SOCIETY OF ACTUARIES

Group LTD Credibility Study

Results from Stage 2

September 2019

SOCIETY OF ACTUARIES

Group LTD Credibility Study

Results from Stage 2

AUTHOR Paul Correia, FSA, MAAA
Milliman, Inc.

Tasha S. Khan, FSA, MAAA
Milliman, Inc.

Adelina Koseva
Milliman, Inc.

SPONSOR SOA Group Long-Term Disability Credibility Experience Committee

Caveat and Disclaimer

CONTENTS

Group LTD Credibility Study

Results from Stage 2

Section 1: Introduction

The research performed in Stage 1 of the GLTD Credibility Study included an analysis of correlation coefficients between historical LTD experience and future experience based on claim cost ratios calculated from industry data. It also included an analysis of the relative error between historical and future experience. The results were summarized across different variables such as length of experience period, LYE group, industry, etc. Results from Stage 1 are published in the following report:

https://www.soa.org/resources/experience-studies/2018/2018-gltd-credibility-study-stage-1/

In Stage 2 of the GLTD Credibility Study, this analysis was expanded using the industry data gathered during Stage 1. As a reminder, policy exposure and claim data were gathered from 14 disability insurers. The data included 300,020 claim records for LTD claims incurred between January 1, 2004 and December 31, 2011, and 102,951 policy records for LTD policies inforce for at least five consecutive years between January 1, 2004 and December 31, 2011. The data was carefully reviewed, and certain data was excluded if there were issues that could impact study results.

The analysis is based on incurred claim costs, calculated as the present value of expected payments on each claim, divided by covered payroll. The present value of expected payments is calculated as of the end of the elimination period, and is based on claim termination rates from the 2012 GLTD Basic Table, the gross benefit amount payable under the policy with no offsets, and a 3.5% interest rate assumption.

First, a modified version of the relative error approach was used to compare several credibility formulas currently in use in the group LTD industry. For comparison, a "data-driven" credibility formula was developed that is designed to minimize the relative error between predicted claim costs and actual observed claim costs in the observation period. The historical experience representing the "experience rate" is determined from the claim cost in a three-year period (CC1). The observed claim cost, or subsequent claim cost, is the claim cost observed for the next two years of the study period (CC2). The relative error is equal to the absolute difference between the predicted claim cost (based on the credibility formula being tested) and the observed claim cost, divided by the predicted claim cost. The relative error is calculated at the policy level, then rolled into a weighted average relative error for each life year of exposure (LYE) group. An overall weighted average relative error was also determined for each of the credibility formulas, where the average is weighted by life years of exposure in the historical period.

Next, predictive modeling methods were used to study the relative importance of different variables for rating LTD cases. Models were developed in R using the xgboost package, which applies decision trees to identify key predictor variables and to determine where to split the data based on those key variables to reduce the heterogeneity of the data. The output from the model includes a scoring of the relative contribution of each variable for predicting future claim costs. These scores were then used to determine the relative importance of each variable in the rating process. Results were validated by comparing predicted claim costs from the xgboost model to predicted claim costs based on traditional LTD rating formulas, using several different comparison methods.

Any reader of this report should keep in mind that the results do not give specific guidance on the level of credibility that should be assumed for pricing LTD cases, because claim costs were calculated based on a simplified approach using expected future benefit payments with no offsets (i.e., the claim costs ignore volatility associated with actual claim termination and variable benefit patterns). Therefore, the results are more informative when interpreted as relativities as opposed to absolute values.

Section 2: Glossary of Credibility Formulas

Throughout this report, references are made to the following credibility formulas commonly used in the industry for pricing LTD insurance products:

- Industry Formula 1 – Relatively simple and somewhat common approach for determining credibility, based loosely on limited fluctuation credibility theory. Assumes 100% credibility at 25,000 life years of exposure.

  - $Z_1 = \text{Min}\left[\,100\%\,,\ \sqrt{\frac{\text{LYE}}{25{,}000}}\,\right]$

- Industry Formula 2 – Typical industry formula based on life years of exposure and expected claims. Assumes 100% credibility at 25,000 life years of exposure.

  - $Z_2 = \text{Max}\left[\,0\%,\ \text{Min}\left[\,100\%\,,\ \frac{(\text{Exp.claims per 1000})\times\frac{\text{LYE}}{1000}}{(\text{Exp.claims per 1000})\times\frac{\text{LYE}}{1000}+25-\frac{\text{LYE}}{1000}}\,\right]\right]$

- Industry Formula 3 – Similar to Industry Formula 2, but based on the number of actual claims observed in the experience period instead of expected claims. Assumes 100% credibility at 25,000 life years of exposure.

  - $Z_3 = \text{Max}\left[\,0\%,\ \text{Min}\left[\,100\%\,,\ \frac{(\text{Actual claims})}{(\text{Actual claims})+25-\frac{\text{LYE}}{1000}}\,\right]\right]$

- Data-driven Formula – We developed a fourth "data-driven" credibility formula based on the experience data used for this study. This formula is designed to minimize the average relative error between predicted and observed claim costs within each LYE group. Note that LYE is the only variable considered in this Data-driven Formula. The resulting formula is shown below:

  - $Z_4 = \text{Max}\left[\,0\%,\ \text{Min}\left[\,100\%,\ 0.1272 * \ln(\text{LYE}) - 0.5657\,\right]\right]$

Section 3: Executive Summary

The following are key observations from the GLTD Credibility Study. Additional details are provided in subsequent sections of this report.

3.1 Key Observations from Stage 1

- An analysis of correlation coefficients between LTD experience from a prior (lookback) period to a subsequent period suggests that more recent experience is more credible for predicting future experience. In general, correlation coefficients are higher when the experience is based on a shorter lookback period and a shorter gap (or no gap) between the lookback and subsequent periods. For example, results imply that one year of experience for a 4,000 life group demonstrates higher correlation than four years of experience for a 1,000 life group. However, for a given case size, results show greater correlations for longer lookback periods (which increase LYE).

- Correlation coefficients decreased when moving from a study based only on claim incidence to our baseline study method based on actual gross benefit amounts and expected claim terminations. Correlation coefficients decreased further when we moved from our baseline study to a study based on actual claim terminations. This result suggests that LTD credibility is lowered by common characteristics of LTD claim blocks such as variable benefit amounts and claim termination activity. Readers should, therefore, keep this in mind when interpreting the results of these studies as important claim characteristics, such as offset activity and actual claim terminations, have not been reflected in most of the study results.

- An analysis of the relative error between LTD case rates (calculated as a credibility-weighted average of manual and experience rates) and subsequent LTD experience suggests that improving the refinement of manual rates leads to a better ability to predict claim costs, allowing for a reduction in the assumed credibility.

- There is significant volatility in LTD claims experience even for very large groups whose experience includes significant exposure. Results from the relative error analysis show that estimates of future experience using past experience are far from perfect even for the largest groups.

3.2 Key Observations from Stage 2

Comparison of Traditional Credibility Methods

- Several credibility formulas currently used in the group LTD industry were compared using the relative error measure. Industry Formula 3, defined in Section 3 of this report and based on life years of exposure and actual claims, produced the lowest overall relative error. This result implies that this formula, on average, produces the most accurate predictions of future claim costs.

- The Data-driven Formula, which is based solely on LYE, is designed to minimize the relative error. However, it does not perform significantly better than the industry formulas and, in fact, produces a higher overall relative error than Industry Formula 3. This result implies that the credibility formulas currently used in the group LTD industry, when compared to a data-driven approach based solely on LYE, produce reasonable credibility weights.

- We observed decreasing relative errors for the highest LYE groups by increasing the full credibility threshold from 25,000 LYE to 35,000 LYE to 45,000 LYE to 55,000 LYE (see Table 6). This result may suggest that an optimal credibility formula would approach, but never reach, full credibility.

Predictive Modeling Approach

- Predictive models were developed in R using random forest (RF) models within the xgboost package. The relative importance of each variable is evaluated using the SHapley Additive exPlanations (SHAP) importance method—a measure that represents the relative contribution of each variable to the model (note that the appendix of this report contains links to articles and papers that discuss the SHAP importance method). The following variables were identified as the most important variables affecting the credibility of LTD claim experience:

Table 1

SHAP IMPORTANCE OF VARIABLES IN RANDOM FOREST MODEL
(CREDIBILITY MODEL WITH KEY VARIABLES)

Feature	SHAP Importance
delta_pct[1]	49.7%
BetterOrWorse[2]	24.0%
Total LYE	16.3%
Claim Count	10.0%
GRAND TOTAL	**100.0%**

- As expected, the most important predictor of future claim experience, according to this model, is prior experience. The indicator for whether historical claim experience is better or worse than the manual ("BetterOrWorse") also shows a high SHAP importance rating. This implies that it may improve credibility methods to develop formulas that vary based on whether the experience rate is higher or lower than the manual rate. Other key variables that seem to affect the predictive power of historical experience are total life years of exposure and actual claim counts in the historical period.

