The Factors of Happiness

Andie Donovan

6/14/2017

Loading Necessary Modules:

```
library(ggplot2) # Data visualization
library(randomForest) #randomForest, treesize
library(tree) #decision tree
library(class) #knn
library(rpart) #rpart, printcp, etc. for decision trees
library(matrixStats) #column medians
library(ROCR) #ROCR curves
library(corrplot) #correlation plots
```

Reading the Data File:

```
df <- read.table('~/Happiness_New.txt',header=T,dec='.',sep=",", stringsAsFactors = T)
```

Exploratory Analysis and Preprocessing Step:

The first step is to check if there was missing or inaccurate values

```
any(is.na(df))
## [1] FALSE
```

It appears that there are no missing values in the dataset.
Next, let’s take a look at a portion of the data and a few of the summary statistics

```
head(df) #Let's take a look at what the data set looks like
##       Country         Region Happiness.Rank Happiness.Score
## 1     Denmark Western Europe              1           7.526
## 2 Switzerland Western Europe              2           7.509
## 3     Iceland Western Europe              3           7.501
## 4      Norway Western Europe              4           7.498
## 5     Finland Western Europe              5           7.413
## 6      Canada  North America              6           7.404
##   Lower.Confidence.Interval Upper.Confidence.Interval
## 1                     7.460                     7.592
## 2                     7.428                     7.590
## 3                     7.333                     7.669
## 4                     7.421                     7.575
## 5                     7.351                     7.475
## 6                     7.335                     7.473
##   Economy..GDP.per.Capita.  Family Health..Life.Expectancy. Freedom
## 1                  1.44178 1.16374                  0.79504 0.57941
## 2                  1.52733 1.14524                  0.86303 0.58557
## 3                  1.42666 1.18326                  0.86733 0.56624
## 4                  1.57744 1.12690                  0.79579 0.59609
## 5                  1.40598 1.13464                  0.81091 0.57104
## 6                  1.44015 1.09610                  0.82760 0.57370
##   Trust..Government.Corruption. Generosity Dystopia.Residual
## 1                       0.44453    0.36171           2.73939
## 2                       0.41203    0.28083           2.69463
## 3                       0.14975    0.47678           2.83137
## 4                       0.35776    0.37895           2.66465
## 5                       0.41004    0.25492           2.82596
## 6                       0.31329    0.44834           2.70485

str(df) # How many observations, variables, variable class types, levels, etc.
## 'data.frame':    157 obs. of  13 variables:
##  $ Country                      : Factor w/ 157 levels "Afghanistan",..: 38 135 58 104
45 26 98 99 7 134 ...
##  $ Region                       : Factor w/ 10 levels "Australia and New Zealand",..:
10 10 10 10 10 6 10 1 1 10 ...
