Prediction for Insurance Premiums Based on Random Forest and Multiple Linear Regression

Tingyuan Zhang *

Department of Mathematics, University of Macau, Macau, China

* db92810@um.edu.mo

Abstract. Insurance benefit forecasting is very important for insurance, and research on insurance benefit forecasting has been going on all the time. This paper aims to find an efficient and simple model for predicting insurance benefits based on multiple linear regression and machine learning scenarios. To be specific, in the process of prediction, random forest and linear regression are used as prediction models. Through the comparison and analysis of the results, it is found that the random forest is more accurate in predicting the results but lacks interpretability. Although linear regression is not as accurate as random forest, it can clearly explain the model and facilitate analysis and discussion. Although the two models have their own advantages. Based on the comparison and analysis, they can provide some help in referring to whether to use these two models. These results shed light on guiding further exploration of predicting insurance benefits in a simple way.

Keywords: Insurance Premiums; Prediction; Random Forest; Linear regression.

1. Introduction

It is generally believed that insurance first appeared in sea. Due to the insecurity of the sea, people needed to make sure their families had life security if they could not survive after sailing. One of the reasons insurances originated at sea is that the sea is the best place to escape feudalism [1]. This shows that insurance is a product of social development. Insurance is believed to boost economic growth. When entrepreneurs are exposed to risks (e.g., fire), insurance can provide protection and facilitate economic activity [2]. Insurance also has other value, it can bring Responsibility, Solidarity, Justice and Truth [3]. Therefore, insurance is a very important industry. For insurance forecasting, using existing data to calculate future demand enables insurance companies to plan, in advance, which is conducive to saving costs and reducing losses [4]. Prediction as part of data mining has many benefits for the insurance industry. It reduces costs, increases revenue, attracts more customers, and continues to innovate for insurers [5].

Currently, there are many methods for predicting insurance benefits. For instance, in predicting insurance benefits using predicted mortality, a classic model is the Lee–Carter model [6]. It uses time series, using the principle of random walk, to make predictions. In health insurance, people use artificial intelligence and machine learning to estimate insurance premiums [7]. Based on the analysis of personal characteristics, use neural network models to analyze and predict. In order to find the relationship between income and premiums, some scholars use time series analysis and cross section [8]. Researchers find relationships by analyzing the degree of correlation in a specific time and differences over time. Others use the panel smooth transition regression (PSTR) model to study the relationship between income and premiums [9]. PSTR is used to verify that the two have a non-linear relationship. In addition, the Markov Model and Generalized Linear Model are used together to predict trends in premiums [10]. Its purpose is to obtain unobservable data. In conclusion, there are many ways to predict premiums, and they are constantly evolving and improving.

This article aims to obtain an effective and relatively accurate model for predicting insurance premiums by observing individual characteristics (e.g., gender, age, smoking) with a relatively simple and low-cost method. Random forest and multiple linear regression model will be used for model building. In section 2, the principles of these two models will be briefly introduced. In section 3, data processing and analysis of model rationality will be carried out. In section 4, the effect of the model will be tested by comparing other statistical values such as R^2 and MSE. The results of the model

will be presented. In section 5, significance and limitations of models will be evaluated. In section 6, there are some conclusions for this paper.

2. Statistical Principleto

2.1 Random Forest

To introduce random forest, Breiman' paper and Biau & Erwan's paper will be refered [11, 12]. Since predictors which are used is surpervised, then this researh just introduce random forests for regression in mathematics. First of all, the means of random in random forest should be known. There are two aspects for randomness. For evry tree which is produced, the sample are chosen randomly. On the other hand, m predictors in every split in every tree are chosen randomly from p predictors ($\mathrm{m} \leq \mathrm{p}$). Now, for observed set $\mathbf{X} \subset \mathbb{R}^p$ where $\boldsymbol{x} \in \mathbf{X}$ is random vector with p predictors, it need uses $\boldsymbol{x}$ to predict $\mathrm{Y} \subset \mathbb{R}$. A regression function $h(\boldsymbol{x})$ should be faound and fatisfies $h(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$. To produce trees, this process need a train set $\mathcal{L}_n = \big((\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \dots, (\boldsymbol{X}_n, Y_n)\big)$ which is independent random valuable for each sample. $\mathrm{h_n}: \mathbf{X} \to \mathbb{R}$ should be found by use $\mathcal{L}_n$, and when $\mathrm{n} \to \infty$, $\mathrm{h_n} \to \mathrm{h}$, because $\mathrm{E}[\mathrm{h_n^2} - \mathrm{h}^2] \to 0$ as $\mathrm{n} \to \infty$.

