**Does Self-Esteem Contribute to Dating Success in a Speed-Dating Scenario?**

**By Meagan Lacroix**

**Prepared for CKME 136**

Researchers have long been interested in the factors which lead to romantic attraction. However, much of the research conducted in this area has relied on self-report surveys or artificially constructed scenarios which lack ecological validity. To address this limitation, some researchers have chosen to study romantic attraction in the context of a speed-dating scenario.  This method has advantages because it allows for observation of dating choices in real-time with real people. Fisman, Iyengar, Kamenica, and Simonson (2006 and 2008) used the speed-dating method  to identify the factors which play a role in mate selection. More specifically, the researchers investigated the gender and racial differences in mate selection, with a focus on individual choices rather than number of final matches.  While their research furthers our understanding of group differences in mate selection, they did not address the potential role that self-esteem may play in initial attraction, nor did they examine the importance of self-perceived romantic desirability. Other research examining romantic attraction in a speed-dating scenario has found that higher self-esteem may lead to greater dating success (Wu, Chen, & Greenberger, 2019). As well, perceived romantic desirability may be an important mediator of self-esteem (Bale & Archer, 2013). The focus of this literature review and subsequent data analysis will be on the importance of self-esteem and perceived romantic desirability in romantic attraction and how this may be explored in the dataset created by Fismen et al (2006 and 2008). This project will focus on using a machine learning model to determine if participant’s self-ratings (a possible indicator of self-esteem and perceived romantic desirability) are important predictors to dating success.

    In the study conducted by Fismen et al (2006 and 2008) a series of speed-dating events took place where participants had 4 minutes to engage in conversation and decide if they want to see each other again. If both partners express interest (aka “match”), they are given each other’s contact information after the event so they may plan future dates. Prior  to the event, participants completed a demographic and personal interests questionnaire and were asked to provide ratings on a series of attributes: attractiveness, sincerity, intelligence, funness, ambition, and shared interests. Participants were asked to rate themselves and their partners on these attributes. They were also asked to rate how much they value each attribute in a potential romantic partner. Using linear probability models, the researchers found that both men and women highly value attractiveness in a potential partner, though men valued this attribute more.  Women, on the other hand, place more weight on intelligence. Men were also found to be less attracted to women whose intelligence or ambition exceeds their own, and more attracted to women with shared interests. Women showed a greater preference for partners of the same race and who are from more affluent neighborhoods. Overall, this study supports the existence of gender differences in mate selection, but does not fully explore the potential role that participant’s self-ratings may play in dating success overall.

Luo and Zhang (2009) also examined romantic attraction using a speed-dating scenario but placed more emphasis on personality characteristics of the participants. Prior to the speed date event, participants completed questionnaires measuring their political attitudes, personal values, interests, personality type, affectivity, attachment style, and self-esteem. At the speed-dating event, participants rated their attraction to each date. Correlations between partner attraction and personal characteristics were computed for men and women using a Social Relations Model (SRM) where variance of the model is partitioned into three components: variance due to the actor (how much the participant likes their partner), variance due to the partner (how much everyone likes the partner), and variance due to the relationship between actor and partner (how much they like each other).. The results showed that men were significantly more attracted to women who were older, lighter, physically attractive, athletic, conservative, extroverted, agreeable, conscientious, and who have high self-esteem. Women, however, were only significantly more attracted to men who were physically attractive and athletic. These findings suggest that positive personality characteristics may be important for deciding a man’s attraction, whereas this is less important for women. The researchers also found that similarity between partners (i.e. how much they have in common) did not significantly predict attraction. These results somewhat support the findings of Fismen et al (2006 and 2008). Both studies found that while the mechanisms for attraction may differ between sexes, attractiveness is the most important predictor of attraction for both groups.

Romantic attraction in a speed-dating scenario was also explored by Wu, Chen, and Greenberger (2019) and focused on a sample of Asian American students. This study measured communal attributes (i.e. attributes that emphasize care for others rather than self-interest), social desirability (i.e. concern with social approval), self-esteem, and narcissism. These attributes were correlated with dating success. It was found that communal attributes and self-esteem were positively correlated with desirability and dating success, while social desirability and narcissism showed no effects. Results were consistent for both men and women. This study somewhat contrasts with the findings of Luo and Zhang (2009) who found that communal attributes (e.g. agreeableness and conscientiousness) and self-esteem increased women’s desirability but not men’s.

A “sociometer” theory of self-esteem has emerged which posits that self-esteem may function as a way to measure one’s value in interpersonal relationships and may therefor be influenced by self-perceptions of desirability (Leary, Tambor, Terdal, & Downs, 1995). Bale  and Archer (2013) examined this theory in the context of romantic relationships and perceived desirability. The researchers measured self-esteem, perceived facial attractiveness, perceived body attractiveness, romantic self-confidence, and self-confidence in attractiveness and found that all variables were significantly positively correlated. Linear regression and mediation analysis also found that self-esteem was significantly predicted by self-perceived facial and bodily attractiveness and self-confidence in attractiveness, and these relationships were also mediated by romantic self-confidence. These findings suggest that self-perceived romantic desirability is connected to self-esteem.

The findings from the above research may inform further analyses of the Fismen et al. (2006 and 2008) dataset which go beyond examining general sex and racial differences in attraction. I propose an examination of the dataset which focuses more on the factors which influence dating success in a machine learning model, with a focus on the possible role for indicators of self-esteem.