##  $ Happiness.Rank               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Happiness.Score              : num  7.53 7.51 7.5 7.5 7.41 ...
##  $ Lower.Confidence.Interval    : num  7.46 7.43 7.33 7.42 7.35 ...
##  $ Upper.Confidence.Interval    : num  7.59 7.59 7.67 7.58 7.47 ...
##  $ Economy..GDP.per.Capita.     : num  1.44 1.53 1.43 1.58 1.41 ...
##  $ Family                       : num  1.16 1.15 1.18 1.13 1.13 ...
##  $ Health..Life.Expectancy.     : num  0.795 0.863 0.867 0.796 0.811 ...
##  $ Freedom                      : num  0.579 0.586 0.566 0.596 0.571 ...
##  $ Trust..Government.Corruption.: num  0.445 0.412 0.15 0.358 0.41 ...
##  $ Generosity                   : num  0.362 0.281 0.477 0.379 0.255 ...
##  $ Dystopia.Residual            : num  2.74 2.69 2.83 2.66 2.83 ...

summary(df) #Summary statistics for all of the variables
##         Country                                 Region    Happiness.Rank
##  Afghanistan:  1   Sub-Saharan Africa             :38   Min.   :  1.00
##  Albania    :  1   Central and Eastern Europe     :29   1st Qu.: 40.00
##  Algeria    :  1   Latin America and Caribbean    :24   Median : 79.00
##  Angola     :  1   Western Europe                 :21   Mean   : 78.98
##  Argentina  :  1   Middle East and Northern Africa:19   3rd Qu.:118.00
##  Armenia    :  1   Southeastern Asia              : 9   Max.   :157.00
##  (Other)    :151   (Other)                        :17
##  Happiness.Score Lower.Confidence.Interval Upper.Confidence.Interval
##  Min.   :2.905   Min.   :2.732             Min.   :3.078
##  1st Qu.:4.404   1st Qu.:4.327             1st Qu.:4.465
##  Median :5.314   Median :5.237             Median :5.419
##  Mean   :5.382   Mean   :5.282             Mean   :5.482
##  3rd Qu.:6.269   3rd Qu.:6.154             3rd Qu.:6.434
##  Max.   :7.526   Max.   :7.460             Max.   :7.669
##
##  Economy..GDP.per.Capita.     Family       Health..Life.Expectancy.
##  Min.   :0.0000           Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.6702           1st Qu.:0.6418   1st Qu.:0.3829
##  Median :1.0278           Median :0.8414   Median :0.5966
##  Mean   :0.9539           Mean   :0.7936   Mean   :0.5576
##  3rd Qu.:1.2796           3rd Qu.:1.0215   3rd Qu.:0.7299
##  Max.   :1.8243           Max.   :1.1833   Max.   :0.9528
##
##     Freedom       Trust..Government.Corruption.   Generosity
##  Min.   :0.0000   Min.   :0.00000               Min.   :0.0000
##  1st Qu.:0.2575   1st Qu.:0.06126               1st Qu.:0.1546
##  Median :0.3975   Median :0.10547               Median :0.2225
##  Mean   :0.3710   Mean   :0.13762               Mean   :0.2426
##  3rd Qu.:0.4845   3rd Qu.:0.17554               3rd Qu.:0.3119
##  Max.   :0.6085   Max.   :0.50521               Max.   :0.8197
##
##  Dystopia.Residual
##  Min.   :0.8179  
##  1st Qu.:2.0317  
##  Median :2.2907  
##  Mean   :2.3258  
##  3rd Qu.:2.6646  
##  Max.   :3.8377  
## 
```