- The predicted values produced by the random forest model tend to be closer to the actual subsequent claim costs, on average, for our dataset than predicted values generated by the industry credibility formulas. There appears to be a significant reduction in the error in our predictions when moving from any of the industry credibility formulas we tested to a predictive model, as shown in Table 2 below (in which the column with the lowest relative error for each LYE group is highlighted, for ease of comparison). This result implies that current pricing methods could potentially be improved upon by employing predictive modeling techniques in the development of LTD case rates.

---

[1] Variable representing the ratio of the experience rate to the manual rate.
[2] Indicator of whether the experience rate is higher or lower than the manual rate.

Table 2

WEIGHTED AVERAGE RELATIVE ERROR BY LYE

LYE Group	RF3 Model	Industry 1	Industry 2	Industry 3
0-99	175.4%	178.9%	179.4%	175.0%
100-499	131.7%	136.8%	137.4%	132.9%
500-999	87.2%	89.6%	90.0%	86.1%
1,000-1,999	65.9%	69.5%	69.9%	66.4%
2,000-2,999	52.6%	56.7%	56.5%	54.4%
3,000-3,999	44.2%	48.8%	48.4%	46.7%
4,000-4,999	42.2%	44.6%	44.3%	42.8%
5,000-7,499	38.9%	40.4%	40.0%	40.3%
7,500-9,999	34.9%	37.1%	38.0%	35.0%
10,000-19,999	28.5%	31.0%	31.9%	29.1%
20,000-29,999	26.3%	28.9%	29.5%	28.5%
30,000-39,999	23.6%	24.8%	24.8%	24.8%
40,000-49,999	24.6%	28.8%	28.8%	28.8%
50,000+	22.3%	25.9%	25.9%	25.9%
GRAND TOTAL	**60.3%**	**63.5%**	**63.8%**	**61.6%**

Note that in the table above, the overall relative error in the final row represents the weighted average relative error for all cases included in the analysis, weighted by LYE.

- We also compared the claim costs produced by each credibility method using alternative comparison methods. These comparisons also support the conclusion that the predictive model produces predicted claim costs that are, on average, closer to actual claim costs than the claim costs produced by the various industry credibility formulas.

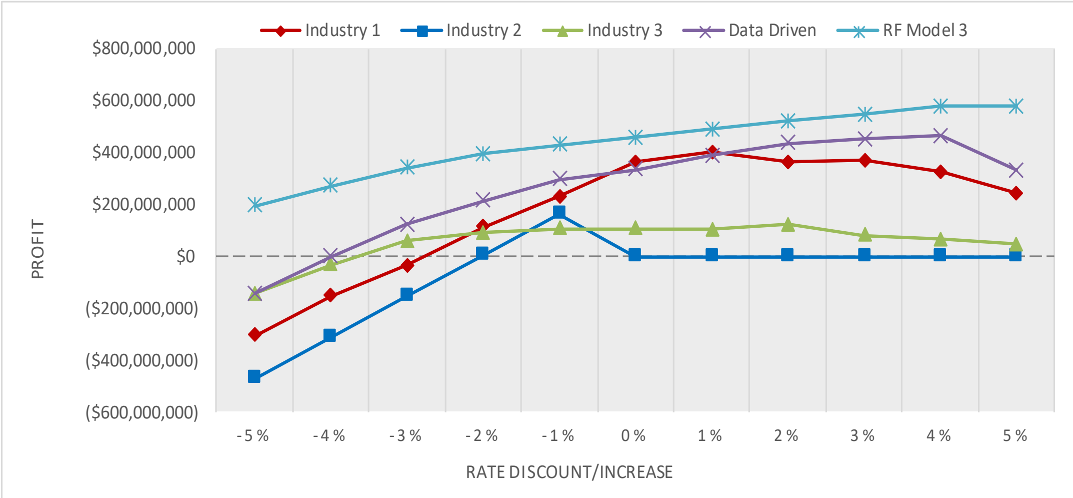
Efficient Frontier Analysis

- The relative error analysis provides insight into which approaches produce the best predictions of future experience. We also performed an efficient frontier analysis to take these results one step further by considering how the accuracy in predictions relates to profitability of LTD business.

- The efficient frontier analysis provides a framework for visualizing the theoretical value of different approaches for estimating future experience by creating a simplified market simulation comparing outcomes of different pricing methods (based on different credibility formulas and on the predictive modeling approach) against a "market rate" to determine which pricing method results in an optimal profit profile.

- Case rates are developed using different credibility formulas and predictive modeling methods and are compared to the market rate. When the case rate is lower than the market rate for a given case, the case is assumed to result in a sale. The different pricing methods were then evaluated based on the profitability of cases that sold – i.e., based on the gains and losses than emerged on those cases over the next two years. We also tested the impact of applying various discounts to the case rates generated by each pricing method. Chart 1 below shows the results from the efficient frontier analysis, in which profit represents the difference between earned premium from cases that sold and incurred claims over a two-year period.

Chart 1
PROFIT VERSUS RATE DISCOUNT / INCREASE



Based on the chart above, the most favorable results correspond to values generated by the predictive model, which are uniformly superior in all rating scenarios. These results imply that business acquired using rates from the predictive model is more profitable than business acquired using other pricing methods, independent of profit targets and pricing strategies. Note that Industry Formula 2 is used as the basis for determining the market rate.

Section 4: Manual Rates

An important aspect of the study was defining the manual rate basis for evaluating the different credibility formulas. We tested three different manual rating bases:

1. Single Manual Rate for all Cases: Based on the overall average claim costs for the entire dataset. Specifically, this simplified manual rate is equal to the present value of future benefits for all policy records included in the study, divided by the total covered payroll for all policy records included in the study. This single manual rate applies to all groups.

2. Refined Manual Rate: Based on expected claim costs that vary by case size, elimination period, industry group, definition of disability, and employer-paid vs. voluntary coverage. Specifically, these manual rates are equal to the present value of future benefits divided by covered payroll for all groups included in any given case size/EP/industry/definition of disability/employer-paid vs. voluntary segment. These refined manual rates do not take into account many important rating variables such as age and gender mix, which was not available in the study data.

3. RF Model Manual Rates: The Refined Manual Rates from (2) above were further refined using an RF model that uses overall industry data and was calibrated using the following variables:

   Independent variables:

   - Industry
   - Region
   - Elimination Period
   - Benefit Percent
   - Benefit Period
   - Voluntary Indicator (employer-paid vs. employee-paid)
   - COLA
   - Definition of Disability
   - Integration with STD
   - Case Size

   Dependent variable: Experience rates from 3-year lookback period.

   The output from the RF model is a unique manual rate for each policy. More detail on the development of manual rates using an RF model is provided in Section 6 of this report (Section 6: Predictive Model 1 – Development of Manual Rates).

We compared the different manual rating bases by calculating relative errors separately for the three bases, using 0% credibility in each case (i.e., the relative error is equal to the absolute difference between the manual rate (based on the basis being tested) and the observed claim cost, divided by the manual rate). A comparison of the relative errors is provided in Table 3 below.

Table 3

WEIGHTED AVERAGE RELATIVE ERROR BY LYE AND MANUAL RATE BASIS
RELATIVE ERROR = ABS [MANUAL RATE – ACTUAL RATE] / MANUAL RATE

LYE Group	Single Manual Rate	Refined Manual Rate	RF Model Manual Rate
0-99	180.9%	189.5%	180.1%
100-499	148.6%	148.5%	139.2%
500-999	105.2%	99.1%	93.0%
1,000-1,999	89.4%	80.9%	76.4%
2,000-2,999	78.1%	69.2%	64.5%
3,000-3,999	75.7%	65.1%	60.6%
4,000-4,999	71.4%	58.2%	54.0%
5,000-7,499	67.1%	56.2%	52.4%
7,500-9,999	61.6%	50.9%	46.4%
10,000-19,999	64.4%	51.4%	46.4%
20,000-29,999	52.6%	47.8%	43.6%
30,000-39,999	52.1%	59.8%	51.5%
40,000-49,999	61.1%	34.9%	31.5%
50,000+	40.1%	40.4%	32.4%
GRAND TOTAL	**84.3%**	**78.4%**	**72.4%**

The results shown above indicate that, before the application of credibility methods, improving the refinement of manual rates leads to a better ability to predict claim costs. When we combine the manual rates with experience rates using credibility factors based on Industry Formula 1[3], we still notice the refined manual rate outperforming the single manual rate, and the RF model manual rate producing the closest predictions overall, as shown below. Note that the relative errors shown below are equal to the absolute difference between the predicted rate (based on the manual rate being tested) and the observed claim cost, divided by the predicted rate.

