Decision trees and ensemble methods: what variable best predicts the divorce outcome of couples?

Cimmaron Yeoman

2023-04-01

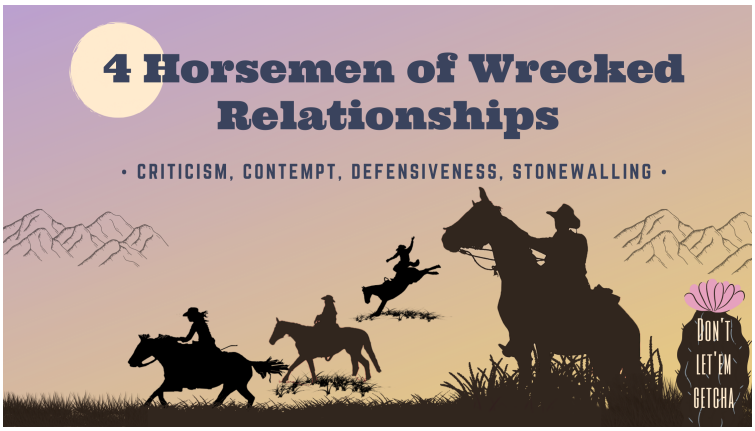
Introduction

Gottman Couples Therapy is a research based form of couples therapy created by Drs. John and Julie Gottman. The method was developed to identify issues in relationships and improve them. This report evaluates a data set containing 54 Divorce Predictor Scale variables, questions derived from Gottman Couples therapy, to assess the relationship of 170 couples across Turkey. Couples included in the study were either divorced or still married, and each pair responded to the predictor variable statements using a 5 level scale (0 = Never, 1 = Seldom, 2 = Averagely, 3 = Frequently, 4 = Always). Sample questions include *"I can use negative statements about my spouse's personality during our discussions (33)"* or *"Our dreams with my spouse are similar and harmonious (15)"*.

The goal of this report was to identify which variable best predicts divorce using decision trees and other ensemble methods. In the data set, **Class** is the response variable, with 0 = still married and 1 = divorced. Each predictor variable is labeled as 'Atr' with the question number attached at the end (ex. **Atr22**).

Figure 1

Gottman's Four Horsemen of the Apocalypse: relationship edition



Importing the data set

The data set was imported into R Studio and examined for any missing cases. No observations were missing and the data set was renamed to **divorce**.

```
library(readr)
divorce <- read.csv("divorce_data.csv", header = TRUE)
divorce <- divorce[complete.cases(divorce), ]
```

Test/train and randomization

A test and train set were made and the response **Class** was changed to a factor. Another factor-free test/train set was saved too (hidden and not in the chunk below).

```
set.seed(20)
divorce2 <- divorce
divorce2$Class <- as.factor(divorce2$Class)
train2_D <- sample(1:nrow(divorce2), 0.5* nrow(divorce2), replace = FALSE)
test2_D <- divorce2[-train2_D, ]
train2_D <- divorce2[train2_D, ]
save(file = "train2_D.rda", train2_D)
save(file = "test2_D.rda", test2_D)
```

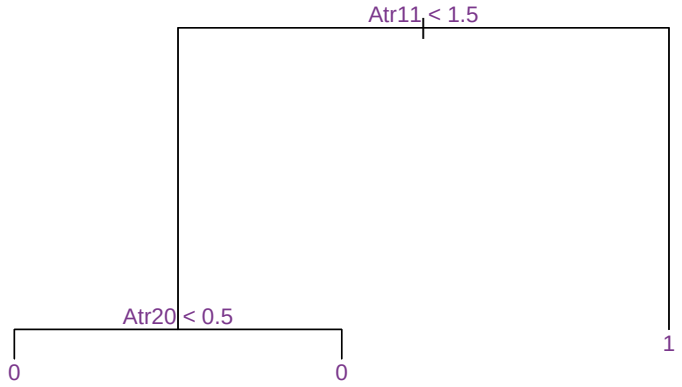
A Simple decision tree

A basic decision tree or classification tree with three terminal nodes was made. This tree selected the **Atr11** (*"... when I look back, I see that my spouse and I have been in harmony with each other"*), and **Atr20** (*"My spouse and I have similar values in trust"*) predictor variables. The model frequently included **Atr11** even if the test/train data split was loaded with a different seed number. It was not possible to get a tree with more than three terminal nodes.

```
tree_S <- tree(Class ~., data = train2_D)
summary(tree_S)
tree_S
```

A simple tree plotted

Figure 2



Training and testing errors:

The training error was **2.35%** and the testing error was **2.35**.

