FINAL REPORT

Team: Significant

CIS 9660: Data Mining for Business Analytics

Professor: Yuanfeng Cai

Thursday 6:05 PM- 9PM

Greg Martin Teo: gregmartin.teo@baruchmail.cuny.edu
Saranda Sopa Azemi: saranda.sopaazemi@baruchmail.cuny.edu
Pajtesa Rexhepi: pajtesa.rexhepi@baruchmail.cuny.edu
Yael Saposh: yael.saposh@baruchmail.cuny.edu

Introduction

United States is one of the countries that has the most expensive healthcare in the world. Many families spend their life in the borderline between the ability to afford health coverage and not qualifying for the free state Medicaid health coverage (for those whose income is below the poverty line)! For this Project we were interested to analyze the factors that contributed mostly to the final charges. After some research we found a dataset that contains variables related to this issue. This dataset is a public dataset and available via: https://www.kaggle.com/mirichoi0218/insurance. The file is downloaded as a csv file and is named insurance.

Exploratory data analysis

We started by performing exploratory data analysis to assess the variables, and how they interacted with each other or with the target variable. After reading through the file and assigning the first row as the header row we used the dim function and noticed that there are seven variables in total three of which are qualitative: sex with two levels ”female”, “male”; smoker with two levels: “yes”, “no”; region with four levels “southeast”, “southwest”, “northeast” and “northwest”; and four quantitative variables: age, bmi, children and charges. The set contains a total of 1338 observations.

In order to avoid an overwhelming number of paired scatterplots we decided to use only the quantitative variables first shown in Figure 1 in the Appendix below. This image shows that the scatterplot using age and charges indicates a positive linear relationship because as the age increases the charges also tend to increase. In the same figure we can find the scatterplot of bmi and charges which indicates no clear relationship between these two variables. And the last scatterplot in this figure between children and charges indicates a negative linear relationship, as the number of children increases the charges tend to decrease.

For the three remaining categorical variables we decided to use boxplots to analyze them. Figure 2 shows the boxplots, and we can confirm that sex and region have little impact on charges while smoker variable has a very large effect. Referring to figure 2, we see that the median charges for a smoker with level “yes” is much higher than the median of charges for a nonsmoker. Consequently, smoker variable also has higher upper and lower quartiles as well as a much higher upper and lower whisker. For sex variable we can notice that although the median is the same for both males and females, we see that the males have a higher upper quartile and higher upper whisker. This suggests that males who are charged above the median tend to be charged a higher rate than females who are charged above median. Next, for the region variables although the mean seems almost equal for all the boxplots, we notice that the southeast and northeast have higher upper quartiles and also higher upper whiskers. This indicates that the regions of northeast and southeast charges that fall above the median tend to be higher than the charges above the median in northwest and southwest. After assessing both boxplots and scatterplots we conclude that age, bmi and smoker variables have a larger impact on charges than children, sex, and region.

Linear Regression

Next, we wanted to create a linear regression model and decided to create the first model with all the variables included which gave us an R squared of 0.7494.

For the second model we wanted to create a model using only variables that we identified as important in exploratory data analysis: age, bmi and smoker. This model gave us a slightly lower adjusted R square of 0.7469.

Next, we decided to create a third model using the same three variables like model two only with one interaction between bmi and smoker and this returned a higher adjusted R squared value of 0.8385.

Furthermore, we wanted to create a fourth model which is model for with the same variables only the interaction is now between age and smoker and this model returns an adjusted R square of 0.7472 almost equal to the first model.

After considering each model we see that the third model is the best one so far. We continue our analysis by using the hold out method.

Hold-out method

To assess the performance of our models we used two types of resampling methods. First, we illustrate how to compare the models using hold-out. We use 80-20 hold-out in this case. We develop models using a training data set and assess the model performance using a test data set. We randomly select 80% of the observations for training and the remaining in the test set.

After that, we used AIC, BIC and adjusted R^2 to find the best model. After running each model AIC, BIC and adjusted R^2 we can see these results:

Model 1- **AIC:** 21649.65, **BIC**: 21699.4, **Adjusted R^2** : 0.7479 ;

Model 2- **AIC:** 21650.84**, BIC**: 21675.72, **Adjusted R^2** : 0.7464.

