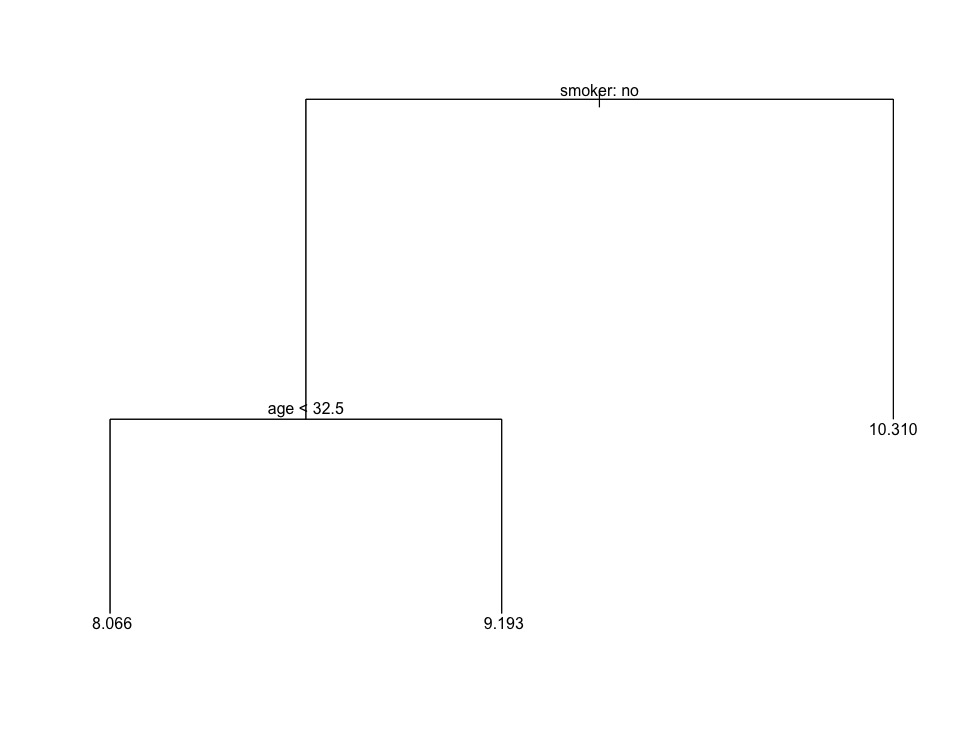
Herbert Borek  
8/25/2019  
Final Project  
BDAT 640  
Section: 01W  
Instructor: Chris Shannon  
File Name: FinalProject\_Borek\_Herbert.docx

