Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/45/20230270

Machine Learning in Medical Insurance Prediction

Qiyun Lei[1,a,*]

[1]Pinghe school, Shanghai Pudong 201206, China
a. scarlettlei2008@hotmail.com
**corresponding author*

Abstract: Nowadays, the trend of an ageing society is more and more obvious. Accompanied with the huge population of the elderly, the medical insurance industry has more prospects and potential. As a result, more service and business operations of insurance companies are in need. With the analysis from past data, computer algorithms help a lot in predicting the new output values, aiding data-driven business decisions, ranking of influential factors and digital computerization. Through machine learning, the insurance companies are able to make a decision flatly in premiums without having unnecessary medical expenditure. The provided models include linear regression, polynomial regression, and random forest. Through the comparation of these three models, with the output data of MAE and other indicators, we can see that polynomial regression is the best model. Within the efficient operation of this method, it can soon be prevalent among the medical industry. Avoiding problems of high cost of labor and inevitable manmade mistakes, polynomial regression aids the technical advance and statistical progress to prosper.

Keywords: machine learning, medical insurance cost, linear regression

1. Introduction

It is seen that the number of adults ages 85 and older, which especially require outer assistance in personal caring, will nearly quadruple between 2000 and 2040 [1]. This data provided an eidetic description of ageing problems. In the high probability of being affected by illness, elderly people will choose medical insurance to get the compensation they want if illness comes. Moreover, human lives and personal health may be endangered by unexpected disasters, worldwide pandemic and shortage of medicine. Unforeseen and inevitable circumstances can be met at any time in life. The US government spent 9.8% of global GDP on health care [1]. Efforts and techniques are made to solve medical problems. As long as higher life expectancy is incorporated in people's mind, health insurance has become a common commodity to people.

Often conventional datasets are built so as to contain data by companies, containing beneficials' basic information. However, only 10-15percent of the data is processed for providing ideas. Hence, the following transformation of data should be concerned for the importance to companies' growth. Using machine learning, companies can solve the problem. With structured datasets, semi-structured and unstructured datasets, the step of preprocessing aids rearranging into structured sets thoroughly. Also, with null data and some mistakes in datasets, machine learning filters out useless data effectively. It predicts the insurance cost in higher accuracy with different models required.

Moreover, for customers, the transparency of charges can be doubted because there is overpricing and prioritizing. These phenomena implicate the dark side of medical systems, such as illegal prioritizing to upper classes or bribery. If computers aren't included in the calculating process, there is much space for some people to manipulate such as intentionally elevate the price during calculating. Scandals in insurance companies aren't impossible. Thus, with the working of machine learning, its self-studying mode can decrease the likelihood of relevant financial crimes. Also, machine learning prevents the calculating or analyzing mistakes done by people. Tiredness and negative emotions are all factors affecting the output result through human. Highly effective machines may decrease this possibility.

In this passage, this essay finds the datasets from Kaggle. It aims to use the attributes to do the medical cost prediction. In fact, the prediction will be used in different methods – simple linear regression, polynomial regression and random forest. Results of different models will be compared at last, and the best model will be shown.

With the strong growth of the insurance industry and personal expenditure on healthcare, professionals' research deepens as well. In [2], Akhil and Kirthy find the correlation between attributes like smoking, age, etc. to medical expense. Then, they apply the most influential factor to the prediction model. They conclude that smoking seems to have more influence. Similar research is also shown on [3], in which Sturm discusses how obesity, smoking and excessive alcohol intake affect medical expenditures. With the comparation, he suggests that obesity affects more. Obesity is a considerable factor in insurance costs. Research like [4] and [5] use statistical calculations and meta-analysis respectively. Both conclusions show how economic burdens rise due to prevalent obesity. Paper as [6] focuses on the inequality in medical system of China. Poverty and high-income possess utterly opposite burdens in medical insurance. In [7], operating with a simulated model, health-care costs were estimated for a group of obese people aged 20 y at least. The conclusion is that despite of effective obesity prevention decreases relevant cost, the cost will still be offset by other disease. In [8], estimates of the medical costs of smoking in USA are compared, in which the result is equal.