**The Data Set**

The dataset was curated by Fismen et al. (2006 and 2008) and obtained from [https://www.kaggle.com/annavictoria/speed-dating-experiment/kernels. Fismen et al. (2006](https://www.kaggle.com/annavictoria/speed-dating-experiment/kernels.%20%20Fismen%20et%20al.%20(2006) and 2008) conducted a series of 21 speed-dating events and recruited 278 men and 276 women from Columbia University. Data was collected across 4 time-points - pre-event (Time 1), mid-event (Time 2), day after the event (Time 3), and 3-4 weeks after the event (Time 4) – and resulted in 195 variables being collected. Because of the large number of variables, only information collected at Time 1 will be used in the analyses. The outcome variable used will be the percentage of positive “yes” responses that each subject receives from their partners at the speed dating event (labeled as “other\_yes\_per” in the dataset). This variable will be considered the measure of dating success for each participant. The data consists of attribute ratings given to each participant by their partners. These ratings reflect the attractiveness of the participant in different domains (physical attractiveness, intelligence, ambition, shared interests, funness, and sincerity) and will be analyzed as an aggregated average rather than individually for each participant-partner interaction. The data also consists of the participant’s individual characteristics (i.e. demographic information, interest in certain activities, dating goals) and self-ratings. The self-ratings measure how participants think they rate in the domains of attractiveness, sincerity, funness, intelligence, and ambition. These self-ratings will be used as an indicator of the participant’s self-esteem. The participants also rated how they think others perceive them on these attributes. These ratings will be used as a measure of perceived attractiveness.

A regression tree machine learning algorithm will be applied to the dataset to determine the conditions under which participants enjoy greater dating success and more specifically, if participant self-ratings play a significant role. This method will be done without prior separation of the data into gender groups in order to determine which variables emerge as important to predicting the outcome variable. Regression trees were chosen as the preferred analysis method because of the ease of interpretability of tree diagrams as well as the flexibility of these models in handling many predictor variables. Different regression tree methods will be tested in R to determine which model shows the best fit. Root mean square error (RMSE) and mean absolute error (MAE) will be used as the performance metrics to evaluate the models.

**Research Questions**

1. How important are participant self-ratings to the prediction of dating success in a regression tree model?
2. How important is the participant’s perception of how others see them when predicting dating success?
3. Overall, which variables contribute the most to the prediction of dating success in a regression tree model?
4. Which factors play a bigger role in predicting dating success – the participant’s own attributes or the partner’s rating of those attributes?

**Variables to be Used**

(\*For a comprehensive list of all the variables collected by Fismen et al. please see attached data key\*)

gender:    Female=0

    Male=1

attr\_o – shar\_o:     Rating by partner the night of the event, for all 6 attributes

like\_o:     Overall, how much do you like this person? (rating by partner)

                        (1=don't like at all, 10=like a lot)

prob\_o:     How probable do you think it is that this person will say 'yes' for you?

(rating by partner)

                       (1=not probable, 10=extremely probable)

age: Age of the subject

field\_cd:     field of study,  coded

1= Law

2= Math

3= Social Science, Psychologist

4= Medical Science, Pharmaceuticals, and Bio Tech

5= Engineering

6= English/Creative Writing/ Journalism

7= History/Religion/Philosophy

8= Business/Econ/Finance

9= Education, Academia

10= Biological Sciences/Chemistry/Physics

11= Social Work

12= Undergrad/undecided

13=Political Science/International Affairs

14=Film

15=Fine Arts/Arts Administration

16=Languages

17=Architecture

18=Other

race: Race of the subject

    Black/African American=1

    European/Caucasian-American=2

    Latino/Hispanic American=3

    Asian/Pacific Islander/Asian-American=4

    Native American=5

    Other=6

goal:

What is your primary goal in participating in this event?

    Seemed like a fun night out=1

    To meet new people=2

    To get a date=3

    Looking for a serious relationship=4

    To say I did it=5

    Other=6

date:

In general, how frequently do you go on dates?

    Several times a week=1

    Twice a week=2

    Once a week=3

    Twice a month=4

    Once a month=5

    Several times a year=6

    Almost never=7

go out:

How often do you go out (not necessarily on dates)?

    Several times a week=1

    Twice a week=2

    Once a week=3

    Twice a month=4

    Once a month=5

    Several times a year=6

    Almost never=7

exphappy:

Overall, on a scale of 1-10, how happy do you expect to be with the people you meet

during the speed-dating event?

expnum:

Out of the 20 people you will meet, how many do you expect will be interested in dating you?

***We want to know what you look for in the opposite sex.***

Waves 6-9: Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 1-5, 10-21: You have 100 points to distribute among the following attributes -- give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.

attr1\_1

Attractive

sinc1\_1

Sincere

intel1\_1

Intelligent

fun1\_1

Fun

amb1\_1

Ambitious

shar1\_1

Has shared interests/hobbies

***Now we want to know what you think MOST of your fellow men/women look for in the opposite sex.***

attr4\_1

Attractive

sinc4\_1

Sincere

intel4\_1

Intelligent

fun4\_1

Fun

amb4\_1

Ambitious

shar4\_1

Shared Interests/Hobbies

***What do you think the opposite sex looks for in a date?***

attr2\_1

Attractive

sinc2\_1

Sincere

int2\_1

Intelligent

fun2\_1

Fun

amb2\_1

Ambitious

shar2\_1

Has shared interests/hobbies

***How do you think you measure up?***

Please rate your opinion of your own attributes, on a scale of 1-10 (be honest!):

attr3\_1

Attractive

sinc3\_1

Sincere

int3\_1

Intelligent

fun3\_1

Fun

amb3\_1

Ambitious

***And finally, how do you think others perceive you?***

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5\_1

Attractive

sinc5\_1

Sincere

int5\_1

Intelligent

fun5\_1

Fun

amb5\_1

Ambitious

Ratings of the subject for each partner they met during the event (measured on a scale of 1-10)

attr

Attractive

sinc

Sincere

int

Intelligent

fun

Fun

amb

Ambitious

shar

Shared interests

like:     Overall, how much do you like this person? (rating by subject)

                        (1=don't like at all, 10=like a lot)

prob:     How probable do you think it is that this person will say 'yes' for you?

