CYO_Final_Project

Matt Harvill

7/27/2021

INTRODUCTION

In 1969 *Psychology Today* published a survey from Ray Fair in which respondents were asked whether they had extra-marital affairs and to rate themselves and their marriages. Multiple variables were assessed in the survey, which has since been referred to as "Fair's Affairs". This report uses the dataset to determine which, if any, predictors are more prevelant in predicting affairs and which of a handful of prediction models might be most effective in predicted outcomes accurately.

OVERVIEW

The dataset consists of 601 observations and 9 variables.
The variables are:

affairs; a numeric value, **0-12**, how often respondaents engaged in extramarital sexual intercourse during the past year

gender; a factor, male or female

age; a numeric variable coding age in years: **17.5** = under 20, **22** = 20–24, **27** = 25–29, **32** = 30–34, **37** = 35–39, **42** = 40–44, **47** = 45–49, **52** = 50–54, **57** = 55 or over

yearsmarried; a numeric variable coding number of years married: **0.125** = 3 months or less, **0.417** = 4–6 months, **0.75** = 6 months–1 year, **1.5** = 1–2 years, **4** = 3–5 years, **7** = 6–8 years, **10** = 9–11 years, **15** = 12 or more years

children; a factor, yes or no

religiousness; a numeric variable coding religiousness: **1** = anti, **2** = not at all, **3** = slightly, **4** = somewhat, **5** = very

education; a numeric variable coding level of education: **9** = grade school, **12** = high school graduate, **14** = some college, **16** = college graduate, **17** = some graduate work, **18** = master's degree, **20** = Ph.D., M.D., or other advanced degree

occupation; a numeric variable coding occupation: **1** = student, **2** = farming, agriculture; semi-skilled, or unskilled worker; **3** = white-collar; **4** = teacher, counselor, social worker, nurse; artist, writers; technician, skilled worker, **5** = managerial, administrative, business, **6** = professional with advanced degree

rating; a numeric variable coding self rating of marriage: **1** = very unhappy, **2** = somewhat unhappy, **3** = average, **4** = happier than average, **5** = very happy

Data Loading

load dataset, necessary packages and libraries

```r
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(readr)) install.packages("readr", repos = "http://cran.us.r-project.org")
if(!require(stringer)) install.packages("stringer", repos = "http://cran.us.r-project.org")
if(!require(AER)) install.packages("AER", repos = "http://cran.us.r-project.org")
if(!require(texreg)) install.packages("texreg", repos = "http://cran.us.r-project.org")

library(tidyverse)
library(caret)
library(data.table)
library(readr)
library(stringr)
library(AER)
library(ggplot2)
library(gridExtra)
library(texreg)
library(rpart)

data("Affairs")
options(digits = 3)
```

Note: this process could take a couple of minutes

METHODS & ANALYSIS

Linear Models and Logistic Regression will be used to find the strongest indicators of infidelity in the survey. Once those factors have been determined predictions will be used to test the magnitude of their effects. Lastly, different prediction models will be tested against the data to try and see which would be best suited for accurate predictions given a more robust (and modern) dataset.

Data Analysis

first 7 rows with header a snapshot of the data

```
##    affairs gender age yearsmarried children religiousness education occupation
## 4        0   male  37        10.00       no             3        18          7
## 5        0 female  27         4.00       no             4        14          6
## 11       0 female  32        15.00      yes             1        12          1
## 16       0   male  57        15.00      yes             5        18          6
## 23       0   male  22         0.75       no             2        17          6
## 29       0 female  32         1.50       no             2        17          5
##    rating
## 4       4
## 5       4
## 11      4
## 16      5
## 23      3
## 29      5
```

basic summary statistics

```
##     affairs         gender         age         yearsmarried   children
##  Min.   : 0.00   female:315   Min.   :17.5   Min.   : 0.12   no :171
##  1st Qu.: 0.00   male  :286   1st Qu.:27.0   1st Qu.: 4.00   yes:430
##  Median : 0.00                Median :32.0   Median : 7.00
##  Mean   : 1.46                Mean   :32.5   Mean   : 8.18
##  3rd Qu.: 0.00                3rd Qu.:37.0   3rd Qu.:15.00
##  Max.   :12.00                Max.   :57.0   Max.   :15.00
##  religiousness    education      occupation       rating
##  Min.   :1.00   Min.   : 9.0   Min.   :1.00   Min.   :1.00
##  1st Qu.:2.00   1st Qu.:14.0   1st Qu.:3.00   1st Qu.:3.00
##  Median :3.00   Median :16.0   Median :5.00   Median :4.00
##  Mean   :3.12   Mean   :16.2   Mean   :4.19   Mean   :3.93
##  3rd Qu.:4.00   3rd Qu.:18.0   3rd Qu.:6.00   3rd Qu.:5.00
##  Max.   :5.00   Max.   :20.0   Max.   :7.00   Max.   :5.00
```