Table 4

WEIGHTED AVERAGE RELATIVE ERROR BY LYE AND MANUAL RATE BASIS
RELATIVE ERROR = ABS [PREDICTED RATE – ACTUAL RATE] / PREDICTED RATE

LYE Group	Single Manual Rate	Refined Manual Rate	RF Model Manual Rate
0-99	178.4%	187.1%	178.9%
100-499	142.4%	144.2%	136.8%
500-999	95.7%	93.9%	89.6%
1,000-1,999	75.0%	72.0%	69.5%
2,000-2,999	62.2%	59.1%	56.7%
3,000-3,999	53.7%	50.2%	48.8%
4,000-4,999	48.7%	45.8%	44.6%
5,000-7,499	45.3%	42.2%	40.4%
7,500-9,999	40.0%	38.8%	37.1%
10,000-19,999	30.6%	31.9%	31.0%
20,000-29,999	28.7%	29.3%	28.9%
30,000-39,999	24.8%	24.8%	24.8%
40,000-49,999	28.8%	28.8%	28.8%
50,000+	25.9%	25.9%	25.9%
GRAND TOTAL	**66.3%**	**66.1%**	**63.5%**

[3] Industry Formula 1 was selected at random for this analysis. The results in Table 4 are similar when different credibility factors are used.

We determined that the RF model manual rates most closely resemble manual rates used by insurance companies versus the other two bases. Therefore, all of the analyses discussed in subsequent sections of this report were performed using the RF model manual rates.

Section 5: Comparison of Traditional Credibility Methods

5.1 Analytical Methods

We began Stage 2 analysis by testing several of the credibility formulas currently being used in the group LTD industry, as well as a data-driven formula developed using the study data. The Data–driven Formula was developed by first calculating relative errors for all policies corresponding to credibility values ranging from 0% to 100%. We then determined the credibility factors that minimize the average relative error for all policies included in a given LYE group, and used them to define a logarithmic formula that best fit these credibility factors and varied by LYE.

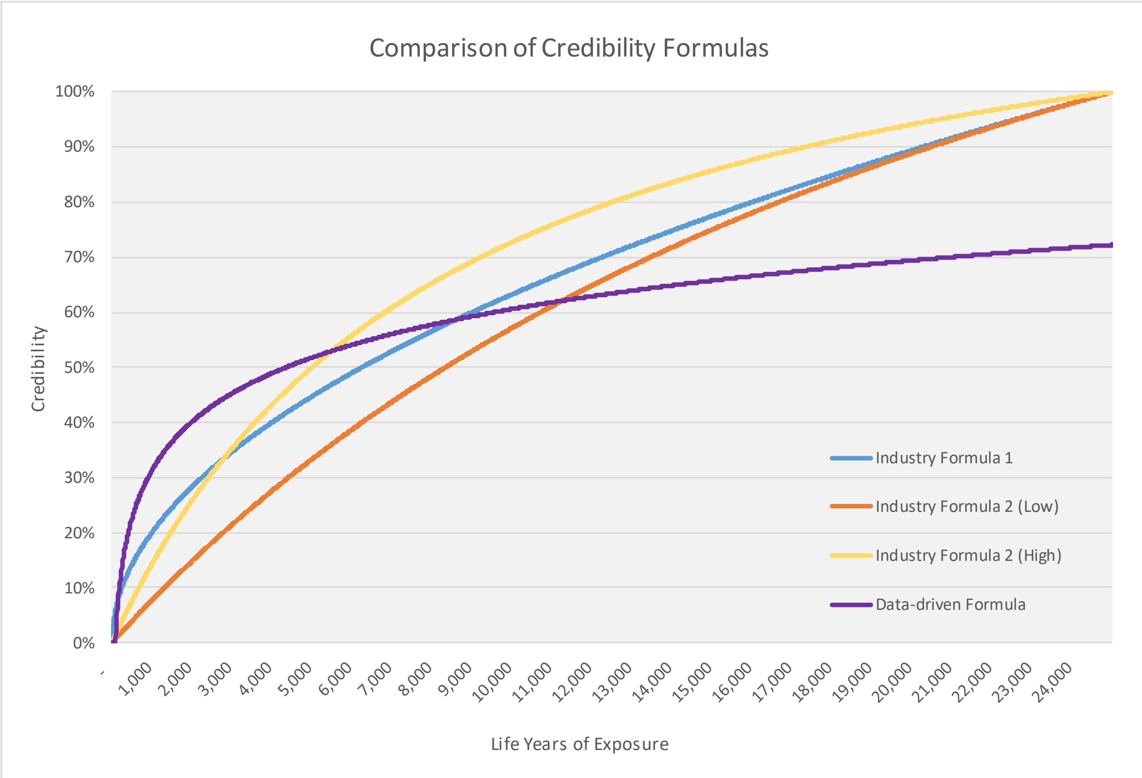
We tested the credibility formulas defined in Section 3 of this report, and restated them below:

- Industry Formula 1:
  - $Z_1 = \text{Min}\left[\,100\%\,,\ \sqrt{\frac{\text{LYE}}{25{,}000}}\,\right]$
- Industry Formula 2:
  - $Z_2 = \text{Max}\left[\,0\%,\ \text{Min}\left[\,100\%\,,\ \frac{(\text{Exp.claims per 1000})\times\frac{\text{LYE}}{1000}}{(\text{Exp.claims per 1000})\times\frac{\text{LYE}}{1000}+25-\frac{\text{LYE}}{1000}}\,\right]\right]$
- Industry Formula 3:
  - $Z_3 = \text{Max}\left[\,0\%,\ \text{Min}\left[\,100\%\,,\ \frac{(\text{Actual claims})}{(\text{Actual claims})+25-\frac{\text{LYE}}{1000}}\,\right]\right]$
- Data-driven Formula:
  - $Z_4 = \text{Max}\left[\,0\%,\ \text{Min}\left[\,100\%\,,\ 0.1272 * \ln(\text{LYE}) - 0.5657\right]\right]$

The following graph provides a visual comparison of Industry Formula 1, Industry Formula 2, and the Data-driven Formula. Note that two curves are shown for Industry Formula 2 based on expected claims of 2 per 1,000 (Low Scenario) and 4 per 1,000 (High Scenario)[4] to illustrate the impact of higher claim levels on the assumed credibility.

[4] The expected claims values are illustrative and were arbitrarily chosen to facilitate the comparison. In reality, each policy record has its own values for expected claims and actual claims.

Graph 1
COMPARISON OF CREDIBILITY FORMULAS



Generally speaking, the Data-driven Credibility Formula assigns higher credibility than the industry formulas for smaller size cases, and lower credibility for larger size cases, with all else being equal. Also, the industry formulas assume full credibility at 25,000 LYE, whereas the Data-driven formula only assumes 72% credibility at 25,000 LYE.

To compare the effectiveness of each formula, we first calculated the predicted claim costs (CCP) as the credibility-weighted average of the historical claim costs, CC1, and the manual rates. For the manual rates, we used the RF model manual rates described in Section 3 (with additional detail on the development of these manual rates provided in Section 7).

We then compared the predicted claim costs, CCP, to the observed claim costs, CC2, using the relative error measure. The relative error for a given policy is equal to the absolute difference between CCP and CC2, divided by CCP. Again, the average relative error, weighted by LYE in the historical period, is calculated for each LYE group and overall.

5.2 Test Results

Table 5 below compares the relative error for each credibility formula.

Table 5

RELATIVE ERROR COMPARISON OF CREDIBILITY FORMULAS

LYE Group	Data-driven	Industry 1	Industry 2	Industry 3
0-99	180.0%	178.9%	179.4%	175.0%
100-499	137.0%	136.8%	137.4%	132.9%
500-999	90.3%	89.6%	90.0%	86.1%
1,000-1,999	69.1%	69.5%	69.9%	66.4%
2,000-2,999	56.0%	56.7%	56.5%	54.4%
3,000-3,999	47.6%	48.8%	48.4%	46.7%
4,000-4,999	44.3%	44.6%	44.3%	42.8%
5,000-7,499	40.1%	40.4%	40.0%	40.3%
7,500-9,999	36.9%	37.1%	38.0%	35.0%
10,000-19,999	30.7%	31.0%	31.9%	29.1%
20,000-29,999	26.2%	28.9%	29.5%	28.5%
30,000-39,999	24.8%	24.8%	24.8%	24.8%
40,000-49,999	24.7%	28.8%	28.8%	28.8%
50,000+	25.8%	25.9%	25.9%	25.9%
GRAND TOTAL	**63.1%**	**63.5%**	**63.8%**	**61.6%**

In the table above, we have highlighted the column with the lowest relative error for each LYE group for ease of comparison. Industry Formula 3 produces the lowest relative error in most LYE segments, and overall, potentially indicating that this formula produces the most accurate predictions of future claim costs. For groups in higher LYE segments, however, the Data-driven Formula often produces the lowest relative error. As we will see below, this may be driven by the fact that all of the industry formulas tested assign full credibility when the LYE reaches or exceeds 25,000, while the Data-driven Formula assigns only partial credibility for these higher LYE groups.

