International Journal of Innovative Research and Scientific Studies, 8(4) 2025, pages: 1013-1021

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Predictive modeling of right-skewed health insurance costs using copula regression and ensemble learning

Dimas Anggara[1,2], Khairil Anwar Notodiputro[1*], Indahwati[1], Anang Kurnia[1]

[1]*School of Data Science, Mathematics and Informatics, IPB University, Bogor 16680, Indonesia.*
[2]*Directorate of Price Statistics, Statistics Indonesia, BPS Jakarta 10710, Indonesia.*

Corresponding author: Khairil Anwar Notodiputro (*Email: khairil@apps.ipb.ac.id*)

Abstract

This study compares four predictive models in the context of a response variable characterized by a right-skewed, non-symmetric distribution, specifically health cost insurance data. The modeling approaches employed include copula-based models (copula regression with logarithmic transformation and standard copula regression) and ensemble learning methods (Random Forest/RF and Extreme Gradient Boosting/XGBoost). The health cost data was partitioned into 80% for training and 20% for testing. Model fitting was conducted using the training data, while model evaluation was performed using the testing data. The performance of each model was assessed based on several evaluation metrics: Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Median Absolute Deviation (MAD). Additionally, this study includes an analysis of outlier prediction, where the constructed models were utilized to predict outliers within the health cost data. The results of the study indicate that the copula regression model with logarithmic transformation is more suitable for response variables exhibiting non-symmetric, right-skewed distributions, such as health expenditure data. This is evidenced by the low values of the MAD and MAPE metrics. Another key finding is that the copula regression and XGBoost models demonstrate superior performance in predicting outliers compared to the other two models evaluated in this study.

Keywords: Copula regression, Extreme gradient boosting, Individual medical cost, Random forest, Skewness.

DOI: 10.53894/ijirss.v8i4.7986
Funding: This study received no specific financial support.
History: Received: 14 April 2025 / **Revised:** 20 May 2025 / **Accepted:** 22 May 2025 / **Published:** 20 June 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

The statistical issue addressed in this study is the violation of the normality assumption in linear regression analysis. Violating the normality assumption in regression analysis affects the validity of statistical inference. The assumption of normality in the error terms allows the use of statistical tests such as the F-test and t-test for hypothesis testing on model parameters. If the error terms do not follow a normal distribution, the results of these statistical tests become invalid. Furthermore, violating the normality assumption can lead to inaccurate confidence interval estimates, meaning that the intervals may not accurately reflect the actual uncertainty.

Although the assumption of normality is crucial, real-world data often deviates from a normal distribution [1]. Examples of such data include income, expenditures, waiting times, number of visits, number of traffic accidents, length of hospital stays, product sales over a given period, number of likes and comments on social media, and many other datasets that tend to be skewed rather than normally distributed. A common approach to handling non-normally distributed data is the use of transformation methods, such as logarithmic, square root, inverse, and other transformations. The purpose of these transformations is to adjust non-normal data so that it approximates a normal distribution [2, 3].

In addition to transformation methods, statistical modeling approaches that do not require the normality assumption can also be employed. These models include copula regression, random forests, and XGBoost, among others. Copula regression is a statistical method that incorporates the concept of copulas to model dependencies between variables [4]. This method has been developed by several researchers in the fields of statistics and mathematics. One of the pioneers of copula theory is Roger B. Nelsen, who authored the book An Introduction to Copulas [5], a fundamental reference in the field. Additionally, Genest has contributed extensively to research on copulas and their applications in statistics, including copula regression [6-9].

On the other hand, machine learning techniques offer alternative approaches to regression modelling that can address violations of the normality assumption and account for nonlinear relationships between variables. Several machine learning methods have been developed specifically for regression analysis. The first published regression tree algorithm was Automatic Interaction Detection (AID), which applies a recursive data partitioning method to create a piecewise-constant model [10]. AID later evolved into THAID; the first classification tree algorithm designed to maximize the number of cases in the dominant category [11]. Subsequently, Chi-squared Automatic Interaction Detection (CHAID) was introduced, utilizing the Chi-squared test to determine the most significant splits, Kass [12]. Breiman [13] introduced the Classification and Regression Trees (CART) algorithm, which became the foundation of many modern decision tree methods [13]. CART employs a greedy search and cross-validation to determine the optimal splits and prune the tree. Ross Quinlan later developed the ID3 [14] and C4.5 [15] algorithms, which use entropy-based criteria to select the best splits and have served as the basis for numerous modern decision tree algorithms.

