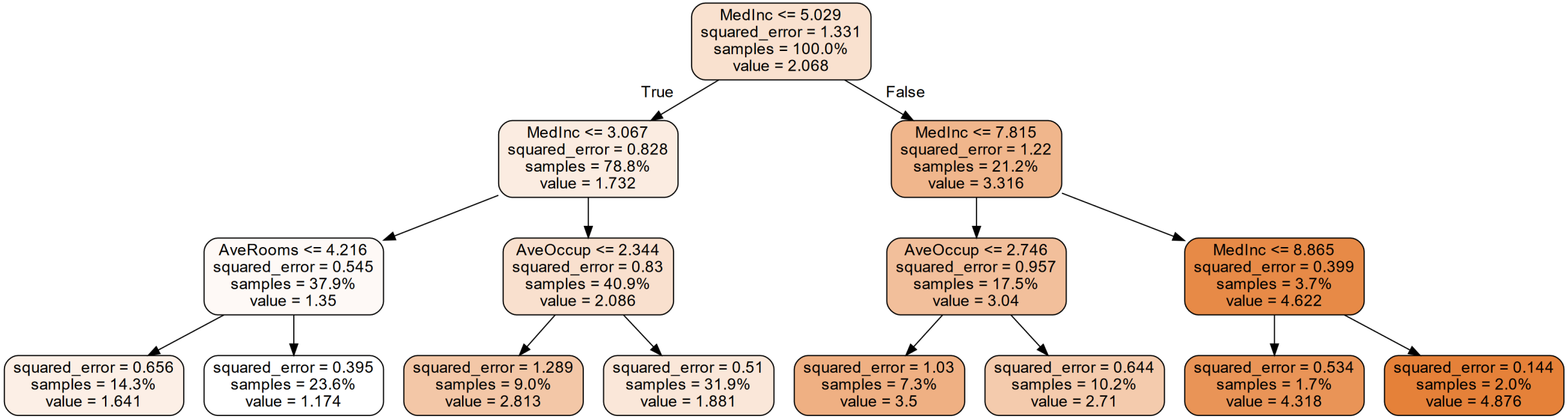
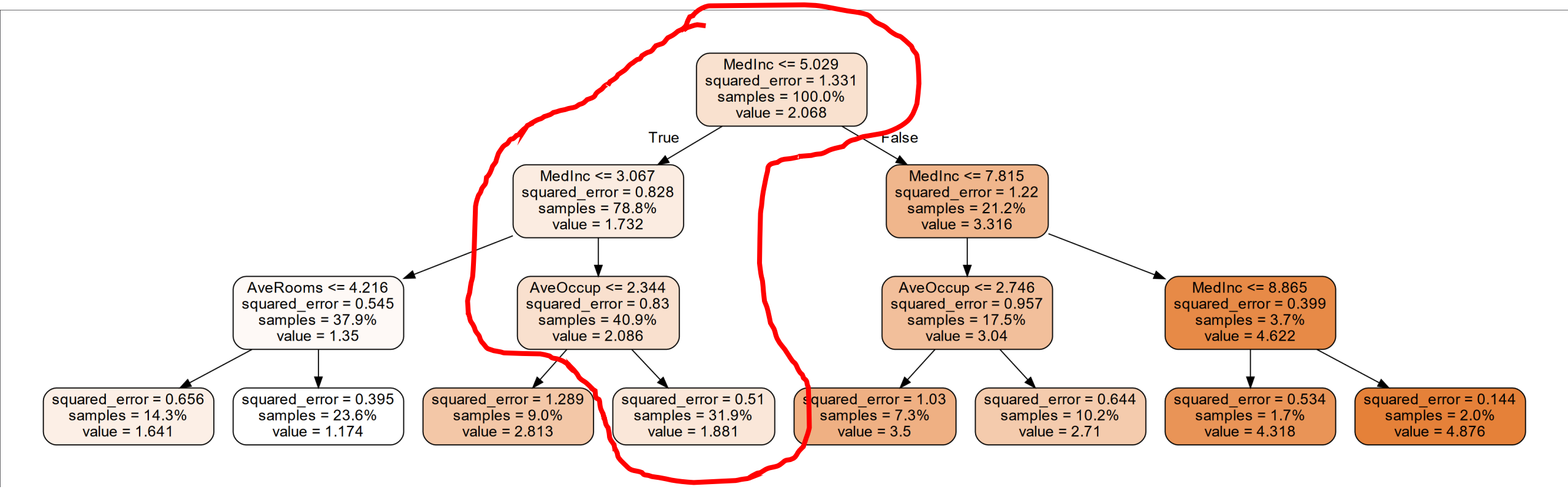
Local explanations

