# *FACTORS IN HEALTHCARE COVERAGE*

Project 1

*Lantz, Emily*

*COSC 6520 | Data Analytics*

## Section 1: Business Need and Importance

The average person spends 41% of their income on health insurance and healthcare annually. However, what factors contribute most to the cost of health insurance? The average person spends $13,000 a year on their health insurance. A public dataset analyzes different factors that impact health insurance costs, including age[i], sex[ii], BMI[iii], smoker[iv], number of children[v], and region[vi]. Names and specific locations are excluded to protect people's privacy and HIPPA. Suppose an individual knows the factors that most impact the cost of health insurance. In that case, they may be able to change factors of their lives to lower their cost, specifically if they will be above or below the national average in the United States. When it comes to individual health insurance plans, there are four primary levels: bronze, silver, gold, and platinum. Annually, bronze costs the least, while platinum expenses the most. Furthermore, insurance companies can better analyze their costs and plan to offer competitive prices, given certain factors on a person. Along with personal empowerment when it comes to picking health insurance, health insurance providers can also give their customers fast predictions on their health insurance tiers and costs based on a few limited factors, allowing for more efficiency and customer bandwidth.

## Section 2: Statistical Methodology

The first supervised data mining technique used was a classification tree. In the dataset, everyone is given an associated annual insurance charge. A classification tree was chosen because it can handle both numerical and categorical variables and the simplicity. According to Forbes[vii], the costs for each tier of insurance are Bronze—$10,680 or less, Silver costs between $10,680 and $13,968, Gold costs between $13,968-$18,180, and Platinum is anything higher than $18,180. With this information, charges can be binned into the four different tiers with the corresponding labels. Charges are also changed into a factor variable to ensure the correct progression of the classification tree. The target variable is the "charges" to determine which tier of health insurance someone would have. Additionally, as a categorical variable, region is changed to a factor variable with four different factors: northeast, northwest, southeast, and southwest. Smoker is changed from "yes" and "no" to "1" and "0", respectively. Using the "rpart"

package in R, the data is classified into a training dataset, 70% of the data, and the testing dataset, 30% of the data, using recursive partitioning. For the full tree, there was a complexity variable of 0. For the pruned tree, there was a complexity variable of 0. 0.00364078. With the packages "gains" and "pROC," the classification tree can be evaluated using acne metrics such as accuracy, specificity, and sensitivity.

The second supervised data mining technique used was K Nearest Neighbor, which was used to predict whether a person's annual insurance cost will be above or below the national average. Since the dataset is nonparametric and relies on numerical data, the simplicity of KKN made it clear that the KNN would be helpful in this analysis; the R packages "caret" was used in this technique to find the optimal number of k through cross-validation. Data management in KKN involved creating bins and factors and scaling the data. In this scenario, charges were binned into two separate categories. 1 if the charges are above $13,000 or 0 if it was below $13,000[viii]. Additionally, KKN works better when variables are scaled because it allows for all variables to be equally important, along with more accurate distance measurements. The categories of age, sex, BMI, children, and smoker are scaled, using R's "Scale" function to normalize the data. The region variable was excluded from this model because it is a categorical variable and transforming it into four separate bins and normalizing it would unfairly weigh the data. On the other hand, sex and smoker were converted into numerical data, 0 and 1, then scaled to complete the KKN model. The dataset was then divided into the training dataset, 80% of the data, and the testing dataset, 20% of the data. With the packages "gains" and "pROC," KNN can be evaluated using accuracy, specificity, and sensitivity metrics.

## Section 3: Results and Interpretation

In the first default tree[ix], the training set has 938 obversions. This model is intentionally not pruned to find the most significant variables. However, this does lead to a high relative error rate and cross-validation error rate of 1.0. Looking further into the summary, we can see that the model can be pruned more as the complexity parameter decreases, leading to a lower cross validation error. Additionally, age, smoking, children, and BMI are the most critical variables that impact "charges," with scores of 47, 46, 3, and 3, respectively. In total, there are 41 nodes in the default tree. Four numbers are below each node in the tree, which describes the number of observations in each tier. For example, node 11 has 0 observations in Bronze, 6 in Silver, 17 in Gold, and 1 in Platinum.

The full tree[x] has a total of 6,9067 nodes, and it is defined as a full tree by the complexity parameter being at 0. Additionally, the parent note is characterized by having at least two but can be further split with at least 1 observation per leaf node. The full tree will help us understand the best complexity parameter for the pruned tree. After creating the total tree, the complexity parameter table[xi] is printed, which we can see after the 7th split, the cross-validation record increases from 0.025349 to 0.025434. Thus, a pruned tree[xii] is created using the complexity parameter of 0.0036407 and the full tree. Once more, the tree comprises 938 observations, with a total of 83 nodes. The variables of importance are age, smoking, children, BMI, and region, with scores of 46, 43, 5, 4, and 1.

