Attraction and Machine Learning: Evaluating MSM Attraction Preferences using Random Forest

Darin Challacombe, Adjunct Instructor, Department of Psychology. djchallacombe@fhsu.edu

Background and Research Purpose

Machine learning tools (Reis et al., 2019) can provide tools like algorithms useful for evaluating data sets where traditional statistical analyses are difficult. They can predict relationships based on the algorithm it already developed (Breiman, 2001; Reis et al., 2019). A random forest is an example of a supervised machine learning technique.

Breiman's (2001) seminal work on a random forest (RF) describe it as a "classifier consisting of a collection of tree-structured classifiers" (p. 6), whereas others describe random RFs as being collections of decision trees (Fife & D'Onofrio, 2022). RFs can evaluate various decisions and find the best one to utilize. Recently, interest in RF has piqued the interest of social science researchers in psychology and attraction. For example, Fife and D'Onofrio (2022) recently suggested using RF in psychology, not as a replacement for traditional analyses but as a way to hone models. Joel et al. (2017) were some of the first to explore RF to attraction data. Other researchers have utilized RF to aid in predicting sexual satisfaction and sexual desire (Vowels et al., 2021, 2022). The current research hopes to add to this field by using RF to examine men who have sex with men (MSM) attraction data.

Study 1

Method. Using previous reported data (Challacombe & Perdomo, 2021), I utilized Python version Python 3.12.1 and the RandomForestClassifier available with scikit-learn 1.4.1 to evaluate the data sets.

Selected Results. From the five participant variables, I found age had the highest accuracy rating (accuracy = 0.85), followed position (0.72) and race (0.71) (see Table 1).

RFs can be displayed graphically. Figure 1 shows the graphical representation of the RF for the 2018 data variable of age.

Table 1

Python Random Forest Results

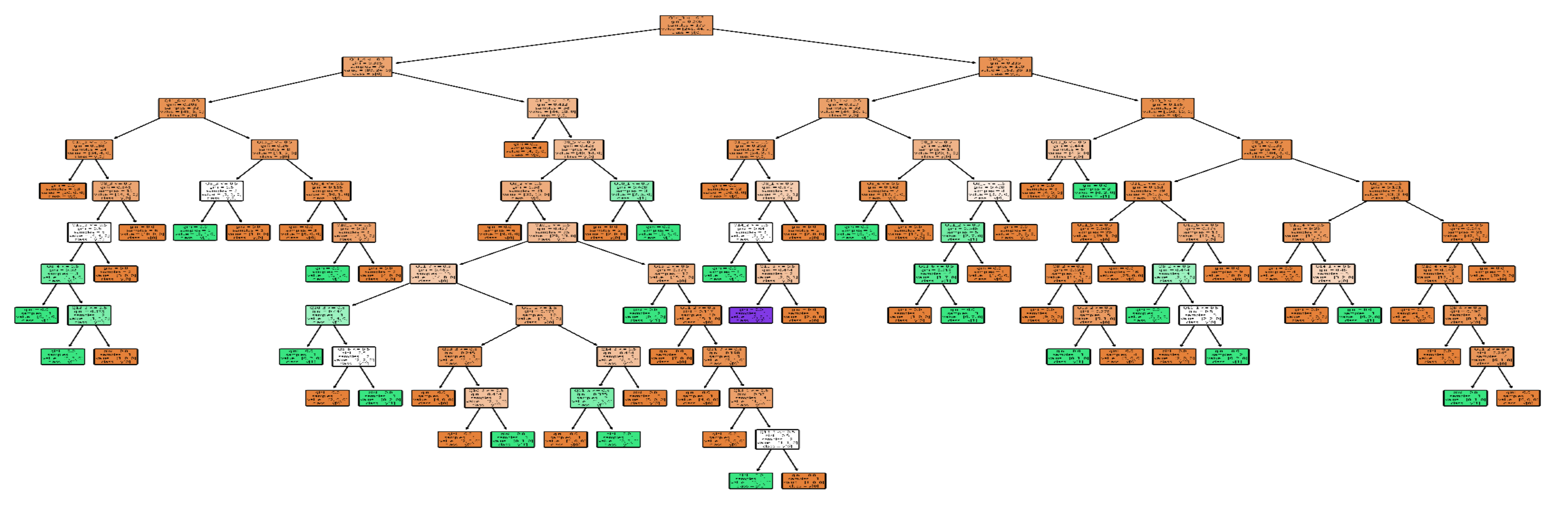
Variable	Description	Accuracy
Q3 (Race)	American Indian or Alaska Native (n=3); Asian (n=45); Black or African American (n=11); Native Hawaiian or Other Pacific Islander (n=40); and, White (n=256).	0.71
Q7 (Body Type)	Athletic (n=96); Average (n=113); Curvy (n=60); Muscular (n=52); Stocky (n=17); Thin (n=20)	0.32
Q9 (Hairiness)	No body hair (n=33); A little body hair (n=144); Average hair (n=135); A lot of hair (n=46)	0.49
Q2 (Age)	Categories (18-29 [n=299]; 30-49 [n=56]; 49+[n=3])	0.85
Q12_1 (My Position)	Categorized (61-100 = giver [n=128]; 40-60 = versatile [n=43]; 0-39 = receiver [n=202])	0.72

