Proceedings of the 3rd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/79/20241754

Comparison and Analysis of the Effectiveness of Linear Regression, Decision Tree, and Random Forest Models for Health Insurance Premium Forecasting

Yaowen Hu[1,a,*]

[1]*College of Liberal Arts, University of Minnesota, Minneapolis, Minnesota, 55414, USA*
a. yaowen_hu01@163.com
**corresponding author*

Abstract: Health insurance is a type of insurance that covers individual and family medical expenses and is important for the health and financial security of individuals and families. To better predict the demand for health insurance, three regression models in machine learning - random forest, linear regression, and decision tree - are widely used for health insurance prediction. Among these three regression models, random forest regression has the best prediction effect with a model score of 0.8564, which is the best prediction effect among the three models. Random forest regression is an integrated learning method that combines multiple decision tree models into a more powerful model that can effectively avoid overfitting problems and can handle large amounts of data. Therefore, random forest regression is a very effective method for health insurance prediction. The next model is the linear regression model with a model score of 0.7584. The linear regression model is a basic regression model that can be used to predict a linear relationship between two variables. In health insurance prediction, linear regression modeling can be used to predict the linear relationship between health insurance costs and related factors such as age, gender, and illness. The worst predictor was the decision tree, with a model score of 0.7097. The decision tree model can be used in Medicare forecasting to predict nonlinear relationships between Medicare costs and related factors such as age, gender, and illness.

Keywords: Linear regression, Decision tree, Random forest

1. Introduction

Health insurance is a type of insurance that protects individuals and families from medical expenses and is very important for the health and financial security of individuals and families [1]. First of all, medical insurance can provide financial security because medical expenses are often a major burden on family finances. Without medical insurance, in the event of illness or accident, families may have to bear high medical expenses and may even fall into economic crisis as a result [2]. And with medical insurance, the family's financial security can be effectively protected. Secondly, medical insurance can improve the quality of medical care [3]. Health insurance can make it easier for people to access healthcare services because they do not have to worry about the cost of healthcare. This can encourage people to seek medical care earlier and treat their illnesses in a timely manner, thus improving the quality of healthcare. In addition, health insurance can reduce medical [4].

DOI: 10.54254/2754-1169/79/20241754

Machine learning has produced many research results on the problem of healthcare cost prediction. By utilizing machine learning algorithms, healthcare costs can be predicted more accurately, thus providing better services to insurance companies [5]. Using neural network algorithms, a researcher analyzed the U.S. National Health and Nutrition Examination Survey dataset and established a neural network-based medical cost prediction model. The model can accurately predict healthcare costs and can be personalized based on different factors [6]; Some researchers used decision tree algorithms to analyze the U.S. health insurance dataset and established a decision tree-based healthcare cost prediction model. The model can predict medical costs based on different factors, such as age, gender, and BMI [7]; some researchers used deep learning algorithms to analyze the U.S. health insurance dataset and built a deep learning-based medical cost prediction model. The model can accurately predict healthcare costs and can be personalized based on different factors [8].

Companies use human labor to predict healthcare costs for insurance companies, which is a very time-consuming, labor-intensive, and often inaccurate process [9,10]. Therefore, in order to improve the accuracy and efficiency of the prediction, this paper modernizes the legacy system based on the insurance dataset available on kaggle in the hope of implementing an automated method for predicting the healthcare costs of insurance companies based on various factors. The method will take into account several factors, such as age, body mass index, smoking habits, number of children, etc., and build a medical cost prediction model through data analysis and machine learning algorithms. The model allows for automated prediction, which not only improves the accuracy and efficiency of the prediction, but also saves significant time and resource costs.

2. Data set introduction

The health insurance dataset on Kaggle is a very classic dataset consisting of 1338 customers' basic information and health insurance costs. The purpose of this dataset is to build a healthcare cost prediction model through the relationship between the basic information of the customers and the insurance cost to provide better service to the insurance companies.