The classification tree can use a confusion matrix[xiii] to predict a person's health insurance tier with 87.25% accuracy. Moreover, looking at the matrix by class, Bronze and Gold are the easiest to correctly classify, and Gold is the most difficult. The gains table[xiv] also confirms that the classification model performs better than random selection because the Lift Index and Cumulative Lift are both high, particularly at the start of the dataset. Finally, the multiclass ROC[xv] score is 0.7859, which is 0.2859 better than the average. In total there was six rocs produced by the multiclass ROC function. The cumulative lift chart[xvi] also support the strength of the model.

With all this information from the classification tree, people are better than random at accurately predicting the level of health insurance they should purchase or the level that will be offered to them from a health insurance company. Moreover, they can see what will directly impact them, making their tiers either increase or decrease depending on the flow of the classification tree.

Next is the K-Nearest Neighbor supervised model, which used "K"s from 1-100 in the first KNN fit model, using cross-validation. With all six variables transformed to numeric and scaled, the optimal model found was k=5 with an accuracy of 89.96%. Using the KNN model and the validation set, it is predicted that the KNN[xvii] will predict with 89.14% whether a person will be above or below the national average health insurance costs[xviii]. Additionally, 72.41% accuracy was used to predict product-positive cases, and 97.22% predicted negative cases. Next, a probability prediction variable is created based on the KNN fit model. With the confusion matrix[xix] for probability, there is 83.52% accuracy that the probability of being over the national average will be the prediction and outcome for an observation. Additionally, 80.46 % accuracy was achieved in predicting product positive cases, and 85.00 % predicted negative cases. Similar to the classification tree, the KKN gain chart[xx], cumulative lift chart[xxi], and decile-wise lift chart[xxii] demonstrate that the model is better than random assignment.

With all this information from the KKN model, given the factors in the dataset, excluding location, people can generally be grouped to determine whether they would be above or below the national average. Insurance companies can use this information quickly for their clients, with minimal information, to give them a general estimate of their annual health insurance costs. Finally, the AUC score from the ROC curve[xxiii] is 88.75% which is 38.75% better than random selection.

## Section 4: Alternative Approaches

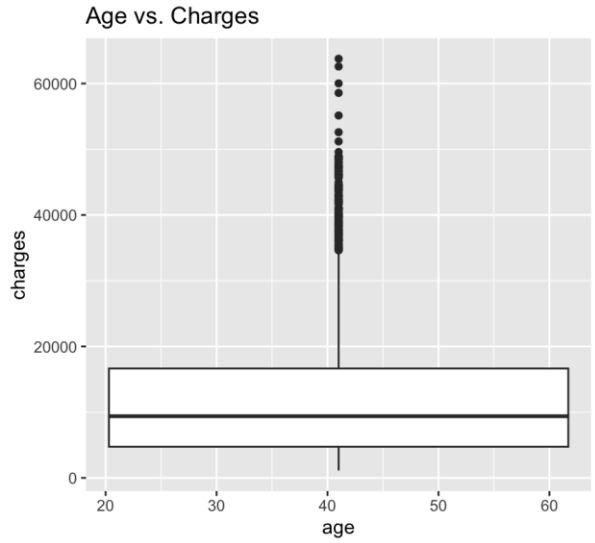
In this approach of using KKN and classification trees, it comes down to the simplicity of the models. Using these models would allow for faster analysis for insurance companies. Thus, they can serve a larger amount of consumers with accurate predictions of cost and health insurance tier. Using other methods, such as Naïve Bayes, would not allow for both categorical and numerical data to be analyzed and assumes independence. With health information, data is clearly not independent such as age and BMI.

