BAREKENG: Jurnal Ilmu Matematika dan Terapan
March 2023 Volume 17 Issue 1 Page 0135–0146
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0135-0146

BAYESIAN ADDITIVE REGRESSION TREE APPLICATION FOR PREDICTING MATERNITY RECOVERY RATE OF GROUP LONG-TERM DISABILITY INSURANCE

Stevanny Budiana[1], Felivia Kusnadi[2*], Robyn Irawan[3]

[1,2,3]*Center for Mathematics and Society, Department of Mathematics, Faculty of Information Technology and Science, Parahyangan Catholic University*
Jl. Ciumbuleuit No. 94, Bandung, 40141, Indonesia

Corresponding author's e-mail: [*]felivia@unpar.ac.id

ABSTRACT

Bayesian Additive Regression Tree (BART) is a sum-of-trees model used to approximate classification or regression cases. The main idea of this method is to use a prior distribution to keep the tree size small and a likelihood from data to get the posterior. By fixing the tree size as small as possible, the approximation of each tree would have a little effect on the posterior, which is the sum of all output from all the trees used. The Bayesian additive regression tree method will be used for predicting the maternity recovery rate of group long-term disability insurance data from the Society of Actuaries (SOA). The decision tree-based models, such as Gradient Boosting Machine, Random Forest, Decision Tree, and Bayesian Additive Regression Tree model, are compared to find the best model by comparing mean squared error and program runtime. After comparing some models, the Bayesian Additive Regression Tree model gives the best prediction based on smaller root mean squared error values and a relatively short runtime.

Article History:

Received: 29th August 2022
Revised: 22nd November 2022
Accepted: 17th January 2023

Keywords:

Bayesian Additive Regression Tree;
Maternity Recovery Rate;
Prior;
Sum-of-Trees.

How to cite this article:

S. Budiana., F. Kusnadi and R. Irawan. "BAYESIAN ADDITIVE REGRESSION TREE APPLICATION FOR PREDICTING MATERNITY RECOVERY RATE OF GROUP LONG-TERM DISABILITY INSURANCE," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0135-0146, March 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

There was an increase in global disability cases by 52% in 2017 compared to 1990, with the majority of disabilities caused by non-communicable diseases such as respiratory infection, cancer, digestive disease, diabetes, musculoskeletal disorders, etc. Institute for Health Metrics and Evaluation (IHME) and The World Health Organization (WHO) also said that the frequency of people with disability is highest in people ranging from 20 to 70 years old. By gender, women are more susceptible to disability than men. One of the reasons why a woman has higher disability cases is pregnancy. Women are more likely to experience disability because of high blood pressure, high blood sugar, and higher body mass index or BMI [1].

Disability insurance is one kind of protection given to the insured if the insured suffers from a certain condition that leads the insured to be unable to work and receive income. The benefit given to the insured is paid periodically per the insurance contract. The disability insurance that will be used for case study is group long-term disability insurance; this insurance gives protection to a group of people from one institute or company for a long time of coverage duration. The amount of benefit is around 60% to 80% of the insured's monthly income and the coverage duration is also diverse from 2 years to pension.

The data used for the case study is Group Long-Term Disability (GLTD) insurance year 2008 from the Society of Actuaries website. The data consist of several disabilities causes such as cancer, diabetes, nervous system, maternity, etc. Of several disability causes, the one to be used is maternity disability. Disability because of maternity can happen in the form of disease or injury that might happen while still in pregnancy or even after giving birth and could last for several months to a lifetime. Most maternity disability case usually comes from excessive bleeding, anemia, infection, damage to organ, hypertension, and depression [2].

A company that will release a disability insurance product needs to know the risks they will bear when some insured makes a claim. The risks can be approximated by predicting the insured's recovery rate if the insured is declared to have suffered an injury or disease that caused disability. The recovery rate is predicted by using maternity disability data and the method will be used to predict the recovery rate is Bayesian Additive Regression Tree (BART). The Bayesian Additive Regression Tree method is a sum-of-trees model, which means the prediction is a sum of outputs from each tree, as illustrated by Tan [3].

The Bayesian Additive Regression Tree method is superior to the other decision tree-based methods. The method uses prior distribution to form a tree so the time to generate a tree is quicker than the others, said to give better prediction accuracy, and the approximated value is close to the real values [4]. Bayesian Additive Regression Tree uses the same prior distribution and step to form a tree as Bayesian Classification and Regression Tree [5] but uses several trees rather than using only one tree. Using several trees at once gives an improvement in accuracy, avoids overfitting, and gives smaller variance than just one tree [6], which is also the main idea of Random Forest [7]. Bayesian Additive Regression Tree also uses several iterations to update the current tree based on information from the former tree, which is similar to how the Gradient Boosting Machine method works [8]. The structure for each tree in every iteration will be formed by doing Bayesian Backfitting Markov Chain Monte Carlo [9] and the predicted recovery rate after the burn-in period using Bayesian Backfitting Markov Chain Monte Carlo will converge to the real value [10]. Bayesian Additive Regression Tree can be used to find an important variable by making variable selection with some threshold such as local threshold, global max threshold, and global SE threshold [11].