This dataset contains many key features such as age, gender, BMI, number of children, and whether the customer is a smoker or not. These features can help us to better understand the needs of our customers and provide them with better health insurance services. For example, age is a very important factor because as people get older, they are more likely to suffer from some diseases and need more health insurance services. And BMI is also a very important factor because too high or too low BMI can lead to health problems and require more health insurance services.

Using this dataset, we can build various healthcare cost prediction models such as those based on algorithms such as linear regression, decision trees, neural networks, and deep learning. These models can predict the health insurance cost of customers based on their basic information and provide better services to insurance companies.

In addition, this dataset can be used for data visualization and exploratory analysis. By visualizing the relationship between customers' basic information and health insurance costs, we can better understand the data and discover some interesting patterns and trends. For example, we can find a positive correlation between age and health insurance costs, a positive correlation between BMI and health insurance costs, and a negative correlation between smoking and health insurance costs, etc.

In conclusion, the health insurance dataset on Kaggle is a very valuable dataset that can help us better understand the needs of our customers and provide better services to insurance companies. By utilizing data analytics and machine learning techniques, we can build various healthcare cost prediction models and discover some interesting patterns and trends. Some of the data are shown in Table 1:

DOI: 10.54254/2754-1169/79/20241754

Table 1: Partial dataset.

Age	Sex	Bmi	Children	Smoker	Region	Id	Charges
24	male	23.655	0	no	northwest	693	2352.96845
28	female	26.51	2	no	southeast	1297	4340.4409
51	male	39.7	1	no	southwest	634	9391.346
47	male	36.08	1	yes	southeast	1022	42211.1382
46	female	28.9	2	no	southwest	178	8823.279

3. Statistical analysis of data

The distribution trends and proportions of age, BMI and premium are shown in Figures 1, 2 and 3, respectively:

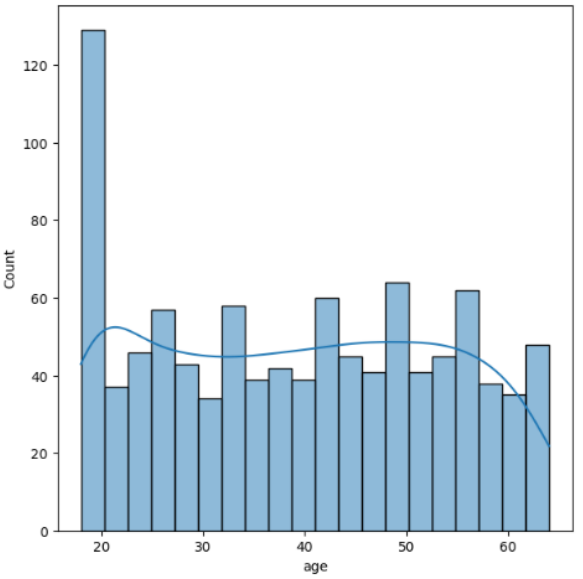

Figure 1. Age.
(Photo credit: Original)

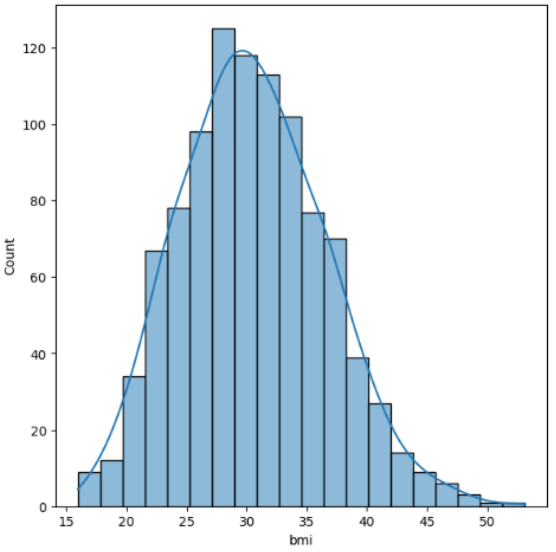

Figure 2: Bmi.
(Photo credit: Original)