Local explanations

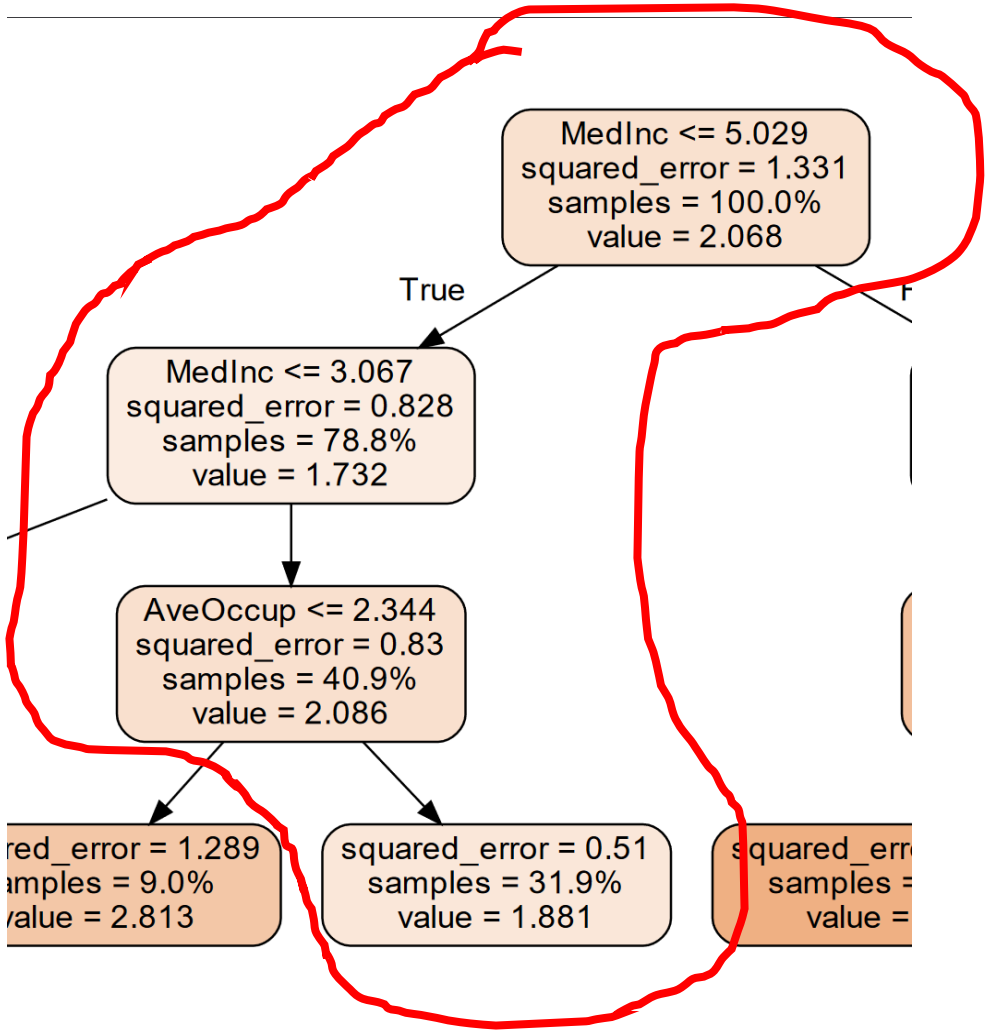


Local explanations

Sample 1

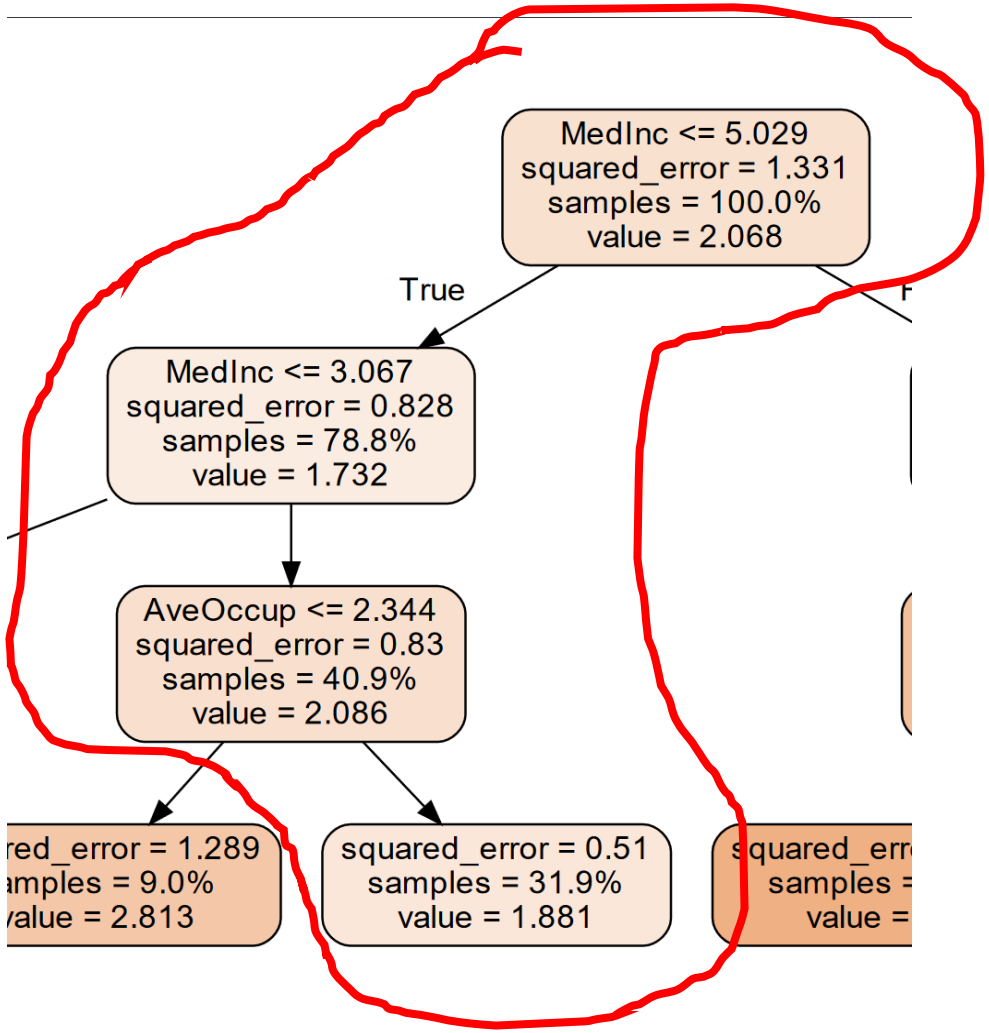


Local explanations



$$\texttt{Prediction} = \texttt{y\_mean} + \sum split\ contribution$$

