Modelo Predictivo de Abandono de Clientes en el Sector de Telecomunicaciones usando Machine Learning en R

Trabajo Personal

Autor:

Juan Gabriel Carvajal Negrete

Información basica de la base de datos

Contexto de la base de datos

Predecir el comportamiento para fidelizar a los clientes. Puede analizar todos los datos relevantes de los clientes y desarrollar programas de fidelización específicos. [Conjuntos de datos de muestra de IBM].Cada fila representa un cliente, cada columna contiene los atributos del cliente descritos en la columna Metadatos.

La base de datos cuenta con **7043, 21** filas y columnas respectivamente,los nombres de las variables de la base de datos son **customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn**, el conjunto de datos incluye información sobre:

- Clientes que se fueron en el último mes: la columna se llama **Churn** (Variable objetivo)
- Servicios a los que se ha suscrito cada cliente: teléfono, líneas múltiples, Internet, seguridad en línea, copia de seguridad en línea, protección de dispositivos, soporte técnico y transmisión de TV y películas.
- Información de la cuenta del cliente: cuánto tiempo ha sido cliente, contrato, método de pago, facturación electrónica, cargos mensuales y cargos totales
- Información demográfica sobre los clientes: género, rango de edad y si tienen parejas y dependientes.

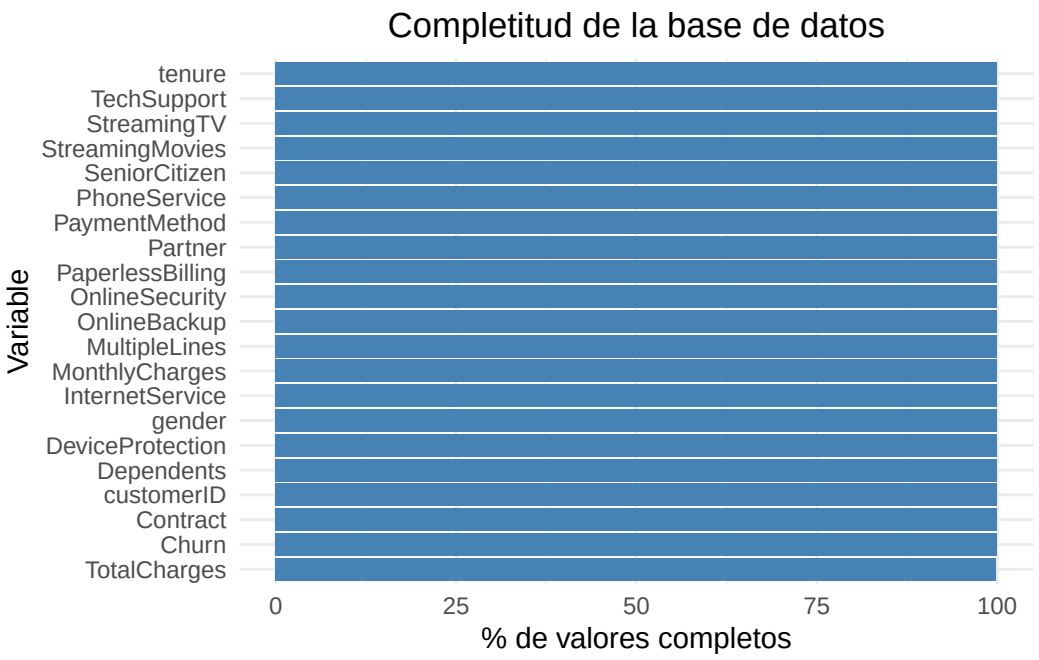
Estas son algunas observaciones básicas de la base de datos ahora continuaremos e iremos mas afondo con nuestro datos, mostrando que tan completas están nuestras variables y la importancia que tendrá cada una de ellas en nuestros futuros modelos.

La base de datos la puedes encontrar dando click aquí

Descripción General y Estadísticas Iniciales de la Base de Datos

Table 1: Porcentaje de completitud de las variables

N°	Variable	Porcentaje Completo (%)
1	customerID	100.00
2	gender	100.00
3	SeniorCitizen	100.00
4	Partner	100.00
5	Dependents	100.00
6	tenure	100.00
7	PhoneService	100.00
8	MultipleLines	100.00
9	InternetService	100.00
10	OnlineSecurity	100.00
11	OnlineBackup	100.00
12	DeviceProtection	100.00
13	TechSupport	100.00
14	StreamingTV	100.00
15	StreamingMovies	100.00
16	Contract	100.00
17	PaperlessBilling	100.00
18	PaymentMethod	100.00
19	MonthlyCharges	100.00
20	TotalCharges	99.84
21	Churn	100.00



De acuerdo con la tabla y el gráfico presentados, la base de datos refleja un alto nivel de completitud, con un porcentaje mínimo de valores faltantes. Esta condición es favorable, ya que garantiza una mayor confiabilidad en los resultados posteriores y disminuye la necesidad de aplicar técnicas de imputación o eliminación de observaciones.