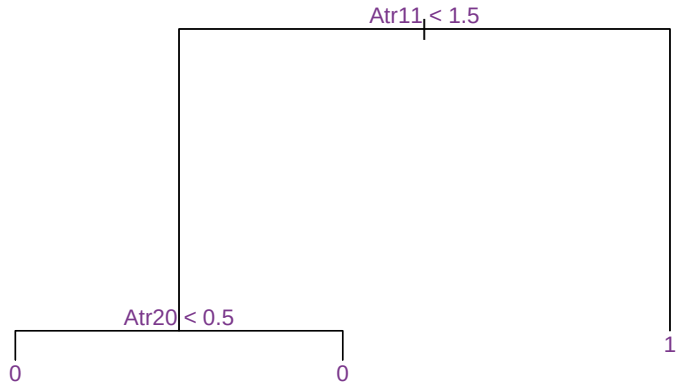
```
##
## tree_S1  0  1
##       0 45  2
##       1  0 38
## [1] 2.352941
##
## tree_S2  0  1
##       0 41  2
##       1  0 42
## [1] 2.352941
```

Adding variables

I tried to run a more complicated decision tree but results were not much different, the plot looked exactly the same.

```
tree_C <- tree(Class ~  Atr11 +Atr20 + Atr33 + Atr34 + Atr35 + Atr40 +Atr31
               +Atr7 +Atr2, data = train2_D)
summary(tree_C)
tree_C
```

Figure 3



Adding more variables did not seem to change the decision tree at all. I included **Atr20** and **Atr11** from the simple tree, as they were continuously selected, regardless of which variables were added. If they were not included, sometimes the tree would gain another terminal node with duplicated variables or continuously select another variable like **Atr40**. I am assuming the variables that the model repeatedly selects are very relevant to either the outcome of divorce **(1)** or the outcome of still married **(0)**.

Training and testing errors:

The training and testing errors were both **2.35%**.

```
##
## complex_pred1  0  1
##             0 45  2
##             1  0 38
## [1] 2.352941
##
## complex_pred2  0  1
##             0 41  2
##             1  0 42
## [1] 2.352941
```

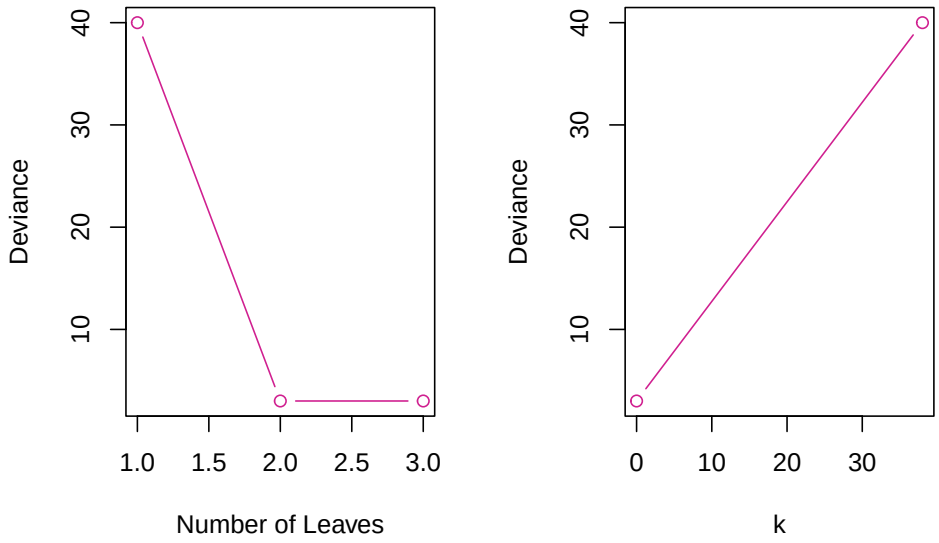
The training and testing errors were exactly the same for the complex decision tree. This indicates that the first simple model has already captured the predictor variable most relevant to the divorce outcome response. I will test a few other methods to try and produce a better or different result.

Prune the tree

Using pruning to cross validate and assess the model for error, the number of least selected drops off at **2**. When this is plotted, only **Atr11** is selected. The testing/training error was **2.35%**, once again.

