Global Scientific and Academic Research Journal of Economics, Business and Management
ISSN: 2583-5645 (Online)
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journals-gsarjebm-home/

Research on Medical Insurance Cost Prediction based on Machine Learning Model

By

Wang Tianmei[1], Ren Yangyang[2]

[1,2]Southwest Petroleum University,No. 8, Xindu Avenue, Xindu District,Chengdu, Sichuan 610500, P. R. China

Article History

Received: 08/08/2025
Accepted: 14/08/2025
Published: 16/08/2025

Vol –4 Issue – 8

PP: -37-42

Abstract

This study aims to utilize machine learning models to predict medical insurance costs. As a prepayment and risk-sharing mechanism, medical insurance plays a significant role in ensuring the health security of individuals and society. However, the factors influencing medical insurance costs are complex and diverse, making accurate prediction a challenge. With the accumulation of medical insurance big data and the development of digital technology, the application of machine learning in predicting medical insurance costs has gradually attracted attention. This study aims to establish machine learning models to identify key factors from historical data and construct models capable of predicting future costs. Through the final results, it provides decision support for policymakers and helps individuals make more informed medical consumption decisions.

Key words: *Medical insurance cost Machine learning Prediction model*

1. Research background

Medical insurance is generally a prepayment and risk-sharing mechanism designed to cover medical expenses incurred due to illness.In China, the payment of medical expenses incurred due to illness is generally covered by basic medical insurance, a social insurance system established to compensate workers for economic losses caused by disease risks. The emphasis placed on health has made the medical field one of the key areas of focus, leading to the emergence of related research on medical insurance, insurance costs, and other aspects that have gradually become crucial research areas in the field of health insurance. However, due to the various factors driving medical insurance costs and their complexity, there are challenges in accurately establishing predictive models. Factors such as demographic information, health status, geographical location, and lifestyle choices can all have a significant impact on the expected cost of medical insurance. Other important factors, such as coverage and plan type, also play a significant role in determining the potential cost of medical insurance. How to reduce patients' medical expenses through these influencing factors is currently a problem that needs to be addressed and dealt with.

With the continuous improvement of the informatization level of the social security system, a massive amount of medical insurance big data has been accumulated for us to explore and utilize. In recent years, digital technology has become increasingly widespread, and machine learning has achieved remarkable results in various fields. Its application in medical insurance expense prediction has also gradually gained attention. Machine learning models can analyze a large amount of historical data to identify key factors affecting medical insurance expenses and construct models capable of predicting future expenses. However, we are still uncertain about the accuracy of machine learning models in predicting medical expenses. Therefore, this study aims to establish a machine learning model for predicting medical insurance expenses, which is of great significance for relevant policymakers to formulate reasonable medical insurance policies. At the same time, individuals can also make more informed medical consumption decisions by understanding the prediction of medical insurance expenses and taking precautions in critical moments.

2. Current research status at home and abroad

2.1 Medical insurance

Generally speaking, medical insurance is a prepayment and risk-sharing mechanism designed to cover medical expenses incurred due to illness. Syed Khurram Azmat et al. (2024) believe that expenses include hospitalization fees, drug fees, and diagnosis and treatment fees. Social and national health insurance have now enabled people to access healthcare services more fairly and protect them from financial risks related to illness[1]. Ugochukwu Orji et al. (2024) elaborated on the specific combinations represented by the three main types of health insurance systems[2]:

1. Health Maintenance Organization (HMO): In this plan, a list of doctors who directly cooperate or have

*Corresponding Author: **Wang Tianmei.**

contracts with the organization is provided for the insured to choose as their primary care physicians.

2. Preferred Provider Organization (PPO): The PPO program operates by providing a list of pre-approved contracted providers. It is common to use a 60/40 split reimbursement, which means that the insurance company pays 60% of the cost and the insured person pays the remaining 40%.
3. High Deductible Health Plan (HDHP) and Health Savings Account (HSA): The working principle of HDHP is to allow the insured to operate a health savings account at the provider's place, deduct treatment expenses from the account, and deduct a possibly lower monthly premium according to an agreed percentage and plan.

