Predicting Divorce

Celeste Chen and Sarah Unbehaun

May 8, 2017

Project Overview

Given the evolving discourse and perspectives on marriage, we seek to predict the probability that a couple will get a divorce based factors such as social views; attitudes about pornography, premarital sex or extra-marital sex; and fundamentalism of religious views. Using data from the General Social Survey, we will train an algorithm on a generational cohort from the years 1996-2014 to determine which types of respondents are most likely to experience divorce. We will test several methods of analyzing the data for predictions: (1) logistic regression, (2) decision trees, (3) random forest and (4) clustering. Based on the accuracy results of each method, we will determine the best model for prediction.

Data

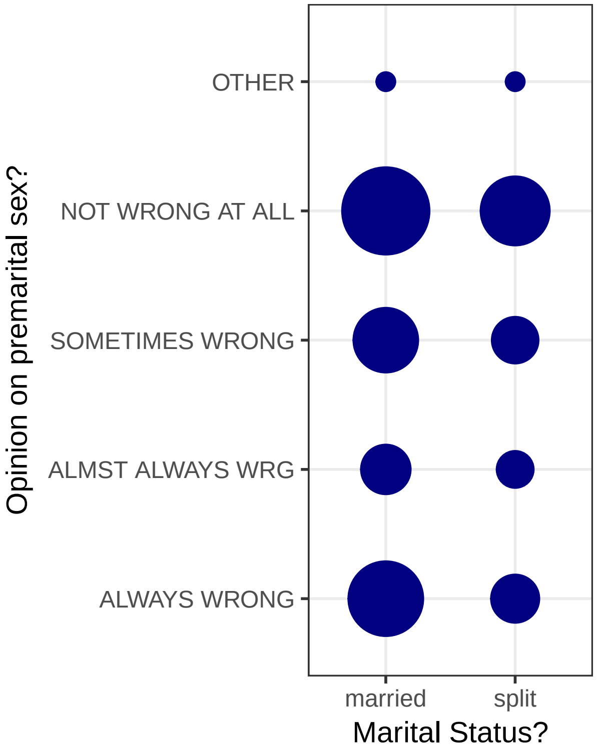
Our dataset will be taken from the General Social Survey (GSS) and focus on observations from 1996 to 2014. According to the GSS website >The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. Altogether the GSS is the single best source for sociological and attitudinal trend data covering the United States.

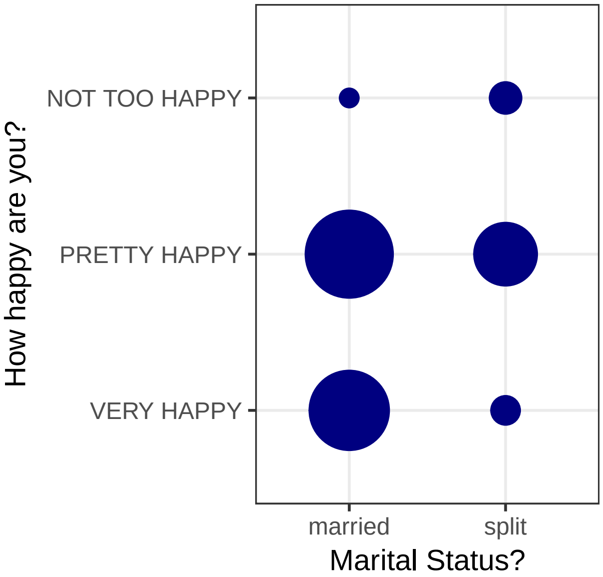
We have extracted the variables on the following characteristics for our analysis: opinions on sex education in the public schools; region of interview; opinions on premarital sex; opinions on extramarital sex; views on pornography laws; marital status; political party ID; how fundamental is respondent; number of children; age of respondent; ever been widowed; highest educational degree achieved; respondent's labor force status; whether the respondent thinks of self as liberal or conservative; general self-reported happiness; can people be trusted; self-reported class; self-reported income; region at age 16; who the respondent lived with at age 16; religion at age 16; frequency of church attendance; does the respondent believe other people would try to take advantage or be fair; does the respondent believe other people usually try to be helpful; size of city or town of residence; confidence in scientific community; self-reported job satisfaction; elf-reported satisfaction with financial situation; whether abortions should be legal for married women who don't want more children; whether abortions should be legal if the woman is single and doesn't want to marry; and whether the respondent has seen an x-rated movie in the last year.

