Analysis Summary

In week 3, the objective was to start with the model developed in week 2 and build variants. These variants were evaluated, including experimentation with boosting, bagging, and composite models. Each of these model variants include significant experimentation. As a result, separate stream files were created to avoid clutter and enhance clarity of results. These streams were organized in an SPSS Modeler project.

Basic checks for feature selection, and removal, had already been done, so these were not included the individual stream. The analysis started by performing an analysis on the week 2 model. This model, shown in the Week 2 Model stream, was built using the Auto Classifier node and resulted in a composite model that included a C5, CHAID, and Tree-AS models. An Analysis and Evaluation nodes were added to evaluate its quality.

The evaluation below showed very similar training and testing results. The gain looked very good up until roughly the 50 percentile mark. In addition, the ROC showed very strong performance until roughly a TP Rate of .79 where performance decreased very significantly. Accuracy was maintained according to the confusion matrix and the lift showed reasonable results.

|  |  |
| --- | --- |
| Gain | Lift |

|  |  |
| --- | --- |
| ROC | Confusion Matrix |

The Auto Classifier was adjusted at this point to build a composite model with 9 models. The purpose of this was to identify individual models that should be exampled more closely. 8 models, including C5, Tree-AS, C&R Tree, CHAID, Logistic Regression, LSVM, Bayesian Network, and Neural Net, scored above 58% accuracy. Each of these models were placed in a new stream, called Base Eval, for an individual model inspection.

The C5 model performed extremely well by itself. The gain and lift charts below show great distance from the 50% propensity threshold and especially gain showed strong alignment with the “blue” best line. ROC also performed well especially up to the high 80s with the TP Rate sensitivity.

|  |  |
| --- | --- |
| Gain | Lift |
| ROC | Confusion Matrix |

This stood in sharp contrast to the other 7 models that were evaluated. Not all of them have been included in this summary, but a C&R Tree is included below for comparison to the C5. The remaining 6 models showed very similar patterns to the C&R Treee. Note how close the results tract to the “red” lines and are not close to the “blue” lines.

|  |  |
| --- | --- |
| Gain | Lift |

|  |  |
| --- | --- |
| ROC | Confusion Matrix |

Boosting and Bagging

Multiple models were adjusted, in the Bag & Boost stream, to evaluate the potential benefits of boosting and bagging. Unfortunately, no real breakthroughs were observed in terms of accuracy, or stability, of the models. In some cases, like the boosted C&R Tree lift chart below, the model doesn’t seem to perform as well.

|  |  |
| --- | --- |
| C&R Tree Boosting | |
| Gain | Lift |

|  |  |
| --- | --- |
| ROC | Confusion Matrix |

|  |  |
| --- | --- |
| C&R Tree Bagging | |
| Gain | Lift |
| ROC | Confusion Matrix |

Bootstrapping

Bootstrapping was performed, in the Bootstrapping stream, to increase the number of available training and testing records from 390,116 to 533,344. This technique did not seem to have a significant impact. The evaluation was done using a C&R Tree.

|  |  |  |
| --- | --- | --- |
| Gain | | Lift |
| ROC | Confusion Matrix | |

Ensemble

In this final experiment, the Ensemble Node was used to combine a C5 and Logistic Regression models. This was done using a confidence-weighted voting which provided strong results. Overall, the results were very similar to the straight C5 model previously discussed.

|  |  |  |
| --- | --- | --- |
| Gain | | Lift |
| ROC | Confusion Matrix | |

In summary, the C5 model showed some of the best results, while the composite models showed similar or strong performance. Bagging, boosting, and bootstrapping didn’t show any significant advantages in this particular situation. However, they are just one set of options to explore when looking for ways to improve classification models.