Performance Evaluation of Financial Industry Related Expense Forecasting Using Various Regression Algorithms for Machine Learning

Liangyong Yao [1, a], Yan Lin [2, *], Yalun Mo [1, b], Feng Wang [1, c]

[1] Hangzhou Wasu Media Television Network Co., Ltd. Zijing Branch, Hangzhou, Zhejiang, China

[2] Zhejiang China Radio and Television Network Co., Ltd.Hangzhou, Zhejiang, China

* Corresponding Author Email: liny@wasu.com, [a] yaoly@wasu.com, [b] 457283362@qq.com, [c] wangfeng5@wasu.com

Abstract. Insurance costs refer to the fees charged by insurance companies to customers to pay for possible risks and losses. Insurance costs are usually based on the personal information of the insured, such as age, gender, occupation, health status and so on. For insurance companies, it is very important to accurately predict insurance costs, because it is directly related to the company's profits and risk control capabilities. The purpose of using regression algorithm to predict insurance expenses is to make insurance companies evaluate customers' risks more accurately and make more reasonable insurance expenses, so as to better manage risks and improve the company's profitability. In addition, for individuals, knowing their own insurance cost forecast results will also help them make better decisions and choose the most suitable insurance products to protect themselves and their families.In order to improve the pricing accuracy and profit rate of insurance companies, this study uses regression algorithm to predict insurance costs. It uses real anonymous data sets, which contain information of the insured from different regions, different ages, different sexes and different smoking status. It uses the comparison algorithm function of regression algorithm, which contains dozens of algorithms and covers all regression algorithms and compare their prediction performance. Our data set takes into account various factors that affect the insurance cost, such as age, gender, body mass index, smoking status and so on. And add them to the model as independent variables. It uses cross-validation to evaluate the generalization ability of the model and R2 index to evaluate the prediction performance. The results show that GBR has the best prediction performance, with R2 of 87%. Our research provides an accurate method for insurance companies to predict insurance costs, which is helpful for insurance companies to formulate more reasonable pricing strategies and improve market competitiveness.

Keywords: Regression, insurance, prediction, machine learning.

1. Introduction

The insurance industry is a kind of risk management business, whose purpose is to provide financial support to customers when they suffer some unforeseen losses. Insurance companies need to formulate insurance fees according to factors such as customers' risk levels and historical data. This will ensure that the insurance company can make a profit when making claims, and at the same time ensure that customers can be guaranteed when taking risks. Using regression algorithm to predict insurance cost is a method based on historical data and risk assessment. Regression algorithm is a method to predict unknown values by establishing mathematical relations between variables, which can be used to predict continuous numerical variables. When forecasting the insurance cost, It can use regression algorithm to forecast the insurance cost of customers. The training process of regression model needs to use a lot of historical data, such as the age, gender, occupation, disease history and other information of customers, as well as the historical claims data of customers. Through the analysis of these data, the regression algorithm can learn the mathematical relationship between variables, thus predicting the insurance cost of customers. The significance of the regression algorithm in forecasting insurance expenses lies in that it can help insurance companies to predict insurance expenses of customers more accurately. In this way, insurance companies can formulate

more reasonable insurance plans and prices and reduce unnecessary losses and risks. At the same time, customers can better understand the risks they need to bear and the corresponding premiums and improve customer satisfaction and loyalty. In addition, for the insurance industry, the use of regression algorithm to predict insurance costs can also help insurance companies improve market competitiveness and economic benefits. From a global perspective, there are many ways to use algorithms to predict insurance costs, and their uses are different. It is a good way to change the operation mode of insurance with mixed model, because it is more transparent and fairer to policyholders, and the details of the amount are clear at a glance [1]. In addition, using data mining technology to price premiums can better control risks [2]. The application of neural network method in automobile insurance is a new attempt, which can predict the problem of claim payment[3]. As mentioned earlier, the effect of mixed mode on insurance is also mentioned. Here, some scholars have studied the important factors that affect the purchase of insurance policies by new users by combining various models [4]. Calculating the premium is another application in the insurance field, and he pays attention to estimating the premium. The value of the final premium can be calculated by using the improved Markov model to improve the market competitiveness of insurance types [5]. Using machine learning to design insurance packages based on demand is a new application, which will attract customers' attention [6]. Using GLM algorithm can reasonably supervise the premium [7]. The analysis of national disease data can predict high-cost medical expenses [8]. In short, various applications in the insurance field provide a good space for machine learning algorithms, and various applications will become more important in the future.

