Machine Learning Approaches for Medical Insurance Cost Prediction

Chenglin Wang

{ chenglin.wang@student.manchester.ac.uk}

School of Natural Science, The University of Manchester, Manchester, United Kingdom

Abstract. The question of medical insurance pricing is a significant concern at the contemporary healthcare level, because effective anticipation of medical costs allows healthcare workers to effectively cope with the risk and to develop a fair insurance policy. This paper uses and compares two regression techniques, Linear Regression and Random Forest, with the use of the Kaggle Medical Cost Personal Dataset. The obtained results indicate that the random forest model is infinitely better than the Linear Regression, making errors in predictions lower and attaining a high value of R2 (0.865). The analysis of the feature importance also indicates that the most significant determinants of medical insurance bills are smoking status, BMI, and then age. This indicates that the benefit of machine learning techniques in improving predictive accuracy and determining the most crucial cost drivers, offering quasi-realistic hints in pricing of insurance and in risk control.

Keywords: Medical Insurance Pricing; Machine Learning; Random Forest; Linear Regression; Feature Importance

1 Introduction

Earlier on, numerous studies used classical statistical models, including linear regression, to forecast medical costs. These are easy-to-comprehend models. They, however, assume a linear relationship between variables [1-3]. The real information tends to be more complicated. These are possible factors that are related. As an example, the combined effect of age and smoking could be stronger. These trends might not be well recognized in the traditional models [4-6]. Although recent studies have presented more sophisticated machine learning techniques that are meant to improve the predictive power, there are studies that concentrate mainly on the improvement of accuracy and there is little concentration on feature performance comparison or clear elucidation of the importance of features. This cannot be a real solution to the problem [7, 8].

Previous studies on medical cost prediction have mainly relied on traditional statistical models, such as linear regression and generalized linear models. These approaches are widely used because they are easy to interpret and implement. However, they typically assume linear

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365157

relationships between variables and may struggle to capture complex interactions among factors such as age, body mass index (BMI), and smoking behavior. Although recent research has introduced more advanced machine learning methods to enhance predictive performance, some studies focus primarily on accuracy improvement while giving limited attention to comparing models systematically or interpreting feature importance in a clear and practical way. This creates a gap between predictive performance and meaningful policy insights [9, 10].

To address these limitations, this study applies and compares both a baseline linear regression model and a more flexible ensemble method, the Random Forest, on the Kaggle Medical Cost Personal Dataset. The models are tested through the MAE, RMSE, and R2 metrics after the preprocessing of the data and the process of exploratory data analysis. Besides a comparison of the predictive performance, the research examines feature importance. The goal is to identify the variables that have the greatest impact on insurance costs. By performing model comparison and interpret analysis, the advancement of the study is not only to enhance the accuracy of prediction, but also gives a better insight into the key predictors of medical insurance cost, specifically the impact of smoking status, body mass index (BMI) and age.

2 Datast

2.1 Dataset information and methods

The dataset used in this project is the Medical Cost Personal Dataset, publicly available on the Kaggle platform. This dataset is widely applied in data analysis and modeling research related to medical insurance cost prediction, and it has good representativeness and practicality. The dataset contains a total of 1,338 valid records, each corresponding to the personal information and insurance cost data of an insured individual. The overall structure is clear and there are no missing values, providing a high-quality data foundation for subsequent modeling.

The dataset includes 7 feature variables and 1 target variable. The specific Information Is as follows: The target variable is charges, which represents the medical insurance cost of the insured individual (unit: USD), and it is used as the prediction target of the model; The feature variables cover the core personal attributes and life-related factors of the insured individuals, including age (a continuous variable), sex (a categorical variable with values of male/female), BMI (body mass index, a continuous variable reflecting the health indicator related to weight and height), children (the number of dependent children, a discrete variable), smoker (smoking status, a categorical variable with values of yes/no), and region (residential area, a categorical variable).

2.2 Dataset pre-processing

For preprocessing, the dataset was cleaned without missing values. One-hot encoding is used on the three categorical variables of sex, smoker, and region in order to transform them into binary numerical features. This approach helps to eliminate the disorder of the categorical variables and prevent it from interfering with the model training. It will make sure that the model is able to process the information of their variables efficiently. Separate the dataset into 80/20 ratio between the training set and test set.

There are some features of data distribution, the interrelations of variables, which this research conducted a systematic exploratory data analysis to clarify. As shown in Figure 1, the distribution of insurance charges exhibits clear right-skewness. As Figure 1 shows, most of the persons make relatively medium medical costs, and only a few percent of people are subjected to extremely high costs. This model is in line with the organization of medical spending in the real world, whereby a relatively small group of people occupies a relatively huge segment of the total expenditure on health care.

