Prediction of Annual Medical Cost

Bixuan LIU

January 14, 2024

Abstract

This project considers a dataset from kaggle, which has some information of clients from a health insurance company and their annual medical costs. It is very important for a health insurance company to predict the annual medical cost of a potential client, based on which the company can determine a reasonable premium. Therefore, in this project I implement four supervised machine learning methods, namely Gaussian Linear Model, Lasso, Regression Tree and Random Forest, to predict the annual medical costs, and select the best model for the task. Moreover, based on the results of this project, I'm able to design a short questionnaire for the potential clients of the health insurance company which will help them to predict the future annual medical costs and determine premiums.

Keywords: Health Insurance, Lasso, Random Forest

Contents

1 Motivation

It's very important for a health insurance company to make correct predictions of a potential client's annual medical cost, which is the foundation of determining a premium. Therefore, there are three main objectives for this project:

1. Select the most appropriate machine learning method to predict the annual costs covered by health insurance of clients.
2. Select the most influential factor(s) for the prediction of annual medical cost.
3. Design a short questionnaire for the potential clients of the health insurance company.

2 Dataset

2.1 Source

This dataset is from kaggle: https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download.

2.2 Description

The data is information collected from the clients of a health insurance company. It contains each client's individual medical costs billed by health insurance and several features of the client which may contribute to the costs. Names and descriptions of the entries are listed below:

- **age**: age of primary beneficiary;
- **sex**: insurance contractor gender, female, male;
- **bmi**: body mass index;
- **children**: number of children covered by health insurance / number of dependents;
- **smoker**: smoking;
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest;
- **charges**: individual medical costs billed by health insurance.

See table 1 for the first three rows of the dataset as an example.

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462

Table 1: First Three Rows of Dataset

2.3 Visualization

From table 1, it's clear that the dataset contains both numerical (age, bmi, and charges) and categorical (sex, children, smoker, and region) entries. Next, I visualize these entries separately.

From figures 1 and 2, we can already observe some patterns about annual medical charges. There are three important columns: age, bmi and smoker, while the others do not show significant relations with the charges.