Local explanations

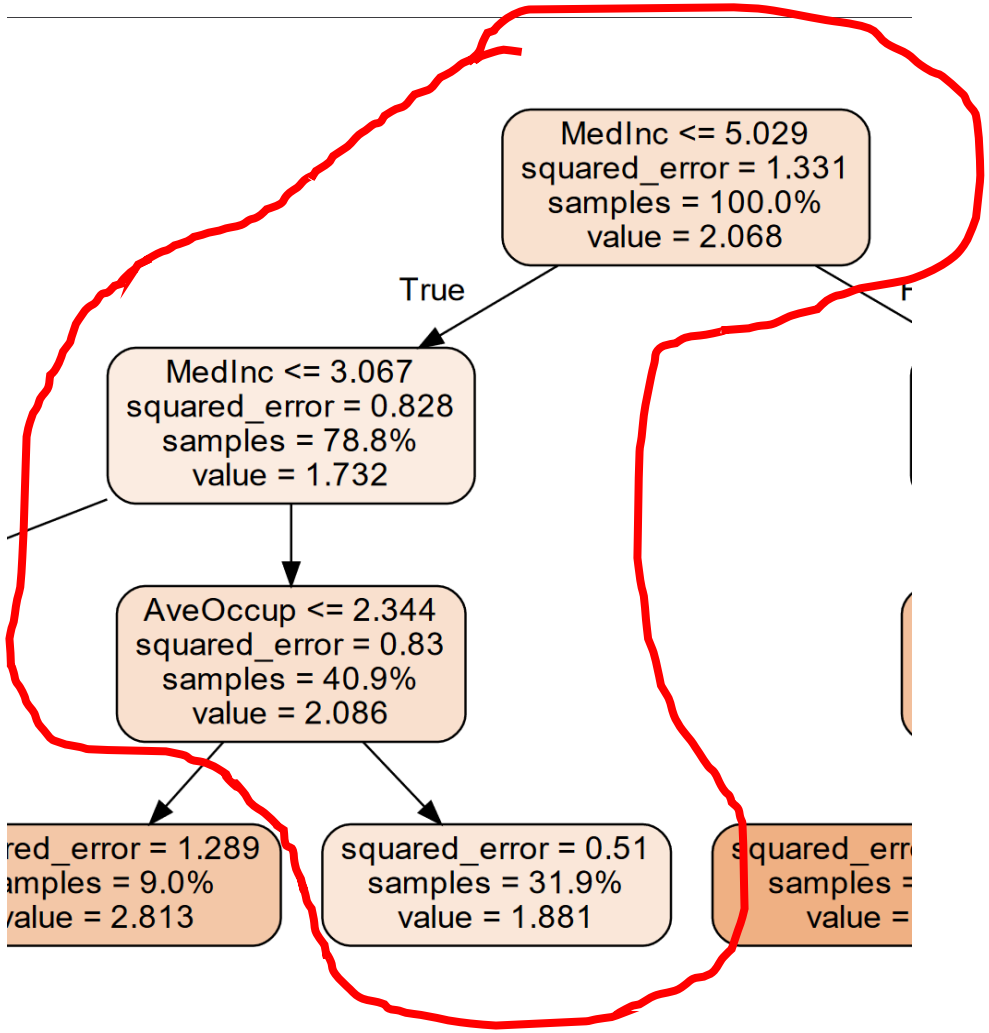


$$\text{Prediction} = \text{y\_mean} + \sum split\ contribution$$

y_mean = 2.068

Split sum = (1.732 – 2.068) +
(2.086 – 1.732) +
(1.881 – 2.086)