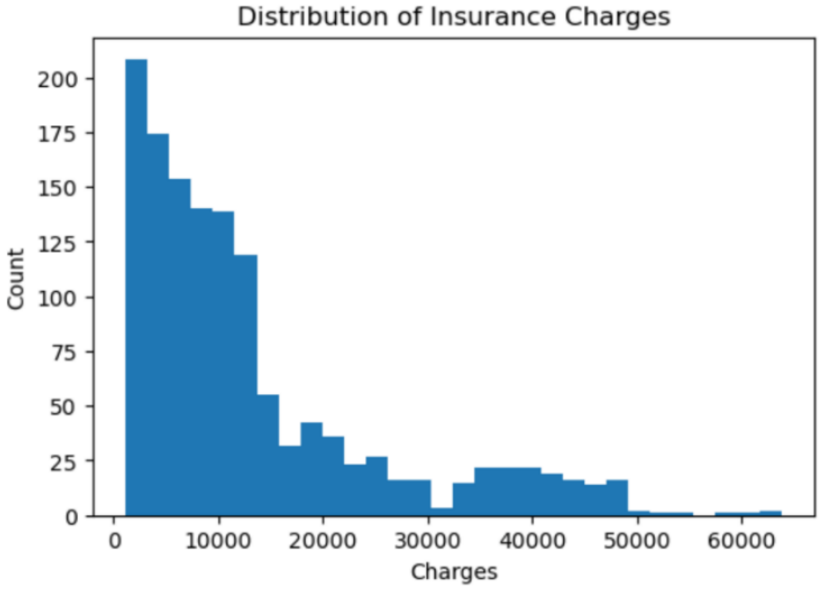

Fig. 1. Distribution of Insurance Charges (Picture credit: Original)

Smoking status is most related to medical charges and its correlation coefficient stands at 0.79, as shown in Figure 2. This implies that the issue of whether an individual is a smoker is highly related to increased insurance prices. The correlation between age and charges (r=0.30) is moderate, which implies that the price of medical care tends to rise with age. There is also a positive relationship between BMI and charges (r=0.20), but the effects are less significant than those of smoking and age. Conversely, the number of children is associated with medical costs only with a very weak level (0.07), which means this factor does not considerably influence the final price. Generally speaking, of all the variables, smoking seems to have the most significant impact in relationship with increased medical charges.

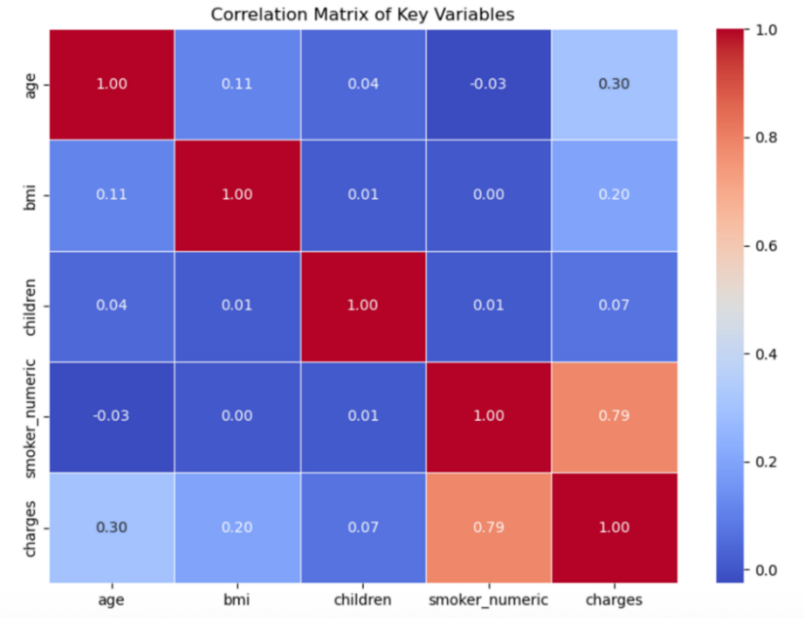

Fig. 2. Correlation Matrix of Key Variables (Picture credit: Original)

2.3 Models

In this study, two regression models were selected for comparison: Linear Regression and Random Forest. These two models were chosen because they represent different modeling ideas and levels of complexity.

Linear Regression is a traditional statistical method that assumes a linear relationship between the independent variables and the dependent variable. It approximates the extent of variation of the target variable when a single predictor varies, holding other variables constant. This is one of the advantages of this model since the coefficients are easy to interpret, and thus, the aspects of age, BMI and smoking status can be understood with respect to their influence on medical costs in this research. Moreover, it may also be used as a reference model to test whether the key associations of the dataset are more or less linear.

In contrast, the Random Forest is a collective learning method based on multiple decision trees. It does not build a single model but uses different subsets of data to construct many trees, and then combines their predictions. In the process, overfitting is minimized and prediction stability is often enhanced. More to the point, the relationship and interaction between variables that are non-linear and can exist in the data of medical cost, can also be captured by the Random Forest.

The effects of the model complexity on prediction results can be easily illustrated by means of comparison between the two.

