PRUEBA TÉCNICA - ANALISTA DE DATOS

FABIO ANDRÉS GONZÁLEZ VILLOTA
Ingeniero Industrial y científico de datos

C-Pocket

Resumen

En el presente informe se desarrolla un caso práctico enfocado en analítica de datos, específicamente en la construcción de un modelo analítico predictivo orientado a reducir la pérdida de clientes en una empresa del sector de telecomunicaciones. Para el desarrollo de este proyecto se empleó programación en Python, aplicando técnicas de ciencia de datos. En el siguiente enlace se encuentra los archivos utilizados para el desarrollo planteado en este documento:
https://colab.research.google.com/drive/11w3ofWCundDpXHmpUc1vGhA0J0IRChHX#scrollTo=vYolouOfONKd

MODELO ANALÍTICO DE CLASIFICACIÓN BINARIA

1. Análisis exploratorio

Para comprender de mejor manera los datos suministrados se analiza el comportamiento de algunas variables. Dado que los datos en su etapa inicial se encuentran normalizados, se dificulta el entendimiento de estos si se realiza un análisis profundo, por lo que sólo se analizarán las variables no normalizadas, las cuales son: 'TotalCharges, 'y la variable objetivo 'Churn'.

Características de los datos: Se cuenta con 7043 registros cada uno con 21 características, de las cuales 20 están normalizadas. En la verificación de datos nulos no se encontraron evidencias de que existen, de igual manera tampoco se encontraron datos duplicados.

Durante el proceso de conversión a tipo numérico de 'TotalCharges' se generó un error, el cual se debió a la presencia de valores faltantes representados como NaN. Dado que estos valores correspondían únicamente a 11 de los 7,043 registros totales (una proporción mínima), se decidió eliminar dichas filas del conjunto de datos.

En el Grafico 1 se observa la distribución de la variable objetivo (Churn), en donde 0 representa quienes no abandonan el servicio, mientras 1 los que sí lo hacen.

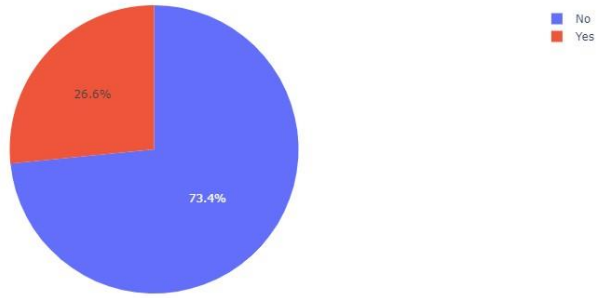

Gráfico 1. Churn

Hay un desequilibrio de clases en la variable objetivo, en donde se encuentran 5163 registros de personas que no abandonan al servicio, mientras que de quienes lo abandonan 1869 registros.

Por otro lado, se observó que clientes con Internet de fibra óptica son más propensos a desertar que los que tienen DSL o no tienen servicio (véase Grafico 2).

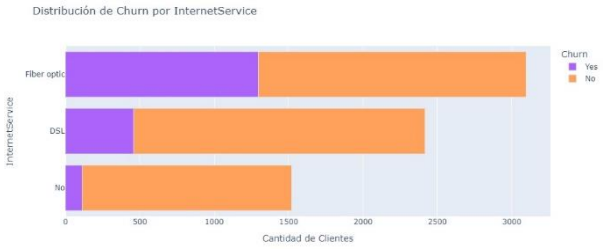

Gráfico 2. Chur vs Servicio de internet

En el repositorio del proyecto se observan otros análisis de interés en el estudio para una comprensión adecuada de los datos.

2. Limpieza y preprocesamiento de datos

La base no requirió una limpieza compleja al no presentaron atípicos y a que la mayoría de sus variables estaban anomalías. Se renombraron registros como, por ejemplo, 'No internet service' y 'No phone service' por 'No'. También se desarrolló una nueva variable con respecto a la permanencia del cliente en años. La variable objetivo se transformó de categórica a numérica.

3. Modelo predictivo

Ya que el objetivo de estudio es realizar un modelo predictivo que permita saber la probabilidad de que un cliente abandone el servicio, se proponen los siguientes algoritmos para la solución del problema analítico:

1. Regresión logística (LR)
2. Clasificador de bosques aleatorios (RF)
3. Clasificador XGBoost (XGB)

Estos algoritmos presentan las cualidades adecuadas para abordar el problema en cuestión, dado que son buenos clasificadores binarios, son fáciles de interpretar, poseen gran capacidad de generalización en los datos y son eficientes al trabajar con grandes cantidades de información.

Dada la cantidad de variables en los datos, es importante realizar un método de selección de variables, con el propósito inicialmente de encontrar las variables que expliquen de mayor manera la variabilidad de las predicciones y por otro lado ahorrar capacidad computacional. La selección de variables se realizó mediante el método wrapper. El resultado obtenido indica las características más adecuadas que mejoran el rendimiento de los algoritmos. En total fueron 6 variables significativas para los modelos.