The 6 main predictor factors were reported for each country on a scale of 1 to 10:

- Gross Domestic Product (GDP)
- Percieved Family Support (Family)
- Quality of Health and Life Expectancy (Health)
- Percieved Freedoms of Religion, Property, Speech, etc. (Freedom)
- Trust in Government / Lack of Corruption (Trust)
- Charitable Donations (Generosity)

Feature Engineering

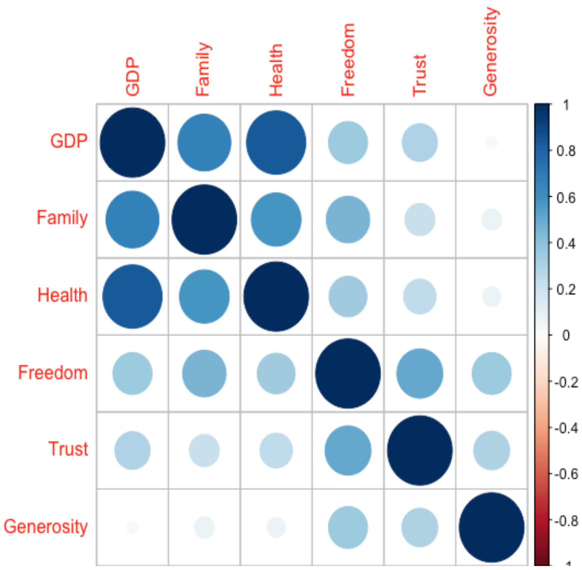
I created a new binary variable called "Happy" in which countries were assigned a value of 1 if their Happiness Score was above the world median of 5.314 and 0 if it was equal to or below the median. I put these six predictor variables and "Happy"" into a new data frame called "happy.new" and then scaled the numerical variables of the data frame and labeled it "happy.s"

```
colnames(df)<-c("Country","Region", "Rank","Score", "LCI","UCI", "GDP","Family", 
"Health", "Freedom", "Trust", "Generosity", "Dystopia") #Renaming the columns/ variables

df = df %>% mutate(Happy=as.factor(ifelse(Score <= median(Score), "0", "1"))) #Creating a 
new binary variable "Happy" based on Happiness score of each country

new.happy = df %>% 
  mutate(Happy=as.factor(ifelse(Score <= median(Score), "No", "Yes"))) %>% 
  select(-Country, -Region, -Rank, -Score, -LCI, -UCI, -Dystopia) #Creating a new data 
set with only the Happy variable and the 6 predictor attributes

scaled<-scale(new.happy[,1:6]) #Scaled predictor variable, makes comparison easier
happy.s<-cbind(scaled, new.happy$Happy) #add Happy variable back
happy.s<-as.data.frame(happy.s)
happy.s= happy.s %>% mutate(Happy=as.factor(ifelse(V7<2, "0", "1")))
happy.s=happy.s[,-7]
```

Further Exploratory Analysis and Basic Data Visualization:

To explore the relationship between the variables, I created a correlation plot using standard Pearson distance.

```
cor.happy<-cor(new.happy[,1:6], use="complete", method="pearson")
corrplot(cor.happy)
```



We see that GDP and Health have the highest correlation while GDP and Generosity have a slightly negative correlation. While the positive relationship between health/life expectancy and income per capita makes sense (higher income means a stable food supply and more money to spend on health care), the relationship between generosity and GDP is counter-intuitive and indicates that people living in wealthier countries may be less inclined to donate to charitable causes. Generosity also has low correlation with Family and Health.

Variable Means and Medians:

Evaluating the means and medians of the variables is useful for identifying whether the data distributions are Normal or skewed

```
new.happy.numeric<-new.happy[,1:6]
new.happy.numeric<-as.matrix(new.happy.numeric)
Medians<-colMedians(new.happy.numeric, na.rm=FALSE)
Means<-colMeans(new.happy.numeric, na.rm = FALSE)
comp<-cbind(Medians,Means)
comp
##            Medians     Means
## GDP        1.02780 0.9538798
## Family     0.84142 0.7936211
## Health     0.59659 0.5576190
## Freedom    0.39747 0.3709939
## Trust      0.10547 0.1376238
## Generosity 0.22245 0.2426349
```

I produced a table of the means and medians to compare central tendencies and search for clues of skewed distributions. From this table, it is very easy to see that means are very close to the medians for all of the variables, giving no indication of skew.

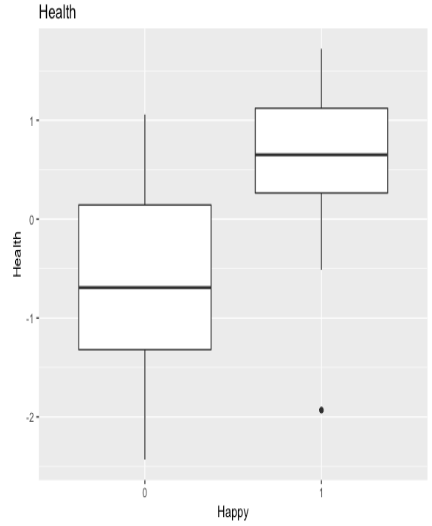
Boxplots:

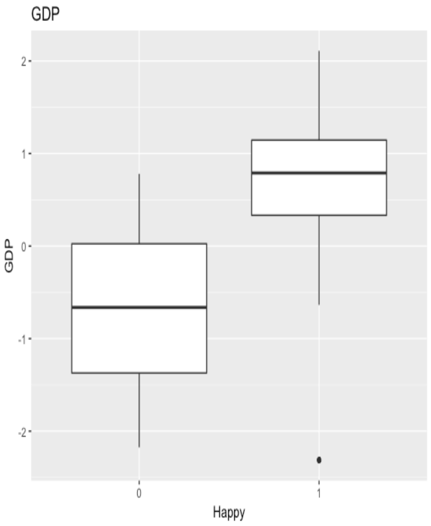
To further illustrate the distributions I plotted the boxplots of my main six attributes for each level of the response variable: “happy” (assigned a value of 1) or “unhappy” (value of 0).

I found that for countries labeled as unhappy, the distributions for health, GDP, Family, and Freedom appeared to be fairly symmetric. Conversely, the distributions for Generosity and Trust seemed skewed to the right. I hypothesized that this could be due to large differences in available income and government transparency and accountability across countries with lower population satisfaction. Unsurprisingly, the scores were lower for every variable for unhappy countries when compared to those of happy countries.

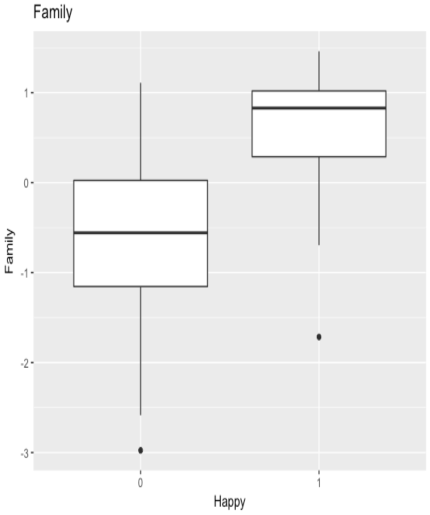
The discrepencies are most noticable in the boxplots for GDP, Family, and Health. Interestingly, the plots for the variable Generosity are almost identical accrross happy and unhappy countries, indicating that perhaps Generosity is either a fairly uninfluential factor in happiness or that people are equally as generous regardless of happiness level.

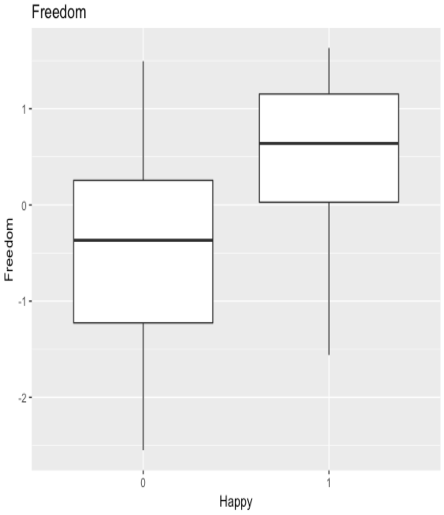
```r
qplot(Happy, Health, data=happy.s, geom="boxplot", main="Health")
qplot(Happy, GDP, data=happy.s, geom="boxplot", main="GDP")
```



