



Insurance Pricing from the Perspective of Healthcare Cost Prediction—— An Empirical Analysis Based on Machine Learning Using UCI Data

Zetao Shi

School of Mathematical Sciences, Zhejiang University of Technology, Hangzhou, 310014, China

302023316006@zjut.edu.cn

Abstract. In the process of continuous development of the medical insurance industry, the pricing mechanism of medical insurance is gradually maturing. In this context, it has become a key issue to carry out more scientific and reasonable medical cost prediction and insurance pricing based on individual risk characteristics. Based on the UCI medical insurance data set, this study carefully selected age, gender BMI, Smoking status, number of children, region and other representative variables. Using linear regression, random forest and XGBoost, three common and effective model building methods, this paper makes an in-depth study of individual annual medical expenses. By comparing and analyzing the results of different models, this paper comprehensively analyzes the influencing factors of individual annual medical expenses and the prediction effect of each model. Finally, it was found that smoking, age and BMI were the core factors affecting medical expenses, and smoking had the most prominent impact on medical expenses. The effect of BMI on medical expenses is nonlinear, and the ability of the number of children to explain the change of medical expenses is relatively weak.

Keywords: Health Insurance Pricing, Medical Cost Prediction, Machine Learning, Xgboost, Algorithmic Fairness.

1 Introduction

The reform of the medical insurance pricing mechanism is an important issue in the current optimization of the medical security system. The existing insurance pricing model generally has a certain degree of "one-size-fits-all" phenomenon. For people with different physical conditions and family structures, insurance companies often charge relatively uniform premiums or rely too much on single risk factors such as age, while ignoring the differences in individual health behaviors. This pricing method is likely to lead to low-risk groups subsidizing high-risk groups to a certain extent, which not only reduces the efficiency of the insurance market but also weakens the incentive mechanism for health behaviors. Rothschild and Stiglitz explained from classical theory that information asymmetry will lead to risk screening

problems in the insurance market [1]. Recent studies have further shown that, in the context of big data and artificial intelligence, insurance companies can identify individual risk differences through more refined risk classification, digital screening and data analysis, and optimize pricing and underwriting decisions accordingly [2]. In addition, with the development of big data technology, the risk classification mechanism in the insurance market is shifting from traditional experience judgment to data-driven refined identification, but this process is also accompanied by real-world controversies on privacy protection and pricing fairness [3].

With the continuous accumulation of medical data, machine learning technology has gradually shown obvious advantages in the field of medical cost prediction, providing a new technical path for differentiated insurance pricing. However, existing studies are mostly focused on improving the accuracy of model prediction, and there is still a lack of sufficient discussion on the fairness constraints and discrimination risks that may be involved in differentiated pricing [4]. Based on the UCI Medical Insurance dataset (1338 samples), this paper constructs linear regression, random forest, and XGBoost models for comparative analysis, identifies the main driving factors of medical expenses, and analyzes the role mechanism of health behavior variables in insurance pricing. The study found that the impact of smoking on medical expenses is significantly heterogeneous among different age groups, suggesting that age may play a moderating role [5]. There is a non-linear relationship between BMI and medical expenses [6]. In contrast, the number of children has a weaker explanatory power for medical expenses. Through the feature interpretation method of machine learning models, this paper further analyzes the contribution of different risk factors and explores their policy implications in insurance pricing.

2 Literature Review

In recent years, research on medical insurance pricing has gradually shifted from traditional pricing models based on demographic characteristics to methods that rely more on behavioral economics and data-driven approaches. Rothschild and Stiglitz emphasized the problem of information asymmetry in classic insurance market theory and pointed out that insurance companies should consider the implicit health risks of policyholders when pricing [1]. However, traditional linear pricing models often rely on single risk factors (such as age or gender) and fail to fully consider factors such as health behavior. With the popularization of medical big data, more and more scholars have begun to try to introduce machine learning algorithms to improve the accuracy of medical cost prediction by modeling large amounts of data. Existing research shows that machine learning methods have strong applicability in medical cost prediction and insurance pricing, especially in more effectively characterizing complex nonlinear features and variable interaction relationships that are difficult to identify by traditional linear models. For example, Drewe-Boss et al. found in their study based on large-scale health insurance claims data that deep learning models have strong applicability in individual medical cost prediction and can more effectively capture the complex nonlinear relationship between medical costs and risk