(rating by subject)

                       (1=not probable, 10=extremely probable)

match\_es:

How many matches do you estimate you will get (a match occurs when you and your partner both check “Yes” next to decision)?: \_\_\_\_\_\_\_\_\_\_\_

**Approach**

**Data Cleaning**

**Case removal:** Cases from wave 12 will be removed. This wave had a different experimental variation compared to the rest of the waves. Cases having more than 50% of data missing will also be removed.

**Variable selection**: Variables will be removed based on their utility to the analyses, prior analysis by Fismen et al. (2006 and 2008), and quality of the data (missing values, inconsistency in measurements, etc), or repetition in the dataset.

**Aggregation:** The dataset will be aggregated using FUN = Mean so that there is only 1 observations per subject

**Variable Creation:** New variables will be created from dichotomous variables which cannot be transformed into the mean (e.g. “match”). Categorical variables will be binarized into dummy variables

**Splitting the Data:** The dataset will be split into a testing and training set using a 70-30 split.

**Out of range and missing values**: Values will be imputed with means or within-range values where possible and appropriate. This data cleaning step will be performed on the training set only.

**Checking Assumptions**

**Correlation:** Correlation matrices will be created to determine the presence of multicollinearity and singularity for any of the variables, as well as linear bivariate correlations with the dependent variables

**Linearity, homoscedasticity, and normality of residuals**: A Q-Q plot, plot of residual vs fitted values, and a residual histogram will be generated to test the assumption of linear regression.

**Shapiro-Wilk test:** A Shapiro-Wilk test will be used to test the null hypothesis that the distribution of other\_yes\_per values comes from a normally distributed population.

**Selection of Important Variables**

**Variable Importance:** Variable importance will be determined using 4 methods: multiple regression with backward elimination, information gain, Boruta, and variable importance with the rpart() function. Variables which are flagged as important by at least 2 of these methods will be used in the regression tree analyses.

**Regression Trees**

Different regression tree methods will be tested in r to determine which model has the best fit. The r functions to be used for creating the trees will be tree(), cv.tree(), rpart() and randomForest. Models will be compared using the performance metrics RMSE and MAE.

**Methods and Results**

**Removing redundant variables and cases**

1. Wave number 12 was removed from the data set. This wave had a different experimental condition compared to the other waves – participants were only allowed to say “yes” to 50% of their dates. This condition would skew the results of the dependent variable “other\_yes\_per”.
2. The complete dataset has 195 variables collected across 4 time points. Only variables collected at time point 1 (the actual speed-date event) were of interest to the analyses. Variables collected at time points 2-4 were removed (k = 84). The remaining variables were then examined for redundancy. The following variables were deemed unimportant to the prediction of the outcome variable: id, idg, condtn, wave, position, positin1, order, partner and pid. Because the analyses are focused on the attributes of the subject and their relation to “yes” decisions, the attributes of the subject’s partner were removed from the dataset. These variables were int\_corr, samerace, age\_o, race\_o, pf\_o\_att to pf\_o\_shar, met\_o, and met. Some of the variables collected be Fismen et al. (2006) were not self-reported but rather best-guess estimates based on region of origin and undergraduate institution. These variables were mn\_sat, tuition, and income. Because these variables may be inaccurate and were not actually reported by the participants, they were removed from the dataset. The variables undergrd, from, and zipcode varied widely and were also removed. The variables imprace and imprelig – measuring the importance of race and religion to the subject when making a date decision – were also removed. Career and career\_c were also removed because these variables were measuring nearly the same thing as field and field\_c. Lastly, after calculating the percentage of yes decisions made by the subject (subject\_yes\_per) and the percentage of yes decisions made about the subject (other\_yes\_per), the variables round, match, dec\_o, and dec were removed from the data set as they were no longer needed. After removing these variables, 75 variables remained.
3. The variables were then examined for missingness. The variables attr4\_1 to shar4\_1 were only answered by waves 6-21. Because of this inconsistency between waves these variables were removed. The variables attr5\_1 to amb5\_1 was found to be missing nearly 50% of cases. These variables were also removed. Lastly, the variable expnum was missing 77% of cases. This variable was removed. After removing these variables, 63 variables remained.
4. The variable field\_cd is the coded version of the variable field. Field\_cd was found to have some missing data for cases where data was present in the field column. Missing field\_cd cases were imputed with 18 = “other” prior to removing the field variable. Sixty-two variables remained.

**Binarizing the Categorical Variables**

In order to perform multiple linear regression, all categorical variables were split into new columns and binarized.