From the summary we can quickly see a few things in the data. 315 of the survey respondents were female, 286 male. 430 of those had children, and the average age is 32.5.

number of affairs by survey respondents

```
##
##   0   1   2   3   7  12
## 451  34  17  19  42  38
```

451 of the 601 did not report having a extra-marital affair. Of the 25% who reported having affairs 6% say they had at least 12 liaisons.

look at female/male proportion

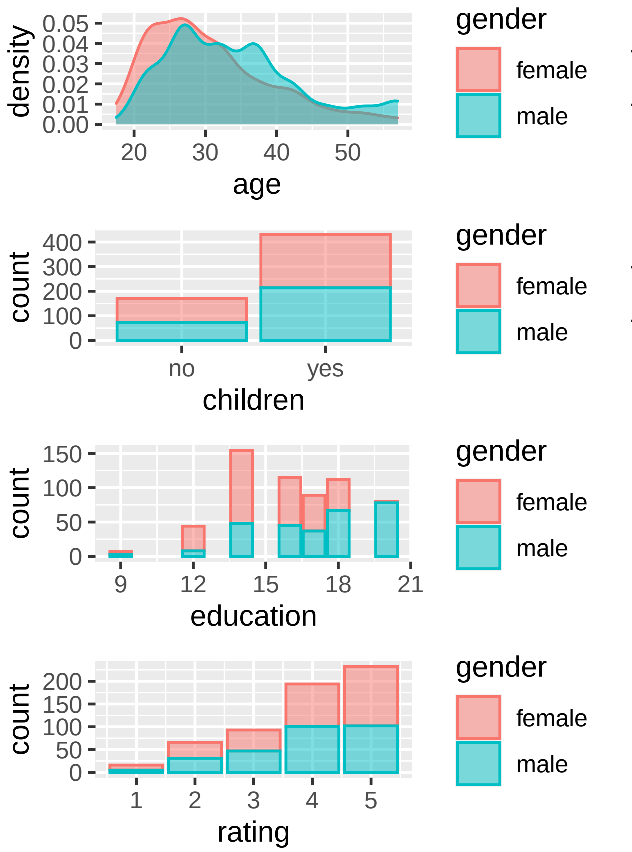
```
##
## female   male
##  0.524  0.476
```

Not quite an even split, but close.

comparing affair tendency by gender

```
## # A tibble: 2 x 3
##   gender average twelve_or_more
##   <fct>    <dbl>          <dbl>
## 1 female    1.42         0.0635
## 2 male      1.50         0.0629
```

Contrary to popular thought, there doesn't seem to be a proclivity for infidelity based on gender.



plots looking at the breakdown of the different affair factors

Here things get a little interesting. Looking at the factors of **age**, **yearsmarried**, **children**, **religiousness**, and **rating** the breakdown seems to pretty even between the genders. Among **education** and **occupation** there are apparent gender inconsistencies. Of the survey respondents who have Ph.d.s or other advanced degrees the *vast* majority are men. Meanwhile for the high school or some college categories, women are the majority. In the occupation category those in the more "prestigious" side are almost all men, while in the student category the affairs are being had by women predominently.

It must be stated that all of this data is self reported, which has inherent problems. One can't help but wonder, though, about the possible correlations of these data.

outcome is binary so transform affairs into yes/no To better see the yes or no answer

```r
Affairs$y_n_affairs[Affairs$affairs > 0] <- 1
Affairs$y_n_affairs[Affairs$affairs == 0] <- 0
Affairs$y_n_affairs <- factor(Affairs$y_n_affairs, levels = c(0,1), labels = c("No", "Yes"))
```

check success

```
##
##  No Yes
## 451 150
```

Looks good!

