What Makes a Perfect Match? Exploring Speed Dating Survey Data

Minwoo Ahn, Adam Badawy, Ruqin Ren, Chubing Zeng

University of Southern California

Abstract

Understanding dating behavior is intrinsically interesting and practically important for every individual in the society. People want to find someone who they want to share stories and emotions, understand and sympathize, commit the rest of life together. Accordingly, people spend a huge part of life finding "the one" or "soul-mate" who they believe potentially maximize their happiness and satisfaction in life. Ironically, they often end up breaking up and saying "He/She was not the right person". Researchers in many areas have studied dating behavior in varying ways to understand why people repeat the vicious circle and still get in there to find the right person. Here, we investigated dating behavior by analyzing relations between multi-aspect variables that include physical and psychological features of individuals and the probability of match in a speed-dating situation. We used theoretical approach and machine learning approach to investigate the pattern of dating behavior and to find the best predictor of match in dataing. For theory driven approach, we used multilevel linear model and multilevel logistic regression. For machine-learning approach, we used learning vector quantization and extreme gradient boosting.

What Makes a Perfect Match? Exploring Speed Dating Survey Data

Introduction

Data

The dataset was gathered from participants in experimental speed dating events Fishman and colleagues from 2002-2004 (Fisman, Iyengar, Kamenica, & Simonson, 2006). During the events, participants had a four-minute date with every other participant of the opposite sex. At the beginning and the end of each wave, participants were asked to answer questionnaire that includes diverse variables (e.g., attributes, gender, race, activities, and etc.), and if they would like to see their date again. For instance, they were asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. The dataset also includes questionnaire data gathered from participants at different points in the process, which includes demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information. There are 8375 observations and 58 variables in the raw dataset. We first dropped a large fraction of the features that we end up not using, fills in missing entries, and drops examples coming from very small events. The original dataset has a row for each participant on a given date rather than a single row for each date. we joined the dataset with itself so that the rows correspond to dates, with all information about both participants present in a single row. After data processing, there are 4184 observations of dates and 53 independent variables.

Multilevel-linear Models

Having a romantic relationship is one of the most exciting and complicated events in our life. It brings us happiness, joy, excitement and sometimes sadness, frustration, and misery. Researcher have strived to investigate factors that contribute to a successful

relationship and high relationship satisfaction. In psychology, researchers have focused on mechanisms involved in mindreading to understand this issues; to have a successful romance, understanding partner's mind is critical. One needs to infer partner's thought, feeling, and needs, and one reflects them into his behavior. These mental strategies have been studied by researchers in psychology for a long time with different terminology (Ross, Greene & House, 1977; Fiske & Neuberg, 1990; Hamilton & Sherman, 1994; Wimmer, & Perner, 1983; Frith, & Frith, 2006; Ames, 2002a; Ames, 2002b; Tamir and Mitchell, 2013).

A few researchers recently proposed cognitive models that postulate perceived similarity as a moderating role of social inferential strategies (Ames, 2002a, 2002b; Tamir and Mitchell, 2013). Ames (2004a, 2004b) proposed 'global perceived similarity' as a mediating factor in mindreading strategies. He found that people use different structure of knowledge when they make inferences on other's mind; they project themselves onto others when they find others similar, whereas use stereotyping to others found to be dissimilar. Similarly, Tamir and Mitchell (2013) found that people use the self as an anchor and a target as a subject to be compared, while adjusting the difference between themselves and the target in social inferences. The general findings from the literature above are mental processes involved in mindreading strategies are not unconscious nor automated processes; rather, they are conscious, deliberate, and effortful mental operations. They are also vulnerable to egocentric biases; one understands other minds based on what they feel right, fluent, and consistent (Epley, Keysar, Van Boven, & Gilovich, 2004; Epley, Morewedge, & Keysar, 2004; Lin, Keysar, & Epley, 2010). Thus, in order to accurately understand other minds, one needs to consciously remove the influence from self-centered biases, stay objective, and take other's perspectives.

Because of this automated operation in the mind, it is not difficult for us to understand why people have little success in a situation like a speed-dating; they're not given enough time to find out partner's thought, feeling, motive, and most importantly, the

discrepancy between their own perception on themselves and partner's perception on them. All of these psychological processes require considerable conscious/deliberate effort to control egocentric biases, maintain objective/accurate perception on the self and others, and fill the gap between those two different perspectives. Thus, it is reasonable to conjecture that one with good mental capacities to successfully process those psychological stages would have a better chance to get matched in a speed-dating. Here, I hypothesize that people with better perception about the discrepancy between their own impression of themselves and a partner's view on them will have a better chance for getting matched (i.e., self-insight; Solomon and Vazire, 2016). In other words, the probability of match is the function of self-insight. If this hypothesis turns out to be valid, I expect a negative relation between self-insight scores of each attribute and the probability of match. I'll describe more details on the model below.