Con esta base sólida, es posible avanzar hacia el análisis exploratorio de cada una de las variables, lo cual permitirá detectar posibles anomalías, patrones inusuales o inconsistencias en los registros. Asimismo, resulta fundamental revisar que el tipo de dato asignado a cada variable corresponda efectivamente con su naturaleza, asegurando que la base se encuentre correctamente estructurada para el modelado y el análisis estadístico posterior.

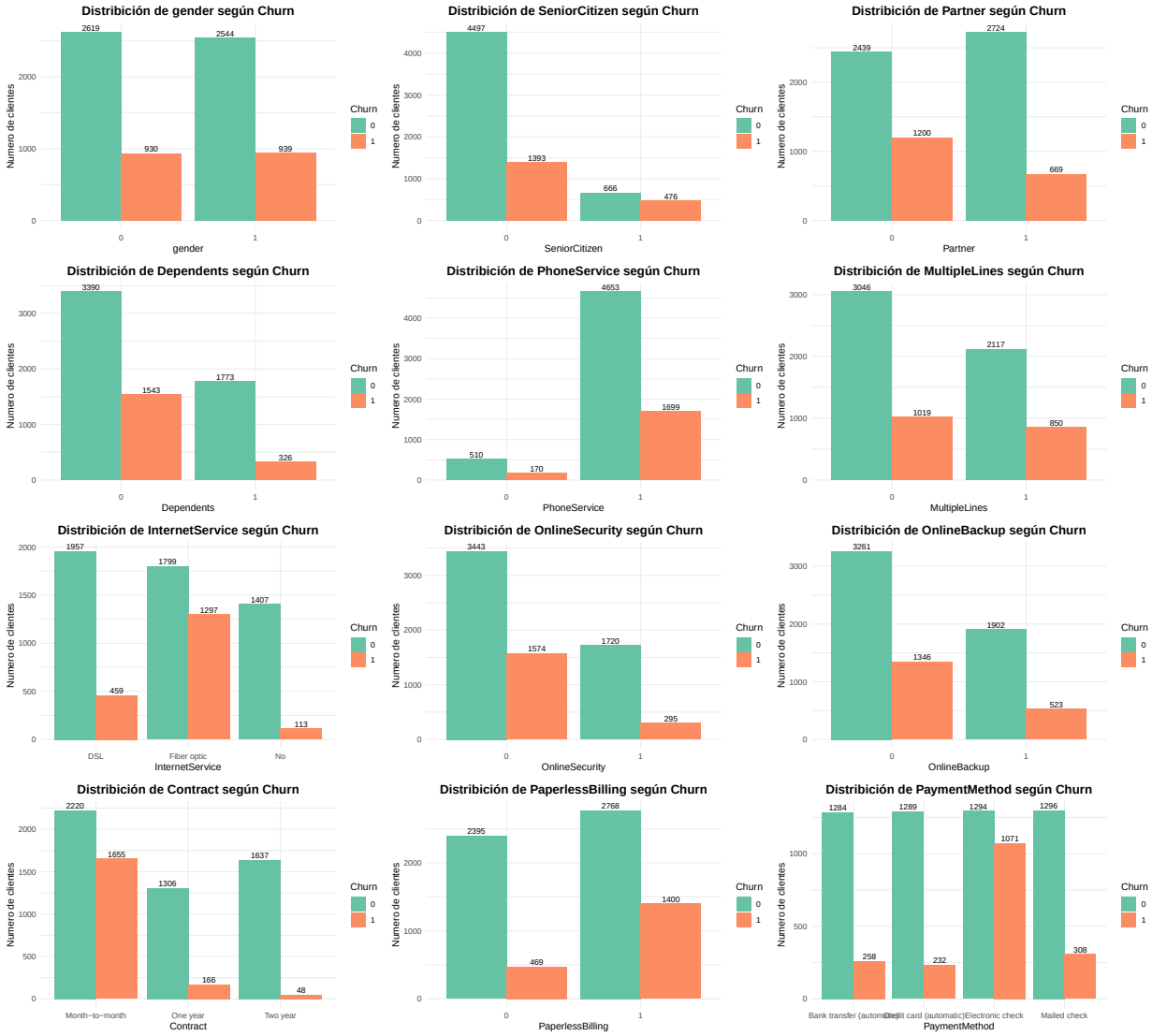
Análisis individual de las variables

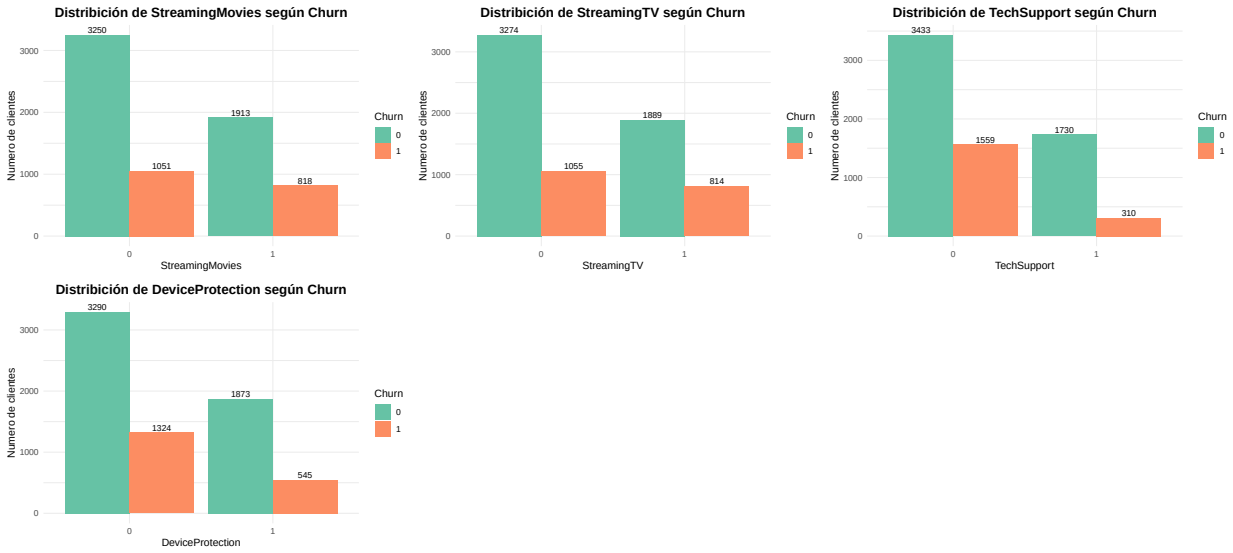