Local explanations



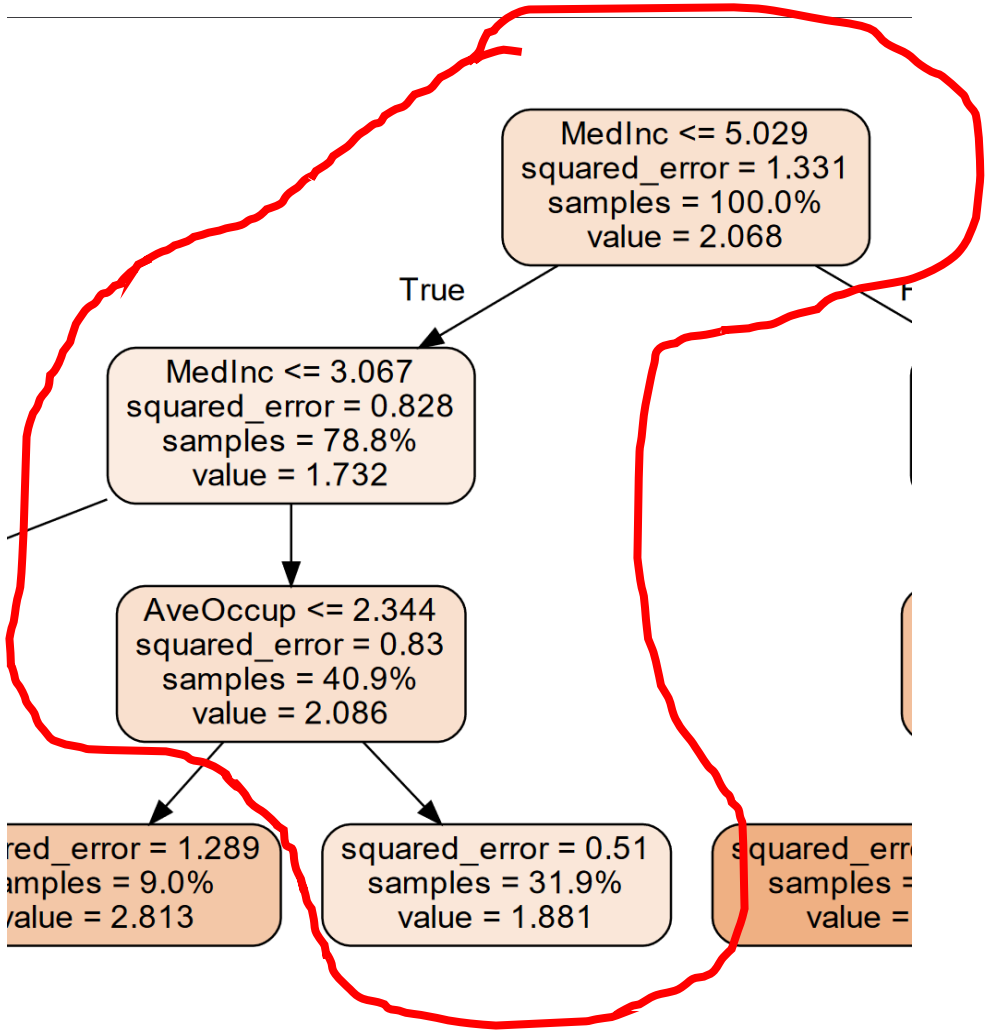
$$\text{Prediction} = \text{y\_mean} + \sum split\ contribution$$

y_mean = 2.068

Split sum = (1.732 – 2.068) +
(2.086 – 1.732) +
(1.881 – 2.086)

