Predicting Sleeping Arrangements for Couples

dwilde, dmwaisya, sboukman, cslogin

Goal

Couples sleep alone for a multitude of reasons, anything from snoring to no longer being physically intimate. Although analyzing these reasons can give us insights into understanding the complexities of modern relationships, we are much more interested in the bigger picture. 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 sleep together or not. How well do various classification models compare to the random guessing of labels?

Data

The data was made publicly available through FiveThirtyEight and was collected from a SurveyMonkey Audience incorporating census data throughout the United States. Our methodology in preprocessing the data began with extracting relevant attributes and cleaning null values. We then encode labels and create dummy variables for each of the relevant attributes. These variables include: gender, household income, age, education, relationship status, relationship length and location. There are 805 data points that include relevant variables after cleaning the data.

- 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

Model & Evaluation Setup

- Naive Bayes Classification: Supervised classification algorithm. Uses the distribution of categories in attributes to make predictions. We create our label distributions for each attribute in the dataset and implement Laplace Smoothing to ensure no attributes have a predicted distribution of 0.
- Binary Logistic Regression: Calculates the ‘odds’ of a binary outcome given the provided explanatory variables. This can be used for prediction by mapping values above a certain threshold to 1 and below to 0. It assumes that observations in the sample are bernoulli-distributed, and that there exists a linear relationship between the explanatory variables and the log-odds that Y=1.

- Classification Decision Tree: The target variable in a decision tree takes a discrete set of values where the leaves in the tree represent class labels and branches represent conjunctions of features that lead to those class labels. The maximum depth of a tree is a measure of how many splits a tree can make before coming to a prediction. Too large of a value may result in overfitting of the data.

Results & Analysis

Claim #1: 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).
Support for Claim #1: Having conducted a t-test for difference of means between the average accuracy prediction of each of our models with respect to random accuracy, we obtain t-statistics which are statistically significant at even the 1% level.

Claim #2: We observe that Naive Bayes performs marginally better than the logit regression and decision tree models,
Support for Claim #2: This is supported by the fact that the Naive Bayes model consistently scores higher than the other two, and yields a confidence interval which is shifted more towards 100% - which we believe indicates a better performing model because all models yield the same standard error, which is the only other influencer of confidence interval size.