Besides Bayesian Additive Regression Tree, other methods will also be used to predict maternity recovery rates, such as Decision Tree, Random Forest, and Gradient Boosting Machine. Those methods will be compared to find a model that gives the most accurate and efficient result by comparing the Root Mean Square Error value and the program runtime for each model. The Decision Tree method is the base of all the methods that will be used in this paper, it only uses one tree decision to predict the recovery rate and when the tree size is too big, there is a chance that the variance of prediction value is also big [6]. To minimize the variance for prediction, we used Random Forest, as explained before, and used several trees to predict the recovery rate. We could also use Gradient Boosting Machine to predict the recovery rate; note that Gradient Boosting Machine will give a better prediction if the chosen loss function and parameters are right [12]. And lastly, we will use Bayesian Additive Regression Tree to predict the recovery rate. The method is said to give better results by using some prior distribution for parameters and using Bayesian Backfitting Markov Chain Monte Carlo to determine tree size.

Some datasets will also be used for building a model; each dataset will consist of different data and have its restrictions. The reason for making many datasets is that the original data for maternity recovery rate

consist of a very large amount of 0 and 1 value. The enormous amount of 0 value might lead to an inaccurate predicted value. After building models on each dataset, it will be proven that Bayesian Additive Regression Tree gives the best prediction results and the analysis of the model will also be shown.

2. RESEARCH METHODS

The data used in this paper is group long-term disability insurance data from the year 2008 and come from the Society of Actuaries (SOA) website [13]. The case study will be done by using only maternity disability data and models will be built by using a tree decision-based method to predict the recovery rate. The root mean squared error and runtime for each model will also be compared to find the best model. Some data will be removed, such as recovery rate with 0 and 1 values, male data, and unknown values.

2.1. Data Description

The group long term disability insurance data consists of several variable such as:

Table 1. Variables Description

Variable Name	Description	Value
Actual Recovery Rate	The insured's recovery rate	[0,1]
Age	The age of insured's in years	[20,70]
Duration	Recovery time in months	[2,36]
Gender	Gender or insured's	Male, Female
Gross Indexed Benefit Amount	Gross monthly benefit	$1,000, $1,000-$1,999, $2,000-$2,999, $3,000-$3,999, $4,000-$4,999, $5,000-$9,999, $10,000-$14,999, $15,000-$19,999, $20,000 and over, Unknown
Integration with STD	Integration between long-term insurance with short-term insurance	Integrated with ASO of Fully-Insured STD, Not Integrated with STD, Unknown
Own Occupation to Any Occupation Transition	A change from own occupation to any occupation	Own+1, Own+0, OwnOther
Taxability Benefit	Taxable benefit from a claim	Non-Taxable, Partial Taxability, 100% Taxable

Own occupation is a condition where someone with a disability unable to work in their current profession. While any occupation means someone with a disability unable to work in any other occupation. Variables described in Table 1 will be proceeded so observed data are from female population and data containing unknown values will be removed. The dependent variable for case study is the actual recovery rate and the independent variables are other variables besides the actual recovery rate. We also use several datasets because the original data contain a large amount of 0 and 1 values. Although data with a large amount of 0 value is said to have no big impact on the tree model [13], we found indications that those values still affect the prediction result and make a biased prediction. The dataset after removing male and unknown values will be called dataset 1. The second dataset, called dataset 2, consists of maternity data without 0 value. Dataset 3 consist of maternity data without 0 and 1 values and dataset 4 consist of maternity data without 1 value. For each dataset, 80% of the data will be classified as training data and the others will be classified as test data [14]. The amount of data for each data can be seen in the following **Table 2**.

Table 2. Dataset for Case Study

Dataset	Value	Number of Data	Training data	Test data
1	[0,1]	6.178	4.942	1.236
2	(0,1]	2.241	1.792	449
3	(0,1)	1.957	1.565	392
4	[0,1)	5.894	4.715	1.174

Decision tree-based models will be built for each dataset from Table 2, and the root mean squared error and runtime for each model will be compared to find the best method and prove that Bayesian Additive Regression Tree gives better prediction accuracy than other methods [4].

2.2. Bayesian Additive Regression Tree

Bayesian Additive Regression Tree is an approximation method using a sum-of-trees model. Sum-of-trees model is an additive model with multivariate components. The main idea to build this sum-of-trees model is by choosing a particular prior distribution so the tree structure will stay small through many iterations. By keeping the tree size as small as possible, the sum-of-trees model will consist of many trees where each tree will have a small effect on the final output approximation and explain just a little information from the data. Sum-of-trees model is denoted by

$$\boldsymbol{Y} = \sum_{j=1}^{m} g_j(x) + \varepsilon$$

with Y symbolizing the output or predictions for the dependent variable, m is the amount of tree, $g_j(x)$ is the information from jth tree, and ε is the error or residual assumed to be normally distributed $N(0, \sigma^2)$. For sum-of-trees model, there are some parameters used to describe the tree structures denoted as T_j and explain a set of values on each terminal node on jth tree denoted as $M_j = \{\mu_{1j}, \mu_{2j}, \dots, \mu_{bj}\}$ with b is the amounts of terminal nodes at one tree. The sum-of-trees model after knowing the tree structure and value on each terminal node is denoted by

$$\boldsymbol{Y} = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon \quad (1)$$

