SOCIETY OF ACTUARIES

Predicting Group Long Term Disability Recovery and Mortality Rates using Tree Models

August 2017

Predicting Group Long Term Disability Recovery and Mortality Rates Using Tree Models

AUTHOR

Mervyn Kopinsky, FSA, EA
Experience Studies Actuary
Society of Actuaries

Caveat and Disclaimer

TABLE OF CONTENTS

Section 1: Purpose and Acknowledgements

1.1 Purpose

Creating decrement rate tables from a comprehensive set of data is a lengthy and time consuming task. Such tables often have numerous inputs (such as Age, Gender, Benefit Amount etc.) and the creation of decrement tables, aside from the work of calculating the appropriate rate tables requires significant effort to ensure that each possible combination of inputs results in a reasonable estimated decrement rate.

We have found that by using the tree modeling capabilities available in most statistical packages (we used R for this study), we were able to quickly develop recovery and mortality rates that appear to fit the data extremely well. Additionally, the tree modeling methodology avoids the possibility of producing unreasonable rates where a traditional approach could potentially be swayed by extreme values in the data.

The use of Predictive Analytic techniques is rapidly expanding within the actuarial community. We hope that this report provides additional insights into how these techniques can be used. We have also provided the R code that was used to develop these models so as to simplify the learning curve for anyone interested in further exploring these methods.

1.2 Scope

The 2004 – 2012 Group Long Term Disability (GLTD) Database provides a rich source of robust data regarding the recovery and mortality experience of about 25 companies who insure Long Term Disability business. This report details the results obtained from fitting a tree model to this data so as to estimate recovery and mortality rates.

The most recent release of this database also included several new variables that were not previously part of this database, namely:

1. Region
2. Case Size
3. Integration with Short Term Disability
4. Industry SIC Code
5. Taxability of Benefits
6. Indexed Monthly Salary
7. Company Size

We developed various models including these new variables to determine whether the new variables would improve the predictive ability of the models.

The bulk of this report is devoted to the modeling of non-maternity related recovery rates. However, the text applies equally to mortality rates and maternity recovery rates – models for these are discussed towards the end of the report.

This study data relates to claims already incurred, and the ultimate recoveries and deaths associated with these claims. The study does not review the incidence of disability claims.

1.3 Acknowledgements

I would like to thank Brian Holland and Rick Leavitt for their peer review of this report. Their comments and insights added significantly to the report's readability. I would also like to thank our Summer Intern Alexis Berg for her help in developing and testing many of the results discussed in this report. Finally I would like to thank Erika Schulty for her help in finalizing this report and getting it ready for publication.

Section 2: Background

2.1 Background

In 2011, the Committee Society of Actuaries published the 2008 Group Long Term Disability Experience Table Report which detailed a mathematical model for determining recovery and mortality rates. The data underlying this report was collected for calendar years 1997 to 2006.

The recovery rate model published in 2008 relied on 7 input variables:

1. Duration
2. Age
3. Gender
4. Disability Category
5. Own to Any Occupation Transition
6. Gross Indexed Benefit Amount
7. Elimination Period

Recovery rates for specific individual characteristics (Age, Duration, Gender etc.) are determined by multiplying together various factors from the model.

The 2008 model overall gives a very good fit when measured against the 2004 to 2012 data. Various webcasts (available under Professional Development on the Society of Actuaries website) have reviewed the Actual to Expected performance of the 2008 model against the 2004 to 2012 data. We have not focused on reviewing the areas in which the Actual to Expected's (based on the 2008 model) differ – rather this report looks at developing recovery and mortality rates from scratch, using this newer dataset. We use the Expected recovery and mortality rates from the 2008 report (which are already included in the data) as a baseline to test how well our model fits the data.

2.2 A Note on Terminology

Throughout this report, we refer to the Expected number of recoveries as being the number of recoveries expected using the rates from the 2008 model, and multiplying them by the applicable exposure for the record. As noted above, these expected recoveries are already calculated and in the data.

We use the term Predicted throughout this report to refer to the recovery and mortality rates developed in this report. The Predicted number of recoveries / deaths is then the predicted recovery / mortality rate multiplied by the applicable exposure.

Section 3: Tree Models

3.1 Why Consider Tree Models?

Recovery rates have significant non-linearity and do not change monotonically due to transition periods – such as the change from Own to Any Occupation and other policy features. These characteristics also suggest that a linear model, even with the often-used logarithmic transformation of the dependent rates, may not be flexible enough to fit the observed rates. As described in detail in this report, we found that tree models appeared to give an excellent fit to the actual recovery rates.

As part of the modeling process, we tried fitting a Generalized Additive Model with various polynomial expressions to the data, and also tried using a number of variable transformations so as to provide additional flexibility to the fitting process. However, we found that tree models appeared to give a better overall fit to the data.

3.2 Introduction to Tree Models

Numerous references are available describing the structure of tree models. For example, chapter 8 of An Introduction to Statistical Learning (James, Witten, Hastie, & Tibshirani, 2013) provides a detailed overview of constructing a tree model. (Chapter 9 of Elements of Statistical Learning (Hastie, Tibshirani, & Friedman, 2008) is also a good reference, providing some different details.) Essentially the tree modeling technique looks at the dependent variable (in this case, say the recovery rate), and considers how it can segment the data based on the independent variables into two nodes in such a way as to get the greatest separation in the mean recovery rates between the two nodes. This process is then repeated recursively at each newly created node, and the user enters the parameters controlling how many times this process will be repeated.

Specifically, tree models partition the data by considering the equation

$$\text{TSS} = \sum_{i=1}^{N} w_i \, (y_i - \bar{y})^2 \qquad (1)$$

Here N is the total number of records included in the input data, w_i is the weight (exposure) for the i'th record, y_i is the recovery rate for the i'th record and $\bar{y}$ is the weighted mean recovery rate for all the records.

The Total Sum of Squares (TSS) represents the magnitude of how much the individual observations differ from their mean. (Conceptually this is similar to the variance of a random variable.) In order to create a split into two new nodes, the model considers what variable segmentation, or combination of categorical variable classes, will cause the greatest reduction in the TSS metric once the split has occurred. Essentially the model calculates the TSS above for every possible segmentation of the input variables, and segments the data by the variable that creates the greatest reduction. (For example, when constructing a tree for recovery rates, a Duration of 8 months or less vs a Duration of 9 months or more provides the maximum TSS reduction, so the first node splits recovery rates by this criterion. In the case of mortality rates, Cancer claims are the most predictive of higher mortality, so the first tree nodes split the Cancer category from the other disability categories. The software also considers combinations of categories at each split

– not only individual categories - that would produce the greatest TSS reduction.) Hence this split can be based on categorical variables such as Disability Category, or ordinal variables such as Duration. After the data has been segmented, there will be 2 classes of recoveries, with two separate means. The table below shows how segmenting the data between Duration less than or equal to 8 and more than 8 minimizes the TSS by Duration for the recovery rate data set used in this study.

Total Sum of Squares: All Records			37,393
Duration	TSS <=Duration	TSS >Duration	Sum
2	1,251	34,842	36,093
3	2,234	32,890	35,124
4	5,064	26,948	32,012
5	6,917	23,048	29,965
6	8,385	20,511	28,896
7	10,278	18,112	28,390
8	11,813	16,533	28,346
9	13,051	15,418	28,469
10	14,139	14,531	28,670
11	15,087	13,854	28,941

This process is then repeated recursively at each new node of the tree. Various parameters control how complex a tree is created and at which point the TSS reduction for splitting a node is no longer considered worthwhile. Depending on the purpose, one my want more or less complex trees, but one should not develop a tree of such complexity that it has been over-fitted to the data. (For a good discussion if overfitting, see for example Chapter 2 of *An Introduction to Statistical Learning.)*

At each node of the tree, the Predicted recovery rate is simply the weighted mean recovery rate for the actual observations in that node. Note that each data record includes an exposure count, as well as an actual number of recoveries and deaths. The exposures are used as weights for all modeling developed in this report.

3.3 Records with Zero Recoveries

One of the concerns with the GLTD data – even when summarized into less granular categories - is that they may still contain a high percentage of records with zero recoveries. This can be a significant issue for parametric models based on assumed distributions for the recovery rates since there is an overabundance of records with zero recovery rates. Traditional statistical distributions do not typically allow explicitly for such high levels of overabundance. This is discussed in detail in Lally & Hartman, 2016.

This overabundance of zero's is not a concern for tree models. As described above, the model splits the data by looking for the data segmentation that creates the largest possible reduction in the Total Sums of Squares. Because the calculation is weighted by the exposure, the correct weighted mean for each tree

split is calculated. Said differently, if one calculated the mean recovery rate for the newly created tree branch by taking

$$\frac{Total\ Recoveries\ in\ branch}{Total\ Exposure\ for\ the\ branch} \quad (2)$$

then the results of such a calculation would equal the weighted mean calculated by the R routine in determining the split. Hence the presence of a large number of zero-recovery records has no impact on the ability of the R software to correctly determine the optimal variable split.

3.4 Implementation of Tree Models in R

We used the R Package "rpart" to prepare all the modelling modeling described in this report. This package implements most of the functionality of the tree model as developed originally developed in 1984. (Breiman, Friedman, Stone, & Stone, 1984) A full listing of the R package can be found here:

https://cran.r-project.org/web/packages/rpart/rpart.pdf

Section 4: Overview of Data and Initial Modeling Considerations

4.1 Overview of Data

Figure 1 and Figure 2 show recovery rates contrasted against Duration and Age. Clearly both these variables influence the recovery rates. Additionally, we note that the recovery rates for maternity claims show a very different pattern from the other categories. In order to prevent maternity claims dominating one branch of the tree, we have removed the maternity claims from initial modeling of the recovery rates and have modeled those separately.

