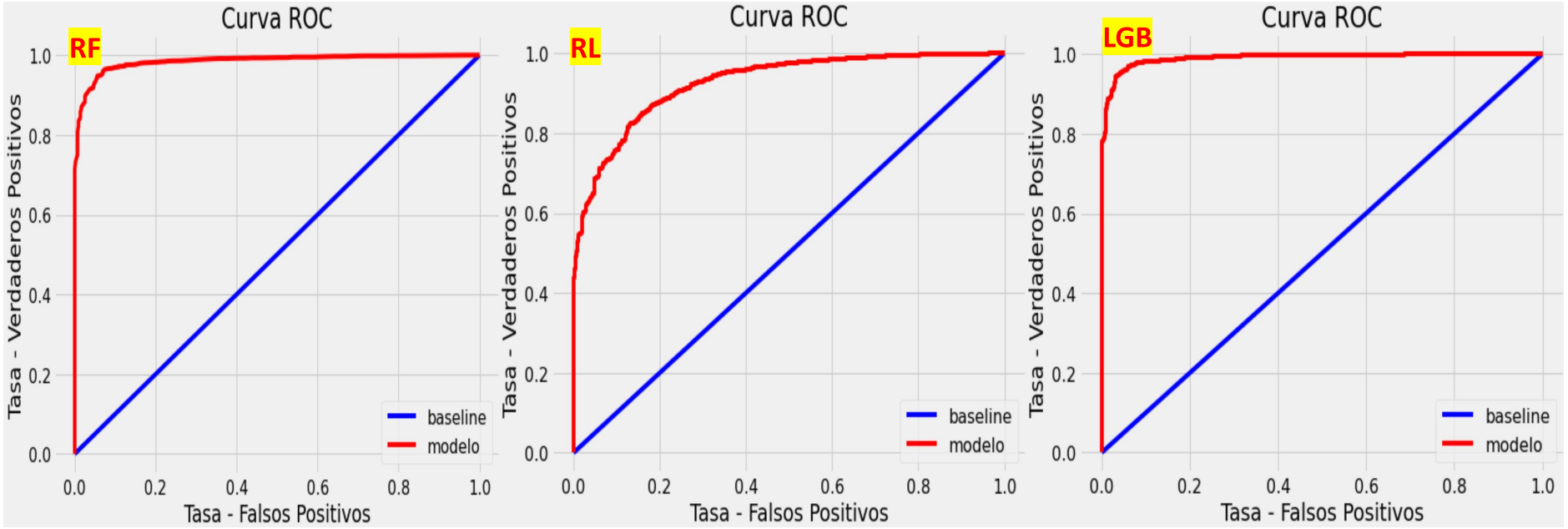
Matriz de confusión:

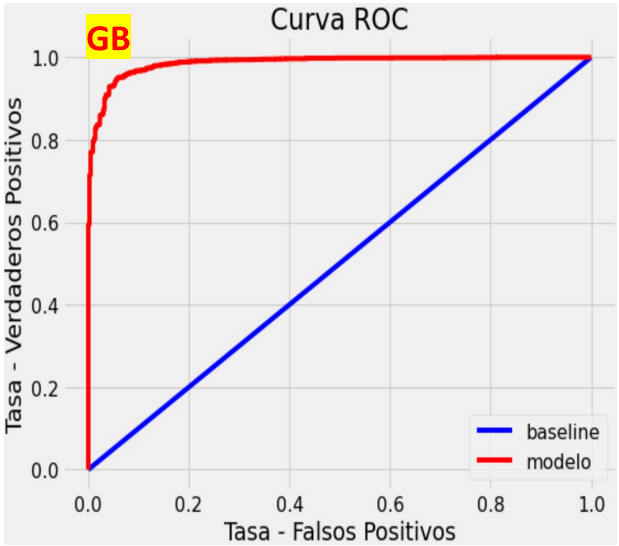
Herramienta muy útil para valorar cómo de bueno es un modelo clasificando basado **en** aprendizaje automático.

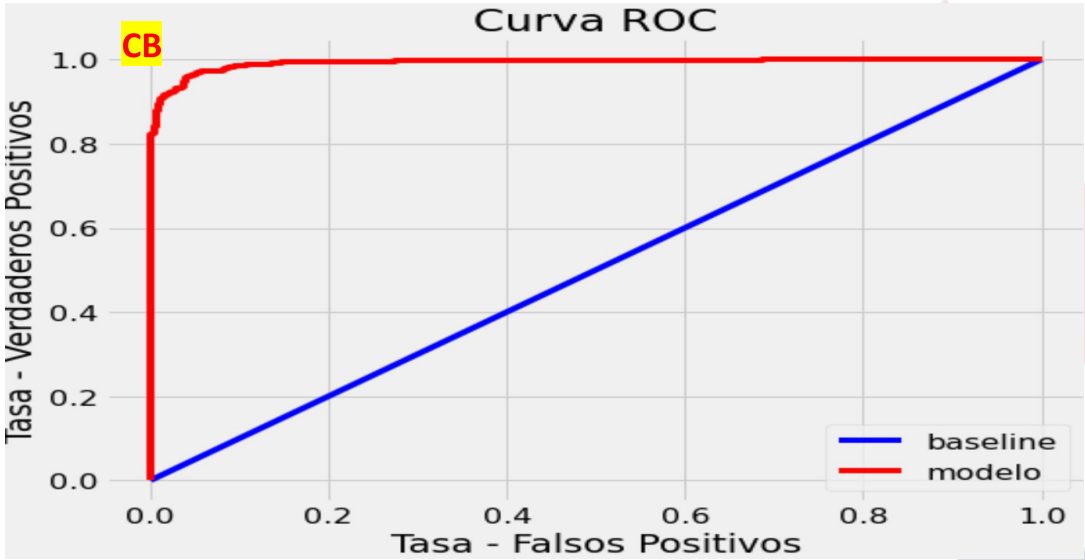
Curva ROC:

Es una representación gráfica de la sensibilidad o razón de verdaderos positivos frente a la especificidad o razón de verdaderos negativos para un sistema clasificador binario.

Comparación de la curva ROC







Comparación de métricas

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

Comparación de métricas

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

PRUEBA DE HIPÓTESIS

Prueba de Hipotesis

```
from statsmodels.stats.proportion import proportions_ztest
# Significancia
Significancia = 0.05
# Entradas
pred_exitos = 2713
total = 2796
# Ho = 0.5
Hipotesis_nula = 0.5
# Ho: p <=0.5
# H1: p > 0.5
stat, p_value = proportions_ztest(count=pred_exitos, nobs=total, value=Hipotesis_nula, alternative='larger')
# reporte
print('z_stat: %0.3f, p_value: %0.3f' % (stat, p_value))
if p_value > Significancia:
   print ("No se rechaza la hipotesis Nula ")
else:
   print ("Se rechaza la hipotesis nula - Se puede afirmar que la hipotesis alterna es verdadera")
```

```
z_stat: 146.531, p_value: 0.000
Se rechaza la hipotesis nula - Se puede afirmar que la hipotesis alterna es verdadera
```

Comentario: Se puedo observar que en efecto lo que buscamos demostrar que un modelo nos ayudará en más de 50% es correcto. Y existen pruebas estadísticas para afirmarlo.

Conclusiones y Recomendaciones

Conclusiones

Luego de obtener los datos, analizar, transformar, modelar y comparar 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).

Data Science

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.

Data Science

www.datascience.pe