The prior for his method is assumed to be independent and symmetric so that the prior can be written as

$$p((T_1, M_1), \dots, (T_m, M_m), \sigma) = \left[\prod_j \prod_i p\left(\mu_{ij} \middle| T_j\right) p(T_j)\right] p(\sigma) \quad (2)$$

By assuming the prior independent and symmetric, the tree components, such as tree structure and value on each terminal node (T_j, M_j), are independent with each other and with σ. Also, each terminal nodes for every tree are independent of each other **[4]**.

2.3. Prior

Prior chosen in this section will affect the tree structure and makes the output of each tree small. Parameters prior that will be used for building the model are tree structure T_j, the value on each terminal node M_j, standard deviation σ, and the amount of tree m. Note that all the parameter chosen in this section is the value recommended after doing several trials and the model built using recommended values is called the default model, while the model built with other values combination by doing cross-validation is called the cross-validation model.

2.3.1. Prior for Tree Structure

To find $p(T_j)$ for **Equation (2)**, we can use

$$\alpha(1+d)^{-\beta}, \ \alpha \in (0,1), \beta \in [0, \infty) \quad (3)$$

Equation (3) is a probability for a node at depth $d = 0, 1, 2, \dots$ is nonterminal. We $\alpha = 0.95$ and $\beta = 2$ as recommended by Chipman **[4]**. By choosing those values, we got prior information that tree sizes of 2 or 3 receive prior probability of 0.55 and 0.28 giving a conclusion that most trees have a depth of 2 or 3. It is also possible for the tree size to be bigger than 3 and grow into a big tree if the data demand so.

2.3.2. Prior for Value on Terminal Nodes

To find $p(\mu_{ij}|T_j)$, we use the conjugate normal distribution $N(\mu_\mu, \sigma_\mu^2)$ as prior distribution. To find the prediction for output which is symbolized as $E(Y|x)$, we sum up m terminal nodes value from all trees so the prediction distribution is $N(m\mu_\mu, m\sigma_\mu^2)$. Also, the $E(x)$ is likely to fall in between y_{min} and y_{max}, the minimum and maximum value of the dependent variable. μ_μ and σ_μ can be found by satisfying these conditions:

$$m\mu_\mu - k\sqrt{m}\sigma_\mu = y_{min},$$

$$m\mu_\mu + k\sqrt{m}\sigma_\mu = y_{max}$$

By choosing $k = 2$, it is said that 95% of prediction will fall in (y_{min}, y_{max}) interval. We also assume $y_{min} = -0.5$ and $y_{max} = 0.5$ and transformed y value to fall inside $(-0.5, 0.5)$. By doing so, we can use $\mu_\mu = 0$ as the center of prior distribution so that the value on each terminal node is small and approximate σ_μ by solving $\sigma_\mu = \frac{0.5}{k\sqrt{m}}$.

2.3.3. Prior for Standard Deviation

To find $p(\sigma)$, we use conjugate inverse chi-square distribution $\sigma^2 \sim v\lambda/\chi_v^2$ as prior distribution. By using a data-informed prior approach, we assign a probability to the plausible values of σ. Let $\hat{\sigma}$ be the prediction of standard deviation, we will choose v and λ so that the qth prior quantile will be located at $\hat{\sigma}$ or $p(\sigma < \hat{\sigma})$. The recommended (v, q) value is (3;0.9). By choosing that combination, we got a distribution where the possible standard deviation value is not too concentrated on small values.

2.3.4. Number of Trees

The value for m is the number of trees used for the sum-of-trees model. The bigger the value is the better the prediction will be, but after some point, the prediction will become stagnant and the performance will get worse. The number of trees to be used is 200.

2.4. Backfitting Markov Chain Monte Carlo and Posterior Inference

The Bayesian Additive Regression Tree method uses Bayesian Backfitting Markov Chain Monte Carlo **[9]** to build a tree structure. The backfitting process will be done 1,250 times to simulate the tree shape until the shapes converge. The tree shapes, values on terminal nodes, variance, and posterior Y at **Equation (1)** will change along with the iterations. The backfitting process uses metropolis-hasting algorithm **[15]** to reject and accept the proposal tree shape using grow, prune, and change procedure to alter the tree shape **[16]**.

