JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 7, No. 4, October 2023, pp. 1221-1236

Prediction of Maternity Recovery Rate of Group Long-Term Disability Insurance Using XGBoost

Felivia Kusnadi[1*], Andry Wijaya[2], Julius Dharma Lesmono[3]
[1,2,3]Center for Mathematics and Society, Department of Mathematics, Parahyangan Catholic University, Bandung, Indonesia
felivia@unpar.ac.id[1], andrywijaya@gmail.com[2], jdharma@unpar.ac.id[3]

Article History:
Received : 27-07-2023
Revised : 13-10-2023
Accepted : 18-10-2023
Online : 19-10-2023

Keywords:
Imbalanced Data;
Maternity Recovery Rate;
XGBoost;
Variable Importance.

ABSTRACT

To help insurers determine insurance rates incorporating maternity factors, it is crucial to understand the maternity recovery rate, which was a metric used by insurance companies to understand how much of the expenses associated with maternity care and related medical services are covered by their policies. This paper employed Extreme Gradient Boosting (XGBoost), a powerful method for handling complex data relationships and preventing overfitting, on North American Group Long-Term Disability dataset obtained from the Society of Actuaries, which listed maternity as one of its categories, to predict the maternity recovery rate. In comparison, other machine learning methods such as Gradient Boosting Machine (GBM) and Bayesian Additive Regression Tree (BART) were used, with Root Mean Squared Error (RMSE) values calculated the difference between predicted and observed maternity recovery rates. Four datasets, 3 imbalanced and 1 fairly-balanced, were created out of the original dataset to test each method's predictive prowess. The study revealed that XGBoost performed exceptionally well on the imbalanced datasets, while BART showed slight superiority in fairly-balanced data. Furthermore, the model identified the duration, exposures, and age of participants in both predicting maternity recovery rates and the underwriting process.

https://doi.org/10.31764/jtam.v7i4.16825

This is an open access article under the CC–BY-SA license

A. INTRODUCTION

Disability insurance is gaining significance on a global scale due to factors, such as an aging population, individuals encountering disabilities after a certain event, or having a family member requiring medical care and attention due to the disability condition (Haberman, S.; Pitacco, 2018). In developed countries such as the US, there has been a notable rise in the proportion of working-age group benefiting from the federal Disability Insurance (DI) program (Deshpande, M.; Lockwood, 2022). This share has escalated from 2.2% in the late 1970s to 4.6% in 2013 (Liebman, 2015). More information is required to compute the necessary premium for disability insurance, one of which is decrement rates. However, constructing decrement rates of a sample drawn from a certain population are rigorous and time-consuming processes (Fong, J. H.; Shao, A. W.; Sherris, 2015; Kopinsky, 2017; London, 1982).

The Institute for Health Metrics and Evaluation (IHME) and The World Health Organization (WHO) had reported that the highest prevalence of disability occurs among individuals aged 20 to 70 years old. Additionally, women are more vulnerable to disabilities compared to men,

partly due to factors such as pregnancy (Budiana et al., (2023). William et al. (2019); William et al. (2018) found that adverse births significantly increase the risk of expensive out-of-hospital expenses related to delivery and postnatal care due to complications for both the baby and mother. This risk applies both during childbirth and the postpartum period. Furthermore, White et al. (2022) concluded that age is one of the main factors of the US maternal mortality economic burden. As a person's age increases, there is a greater economic burden associated with maternal complications. This may be the result of older women tend to face more complex health issues related to pregnancy and childbirth, or it could be related to factors such as the financial burden on the family if the mother dies during or after childbirth. Age, in this context, is highlighted as a key factor in understanding the economic implications of maternal mortality.

The data used for this study is the 2008 Group Long Term Disability (GLTD) recovery data from the Society of Actuaries (SOA) website. There are numerous categories of disabilities, such as disabilities to the mental and nervous system, back, digestive system, respiratory system, and musculoskeletal, as well as cancer, diabetes, and maternity, etc. Kopinsky (2017) did exploratory data analysis and found that the pattern shown by the maternity differed from the rest of the disability categories. Thus, we dug deeper into the maternal disability category. Maternity-related disability can manifest because of illnesses or injuries occurring during pregnancy or after childbirth, potentially lasting from several months to a lifetime. The majority of such cases are linked to complications like excessive bleeding, infections, organ damage, hypertension, and can be linked from depression (AbouZahr, 2003).

Healthcare datasets commonly exhibit highly imbalanced data, where the majority and minority classifiers lack balance, leading to inaccurate predictions when processed by the classifiers. Another prevalent characteristic of healthcare datasets is the presence of missing values (Jothi et al., 2015). In most statistical software, tree-based methods are preferred due to their robustness against unbalanced datasets Hassan et al. (2016); Krawczyk et al. (2014); Singhal et al. (2018), such as disease detection and fraud diagnosis which mainly exist in classification problems. These methods adapt well to the imbalances by splitting data based on features, making them more robust and versatile. Decision trees and random forests were used in predicting the maternity recovery rates provided on the SOA GLTD dataset obtained by Kopinsky (2017), with random forests resulted in worse prediction by using MSE as its evaluation metric. Gradient Boosting Machine (GBM) and Bayesian Additive Regression Tree (BART) were also used in Budiana et al. (2023) to the same dataset, with BART being the best model to predict the outcome by using RMSE as its evaluation metric as the values of the maternity recovery rates are real numbers ranging between 0 and 1. Thus, by incorporating RMSE instead of MSE, the error values are more amplified and enlarged.

Despite being a strong predictor, GBM often suffers from overfitting. As a remedy, Chen et al. (2016) introduced a more recent version known as Extreme Gradient Boosting (XGBoost). It is a machine learning algorithm that employs an ensemble approach, utilizing decision trees within a gradient boosting machine (GBM) framework. This combination allows XGBoost to achieve excellent model performance and high-speed processing. Liu et al. (2023) utilized XGBoost to enhance the predicted outcomes developed from conventional machine learning algorithms and resulted in an increased F1-score of 6.13%. On imbalanced dataset such as

personal credit evaluation, XGBoost performed better than the other tree-based models and logistic regression (Li et al., 2020).

In another studies, XGBoost performed classifications better than other machine learning models. XGBoost performs better than Support Vector Machines (SVM) in discriminating certain diseases in patients from healthy controls, using confusion matrix as its evaluation metric (Binson et al., 2021; Ogunleye et al., 2019). On the soil liquefaction prediction, whose data are sampled using different techniques, the study found that XGBoost perform better than Random Forests and SVM Demir et al. (2022), data undergoing transformation Sahin (2023), better than random forest and gradient boosting machine on landslide data using RMSE as the evaluation metric Sahin (2020), better than logistic regression, Bayesian Additive Regression Tree (BART), random forest, and SVM on tumor classification problem Zhang et al. (2023), better than SVM and K-nearest neighbour (KNN) on company bankruptcy classification problem Muslim et al. (2021), and on surface water flooding data that stated XGBoost had a better generalization ability than SVM to improve prediction accuracy (Wang et al., 2021).

