Comparative Analysis of Machine Learning Models for Health Insurance Premium Prediction Using R

Xiaowei Fang*, Zhuoxuan Zhang

Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China

Abstract: This study explores methodologies for forecasting health insurance premiums, focusing on predictive accuracy and reliability. Using a dataset with variables such as age, gender, BMI, and diseases, we apply multiple techniques-including the K-Nearest Neighbors (KNN) algorithm, voting methods, and other machine learning algorithms-to predict premiums. A comparative analysis highlights each method's strengths and limitations, offering insights into which approach provides the most accurate and practical predictions. The findings aim to guide insurers in selecting effective forecasting methods to enhance premium pricing strategies and improve risk management.

***Keywords*:** Health Insurance; Machine Learning Models; Random Forest; XGBoost; Gradient Boosting; Predictive Modeling.

1. Introduction

As society continues to advane, increasing attention is being paid to human health. In modern society, the effective management of health insurance has become crucial for safeguarding public health and controlling medical costs, in order to mitigate the economic impact of major diseases. With the continuous development of data science and machine learning technologies, leveraging big data analysis to identify and predict the characteristics of health insurance has emerged as a prominent area of research.

Excessive healthcare expenses, which are closely tied to economic hardship, have become an escalating challenge in lower-income nations, particularly among marginalized rural populations. Households with limited financial means face disproportionate impacts due to restricted access to essential resources required for health risk prevention and the adoption of beneficial health practices. In addition, health risks can lead to unexpected expenditures that seriously increase the financial burden on families (Su, Sha and Liu, 2023). To solve these pressing problems, governments around the world have introduced health insurance policies. Medical insurance is a contract in which an insurance company owes the insured the financial risk of medical expenses and provides financial protection in the event of injury, surgery, or other unfortunate events (Najar, 2024). It is critical for the health sector to use predictive models to accurately predict individual healthcare costs. Accurate estimates of cost can help health insurers, and a growing number of healthcare delivery organizations, plan for the future limited healthcare management resources. In addition, knowing their likely future expenses in advance can help patients choose an insurance plan with appropriate deductibles and premiums (Hassan, 2021). These factors play an important role in the formulation of insurance policies.

Classical techniques like the Bornhuetter-Ferguson method and chain-ladder models remain foundational for reserve estimation. However, these methods often struggle with complex dependencies (e.g., seasonality, loss distributions) and large datasets. Actuarial packages in R, such as actuar and lme4, provide robust implementations of these models. The actuar package supports loss distribution modeling, risk theory, and credibility analysis, while lme4 enables mixed-effects modeling for hierarchical data (Bates et al., 2014). Charpentier (2014) further emphasizes R versatility in actuarial workflows, including survival analysis and portfolio allocation.

Machine learning algorithms, especially ensemble-based predictive modeling techniques including XGBoost and random forest architectures, have outperformed conventional approaches in insurance liability estimation. Dash et al. (2024) integrated several regression techniques, including logistic regression, random forest regression, decision tree regression, gradient boosted regression, and linear regression. Their prediction model exhibited high accuracy, with an average absolute error of less than 5% across all samples. The research highlighted age and BMI as primary predictors of insurance rates, while gender and location had relatively minor effects. Padiear et al. (2023) used three models — linear regression, decision tree regression, and random forest regression to estimate insurance costs, considering factors like age, BMI, smoking status, children, charges, gender, and region. They found smoking to be the most significant factor. The accuracy rates were 74.26% for linear regression, 78.48% for decision tree, and 86.29% for random forest. Kandula et al. (2024) developed an algorithm combining multiple regression models, achieving an accuracy of 88.98% with the random gradient enhancement technique.

In this study, we used machine learning to compare the accuracy of various regression models, including Random Forest, XGBoost, Voting, Gradient Boosting, K-Nearest Neighbors. The algorithmic performance underwent rigorous multidimensional validation through predictive accuracy (R2 coefficient) and error magnitude (MAE/RMSE), then discussed the results and summarized future research directions.

2. Methodology

To calculate the cost of the model prediction, we acquired the data set from the Kaggle website. The platform provides detailed descriptions and metadata for each dataset to help users understand its characteristics and potential applications.