Method

Participant

551 participants volunteered for the experiment. Among 551, 130 participants were removed due to their incomplete and random responses on the survey. After cleaning outliers, female participants were 214, and male participants were 207. Model description

To test and verify the hypothesis, self-insight scores on five attributes (i.e., attractive, fun, intelligent, ambitious, and sincere) were calculated and used as independent variables:

$$SIij = \beta_{0ij} + (\beta_{0ij} - \beta_{1i}Pperception_j) + (\beta_{0ij} - \beta_{2i}Gperception_j)$$

In the equation (1), self-insight (SI) of each participant $_i$ on each attribute $_j$ is the sum of the average of self-measure on each attribute (β_{0ij}), the difference between the self-measure and a partner's perception on each attribute ($\beta_{0ij} - \beta_{1i}Pperception_j$), and the different between the self-measure and a partner's perception on each attribute ($\beta_{0ij} - \beta_{2i}Gperception_j$). The dependent variable was the conditional probability of match:

$$P(dec|match) = P(yes|dec)P(yes|match)$$

In the equation, P(yes|dec) is the probability of saying yes on a question, "Do you want to see this person again?", and P(yes|match) is the probability of getting matched with the partner. This score was calculated and rounded up by two decimal points.

Results

In the analysis, I conceptualized the data as a two level structure; each attributes were nested within individuals, and the individuals were nested within gender. In this data, participants meet a number of potential mates (between 9 and 21) for four minutes each, and have the opportunity to accept or reject each partner. Also, the data includes individual ratings and preferences on various items (e.g., attributes, activities, races, gender, majors, and etc.).

Analysis: within self-insight measure between gender.

In this analysis, I examined the relationship between within self-insight scores between gender by using lmer package in R. The results are as follow:

Attractive

The expected probability of match (jMatch) for an individual self-insight attractive score (siAttr) across gender is 0.44 with a standard deviation of 0.057. The estimated gain in jMatch per point of siAttr is -0.03, with a standard deviation of 0.014.

Fun

The expected jMatch for an siFun between gender is 0.4 with a standard deviation of 0.066. The estimated gain in jMatch per point of siFun is -0.017, with a standard deviation of 0.016.

Ambitious

The expected jMatch for an siAmb between gender is 0.39 with a standard deviation of 0.078. The estimated gain in jMatch per point of siFun is -0.014, with a standard deviation of 0.022.

Sincere

The expected jMatch for an siSinc between gender is 0.41 with a standard deviation of 0.001. The estimated gain in jMatch per point of siSinc is -0.015, with a standard deviation of 2.2.

Intelligent

The expected jMatch for an siIntel between gender is 0.44 with a standard deviation of 0.12. The estimated gain in jMatch per point of siIntel is -0.014, with a standard deviation of 0.02.

ANOVA was conducted to see whether there is significant difference of each self-insight scores and the probability of match between gender. Except the attractive variable, all the other attributes were significantly different between gender at $p < 0.001$.

Theory-inspired Models

Mixed-effects Models

In my prediction of matching, I used different models that were theoretically motivated. I wanted to compare how well individual and mixed-effects models with theoretically important predictors would perform in comparison to machine learning models. The first model I used is logistic regression, to serve as the baseline, which we compare our predictions to. The second two models are random intercept mixed-effects and

random intercept and random slope mixed-effects models. For my part, I did some preprocessing for the variables I am using in my model. First, I created a new column("diffage") that basically take the absolute difference between the participant and the partner. Second, I logged the age values of the dyad. Third, I logged the value of wealth. Furthermore, I divided the dataset into two, the first one to be used as the training set and the second part as the testing set. I did the same as well for my response ("match") to use it for prediction purposes. For the individual level logistic regression model, I used match (decision of the respondent and decision of the partner) as the response and the following as the predictors:

1. Attraction: rating by the respondent of the potential partner
2. Wealth: wealth of the respondent
3. Same-race: 1 if yes, 0 if no
4. Difference in age: absolute value of the difference in age between the respondent and partner

My prediction accuracy for this model is 83%, and my AUC is 50.5%, which is not much better than a random guess. This clearly shows that we need a more sophisticated model to predict matching than a simple logistic regression with few important features or predictors.

As for statistical significance, all the variables show strong statistical significance, which is not surprising, except for samerace (Table 1, Column 1). This can be interpreted in two ways: either that difference of race does not influence the probability of matching or that since in this experiment, the subjects are all college students (either undergraduate or graduate students) that for this kind of sample, difference in race is highly tolerated in comparison to the larger public.

For mixed-effects model logistic regression model, I used match as the response, different intercept for each person and the following as the predictors:

1. Attraction by the respondent

2. Attraction by the partner

3. Wealth: wealth of the respondent

4. Same-race: 1 if yes, 0 if no

5. Difference in age: absolute value of the difference in age between the respondent and partner

My prediction accuracy for this model is 84.8%, and my AUC is 58.5%, which is better than the performance of the individual level logistic regression better than a random guess.

As for statistical significance, all the variables show strong statistical significance, except for samerace (Table 1, Column 3).

I tried a different mixed-effects model with the same response and predictors, but with having a different intercept for each individual and a different effect of same race for each individual (different slope). The results were quite close to the random intercept model, without showing much improvement. The prediction accuracy for this model was 84.8%, and my AUC is 58.4%, which is better than the performance of the individual logistic regression,but again not much better than the simpler random intercept model.

As for statistical significance, all the variables show strong statistical significance, except for samerace again (Table 1, Column 2).

Effect of race on gender

There has been considerable literature written on how race has a different effect on women vs. men. Iyengar et al. 2006 shows results that point to the notion that "women exhibit stronger racial preferences than men." According to them, a possible explanation for this that each gender has different goal when it comes to dating, that women are more interested in forming a relationship, so race becomes an issue from the start. On the other hand, men are initially more interested in casual sex, so even if they care about the race of their potential partner, it will not come up earlier in the relationship. In another article by the same authors, they show that women are 14% points more likely to accept a partner of their own race ("Given the underlying YesRate of 38 percent"), which is a large effect, while men showed no significant racial preference (Fisman, Iyengar, Kamenica, & Simonson, 2008). Thus, I wanted to explore this relationship.