XGBoost and other tree models were commonly employed for classifying diseases, mapping geography, and predicting bankruptcy. However, their usage in solving regression problems, especially in insurance-related scenarios, has been limited. This study aims to develop tree-based models to predict regression outcomes, specifically maternity recovery rates based on input data. These recovery rates play a vital role in determining the inclusion of maternity care costs in insurance policies and calculating group health insurance premiums. The dataset used contains maternity recovery rates ranging from 0 to 1, with "0" indicating no recovery and "1" indicating full recovery. The dataset is divided into four parts to analyze the impact of these two dominant values. Each section is further split into training and testing data. Training data are used to build models for each section, while testing data assesses the models' effectiveness. The models used in this research include XGBoost, GBM, and BART. Additionally, due to its complicated nature, we identify the significant variables that contribute to the establishment of the tree model as shown in (Quan, Z.; Valdez, 2018). These material variables are then used as important factors to predict the recovery rates, as well as in ensuring the underwriting process performs better in selecting potential groups to insure and to correctly determine premiums of the group insurance policies.

B. METHODS

The methodology of the research conducted in this paper is as shown in the flowchart in Figure 1. Initially, we gather and pre-process the dataset. Next, we divide the data into four separate subsets. For each of these subsets, we further divide them into training and testing sets. We then train each tree-based model on the training data and tune their hyperparameters to minimize the RMSE on the testing data. Once we've obtained satisfactory results, we proceed to compare the models and select the best-performing tree-based model. Additionally, we assess variable importance to identify the key factors that contributed to the construction of the most effective model, as shown in Figure 1.

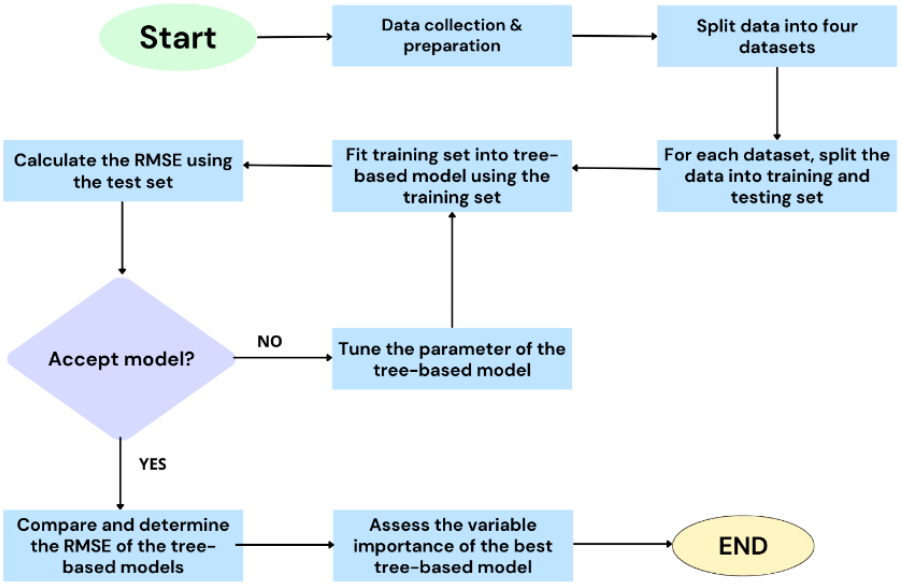

Figure 1. Flowchart of the Methodology

1. Data Source

This paper uses the GTLD recovery data, excerpted from (Kopinsky, 2017). The initial dataset, sourced from the 2004 - 2012 North American Group Long Term Disability (GLTD) Database and comprising 46 million records across 25 companies, underwent filtration by the author. This filtering process reduced the data to 818,941 rows, aimed at improving execution times and avoiding inefficiencies associated with processing the entire dataset. The author specifically extracted subsets of data that contained the essential variables required for the model. The disability categories comprised of "Back", "Cancer", "Circulatory", "Diabetes", "Digestive", "Ill-defined and Misc Conditions", "Injury other than back", "Maternity", "Mental and Nervous", "Nervous System", "Other", "Other Musculoskeletal", and "Respiratory". The recovery rate of maternity, the yellow-greenish line, shows huge discrepancies from the other disabilities, as shown in Figure 2. The recovery rate at ages of 35 to 60 shows significant progress. This makes sense since women give birth during their reproductive years, which is around 15 years old. Young women tend to not have complications of childbirth, whereas women with older ages are more prone to aftereffects of labour. As age increases, the chance of recovering decreases. Similarly, other forms of disabilities are more likely to see recovery at younger ages. However, the ages are typically between 15 and 23 years old, as shown in Figure 2.

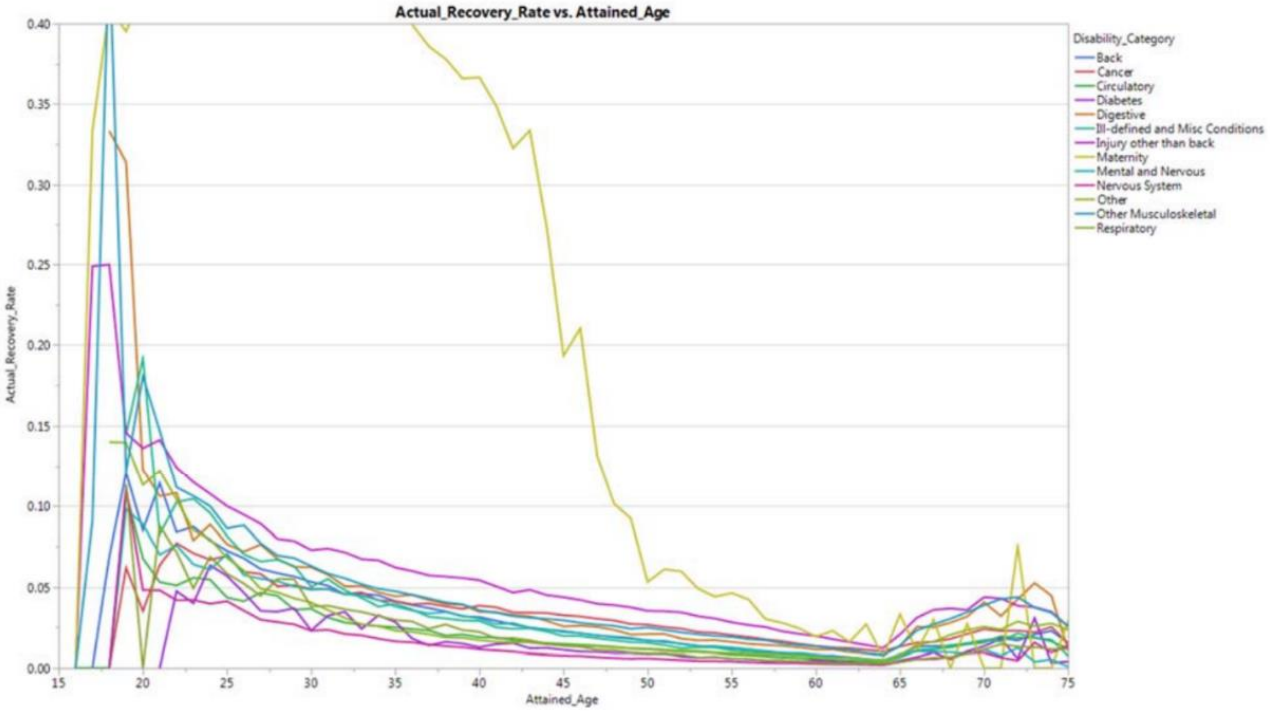

Figure 2. Recovery rates compared to Age (Kopinsky, 2017)