2.1 Dataset

There are 15000 data in the dataset. The data set used in this study comprises 11 variables that are crucial for predicting insurance claims. These variables are defined as follows:

Table 1. The variables of insurance dataset

Feature	Description
Age	The age of the policyholder (Numeric).
Sex	The gender of the policyholder (Categorical).
Weight	The weight of the policyholder (Numeric).
BMI	Body Mass Index, calculated as weight in kilograms divided by the square of height in meters (kg/m²). It provides an objective measure of body weight relative to height (Numeric).
Hereditary_diseases	A policyholder suffering from a hereditary diseases or not (Categorical)
No_of_Dependents	The number of individuals financially dependent on the policyholder (Numeric).
Smoker	Indicates whether the policyholder is a smoker (1) or non-smoker (0) (Categorical).
Bloodpressure	The blood pressure reading of the policyholder (Numeric).
Diabetes	Indicates whether the policyholder has diabetes (1) or not (0) (Categorical).
Regular-ex	Indicates if the policyholder engages in regular exercise (1) or not (0) (Categorical).
Claim	The monetary amount claimed by the policyholder (Numeric).

2.2 Exploratory data analysis

Data pre-processing is the first step in health insurance premium prediction , there were 396 missing values in the age variable and 956 missing values in the BMI variable. The missing values and outliers are removed. For all models, the data set was split into two parts: 70% of the data was used for training, and 30% for testing. This structured data serves as the foundation for developing predictive models aimed at enhancing the accuracy of insurance claim predictions. The data processing and implementation steps are shown in Figure 1.

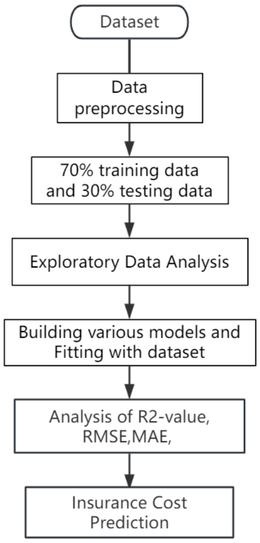
Dataset
Data preprocessing
70% training data and 30% testing data
Exploratory Data Analysis
Building various models and Fitting with dataset
Analysis of R2-value, RMSE,MAE,
Insurance Cost Prediction

Figure 1. Step for data processing

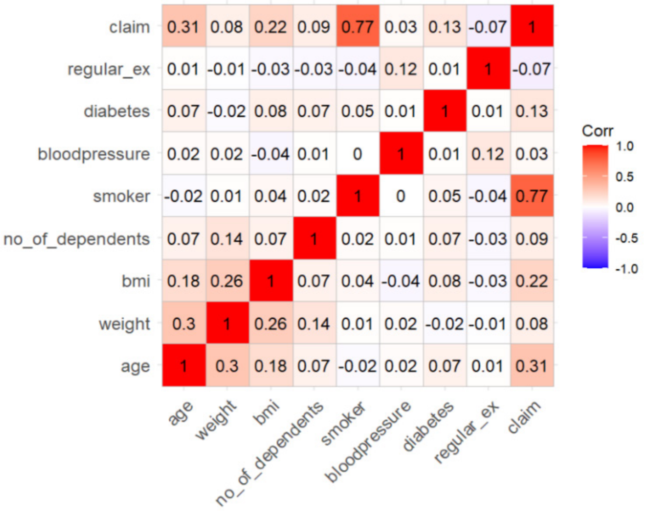

Figure 2. Correlation matrix of numeric variables

The correlation matrix reveals important relationships between the predictor variables and the health insurance claim amount. The smoker variable exhibits the strongest correlation with the insurance claim, with a coefficient of 0.77, indicating a substantial positive association between smoking status and higher insurance premiums. Following this, age shows a moderate positive correlation with the claim amount, with a correlation coefficient of 0.31, suggesting that older individuals tend to have higher insurance premiums. The BMI variable also demonstrates a positive correlation of 0.22 with the insurance claim, though this relationship is weaker. Other variables in the dataset, however, show relatively low correlations with the insurance claim, indicating that they have a less significant impact on predicting premium amounts. The results are shown in Figure 2.

2.3 Models

This study utilized a variety of machine learning models, including Random Forests, Gradient Boosting, K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), and Voting models to predict insurance payments. Individual health data were analyzed using regression models to estimate insurance costs, considering factors influencing premiums.

2.3.1 Random Forest

The Random Forest model is a powerful ensemble learning method for classification and regression tasks. It builds multiple decision trees during training and outputs the mode (classification) or mean (regression) of the individual trees. This method excels with large, high-dimensional datasets and handles missing data effectively.