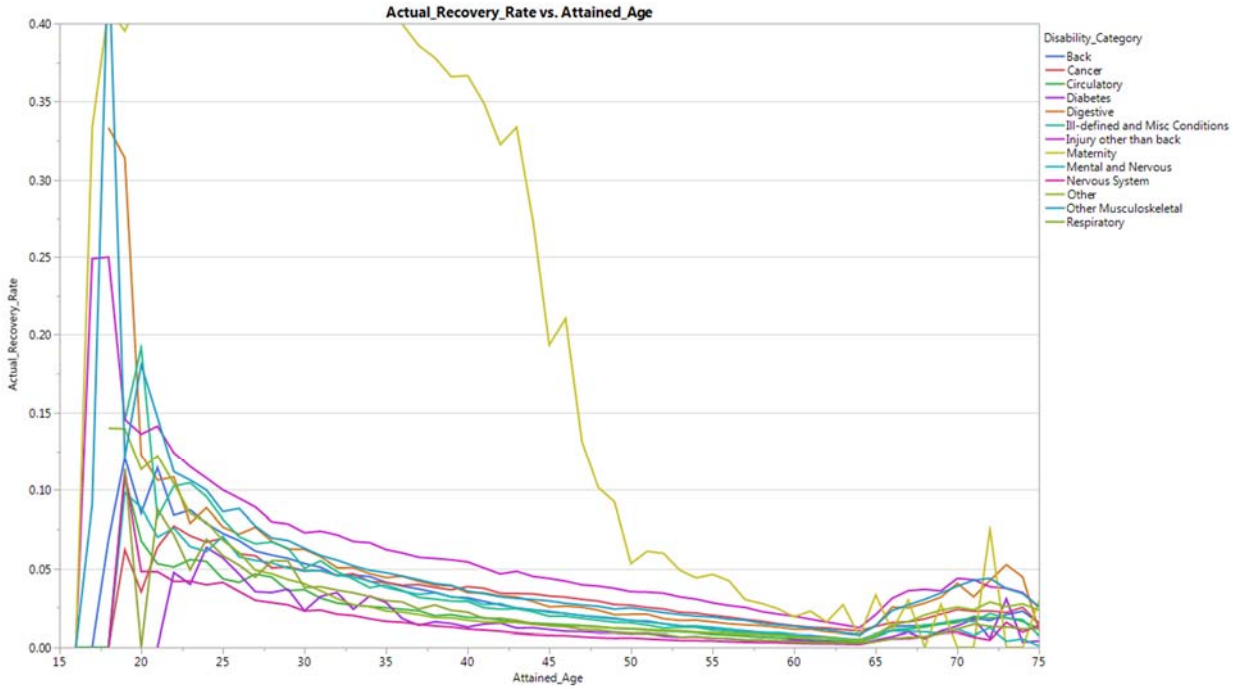

Figure 1. Comparison of recovery rates by Attained Age, including maternities.

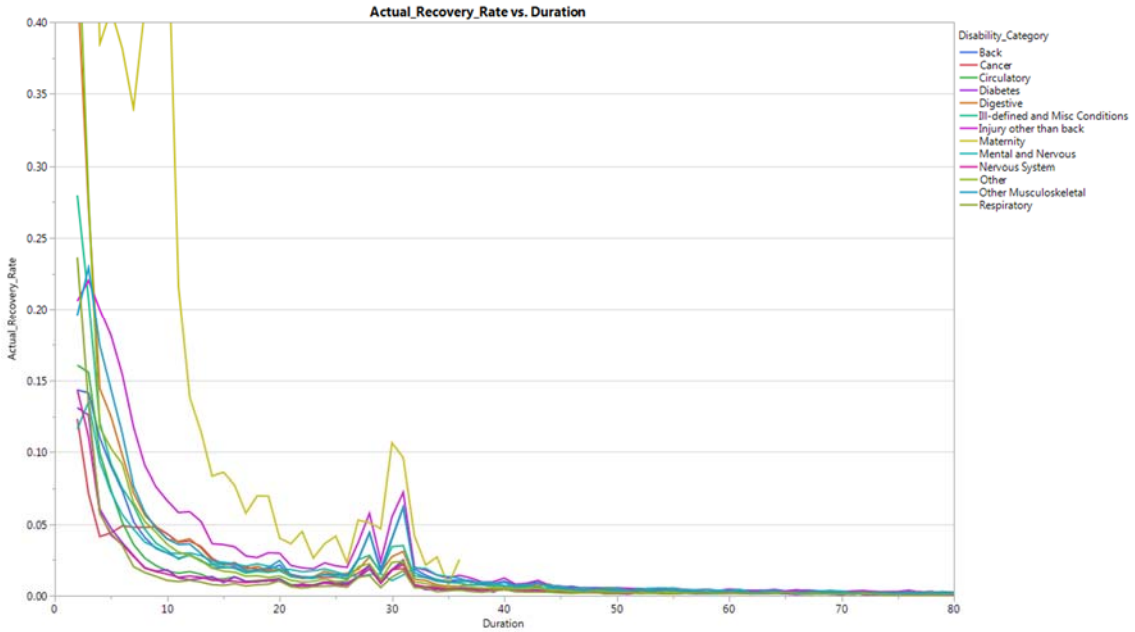

Figure 2. Comparison of recovery rates by Claim Duration, including maternities.

4.2 Data Consolidation and Training vs Test Data

The entire GLTD dataset contains over 46 million records. While the entire file can be read into R, running models against the entire dataset proved to be too unwieldy. So as to speed execution times, we extracted subsets of the data that contained the critical variables needed for the model. This generally led

to files with between 500,000 and 3,000,000 records, and analyses of these files was a lot more manageable and produced good results. These smaller files captured all the exposures and recoveries from the large dataset – they were just summarized sets of the data, similar to multidimensional pivot tables in Excel.

After constructing these summary files, they were further split as follows:

- Maternity records were removed
- A training set for model development was created, using 70% of the data
- A testing set for model testing was created with the remaining 30% of the data

Following standard modeling convention, the models were constructed based only on the training data, without any consideration given to the testing data. The testing data set was then read into the model so that the model could make recovery rate predictions based on the input variables included in the records of the testing data. The models were then tested for goodness of fit only on the testing data. In this way, the model is never tested for fit based on the same data that was used to construct the model. This approach reduces the chances of overfitting the model.

4.3 Testing for Goodness of Fit

At each step of fitting the model, we compared the Mean Square Error (MSE) of the Predicted Recoveries vs the MSE of the Expected Recoveries, using the testing data after fitting the model on the training data. (Recall that Section 2.2 clarified the Predicted vs Expected terminology used here.) This is straightforward to calculate because the Expected Recoveries are already computed and part of the GLTD database.

$$\text{MSE\_Predict} = \frac{1}{N}\sum_{i=1}^{N}(PredictedRecoveries_i - ActualRecoveries_i)^2 \quad (3)$$

$$\text{MSE\_Expected} = \frac{1}{N}\sum_{i=1}^{N}(ExpectedRecoveries_i - ActualRecoveries_i)^2 \quad (4)$$

These Expected Recoveries are based on the 1997 to 2006 data: since the data being analyzed is now more up to date, we would hope that the result of this modeling would give a substantially better fit than the 2008 model.

Data Considerations and Testing the Fit

The consolidated data sets used for modeling vary by size depending on the number of variables (and possible values for those variables). When the number of possible discrete combinations of variables and their possible values gets larger, the Actual, Expected and Predicted Recoveries per combination decreases. Per the MSE equations above, this will by necessity drive a lower MSE between different models.

It would hence be a mistake to focus purely on the MSE value – one could reduce the model MSE purely by generating a larger data set using variables that are not part of the final model, even if the fit was not as good. (Note – the fit of the ultimate model and construction of the tree should not be impacted – it is merely the comparison of different MSE's that can be problematic. This occurs only because we are using

summarized data which is not typical for most statistical applications.) Hence, in comparing different tree models, we also considered the percentage reduction in the MSE calculated from the Predicted values vs the MSE calculated from the Expected's. Generally a model was seen as fitting better based on the percentage reduction achieved by the Predicted MSE vs the Expected MSE, using the test data. Some judgement was needed here, but this appeared to provide a good comparative basis.

We note that it may be valuable to explore some type of normalization of the MSE measures to make the different models more directly comparable. This is an area for future research.

Section 5: Constructing the Model

5.1 Variable Selection

The 2008 report detailed that 7 variables noted above under Background (or derivations from them) were critical variables impacting recovery rates. Additionally, the updated 2004 - 2012 data set includes 7 new variables, namely

1. Region
2. Case Size
3. Integration with Short Term Disability
4. Industry SIC Code
5. Taxability of Benefits
6. Indexed Monthly Salary
7. Company Group

Part of this study reviewed these new variables and how effective they were in improving the model fit. We also relied on the results of others who looked at which of the new variables were correlated with the recovery rates (Davis & Correia, 2017).

In order to aid computation time and file manipulation, some summarization of the variables was needed. For example, Figure 1 suggests that Age in 5 - year age groups should provide sufficient granularity to adequately explain the observed recovery rate. However the steep decline in recovery rates by Duration shown in Figure 2 implies that Duration without any grouping be used at least for the first 40 months. We found that using Duration in integral months up to a maximum of 49 months provided sufficient explanatory control to allow an excellent fit.

Independent Variables Included in the Final Model

A brief summary of the independent variables that are included in the model, as well as explanations for a few exclusions, is provided below.

Duration. In every model we built, Duration was the most powerful predictor of recovery rates. We tried Duration as supplied as well as (Duration – Elimination Period) and providing Elimination Period as a separate variable. Our modeling found that Duration as supplied provided the best fit. Duration varies from 2 to 579, with more than half of the exposures being for durations less than 50.