2. Data Preparation

Initially, we conducted further data cleaning by isolating only maternity-related entries and excluding any observation of the "male" gender from the dataset. This dataset is then called Dataset 1. Subsequently, we remove the gender and disability category variables from the dataset as they each only contain one entry, namely "female" and "maternity", respectively. The variable Actual Deaths is also omitted from the dataset as it is unnecessary to compute the mortality probability. The remaining variables descriptions are as follow.

a. Duration_12_49 signifies the participant's time to recovery, ranging from 2 to 49 months.
b. AgeBand which denotes the participant's age banded into groups of 5, from 20 to 70-year-old.
c. OwnOccToAnyTransition_MOD categorizes the type of change from own occupation (the initial job before disability occurred) to any occupation (the ongoing work that is being undertaken due to the disability occurrence).
d. Integration_with_STD denotes whether the participant's current plan includes integration with short-term disability insurance (Contreary, K.; Ben-Shalom, Y.; Gifford, 2018).
e. Taxability_Benefits include income or financial benefits that may or may not be taxed. Some benefits are tax-exempt, reducing the recipient's tax burden, while others are taxable and need to be reported as part of the taxable income.
f. Gross_Indexed_Benefit_Amount is the categorical variable that represents the grouping of the original benefit amounts of the insurance policy which gradually increases overtime to adjust for the cost of living and inflation.
g. Exposures refer to the typical actuarial measure of participants in the dataset. The values are not whole numbers due to the inclusion of fractional years, and adjustments

have been applied to mitigate the overwhelming influence of extremely large contributing companies on the data.

h. Actual_Recoveries represent the number of participants who have recovered within the dataset. To prevent the dominance of the results by the largest companies' experiences, a dampening factor is applied to their data, leading to non-integer recovery figures.

Additionally, we created a new variable called Actual_Recovery_Rate, which is the rate at which participants died, i.e., Actual_Recoveries divided by Exposures, ranging from 0 to 1. In this context, when the probability value is "0," it signifies that the participant is unable to recover, whereas a probability value of "1" indicates the participant's certain recovery. Dataset 1 consists of massive amounts of “0” and “1” values, which creates imbalanced in the dataset. Therefore, we filter out the 0 values and call it Dataset 2. Dataset 3 is constructed by removing the “1” values from Dataset 2. The last dataset is obtained by filtering out the “1” values from Dataset 1, called Dataset 4. The datasets are further examined at Table 1. To construct the models, the data are split into 70% for training observations and the other 30% for testing observations, as shown in Table 1.

Table 1. Datasets for Analysis

Dataset	Value	Observations
1	[0,1]	6,178
2	(0,1]	2,241
3	(0,1)	1,957
4	[0,1)	5,894

3. GBM Model

Boosting, a technique from machine learning, improves the accuracy of a weak classifier by combining multiple instances for better predictions. This approach was applied to statistical modeling with models like AdaBoost and GBM. GBM gradually constructs a strong predictive model by incrementally adding weak learners, often decision trees. Unlike starting with a small "stump" tree, GBM begins with a single leaf as an initial guess for all observations. It then builds larger trees based on previous errors, with pruning to prevent overfitting. GBM continues this process, scaling and creating trees based on errors, until the desired number of trees is reached, or further improvement isn't observed. Additionally, GBM requires a differentiable loss function $L(y_i, F(x_i))$ where y_i is the actual value and $F(x_i)$ is the predicted value of x_i, which is used to assess how well the model is performing. It can be viewed as a measurement of the total error resulted from the model (Friedman, 2001, 2002). The loss function used is least-squares function:

$$L(y_i, F(x_i)) = \frac{1}{2}[y_i - F(x_i)]^2. \quad (1)$$

The loss function is minimized using the gradient descent technique, in which the local minimum of a function is calculated using gradients by Algorithm 2.1 as follows.

Algorithm 2.1. Gradient Boosting

Input:
A training dataset $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, containing $X_i = (x_{i,1}, \ldots, x_{i,p})^T$ which denotes the p predictors, the responses Y_i, the number of iterations M, and the learning rate ω.

Steps:

a. Initialize the model with a constant value: $F_0(x) = \bar{Y}$.
b. For $m = 1, \ldots, M$ do:
   1) For $i = 1, \ldots, n$, compute:
   $$r_{i,m} = -\left[\frac{\partial L(Y_i, F_{m-1}(X_i))}{\partial F_{m-1}(X_i)}\right] = -\frac{\partial}{\partial F_{m-1}(X_i)}\left[\frac{1}{2}[Y_i - F_{m-1}(x_i)]^2\right] = Y_i - F_{m-1}(x_i)$$
   2) Fit a regression tree to the $r_{i,m}$ values and create split regions $R_{j,m}$ for $j = 1, \ldots, J_m$
   3) For $j = 1, \ldots, J_m$, compute
   $$\gamma_{j,m} = \min_{\gamma} \sum_{X_i \in R_{j,m}} L(Y_i, F_{m-1}(X_i) + \gamma) = \frac{1}{n} \sum_{X_i \in R_{j,m}} Y_i - F_{m-1}(x_i)$$
   4) Update with learning rate ω: $F_m(X) = F_{m-1}(X) + \omega \sum_{j=1}^{J_m} \gamma_{j,m} \cdot 1\{X \in R_{j,m}\}$

Output:
The values of $F_m(X)$ for every $m = 1, \ldots, M$.
The outcomes are subsequently fed into the RMSE function. If the outcomes do not meet the desired criteria, adjust the parameters by experimenting with various values.

4. XGBoost Model

As mentioned previously, GBM tends to overfit. To remedy this, (Chen, T., & Guestrin, 2016) enhanced the GBM algorithm into a more robust algorithm.

