Project Midterm Report – How Couples Meet and Stay Together

Catherine Weldon, Angie Pinilla, and Josh Robbins

Our data set seeks to find factors that determine the success of a relationship/marriage. The data set has 5 waves (one survey/year for five years). The first survey has 35 questions – many questions having multiple parts – and focuses on demographic information such as income, race, age, sexual orientation, and various other questions we hope can provide insight into how successful a relationship will be. The metric we are using to measure the success of the relationship is the quality of the relationship where participants rate their relationship on a scale of 1 - 5. The other waves provide follow-up data for most of the participants, detailing if they broke up, and if so, why. We plan to use these follow-up waves to further analyze factors that contribute to break-ups.

Because our sample size is relatively small (4,002 participants) over-fitting is our main concern. To avoid over-fitting, we will not choose a model that is too flexible and we will look at the variance and bias to determine an appropriate tradeoff. If it still seems like over-fitting could be an issue, we will bootstrap to artificially create more data to analyze our model.

In our preliminary analyses, when using a classification model, we will test the effectiveness of the learned model by evaluating an unbiased estimate of the accuracy on our unseen data (i.e. the test set we set aside). We can calculate the test error rate: the average number of incorrect predictions on the test set when compared to the true class labels of the test set. Additionally, we can draw a comparison between the train error rate and the test error rate to identify the model?s underfitting/overfitting behavior.

In our preliminary models, we aimed to classify our data on the response variable ”Q34”: a partner?s perceived measure of quality of his/her relationship, on a scale 1-5. Thus we used this scale as the class labels for our data. To perform this classification, we first used a ”Multiclass learning model using One-vs-Rest” (train error-rate = 0.55, test error-rate = 0.56), and then implemented a ”RandomForestClassifier” (train error-rate = 0.01, test error-rate = 0.12) based on the previous model?s low performance accuracy. In addition, with future models, we will implement a K-fold cross-validation to compute a performance measure such as accuracy from the resulting model?s performance on the remaining part of the data not used in the k-1 folds marked for training.