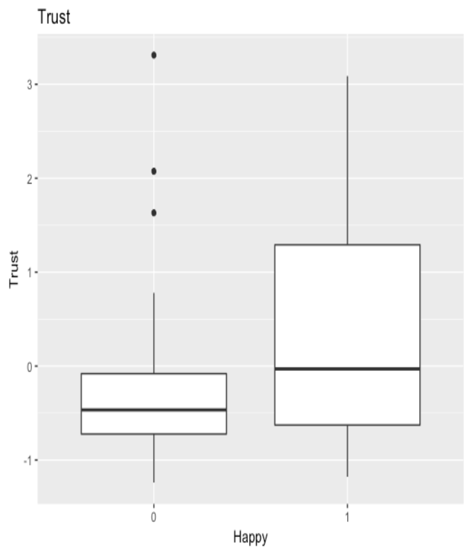


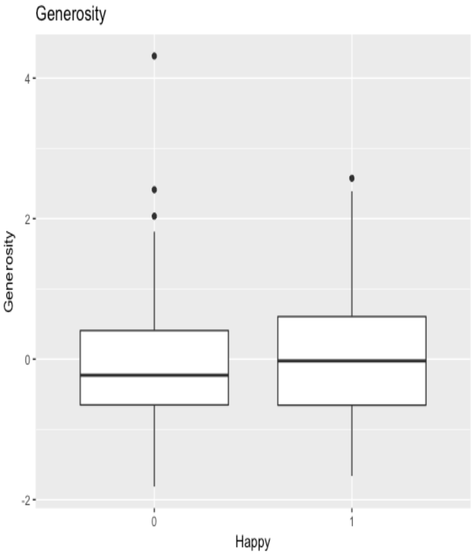
```r
qplot(Happy, Family, data=happy.s, geom="boxplot", main="Family")
qplot(Happy, Freedom, data=happy.s, geom="boxplot", main="Freedom")
```





```r
qplot(Happy, Trust, data=happy.s, geom="boxplot", main="Trust")
qplot(Happy, Generosity, data=happy.s, geom="boxplot", main="Generosity")
```





From the boxplots above, it appears that there is a slight positive skew for the Trust variable within countries labeled as “Happy. Additionally, there appears to be a few outliers on the boxplot of Trust scores in the the”unhappy" countries. A possible future step would be to Normalize the data to reduce noise and bias. Lastly, for every variable except Generosity, the scores are significantly higher for countries labeled as Happy. This could indicate that countries in which the population reports higher levels of Family, Trust, GDP, Freedom, and Health also tend to rate their overall happiness and well-being as high.

Data Mining Techniques

Logistic Regression:

I begin my data mining analysis by fitting a General Linear Model on the regressor `Happy`. This is a logistic regression (aka “logit model”) since the response variable is binary, and therfore set the model parameter `family` to “binomial” to indicate the logit link function. I will use the scaled data for consistency.

```
happy.glm.scaled<-glm(Happy~., data=happy.s, family="binomial") # run logistic regression
on Happy variable

summary(happy.glm.scaled)
##
## Call:
## glm(formula = Happy ~ ., family = "binomial", data = happy.s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2351  -0.3426  -0.0141   0.4072   3.1711
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1148     0.3047  -0.377   0.7064
## GDP           1.2564     0.6389   1.966   0.0492 *
## Family        1.1886     0.4537   2.620   0.0088 **
## Health        1.1128     0.5529   2.013   0.0441 *
## Freedom       0.7685     0.3701   2.077   0.0378 *
## Trust         0.7556     0.4048   1.866   0.0620 .
## Generosity   -0.1723     0.3100  -0.556   0.5783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 217.642  on 156  degrees of freedom
## Residual deviance:  92.222  on 150  degrees of freedom
## AIC: 106.22
##
## Number of Fisher Scoring iterations: 7

pred.glm = predict(happy.glm.scaled, type="response") #use predict function to get our
classifications from estimated probabilities (log odds)
pred.glm=round(pred.glm, digits=2) #round probabilities
df = df %>%
mutate(predHappy=as.factor(ifelse(pred.glm<=0.5, "0", "1"))) #predicted class labels from

table(pred=df$predHappy, true=df$Happy) #Creating a confusion Matrix
##     true
## pred  0  1
##    0 67  8
##    1 12 70
pred.glm2 <- ifelse(pred.glm > 0.5,1,0)
MisClasErr <- mean(pred.glm2 != df$Happy)

print(paste('Accuracy Rate:',round(1-MisClasErr, 4)))
## [1] "Accuracy Rate: 0.8726"
```

From the output, it is important to note that GDP has the highest coefficient, followed by Family and then Health, meaning that their effects on predicted happiness are estimated to be the highest. I would also like to highlight the large p-value for Generosity, which indicates that it is not a significant variable in our model and can possibly be removed. At an alpha of 0.05, we would also find trust to be insignificant.

