Predicting Frequency of Extramarital Affairs

Sarah Li (#60136959), Jun Won (Lakon) Park (#79453940)

What influences married individuals to become involved in affairs? There is speculation about what influences how likely someone will cheat on their spouse, and we would like to confirm speculations with real data. With this project, we are aiming to predict whether an individual is involved in extramarital affairs. If they are involved, we would like to also predict the frequency of involvement. In order to build our predictive model, we will also explore how each predictor influences the chances of infidelity in a marriage.

1 Introduction

Data

The "Affairs" data is from a survey conducted by Psychology Today in 1969. The response variable is **affairs**, representing how often the individual engaged in extramarital sexual involvements in the past year. There are 5 ordinal, 1 continuous, and 2 categorical predictor variables. The survey conducted collected information in ranges for simplicity.

Value	Frequency of affairs
0	None
1	Once
2	Twice
3	3 times
7	4-10 times
12	Monthly or more often

Predictor variables:

- Sex (0 = female, 1 = male)
- Children (0 = no, 1 = yes)
- Age (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)
- 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)
- How religious (5 = very, 4 = somewhat, 3 = slightly, 2 = not at all, 1 = anti)
- 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 rating (1-7 according to **Hollingshead classification scale**, reverse numbered)
- Marriage rating (5 = very happy, 4 = happier than average, 3 = average, 2 = somewhat unhappy, 1 = very unhappy)

Areas of Interest

Many may assume that marriage stability, or "sunk-cost" of a marriage decreases the likelihood of affairs. In that sense, it is worth investigating whether years married, having children, or marriage satisfaction are influential. Moreover, variables which may not obviously contribute to infidelity, such as religiousness and socioeconomic status (contributed to by education level and occupation rating), might have some influence.

In order to find out which predictors play the biggest role in determining if a married individual is involved in an affair, we would like to explore association between variables and determine if we should use those variables with the strongest correlation in our predictive model.

Finally, we are interested in creating a model that can predict if an individual has been involved in extramarital affairs using the variables found to contribute to infidelity. To build on this, we will also try to predict the frequency of affairs.

2 Exploratory Data Analysis

In order to investigate what influences cheating in a marriage, ignoring frequency, we define **cheat** as a binary variable based on the **affairs** column.

```r
data("Affairs", package="AER")
df <- Affairs
rownames(df) <- NULL
# cheat = 0 if frequency of affairs = 0
# cheat = 1 if frequency of affairs > 0
df$cheat <- ifelse(df$affairs > 0, 1, 0)
df$cheat <- factor(df$cheat)

table(df$affairs)
```