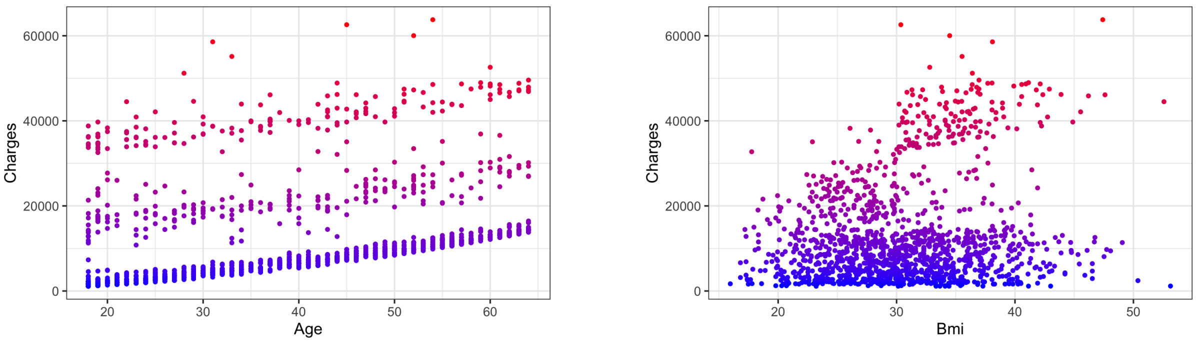

Figure 1: Age and Bmi vs Charges

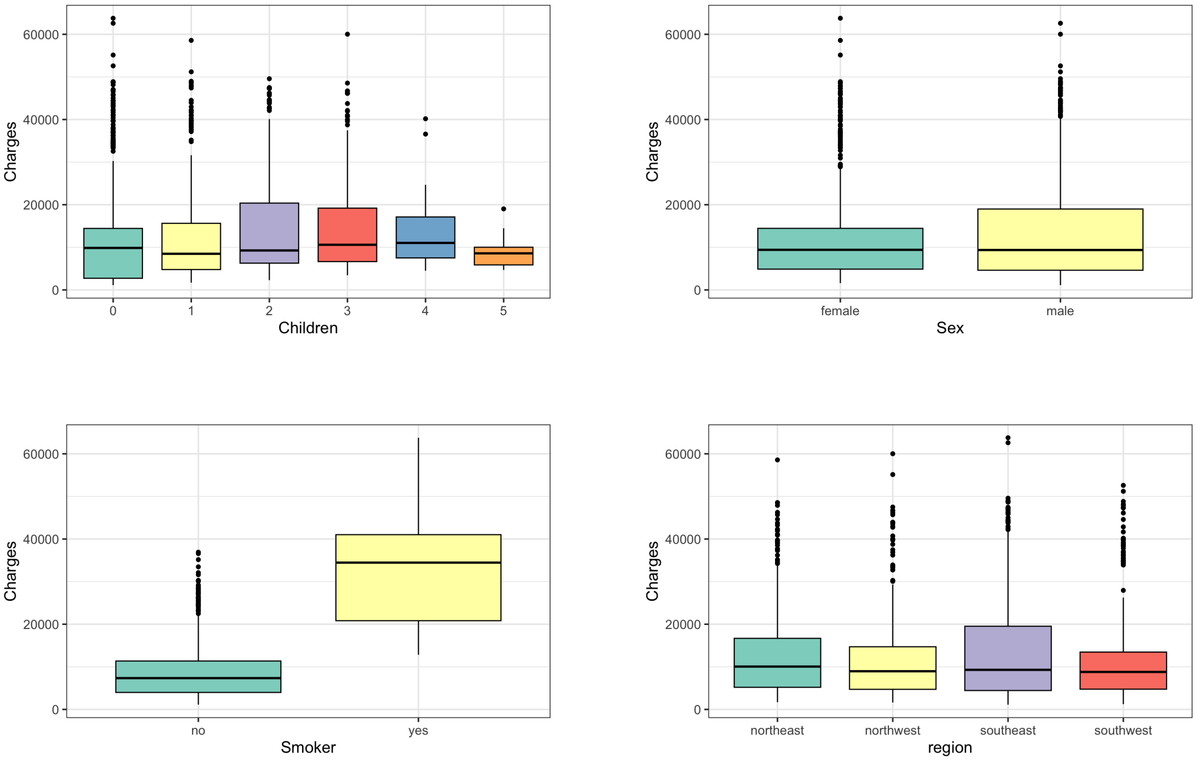

Figure 2: Children, Sex, Smoker and Region vs Charges

3 Model Selection

In this section, I consider the following four machine learning methods:

- Gaussion Linear Model;
- Lasso;
- Regression Tree;
- Random Forest.

The purpose of this section is to select the best model for the prediction of annual charges and selection of important variable(s) that contribute(s) the most to the prediction.

3.1 Steps for Fitting each Model to the Data

- **Step 1,** split the dataset into 80% of training and 20% of testing set randomly.
- **Step 2,** fit the model to training set.
- **Step 3,** calculate the studendized residuals of training set, detect outliers and remove them.
- **Step 4,** fit again the model to cleaned training set and evaluate the model.
- **Step 5,** calculate RMSFE (Root Mean Squared Forecast Error) on the test set.

3.2 Gaussian Linear Model

Gaussian Linear Model is generally represented as:

$$Y = X\theta + \epsilon \tag{1}$$

where Y is of dimension $n \times 1$, representing the response variable (variable to be predicted) of n observations. Y_i's are assumed to be independent and Gaussian distributed. X is of dimension $n \times (p+1)$, representing n observations of p explanatory variables and 1 constant. θ is of dimension $(p+1) \times 1$, representing fixed and unknown parameters. ϵ is of dimension $n \times 1$, representing the residuals. Here, we assume ϵ's are independently, identically distributed as $\mathcal{N} \sim \left(0, \sigma^2\right)$.

When applying Gaussian Linear Model to the dataset, categorical columns are transfered into numerical using one-hot-encoding. After fitting the model directly to training data, we get R^2 equals 74.32%, which is fairly good. Next, I use studentized residuals to detect outliers (see figure 3). 57 outliers are detected.