Split sum = -0.336 + 0.354 – 0.205 = -0.187

Local explanations

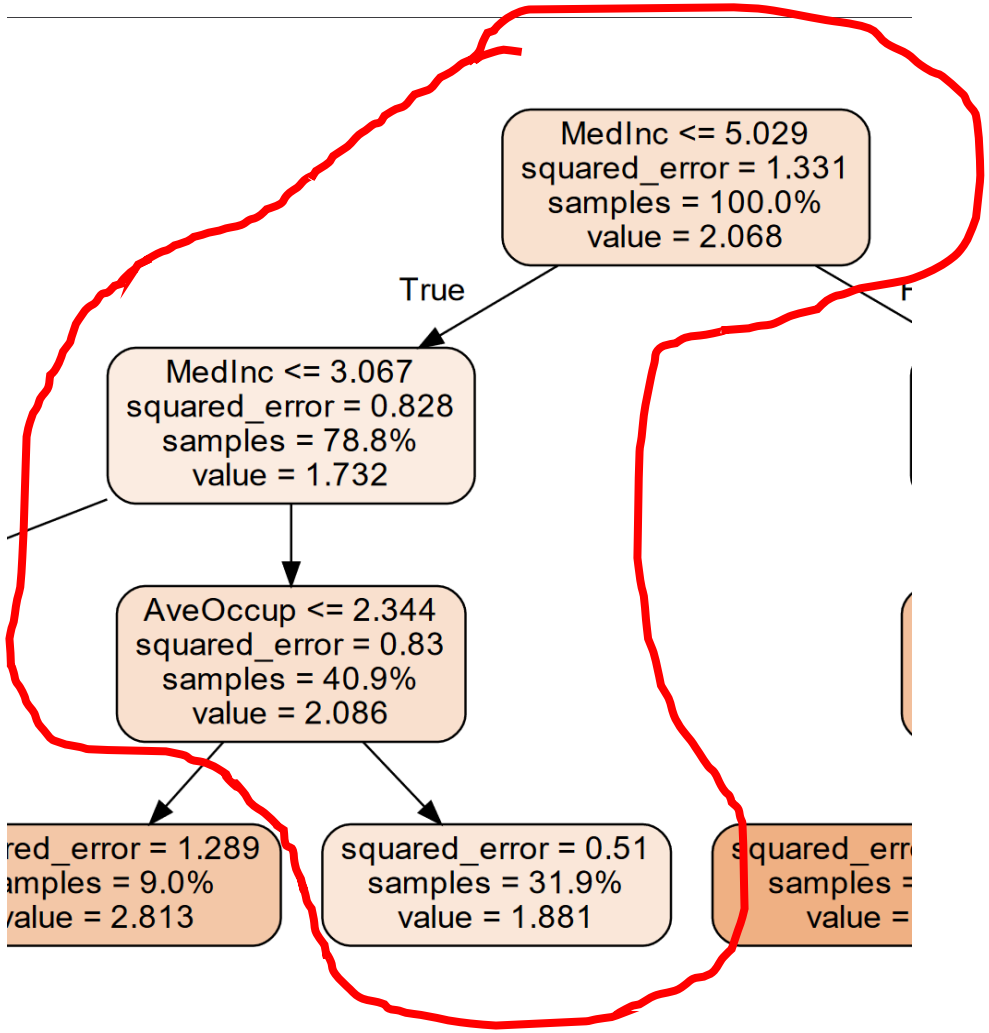


$$\text{Prediction} = \text{y\_mean} + \sum split\ contribution$$

$$\text{Prediction} = \text{y\_mean} + \text{split sum}$$

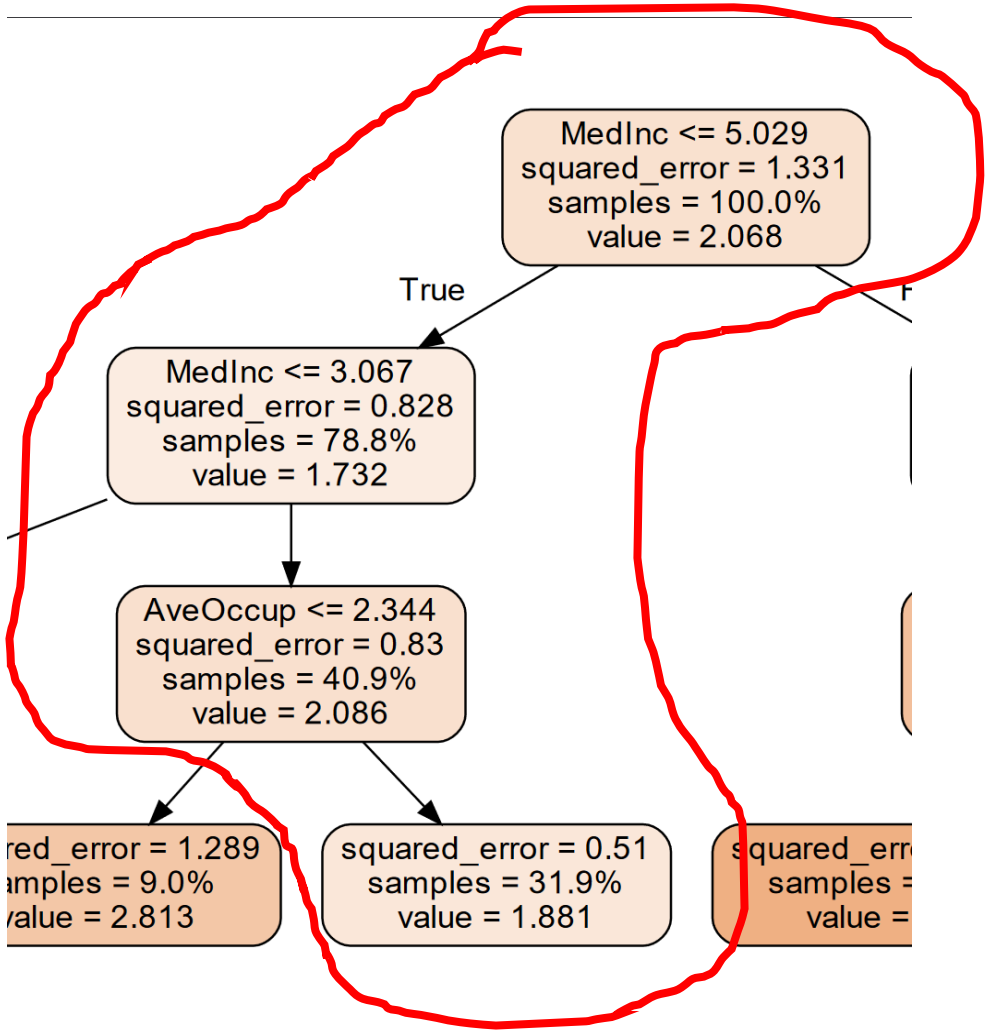
$$\text{prediction} = 2.068 - 0.187 = 1.881$$