Not all of these questions were asked of all respondents. Three ballots were administered, so our data had to be analyzed in groups corresponding to each ballot. We also restricted our data to those who were married or "split," split being those who were either divorced or separated. The number of observations for each category are shown below:

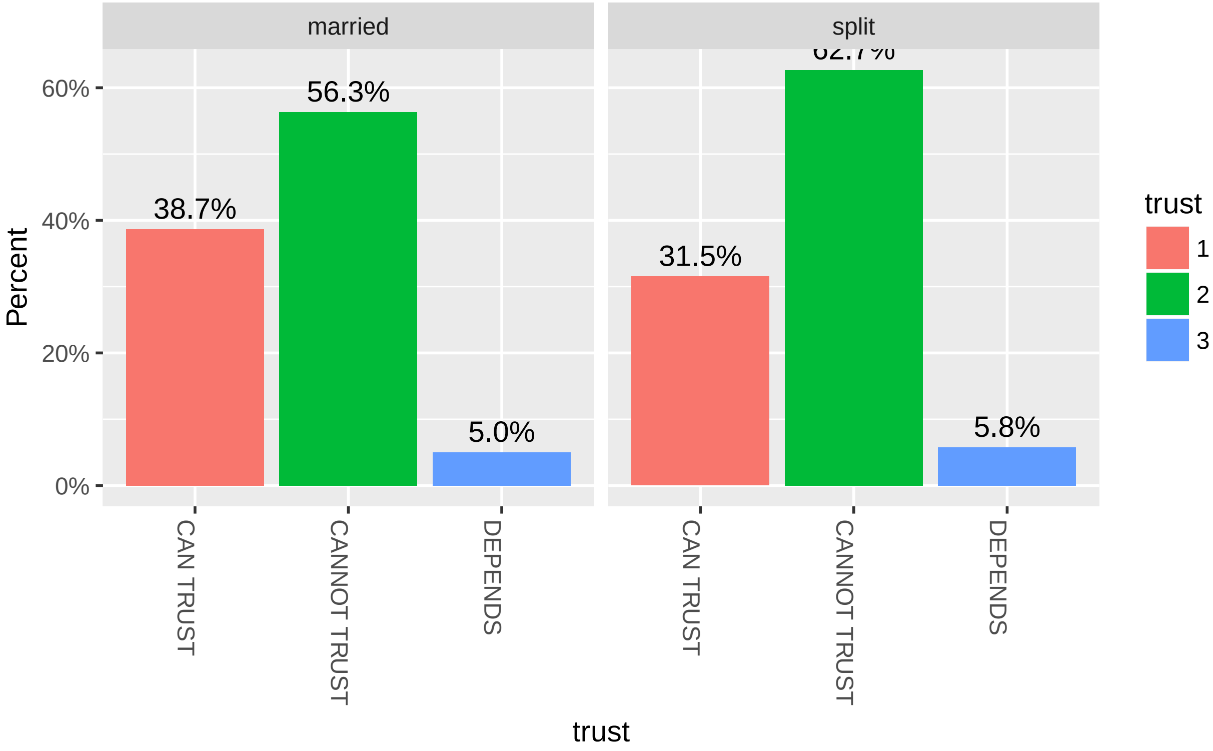
Ballot	A	B	C	Total
Married	4008	4025	4054	12853
Split	1607	1665	1667	5238
Total	5615	5700	5721	18091