From 1,250 iterations, some parts of the iterations will be removed because of their inconsistency or instability. This part is called the burn-in period, where the result of the said period is still oscillating and has yet to reach the point of convergence. We assume the first 250 iterations as the burn-in period and the rest is categorized as the post-burn-in period **[17]**. We will find the posterior Y by analyzing the pattern of the iteration from the post-burn-in period to find the convergence of the posterior. Posterior Y from post burn-in period is denoted by

$$f^*(.) = \sum_{j=1}^{m} g(.; T_j^*, M_j^*),$$

with $f^*(.)$ is the post-burn-in period, m is the number of trees, T_j^* is tree structure post- burn-in period, and M_j^* is value on terminal nodes post-burn-in period. Backfitting is used to find the next tree structure based on the current tree structure. Assume that we already know the structure of previous $m-1$ trees, assumes $T_{(j)}$ and $M_{(j)}$ as information about previous trees and we want to know the mth tree structure based on previous information, then the mth tree structure can be denoted as $(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y$.

Standard deviation σ is assumed to be drawn from the inverse gamma distribution and (T_j, M_j) depends on partial residuals denoted as

$$R_j \equiv y - \sum_{k \neq j} g(x; T_k, M_k),$$

which is the residual between observed data and $m-1$ previous trees. Also, the tree structure T_j can be obtained by using the metropolis-hastings algorithm and the value on terminal nodes can be found after T_j is known.

2.5. Other Decision Tree-Based Method

Besides Bayesian Additive Regression Tree, other methods such as decision tree, random forest, and gradient boosting machine are also used to compare the performance of each method and find a model with the best performance. Each method will be used to build a model on every dataset explained in this subsection.

2.5.1. Decision Tree

A decision tree uses only one tree to make a prediction. By using only one tree, the size of the tree will be big, and the variance of prediction will be large. Predictions using a decision tree can be approximated using

$$\widehat{f_i}(x) = \frac{1}{N_i} \sum_{n_i=1}^{N_i} y_{n_i}(x),$$

with $\widehat{f_i}(x)$ is prediction and equal to the average of data in nodes terminal, N_i is the total amount of data on ith terminal node, and y_{n_i} is the data value on ith terminal node.

2.5.2. Random Forest

Random Forest uses more than one tree to make a prediction. By using more than one tree, the accuracy of prediction is much better and results in smaller prediction variance than the decision tree method. The final prediction for random forest is the average of prediction from each tree and is calculated by

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x),$$

with $\hat{f}(x)$ is the final prediction, B is the number of trees used, and $\hat{f}_b(x)$ is the prediction of bth tree.

2.5.3. Gradient Boosting Machine

Same as random forest, the gradient boosting machine also uses more than one tree. Though, rather than using many trees at once like the random forest method, the gradient boosting machine uses only one tree for every iteration and uses the recent trees to improve the next tree. Prediction using gradient boosting machine is the sum of weighted prediction from each iteration. The formula for prediction is

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}_b(x),$$

with $\hat{f}(x)$ is the final prediction, B is the number of iterations, λ is the shrinkage constant, and $\hat{f}_b(x)$ is the prediction from bth iteration.

2.6. Root Mean Squared Error

To know whether a model gives an accurate prediction, we use root mean squared error (RMSE) to test the accuracy. Root mean squared errors, as its name already states, is the root of mean squared error that is usually used to calculate accuracy between observed data and prediction. We use RMSE and not MSE to calculate accuracy because the data used is varied around 0 and 1. The smaller the RMSE value is, the better the accuracy will be. The root mean squared error can be calculated using

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2},$$

with $RMSE$ is the root mean squared error value, n is the amount of data, y_i is the observed data, and $\hat{f}(x_i)$ is the prediction for observed data.

2.7. Model Analysis

Model analyses that will be used for Bayesian Additive Regression Tree are normality test, heteroscedasticity test, and variable selection.

2.7.1. Normality Test

A normality test is used to determine whether a model is built using normally distributed data or not. The hypothesis used for this test are:

H_0 : The model is built using normally distributed data.

H_1 : The model is built not using normally distributed data.

Other than the hypothesis test, the normality test can also be done using a qq plot [18]. Quantile-quantile plot or qq plot is a plot between the quantile of observed data and the theoretical quantile. If the points in the qq plot form a straight line, then the model is built using normally distributed data. If the line is not straight, then we can do some transformation to the dependent variable and build a new model using the transformed data.

2.7.2. Heteroscedasticity Test

The heteroscedasticity test is used to determine the value of error variance. Heteroscedasticity is a condition where the error for each prediction is irregular or not constant. For a model to be called a good model, the heteroscedasticity test shall be rejected to indicate that the prediction errors are constant or homoscedastic [6]. When a model has a heteroscedastic error variance, we need to build a new model that satisfies the null hypothesis. The hypothesis used for the heteroscedasticity test are:

H_0 : Error variance is constant.

H_1 : Error variance is not constant.

Other than the hypothesis test, the visual test can be done using a scatter plot. If the scatter plot is dispersed without any pattern, then the error variance is constant. On the contrary, if the scatter plot shows any pattern, such as increasing, decreasing, or expanding, then the error variance is not constant.

2.7.3. Variable Selection

Variable selection is used to find some variables that are significant to the prediction. The significant variable can be determined by the variable inclusion proportion denoted as p_K. The variable inclusion proportion or p_K is the proportion for Kth variable. Variable selection for the Bayesian Additive Regression Tree model uses the permuted dependent variable as the response variable rather than just using a normal dependent variable. Permuted data is used to remove the initial correlation between a dependent and independent variables that might exist when using the original data [11].