Age. We found that the model fitted well with ages in quinquennial groups. Age proved to have a smaller impact on the Predicted recovery rates than expected so we don't think that preparing a model using integral ages would provide an improved fit.

Integration with STD. This is one of the new variables in the dataset. A traditional calculation of the recovery rate by the different STD categories clearly shows that the recovery rate is impacted by any STD integration. Similarly an AE analysis, using the 2008 tables as the basis for the Expected's, highlights that this variable adds to the explanatory ability of the model.

Indexed Monthly Salary and Gross Indexed Monthly Amount. The 2008 study found that Gross Indexed Monthly Amount impacted recovery rates. The new variable Gross Indexed Monthly Salary is also clearly related to recovery rates (Davis & Correia, 2017). We tried running the model with each of these variables included separately and in combination. We found that the Gross Indexed Monthly Amount provided the largest improvement to the fit.

Disability Category. The recovery rates in the above charts show that this is clearly a significant variable in determining recovery rates. We did not try to combine factors *a priori* – the rpart package automatically looks at all combinations of categorical variables to determine the combination having the greatest impact on the sum of square errors.

Own to Any Occupation Transition. An analysis of recovery rates by the Own Occ variable showed that the greatest recovery rates were at OwnOcc + 0 and OwnOcc + 1. Accordingly we used these two values, and set all the others to the same value.

Gender. Gender was included in the model, and reviewing predicted recovery rates from the model shows a good fit between Predicted and Actual recovery rates when split by Gender. However, it should be noted that Gender played a very minor role in the fitting process. (See Variable Importance below.)

Variables Excluded from the Final Model

The following variables did not significantly improve the predictions of the model so have not been included in the final model produced.

Taxability of Benefits. An analysis similar to that described under Integration with STD shows that this variable does help predict recovery rates. However, the highest recovery rates were associated with an "Unknown" taxability category, while the recovery rates for the other categories were overall roughly similar. Hence, this variable appears to have limited practical use, so has been excluded.

Elimination Period. Elimination Period is included into the model by proxy, but not directly. By including Duration, there is some implicit recognition of the Elimination Period. (The Duration in the data starts one month after the Elimination Period, so a claim with a 6 month Elimination Period would start with a Duration of 7). Additionally, the inclusion of the Own to Any Occupation variable recognizes the Duration since the end of the Elimination Period. As part of our modeling we constructed a variable Duration Since Elimination Period, but this variable provided a poorer fit to the data than using the supplied value of Duration.

Region, Case Size, Industry, Company Group. The study by Davis & Correia (2017) showed some relationship between some of these variables and the actual recovery rates. However, various attempts to include these variables in the modeling did not materially improve the fit. Essentially the other variables included in the model explained such a large part of the variations that these variables were unable to meaninfully provide additional explanatory value. In addition to modeling Industry using the provided SIC code, we also mapped SIC code to Collar color, but did not find an improved fit as a result of this mapping. (Note that Collar did improve the fit for the mortality rate development shown later.)

5.2 Building the Tree

We setup the basic tree model parameters and then ran multiple iterations to increase the depth of the tree and determine whether the increased depth provided a meaningful reduction in the MSE. Sample R code for the models is shown in Appendix A.

We show the first 3 stages of building a tree model below. The MSE when comparing the Actual recovery rates for each summarized record in the data against the Expected's is 2.15 (Equation 4 above).To calculated the MSE for the predicted number of recoveries prior to creating any splits in the tree, we note that the mean recovery rate for all records in the database is 1.5% (shown in the root node of figure 3). If we set the Predicted recovery rate for every record equal to 1.5% times the exposure in the record and calculate the MSE, we get that the Predicted MSE for the entire dataset is 85.16 (Equation 3 above).

Figure 3 below shows the first split of the node – as shown, the variable that causes the greatest reduction in the TSS is Duration. Specifically, splitting the data into durations less than or equal to 8 months and greater than 8 months causes the greatest TSS reduction. Given the rapid reduction in recovery rates with increasing duration, this result is not surprising. The information below shows that 92% of the exposures have duration 9 or greater, and that the recovery rate for those records is 0.95%. For the records with duration 8 or less, the recovery rate is 7.6%.

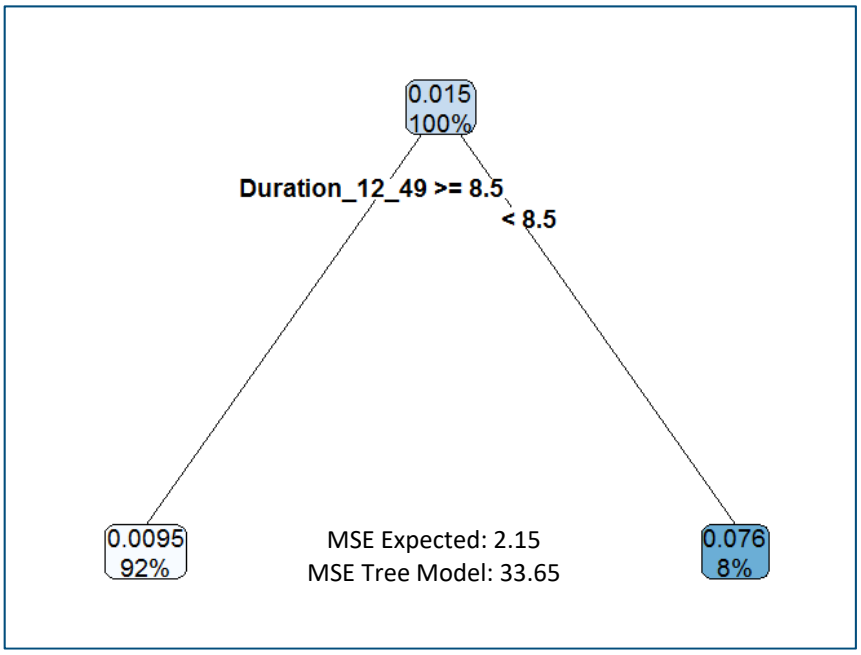

Figure 3. Tree model showing only the first node split

Figure 4 illustrates how the model branches are created when a more complex tree is built. (The control for the complexity of the tree is discussed below in rpart Parameters.) As noted, a relatively simple tree has dramatically reduced the Mean Square Error (MSE). Most of the data variance has now been explained by creating only 10 leaves (or ending nodes). Note that the left-most leaf, for Durations greater than 32, still contains 58% of the exposures. The observations in that node have a recovery rate of only 0.25% versus a recovery rate of 27% shown for the leaf on the right hand side. (This leaf corresponds to durations less than 3.5 and disability categories of Digestive, Injury other than Back, Other and Other Musculoskeletal.)

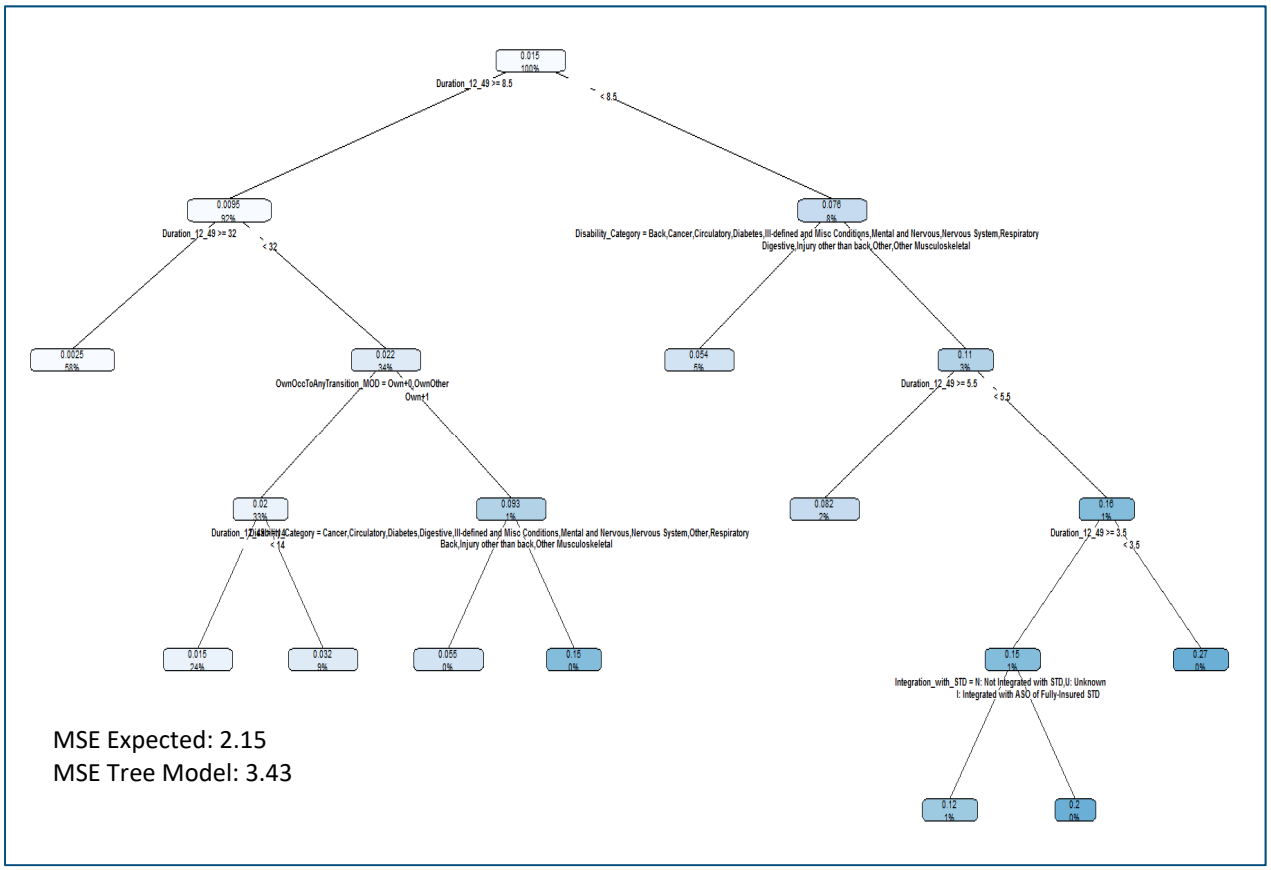

Figure 4. Development of a more complex tree with 10 leaves.