DOI: 10.54254/2754-1169/79/20241754

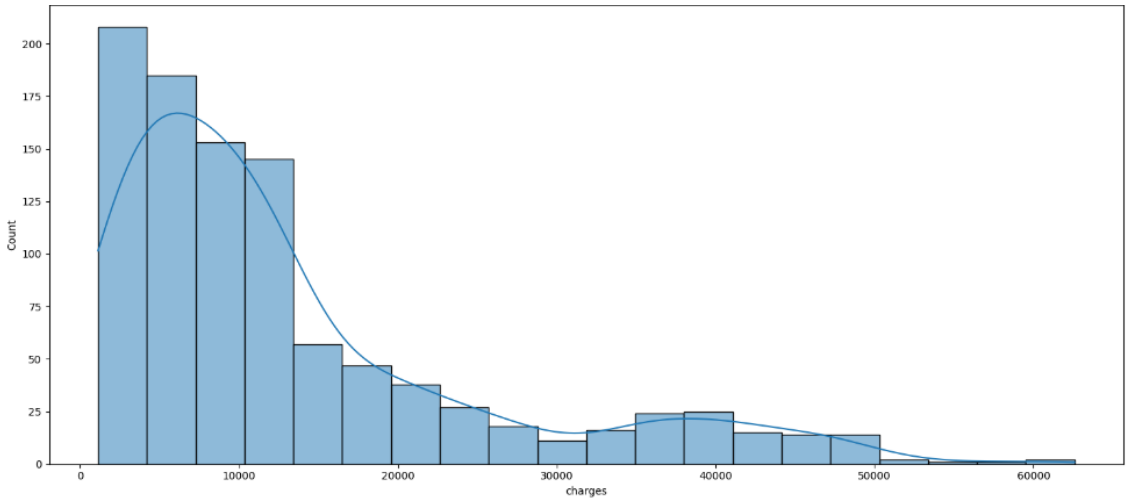

Figure 3: Charges.
(Photo credit: Original)

From the above figure it can be observed that the dataset includes all age groups and is well balanced across the age groups. BMI is normally distributed in the dataset in the fee column most of the premiums are below 13500.

4. Statistical analysis of data

4.1. Linear regression model

Linear regression is a statistical method used to model and predict relationships between continuous variables. It is one of the simplest methods of regression analysis and is commonly used for exploratory data analysis and predictive analysis. A linear regression model is based on a linear relationship between one or more independent variables and a continuous dependent variable, and the line of best fit is determined by minimizing the sum of squared residuals. The model can be used to predict new data points and can be used to explain relationships between variables.

In linear regression, we assume that there is a linear relationship between the dependent and independent variables, i.e., the dependent variable can be expressed as a linear combination of the independent variables, plus an error term. This error term is usually assumed to follow a normal distribution with a mean of 0 and a constant variance. The goal of linear regression is to find the optimal regression coefficients such that the model minimizes the sum of squared residuals. The residual is the difference between the observed and predicted values. To find the optimal regression coefficients, we can use the method of least squares. The goal of the least squares method is to minimize the sum of squared residuals, i.e. to find the best regression coefficients such that the sum of squared residuals is minimized.

The performance of a linear regression model can be assessed by a variety of metrics, including the R-squared value, the mean square error (MSE), the root mean square error (RMSE), etc. The R-squared value is a value between 0 and 1, which indicates the proportion of the variability in the dependent variable that can be accounted for by the model. The MSE and the RMSE are measures of the predictive performance of the model, and they indicate the average error between the predicted and actual values of the model.

The linear regression model is a simple but powerful predictive model that can be used in many fields, including economics, finance, medicine, and biology. It can help us understand the relationships between variables and can be used to predict new data points.

DOI: 10.54254/2754-1169/79/20241754

4.2. Decision tree model

Decision tree regression is a regression analysis method based on tree structure. Unlike traditional linear regression, decision tree regression can handle nonlinear relationships and can handle multiple independent and categorical variables simultaneously. Decision tree regression is a simple but powerful predictive model that can be used in many fields, including finance, medicine, biology, and more.