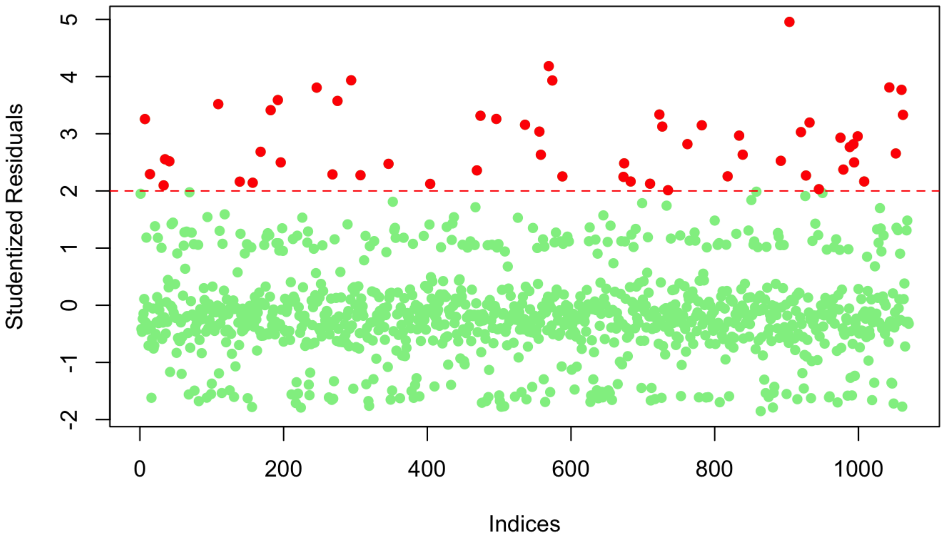

Figure 3: Outlier Detection for Gaussian Linear Model

After removing the outliers in the training set, I fit again the model to the cleaned training set. The new R^2 value is 84.13%, which is improved compared to the previous model. The coefficients are listed in table 2. See figure 4 for the distribution of standardized residuals. The distribution of standardized residuals is fairly close to a normal distribution, which means that Gaussian Linear Model generally captures the relations between the variables. But clearly linear models are not the best choice since there clearly are some nonlinear patterns in the data.

Attribute	Estimated Coefficient
Intercept	-12931.42
age	256.63
sexfemale	-235.26
bmi	338.98
children	256.63
smokeryes	256.63
regionsoutheast	-940.62
regionnorthwest	-341.06
regionsouthwest	-429.29

Table 2: Coefficients Estimated by Gaussian Linear Model

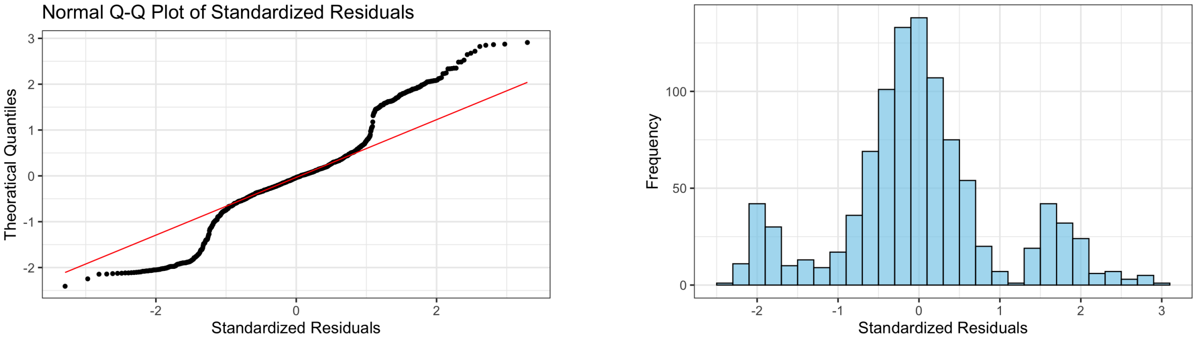

Figure 4: Distribution of Standardized Residuals of Gaussian Linear Model

3.3 Lasso

Lasso is one of the most influential penalized methods in machine learning. Lasso estimator is the solution of the following optimization problem:

$$\hat{\beta}_L = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{i,j}\beta_j \right)^2 + \lambda \|\beta\|_1 \right\} \tag{2}$$

By penalizing the loss function with l_1-norm, Lasso plays an important rule not only in prediction, but also in variable selection.

Since Lasso is particularly sensitive to the scale of the features, first I scale the columns: age, bmi and children to $0 \sim 1$, and then fit the model to training data. Since Lasso is another linear model, here I don't list the detailed results of coefficients and distributions of residuals anymore, and focus on variable selection instead.