## Section 5: Conclusions

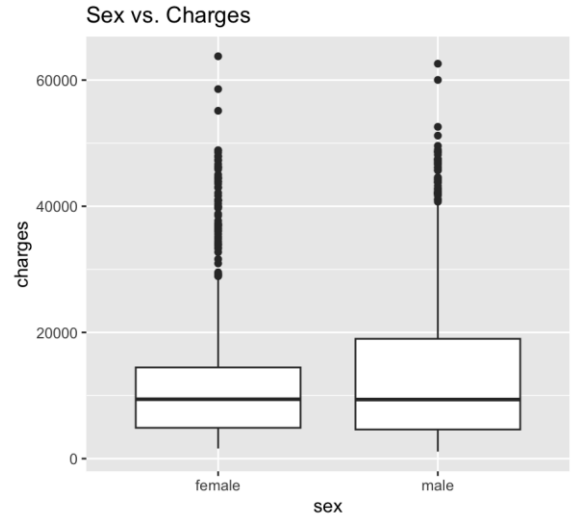
In today's world, companies continually pride themselves on their efficiency and privacy protections. When it comes to health insurance, people must provide very personal information to get an accurate prediction of their healthcare costs. This dataset shows that companies and people only need a few factors to make accurate predictions on their healthcare prices and health insurance tiers. With limited factors, people's information will stay private. The impact of these findings will allow companies to make quicker decisions about the type of plan to offer a customer. Furthermore, it will enable people to make empowered decisions to save money on health insurance. They can change factors about their lives, such as deciding the number of children to have, whether to smoke, and their BMI score. Efficiency, privacy, and empowerment will allow people to take control of their healthcare coverage in their lives.