In China, generally speaking, it refers to the basic medical insurance system, which is a social insurance system established to compensate workers for economic losses caused by illness. In terms of the main parties involved, Li Lele (2020) pointed out that medical insurance involves three main parties: the insured, designated medical institutions, and medical insurance management departments[3]. In terms of the current situation, Zhu Minglai (2021) and He Wenjiong (2013) noted that China's universal medical insurance system has gradually improved in recent years, and medical service capabilities have significantly increased[45]. However, Yang Heyi et al. (2023) stated that the increasing medical costs are currently exacerbating the economic burden on patients and the pressure on medical insurance fund payments. Therefore, further exploration is needed in the research on medical insurance costs[6].

2.2 Machine learning

Machine Learning (ML), as a subset of Artificial Intelligence (AI), focuses on creating intelligent systems capable of autonomous learning from large datasets. It primarily falls into three broad categories:

1. Supervised learning: Qayyum (2020) pointed out that supervised learning maps the association between the input and output of a set of labeled training data[7].
2. Unsupervised Learning: Unlike supervised machine learning, unsupervised machine learning lacks predefined labels. It autonomously discovers patterns in data, enabling it to effectively identify hidden trends. Common applications include clustering using algorithms such as K-Means and DBSCAN.
3. Semi-supervised learning: Van Engelen Hoos (2020) pointed out that semi-supervised learning (SSL) involves utilizing both labeled and unlabeled data for classification or clustering tasks[8].

In the research on machine learning systems, Carleo et al. (2019) emphasized that although AI encompasses various technologies, including expert systems, deep learning, and robotics, machine learning specifically revolves around data-driven learning[9]. According to Adibimanesh et al. (2023), the widespread adoption of machine learning technology across various industries can be attributed to technological advancements, the computational power of modern GPUs, and the availability of various datasets[10]. However, Panesar (2019) pointed out that despite these achievements, it is worth noting that the full potential of AI and machine learning has not yet been realized, and research in this field is ongoing[11].

2.3 Application of machine learning in the field of medical insurance

Machine learning tools and techniques assist decision-making through data-based predictions and forecasts. Upon reviewing and analyzing relevant literature, it has been found that numerous studies have been conducted on machine learning systems in the field of medical insurance. For instance, foreign scholars such as Ul Hassan et al. (2021) employed various methods including linear regression, support vector regression, Ridge regression, Stochastic Gradient Boosting (SGB), XGBoost, decision tree, random forest regression, multiple linear regression, and k-nearest neighbors to analyze the medical insurance dataset obtained from the Kaggle database. The results indicated that SGB achieved the highest accuracy rate of 86%, with a Root Mean Squared Error (RMSE) of 0.340[12]. Meanwhile, Vimont et al. (2022) compared simple neural networks, random forests, and generalized linear models for predicting individual-level medical expenses. Their findings revealed that the RF model proposed by Vimont et al. exhibited the best performance, with a coefficient of determination (R2) of 47.5%, and a Hit Rate (HiR) of 67%[13]. Domestically, Liu Peng (2019) conducted an in-depth study on the application of association rule mining algorithms in electronic medical record text expression data using Spark machine learning technology, and based on this, established the core logic of a medical insurance fraud behavior audit system[14]. Scholar Li Bin (2022) proposed the use of customized scheduling and P2P-sidecar optimization methods in the medical field through Kubernetes machine learning, satisfying the frequent container change requirements[15].

2.4 Commentary

Whether in single fields or cross-disciplinary areas, scholars both domestically and internationally have conducted numerous studies and achieved abundant research outcomes. However, regarding the prediction of medical insurance expenses, there has not been much elaboration on its accuracy. Therefore, this paper conducts a predictive study on medical insurance expenses based on machine learning models, further exploring its influencing factors, prediction accuracy, and data fitting degree, with the aim of providing useful suggestions for government policymakers and individuals.

3. Research questions

The article explores two issues: first, it seeks to identify the main factors that affect medical costs; second, it evaluates the fitting degree and accuracy of various machine learning models in predicting medical costs.