Breiman further developed Random Forest, an ensemble learning method that aggregates multiple decision trees to improve predictive accuracy and mitigate overfitting Breiman et al. [16]. Loh [17] introduced GUIDE (Generalized, Unbiased, Interaction Detection and Estimation), which addresses some limitations of earlier algorithms and allows for linear separation within subsets of variables [17]. More recently, XGBoost (Extreme Gradient Boosting) has emerged as a highly efficient and fast boosting algorithm. XGBoost utilizes optimized gradient boosting techniques designed for high performance and scalability while supporting parallel and distributed computing [18].

Given the rapid advancements in regression-based machine learning methods, it is of particular interest to assess their performance in predicting outcomes for skewed distributions. This study aims to model individual health costs covered by health insurance, which are response-distributed, asymmetric, and skewed to the right. Additionally, it compares models-based regression such as copula regression, random forests, and extreme gradient boosting (XGBoost). By evaluating these models, this study seeks to identify their respective advantages and limitations and determine which method provides the most accurate predictions for right-skewed data distributions.

2. Methodology

2.1. Classical Regression

Classical Regression referred to in this study is linear regression, either simple or multiple. Linear regression is a statistical method used to model the relationship between one response variable (Y) and one or more independent variables ($X_1, X_2, \ldots, X_n$). The main purpose of linear regression is to understand how independent variables affect the response variable or to predict the response variable from known independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

Where, Y is the response variable to be predicted, $X_1, X_2, \ldots, X_n$ are independent variables/predictors, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are parameters or coefficients of regression, and ε is an error model.

The estimation of regression parameters, β, uses the Ordinary Least Squares (OLS) method. It should be noted that the response variables used are skewed to the right, so modelling with classical regression in this study cannot be done due to a violation of the normality assumption of errors.

2.2. Copula Regression

Copula regression is a statistical modeling technique designed to handle complex dependencies between variables, particularly when the data exhibits non-normal, skewed, or heavy-tailed distributions. Unlike traditional regression methods that assume a specific joint distribution (often normal), copula regression separates the modeling of marginal distributions

from the dependence structure among variables. This is achieved using a copula function, which links the marginal distributions into a joint distribution, allowing for greater flexibility in capturing nonlinear and asymmetric relationships.

The primary utility of copula regression lies in its ability to model data that violate the assumptions of classical linear regression, such as right-skewed health insurance costs or financial losses. It is especially useful in fields like finance, insurance, and health economics, where extreme values and tail dependencies are common. By accurately capturing the dependence structure and allowing for different types of marginal distributions, copula regression provides more reliable and interpretable predictions in complex real-world scenarios.

Copula regression uses Sklar's theorem to see the relationship between the independent variables and the response variables. The relationship between the distribution function and the copula function is explained by Sklar [19] as follows.

$$F_{Y,X}(y,x) = C\big(F_Y(y), F_X(x)\big) \quad (2)$$

where Y and X are random variables and $C(.)$ It is a copula function. Prediction value can be obtained by applying the concept of the expected value of Y given X. The marginal function of the response variables Y and X is a non-decreasing function. The copula regression model can be written with the following equation.

$$Y = E[Y|X] + (Y - E[Y|X]) \quad (3)$$

$$Y = E[Y|X] + \varepsilon \quad (4)$$

where Y is response variables, X is covariate or independent variables, and ε is error terms.

2.3. Random Forest (RF) Regression

The random forest algorithm is a method that uses several learning algorithms simultaneously and then combines them to obtain more accurate modeling results. This method is an extension of the regression tree method by applying the bootstrap aggregating method and random feature selection. The random forest algorithm can be used for classification and regression problems. If the desired output is a discrete data type, the classification method is used, but if the output is continuous data, the regression method, also known as random forest regression, is used [16]. Random forest makes predictions by combining the results of each regression tree by taking the average value for random forest regression [20].

Random forest begins with many bootstrap samples taken randomly with replacement from the original training data. A regression tree is fitted to each bootstrap sample. Bootstrapping is the process of randomly sampling a subset of the dataset for a certain number of iterations and a certain number of variables. Then, the sample is returned to the dataset so that it can be re-selected in further analysis. A set of input variables is selected from the total set considered randomly as a binary partition for each node in each tree. The predicted value of the observation results is the average value of all trees, which is called aggregation.

The hyperparameters that must be optimized in the random forest method are the number of regression trees and the number of input variables at each node [21]. The following are the stages of the random forest regression algorithm.