# Appendix

---

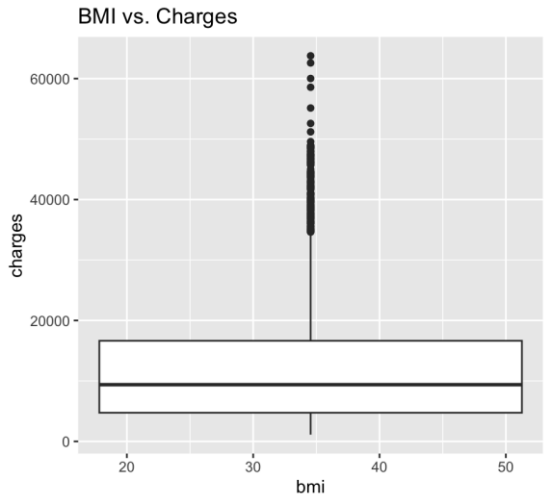



i

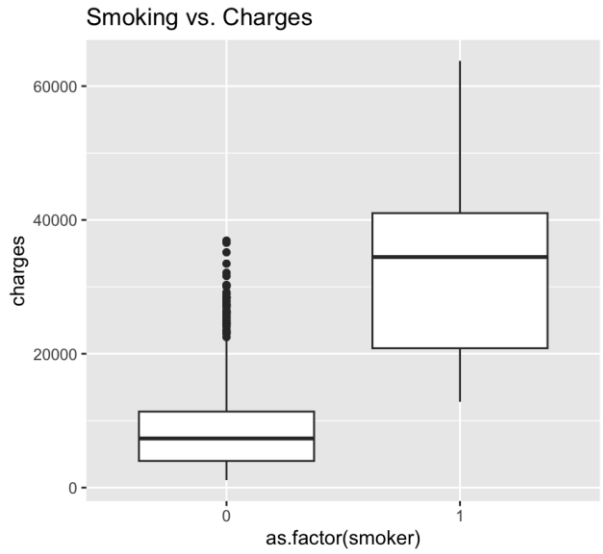


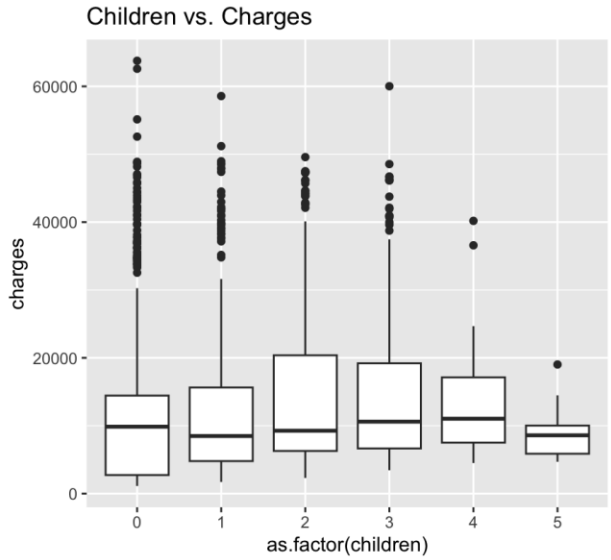

ii

iii

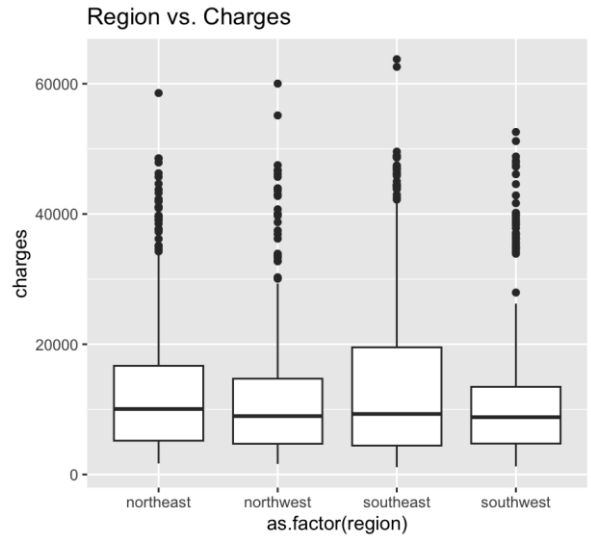

BMI vs. Charges
charges
60000
40000
20000
0
20
30
40
50
bmi

iv


Smoking vs. Charges
charges
60000
40000
20000
0
0
1
as.factor(smoker)



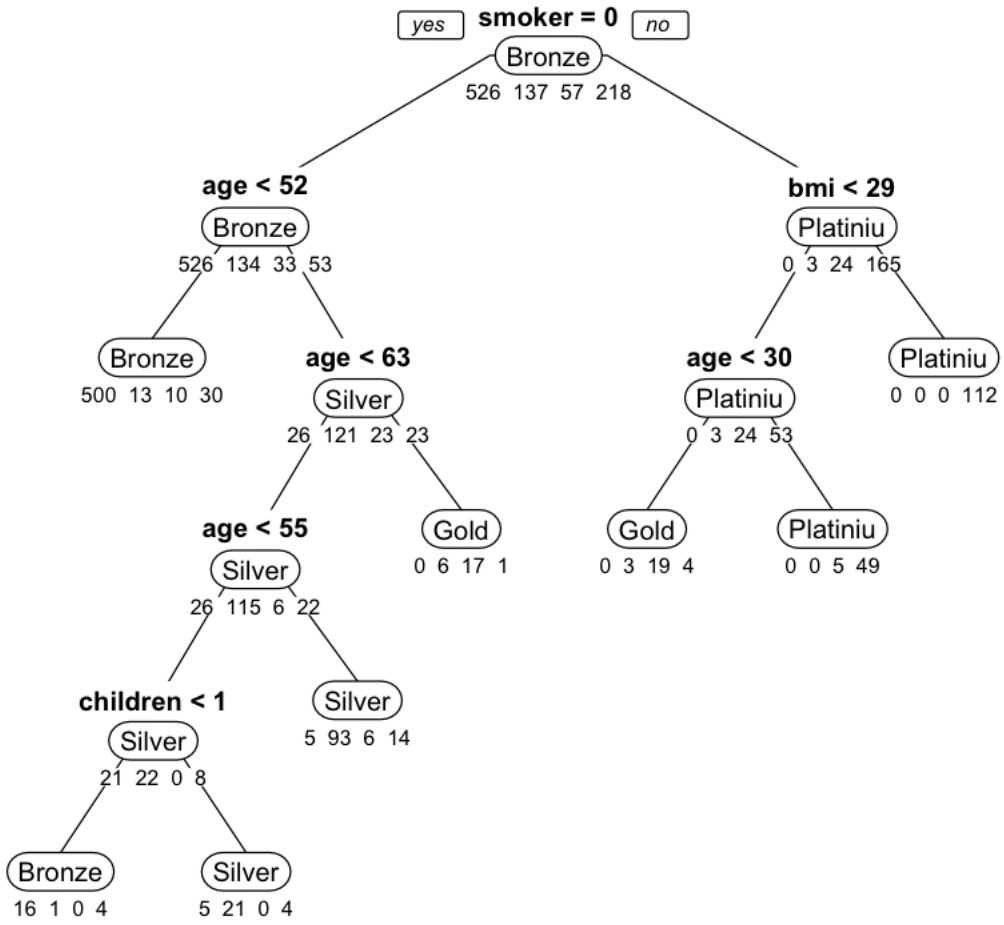

[v]

