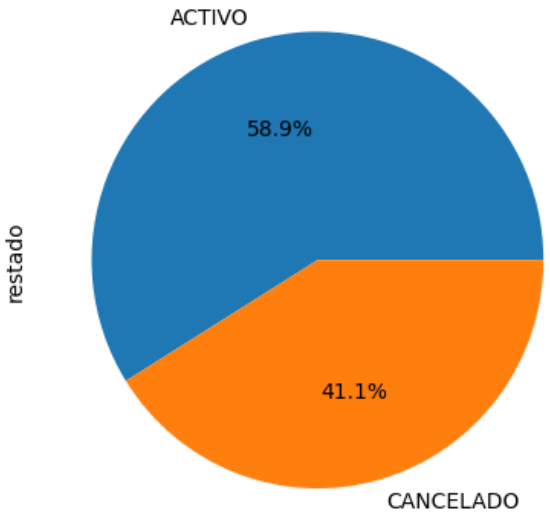
Introducción

Para estimar un modelo predictivo que permita determinar la probabilidad de que un cliente cancele su servicio de telefonía se utilizó una base de datos compuesta por una muestra de 9702 clientes de una compañía telefónica seleccionados aleatoriamente. La misma se encuentra compuesta de 14 variables, las cuales se encuentran divididas en **categóricas** y **numéricas**.

Variables numéricas	Código	Variables categóricas	Código
Promedio consumo vida activa	prom_act	Prom. Cons 2/4 meses 1. Prom consumo entre 0,02 y 1.2487 2. Prom consumo de más de 1.2487 hasta 500 3. Prom consumo mayor a 500	tprom4_2
Antigüedad del cliente en días	d_edad	Deuda Total 1. Sin deuda 2. Deuda total hasta 157,25 3. Mayor de 157,25	tdeuda
Tiempo que resta para el fin de contrato	d_fincon	Región 1. Bs As, Sta Fe y La Pampa 2. Cuyo, Litoral, Mediterraneo y NOA 3. Patagonia	region_3
Historia de pago	hist_pag	Forma de adquisición 1. Comodato 2. Stock distribuido y Leasing 3. Propio	adquisi1
Porcentaje de desalocación	porc_des	Tipo de cuenta 1. Negocios 2. Otros 3. Personal 4. Top	rtip_cta
Duración del contrato	d_durac	Modelo 1. Motorola 2. Nokia 3. Otros	mmodelo
		Estado del Agente 1. Activo 2. Baja 3. Trámite baja y Otros	rest_ag1

Los clientes fueron segmentados en dos grupos:

- Clientes que se encuentran activos en la empresa con un total de 5712 casos.
- Clientes que han cancelado su servicio telefónico con un total de 3990 casos.



La proporción de clientes que se encuentran activos es del 58.9%, por lo cual existe una probabilidad a priori del 41,1 % de que un cliente cancele su servicio en la empresa.

Modelo de regresión logística - Explicación

Se aplicó un modelo de Regresión Logística para explicar la influencia de las variables regresoras en relación a la variable respuesta de cliente en estado **Cancelado.**

Logit Regression Results

Dep. Variable:	restado	No. Observations:	9702
Model:	Logit	Df Residuals:	9680
Method:	MLE	Df Model:	21
Date:	Fri, 03 Nov 2023	Pseudo R-squ.:	0.5366
Time:	16:31:19	Log-Likelihood:	-3045.2
converged:	True	LL-Null:	-6571.3
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.8316	0.385	7.350	0.000	2.076	3.587
tprom4_2[T.Mayor a 500]	1.8261	0.087	21.004	0.000	1.656	1.997
tprom4_2[T.Más de 1,2487 a 500]	0.4519	0.079	5.688	0.000	0.296	0.608
mmodelo[T.Nokia]	0.3880	0.079	4.901	0.000	0.233	0.543
mmodelo[T.Otros]	-0.0385	0.094	-0.409	0.682	-0.223	0.146
tdeuda[T.Sin deuda]	-1.0577	0.113	-9.322	0.000	-1.280	-0.835
tdeuda[T.hasta 157,25]	-1.7911	0.117	-15.325	0.000	-2.020	-1.562
region_3[T.Patagonia]	0.5100	0.135	3.788	0.000	0.246	0.774
region_3[T.Resto]	0.2324	0.079	2.936	0.003	0.077	0.388
adquisi1[T.Dist.y Leasing]	-0.2134	0.097	-2.193	0.028	-0.404	-0.023
adquisi1[T.Propio]	0.4687	0.088	5.315	0.000	0.296	0.642
rtip_cta[T.Otros]	-0.6957	0.340	-2.044	0.041	-1.363	-0.029
rtip_cta[T.Personal]	-0.2529	0.071	-3.556	0.000	-0.392	-0.114
rtip_cta[T.Top]	-1.6299	0.200	-8.167	0.000	-2.021	-1.239
rest_ag1[T.Baja]	0.7560	0.079	9.585	0.000	0.601	0.911
rest_ag1[T.Tramita baja y OE]	0.4322	0.102	4.242	0.000	0.232	0.632
prom_act	0.0008	0.000	3.738	0.000	0.000	0.001
d_edad	-0.0064	0.000	-20.052	0.000	-0.007	-0.006
d_fincon	-0.0012	0.000	-3.888	0.000	-0.002	-0.001
porc_des	0.3045	0.057	5.339	0.000	0.193	0.416
hist_p	0.6318	0.026	23.866	0.000	0.580	0.684
d_durac	-0.7600	0.067	-11.397	0.000	-0.891	-0.629