2. Methodology

2.1. Gradient Boosting Regressor

An ensemble learning algorithm based on decision tree is used to solve regression problems. Based on the idea of gradient lifting, it trains a new model each time to fit the residual of the previous model, and finally weights and sums the prediction results of all models to get the final prediction result. It usually has better prediction performance and generalization ability than a single decision tree. In GBR, each model is a decision tree, and each decision tree contains several nodes and leaf nodes, and each node corresponds to a feature and a threshold[9]. By comparing the eigenvalues and thresholds of samples, samples are assigned to the left subtree or the right subtree. The leaf node corresponds to an output value, that is, the prediction value of the model to the input features.

$$\hat{y} = \sum_{i=1}^{n} f_i(x) \tag{1}$$

Where $\hat{y}$ is the predicted value of the model, n is the number of models, and $f_i(x)$ is the predicted value of the i model to the input sample x. In the training process, the predicted value of each model is the residual of the previous model, namely:

$$r_{ij} = y_j - \sum_{k=1}^{i-1} f_k(x_j) \tag{2}$$

Where y_j is the real output value of the sample j, i is the number of the current model, $f_k(x_j)$ is the predicted value of the k model to the sample j, and r_{ij} is the i model to the sample j. Every new model must fit the residual r_{ij}, and finally get a new model fi(x), which is added to the current model sequence. The parameters of GBR include decision tree depth, learning rate, sub-sampling and so on, which can be optimized by cross-validation and other methods.

2.2. Random Forest Regressor

The decision tree integration method is used for regression analysis. It is a powerful algorithm and is suitable for many different types of regression problems. Random Forest Regressor is used to predict the value of a target variable, which is composed of multiple input variables. For example, if we want to predict the price of a person's house, it can use Random Forest Regressor to predict the

price, which may be affected by many factors such as the location, size and building year of the house[10]. The calculation formula of Random Forest Regressor is relatively complex, and it is an integrated method based on decision tree. In Random Forest, It creates multiple decision trees and combine them to produce the final prediction result. Each decision tree is trained on a random subset of data, which helps to prevent over-fitting and improve the robustness of the model. Random Forest's formula is as follows:

$$\hat{y}i = \frac{1}{N}\sum j = 1^{N} T_j(x_i) \tag{3}$$

Where $\hat{y}i$ is the predicted value of the target variable, $ N $ is the number of decision trees, and $T_j(x_i)$ is the predicted value of the j decision tree pair input xi. Random Forest uses the average value to calculate the predicted values of multiple decision trees, so as to get the final prediction result.

2.3. Residuals

Refers to the difference between the observed value and the predicted value of the model, which is a common way to evaluate the goodness of fit of the model. Residual is a measure of the model's ability to explain data. If the model can explain data well, the residual should be small, otherwise it will be large. Usually, it uses residuals to evaluate whether the model can fit the data well, that is, whether the model can explain most of the variance in the data. If the residual distribution of the model is uneven or there are obvious patterns, it may indicate that the assumptions of the model are not in line with the actual situation and the model needs to be adjusted or re-selected. In regression analysis, residuals are often used to evaluate the goodness of fit of models. Generally speaking, the smaller the residual is, the better the model fits[11]. In addition, the normality and independence of residuals are also important indicators for evaluating regression models. Residual is an important tool to evaluate the goodness of fit of the model. By analyzing and testing the residual, it can help us to judge whether the model conforms to the actual situation and whether it needs to be improved and optimized. The formula is:

$$e(i) = y(i) - \hat{y}(i) \tag{4}$$

Where e(i) represents the residual of the i observation, y(i) represents the actual value of the i observation, and (i) represents the predicted value of the i observation. This formula shows that the residual is the difference between the observed value and the predicted value, that is, the error between the actual value and the predicted value. The smaller the residual is, the better the prediction ability of the model is.