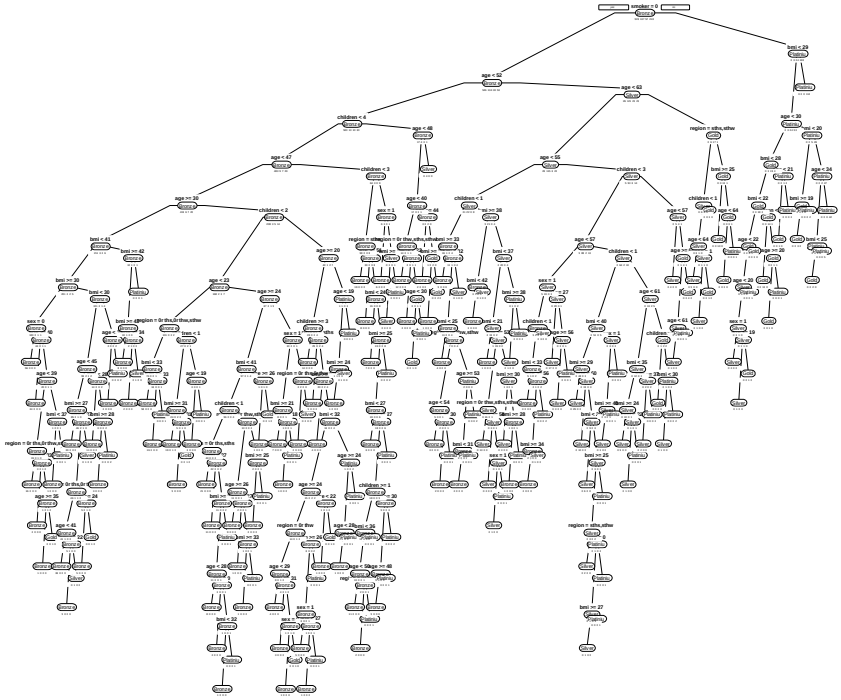



[vi]

[vii] https://www.forbes.com/advisor/health-insurance/how-much-does-health-insurance-cost/

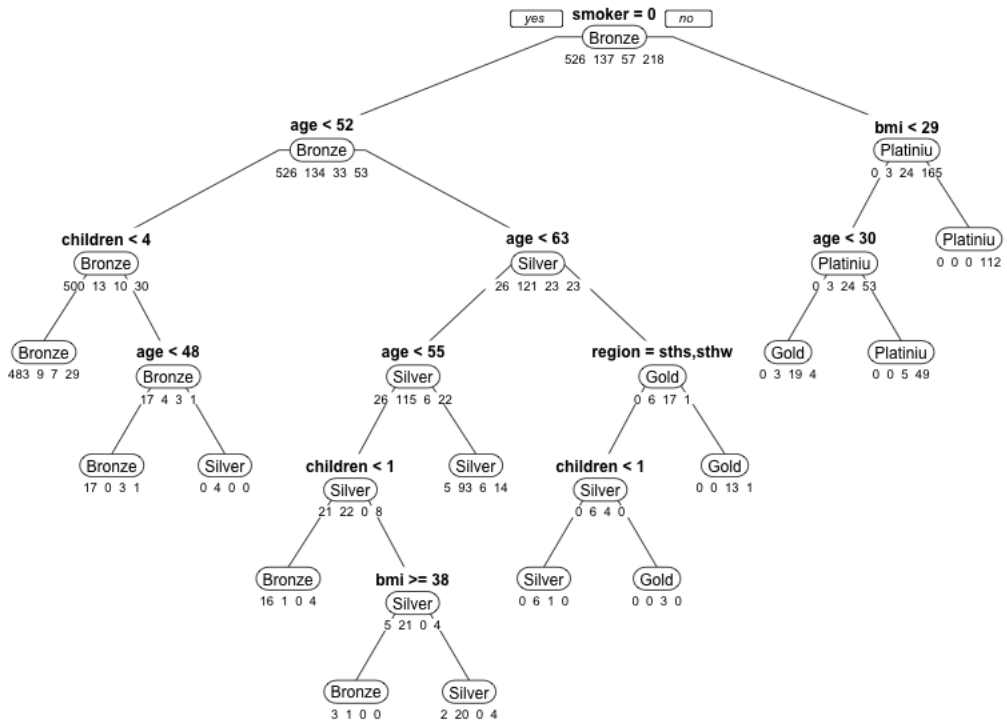
[viii] https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical


yes
smoker = 0
no
Bronze
526 137 57 218
age < 52
Bronze
526 134 33 53
bmi < 29
Platiniu
0 3 24 165
Bronze
500 13 10 30
age < 63
Silver
26 121 23 23
age < 30
Platiniu
0 3 24 53
Platiniu
0 0 0 112
age < 55
Silver
26 115 6 22
Gold
0 6 17 1
Gold
0 3 19 4
Platiniu
0 0 5 49
children < 1
Silver
21 22 0 8
Silver
5 93 6 14
Bronze
16 1 0 4
Silver
5 21 0 4