*Corresponding Author: **Wang Tianmei.**

4. Method description

This article primarily employs the following methods: (1) Logistic regression analysis: a generalized linear regression analysis model commonly used in data mining, automatic disease diagnosis, economic forecasting, and other fields. Logistic regression estimates the probability of an event occurring based on a given set of independent variable data. By constructing a regression model, it analyzes the impact of factors such as age, gender, BMI, and smoking status on medical insurance costs and predicts future cost trends. (2) Ridge regression: (essentially: an improved least squares estimation method). By sacrificing the unbiasedness of the least squares method, it obtains regression coefficients that are more realistic and reliable at the cost of losing some information and reducing accuracy. (3) Decision tree: Based on the known probabilities of various scenarios of medical insurance costs, a decision tree is constructed to calculate the probability that the expected value of net present value is greater than or equal to zero, thereby evaluating the fit of the data. (4) Random forest and gradient boosting regression: Based on the decision tree, further construct random forest models and gradient boosting regression models to predict medical insurance costs.

By establishing the aforementioned models to predict medical insurance expenses, we observe and compare the R2 scores of various models to assess their degree of fit to the data. The closer the result is to 1, the stronger the explanatory power of the model's variables on the dependent variable, and the better the data fit.

Subsequently, by establishing four types of models, namely logistic regression, K-nearest neighbor algorithm, support vector machine, and naive Bayes, and training them to classify individuals into two categories based on their expenses: those above and below the median, the medical insurance expenses were classified and predicted, and the prediction accuracy of each model was evaluated.

5. Experimental results

5.1 Dataset Overview

This dataset is sourced from Kaggle, with a total of 2,772 entries, including 7 items such as age, gender, BMI, region, smoking status, whether there are children in the household and their number, and personal medical expenses.

5.2 Descriptive statistics

According to the results of descriptive statistical analysis, this dataset covers a wide range of ages, especially with a large number of individuals aged between 18 and 22.5 years old; the distribution of males and females, as well as regions (northeast, northwest, southeast, and southwest), is relatively balanced; most people have fewer than three children; most fall within the BMI range, indicating overweight to moderate obesity (29.26 to 31.16); and there are relatively few smokers. The specific situation is shown in Figure 1 below.

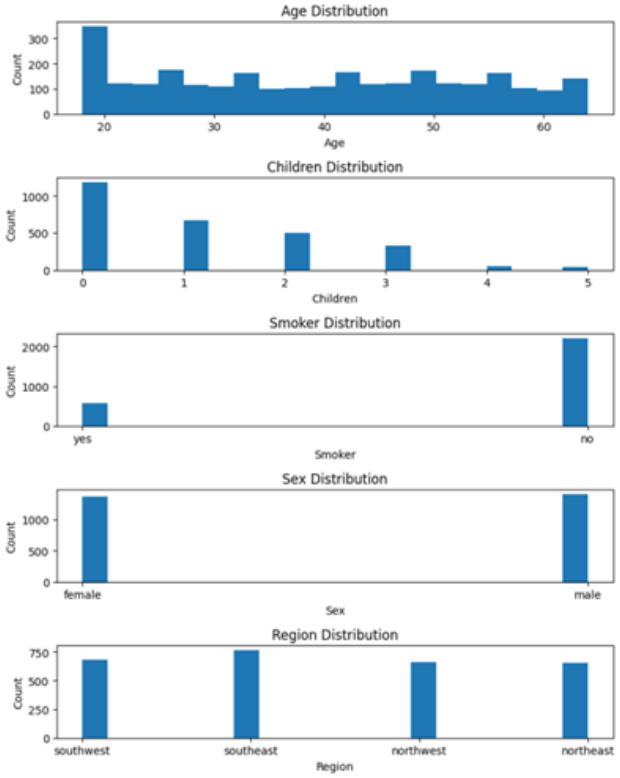
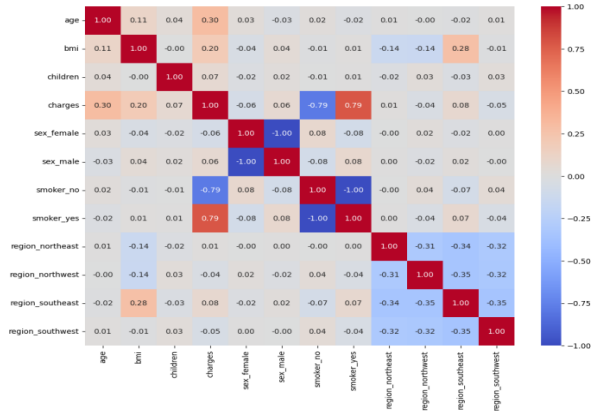