1. field\_c: this variable was originally divided into 18 categories. Five new variables were created using these 18 categories. They were 1) hard\_sci (math, medical science, engineering, biological sciences), 2) soft\_sci (social science, social work, architecture), 3) arts (English, history, education, film, fine arts, languages), 4) bus\_pol (law, business, political science), and 5) other\_undec (other/undecided). Responses for these new variables were coded as 1 = yes and 0 = no. The original field\_c variable was then deleted.
2. race: this variable was originally divided into 6 categories. One category – Native American – had no observations and so this category was not used. Five new variables were created to reflect these categories: race\_black, race\_white, race\_hisp, race\_asian, and race\_other. Responses for these new variables were coded as 1 = yes and 0 = no. The original race variable was then deleted.
3. goal: this variable originally had 6 levels. A new variable (goal\_date) was created to reflect participants whose main goal was “To get a date” or “Looking for a serious relationship” (coded as 1) and all other goals (coded as 0). The original goal variable was then deleted.
4. date: this variable originally had 7 levels. Three new variables were created to reflect frequent daters (date\_freq)), moderate daters (date\_mod) and infrequent daters (date\_infreq). Responses for these new variables were coded as 1 = yes and 0 = no. The original date variable was then deleted.
5. go\_out: this variable originally had 7 levels. Three new variables were created to reflect participants who frequently go out (out\_freq)), go moderately (out\_mod) and infrequently go out (out\_infreq). Responses for these new variables were coded as 1 = yes and 0 = no. The original go\_out variable was then deleted.

After creating these new binarized variables, 74 variables are counted in the dataset.

**Aggregating the Data and Checking Skewness**

Prior to aggregating the cases into one row per subject, the distribution of each of the continuous variables was checked for normality using skewness. A variable is considered highly skewed if it has a skewness value +- >1.69. Only 1 variable – attr1\_1 – exceeded this skewness threshold (skew = 2.22). Because the majority of the variables were normally distributed, the mean was used to aggregate observations. The resulting dataset had 523 observations.

After this step, NA values were again counted for all of the variables. This was done because the aggregation process of computing the mean for each group of subject observations would reduce the number of NAs seen in the original un-aggregated dataset. match\_es was found to have the highest number of missing cases (NA = 72). Because of the high missingness of match\_es and the high skewness of attr1\_1, these variables were removed from the dataset rather than imputing and transforming. The iid variable was also removed at this step as it is no longer needed after aggregation. The resulting dataset now has 71 variables.

**Splitting the Data**

The data was next split into a training and test set using a 70:30 split. The training set has 366 observations and the test set has 157 observations. Further data cleaning procedures were next applied to the training set only.

**Missing Data and Out of Range Values**

Missing data was counted across all columns for each participant. Participants with more than 50% data missing were removed from the train set. Five cases were removed, bringing the total number of cases down to 361.

One case was missing a value on the age variable. This was imputed using the mean.

Some out of range values were found for the “interests” variables hiking, gaming, reading, and yoga. Out of range values were imputed with the nearest within-range value.

One NA value was found for exphappy. This was imputed using the mean.

A very small number of NA values were found for the attribute rating variables – specifically shar1\_1, sinc, fun, intel, prob, and shar. Because of the small number of missing values, these were imputed using the mean of the respective variable.

One NA value was found for date\_freq, date\_mod, and date\_infreq. These missing values were imputed with 0.

**Checking Assumptions for Linear Regression – Residual Diagnostics**

A qq-plot was generated in order to check the assumption of normality of the distribution of errors in a linear model. As can be seen in Figure 1, the residuals follow an approximately straight line, suggesting normality.

Next, tests for the assumption of normality were generated (see Figure 2). Of particular interest is the Shapiro-Wilk test which tests the null hypothesis that the data comes from a population which is normally distributed. The Shapiro-Wilk test generated a p-value of .82, indicating normality of the data.

The correlation between the observed residuals and expected residuals under normality was next assessed and found to be highly correlated (.99).

A residual vs fitted values plot was generated to detect non-linearity, unequal variances, and outliers (see Figure 3). It can be seen that the residuals are spread randomly around the 0 line, indicating that the relationship between the predictor and outcome variables is linear. The residuals also form a roughly horizonal around the 0 line, indicating homogeneity of error variance. As well, no one residual appears drastically far away from the random distribution of residuals, indicating there are no outliers in the dataset.

Lastly, a residual histogram supports the assumption of normality (see Figure 4). The residuals are evenly distributed around zero and have an approximately bell shape.

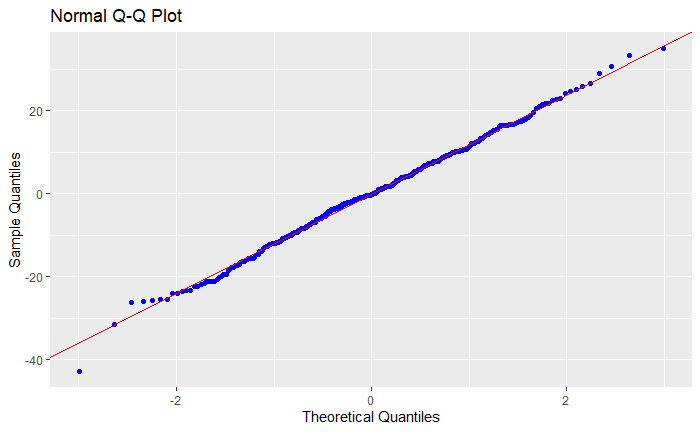


Figure 1. QQ plot showing the distribution of residuals in a linear model with other\_yes\_per as the dependent variable