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

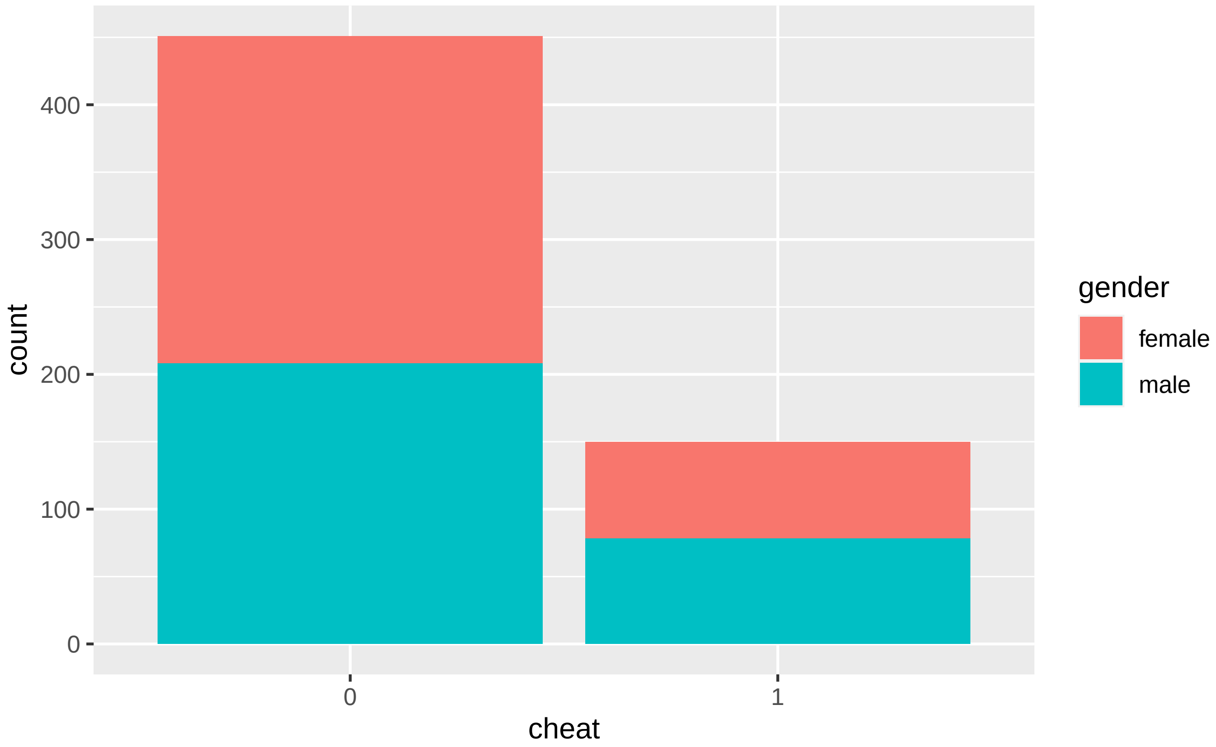


Figure 1: Cheat = 1 shows the number of females and males in the group of respondents who have cheated within a year before taking the survey. Cheat = 0 shows the number of those who did not.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  as.numeric(df$cheat) by df$gender
## W = 43056, p-value = 0.106
## alternative hypothesis: true location shift is less than 0
```

- From Figure 1, the proportion of individuals involved in affairs by gender seems to be equal which suggests no apparent relationship between gender and affairs.
- This is contradictory to many numerous sample surveys[1] that say more men cheat than women.