Figure 1 Distribution of survey respondents by age, number of children in the household, smoking status, gender, and region

5.3 Analysis of influencing factors

As shown in Figure 2 below, there is a significant correlation between smoking status and insurance premiums, with a correlation coefficient as high as 79%. This indicates that the insurance premiums of smokers increase significantly due to higher health risks. This is consistent with established knowledge about the harmful effects of smoking on health. In addition, age and BMI also have a positive correlation with premiums, with correlation coefficients of 30% and 20% respectively, reflecting the expected higher medical costs associated with increasing age and BMI levels. Although regional differences have a certain correlation with premiums, their impact is less significant compared to smoking, age, and BMI. Gender and the number of children have the smallest correlation with premiums (6% and 7% respectively), indicating that these factors have little impact on premiums in this dataset.



*Corresponding Author: **Wang Tianmei.**

Figure 2 Correlation result chart of the 7 survey items for the survey respondents

5.3.1Smoking - Costs

From the scatter plot analysis below, two distinct clusters can be observed: the smokers' cluster and the non-smokers' cluster. The clear divergence indicates that, compared to non-smokers, smokers consistently incur higher costs, suggesting that smoking increases medical expenses. Furthermore, significant differences exist within each cluster, particularly among smokers. The wider distribution of cost ranges suggests that factors other than smoking, such as age, BMI, and comorbidities, also significantly affect individual insurance premiums.

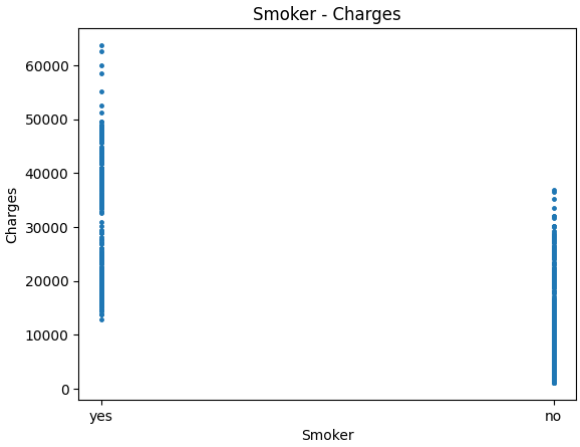


Figure 3 Smoking - Cost Scatter Plot

Table 1 Smoking status and average medical insurance expenses of survey respondents

Does the survey respondent smoke	Smoking	No Smoking
Average medical insurance expenses	$32,223.14	$8417.87

From Table 1, it can be seen that the average medical insurance cost for smokers is $32,223.14, while the average medical insurance cost for non-smokers is $8,417.87. The analysis reveals a significant difference in costs between smokers and non-smokers. Overall, both Figure 3 and Table 1 confirm the hypothesis that smoking status is a key predictor of increased medical costs.

5.3.2Age - Cost

As shown in Figure 4, the dataset exhibits a wide range of age distribution, and the correlation between age and expenditure is generally even, except for a few data points. The data reveals that the expenditure at the age of 21 is the lowest, while those at the ages of 28, 31, 33, 45, 52, and 54 have higher expenditures, with no apparent pattern.

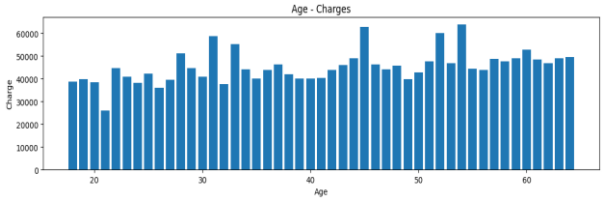


Figure 4 Bar chart of age-cost for survey respondents

5.4 Other data analysis

As shown in Figure 5, the distribution of BMI among different age groups of survey respondents is relatively even, indicating a weak correlation between BMI and age. Additionally, it can be observed from the figure that the BMI of each age group in the survey population is greater than 25, indicating that most of the surveyed population falls into the overweight category and should pay more attention to their health in the future. Furthermore, there is a significant variation in the distribution of the number of children owned by individuals of different age groups in the survey population.