Using an individual logistic regression model, I put decision of the respondent as the response and the following as the predictors:

1. Attraction by the respondent
2. Wealth: wealth of the respondent
3. Difference in age: absolute value of the difference in age between the respondent and partner
4. Gender: 1 if male, 0 if female
5. Same-race: 1 if yes, 0 if no
6. Same-race*Gender: interaction terms of samerace and gender

As for statistical significance, all the variables show strong statistical significance, except for samerace and the interaction term of samerace and gender. This goes against

my hypothesis, but again, it may be explained by the high education of the subjects and that we may see such effect with a sample that is more representative of the population.

Machine Learning Approach

Aside from the theory-driven approaches, it is also beneficial to try out machine learning approaches to make predictions. Next section discusses several machine learning algorithms, including Learning Vector Quantization (LVQ) method for model training, univariate filtering and recursive feature selection for diminishing feature dimensions.

This problem can be treated as a binary classification problem, with 0/1 two possible match outcomes and 52 features for prediction.

Learning Vector Quantization

Learning Vector Quantization (LVQ) is a special case of artificial neural network. Assume that a number of reference vectors w_k are placed in the input space. Usually, several reference vectors are assigned to each class. An input vector x is decided to belong to the same class to which the nearest reference vector belongs. There are several distance measuring methods to calculate which reference vector is the nearest, and one commonly used distance measure is Euclidean distance. Starting with properly defined initial values, the reference vectors are then updated as follows according to LVQ2.1 algorithm in Kohonen (1995):

$$w_i(t+1) = w_i(t) - \alpha(t)(x - w_i(t))$$

$$w_j(t+1) = w_j(t) - \alpha(t)(x - w_j(t))$$

where $0<\alpha\ (\mathrm{t}) < 1$, and $\alpha\ (\mathrm{t})$ is a learning parameter that can be set by the researcher to reflect the speed of learning. The two reference vectors w_i and W_j are the nearest to x; x and w_j belong to the same class, while x and w_i belong to different classes. This can be understood simply as that the updating process shortens the distance between

x and w_j and furthers out the distance between x and w_i. Essentially, this is an iterative process of updating the reference vectors. Additionally, there are several tuning parameters: number of closest vectors that determine the classification result (usually set to 1 but could be larger than 1), sizes of reference vectors, distance measure, learning speed rate α.

The LVQ method can be implemented in *caret* package in R. Using our data set, the packages automatically selects the best model to be size of reference vectors = 22, k nearest vectors = 1. The best model generates testing set AUC of 53.11% and training set accuracy rate of around 83%, which is still not ideal and did not beat the results of random chance. The results are shown in figure 4 and figure 5.

Additionally, to compare with the theory-driven approach in finding the most predictive features among the 52 of them. LVQ based importance ranking is also presented here in figure 6.

Feature Selection: Univariate Filtering

The LVQ method that uses all 52 features did not yield good enough results, so we further explore if some feature selection algorithms could help refine the model. First, univariate filtering is tested.

Univariate filtering methods select a small number of features based on univariate statistics assessing the potential of individual features for class prediction. It does so in a one-by-one fashion, then removes all the features that are individually less correlated with the prediction outcome (Miller et al., 2009). Then the training is conducted using random forest method.

This method is also implemented in *caret* package. The method helps to conduct 10-fold cross validation, and it also tunes the models to find a best model. On average, 21.2 variables were selected using univariate filtering (min = 18, max = 25). The best model uses 25 features, about half of total 52 available ones. The best model generates testing set

area under the curve of 72.95%. This is a much better result than using all 52 features in LVQ method. The ROC curve is presented in figure 7.

Feature Selection: Recursive Feature Elimination

To try out another feature selection method, we also tested recursive feature elimination (RFE) algorithm. Guyon, Weston, Barnhill, and Vapnik (2002) proposed variable selection by means of recursive feature elimination (in companion to Support Vector Machine) for selection of genes in micro array data. The aim of RFE is to identify a subset of predetermined size m $<$ p of the set of p available variables for inclusion in the classifier. The procedure starts with all p variables in the classifier, and then sequentially removes one variable at a time until m variables remain (Louw & Steel, 2006). The number of m is a parameter that can be tuned.

The result of RFE feature selection and bag of trees classifier is presented in figure 8. The best model generates testing set ROC curve of 71.94%, which is slightly worse than the previous model.

Extreme Gradient Boosting

XGboost is short for "Extreme Gradient Boosting". The model of XGboost is tree ensembles, just like in random forest. The tree ensemble model is a set of classification and regression trees (CART). We classify the members of a family into different leaves, and assign them the score on corresponding leaf, and a real score is associated with each of the leaves. Usually, a single tree is not strong enough to be used in practice. Therefore, we use tree ensemble model, which sums the prediction of multiple trees together. The final score of each observation is the sum of the prediction score of each individual tree. An important fact is that the two trees try to complement each other. Boosted trees and random forests are not different in terms of model, the difference is how we train them.

Mathematically, we can write our model in the form: $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$, Where f_k is the set of all possible CARTs. Our objective is to optimize the objective function: $obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} G(f_k)$.