Model 3- **AIC:** 21177.29**, BIC**: 21207.14, **Adjusted R^2** : 0.8373;

Model 4- **AIC:** 21647.09**, BIC**: 21676.94, **Adjusted R^2** : 0.7475;

From the result above, we can conclude that the model with the lowest AIC, BIC and Adjusted R^2 is: Model 3**.** This model does a better job than the other models at predicting the test data.

Cross-Validation

Now we will illustrate how to compare the models using cross-validation. We used 5-fold cross validation along with MSE (Mean Squared Error) and RMSE (Root Mean Squared Error).

We run MSE and RMSE for each model from 1 to 4. And we wanted to compare which model is better in terms of reflecting performance when dealing with large error values. Therefore, we decided to include RMSE in our analysis along with MSE. Again, Cross Validation confirms that the best model is Model 3 with the lowest MSE and RMSE.

Model 1- **MSE:** 37154801.67 ; **RMSE**: 6095.47

Model 2- **MSE:** 37450507.38 **; RMSE**: 6119.68

Model 3- **MSE:** 24333022.64 **; RMSE**: 4932.85

Model 4- **MSE:** 37447756.43 **; RMSE**: 6119.46

Regression Tree

First, we will start with cv.tree function to perform 10-fold cross validation to find the best subtree. After running the tree function and running the summary function after it we see not all the predictors are used from the original data set, the only variables that are used in the tree construction are: smoker, age and bmi. Moreover, we can see that the number of terminal nodes is 6, and the mean of squared regression tree is 5265.331. The tree in figure 3 shows that for a non-smoker who is less than 44.5 years old charges are 5319 while for nonsmokers above 44.5 years old charges are 12510. Next for a smoker whose bmi is less than 30.1 and that are younger than 39.5 charges are 18030 and those who are older than 39.5 get charged 25680. Furthermore, smokers who have a bmi higher than 30.1 and are older than younger than 41.5 charges are 36420 and for those who are older than 41.5 charges are 46350.

Bagging

Next, we used Bagging, which is variation of Random Forest that uses all predictors. Where we get the Mean of squared residuals: 20061469 with 85.74 % of the variation explained by the bagged regression tree.

After, we evaluated the performance of bagging by fitting it to the testing dataset (“insurance [-train]”) and then we calculated RMSE which is 5122.742. Going back to the first linear regression model that we created earlier which had an RMSE of 6095.47 we see that bagging has actually lowered the RMSE significantly. Therefore, we can claim that the predictive performance of the tree has improved substantially.

Random Forest

The next step is using the Random Forest, where we will use only three predictors mtry=3, and the importance=TRUE which indicates that the importance of predictors is assessed. The RMSE of random forest is: 4983.785, and each variable is ranked by importance. This value has improved if we consider the bagging method. However, we must remember that while bagging uses all predictors to calculate the RMSE Random Forest uses the square root of predictors which in our case is √6 = 2.45 or 3 random predictors at each split. The generated tree type is regression and there are 500 trees generated. Figure 4 shows that smoker has the highest importance with 160.997, and sex has lowest importance variable with -0.904632. This confirms our earlier assessment of important predictor variables in the pairwise scatterplots where we identified smoker as a variable with a strong effect in the response variable.

Conclusion

Finally, we can conclude that the lowest RSME is that of linear regression Model 3, with RMSE of 4932.851 and following is random forest with an RMSE of 4983.785. After using these techniques to asses the predictability of the models we learned that there are many tests we need to take to evaluate and improve a model in order to get a clear understanding of how well the model is performing and which one to choose as a final model to base our predictions on. Based on our results we should use Random Forest to predict future charges since this model has the lowest RMSE that is generated after testing a large number of splits to improve the basic linear regression model. In the end we believe that those who find themselves unable to find affordable healthcare such as part time workers or freelancers should really consider quitting smoking if they are and if they have a large bmi they might consider to get in shape as well. This will benefit them in two ways it will significantly improve their quality of life as well as increase chances of finding affordable healthcare. Next, these models can also be used by insurance companies to analyze their customers and create adequate pricing strategies. Finally, lawmakers can also benefit from such models by using them to regulate health insurance prices in the market.