Continuing this process, we developed a very complex tree, as shown in Figure 5. This is the final model we created before some minor pruning of the tree, as described below. The R code used to develop this tree, and do the pruning, is shown in Appendix A. Developing a more complex tree than shown resulted in an increase in the MSE calculated for the testing data – a clear sign that overfitting was becoming a concern.

While interpretations off this tree are impossible given the complexity and difficulty of showing the labels, we believe that the less detailed tree shown in Figure 4 provides more useful information for inference purposes. Amongst the interesting observations from Figure 4, are that Age and Gender have not yet provided significant enough information to be included in the split. (This does not mean Age and Gender don't have predictive utility – it means that other variables are more important, and have potentially already explained some of the impact of Age and Gender. This is discussed below in Confounding and Colinearity.)

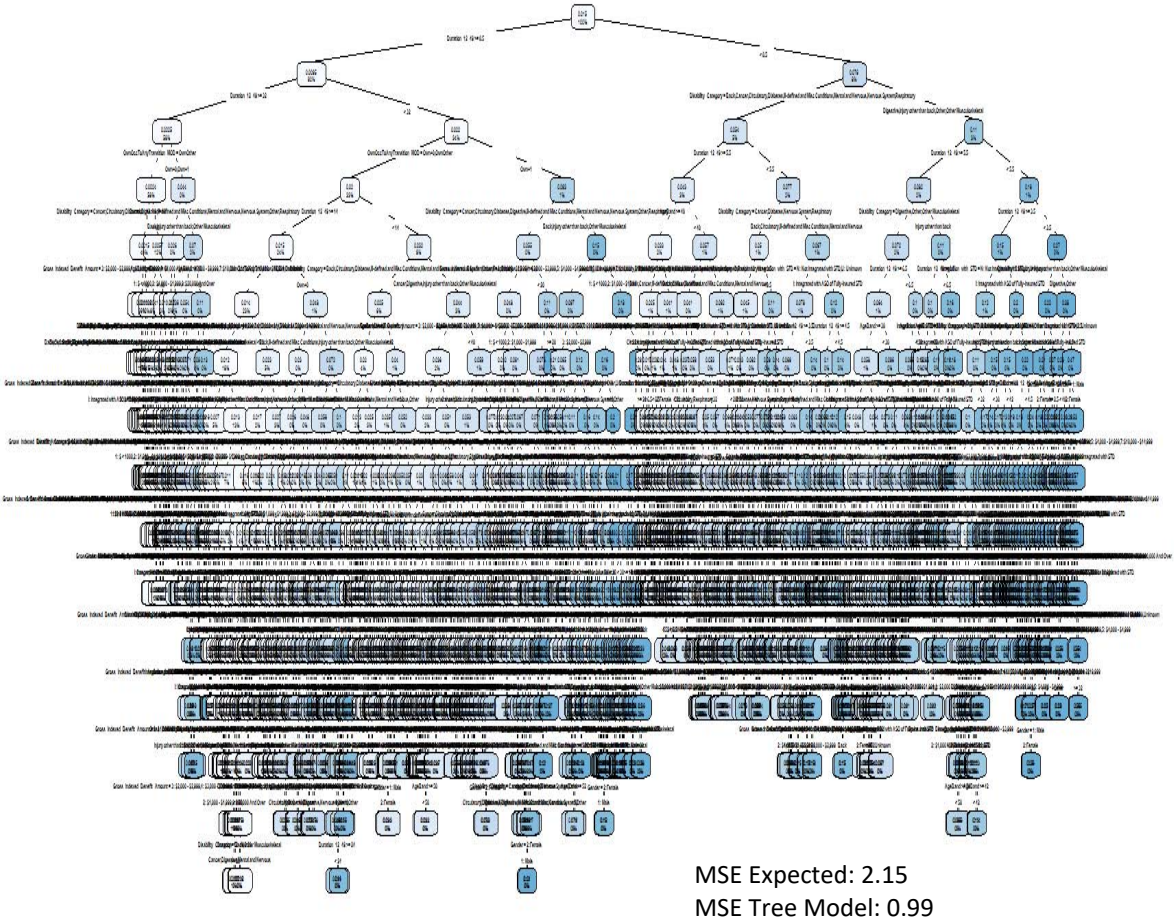

Figure 5. Complex tree before clear signs of overfitting begin to appear.

Figures 6 - 8 below demonstrate a few different views, including splits by Gender and Age Band, of the final tree model. The graphs show the fit to the testing data, after building the model on the training data. As shown, the Predicted recovery rates fit very closely to the actuals.

In the 3 figures below, and similar to Figure 2, there is a noticeable spike in recoveries at 27 to 28 months, and 30 to 31 months. This arises in the actual recovery rates since most policies have an elimination period of 3 months or 6 months, combined with an Own Occupation to Any Occupation limitation of 24 months. Accordingly, the Own to Any Occupation limitation begins to apply in most cases at 27 months and 30 months. As noted above under Variable Selection, most recovery rates spike at the transition month or the month after – i.e. months 27 and 28, or months 30 and 31.

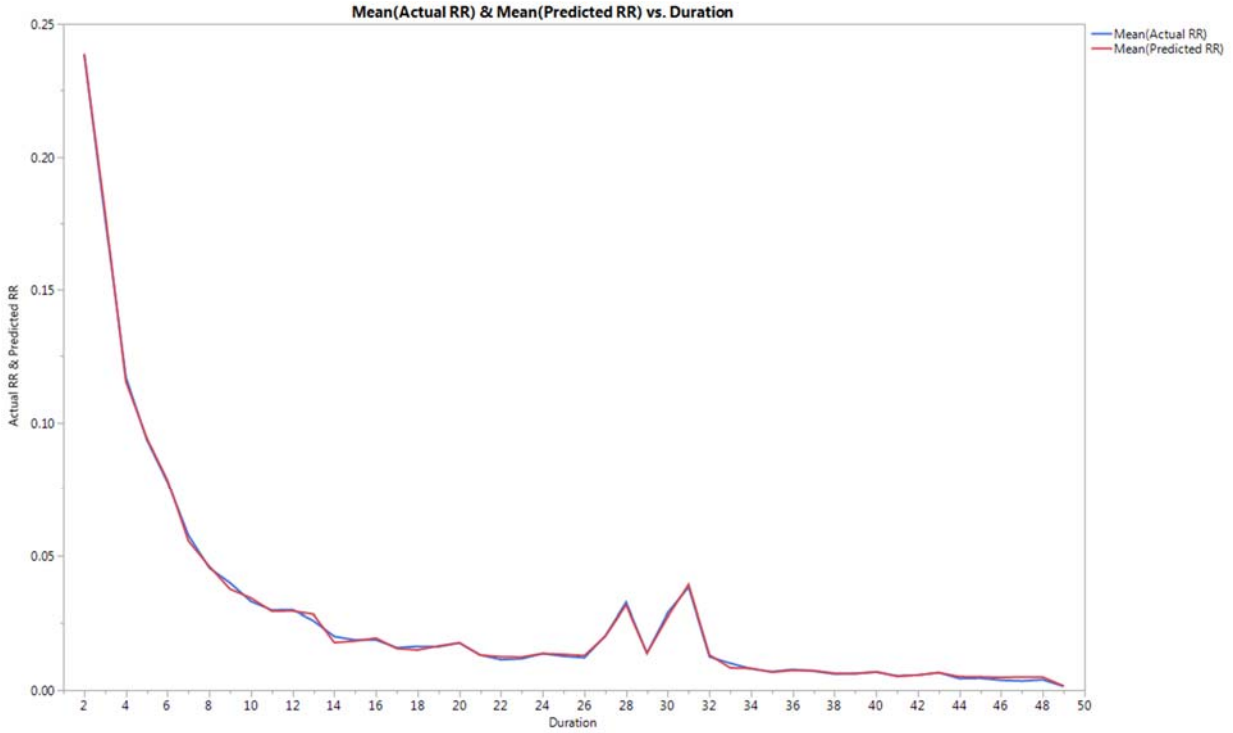

Figure 6. Comparison of Actual and Predicted recovery rates – testing data.