Se obtuvo un Pseudo R^2 de 0.5366, lo cual indica que el modelo explica el 53,6% de la variabilidad de la variable dependiente, también se obtuvo un valor de log verosimilitud de -3045, lo cual indica que este modelo explica de mejor manera cómo se relacionan las variables independientes con la variable dependiente en función de los datos ingresado, en comparación con los anteriores modelos que se obtuvieron resultados de -4146 (modelo con variables categóricas) y -3590 (modelo con variables numéricas).

En este modelado se han utilizado como categorías de referencia a tprom4_2[Prom consumo entre 0.02 y 1.2487], tdeuda[Mayor de 157,25], region_3[Bs As, Sta Fe y La Pampa], adquisi1[Comodato], rtip_cta[Negocios], mmodelo[Motorola], rest_ag1[Activo].

Se puede observar que la mayoría de las variables resultaron significativas en el modelo, a excepción de la variable mmodelos_Otros. Sin embargo, al tratarse solamente de

una categoría de la variable mmodelo se optó por mantenerla ya que no modifica significativamente los resultados obtenidos.

En cuanto a las variables numéricas, podemos observar en la columna de **coef** que funcionan como factores protectores de la variable dependiente, es decir que frente a un aumento unitario de cualquiera de ellas disminuyen las posibilidades de que el usuario cancele el servicio. Mientras que en las variables categóricas, todas funcionan como factores protectores en relación a la categoría de referencia a excepción de la categoría "Mayor a 500" de la variable "tprom4_2", esta última indica que frente al cambio de la categoría de referencia a "Mayor a 500" aumenta la posibilidad de que la persona cancele el servicio.

Para finalizar con la explicación que ofrece este modelo de Regresión logística se calcularon los Odds ratios de las variables independientes.

Intercept	16.971765
tprom4_2[T.Mayor a 500]	6.209912
tprom4_2[T.Más de 1,2487 a 500]	1.571297
mmodelo[T.Nokia]	1.474092
mmodelo[T.Otros]	0.962189
tdeuda[T.Sin deuda]	0.347252
tdeuda[T.hasta 157,25]	0.166780
region_3[T.Patagonia]	1.665268
region_3[T.Resto]	1.261659
adquisi1[T.Dist.y Leasing]	0.807852
adquisi1[T.Propio]	1.597906
rtip_cta[T.Otros]	0.498709
rtip_cta[T.Personal]	0.776573
rtip_cta[T.Top]	0.195950
rest_ag1[T.Baja]	2.129638
rest_ag1[T.Tramita baja y OE]	1.540625
prom_act	1.000756
d_edad	0.993598
d_fincon	0.998775
porc_des	1.355946
hist_p	1.881066
d_durac	0.467656

Al analizar los mismos nos encontramos con que si un cliente tiene un promedio de consumo "mayor a 500" tiene 6.2 más probabilidades de cancelar su servicio que un cliente que consuma "entre 0.02 y 12487", mientras que un cliente que su "Agente esté dado de baja**"** tiene una probabilidad de cancelar el servicio 2.12 más altas que quienes tienen su "Agente activo". Además nos encontramos con que los clientes que tienen un "consumo promedio entre 12487 a 500", un modelo de teléfono "Nokia", sean de la "Patagonia o el resto del país", tengan teléfono "Propio", o su Agente se encuentre "Tramitando la baja" tienen una probabilidad de cancelar el servicio entre 1.26 y 1.66 más altas que quienes se encuentran en sus categorías de referencia. Observamos también que frente al aumento unitario de porcentaje de desalocación e historial de pago aumentan las chances de que el cliente cancele el servicio en 1.35 y 1.88 veces respectivamente.