- The Mann–Whitney U test[2] (Wilcoxon rank sum test with continuity correction) can be conducted by treating **cheat** as an ordinal variable, since we are not interested at frequency at this point, to confirm if there is a difference between males and females. The alternative used is "less" because we might guess that women cheat less than men. The p-value is found to be greater than 0.1, so the difference between genders is insignificant.

[1] Reference: Summary of results from General Social Survey
[2] Reference: Mann-Whiteney U test in R

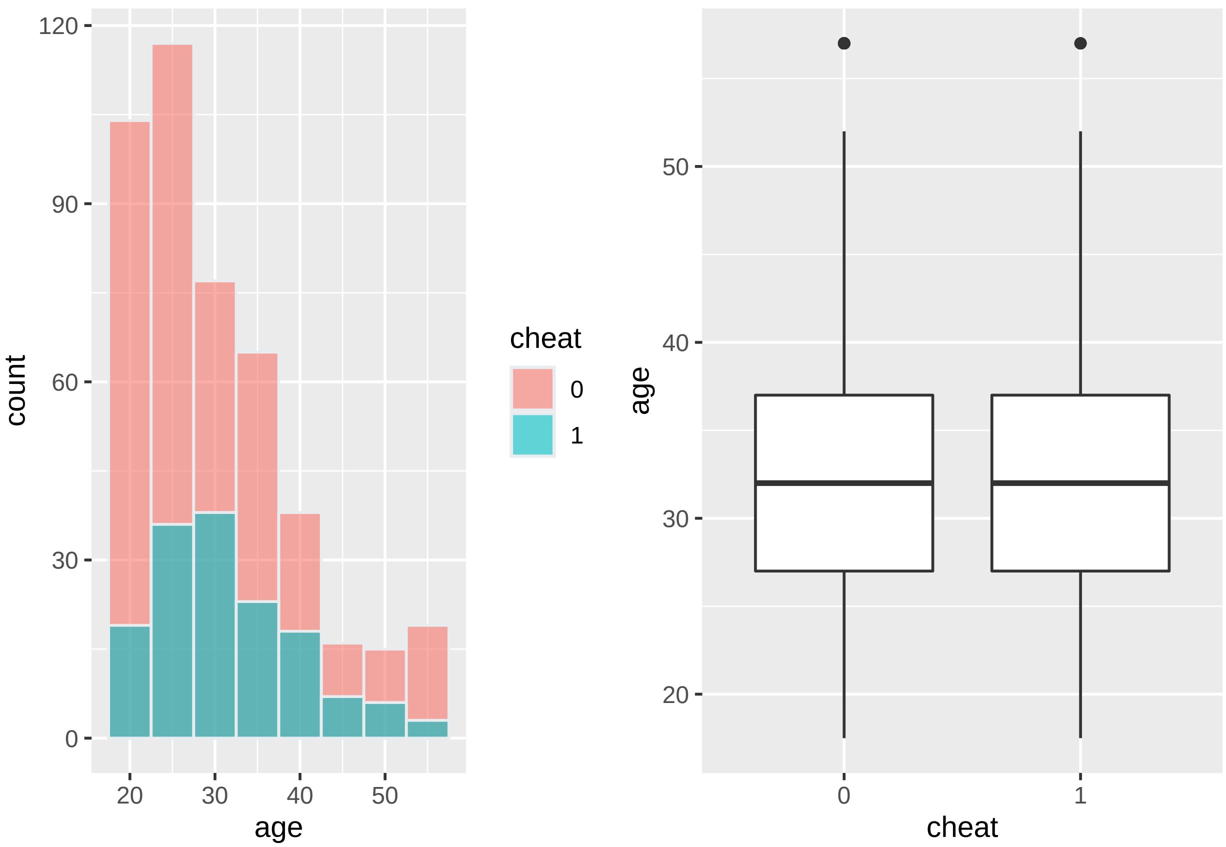


Figure 2: (Left) The proportion of cheating individuals distributed over age. (Right) Compares the distribution of ages in the cheating group and non-cheating group.

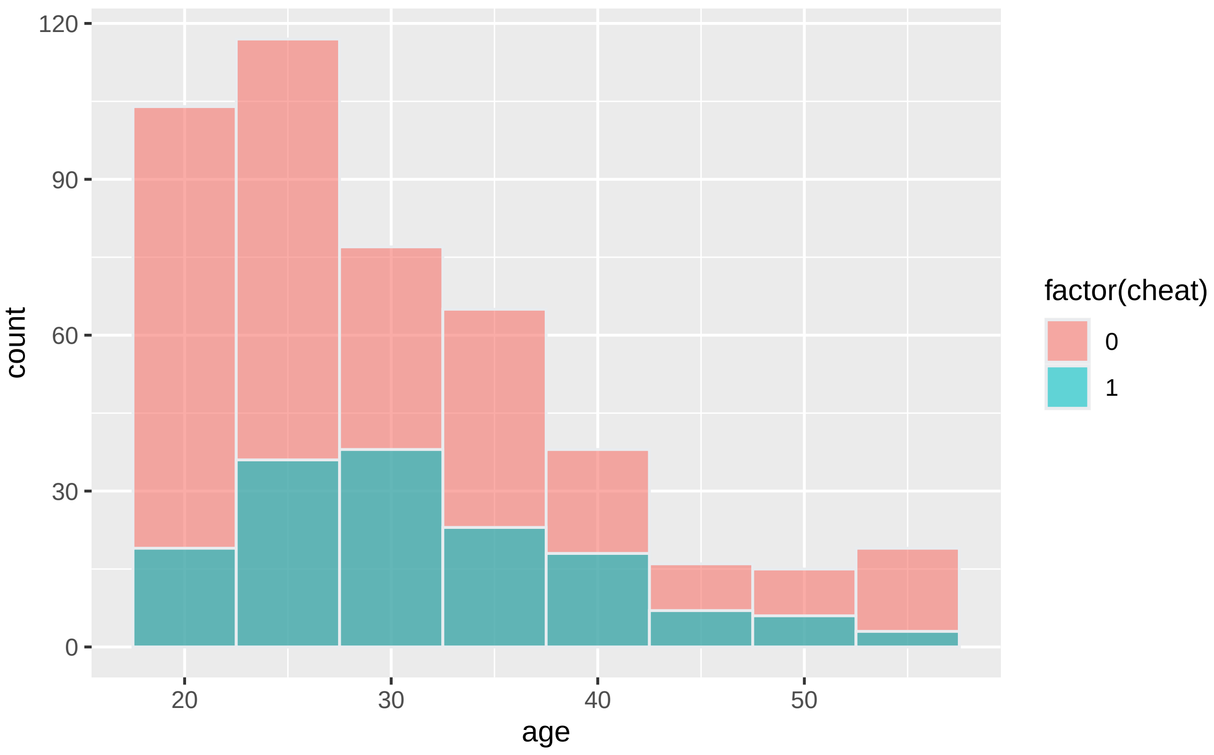


Figure 3: Distribution of individuals involved in affairs over all ages of respondents.