Random Forests use bootstrap aggregating (bagging), creating data subsets with replacement. For each subset, a decision tree is built, selecting a random subset of features at each node to determine splits. This randomness reduces overfitting, especially with many features. A key strength of Random Forests is featuring importance estimation, identifying influential factors in predictions. This is valuable in contexts like insurance claim prediction, where variables such as age, BMI, and smoking impact risk assessment and premiums. Random Forests capture complex, non-linear relationships and ensure high accuracy through ensemble averaging, reducing variance. In this study, the model predicts insurance claim costs, with hyperparameters optimized by cross-validation. Its performance is evaluated using metrics like MSE, demonstrating its effectiveness in predicting claim amounts accurately.

2.3.2 Gradient Boosting

Gradient Boosting is an advanced ensemble learning method for classification and regression tasks. It operates through sequential error correction, iteratively training weak learners to focus on residual errors from prior iterations. The method builds weak learners sequentially, starting with a base predictor (e.g., a shallow regression tree) to establish initial approximations. Subsequent models optimize residual errors using adaptive weighting, continuing until convergence criteria are met. Predictive outputs are formed by combining models, with contributions modulated by hyperparameters like learning rate.

Gradient Boosting's flexibility lies in its ability to optimize differentiable loss functions, making it suitable for diverse problems. In insurance claim prediction, it effectively models non-linear relationships between features like age, BMI, and smoking status, and claim amounts.

2.3.3 Voting Model

The Voting model is an ensemble learning technique that combines predictions from multiple models to improve accuracy. It leverages the strengths of diverse algorithms, making it effective for tasks like insurance claim prediction by capturing varied data patterns. This approach can integrate fundamentally different models, such as decision trees, support vector machines, and neural networks, reducing overfitting by combining uncorrelated errors. For insurance claim prediction, combining models like Random Forest (for non-linear relationships) and linear regression (for linear patterns) ensures balanced predictions.

The Voting model involves training individual base models and aggregating their predictions using hard or soft voting. Hyperparameter tuning for each base model is performed via cross-validation to optimize their contributions to the ensemble. Performance is evaluated using metrics like mean squared error (MSE) for regression tasks, assessing its predictive accuracy.

2.3.4 K-Nearest Neighborhood

The K-Nearest Neighbors (KNN) algorithm is a simple yet robust non-parametric method for classification and regression tasks. It predicts outcomes based on the "K" nearest neighbors in the feature space, making it effective for non-linear or ambiguous feature-target relationships, such as in insurance claim prediction with high-dimensional data.

For regression tasks, KNN predicts by averaging the target values of the "K" nearest neighbors. The choice of "K" and the distance metric (e.g., Euclidean, Manhattan) are critical hyperparameters that influence performance. Smaller "K" values may lead to overfitting, while larger values may over smooth predictions, requiring careful tuning through cross-validation.

In this study, KNN is applied to predict insurance claim costs, with hyperparameters optimized to balance bias and variance. Performance is evaluated using metrics like mean squared error (MSE). Despite its computational limitations, KNN serves as a baseline model, offering insights into feature relationships and demonstrating its value in handling non-linear patterns in predictive modeling.

2.3.5 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an advanced gradient boosting implementation known for its high predictive power and computational efficiency. It is widely used for classification and regression tasks, including complex problems like insurance claim prediction.

XGBoost builds decision trees sequentially, correcting errors from previous iterations. It incorporates L1 (Lasso) and L2 (Ridge) regularization to control overfitting, balancing model complexity and accuracy. The model also handles missing data internally, automatically determining the best approach for missing values without extensive preprocessing.

In this study, XGBoost predicts insurance claim costs, with hyperparameters (e.g., number of trees, max depth, learning rate, regularization parameters) optimized via cross-validation. Performance is evaluated using metrics like mean squared error (MSE), demonstrating its accuracy and reliability.

3. Results

By fitting various models to the dataset, we obtained the final results. We use the coefficient of determination, pronounced R-squared score as the accuracy of the model. It is the proportion of the variation in the dependent variable that is predictable from the independent variables. The root mean squared error (RMSE) and mean absolute error (MAE), both shows the differences between predicted value and true value. The smaller they are, the more accurate the model is. Based on the algorithm, the calculation results show in Table 2.

Table 2. Accuracy comparison of different models

Models	R2 score	RMSE	MAE
Random Forest	0.9688	2133.41	539.08
XGBoost	0.9668	2250.29	695.46
Voting	0.9392	2909.64	1566.83
Gradient Boosting	0.8792	4248.14	2613.84
K-nearest Neightbors	0.8325	4716.81	2841.68

In Table 2 we can know the model of Random Forest is the best. The R-squared score of the Random Forest model is closest to one, and it has the smallest MAE and RMSE. After hyperparameter tuning, it is determined that when ntree is 200 and mtry is 4, the model achieves the best performance, highest efficiency and accuracy, and lowest error. The prediction results are shown in Figure 3.

