Lachlan Sinclair

2/9/2020

Week 5 Assignment 1

Reflections

A. This question was far more complicated than it initially seemed. First for the binary target lets assume that one of the possible values is the designated value and the other is not.

   Case 1: When the prediction of the binary target matches the designated value of the target, the estimated probability and raw propensity are equal. To be clear I do not mean when the prediction is matching the actual value.

   Case 2: When the estimated prediction of the binary target does not match the designated target value, the raw propensity is equal to one minus the estimated prediction.

   I got a bit confused since the "confidence" of the prediction does not care about what the designated value is, but the raw propensity does. The confidence is only an evaluation of the how likely the prediction value will match the actual value. The general way I have been thinking about the raw propensity is how likely is the data instance to be a part of the designated binary value. Like the second reflection question I had to refer to the IBM's knowledge center to help me clarify what we were doing in the modeler. From what I gathered the confidence was represented as the $RC-LEAVE values and the raw propensity was represented as the $RPP-LEAVE values.

B. I was not able to find anything in the text or the lectures directly referencing this, however there was some information on this in the additional resources section. According to IBM's knowledge centers website there are a few ways in which their trees handle missing values, so there doesn't appear to be a single "correct" method used to handle such situations. Some types of decision trees use substitute values for missing categorical attributes, using a substitute allows that specific data instance to move down the tree until it reaches a leaf node. Another option is to allow the tree to treat missing values as a separate category, this allows missing values to be built directly into the tree structure. The most interesting option listed is used by the C5.0 tree, it uses a fractioning method to pass factional parts of a record down each branch from the node where the split occurs on the missing value. This final option is a bit confusing, conceptually I understand splitting the data instance into smaller sections, however I am unsure how this affects the conclusion. I assume each fractional section has its final assigned value added into some prediction based on the weighting of each section. I guess this one doesn’t fall into the “easily” portion of the question but is interesting none the less.