It is interesting that the Data-driven Formula, which is designed to minimize the relative error, does not perform significantly better than the industry formulas and, in fact, produces a higher overall relative error than Industry Formula 3. This result implies that the credibility formulas currently used in the group LTD industry, when compared to a data-driven approach based solely on LYE, produce reasonable credibility weights.

The industry formulas represented in Table 5 reflect a full credibility threshold of 25,000. We also tested the formulas using alternative maximum credibility thresholds. The results for Industry Formula 3 are summarized in the table below.

Table 6

RELATIVE ERROR COMPARISON - ALTERNATIVE MAXIMUM CREDIBILITY THRESHOLDS FOR INDUSTRY FORMULA 3

LYE Group	Maximum Credibility Threshold			
	25,000	35,000	45,000	55,000
0-99	175.0%	176.0%	176.6%	177.1%
100-499	132.9%	134.0%	134.8%	135.3%
500-999	86.1%	86.1%	86.1%	88.2%
1,000-1,999	66.4%	67.4%	68.2%	68.8%
2,000-2,999	54.4%	55.3%	56.1%	56.8%
3,000-3,999	46.7%	46.7%	46.7%	49.9%
4,000-4,999	42.8%	42.8%	42.8%	44.4%
5,000-7,499	40.3%	40.3%	40.3%	41.8%
7,500-9,999	35.0%	35.0%	35.2%	35.5%
10,000-19,999	29.1%	29.1%	28.9%	29.3%
20,000-29,999	28.5%	28.5%	28.5%	25.8%
30,000-39,999	24.8%	24.8%	24.8%	23.9%
40,000-49,999	28.8%	28.8%	28.8%	23.9%
50,000+	25.9%	25.9%	25.9%	25.8%
Weighted Average	61.6%	61.8%	62.2%	62.5%

Overall, the 25,000 maximum credibility threshold produces the lowest average relative error. This is true for LYE groups up to 5,000-7,499. At higher LYE groups, higher maximum credibility thresholds produce lower relative errors. This result seems to imply that an optimal credibility formula would approach, but never reach, full credibility. It is also interesting to note that, when comparing Tables 5 and 6, even for larger LYE groups, the higher thresholds often produce better results than the Data-driven formula based solely on LYE.

We also tested an alternative Data-driven formula that includes a variable indicating whether the experience rate is higher or lower than the manual rate (which is referred to as the "BetterOrWorse" variable in this report) in addition to LYE. Table 7 below shows the relative errors for data-driven formulas that vary by LYE only versus LYE and BetterOrWorse:

Table 7

RELATIVE ERROR COMPARISON - ALTERNATIVE DATA-DRIVEN FORMULAS

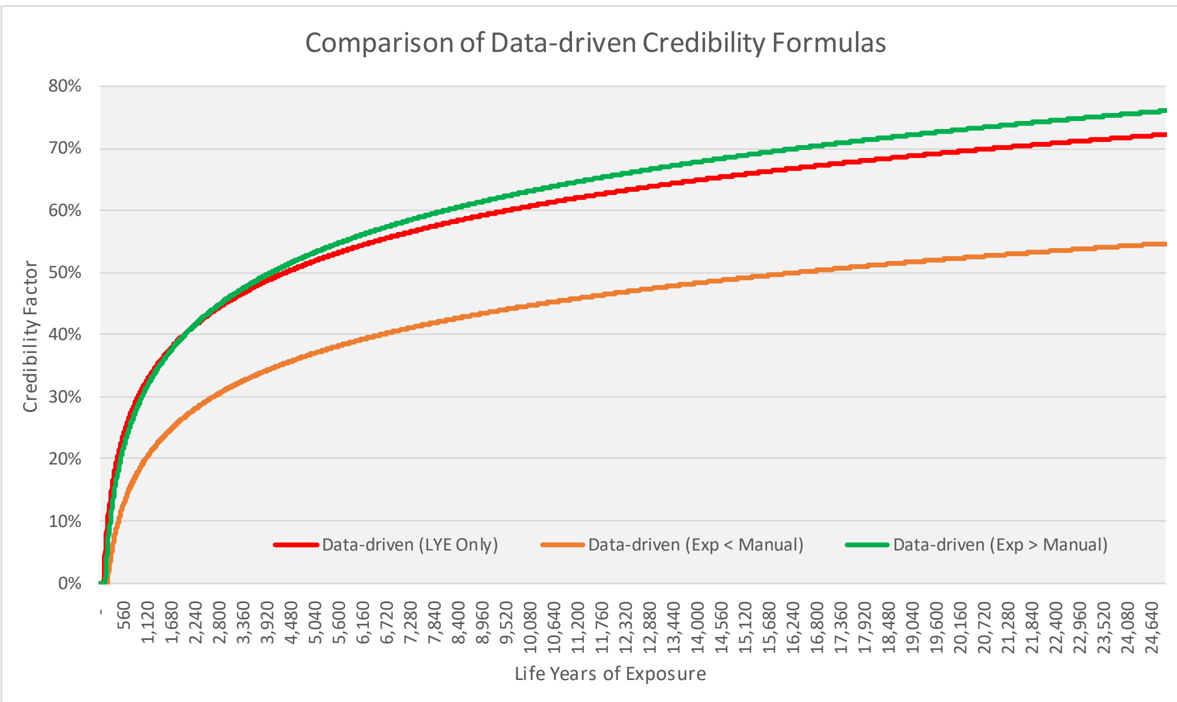
LYE Group	Variables Included in Credibility Formula	
	LYE only	LYE and BetterOrWorse
0-99	180.0%	180.1%
100-499	137.0%	134.6%
500-999	90.3%	87.3%
1,000-1,999	69.1%	67.3%
2,000-2,999	56.0%	54.8%
3,000-3,999	47.6%	47.3%
4,000-4,999	44.3%	43.3%
5,000-7,499	40.1%	40.2%
7,500-9,999	36.9%	35.8%
10,000-19,999	30.7%	29.8%
20,000-29,999	26.2%	26.3%
30,000-39,999	24.8%	25.0%
40,000-49,999	24.7%	24.2%
50,000+	25.8%	23.7%
Weighted Average	63.1%	61.9%

The Data-driven formula that varies by LYE and BetterOrWorse produces the lowest overall relative error, and the lowest relative errors in most LYE segments, suggesting that an effective approach for pricing LTD may be to use different formulas depending on whether the experience rate is greater than or less than the manual rate.

The equations for the Data-driven formulas are provided below, followed by the graphs of the equations:

- Data-driven formula that varies by LYE only:
  - Z = Max [0%, Min [100% , 0.1272 * ln(LYE) – 0.5657]]
- Data-driven formula that varies by LYE and BetterOrWorse:
  - Experience < Manual: Z = Max [0%, Min [100% , 0.1104 * ln(LYE) – 0.5710]]
  - Experience > Manual: Z = Max [0%, Min [100% , 0.1425 * ln(LYE) – 0.6825]]

Graph 2
COMPARISON OF DATA-DRIVEN CREDIBILITY FORMULAS



The Data-driven formula that varies by LYE and BetterOrWorse assigns higher credibility to experience in cases where the experience rate is higher than the manual rate, and vice versa. This is somewhat intuitive from the standpoint that cases with worse experience generally have more claims, so the experience includes more data.

Section 6: Predictive Model 1 – Development of Manual Rates

6.1 Analytical Methods

The predictive models were designed to study relationships among different variables in order to determine the relative importance of each variable for predicting future experience based on historical experience. These variables included group and case-level characteristics (industry, region, elimination period, etc.), manual rates, experience rates, and other variables that describe relationships between the manual rate and experience rate (for example, whether the experience rate is greater than the manual rate).

The models were developed in R using the xgboost package to build relationships among the variables. These models create independent decision trees for every policy record. Each tree is trained independently using random samples of data—i.e., each decision tree considers a random subset of variables and uses a random set of the training data. The model then takes an average of all of the individual decision trees to adjust the starting expectation (i.e., manual rate) in order to estimate future claim costs.

In the xgboost model, all of the independent variables need to be numerical; therefore, the categorical data, such as industry and region, were translated into indicator variables with numerical values (e.g., 1 if manufacturing industry and 0 otherwise). The data was segmented into a training set (80%) and a test set (20%) to validate model output, where the test set is used to test the decision trees on a different dataset than the training data. The training set and test set were randomly selected within the model.