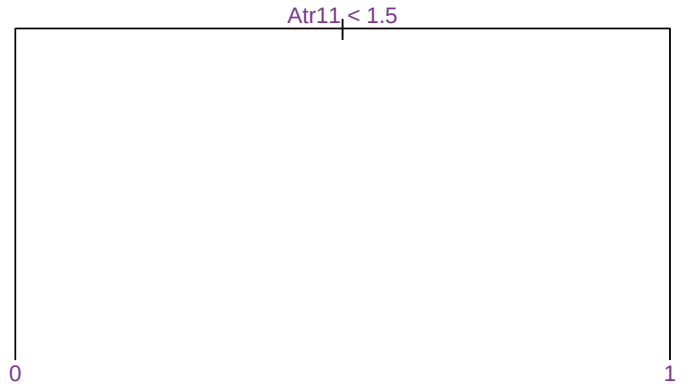
Ideal number of leaves

Figure 4



Pruned model

Figure 5



Training and testing errors:

The results and code not were included below as they once again, had the exact same errors of **2.35%**.

Random forest

A Random Forest model was attempted with the variables **At11** and **Atr20**, along with a few more of the variables from the more complicated tree tested earlier.

```
set.seed(20)
rf_D <- randomForest(Class ~ Atr11 + Atr20 + Atr33 + Atr35 + Atr40
                     +Atr31 +Atr2, data = train2_D, mtry = 3, type = "class",
                     importance = TRUE)
```

Training and testing errors:

The Random Forest produced a testing and training error of **0%** with these variables, which was strange, and perhaps a sign of underfitting?

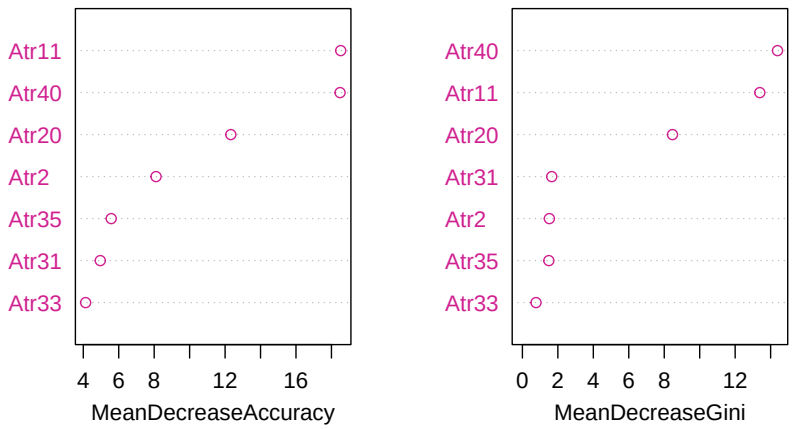
```
##
## pred_Drf1  0  1
##         0 45  0
##         1  0 40
## [1] 0
##
## pred_Drf2  0  1
##         0 41  0
##         1  0 44
## [1] 0
```

Random forest variable importance analysis

The mean decrease accuracy found that model accuracy would decrease by about **18%** for **Atr40** and **Atr11**. The **Atr20** variable followed with just over **12%**. In table below, note these three variables were also listed with the highest mean decrease gini scores, emphasizing their importance. If my assumption is correct, the matrix to the left of the these scores indicates the mean decrease accuracy for the married **(0)** versus divorce **(1)** outcome. If so, **Atr40** could be the best predictor of the divorce response and **Atr11** could be the best predictor of the still married response.

```
##                0         1 MeanDecreaseAccuracy MeanDecreaseGini
## Atr11 19.062291  9.400398             18.522981       13.3902241
## Atr20 11.058352  6.902568             12.324028        8.4651823
## Atr33  2.467082  3.156247              4.131768        0.7679234
## Atr35  4.185185  3.446342              5.577554        1.4907795
## Atr40 17.557053 12.938866             18.486111       14.3891386
## Atr31  4.872235  1.256078              4.960783        1.6534080
## Atr2   9.466975 -1.664134              8.105197        1.5201812
```

Figure 6



Experimenting with subsets

Variables **Atr40**, **Atr11**, and **Atr20** were all separated into their own subset with the **Class** response variable.

```
Atr40_level <- divorce[, c("Atr40", "Class")]
Atr11_level <- divorce[, c("Atr11", "Class")]
Atr20_level <- divorce[, c("Atr20", "Class")]
```

When the response variable equaled 1 or divorced, I compared it against the predictor variable scale response of "frequently" **(3)** or more. I did the same for when the response variable equaled 0 or still married, comparing it to the predictor variable scale score of "averagely" **(2)** or less. I also tried another scale level for the married group of "never" **(0)**. The output provided me with the number of divorced or married couples who matched each scale score of the predictor variables.

