Proceedings of CONF-BPS 2026 Symposium: Innovation, Finance, and Governance for Sustainable Global Growth
DOI: 10.54254/2754-1169/2026.BJ31262

Optimizing Premium Pricing Strategies and Predicting Premiums by Leveraging Insurance Data

Jialiang Zheng

Faculty of Computing, University Technology Malaysia, Johor Bahru, Maylasia
jialiang20@graduate.utm.my

Abstract. To optimize the premium pricing strategy of medical insurance and enhance the accuracy of premium prediction, this study utilized the insurance data of 1,338 customers from the Kaggle platform as the research sample. After data preprocessing, four machine learning models - CatBoost, XGBoost, Gradient Boosting, and LightGBM - were adopted to build advanced predictive models, and their performances were evaluated through comparative analysis. Initially, the key factors influencing medical expenses and premiums were identified through distribution density plots, scatter plots, violin plots, and correlation heatmaps. The model training results indicated that the XGBoost model outperformed the others, with a recall rate of 99% for standard premium customers and an accuracy rate of 98% for high-premium customers. However, when applying this model to predict the results of sample customers, certain errors still occurred. This study provides data-driven support for insurance institutions to formulate scientific premium pricing strategies. In the future, by expanding the sample size and optimizing model parameters, the prediction accuracy can be further improved.

Keywords: Medical Insurance, Premium Pricing, Machine Learning, XGBoost, Premium Prediction.

1. Introduction

For insurance institutions, formulating and continuously improving premium pricing strategies is not only related to operational efficiency, but also the core driving force for long-term sustainable development. At a time when customers demand greater transparency and fairness and regulatory requirements are increasingly stringent, well-designed pricing strategies can make insurance companies stand out, retain existing customers and attract new customers. However, this process requires a lot of time and manpower to identify the factors that have a significant impact on the premium. This traditional manual method will not only delay the launch of the updated pricing scheme, but also may ignore subtle but key variables, resulting in inaccurate pricing, thereby reducing the profit margin or scaring away potential customers. This study uses structured data set analysis to automatically identify the key indicators that affect the premium, and establishes a prediction model for potential customers, which significantly simplifies the analysis workflow and solves this inefficient problem.

DOI: 10.54254/2754-1169/2026.BJ31262

Medical insurance plays a crucial role in safeguarding citizens' health and maintaining social stability, serving as a financial safety net to alleviate the economic burden caused by sudden illnesses. It compensates for the economic losses of workers due to disease risks and is an effective means to narrow the gap between urban and rural areas and promote urbanization [1], as well as a key factor in reducing the vulnerability of families to poverty [2]. From an empirical perspective, there is a statistically significant positive correlation between the coverage rate of comprehensive medical insurance and population health outcomes [3]. Under the impetus of strong market demand (driven by population aging and enhanced health awareness) and the large and diverse customer base of the global insurance industry, medical insurance has become a hot topic and recurring keyword in academic research in fields such as public health, economics, and data science.

The medical insurance systems of various countries in the world are different, but the institutions providing insurance services are mainly divided into two categories: public sector and private companies. Public medical insurance provides basic medical security; However, academic research shows that obtaining public health insurance requires a long waiting time, and the quality of services provided is usually poor [4]. The public medical insurance system has a positive impact on the health status of workers and labor supply. In addition, improving the public medical insurance system can reduce the negative impact of population aging and epidemic on labor supply [5].

Public health insurance plays a vital role in the health insurance system. It mainly undertakes the task of medical security for low-income groups and vulnerable groups. However, as a welfare fund uniformly planned and managed by the public sector, it is restricted by budget restrictions and standardized policies and cannot cover all aspects of citizens' medical treatment process in hospitals, such as elective surgery, expensive drugs or personalized care plans. Therefore, only relying on public health insurance can not meet the diversified medical insurance needs of all people, especially those who have higher health expectations or specific medical needs. As a supplement to public health insurance, private health insurance can meet individual needs for higher quality and more diversified medical services. For example, Fu et al. (year) showed that private health insurance played a positive role in reducing family financial risk [6], while hullegie, P. et al. Believed that private health insurance had a negative impact on the frequency of medical consultation, but had a positive impact on health outcomes [7].