The SHAP method was used to evaluate the relative importance of each variable. SHAP importance is a relatively new and unbiased method recently added to xgboost. The SHAP importance represents the relative contribution of a variable to the model, where a higher value of SHAP importance means the variable is more important for generating a prediction.

Several links to additional information on the xgboost model, random forests, and the SHAP importance measure are provided in the Appendix.

6.2 Variables

The following variables were included in the analysis:

- Manual rate
- Experience rate
- Industry
- Region
- Elimination Period
- Benefit Percent
- Benefit Period
- Voluntary Indicator (employer-paid vs. employee-paid)
- COLA
- Definition of Disability
- Integration with STD
- Case Size

Experience rates were calculated for all policies as the ratio of incurred claims divided by covered payroll using experience from a three-year experience period. The incurred claims in the numerator represent the present value of expected benefits as of the end of the elimination period for all claims incurred in the period, using the 2012 GLTD Basic Table for projecting future benefits and a 3.5% discount rate. This approach ensures the experience rates were calculated consistently for every policy, and are not biased by different approaches for estimating incurred claims among companies that participated in the study, or by different claim management practices among those companies.

The manual rates were calculated in two stages. First, a preliminary set of rates was developed from overall industry experience by computing average incurred claim cost ratios across broad segments. Covered payroll was again used in the denominators of the ratios so that the experience and manual rates would be on the same basis. These preliminary rates varied by elimination period, voluntary indicator, definition of disability, and industry. Next, the preliminary rates were refined using a random forest (RF) model that uses overall industry data and was calibrated using the following variables:

Training variable: Preliminary rate that varies by elimination period, voluntary indicator, definition of disability, and industry.

Independent variables:

- Industry
- Region
- Elimination Period
- Benefit Percent
- Benefit Period
- Voluntary Indicator (employer-paid vs. employee-paid)
- COLA
- Definition of Disability
- Integration with STD
- Case Size

Dependent variable: Experience rates

The output from the model is a unique manual rate for each policy based on the manual rating formula as defined by the model. Values for SHAP importance were also generated by the model and used to determine the relative contribution of each variable in estimating the manual rates.

6.3 Test Results – Manual Rates

The output of the first random forest model is a manual rate for each record based on the manual rating formula as defined by the RF model. A higher value of SHAP importance means the variable is more important for generating a manual rate.

Table 8
SHAP IMPORTANCE OF VARIABLES IN RF1 MODEL
(MANUAL RATE BUILD)

Variable	SHAP Importance
STD Integration	53.5%
Industry	15.3%
Region	10.1%
COLA	5.7%
Case Size	4.2%
Definition of Disability	3.4%
Voluntary Indicator Group	3.1%
Elimination Period	2.6%
Benefit Percent	1.9%
Benefit Period	0.1%
GRAND TOTAL	100%

Note that the training variable for this model (i.e. the preliminary rate) is the Refined Manual Rate described in Section 3, so the variability in claim experience based on features such as elimination period and definition of disability were already reflected in the refined manual rates. These features, therefore, show relatively low SHAP importance in the table above. Other variables, such as integration with STD and Region, on the other hand, were not reflected in the refined manual rate and, therefore, indicate a higher SHAP importance.

The output of this random forest model is a manual rate for each policy record. This manual rate is then used in the second random forest model.

Section 7: Predictive Model 2 – Identification of Important Variables

7.1 Analytical Methods

The manual rates developed from the first RF model were used in a second RF model to identify the key variables that affect future claim cost predictions.

7.2 Variables

The new manual rates developed in the first RF model were used in a second RF model to predict future claim rates for all policies. In this model, the training variable is the new manual rate.

Training variable: New manual rate output from the first random forest model

Independent variables:

- Total LYE (exposure within three-year lookback experience period)
- BetterOrWorse (indicator for whether experience rate is higher or lower than the manual rate)
- Delta_pct (variable representing the ratio of the experience rate to the manual rate)
- Industry
- Region
- Elimination Period
- Benefit Percent
- Benefit Period
- Case Size
- Voluntary Indicator
- COLA
- Definition of Disability
- Integration with STD
- Calendar Year
- Actual Claim Count

Dependent variable: Claim rate from subsequent two-year experience period immediately following the three-year lookback experience period

The output from the second RF model is a scoring of the relative importance of each of the variables for predicting future claim rates.

7.3 Test Results – Key Variables

Because the manual rates were developed using a predictive model that incorporates the key variables expected to affect underlying claim experience, we can now isolate the impact of these and other variables on the predictive power of historical LTD claim experience.

Table 9
SHAP IMPORTANCE OF VARIABLES IN RF2 MODEL
(CREDIBILITY RATE BUILD)

Variable	SHAP Importance
delta_pct	36.4%
BetterOrWorse	15.6%
Benefit Percent	10.2%
Claim Count	7.4%
Region	6.5%
Total LYE	6.0%
Industry	5.8%
Case Size	5.7%
Definition of Disability	1.8%
STD Integration	1.3%
COLA	1.1%
Voluntary Indicator Group	0.9%
Calendar Year	0.9%
Elimination Period	0.6%
Benefit Period	0.0%
GRAND TOTAL	100.0%

As expected, the most important predictor of future claim experience, according to this model, is prior experience. The indicator for whether historical claim experience is better or worse than the manual also shows a high SHAP importance rating. This implies that it may improve credibility methods to produce credibility formulas that vary depending on whether the experience rate is higher or lower than the manual rate.

The variables shown above were then tested further for their contributions toward improving the prediction of future experience through a stepwise comparison of relative errors. For example, predicted values were developed separately using RF models that feature (1) LYE only, and (2) LYE and case size. We then calculated relative errors for the predicted values based on LYE only and for those based on LYE and case size, and we performed a side-by-side comparison of the relative errors to see if they improved from adding a new variable. Not surprisingly, the relative errors are very similar for the predicted values based on LYE only and those based on LYE and case size, because LYE and case size are closely related.

We performed these comparisons using all of the variables in Table 9 to determine whether the variables considered important by the SHAP method do, in fact, provide meaningful improvements in predicting future claim costs. The results from these comparisons suggest that the most important variables for predicting future experience are LYE, delta_pct, BetterOrWorse, and claim count. Including other variables—such as benefit percent, region, etc.—did not significantly improve predictions. Therefore, we considered the independent variables LYE, delta_pct, BetterOrWorse, and claim count to be the most important variables based on the reduction in relative errors observed when these variables are included. The table below provides results from the stepwise comparison for these four variables, in which we observed significant improvement in relative errors.

Table 10

RELATIVE ERROR COMPARISONS

LYE Group	Variables Included in RF Model			
	LYE Only	LYE and Delta_pct	LYE, Delta_pct, and Claim Count	LYE, Delta_pct, Claim Count, and BetterOrWorse
0-99	174.4%	174.6%	176.4%	175.4%
100-499	133.1%	131.4%	130.5%	131.7%
500-999	90.7%	87.8%	85.6%	87.2%
1,000-1,999	74.8%	66.6%	65.4%	65.9%
2,000-2,999	62.6%	53.1%	52.2%	52.6%
3,000-3,999	59.1%	45.1%	44.7%	44.2%
4,000-4,999	52.8%	42.4%	42.0%	42.2%
5,000-7,499	51.2%	39.9%	39.9%	38.9%
7,500-9,999	45.7%	35.8%	35.4%	34.9%
10,000-19,999	45.1%	30.3%	30.0%	28.5%
20,000-29,999	43.0%	28.0%	28.3%	26.3%
30,000-39,999	49.9%	25.8%	25.5%	23.6%
40,000-49,999	31.7%	25.6%	27.0%	24.6%
50,000+	31.8%	23.6%	23.6%	22.3%
Weighted Average	**70.2%**	**61.2%**	**60.8%**	**60.3%**

Adding the Delta_pct variable reduced the overall relative error from 70.2% to 61.2%. As expected, the predictions for larger-sized cases benefitted significantly from the inclusion of Delta_pct (i.e., prior experience). Adding the variables Claim Count and BetterOrWorse provided more modest improvements in relative errors. Furthermore, adding the Claim Count variable reduced relative errors primarily for smaller-sized cases (<5,000 LYE), and adding the BetterOrWorse variable reduced relative errors primarily for larger-sized cases (>5,000 LYE).

We tested other variables and noticed very little improvement in the relative errors when including additional variables beyond LYE, delta_pct, BetterOrWorse, and claim count.

Section 8: Predictive Model 3 – Generation of Predicted Claim Costs

8.1 Analytical Methods

We set up the third random forest model to generate predicted claim costs based on the key variables identified in Section 7 above. The predicted claim costs generated by this model were then tested against case rates calculated from other credibility methods (i.e., the industry formulas and Data-driven formula).