Now, some decisions need be mad, producing m trees, and genrating a set of random vedctors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_\mathrm{m})$ where $\boldsymbol{\theta}_k$ is a random vextor for kth tree($1 \leq \mathrm{k} \leq \mathrm{m}$). $\boldsymbol{\theta}$ is used to be a prior for train set to grow different trees and choose predictors for splitting. For kth tree, it can be expressed as

$$h_n(\boldsymbol{x}; \boldsymbol{\theta}_k, \mathcal{L}_n) = \sum_{i \in \mathcal{L}_n^*(\boldsymbol{\theta}_k)} \frac{1_{\boldsymbol{X}_i \in A_{n(\boldsymbol{x};\boldsymbol{\theta}_k,\mathcal{L}_n)}} Y_i}{N_{n(\boldsymbol{x};\boldsymbol{\theta}_k,\mathcal{L}_n)}} \tag{1}$$

where $\mathcal{L}_n^*(\boldsymbol{\theta}_k)$ is the number of the data set which was choosen by $\boldsymbol{\theta}_k$, $\boldsymbol{X}_i \in A_{n(\boldsymbol{x};\boldsymbol{\theta}_k,\mathcal{L}_n)}$ means the $\boldsymbol{X}_i$ belongs to the data set which was choosen by $\boldsymbol{\theta}_k$, $N_{n(\boldsymbol{x};\boldsymbol{\theta}_k,\mathcal{L}_n)}$ is the number of the $\boldsymbol{X}_i$ belongs to $A_{n(\boldsymbol{x};\boldsymbol{\theta}_k,\mathcal{L}_n)}$. Then, one can combine m trees into one function

$$h_{m,n}(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{L}_n) = \frac{1}{m} \sum_{k=1}^{m} h_n(\boldsymbol{x}; \boldsymbol{\theta}_k, \mathcal{L}_n) \tag{2}$$

When $\mathrm{m} = \infty$, $h_{m,n}(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{L}_n) = E_\theta[h_n(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{L}_n)]$. Hence if $\mathrm{m} \to \infty$, one derives:

$$\lim_{m \to \infty} h_{m,n}(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{L}_n) = E_\theta[h_n(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{L}_n)] \tag{3}$$

Therefore, in theory, the more trees are generated, the more accurate the model will be.

2.2 Linear Regression

Linear regression is a very sample method to predicte a quantitative response Y. According to the number of predictors X, it is divided into simple linear regression with one predictor and multiple linear regression with more than one predictors. For simple linear regression, the linear relationship is

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \varepsilon \tag{4}$$

where ε is a random variable has a normal distribution with mean 0. It can be estimated it using odserved data $(\mathrm{x_1}, \mathrm{y_1}), (\mathrm{x_2}, \mathrm{y_2}), \dots, (\mathrm{x_n}, \mathrm{y_n})$, the function can be written as

$$\widehat{\mathrm{y_i}} = \widehat{\beta_0} + \widehat{\beta_1} \mathrm{x_i} \tag{5}$$

where $\hat{\cdot}$ means it is estimated value for unknow value. Now, if to compute $\widehat{\beta_0}$ and $\widehat{\beta_1}$, the residual $\mathrm{e_i} = \mathrm{y_i} - \widehat{\mathrm{y_i}}$ which is the difference between observed responce $\mathrm{y_i}$ and predicted responce $\widehat{\mathrm{y_i}}$ should be used. Then, one uses the residual to define the residual sum of squares (RSS), which is

$$\mathrm{RSS} = \mathrm{e_1^2} + \mathrm{e_2^2} + \cdots + \mathrm{e_n^2} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e_i} \tag{6}$$