3.1. Tree Ensemble Model

Given a dataset D with n rows of observations and m features that is denoted by $D = \{(\boldsymbol{x_i}, y_i)\}$ with $\boldsymbol{x_i} = (x_{i,1}, \ldots, x_{i,m})^T$ that signifies a predictor variable on i^{th} row of observation and v^{th} feature with $i = 1,2, \ldots, n$ and $v = 1,2, \ldots, m$, meanwhile y_i denotes the response variable on the i^{th} row of observation with $y_i \in \mathbb{R}$. A tree ensemble model uses K additive functions to predict the model:

$$\widehat{y_i} = \theta(\boldsymbol{x_i}) = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}). \quad (2)$$

with f denotes the predictive function on the k$^{\text{th}}$ decision tree.

3.2. Regularized Objective Function

Some of the enhancements include the regularization term, Lasso or Ridge regression (Melkumova, L. E.; Shatskikh, 2017), incorporated into the objective function which reduces overfitting. $L(\theta)$ is a regularized (ridge regression) objective function that measures the performance of the predictive model to assess its accuracy.

$$\mathrm{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k(\boldsymbol{x_i})), \quad (3)$$

$$\Omega(f_k(\boldsymbol{x_i})) = \gamma T + \frac{\lambda}{2} \cdot ||f_k(\boldsymbol{x_i})||^2. \quad (4)$$

with l denotes the loss function, T denotes the set of leaves in a tree, λ denotes the hyperparameter regularization, and γ denotes the pseudo-regularization hyperparameter or minimum split loss reduction.

3.3. Loss Function

For outcomes ranging between 0 and 1 (Shen, 2005), the logistic loss function is used:

$$l(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2, \quad (5)$$

3.4. Gradient Tree Boosting

The optimal solution of a tree ensemble model on a regularized objective function is obtained using additive model. Assume that the predictive value $\hat{y}_i^{(K)}$ is the prediction in the i^{th} observation and K^{th} tree, defined as follows:

$$\hat{y}_i^{(K)} = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}) = \hat{y}_i^{(K-1)} + f_K(\boldsymbol{x_i}), \text{ with } \hat{y}_i^{(0)} = 0 \quad (6)$$

Substituting (6) into (3) gives the regularized objective function on the K^{th} tree:

$$L^{(K)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(K-1)} + f_K(\boldsymbol{x_i})) + \Omega(f_K(\boldsymbol{x_i})). \quad (7)$$

Algorithm 2.2. Extreme Gradient Boosting

Input:

A training dataset $D = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)\}$, containing $\boldsymbol{x_i} = (x_{i,1}, \ldots, x_{i,m})^T$ which denotes the m predictors, the responses y_i, the number of iterations K, hyperparameter regularization λ, pseudo-regularization hyperparameter γ, and the learning rate ω.

Steps:

a. Initialize the model with a constant value: $\hat{y}_i^{(0)} = 0.5$. The initial value must not be the mean value of the data.
b. For $i = 1, \ldots, n$ do:
   1) Compute the first derivative of the loss function:
   $$g_i = -(y_i - \hat{y}_i^{(K-1)})$$
   2) Compute the second derivative of the loss function, $h_i = 1$.
   3) Determine the best splitting point candidate by splitting the data into percentile or quartile.
   4) Construct regression tree that consists of g_i on each leaf node.
   5) Calculate the value of the loss reduction after splitting for every splitting point until the best splitting point is gained, which has the maximum L_{split}, the splitting process is stopped when L_{split} value is negative or when there remains only one g_i on the leaf:

$$L_{split} = \frac{1}{2}\left[\sum_{j=1}^{T}\frac{(\sum_{i\epsilon I_{left}} g_i)^2}{\lambda + \sum_{i\epsilon I_{left}} h_i} + \sum_{j=1}^{T}\frac{(\sum_{i\epsilon I_{right}} g_i)^2}{\lambda + \sum_{i\epsilon I_{right}} h_i} - \sum_{j=1}^{T}\frac{(\sum_{i\epsilon I} g_i)^2}{\lambda + \sum_{i\epsilon I} h_i}\right] - \gamma T$$

6) Calculate the value of the optimal weight,
7)

$$w_j * = -\frac{\sum_{i\epsilon I_j} g_i}{\lambda + \sum_{i\epsilon I_j} h_i}.$$

8) Update with learning rate ω: $\widehat{y_i}^{(K)} = \widehat{y_i}^{(K-1)} + \omega \cdot w_j *$.

Output:
The values of $\widehat{y_i}^{(K)}$ for every $i = 1, \ldots, n$.
The results are then entered into the RMSE function, and if they fail to meet the desired standards, you should tune the parameters by testing different values.

5. Root Mean Squared Error

The root mean squared error (RMSE) evaluates the performance of a specified predictive model, measuring the average differences between the actual and predicted values. The RMSE is calculated as follows (Wang, W.; Lu, 2018):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}. \quad (8)$$

with n is the number of observations, y_i is the actual value, and $\widehat{y_i}$ is the predicted value. A more favourable model is indicated by a smaller RMSE value. The RMSE metric is used as the output values from our model ranges from 0 to 1, with “0” and “1” values dominating the dataset. Initially, we applied MSE to the model, but the results showed very small values. Consequently, we opted for RME to magnify the error values, as it is important to understand that taking the square root of a fraction ranging from 0 to 1 amplifies the outcome. Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE) calculate the difference between actual and predicted values divided by the actual value. These metrics are not suitable for evaluation as the dataset contains a dominant value of "0." Mean Error (ME) and Mean Absolute Error (MAE) are similar in nature with RMSE, thus it is pointless to use these metrics.

C. RESULT AND DISCUSSION

1. Model Run on the Datasets

The datasets are run through the XGBoost, GBM, and BART algorithms by the R software. After conducting several parameters tuning to obtain the lowest RMSE value, the resulting optimal XGBoost tree models are shown in Figures 4, 5, 6, and 7 for each dataset. Branches with the symbol (< split value) mean "yes" for the associated splitting criteria, while those without it mean "no." There were three components on the leaf node, namely cover, gain, and value. *Cover* is the total of the second derivatives of the loss function on the training dataset classified

to the leaf. The deeper the node of the tree, the smaller the value of the cover is. *Gain* represents the metric for loss function reduction by performing a split into the internal node. *Value* denotes the mean of the target variable if the observation ends up in that associated leaf node. To obtain the maternity recovery rate, the *Value* from the leaf node needs to be transformed using the sigmoid function to have $\hat{y}_i$ values ranging from 0 to 1, as in the original dataset.

For illustration purpose, let's examine the two lowest leaf nodes of the 99th tree of the XGBoost model in Figure 3. If the AgeBand is greater than 67.5, the bottom leaf node has *Cover* of 1.86761832 and *Value* of -0.0592818633. On the other hand, if the AgeBand falls between 62.5 and 67.5, the second lowest leaf node has *Cover* of 1.84421849 and *Value* of -0.103858821. Consequently, if an observation lands on the bottom leaf node, the predicted maternity recovery rate is approximately $\frac{1}{1+e^{-0.0592818633}} = 0.514816127$. If it's assigned to the second lowest leaf node, the resulting maternity recovery rate is roughly $\frac{1}{1+e^{-0.103858821}} = 0.525941391$, as shown in Figure 3, Figure 4, Figure 5, and Figure 6.