```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.50   27.00   32.00   32.49   37.00   57.00
```

```
summary(df$age[df$cheat == 0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.50   27.00   32.00   32.18   37.00   57.00
```

```
summary(df$age[df$cheat == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.50   27.00   32.00   33.41   37.00   57.00
```

- The age of individuals involved in affairs mostly distributed between 22-33.
- It is interesting to note that the age distribution of people involved in affairs is almost identical to the age distribution of people not involved in affairs. The average age of people involved in affairs is slightly higher than those who are not.
- The overall age distribution of the sample resembles a right-skewed, log-normal distribution.

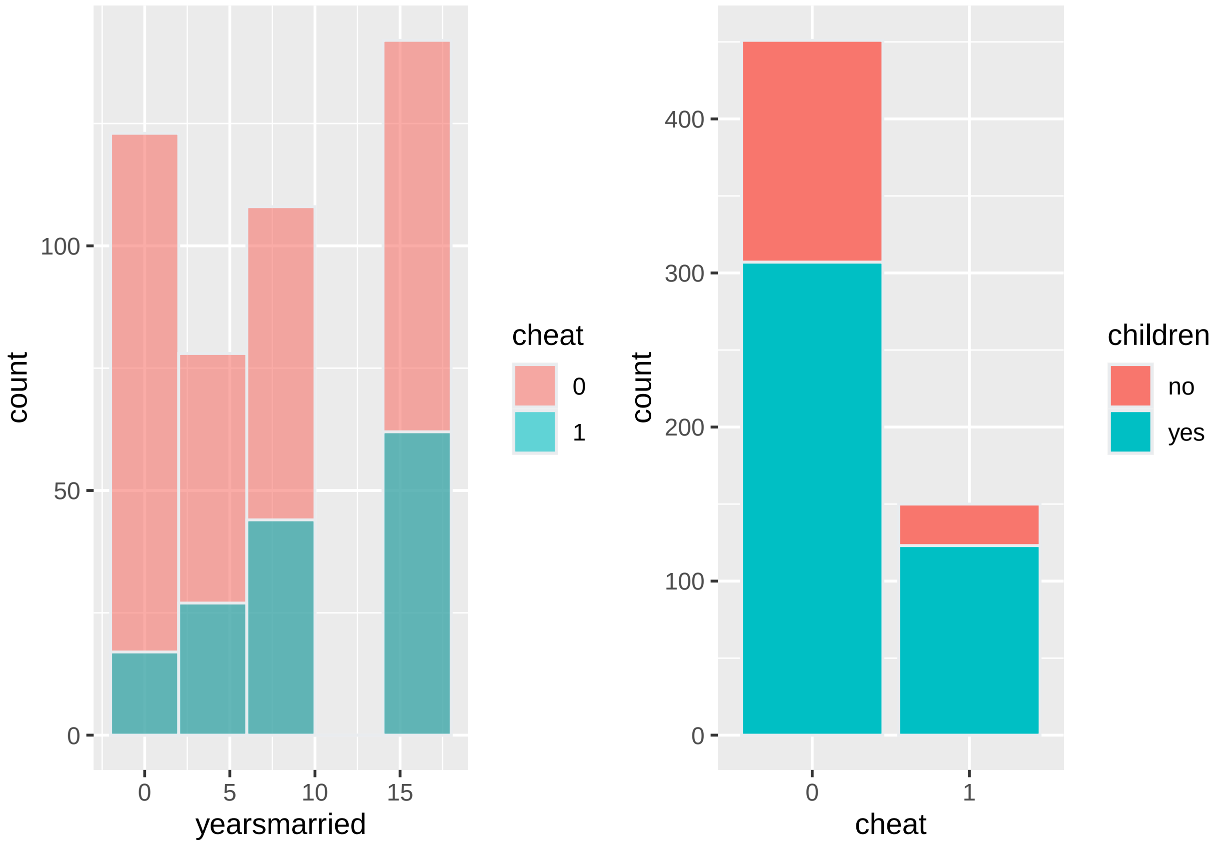


Figure 4: (Left) The proportion of cheating individuals distributed over number of years married. (Right) The proportion of individuals with children in the cheating group (cheat=1) and non-cheating group (cheat=0).

- The proportion of individuals involved in affairs increases as number of years married increases.
- The proportion of individuals involved in affairs most likely have children.

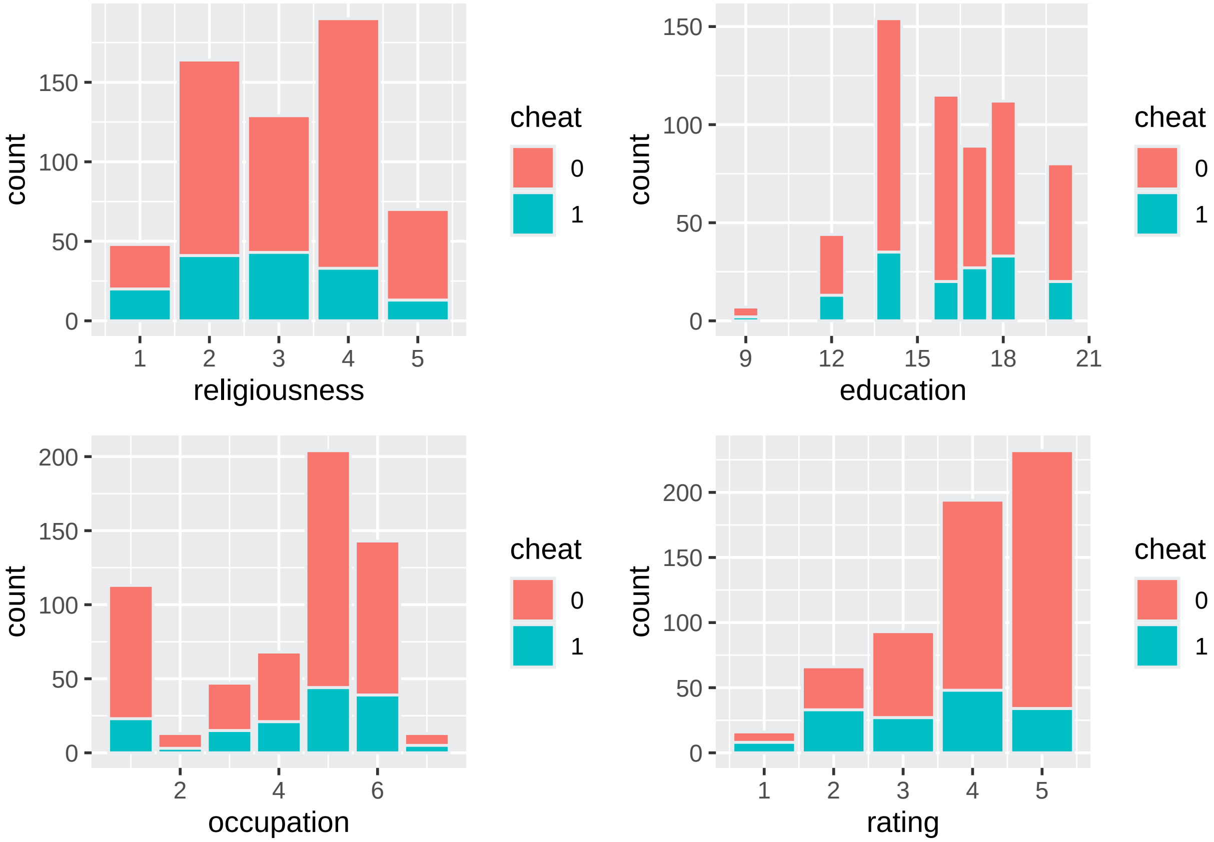


Figure 5: We use these four histograms to inspect whether the proportion of cheating individuals distributed over other possible predictor variables. There is no evident pattern between affairs and education, occupation and rating. Religiousness stands out as a potentially useful predictor of extramarital affairs.

- The distribution of individuals involved in affairs by religiousness form an approximate normal distribution.
- Note that the proportion of individuals that are involved in affairs is significantly lower for the most religious group.

3 Deeper analysis

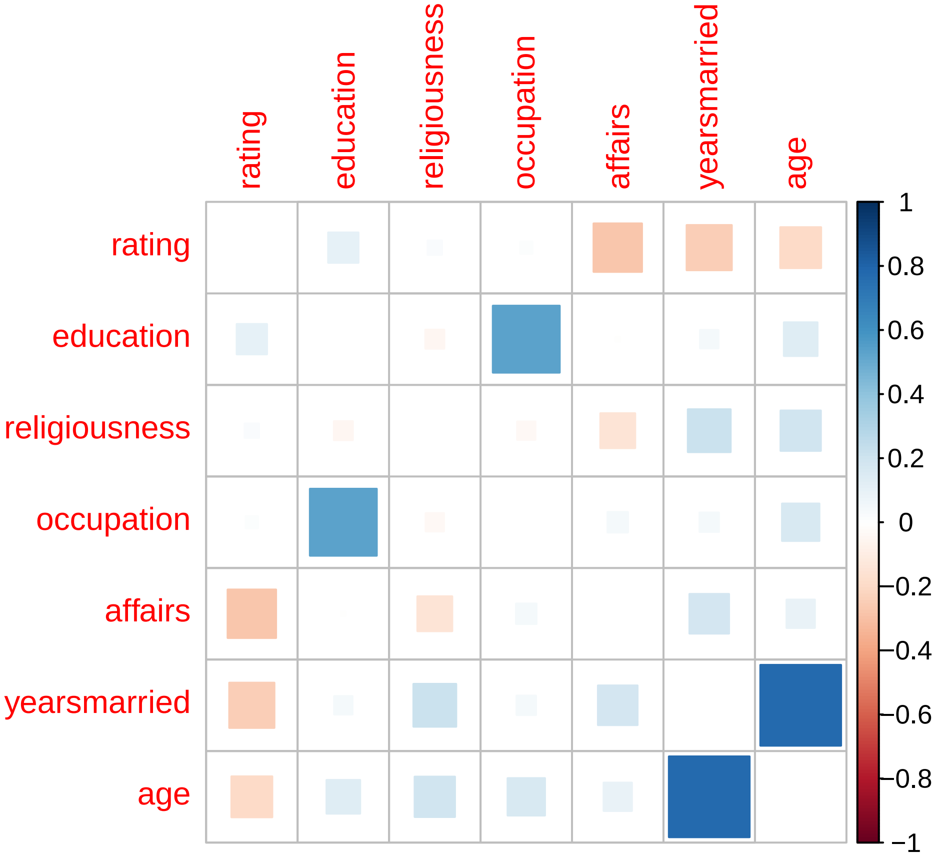


Figure 6: Correlation plot of all ordinal variables in the ‘affairs’ dataset with blue indicating positive correlation and red indicating negative correlation.

The correlation plot suggests that the marriage rating has the greatest negative correlation with affairs. Hence, you are more likely to cheat on your spouse if the marriage rating is low. This makes sense since people cheat when they are unsatisfied with their current partner or they wouldn’t cheat.

