Ejercicio 1

Zoe Borrone Luca Mazzarello Ignacio Pardo

2023-08-21

source ("sample_exp.R")

Introducción

A partir de los dos datasets incluidos con la consigna (Churn y Hearts) y dos datasets que obtuvimos nosotros en internet (flujo de autos y performance de estudiantes) realizamos un análisis exploratorio de los datos, y evaluamos la performance de distintos modelos de arboles de decisión para predecir la variable objetivo de cada dataset.

Datasets

Churn

El dataset Churn contiene información sobre clientes de una compañía de telecomunicaciones de Irán. Cada fila representa un cliente y cada columna una variable. Las variables son:

`call_failure, complains, subscription_length, charge_amount, seconds_of_use, frequency_of_use, frequency_of_sms, distinct_called_numbers, age_group, tariff_plan, status, age, customer_value`

La columna a predecir es la variable `churn` que indica si el cliente se dio de baja o no.

Link: https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset

Hearts

El dataset Hearts contiene información sobre pacientes y su condición respecto a Enfermedad del Corazón.

Las variables del dataset son:

`Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG`
`MaxHR, ExerciseAngina, Oldpeak, ST_Slope, HeartDisease`

La columna a predecir es la variable `HeartDisease` que indica si el paciente tiene o no Enfermedad del Corazón.

Link: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Flujo de autos

El dataset Flujo de autos contiene información sobre cantidad de autos al ingreso o egreso de la ciudad de Buenos Aires.

El dataset originalmente contenía las columnas:

`CODIGO_LOCACION,HORA,CANTIDAD,SENTIDO,LATITUD,LONGITUD`

Para convertirlo en un problema de predicción binaria, se agregó una columna `greater_mean` la cual contiene `True` si la cantidad de autos es mayor al promedio de autos de ese sentido y hora, y `False` en caso contrario.

Ademas para facilitar el análisis exploratorio, se agregaron las columnas `anio`, `mes`, `dia` y `hr` que contienen el año, mes, dia y hora respectivamente. Los minutos y segundos siempre eran 0, por lo que no se incluyeron. Por último se eliminó la columna `CODIGO_LOCACION` ya que no aportaba información relevante.

El dataset resultante contiene las columnas:

`SENTIDO,LATITUD,LONGITUD,anio,mes,dia,hr,greater_mean`

La columna a predecir es la variable `greater_mean` que indica si la cantidad de autos es mayor al promedio de autos de ese sentido y hora.

Al calcular la media de cantidad de autos, nos encontramos con el problema de cometer **Data Leakage**, ya que estabamos promediando la columna cantidad sin separar previamente en datos de Train y de Validación. Para solucionar este problem se podrían separar los datos en Train y Validación, y luego calcular la media de cantidad de autos de cada sentido y hora en el conjunto de Train. Por como está planteado el código esto no fue posible, por lo que solo usamos este dataset para observar su comportamiento en el experimento de ejemplo pero no en el resto de los experimentos.

Como el dataset contenía 189.815 filas, decidimos tomar una muestra aleatoria de 10.000 filas para poder trabajar con el mismo.

Link: https://data.buenosaires.gob.ar/dataset/flujo-vehicular-anillo-digital

Performance de estudiantes

Este dataset contiene información sobre muchos estudiantes en dos escuelas secundarias de Portugal, y su performance en dos materias: Matemática y Portugués. Para llevar a cabo el experimento, nos quedamos con el dataset de Portugal ya que contiene más datos que el de Matemática (649 vs 395).

Los atributos son los siguientes 30:

- 1 school - escuela del estudiante (binario: “GP” - Gabriel Pereira o “MS” - Mousinho da Silveira)
- 2 sex - sexo del alumno (binario: “F” - mujer o “M” - hombre)
- 3 age - edad del estudiante (numérico: de 15 a 22 años)
- 4 address - tipo de domicilio del estudiante (binario: “U” - urbano o “R” - rural)
- 5 famsize - tamaño de la familia (binario: “LE3” - menor o igual a 3 o “GT3” - mayor de 3)
- 6 Pstatus - estado de convivencia de los padres (binario: “T” - viven juntos o “A” - separados)
- 7 Medu - educación de la madre (numérico: 0 - ninguna, 1 - educación primaria (4º grado), 2 - 5º a 9º grado, 3 - educación secundaria o 4 - educación superior)
- 8 Fedu - educación del padre (numérico: 0 - ninguna, 1 - educación primaria (4º grado), 2 - 5º a 9º grado, 3 - educación secundaria o 4 - educación superior)
- 9 Mjob - trabajo de la madre (nominal: ‘maestra’, relacionado con la atención sanitaria, ‘servicios’ civiles (por ejemplo, administrativo o policía), ‘en_casa’ u ‘otro’)
- 10 Fjob - trabajo del padre (nominal: profesor’, ‘sanitario’, ‘servicios’ civiles (por ejemplo, administrativo o policía), ‘en_casa’ u ‘otro’)
- 11 reason - razón para elegir este centro (nominal: cerca de “casa”, “reputación” del centro, preferencia de “curso” u “otro”)
- 12 guardian - tutor del alumno (nominal: “madre”, “padre” u “otro”)
- 13 traveltime - tiempo de viaje de casa al colegio (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, o 4 - >1 hora)
- 14 studytime - tiempo de estudio semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, o 4 - >10 horas)
- 15 failures - número de fracasos de clase anteriores (numérico: n si 1<=n<3, si no 4)
- 16 schoolsup - apoyo educativo adicional (binario: sí o no)
- 17 famsup - apoyo educativo familiar (binario: sí o no)
- 18 paid - clases pagadas extra dentro de la asignatura del curso (Matemáticas o Portugués) (binario: sí o no)
- 19 activities - actividades extraescolares (binario: sí o no)

- 20 nursery - asistió a la guardería (binario: sí o no)
- 21 higher - desea cursar estudios superiores (binario: sí o no)
- 22 internet - acceso a internet en casa (binario: sí o no)
- 23 romantic - con una relación romántica (binario: sí o no)
- 24 famrel - calidad de las relaciones familiares (numérico: de 1 - muy malas a 5 - excelentes)
- 25 freetime - tiempo libre después de la escuela (numérico: de 1 - muy bajo a 5 - muy alto)
- 26 goout - salir con amigos (numérico: de 1 - muy bajo a 5 - muy alto)
- 27 Dalc - consumo de alcohol en días laborables (numérico: de 1 - muy bajo a 5 - muy alto)
- 28 Walc - consumo de alcohol en fin de semana (numérico: de 1 - muy bajo a 5 - muy alto)
- 29 health - estado de salud actual (numérico: de 1 - muy malo a 5 - muy bueno)
- 30 absences - número de faltas de asistencia a la escuela (numérico: de 0 a 93)

13 binarias, 4 nominales contra 13 numericas

El dataset tambien contiene las siguientes 3 variables de notas:

- 31 G1 - nota del primer trimestre (numérico: de 0 a 20)
- 31 G2 - nota del segundo trimestre (numérico: de 0 a 20)
- 32 G3 - nota del tercer trimestre (numérico: de 0 a 20)

Para convertirlo en un problema de decisión binaria, agregamos de la columna "GM" que contiene "True" si el promedio de las notas G1, G2 y G3 es mayor a 12, y "False" en caso contrario.

Link: https://archive.ics.uci.edu/dataset/320/student+performance

Experimento de ejemplo

El experimento de ejemplo, busca ver la performance de los distintos modelos de arboles de decisión para predecir una variable de salida de cada dataset. Ademas, por cada dataset se plantean dos experimentos: uno con el dataset original y otro con el dataset preprocesado.

En el preprocesamiento, se eliminan datos con una proporción `prop_NA` a cada atributo del dataset.

Resultados

En la figura 1 se puede observar la performance de los distintos modelos de arboles de decisión con la métrica AUC (Area Under the Curve) para cada dataset y cada experimento.

Se puede observar como para los 4 datasets, el modelo con `prop_NA = 0` muestra alcanzar un pico de performance inicialmente, y luego ir decayendo a medida que se aumenta la profundidad del árbol. Aproximadamente esto comienza a ocurrir para todos los datasets cuando el `max_tree_depth >= 15`. Esto se debe a que el modelo se vuelve muy flexible y se va sobreajustando a los datos de entrenamiento, y por lo tanto no generaliza bien a los datos de validación.

Cabe destacar que para el dataset Churn, la caida es muy leve, manteniendose casi constante a medida que aumenta la profundidad del árbol. Por el otro lado, para el dataset Heart, se alcanza un valor de AUC casi igual a 1 pero luego al aumentar poco el tamaño del arbol este cae abruptamente.