In the private insurance sector, practices such as medical fraud and overcharging have garnered considerable attention. Medical insurance fraud has continuously evolved more sophisticated schemes to evade existing detection methods [8]. For example, in the United States, overcharging (that is, health care providers charge insurance institutions fees exceeding the allowable limit) is very common [9]. In the UAE, fraud in the field of medical insurance has adversely affected access to high-quality medical services and the broader economic and social structure [10]. Therefore, it is particularly important to formulate a reasonable premium pricing strategy.

Many factors will affect medical expenses and insurance premiums. This complexity makes the pricing process of the customer's insurance rate extremely complicated and cumbersome, because each factor needs to be weighed and verified to avoid deviation. This study aims to use insurance data as a reference: after data preprocessing and optimization, four different machine learning models —XGBoost, Catboost, Gradient Boosting and LightGBM —are used to predict the potential insurance premium of future customers. The accuracy and performance of these models are systematically compared in order to provide data-based support for insurance institutions to optimize their premium pricing strategies. This study determines the most robust model which is most suitable for practical application. Finally, this research not only helps insurance companies reduce the cost of time and human resources, but also provides them with scalable tools to adjust

DOI: 10.54254/2754-1169/2026.BJ31262

pricing strategies according to changing market conditions, so as to improve operational efficiency and customer trust.

2. Methods

2.1. Data description

The data of 1338 customers used in this study were obtained from kaggle, including 8 independent predictive variables, such as age, gender and BMI (body mass index). This data set contains detailed personal information, medical expenses and insurance premiums of each customer. Table 1 shows the data field types in the dataset and their corresponding explanations.

Table 1. Variable description

Column Name	Type	Explanation of Customer Information	Distribution Range
age	Int	Age	18 - 64
gender	String	Gender	male or female
bmi	Float	BMI index	16 - 53.1
children	Int	Number of children owned	0 – 5
discount_eligibility	Int	Eligibility for discount	1 or 0
region	String	Region	northeast - southwest
expenses	Float	Medical expenses	1127.81 - 63770.43
premium	Float	premium	11.2187-1983.1064

2.2. Sample selection

During the Python-based data preprocessing phase, this study first performed preprocessing operations (e.g., removing duplicate rows) on the dataset. Subsequently, descriptive statistics were computed for each indicator, and their distribution density plots were generated. Figures 1 to 4 illustrate the distribution densities of selected key indicators.

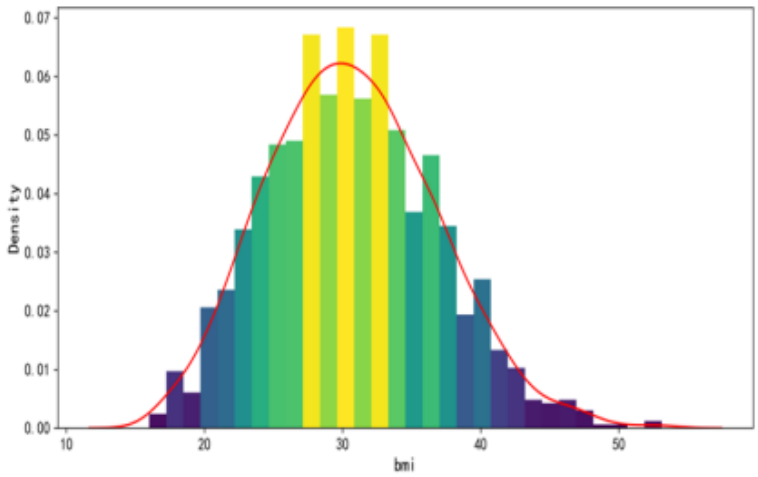

Figure 1. Allocation of bmi (picture credit: original)