- Bootstrapping
- Random data samples are selected as many as k from the original data set with replacement. The purpose of replacement is so that the sample can be reselected in the next iteration.
- Random feature selection
- The regression tree is built until it reaches the maximum size. The stages of creating a regression tree are as follows: first, the data for each variable X is partitioned into two parts. The first partition contains the values of variable X that are less than the average, and the remaining values form the second partition. The value of variable Y is modeled using X in both partitions. This modeling produces the predicted results for each partition. The error for both is calculated using the criterion function, comparing the predicted value with the original Y. This process is repeated with other variables X. Then, from all the variables, the one that produces the smallest error measurement is selected.
- Steps (1) and (2) above are repeated k times to obtain k regression trees.
- The joint estimation of k regression trees is done by using the average value of each output in each regression tree [22].

The criterion function is used to measure the quality of separation on an attribute. Here are some criterion functions for continuous response variables: squared error loss, absolute loss, Friedman mean square error, Poisson.

2.4. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm based on ensemble learning that employs the boosting technique to enhance predictive accuracy. It is an advanced implementation of traditional gradient boosting, incorporating various optimizations that make it faster, more efficient, and more accurate. XGBoost builds models sequentially, where each new model aims to correct the errors made by its predecessors. This algorithm is particularly useful for both classification and regression tasks involving complex, high-dimensional, and large-scale datasets. Its ability to handle imbalanced data, capture nonlinear feature interactions, and support regularization makes it a powerful tool in predictive modeling, especially in applications such as cost prediction, fraud detection, and risk assessment.

Chen mentioned that XGBoost is a well-designed gradient boosting-based decision tree ensemble, only different in the presence of an extension of the objective function minimizing the loss function with the following formulation [18].

$$L_{XGB} = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{j=1}^{m} \Omega(h_j) \quad (5)$$

$$\Omega(h_j) = \gamma T_j + \frac{1}{2}\lambda \|w_j\|^2 \quad (6)$$

L_{XGB} is the loss function of XGBoost, Ω is the regularization function applied to each j^{th} regression tree (h_j), m is the number of trees, n is the number of observations. γ is an internal node, T_j is the number of leaves on the j^{th} tree, λ adalah nilai learning rate, and w_j is the output score for each leaf in the j^{th} tree.

2.5. Model Evaluation

Of the three methods, researchers believe that each has its own advantages and disadvantages. Table 1 presents the strengths and weaknesses of these three methods. The model used in this study will be evaluated through its error distribution. In addition, the Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Median Absolute Deviation (MAD) values will also be calculated. Smaller MSE, MAPE, MAE, and MAD values will indicate better model predictions.

Table 1.
Advantages and disadvantages of three methods that regression-based models.

Method	Main Advantages	Main Disadvantages
Copula Regression	Handles nonlinear dependencies & non-normal distributions	Difficult to implement & requires careful model selection
Random Forest	Simple, robust to noise, no assumptions needed	Less interpretable & slow prediction
XGBoost	Highly accurate, efficient, suitable for big data	Complex and requires extensive parameter tuning, prone to overfitting

3. Application to Health Insurance Cost Data and Discussion

The insurance data utilized in this study was sourced from Bret Lanz's book, Machine Learning Using R. The response variable analyzed was the individual medical costs billed by health insurance (in thousand US dollars). The variables used in the study are presented in Table 2. The dataset comprises 1,338 observations, with six independent variables.

Table 2.
Independent and Response Variables of Insurance Data.

Variables	Description	Type
Age	Age of primary beneficiary	numeric
Sex	insurance contractor gender (female or male)	categoric
BMI	Body mass index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9	numeric
Children	Number of children covered by health insurance / Number of dependents	numeric
Smoker	Smoking (yes or no)	categoric
Region	The beneficiary's residential area in the US (northeast, southeast, southwest, northwest).	categoric
Charges	Individual medical costs billed by health insurance (as a response variable)	numeric

Source: Machine learning using R [23].

The response variable is the individual medical costs billed by health insurance (hereinafter referred to as insurance costs or charges); this distribution is skewed (Figure 1). Logarithmic transformation successfully changes the distribution to be quite symmetrical, even though, when tested with the Shapiro-Wilk test statistic, it does not follow a normal distribution (see Figure 1). The distribution of response variables that are not normal can indicate that the modeling to be carried out has an error that is not normally distributed as well.