Appendix

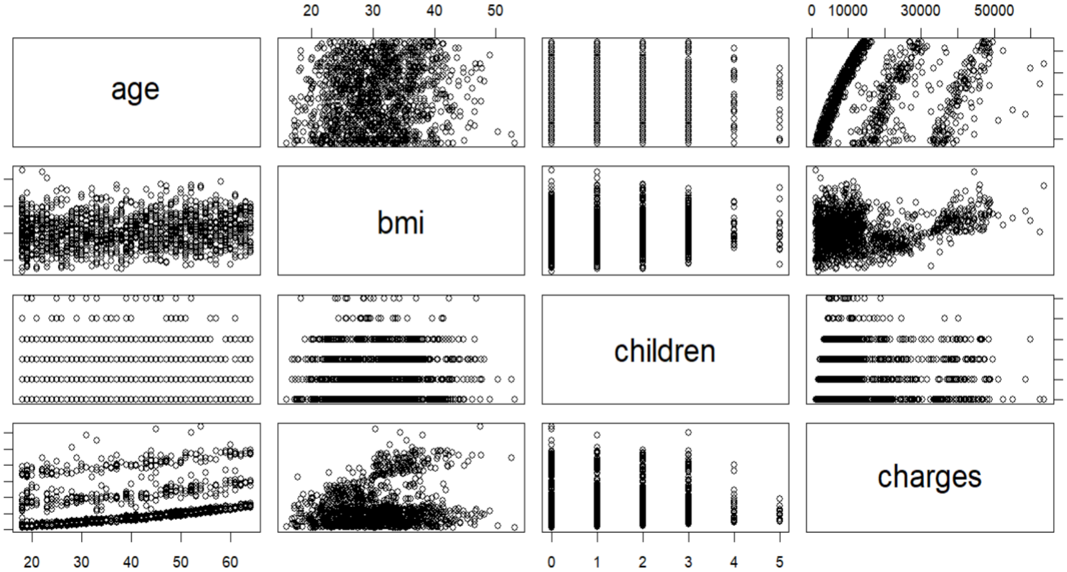

Figure 1. Scatterplot for variables: age, bmi, children and charges.

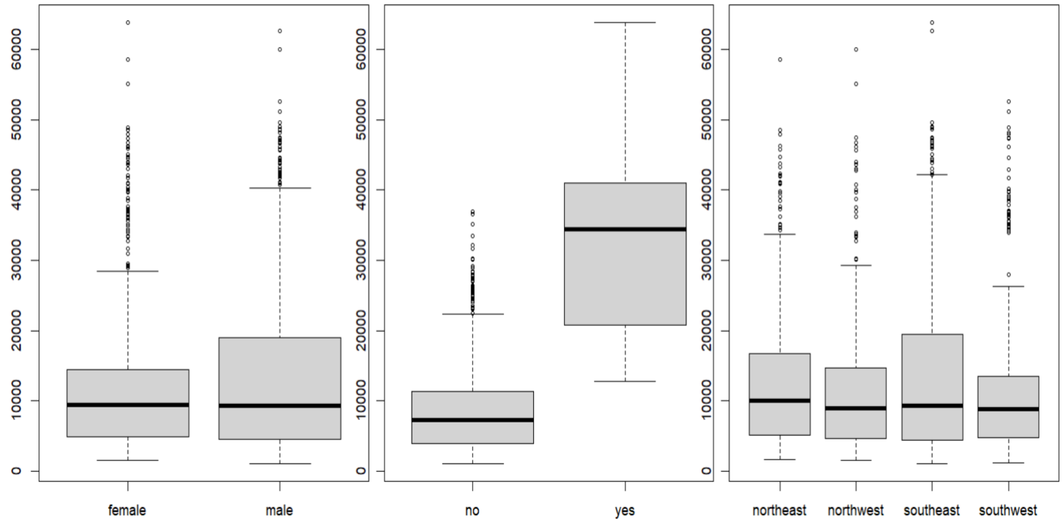

Figure 2. Box plot for three categorical variables: sex, smoker, and region.