factors, thereby improving prediction accuracy [6]. Recent studies have also shown that the relationship between BMI and medical expenses or medical resource utilization is often not a simple linear one, but may exhibit a J-shaped or other nonlinear characteristics [7]. In addition, recent studies have further shown that machine learning can not only be used for individual medical expense prediction, but also for the construction of risk-adjusted payment formulas, thus providing more refined technical support for insurance pricing and payment mechanism design [8]. Gabrielli and Wüthrich showed from the perspective of individual claims history modeling that deep learning methods can more meticulously depict the individual-level claims process, providing new ideas for insurance claims prediction [9]. However, the application of machine learning models in insurance pricing is not without controversy. Dwork et al. raised the issue of "fairness", that is, in insurance pricing, the algorithm should avoid unreasonable pricing discrimination based on non-intervention factors such as gender and region [4]. In the context of insurance analysis, prediction accuracy, model interpretability and fairness have been regarded as core issues that need to be balanced simultaneously. Overly pursuing refined pricing may bring new discrimination and governance risks [10]. Therefore, how to reasonably distinguish between modifiable risk factors such as smoking and obesity and non-modifiable characteristics such as age and gender in insurance pricing, and how to construct a risk differentiation mechanism that balances pricing accuracy and fairness, has become an important topic of current research.

3 Research Design

3.1 Data Description

This paper uses the UCI Medicare dataset, which contains 1,338 records documenting the insured's basic demographic characteristics, health behaviors, family structure, and corresponding annual medical expenses. Variables in the dataset include: age, sex, body mass index (BMI), smoking status, number of children, and region. The target variable is the individual's annual medical expenses (charges), which exhibits a right-skewed distribution, with some high-risk individuals experiencing extreme medical expenses.

3.2 Variable Specification

This paper uses the individual's annual medical expenses (charges) as the dependent variable to measure the insured's level of medical expenditure. Explanatory variables mainly include demographic characteristics, health behaviors, and family structure. Demographic variables include age and sex, where age reflects the trend of individual health risk changes over the life cycle, and sex is used to examine its potential impact on medical expenses. Health behavior variables primarily include Body Mass Index (BMI) and smoking status. BMI measures an individual's weight health status, while smoking is considered a significant and modifiable risk factor. Family structure variables include the number of children, reflecting the potential impact of family

structure on health behavior and healthcare needs. Furthermore, this paper introduces a region variable to characterize the impact of differences in healthcare resources and lifestyles across different regions on healthcare costs. To further identify nonlinear relationships and interaction effects among variables, this paper constructs two extended variables: a smoking-age interaction term (smoker_age_interact) to examine the differences in the impact of smoking behavior on healthcare costs at different age stages, and a BMI squared term (bmi²) to capture the potential nonlinear relationship between BMI and healthcare costs.

4 Model Selection

4.1 Linear Regression Model

First, a linear regression model is used as the baseline model for training. This paper inputs the data from the training set into the regression model, outputs the prediction results, and evaluates the model performance. The specific regression equation is as follows:

$$\text{Charges}_i = b_0 + b_1 \text{age}_i + b_2 \text{BMI}_i + b_3 \text{Smoker}_i + b_4 \text{Children}_i + b_5 \text{Region}_i + b_6 \text{Sex}_i + \varepsilon \quad (1)$$

In this model, an interaction term between smoking and age is introduced to examine the impact of smoking on healthcare costs in the elderly population.

4.2 Random Forest Regression Model

To capture the nonlinear relationships in the data, this paper uses a random forest regression model. The specific regression equation is as follows:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (2)$$

Random forest is an ensemble learning method that reduces overfitting and improves the model's generalization ability by constructing multiple decision trees and averaging their results. In this paper, 100 decision trees were used during model training, and hyperparameters were tuned to improve prediction performance.

4.3 XGBoost Regression Model

XGBoost is an optimization algorithm based on Gradient Boosting Trees (GBDT), particularly suitable for handling large-scale data and achieving high prediction accuracy. The specific regression equation is as follows:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (3)$$

This paper uses XGBoost to train the model on the training set and evaluates its performance on the test set. To improve model performance, hyperparameters such as 'max_depth' and 'learning_rate' are adjusted, and a base model of 200 trees is used.

4.4 Model Evaluation

In the model evaluation phase, this paper uses multiple metrics to comprehensively compare the predictive performance of different models. First, the coefficient of determination (R²) is used to measure the model's ability to explain the variation in

sample data; a higher R^2 value indicates a better fit to the data. Second, the root mean square error (RMSE) is used to measure the deviation between predicted and true values. This metric is sensitive to larger errors and reflects the model's overall predictive stability. Simultaneously, the mean absolute error (MAE) is used as an auxiliary metric to measure the average level of prediction error; a smaller MAE indicates that the model's prediction is closer to the true value. By comprehensively comparing metrics such as R^2 , RMSE, and MAE, this paper systematically evaluates the performance of different models in medical cost prediction and further analyzes the advantages and disadvantages of each model in terms of prediction accuracy and stability.