Por otro lado, nos encontramos con que los clientes que tienen Tipo de cuenta "Top, Personal u Otros", se encuentren "Sin deuda o con una deuda hasta 157,25", hayan adquirido su equipo por medio de "Stock distribuido y Leasing" y tengan "mayor duración de

contrato” en términos unitarios, tienen mayor probabilidad de mantener activo el servicio en la compañía que quienes se encuentran en sus categorías de referencia.

Modelo de Regresión Logística - Predicción

Para la creación de un modelo predictivo que permita saber quiénes son los clientes que tienen mayor probabilidad de cancelar su servicio en la compañía se utilizó la Regresión Logística.

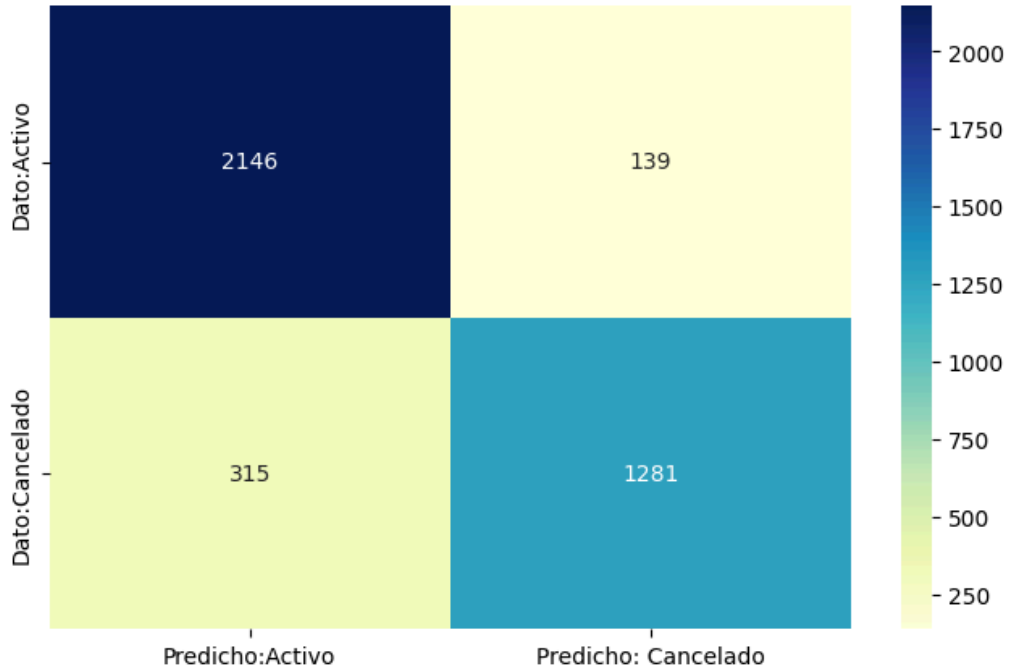
Se dividió la base de datos en una muestra de entrenamiento y una muestra de testeo (60%-40%).

	Muestra de entrenamiento	Muestra de testeo
Activos	3427	2285
Cancelados	2394	1596