2. Methodology

2.1. Simple Linear Regression

Among the model selection, the first one is linear regression. The target variable(Y) is associated to independent variables(X). In simple linear regression, there is merely a single independent variable. It is believed that because both X and Y have numerical values, there is calculated correlation between the two variables. The equation is given by:

$$Y = a + bx \tag{1}$$

where "a" and "b" are the coefficients of the models. "a" is the value of intercept when X is equal to 0. In addition, "b" is the slope of X. "b" suggests how Y will affect during the changing of X. However, the models cannot be that accurate. The deviations of predicted values and actual values may exist. Mostly there will be differences between the two values. Thus, introducing the error term can make the output Y more critical. Error term is independent that it depends on the residual of each value of X. The equation is given as:

$$Y = a + bX + \epsilon \tag{2}$$

2.2. Polynomial Regression

Like linear regression, polynomial regression shares the same idea. However, in linear regression, there are only a single independent variable and a single dependent variable. But polynomial regression contains numerous predictor variables, and the calculation of Y is based on the predictor variables. Moreover, predictor variables are independent during the assumption. Assume that target variable hinges on "n" independent variables. The model is turned into:

$$Y = a + b1X1 + b2X2 + b3X3 + \cdots + bnXn + \epsilon \tag{3}$$

Thus, in the application of medical insurance cost prediction, one option is to run several independent simple linear regressions, in which they use different features from data as a predictor. Another option is to extend the simple linear regression to polynomial regression so that the models can directly accommodate multiple predictors.

Normally polynomial regression works better than simple linear regression. It is unclear how to make a single prediction of charges given the multiple features of beneficials in medical insurance, as each of the charges is linked to a separate regression equation. Also, each of the three regression equations ignores the other features in forming estimates for the regression coefficients. At the meantime, polynomial regression tends to form a more compact equation for it contains deep correlations between multiple factors simultaneously.

2.3. Random Forest

Random forest Algorithm originates from the basis of bagging method which generates random subsets of data's features. Including node size, the number of trees, and the number of features sampled, it can be used in both classifications and regression problems. It consists of loads of decision trees, in which each tree in the ensemble contains a data sample drawn from a training set with replacement. For regression problems, individual decision trees will be averaged.

2.4. Indicators

The results of three models are identified using MSE, MAE,R^2 and accuracy score. Mean squared error (MSE) measures the error of model. It observes the average squared difference between predicted value and actual value. The formula is given as:

$$MSE = \frac{\Sigma(yi - \hat{y}\mathrm{i})^2}{n} \tag{4}$$

where yi means the actual value and ŷi is the corresponding predicted value. "n" is the number of samples. As a result, the model is more accurate when MSE is lower.
While MAE means the mean absolute error and the formula is

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|xi - x| \tag{5}$$

Moreover, R^2score is also a typical indicator to examine linear regression model. But of course, it can be applied in other models as well. The formula is:

$$= 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}\mathrm{i})^2}{\sum_{i=1}^{n}(yi - \bar{y}\mathrm{i})^2} \tag{6}$$

Linear regression model aims to find the relationship between factors. But unlike functions, there isn’t accurate numerical result. Thus, examining the predicted value and actual value is very crucial. Sum of each square of the difference of predicted value and counterpart true value shows the deviation of data. However, merely this cannot provide a critical judgement. For example, datasets of students’ weight are much smaller than the dataset of astronomical measurement, whereas people can’t draw a conclusion easily based on two data. It is important to compare it to the average score. Thus, it must be divided as $\frac{\sum_{i=1}^{n}(yi-\hat{y}\mathrm{i})^2}{\sum_{i=1}^{n}(yi-\bar{y}\mathrm{i})^2}$. However, the ultimate formula 1-$\frac{\sum_{i=1}^{n}(yi-\hat{y}\mathrm{i})^2}{\sum_{i=1}^{n}(yi-\bar{y}\mathrm{i})^2}$ is a manner to standardize, preventing too large number.

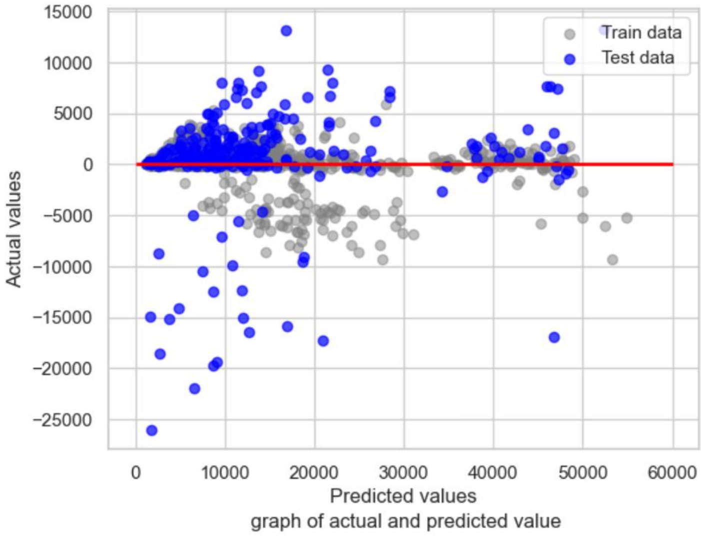

Figure 1: This graph depicts the result of Random Forest Regressor. Deviation of train and test data can be shown here.