The significant variable is chosen by using the free-model approach to find a correlation between predictions and the dependent variable after being permuted [4]. The model-free approach uses backward elimination to find the best prediction by eliminating insignificant variables one by one [19]. After the inclusion proportion is found, the next step to do is to find the threshold for each variable. There are several thresholds, such as local threshold, global max threshold, and global SE threshold, with each of them having different stringency. A variable with a proportion value exceeding the threshold is called a significant variable. For different data, the threshold to be used is different and can be chosen by using a cross-validation procedure [11].

3. RESULTS AND DISCUSSION

3.1. Case Study Result

The case study is done by building decision tree-based models such as Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), and Bayesian Additive Regression Tree (BART and BART-CV) using several datasets. For each model, the root mean squared error (RMSE) and the runtime will be shown and compared to find a model with the best performance.

Table 3. RMSE of Every Model on Different Dataset

RMSE	Dataset 1	Dataset 2	Dataset 3	Dataset 4
DT	0.2215	0.2180	0.1415	0.1329
RF	0.2327	0.3125	0.1353	0.1330
GBM	0.2199	0.2150	0.1394	0.1318
BART	0.2167	0.1906	0.1266	0.1263
BART-CV	0.2169	0.1920	0.1277	0.1260

From Table 3, the model with the best performance for every dataset is either BART or BART-CV. This also proves that the Bayesian Additive Regression Tree method gives better prediction accuracy than other methods [4].

Table 4. Runtime for Every Model on Different Dataset in Minutes

RMSE	Dataset 1	Dataset 2	Dataset 3	Dataset 4
DT	0.0303	0.0474	0.0445	0.0244
RF	23.5676	5.6763	2.7333	20.8417
GBM	0.1784	0.2556	0.2194	0.1648
BART	0.6999	0.4958	0.4992	0.7011
BART-CV	68.6231	22.9564	22.3574	82.0315

From **Table 4**, we can see that model BART-CV takes much time to build or is computationally expensive, and so we conclude that BART is the model with the best performance because not only does it give a smaller RMSE than other models, but it also takes relatively short times to build. By comparing the RMSE of each model, we got those models on dataset 3 and dataset 4 give similar RMSE, but if we also compare the runtime, dataset 4 takes more time than dataset 3. So, for model analysis, we will use the BART model on dataset 3.

3.2. Bayesian Additive Regression Model Analysis

The model to be analyzed is the BART model from dataset 3 with parameter combination $(k, q, v, m) = (2, 3, 0.9, 50)$ or the default model. The R Squared value is 0.5371. For residual analysis, the p-value for the normality test using the Shapiro-Wilk test is 0, and the p-value for zero-mean noise is 0.96884. Based on the p-value and visual test in **Figure 1**, we conclude that the model is not built using normally distributed data and the error variance is constant. The explanation for the normality test in subsection 2.7.1 says that when the data is not normally distributed, transformation on observed data should be done, but after doing transformation, the result still does not improve much, so we will use the normal data without any transformation.

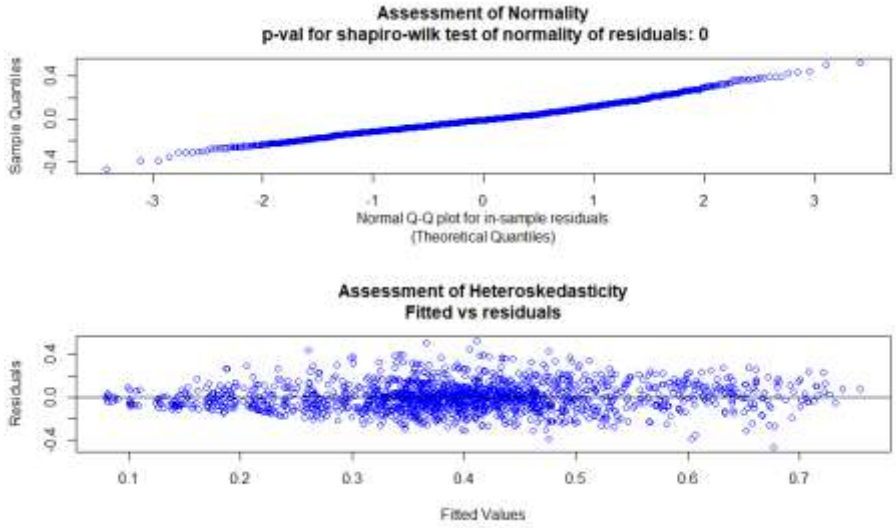

Figure 1. Residual or Error Analysis for BART Model

Figure 2a is partitioned into five parts by grey vertical lines where the left part is the burn-in period and the other four is the post-burn-in period. The blue horizontal line is the average error variance which converges to 0.017. **Figure 2b** is also divided into two parts where the left part is the burn-in period and the right part is the post-burn-in period. It shows the percentage of acceptance from grow, prune, and change proposals with initial probability of 28%, 28%, and 44% sequentially **[16]**. **Figure 2c** illustrates the number of nodes for each tree on every iteration after the burn-in period, with the blue line being the average of nodes on every tree which is 5 nodes. Lastly, **Figure 2d** is the depth for every tree after the burn-in period and the blue line is the average depth which is 2.