Investigating the accuracy of this model:

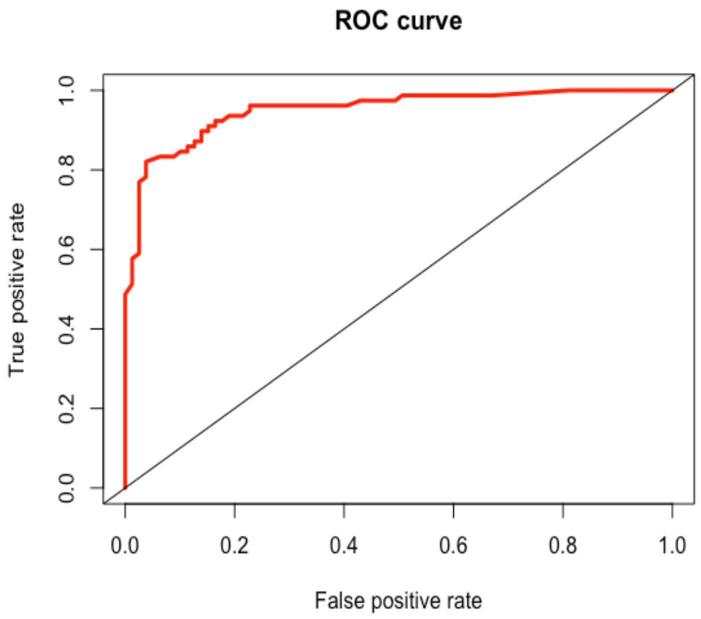
- Out of 157 cases (countries), our model classified 72+67= 139 correctly (this is 88.535%)
- Out of the 79 “not happy” countries, the model classified 72 (91.139%) correctly.
- Out of 78 “happy” countries, the model classified 67 (85.897%) correctly.

Accuracy Rate = (67+70) / (67+8+12+70) = 87.21%

- False Positive Rate (FPR): 11/ 78 (14.10256%)
- False Negative Rate (FNR): 7/ 79 (8.860759%)

The GLM model obtained an accuracy rate of 87.12%, which is fairly accurate. We are able to visualize the performace of the model with a ROC curve, which maps the True Positive Rate against the False Positive Rate.

```
pred = prediction(pred.glm, df$Happy)
perf = performance(pred, measure="tpr", x.measure="fpr")

plot(perf, col=2, lwd=3, main="ROC curve")
abline(0,1)
```



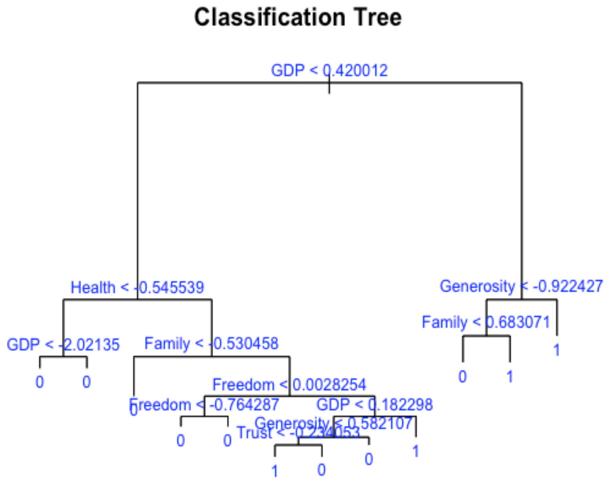
```
auc = performance(pred, "auc")@y.values
#Calculate Area under the curve to evaluate performance of the glm model
print(paste('AUC:',auc))
## [1] "AUC: 0.95042194092827"
```

The high corresponding area under the curve (AUC) amount of .945, indicates that the GLM is an accurate model for the data.