simple linear regession models

Running a linear regression model on each factor individually to look for potential obvious predictors

```r
age_lm <- lm(affairs ~ age, data = Affairs)
yearsmarried_lm <- lm(affairs ~ yearsmarried, data = Affairs)
children_lm <- lm(affairs ~ children, data = Affairs)
religiousness_lm <- lm(affairs ~ religiousness, data = Affairs)
education_lm <- lm(affairs ~ education, data = Affairs)
occupation_lm <- lm(affairs ~ occupation, data = Affairs)
rating_lm <- lm(affairs ~ rating, data = Affairs)

screenreg(list(age_lm, yearsmarried_lm, children_lm, religiousness_lm,
               education_lm, occupation_lm, rating_lm))
```

```
##
## ===========================================================================================
##                Model 1   Model 2     Model 3     Model 4     Model 5  Model 6    Model 7
## -------------------------------------------------------------------------------------------
## (Intercept)      0.36      0.55 *      0.91 ***    2.73 ***    1.51     1.08 **    4.74 ***
##                 (0.49)    (0.24)      (0.25)      (0.38)      (0.92)   (0.34)     (0.48)
## age              0.03 *
##                 (0.01)
## yearsmarried               0.11 ***
##                           (0.02)
## childrenyes                            0.76 *
##                                       (0.30)
## religiousness                                    -0.41 ***
##                                                   (0.11)
## education                                                    -0.00
##                                                               (0.06)
## occupation                                                              0.09
##                                                                        (0.07)
## rating                                                                            -0.84 ***
##                                                                                   (0.12)
## -------------------------------------------------------------------------------------------
## R^2              0.01      0.03        0.01        0.02        0.00     0.00       0.08
## Adj. R^2         0.01      0.03        0.01        0.02       -0.00     0.00       0.08
## Num. obs.      601       601         601         601         601      601        601
## ===========================================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

It would appear from these models that children, rating and religiousness are the biggest factors on whether or not a person will have an affair; having children seems to have a positive correlation, while rating and religiousness are negative.

However, this method fails to account for the interaction of said factors.

logistic regression

Using Logistic Regression to fit the model

```
fit_all <- glm(y_n_affairs ~ gender + age + yearsmarried + children + religiousness + education +
               occupation + rating, data = Affairs, family = binomial())
summary(fit_all)
```

```
##
## Call:
## glm(formula = y_n_affairs ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.571  -0.750  -0.569  -0.254   2.519
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.3773     0.8878    1.55   0.1208
## gendermale      0.2803     0.2391    1.17   0.2411
## age            -0.0443     0.0182   -2.43   0.0153 *
## yearsmarried    0.0948     0.0322    2.94   0.0033 **
## childrenyes     0.3977     0.2915    1.36   0.1725
## religiousness  -0.3247     0.0898   -3.62   0.0003 ***
## education       0.0211     0.0505    0.42   0.6769
## occupation      0.0309     0.0718    0.43   0.6666
## rating         -0.4685     0.0909   -5.15  2.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 609.51  on 592  degrees of freedom
## AIC: 627.5
##
## Number of Fisher Scoring iterations: 4
```

From here it can be seen that the variables of **age**, **yearsmarried**, **religousness**, and **rating** seem to be the most relevant factors. The focus will be on those factors going forward.

Here is a new fit model using the strongest indicators.

```
fit_fewer <- glm(y_n_affairs ~ age + yearsmarried + religiousness + rating, data = Affairs,
                 family = binomial())
summary(fit_fewer)
```

```
##
## Call:
## glm(formula = y_n_affairs ~ age + yearsmarried + religiousness +
##     rating, family = binomial(), data = Affairs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.628  -0.755  -0.570  -0.262   2.400
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.9308     0.6103    3.16  0.00156 **
## age            -0.0353     0.0174   -2.03  0.04213 *
## yearsmarried    0.1006     0.0292    3.44  0.00057 ***
## religiousness  -0.3290     0.0895   -3.68  0.00023 ***
## rating         -0.4614     0.0888   -5.19  2.1e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 615.36  on 596  degrees of freedom
## AIC: 625.4
##
## Number of Fisher Scoring iterations: 4
```

```
coef(fit_fewer)
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##        1.9308       -0.0353        0.1006       -0.3290       -0.4614
```

Now the overall effects are becoming more clear. The length of time one is married *increases* the likelihood of an affair, while happiness in the marriage and religiousness *decrease* the likelihood. In addition as one gets older there is less proclivity to engage in affairs, which inherently contradicts the years married factor, but doesn’t outweigh it.

probabilites based on stongest factors

Here the testing of these factors individually to see the probabilities at given intervals using the averages of those not being tested.