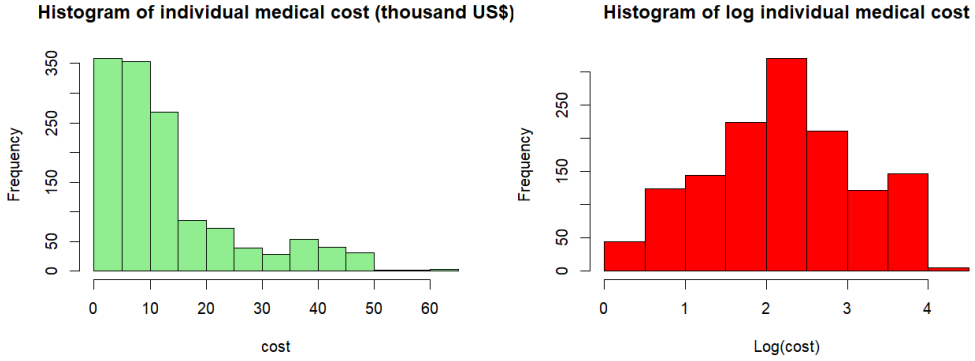

Figure 1.
Histogram of Insurance Cost.

The next analysis wants to see the relationship between variables using the following Pearson correlation as in Table 3 and Figure 2. Since out of the 7 variables, 3 are categorical and the remaining 4 are numerical, the Pearson correlation is calculated only among the numerical variables, with charges as the response variable, and age, BMI, and number of children as the covariates.

Table 3.
Correlation Pearson of numeric variables.

	Charges	Age	Bmi	Children
Charges	1.00			
Age	0.30	1.00		
Bmi	0.20	0.11	1.00	
Children	0.07	0.04	0.01	1.00