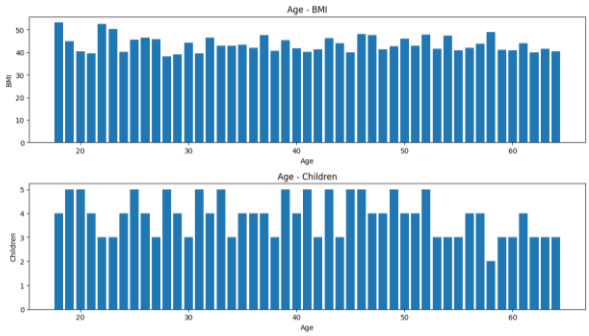


Figure 5 Bar chart of survey respondents' age-BMI and age-number of children distribution

5.5 Fitness and prediction accuracy of machine learning models

5.5.1Model fitting results

Using cost as the target variable, five models were established, namely Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression, to evaluate various regression models for predicting medical insurance costs.

The results are presented in Table 2. Decision tree and random forest regression outperformed linear regression and ridge regression, with random forest achieving a higher R2 score of 0.95 and a lower error metric. Gradient boosting regression achieved a balance between performance and error metrics, with an R2 score of 0.87.

*Corresponding Author: **Wang Tianmei.**

Table 2 Survey respondents' smoking status and average medical insurance expenses

model	Linear regression	Ridge regression	Decision tree regression	Random forest regression	Gradient boosting regression
R2 score	0.73981661775643	0.7397837206133645	0.9480383258895239	0.9504527126517399	0.8745359418641633
MSE score	39933194.54805148	39938243.63305753	7975127.478947112	7604565.0876889415	19256343.733882822
MAE score	4160.247974762996	4162.65122385544	578.5133876936937	1315.5737802157662	2304.7186285469547

5.5.2Model prediction accuracy results

The aforementioned logistic regression, K-nearest neighbor (KNN) classification, support vector machine (SVM) classification, and naive Bayes classification models were trained to categorize individuals into two groups based on their expenses: above or below the median. The objective is to predict the accuracy of the machine learning model.

The results are presented in Table 3. Logistic regression exhibits the highest accuracy rate, reaching 91%, followed by KNN classification (87%), SVM classification closely behind, and Naive Bayes classification having the lowest accuracy rate. The aforementioned results indicate that logistic regression boasts the highest accuracy rate, suggesting a strong linear relationship. This finding is beneficial for making better adjustments in the subsequent classification of medical insurance expenses.

Table 3 Survey respondents' smoking status and average medical insurance expenses

Classification model	Logistic regression	K-Nearest Neighbor algorithm	Support Vector Machine	Naive Bayes
accuracy	0.91	0.87	0.81	0.69

5.5.3ROC curve analysis results

As shown in Figure 6 below, the insurance costs are classified into ROC curves above or below the median. Logistic regression and KNN classification demonstrate high AUC scores, indicating good performance in effectively distinguishing between cost categories. However, SVM classification performs slightly worse, while Naive Bayes classification performs the least well.

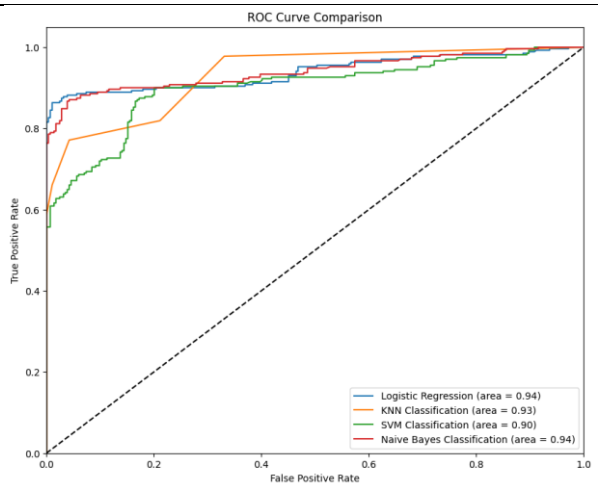


Figure 6 ROC curve graphs of Logistic, KNN, SVM, and Naive Bayes

6. Summary and suggestions

Based on the above data analysis and mathematical simulation results, the main factor affecting medical expenses is smoking, with a correlation of 79%, followed by age and BMI.