-----------------------------------------------

Test Statistic pvalue

-----------------------------------------------

Shapiro-Wilk 0.9973 0.8223

Kolmogorov-Smirnov 0.0356 0.7504

Cramer-von Mises 29.2879 0.0000

Anderson-Darling 0.2827 0.6336

-----------------------------------------------

Figure 2. Test statistics generated to test the assumption of normality

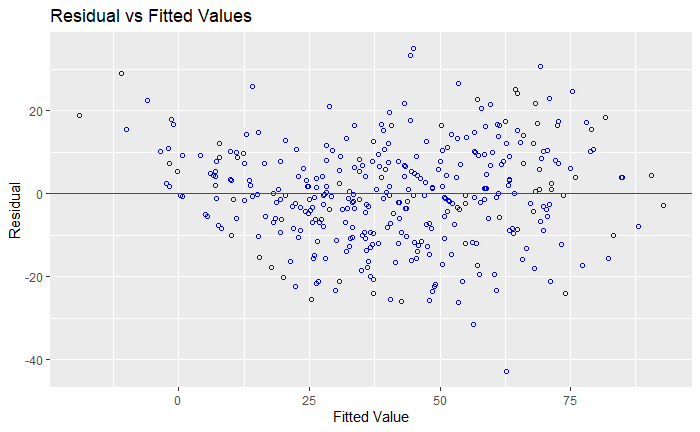


Figure 3. Residual vs fitted values plot to determine the presence of non-linearity, unequal variance, and outliers

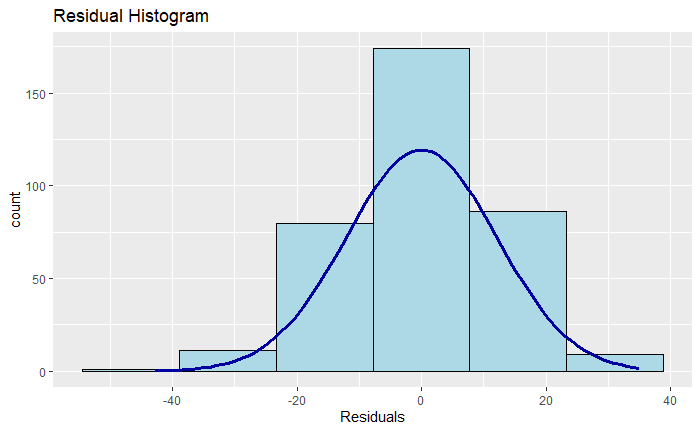


Figure 4. Histogram of residuals

**Checking for Multicollinearity and Singularity**

A Pearson’s r correlation table was generated in order to determine the presence of multicollinearity or singularity among the independent variables. Because of the large number of variable pairs (n = 789), the table is not shown here. A variable pair was considered to show multicollinearity/singularity if the correlation between them exceeded .90. No correlation pair was observed to exceed .87, indicating absence of multicollinearity and singularity.

**Correlations Between IVs and DV**

Pearson’s correlations were computed for the relationship between the dependent variable other\_yes\_per and all of the independent variables. A significant correlation (p < .05) was found for 33 of the 70 independent variables (see Table 1).

The highest correlations were observed to be between other\_yes\_per and the partner rating attributes (attr\_o to prob\_o). These correlations were all positive and indicate that as average partner ratings increase, so too does the percentage of “yes” responses for the subject.

Only two of the subject self-ratings (attr3\_1 and fun3\_1) were significantly positively correlated with other\_yes\_per. As subject ratings of their own attractiveness and funness increase, their percentage of “yes” responses from partners increases.

Table 1

*Significant Pearson’s Correlation Coefficients Between the Dependent Variable other\_yes\_per and the Independent Variables*

|  |  |  |
| --- | --- | --- |
| Independent Variable | Pearson’s r coefficient | P value |
| gender | -0.24 | 0.000 |
| age | -0.11 | 0.046 |
| attr\_o | 0.79 | 0.000 |
| sinc\_o | 0.23 | 0.000 |
| intel\_o | 0.27 | 0.000 |
| fun\_o | 0.65 | 0.000 |
| amb\_o | 0.25 | 0.000 |
| shar\_o | 0.65 | 0.000 |
| like\_o | 0.78 | 0.000 |
| prob\_o | 0.47 | 0.000 |
| sinc1\_1 | -0.13 | 0.016 |
| shar1\_1 | -0.17 | 0.002 |
| attr2\_1 | 0.13 | 0.013 |
| sinc2\_1 | -0.12 | 0.028 |
| attr3\_1 | 0.20 | 0.000 |
| fun3\_1 | 0.21 | 0.000 |
| sinc | 0.11 | 0.043 |
| prob | 0.19 | 0.000 |
| exercise | 0.17 | 0.001 |
| hiking | 0.11 | 0.046 |
| shopping | 0.11 | 0.042 |
| yoga | 0.16 | 0.002 |
| tv | -0.10 | 0.047 |
| field\_hard\_sci | -0.11 | 0.046 |
| field\_bus\_pol | 0.14 | 0.008 |
| race\_white | 0.18 | 0.001 |
| race\_asian | -0.21 | 0.000 |
| date\_freq | 0.14 | 0.007 |
| date\_infreq | -0.16 | 0.003 |
| out\_freq | 0.21 | 0.000 |
| out\_mod | -0.17 | 0.012 |
| out\_infreq | -0.13 | 0.017 |
| subject\_yes\_per | -0.28 | 0.000 |

**Determining Variable Importance**

Variable importance was determined using 4 methods: multiple linear regression with backward selection, Boruta, information gain, and rpart. For the linear regression and Boruta methods, a p-value of < .05 was used to select important features.