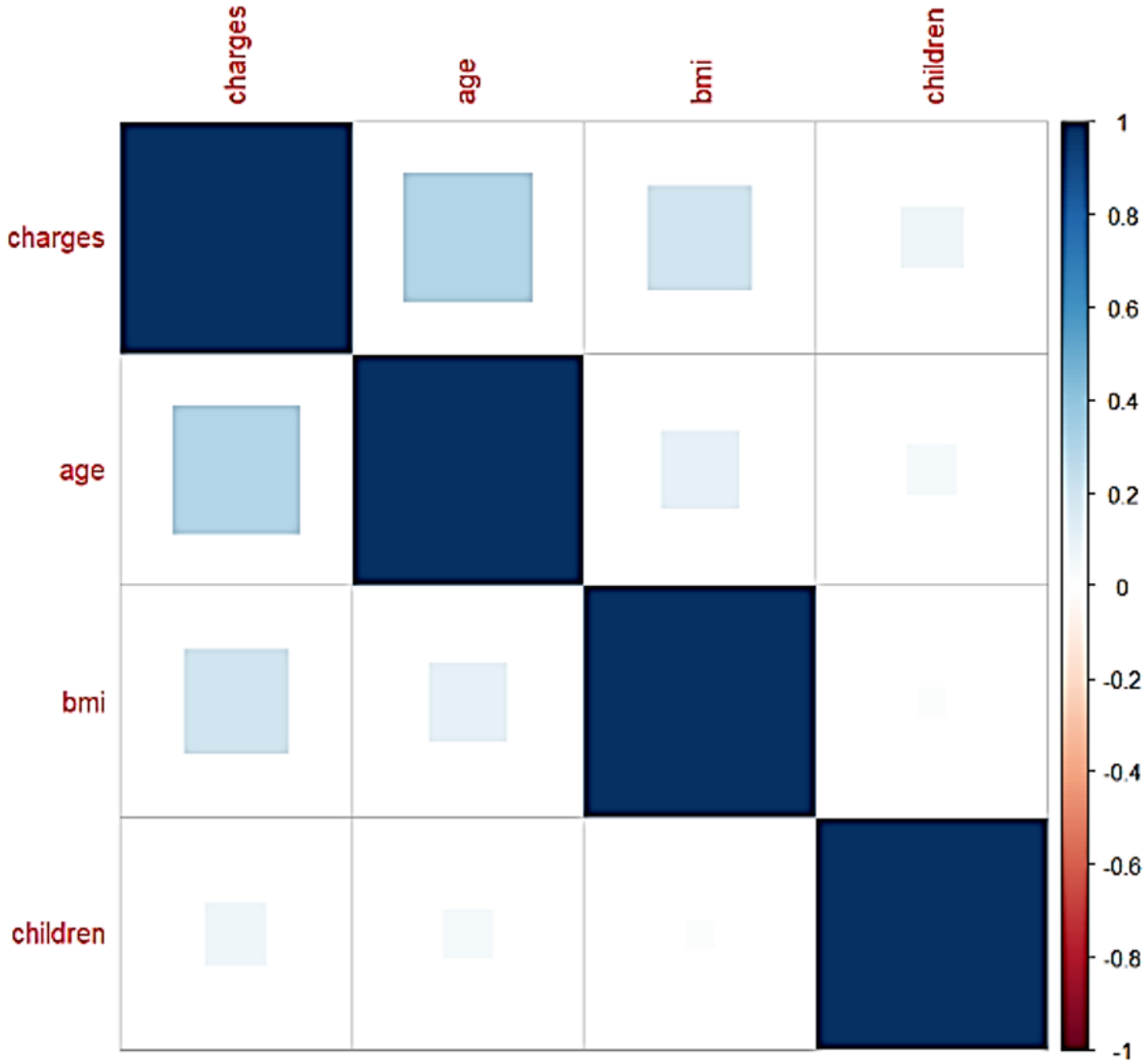

Figure 2.
Heatmap correlation of numeric variables used.

Several regression models will be developed and compared, as outlined in Table 4. For the log-copula regression model (Model 2), the response variable used is the logarithm of insurance costs. Consequently, the predictions made need to be back-transformed to the original scale of the variables.

Table 4.
The modeling was performed using health insurance cost data.

Model	Description
Model 1	Copula regression
Model 2	Log-copula regression
Model 3	Random forest
Model 4	XGBoost

First, the data is divided into two subsets: training data and testing data, with proportions of 80% and 20%, respectively. The training data is used to develop models, while the testing data is used for prediction and model evaluation. Predictions are made using the testing data, allowing for the calculation of errors by comparing the fitted values (response estimates) to the actual response variables from the testing data.

