# Lab Evaluations

## Testing Dataset

## Model Selection

As listed within Chapter 7, page 266 of our textbook, “Each of the candidate subtrees is used to classify the records or estimate values in the validation set. The tree that performs this task with the lowest overall error is declared the winner.” (Berry, 2010). After reviewing each of the performance measures, risk charts, and overall error values….I will be running the Section 8 model (Default Settings, Input Variable Modification) using the testing dataset. The following steps were completed prior to running on testing dataset: (1) Click on the Transform tab, (2) select the INCOME numeric variable (3) transform using Recode, KMeans, Number=6 (4) Go to Data tab and (5) confirm INCOME numeric variable is set to Ignore and BK6\_Income is set to Input.

## Testing Set Results – Error Matrix, Performance Measures, and Risk Chart

Error Matrix

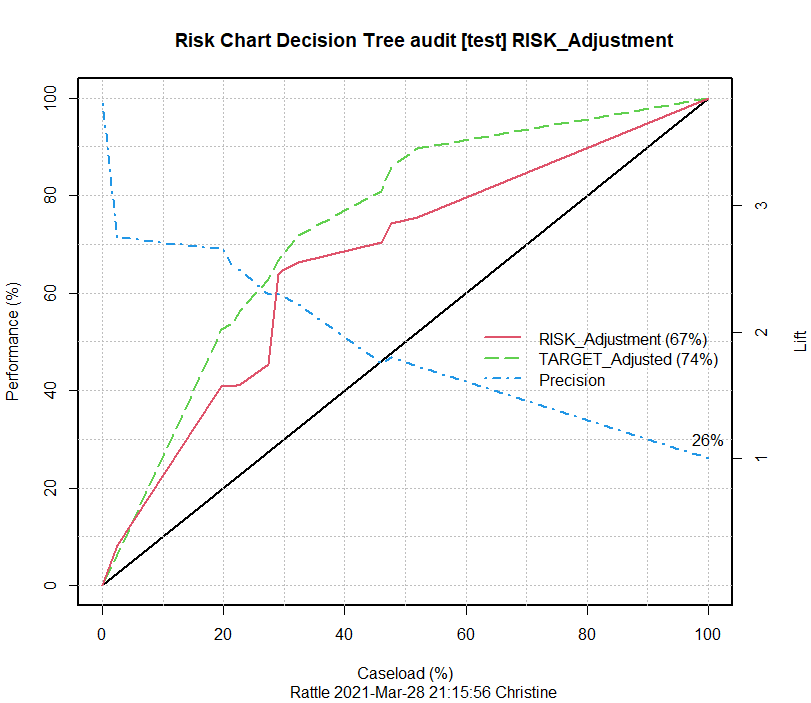
Graphical user interface, text, application, email

Description automatically generated

Performance Measures

|  |  |
| --- | --- |
| Precision | TP/(TP+FP) = 49/(49+33) = 49/82 = .60 = **60%** |
| Sensitivity | TP/(TP+FN) = 49/(49+29) = 49/78 = .63 = **63%** |
| Specificity | TN/(TN+FP) = 189/(189+33) = 189/222 = .85 = **85%** |

Risk Chart



## Evaluation Result

The Specificity performance measurement value shows that the model found 85% of the true negatives within the Testing Dataset. When reviewing the Sensitivity performance measure, this model is finding 63% of the actual positives and the Precision performance measure has a 60% accuracy in predicting positive targets. Lastly, the Overall Error and Averaged Class Error measures came in higher when applied to the Testing Dataset versus the training/validation datasets, 20.7% and 26.05% respectively. As our textbook authors mentioned, this seems as though the model accuracy of the testing dataset was slightly overstated because of the training and validation datasets being a part of the model creation. When the caseload (or percentage of taxpayers that we sampled) is between 20% and 50% the use of this model on the Testing Dataset will provide 1.5 times the performance.

## Summary Report for Senior Management

<*refer to next page for report*>

Prepare a half page summary, not including any illustrations you might want to use, of these activities that you’ve just done, that you’ve gone through, suitable for distribution to senior management. So, keep these statistics and terminology obviously to a minimum, and

explain in basic terms what’s going on here, in your opinion regarding this model, and specific you can put in recommendations. Should we even use these – this recommended model? Is it good enough to actually put out there and to use it in lieu of sampling all of of everybody that’s – and what kind of considerations might we have to consider to make that kind of judgment call.

# Summary Report

At management’s request, data mining techniques were initiated and completed for the company’s data on 2000 randomly chosen data records to help prepare an accurate and reliable prediction model to help understand who is more likely to get audited by the Internal Revenue Service (IRS) incurring penalties and those that are not likely to get audited or incur penalties. This was an extensive process that consisted of four major phases: (1) Process Overview/Data Setup, (2) Decision Tree Model Generation/Interpretation, (3) Decision Tree Model Evaluation and (4) Decision Tree Model Testing.

The first phase was a high-level overview of the process and steps needed for setting up and preparing the datasets which can also involve cleaning/prepping the data. There are many programs available to help with this work, but we focused on utilizing the Rattle graphical user interface as it is a powerful and versatile application focused on data mining, analytics, and statistics. A very important technique for data mining is separating a dataset into three separate partitions….Training Dataset, Validation Dataset, and Testing Dataset. Each of these partitions are of utmost importance to ensure the best prediction models are created, tested, and implemented.

Next phase was focused on the generation and interpretation of the Decision Tree Model(s) as well as using the Training Dataset partition. In this step, the analysts are getting more involved with the data and talking with business experts, management and asking clarifying questions to better understand the problem that the company/management is trying to solve.

Moving into the third phase consists of evaluating the Decision Tree Models by utilizing the Validation Dataset partitions. Analysts are also working the decision tree models and continuing to determine the best fit models for the solution to be solved. There are times in this phase where analysts may need to go back to the ‘drawing board’ and choose a different type of data modeling process.

We just finished with the final phase which involved evaluating the Best Decision Tree Models by utilizing the Testing Dataset partition. This final partition is a portion of the original dataset that has not been corrupted/touched, more aptly, used in any of the other testing which helps the analysts ensure there is a clean dataset partition for a final evaluation of the prediction model. Recommendations, results, and next steps will be communicated to the management team in the upcoming week.

# 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/