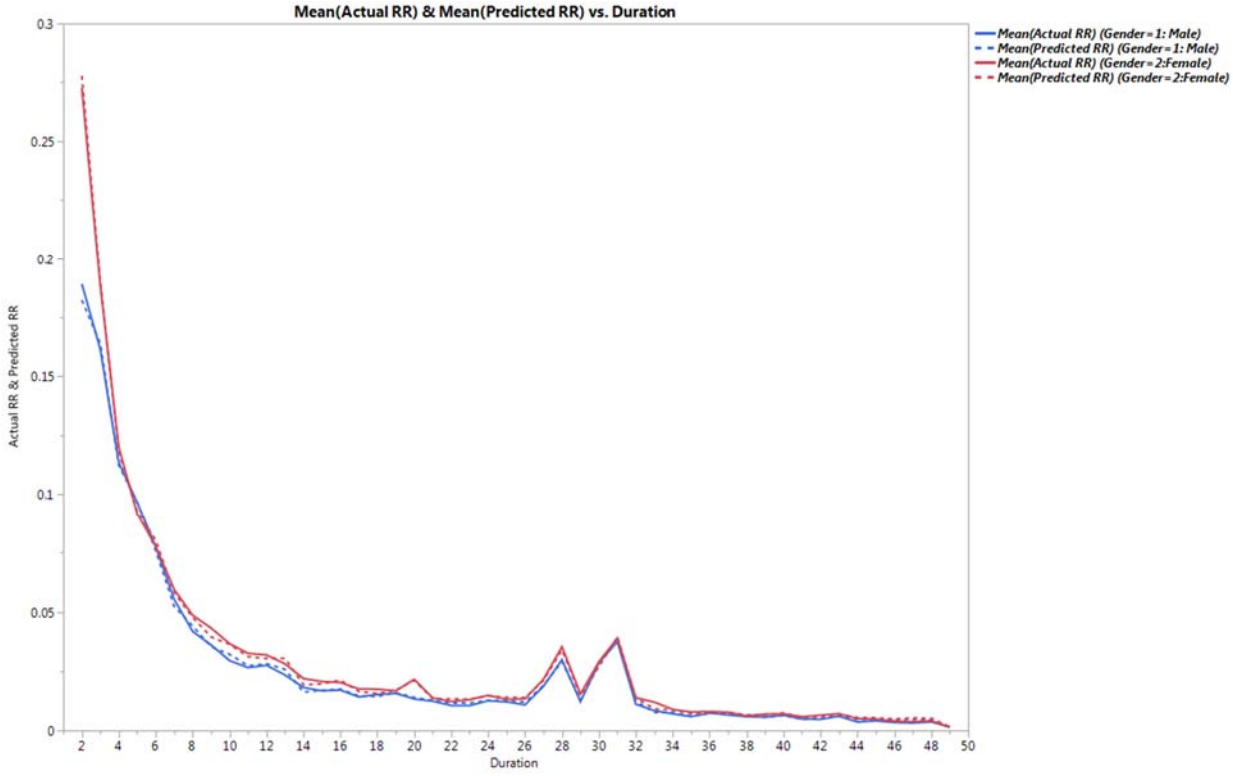

Figure 7. Comparison of Actual and Predicted recovery rates split by Gender – testing data

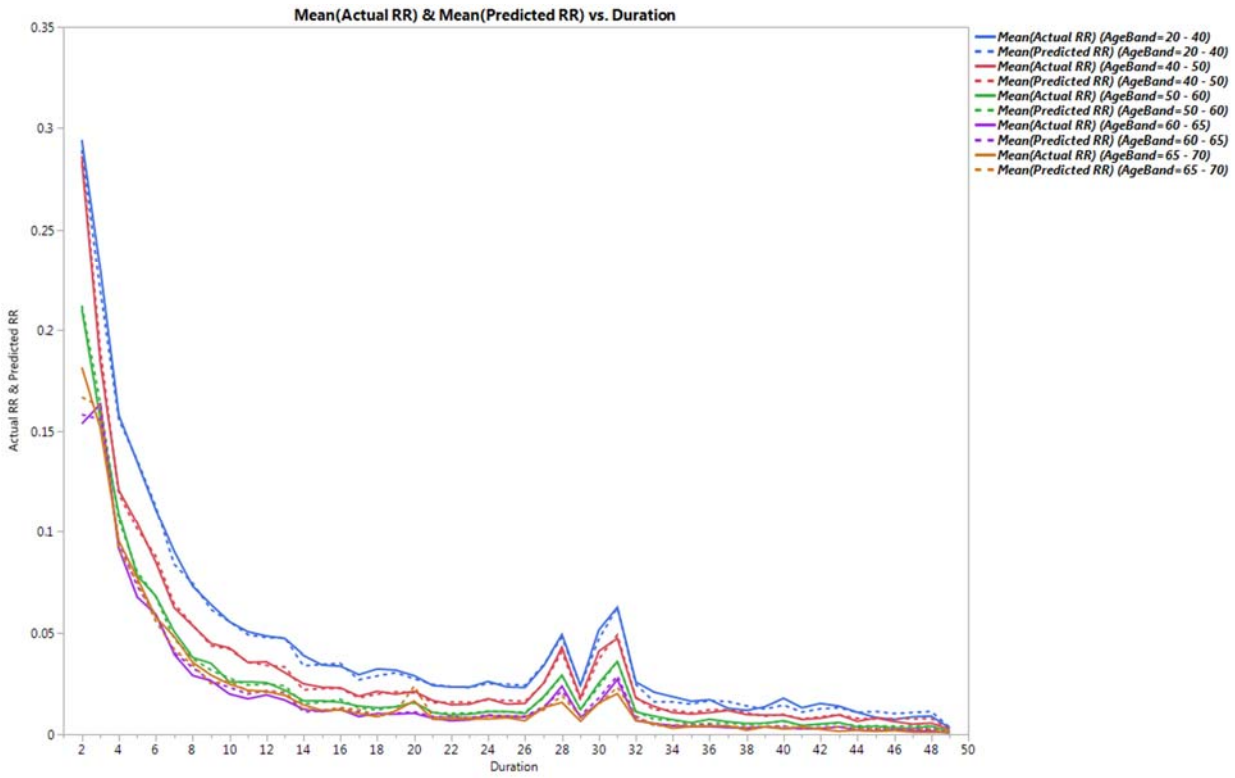

Figure 8. Comparison of Actual and Predicted recovery rates split by Age Band – testing data

5.3 rpart Parameters

rpart Has a number of parameters that control the structure of the tree. The key parameters are:

- *maxdepth:* This parameter sets the maximum depth of any node of the tree, with the root node counting as zero.
- *minsplit:* the minimum number of observations for the software to attempt to split any node
- *complexity parameter (cp):* A split will only occur if the fit improves by the specified factor. In the models described here, the value of R Squared must have increased by at least cp in order for the split to occur.

In practice we found that the last parameter was the one that had the greatest impact on the fitted model. Relatively small changes in the cp parameter would results in substantial changes in the MSE of the final fit. Accordingly, it took some trial and error to settle on the value of cp so that it decreased the MSE noticeably, whereas reducing the cp further had little impact (or possibly increased) the MSE using the testing data. (An increase in the MSE on the testing data, which happened with smaller values of cp, indicates overfitting to the training data.)

5.3 Variable Importance

rpart Provides information about the variable importance – it shows, as a percentage, how much each variable contributed to the reduction in the MSE. For the final model, the variable importance was as follows:

Variable	Importance (%)
Duration	63
Disability Category	17
Age Band	6
Own to Any Occ Transition	6
STD Integration	5
Gross Indexed Benefit Amount	2
Gender	1

The above results are somewhat surprising, given the traditional actuarial approach of starting analyses based on Duration, Age and Gender.

Confounding and Collinearity

Figure 7 shows the Actual and Predicted results split by Gender so as to clarify a common interpretation difficulty with models in general – not just tree models. The Gender importance of 1% needs to be interpreted with care (as would a similar result with a GLM or other model). From Figure 7, it's clear that the Predicted and Actual rates by Gender show material differences. When constructing a mathematical model on data such as this, it's always possible for one variable to mask the impact of another. So Gender differences in Disability Category, Gross Indexed Benefit Amount etc. may already influence the Predicted recovery rates, so that the inclusion of Gender directly as a variable only explains a small residual amount of difference. In some situations, this can be due to variable confounding, and in other situations, it can be due to variable collinearity. Both of these create difficulty for the modeler, and require extra care when interpreting results.

5.4 Pruning to Avoid Overfitting

Once a tree had been produced, we pruned the tree back, again using the rpart routine. Pruning is used to avoid overfitting, while still providing almost the same level of fit as the original tree. The idea behind pruning is discussed in James, Witten, Hastie, & Tibshirani, 2013.

Given the complex nature of the data that is being modeled, pruning did not significantly simplify the tree beyond the one shown in Figure 5. However, since the pruning should reduce any overfitting of the model, we consider this to be superior to the unpruned tree. The Mean square error did not change meaningfully from that shown above.

5.5 Reviewing Results for Reasonability and Artifacts

Another way of looking at the reasonability of the results is to consider which records in the testing data set have the largest deviations between the Actual number of recoveries and the Predicted number of recoveries. These largest deviations would be expected to occur in records with higher exposures and hence higher numbers of Predicted, Expected and Actual recoveries. For comparative purposes, we again compare the Predicted model deviances with the Expected deviances shown in the data.

Figures 9 shows, for each Duration modeled, a comparison over all variable combinations for that duration of the maximum values of:

- |Predicted Recoveries – Actual Recoveries| and
- |Expected Recoveries – Actual Recoveries|

(This is the same as showing the square root of the maximum square error.) These comparisons are shown for the testing data, after building the model on the training data.

The graph indicates that the Predicted recoveries appear to avoid significant prediction errors across all durations, with no substantial shortcomings when compared against the Expected recoveries. This same picture emerges if any of the other input variables are used on the X axis instead of Duration.

The 2008 model was constructed from older data, so it is not surprising that there are some larger variances here versus the tree model which was constructed using the newer data (albeit a training set, but still constructed from the same data.) The robustness of the tree method does make intuitive sense - since the Predicted value is the mean of the observations in the applicable node, and by construction these nodes must have a minimum number of observations, the possibility of extreme values being predicted is virtually eliminated.

