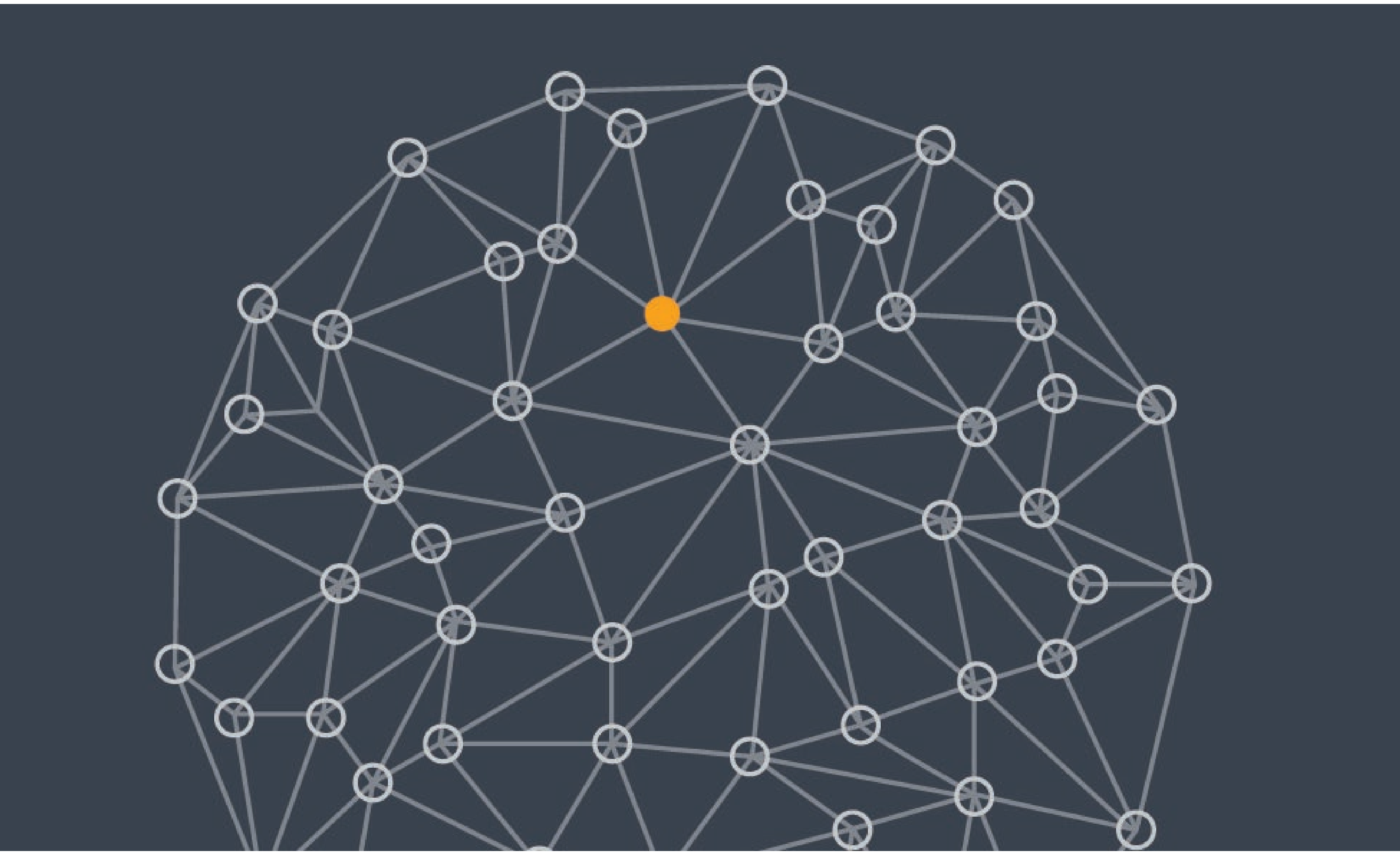
MILLIMAN RESEARCH REPORT

The Feasibility of using Machine Learning Techniques for Health Insurance Pricing in India