The XGboost method can be implemented in *xgboost* package in R. Applying XGboost to our datasets, there are 756 CART in the tree ensembles. Figure 9 and Figure 10 shows two examples of CARTs. From Figure 9, if the male participant in the pair is not funny (fun rate smaller than 55), and is not physically attractive, then the prediction score would be low. On the other hand, if the male participant is funny, physically attractive, and likes physical activities, then the the prediction score is higher. Figure 10 is another plot of the individual decision tree. One important fact is the the two trees try to complement each other. For example, intelligence is not on individual decision tree 1 but is an important attribute on individual decision tree 2. Similarly, if the male participant in the group is not smart and the female participant has high preference for partner's attractiveness, then the prediction score is low. And if the male participant is smart and attractive, and the female participant expect to be happy with the speed dating event, then in general the prediction score is high.

The result of ROC curve of XGboost is presented in figure 11. The AUC score for XGboost model is 78%.

Feature Selection: XGboost

Figure 12 shows the feature importance score for each variable. From this figure, the five most important features include physical attraction of male participant, physical attraction preference of male participant, funniness of male participant, age of male participant and sincerity preference of male participant. The five least important features include intelligence level of female participant, how much the female participant likes shopping, expected number of people who will be interested in dating you and same race.

Figure 13 is the correlation matrix of 20 most important features identified by XGboost. Big circles are for high positive correlations and big orange circles are for high negative correlations. High positive correlation includes fun vs. attractiveness, sincerity preference vs. date frequency, and physic activity vs. attractiveness preference. High negative correlation includes attractiveness preference of male vs. intelligence preference, sincerity preference, and ambition preference, intelligence preference of women vs. ambition preference and funniness preference, attractiveness preference vs. ambition preference and fun preference.

Models Comparison

After discussing each one of the predicting models, table 3 summarizes the results and compares model fits by area under ROC curve metric. The two theory driven models are not performing well as mentioned earlier, partly because of the limited number of variables selected. Machine learning techniques such as learning vector quantization, univariate filtering and recursive feature elimination improved the results to nearly 72% area under ROC in testing sets. The best model is XGBoost, which uses ensemble learning based on decision tree methods.

The project also highlights the importance of finding the correct features. First, despite the conventional wisdom of "it depends on the research question" or "it should be theory-driven," our results indicate that purely theory-driven methods did not outperform "blind" machine learning that does not care much about exactly are the features that go into the model. Second, even among machine learning techniques, classifiers perform much better when they are helped with some sort of feature selection methods. It is very important for any kind of model, theory driven or machine learning based, to take in only the most important features. It teaches an important less for researchers to not only care about finding the correct theory-valid variables, but also consider broadly of all possible

predictors, because many of them may be useful in ways that are not intuitively understandable.

References

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*, 673–697.

Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2008). Racial preferences in dating. *The Review of Economic Studies*, *75*(1), 117–132.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389–422.

Kohonen, T. (1995). Learning vector quantization. In *Self-organizing maps* (pp. 175–189). Springer.

Louw, N., & Steel, S. (2006). Variable selection in kernel fisher discriminant analysis by means of recursive feature elimination. *Computational Statistics & Data Analysis*, *51*(3), 2043–2055.

Miller, W. R., Larionov, A., Renshaw, L., Anderson, T. J., Walker, J. R., Krause, A., ... Dixon, J. M. (2009). Gene expression profiles differentiating between breast cancers clinically responsive or resistant to letrozole. *Journal of Clinical Oncology*, *27*(9), 1382–1387.

Table 1

	Dependent variable:		
	match		
	logistic	*generalized linear mixed-effects*	
	(1)	(2)	(3)
attr	0.452***	0.551***	0.552***
	(0.027)	(0.034)	(0.034)
attr_o		0.508***	0.508***
		(0.032)	(0.032)
wealth	0.530**	0.407	0.419
	(0.235)	(0.314)	(0.315)
samerace	−0.117	−0.133	−0.204*
	(0.090)	(0.118)	(0.106)
diff_age	−0.069***	−0.066***	−0.066***
	(0.016)	(0.018)	(0.018)
Constant	−6.236***	−9.995***	−10.009***
	(0.848)	(1.156)	(1.161)
Observations	4,189	4,189	4,189
Log Likelihood	-1,684.197	-1,491.628	-1,492.477
Akaike Inf. Crit.	3,378.394	3,001.256	2,998.954
Bayesian Inf. Crit.		3,058.318	3,043.335
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 2
Interaction-Term Model

	Dependent variable:
	dec
attr	0.692***
	(0.018)
wealth	0.461***
	(0.132)
samerace	−0.008
	(0.076)
diff_age	−0.018**
	(0.009)
gender	0.248***
	(0.066)
samerace:gender	0.002
	(0.105)
Constant	−6.409***
	(0.480)
Observations	8,378
Log Likelihood	-4,533.609
Akaike Inf. Crit.	9,081.219
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3
Models Comparison

Classifier/Training Model	**Feature Selection Method**	**Area Under ROC**
Theory1		
Theory2		
Learning Vector Quantization	None	53.11
Random Forest	Univariate Filtering	72.95
Bagging Trees	Recursive Feature Elimination	71.94
XGboost	Adaptive Training	78