5 Experimental Procedure

First, this paper imports the UCI medical insurance dataset and conducts basic exploratory data analysis. The shape, missing values, and data types of the data were examined. First, the missing values and variable types were checked, and appropriate preprocessing was performed as necessary. For categorical variables (such as gender, smoking status, and region), they were coded and converted into numerical data suitable for machine learning algorithms.

In terms of feature construction, this paper further introduced several extended variables based on the research hypotheses. First, to examine the differences in the impact of smoking behavior on medical expenses at different age stages, an interaction term between smoking status and age was constructed. Second, considering the potential non-linear relationship between BMI and medical expenses, a BMI squared term was added to the original BMI variable, and an obesity indicator variable was set to identify the additional cost risks that high BMI individuals may face. Third, to examine the potential impact of differences in family structure on medical expenses, the original number of children variable was grouped to more clearly compare changes in medical expenses under different family burden levels. Through these processing steps, this paper enhanced the model's ability to identify interaction effects and non-linear relationships while preserving the information of the original variables.

6 Empirical Results and Discussion

6.1 Descriptive Statistics

Figure 1 shows that medical charges exhibit a right-skewed distribution. Smokers had significantly higher average medical charges than non-smokers, suggesting that smoking may be a significant risk factor for medical costs. Descriptive statistics on BMI show that most samples had BMIs within the normal range, but a certain percentage of individuals had BMIs exceeding 30, indicating the potential impact of obesity on medical costs.

As shown in Figure 2, to ensure model stability and improve training efficiency, this paper standardizes numerical variables and encodes categorical variables using OneHotEncoder. The data is divided into a training set (80%) and a test set (20%) to ensure the objectivity and accuracy of model training and evaluation.

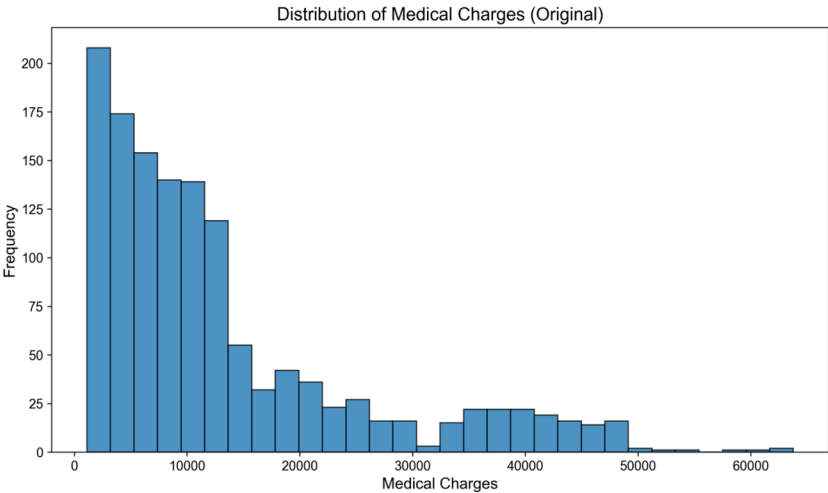


Fig. 1. Distribution of Medical Charges (a) (Picture credit: Original)

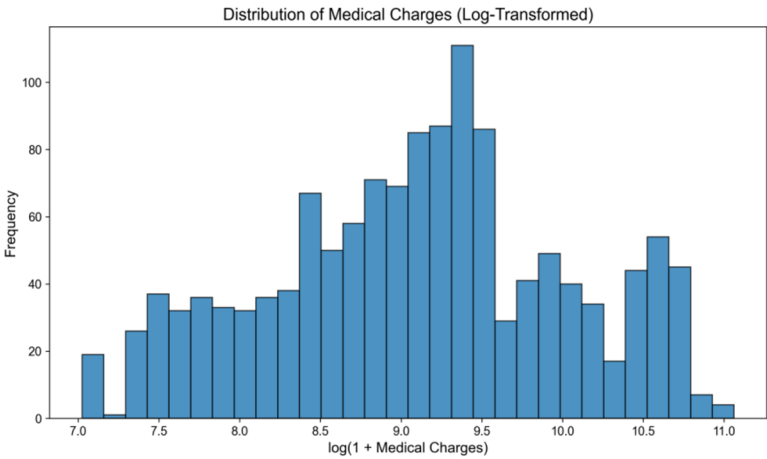


Fig. 2. Distribution of Medical Charge (Log-Transformed) (Picture credit: Original)