Table 2: Tipos de variables en el dataset

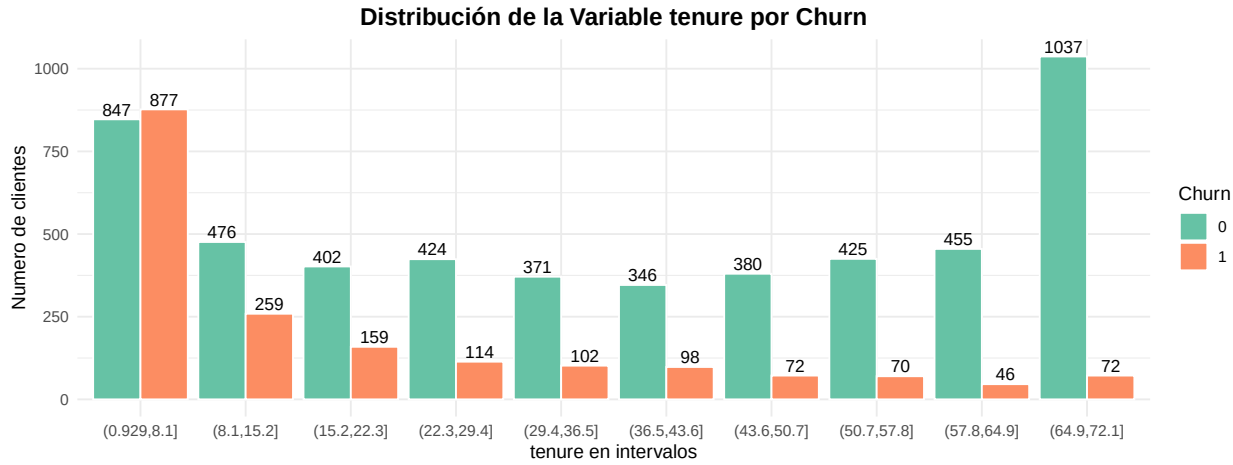
	Variables	**Tipo**
1	customerID	factor
2	gender	factor
3	SeniorCitizen	factor
4	Partner	factor
5	Dependents	factor
6	tenure	integer
7	PhoneService	factor
8	MultipleLines	factor
9	InternetService	factor
10	OnlineSecurity	factor
11	OnlineBackup	factor
12	DeviceProtection	factor
13	TechSupport	factor
14	StreamingTV	factor
15	StreamingMovies	factor
16	Contract	factor
17	PaperlessBilling	factor
18	PaymentMethod	factor
19	MonthlyCharges	numeric
20	TotalCharges	numeric
21	Churn	factor

