Laboratorio 2
Predicción de Churn en Servicios de Suscripción Utilizando Modelos de Clasificación

Objetivo:

Aplicar técnicas de clasificación para predecir el *churn* (deserción de clientes) en un servicio de suscripción basado en un conjunto de datos de una empresa. Se debe explorar los datos, procesarlos, construir modelos de clasificación, evaluar su rendimiento e interpretar los resultados.

Descripción del Problema:

El *churn*, es la tasa a la que los clientes cancelan sus suscripciones, es una métrica vital para los negocios que ofrecen servicios de suscripción. Se emplearán técnicas de análisis predictivo para anticipar qué clientes tienen mayor probabilidad de churn, permitiendo a la empresa tomar medidas proactivas para retener a los clientes. El dataset contiene información anonimizada sobre las suscripciones de los clientes y su interacción con el servicio, incluyendo tipo de suscripción, método de pago, preferencias de contenido, interacciones con el soporte al cliente y otros atributos que podrían ser relevantes.

Requerimientos:

1. **Exploración y Visualización de Datos**
   - Cargar los datos de los archivos Info_lab2_parte1.parquet y Info_lab2_parte2.parquet que se encuentra en el GES, concatenándolos por filas.
   - Verificar si hay valores faltantes o errores en los datos.
   - Visualizar las relaciones entre las variables predictoras y la variable de respuesta. Utilizar gráficos apropiados, como scatterplots, boxplots o gráficos de línea.
   - Identificar y tratar posibles valores atípicos.
   - Detectar y corregir inconsistencias en caso sea necesario.
2. **Preprocesamiento de Datos**
   - Codificar las variables categóricas apropiadamente para su uso en un modelo de clasificación.
   - Normalizar o estandarizar las variables numéricas si se considera necesario.
3. **Ajuste de Modelos de Clasificación**
   - Ajustar multiples modelos de regresión logística para predecir el churn usando las variables predictoras. Utilizar `sklearn.linear_model.LogisticRegression`.
   - Considerar añadir interacciones entre variables. Se puede usar `PolynomialFeatures` para crear términos de interacción.
   - Explorar relaciones potencialmente no lineales y considerar transformaciones de variables.

4. **Evaluación de Modelos**
   - Calcular la matriz de confusión, la precisión, el recall, el F1-score y el AUC-ROC de cada modelo. Utilizar `sklearn.metrics`.
   - Utilizar validación cruzada de 5-fold para obtener estimaciones más robustas del rendimiento de cada modelo. Se puede usar `sklearn.model_selection.cross_val_score`.
   - Comparar diferentes modelos de clasificación (por ejemplo, con diferentes combinaciones de predictores) y seleccionar el mejor basándose en las métricas de rendimiento.
5. **Interpretación de Resultados**
   - Examinar los coeficientes del mejor modelo e interpretar su significado en el contexto de la predicción de churn.
   - Visualizar los churn predichos vs los churn reales por el mejor modelo.
   - Discutir las implicaciones de los resultados. ¿Qué factores parecen influir más en el churn? ¿Hay algunas limitaciones del modelo?
6. **Preguntas de Análisis y Modelo**
   - Existe la creencia que las variables MonthlyCharges y ViewingHoursPerWeek tienen gran importancia para saber si una persona se dará de baja del servicio, Su modelo puede corroborar esa creencia? ¿Existe alguna tendencia observable?
   - ¿Qué variables tienen una mayor asociación con la variable Churn? ¿Qué insights se pueden obtener de esta relación?
   - ¿Hay alguna relación entre las variables demográficas y la probabilidad de churn?
   - ¿Cuales serían la métrica más adecuadas para evaluar el rendimiento del modelo? La importancia de saber esta métrica única es para agregarla al dashboard de producción de la empresa, y mantener una visualización constante del comportamiento del modelo, ante nuevas tendencias de los clientes.
   - ¿Cómo se interpreta la matriz de confusión obtenida del modelo?
   - Explique la curva ROC y el AUC en la evaluación del modelo.
   - ¿Cuáles son las variables más importantes para predecir el churn según el modelo ganador?
   - ¿Cómo se pueden interpretar los coeficientes de las variables del modelo ganador?
   - Un experto con muchos años de experiencia en la empresa cree que la clave de todo el problema es entender la relación entre EngagementScore y Churn. ¿Esa relación existe?, ¿cómo se refleja en el modelo?
   - ¿Existen posibles variables de confusión en el dataset? Explique como las identifico que posibles conclusiones erróneas podemos obtener de ellas.

Entregables:

- Un archivo .zip con un notebook de Jupyter documentado (código y análisis) que muestre el trabajo realizado en cada uno de los requerimientos anteriores.
- Resumen de los hallazgos, interpretación de los resultados y discusión de las implicaciones.
- Explore técnicas de regularización como Ridge o Lasso para controlar la complejidad del modelo y evitar el sobreajuste. Compare su rendimiento con el modelo no regularizado.