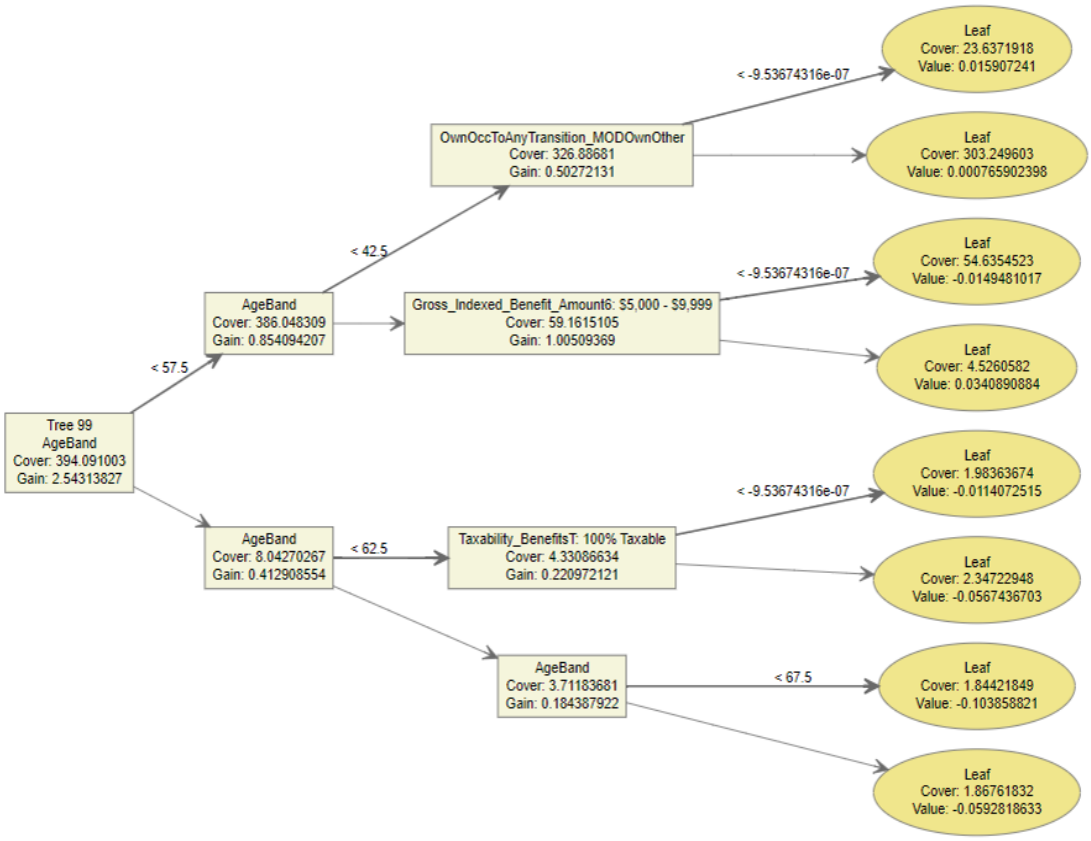

Figure 3. Non-optimal XGBoost model for Dataset 1

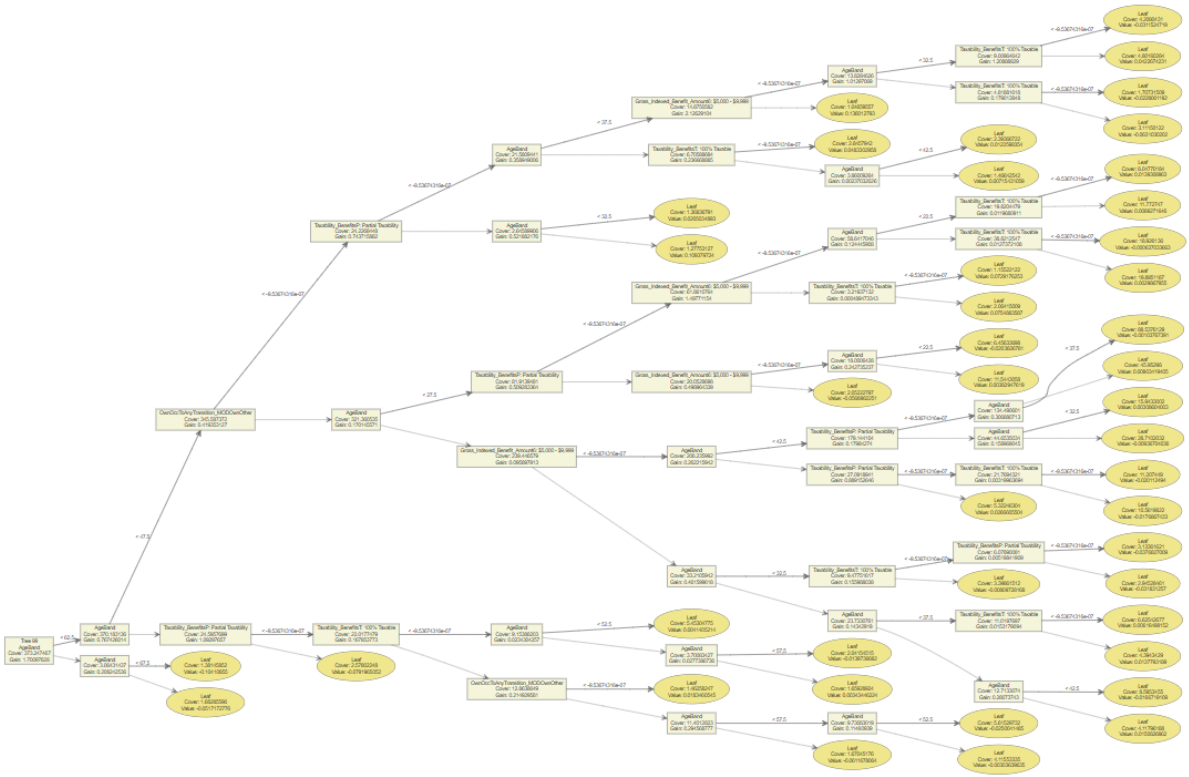

Figure 4. Optimal XGBoost tree model for Dataset 1

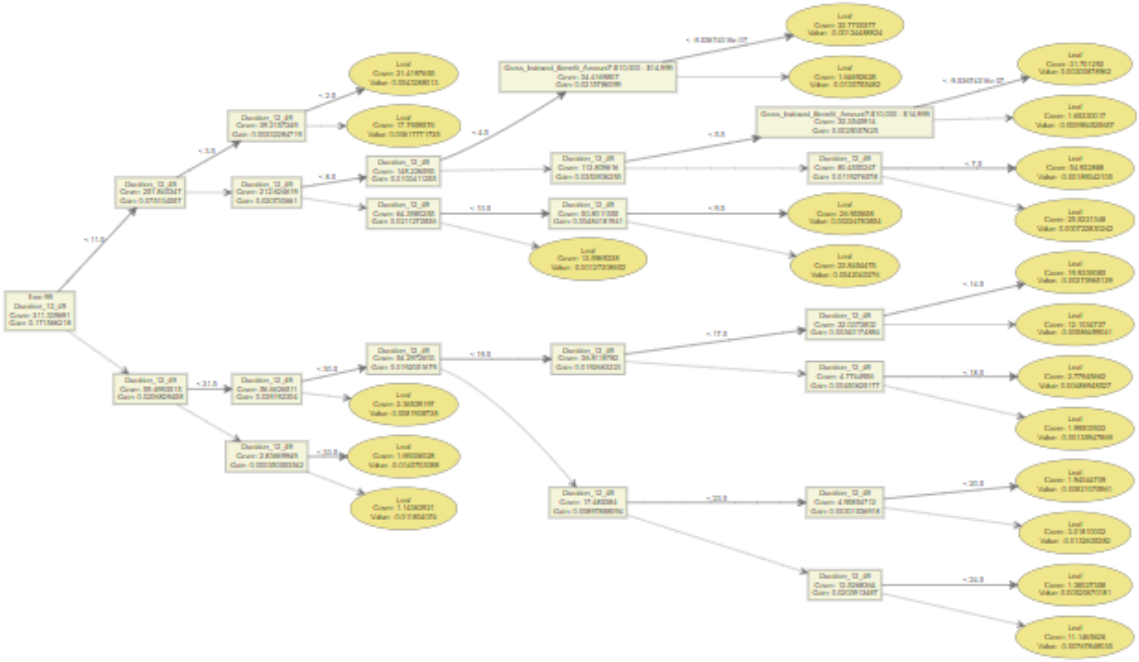

Figure 5. Optimal XGBoost model of Dataset 2

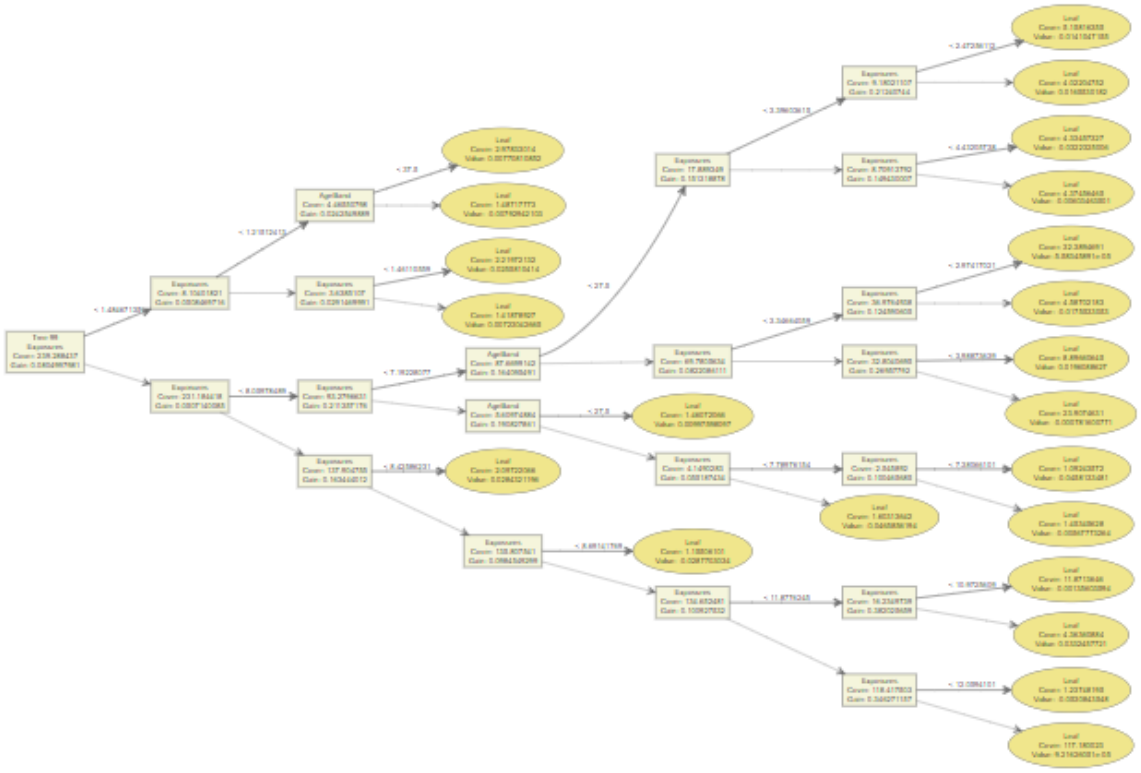

Figure 6. Optimal XGBoost model of Dataset 3

RMSE of the models are shown in Table 2, with RMSE for GBM and BART are explained in (Budiana, S.; Kusnadi, F.; Irawan, 2023). RMSE values shown by the four datasets indicate that XGBoost outperforms both models, except in the third dataset. Dataset 1 contains many observations with both "0" and "1" values. Dataset 2 is predominantly comprised of "1" values, while Dataset 4 mainly consists of "0" values. XGBoost performs exceptionally well on imbalanced data, whereas BART is most effective on Dataset 3 which comprises of fairly-balanced data, in which the all the majority values of “0” and “1” had been eliminated, leaving behind values between 0 and 1, as shown in Figure 7 and Table 2.