In figure 1, the predicted value and actual value of both training and testing datasets can be shown graphically. We can see that dots are more accumulated near the red line.

3. Data

From the Kaggle site [9], I obtain datasets about medical insurance cost in US. The dataset contains 7 attributes and 1338 rows. Three of the attributes are categorical and the rest are numerical. The data is split into training data and testing data. The ratio between training data and testing data is 80:20. Training data aids to build the model, train the machine and increase its accuracy. The more training data fit into the model, the more mature it is. Testing data is like doing an exam to the model. It evaluates the model.

Among the introduction of 7 features, there is table 1 below to show.

Table 1: Lists of 7 attributes.

Name	Description
age	Age of primary beneficiary
sex	Gender of beneficiary
bmi	Body mass index, giving an understanding of body, weights that are relatively high or low to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
Children	Number of children covered by beneficiary
smoking	Smoking or not
Region	Northwest, northeast, southwest, southeast in USA
Charges	Charges of medical insurance

We can see the attributes above. In the following figure 2, the distribution of charges can be seen. The graph of charges follows Gaussian distribution.

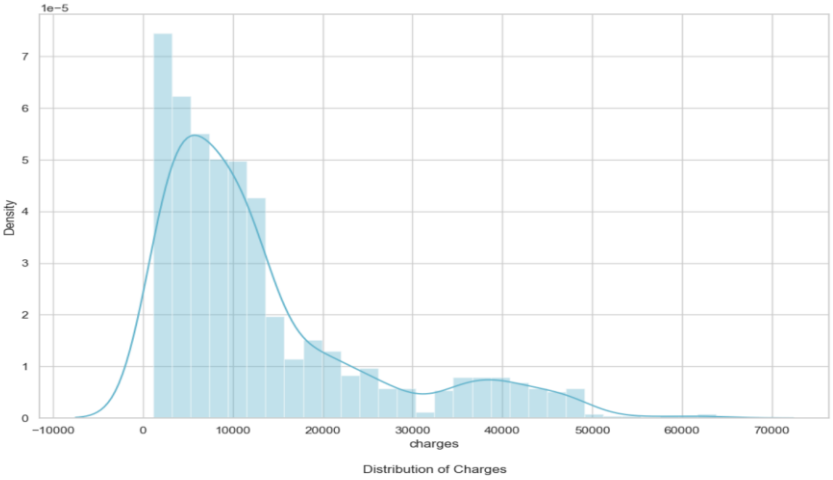

Figure 2: Distribution of charges.

From figure 3, the data should be shuffled into structured, complete, and properly applied one. Firstly, the datasets should be checked with missing values. Luckily in this dataset there is no missing value. However, normally, the column with missing value will use the mean of two data next to it. Also, because the regression models only accept numerical data, I must use label coding to transfer the categorical columns. As a result, the data is ready to be analyzed.

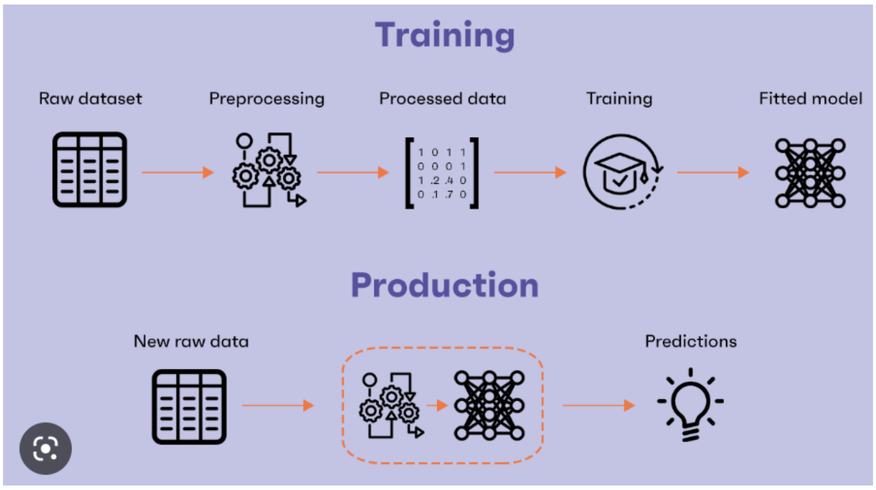

Figure 3: Data preprocessing procedure (Source: https://www.clearbox.ai/blog/2021-02-10-automating-data-preparation-and-preprocessing-in-production-ready-ml-models/).