The basic idea of decision tree regression is to predict the value of a target variable by dividing the data into different subsets, each with similar characteristics, and then constructing a simple model in each subset. These subsets are represented by a tree structure where each node represents an independent variable, each branch represents a value taken by that independent variable, and each leaf node represents a predicted value. When predicting a new data point, we start from the root node and assign the data point to the corresponding subset based on the characteristics of each node until we reach the leaf node, and then use the predicted value of that leaf node as the prediction result.

The decision tree regression construction process can be accomplished by recursively splitting the dataset into different subsets. At each split, we choose an independent variable and a split point to divide the dataset into two subsets. The selection of the independent variable and split point can be done using different algorithms, including those based on information entropy, Gini index, etc. In each subset, we can continue to split recursively until a predetermined stopping condition is reached, such as reaching a maximum depth, subset size less than a certain threshold, etc.

The performance of decision tree regression can be evaluated by a variety of metrics, including mean square error (MSE), root mean square error (RMSE), etc. These metrics represent the average error between the predicted and actual values of the model. Typically, we use cross-validation to assess the performance of decision tree regression to avoid overfitting.

Advantages of decision tree regression include ease of understanding and interpretation, ability to handle nonlinear relationships, ability to handle multiple independent and categorical variables at the same time, and applicability to large-scale datasets. Disadvantages include easy overfitting, sensitivity to noise, need to choose appropriate stopping conditions, etc.

4.3. Random forest model

Random forest regression is an integrated learning method based on decision trees, which improves prediction performance by combining multiple decision trees. Random forest regression can deal with nonlinear relationships and multiple independent and categorical variables at the same time, with high prediction accuracy and robustness, and is widely used in finance, medicine, biology and other fields.

The basic idea of random forest regression is to predict the values of the target variables by constructing multiple decision trees, and then average or vote the predictions of these decision trees to get the final prediction. In constructing each decision tree, we use different random samples and random features to train the model to increase model diversity and reduce overfitting. Specifically, we randomly draw a portion of samples and a portion of features from the original dataset and then use these samples and features to train the decision tree. In this way, each decision tree models a different part of the dataset, thus increasing the generalization ability and robustness of the model.

The construction process of random forest regression can be divided into two stages. First, we use random samples and random features to construct multiple decision trees. In constructing each decision tree, we use different algorithms and parameters to train the model, such as algorithms based on information entropy, Gini index, and so on. Next, we average or vote the predictions of these decision trees to get the final prediction. The averaging method averages the predictions of all the decision trees, while the voting method votes the predictions of all the decision trees to get the final prediction.

DOI: 10.54254/2754-1169/79/20241754

The performance of random forest regression can be evaluated by a variety of metrics, including mean square error (MSE), root mean square error (RMSE), and so on. These metrics represent the average error between the predicted and actual values of the model. Typically, we use cross-validation to assess the performance of random forest regression to avoid overfitting.

Advantages of random forest regression include easy to understand and interpret, can deal with nonlinear relationships, can deal with multiple independent and categorical variables at the same time, applicable to large-scale datasets, and so on. Disadvantages include the need to choose appropriate parameters and algorithms, the need for longer training time, etc. In practice, random forest regression is usually used together with other techniques, such as gradient boosting, neural networks, etc., to improve prediction performance.

5. Result

The training set, validation set and test set are divided according to 6:2:2, the training set is used to train the model, the validation set is used to validate the results of the training, and the test set is used for the testing of the model, and the results are shown in Table 2 and Figure 4:

Table 2: Model evaluation.

Model	Training Set R2 Scores	Test Set R2 Scores
Linear Regression	0.7486	0.7584
Decision Tree Regressor	0.9872	0.7097
Random Forest Regressor	0.9665	0.8564

Figure 4: Model evaluation.
(Photo credit: Original)

From the above figure, it can be seen that the random forest regression has the best prediction effect, with a model score of 0.8564, which is the best prediction effect of the three models; followed by the linear regression model, with a model score of 0.7584; and the worst prediction effect is the decision tree, with a model score of 0.7097.