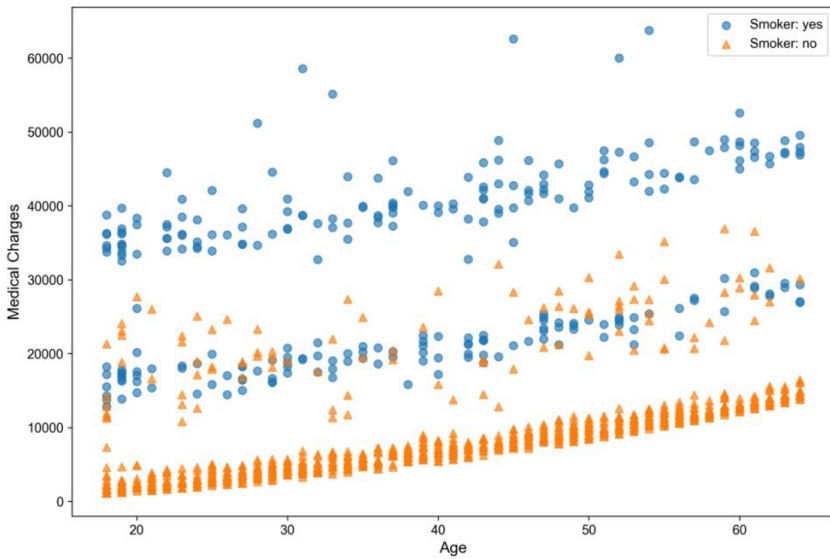


Fig. 3. The interaction effect of smoking and age (Picture credit: Original)

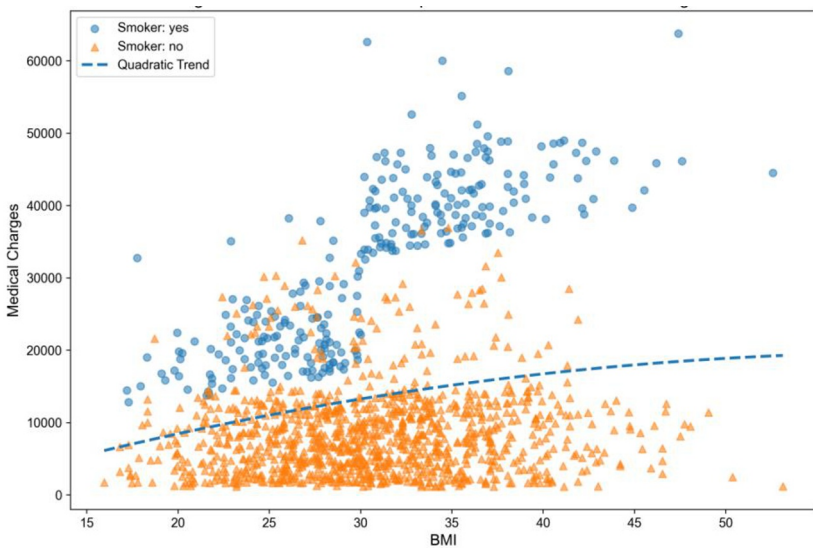


Fig. 4. Nonlinear relationship Between BMI and Medical Charges (Picture credit: Original)

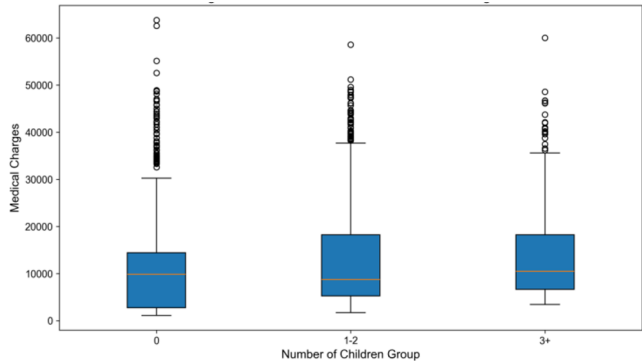


Fig. 5. Number of Children and Medical Charges (Picture credit: Original)

6.2 Model Results

As shown in Figure 3, smoking status has a significant positive impact on medical expenses. As shown in Figure 4, the predicted medical expenses for smokers are significantly higher than those for non-smokers. Furthermore, there is a non-linear relationship between BMI and medical expenses. Higher BMI significantly increases medical expenses. The interaction term between age and smoking also indicates that smoking has a greater impact on medical expenses for the elderly.