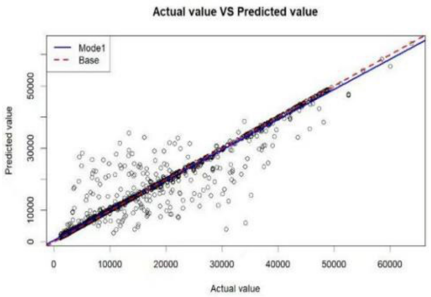

Figure 3. Result of RF model

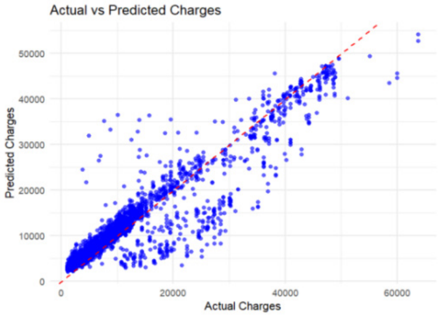

Figure 4. Result of Gradient Boosting model

Individual Conditional Expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when a feature changes. Figure 5 shows the C-ICE plot for the random forest model. The results displayed in the figures align closely with the predicted outcomes. It is evident that smoking has a significant impact on insurance premiums, with smoking substantially increasing the premium amount. In contrast, age and BMI exhibit a weak positive correlation with premiums, while weight shows a slight negative correlation. Interestingly, when age exceeds 43 years, BMI surpasses 32, or blood pressure exceeds 90, the rate of increase in premiums accelerates. This suggests that at certain thresholds — specifically advanced age, obesity, and hypertension — insurance premiums rise significantly.

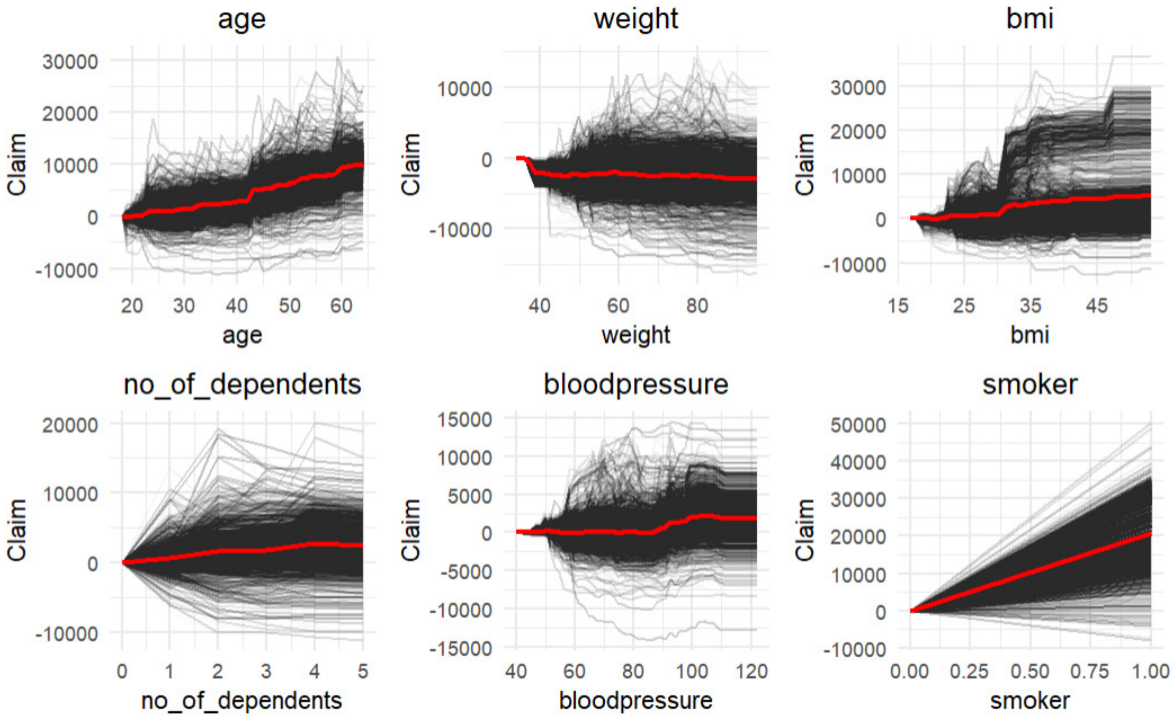

Figure 5. C-ICE plot for RF model

The second-best model is the XGBoost model. Through hyperparameter tuning, optimal model performance was achieved when the number of boosting rounds (n-rounds) was set to 500. At this configuration, the model demonstrated the highest efficiency and accuracy, with minimal error.