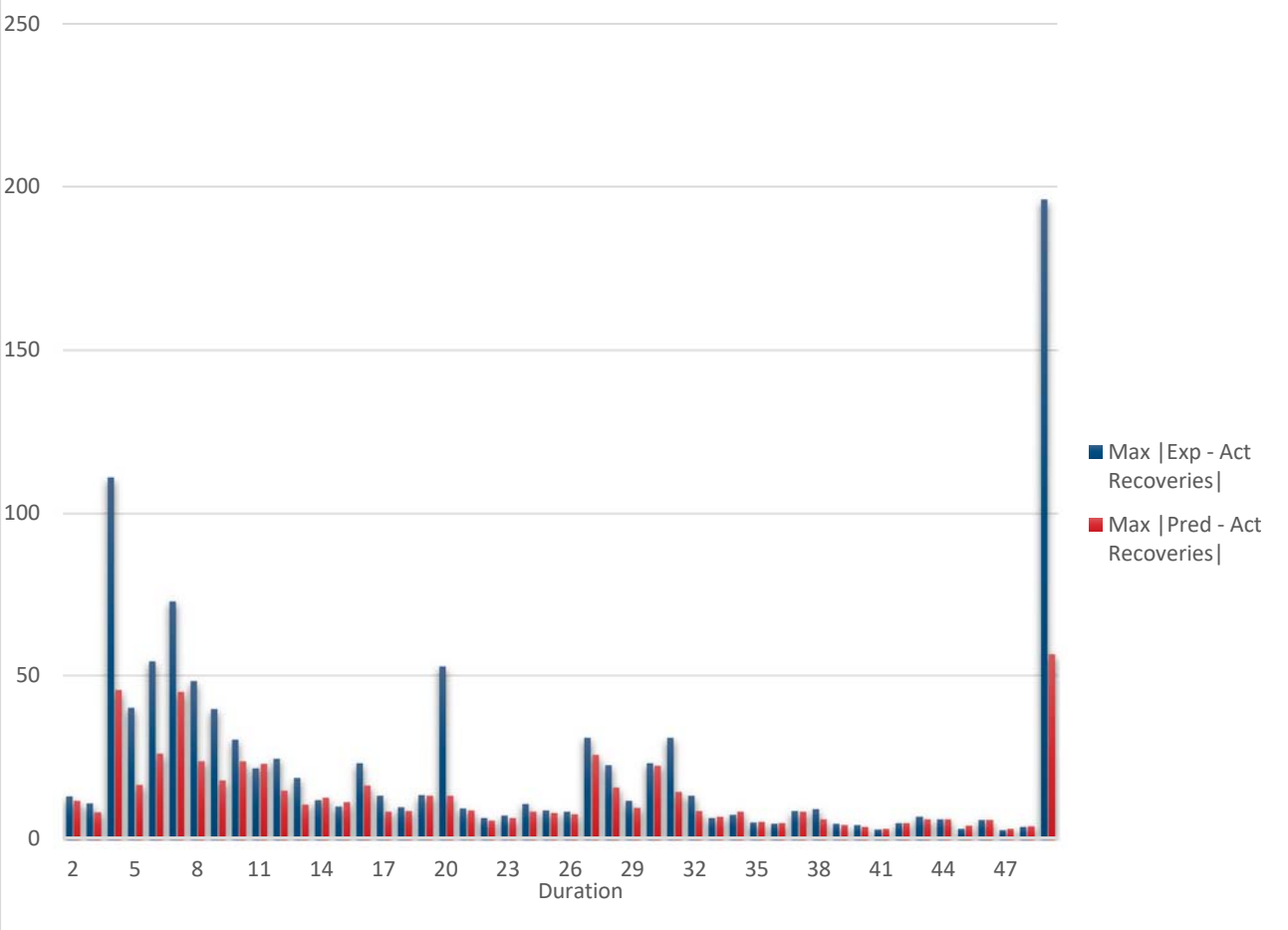

Figure 9. Comparison of testing data differences between Expected (blue) and Predicted (red) recoveries, by Duration.

5.6 Reproducing Results

Appendix A contains the R code and links to the input data files that were used to generate the results included in this report. By reading in the input files provided and running the R code in a suitable R interface such as RStudio, those interested should be able to duplicate our results.

Section 6: Extracting Recovery Rates for Valuation or Other Use

6.1 Extracting Decrement Rates from the Model

In order for the tree model to have a more direct practical application, it needs to be able to provide rates that can be used in other systems. Because there are 7 independent variables, there are a multitude of possible variable combinations that impact the recovery rates. Each recovery rate is a function of the 7 variables discussed above under "Variables Included".

In order to extract these rates, we developed a CSV file that included all possible combinations of these variables. The code used to develop this CSV file is shown in Appendix A.4. It is then straightforward to read this CSV file into R and have R calculate the Predicted recovery rate for every set of variable combinations.

Figures 10 and 11 below show the recovery rates for a 12 sample individuals (one for each Disability Category), assuming they became disabled at age 20. The graphs appear stable and free of strange artifacts. This is not surprising given that the model always regresses to the mean in those areas where there are fewer exposures. No graduation has been applied to these rates.

6.2 Graduation

There are numerous situations where an actuary may prefer to have graduated decrement rates. As would be the case with any model for developing decrement rates, raw rates would first be generated and then a suitable graduation technique applied. The tree model described here has developed these raw rates; if graduation is desired, a suitable graduation technique could be applied to these rates.

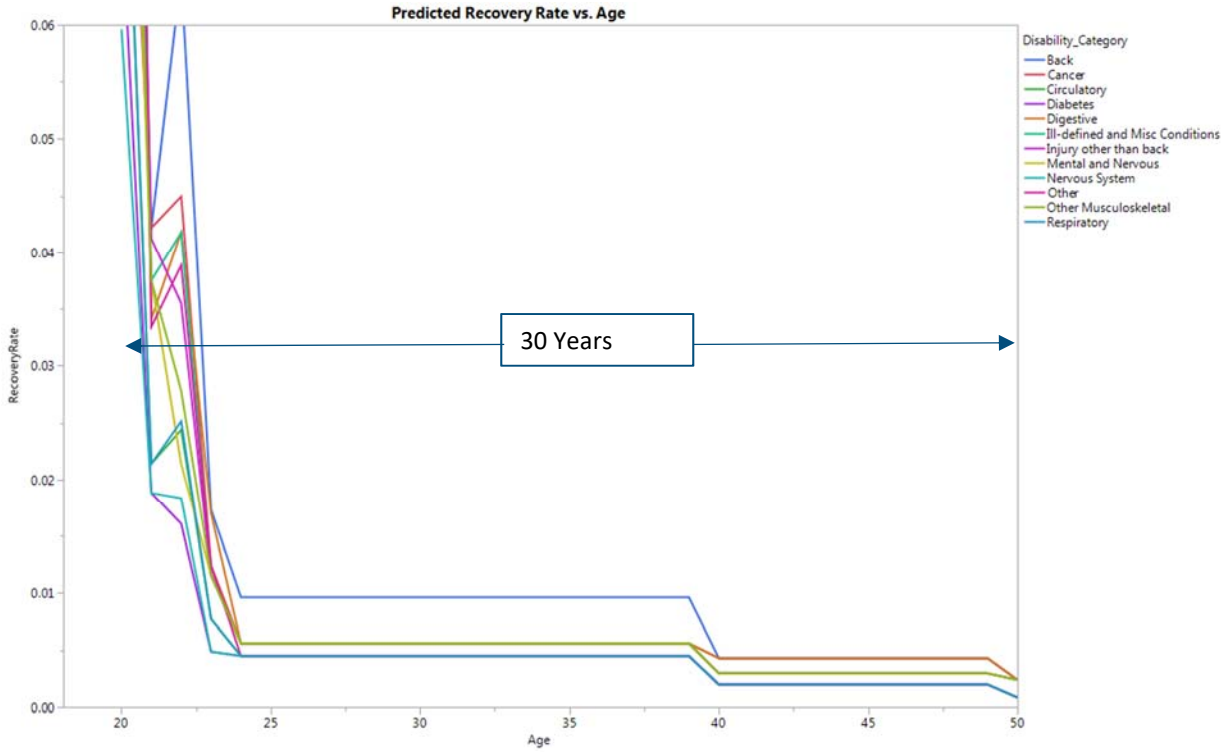

Figure 10. Recovery rates for 12 sample individuals who become disabled at age 20

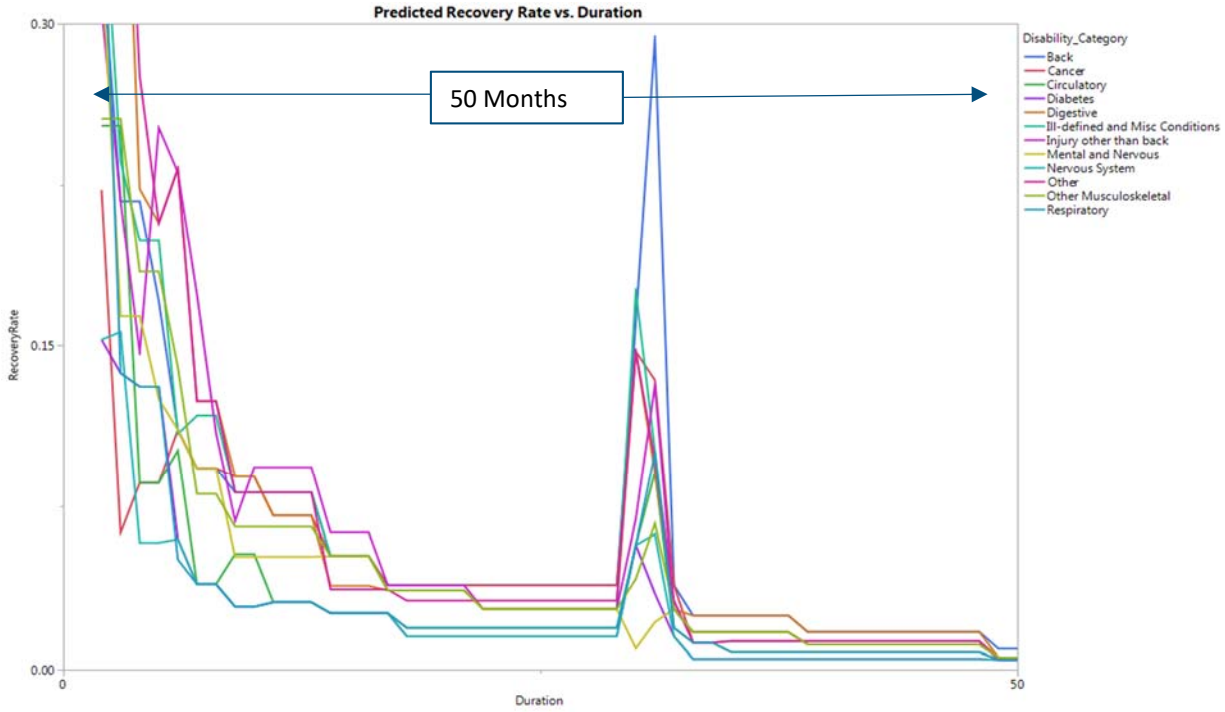

Figure 11. Same as figure 10, but focus on the first 50 months

Section 7: Maternity and Mortality Models

7.1 Maternity Recovery Rates

We built a new recovery rate tree model, honing in only on maternity claims. The process was as described above for recovery rates, but including only maternity claims in the data.