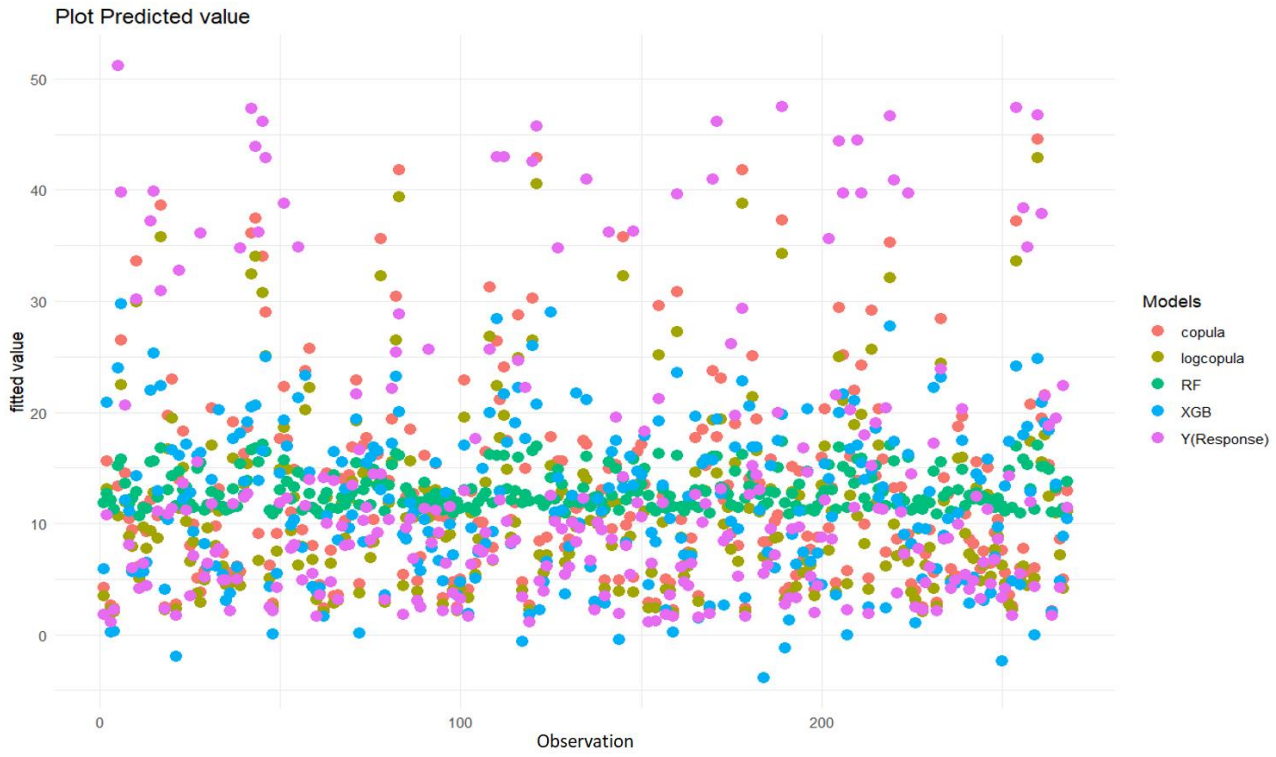

Figure 3.
Fitted value from all models.

In Figure 3, the predicted value of the random forest (RF) model, the green dots appear to be more clustered around the value of 10-15. While the copula and log-copula models can better predict the original Y value, the violet dots have high values (greater than 30). The red dots (predicted by the copula model) and gold dots (predicted by the log-copula model) follow the extreme response values.

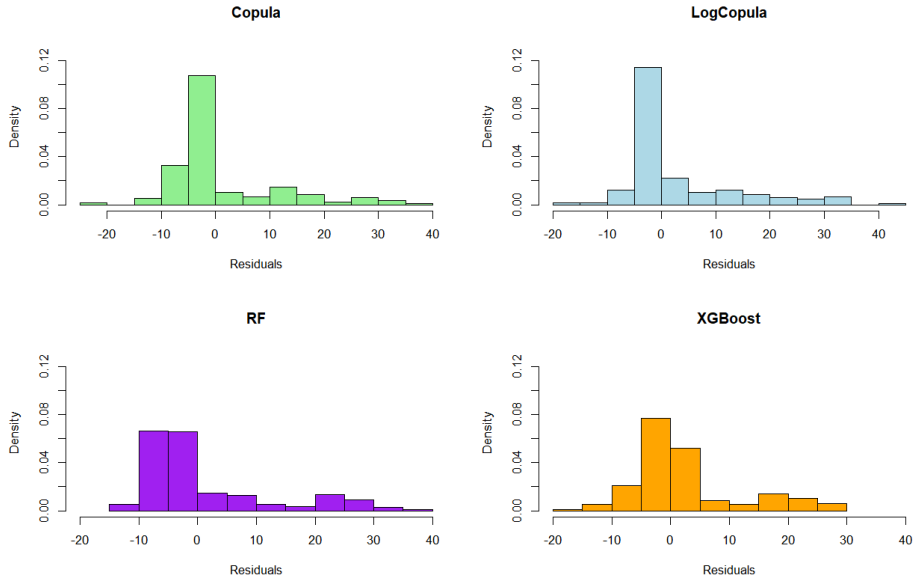

Figure 4.
Residuals Distribution.

Model evaluation can be conducted by examining the error distribution. The errors produced by the four models exhibit different distributions. Figure 4 illustrates the error distribution of the four models. Almost all error values are concentrated around zero, indicating that the four models possess good accuracy.

The error boxplots of the four models are presented in Figure 5. The more compressed the error, the better the model is at predicting the actual data. The Copula, Log-Copula, and XGBoost models exhibit smaller and more compressed errors, followed by the Random Forest model, which also appears compressed with a mean value close to zero. The next model evaluation is conducted using the MSE, MAPE, MAE, and MAD values. The smallest values indicate better predictions for modeling with this insurance data.

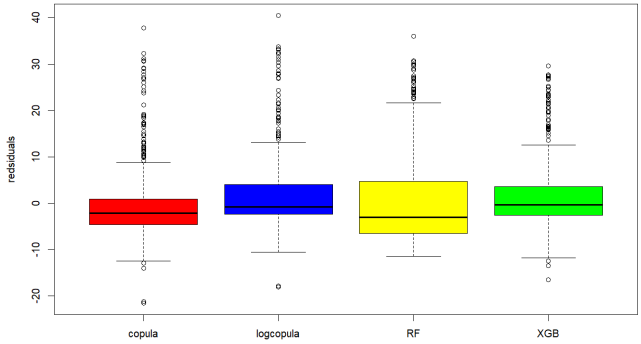

Figure 5.
Boxplot of Residuals.