December 2023

Rachin Aggarwal, MBA
Lalima Chakravarty, MBA



Milliman

Table of Contents

1. INTRODUCTION 1

2. CASE STUDY – MACHINE LEARNING TECHNIQUES FOR PRICING 2

UNDERLYING DATA AND APPROACH 2

ANALYSIS 2

3. SURVEY - ADOPTION LEVEL OF ML ACROSS INDUSTRY 7

4. CONCLUSION 8

5. APPENDICES 9

APPENDIX A – LIST OF VARIABLES IN THE DATA 9

APPENDIX B – SEARCH GRID FOR TUNING PARAMETERS 9

APPENDIX C – POISSON AND GAMMA DEVIANCE 9

1. INTRODUCTION

The health insurance in India is a growing segment of India's economy. It is one of the major contributors of growth of general insurance industry in India. Rise in middle class, higher hospitalization cost, expensive health care, digitization and increase in awareness level are some of the important drivers for the growth of health insurance market in India.

One of the biggest challenges faced by the health insurance industry in India, and particularly actuaries, is determination of a comprehensive, fair, and adequate price for insuring different sets of benefits. If premiums are set too high, policyholders may switch to other companies and if premiums are underpriced, companies may have inadequate funds to cover claims.

Pricing actuaries tend to operate within the classical framework of Generalised Linear Models (GLMs) for retail products. The traditional actuarial pricing methods sometimes result in situations where the pricing is not reflective of the underlying risk.

With the emergence of machine learning, computing power, and development of predictive modeling tools and techniques, it has now become easier to analyze big data comprehensively and draw conclusions that allow insurance companies to form profitable diverse portfolios and minimize potential adverse risks. Machine learning offers new ways to the companies to make more precise pricing, giving them greater certainty that their pricing is aligned to the risks underlying their portfolio.

Insurance claim data is usually characterized by highly imbalanced count data with excessive zeros and varying exposure-to-risk on the frequency side, combined with long tailed continuous data on the severity side. Thus, claim frequency and claim severity are skewed distributions and are therefore generally modelled using Poisson and Gamma distributions respectively across the industry. In some cases, depending on the data, other models such as Zero-inflated Poisson models, Quassipoisson and Negative Binomial are also used.

In this paper, we investigate how relatively new techniques such as tree-based machine learning models perform compared to the classical actuarial pricing approach of GLMs, along with capturing the current adoption level of such techniques across the health insurance industry in India through a survey.

Tree-based models use a chain of *if-then* rules to generate predictions from one or more decision trees. The predictive power of a single decision tree is low and we therefore resort to more advanced ensemble-based tree models such as Random Forest or Gradient Boosting Model (GBM) for the purpose of this research.

Tree-based models can be visualized and are easy to understand. On the other hand, ensemble-based models are better at handling noise and imbalanced data. While these ML models are good at handling complex, non-linear relationships, there are some limitations:

- Predictions from single decision tree are prone to overfitting and are comparatively weak compared with more advanced machine learning models such as support vectors and neural networks.
- Insurance pricing models are heavily regulated and must meet all regulatory and filing requirements before being deployed in practice.

Therefore, machine learning models must be transparent with appropriate reasoning and justification for any differences in premium calculations.

Note that this paper is intended solely for educational purposes and presents information of a general nature. The underlying data and analysis were reviewed on this basis. This report is not intended to guide or determine any specific individual situation, and readers should consult qualified professionals before taking specific actions.

2. CASE STUDY – MACHINE LEARNING TECHNIQUES FOR PRICING

The purpose of this research is to study the feasibility of machine learning techniques other than GLM for health insurance pricing. Machine learning techniques such as regression trees, random forest etc. are getting increasingly popular for pricing insurance products. We also aimed to analyse the adoption levels of these machine learning techniques for pricing through industry survey.

UNDERLYING DATA AND APPROACH

We have used claims and membership data from a health insurer for the purpose of this research. In addition, we have referred to research papers and case studies[1] to refine our approach and methodology.

We have used the following three models / statistical methods for our study.