DOI: 10.54254/2754-1169/79/20241754

6. Conclusion

Random forest, linear regression and decision tree are three regression models commonly used in machine learning. On the prediction of medical insurance, random forest regression has the best prediction effect, with a model score of 0.8564, which is the best prediction effect of the three models; followed by linear regression model, with a model score of 0.7584; and the worst prediction effect is the decision tree, with a model score of 0.7097. The principles and application scenarios of the three models are different, and thus they differ in their prediction effects.

Random forest is an integrated learning method that improves prediction accuracy by combining multiple decision trees. Each decision tree is trained on randomly selected samples and features, which avoids the overfitting problem. Random forests perform well in dealing with high-dimensional data and nonlinear relationships, so they are widely used in many real-world problems. In this prediction, the random forest model scored the highest, indicating that the model can fit the data well and has strong generalization ability.

Linear regression model is a kind of regression model based on linear relationship, which assumes that there is a linear relationship between the independent variable and the dependent variable, and solves the model parameters by the least squares method. The linear regression model encounters dimensional catastrophe problems when dealing with high-dimensional data, but performs better in low-dimensional data and data with more obvious linear relationships. In this prediction, the linear regression model scored the next highest score, indicating that the model is able to capture the linear relationship in the data and has some predictive ability.

The decision tree model is a regression model based on a tree structure, which constructs a decision tree by recursively dividing the data set into multiple subsets, each corresponding to a leaf node. The decision tree model performs better in dealing with nonlinear relationships, but is prone to overfitting problems. In this prediction, the decision tree model scored the lowest, indicating that the model cannot fit the data well and may have overfitting problems.

In summary, the random forest regression model scored the highest, indicating that this model performed best in this prediction. The linear regression model was the next highest and the decision tree model performed the worst.

References

[1] *Ugochukwu O ,Elochukwu U .Machine learning for an explainable cost prediction of medical insurance[J].Machine Learning with Applications,2024,15100516-.*

[2] *Nurahmed H T ,Mekashaw E B .Dropout rate and associated factors of community-based health insurance beneficiaries in Ethiopia: a systematic review and meta-analysis[J].BMC Public Health,2023,23(1):2425-2425.*

[3] *Nurahmed H T ,Mekashaw E B .Dropout rate and associated factors of community-based health insurance beneficiaries in Ethiopia: a systematic review and meta-analysis[J].BMC Public Health,2023,23(1):2425-2425.*

[4] *C R K V ,A J C R V V ,Michel O .Risk Adjustment in Health Insurance Markets: Do Not Overlook the Real Healthy.[J].Medical care,2023,*

[5] *Siegfried G ,Juliane T ,Stefanie S , et al.Decreasing COPD-related incidences and hospital admissions in a German health insurance population[J].Scientific Reports, 2023,13(1): 21293-21293.*

[6] *Vicky V ,L J W ,Michelle K .Experiences of low-income college students in selection of health insurance, access, and quality of care.[J].Journal of American college health : J of ACH,2023,11-10.*

[7] *S. A M ,Katia B ,Michelle K , et al.Access to Specialized Care Across the Lifespan in Tetralogy of Fallot[J].CJC Pediatric and Congenital Heart Disease,2023,2(6PA):267-282.*

[8] *Jungtaek L .Effects of private health insurance on healthcare services during the MERS Pandemic: Evidence from Korea[J].Heliyon,2023,9(12):e22241-e22241.*

[9] *Okensama L ,E B A ,A K H , et al.United States insurance coverage of immediate lymphatic reconstruction.[J].Journal of surgical oncology,2023,*

[10] *HealthPartners Debuts 2024 Medicare Advantage Plans[J].Manufacturing Close - Up,2023,*