Predict the Insurance charges (Most Parsimonious Model)

Sanzida Parvin

Executive Summary

The goal of this project is to find out the most parsimonious model for predicting the insurance cost of an individual based on some independent variables like age, sex, region, number of children etc. By doing this analysis the target audiences of this project, any insurance company, will easily find out the most appropriate model to calculate their insurance charges to their customers and make any increase or decrease in charges if needed to maximize their sell and better serve their customers.

The data set for this project was collected from a popular website for data mining competitions named Kaggle (https://www.kaggle.com/mirichoi0218/insurance). It has six predictor variables and one response variable, based on what the model will be structured.

The variables are consisting of both categorical and quantitative, but the response variable is quantitative. So, the two data mining techniques that I choose to find out the most appropriate model are linear regression using regsubsets function and the artificial neural network.

First Technique

After exploring the data set, it was found that there were no missing values but the response variable ‘charges’ is highly right skewed

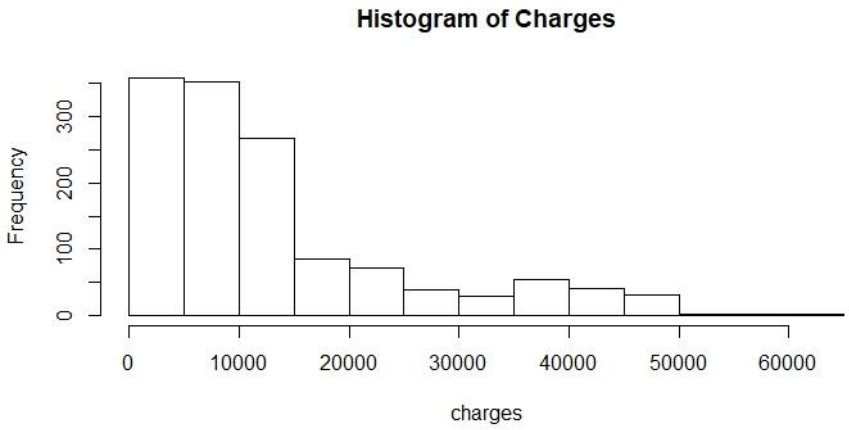

Fig1: Histogram of the variable charges

To treat this skewness, log transformation of the variable was used for linear regression analysis. The residual plots for linear regression are shown below,

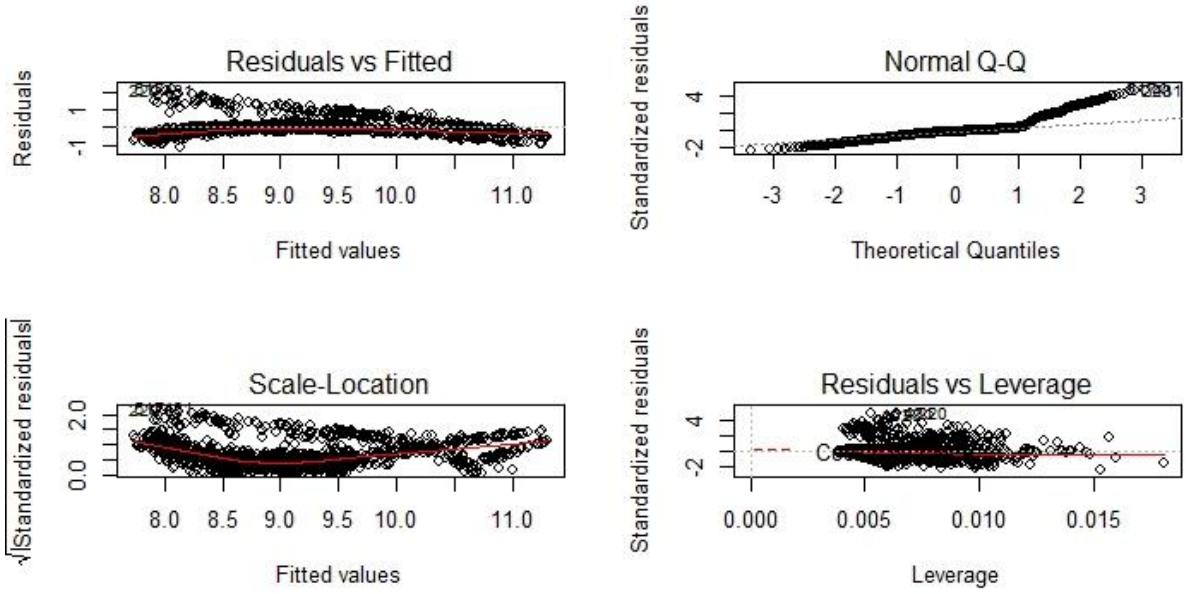

Fig2: Residual plots of insurance data

The regsubsets function find out a model with all six variables is suitable for the best analysis. Both the BIC and adjr2 shows that the full model will give the best predictions.