x

```
           CP nsplit rel error  xerror     xstd
1  0.40048544      0 1.0000000 1.00000 0.036893
2  0.23058252      1 0.5995146 0.59951 0.032741
3  0.02669903      2 0.3689320 0.37621 0.027609
4  0.01820388      3 0.3422330 0.37864 0.027680
5  0.00606796      7 0.2694175 0.32282 0.025931
6  0.00485437      9 0.2572816 0.31068 0.025518
7  0.00364078     12 0.2427184 0.30583 0.025349
8  0.00323625     16 0.2281553 0.30825 0.025434
9  0.00242718     19 0.2184466 0.30340 0.025264
10 0.00161812     45 0.1553398 0.30340 0.025264
11 0.00145631     62 0.1262136 0.34223 0.026567
12 0.00121359     70 0.1140777 0.42718 0.029023
13 0.00104022    118 0.0485437 0.44175 0.029397
14 0.00097087    125 0.0412621 0.45388 0.029699
15 0.00080906    132 0.0339806 0.46359 0.029935
16 0.00060680    147 0.0218447 0.46359 0.029935
17 0.00000000    169 0.0024272 0.46602 0.029993
```

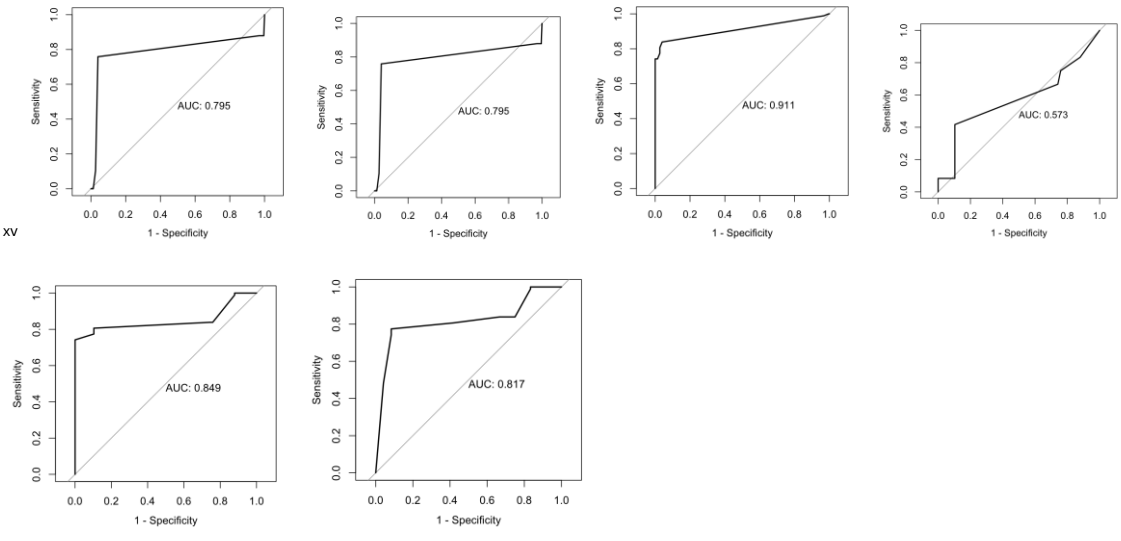

yes
smoker = 0
no
Bronze
526 137 57 218
age < 52
Bronze
526 134 33 53
bmi < 29
Platiniu
0 3 24 165
children < 4
Bronze
500 13 10 30
age < 63
Silver
26 121 23 23
age < 30
Platiniu
0 3 24 53
Platiniu
0 0 0 112
Bronze
483 9 7 29
age < 48
Bronze
17 4 3 1
age < 55
Silver
26 115 6 22
region = sths,sthw
Gold
0 6 17 1
Gold
0 3 19 4
Platiniu
0 0 5 49
Bronze
17 0 3 1
Silver
0 4 0 0
children < 1
Silver
21 22 0 8
Silver
5 93 6 14
children < 1
Silver
0 6 4 0
Gold
0 0 13 1
Bronze
16 1 0 4
bmi >= 38
Silver
5 21 0 4
Silver
0 6 1 0
Gold
0 0 3 0
Bronze
3 1 0 0
Silver
2 20 0 4

```
Confusion Matrix and Statistics

          Reference
Prediction  Bronze Silver Gold Platinium
  Bronze       219      9    2        15
  Silver         6     48    7         6
  Gold           0      1   13         3
  Platinium      0      0    2        69

Overall Statistics

               Accuracy : 0.8725
                 95% CI : (0.8358, 0.9036)
    No Information Rate : 0.5625
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7829

 Mcnemar's Test P-Value : 8.249e-05

Statistics by Class:

                     Class: Bronze Class: Silver Class: Gold Class: Platinium
Sensitivity                 0.9733        0.8276      0.5417          0.7419
Specificity                 0.8514        0.9444      0.9894          0.9935
Pos Pred Value              0.8939        0.7164      0.7647          0.9718
Neg Pred Value              0.9613        0.9700      0.9713          0.9271
Prevalence                  0.5625        0.1450      0.0600          0.2325
Detection Rate              0.5475        0.1200      0.0325          0.1725
Detection Prevalence        0.6125        0.1675      0.0425          0.1775
Balanced Accuracy           0.9124        0.8860      0.7655          0.8677
```