Then, one uses the least squares approach to find $\widehat{\beta_0}$ and $\widehat{\beta_1}$, which can minimize RSS. Let

$$\begin{cases} \frac{\partial RSS}{\partial \beta_0} = 0 \\ \frac{\partial RSS}{\partial \beta_1} = 0 \end{cases} \tag{7}$$

solving the equaton, one derives

$$\begin{cases} \widehat{\beta_1} = \frac{\sum(x_i-\bar{x})(y_i-\bar{Y})}{\sum(x_i-\bar{x})^2} \\ \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \end{cases} \quad (8)$$

For multiple linear regression, the linear relationship with p predictors is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (9)$$

The least squares approach also be used to find $\widehat{\beta_0}$, $\widehat{\beta_1}$,…, $\widehat{\beta_p}$ which can minimize RSS. It is the same as simple linear regression. By the way, polynomial regression can be seen as a special multiple regression, using different degrees of the polynomial as predictors.

3. Data and Method

3.1 Data

For the data set this paper used is found from Kaggle, called Medical Cost Personal Datasets. Here are some basic information about the data. The response is charges, which corresponds to individual medical costs billed by health insurance. The six predictors are given as follows:

Age: age of primary beneficiary

Sex: gender of insurance contractor

BMI: body mass index, providing an understanding of body, weights that are relatively high or low relative 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 health insurance / Number of dependents

Smoker: smoking

Region: the beneficiary's residential area in the US, northeast (NE), southeast (SE), southwest (SW), northwest (NW)

This data set has 1338 samples without any missing and the statistical descriptions are summarized in Table. 1 and Table. 2.

Table 1. Details for quantitative valuables

Valuable	Valid number	Mean	Std. Deviation	Quantiles: Min	25%	50%	75%	Max
Age	1388	39.2	14	18	27	39	51	64
BMI	1388	30.7	6.1	16	26.3	30.4	34.7	53.1
Children	1388	1.09	1.21	0	0	1	2	5
Charges	1388	13.3k	12.1k	1.12k	4.74k	9.39k	16.7k	63.8k

Table 2. Details for qualitative valuables

Valuable	Valid number	Proportion
Sex	1388	Male:51% Femal:49%
Smoker	1388	Yes:80% No:20%
Region	1388	NE:25% SE:28% SW:25% NW:25%

3.2 Feture Selection

3.2.1 Pearson correlation

During prediction, there are some predictors that have little or no effect on the response. When building a model, in order to reduce the complexity of the model and reduce the amount of computation, these variables will be removed. For how to find these variables, it is necessary to introduce filtering method. For this paper, Pearson correlation are used to filter feture. Pearson correlation [13], one of the most commonly used correlation coefficients. This statistic value is used to describe the degree of linear correlation between two variables, ranging from -1 to +1, "-" means

negative correlation, "+" means positive correlation. The formula for Pearson correlation $\rho(x,y)$ or r_{xy} can be expressed as

$$\rho(x,y) = r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{10}$$

where s_{xy} is the covariance between two variables x and y, s_x and s_y are standard deviations for X and Y.

3.2.2 Point biserial & eta correlation

Since Pearson correlation is used to express the degree of correlation between two numerical valuable, then one should find a statistic value to express correlation between numerical valuable and categorical variable. And for ordinal variable, Spearman correlation is used. But for valuable in this data, those are disordered. So, point biserial correlation and eta correlation will be concerned. Point biserial correlation [14] is use for categorical variable which is categorical variable including gender, 0/1 variable, etc. The point biserial correlation r_{pb} is

$$r_{pb} = \frac{\overline{Y_1} - \overline{Y_0}}{s_y} \sqrt{\frac{N_0 N_1}{N(N-1)}} \tag{11}$$

where $\overline{Y_0}$ and $\overline{Y_1}$ are means of each category, N_0 and N_1 are number of each category, N is total number of Y, s_y is standard deviation for Y. Since r_{pb} is generated by Pearson's product moment correlation, the result of r_{pb} is equivalent to the Pearson correlation.