A table showing the identification of important variables for each method is shown below (Table 2). A variable was selected as an important feature if it was identified by at least 2 of the feature selection methods. Using this criteria, the variables to be used in the regression tree models are gender, attr\_o, fun\_o, shar\_o, like\_o, prob\_o, and subject\_yes\_per.

Table 2

*Variables Selected as Important Using Four Feature Selection Methods*

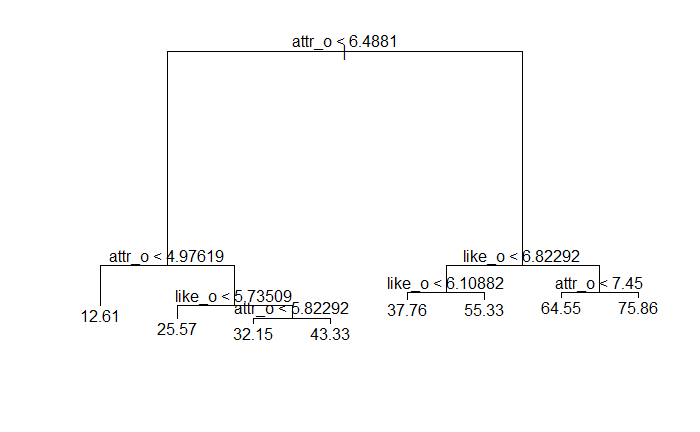
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Multiple Linear Regression | Information Gain Method | Boruta Method | rpart |
| gender | \* |  | \* |  |
| attr\_o | \* | \* | \* | \* |
| fun\_o |  | \* | \* | \* |
| shar\_o |  | \* | \* | \* |
| intel\_o |  |  | \* |  |
| sinc\_o | \* |  |  |  |
| like\_o | \* | \* | \* | \* |
| prob\_o | \* | \* | \* | \* |
| tvsports | \* |  |  |  |
| fun1\_1 | \* |  |  |  |
| shar1\_1 | \* |  |  |  |
| sinc2\_1 | \* |  |  |  |
| attr3\_1 | \* |  |  |  |
| fun3\_1 | \* |  |  |  |
| subject\_yes\_per | \* | \* | \* |  |
| field\_soft\_sci | \* |  |  |  |

**Building the Regression Tree Models**

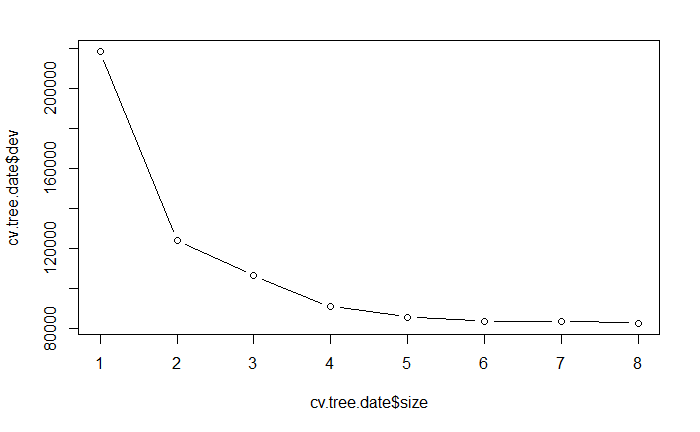
Tree models were chosen to identify the conditions under which participants receive high “yes” decisions from their partners. This method was chosen due to the ease of interpretability of the models as well as the wide range of tree-building packages available in r.

*tree() and cv.tree()*

The first tree model was built using tree() in the “caret” package in r. This model chose only 2 variables for inclusion – attr\_o and like\_o – and had 8 terminal nodes. The model shows that subjects who have an average attr\_o score greater than 6.5 and an average like\_o score of greater than 6.8 have partner “yes” responses of ~65% and higher (see Figure 5). The MAE and RMSE were 10.67 and 13.31, respectively. The RMSE indicates that the average difference between the predicted and observed values of other\_yes\_per in the model is 13.31. The MAE indicates that the average absolute value of the prediction error is 10.67.

Figure 5. Regression tree model using tree() function in r with parameters set to default

Next, cv.tree() was used to run a 10-fold cross validation model to determine the optimal number of nodes for the tree based on cross-validation error. The lowest cross-validation error was found for the tree with 8 nodes (see Figure 6), suggesting that pruning this model would not lead to better performance.

Figure 6. Level of cross-validation error in the regression tree model across number of terminal nodes used

*rpart()*

A regression tree using rpart() in the rpart package was next created to determine if better performance could be achieved using a different regression tree method. This model was found to be identical to the previous tree model, with 8 terminal nodes and similar splits across the attr\_o and like\_o variables (see Figure 7). Tuning of the complexity parameter (cp) was attempted in order to improve performance of the tree. Similar to the tree() method, optimal cp was selected based on the lowest cross-validation error. Unlike the tree() model, this model was found to have the lowest cross validation error for a tree with 7 splits. A new pruned model using 7 splits was run and the MAE and RMSE were calculated (see Table 3). These performance metrics show poorer performance compared to the first unpruned tree() model.

A series of other hyperparameters were tested for the rpart() model using a grid search. A sequence of values for minsplit and maxdepth were first specified and then a data frame containing all possible combinations was created. An rpart() model was created for each of these minsplit and maxdepth combinations, and RMSE and MAE values were computed. The best performing model had a minsplit of 50 and maxdepth of 4. Again, only the attr\_o and like\_o variables were used, and 8 terminal nodes were created (see Figure 8). The RMSE and MAE for this model showed slightly better performance than the original rpart() tree (see Table 3).