The next predictor that is most correlated with affairs is years marriaged with positive correlation. The aligns with our insight from plots earlier. A plausible explanation might be as the years of marriage increases, the relationship becomes more boring and people are likely to draw their attention to other potential partners.

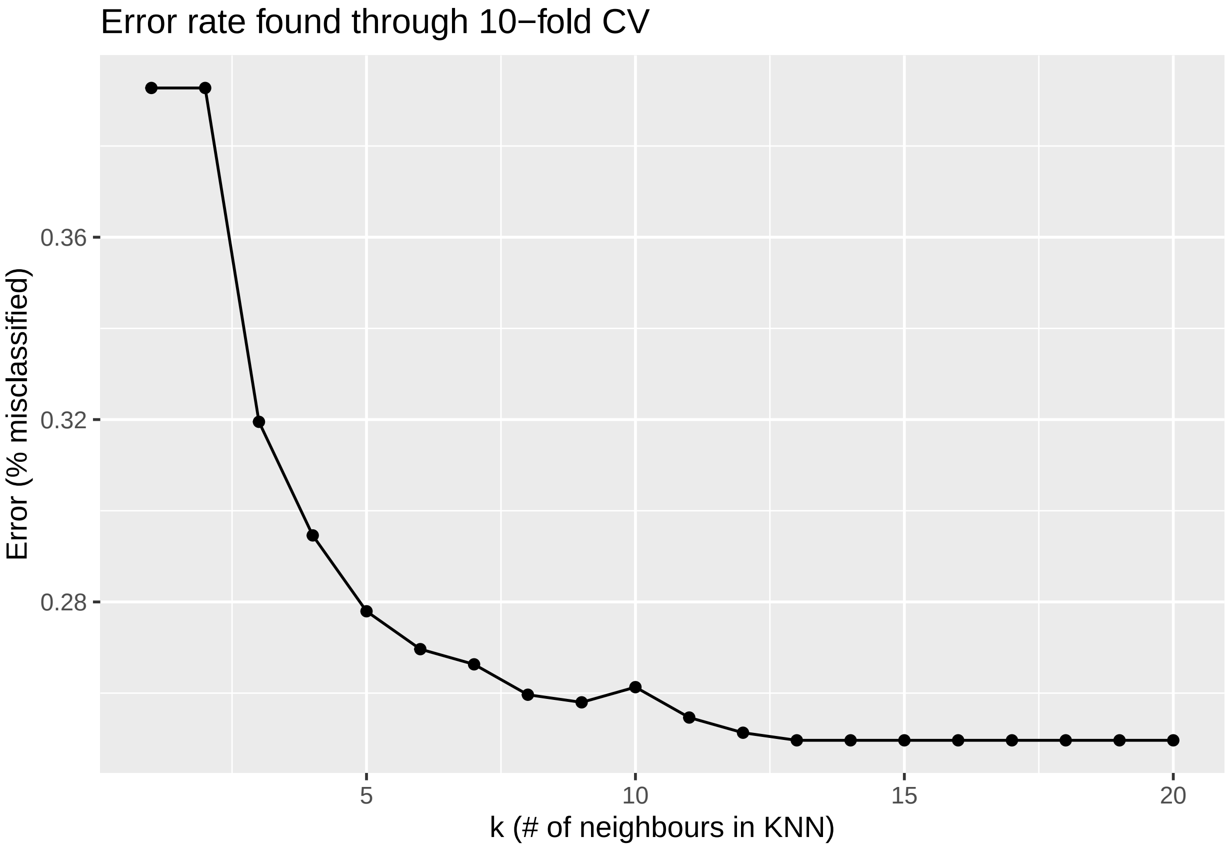
The negative correlation between religiousness and affairs suggests that the more religious you are, the less likely you are going to cheat. A possible reasoning behind this is that oftentimes, cheating is condemned in religion. The correlation of other predictors are too insignificant.

4 Classification

Since the frequency of extramarital affairs is split into ranges, this can be considered a multi-class classification problem. If our model can effectively predict the frequency class of affairs, it can also predict whether or not an individual would be involved in an affair (since the frequency would be 0 or greater than 0). Hence, a model which can predict frequency is desirable. We also do not know the shape of the decision boundaries for certain when we have 8 dimensions to consider.

KNN

The KNN classification algorithm is capable of classifying similar samples together. However, since our data involves categorical features, we cannot use euclidean distance. Hence, we propose to use gower's distance[3], which is a measure of how different two observations are. The smaller the distance, the closer you are and vice-versa. The mode[4] of the neighbouring classes will be used to classify new observations.



Using the elbow method[5], $k = 13$ is the optimal number of nearest neighbors to use with CV score of around 0.2496. One reasoning behind unsatisfactory performance of the model may be due to the unequal distribution of affairs which creates an imbalanced dataset. Only 150 out of the 601 surveyed reported to be in an extramarital affair in the past year. If more cheating individuals are surveyed, we may have more useful observations for making predictions.

Due to the imbalanced nature of the data, we will use a modern sampling technique to see if we can improve our classification accuracy. First, we will formulate our problem into a binary classification problem such that we will try to predict if an individual is involved in an affair or not. In this way, we can try to model the