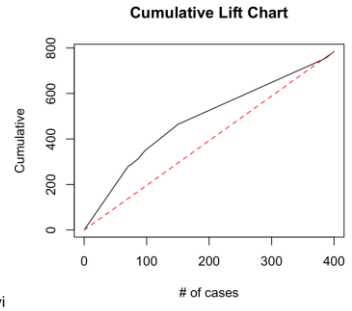
xiii

| Depth of File | N | Cume N | Mean Resp | Cume Mean Resp | Cume Pct of Total Resp | Lift Index | Cume Lift | Mean Model Score |
|---|---|---|---|---|---|---|---|---|
| 12 | 46 | 46 | 3.98 | 3.98 | 23.3% | 203 | 203 | 1.00 |
| 18 | 25 | 71 | 3.96 | 3.97 | 35.9% | 202 | 202 | 0.91 |
| 18 | 3 | 74 | 1.00 | 3.85 | 36.3% | 51 | 196 | 0.19 |
| 22 | 12 | 86 | 2.25 | 3.63 | 39.7% | 115 | 185 | 0.15 |
| 24 | 11 | 97 | 3.27 | 3.59 | 44.3% | 167 | 183 | 0.15 |
| 36 | 49 | 146 | 2.18 | 3.12 | 58.0% | 111 | 159 | 0.12 |
| 37 | 3 | 149 | 2.67 | 3.11 | 59.0% | 136 | 158 | 0.07 |
| 95 | 232 | 381 | 1.23 | 1.96 | 95.3% | 63 | 100 | 0.05 |
| 97 | 7 | 388 | 1.43 | 1.95 | 96.6% | 73 | 100 | 0.05 |
| 100 | 12 | 400 | 2.25 | 1.96 | 100.0% | 115 | 100 | 0.00 |

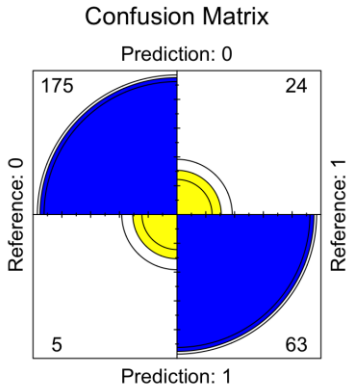
xiv




xv




xvi

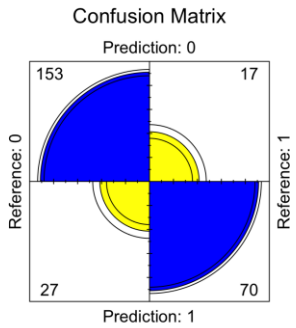



xvii

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 175  24
         1   5  63

               Accuracy : 0.8914
                 95% CI : (0.8477, 0.926)
    No Information Rate : 0.6742
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.738

 Mcnemar's Test P-Value : 0.0008302

            Sensitivity : 0.7241
            Specificity : 0.9722
         Pos Pred Value : 0.9265
         Neg Pred Value : 0.8794
             Prevalence : 0.3258
         Detection Rate : 0.2360
   Detection Prevalence : 0.2547
      Balanced Accuracy : 0.8482