We found the following percentage measures of variable importance:

Variable	Importance (%)
Duration	76
Age Band	12
Integration with STD	6
Gross Indexed Benefit Amount	5
Own Occ to Any Occ	2

The MSE for this model was 76.7 vs 122.2 for the Expected's. (As noted in Section 4.3, these much larger MSE's than for the recovery model for the other categories does not necessarily mean the fit is worse. The maternity category has more exposures per record (more concentration in age bands and durations) so that the error per summarized record is larger, but the fit is not necessarily better or worse.)

We note too that given the substantially smaller number of exposures than when all disability categories are combined, the model is more susceptible to smaller changes in the rpart parameters. We could get some better and worse fits to the training data by adjusting some of the parameters; we show here a model that produced middle-of-the-road fits.

7.2 Mortality Rates

We reviewed mortality rates by Age and Duration, and it was clear that these variables affect mortality rates differently than recovery rates. Based on some preliminary modeling and analysis, we found that the following variables, and groupings of those variables, appeared to give the best fit to the observed mortality rates:

- Duration in integral months to 84, then in increments of 12 to a max of 264 (This was the approach previously used for the 2008 Study).
- Ages in 5-year increments from 35 to 60, then integral ages thereafter. This choice was made based on the steepening mortality curve after age 60.
- Collar mapped according to SIC codes, per standard industry mappings.

We found the following percentage measures of variable importance:

Variable	Importance (%)
Disability_Category	73
Duration_12	20
AgeBand_3060	3
Gender	2
Collar	2

The MSE for this model was 0.284 vs 0.413 for the Expected's, so this new model demonstrates a meaningful improvement in fit.

The final pruned tree is again complex. However, we thought it would be instructive to illustrate how far back we would need to prune the complex tree so as to get the MSE for the Predicted values (0.403) to approximate that of the Expected values (0.413). (Once can adjust the level of pruning by altering the same cp parameter previously discussed.) We find it interesting and counter-intuitive that such a simple tree, with only 12 leaves, can produce a MSE comparable with that of a much more complex table of mortality rates. (D_12 refers to duration and AB_3 to age band).

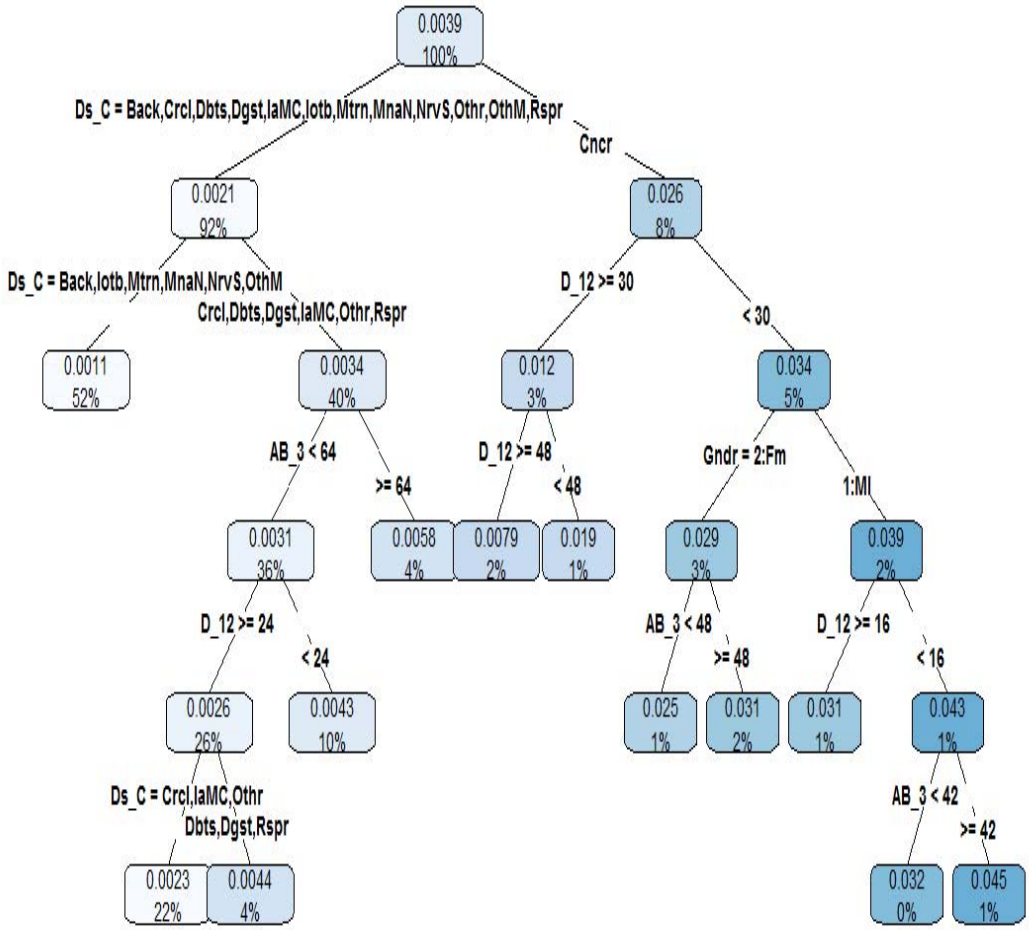

Figure 12. Mortality rate tree pruned back to the point where the Predicted MSE (0.403) approximates the Expected MSE (0.413). This is not what we consider the final model – illustrative only of how a relatively simple tree has surprisingly powerful predictive ability.

Section 8: Addendum and Future Work

8.1 A Note on Random Forests

While preparing this study, we made numerous attempts to improve upon the results using Random Forests. Their reputation suggests that they should be better at predictions than regular tree models. (See for example Section 9.3 of Breiman, Statistical Modeling: The Two Cultures, 2001.) However, for this purpose, we found that the Random Forest predictions did not fit the test data as well as the tree model. We believe this is because we are working from a very large, robust set of data, with the number of records about million-fold as great as the number of variables included in the data. As such, the randomness of the Forest model is not needed to build a better predictor. Random Forests excel where there are a large number of predictor variables with fewer observations. However, in this case there are few predictor variables and comparatively many observations.

8.2 Areas for Additional Research

We believe the following areas could provide useful additional research:

- Stratify the data by calendar year and build a model on the earlier year's data and apply it to the later data (as opposed to a purely random selection of data, spread across all years). Initial work in this area indicates that the improvement in the Predicted MSE to the Expected MSE is not as large as shown in the above models, but still substantial.

- Develop a methodology to normalize the MSE comparators so that different models with different numbers of possible variable values / combinations will be able to be compared directly

- Application of neural network algorithms to predict recovery / mortality rates

- Application of Gradient Boosting Machine algorithms

- Application of tree models to Actual to Expected rate analyses to highlight areas of maximum divergence between the Actual's and Expected's

References

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*.

Breiman, L., Friedman, J., Stone, C. J., & Stone, C. J. (1984). *Classification and Regression Trees.* Boca Raton, FL: CRC.

Davis, K., & Correia, P. (2017, April 20). *GLTD 2016 Termination Study: Introduction to New Variables.* Retrieved from www.soa.org: https://www.soa.org/prof-dev/events/2017-gltd-intro-variables/

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning.* Stanford: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.

Appendix A

The data required to run the R code below has been published together with this report. Two datasets are provide – one for the recovery and maternity models (A.1 and A.2) and a separate input file for the mortality model (A.3).