Figures 3 and **4** show the coverage probability of credible interval and prediction interval **[20]** for the BART Model. The coverage probability of the credible interval is 38.34%, which means 38.34% of observed

data will fall inside the credible interval. The small coverage probability for credible intervals might be because the data is not normally distributed. The coverage probability of the prediction interval is 95.59% means 95.59% of the predicted future value will fall inside the prediction interval.

Figure 5 shows the variable inclusion proportion for every independent variable from highest to lowest proportion. Duration and age band are the variables mostly used as splitting variables and followed by other variables. These variables are deemed as material and need to be incorporated into the underwriting process. To process this even further, we investigate which splitting variables are chosen due to the correlation between variables; we make variable selection and the result is shown in **Figure 6**.

In **Figure 6a**, the green line denotes the local threshold, the red line indicates the global maximum threshold, and the blue line signifies the global SE threshold, whereas in **Figure 6b**, solid dots represent the variable inclusion proportion exceeds all the threshold, hollow dots mean the variable inclusion proportion does not exceed all the threshold, and the asterisk mark the variable inclusion proportion exceeds the global SE threshold. The thresholds chosen after doing cross-validation is the local threshold and the independent variable that significantly affects the prediction are duration, $4,000-$4,999, $<1,000, and $1,000-$1,999.