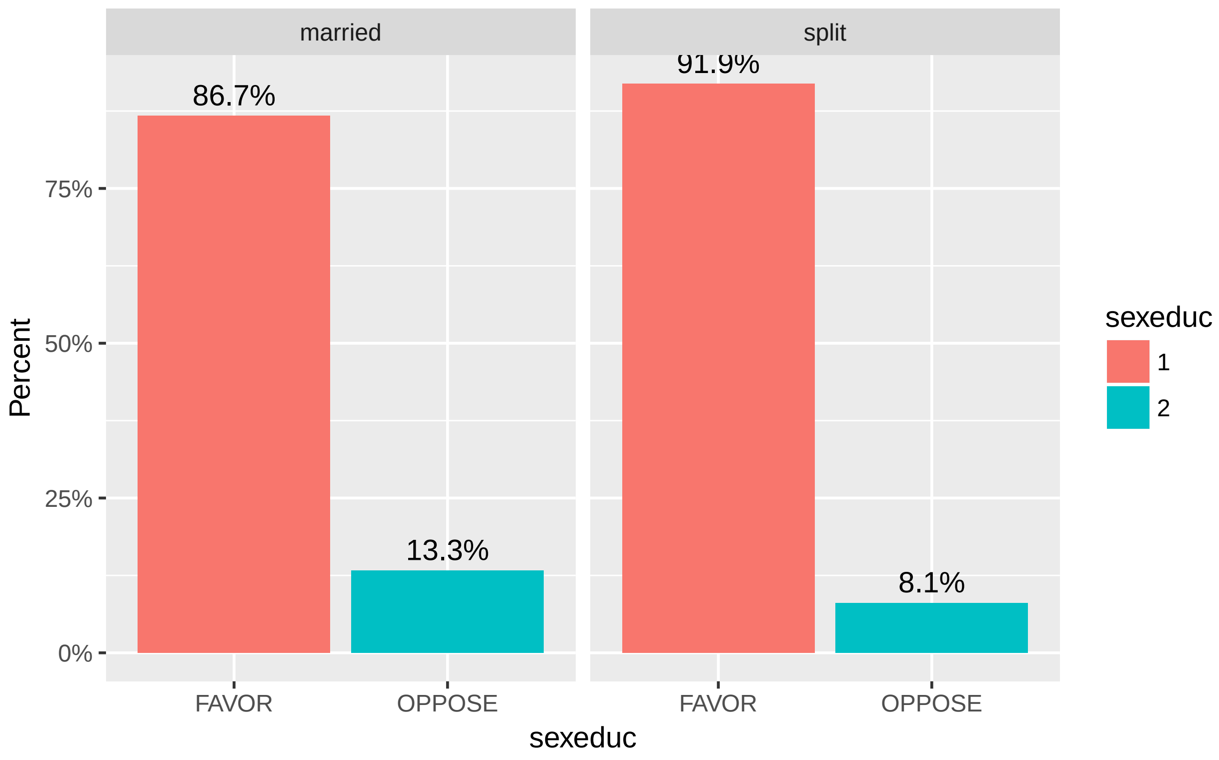
A few examples of the differences between the married and split groups can be seen in the graphs below. First, respondents who are split appear slightly more likely to say that premarital sex is not wrong at all. They are significantly less likely to say the they are very happy.





Also, respondents who are split are more likely to say that others cannot be trusted and slightly more likely to support sexual education in schools.





Analysis Method 1: Logistic Regression

We first tried logistic regression models, using all relevant variables for each ballot group as well as trying models using subsets of variables: demographics only, social/political views only, and removing non-statistically significant variables based on the the complete regression. Each ballot group was separated into train, validate and test groups. The overall accuracy for each model based on the test data, calculated as the total correct predictions divided by the total number of observations, is reported below. The accuracy was fairly similar across models.

Ballot	Model	Overall Accuracy
A	all variables	0.75
A	demographics	0.76
B	all variables	0.74
B	stat. signif.	0.74
C	all variables	0.72
C	pol/social	0.71