```
new_ages <- tibble(age = c(17, 27, 37, 47, 57), yearsmarried = mean(Affairs$yearsmarried),
                      religiousness = mean(Affairs$religiousness), rating = mean(Affairs$rating))
new_ages$probability <- predict(fit_fewer, newdata = new_ages, type = "response")
new_ages
```

```
## # A tibble: 5 x 5
##     age yearsmarried religiousness rating probability
##   <dbl>        <dbl>         <dbl>  <dbl>       <dbl>
## 1    17         8.18          3.12   3.93       0.335
## 2    27         8.18          3.12   3.93       0.262
## 3    37         8.18          3.12   3.93       0.199
## 4    47         8.18          3.12   3.93       0.149
## 5    57         8.18          3.12   3.93       0.109
```

A steady decline in the probability of infidelity as one ages. A person in the youngest age range is over 3 times more likely to have an affair than someone in the highest range.

```
new_years <- tibble(age = mean(Affairs$age), yearsmarried = c(0.125,0.417,0.75,1.5,4,7,10,15), religiou
                    rating = mean(Affairs$rating))
new_years$probability <- predict(fit_fewer, newdata = new_years, type = "response")
new_years
```

```
## # A tibble: 8 x 5
##     age yearsmarried religiousness rating probability
##   <dbl>        <dbl>         <dbl>  <dbl>       <dbl>
## 1  32.5        0.125          3.12   3.93       0.115
## 2  32.5        0.417          3.12   3.93       0.118
## 3  32.5        0.75           3.12   3.93       0.121
## 4  32.5        1.5            3.12   3.93       0.130
## 5  32.5        4              3.12   3.93       0.161
## 6  32.5        7              3.12   3.93       0.206
## 7  32.5       10              3.12   3.93       0.260
## 8  32.5       15              3.12   3.93       0.367
```

Looking at a table of years married, as expected a reverse effect is seen. The longer one is married the more probable an affair is to occur, at an almost geometric increase of likelihood.

```
new_rlgn <- tibble(age = mean(Affairs$age), yearsmarried = mean(Affairs$yearsmarried),
                   religiousness = c(1,2,3,4,5), rating = mean(Affairs$rating))
new_rlgn$probability <- predict(fit_fewer, newdata = new_rlgn, type = "response")
new_rlgn
```

```
## # A tibble: 5 x 5
##     age yearsmarried religiousness rating probability
##   <dbl>        <dbl>         <dbl>  <dbl>       <dbl>
## 1  32.5         8.18             1   3.93       0.369
## 2  32.5         8.18             2   3.93       0.296
## 3  32.5         8.18             3   3.93       0.233
## 4  32.5         8.18             4   3.93       0.179
## 5  32.5         8.18             5   3.93       0.136
```

A person who reports to be strongly religious is much less likely to engage in an extra-marital affair.

```
new_rating <- tibble(age = mean(Affairs$age), yearsmarried = mean(Affairs$yearsmarried),
                   religiousness = mean(Affairs$religiousness), rating = c(1,2,3,4,5))
new_rating$probability <- predict(fit_fewer, newdata = new_rating, type = "response")
new_rating
```

```
## # A tibble: 5 x 5
##     age yearsmarried religiousness rating probability
##   <dbl>        <dbl>         <dbl>  <dbl>       <dbl>
## 1  32.5         8.18          3.12      1       0.530
## 2  32.5         8.18          3.12      2       0.416
## 3  32.5         8.18          3.12      3       0.310
## 4  32.5         8.18          3.12      4       0.220
## 5  32.5         8.18          3.12      5       0.151
```

Those that report happiness in their marriage are *far* less likely to have an affair. This makes intuitive sense.

The strongest indicator appears to be **rating**. To test some different prediction models I will use rating as the training variable in the following section.

knn

Begin by testing KNN model.

```
Affairs %>% as_tibble()
```

```
## # A tibble: 601 x 10
##    affairs gender   age yearsmarried children religiousness education occupation
##      <dbl> <fct>  <dbl>        <dbl> <fct>            <int>     <dbl>      <int>
##  1       0 male      37        10    no                   3        18          7
##  2       0 female    27         4    no                   4        14          6
##  3       0 female    32        15    yes                  1        12          1
##  4       0 male      57        15    yes                  5        18          6
##  5       0 male      22         0.75 no                   2        17          6
##  6       0 female    32         1.5  no                   2        17          5
##  7       0 female    22         0.75 no                   2        12          1
##  8       0 male      57        15    yes                  2        14          4
##  9       0 female    32        15    yes                  4        16          1
## 10       0 male      22         1.5  no                   4        14          4
## # ... with 591 more rows, and 2 more variables: rating <int>, y_n_affairs <fct>
```