[3] Distance between two observations is obtained as a weighted sum of differences for each variable. By default, the weights are all equal to 1. Reference: KNN with Gower distance

[4] Reference: How to find the statistical mode?

[5] We choose k based on the largest value at which the error rate decreases. Reference: Elbow Method in Supervised ML

pattern for people involved in affairs.

Furthermore, we will use other types of predictive models which can give us an idea of which factors contribute the post to predicting affairs. With KNN, it is unclear which variables had the greatest influence.

5 SMOTE Sampling

SMOTE (Synthetic Minority Over-sampling Technique)[6] is an oversampling technique which oversamples from the minority class using nearest neighbours to generate a balanced dataset. We expect the algorithm to perform much better with the newly generated, balanced dataset.

Before SMOTE balancing:

Did not cheat	451
Did cheat	150

After SMOTE balancing:

Did not cheat	451
Did cheat	450

The new dataset contains 450 observations of class 1 which is 3 times the original number of observations we had before so the dataset is now balanced.

6 Random Forests (Binary)

Random Forest is an ensemble classification algorithm which is capable of classifying similar data together. It can handle categorical variables well and has lower computation time than KNN, as we do not need to calculate the distance between observations.

The cross-validation score is about 16.65% with out-of-bag error of approximately 16%. So far, we can more accurately determine whether an individual cheats, rather than the individual's frequency of affairs. In the next section, we will try using a random forest model for multi-class classification to compare with our KNN model.

From the variable importance plot below (Figure 7), we see that rating and religiousness are the best in the sense that they have the highest mean decrease in gini index[7]. This aligns with our initial analysis that rating and religiousness would be highly correlated with cheating. Children and gender have the lowest mean decrease in gini index. This conclusion aligns with our initial analysis. We observed an evident pattern and correlation between rating and religiousnes while we did not see any for children and gender.