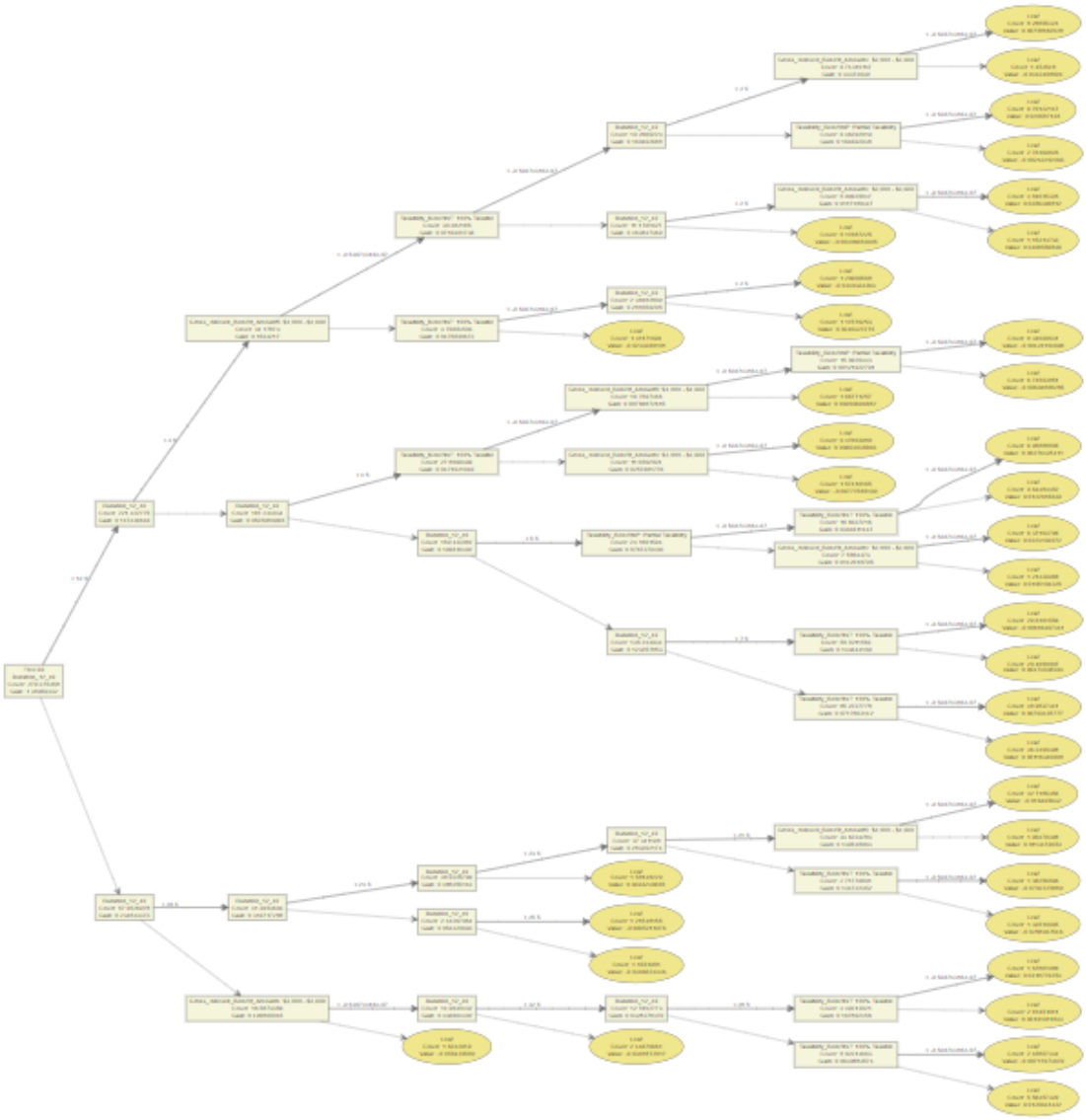

Figure 7. Optimal XGBoost model of Dataset 4

Table 2. RMSE comparison of the three models

RMSE	Dataset 1	Dataset 2	Dataset 3	Dataset 4
XGBoost	0.2113	0.1386	0.1349	0.1208
GBM	0.2199	0.2150	0.1394	0.1318
BART	0.2167	0.1906	0.1266	0.1263

2. Variable Importance of the XGBoost Model

As XGBoost is a model with high predictive capability, it sacrifices its interpretability capacity due to its complexity nature. To overcome this model opacity, one of the most renown methods is by analysing variable importance (Wei, P.; Lu, Z.; Song, 2015). By using this, it analyses the model structure and ranks the information gain by each feature. From Figure 2, the *gain* metric is the information gained for each split which denotes the significance of a certain node. *Cover* is as described before. *Frequency*, also known as the weight, signifies the percentage indicating the proportion of occurrences of a specific feature among all the trees in the model.

In practical terms, we are applying Dataset 1 for observation. Duration becomes the biggest influencer in the model with 44% information gain, while Exposures and AgeBand are the other contributing factors for the model. All the other features are considered insignificant to the contributing factor of the model. The three variables are entirely logical as the participant's recovery time, the number of participants being observed, and each of the participant's age all significantly contribute to the prediction of maternity recovery rate. In particular, the age variable is in line with the study by White et al. (2022), although not necessarily the most important, as shown in Figure 8.

Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
Duration_12_49	0.44131543	0.23691839	0.28041757
Exposures	0.24685292	0.24227606	0.28347732
AgeBand	0.14149940	0.15497782	0.12383009
Integration_with_STDN: Not Integrated with STD	0.02164182	0.03923190	0.04535637
OwnOccToAnyTransition_MODOwnOther	0.02152752	0.03353784	0.01241901
Taxability_BenefitsT: 100% Taxable	0.02134593	0.03130189	0.04409647

Figure 8. Variable importance of the Dataset 1 model

D. CONCLUSION AND SUGGESTIONS

This paper examines the pre-processing of maternity recovery data before applying three machine learning techniques to the dataset. Among the three methods, XGBoost demonstrates superior performance, particularly in handling imbalanced data commonly found in health data, as indicated by its lowest RMSE values. Moreover, variables such as duration, exposures, and age are considered crucial factors in addressing product innovation and the underwriting process. To develop this research further, one can prepare the data better by applying a combination of oversampling and undersampling to remedy the imbalance nature of the data, i.e. by using ROSE and SMOTE (Selamat, N.A.; Abdullah, A.; Diah, 2022). One can also improve the prediction rate by using deep learning methods.

REFERENCES

AbouZahr, C. (2003). Global Burden of Maternal Death and Disability. *British Medical Bulletin*, *67*(1), 1–11. https://doi.org/10.1093/bmb/ldg015

Binson, V. A.; Subramoniam, M.; Sunny, Y.; Mathew, L. (2021). Prediction of Pulmonary Diseases with Electronic Nose Using SVM and XGBoost. *IEEE Sensors Journal*, *21*(18), 20886–20895. https://doi.org/10.1109/JSEN.2021.3100390

Budiana, S.; Kusnadi, F.; Irawan, R. (2023). Bayesian Additive Regression Tree Application for Predicting Maternity Recovery Rate of Group Long-Term Disability Insurance. *Barekeng: Jurnal Ilmu Matematika Dan Terapan*, *17*(1), 135–146. https://doi.org/10.30598/barekengvol17iss1pp0135-0146

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Contreary, K.; Ben-Shalom, Y.; Gifford, B. (2018). Using Predictive Analytics for Early Identification of Short-Term Disability Claimants who Exhaust Their Benefits. *Journal of Occupational Rehabilitation*, *28*, 584–596. https://doi.org/10.1007/s10926-018-9815-5

Demir, S.; Şahin, E. K. (2022). Liquefaction Prediction with Robust Machine Learning Algorithms (SVM, RF, and XGBoost) Supported by Genetic Algorithm-Based Feature Selection and Parameter Optimization from the Perspective of Data Processing. *Environmental Earth Sciences*, *81*(18), 459. https://doi.org/10.1007/s12665-022-10578-4

Deshpande, M.; Lockwood, L. M. (2022). Beyond Health: Nonhealth Risk and the Value of Disability Insurance. *Econometrica*, *90*(4), 1781–1810. https://doi.org/10.3982/ECTA19668

Fong, J. H.; Shao, A. W.; Sherris, M. (2015). Multistate Actuarial Models of Functional Disability. *North American Actuarial Journal*, *19*(1), 41–59. https://doi.org/10.1080/10920277.2014.978025

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, *29*(5), 1189–1232. https://www.jstor.org/stable/2699986

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Haberman, S.; Pitacco, E. (2018). *Actuarial Models for Disability Insurance*.