```
table(Affairs$rating)
```

```
##
##   1   2   3   4   5
##  16  66  93 194 232
```

Viewing a table of those *not* having affairs for each rating

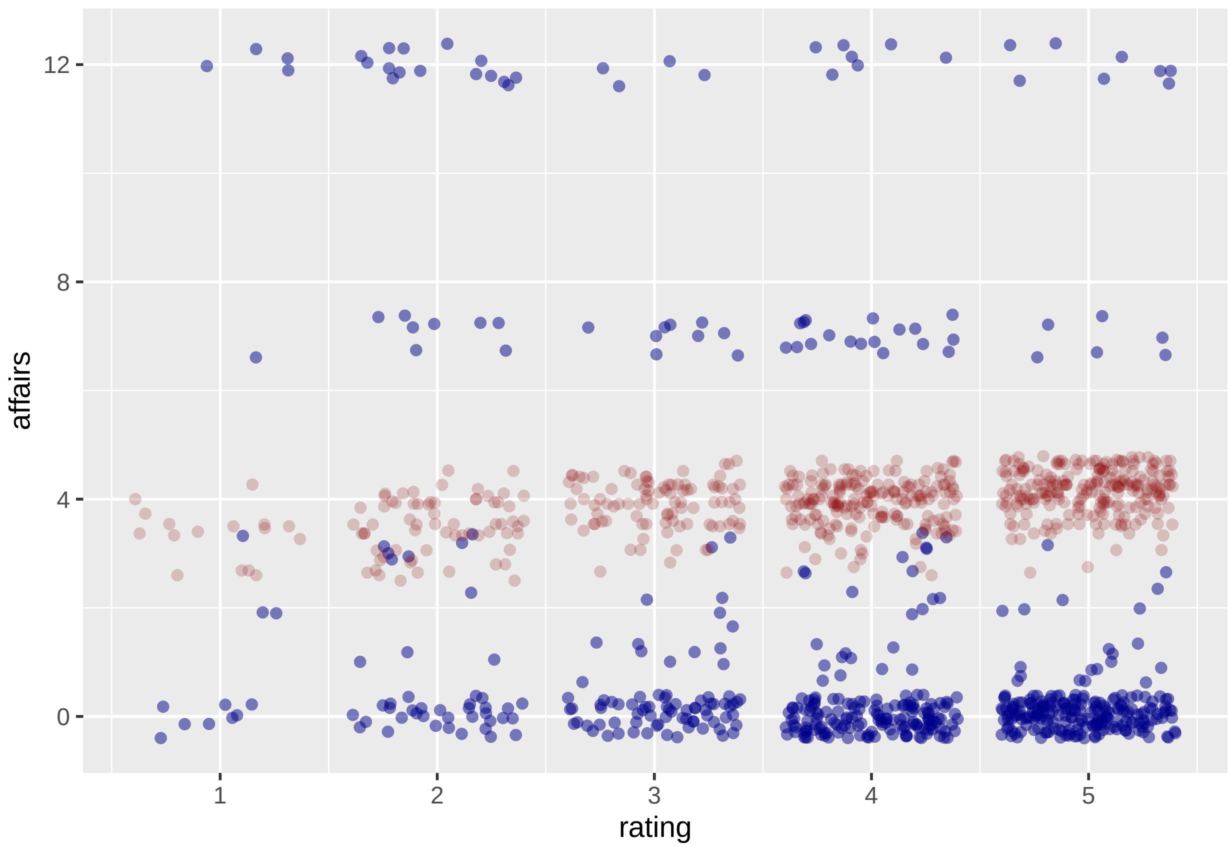
```
Affairs <- select(Affairs, -gender, -children, -education, -occupation)
```

Removing the least relevant factors to simplify the process

```
set.seed(1983, sample.kind = "Rounding")
```

```
fit <- train(rating ~ .,  method = "knn",
             tuneGrid = data.frame(k = seq(1, 15, 2)),
             data = Affairs)
```