2.4. Prediction error

Difference or error between the predicted value and the actual observed value. The smaller the prediction error, the higher the accuracy of the model. Data sets are usually divided into training sets and test sets. In the training process, the model is trained by using the training set data, and the performance of the model is evaluated by the test set data. The prediction error is to measure the performance of the model by comparing the difference between the actual observed values of the test set and the predicted values of the model. Various indicators are usually used to calculate the prediction error, including mean square error (MSE), mean absolute error (MAE) and log loss. Taking the R^2 as an example, the calculation formula is as follows:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{5}$$

Where SS_{res} is the sum of squares of residuals, indicating the difference between the predicted values of the model and the actual observed values; S_{tot} is the sum of total squares, which indicates the difference between the dependent variable and its mean. The numerator in the formula represents the degree of variation of the dependent variable that cannot be explained by the model, and the denominator represents the total degree of variation of the dependent variable[12]. It can be seen that R^2 measures the fitting degree of the model by calculating the ratio of the degree of variation of the

dependent variable that can be explained by the model to the total degree of variation. When the predicted value of the model is very close to the actual observation value, SS_{res} will become smaller, and R^2 will be close to 1, indicating that the model fits well. On the other hand, when there is a big gap between the predicted value of the model and the actual observed value, SS_{res} will increase, and R^2 will be close to 0, indicating that the model is poorly fitted.

2.5. Feature importance Plot

A chart for visualizing the importance of features. In data analysis and machine learning, multiple features are usually used to train models. The importance of features can help us understand which features have the most influence on the prediction results of the model. In the feature importance diagram, the influence of each feature on the prediction accuracy of the model is usually displayed. Generally speaking, the higher the feature importance score, the more important the feature is to the prediction results of the model. This kind of chart can help us identify the most important features, so as to better understand how the model predicts.

3. Results

3.1. Data set

An insurance basic data from an insurance company, this data set contains seven data fields, such as age, gender, bmi, number of children in the family, smoking or not, region and charges. The total data volume is 1038. It will use 70% of the data for algorithm training and 30% for testing. The target field of forecast is insurance expense. Table 1 below will show some data:

Table 1. Date set

Age	Sex	Bmi	Children	Smoker	Region	Charges
36	male	27.55	3	no	northeast	6746.7425
53	female	22.61	3	yes	northeast	24873.3849
56	female	37.51	2	no	southeast	12265.5069
…	…	…	…	…	…	…

3.2. Compare models

By comparing the algorithm functions, the best performance algorithm is obtained in the same data set. In this paper, the performance of R^2 is compared. Table 2 is the result data of the best algorithm finally obtained.

Table 2. Performance comparison of algorithm models

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Gradient Boosting regressor	2541.0830	21037135.7904	4452.2000	0.8453	0.4536	0.3202
Random Forest Regressor	2621.3262	22649147.1096	4642.6662	0.8348	0.4797	0.3421

3.3. Analyze model

In this paper, residuals, prediction error and feature importance plot will be used to analyze the Gradient Boosting regressor graphically. The first is residuals, you can see figure1:

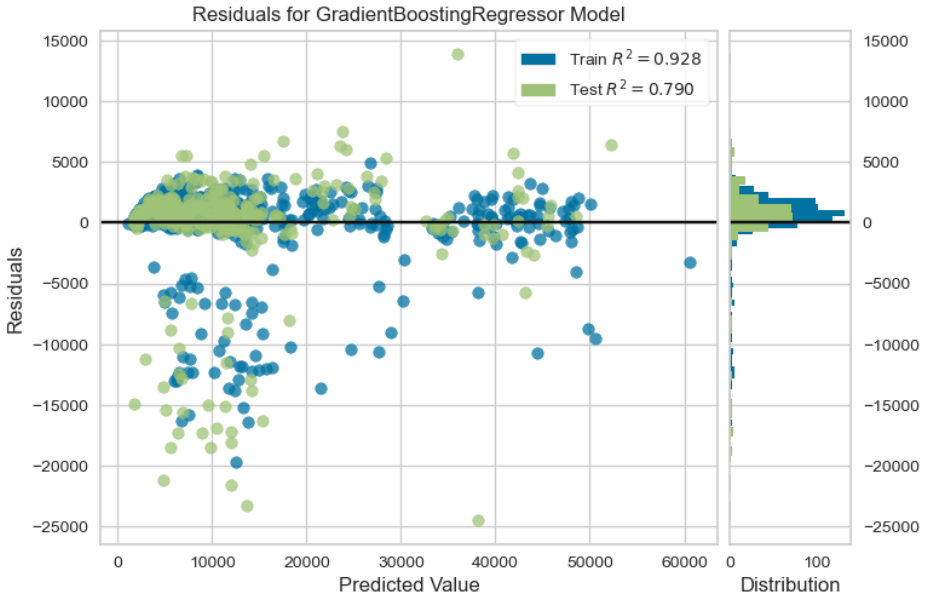

Fig. 1. Residuals for gradient boosting regressor model

Although R^2 reaches 92.8% in the training set, it is only 79% in the test set, which shows that the generalization ability of the algorithm is not strong, and there is a certain gap between the training and test results. In addition, over-fitting may also lead to differences in residuals between the training set and the test set. If the model is over-fitted on the training set, it may over-match the noise and details in the training set, which may lead to poor performance on the test set. Therefore, when evaluating the prediction performance of the model, it is necessary to evaluate the performance of the model on different data sets by cross-validation, so as to understand the generalization ability of the model more accurately.

