Jack Beautz and Caleb Berman

Jpb375 and Ckb65

Physical Attractiveness and Speed Dating Midterm Report

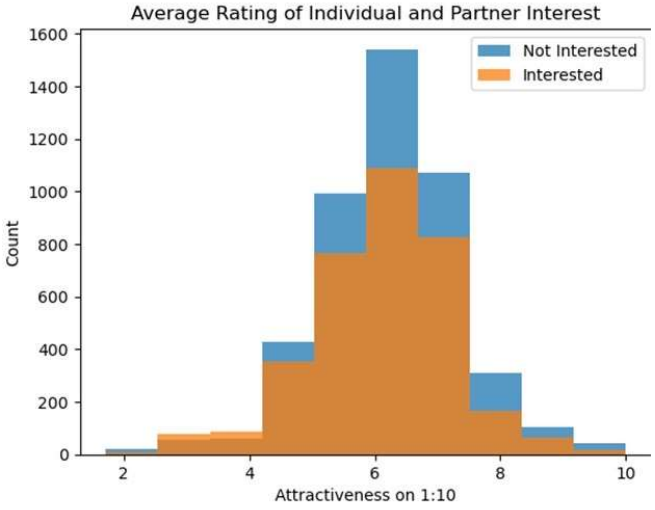
The Data

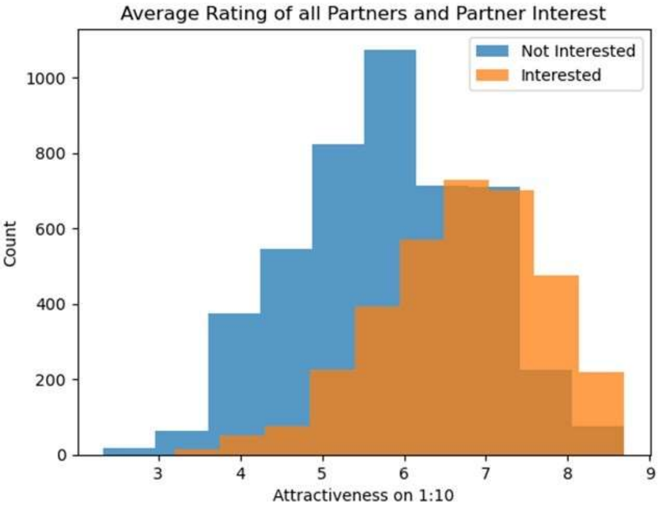
All data is derived from multiple rounds of a speed dating event proctored by Columbia Professors Ray Fisman and Sheena Iyengar. The data from each permutation of partners was collected via a survey after a 4-minute speed date detailing in which participants were asked to rate their partner on 1-10 scales on Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. Other lifestyle information was derived from a questionnaire at the onset of the experiment.

Each participant of the speed dating study was given a unique iid key. The rows of the dataset contain every possible pairing between two people: an individual's iid with their partner's key, pid. While the number of rows in this data set is relatively small, the data is very feature rich.

There are two binary features, dec_o and match, which indicate the decision to match or not by the partner and whether or not both people decided to match in the pairing. There is also one real-valued feature attr_o which contains these features will be the objectives of this analysis.

In total there are 195 columns and 8378 rows to this dataset. After removing entries which have missing value for dec_o, match, and attr_o the dataset has 8166 rows.





Avoiding Overfitting and Underfitting

Since our dataset is very feature-rich we were concerned about the possibility of overfitting our data. To avoid overfitting we utilized an 80/20 test train split so that we had a test set to evaluate our error on. For further models we plan on using additional techniques such as cross validation and principal component analysis.

Models

For the midterm report we have chosen to focus on the importance of physical attractiveness, relative to a few other key features. Because we only considered aggregations of this one key feature there were no missing values. attr_o is a numerical rating on a scale of one to ten of the attractiveness of the individual's partner in a pairing. This value was stored as a String in the CSV but was converted to a Float64 when the data was cleaned.

Using attr_o we created two new features. Avg_attr_self is the average rating of the individual associated with a given iid by their partners in all pairings. Roughly speaking this feature measures how attractive an individual is perceived by others. Avg_attr_others is the average attractiveness rating given by the individual. It measures how this person, on average, rates potential partners.

The objective of the modeling is the dec_o column of the dataset. This column is one if their partner is interested in matching and zero if the partner does not want to match.

For preliminary models the data set has 8166 rows, 3 features (avg_attr_self, avg_attr_others, attr_o) and one objective (dec_o).

One concern with our data was that each speed dating trial would vary in it's match rate given that the individual trial groups were not very large. To ensure a fair model using data across all trials, we compared match rates for small and large groups of Speed Daters.The correlation between match percentage and number of people met is -0.02762, so fairly small and thus we concluded that aggregating all of the speed dating data was a fine choice.

Linear Regression

The objective was defined as a match_rate by averaging all of an individual's partner's decisions to match or not. For example if an individual had two partners and only one would like to match with them, the individual's match rate would be 0.5. This redefinition of dec_o created a continuous objective which could be predicted using linear regression. A simple model was built using two aggregate features avg_attr_self and avg_attr_others. The resulting MSE was -3e-16 on the training set and 1.2e-3 on the test set.

Logistic Regression

Scikit-learn’s logistic model was used to predict the decision to match by the partner in each pairing based solely on the attractiveness features. The model correctly predicted matches on the training set 72.3% of the time and 74.7% on the test set. We hope to improve this prediction by utilizing methods outlined in the Next Steps section.

Next Steps

Further features

We hope to increase the features we’re adding to our models to include the average income of the person’s neighborhood as well as one-hot encode all of the career information we have on hand.

Other Classification Loss functions

We hope to implement additional loss functions beyond logistic loss. In particular, we believe that utilizing a support vector machine might be helpful in creating a safety margin where one could reasonably expect they would be matched.

Random Forest

We hope to explore how nonlinear models perform on this dataset by using a Random Forest to predict the partner’s decision to match. We can build a decision tree on the training set of the data and then run the model on the test set. In addition, the factor importance attribute of decision tree packages will numerically determine the relative importance of physical attributes in the dataset so we can select out the most important attributes in speed dating.