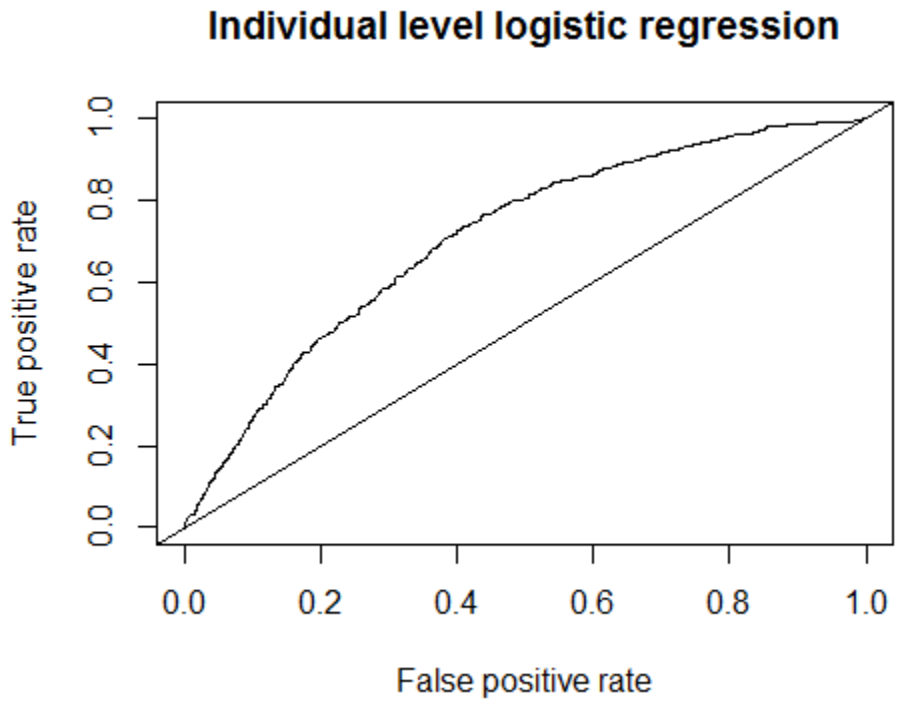

Figure 1. ROC Curve

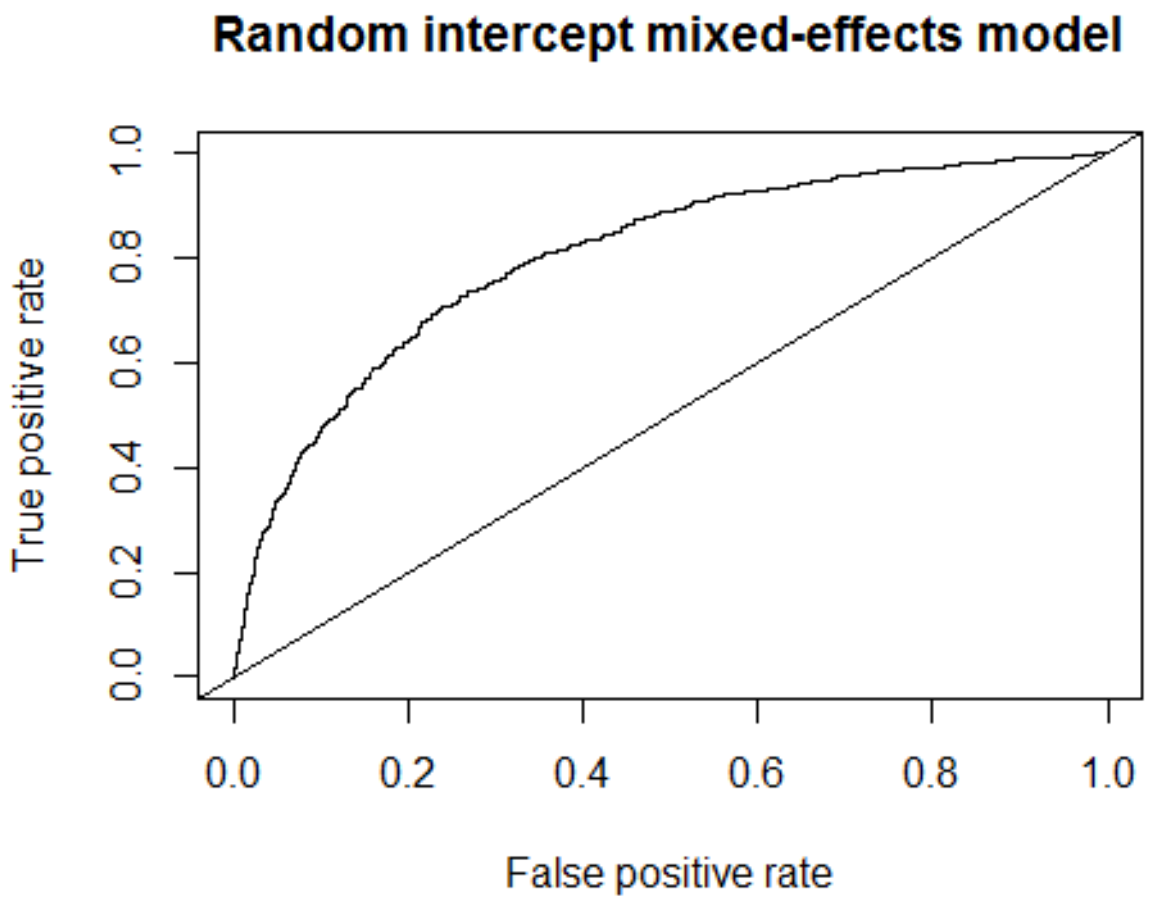

Figure 2. ROC Curve

Figure 3. ROC Curve

```
Pre-processing: scaled (52)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1883, 1883, 1882, 1883, 1882, 1883, ...
Resampling results across tuning parameters:

  size  k   Accuracy   Kappa
  22     1  0.8350832  0.03368475
  22     6  0.8333295  0.04107645
  22    11  0.8330166  0.03917646
  33     1  0.8314195  0.02016958
  33     6  0.8296643  0.02385939
  33    11  0.8323787  0.04051632
  44     1  0.8307815  0.03994952
  44     6  0.8336516  0.06808597
  44    11  0.8323764  0.06714789
```