Analysis Method Two: Decision Trees

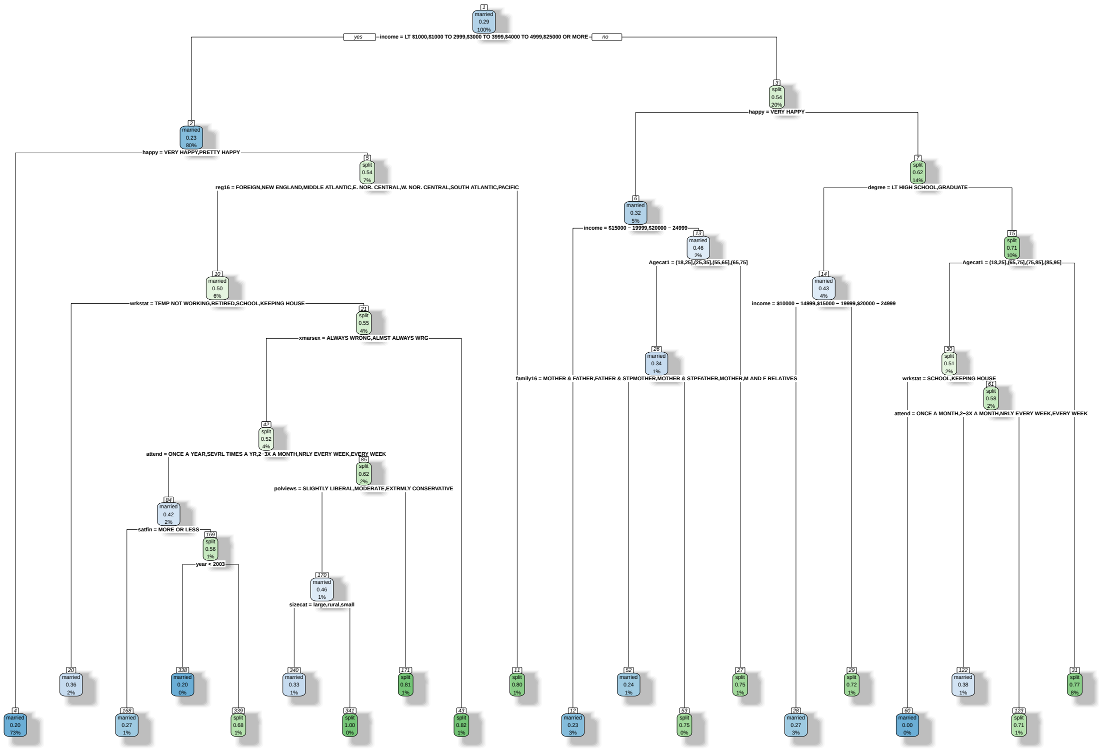
For both decision trees and random forests (below) we tested using both the complete cases and imputing data to fill in missing values. In the cases of both the complete cases and imputed data, decision tree models for all 3 ballots were more accurate than their random forest counterparts. For these initial runs, we tested our models on their corresponding training sets.

Model: Decision Trees with Complete Cases: Initial Tests on Complete Cases Training Sets

We trained decision tree models on data from complete observations from Ballots A, B and C, and we then tested these models on their corresponding testing sets (and then compared the accuracy rates to those of the imputed models). As an example, the model based on the complete observations for Ballot A had an accuracy rate of 75.38%, with an accuracy rate in predicting married statuses of 78.31% versus an error in predicting split statuses of 59.76%.

Model: Decision Trees with Imputed Values: Initial Tests on Imputed Data Training Sets

We trained decision tree models on imputed data for Ballots A, B, C and then tested them. As with the random forest models, the decision model based on Ballot C data with imputed observations was most accurate. It had an overall error rate of 78.73%. The tree below shows that the splits were based on income, happiness, degree, age, region at age 16, work status, financial satisfaction, family life at age 16, church attendance, and political views.