Looking at Figure 2, the best fit line is basically consistent with the verification line, R^2=79%. and most of the points are on the diagonal, but some points are far from the diagonal.

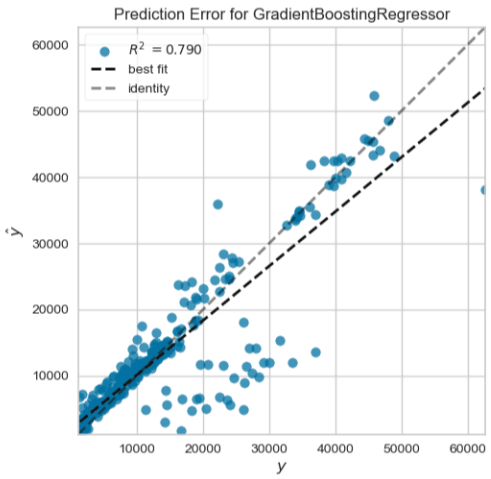

Fig. 2. Prediction error for Gradient Boosting regressor

Look at the Figure 3, smoking has the greatest influence on the final premium, and the second influence is bmi, because smoking will lead to health, which will lead to the increase of premium.

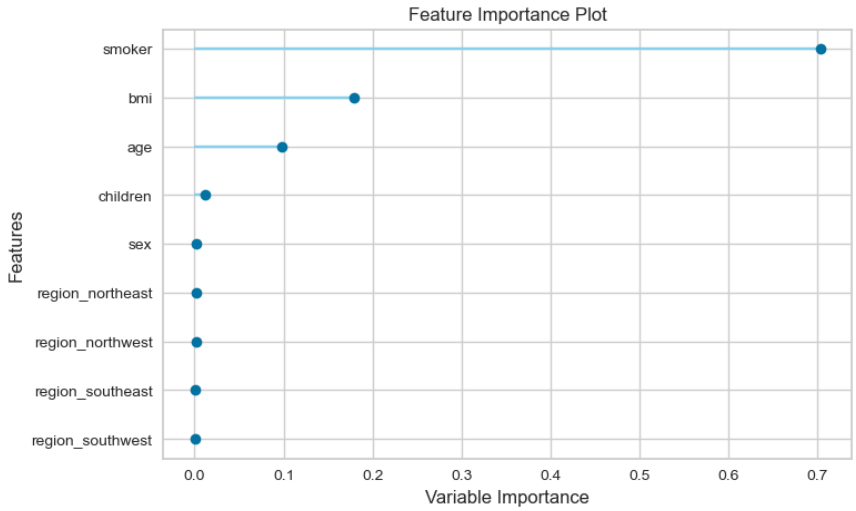

Fig. 3. Feature importance plot

3.4. Prediction and Test data

Through a series of operations such as parameter optimization of the algorithm, It is finally obtained that R2 is 87% as shown in the table 3

Table 3. Optimized Gradient Boosting regressor

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Gradient Boosting regressor	2401.6756	15038735.6142	4021.25631	0.8710	0.4012	0.2973

Table 4 shows the comparison between the predicted value and the actual value in the training set. Although the two values are very close, there are still differences.

Table 4. Comparison between predicted value and actual value of training set

Age	Sex	Bmi	Children	Smoker	Region	Charges	Prediction_label
36	male	27.55	3	no	northeast	6746.7425	8345.66
53	female	22.61	3	yes	northeast	24873.3849	24613.59
56	female	37.51	2	no	southeast	12265.5069	12380.81
…	…	…	…	…	…	…	…

Table 5 shows the comparison between the predicted value and the actual value in the test set, and the comparison result is worse than that in the training set.