Figure 4. Learning Vector Quantization Model Tuning

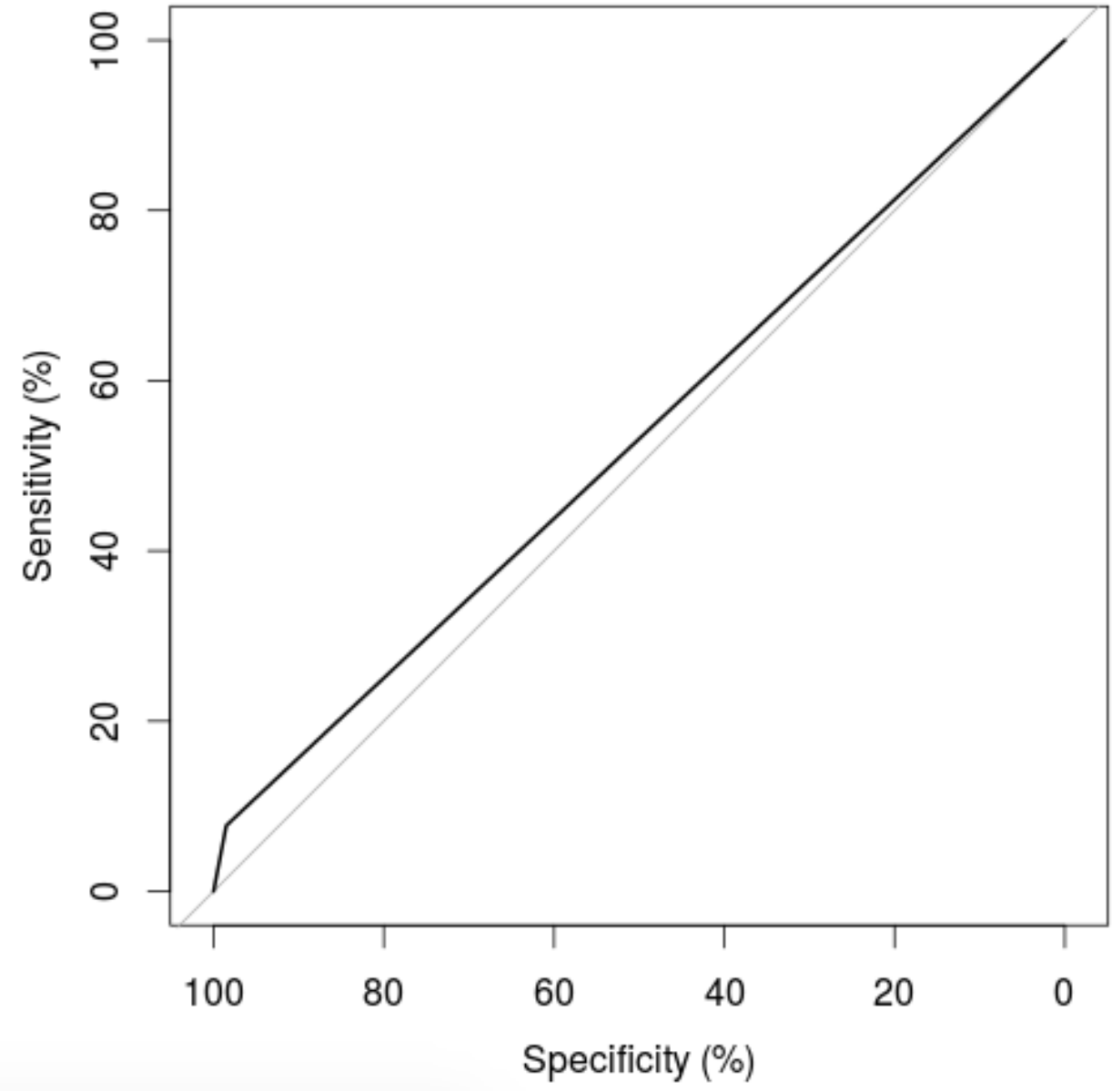

Figure 5. Learning Vector Quantization ROC Curve

