# Exploring and Generating the Rattle Model Tab

## Training Dataset – Results Table and Diagram (Default Settings)

|  |
| --- |
| Figure 1‑1Performance Evaluation Table |
| Table  Description automatically generated |
|  |
| Figure 1‑2Decision Tree Diagram |
|  |

## Training Dataset – Performance Evaluation Table and Diagram (Reduced Number of Splits)

|  |
| --- |
| Figure 2‑1Performance Evaluation Table |
| A picture containing text  Description automatically generated |
|  |
| Figure 2‑2Decision Tree Diagram |
|  |

## Comparison/Findings from Figure 1‑1 (Default) and Figure 2‑1 (Reduced Splits)

After reviewing Figure 1‑1, produced with Rattle’s default settings, there are several areas of concern such as over-fitting, model instability, and degree of complexity. First, as you move down the complexity parameter (CP) column the values move closer to zero which signifies the risk of over-fitting. In addition, within the Relative Error (rel error) column, the last three rows in the table each have a significant increase when compared to their Cross-validation Error column (xerror) counterpart increasing the possibility of model instability. Lastly, having the Max Depth field set to 30 adds more branches and is likely to increase the level of complexity on this training dataset model. Turning attention to Figure 2‑1, the previously mentioned concerns were mostly addressed. Changing the number of levels/depth within the Max Depth field from 30 (Rattle’s default value) to 5 helped reduce complexity by removing the Employment category from the variables used in the tree construction as well as reduce the number of branches/leafs. There is also a significant increase in the Relative Error and Cross-validation Error values for Level 4 that will need to be watched as we continue this analysis.

## Training Dataset – Results Table and Diagram (Default, Input Variable Modification)

|  |
| --- |
| Figure 4‑1Performance Evaluation Table |
| Table  Description automatically generated with medium confidence |
|  |
| Figure 4‑2Decision Tree Diagram |
|  |

## Comparison/Findings from Figure 1‑1 (Default) and Figure 4‑1 (Default Settings, Input Variable Modification for Income)

After changing the INCOME input variable, the model used for Figure 1‑1 was ran again and labeled as Figure 4‑1. With this run, it was determined that Figure 4‑1 will likely have similar overfitting issues as well as model stability problems. This new figure (4.1) illustrates a minimal split value equal to 20, a minimum bucket of 7, a maximum depth of 30 and a complexity value of 0.01. In addition, changing the Max Depth field to a lower limit will assist in narrowing down/pruning extraneous entries that could adversely affect the model output. The Root node error and first two levels did not change between the two models; however, Cross-validation Error in Level 3 had a minimal decrease from 79% to 78%. A new split was added to the nsplit column for Figure 4‑1 and starting with Level 4 the differences between the Relative and Cross-validation Errors has increased.

# References

R Core Team (2021). R: A language and environment for statistical  
computing. R Foundation for Statistical Computing, Vienna, Austria.   
URL https://www.R-project.org/.

Berry, G.S.L.M. J. (2010). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. [devry]. Retrieved from https://devry.vitalsource.com/#/books/9781118275603/