Hassan, A. K. I.; Abraham, A. (2016). Modeling Insurance Fraud Detection Using Imbalanced Data Classification. *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015)*, 117–127. https://doi.org/10.1007/978-3-319-27400-3_11

Jothi, N.; Husain, W. (2015). Data Mining in Healthcare–A Review. *Procedia Computer Science*, *72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Kopinsky, M. (2017). Predicting Group Long Term Disability Recovery and Mortality Rates using Tree Models. In *Society of Actuaries*. https://www.soa.org/globalassets/assets/Files/Research/Projects/2017-gltd-recovery-mortality-tree.pdf

Krawczyk, B.; Woźniak, M.; Schaefer, G. (2014). Cost-Sensitive Decision Tree Ensembles for Effective Imbalanced Classification. *Applied Soft Computing*, *14*, 554–562. https://doi.org/10.1016/j.asoc.2013.08.014

Li, H.; Cao, Y.; Li, S.; Zhao, J.; Sun, Y. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems*, *35*(3), 52–61. https://doi.org/10.1109/MIS.2020.2972533

Liebman, J. B. (2015). Understanding the Increase in Disability Insurance Benefit Receipt in the United States. *Journal of Economic Perspectives*, *29*(2), 123–150. https://doi.org/10.1257/jep.29.2.123

Liu, J.; Xu, K.; Cai, B.; Guo, Z. (2023). Fault Prediction of On-Board Train Control Equipment Using a CGAN-Enhanced XGBoost Method with Unbalanced Samples. *Machines*, *11*(1), 114. https://doi.org/10.3390/machines11010114

London, R. L. (1982). An Overview of Actuarial Decrement Rate Estimation. *Actuarial Research Conference of the Society of Actuaries*, *17*, 1–10. https://www.soa.org/globalassets/assets/library/research/actuarial-research-clearing-house/1978-89/1983/arch-2/arch83v23.pdf

Melkumova, L. E.; Shatskikh, S. Y. (2017). Comparing Ridge and LASSO Estimators for Data Analysis. *Procedia Engineering*, *201*, 746–755. https://doi.org/10.1016/j.proeng.2017.09.615

Muslim, M.A.; Dasril, Y. (2021). Company Bankruptcy Prediction Framework Based on the Most Influential Features Using XGBoost and Stacking Ensemble Learning. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(6), 5549–5557. https://doi.org/10.11591/ijece.v11i6.pp5549-5557

Ogunleye, A.; Wang, Q. G. (2019). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(6), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

Quan, Z.; Valdez, E. A. (2018). Predictive Analytics of Insurance Claims Using Multivariate Decision Trees. *Dependence Modeling*, *6*(1), 377–407. https://doi.org/10.1515/demo-2018-0022

Sahin, E. K. (2020). Assessing the Predictive Capability of Ensemble Tree Methods for Landslide Susceptibility Mapping Using XGBoost, Gradient Boosting Machine, and Random Forest. *SN Applied Sciences*, *2*(7), 1308. https://doi.org/10.1007/s42452-020-3060-1

Sahin, E. K. (2023). Implementation of Free and Open-Source Semi-Automatic Feature Engineering Tool in Landslide Susceptibility Mapping Using the Machine-Learning Algorithms RF, SVM, and XGBoost. *Stochastic Environmental Research and Risk Assessment*, *37*(3), 1067–1092. https://doi.org/10.1007/s00477-022-02330-y

Selamat, N.A.; Abdullah, A.; Diah, N. M. (2022). Association Features of SMOTE and ROSE for Drug Addiction Relapse Risk. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7710–7719. https://doi.org/10.1016/j.jksuci.2022.06.012

Shen, Y. (2005). *Loss Functions for Binary Classification and Class Probability Estimation*. https://www.proquest.com/openview/ff8caed03c746ebca2d686ec5b385710/1?pq-origsite=gscholar&cbl=18750&diss=y

Singhal, Y.; Jain, A.; Batra, S.; Varshney, Y.; Rathi, M. (2018). Review of Bagging and Boosting Classification Performance on Unbalanced Binary Classification. *IEEE 8th International Advance Computing Conference (IACC)*, 338–343. https://doi.org/10.1109/IADCC.2018.8692138

Wang, W.; Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conference Series: Materials Science and Engineering*, *324*, 012049. https://doi.org/10.1088/1757-899X/324/1/012049

Wang, X.; Fu, D.; Wang, Y.; Guo, Y.; Ding, Y. (2021). The XGBoost and the SVM-Based Prediction Models for Bioretention Cell Decontamination Effect. *Arabian Journal of Geosciences*, *14*, 1–11. https://doi.org/10.1007/s12517-021-07013-6

Wei, P.; Lu, Z.; Song, J. (2015). Variable Importance Analysis: A Comprehensive Review. *Reliability Engineering & System Safety*, *142*, 399–432. https://doi.org/10.1016/j.ress.2015.05.018

White, R. S.; Lui, B.; Bryant-Huppert, J.; Chaturvedi, R.; Hoyler, M.; Aaronson, J. (2022). Economic Burden of Maternal Mortality in the USA. *Journal of Comparative Effectiveness Research*, *11*(13), 927–933. https://doi.org/10.2217/cer-2022-0056

William, J.; Chojenta, C.; Martin, M. A.; Loxton, D. (2019). An Actuarial Investigation Into Maternal Out-of-Hospital Cost Risk Factors. *Annals of Actuarial Science*, *13*(1), 1–35. https://doi.org/10.1017/S1748499518000015

William, J.; Martin, M. A.; Chojenta, C.; Loxton, D. (2018). An Actuarial Investigation Into Maternal Hospital Cost Risk Factors for Public Patients. *Annals of Actuarial Science*, *12*(1), 106–129. https://doi.org/10.1017/S174849951700015X

Zhang, Y.; Wang, J.; Liang, B.; Wu, H.; Chen, Y. (2023). Diagnosis of Malignant Pleural Effusion with Combinations of Multiple Tumor Markers: A Comparison Study of Five Machine Learning Models. *The International Journal of Biological Markers*, *38*(2), 03936155231158125. https://doi.org/10.1177/03936155231158125