8.2 Variables

The third random forest model uses the following variables:

Training variable: new manual rate output from RF Model 1

Independent variables:

- Total LYE (exposure within 3-year lookback experience period)
- Actual Claim Count
- BetterOrWorse (indicator for whether experience is more or less favorable than the manual)
- Delta_pct (variable representing the ratio of the experience rate to the manual rate)

Dependent variable: Claim rate from subsequent period (two-year period following the lookback experience period)

Again, for testing results, the data is segmented into a training set (80%) and a test set (20%), where the test set is used to test the random forest on a different dataset than the training data. The training set and test set are randomly selected within the data.

8.3 Test Results – Predicted Claim Costs

Values for SHAP importance from the RF3 Model are summarized in Table 11 below.

Table 11
SHAP IMPORTANCE OF VARIABLES IN RF3 MODEL
(CREDIBILITY MODEL WITH KEY VARIABLES)

Feature	SHAP Importance
delta_pct	49.7%
BetterOrWorse	24.0%
Total LYE	16.3%
Claim Count	10.0%
GRAND TOTAL	100.0%

Once we have the predicted claim costs from RF Model 3, we compare the predicted rates to experience in the subsequent period (CC2) using the relative error measure. Results for each LYE group, and in total, are shown in the table below. For comparison, we have also included the same relative error measures for the industry credibility formulas.

Table 12

WEIGHTED AVERAGE RELATIVE ERROR BY LYE AND RATING METHOD

LYE Group	RF3 Model	Industry 1	Industry 2	Industry 3
0-99	175.4%	178.9%	179.4%	175.0%
100-499	131.7%	136.8%	137.4%	132.9%
500-999	87.2%	89.6%	90.0%	86.1%
1,000-1,999	65.9%	69.5%	69.9%	66.4%
2,000-2,999	52.6%	56.7%	56.5%	54.4%
3,000-3,999	44.2%	48.8%	48.4%	46.7%
4,000-4,999	42.2%	44.6%	44.3%	42.8%
5,000-7,499	38.9%	40.4%	40.0%	40.3%
7,500-9,999	34.9%	37.1%	38.0%	35.0%
10,000-19,999	28.5%	31.0%	31.9%	29.1%
20,000-29,999	26.3%	28.9%	29.5%	28.5%
30,000-39,999	23.6%	24.8%	24.8%	24.8%
40,000-49,999	24.6%	28.8%	28.8%	28.8%
50,000+	22.3%	25.9%	25.9%	25.9%
Weighted Average	**60.3%**	**63.5%**	**63.8%**	**61.6%**

The predicted values produced by the random forest model tend to be closer to actual future claim costs, on average, for our dataset. There appears to be a significant reduction in the error in our predictions when moving from any of the industry credibility formulas we tested to a predictive model. This result implies that current credibility methods could potentially be improved upon by employing predictive modeling techniques in the development of LTD case rates.

We considered an alternative approach for comparing the predicted values from the RF Model 3 to the predicted values calculated from traditional credibility methods. For every case in the dataset, we assigned a score of 1 to the method that produced a rate that is closest in absolute value to the actual future claim rate, and a score of 0 to all of the other methods. We then tallied the scores to determine which method produced the closest rate most often. The results are shown below:

Table 13

PERCENTAGE OF CASES WITH CLOSEST PREDICTED VALUES

LYE Group	RF3 Model	Industry 1	Industry 2	Industry 3
0-99	76%	5%	11%	8%
100-499	62%	5%	18%	15%
500-999	55%	3%	16%	25%
1,000-1,999	56%	3%	15%	26%
2,000-2,999	55%	9%	11%	26%
3,000-3,999	55%	18%	7%	21%
4,000-4,999	40%	21%	15%	24%
5,000-7,499	27%	27%	28%	19%
7,500-9,999	30%	24%	25%	21%
10,000-19,999	36%	22%	22%	20%
20,000-29,999	33%	22%	22%	23%
30,000-39,999	28%	24%	24%	24%
40,000-49,999	24%	25%	25%	25%
50,000+	22%	26%	26%	26%
Total 0 – 999	**70%**	**5%**	**14%**	**12%**
Total 1,000 +	**48%**	**12%**	**16%**	**24%**
GRAND TOTAL	**68%**	**5%**	**14%**	**13%**

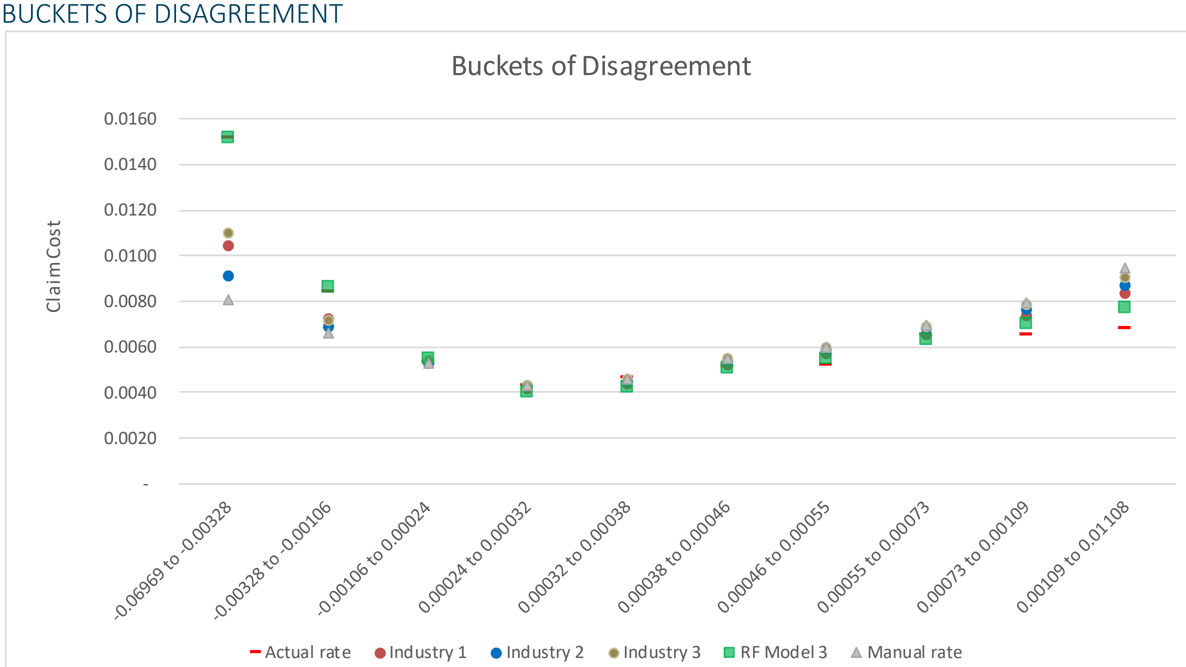
Based on the results shown above, the predicted claim costs from the RF3 model were closer to actual claim costs for most policy records. For example, for policies within the 0-99 LYE group, the predicted claim costs from the RF3 model were closest on 76% of the cases, followed by predicted claim costs based on Industry Formula 2 (11% of cases), Industry Formula 3 (8% of cases), and Industry Formula 1 (5% of cases).

Furthermore, based on the results in Table 13, the predicted claim costs from the RF3 model are significantly better for smaller-sized cases. Because the predicted values from the RF3 model are based, in part, on experience characteristics (e.g., claim count and BetterOrWorse), the results may suggest that there is value in considering historical experience even for smaller-sized groups. They may also suggest that traditional approaches for estimating credibility may not assign enough credibility to smaller-sized cases, and this is consistent with the data-driven approach, which also assigns higher credibility to smaller-sized cases than the industry formulas, generally speaking.

8.4 Buckets of Disagreement

Another approach used by actuaries to evaluate model predictions is known as the "buckets of disagreement" comparison. First, the difference between the manual rate and the predicted value (generated by the RF3 model) is calculated for each observation, and then the observations are sorted from smallest to largest difference. The observations are then divided into 10 equal buckets. For example, if there are 100 observations, then the first bucket would contain the 10 observations with the smallest difference between the manual and predicted value (i.e., the first bucket would typically be one of the two buckets with the largest absolute differences in which the differences are negative). Within each bucket, we calculate the average manual rate, the average predicted value from the RF model, the average actual rate from the subsequent period (CC2), and the average case rates using different credibility formulas. We can then compare these values for each bucket to determine which approach produces rates that are closer to the actual rate from the subsequent period (i.e., the value we are trying to predict). The graph below shows the results of this analysis, where the predicted values are generated by the RF Model 3 approach described above. Note that using alternative credibility formulas to determine the buckets does not produce results that are significantly different from the results presented in this graph.