```
Affairs %>%
  mutate(y_hat = predict(fit)) %>%
  ggplot() +
  geom_jitter(aes(rating, affairs), col = "darkblue",alpha = 0.5) +
  geom_jitter(aes(rating, y_hat), col="darkred", alpha = 0.2)
```



Comparing the actual data (blue) vs the KNN predictive model (red). It doesn't seem to match the data very well at all.

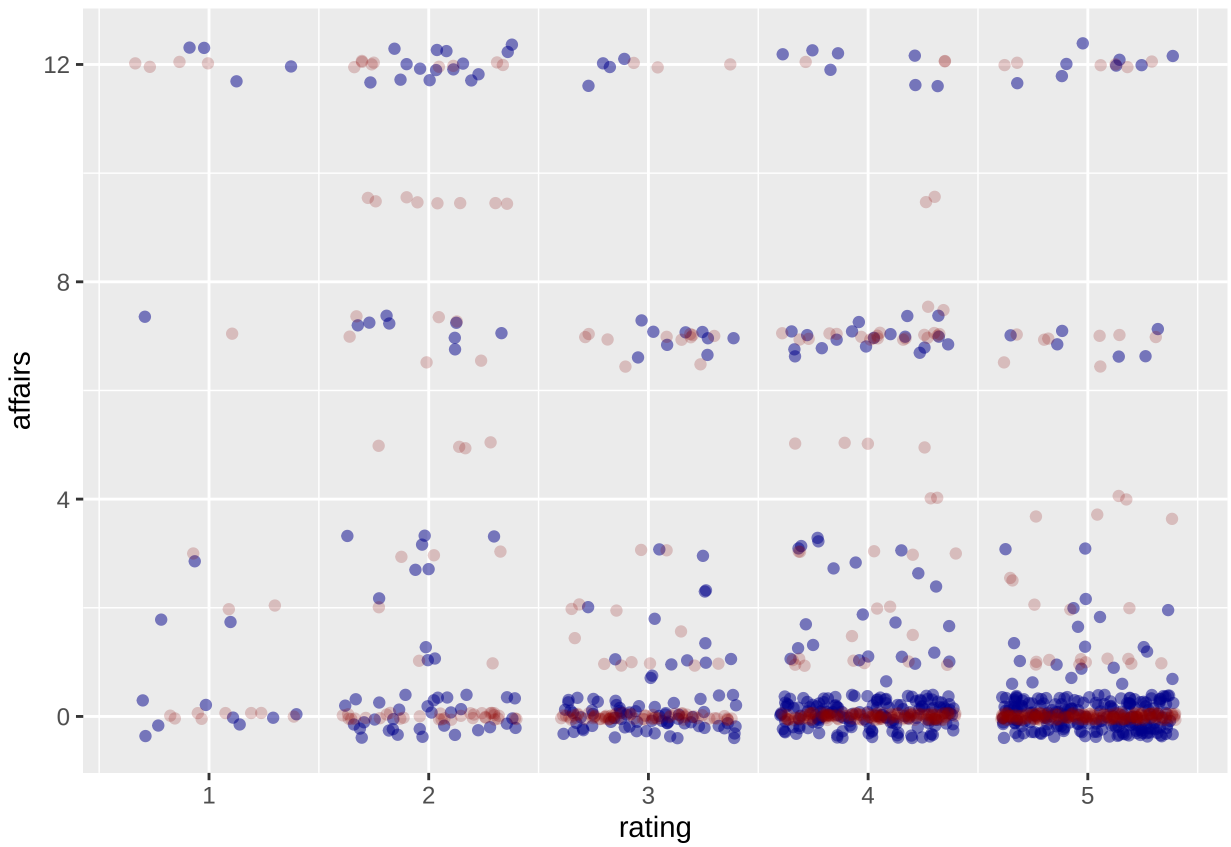
regression tree

See how well the Regression Tree predicts the correct outcomes.

```
fit <- rpart(affairs ~ ., data = Affairs)

fit <- rpart(affairs ~ ., data = Affairs, control = rpart.control(cp = 0, minsplit = 2))
```

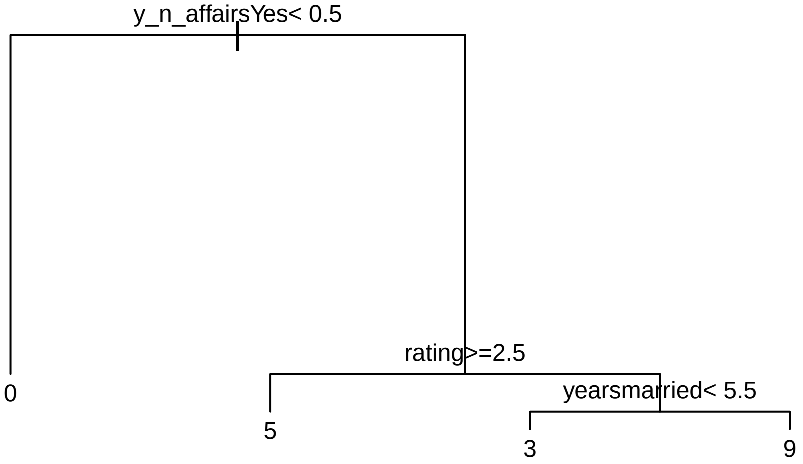
```
Affairs %>%
  mutate(y_hat = predict(fit)) %>%
  ggplot() +
  geom_jitter(aes(rating, affairs), col = "darkblue", alpha = 0.5) +
  geom_jitter(aes(rating, y_hat), col="darkred", alpha = 0.2)
```



The resulting graph lines up with the data very well, in both frequency and spacing.