As for categorical variable which is larger than two levels, eta correlation [15] is concerned. Eta correlation is represented by η and formula is

$$\eta^2 = 1 - \frac{\sigma^2_{Y_c}}{\sigma^2_Y} = \frac{\sum_c N_c (\overline{y_c} - \overline{y})^2}{\sum (y_i - \overline{y})^2} \tag{12}$$

where σ^2_Y is variance of Y, $\sigma^2_{Y_c}$ is variance of Y which is divide by different categories, N_c is number of different categories and $\overline{y_c}$ is mean of different categories. This η^2 can be seen as R^2 in multiple linear regression. Hence, it's range is [0,1], the closer to 0 the smaller the correlation, the closer to 1 the greater the correlation.

3.3 Model & Metric

This paper use two model, random forest and multiple regression. The advantage of random forest is that it can effectively reduce variance without increasing bias. It is a very commonly used method in machine learning. As for using multiple regression instead of polynomial regression, it's because of bias. Also, it will increase the complexity of the model, and the added variables sometimes don't have much impact. Since polynomial regression is seen as a special multiple regression, then using correlation to filter variable. For quantitative valuables, Pearson correlation will be the standard for measuring variables of different times, and the largest one will be used as a variable. To compare accuracy of models, the coefficient of multiple determination R^2 will be used [16]. R^2 can be a good assessment of the predictive ability of the model. R^2 is

$$R^2 = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2} \tag{13}$$

For reference, the mean squared error (MSE) will also be calculated.

$$MSE = \frac{\sum |\hat{y}_i - y_i|^2}{n} \tag{14}$$

4. Results and Discussion

4.1 Feature Selection

Primarily, for quantitative valuables, the Pearson correlation for each predictor in different power. Fig. 1, Table. 3 and Table. 4 are the results of Pearson correlation for different predictors. From result, quadratic valuable for age is best, linear is beat for BMI and children. Then using point biserial correlation and eta correlation to test qualitative valuables. If result is very small, this predictor may

be considered to giving up. According to the analysis, smoker has high correlation, but sex and region show low correlation. So, sex and region will be given up. Hence, the feature selection get quadratic for age, linear for BMI and children and qualitative smoker.

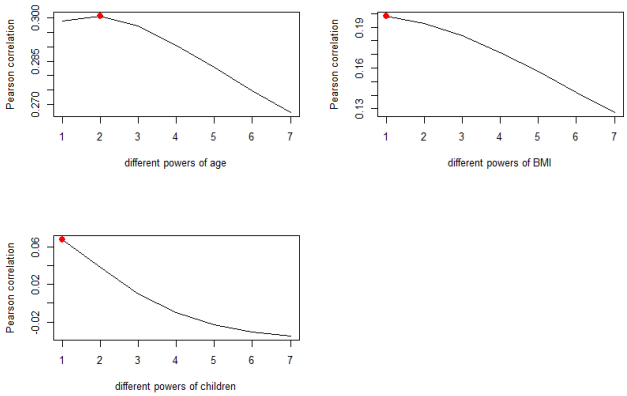

Fig. 1 The results of Pearson correlation for different predictors.