Table 5. Comparison between predicted value and actual value of test set

Age	Sex	Bmi	Children	Smoker	Region	Charges	Prediction_label
33	female	18.5	1	no	southeast	4766.0219	5557.3022
26	male	27.05	0	yes	northeast	17043.34	21117.7366
64	female	31.8	2	no	northeast	16069.08	14459.0491
…	…	…	…	…	…	…	…

4. Conclusion

Regression algorithm can use known insurance data to predict future insurance costs, thus helping insurance companies to better evaluate risks and formulate insurance strategies. In this paper, the age, sex, smoking history and health index of the insured are analyzed by regression algorithm to predict the insurance cost, although the performance is 87% in the end. However, there is still a certain gap between the predicted value and the actual value, especially in the test set. Therefore, this point is the biggest deficiency in this paper, and it is necessary to use a larger number of samples and other algorithms such as neural network algorithm in the future research to verify whether it can get better results. In a word, using regression algorithm to predict insurance costs can provide valuable information for insurance companies to help them better manage risks and formulate insurance strategies. However, any prediction result should be further analyzed and evaluated to ensure its accuracy and reliability.

References

[1] R. Cahyandari, R. L. Ariany, Sukono, and Y. S. Perkasa, "The Hybrid Model Algorithm on Sharia Insurance," J. Phys.: Conf. Ser., vol. 1090, p. 012080, 2018, doi: 10.1088/1742-6596/1090/1/012080.

[2] Amela Omerašević and Jasmina Selimović, "Risk factors selection with data mining meth-ods for insurance premium ratemaking," Zb. rad. Ekon. fak. Rij., vol. 38, no. 2, 2020, doi: 10.18045/zbefri.2020.2.667.

[3] G. Tzougas and K. Kutzkov, "Enhancing Logistic Regression Using Neural Networks for Classification in Actuarial Learning," Algorithms, vol. 16, no. 2, p. 99, 2023, doi: 10.3390/a16020099.

[4] Y.-S. Chen, C.-K. Lin, Y.-S. Lin, S.-F. Chen, and H.-H. Tsao, "Identification of Potential Valid Clients for a Sustainable Insurance Policy Using an Advanced Mixed Classification Model," Sustainability, vol. 14, no. 7, p. 3964, 2022, doi: 10.3390/su14073964.

[5] Y. Antonio, S. W. Indratno, and S. W. Saputro, "Pricing of cyber insurance premiums using a Markov-based dynamic model with clustering structure," PLoS ONE, vol. 16, no. 10, p. e0258867, 2021, doi: 10.1371/journal.pone.0258867.

[6] I. Matloob, S. A. Khan, F. Hussain, W. H. Butt, R. Rukaiya, and F. Khalique, "Need-Based and Optimized Health Insurance Package Using Clustering Algorithm," Appl. Sci., vol. 11, no. 18, p. 8478, 2021, doi: 10.3390/app11188478.

[7] S. Xie and R. Luo, "Measuring Variable Importance in Generalized Linear Models for Mod-eling Size of Loss Distributions," Mathematics, vol. 10, no. 10, p. 1630, 2022, doi: 10.3390/math10101630.

[8] Y. Choi, J. An, S. Ryu, and J. Kim, "Development and Evaluation of Machine Learning-Based High-Cost Prediction Model Using Health Check-Up Data by the National Health In-surance Service of Korea," IJERPH, vol. 19, no. 20, p. 13672, 2022, doi: 10.3390/ijerph192013672.

[9] N. Bagalkot, A. Keprate, and R. Orderløkken, "Combining Computational Fluid Dynamics and Gradient Boosting Regressor for Predicting Force Distribution on Horizontal Axis Wind Turbine," Vibration, vol. 4, no. 1, pp. 248–262, 2021, doi: 10.3390/vibration4010017.

[10] W. Ding and X. Qie, "Prediction of Air Pollutant Concentrations via RANDOM Forest Re-gressor Coupled with Uncertainty Analysis—A Case Study in Ningxia," Atmosphere. (Ba-sel)., vol. 13, no. 6, p. 960, 2022, doi: 10.3390/atmos13060960.

[11] N. Mu, "Research on Injury Causes and Prevention Effect of College Rowing Athletes Based on Multiple Regression and Residual Algorithm," J. Environ. Public Health, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/4896336.

[12] J. B. Holmes, K. G. Dodds, and M. A. Lee, "Estimation of genetic connectedness diagnos-tics based on prediction errors without the prediction error variance–covariance matrix," Genet Sel Evol, vol. 49, no. 1, 2017, doi: 10.1186/s12711-017-0302-9.