To get a more authentic and parsimonious model, a 10-fold cross validation was used. And the standard error of the cross-validation errors indicates that the model with 3,4,5 and 6 variables will give the best model.

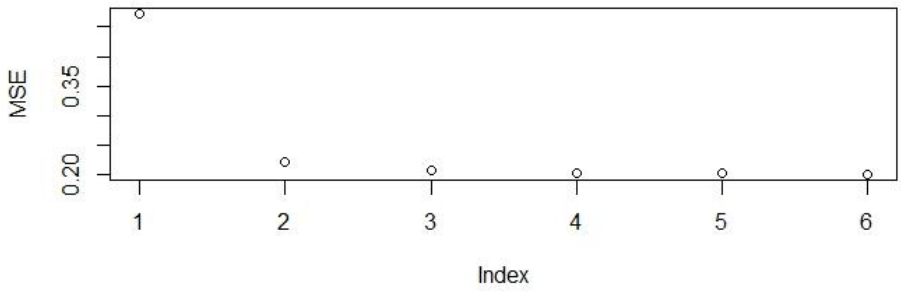

Fig3: Standard errors of the cross-validation

To avoid the complexity of the model, I have selected the three variables as the most parsimonious model. The simplified model with three variables would be

y_hat = 7.28772342 + 0.03528491*age + 0.10163109*children + 1.54427238*smokeryes

Second Technique

The artificial neural network is working in a similar way of real neural network. Choosing the appropriate number of hidden nodes will give the most appropriate model of the analysis. The figure of the neural network model fit of one hidden node is showing below.

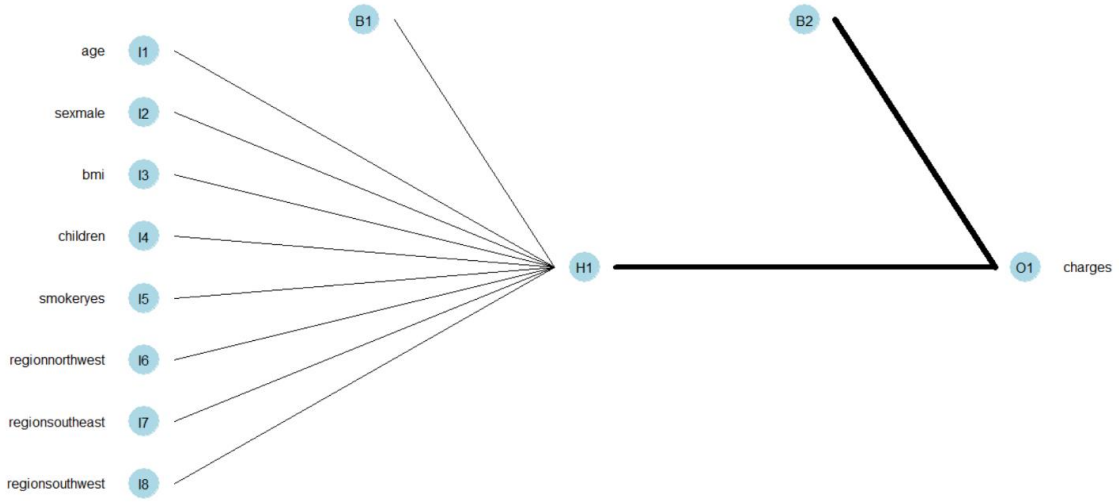

Fig4: Neural network plot with one hidden node

This figure didn't tell us which variables are most important to calculate the charges of the insurance. To find out the best model, a 10-fold cross validation with iterate over 1-8 hidden nodes was used to select the best number of hidden nodes for the best model. And the cross-validation result shows that with 4 hidden nodes will give the best model to calculate the charges of the insurance.