The predictive performance of the random forest and XGBoost models is significantly better than that of the linear regression model. As shown in Figure 6, in the XGBoost model, the interaction effects of smoking status, BMI, and age are captured. These factors have a much greater impact on medical expenses than variables such as gender or number of children, as shown in Figure 5. The model comparison results are shown in Table 1. The R^2 value of the XGBoost model reaches 0.89, significantly higher than that of the linear regression model and the random forest model.

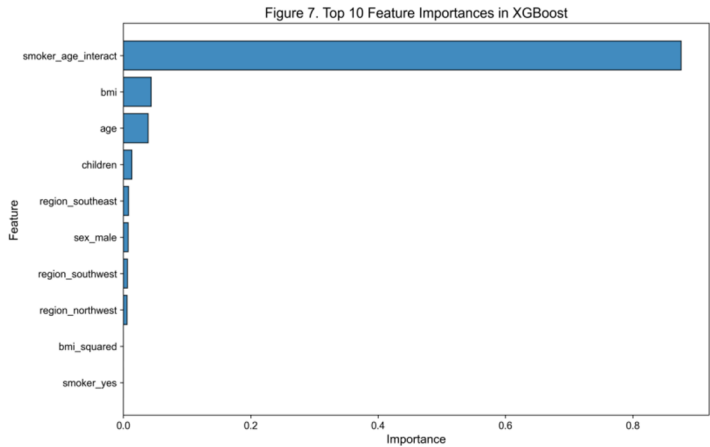


Fig. 6. XGBoost Feature Importance Ranking (Picture credit: Original)

Table 1. Comparison of Results

Model	R ²	RMSE
Linear Regression	0.68	4700
Random Forest	0.85	2900
XGBoost	0.89	2550

6.3 Interpretability Analysis

Feature importance analysis using the XGBoost model identified smoking status and BMI as the two most important features. SHAP value analysis further showed that smoking status has a consistently positive impact on medical costs, while BMI exhibits a typical non-linear effect. Smokers have higher medical costs, and individuals with higher BMI experience a greater increase in medical costs.

7 Conclusion

This paper empirically analyzes the influencing factors of medical costs using the UCI Medicare dataset and employing linear regression, random forest, and XGBoost models. The results indicate that smoking status, age, and BMI are the main factors affecting medical costs, with smoking having the most significant impact and exhibiting a clear interaction effect with age. Compared to traditional linear regression models, the XGBoost model performs better in terms of prediction accuracy and can more effectively capture the non-linear relationship between medical costs and risk factors.

This research suggests that insurance pricing should pay more attention to modifiable risk factors such as health behaviors, including smoking and obesity, and improve pricing accuracy through data-driven methods. Simultaneously, in promoting differentiated pricing, it is also necessary to address algorithmic fairness issues to avoid unreasonable pricing differences due to factors such as gender or region. Future research could incorporate larger-scale medical insurance data for analysis to enhance the model's practical application value.

References

1. Rothschild, M., Stiglitz, J.: Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *The Quarterly Journal of Economics* 90(4), 629-649 (1976)
2. Murphy, F., Xu, W., Xu, X.: The risk screening effect of digital insurance distribution. *Journal of Risk and Insurance* 91(4), 841-866 (2024)
3. Eling, M., Gemmo, I., Guxha, D., Schmeiser, H.: Big data, risk classification, and privacy in insurance markets. *The Geneva Risk and Insurance Review* 49(1), 75-126 (2024)

4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
5. Gu, D., Sung, H., Calfee, C.S., Wang, Y., Yao, T., Max, W.: Smoking-attributable health care expenditures for US adults with chronic lower respiratory disease. *JAMA Network Open* 7(5), e2413869 (2024)
6. Drewe-Boss, P., Enders, D., Walker, J., Ohler, U.: Deep learning for prediction of population health costs. *BMC Medical Informatics and Decision Making* 22(1), 32 (2022)
7. Edwards, C.H., Bjørngaard, J.H., Kinge, J.M., Vie, G.Å., Halsteinli, V., Ødegård, R., Kulseng, B., Bjørnelv, G.W.: The healthcare costs of increased body mass index: evidence from The Trøndelag Health Study. *Health Economics Review* 14(1), 36 (2024)
8. Andriola, C., Ellis, R.P., Siracuse, J.J., et al.: A novel machine learning algorithm for creating risk-adjusted payment formulas. *JAMA Health Forum* 5(4), e240625 (2024)
9. Gabrielli, A., Wüthrich, M.V.: An individual claims history simulation machine. *Risks* 6(2), 29 (2018)
10. Aas, K., Charpentier, A., Huang, F., Richman, R.: Insurance analytics: prediction, explainability, and fairness. *Annals of Actuarial Science* 18(3), 535–539 (2024)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