1. GLM

   GLM is an extension of classical linear model and introduces a link function around the linear combination of the explanatory variables, is defined by the following equation:

   $g(E[Y|X]) = X\beta$,

   where $E[Y|X]$ is the expected value of Y conditional on X,

   $X\beta$ is the linear predictor, a linear combination of unknown parameters β, and

   g is link function

2. Random Forest (RF)

   Random forest model is a collection of multiple decision trees that are created independently, using a random sample of the data. Only a subset of features is then selected at random out of the total and the best split feature from the subset is used to split each node in a tree.

3. Gradient Boosting Model (GBM)

   Gradient boosting trees construct a multitude of decision trees that are dependent on each another, i.e., GBM uses decision trees as base learners and combines multiple weak learners into one powerful predictor. Gradient Boosting Model reduces error mainly by reducing bias while Random Forest model reduces error by reducing variance.

In carrying out our analysis and producing this research report, we have relied on the de-identified data and information available to us from our pricing assignments. We did not audit or verify this data or other information. If the underlying data or information is inaccurate or incomplete, the results of our analysis may likewise be inaccurate or incomplete.

We performed a limited review of the data used directly in our analysis for reasonableness and consistency and did not find material defects in the data. It should be noted that, in some cases, errors were spotted in the underlying data. We ran reasonable data checks and made minor changes to the data (modeling variables) to ensure consistency. A summary of modeling features is listed in Appendix A.

ANALYSIS

Step 1: Importing Data

We started the analysis by importing claims and membership data using R programming language.

Table 1 shows the data description of the underlying data. The libraries and functions used in reading the data along with reading time summary are shown in Table 2.

[1] Thesis "Boosting insights in insurance tariff plans with tree-based machine learning methods" written by Roel Henckaerts, Marie-Pier Cote, Katrien Antonio, and Roel Verbelen
Thesis "Towards Machine Learning: Alternative Methods for Insurance Pricing – Poisson-Gamma GLM's, Tweedie GLM's and Artificial Neural Networks" written by Navarun Jain and published by IFoA

TABLE 1: DATA DESCRIPTION

DATASET	ROWS	COLUMN	FILE SIZE
CLIENT DATA.CSV	10,667,365	11	1.14 GB
CLIENT DATA.RDS	10,667,365	11	69.4 MB

TABLE 2: READING TIME SUMMARY

LIBRARY	FUNCTION	READING TIME (IN SECONDS)
UTILS	READ.CSV()	35.7
BASE	READRDS()	15.5
DATA.TABLE	FREAD()	5.6

Step 2: Data Processing

We performed the following data checks before proceeding with the modelling step.

- Validated data types to ensure consistent variable formats
- Checked the range of values (min/max) and its reasonability
- Checked for missing values in the data
- Checked for duplicate records
- Checked negative values in the fields for Count and Net Paid Amount, as these fields are used in modeling variables.

Based on the data checks, we cleaned and prepared the data for analysis.

We partitioned the data in to six subsets (also called “folds”). We used these folds for applying cross-validation and checking model stability. For cross fold validation, we trained the model on the first five folds of data and tested on the sixth fold to obtain the modeling results. This process was iteratively carried out for different fold combinations for training and test sets. The cross-fold validation process is explained briefly in the next step.

We also identified extreme outliers for each target variable i.e., Frequency and Severity. Outliers are defined as any data values which lie more than 3 times the interquartile range below the first quartile or above the third quartile. This constituted around 12% of the data. We removed these outliers from the processed data for further analysis.

Step 3: Data Modeling

In this step, we fitted machine learning models on the processed data.

We partitioned the data into six subsets (also called “folds”) following a cross-fold approach as explained below:

- The test data and the train data are partitioned in such a way that test data has one fold and train data has five folds.
- The five-fold train data is split further into validation data and train data in the ratio of 1:3.

The train data, validation data and test data can be defined as:

1. Train data - dataset for fitting the model with predictors and target variable.
2. Validation data - a sample of train data, used for estimating and tuning the model’s hyper-parameters.
3. Test data - dataset for model testing with only predictor variables.