The figure of the neural network model with four hidden nodes (most appropriate model) is showing below

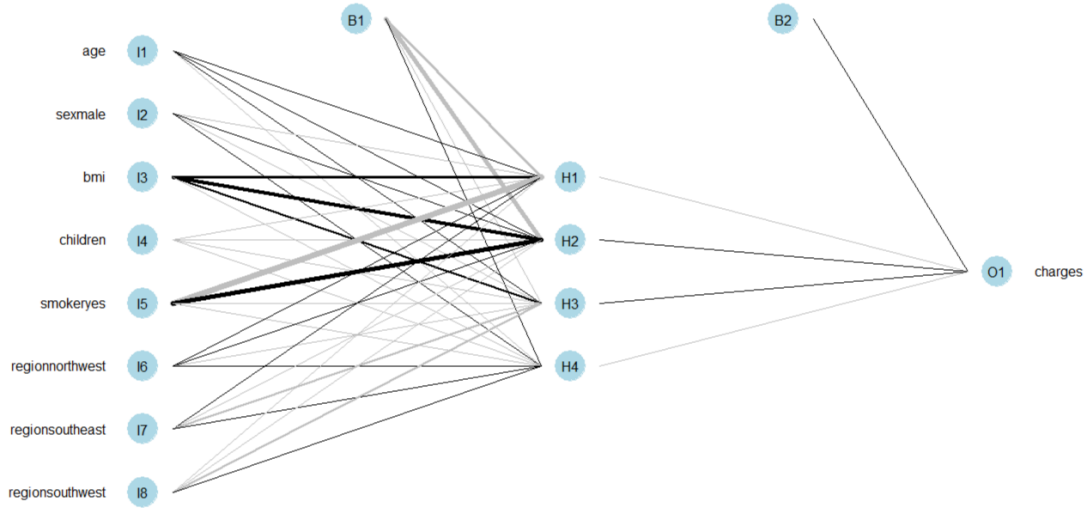

Fig5: Neural network plot with four hidden nodes

Here, the black edges represent positive weights. Gray edges represent negative weights, and the thicker edges represent weights with larger absolute values. From the plot of the neural network, we can see that the smokeryes has the largest positive weight.

Figure 5 shows the most important variables for the best model selection using four hidden nodes.

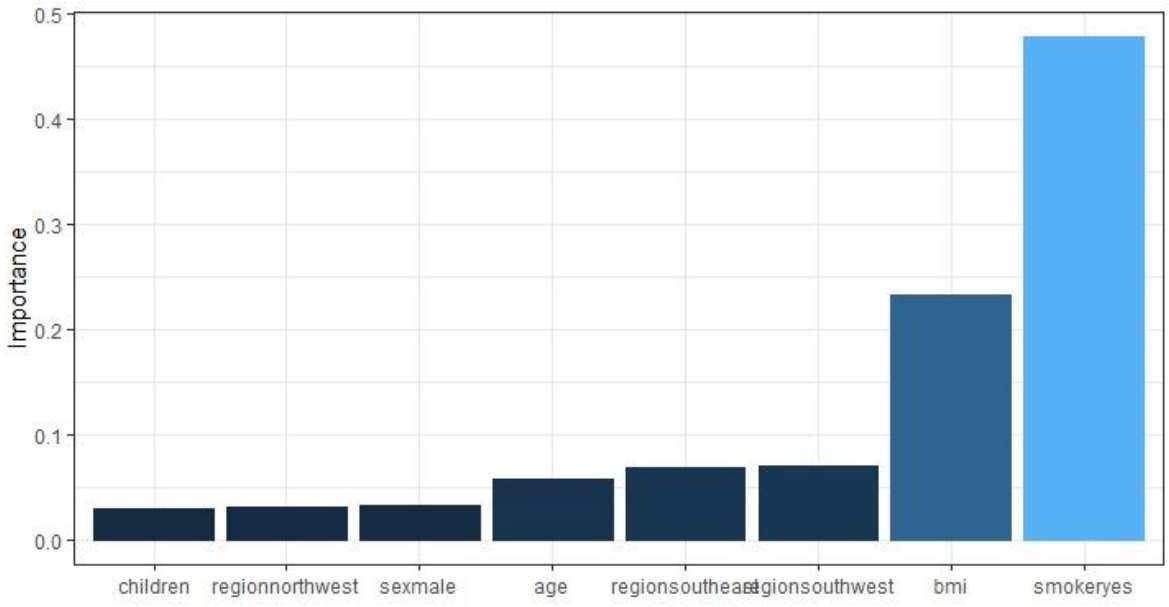

Fig6: Important variables of the neural network model

From the figure we can clearly see that the smokeryes and bmi are the most two important variables to predict the charges of the insurance for the customers.

Discussion

Both of the techniques show the most appropriate model to calculate the charges of the insurance for the target audience (insurance company). If more variables will add in future to analyze more complex situation, then I believe those techniques will also do that and give accurate results. Because the linear regression with regsubsets function and the neural network analysis are both robust type of techniques in the field of data mining.

References

1. https://www.kaggle.com/mirichoi0218/insurance

2. Course materials