Local explanations



$$\text{Prediction} = \text{y\_mean} + \sum feature\ contribution$$

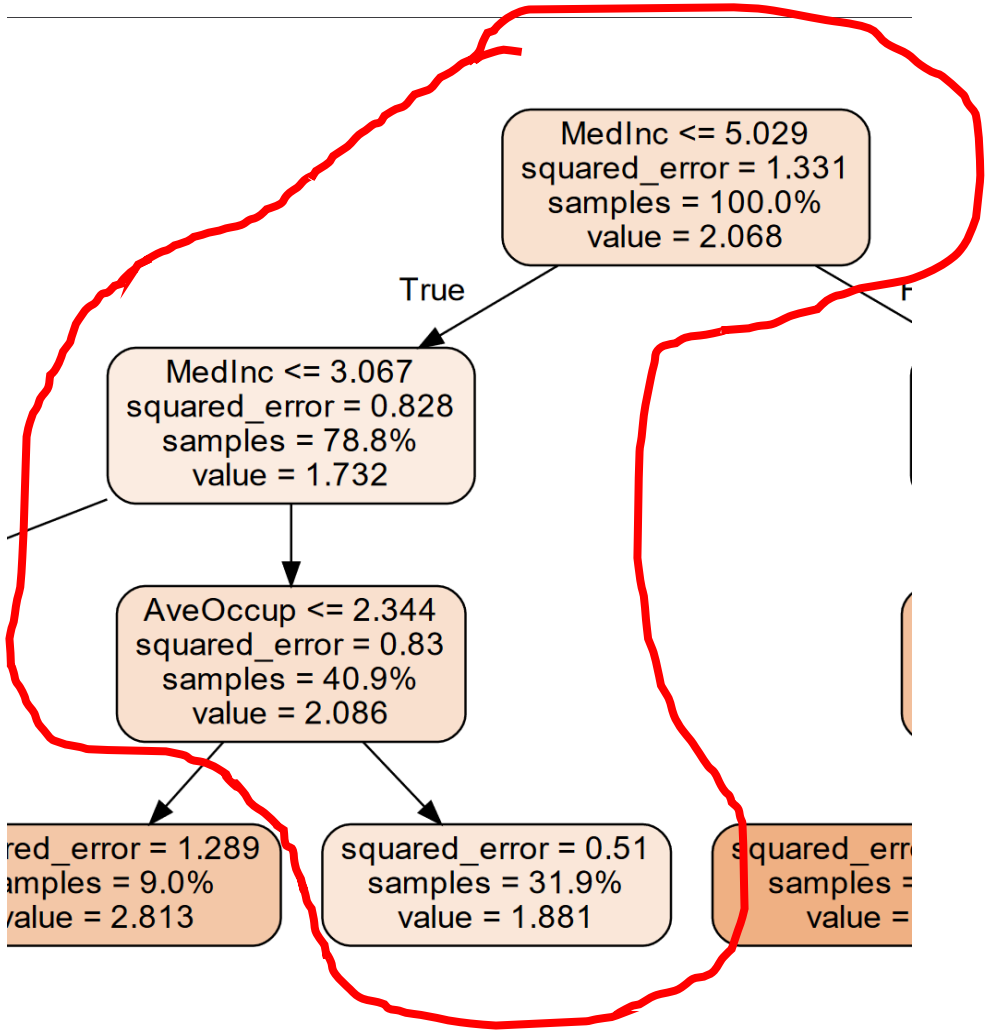
Local explanations



$$\text{Prediction} = \text{y\_mean} + \sum feature\ contribution$$

$$\begin{aligned} \text{Split sum} = {} & (1.732 - 2.068) + && \text{MedInc} \\ & (2.086 - 1.732) + && \text{MedInc} \\ & (1.881 - 2.086) && \text{AveOccup} \end{aligned}$$