1 married 0.29 100%
yes
income = LT $1000,$1000 TO 2999,$3000 TO 3999,$4000 TO 4999,$25000 OR MORE
no
3 split 0.54 20%
happy = VERY HAPPY
2 married 0.23 80%
happy = VERY HAPPY,PRETTY HAPPY
5 split 0.54 7%
reg16 = FOREIGN,NEW ENGLAND,MIDDLE ATLANTIC,E. NOR. CENTRAL,W. NOR. CENTRAL,SOUTH ATLANTIC,PACIFIC
7 split 0.62 14%
degree = LT HIGH SCHOOL,GRADUATE
6 married 0.32 5%
income = $15000 − 19999,$20000 − 24999
13 married 0.46 2%
Agecat1 = (18,25],(25,35],(55,65],(65,75]
15 split 0.71 10%
Agecat1 = (18,25],(65,75],(75,85],(85,95]
10 married 0.50 6%
wrkstat = TEMP NOT WORKING,RETIRED,SCHOOL,KEEPING HOUSE
14 married 0.43 4%
income = $10000 − 14999,$15000 − 19999,$20000 − 24999
21 split 0.55 4%
xmarsex = ALWAYS WRONG,ALMST ALWAYS WRG
26 married 0.34 1%
family16 = MOTHER & FATHER,FATHER & STPMOTHER,MOTHER & STPFATHER,MOTHER,M AND F RELATIVES
30 split 0.51 2%
wrkstat = SCHOOL,KEEPING HOUSE
61 split 0.58 2%
attend = ONCE A MONTH,2−3X A MONTH,NRLY EVERY WEEK,EVERY WEEK
42 split 0.52 4%
attend = ONCE A YEAR,SEVRL TIMES A YR,2−3X A MONTH,NRLY EVERY WEEK,EVERY WEEK
85 split 0.62 2%
polviews = SLIGHTLY LIBERAL,MODERATE,EXTRMLY CONSERVATIVE
84 married 0.42 2%
satfin = MORE OR LESS
169 split 0.56 1%
year < 2003
170 married 0.46 1%
sizecat = large,rural,small
20 married 0.36 2%
338 married 0.20 0%
340 married 0.33 1%
171 split 0.81 1%
11 split 0.80 1%
52 married 0.24 1%
27 split 0.75 1%
29 split 0.72 1%
122 married 0.38 1%
31 split 0.77 8%
4 married 0.20 73%
168 married 0.27 1%
339 split 0.68 1%
341 split 1.00 0%
43 split 0.82 1%
12 married 0.23 3%
53 split 0.75 0%
28 married 0.27 3%
60 married 0.00 0%
123 split 0.71 1%

Decision Trees: Comparing Predictive Accuracy of Imputed and Complete Case Decision Tree Models on Complete Case Test Sets

The most accurate model that predicts split (divorce/separate) outcomes is the imputed decision tree model based on Ballot C. The model's overall accuracy rate in predicting marital status is 75.51%. It has a 78.16% accuracy rate at predicting married statuses, and a 60% accuracy rate at predicting splits.

imputed Ballot C	predicted marriage	predicted split	accuracy

imputed Ballot C	predicted marriage	predicted split	accuracy
married	297	83	0.78
split	26	39	0.60

Analysis Method Three: Random Forests

Random Forests with Complete Cases

We trained a set of random forest models on data from complete observations from Ballots A, B and C, and we then tested these models on their corresponding validation sets. The model based on the complete observations for Ballot C had an overall error rate of 23.79%, with an error in predicting married statuses of 5.64% vs. an error in predicting split statuses of 67.41%.

Random Forests with Imputed Values

Just as we did with the models based on the complete observations, we trained random forest models on imputed data for Ballots A, B, C and then tested them on their corresponding validation sets. Again, the model based on Ballot C data with imputed observations was most accurate. It had an overall error rate of 24.6%, with an error in correctly predicting married respondents of 5.5% and an error in correctly predicting split respondents of 70.5%.