We built and tested the machine learning models on all possible training and test dataset combinations. The model with the most stable predictions across all folds is considered as the best fit model.

Tuning hyper-parameters is an important step in determining the best fit model. This is done using the h2o framework in R which allows for cartesian and random grid searches for parameter tuning. Details of tuning grid of hyper-parameters along with search criteria to determine best tuned model is provided in Appendix B.

Table 3 below summarizes hyper-parameters used for three models – GLM, Random Forest and Gradient Boosting Model (GBM) below:

TABLE 3: PARAMETERS DESCRIPTION AND THE TUNING RANGE/VALUES USED FOR EACH PARAMETER

MODEL	LIBRARY	FUNCTION	PARAMETERS	VALUE USED / TUNING RANGE	PARAMETER DESCRIPTION
GLM	STATS	GLM()	FAMILY	FREQUENCY: POISSON (LINK = 'LOG') SEVERITY: TWEEDIE (VAR.POWER, LINK.POWER)	PRODUCES A GENERALIZED LINEAR MODEL FAMILY OBJECT.
			VAR.POWER	2	INDEX OF POWER VARIANCE FUNCTION.
			LINK.POWER	0	INDEX OF POWER LINK FUNCTION. DEFAULTS TO 1-VAR.POWER.
RANDOM FOREST (RF)	H2O	H2O.GRID (ALGORITHM = "RANDOMFOREST")	NTREES	POSITIVE INTEGER (50 ≤ N ≤ 1000), STEP INCREASE BY 50.	INTEGER SPECIFYING THE NUMBER OF TREES IN THE MODEL.
			MTRIES	POSITIVE INTEGER (1 ≤ N ≤ NUMBER OF FEATURES), STEP INCREASE BY 1.	INTEGER SPECIFYING THE NUMBER OF RANDOMLY CHOSEN VARIABLES TO CONSIDER AT EACH NODE TO FIND THE OPTIMAL SPLIT.
			SAMPLE_RATE	POSITIVE NUMBER (0.2 ≤ P ≤ 1), STEP INCREASE BY 0.05.	ROW SAMPLING RATE FOR EACH INDIVIDUAL TREE IN THE MODEL.
GRADIENT BOOSTING MODEL (GBM)	H2O	H2O.GRID (ALGORITHM = "GBM")	NTREES	POSITIVE INTEGER (50 ≤ N ≤ 1000), STEP INCREASE BY 50.	INTEGER SPECIFYING NUMBER OF TREES IN THE MODEL.
			MAX_DEPTH	POSITIVE INTEGER (2 ≤ N ≤ NUMBER OF FEATURES), STEP INCREASE BY 1.	DEPTH OF EACH TREE OR MAXIMUM NUMBER OF TIMES SPLITTING IS DONE.
			LEARN_RATE	VALUES RANGE FROM 0-1. RANGE = C(0.001, 0.01, 0.1)	LEARNING RATE CONTROLS THE LEARNING SPEED BASIS PREVIOUS WEAK LEARNERS (TREES).
			SAMPLE_RATE	POSITIVE NUMBER (0.2 ≤ P ≤ 1), STEP INCREASE BY 0.1.	SAMPLES SPECIFIED PERCENTAGE OF ROWS PER SPLIT.
			COL_SAMPLE_RATE	POSITIVE NUMBER (0.2 ≤ P ≤ 1), STEP INCREASE BY 0.1.	SAMPLES SPECIFIED PERCENTAGE OF COLUMNS PER SPLIT.

We also faced a few constraints while modelling the data:

- Maximum tuning time was set to one hour for the purpose of this analysis, which may not provide sufficient modeling iterations to conclude best fit model.
- Certain models failed to deploy due to memory allocation error (e.g., for random forest, if number of trees parameter had a higher value, (say 2,000) model deployment fails). In our study, 82 models failed in the Random Forest technique due to memory allocation error in Fold 1.