Local explanations



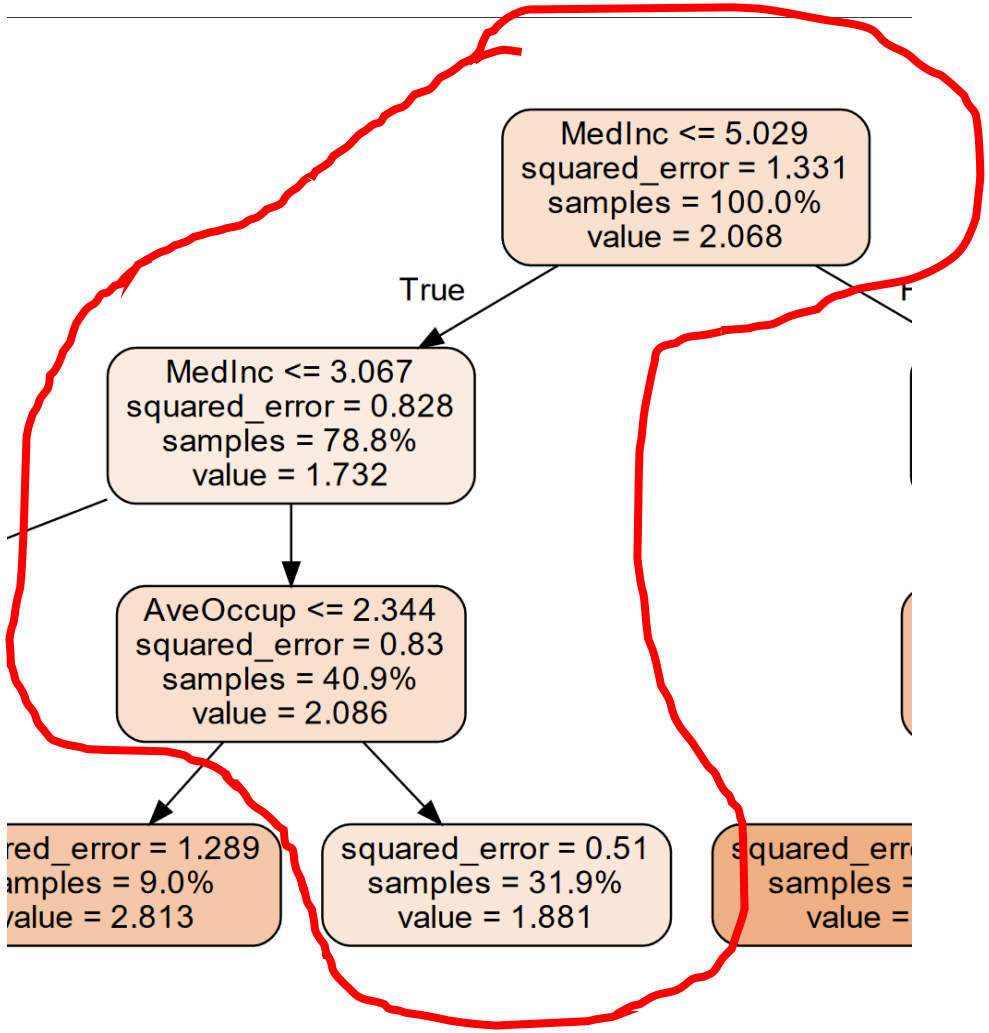
$$\text{Prediction} = \text{y\_mean} + \sum feature\ contribution$$

$$\begin{aligned}\text{Split sum} = {} & (1.732 - 2.068) + && \text{MedInc} \\ & (2.086 - 1.732) + && \text{MedInc} \\ & (1.881 - 2.086) && \text{AveOccup}\end{aligned}$$

$$\begin{aligned}\text{MedInc} = {} & (1.732 - 2.068) + \\ & (2.086 - 1.732) +\end{aligned}$$