Sin embargo, para los 4 datasets, el modelo con `prop_NA = 0.7` muestra una performance casi constante a medida que aumenta la profundidad del árbol. Esto se debe a que al eliminar la mas de la mitad de los datos, el modelo no tiene suficiente información para aprender, por lo que no se sobreajusta a los datos de entrenamiento y generaliza mejor a los datos de validación.

En el dataset de Student Performance, se ve mayor variabilidad en el corto plazo con el incremento de la profundidad del árbol. Esto se debe a que el dataset contiene muchos datos, por lo que el modelo puede aprender mejor y por lo tanto sobreajustarse a los datos de entrenamiento.

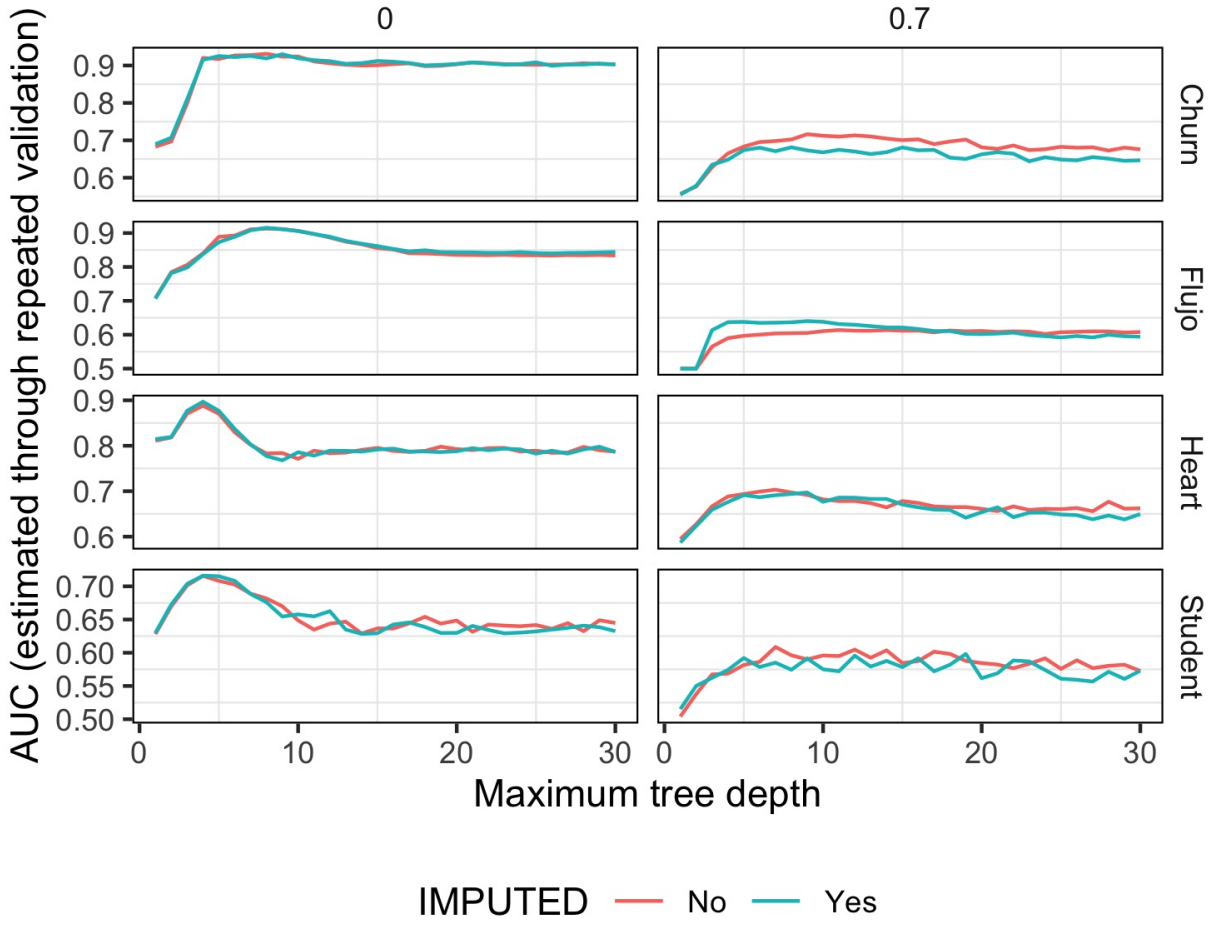

Figure 1: Output Experimento de Ejemplo