Table 3. Pearson correlation for different predictors

Valuable	Linear	Quadratic	Cubic	Quartic	Quadrillion	Sixth	Seventh
Age	0.2990	0.3007	0.2971	0.2905	0.2827	0.2747	0.2669
BMI	0.1983	0.1929	0.1837	0.1715	0.1573	0.1421	0.1269
Children	0.0679	0.0386	0.0101	-0.0102	-0.0230	-0.0308	-0.0354

Table 4. Point biserial correlation and eta correlation for qualitative valuables

Valuable	Point biserial correlation	Eta correlation
Sex	-0.0572	-
Smoker	0.7872	-
Region	-	0.0814

4.2 Random Forest

For random forest, there are 500 trees are built, and 2 valuables for each split. Here are some information for this model. According to the results of R^2 in Table. 5, the fitting degree of random forest is still relatively high. Random forest is able to derive the importance of each variable by calculating the average decrease in the Gini coefficient. The results are shown in Fig. 2. Apparently, age, bmi and smoker are very important. It also proves that the result of feature selection is correct.

Table 5. Accuracy of random forest

Model	MSE	R^2
Random foreset	21963824	0. 8501

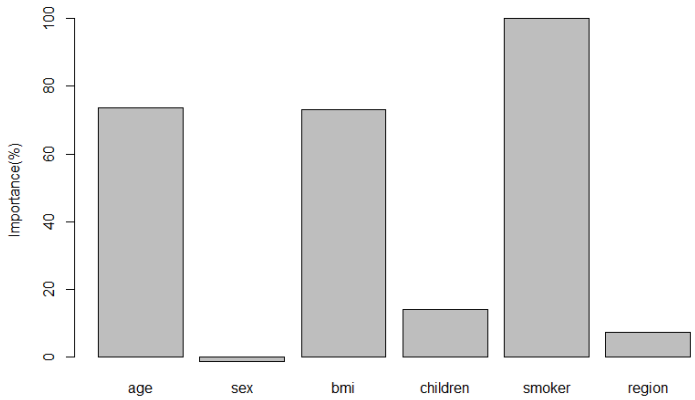

Fig. 2 The results of the importance of each variable

4.3 Linear Regression

For linear model, there are four valuables, quadratic for age, linear for BMI and children and qualitative smoker. Since it is a regression model, then qualitative smoker is transformed into dummy variable. Results of the model are listed in Table 6. and Table 7. Since the p-value which be shown by $Pr(>|t|)$ is very small, then the features selected are fitting. Although MSE and R^2 are not particularly good, it can also be considered a relatively fitted model.

Table 6. Result of linear regression

Valuable	Coefficient	Standard error	$Pr(>\|t\|)$
Intercept	-12102.77	941.98	< 2e-16
Age	257.85	11.90	< 2e-16
BMI	321.85	27.38	< 2e-16
Children	473.50	137.79	0.000608
Smoker	23811.40	411.22	< 2e-16

Table 7. Accuracy of linear regression

Model	MSE	R^2
Linear regression	36680456	0.7497

4.4 Comparison and Application

Comparing two models, each model has advantages and disadvantages. The MSE of random forest is smaller and R^2 is larger, so the accuracy is better and the model fits better. However, there is a serious problem with random forest, that this model is a black box, interpretability is given up for better accuracy, random forest cannot explain the relationship between predictors and response. For linear regression, though the accuracy is not as good as random forest, all its factors are obvious. It is easy to obtain a lot of information through this model, and even some unobservable information can be obtained by calculation.

Therefore, random forest is more suitable for analysis that focuses on predicting results and does not care about how the model itself is constructed, and the relationship between factors is not important. Although linear regression is slightly less accurate, it has strong explanatory power for each factor. Thus, it is suitable for decision analysis, i.e., adjusting the plan according to the relationship between the predictor and the response. Hence, for predicting insurance premiums, may be linear regression is better.

5. Limitations and Future Outlooks

First of all, for data, there are not many predictors collected, because a person's health is actually related to many aspects, not just 6 factors. Therefore, when the model is established, there are many factors that are not observable, which will make the model fitting difficult, and there are also some difficult-to-interpret aspects when interpreting the model. Second, both models also have some limitations. Random forest has high accuracy for prediction, but the fitting process cannot be known, so it is very difficult to explain the model. Random forests cannot tell the relationship between specific variables. The limitations of linear regression are also obvious. Its accuracy is not as good as random forest. In essence, linear regression mainly builds a linear model, but the real situation cannot be arranged linearly, so the model will produce large errors. In the process of factor selection, in order to avoid overfitting of the model, some predictors are abandoned. Therefore, the model also loses the ability to explain these predictors. Although the correlation between these predictors and the response is not large, they still have a certain impact on the response.