DOI: 10.54254/2754-1169/2026.BJ31262

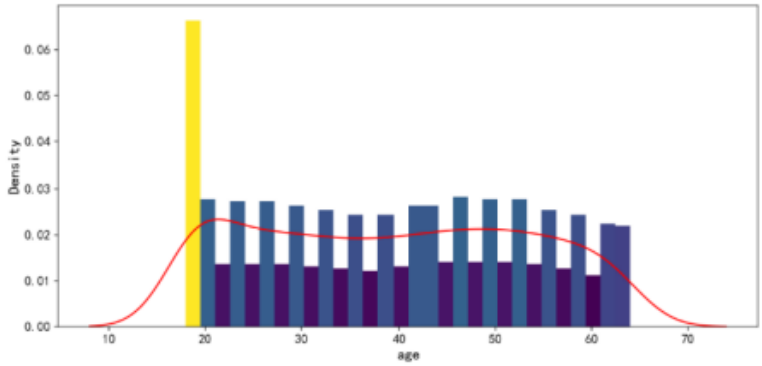

Figure 2. Allocation of age (picture credit: original)

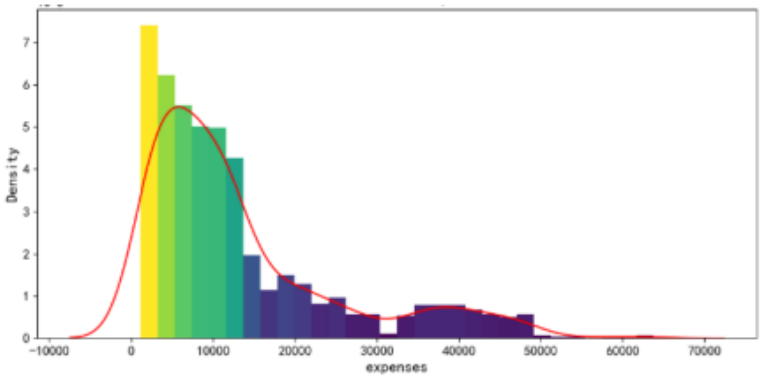

Figure 3. Allocation of expenses (picture credit: original)

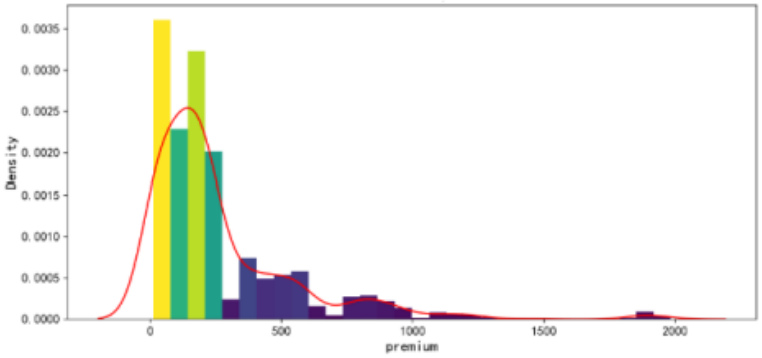

Figure 4. Allocation of premium (picture credit: original)

The gender variable shows an approximately balanced distribution. Specifically, there are about 660 female customers and about 670 male customers, with a relatively small difference in the proportion between the two genders. In terms of discount eligibility, the sample presents a significant imbalance: the vast majority of customers are not eligible for discounts, with a number of about 1,050; only about 280 customers are eligible for discounts, accounting for a relatively low proportion. These sample customers are distributed in four regions: Southwest, Southeast, northwest and northeast. Among them, the southeast region has the largest number of customers, about 360 people; The southwest region ranks second, with about 325 customers; The number of customers in the northwest and northeast regions is relatively close, both about 325.

DOI: 10.54254/2754-1169/2026.BJ31262

According to these results, it is obvious that there are no missing values in this dataset. The age range of the sample is between 18 and 64 years old, and the average BMI (body mass index) is about 30, indicating that most customers are overweight. Gender may have a relatively small impact on medical expenses, while discount policy may affect premium pricing. In addition, regional factors may affect both medical expenses and insurance premiums.