The Voting model also shows a high level of accuracy and a strong fit to the data, as indicated by an R2 value close to 1. The relatively low MAE and RMSE values suggest that the model is highly effective at predicting claim amounts. Additionally. we conducted an analysis of feature importance for each individual model to understand which variables were most influential in predicting claims. Interaction features like age_bmi, weight_smoker, and age_bloodpressure emerged as significant predictors, highlighting the importance of capturing complex relationships in the data. For the Gradient Boosting model (Figure 4) and the KNN model, the R-squared scores are not close enough to one. However, we made some other important discoveries. The results of the Gradient Boosting model show that smoking status and BMI have the greatest influence on premiums.

4. Conclusion

This study embarked on a deep dive into the world of health insurance cost prediction, testing the performance of various machine learning models to identify which could most accurately forecast future expenses. The result shows that the Random Forest, XGBoost, and Voting models emerged as the top performers, demonstrating exceptional precision and reliability. These models, with their sophisticated algorithms, are particularly adept at uncovering complex patterns in data, making them invaluable tools for the insurance industry.

For the broader health insurance ecosystem, these findings are innovative. Insurers now have the power to refine their pricing strategies, not just to maximize profit but to foster transparency and fairness. In a world where health costs are rising, this level of predictive power could also help alleviate some of the financial strain on policyholders, making insurance premiums more predictable and justifiable. Policyholders, on their end, can better understand how their choices — whether it's weight management or smoking habits — directly affect their premiums, empowering them to make more informed, health-conscious decisions.

Looking ahead, this research paves the way for even more refined models. While smoking status and BMI are significant, there are countless other factors such as genetics, medical history, that could further enhance prediction accuracy. The next frontier lies in real-time data integration: continuously updated health metrics could enable insurers to adjust premiums dynamically, offering a level of adaptability that static models can't match.

References

[1] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects models usinglme4. Journal of Statistical Software, 67(1). https://doi.org/10.18637/jss.v067.i01

[2] Charpentier, A. (2014). Computational Actuarial Science with R. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/b17230

[3] Dash, S., Panigrahi, B. S., Sanikommu, V. V. B. R., Madhavi, B. K., & Sahoo, S. K. (2024). A Comparative Analysis of Different Machine Learning Techniques For Medical Insurance Premium Prediction. Dash, S. Et Al., 1–6. https://doi.org/10.1109/ic-cgu58078.2024.10530731

[4] Hassan, C. a. U., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. a. A., & Ullah, S. S. (2021). A computational intelligence approach for predicting medical insurance cost. Mathematical Problems in Engineering, 2021, 1–13. https://doi.org/10.1155/2021/1162553

[5] Kandula, A. R., Kalyanapu, S., Rayapalli, S. N., Veerabathina, K. R., Modugumudi, V., & Kanikella, S. R. (2024). Medical Insurance Predictive Modelling: An Analysis of Machine Learning Methods. Kandula, A.R. Et Al., 1–5. https://doi.org/10.1109/iatmsi60426.2024.10502643

[6] Manathunga, V., & Zhu, D. (2022). Unearned premium risk and machine learning techniques. Frontiers in Applied Mathematics and Statistics, 8. https://doi.org/10.3389/fams.2022.1056529

[7] Najar, P.A. (2024). Adopting Global Health Insurance Models for Medical Tourists in India: Implications for Stakeholders in Digitalized Era. Journal of the Insurance Institute of India, 11(3), 102–111.

[8] Patidar, S., Dudi, S., & Rohit, N. (2023). Estimating Medical Insurance Cost using Linear Regression with HyperParameterization, Decision Tree and Random Forest Models. 2022 12th International Conference on Cloud Computing, Data Science &Amp; Engineering (Confluence), 504–508. https://doi.org/10.1109/confluence56041.2023.10048836

[9] Su, L., Sha, M., & Liu, R. (2023). Medical insurance, labor supply, and anti-poverty initiatives: Micro-evidence from China. International Studies of Economics, 19(2), 268–292. https://doi.org/10.1002/ise3.70

[10] Von Ulmenstein, U., Tretter, M., Ehrlich, D. B., & Von Peharnik, C. L. (2022). Limiting medical certainties? Funding challenges for German and comparable public healthcare systems due to AI prediction and how to address them. Frontiers in Artificial Intelligence, 5. https://doi.org/10.3389/frai.2022.913093