

Assignment 4 - Decision Tree

Keith G. Williams - 800690755

Friday, June 12, 2015

Problem Description

A company sent out some promotion to various houses and recorded a few facts about each house and also whether the people responded or not. Please create a Decision Tree (similar to one discussed in class) for the dataset below.

##	district	house_type	income	previous_customer	outcome
## 1	suburban	detached	high	no	nothing
## 2	suburban	detached	high	yes	nothing
## 3	rural	detached	high	no	responded
## 4	urban	semi-detached	high	no	responded
## 5	urban	semi-detached	low	no	responded
## 6	urban	semi-detached	low	yes	nothing
## 7	rural	semi-detached	low	yes	responded
## 8	suburban	terrace	high	no	nothing
## 9	suburban	semi-detached	low	no	responded
## 10	urban	terrace	low	no	responded
## 11	suburban	terrace	low	yes	responded
## 12	rural	terrace	high	yes	responded
## 13	rural	detached	low	no	responded
## 14	urban	terrace	high	yes	nothing

Information Content $I(C;F)$ Calculations

Each node will be split based on the information content calculation:

$$I(C;F) = \sum_{i=1}^m \sum_{j=1}^d P(C = c_i, F = f_j) \log_2 \frac{P(C = c_i, F = f_j)}{P(C = c_i)P(F = f_j)}$$

where C is the class (in this cases outcome), and F is the feature matrix (in this case district, house_type, income, previous_customer)

First Node

$I(\text{outcome}, \text{district}) =$

$$\begin{aligned}
 & P(\text{outcome} = \text{nothing}, \text{district} = \text{suburban}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{district} = \text{suburban})}{P(\text{outcome} = \text{nothing})P(\text{district} = \text{suburban})} \\
 & + P(\text{outcome} = \text{nothing}, \text{district} = \text{rural}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{district} = \text{rural})}{P(\text{outcome} = \text{nothing})P(\text{district} = \text{rural})} \\
 & + P(\text{outcome} = \text{nothing}, \text{district} = \text{urban}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{district} = \text{urban})}{P(\text{outcome} = \text{nothing})P(\text{district} = \text{urban})} \\
 & + P(\text{outcome} = \text{responded}, \text{district} = \text{suburban}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{district} = \text{suburban})}{P(\text{outcome} = \text{responded})P(\text{district} = \text{suburban})} \\
 & + P(\text{outcome} = \text{responded}, \text{district} = \text{rural}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{district} = \text{rural})}{P(\text{outcome} = \text{responded})P(\text{district} = \text{rural})} \\
 & + P(\text{outcome} = \text{responded}, \text{district} = \text{urban}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{district} = \text{urban})}{P(\text{outcome} = \text{responded})P(\text{district} = \text{urban})} \\
 & = \frac{3}{14} \log_2 \frac{\frac{3}{14}}{\frac{5}{14} \frac{5}{14}} + \frac{0}{14} \log_2 \frac{\frac{0}{14}}{\frac{5}{14} \frac{4}{14}} + \frac{2}{14} \log_2 \frac{\frac{2}{14}}{\frac{5}{14} \frac{5}{14}} + \frac{2}{14} \log_2 \frac{\frac{2}{14}}{\frac{9}{14} \frac{5}{14}} + \frac{4}{14} \log_2 \frac{\frac{4}{14}}{\frac{9}{14} \frac{4}{14}} + \frac{3}{14} \log_2 \frac{\frac{3}{14}}{\frac{9}{14} \frac{5}{14}} \\
 & = 0.247
 \end{aligned}$$

$I(\text{outcome}, \text{housetype}) =$

$$\begin{aligned}
 & P(\text{outcome} = \text{nothing}, \text{housetype} = \text{detached}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{detached})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{detached})} \\
 & + P(\text{outcome} = \text{nothing}, \text{housetype} = \text{semidetached}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{semidetached})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{semidetached})} \\
 & + P(\text{outcome} = \text{nothing}, \text{housetype} = \text{terrace}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{terrace})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{terrace})} \\
 & + P(\text{outcome} = \text{responded}, \text{housetype} = \text{detached}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{detached})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{detached})} \\
 & + P(\text{outcome} = \text{responded}, \text{housetype} = \text{semidetached}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{semidetached})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{semidetached})} \\
 & + P(\text{outcome} = \text{responded}, \text{housetype} = \text{terrace}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{terrace})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{terrace})} \\
 & = \frac{2}{14} \log_2 \frac{\frac{2}{14}}{\frac{5}{14} \frac{4}{14}} + \frac{1}{14} \log_2 \frac{\frac{1}{14}}{\frac{5}{14} \frac{5}{14}} + \frac{2}{14} \log_2 \frac{\frac{2}{14}}{\frac{5}{14} \frac{5}{14}} + \frac{2}{14} \log_2 \frac{\frac{2}{14}}{\frac{9}{14} \frac{4}{14}} + \frac{4}{14} \log_2 \frac{\frac{4}{14}}{\frac{9}{14} \frac{5}{14}} + \frac{3}{14} \log_2 \frac{\frac{3}{14}}{\frac{9}{14} \frac{5}{14}} \\
 & = 0.050
 \end{aligned}$$