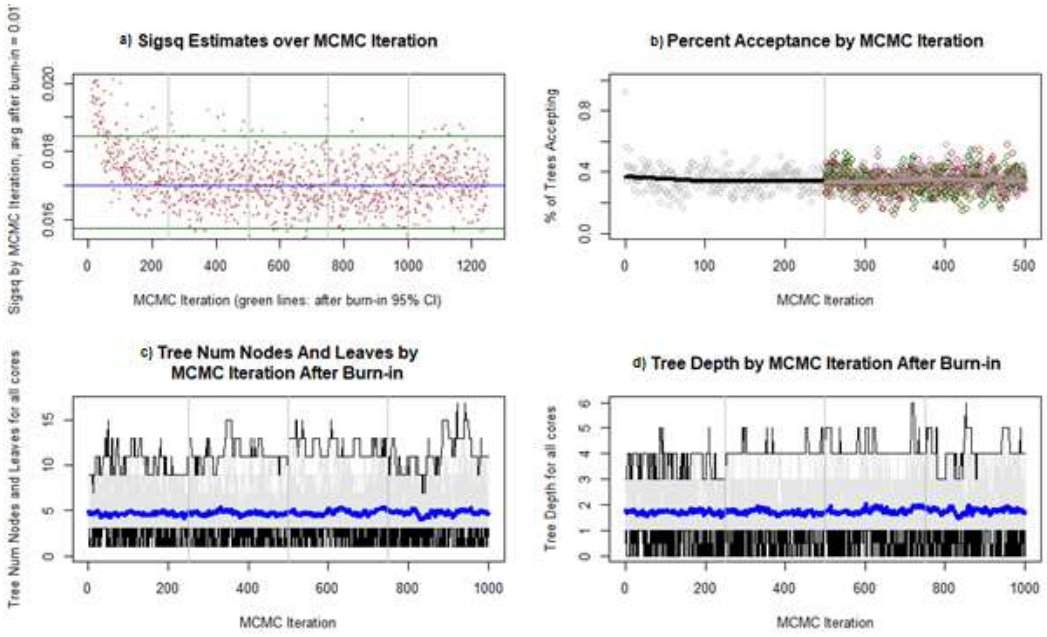

Figure 2. Convergence of BART Model

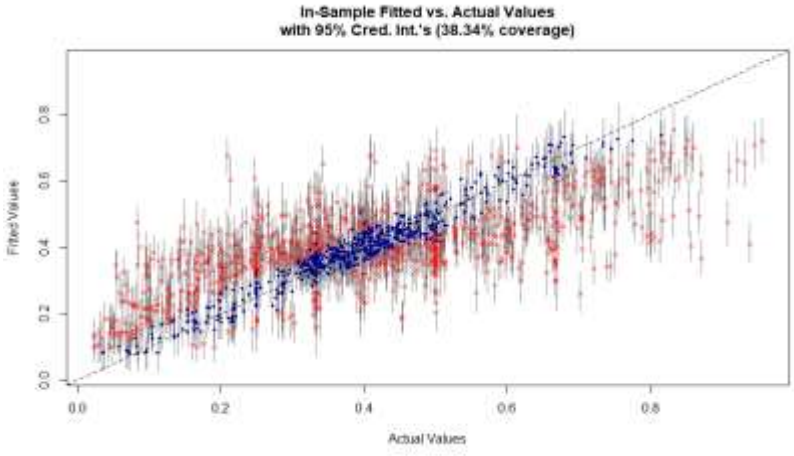

Figure 3. Credible Interval for BART Model

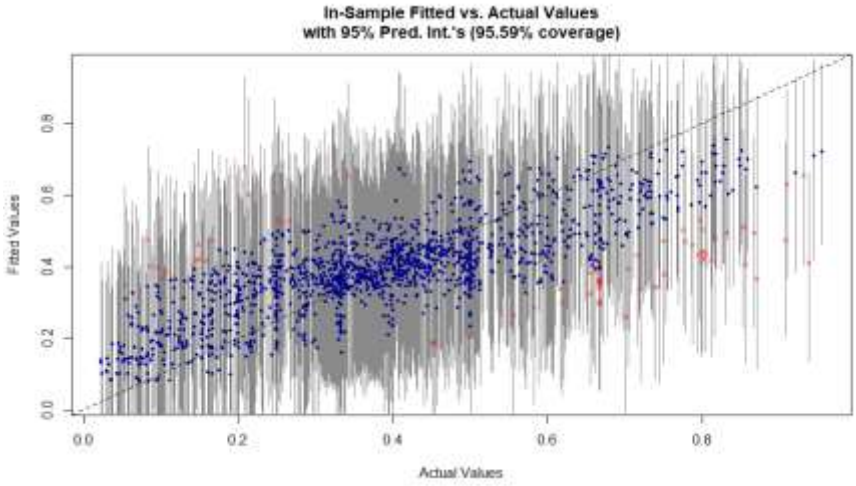

Figure 4. Prediction Interval for BART Model

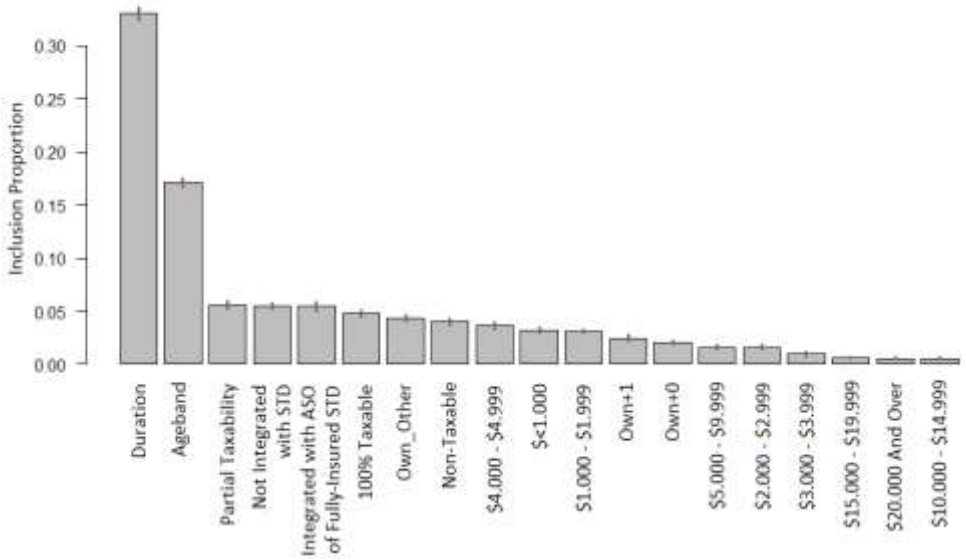

Figure 5. Variable Importance for BART Model

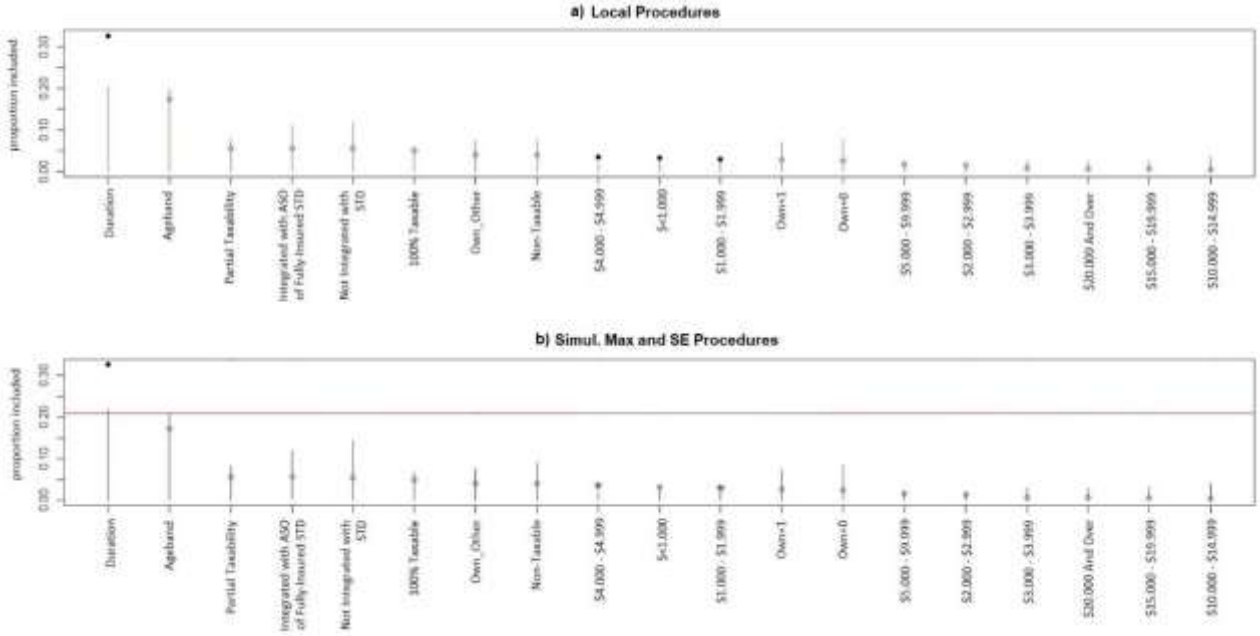

Figure 6. Variable Selection for BART Model

4. CONCLUSIONS

Based on the results obtained, the lowest RMSE values are mostly produced by the BART model. The computational time given is also relatively low. These signify the BART model is the best predictor for this given case. Variable importance and variable selection shown by the BART model also indicate which variables are deemed material to be incorporated in the underwriting process, which are duration, age, and gross indexed benefit amount.

REFERENCES

[1] I. for H. M. and Evaluation, "Findings from the Global Burden of Disease Study 2017," *Online*, 2018. www.healthdata.org.
[2] C. AbouZahr, "Global Burden of Maternal Death and Disability," *Br. Med. Bull.*, vol. 67, no. 1, pp. 1–11, 2003, doi: 10.1093/bmb/ldg015.
[3] J. Tan, Y.V.; Roy, "Bayesian Additive Regression Trees and The General BART Model," *Stat. Med.*, vol. 38, no. 25, pp. 5048–5069, 2019.
[4] R. E. Chipman, H.A.; George, E.I.; McCulloch, "BART: Bayesian Additive Regression Trees," *Ann. Appl. Stat.*, vol. 6, no. 1, pp. 266–298, 2012.
[5] R. E. Chipman, H.A.; George, E.I.; McCulloch, "Bayesian CART Model Search," *J. Am. Stat. Assoc.*, vol. 93, no. 443, pp. 935–948, 1998.
[6] R. Gareth, J.; Witten, D.; Hastie, T.; Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. New York: Springer, 2013.