An excerpt of the Fold 1 analysis description and the tuned parameter values based on input arguments is shown in Table 4:

TABLE 4: FOLD 1 ANALYSIS DESCRIPTION AND TUNED PARAMETER VALUES

	FOLD 1 ANALYSIS					
BASIS	GLM		RANDOM FOREST		GBM	
MODELING VARIABLE	COUNT_COMPLETE	SEVERITY	COUNT_COMPLETE	SEVERITY	COUNT_COMPLETE	SEVERITY
WEIGHTS VARIABLE	M_NEW_EXPOSURE	COUNT_COMPLETE	M_NEW_EXPOSURE	COUNT_COMPLETE	M_NEW_EXPOSURE	COUNT_COMPLETE
PARAMETERS	FAMILY	FAMILY, VAR.POWER, LINK.POWER	NTREES, MTRIES, SAMPLE_RATE	NTREES, MTRIES, SAMPLE_RATE	NTREES, MAX_DEPTH, LEARN_RATE, SAMPLE_RATE, COL_SAMPLE_RATE	NTREES, MAX_DEPTH, LEARN_RATE, SAMPLE_RATE, COL_SAMPLE_RATE
VALUE	POISSON(LINK = 'LOG')	TWEEDIE, 2, 0	450, 3, 0.45	250, 3, 0.4	510, 8, 0.01, 0.9, 0.5	410, 8, 0.01, 0.6, 1
TUNING TIME	< 5 MINUTES	< 5 MINUTES	1 HOUR	1 HOUR	1 HOUR	1 HOUR
MODELING ITERATIONS	NA	NA	6 MODELS	18 MODELS	13 MODELS	84 MODELS

The final step in data analysis is to calculate the predicted frequency and severity. Using these, we derived the following:

- Expected Risk Premium = [Predicted Frequency] x [Predicted Severity]; and
- Expected Model Loss = [Expected Risk Premium] x [Exposure]

Step 4: Evaluation

Modeling Deviance is one of the best stopping metrics for assessing insurance claims frequency / severity dataset. We used Poisson Deviance for Frequency models and Gamma Deviance for evaluating Severity models. The lower the deviance metric, the better the model fit.

We have also compared the Expected Model Loss with Actual Loss to determine accuracy relativity. The results are aggregated at a portfolio level by averaging the predicted losses over all six folds.

The model with the lowest deviance, highest accuracy (relative to actual loss), and the most stable result across all folds is selected as the optimal model. For our study, Random Forest model turned out to be the most optimal tree-based model with lowest Poisson and Gamma deviances and an accuracy score between 99.2% - 99.7% (with a range of 0.5% based on further parameter tuning) respectively at a portfolio level.

Table 5 below summarizes the modeling predictions calculated using all the three methods across all six folds and compares them with the actual loss.

TABLE 5: MODEL PREDICTIONS AND ACTUAL LOSS ACROSS ALL 6 FOLDS

	PREDICTED AND ACTUAL LOSS AMOUNTS			
DATA FOLD	GLM	RF	GBM	ACTUAL LOSS
1	18,805,394	18,788,435	18,743,218	18,911,899
2	18,792,662	18,759,662	18,708,486	18,914,109
3	18,778,749	18,763,180	18,713,389	18,921,344
4	18,792,057	18,763,379	18,691,481	18,923,747
5	18,796,294	18,788,504	18,705,237	18,922,171
6	18,804,341	18,776,507	18,697,793	18,920,281
PORTFOLIO	**18,794,916**	**18,773,278**	**18,709,934**	**18,918,925**

We have used three different methodologies to assess and compare each model:

- Test data Accuracy – this is calculated as the ratio of Predicted Modeling Loss to Actual Loss.
- Poisson deviance – this measure how well the Poisson model fits the data. Model with lowest Poisson deviance implies best fit.
- Gamma deviance – this evaluates the goodness of fit for severity model. Smaller deviance value implies better model fit.

Table 6 below summarizes model evaluation metrics across all six folds.