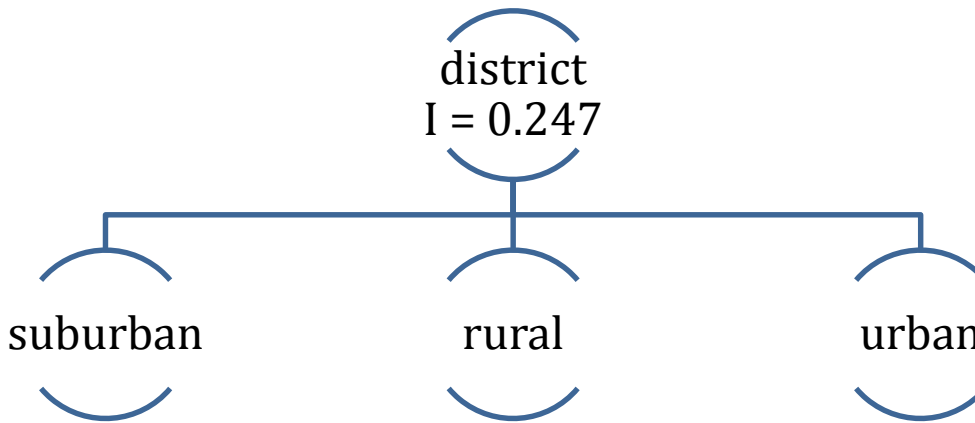
$I(\text{outcome}, \text{income}) =$

$$\begin{aligned}
 & P(\text{outcome} = \text{nothing}, \text{income} = \text{high}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{income} = \text{high})}{P(\text{outcome} = \text{nothing})P(\text{income} = \text{high})} \\
 & + P(\text{outcome} = \text{nothing}, \text{income} = \text{low}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{income} = \text{low})}{P(\text{outcome} = \text{nothing})P(\text{income} = \text{low})} \\
 & + P(\text{outcome} = \text{responded}, \text{income} = \text{high}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{income} = \text{high})}{P(\text{outcome} = \text{responded})P(\text{income} = \text{high})} \\
 & + P(\text{outcome} = \text{responded}, \text{income} = \text{low}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{income} = \text{low})}{P(\text{outcome} = \text{responded})P(\text{income} = \text{low})} \\
 & = \frac{4}{14} \log_2 \frac{\frac{4}{14}}{\frac{5}{14} \frac{7}{14}} + \frac{1}{14} \log_2 \frac{\frac{1}{14}}{\frac{5}{14} \frac{7}{14}} + \frac{3}{14} \log_2 \frac{\frac{3}{14}}{\frac{9}{14} \frac{7}{14}} + \frac{6}{14} \log_2 \frac{\frac{6}{14}}{\frac{9}{14} \frac{7}{14}} \\
 & = 0.152
 \end{aligned}$$

$$I(\text{outcome}, \text{previouscustomer}) =$$

$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{no}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{no})}{P(\text{outcome} = \text{nothing})P(\text{previouscustomer} = \text{no})} \\
& + P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{yes}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{yes})}{P(\text{outcome} = \text{nothing})P(\text{previouscustomer} = \text{yes})} \\
& + P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{no}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{no})}{P(\text{outcome} = \text{responded})P(\text{previouscustomer} = \text{no})} \\
& + P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{yes}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{yes})}{P(\text{outcome} = \text{responded})P(\text{previouscustomer} = \text{yes})} \\
& = \frac{2}{14} \log_2 \frac{\frac{2}{14}}{\frac{5}{14} \frac{8}{14}} + \frac{3}{14} \log_2 \frac{\frac{3}{14}}{\frac{5}{14} \frac{6}{14}} + \frac{6}{14} \log_2 \frac{\frac{6}{14}}{\frac{9}{14} \frac{8}{14}} + \frac{3}{14} \log_2 \frac{\frac{3}{14}}{\frac{9}{14} \frac{6}{14}} \\
& = 0.048
\end{aligned}$$

The greatest information content is $I(\text{outcome}, \text{district}) = 0.247$, so the first internal node will split outcome three ways on district.