In the future research, the data collection will be more detailed and systematic. For responses with more influencing factors, as many predictors as possible will be collected. For random forests, research in recent years has been optimizing this model, and while improving its accuracy, it is also

hoped to improve its explanatory power. Linear regression has always been a classic model, and then it needs to be optimized and adjusted between overfitting and accuracy.

6. Conclusion

In the insurance industry, the prediction of premiums is very important. To sum up, this paper compares the effects of two models, random forest and linear regression, on the prediction of premiums. By comparison, it is found that the accuracy of random forest is higher, but the relationship between predictors and responses cannot be explained for the model. Although linear regression is not as accurate as random forest, the relationship between predictors and responses can be discussed. Based on the characteristics of the insurance industry, it is not only necessary to predict the results, but also need to adjust the decision through the analysis model. Therefore, linear regression may be more suitable for the prediction of insurance benefits. Nevertheless, both models have some limitations. Random forest cannot explain the model, and linear regression reduces some predictors because of its linear characteristics, but it also sacrifices the information contained in these predictors. In the future, data collection should more be more detailed. Random forest may be optimized in explanation. Linear regression may be more adaptable for many models. Overall, by comparing the two models, the differences between the two models in the prediction of insurance benefits are shown, and they offer a guideline for the prediction and analysis of insurance benefits.

References

[1] Clark G W. Betting on lives: life insurance in English society and culture, 1695-1775. Princeton University, 1993.

[2] Masci P. The history of insurance: risk, uncertainty and entrepreneurship. Business and Public Administration Studies 2021, 6.1, 25-25.

[3] Ewald F. The values of insurance. Grey Room 2019, 74: 120-145.

[4] Mantis G., and Richard N. F. Demand for life insurance. Journal of Risk and Insurance, 1968: 247-256.

[5] Umamaheswari K. and Janakiraman S. Role of data mining in insurance industry. Int J Adv Comput Technol 2014, 3: 961-966.

[6] Lee R. D., and Lawrence R. C. Modeling and forecasting US mortality. Journal of the American statistical association, 1992, 87.419: 659-671.

[7] Kaushik K., et al. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums." International Journal of Environmental Research and Public Health 2022, 19.13: 7898.

[8] Beenstock Michael, Gerry Dickinson, and Sajay Khajuria. The relationship between property-liability insurance premiums and income: An international analysis. Journal of risk and Insurance, 1988: 259-272.

[9] Lee Chien-Chiang, and Chiu Yi-Bin. The impact of real income on insurance premiums: Evidence from panel data." International Review of Economics & Finance 2012, 21.1: 246-260.

[10] Berry Lucas. Hybrid Hidden Markov Model and Generalized Linear Model for Auto Insurance Premiums. Diss. Concordia University, 2016.

[11] Breiman Leo. Random forests. Machine learning, 2001, 45.1: 5-32.

[12] Biau Gérard, and Erwan Scornet. A random forest guided tour. Test 2016, 25.2: 197-227.

[13] Goodwin Laura D., and Nancy L. L. Understanding correlation: Factors that affect the size of r. The Journal of Experimental Education 2006, 74.3: 249-266.

[14] Kornbrot Diana. Point biserial correlation. Wiley StatsRef: Statistics Reference Online, 2014.

[15] Wherry R. J., and Erwin K. T. The relation of multiserial eta to other measures of correlation. Psychometrika 1946, 11.3: 155-161.

[16] Lewis-Beck Michael S., and Andrew S. The R-squared: Some straight talk. Political Analysis 1990, 2: 153-171.