Graph 3
BUCKETS OF DISAGREEMENT



In the first bucket, representing policies for which the manual rate is lower than the predicted value, we see that the claim costs generated by the RF model are much closer to the subsequent actual claim costs (CC2) than the claim costs produced by the industry credibility formulas. We see the same result for the second bucket. For buckets 3-8, there is much less variation between the claim costs generated by the predictive model and the claim costs produced by the industry credibility formulas. In buckets 9 and 10, the RF claim costs are again closer to CC2 than the claim costs produced by the industry formulas.

Overall, this analysis supports the conclusion that the random forest model produces predicted claim costs that are, on average, closer to actual future claim costs than the estimates produced by the various industry credibility formulas.

8.5 Test for Overfitting

In order to evaluate the potential for overfitting the model when developing the predicted case rates (in the third RF Model), we also looked at the results for only the test dataset. The test dataset represents the 20% of records that were selected randomly by the model and held back from the model calibration. For the test dataset, the total weighted average relative error is 58.6%. Because this is lower than the overall total relative error of 60.3%, we conclude that the model does not seem to have been over-fit to the data.

Section 9: Efficient Frontier Analysis

9.1 Analytical Methods

The primary objective of this analysis is to evaluate different pricing methods through a model that projects future LTD sales and profitability. These methods were used to calculate case rates for all of the policies included in the analysis. Four of the approaches involved computing the credibility weighted average of the manual and experience rates based on the industry and data-driven credibility formulas described above. The fifth approach uses predicted rates from RF Model 3.

Case rates were then benchmarked against "market rates" to determine the likelihood of cases selling. Any pricing method that produced a case rate lower than the market rate for a given case was assumed to result in a sale. The pricing methods were then evaluated based on the profitability of cases that sold—i.e., based on the gains and losses that emerged on those cases over the next two years.

For every case, we calculated a "market rate" that represented a hurdle for selling the case. For example, if the market rate for a given case is $0.50, we assumed that any pricing method resulting in a case rate below $0.50 would result in a sale. Conversely, we assumed that any method resulting in a case rate equal to or greater than $0.50 would not result in a sale.

The market rates are based on a credibility-weighted average of the predicted manual rates (i.e., rates generated by the first RF Model) and the experience rates, where the underlying credibility is based on Industry Formula 2.

Case rates corresponding to the different industry credibility formulas were calculated based on traditional methods for determining the credibility-weighted average of the experience and manual rates (i.e., case rate = Z x Experience Rate – (1 – Z) x Manual rate). We calculated separate case rates based on Industry Formula 1, Industry Formula 2, Industry Formula 3, and the Data-driven formula that varied by LYE and the BetterOrWorse variable. We also included case rates based on the predicted rates from RF Model 3 (which do not reflect traditional pricing methods).

One way to evaluate the various case rate predictions is by comparing the gains and losses for cases that sold. In this analysis, the gains and losses were calculated as follows:

1. For every case that sold, we estimated earned premium over the next two years (subsequent period) according to the formula:
    - Premium = Case Rate x Covered Payroll in subsequent period
2. We determined the incurred claims amount in the subsequent period for all cases that sold.
3. We calculated gains and losses by taking the difference between the earned premium and incurred claims:
    - Gain/Loss = Premium from Step 1 minus Incurred Claims from Step 2

Note that, in the analysis, we did not include any assumptions around administrative expenses or other non-claim cost items.

9.2 Test Results

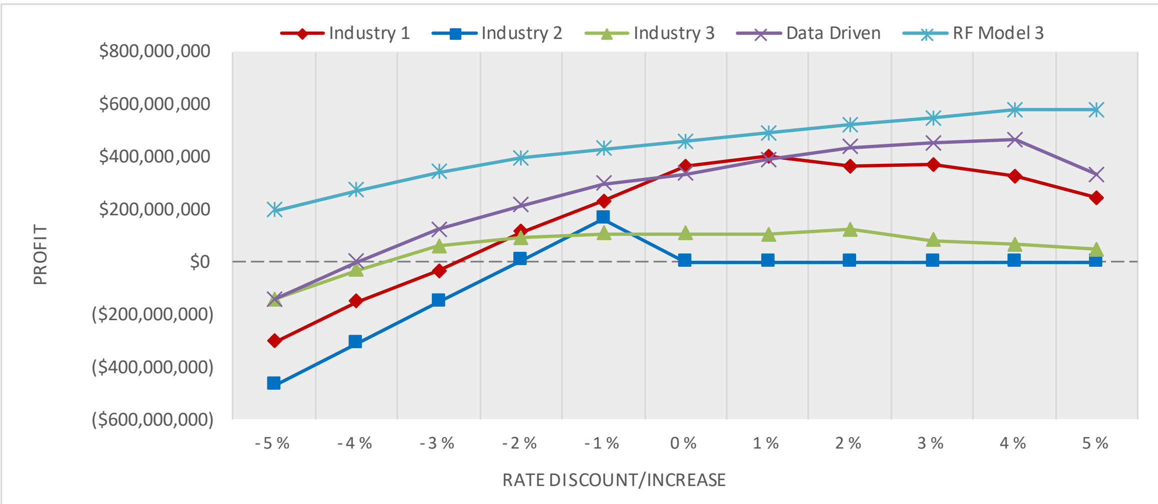
We assumed that a case would sell if the case rate was lower than the market rate. We also considered the impact of discounting case rates by 1% to 5% (e.g., we discounted the case rate by 5%, then compared the discounted case rate to the market rate to determine if the case sold). Similarly, we considered the impact of increasing case rates by 1%

to 5%. We generated results separately for each of the discount and increase scenarios because the number of cases sold and profitability depend on the magnitude of the discount or increase.

The results are summarized in graphical form below. The first graph shows total profit for cases that sold using the five different pricing methods. The second graph shows profit as a percentage of premium. The third graph shows the number of cases that sold for each of the pricing methods. The fourth graph shows a distribution of gains and losses by LYE group, based on the scenario that uses the strict case rate (i.e., no discounting or increasing).

Chart 2

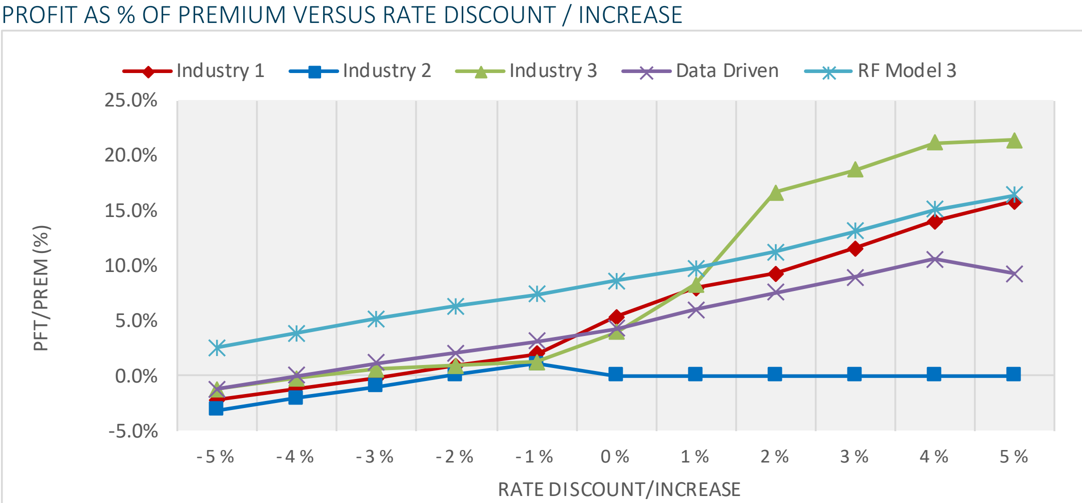
PROFIT VERSUS RATE DISCOUNT / INCREASE



In the chart above, we can see that the predicted values from the RF3 model perform better than the industry formulas and the Data-driven Formula for every rate discount/increase scenario.