Answers

1. Data Preparation

a-f. See Code

1. Build a multiple linear regression model.
   1. Perform multiple linear regression with **charges** as the response and the predictors are **age**, **sex**, **bmi**, **children**, **smoker**, and **region**. Print out the results using the summary() function. **Use the training data set created in step 1.e to train your model.**
      1. > summary(fit)
      2. Call:
      3. lm(formula = charges ~ age + sex + bmi + children + smoker +
      4. region, data = train)
      5. Residuals:
      6. Min 1Q Median 3Q Max
      7. -0.97392 -0.18835 -0.04645 0.05153 2.18141
      8. Coefficients:
      9. Estimate Std. Error t value Pr(>|t|)
      10. (Intercept) 7.024086 0.085780 81.885 < 2e-16 \*\*\*
      11. age 0.035794 0.001034 34.609 < 2e-16 \*\*\*
      12. sexmale -0.091494 0.028984 -3.157 0.001650 \*\*
      13. bmi 0.012142 0.002489 4.878 1.27e-06 \*\*\*
      14. children 0.100402 0.012339 8.137 1.37e-15 \*\*\*
      15. smokeryes 1.597208 0.035783 44.636 < 2e-16 \*\*\*
      16. regionnorthwest -0.044722 0.041966 -1.066 0.286863
      17. regionsoutheast -0.155460 0.042026 -3.699 0.000230 \*\*\*
      18. regionsouthwest -0.149159 0.041953 -3.555 0.000397 \*\*\*
      19. ---
      20. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
      21. Residual standard error: 0.4299 on 883 degrees of freedom
      22. Multiple R-squared: 0.7865, Adjusted R-squared: 0.7846
      23. F-statistic: 406.6 on 8 and 883 DF, p-value: < 2.2e-16
   2. Is there a relationship between the predictors and the response?
      1. Yes. The p-value is less than the cutoff value of .05.
   3. Does **sex** have a statistically significant relationship to the response?
      1. Yes. The p-value for that particular variable is less than .05.
   4. Perform best subset selection using the stepAIC() function from the MASS library, choose best model based on AIC. For the "direction" parameter in the stepAIC() method, set direciton="backward"
      1. > summary(fit.best)
      2. Call:
      3. lm(formula = charges ~ age + sex + bmi + children + smoker +
      4. region, data = train)
      5. Residuals:
      6. Min 1Q Median 3Q Max
      7. -0.97392 -0.18835 -0.04645 0.05153 2.18141
      8. Coefficients:
      9. Estimate Std. Error t value Pr(>|t|)
      10. (Intercept) 7.024086 0.085780 81.885 < 2e-16 \*\*\*
      11. age 0.035794 0.001034 34.609 < 2e-16 \*\*\*
      12. sexmale -0.091494 0.028984 -3.157 0.001650 \*\*
      13. bmi 0.012142 0.002489 4.878 1.27e-06 \*\*\*
      14. children 0.100402 0.012339 8.137 1.37e-15 \*\*\*
      15. smokeryes 1.597208 0.035783 44.636 < 2e-16 \*\*\*
      16. regionnorthwest -0.044722 0.041966 -1.066 0.286863
      17. regionsoutheast -0.155460 0.042026 -3.699 0.000230 \*\*\*
      18. regionsouthwest -0.149159 0.041953 -3.555 0.000397 \*\*\*
      19. ---
      20. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
      21. Residual standard error: 0.4299 on 883 degrees of freedom
      22. Multiple R-squared: 0.7865, Adjusted R-squared: 0.7846
      23. F-statistic: 406.6 on 8 and 883 DF, p-value: < 2.2e-16
   5. Compute the test error of the best model in #3d based on AIC using LOOCV using trainControl() and train() from the caret library. Report the MSE by squaring the reported RMSE.
      1. The MSE is: 0.1988249
   6. Calculate the test error of the best model in #3d based on AIC using 10-fold Cross-Validation. Use train and trainControl from the caret library. Refer to model selected in #3d based on AIC. Report the MSE.
      1. The MSE is: 0.1979252
   7. Calculate and report the test MSE using the best model from 2.d and test data set created in step 1.e.
      1. The MSE is: 0.2249572
   8. Compare the test MSE calculated in step 2.f using 10-fold cross-validation with the test MSE calculated in step 2.g. How similar are they?
      1. The MSE calculated in step 2.f using 10-fold cross-validation is 0.1979252, while the one calculated in step 2.g is 0.2249572. The best model MSE has a greater MSE by 0.02703197.
2. Build a regression tree model.
   1. Build a regression tree model using function tree(), where **charges** is the response and the predictors are **age**, **sex**, **bmi**, **children**, **smoker**, and **region**.
      1. See Code
   2. Find the optimal tree by using cross-validation and display the results in a graphic. Report the best size.
      1. A close up of a map

         Description automatically generated
      2. The best size is three.
   3. Justify the number you picked for the optimal tree with regard to the principle of variance-bias trade-off.
      1. The way I balanced bias and variance was by picking a number of nodes that had a low deviance, while keeping the complexity of the model on the lower side. I picked the model by using the prior visualization.
   4. Prune the tree using the optimal size found in 3.b
      1. See Code
   5. Plot the best tree model and give labels.
      1. 
   6. Calculate the test MSE for the best model.
      1. The test MSE is 0.2776814.
3. Build a random forest model.
   1. Build a random forest model using function randomForest(), where **charges** is the response and the predictors are **age**, **sex**, **bmi**, **children**, **smoker**, and **region**.
      1. See Code
   2. Compute the test error using the test data set.
      1. The test error is 0.1656417.
   3. Extract variable importance measure using the importance() function.
      1. %IncMSE IncNodePurity
      2. age 109.259617 243.402854
      3. sex 4.102053 6.986571
      4. bmi 21.213321 62.102844
      5. children 26.551903 28.974102
      6. smoker 145.993369 328.855393
      7. region 6.467242 17.073537
   4. Plot the variable importance using the function, varImpPlot(). Which are the top 3 important predictors in this model?
      1. A screenshot of a cell phone