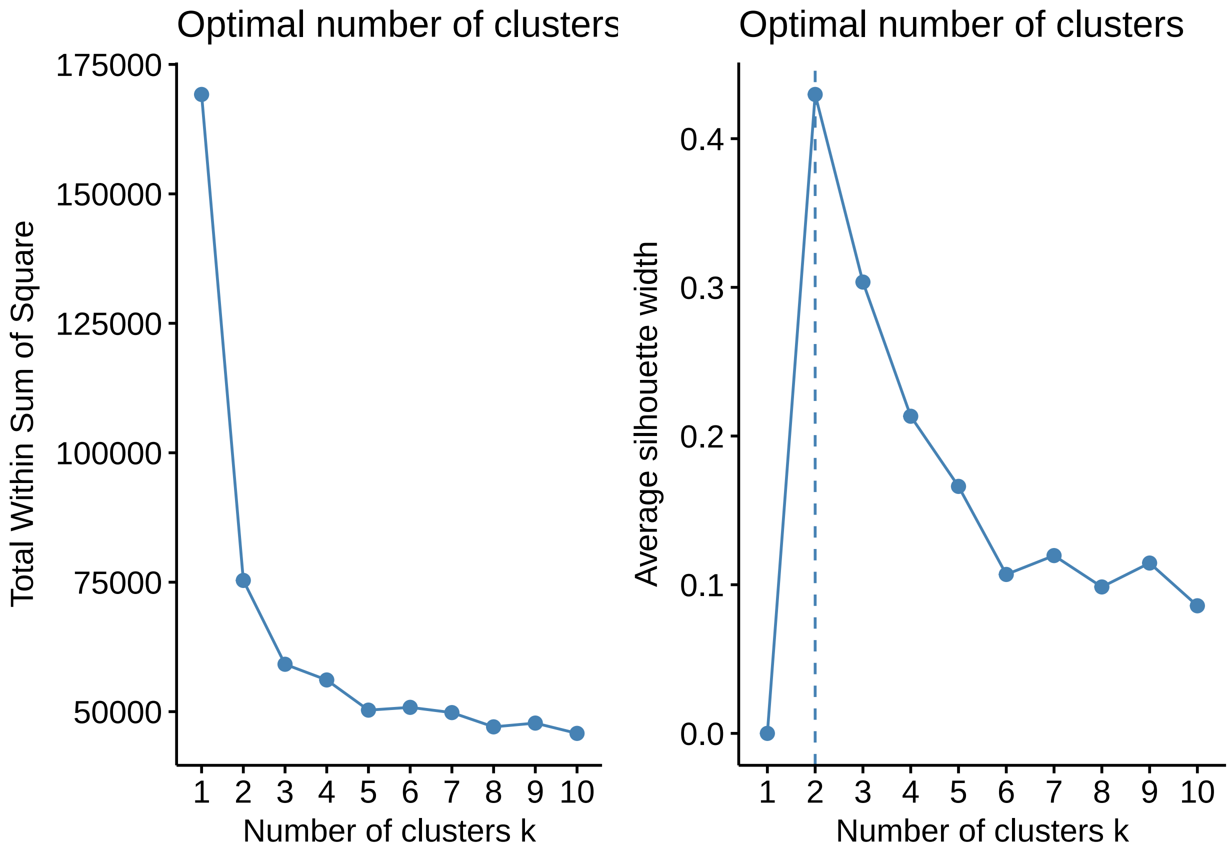
Random Forests: Comparing Results

Finally, we compared our random forest models by testing them on the test (rather than validation) sets corresponding to the complete observations for Ballots A, B and C. Here, the random forest model based on imputed data for Ballot C was most accurate. It had an overall error rate of 23.87% – overally, it predicted marital status with 77% accuracy. It predicted married respondents with a 95.5% accuracy rate (4.5% error rate) and split respondents with a 29.5% accuracy rate (70.5% error rate). From these results, our random forest model based on imputed data for Ballot C could be a useful tool for predicting whether/not a respondent is married (and perhaps be/stay married), but not very powerful with respect to divorce.

Analysis Method Four: Clusters

We also tried analyzing the data using clustering methods. Though the initial tests to determine the optimal number of clusters (2) seemed promising, the clusters did not accurately predict married or split status. Clustering was ineffective regardless of the clustering and distance methods used – kmeans, pam (partitioning around medoids), hierarchical, Ward, single, complete, maximum, Manhattan, etc.

For example, using the data from Ballot A, an elbow plot and a silhouette plot both indicate that the number of clusters should be 2.



However, none of kmeans, pam, and hierarchical clustering methods correctly categorizes by married and split. The following table shows the "accuracy" as calculcated by the number of married observations assigned to cluster 1 and the number of split observations assigned to cluster 2.

Method	Overall accuracy
kmeans	0.55
hierarchical	0.48
PAM	0.48

As another test, average silhouette widths for the hierarchical clusters was calculated, but were only 0.44 and 0.42 (close to 1 is ideal). The results from Ballots B and C were not any more accurate, indicating that clustering is not the optimal method for this analysis.

Conclusion

In the end, we selected two different models as being the best predictors. In terms of predicting the likelihood of being married, the random forest model based on imputed data for Ballot C was most accurate. In terms of predicting the likelihood of being divorced, the decision tree model based on imputed data for Ballot C was most accurate.

Our analysis has limitations in its accuracy and its usefulness for policymakers or marriage counselors. Its accuracy is limited in that we only use data from one half of a couple. Our predictions would likely be much more accurate if we had data on both partners. Its usefulness may be limited in its application to younger generations if their patterns of marriage and divorce differ significantly from those of the people in our dataset. Despite this, our analysis provides some insights into characteristics that are related to an individual's likelihood of getting a divorce.