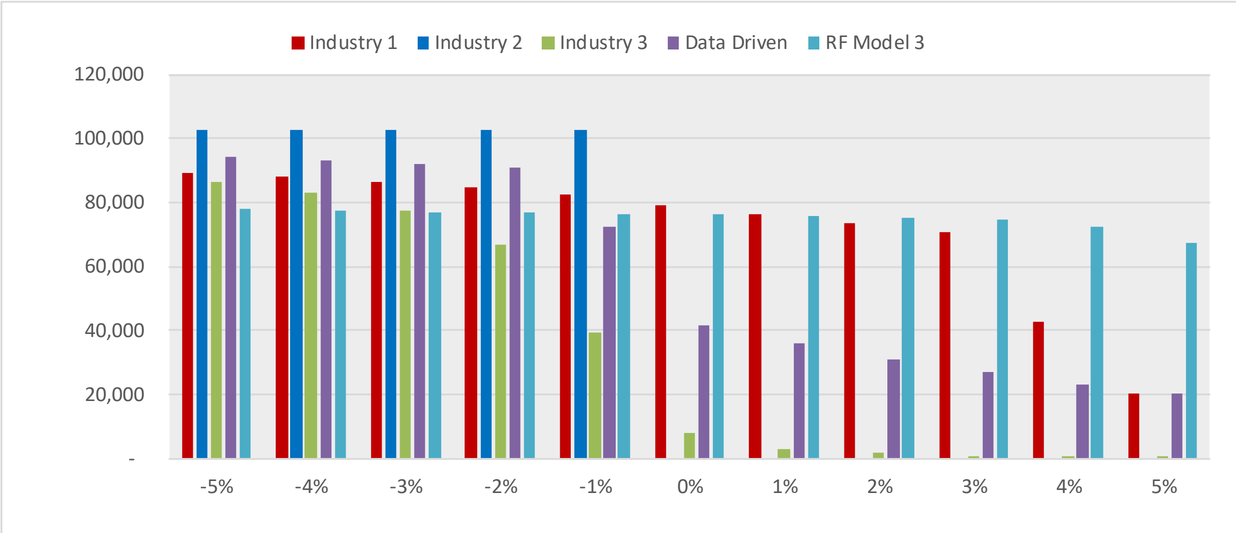
Chart 3

PROFIT AS % OF PREMIUM VERSUS RATE DISCOUNT / INCREASE



When we measure profit as a percentage of total premium sold, the predicted values from the RF3 model still produce favorable results relative to the industry formulas; however, Industry Formula 3 results in greatest profitability when the case rate is increased by 2-5%. The number of cases sold, however, is very low for Industry Formula 3, unless discounts are applied to the case rate.

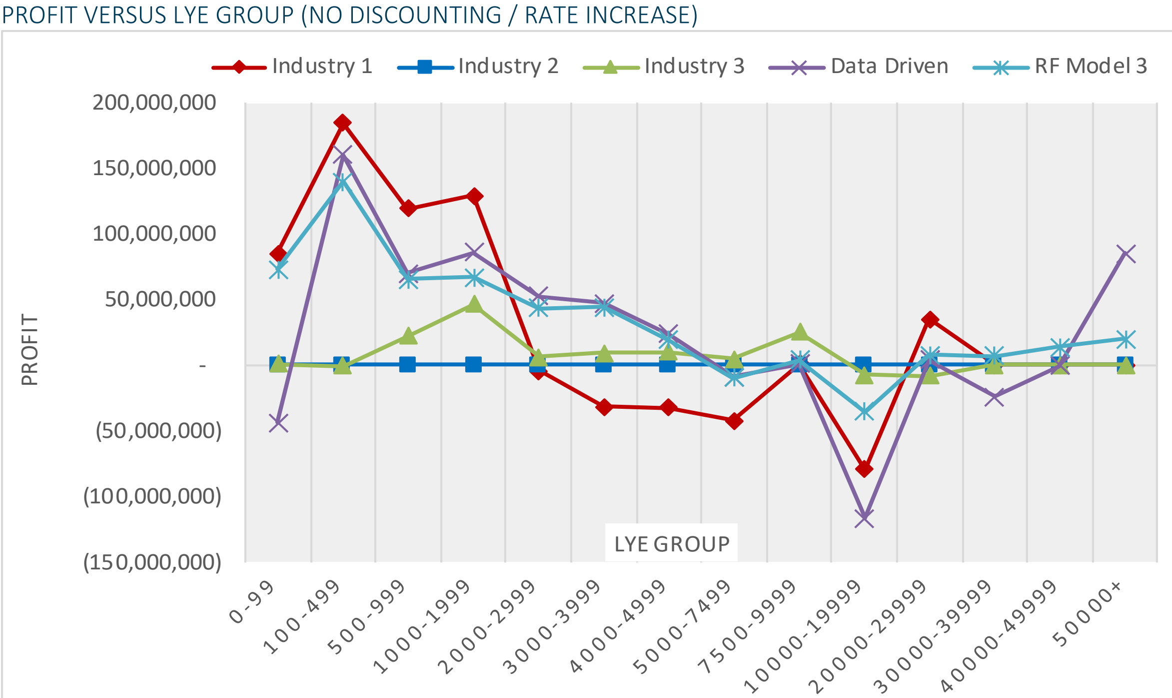
Chart 4
NUMBER OF CASES SOLD



Interestingly, the number of cases sold is relatively stable for the predicted rates from the RF3 model across all rate discount/increase scenarios.

Finally, based on Chart 5 below, Industry Formula 3 produced the lowest volatility between different LYE groups, likely because Industry Formula 3 is the most similar to Industry Formula 2, which is used to develop the market rates. In general, profit is highest for the lowest LYE groups.

Chart 5
PROFIT VERSUS LYE GROUP (NO DISCOUNTING / RATE INCREASE)



Section 10: Acknowledgements

The researchers' deepest gratitude goes to those without whose efforts this project could not have come to fruition: the Project Oversight Group and others for their diligent work overseeing questionnaire development, analyzing and discussing respondent answers, and reviewing and editing this report for accuracy and relevance.

Project Oversight Group Members

Avtar Singh, FSA, MAAA

Bram Spector, FSA, MAAA

Doug Vrooman, FSA, MAAA

Jiayu Guo, FSA, MAAA

Jinn Lin, FSA, MAAA

Julie Shuman, FSA, MAAA

Kari Stokely, FSA, MAAA

Kathy Davis, FSA, MAAA

Mark Mortensen, FSA, MAAA

Matthew Desfosses, ASA, MAAA

Matthew Sawyer, FSA, MAAA

Michael Jiang, FSA, MAAA

Rick Leavitt, ASA, MAAA

Scott Carter, FSA, MAAA

Tom Corcoran, FSA, MAAA (Co-chair)

Warren Cohen, FSA, MAAA (Co-chair)

Zheng Bai, FSA, MAAA

Other Resources

The SOA contracted with MIB's Actuarial and Statistical Research Group to collect, validate, and compile the data for this report. Erika Schulty, SOA Research Associate, and Pete Miller, SOA Experience Studies Actuary, supplied project management support.

Section 11: List of Participating Companies

The Society of Actuaries would like to thank the following 14 companies who contributed data to this study:

- AIG
- Anthem
- Cigna
- Guardian
- Liberty Mutual
- MetLife
- Mutual of Omaha
- Northwestern Mutual
- OneAmerica
- Prudential
- Reliance Standard
- Standard
- The Hartford
- Unum

Section 12: Reliance and Limitations

12.1 Reliance

In conducting the analysis, researchers relied upon the database developed by MIB specifically for the GLTD Credibility Experience Study. Unless otherwise described, researchers did not audit or independently verify any of the information furnished, except for a high level review of the data for reasonableness and consistency. To the extent that any of the data or other information supplied was incorrect or inaccurate, the results of this analysis could be materially affected.

12.2 Limitations on Use and Distribution of Report

This report is intended for the benefit of the Society of Actuaries. Although the authors understand that this report will be made widely available to third parties, Milliman does not assume any duty or liability to such third parties with its work. This report should be distributed and reviewed only in its entirety.

The results in this report are technical in nature and are dependent on certain assumptions and methods. No party should rely upon these results without a thorough understanding of those assumptions and methods. Such an understanding may require consultation with qualified professionals.

The underlying analysis was performed using assumptions about future LTD claim costs. Differences between claim cost projections and actual claim cost amounts depend on the extent to which future experience conforms to the assumptions made for this analysis. It is certain that actual experience will not conform exactly to the assumptions used in this analysis. Actual claim costs will differ from projected claim costs to the extent that actual experience deviates from expected experience.

We, Paul Correia and Tasha Khan, are Consulting Actuaries with Milliman and members of the American Academy of Actuaries. We meet the qualification standards of the American Academy of Actuaries for rendering the actuarial opinion contained in this report.

Appendix: Links to Documentation of xgboost and SHAP Importance

The following links provide useful documentation of the xgboost package for R and of the method for evaluating variables using SHAP importance:

- https://cran.r-project.org/web/packages/xgboost/xgboost.pdf - documentation on how to create a random forest type model with xgboost.

- https://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html# - tutorial on how to load data into xgboost.

- https://xgboost.readthedocs.io/en/latest/parameter.html - discussion of the hyperparameters in xgboost.

- https://github.com/slundberg/shap - summary of SHAP method including creating a feature importance ranking that is consistent with the model's output.

- https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27 - article on SHAP importance that discusses why SHAP importance ranking is considered to be more robust than the default "importance" method output in xgboost.

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 more than 32,000 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 nonactuaries 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