La medición del desempeño de los modelos se realizó tanto con todas las variables a disposición como con las variables seleccionadas, con el objetivo de saber si los modelos tienen mejor rendimiento con menos variables y si estas son capaces de predecir la variabilidad de la variable objetivo. Cabe aclarar que la métrica de desempeño seleccionada para evaluar los modelos es 'Recall, ya que en la gráfica 1 vemos que los datos no se encuentran balanceados.

Se realizaron 4 interacciones para observar el desempeño de los modelos y, los resultados obtenidos indican que la selección de variables no tiene un impacto positivo en el rendimiento del modelo; de hecho, en ciertos escenarios, incluso lo empeora.

Ahora veamos un gráfico comparativo para el desempeño de los modelos con todas las variables.

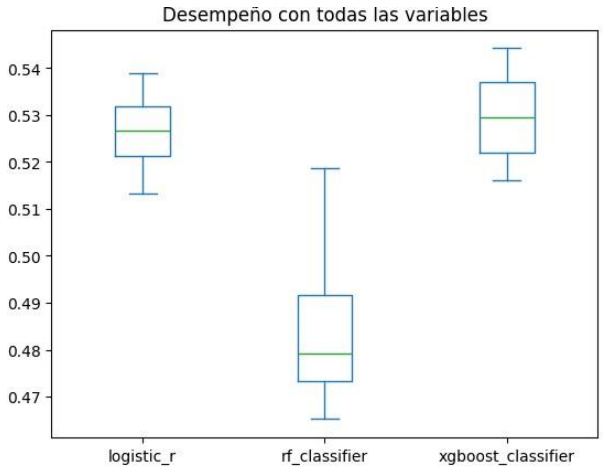

Gráfico 3. Distribución del desempeño de los modelos.

En el Gráfico 3 muestra que XGBClassifie con todas las variables presenta mayor estabilidad frente a los demás modelos, además de alcanzar un desempeño más alto que los otros dos modelos.

Se compararon las matrices de confusión del modelo con y sin afinamiento de hiperparámetros. El modelo afinado mostró una mejor sensibilidad para detectar la deserción, mientras que el no afinado fue más conservador, con mayor precisión y exactitud. Se eligió el modelo afinado por su mejor desempeño en la identificación de clientes que abandonan el servicio.

4. Predicciones

Al realizara las predicciones de la probabilidad de que un cliente abandone el servicio, se obtiene los siguientes resultados:

a. La deserción general es del 26.6%. De 1,407 clientes en el conjunto de prueba, 374 desertaron.
b. El tipo de contrato influye fuertemente en la deserción, los clientes con contrato de dos años tienen una tasa de deserción muy baja (2.44%).
c. Los Clientes con fibra óptica e Internet están más propensos a desertar (véase Gráfico 4).
d. El soporte técnico tiene efecto en la deserción. los clientes sin soporte técnico tienen mayores tasas de deserción.

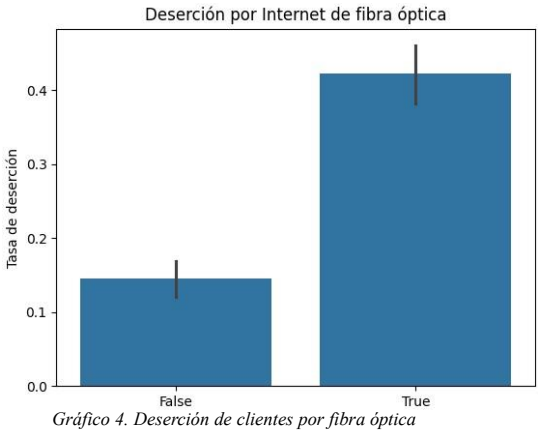

Gráfico 4. Deserción de clientes por fibra óptica

El plan de acción propuesto para combatir la deserción de los clientes es el siguiente.

- Diseñar promociones exclusivas que incentiven a los clientes actuales a migrar desde contratos mensuales hacia planes de 12 o 24 meses. Es clave resaltar beneficios como la estabilidad en el precio, el acceso a atención prioritaria o la posibilidad de obtener ahorros a largo plazo.
- Segmentar los clientes con mayor riesgo de cancelación, especialmente aquellos que tienen contrato mensual, no cuentan con soporte técnico o utilizan el servicio de fibra óptica. A partir de estos perfiles, se deben aplicar campañas personalizadas de retención que ofrezcan soluciones ajustadas a sus necesidades.
- El soporte técnico es un factor clave en la experiencia del cliente. Se sugiere incentivar su uso desde los primeros 90 días del servicio para fortalecer la confianza y resolver dudas o problemas iniciales. Además, incluir TechSupport como parte de beneficios exclusivos en planes anuales o en programas de fidelización puede mejorar la percepción general del servicio.
- Investigar a fondo las razones por las que algunos clientes abandonan el servicio de fibra, ya sea por precio, fallos técnicos o falta de información. Comunicar de forma efectiva las ventajas del servicio, como su velocidad, estabilidad y rendimiento, ayudará a reforzar su valor percibido. También es clave fortalecer la atención al cliente con tiempos de respuesta rápidos y soluciones efectivas.