2.3. Method introduction

This study uses four models for simulation and prediction, namely catboost, xgboost, gradient lifting and lightgbm. After configuring parameters for each model, the four models are trained using data sets, and then their accuracy and performance are compared. Select the model with the best performance to build the premium prediction model.

Xgboost is a scalable end-to-end tree lifting system, which has achieved the most advanced results in many machine learning challenges [11]. For a dataset with n samples and M dimensions, the mathematical expression of xgboost model is as follows.

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (i = 1,2,\ldots n) \tag{1}$$

Catboost introduces two important algorithm improvements, namely, the implementation of the ordered lifting algorithm (a method based on permutation to replace the classical algorithm) and an innovative method for processing classification features [12]. The mathematical formula of catboost model is as follows:

$$\widehat{x_k^i} = \frac{\sum_{j=1}^{i-1}\left[x_{\sigma_j,k}=x_{\sigma_i,k}\right]Y_{\sigma_j}+a\cdot p}{\sum_{j=1}^{i-1}\left[x_{\sigma_j,k}=x_{\sigma_i,k}\right]+a} \tag{2}$$

In Gradient Boosting, the gradient boosting of regression trees can provide competitive, highly robust, and interpretable procedures for both regression and classification tasks, especially suitable for mining unclean data [13]. The idea of the Gradient Boosting algorithm is derived from the gradient descent method. Its basic principle is to train the newly added weak classifier based on the negative gradient information of the loss function of the current model, and then integrate the trained weak classifier into the existing model in an accumulative manner.

As an implementation of a new Gradient Boosting Decision Tree (GBDT) algorithm, LightGBM combines Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). With almost the same accuracy, lightgbm can speed up the training process of traditional gbdt by more than 20 times [14].

3. Results and discussion

This chapter systematically presents and analyzes the key results of the research, focusing on three core dimensions: determining the factors affecting medical expenses and insurance premiums, evaluating the training performance of different machine learning models, and verifying the actual prediction effect of the optimal model. Firstly, through the data visualization tool, the correlation between various customer characteristic indicators and fees is studied and discussed, and the factors that play a leading role in premium pricing are clarified. Second, by conducting comparative experiments on four mainstream machine learning models (XGBoost, Catboost, Gradient Boosting, and LightGBM), the study screens out the model with the highest accuracy and reliability, and

DOI: 10.54254/2754-1169/2026.BJ31262

further validates its performance through key evaluation metrics. Finally, four models are applied to actual customer data for premium prediction, and the deviation between predicted and actual values is discussed, while targeted improvement directions are proposed.

3.1. Factors affecting medical expenses and insurance premiums

To further identify the indicators that may affect expenses and insurance premiums, this study generates a correlation heatmap, which is presented in Figure 5.

Figure 5. Correlation heatmap of various indicators (picture credit: original)

The results show that age is highly positively correlated with expenses. The possible reasons are as follows: with the increase of age, the human body's health level declines, and immunity and resistance weaken. The BMI index has a certain relationship with expenses, which may be due to the increased risk of illness when the BMI index is in an abnormal state. Medical expenses are almost completely correlated with insurance premiums, probably because insurance premiums are determined by medical expenditures. Gender, the number of children one has, and region have a weak relationship with expenses, and these indicators of customers have little impact on the final price of insurance premiums.

3.2. The result of model training

Simulated tests were conducted on four models, and the histogram of the accuracy rate for each simulation is shown in Figure 6.