Figure 3 shows the procedure of training and producing data [10]. It discloses the basic steps of model training with data. It shows the importance of preprocessing data, for data must be trained into correct mode to help fulfill model training.

My basic steps are importing the data, clean the data, split the data into training and testing set, creating a model, training the model, making predictions and evaluating with improving. In the simple linear regression, during the training set, utilizing the drop function on “charges” and fitting the model deduct intercepts and coefficients of model. By the import of metrics from sklearn, the accuracy score shows in here.

As using polynomial regression, I drop “charges” “sex” “region” to form the multiple predictors. Then, likewise, I fit the model and predict with both training set and testing set. Thus, the difference between actual and predicted value can be seen, inferring the accuracy score.

Among the application of random forest, I initialize the model, fit the model and do the same prediction and evaluation. Plotting a graph about these two types of data helps to examine the model more directly.

4. Results

From the results (Table 2), we can see that polynomial regression perform better in accuracy score, while Random Forest seems to be better in R^2 score. However, through the overall evaluation in MSE and other indicators, polynomial regression is the best model. In fact, accuracy scores play an important role in ultimate judgement. Simple linear regression and polynomial regression all have same value in MSE, MAE and R^2 ,whereas accuracy score is utterly different with the difference of 8%.

Table 2: The ultimate results of different models.

Model	MSE	Accuracy score	MAE	R^2
Simple linear regression	18895160.09	80%	2824.49	R2 train data: 0.643, R2 test data: 0.725
Polynomial regression	18895160.09	88%	2824.49	R2 train data: 0.643, R2 test data: 0.725
Random forest	MSE train data: 373 8781.747, MSE test data: 21893323.009	[the evaluation shows on the fig2]	[the evaluation shows on the fig2]	R2 train data: 0.9 71, R2 test data: 0.877

5. Conclusions

This passage mainly talks about the prediction of medical insurance cost using machine learning. The main thread is first applying the data from Kaggle and preprocessing it. Then, it is to train different models to compare them. It is concluded that polynomial regression is the best model. Correctly realize the perfect model, its value is to be used to enhance efficiency and accuracy among insurance industry. However, this research is still immature because it lacks validation. In fact, manners like cross-validation and ANOVA can be applied.

References

[1] *National Health Accounts, National Health Systems Resource Centre, https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.*

[2] *Research Gate, https://www.researchgate.net/publication/339416026_Linear_regression_model_for_predicting_medical_expenses_based_on_insurance_data.*

[3] *Roland S., Ruopeng A., Maroba J., Patel D.: The effects of obesity, smoking, and excessive alcohol intake on healthcare expenditure in a comprehensive medical scheme, South African Medical Journal 103, no. 11 (2013).*

[4] *Wolf, Anne M., and Graham A. Colditz: Current estimates of the economic cost of obesity in the United States, Obesity research 6, no. 2 (1998).*

[5] *Kim, David D., Basu A.: Estimating the medical care costs of obesity in the United States: systematic review, meta-analysis, and empirical analysis, Value in Health 19, no. 5 (2016).*

[6] *Wang, Houli, Xu T., Xu J., Factors contributing to high costs and inequality in China's health care system, JAMA 298, no. 16 (2007).*

[7] *Baal V., Pieter H. M., Johan J., et al.: Lifetime medical costs of obesity: prevention no cure for increasing health expenditure, PLoS medicine 5, no. 2 (2008).*

[8] *Warner, Kenneth E., Thomas A. Hodgson, and Caitlin E. Carrol., Medical costs of smoking in the United States: estimates, their validity, and their implications, Tobacco control 8, no. 3 (1999).*

[9] *Medical Cost Prediction Dataset, https://www.kaggle.com/hely333/eda-regression/data.*

[10] *Andrea Minieri, Automating data preparation and preprocessing in production-ready ML models - Clearbox AI.*