Decision Tree

Next, I decided to use a decision tree to model the data. To do this I regressed Happy on the explanatory variables using the scaled data as before and plugged that into tree function. I then plotted the tree and added the labels for each branch.

```
tree.happy = tree(Happy~., data = happy.s)
plot(tree.happy)
text(tree.happy, pretty = 0, cex = .8, col = "blue")
title("Classification Tree")
```

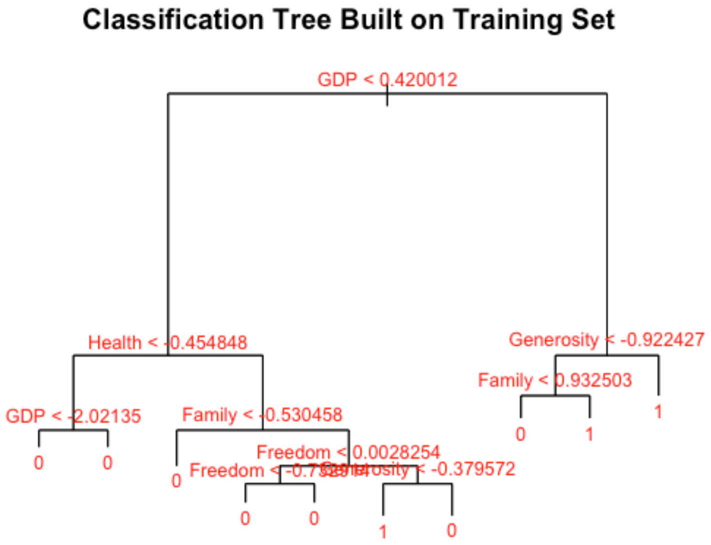


```
summary(tree.happy)
##
## Classification tree:
## tree(formula = Happy ~ ., data = happy.s)
## Number of terminal nodes:  12
## Residual mean deviance:  0.2708 = 39.26 / 145
## Misclassification error rate: 0.05732 = 9 / 157
```

From the summary of the tree, we see that the optimal number of nodes for the tree is 12 and that the model has a misclassification error rate of 0.05732.

To test the accuracy of the model, I created a training and test set by randomly partitioning the data. By fitting the model on the training data, I could then compare the model's predictions with the test data (which contains the true class labels). I followed the same process as before, but this time using the training set "trainSet". I then computed the test error rate and accuracy.

```
set.seed(1) #setting seed for random sample
trainSet = sample(1:nrow(happy.s), 0.75*dim(happy.s)[1]) #75% of observations set as
training set
happy.test = happy.s [-trainSet,] #the remaining 25% of observations set as the test set
Happy.true = happy.test$Happy #true classifications of the test set

#Fit the tree on training set:
tree.training = tree(Happy~., data = happy.s, subset = trainSet)

plot(tree.training)
text(tree.training, pretty = 0, cex = .8, col = "red")
title("Classification Tree Built on Training Set")
```



```
summary(tree.training)
##
## Classification tree:
## tree(formula = Happy ~ ., data = happy.s, subset = trainSet)
## Variables actually used in tree construction:
## [1] "GDP"         "Health"      "Family"      "Freedom"     "Generosity"
## Number of terminal nodes:  10
## Residual mean deviance:  0.3291 = 35.21 / 107
## Misclassification error rate: 0.09402 = 11 / 117
```

We can see that for this tree, the number of nodes was set to 10 and the misclassification error rate has increased from about 6% to 9.4%.