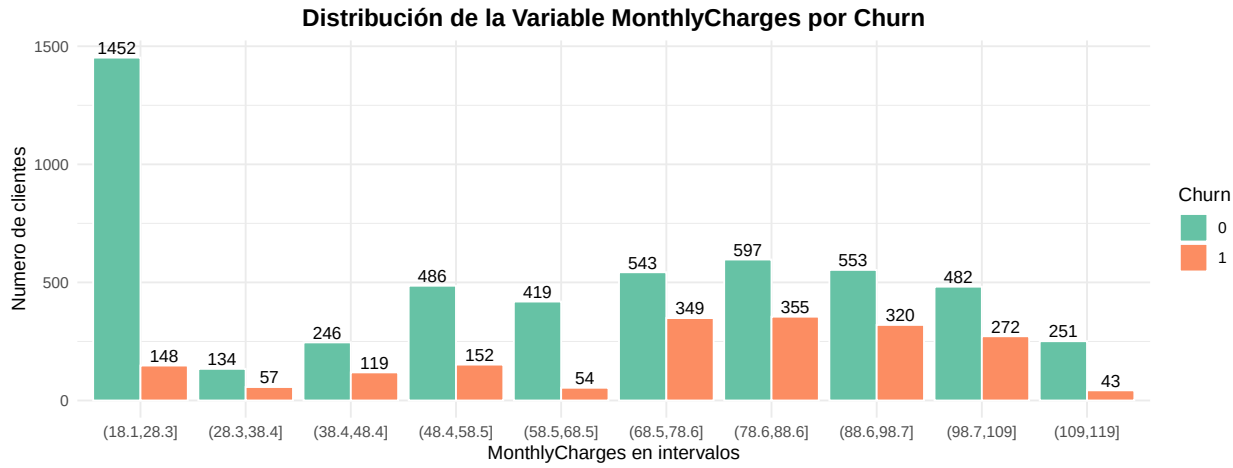
Luego de realizar un análisis detallado de cada variable pudimos identificar que la mayoría de ellas el programa las reconocía como variables **character** cuando en realidad debían ser consideradas como tipo **factor**, ademas también pudimos identificar errores en la digitación de la información.

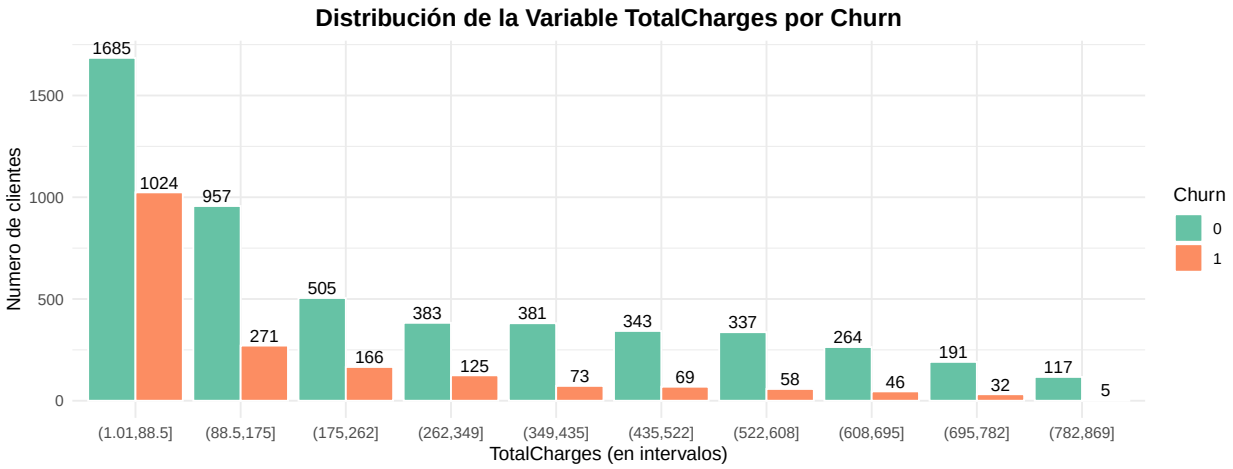
Gráficas descriptivas individuales

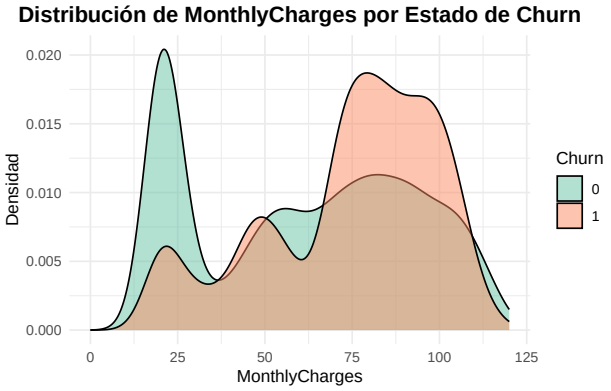


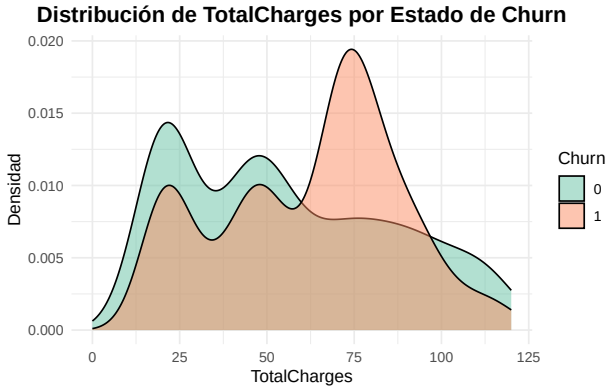

Distribición de StreamingMovies según Churn
Numero de clientes
3250
1051
1913
818
StreamingMovies
Churn
0
1
Distribición de StreamingTV según Churn
3274
1055
1889
814
StreamingTV
Distribición de TechSupport según Churn
3433
1559
1730
310
TechSupport
Distribición de DeviceProtection según Churn
3290
1324
1873
545
DeviceProtection