Second Node Layer

- `district == suburban`

$$I(\text{outcome}, \text{housetype} \mid \text{district} = \text{suburban}) =$$

$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{housetype} = \text{detached}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{detached})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{detached})} \\
& + P(\text{outcome} = \text{nothing}, \text{housetype} = \text{semidetached}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{semidetached})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{semidetached})} \\
& + P(\text{outcome} = \text{nothing}, \text{housetype} = \text{terrace}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{terrace})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{terrace})} \\
& + P(\text{outcome} = \text{responded}, \text{housetype} = \text{detached}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{detached})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{detached})} \\
& + P(\text{outcome} = \text{responded}, \text{housetype} = \text{semidetached}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{semidetached})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{semidetached})} \\
& + P(\text{outcome} = \text{responded}, \text{housetype} = \text{terrace}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{terrace})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{terrace})} \\
& = \frac{2}{5} \log_2 \frac{\frac{2}{5}}{\frac{3}{5} \frac{2}{5}} + \frac{0}{5} \log_2 \frac{\frac{0}{5}}{\frac{3}{5} \frac{1}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{3}{5} \frac{2}{5}} + \frac{0}{5} \log_2 \frac{\frac{0}{5}}{\frac{2}{5} \frac{2}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{1}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{2}{5}} \\
& = 0.571
\end{aligned}$$

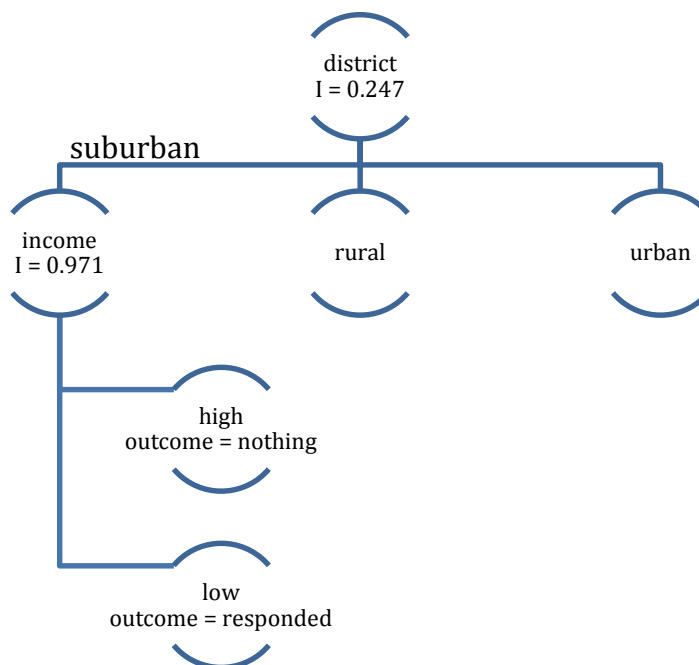
$$I(\text{outcome}, \text{income} \mid \text{district} = \text{suburban}) =$$

$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{income} = \text{high}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{income} = \text{high})}{P(\text{outcome} = \text{nothing})P(\text{income} = \text{high})} \\
& + P(\text{outcome} = \text{nothing}, \text{income} = \text{low}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{income} = \text{low})}{P(\text{outcome} = \text{nothing})P(\text{income} = \text{low})} \\
& + P(\text{outcome} = \text{responded}, \text{income} = \text{high}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{income} = \text{high})}{P(\text{outcome} = \text{responded})P(\text{income} = \text{high})} \\
& + P(\text{outcome} = \text{responded}, \text{income} = \text{low}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{income} = \text{low})}{P(\text{outcome} = \text{responded})P(\text{income} = \text{low})} \\
& = \frac{3}{5} \log_2 \frac{\frac{3}{5}}{\frac{3}{5} \frac{3}{5}} + \frac{0}{5} \log_2 \frac{\frac{0}{5}}{\frac{3}{5} \frac{2}{5}} + \frac{0}{5} \log_2 \frac{\frac{0}{5}}{\frac{2}{5} \frac{3}{5}} + \frac{2}{5} \log_2 \frac{\frac{2}{5}}{\frac{2}{5} \frac{2}{5}} \\
& = 0.971
\end{aligned}$$

$$I(\text{outcome}, \text{previouscustomer} \mid \text{district} = \text{suburban}) =$$