         Description automatically generated
      2. The top three predictors in this model are smoker, age, and children, though children does not have quite as big an MSE as the other two.
4. Build a support vector machine model
   1. The response is **charges** and the predictors are **age**, **sex**, **bmi**, **children**, **smoker**, and **region**. Please use the svm() function with radial kernel and gamma=5 and cost = 50.
      1. > summary(fit.svm)
      2. Call:
      3. svm(formula = f, data = train, kernel = "radial", gamma = 5, cost = 50)
      4. Parameters:
      5. SVM-Type: eps-regression
      6. SVM-Kernel: radial
      7. cost: 50
      8. gamma: 5
      9. epsilon: 0.1
      10. Number of Support Vectors: 722
   2. Perform a grid search to find the best model with potential cost: 1, 10, 50, 100 and potential gamma: 1,3 and 5 and potential kernel: "linear","radial" and "sigmoid". And use the training set created in step 1.e.
      1. See Code
   3. Print out the model results. What are the best model parameters?
      1. > summary(insurance.tune)
      2. Parameter tuning of ‘svm’:
      3. - sampling method: 10-fold cross validation
      4. - best parameters:
      5. cost gamma kernel
      6. 1 1 radial
      7. - best performance: 0.1863861
      8. - Detailed performance results:
      9. cost gamma kernel error dispersion
      10. 1 1 1 linear 1.999842e-01 3.957645e-02
      11. 2 10 1 linear 1.999914e-01 3.957732e-02
      12. 3 50 1 linear 1.999856e-01 3.955185e-02
      13. 4 100 1 linear 2.000457e-01 3.958024e-02
      14. 5 1 3 linear 1.999842e-01 3.957645e-02
      15. 6 10 3 linear 1.999914e-01 3.957732e-02
      16. 7 50 3 linear 1.999856e-01 3.955185e-02
      17. 8 100 3 linear 2.000457e-01 3.958024e-02
      18. 9 1 5 linear 1.999842e-01 3.957645e-02
      19. 10 10 5 linear 1.999914e-01 3.957732e-02
      20. 11 50 5 linear 1.999856e-01 3.955185e-02
      21. 12 100 5 linear 2.000457e-01 3.958024e-02
      22. 13 1 1 radial 1.863861e-01 5.069635e-02
      23. 14 10 1 radial 2.519722e-01 6.881172e-02
      24. 15 50 1 radial 3.843507e-01 1.033024e-01
      25. 16 100 1 radial 4.989574e-01 1.714084e-01
      26. 17 1 3 radial 3.747778e-01 5.947665e-02
      27. 18 10 3 radial 4.032226e-01 6.763976e-02
      28. 19 50 3 radial 4.513373e-01 9.300087e-02
      29. 20 100 3 radial 4.640181e-01 1.011458e-01
      30. 21 1 5 radial 4.786973e-01 6.944814e-02
      31. 22 10 5 radial 4.761719e-01 6.608535e-02
      32. 23 50 5 radial 4.851921e-01 6.946078e-02
      33. 24 100 5 radial 4.890745e-01 6.756874e-02
      34. 25 1 1 sigmoid 2.270118e+03 3.899444e+02
      35. 26 10 1 sigmoid 2.199598e+05 3.743322e+04
      36. 27 50 1 sigmoid 5.797503e+06 1.084097e+06
      37. 28 100 1 sigmoid 2.412287e+07 3.996210e+06
      38. 29 1 3 sigmoid 4.359057e+03 4.641865e+02
      39. 30 10 3 sigmoid 4.379049e+05 5.642319e+04
      40. 31 50 3 sigmoid 1.123735e+07 1.197997e+06
      41. 32 100 3 sigmoid 4.475338e+07 4.601095e+06
      42. 33 1 5 sigmoid 4.669617e+03 4.234287e+02
      43. 34 10 5 sigmoid 4.725087e+05 6.851653e+04
      44. 35 50 5 sigmoid 1.186389e+07 1.297003e+06
      45. 36 100 5 sigmoid 4.774610e+07 7.010079e+06
      46. The best model parameters are with cost=1, gamma=1, and kernel=radial
   4. Forecast **charges** using the test dataset and the best model found in c).
      1. See Code
   5. Compute the MSE (Mean Squared Error) on the test data.
      1. The MSE is 0.226569
5. Perform the k-means cluster analysis.
   1. Use the training data set created in step 1.f and standardize the inputs using the scale() function.
      1. See Code
   2. Convert the standardized inputs to a data frame using the as.data.frame() function.
      1. See Code
   3. Determine the optimal number of clusters, and use the gap\_stat method and set iter.max=20. Justify your answer. It may take longer running time since it uses a large dataset.
      1. A close up of a map