Distribución de la Variable tenure por Churn
Numero de clientes
847
877
476
259
402
159
424
114
371
102
346
98
380
72
425
70
455
46
1037
72
(0.929,8.1]
(8.1,15.2]
(15.2,22.3]
(22.3,29.4]
(29.4,36.5]
(36.5,43.6]
(43.6,50.7]
(50.7,57.8]
(57.8,64.9]
(64.9,72.1]
tenure en intervalos
Churn
0
1


Distribución de la Variable MonthlyCharges por Churn
Numero de clientes
1452
148
134
57
246
119
486
152
419
54
543
349
597
355
553
320
482
272
251
43
(18.1,28.3]
(28.3,38.4]
(38.4,48.4]
(48.4,58.5]
(58.5,68.5]
(68.5,78.6]
(78.6,88.6]
(88.6,98.7]
(98.7,109]
(109,119]
MonthlyCharges en intervalos
Churn
0
1







Las gráficas anteriores muestran la cantidad de clientes que han abandonado el servicio y las personas que han permanecido con el en las diferentes variables de tipo factor que se tienen en la base de datos, ademas también presentamos gráficas con variables numéricas tomadas en intervalos para seguir observando el comportamiento de los cientes, una observación general es que son mas los clientes que deciden continuar con el servicio que los que deciden dejarlo. por ultimo revisemos un gráfico de densidad para las variables mas numéricas de nuestro dataset.

Modelación (Datos supervisados)

Modelo de regresión logístico.

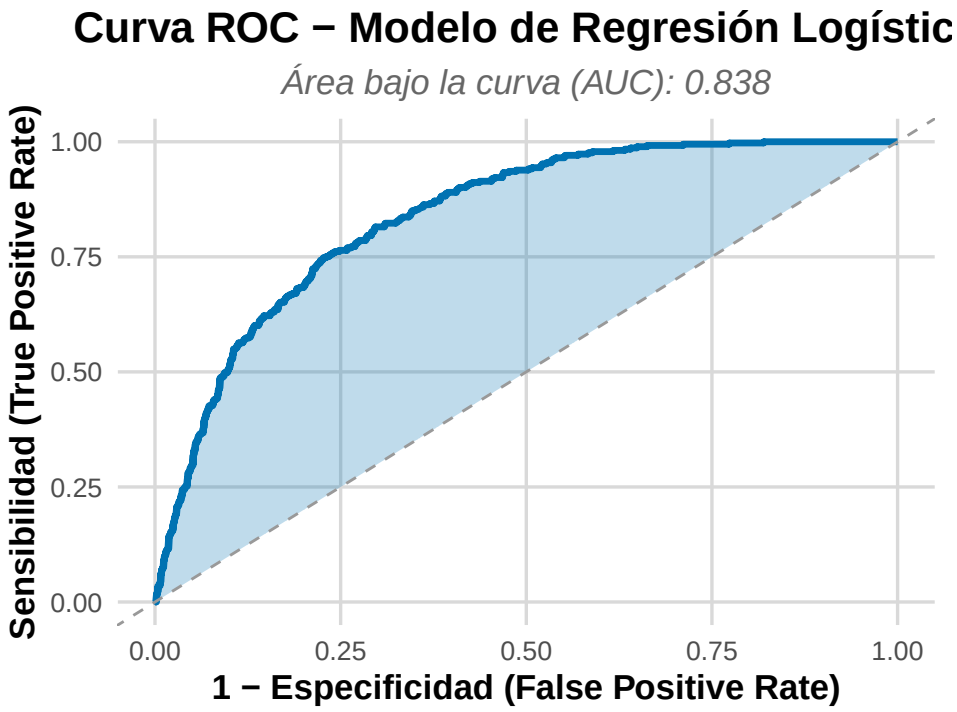
El primer modelo que se va a implementar en este proyecto es el modelo de regresión logística aplicando el método de selección stepwise (paso a paso) y esto fueron los resultados que se obtuvieron.

Table 3: **Resumen de coeficientes del modelo logístico reducido**

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.12	0.40	5.34	0.00
SeniorCitizen1	0.26	0.09	2.80	0.01
tenure	-0.06	0.01	-8.78	0.00
PhoneService1	1.01	0.30	3.43	0.00
MultipleLines1	0.62	0.11	5.85	0.00
InternetServiceFiber optic	2.79	0.31	8.90	0.00
InternetServiceNo	-2.81	0.38	-7.34	0.00
OnlineBackup1	0.28	0.11	2.64	0.01
DeviceProtection1	0.40	0.11	3.64	0.00
StreamingTV1	1.04	0.16	6.69	0.00
StreamingMovies1	0.92	0.15	5.94	0.00
ContractOne year	-0.74	0.12	-6.14	0.00
ContractTwo year	-1.33	0.20	-6.74	0.00
PaperlessBilling1	0.36	0.08	4.33	0.00
PaymentMethodCredit card (automatic)	-0.07	0.13	-0.52	0.60
PaymentMethodElectronic check	0.28	0.11	2.59	0.01
PaymentMethodMailed check	-0.08	0.13	-0.60	0.55
MonthlyCharges	-0.08	0.01	-6.39	0.00
TotalCharges	0.00	0.00	3.98	0.00

```
          Reference
Prediction   0   1
         0 919 167
         1 113 206
```

```
La precision con la que el modelo acierta es de  0.8007117
```