$$\text{AveOccup} = (1.881 - 2.086)$$

Local explanations

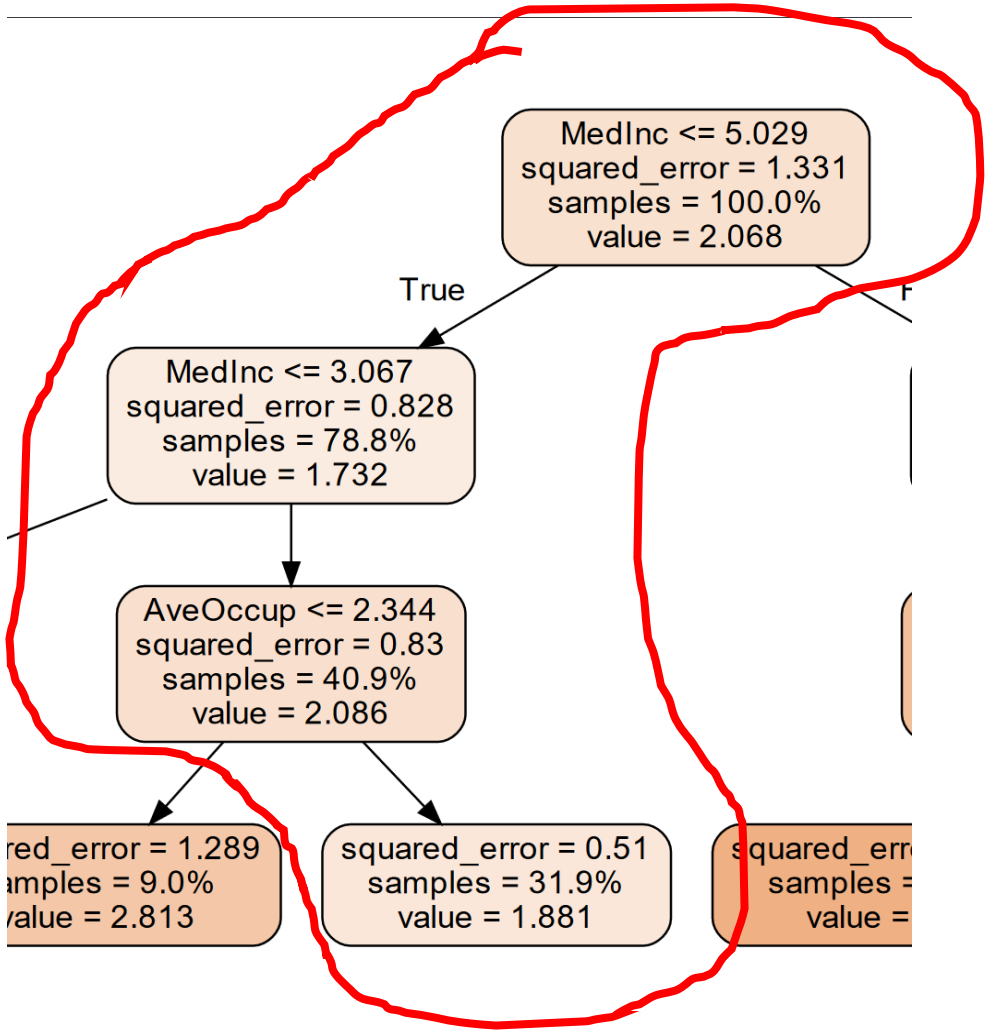


$$\text{Prediction} = \text{y\_mean} + \sum feature\ contribution$$

MedInc = 0.018

AveOccup = -0.205

Local explanations



$$\text{Prediction} = \text{y\_mean} + \sum feature\ contribution$$

MedInc = 0.018

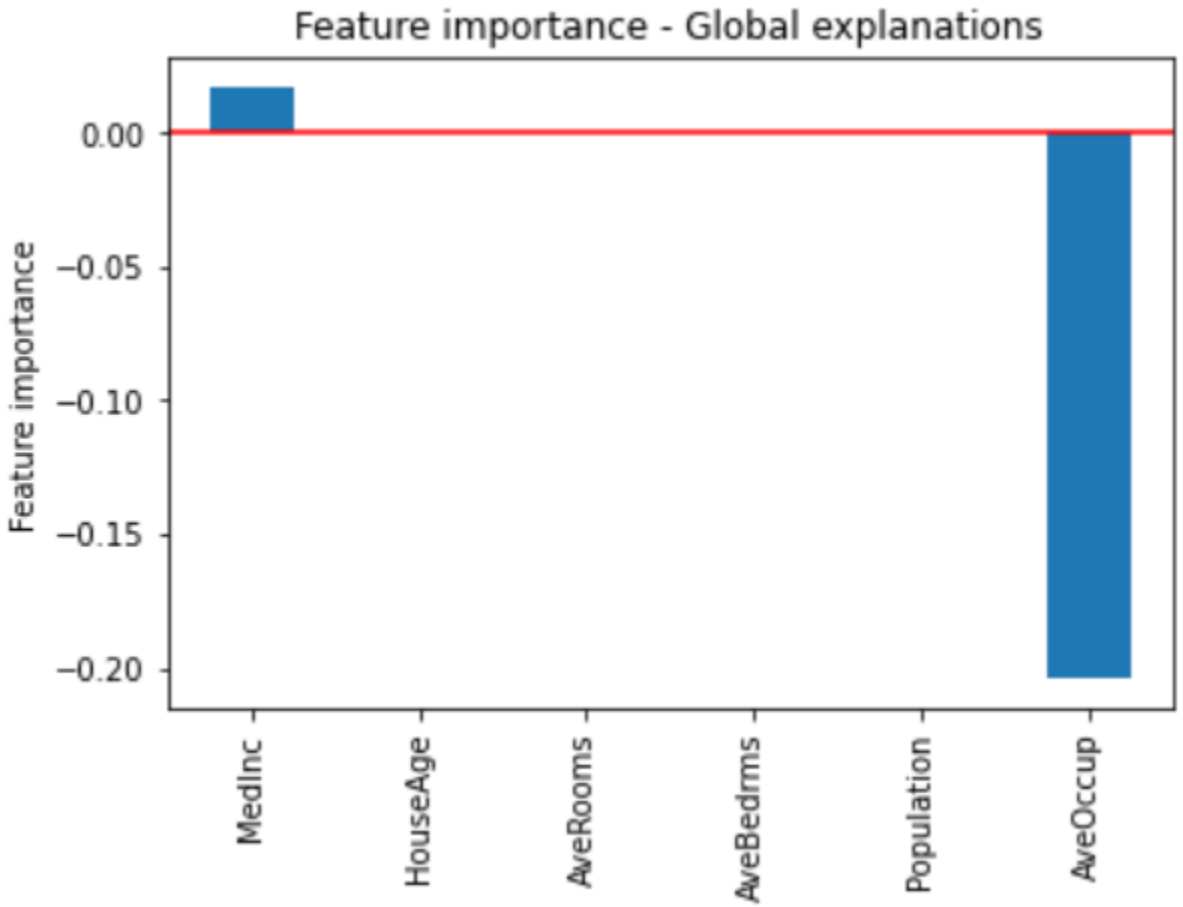
AveOccup = -0.205

Prediction = y_mean + feature sum

Prediction = 2.068 + (0.018-0.205)

Prediction = 1.881

Feature importance



- The change in value respect to the baseline induced by each feature.

Train In Data

THANK YOU

www.trainindata.com