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.

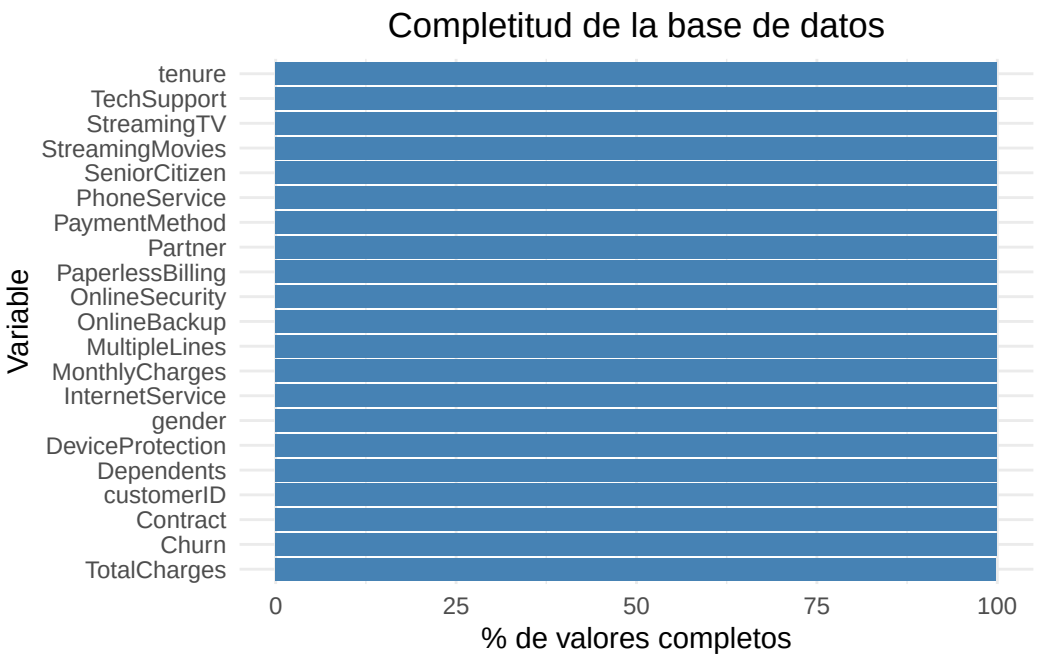
Estas son algunas observaciones básicas de la base de datos ahora continuaremos e iremos mas afondo con nuestro datos, mostrando que tan completas están nuestras variables y la importancia que tendrá cada una de ellas en nuestros futuros modelos.

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

	variables	tipo
gender	gender	factor
SeniorCitizen	SeniorCitizen	factor
Partner	Partner	factor
Dependents	Dependents	factor
tenure	tenure	integer
PhoneService	PhoneService	factor
MultipleLines	MultipleLines	factor
InternetService	InternetService	factor
OnlineSecurity	OnlineSecurity	factor
OnlineBackup	OnlineBackup	factor
DeviceProtection	DeviceProtection	factor
TechSupport	TechSupport	factor
StreamingTV	StreamingTV	factor
StreamingMovies	StreamingMovies	factor
Contract	Contract	factor
PaperlessBilling	PaperlessBilling	factor
PaymentMethod	PaymentMethod	factor
MonthlyCharges	MonthlyCharges	numeric
TotalCharges	TotalCharges	numeric
Churn	Churn	factor