THE BRIDGE

Árboles de Decisión

Construcción (II): Variables no binarias

Árboles de decisión: Variables no binarias

- Hasta ahora hemos trabajado con variables de tipo binario (Si/No, Verdadero/Falso, etc)

- ¿Cómo se calcula el índice Gini de una variable no binaria? ¿Puede volver a aparecer en otros nodos inferiores?

Árboles de decisión: Variables no binarias

- Hasta ahora hemos trabajado con variables de tipo binario (Si/No, Verdadero/Falso, etc)
- ¿Cómo se calcula el índice Gini de una variable no binaria? ¿Puede volver a aparecer en otros nodos inferiores?
- En nuestro ejemplo, ¿cómo se trabaja con la variable "Edad"?

Cliente	Edad	Trabaja	Hipoteca	Ingresos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
C	48	NO	NO	Bajos
D	67	NO	SÍ	Bajos
E	18	SÍ	NO	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor

Edad	Ingresos
18	Bajos
25	Altos
32	Altos
48	Bajos
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes

Edad	Ingresos
18	Bajos
25	Altos
32	Altos
48	Bajos
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes

Edad	Ingresos
18	Bajos
25	Altos
32	Altos
48	Bajos
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes

Edad	Ingresos
18	Bajos
25	Altos
32	Altos
48	Bajos
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes

Edad	Ingresos
18	Bajos
25	Altos
32	Altos
48	Bajos
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes

Edad	Ingresos
18	Bajos
21.5	
25	Altos
28.5	
32	Altos
40	
48	Bajos
57.5	
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes
3. Se trata cada media como un punto de corte (una pregunta) y se obtiene el Gini para cada una

Edad	Ingresos
18	Bajos
21.5	
25	Altos
28.5	
32	Altos
40	
48	Bajos
57.5	
67	Bajos

Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes
3. Se trata cada media como un punto de corte (una pregunta) y se obtiene el Gini para cada una

Edad	Ingresos
18	Bajos
21.5	
25	Altos
28.5	
32	Altos
40	
48	Bajos
57.5	
67	Bajos

Gini (21.5) = 0.3
Gini (28.5) = 0.47
Gini (40) = 0.27
Gini (57.5) = 0.4

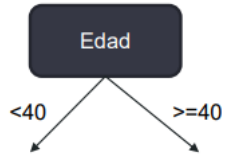
Árboles de decisión: Variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

- Tenemos que determinar cual es el corte (el valor de edad en este caso) mejor para dividir el target (ingresos), ¿es tener más de 25 años? ¿menos de 67?
- Existe un procedimiento:

1. Se ordenan los valores de menor a mayor
2. Se obtienen las medias de los valores de pares adyacentes
3. Se trata cada media como un punto de corte (una pregunta) y se obtiene el Gini para cada una
4. Se escoge el corte con menor Gini

Edad	Ingresos
18	Bajos
21.5	
25	Altos
28.5	
32	Altos
40	
48	Bajos
57.5	
67	Bajos



Árboles de decisión: Variables categóricas

- ¿Y si la variable no es numérica, ni binaria? Supongamos que tuviéramos una variable adicional "Estado Civil"

Árboles de decisión: Variables categóricas

- ¿Y si la variable no es numérica, ni binaria? Supongamos que tuviéramos una variable adicional "Estado Civil"

Estado civil	Ingresos
Soltero	Altos
Casado	Altos
Viudo	Bajos
Casado	Bajos
Soltero	Bajos

Árboles de decisión: Variables categóricas

- ¿Y si la variable no es numérica, ni binaria? Supongamos que tuviéramos una variable adicional "Estado Civil"

Estado civil	Ingresos
Soltero	Altos
Casado	Altos
Viudo	Bajos
Casado	Bajos
Soltero	Bajos

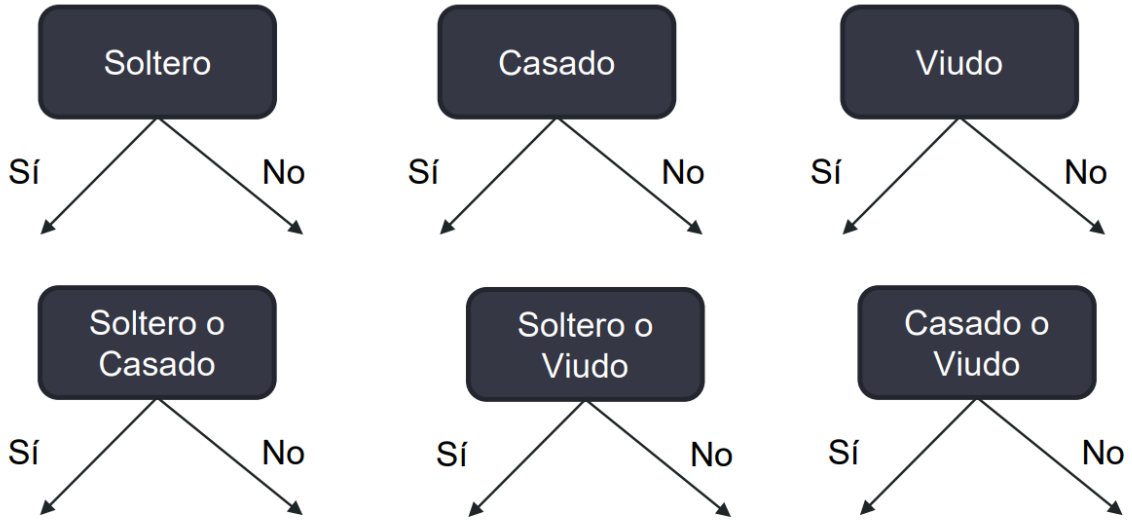
- Cada posible valor de la variable categórica se trata como una variable binaria y se calcula el Gini para estas

Árboles de decisión: Variables categóricas

- ¿Y si la variable no es numérica, ni binaria? Supongamos que tuviéramos una variable adicional "Estado Civil"

Estado civil	Ingresos
Soltero	Altos
Casado	Altos
Viudo	Bajos
Casado	Bajos
Soltero	Bajos

- Cada posible valor de la variable categórica se trata como una variable binaria y se calcula el Gini para estas




THE
BRIDGE