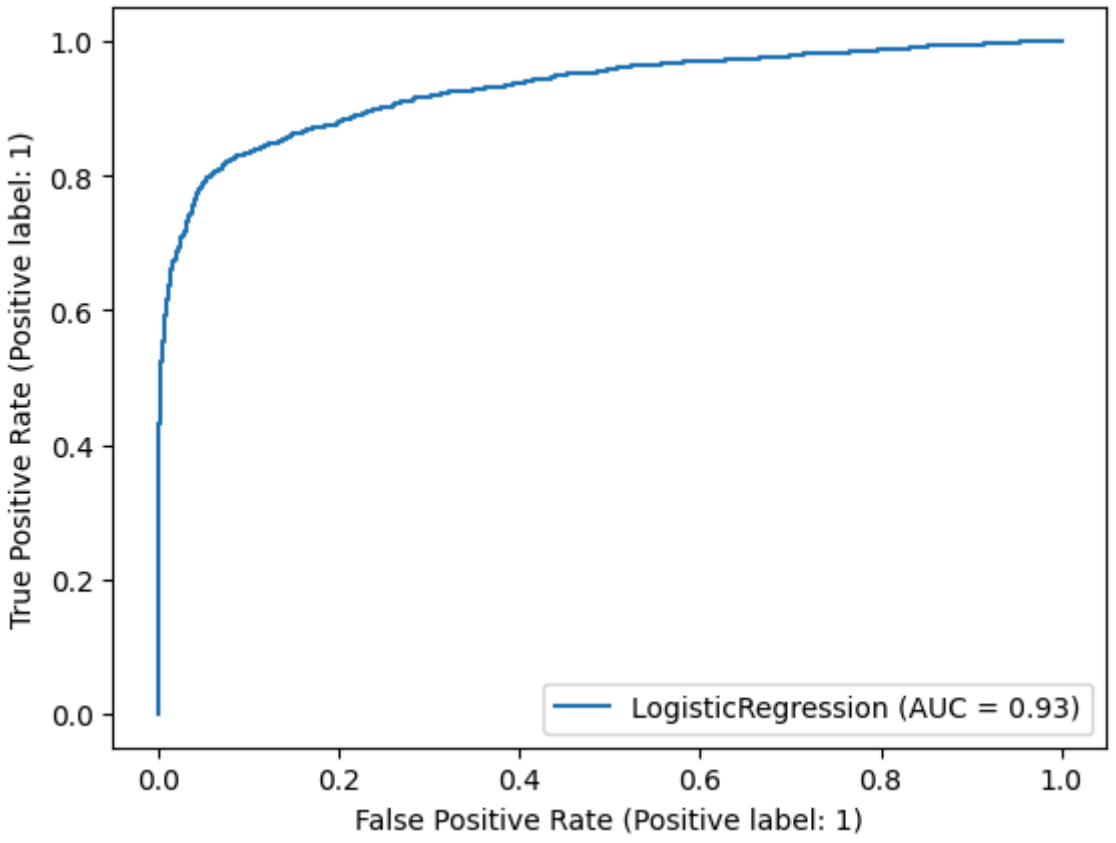
Con la primera muestra se ajustó el modelo de regresión y luego se comprobó su funcionamiento con la segunda muestra.

```
--------------------
Tabla de reporte completa
------------------
              precision    recall  f1-score   support

           0      0.872     0.939     0.904      2285
           1      0.902     0.803     0.849      1596

    accuracy                          0.883      3881
   macro avg      0.887     0.871     0.877      3881
weighted avg      0.884     0.883     0.882      3881

--------------------
Tabla de confusión
------------------
[[2146  139]
 [ 315 1281]]
--------------------
Reporte de medidas de desempeño
------------------
accuracy:  0.8830198402473589
```

Como podemos observar en el reporte, el modelo presenta una sensibilidad del 80% y una especificidad del 93%, por lo cual podemos decir que el modelo es efectivo para identificar la mayoría de los clientes que van a cancelar su servicio como también quienes se van a mantener activos. Con estos índices podemos obtener el grado de exactitud (accuracy) que presenta el modelo, siendo en este caso del 88%.



En la matriz de confusión podemos observar cómo se han predicho los clientes de la muestra de testeo, verificando que ha clasificado de manera correcta 1281 clientes que cancelaron su servicio de telefonía y 2146 clientes que continúan activos como clientes de la compañía. De esta manera corroboramos los índices de especificidad y sensibilidad que se presentaron anteriormente en la tabla del reporte.



La curva ROC presenta un valor de AUC de 0.93 lo cual indica que es un modelo con un rendimiento muy bueno y puede discriminar eficazmente entre los clientes que van a cancelar el servicio de telefonía y aquellos que se van a mantener activos

Para finalizar, tal como vimos en las pruebas anteriores, podemos concluir que el modelo cumple con el objetivo de predecir clientes que potencialmente van a cancelar su servicio, permitiendo que la compañía pueda generar intervenciones que prevengan la cancelación del servicio y por ende la pérdida de clientes.