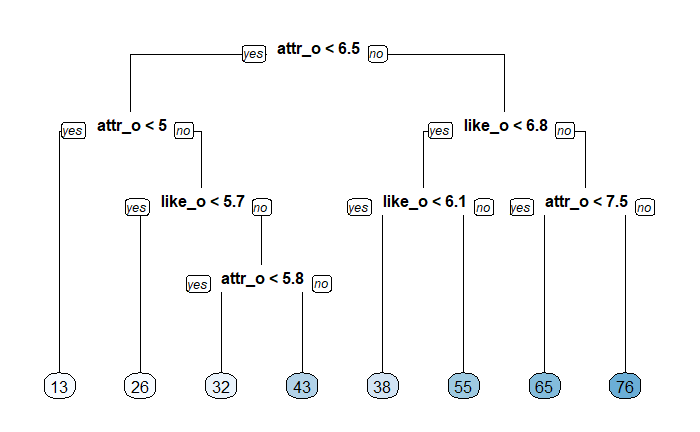


Figure 7. Original rpart() tree

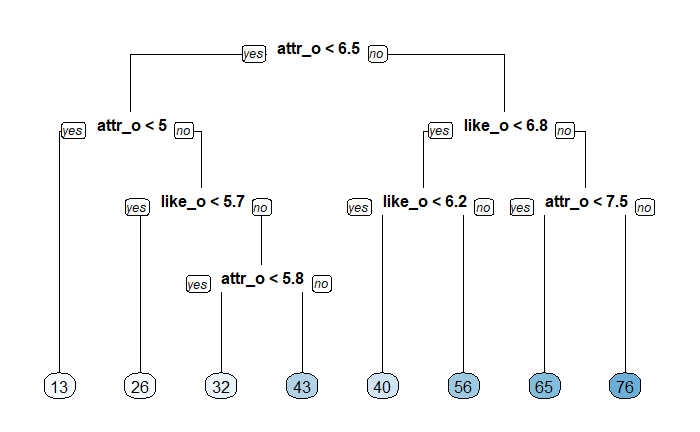


Figure 8. rpart() tree with hyperparameters tuned using grid search

*randomForest()*

A random forest model was next built using the default values in the randomForest package. Five hundred trees were generated, with 2 variables tried at each split. This model explained 66% of the variance in the dependent variable. RMSE and MAE values showed the best performance of all the tree methods tested so far (see Table 3).

A plot of variable importance was generated and shows that for the random forest method, attr\_o and like\_o remain the two most important features in the model (see Figure 9). A plot was created to determine the error across number of tree generated (see Figure 10). This plot shows that the error in the model was greatly reduced after generating more than 100 trees.

tuneRF was next used to determine the optimal value of mtry for the model. This function supported the findings of the original random forest model and determined that mtry = 2 is optimal.

A second random forest model was created to determine if increasing the number of trees generated and specifying mtry = 2 would improve performance. This model resulted in a slight improvement of RMSE and MAE (see Table 3).

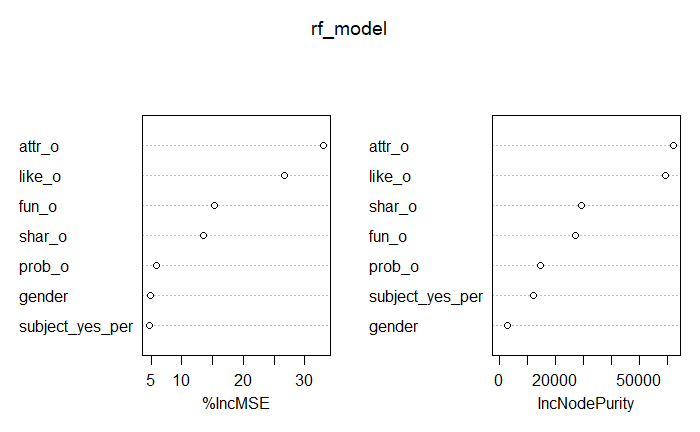


Figure 9. Variable importance according to the random forest model

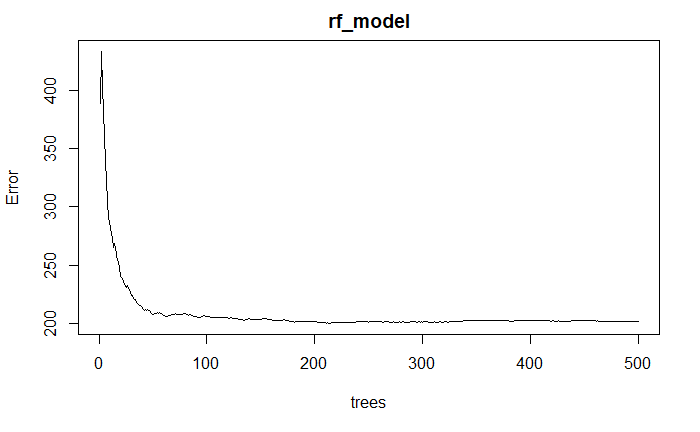


Figure 10. Error in the random forest model across number of tree generated

*Random Forest Grid Search Using trainControl() and train() Functions*

A random forest grid search was customized using the trainControl() and train() functions in r. The grid search was done using the hyperparameters maxnodes and ntree. The best tuned model was found to be for maxnodes = 8 and ntree = 2000. A plot of the RMSE values per maxnode and ntree condition is shown below (see Figure 11). This is consistent with the findings of the other tree models where the best model resulted in a tree with 8 terminal nodes. The performance metrics of this model, however, indicate poorer performance compared to the other models.