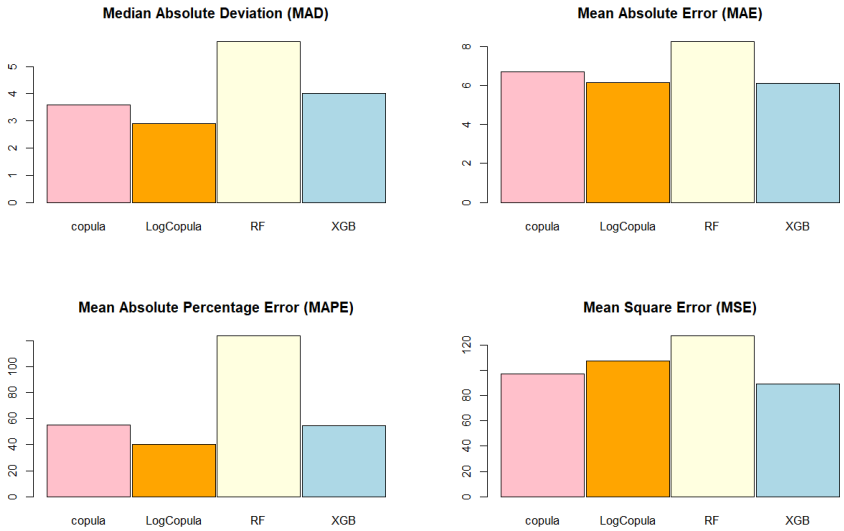

Figure 6.
Model evaluation using MSE, MAPE, MAE, and MAD values.

It can be observed in Figure 6 that, based on the MSE value, the XGBoost method yields the lowest value. When considering the MAPE and MAD values, the LogCopula method exhibits the lowest values. For the MAE value, both the LogCopula and XGBoost methods show the lowest values. The smaller the value of these indicators, the more accurate the predictions made on the testing data.

In this study, it can be concluded that the predictions from the LogCopula and XGBoost methods are superior to those from other methods examined. The random forest regression method is less suitable for data with many outliers and extreme values. This is because the predictions made by the random forest method are derived from averaging the predictions of each leaf on the regression tree and then averaging all regression trees. Consequently, predictions tend to approximate the mean value.

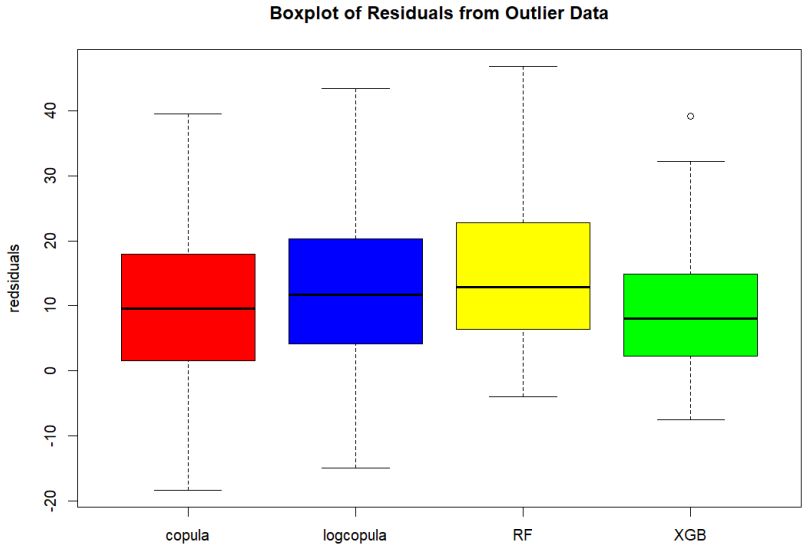

Figure 7.
Error Boxplot of Outlier Data Prediction.

The models developed will be used to predict response variables categorized as outliers. The predictions are compared with the actual response variable data to assess the accuracy of the models for outlier data. Figure 7 presents the error boxplot of the outlier data predictions. It can be observed that the data distribution is predominantly above zero, whereas accurate predictions would exhibit error values close to zero. Several predictions from the copula regression and extreme gradient boosting (XGBoost) methods have errors close to zero, indicating that these models are more effective at predicting response variables in outlier data compared to other models in this study.

4. Conclusion

It can be observed from the error distribution, as well as the MAPE and MAD values, that the copula regression method with logarithmic transformation is more suitable for response variable data that is right-skewed. Therefore, if the distribution

of the data under study is right-skewed, the alternative prediction model recommended by the researcher is copula regression with logarithmic transformation.

