Can We Predict if You Sleep Alone?

Solomon Boukman, Shady Yassin, Drake Wilde, Dybe Mwaisyange

Introduction

Background:
Couples sleep alone for a multitude of reasons from snoring, to no longer being physically intimate. Although analyzing these reasons gives insights in understanding modern relationships, we are more interested in a bigger picture. We believe a more interesting question to ask is whether or not using basic explanatory variables allows us to make an accurate prediction of whether or not a couple sleeps together or not.

In short, **our specific prediction task**:

- Given this data we would like see if we can accurately predict whether or not a couple sleeps together or not.
- We specifically want to analyze how well our tested classification models do in comparison to random guessing of labels.

Data and Preprocessing

How we collected the data:
We downloaded the raw GitHub CSV file from the FiveThirtyEight data repository. As stated above however it was originally collected from a SurveyMonkey Audience.

Attributes Used:
We decided to only use the relevant attributes of 'Gender', 'Household Income' 'Age' , 'Education', 'Relationship Status', 'Relationship Length.', and 'Location'. These were the only attributes that we could possibly use for our specified classification task.
Various attributes are left null throughout the dataset. After cleaning the rows that include these values we are left with 805 data points.

Attribute Distributions:
Age: Equally Distributed for those over 29, while those under 29 make up about 8% of the data
Gender: Equally Distributed
Household income: Approximately normally distributed
Education: Equally distributed for those who have a higher degree of education, ~8 percent of data only have high school education or none
Relationship Length: 50% married for more than 20 years, equally distributed for other relationship lengths
Relationship Status: heavily skewed, mainly married people
Location: heavily skewed, mainly people who live in the Northeast

Labels were encoded to ones and zeros we considered sleeping alone once a year or less as sleeping together

When both you and your partner are at home,how often do you sleep in separate beds?
Once a year or less
Never
Every night
A few times per month
Once a month or less
A few times per month

⇨

0
0
1
1
1
1

N - 1 dummy variables were created for each attribute then concatenated into one table

Age
18-29
45-60
30-44
>60

⇨

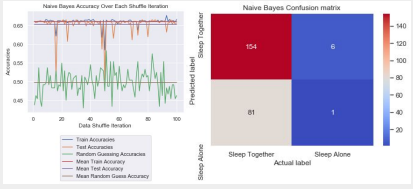
30-44	45-60	>60
0	0	0
0	1	0
1	0	0
0	0	1

Methodology

1. Clean and preprocess data.
2. We extract our relevant attributes (no null values included)
1. Next encode labels and create dummy variables for each of the relevant attributes
2. Pass random train test splits an into each of our models for 100 iterations and record the train and test accuracy.We retain the distributions of the labels in the train and test sets
3. Analyze the results

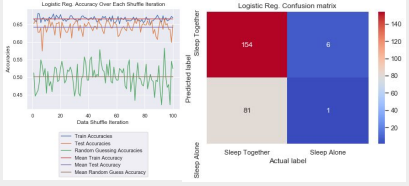
Models

Naive Bayes



Mean Training Error: 0.6601
Mean Testing Error: 0.653

Logistic Regression



Mean Training Error: 0.6658
Mean Testing Error: 0.643

Decision Tree

Depth (Hyperparameter): 3
Training Error: 0.6769
Testing Error: 0.648

Results and Discussion

Model	Test Accuracies Over 100 Iterations	T-Statistic	95% Confidence Interval
Random Guessing (Control)	0.501	-	
Naive Bayes	0.653	38.579 (0.004)	[0.645, 0.661]
Logistic Reg.	0.643	38.945 (0.004)	[0.636, 0.650]
Decision Tree	0.648	37.282 (0.004)	[0.640, 0.655]

*Standard errors in parentheses

On average, our models can predict about 65% of the time whether or not a couple sleeps together, a significant improvement from random guessing (which converges to 50% accuracy). Additionally, we observe that Naive Bayes performs marginally better than the logit regression and decision tree models,

To improve our models, we can consider whether or not reformatting our categorical variables, and perhaps reducing the number of controls could positively impact the predictive accuracy of our models. Perhaps this is contributing to the undepredictive nature of our estimates.