Suggestions: (1) The state and government should formulate and implement strict tobacco control regulations to reduce the number of smokers and lower the smoking rate. (2) The government and community should strengthen tobacco control publicity and education to raise public awareness of the dangers of smoking. (3) Medical institutions should provide customized health management and prevention measures for specific health issues of different age groups; for some routine physical examination items (such as blood glucose and blood pressure tests), free physical examination machines should be added, personnel should be allocated, and free physical examinations should be provided to the public. (4) For individuals, it is important to enhance their own health awareness, strengthen exercise, and improve their ability to resist diseases.

In assessing the goodness of fit of various machine learning models in predicting medical costs, random forests and decision trees demonstrate the best performance, with a goodness of fit exceeding 94%. When evaluating the accuracy of these models in predicting medical costs, logistic regression exhibits the highest accuracy and the most excellent ROC curve.

*Corresponding Author: **Wang Tianmei.**

Suggestions: (1) Random forest and decision tree exhibit excellent performance in terms of fitting, and these models can be considered for priority use in medical expense prediction systems to more accurately estimate the medical expenses of individuals or groups. (2) In situations where it is necessary to clearly distinguish between high and low medical expenses or predict specific categories of medical expenses, the logistic regression model can be considered as a priority. (3) In the future, by collecting more diverse and comprehensive data, improving model algorithms and parameter optimization, and introducing more advanced machine learning techniques such as deep learning, the prediction accuracy in medical insurance expenses can be further enhanced.

7. References

1. Azmat S K, Thom E M, Arshad M, et al. A study protocol for integrating outpatient services at the primary health care level as part of the universal health coverage benefit package within the national health insurance program of Pakistan through private health facilities[J]. Frontiers in Public Health, 2024, 12: 1293278.
2. Orji U, Ukwandu E. Machine learning for an explainable cost prediction of medical insurance[J]. Machine learning with applications, 2024, 15: 100516.
3. 李乐乐.基于博弈理论和激励相容原理的医疗保险相关主体行为研究[J].大连理工大学学报(社会科学版),2020,41(06):67-74.
4. 朱铭来,胡祁,赵轶群.关于实现基本医疗保险全民参保的若干思考[J].中国卫生经济,2021,40(01):45-48.
5. 何文炯.从“广覆盖”到“全覆盖”——中国社会医疗保险三大关键[J].中国医疗保险,2013,(02):11-13.
6. 杨赫祎,冯玉,李天俊,等.基于特征筛选与机器学习的医疗保险报销比例预测研究[J].中国循证医学杂志,2023,23(04):373-378.
7. Qayyum A, Qadir J, Bilal M, et al. Secure and robust machine learning for healthcare: A survey[J]. IEEE Reviews in Biomedical Engineering, 2020, 14: 156-180.
8. Van Engelen J E, Hoos H H. A survey on semi-supervised learning[J]. Machine learning, 2020, 109(2): 373-440.
9. Carleo G, Cirac I, Cranmer K, et al. Machine learning and the physical sciences[J]. Reviews of Modern Physics, 2019, 91(4): 045002.
10. Adibimanesh B, Polesek-Karczewska S, Bagherzadeh F, et al. Energy consumption optimization in wastewater treatment plants: Machine learning for monitoring incineration of sewage sludge[J]. Sustainable energy technologies and assessments, 2023, 56: 103040.
11. Panesar A. Machine learning and AI for healthcare[M]. Coventry, UK: Apress, 2019.
12. ul Hassan C A, Iqbal J, Hussain S, et al. A computational intelligence approach for predicting medical insurance cost[J]. Mathematical Problems in Engineering, 2021, 2021(1): 1162553.
13. Vimont A, Leleu H, Durand-Zaleski I. Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France[J]. The European Journal of Health Economics, 2022, 23(2): 211-223.
14. 刘鹏.基于Spark机器学习实现医疗保险关联频繁模式的欺诈行为挖掘技术探讨[J].中国数字医学,2019,14(05):15-18.
15. 李斌,高振宇.基于Kubernetes的医疗领域机器学习节点部署优化[J].中国信息化,2022,(11):34-36+33.

*Corresponding Author: **Wang Tianmei.**