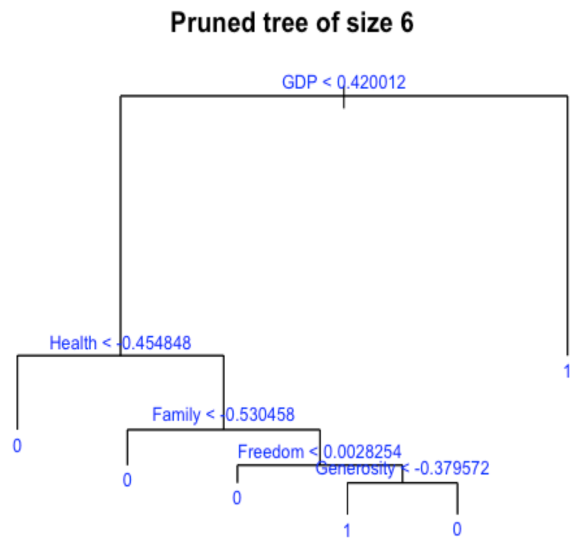
```
#Computing accuracy and test error rate:
TrainingTree.pred = predict(tree.training, happy.test, type="class") # make predictions
on test set

error = table(TrainingTree.pred, Happy.true)
print("Confusion Matrix: ")
## [1] "Confusion Matrix: "
error
##                    Happy.true
## TrainingTree.pred  0  1
##                  0 19  6
##                  1  1 14
accuracy2=sum(diag(error))/sum(error) #Test accuracy rate
print(paste('Accuracy Rate:',accuracy2))
## [1] "Accuracy Rate: 0.825"
class_error=1-sum(diag(error))/sum(error) # Test error rate (Classification Error)
print(paste('Test error:',class_error))
## [1] "Test error: 0.175"
```

The model produced an accuracy rate of 82.5%, which is lower than that of the GLM model.

```
set.seed(1) # for reproducibility
cv = cv.tree(tree.training, FUN=prune.misclass) #10 fold cross validation (by default) to
determine optimal tree complexity
summary(cv)
##        Length Class  Mode
## size   6      -none- numeric
## dev    6      -none- numeric
## k      6      -none- numeric
## method 1      -none- character
cv
## $size
## [1] 10  8  6  5  2  1
##
## $dev
## [1] 29 29 26 27 26 77
##
## $k
## [1] -Inf  0.0  0.5  1.0  2.0 39.0
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"

best.cv = cv$size[which.min(cv$dev)]
best.cv
## [1] 6
```

The optimal tree size, as determined by cross validation is 6 leaves. Therefore we will cut the tree, plot it, and extract the new confusion matrix and accuracy rate:

```
prune.cv = prune.misclass (tree.training, best=best.cv) #prune our decision tree built on
the training set
plot(prune.cv)
text(prune.cv, pretty=0, col = "blue", cex = .8)
title("Pruned tree of size 6")
```



```
pred.cv.prune = predict(prune.cv, happy.test, type="class")
# confusion matrix
err.cv.prune = table(pred.cv.prune, Happy.true)
err.cv.prune
##               Happy.true
## pred.cv.prune  0  1
##              0 18  5
##              1  2 15
acc3=sum(diag(err.cv.prune))/sum(err.cv.prune)
print(paste("Misclassification Error: ", 1-sum(diag(err.cv.prune))/sum(err.cv.prune)))
## [1] "Misclassification Error:  0.175"
print(paste('Accuracy Rate:',acc3))
## [1] "Accuracy Rate: 0.825"
```

The pruned tree provides the same accuracy rate as our intital model but with less complexity, and is therefore preffered. However, the GLM model still remains the most accurate model for predicting happiness.

Conclusion:

Of the Machine Learning models utilized, the Logistic Regression model performed the best with was an accuracy rate of 87.21%, a Type I Error rate of 5.1% and a Type II Error Rate of 7.6%. This was significantly higher than the accuracy rate of the decision tree (82.5%). Additionally, from the GLM model and the initial exploratory analysis, I concluded that GPD per capita had the highest influence on happiness level while Generosity had the lowest. Although in this case it appears that Gross Domestic Product—a money related measurement—was in fact the most telling predictor of whether a country's population reported themselves as "happy" or not, finding new and more complete methods of analyzing a society's well being and growth is an important part of understanding international development. Future work could include utilizing more objective variables such as literacy rates, percent of population with access to water, electricity, and the Internet, the population age distribution, and unemployment rates to paint a more holistic picture of a country's status.