[6] In binary prediction, we oversample the cheating group which is a straightforward usage of the SMOTE function. For multi-class prediction, we perform one vs. all balancing for each nonzero 'affairs' group. Reference: Multi-class SMOTE example

[7] Variables with a higher mean gini decrease lead to tree splits with descendant nodes that are more "pure", where fewer observations are misclassified. Reference: Gini importance

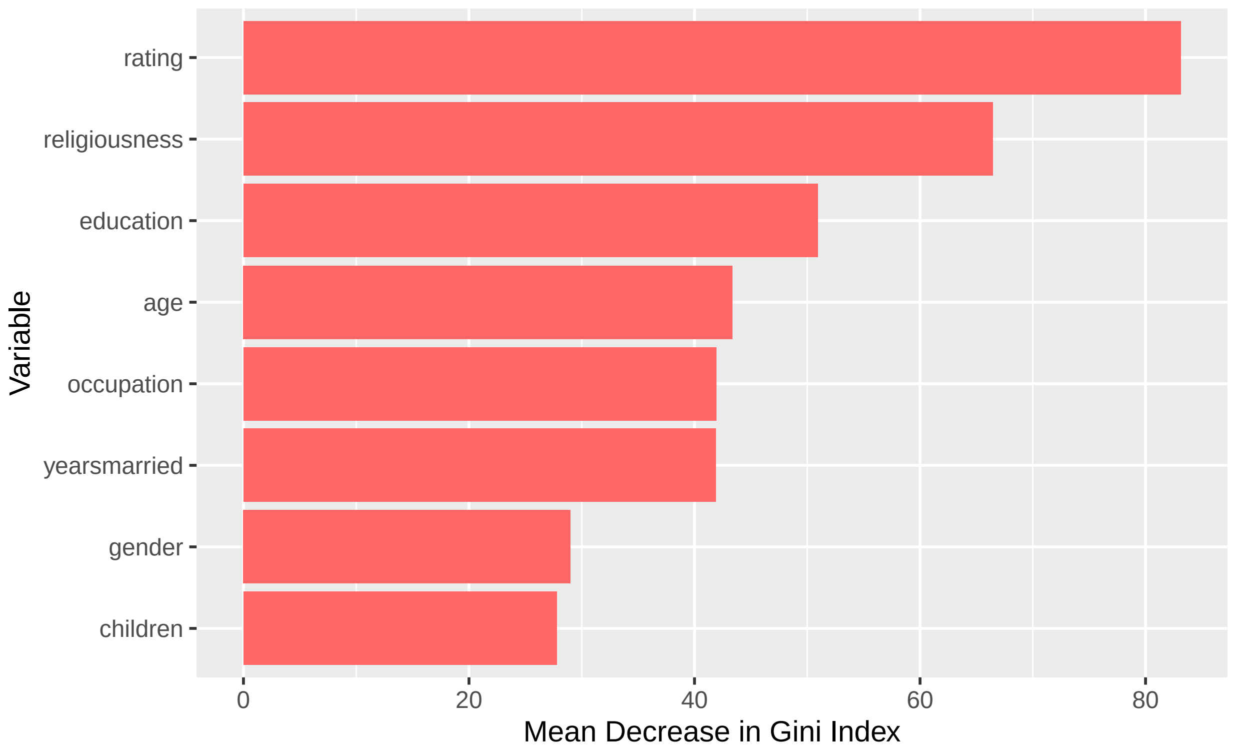


Figure 7: Importance of predictors in the binary random forest model measured by mean decrease in Gini index.

7 Random Forests (Multiclass)

We now perform smote oversampling technique on the variable affairs to solve our original imbalanced dataset problem and perform random forest classification to predict the frequency class of affairs.

Before SMOTE balancing:

Frequency of affairs	0	1	2	3	7	12
# samples	451	34	17	19	42	38

After SMOTE balancing:

Frequency of affairs	0	1	2	3	7	12
# samples	451	442	459	456	462	456

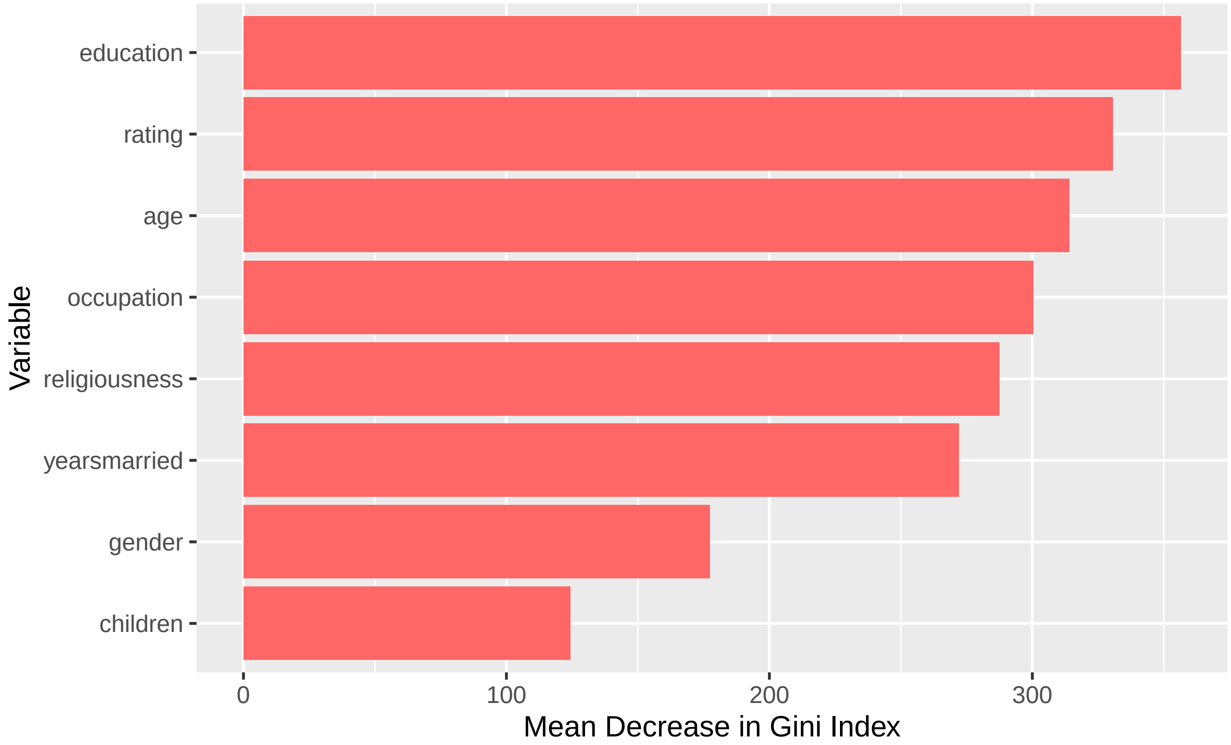


Figure 8: Importance of predictors in the multi-class random forest model measured by mean decrease in Gini index.

The cross-validation score is about 6.5% with out-of-bag error of approximately 6.8%. This is a substantial improvement from the previous KNN algorithm.

Surprisingly, for multiclass classification, the best predictor with the most mean decrease in gini index is education. Hence, we can assume that education is an importance variable in predicting frequency of affairs.

Our conclusion about rating still holds true for multi-class classification as it did for binary classification. However, religiousness has dropped from being the second most important variable from our binary classification model to being the fifth most important in our multi-class classification model. This suggests that religiousness is only useful in predicting whether an individual has cheated.

8 Conclusion

From our analysis, we have discovered that rating and religiousness are the most important factors in determining whether an individual will be involved in an affair. In addition to these variables, education is highly associated with the frequency of affairs. Gender and whether or not one has children are not particularly important in predicting affairs.