DOI: 10.54254/2754-1169/2026.BJ31262

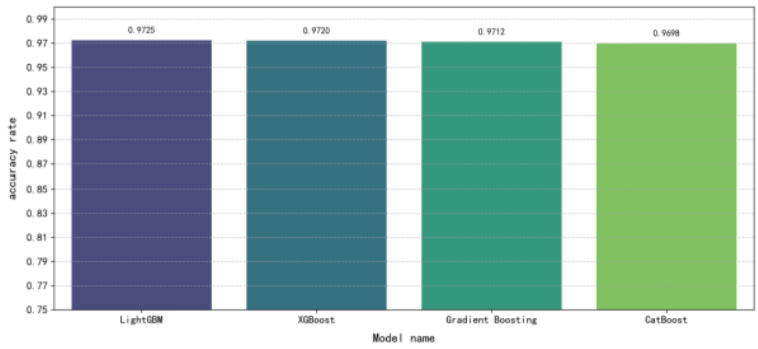

Figure 6. Comparison of accuracy rates among the three models (picture credit: original)

From the above results, it can be seen that the LightGBM model achieves the highest accuracy. Further training was conducted on the LightGBM model, and the results are shown in Table 2.

Table 2. Training results of the LightGBM model

	Precision	Recall	F1-score	support
0	0.96	0.99	0.98	195
1	0.98	0.89	0.94	73
Accuracy			0.97	268
Macro avg	0.97	0.94	0.96	268
Weighted avg	0.97	0.97	0.97	268

In summary, the optimal model is LightGBM. It demonstrates high accuracy in identifying regular premium customers, achieving a recall rate of 99%, which helps avoid customer churn caused by misjudgment. For high-premium customers, its recognition precision is also very high, with a precision rate of 98%. However, the model also has some shortcomings: for instance, its miss detection rate may reach 11%, leading to the omission of some potential customers.

3.3. The prediction results of the premium prediction model

Predictive models were constructed using four different algorithms respectively. Inputs were also formulated, and the customer information corresponding to the premiums of $168.85, $44.4946, and $77.3372 in the dataset was used as the input. The resulting premium prediction outcomes are presented in Table 3.

Table 3. Results of prediction

Actual Premium	Gradient Boosting	XGBoost	CatBoost	LightGBM
168.85	189.56	181.34	178.17	190.86
44.4946	62.08	60.45	66.51	60.78
77.3372	41.06	41.23	44.33	41.40

From the prediction results in Table III, all four models show a consistent trend of overprediction for low premiums and underprediction for medium premiums, which may be related to the sample distribution of the dataset and other factors.

DOI: 10.54254/2754-1169/2026.BJ31262

3.4. Discussion on the results

The model constructed in this study can initially predict a customer's premium based on their personal information. However, there is a certain deviation between the predicted results and the actual premiums, which may be attributed to multiple factors. First, the insufficient sample size fails to capture niche customer patterns, thereby reducing its performance. Second, prioritizing simple interpretation (single GBDT) sacrifices 10–15% accuracy compared to ensemble models like XGBoost. Third, Predictions may violate company thresholds, making results inapplicable.

Future improvements to the model can be advanced in phases: In the short term, priority will be given to expanding the sample size to over 5,000 through multi-source data integration, collecting data across 4 quarters to capture seasonal characteristics, and adjusting the parameters of the GBDT model using Bayesian optimization—with 5-fold cross-validation conducted to reduce prediction deviations. In the long term, business rules will be integrated into the prediction results, and a visual dashboard usable by agents will be built. Ultimately, these efforts will transform the model from a technical prototype into a practical tool.

4. Conclusion

This study is based on a public insurance customer dataset from Kaggle, from which 1,338 valid customer samples were selected for in-depth analysis. The dataset includes 8 core independent predictor variables, specifically including customers' age, gender, BMI (Body Mass Index), number of children, region of residence, smoking status, insurance coverage period, and past medical records. These variables basically cover the key personal attributes and behavioral characteristics commonly considered by the insurance industry in premium pricing, providing comprehensive foundational data support for subsequent research.

In the data analysis phase, the research team comprehensively applied various data processing and visualization techniques, and constructed a multi-dimensional analytical framework to accurately identify the core indicators influencing premium pricing. First, descriptive statistical analysis was used to sort out the distribution characteristics of each variable—for example, calculating the proportion of customers in different age groups and the average BMI of customers in each region—to gain a preliminary understanding of the overall data situation. Second, correlation analysis models were used to quantify the degree of association between variables. Combined with visualization tools such as heat map and scatter diagram, the relationship between various factors and premium cost can be visualized.

