Authors: Natasha Gandhi & Ishana Ram

Exploring Partner Compatibility and Attraction using Speed-Dating Data

Abstract:

In this project, we are using a speed-dating dataset to better understand the qualities and personalities that foster connections and mutual attraction. We used several machine learning models to determine the most important factors that influence how much someone likes their partner on a scale of 1-10 after their speed dating conversation. We first performed exploratory data analysis to select our variables of interest and learn more about our dataset. We then used random forest with bagging on the entire dataset to get a good overview of some of the important variables. After this, we once again used random forest to test if the factors that influence our decisions about our partner are the same factors that influence our partners decision about us, and we found this to be true. Following this, we assessed the most important features using best subset selection, lasso regularization, and a pruned decision tree. Lastly, we converted our decision tree into an interactive web application. According to all of our models, the top three most important predictors of how much someone likes their partner include their partner's attractiveness, sense of humor, and shared interests. We assessed the performance of the models by comparing their test MSE's. We found that the lasso performed best, followed by best subset selection, and the pruned decision tree.

Introduction:

As young adults navigating modern relationships, we hope to gain insights that enrich our experiences and help us build more meaningful and fulfilling relationships. With modern relationships moving towards speed dating and online dating platforms, understanding the dynamics of making a compelling first impression is very important, and we feel that utilizing a speed dating dataset will help us to uncover the factors that contribute to attracting others within a brief encounter. There are several compatibility tests that exist, such as Emory University's Marriage Pact, in which students answer a series of questions about themselves and their preferences in a partner, and are then paired with another student. However, we would like to dive a bit deeper and learn more about the specific characteristics that attract romantic partners. While there are a handful of machine learning methods we can use to do this, we feel that a focus on feature selection methods will allow us to really grasp the important factors. We want to use a mix of linear and nonlinear methods in order to introduce flexibility and differing relationships between variables. For all of these methods, we will be using cross validation or pruning methods to build the most robust models possible. Perhaps we could use a binary target variable in future analyses and utilize classification methods as well. In terms of limitations in our approach, the dataset unfortunately does not include a like variable on the partner side (how much does your partner like you on a scale of 1-10?), so we cannot assess match predictability. Additionally, we need to keep in mind that there are several other factors that influence people's attraction to others, and this dataset only scratches the surface.

Setup:

For this project, we are using the OpenML Speed Dating Dataset which includes data from experimental 4-minute speed dating events from 2002-2004. It includes 123 features such as participant demographic information, partner characteristics and compatibility and 8378 observations. We used python as our programming language and utilized VS Code and Jupyter Notebook. Off the bat, we created a numerical subset that includes only the numerical columns in the dataset. From there, we used our exploratory data analysis findings to clean our data and reduce our main data subset into 17 features, with 'like' as our target variable, and named this dataset 'rating_partner.' We ran the initial random forest algorithm with bagging on the entire dataset with 'like' as our target variable to determine the most important variables. From there, we compared random forest with decision and decision_o as the target variables in order to conclude that what we value in a partner is what the partner values in us, and this led us to focus solely on the rating_partner dataset for the remainder of our analysis. The optimal number of estimators was determined by the lowest test error. Next, we randomly split the dataset into 80% training and 20% testing, and performed best subset selection with the best model having the lowest MSE, lasso regularization with standardization and an optimal alpha value determined by cross-validation, and a pruned regression tree with the parameters selected by cross-validation. We then used streamlit and a rendering platform to convert the pruned decision tree into an interactive web application.

Results:

Model	Training MSE	Test MSE
Random Forest	0.0186	0.1393
Lasso	1.0312	1.1564
Best Subset Selection	1.1534	1.3684
Pruned Decision Tree	0.7582	1.7102

*Random Forest target variable = ‘decision’ (0/1)
*All other models target variable = ‘like’ (0-10)

All three of our models gave us the same top variables, which were funny_partner, attractive_partner, and shared_interests_partner. These variables indicate how a person rates their partner's funniness, attractiveness, and shared interests on the night of the event. The first model we ran was a random forest model. We used the optimal M value for all three random forest models by running the model for each M value and finding the one with the lowest test MSE. Running a random forest with “like” as our target variable and all the other variables in numerical_subset_clean as our features yielded a test MSE of around 1. When running it with decision and decision_o as our target variable, a much lower MSE of .11 and .14 were yielded.. For the decision tree model, running it without pruning yielded a test MSE of 2.21. After cost-complexity pruning, it had an MSE of 1.71, showing that pruning improved the model's accuracy. We ran cross-validation in order to discover the optimal alpha value for lasso and used this to run the model. Out of lasso, best subset selection, and the pruned decision tree, lasso yielded the lowest test MSE. Despite the fact that the decision tree had the highest test MSE, especially in comparison to random forest, we still used it for our algorithm due to convenience.

We exported the text from the decision tree and used if/else statements to turn it into an interactive decision tree.

Discussion:

Our results highlight that the lasso regression performed the best, in terms of MSE. By penalizing the absolute size of the regression coefficient, lasso helped in reducing overfitting, which is particularly beneficial given the complex nature of our data, where some predictors may not have a significant impact on the outcome.

The decision tree, while intuitive and easy to understand, did not perform as well, yielding the highest test MSE among the models. This suggests that despite the pruning, there was probably overfitting to the training data. Decision trees are typically sensitive to small changes in the data, and even with pruning, they might capture noise as signal, especially in a dataset as varied as one from speed-dating events. Looking at the pruned tree, it is possible that variables that were less "important" were used, leading to a higher test MSE.

Comparatively, the best subset selection model, which aims to find the most predictive subset of features, performed moderately well but still did not match the effectiveness of the lasso model. This could be due to its nature of not accounting for the interdependencies between variables as effectively as lasso, which through regularization, inherently considers both feature selection and model complexity.

In reviewing literature and existing models, our approach aligns with current trends in machine learning where regularization and feature selection are key to enhancing model generalizability. Studies such as those by Hastie, Tibshirani, and Friedman highlight the effectiveness of lasso in scenarios similar to ours, where the goal is to predict outcomes from a large set of potential explanatory variables.

Conclusion:

Our project utilized a comprehensive speed-dating dataset to analyze what factors influence mutual attraction and partner compatibility. We implemented and compared four different machine learning models: random forest, lasso, best subset selection and a decision tree, with a focus on understanding which features are most influential in predicting how much someone likes their partner after a speed-dating event. Our findings indicate that attributes such as a partner's sense of humor, attractiveness, and shared interests are pivotal. The lasso model's superior performance underscores the importance of feature selection and regularization in predictive modeling, especially in datasets with numerous predictors and potential multicollinearity. While the decision tree provided valuable insights and an interactive application, its higher MSE suggests limitations in its applicability for predictive accuracy compared to lasso. This project adds to our knowledge on partner attraction and what fosters connections. It provides a framework for further exploration into the dynamics of first impressions, which could be enhanced by incorporating more nuanced behavioral data and applying more sophisticated machine learning techniques in future studies.

References:

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1.

Greitemeyer, Tobias. “What Do Men and Women Want in a Partner? Are Educated Partners Always More Desirable?” Journal of Experimental Social Psychology, Academic Press, 29 Mar. 2006, www.sciencedirect.com/science/article/abs/pii/S0022103106000345.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.