UNIDAD TEMÁTICA 7: Ajuste, evaluación y sintonía de modelos

Trabajo de Aplicación 1

ESCENARIO

El objetivo del modelo en este problema (utilizaremos Naive Bayes) es predecir si una persona respondería o no a una campaña directa por email, en base a atributos demográficos (edad, estilo de vida, ingresos, tipo de auto, estatus familiar y afinidad deportiva)

Paso 1 – preparación de los datos

- Crea un dataset con 10000 ejemplos utilizando el operador “Generate Direct Mailing Data”
  seleccionando una semilla aleatoria local para asegurar la repetibilidad del ejercicio.
- Convierte el atributo “label” a binomial. Esto te permitirá seleccionar métricas de rendimiento específicas para clasificación binomial.
- Divide el dataset con Split Data en dos conjuntos, 80% para entrenamiento y 20% para testeo.
- Conecta la salida de 80% a un operador “Split Validation” y configura éste con una proporción de 0.7/0.3, y muestreo aleatorio.

Paso 2 – Operador de modelado y parámetros

Inserta un operador “Naive Bayes” dentro del proceso “Split Validation”, seguido del “Apply Model” y un operador “Performance (Binomial Classification)”. Configura las siguientes opciones en este último:

- accuracy,
- false positive,
- false negative,
- true positive,
- true negative,
- sensitivity,
- specificity, y
- AUC

Paso 3 Evaluación

Agrega otro operador Apply Model fuera del Split Validation, conéctale el modelo a su puerto "mod", y el dataset de test (20%) a su puerto "unl".

Agrega un operador "Create Lift Chart" con las siguientes opciones:

- target class = response,
- binning type =frequency, y
- number of bins = 10.

Paso 4 Ejecución e interpretación

Al ejecutar el modelo se generará la matriz de confusión y la curva ROC para la muestra de validación (30% del original 80%), mientras que generaremos una "lift curve" para la muestra de test (20%). Podríamos agregar otro "Performance (Binomial Classification) para el dataset de test.

Registra los valores de TP, TN, FP, FN

Observa los resultados de los parámetros seleccionados y verifica los cálculos:

Term	Definition
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

Observa que RapidMiner hace una distinción entre las dos clases cuando calcula la precisión y recall.

Para calcular el recall para "no response", la clase positiva tomada es "no response". ¿Cuál es el TP correspondiente? Y el FN? Cuánto el recall entonces para "no response"? Ver que esto contrasta con el valor que nosotros calculamos, pues asumimos que "response" era la clase "positiva".

El recall de la clase es una métrica muy importante, a tener particularmente en cuenta cuando tratamos con datos muy desbalanceados. Se considera que los datos están muy desbalanceados si la proporción de las clases está sesgada.

Al entrenar un modelo con datos desbalanceados, los valores de recall de las clases resultantes también tienden a quedar sesgados. Por ejemplo, en un dataset en el que hubiera sólo 2% de "responses" el modelo resultante puede tener un muy alto valor de recall para "no responses" pero un valor muy pequeño de recall de clase para "responses".

Este sesgo no se aprecia en la exactitud general del modelo, pero luego utilizar este modelo sobre datos no vistos puede resultar en severos errores de clasificación.

La solución para este problema puede ser o bien **balancear los datos de entrenamiento** para tener **una proporción más o menos similar de las clases,** o **insertar penalidades** o costos sobre las clasificaciones erróneas utilizando un operador **"Metacost"**. Analiza el funcionamiento de este operador y evalúalo.

El valor **AUC** ("Area Under Curve") se muestra junto con la curva **ROC**.

Valores de AUC cercanos **a 1** son indicativos de un **buen modelo**.

Mientras la predicción es correcta para los ejemplos, la curva da un paso hacia arriba (TP incrementado). Si la predicción es errónea, la curva da un paso hacia la derecha (FP incrementado).

Nota que RapidMiner puede mostrar dos curvas más AUC: "optimistic" y "pessimistic". Investiga y documenta qué hacen estas opciones.

Analiza la curva y documenta los hallazgos. ¿Cómo se puede interpretar en función de los datos y del modelo?

EJERCICIO DOMICILIARIO: ANALIZAR LA SALIDA DE LIFT CHART Y DOCUMENTARLA