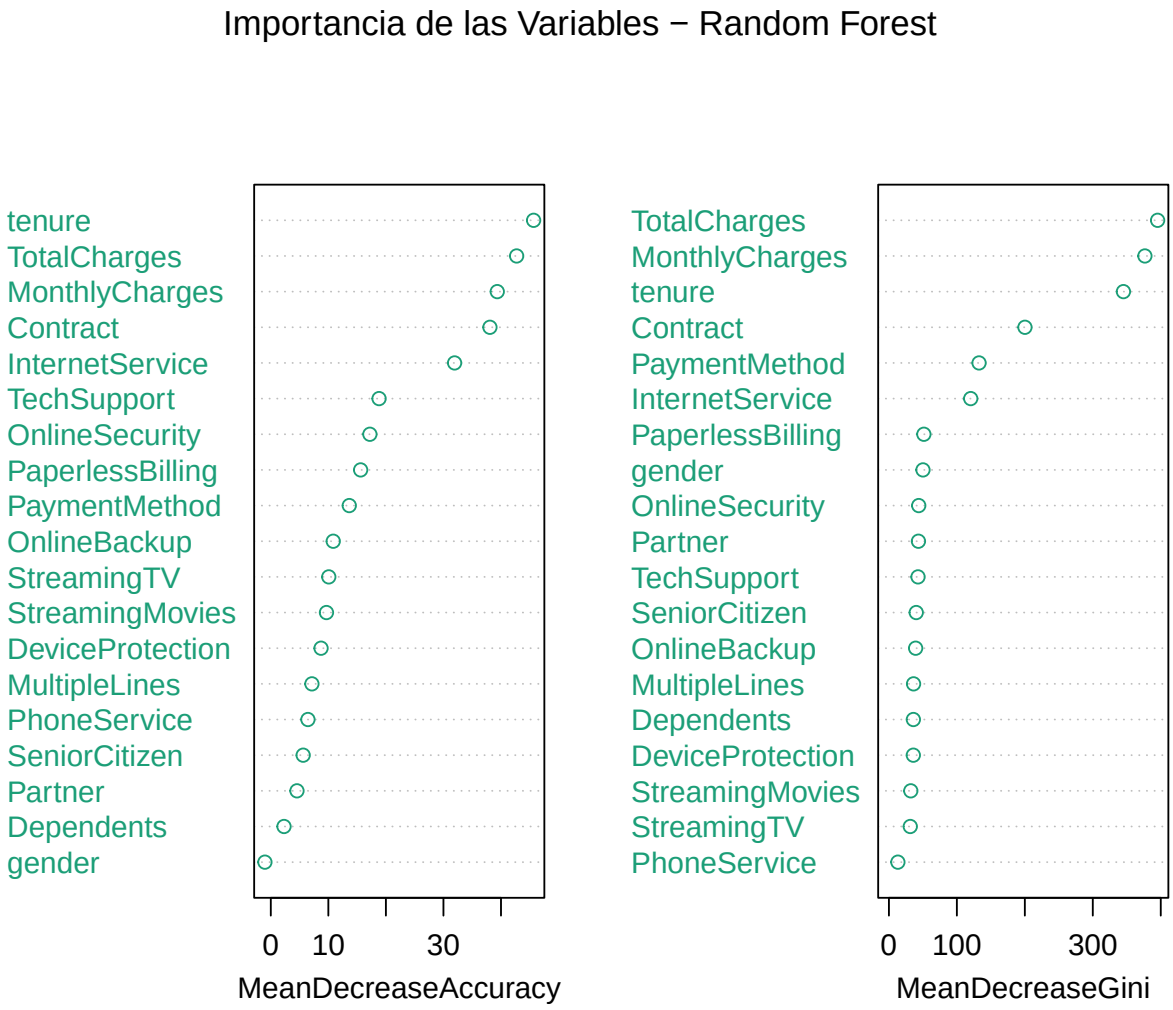
Modelo de Random Forest.

```
          Reference
Prediction   0   1
         0 933 193
         1  99 180
```

```
La precision con la que el modelo acierta es de 0.7921708
```

Hasta este punto ambos modelos dan un precisión casi igual, no se evidencia una mejora en el modelo de **Random Forest** pero, haremos un análisis de las variables e intentaremos mejorar mas las predicciones de nuestro modelo.