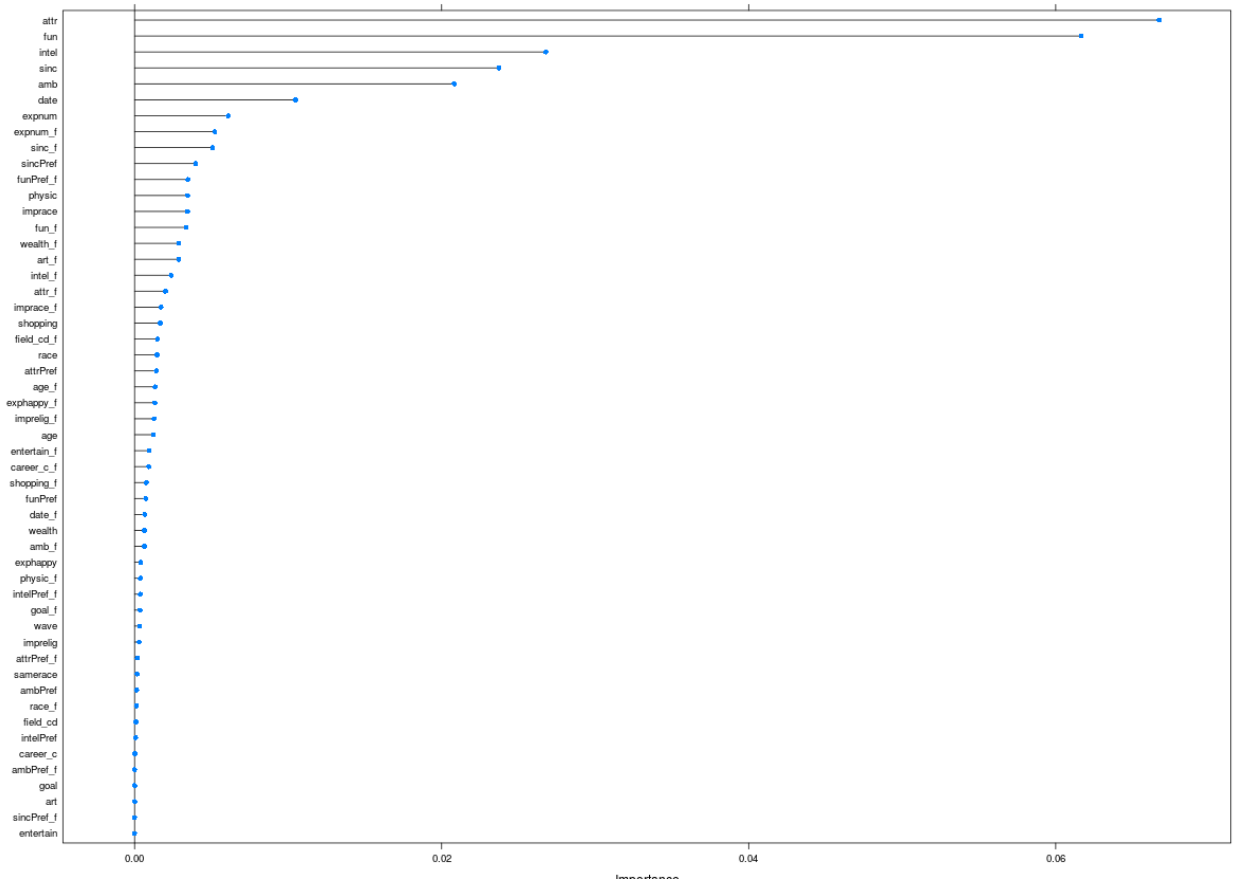

Figure 6. Learning Vector Quantization Feature Importance Ranking

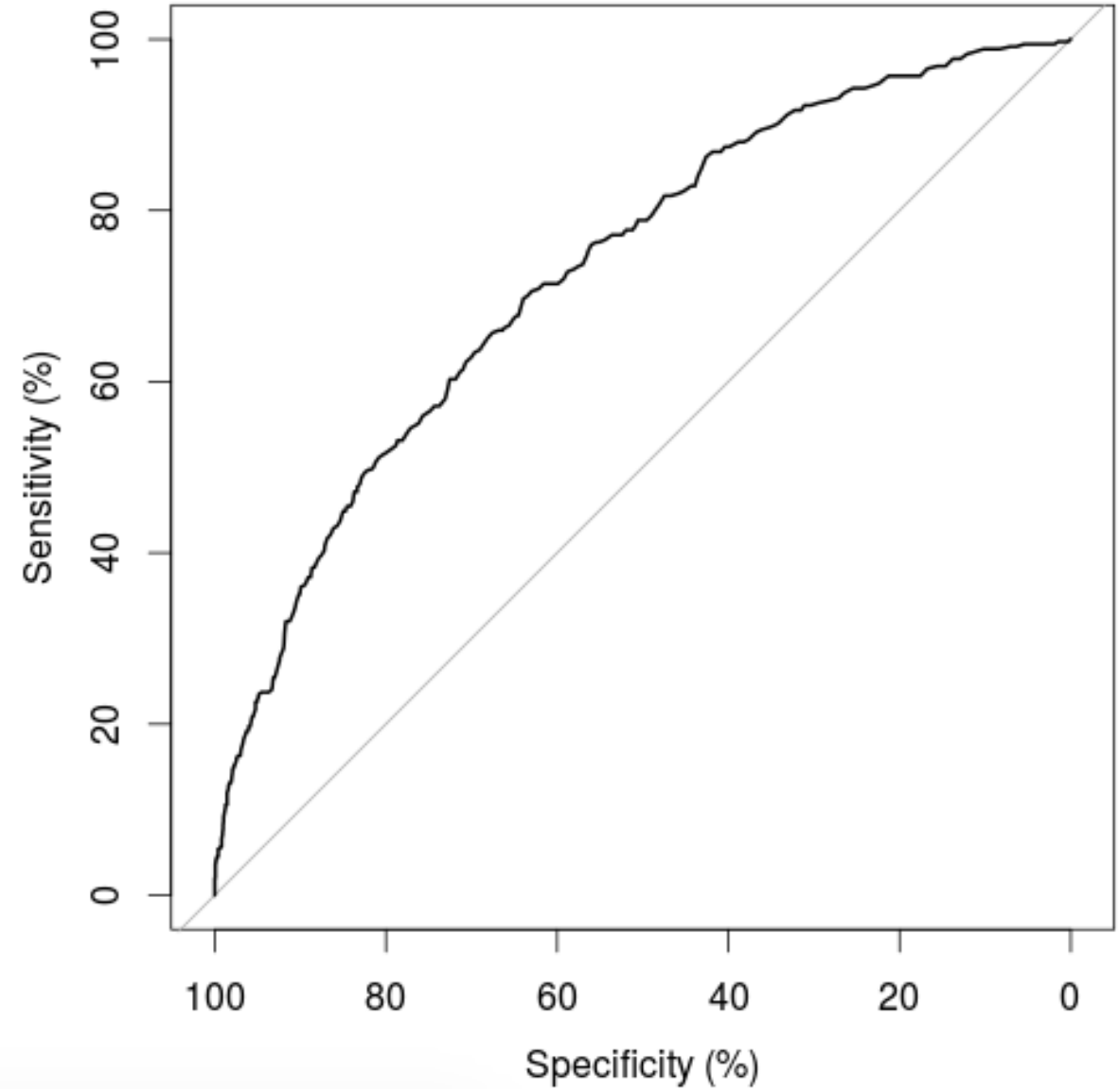