         Description automatically generated
      2. The optimal number of clusters is three, as that is what the graph chose.
   4. Perform k-means clustering using the optimal number of clusters found in step 6.c. Set parameter nstart = 25
      1. See Code
   5. Visualize the clusters in different colors, setting parameter geom="point"
      1. A close up of a map

         Description automatically generated
6. Build a neural networks model.
   1. Using the training data set created in step 1.f, create a neural network model where the response is **charges** and the predictors are **age**, **sexmale**, **bmi**, **children**, **smokeryes**, **regionnorthwest**, **regionsoutheast**, and **regionsouthwest**. Please use 1 hidden layer with 1 neuron. **Do not scale the data.**
      1. See Code
   2. Plot the neural networks.
      1. A close up of a map

         Description automatically generated
   3. Forecast the **charges** in the test dataset.
      1. See Code
   4. Compute test error (MSE).
      1. The MSE is 0.821267.
7. Putting it all together.
   1. For predicting insurance charges, your supervisor asks you to choose the best model among the multiple regression, regression tree, random forest, support vector machine, and neural network models. Compare the test MSEs of the models generated in steps 2.g, 3.f, 4.b, 5.e, and 7.d. Display the names for these types of these models, using these labels: **Multiple Linear Regression**, **Regression Tree**, **Random Forest**, **Support Vector Machine**, and **Neural Network** and their corresponding test MSEs in a data.frame. Label the column in your data frame with the labels as **Model.Type**, and label the column with the test MSEs as **Test.MSE** and round the data in this column to 4 decimal places. Present the formatted data to your supervisor and recommend which model is best and why.
      1. Model.Type Test.MSE
      2. 1 Multiple Linear Regression 0.2250
      3. 2 Regression Tree 0.2777
      4. 3 Random Forest 0.1656
      5. 4 Support Vector Machine 0.2266
      6. 5 Neural Network 0.8213
      7. The model with the lowest Mean Squared Error (and therefore the best model) is the random forest with an MSE of 0.1656. The one with the highest MSE is the neural network, with an MSE of 0.8213.
   2. Another supervisor from the sales department has requested your help to create a predictive model that his sales representatives can use to explain to clients what the potential costs could be for different kinds of customers, and they need an easy and visual way of explaining it. What model would you recommend, and what are the benefits and disadvantages of your recommended model compared to other models?
      1. The model I would recommend would be the regression tree, as it shows the prediction based upon a couple of variables in a very visual way and in a way that would be easy to understand for a layperson who does not know anything about neural networks or support vector machines.
      2. The disadvantages would be that it does have a tendency to oversimplify results
   3. The supervisor from the sales department likes your regression tree model. But she says that the sales people say the numbers in it are way too low and suggests that maybe the numbers on the leaf nodes predicting charges are log transformations of the actual charges. You realize that in step 1.b of this project that you had indeed transformed charges using the log function. And now you realize that you need to reverse the transformation in your final output. The solution you have is to reverse the log transformation of the variables in the regression tree model you created and redisplay the result.   
      Follow these steps:
      1. Copy your pruned tree model to a new variable.
         1. See Code
      2. In your new variable, find the data.frame named "frame" and reverse the log transformation on the data.frame column **yval** using the **exp()** function. (If the copy of your pruned tree model is named **copy\_of\_my\_pruned\_tree**, then the data frame is accessed as **copy\_of\_my\_pruned\_tree$frame**, and it works just like a normal data frame.).
         1. See Code
      3. After you reverse the log transform on the **yval** column, then replot the tree with labels.
         1. A picture containing screenshot

            Description automatically generated