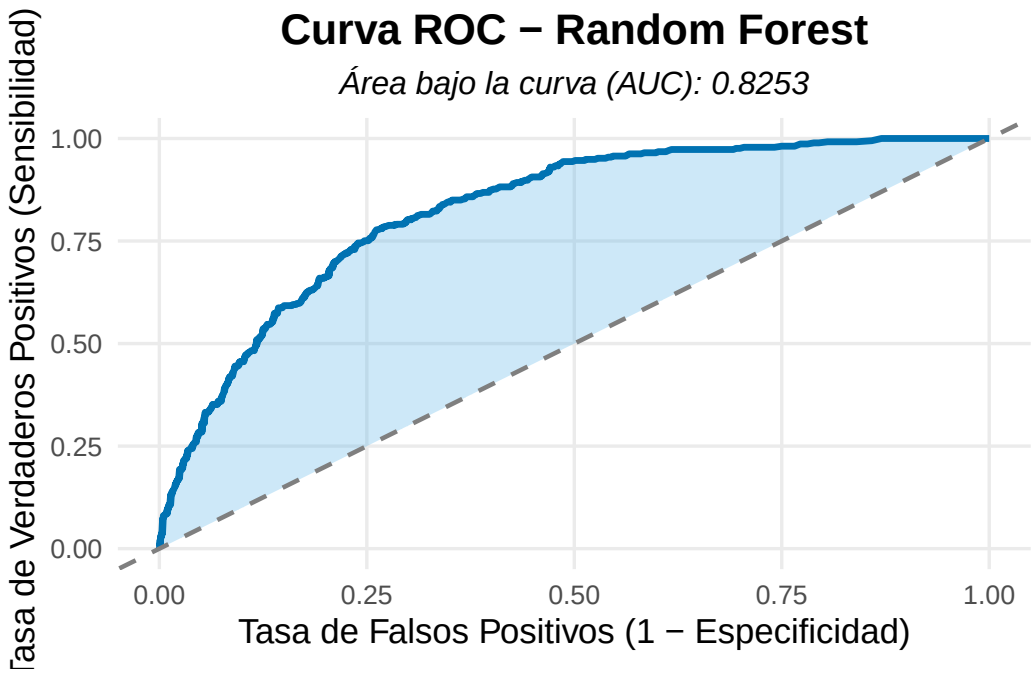
Visualización de importancia relativa de cada variable del modelo



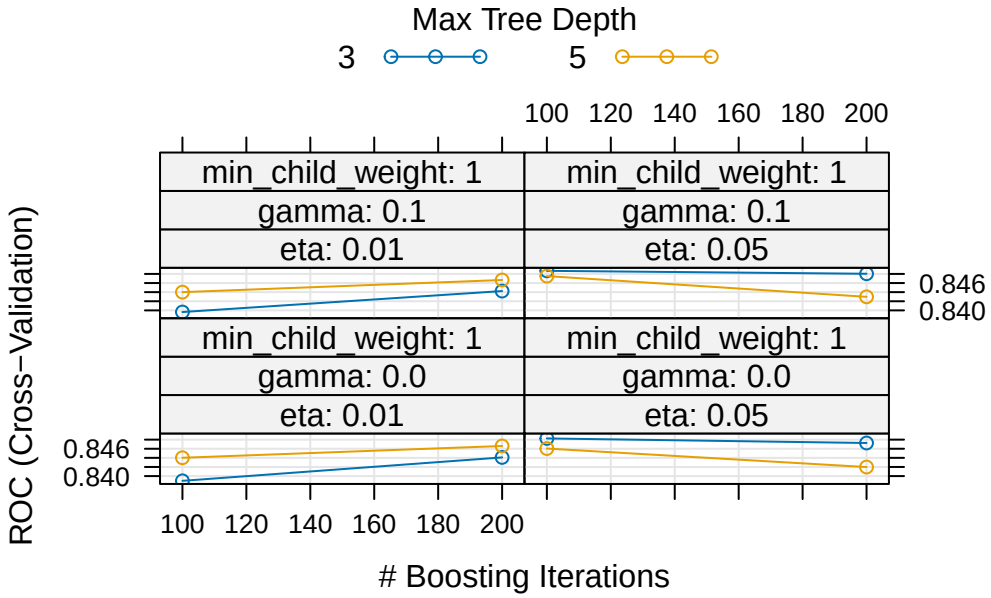
Entonces las variables que vamos a considerar dejar en el modelo son TotalCharges, MonthlyCharges,tenure,Contract, InternetService, PaymentMethod,PaperlessBilling TechSupport, OnlineSecurity, PaperlessBilling, PaymentMethod, OnlineBackup, StreamingTV.