```
train_rpart <- train(affairs ~ ., method = "rpart", tuneGrid = data.frame(cp = seq(0, 0.05, len = 25)),
```

Building a decision tree.

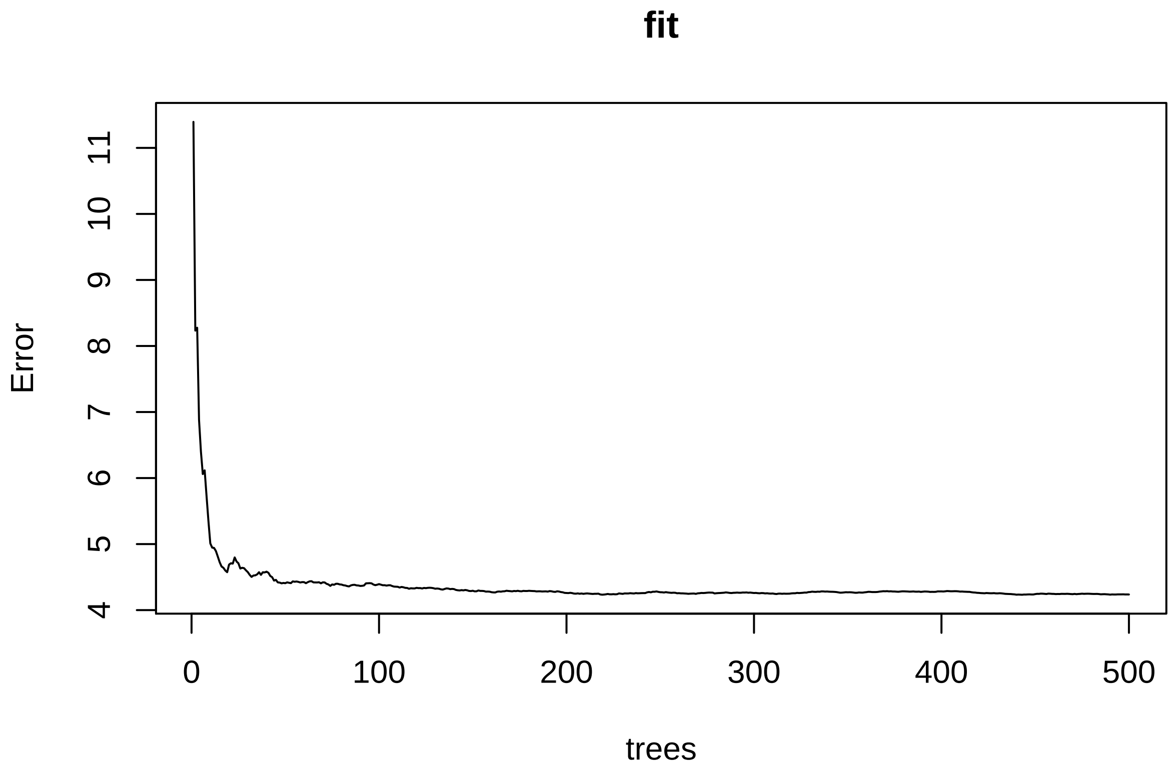


If “Yes” is greater than 0.5. the have an affair branch is followed. After that, if one’s marriage rating is lower than 2.5 they end up averaging 5 liaisons. A higher rating than 2.5 leads to using years married as the next factor, showing that of that group 6 years or more of marriage leads to an even higher number of infidelities.

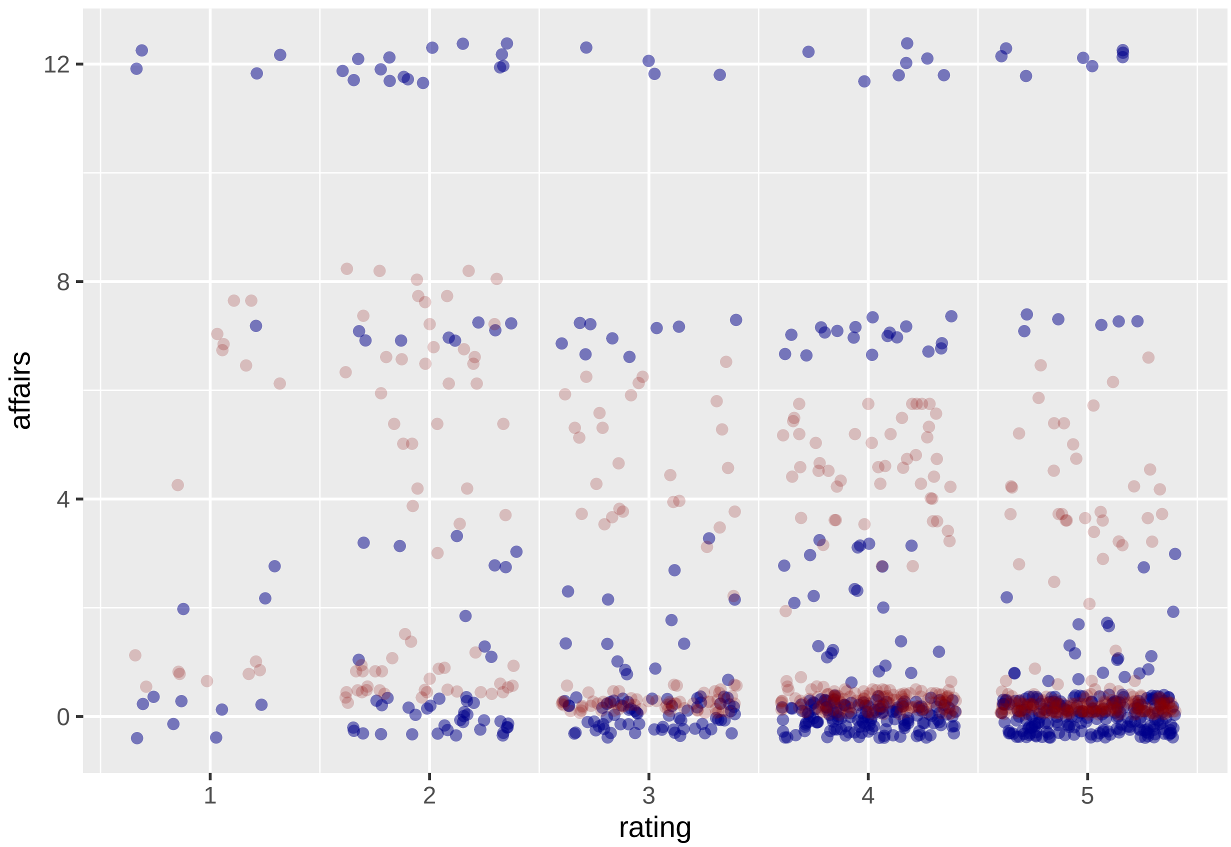
random forest

Testing Random Forest