Study 2

Method. Using the same previous reported data (Challacombe & Perdomo, 2021), I utilized R version 4.3.2 and the randomForest available to evaluate the data sets. I patterned the analysis to Joel et al. (2017) to some degree.

Selected Results. From the five participant variables, I found both age and sexual position were the only two with an actual accuracy score, 0.018. The other three variables showed an accuracy score of 0 using R. Table 5 shows the accuracy scores of the participant variables.

Figure 1



Discussion

Fife and D'Onofrio (2022) recently recommended social science researchers utilize machine learning techniques like RF for advance data analysis. The Challacombe and Perdomo (2021) study provided enough attraction data to utilize the RF tool. Different from other researchers (cf. Joel et al., 2017), the current study approached RF by utilizing two different programming languages.

While the accuracy results from the two different programming languages do not match up, there is considerable overlap with the findings. In both studies, the importance of the top partner-specific variables remained about the same regardless of the programming language used. This indicates that the variables of sexual position (in both studies) and dominance (in studies 1B and 2B) are important MSM attraction factors. This finding supports the original findings for this data.

One of the original intents for this research was to develop out a tool that could predict attraction based on participant variables (e.g., if someone is this race, has this much hair, this age, and this position, then they will be more likely to be attracted to XYZ). While this may be something more possible in the future, the current study did not support the ability to do that. The participants variables, especially for the second study, did not fair well in being accurate.

Limitations

The largest limitation of this study was my own ability to perform the analyses using the programming languages. This created significant challenges, and it is possible that the results of these studies would be more robust of I had a better understanding of the processes. Instead, I thought it would be useful for the scientific community to show that someone, even with not the strongest background in statistics or programming, could utilize open source tools to perform machine learning analyses.

As mentioned in Challacombe and Perdomo (2021), the most significant limitation of the original data is the self-report and subjective nature. A participant's view of how much body hair is "average" may differ than other participants' views. This allows for this data to be skewed and inaccurate.

Next Steps

As with most analyses, there are a variety of ways to accomplish the same goal. My programming approach may not be the most efficient, effective way to accomplish it. Therefore, future research should focus on utilizing other programming tools and techniques to approach machine learning and RF.

References

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. Challacombe, D. J., & Perdomo, N. (2021). Attraction Preferences of Men who have Sex with Men. Sexuality & Culture, 25(6), 2010–2034. https://doi.org/10.1007/s12119-021-09861-6. Fife, D. A., & D'Onofrio, J. (2022). Common, uncommon, and novel applications of random forest in psychological research. Behavior Research Methods, 55(5), 2447–2466. https://doi.org/10.3758/s13428-022-01901-9. Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is Romantic Desire Predictable? Machine Learning Applied to Initial Romantic Attraction. Psychological Science, 28(10), 1478–1489. https://doi.org/10.1177/0956797617714580. Reis, I., Baron, D., & Shahaf, S. (2019). Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. The Astronomical Journal, 157(1), 16. https://doi.org/10.3847/1538-3881/aaf101. Vowels, L. M., Vowels, M. J., & Mark, K. P. (2021). Uncovering the Most Important Factors for Predicting Sexual Desire Using Explainable Machine Learning. The Journal of Sexual Medicine, 18(7), 1198–1216. https://doi.org/10.1016/j.jsxm.2021.04.010. Vowels, L. M., Vowels, M. J., & Mark, K. P. (2022). Identifying the strongest self-report predictors of sexual satisfaction using machine learning. Journal of Social and Personal Relationships, 39(5), 1191–1212. https://doi.org/10.1177/02654075211047004