A.1 R Code to Construct Recovery Model (without Maternities)

```
library(rpart.plot)
library(rpart)

gltdWithMaternity=read.csv("filename&path")   #Read data extracted from Appendix A.1 above
maternityYN<-as.logical(ifelse(gltdWithMaternity$Disability_Category=="Maternity",0,1))
gltd<-gltdWithMaternity[maternityYN,]  # Exclude Maternity records

# Select Training Data on 70% of data
set.seed(1)
Train<-sample(nrow(gltd),floor(0.7* nrow(gltd)),replace=FALSE) #Setup vector with random #'s representing 70% of gltd data
gltdTrain<-gltd[Train,]
gltdTest <-gltd[-Train,]

# Build model
rpart.fit<-rpart(Actual_Recovery_Rate
      ~Disability_Category+AgeBand+Duration_12_49+OwnOccToAnyTransition_MOD+Integration_with_STD+Gender+Gross_Ind
      exed_Benefit_Amount, data=gltdTrain,maxdepth=15, minsplit=110,cp=0.0000001, weights =
      gltdTrain$Exposures,method="anova")

# Prune back model to minimize cross-validation errors
fit.pruned<- prune(rpart.fit, cp = rpart.fit$cptable[which.min(rpart.fit$cptable[,"xerror"]),"CP"] )

# Write out details of R model to text file
sink(file=" filename&path .txt")
fit.pruned
print(" ")
summary(fit.pruned)
sink()


#Use rpart.plot to create tree plot
rpart.plot(fit.pruned, extra=100,cex=0.8,type=4,branch=0,fallen.leaves=FALSE,tweak=0.6,varlen=3, faclen=2,
space=0,compress=FALSE,ycompress=FALSE)

# Test model fit – calculated predicted recoveries in test data based on predicted recovery rates
PredictRpartOut<-predict(fit.pruned,newdata=gltdTest)
ExpectedRR<-gltdTest$Expected_Recoveries / gltdTest$Exposures
PredictedRR<-PredictRpartOut
PredictedRecoveries<-PredictRpartOut * gltdTest$Exposures
Model1_1Predictions_Pruned<-cbind(gltdTest,ExpectedRR,PredictedRR,PredictedRecoveries)

write.csv(Model1_1Predictions_Pruned ,file="filename&path ")   #Write output to view in graphics package

#Calculate and print out expected and predicted Mean Square Errors
MSE_E<- sum((Model1_1Predictions_Pruned$Expected_Recoveries-Model1_1Predictions_Pruned$Actual_Recoveries)^2) /
      nrow(Model1_1Predictions_Pruned)
```

```
MSE_P<- sum((Model1_1Predictions_Pruned$PredictedRR * gltdTest$Exposures -
      Model1_1Predictions_Pruned$Actual_Recoveries)^2) / nrow(Model1_1Predictions_Pruned)
MSE_E
MSE_P
```

A.2 R Code to Construct Maternity Recovery Model

```
library(rpart.plot)
library(rpart)

gltdWithMaternity=read.csv("filename&path")#Read data extracted from Appendix A.1 above

maternityYN<-as.logical(ifelse(gltdWithMaternity$Disability_Category=="Maternity",1,0))
gltd<-gltdWithMaternity[maternityYN,]   # Include Maternity records only  #[maternityYN,]
maleYN<-as.logical(ifelse(gltd$Gender=="1: Male",0,1))  # Exclude Male records – a few questionable data records
gltd<-gltd[maleYN,]

# Select Training Data on 70% of data
set.seed(1)
Train<-sample(nrow(gltd),floor(0.7* nrow(gltd)),replace=FALSE) #Setup vector with random #'s representing 70% of gltd data
gltdTrain<-gltd[Train,]
gltdTest <-gltd[-Train,]

rpart.fit<-rpart(Actual_Recovery_Rate
      ~Disability_Category+AgeBand+Duration_12_49+OwnOccToAnyTransition_MOD+Integration_with_STD+Gross_Indexed_Be
      nefit_Amount, data=gltdTrain,maxdepth=15, minsplit=100,cp=0.00001, weights = gltdTrain$Exposures,method="anova")

fit.pruned<- prune(rpart.fit, cp = rpart.fit$cptable[which.min(rpart.fit$cptable[,"xerror"]),"CP"] )

# Write out details of R model to text file
sink(file=" filename&path .txt")
fit.pruned
print(" ")
summary(fit.pruned)
sink()

rpart.plot(fit.pruned,extra=100,cex=0.8,type=4,branch=0,fallen.leaves=FALSE,tweak=0.6,space=0,compress=FALSE,ycompress=F
ALSE)

# Calculate and print out expected and predicted Mean Square Errors
PredictRpartOut<-predict(fit.pruned,newdata=gltdTest)
ExpectedRR<-gltdTest$Expected_Recoveries / gltdTest$Exposures
PredictedRR<-PredictRpartOut
PredictedRecoveries<-PredictRpartOut * gltdTest$Exposures
Model1_1Predictions_Maternity<-cbind(gltdTest,ExpectedRR,PredictedRR,PredictedRecoveries)
write.csv(Model1_1Predictions_Maternity ,file="filename&path ")
MSE_E<- sum((Model1_1Predictions_Maternity$Expected_Recoveries-Model1_1Predictions_Maternity$Actual_Recoveries)^2) /
      nrow(Model1_1Predictions_Maternity)
MSE_P<- sum((Model1_1Predictions_Maternity$PredictedRR * gltdTest$Exposures -
      Model1_1Predictions_Maternity$Actual_Recoveries)^2) / nrow(Model1_1Predictions_Maternity)
MSE_E
MSE_P
```

A.3 R Code to Construct Mortality Model

```
library(rpart.plot)
library(rpart)

gltdWithMaternity=read.csv("filename&path")  # Read in data extracted in A.4 above
gltdWithMaternity$Actual_Death_Rate<-gltdWithMaternity$Actual_Deaths / gltdWithMaternity$Exposures

# Select Training Data on 70% of data
set.seed(1)
Train<-sample(nrow(gltdWithMaternity),floor(0.7* nrow(gltdWithMaternity)),replace=FALSE) #Setup vector with random #'s
      representing 70% of gltd data
gltdTrain<-gltdWithMaternity[Train,]
gltdTest <-gltdWithMaternity[-Train,]

# Fit tree model
rpart.fit<-rpart(Actual_Death_Rate
      ~Disability_Category+AgeBand_3060+Duration_12+Gender+Collar+Indexed_Monthly_Salary_Mod,
      data=gltdTrain,maxdepth=15,minsplit=120,cp=0.00001,weights = gltdTrain$Exposures,method="anova")

#Prune back model to minimize cross validation errors
fit.pruned<- prune(rpart.fit, cp = rpart.fit$cptable[which.min(rpart.fit$cptable[,"xerror"]),"CP"] )

# Write out details of R model to text file
sink(file=" filename&path .txt")
fit.pruned
print(" ")
summary(fit.pruned)
sink()

# Plot tree
rpart.plot(fit.pruned,extra=100,cex=0.9,type=4,branch=0,fallen.leaves=FALSE, varlen= 2, faclen = 2,
      tweak=0.6,space=0,compress=FALSE,ycompress=FALSE)

#Calculate and print out expected and predicted Mean Square Errors
PredictedDR<-predict(fit.pruned,newdata=gltdTest)
PredictedDeaths<-PredictedDR * gltdTest$Exposures
ExpectedDR<-gltdTest$Expected_Deaths/gltdTest$Exposures
Model3_1Predictions_Pruned<-cbind(gltdTest,PredictedDR,PredictedDeaths,ExpectedDR)
write.csv(Model3_1Predictions_Pruned ,file="filename&path")
MSE_E<- sum((Model3_1Predictions_Pruned$Expected_Deaths-Model3_1Predictions_Pruned$Actual_Deaths)^2) /
      nrow(Model3_1Predictions_Pruned)
MSE_P<- sum((Model3_1Predictions_Pruned$PredictedDeaths - Model3_1Predictions_Pruned$Actual_Deaths)^2) /
      nrow(Model3_1Predictions_Pruned)
MSE_E
MSE_P
```

A.4 Create CSV File for all Possible Variable Combinations for Recovery Model

Note: This code runs very slowly – expect to leave your machine running overnight.

```
Counter=0
MinDuration=1
MaxDuration=49
```

```
#Variable definitions below exactly match original database;  “1: Male” and “2:Female” are correct, not typo’s
TableOut<-data.frame(matrix("",ncol=7),stringsAsFactors = FALSE)
Gender<-c("1: Male","2:Female")
Disability_category<-c("Back","Cancer","Circulatory", "Digestive", "Ill-defined and Misc Conditions","Injury other than back",
"Mental and Nervous", "Nervous System", "Other", "Other Musculoskeletal", "Respiratory", "Diabetes")
GIBAs<- c("1: $ < 1000", "2: $1,000 - $1,999","3: $2,000 - $2,999", "4: $3,000 - $3,999", "5: $4,000 - $4,999", "6: $5,000 -
$9,999", "7:$10,000 - $14,999", "8:$15,000 - $19,999", "9:$20,000 And Over")
Integration_with_STD <- c("I: Integrated with ASO of Fully-Insured STD", "N: Not Integrated with STD", "U: Unknown")
OwnOccToAnyTransition_MOD <- c("OwnOther", "Own+1", "Own+0")
colnames(TableOut)<-c("Duration","AgeBand","Disabiltiy_Category","OwnToAnyOcc","STD_Integration","Gender","GIBA")
for( ig in Gender) { print(ig)
  for (Integration in Integration_with_STD) { print(Integration)
    for(OwnOcc in OwnOccToAnyTransition_MOD ){ print(OwnOcc)
      for (GIBA in GIBAs) { print(GIBA)
        for(Disability in Disability_category) { print(Disability)
          for(ia in seq(20,70,5))  {
            for(id in MinDuration:MaxDuration)    {
              Counter=Counter+1
               TableOut[Counter,]=c(id,ia, Disability, OwnOcc,Integration, ig, GIBA)
            }
          }
        }
      }
    }
    }
 }
write.csv(TableOut, file="filename&path ")
```

About The Society of Actuaries

The Society of Actuaries (SOA), formed in 1949, is one of the largest actuarial professional organizations in the world dedicated to serving 24,000 actuarial members and the public in the United States, Canada and worldwide. In line with the SOA Vision Statement, actuaries act as business leaders who develop and use mathematical models to measure and manage risk in support of financial security for individuals, organizations and the public.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policymakers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public policymakers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement, and other topics. The SOA's research is intended to aid the work of policymakers and regulators and follow certain core principles:

Objectivity: The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

Quality: The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and non-actuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

Relevance: The SOA provides timely research on public policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

Quantification: The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

SOCIETY OF ACTUARIES
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org