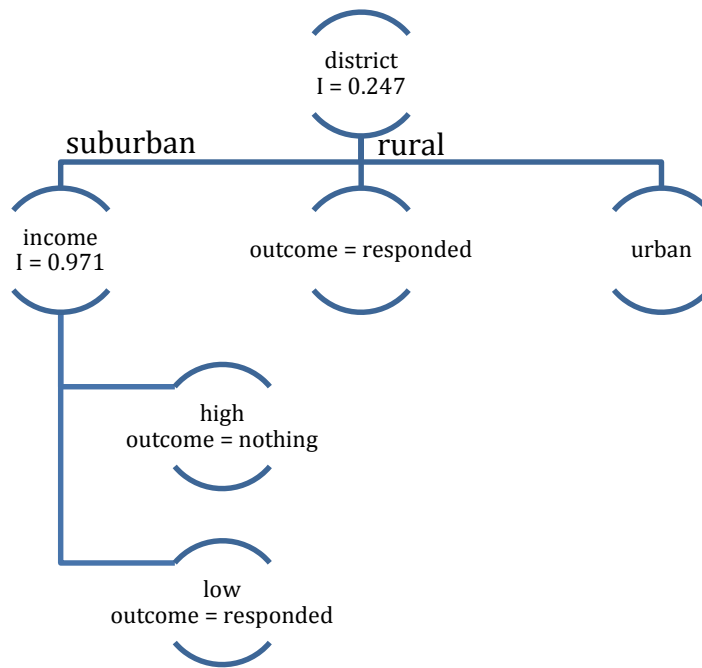
$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{no}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{no})}{P(\text{outcome} = \text{nothing})P(\text{previouscustomer} = \text{no})} \\
& + P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{yes}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{yes})}{P(\text{outcome} = \text{nothing})P(\text{previouscustomer} = \text{yes})} \\
& + P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{no}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{no})}{P(\text{outcome} = \text{responded})P(\text{previouscustomer} = \text{no})} \\
& + P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{yes}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{yes})}{P(\text{outcome} = \text{responded})P(\text{previouscustomer} = \text{yes})} \\
& = \frac{2}{5} \log_2 \frac{\frac{2}{5}}{\frac{3}{5} \frac{3}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{3}{5} \frac{2}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{3}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{2}{5}} \\
& = 0.020
\end{aligned}$$

The greatest information content is $I(\text{outcome}, \text{income} \mid \text{district} = \text{suburban}) = 0.971$, so the suburban internal node will split outcome on income. In fact, this split gives perfect separation, so this split will result in two leaf nodes.



- `district == rural`

Since $P(\text{outcome} = \text{responded} \mid \text{district} = \text{rural}) = 1$, no information can be gained by splitting this node further, so this path will terminate in a leaf node.



- `district == urban`

$$I(\text{outcome}, \text{housetype} \mid \text{district} = \text{urban}) =$$

$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{housetype} = \text{semidetached}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{semidetached})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{semidetached})} \\
& + P(\text{outcome} = \text{nothing}, \text{housetype} = \text{terrace}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{housetype} = \text{terrace})}{P(\text{outcome} = \text{nothing})P(\text{housetype} = \text{terrace})} \\
& + P(\text{outcome} = \text{responded}, \text{housetype} = \text{semidetached}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{semidetached})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{semidetached})} \\
& + P(\text{outcome} = \text{responded}, \text{housetype} = \text{terrace}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{housetype} = \text{terrace})}{P(\text{outcome} = \text{responded})P(\text{housetype} = \text{terrace})} \\
& = \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{3}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{2}{5}} + \frac{2}{5} \log_2 \frac{\frac{2}{5}}{\frac{3}{5} \frac{3}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{3}{5} \frac{2}{5}} \\
& = 0.020
\end{aligned}$$