```
library(randomForest)
fit <- randomForest(affairs ~., data = Affairs)
plot(fit)
```



```
Affairs %>%
  mutate(y_hat = predict(fit, newdata = Affairs)) %>%
  ggplot() +
  geom_jitter(aes(rating, affairs), col = "darkblue", alpha = 0.5) +
  geom_jitter(aes(rating, y_hat), col="darkred", alpha = 0.2)
```



The Random Forest model seems to work relatively well *except* that it does not accurately predict the higher end of the data.

RESULTS

We were able to find the strongest indicators of potential infidelity, with how one rates their happiness in their marriage being the most significant factor. Also, using the data we could show probabilities of affairs over the major factors at multiple intervals. The longer a person is married, the more likely they are to have an affair, but religion and age are also mitigating factors. As far as potential prediction models, overall the Regression Tree appears to work best at being an accurate predictor of the likelihood of affairs.

CONCLUSION

The dataset and its revelations are a fun exploration, but there are inherent flaws trying to use it for any sort of real world predictions. The biggest flaw is that all of the data is self-reported, which creates built in bias. Even reporting anonymously doesn't guarantee honesty in all the answers. It is also data gathered 50 years ago. Many societal shifts have occurred in that time and perhaps some of the factors are not reflective of modern standards. Lastly, it is a very small dataset and may not be truly representative of the general population. This is the main reason no attempt was made to train and test predictions, though perhaps with a more robust dataset this could be done effectively.