The final conclusion is as follows: there is a significant positive correlation between age and cost; BMI is related to cost; The cost is almost completely positively correlated with the premium; However, the relationship between gender, number of children, location and cost is weak. In this study, four models—xgboost, catboost, gradient lifting and lightgbm—are used to predict the premium of customers. The prediction results of the model show that lightgbm performs better than other models in simulating the original data.

However, due to insufficient sample size and insufficient optimization of model parameters, there is a certain degree of deviation between the final prediction results and the expected results. For future research, efforts should be made to prepare data sets with larger sample size in the short term, and optimize model parameters to improve the accuracy and applicability of the prediction model. In the long term, business rules should be integrated into the prediction results, and a visual dashboard usable by agents should be built. Ultimately, this will transform the model from a technical prototype into a practical tool. This will not only provide data support for insurance institutions to

DOI: 10.54254/2754-1169/2026.BJ31262

optimize their premium pricing strategies, but also reduce the time costs and labor costs incurred by insurance institutions in calculating customers' premiums.

References

[1] Huang, P., Sun, Z., Li, L., Li, J. (2024) Has the Integrated Medical Insurance System promoted return-to-hometown entrepreneurship among migrant workers? Evidence from China. Front Public Health. 12: 1323359. doi: 10.3389/fpubh.2024.1323359. PMID: 38371234; PMCID: PMC10870983.

[2] Zhou, X., and Yang, X. (2024) Medical insurance, vulnerability to poverty, and wealth inequality. Front. Public Health 12: 1286549. doi: 10.3389/fpubh.2024.1286549

[3] Sheng, W., Fuchong, L. (2024) Medical insurance, livelihood capital and public health in China. Cost Eff Resour Alloc 22, 48. https: //doi.org/10.1186/s12962-024-00554-z

[4] Guariglia, A., Rossi, M. (2004) Private medical insurance and saving: evidence from the British Household Panel Survey, 11.002; Journal of Health Economics, Volume 23, Issue 4, Pages 761-783, ISSN 0167-6296. https: //doi.org/10.1016/j.jhealeco

[5] Bai, B., Zhang, Y., Liu, Y. (2021) Influences of Public Medical Insurance System on Labor Health Status and Supply. Iran J Public Health. 50(8): 1658-1667.

[6] Fu, X. (2021) Financial protection effects of private health insurance: experimental evidence from Chinese households with resident basic medical insurance. Int J Equity Health. 20(1): 122. doi: 10.1186/s12939-021-01468-5. PMID: 34001149; PMCID: PMC8130397.

[7] Hullegie, P. and Klein, T. J. (2010) The effect of private health insurance on medical care utilization and self-assessed health in Germany. Health Econ., 19: 1048-1062. https: //doi.org/10.1002/hec. 1642.

[8] Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., Alshemaimri, B. (2024) Healthcare insurance fraud detection using data mining. BMC Med Inform Decis Mak. 24(1): 112. doi: 10.1186/s12911-024-02512-4. PMID: 38671513; PMCID: PMC11046758.

[9] Sagnika, S., Amit, V. D. (2022) Toward understanding variations in price and billing in US healthcare services: A predictive analytics approach, Expert Systems with Applications, Volume 209, 118241, ISSN 0957-4174, https: //doi.org/10.1016/j.eswa.2022.118241.

[10] Harichandran, H. P. (2023) An Insight on Medical Insurance Malpractices Prevailing in the Healthcare Industry and its Impact on Socio Economic Background – Special Reference to UAE Private Healthcare Industry. International Journal of Professional Business Review, 8(8), e03634. https: //doi.org/10.26668/businessreview/2023.v8i8.3634.

[11] Chen, T., Guestrin, C. (2016) XGBoost: A scalable tree boosting system. Ithaca. doi: https: //doi-org.ezproxy.utm.my/10.1145/29396 72.2939785

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018) CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems 31.

[13] Jerome, H. F. (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics, Ann. Statist. 29(5), 1189-1232.

[14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.