- The **Atr40** variable was *"We're just starting a discussion/argument before I know what's going on"*
- The **Atr11** variable was *"I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other"*.
- The **Atr20** variable was *"My spouse and I have similar values in trust"*.

I will assume these variables are most relevant in the divorce outcome response.

Variable Atr40

There were **82** divorced couples who had sudden arguments averagely or more, and **79** who had arguments frequently or more. There were **86** married couples who had sudden arguments averagely or less, and **71** couples who never had sudden arguments.

```
#Example of code used to find variables with specific conditions
sum(Atr40_level$Class == 1 & Atr40_level$Atr40 >=2)
## [1] 82
sum(Atr40_level$Class == 1 & Atr40_level$Atr40 >=3)
## [1] 79
sum(Atr40_level$Class == 0 & Atr40_level$Atr40 <=2)
## [1] 86
sum(Atr40_level$Class == 0 & Atr40_level$Atr40 ==0)
## [1] 71
```

Variable Atr11

Oddly, there were a large number of divorced couples who answered the question regarding if they felt in harmony, with **76** feeling in harmony frequently or more, and **80** feeling in harmony averagely or more. For married couples, **69** answered they never felt in harmony, and in total, **86** couples answered either never or seldom. This seems like the data is incorrect, and not what I would expect, but perhaps something else is going on. Potentially, married couples may not feel in harmony but this fact does not negatively impact their marriage.

```
## [1] 76
## [1] 80
## [1] 86
## [1] 69
```

Variable Atr20

The **Atr20** variable seemed more relevant for predicting a couple is married, as the **86** married couples answered they seldom or never had the same trust values, with **80**couples never having the same trust values. There were **79** divorced couples who had the same trust values averagely or more, with only **5** couples giving scores of seldom or never. Again, like **Atr11** this was not the expected answer. I see a pattern.

```
## [1] 5
## [1] 79
## [1] 86
## [1] 80
```

BART

The train/test sets were slightly modified before running Bayesian Additive Regression Trees (BART).

```
xtrainD <- model.matrix(Class ~ ., data = train2_D)[, -1] # drop intercept
xtestD <- model.matrix(Class ~ ., data = test2_D)[, -1] # drop intercept
ytrainD <- train_D[, "Class"]
ytestD <- test_D[, "Class"]
```

```
set.seed(20)
Dbart_fit <- gbart(xtrainD, ytrainD, x.test = xtestD, type = "lbart" )
```

Training and testing errors:

BART had a training error of **1.18%** and a testing error of **2.35%**.

```
## [1] 2.352941
## [1] 1.176471
```

How often variables were selected

Interestingly, **Atr26** appeared the most times in the collection of trees from BART, with **Atr40** following closely as second most frequently occurring. All of the variables had a mean occurrence of less than 1.5 though. The **Atr11** variable was also selected less frequently at 1.029. Possibly BART does not work here and could be underfitting. The results from the original simple trees seem to make the most sense for this data set and predicting which variables result in divorce.

```
ord_B <- order(Dbart_fit$varcount.mean, decreasing = TRUE)
Dbart_fit$varcount.mean[ord_B]
```

Summarizing results

For the decision trees and ensemble methods tested on the divorce data set split into test/train, these are all of the predictions:

```
##           Model TrainingError TestingError
## 1    Basic Tree           2.4          2.4
## 2  Complex Tree           2.4          2.4
## 3   Pruned Tree           2.4          2.4
## 4 Random Forest           0.0          0.0
## 5          BART           2.4          1.2
```

Conclusion

After testing several different models, it appears that the original simple decision tree probably was the best prediction model. I believe **Atr11**, **Atr20**, and **Atr40**, were the questions which best predicted a couples relationship class of married or divorced. Overall, **Atr11** was the most relevant and frequently selected predictor variable for predicting the divorce outcome, after analyzing various simple decision trees and their testing/training errors. I believe that BART and Random Forest may be underfitting the data. The data set included answers from couples which I found were unexpected for the still married and divorced outcomes, perhaps more context on the couples lifestyle or insight from the original study would demystify this.