When I set λ in equation 2 big enough, 500 for example (table 3), Lasso is able to select the following 4 variables: age, bmi, children, smokeryes, and set the coefficients of other features to 0. It's also worth mentioning that from the absolute values of the coefficients, we can conclude that in the four selected features, smoker is the most influential while children is the least.

3.4 Regression Tree and Random Forest

Unlike above mentioned methods, regression tree builds a decision tree with certain nodes and leaves that can capture non-linear relationships. In this section, I only show the results with cleaned training set, the process of detecting outliers are omitted.

Attribute	Estimated Coefficient
Intercept	8715.7824
age	3276.0246
sexfemale	0
sexmale	0
bmi	1489.3908
children	127.6906
smokeryes	22385.9093
regionsoutheast	0
regionnorthwest	0
regionsouthwest	0

Table 3: Coefficients Estimated by Lasso with $\lambda = 500$

Figure 5 gives an instance of the first two layers of the regression tree built on the cleaned training set. The model has R^2 equals 93.42%, which is fairly high, suggesting that the model fits very well with the training data.

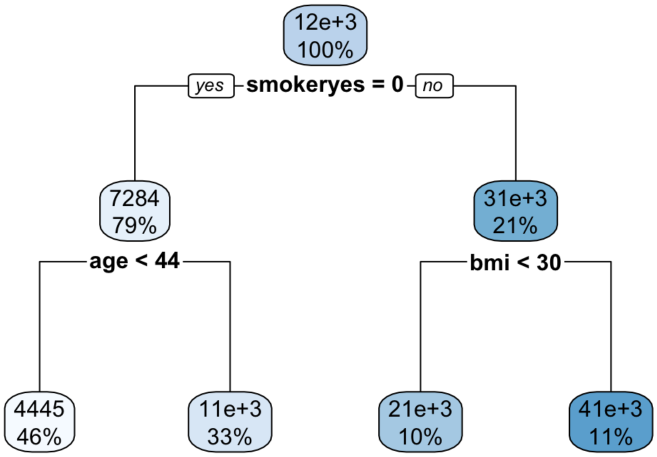

Figure 5: Instance of Regression Tree Structure

Random forest is a combination of several regression trees, which should behave better than simple regression tree. After fitting a random forest with 100 regression trees to the training set, R^2 becomes 95.70%. Figure 6 shows the behavior of the residuals.

4 Conclusion

Table 4 shows the comparison of all models. It's clear that Random Forest is the best among the four methods considered in this project. And it reaches a lowest root mean squared forecast error of 0.32.

Based on all the results above, if I need to design a questionnaire for the insurance company, the three questions that should be asked are:

1. How old are you?

2. Do you smoke?

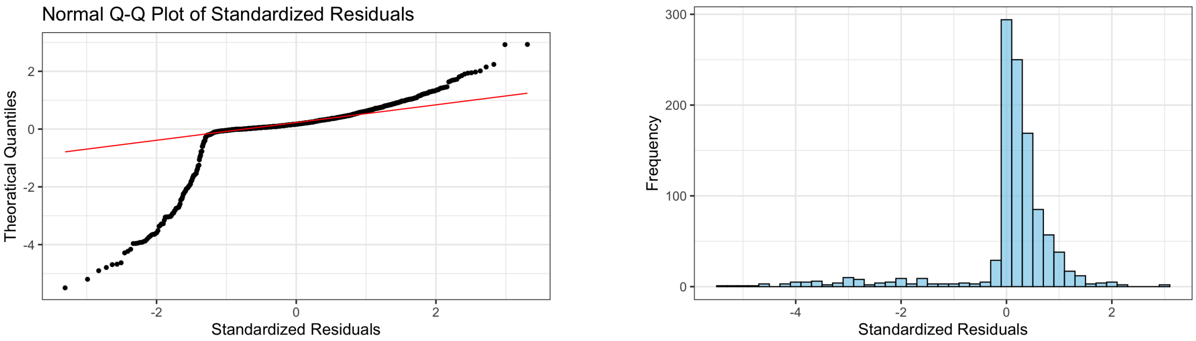

Figure 6: Distribution of Standardized Residuals of Random Forest

Model	GLM	Lasso	Regression Tree	Random Forest
R^2 (%)	74.32	75.62	93.41	95.69
$RMSFE$	0.43	0.43	0.37	0.32

Table 4: Comparison of All Models

3. What's your bmi?

With the answers to the above questions, the insurance company will be able to make a fair prediction on the potential client's annual medical cost and define a premium.