[7] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[8] E. Klug, Maximiliano; Barash, Yiftach; Bechler, Sigalit; Resheff, Yehezkel S; Tron, Talia; Ironi, Avi; Soffer, Shelly; Zimlichman, Eyal; Klang, "A Gradient Boosting Machine Learning Model for Predicting Early Mortality in The Emergency Department Triage: Devising A Nine-Point Triage Score," *J. Gen. Intern. Med.*, vol. 35, no. 1, pp. 220–227, 2020.
[9] R. Hastie, T.; Tibshirani, "Bayesian Backfitting," *Stat. Sci.*, vol. 15, no. 3, pp. 196–223, 2000.
[10] G. Chopin, N.; Ducrocq, "Fast Compression of MCMC Output," *Entropy*, vol. 23, no. 8, p. 1017, 2021.
[11] S. T. Bleich, J.; Kapelner, A.; George, E.I.; Jensen, "Variable Selection for BART: An Application to Gene Regulation," *Ann. Appl. Stat.*, vol. 8, no. 3, pp. 1750–1781, 2014.
[12] J. Diana, A.; Griffin, J. E.; Oberoi, J. S.; Yao, "Machine-Learning Methods for Insurance Applications – A Survey," 2019.
[13] M. Kopinsky, "Predicting Group Long Term Disability Recovery and Mortality Rates using Tree Models," 2017. [Online]. Available: https://www.soa.org/globalassets/assets/Files/Research/Projects/2017-gltd-recovery-mortality-tree.pdf.
[14] K. P. Murphy, *Machine Learning A Probabilistic Perspective*. Massachusetts: MIT Press, 2012.
[15] C. L. V. Lawson, A. B.; Browne, W. J.; Rodeiro, *Disease Mapping with WinBUGS and MLwiN*. John Wiley & Sons, 2003.
[16] J. Kapelner, A.; Bleich, "bartMachine: Machine Learning with Bayesian Additive Regression Trees," 2013.
[17] J. Bleich, *Extensions and Applications of Ensemble-of-trees Methods in Machine Learning*. University of Pennsylvania, 2015.
[18] S. Ghasemi, A.; Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians," *Int. J. Endocrinol. Metab.*, vol. 10, no. 2, pp. 486–489, 2012.
[19] C. J. Li, L.; Cook, R.D.; Nachtsheim, "Model-Free Variable Selection," *J. R. Stat. Soc.*, vol. 67, no. 2, pp. 285–299, 2005.
[20] B. T. Hespanhol, L.; Vallio, C.S.; Costa, L.M.; Saragiotto, "Understanding and Interpreting Confidence and Credible Intervals Around Effect Estimates," *Brazilian J. Phys. Ther.*, vol. 23, no. 4, pp. 290–301, 2019.