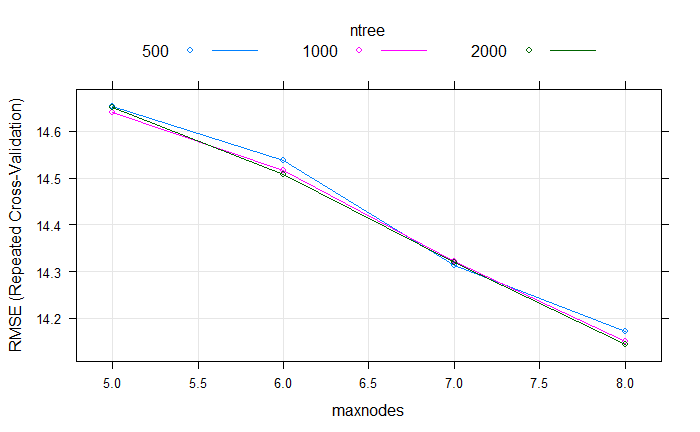


Figure 11. RMSE values for each condition in the random forest grid search

Table 3

*RMSE and MAE Values Found for Each Regression Tree Model*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | tree() unpruned | rpart default | rpart pruned | rpart grid search | random forest default | random forest mtry = 2, ntree = 1000) | random forest grid search |
| RMSE | 13.305 | 13.305 | 13.467 | 13.43 | 13.102 | 13.035 | 13.48 |
| MAE | 10.669 | 10.669 | 10.76 | 10.381 | 10.176 | 10.15 | 10.67 |

**Discussion**

The results of the above analyses suggest that in a regression tree model, attr\_o and like\_o are the most important predictors of other\_yes\_per. This means that the physical attractiveness and likeability ratings that a participant receives from their partner are important to predicting how many “yes” decisions they receive at the speed-dating event. These results are not surprising given that in a 4 minute interaction, other qualities of the participant may not have time to shine through, whereas that person’s physical attractiveness is a quality immediately available for assessment. As well, it seems logical that likeability would predict partner decisions to say “yes” to a potential future date with the participant. The results support the findings of Fismen et al. (2006 and 2008) and Luo and Zhang (2009) which found that physical attractiveness plays an important role in romantic attraction regardless of gender group.

The analyses did not support the theory that participant perceptions of how others see them would play an important role in predicting dating success in a regression tree model. Because of the large amount of missing data for the attr5\_1 to amb5\_1 variables (participant perception of how others see them), these variables were not able to be included in the analysis. A future analysis of the data set could use a smaller subset of data which only includes cases having data present in these fields.

The analyses also did not support the theory that participant self-ratings would play a role in predicting dating success. Only two of the self-rating variables – attr3\_1 and fun3\_1 (self-ratings of attractiveness and funness) – emerged as potentially important features when conducting the multiple linear regression with backwards deletion. These variables, however, were not flagged as important by the other feature selection methods used. This discrepancy is likely due to the different computational methods used to select the important features. The backwards deletion method works by removing features which are insignificantly correlated with the dependent variable. This is done over multiple iterations until there is no longer improvement to the model. Though attr3\_1 and fun3\_1 were included in the final regression model, these variables were both only significant at p = .03, while attr\_o and like\_o were significant at p < .0001. This suggests that there is a linear relationship between attr3\_1, fun3\_1, and other\_yes\_per, but the magnitude of this relationship is not large compared to the other variables. This relationship was therefore likely not captured by the other feature selection methods which use recursive partitioning (rpart), an entropy-based filter (information gain), and a wrapper built around random forest classification (Boruta).

Overall, the results of the regression tree analyses do not provide much useful information about the conditions under which participants enjoy greater dating success. It is interesting to note that though differences in dating preferences between gender groups is well established in the literature, this variable was not used as a splitting criterion by the regression trees generated. Future analyses of the data set may benefit by separation of the data by gender group prior to running the regression tree models. Some features may emerge as more important depending on which gender group is being analysed. This may lead to more informative regression tree models.

**References**

Bale, C. & Archer, J. (2013), “Self-Perceived Attractiveness, Romantic Desirability and Self-Esteem: A

Mating Sociometer Perspective”, *Evolutionary Psychology*, 11(1), 68 - 84.

Campbell, W. K., Rudich, E. A., & Sedikides, C. (2002), “Narcissism, self-esteem, and the positivity of

self-views: Two portraits of self-love”, *Personality and Social Psychology Bulletin*, 28, 358-368.

Fisman, R., Iyengar, S. S., Kamenica, E. & Simonson, I. (2006), “Gender Differences in Mate Selection:

Evidence from a Speed Dating Experiment”, *The Quarterly Journal of Economics*, 121, 673 – 697.

Leary, M. R., Tambor, E. S., Terdal, S. K. & Downs, D. L. (1995), “Self-esteem as an interpersonal monitor:

The sociometer hypothesis”, *Journal of Personality and Social Psychology*, 68, 518-530.

Luo, S. & Zhang, G. (2009), “What leads to romantic attraction: Similarity, reciprocity, security, or

beauty? Evidence from a speed-dating study”, *Journal of Personality,* 77(4), 933-963.

Wu, K., Chen, C. & Greenberger, E. (2019), “Nice guys and gals can finish first: Personality and speed-

dating success among Asian Americans”, *Journal of Social and Personal Relationships*, 36(8)

2507-2527.