TABLE 6: MODEL EVALUATION - TEST DATA ACCURACY, POISSON, AND GAMMA DEVIANCE

ACCURACY			POISSON DEVIANCE			GAMMA DEVIANCE		
GLM	RF	GBM	GLM	RF	GBM	GLM	RF	GBM
99.4%	99.3%	99.1%	0.5864	0.5846	0.5851	0.2601	0.2587	0.2587
99.4%	99.2%	98.9%	0.5865	0.5848	0.5851	0.2599	0.2586	0.2585
99.2%	99.2%	98.9%	0.5862	0.5846	0.5851	0.2600	0.2591	0.2587
99.3%	99.2%	98.8%	0.5866	0.5849	0.5858	0.2601	0.2587	0.2586
99.3%	99.3%	98.9%	0.5862	0.5840	0.5866	0.2597	0.2585	0.2732
99.4%	99.2%	98.8%	0.5862	0.5839	0.5857	0.2606	0.2592	0.2592
99.3%	**99.2%**	**98.9%**	**0.5864**	**0.5845**	**0.5855**	**0.2601**	**0.2588**	**0.2612**

3. SURVEY - ADOPTION LEVEL OF ML ACROSS INDUSTRY

We undertook a limited survey to understand the adoption level of Machine Learning (ML) across the Indian health and general insurance industry. The purpose of the survey was to understand the usage of ML techniques for pricing and other types of analyses in the Health and General insurance market. The survey was taken by six respondents with varying backgrounds (actuaries, data scientists, etc.) and all of them work on pricing and related projects at Indian insurance companies. The survey also intends to assess the effectiveness of ML techniques, based on their experience.

The results of the survey are summarized below:

1. Actuarial pricing exercises are generally performed more than once a year.
2. The time required for conducting a pricing exercise for a product and its variants is around 10-20 average person-days.
3. The most common methods used for pricing are Burning Cost[2] and GLM.
4. Most respondents were willing to use ML modelling for pricing insurance contracts in the future, and by using R and SAS as their preferred programming languages.
5. However, most actuaries have not yet used ML techniques in their team. The top challenges reported in adopting ML techniques are:
   i) Interpretation of results
   ii) Unavailability of resources skilled to use the different ML techniques
   iii) Regulatory requirements
6. All respondents believe that their current methodology has the required predictive abilities to price the insurance contracts appropriately, but it might not give appropriate results where data is scarce for a cohort.
7. The few respondents who currently use ML techniques in their team commented that:
   i) they use an in-house model based on R and Python programming languages.
   ii) they use ML techniques not only for pricing but also for the following:
      - Lapse/Renewal analysis
      - Analysis for identification of large claims
      - Fraud, Waste and Abuse analysis
   iii) they use following ML models:
      - Decision trees/ Random forests
      - Boosting methods (GBM, XGBoost etc.)
      - Deep Learning (Neural Networks etc.)
   iv) they are not comfortable in communicating ML output to the internal stakeholders as well as the regulators.
   v) they validate the models and explain the output by comparing the actual with estimated claim cost.
   vi) they are concerned about regulatory requirements while using ML for pricing. They feel that ML techniques might not have all the required transparency and regulatory approval for pricing.

[2] Burning Cost method (also referred as Frequency-Severity method) is an approach for estimating risk premium as the expected claim cost multiplied by the expected average number of claims in the period.

4. CONCLUSION

There is a significant role of big data and data science methods (such as ML models) in the insurance risk management. Application of ML techniques in insurance will help optimize marketing strategies, improve business decision making, boost income and reduce costs.

Through the research conducted for this paper, we also learnt that it is possible to substantiate the accuracy of a given method, model performance, stability and suitability of the underlying assumptions including choice of loss function through statistical methods. In our analysis, the Random Forest ML model was consistently selected as best modeling approach in terms of highest predictive power on test set and lowest deviance error across all folds. The final accuracy score is subject to an optimal parameter tuning process across different folds. Unlike GLM, RF and GBM have three and five tuning parameters respectively to compute model predictions on unknown test set. Based on the current approach for parameter tuning, tree-based model's performance can still be further improved. GLM, on the other hand does not currently have tuning parameters to optimize. Overall, it can be concluded that autonomous machine learning algorithms such as Random Forest and Gradient Boosting hold great potential for actuaries in insurance pricing. Similarly, more research can be done on ways to combine two distinct models: for example, GLM's ease of interpretability can be combined with predictive power of Random Forest or Gradient Boosting Model to improve model fit.