We can make inferences about why our most important predictors impact the frequency of affairs. Rating was shown to have the greatest negative correlation with number of affairs, which aligns with our results. This suggests that individuals who report less satisfaction with their marriage are more likely to, and if so, more frequently cheat on their spouse. Religiousness has a small negative correlation with number of affairs, so non-religious people simply may be more likely to cheat. The importance of education could be explained if we assume that more academically-inclined individuals have better foresight and can predict the consequences of being involved in an affair.

On the other hand, number of years married has the next greatest positive correlation with number of affairs, but ‘yearsmarried’ does not have significant predictive power in our random forest models. Besides this discrepancy with ‘yearsmarried’, the correlation plot gave good indicators of which variables would be important in predicting affairs.

Using the Random Forest algorithm, we were able to build a predictive model with cross-validation score of 16% for determining whether an individual has been involved in an extramarital affair. When using Random Forest to predict the frequency of affairs, we achieved an error of approximately 6.5%. Considering we built the model based on 601 samples with smote sampling, the result is notable.

There is a lot of room for improvement for our model. For instance, if more real data is collected from more individuals, we are certain that we can improve the accuracy of our classification model. We can also try to use deep learning to see if it yields a better model, given that we have enough computational power and data. A caveat of this method would be losing the interpretability of variables which we have with random forests. Moreover, if we are able to collect a balanced sample of data the model may be more accurate compared to our usage of oversampled data produced by SMOTE. We could use a stratified sampling method which ensures that we collect an equal number of each range of frequencies. Since our goal is classification, this will improve predictive performance for the minority classes.