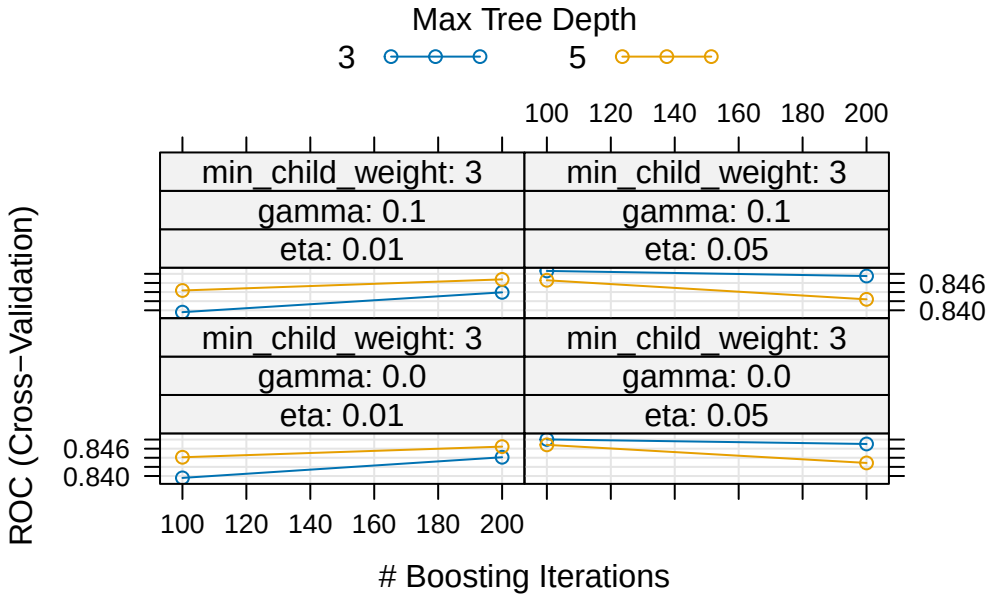
```
          Reference
Prediction   0   1
         0 923 196
         1 109 177
```

```
La precision con la que el modelo acierta es   de 0.7829181
```



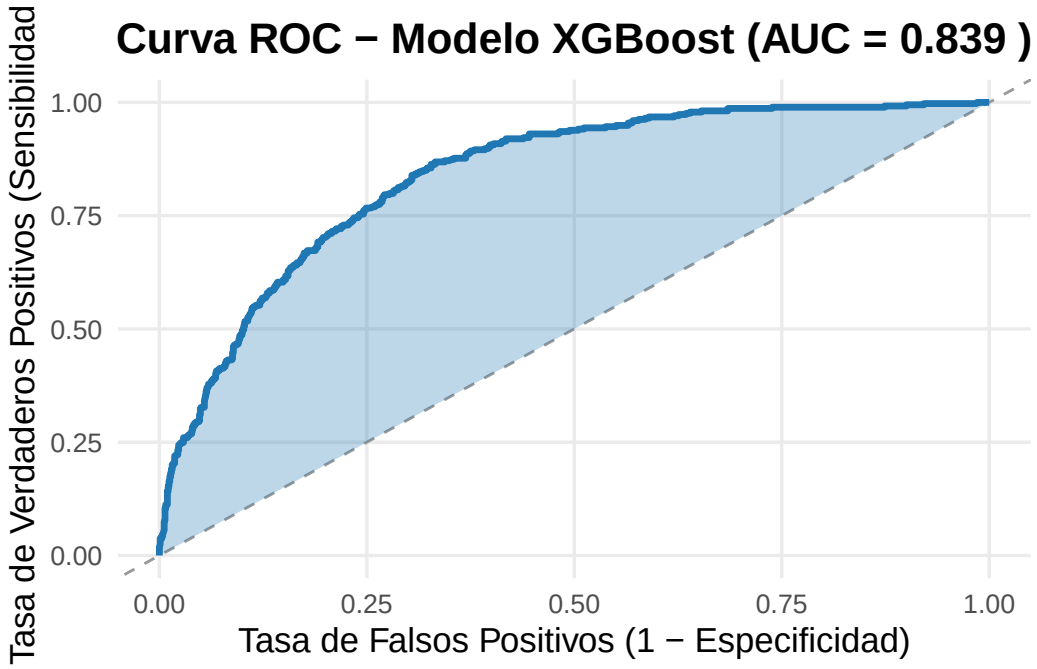
Modelo XGBOOST






```
          Reference
Prediction  No Yes
       No  926 184
       Yes 106 189
```

```
La precision con la que el modelo acierta es   de 0.7935943
```



Conclusion

Tras la comparación de los tres modelos de clasificación —Regresión Logística, Random Forest y XGBoost— se observó que todos presentan un desempeño similar en términos de métricas de evaluación (AUC, precisión, recall y matriz de confusión).

Si bien el modelo XGBoost obtuvo el mejor rendimiento, superando ligeramente a los otros dos, la diferencia fue mínima, lo que indica que los tres algoritmos logran capturar de forma consistente los patrones de abandono (churn) presentes en los datos.

Esto sugiere que el conjunto de variables utilizadas tiene un poder predictivo estable, y que la mejora de la precisión podría depender más del ajuste fino de hiperparámetros o de la incorporación de nuevas variables (por ejemplo, variables de comportamiento o uso del servicio) que del tipo de modelo en sí.