

Predicción de Cancelación de Clientes en el Sector de Telecomunicaciones

1. Resumen de la estrategia de solución

El objetivo del proyecto fue desarrollar un modelo de clasificación que permita predecir si un cliente de una empresa de telecomunicaciones cancelará el servicio. Para lograrlo, se utilizó el conjunto de datos **Telco Customer Churn**, el cual contiene información sobre los clientes, sus contratos, servicios, entre otros.

La estrategia de solución planteada consistió en las siguientes etapas:

- **Exploración de datos:** análisis general de las variables y detección de valores faltantes y tipos de datos.
- **Preprocesamiento:** imputación de valores faltantes, codificación de variables categóricas, escalado de variables numéricas y preparación del conjunto para el modelado.
- **Entrenamiento de modelos:** se entrenaron dos modelos supervisados los cuales fueron random forest y decision tree, y se utilizó GridSearchCV para encontrar los mejores hiperparámetros.
- **Evaluación y comparación:** los modelos fueron evaluados con métricas como *accuracy*, *precision*, *recall* y *F1-score*. Se construyeron curvas de aprendizaje y matrices de confusión, y se realizó un análisis de concordancia entre las predicciones.

2. Resultados obtenidos

Tras aplicar los modelos dos modelos de Random Forest y Árbol de Decisión, se observa que ambos alcanzaron un *accuracy* general del 79%. Sin embargo, el comportamiento interno de cada modelo fue diferente. El Random Forest, con los hiperparámetros seleccionados mostró una alta precisión en el conjunto de entrenamiento, lo que demuestra overfitting. Aun así, logró un *f1-score* de 0.52 para la clase “Churn”, superando al Árbol de Decisión en la detección de cancelaciones, aunque con una tendencia a equivocarse al predecir quiénes realmente se irían.

Por otro lado, el Árbol de Decisión, al tener una estructura más simple, presentó una curva de aprendizaje más equilibrada y un proceso más óptimo, aunque su rendimiento en la clase crítica “Churn” fue ligeramente inferior, con un *f1-score* de 0.49. Su matriz de confusión mostró una mayor cantidad de errores al clasificar clientes que efectivamente se fueron, detectando solo 146 de los 373 casos reales. En resumen, aunque ambos modelos ofrecen un rendimiento similar en términos

generales, el Random Forest demostró una mejor capacidad para detectar cancelaciones, lo que lo convierte en la opción más sólida en este escenario.

Reporte de random forest:

Reporte de clasificación:				
	precision	recall	f1-score	support
0	0.82	0.92	0.87	1036
1	0.67	0.43	0.52	373
accuracy			0.79	1409
macro avg	0.74	0.68	0.69	1409
weighted avg	0.78	0.79	0.78	1409

Reporte de decisión tree:

Reporte de clasificación:				
	precision	recall	f1-score	support
0	0.81	0.93	0.86	1036
1	0.66	0.39	0.49	373
accuracy			0.79	1409
macro avg	0.74	0.66	0.68	1409
weighted avg	0.77	0.79	0.77	1409

3. Conclusiones

- Ambos modelos lograron un accuracy del 79%, sin embargo, difieren en su comportamiento con respecto a la clase Churn, que es la clase más crítica para el negocio.
- Random forest mostró mejor balance general y un mayor F1-score para la clase Churn (0.52), lo que sugiere un mejor rendimiento para detectar cancelaciones reales.
- Árbol de decisión, aunque más simple e interpretativo, tuvo una menor capacidad de detección de Churn ($F1 = 0.49$) y mostró más errores al clasificar clientes que se iban.
- La curva de aprendizaje de RF sugiere overfitting, mientras que el árbol presenta una generalización más controlada.
- En términos de estrategia empresarial, Random forest sería la opción recomendada, aunque se recomienda ajustarlo o combinarlo con otras técnicas por ejemplo, balanceo de clases, para mejorar su recall en la clase positiva.