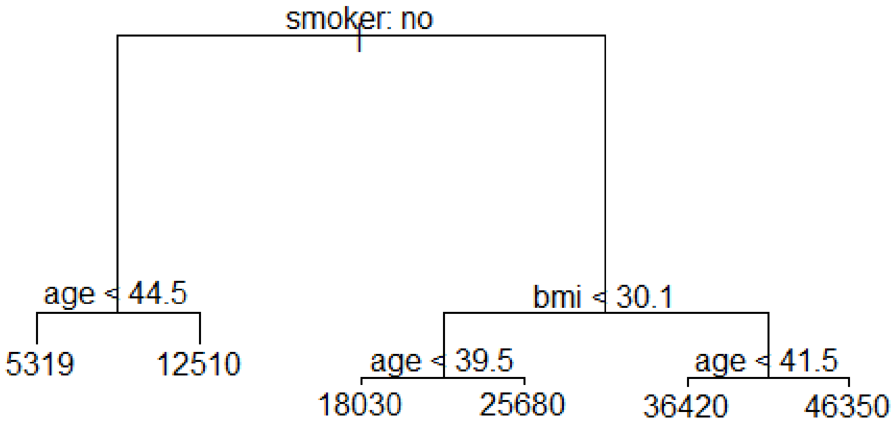

Figure 3. Regression tree

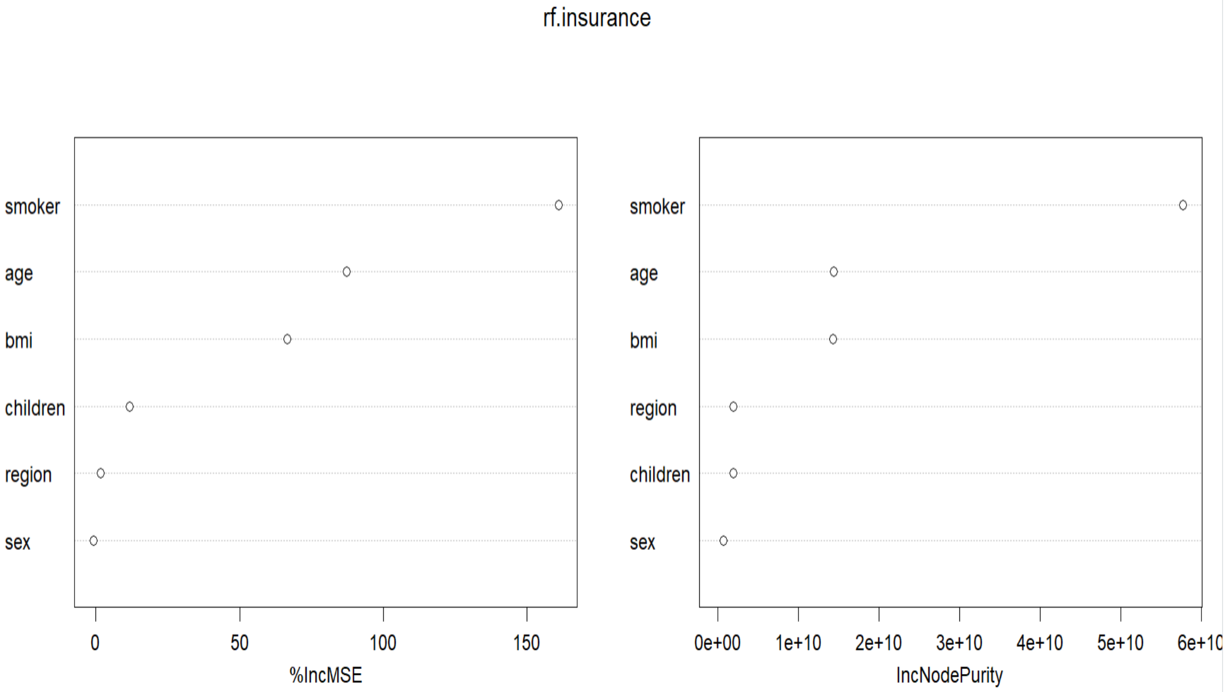

Figure 4. Random forest importance