       'Positive' Class : 1
```

xviii

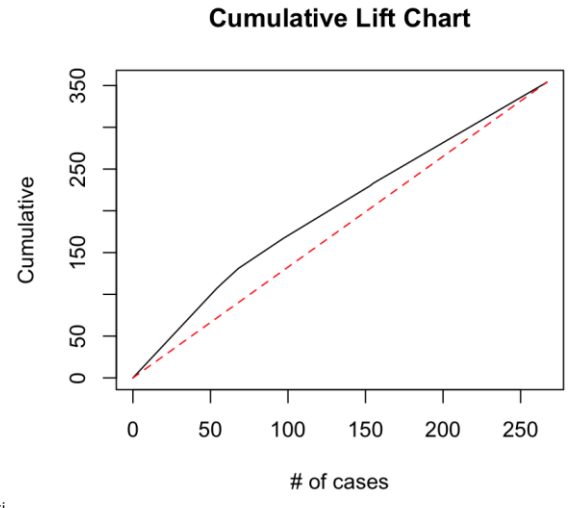
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 153  17
         1  27  70

               Accuracy : 0.8352
                 95% CI : (0.7852, 0.8776)
    No Information Rate : 0.6742
    P-Value [Acc > NIR] : 2.087e-09

                  Kappa : 0.6357

 Mcnemar's Test P-Value : 0.1748

            Sensitivity : 0.8046
            Specificity : 0.8500
         Pos Pred Value : 0.7216
         Neg Pred Value : 0.9000
             Prevalence : 0.3258
         Detection Rate : 0.2622
   Detection Prevalence : 0.3633
      Balanced Accuracy : 0.8273

       'Positive' Class : 1
```



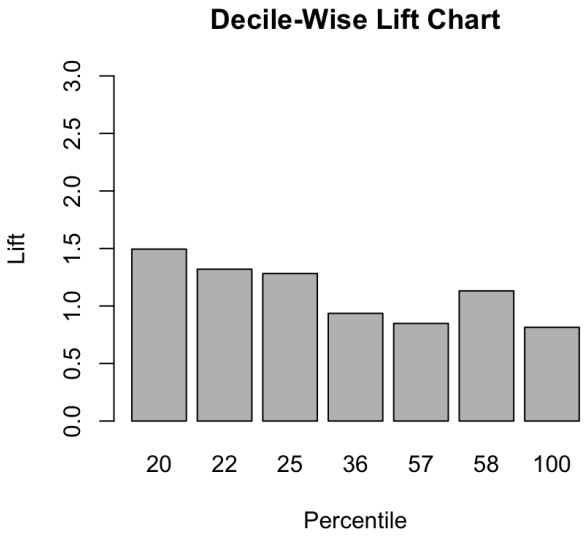

xix

| Depth of File | N | Cume N | Mean Resp | Cume Mean Resp | Cume Pct of Total Resp | Lift Index | Cume Lift | Mean Model Score |
|---|---|---|---|---|---|---|---|---|
| 20 | 54 | 54 | 1.98 | 1.98 | 30.2% | 149 | 149 | 1.00 |
| 22 | 4 | 58 | 1.75 | 1.97 | 32.2% | 132 | 148 | 0.80 |
| 25 | 10 | 68 | 1.70 | 1.93 | 37.0% | 128 | 145 | 0.60 |
| 36 | 29 | 97 | 1.24 | 1.72 | 47.2% | 94 | 130 | 0.40 |
| 57 | 56 | 153 | 1.12 | 1.50 | 65.0% | 85 | 113 | 0.20 |
| 58 | 2 | 155 | 1.50 | 1.50 | 65.8% | 113 | 113 | 0.17 |
| 100 | 112 | 267 | 1.08 | 1.33 | 100.0% | 81 | 100 | 0.00 |

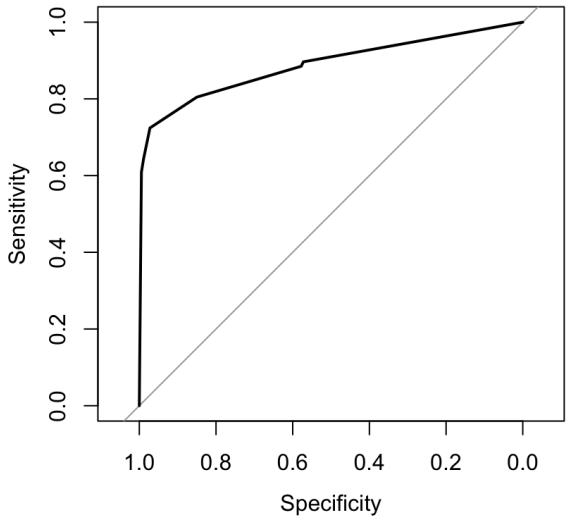
xx




xxi


Decile-Wise Lift Chart
Lift
0.0
0.5
1.0
1.5
2.0
2.5
3.0
20
22
25
36
57
58
100
Percentile

[xxii]


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

[xxiii]