$$I(\text{outcome}, \text{income} \mid \text{district} = \text{suburban}) =$$

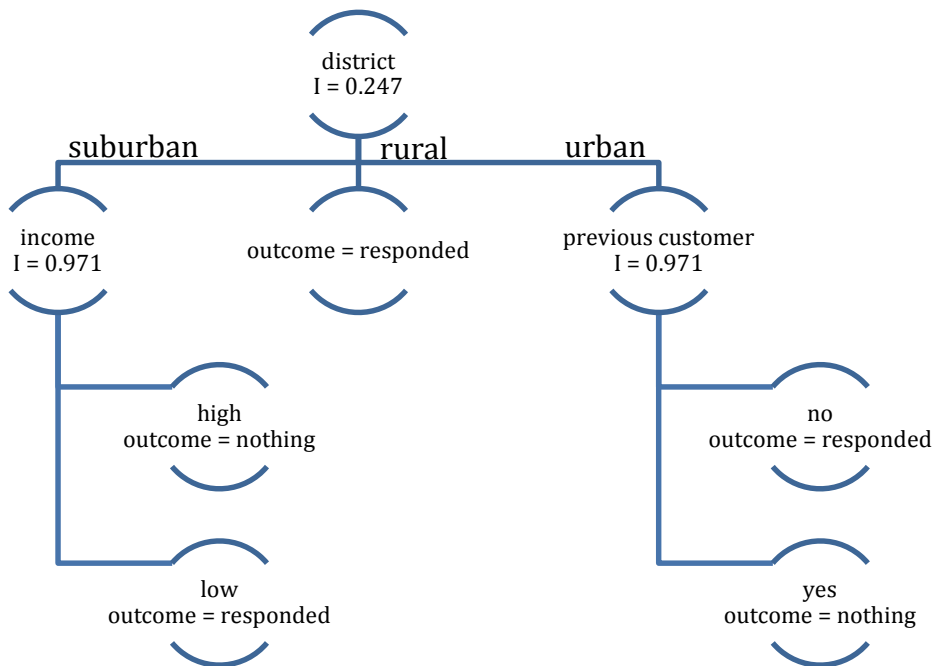
$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{income} = \text{high}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{income} = \text{high})}{P(\text{outcome} = \text{nothing})P(\text{income} = \text{high})} \\
& + P(\text{outcome} = \text{nothing}, \text{income} = \text{low}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{income} = \text{low})}{P(\text{outcome} = \text{nothing})P(\text{income} = \text{low})} \\
& + P(\text{outcome} = \text{responded}, \text{income} = \text{high}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{income} = \text{high})}{P(\text{outcome} = \text{responded})P(\text{income} = \text{high})} \\
& + P(\text{outcome} = \text{responded}, \text{income} = \text{low}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{income} = \text{low})}{P(\text{outcome} = \text{responded})P(\text{income} = \text{low})} \\
& = \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{2}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{2}{5} \frac{3}{5}} + \frac{1}{5} \log_2 \frac{\frac{1}{5}}{\frac{3}{5} \frac{2}{5}} + \frac{2}{5} \log_2 \frac{\frac{2}{5}}{\frac{3}{5} \frac{3}{5}} \\
& = 0.020
\end{aligned}$$

$$I(\text{outcome}, \text{previouscustomer} \mid \text{district} = \text{suburban}) =$$

$$\begin{aligned}
& P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{no}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{no})}{P(\text{outcome} = \text{nothing})P(\text{previouscustomer} = \text{no})} \\
& + P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{yes}) \log_2 \frac{P(\text{outcome} = \text{nothing}, \text{previouscustomer} = \text{yes})}{P(\text{outcome} = \text{nothing})P(\text{previouscustomer} = \text{yes})} \\
& + P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{no}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{no})}{P(\text{outcome} = \text{responded})P(\text{previouscustomer} = \text{no})} \\
& + P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{yes}) \log_2 \frac{P(\text{outcome} = \text{responded}, \text{previouscustomer} = \text{yes})}{P(\text{outcome} = \text{responded})P(\text{previouscustomer} = \text{yes})} \\
& = \frac{0}{5} \log_2 \frac{\frac{0}{5}}{\frac{2}{5} \frac{3}{5}} + \frac{2}{5} \log_2 \frac{\frac{2}{5}}{\frac{2}{5} \frac{2}{5}} + \frac{3}{5} \log_2 \frac{\frac{3}{5}}{\frac{3}{5} \frac{3}{5}} + \frac{0}{5} \log_2 \frac{\frac{0}{5}}{\frac{3}{5} \frac{2}{5}} \\
& = 0.971
\end{aligned}$$

The greatest information content is $I(\text{outcome}, \text{previous} \mid \text{district} = \text{urban}) = 0.971$, so the urban internal node will split outcome on previous customer. In fact, this split gives perfect separation, so this split will result in the final two leaf nodes.

Final Decision Tree



Note that house type is not used in the decision tree, since it is less discriminating than the other features at each node layer. From the final decision tree, it can be concluded that the company should send promotions to:

- low income, suburban residents
- rural residents
- urban residents who are not previous customers