3 Experiment

3.1 Evaluation index

To evaluate the predictive performance of the model, this study employs three commonly used regression model evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). Among them, MAE represents the average absolute deviation between the predicted values and the actual values, which is simple to calculate and insensitive to outliers; RMSE penalizes larger deviations by squaring the errors, better reflecting the model's prediction effect on extreme values; R^2 is used to measure the model's ability to explain the variation in data, with a range of [0,1]. The closer It Is to 1, the better the model's fitting effect.

3.2 Experiment results and analysis

The performance evaluation results of Linear Regression and Random Forest are shown in Table 1.

Table 1. Performance Comparison of Linear Regression and Random Forest

Model	MAE	RMSE	R2
Linear Regression	4181.194474	5796.284659	0.783593
Random Forest	2550.241127	4576.287233	0.865104

As shown in Table 1, all performance indicators of the random forest regression model are superior to those of the linear regression model: MAE has decreased from 4181.19 to 2550.24, a reduction of approximately 39%; RMSE has decreased from 5796.28 to 4576.29, a reduction of approximately 21%; R^2has increased from 0.784 to 0.865.

3.3 Feature important analysis

Based on the feature importance assessment results shown in Table 2, the key factors influencing insurance costs and their relative importance can be clearly identified:

Table 2. Feature Importance Scores from Random Forest Model

Feature	Value
Smoker_yes	0.6086
BMI	0.2166
Age	0.1342
Children	0.0194
Sex_male	0.0064
Region_northwest	0.0056
Region_southeast	0.0053
Region_southwest	0.0039

As illustrated in Table 2, the feature importance analysis shows that:

First, the smoking status is the most crucial factor influencing insurance costs, with an importance score of 0.6086, accounting for over 60% of the total importance. This is consistent with the conclusion from the exploratory analysis that "smokers have significantly higher costs", confirming the strong driving effect of smoking behavior on medical costs; Second, BMI (importance 0.2166) and age (importance 0.1342) are the key factors after

smoking status, respectively reflecting the impact of health status and physiological age on medical needs, which is in line with the actual impact logic of medical costs; Last, the importance scores of the number of children, gender, and place of residence are all below 0.02, having a weak impact on insurance costs. This might be because these factors have a relatively weak direct correlation with the consumption of medical resources.

Overall, lifestyle (smoking), health indicators (BMI), and demographic characteristics (age) are the core driving factors determining medical insurance costs, while the influence of other variables can be disregarded.

4 Conclusion

This study aimed to predict medical insurance charges and identify the key factors that affect them. Two models were used: linear regression and random forest. The results show that the random forest model performs better. It has lower MAE and RMSE. It also has a higher R^2 value. This means it can predict insurance charges more accurately.

It is also seen in the analysis that smoking is the most significant factor. The rates of smokers are considerably higher in comparison to those of non-smokers. There are other significant ones of BMI and age, although they are not as substantial as smoking. The other variables that are influential are region and gender. These findings correlate with the exploratory data analysis, which serves to further comprehend the cause of high medical prices.

However, this project still has some limitations. First, the shape of the dataset is relatively small. It only includes 1338 records. Second, the variables are limited. Some important factors, such as medical history or income level, are not included. Third, only two models were compared. There may be other models that perform even better. In the future, more data can be collected to improve the stability of the model, and a wider range of machine learning methods can be tested. In addition, a more in-depth analysis of feature interactions will be explored.

Overall, this study shows that machine learning is very useful for predicting insurance charges. It is also used in determining important cost drivers. The results might be useful in improving risk management and pricing in medical insurance.

References

[1] Keehan, S.P., Cuckler, G.A., Sisko, A.M., et al.: National health expenditure projections, 2014–24: spending growth faster than recent trends. Health Affairs. pp. 1407-1417 (2015)

[2] Beam, A.L., Kohane, I.S.: Big data and machine learning in health care. Jama. pp. 1317-1318 (2018)

[3] Badawy, M., Ramadan, N., Hefny, H.A.: Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology. pp. 40 (2023)

[4] Orji, U., Ukwandu, E.: Machine learning for an explainable cost prediction of medical insurance. Machine learning with applications. pp. 100516 (2024)

[5] Blier-Wong, C., Cossette, H., Lamontagne, L., et al.: Machine learning in P&C insurance: A review for pricing and reserving. Risks. pp. 4 (2020)
[6] Grize, Y.L., Fischer, W., Lützelschwab, C.: Machine learning applications in nonlife insurance. Applied Stochastic Models in Business and Industry. pp. 523-537 (2020)
[7] Zou, S., Chu, C., Shen, N., et al.: Healthcare cost prediction based on hybrid machine learning algorithms. Mathematics. pp. 4778 (2023)
[8] McDonnell, K., Murphy, F., Sheehan, B., et al.: Deep learning in insurance: Accuracy and model interpretability using TabNet. Expert Systems with Applications. pp. 119543 (2023)
[9] Orji, U., Ukwandu, E.: Machine learning for an explainable cost prediction of medical insurance. Machine learning with applications. pp. 100516 (2024)
[10] Laub, P.J., Pho, T., Wong, B.: An Interpretable Deep Learning Model for General Insurance Pricing. arXiv preprint arXiv:2509.08467. (2025).