For data containing many outliers and extreme values, and where the primary goal is accurate prediction, this study recommends log-copula, copula regression, and extreme gradient boosting as the models of choice. The predictions from these three models can more accurately predict outlier data compared to other methods examined in this study. Random Forest showed the worst performance across all metrics. Therefore, in cases involving right-skewed and asymmetric data, such as health insurance cost data, the Random Forest model is less suitable for modeling, particularly for prediction purposes. Future research is planned to develop copula or log-copula regression models that can be applied to clustered data and capture spatial dependencies for variables exhibiting asymmetric and heavy-tailed characteristics. Such variables are commonly encountered in real-world contexts, particularly in fields such as economics, insurance, and healthcare.

References

[1] N. Hasanah, K. A. Notodiputro, and B. Sartono, "Performance of copula and nested error regression models in estimating per capita expenditure of sub-district in Pidie Regency," *Jurnal Natural*, vol. 23, no. 2, pp. 64-71, 2023. https://doi.org/10.24815/jn.v23i2.31095

[2] D. Handayani, K. A. Notodiputro, A. Saefuddin, I. W. Mangku, and A. Kurnia, "Spatial empirical best predictor of small area poverty indicator," *International Journal of Advances in Soft Computing & Its Applications*, vol. 16, no. 2, pp. 103–122, 2024.

[3] H. A. Zainuddin, K. A. Notodiputro, and K. Sadik, "A simulation study of logarithmic transformation model in spatial empirical best linear unbiased prediction (SEBLUP) method of small area estimation," *Forum Statistika dan Komputasi*, vol. 20, no. 2, pp. 78–84, 2015.

[4] M. K. Najib, S. Nurdiati, and A. Sopaheluwakan, "Multivariate fire risk models using copula regression in Kalimantan, Indonesia," *Natural Hazards*, vol. 113, no. 2, pp. 1263-1283, 2022. https://doi.org/10.1007/s11069-022-05346-3

[5] R. B. Nelsen, *An introduction to copulas*, 2nd ed. New York: Springer Science+Business Media, 2006.

[6] N. Beck, C. Genest, J. Jalbert, and M. Mailhot, "Predicting extreme surges from sparse data using a copula-based hierarchical Bayesian spatial model," *Environmetrics*, vol. 31, no. 5, p. e2616, 2020. https://doi.org/10.1002/env.2616

[7] C. Genest, K. Ghoudi, and L.-P. Rivest, "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions," *Biometrika*, vol. 82, no. 3, pp. 543-552, 1995. https://doi.org/10.1093/biomet/82.3.543

[8] C. Genest and A.-C. Favre, "Everything you always wanted to know about copula modeling but were afraid to ask," *Journal of Hydrologic Engineering*, vol. 12, no. 4, pp. 347-368, 2007.

[9] C. Genest and L.-P. Rivest, "Statistical inference procedures for bivariate Archimedean copulas," *Journal of the American statistical Association*, vol. 88, no. 423, pp. 1034-1043, 1993. https://doi.org/10.2307/2290796

[10] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 415-434, 1963. https://doi.org/10.1080/01621459.1963.10500855

[11] R. Messenger and L. Mandell, "A modal search technique for predictive nominal scale multivariate analysis," *Journal of the American Statistical Association*, vol. 67, no. 340, pp. 768-772, 1972. https://doi.org/10.1080/01621459.1972.10481290

[12] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, no. 2, pp. 119-127, 1980. https://doi.org/10.2307/2986296

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[14] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[15] J. R. Quinlan, "Combining instance-based and model-based learning," in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 236-243.

[16] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470, 2017.

[17] W.-Y. Loh, "Regression tress with unbiased variable selection and interaction detection," *Statistica Sinica*, vol. 12, no. 2, pp. 361-386, 2002.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[19] M. Sklar, "Distribution functions with dimensions and their margins," *Annales De l'ISUP*, vol. 8, no. 3, pp. 229-231, 1959.

[20] L. Capitaine, R. Genuer, and R. Thiébaut, "Random forests for high-dimensional longitudinal data," *Statistical Methods in Medical Research*, vol. 30, no. 1, pp. 166-184, 2021.

[21] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad002, 2023. https://doi.org/10.1093/bib/bbad002

[22] C. Longani, S. Prasad Potharaju, and S. Deore, *Price prediction for pre-owned cars using ensemble machine learning techniques. In Recent Trends in Intensive Computing*. Amsterdam, Netherlands: IOS Press, 2021.

[23] B. Lantz, *Machine learning with R: Expert techniques for predictive modeling*. Birmingham, UK: Packt Publishing Ltd, 2013.