However, the current adoption levels of ML techniques, especially for pricing, is low due to multiple reasons ranging from scarcity of skilled resources to lack of transparency, both in terms of modeling process and the output. So, it is essential to train resources in this emerging field to adapt and use these new techniques to their best potential, and through consistent effort by the industry, we may derive advantages of using these techniques over traditional methods.

5. APPENDICES

APPENDIX A – LIST OF VARIABLES IN THE DATA

TABLE A.1: DESCRIPTION OF VARIABLES IN THE DATA

CLAIMS AND EXPOSURE-TO-RISK VARIABLES:	
COUNT_COMPLETE	18,805,394
NET_PAID_AMOUNT_COMPLETE	18,792,662
M_NEW_EXPOSURE	18,778,749
MODELING FEATURES:	
M_PAYER	INSURER NAME - CATEGORICAL VARIABLE WITH 5 LEVELS.
M_MARITALSTATUS	MARITAL STATUS OF THE POLICYHOLDER.
M _NATIONALITY	NATIONALITY OF THE POLICYHOLDER.
M _NETWORK	NETWORK TYPE FOR EACH POLICYHOLDER.
M_GENDER_NEW	GENDER OF THE POLICYHOLDER MAPPED WITHIN SAS.
M_RELATIONSHIP_NEW	RELATIONSHIP OF POLICYHOLDER - EITHER ONE OF DEPENDENT, PRINCIPAL OR SPOUSE.
M_PERIOD_TYPE	CALENDAR YEAR IN WHICH POLICY WAS ACTIVE.
M_AGE_BRACKET	AGE BRACKETS FOR POLICYHOLDERS RANGING FROM 0-17 TO 66+.

APPENDIX B – SEARCH GRID FOR TUNING PARAMETERS

TABLE B.1: SEARCH GRID FOR THE TUNING PARAMETERS IN DIFFERENT TREE-BASED MACHINE LEARNING MODELS.

ARGUMENT	VALUE
STRATEGY	"RANDOMDISCRETE"
MAX_MODELS	150
MAX_RUNTIME_SECS	3600
STOPPING_METRIC	"DEVIANCE"
STOPPING_ROUNDS	20
STRATEGY	"RANDOMDISCRETE"
MAX_MODELS	150

MODEL	PARAMETER RANGE
RANDOM FOREST (RF)	NTREES ∈ {50, 100, 150,…, 1000}
	MTRIES ∈ {1, 2, 3, 4, 5, 6, 7, 8}
	SAMPLE_RATE ∈ {0.20, 0.25, 0.30,…, 1}
GRADIENT BOOSTING MODEL (GBM)	NTREES ∈ {50, 100, 150,…, 1000}
	MAX_DEPTH ∈ {2, 3, 4, 5, 6, 7, 8}
	LEARN_RATE ∈ {0.001, 0.01, 0.1}
	SAMPLE_RATE ∈ {0.20, 0.30, 0.40,…, 1}
	COL_SAMPLE_RATE ∈ {0.20, 0.30, 0.40,…, 1}

APPENDIX C – POISSON AND GAMMA DEVIANCE

Poisson Deviance is defined as:

$$D = 2\sum\left[y_i\, log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

where y_i= Actual Response Variable,

$\hat{\mu}_i$= Fitted Response

Gamma Deviance is defined as:

$$2\sum\left[-\,log\left(\frac{y_i}{\mu_i}\right)+\frac{(y_i-\mu_i)}{\mu_i}\right]$$

where y_i= Actual Response Variable,

μ_i= Fitted Response

For more information about Milliman India, please visit us at:

in.milliman.com

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Rachin Aggarwal
rachin.aggarwal@milliman.com

Lalima Chakravarty
lalima.chakravarty@milliman.com