Figure 7. Univariate Filtering and Random Forest Prediction ROC Curve

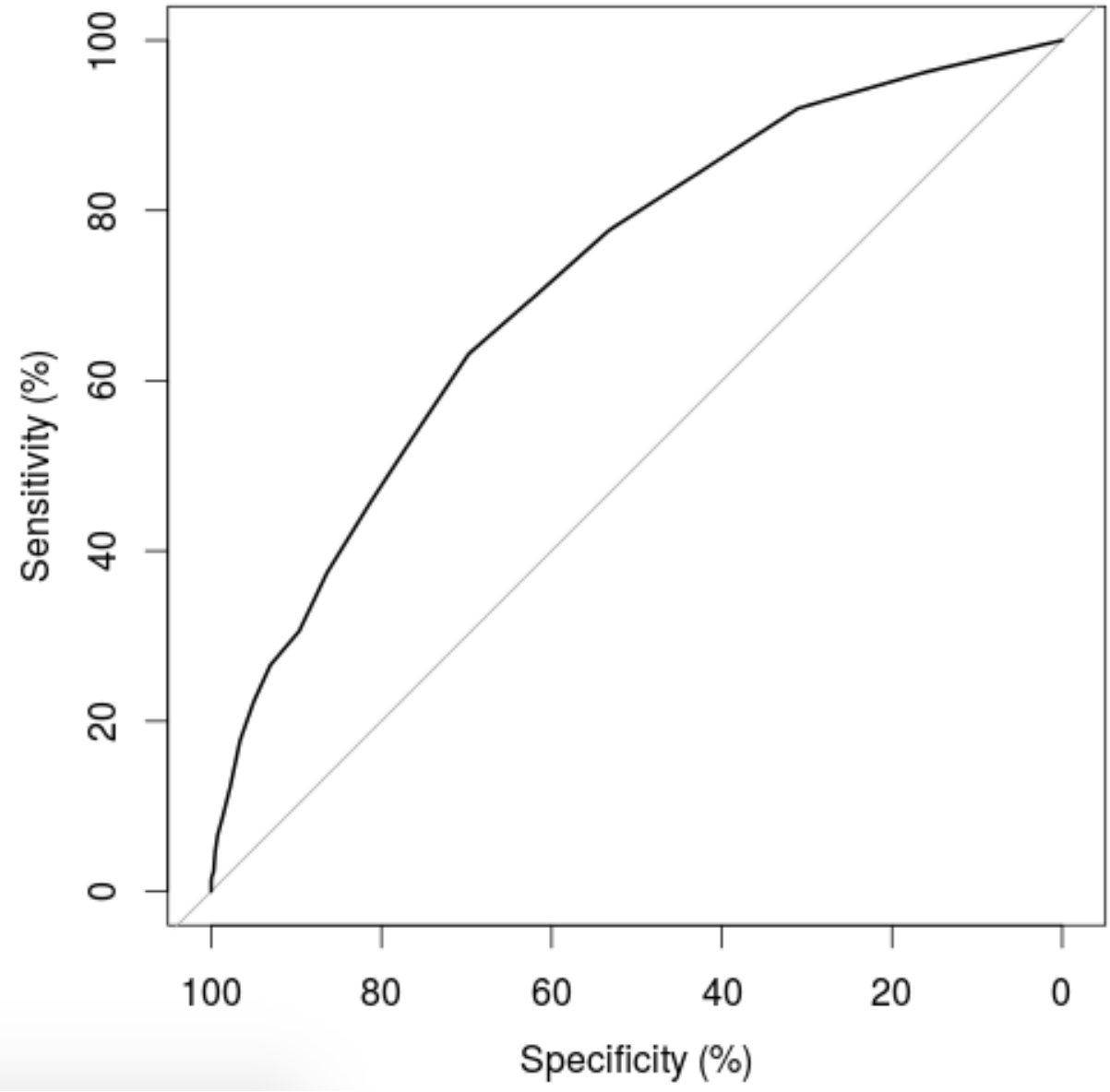

Figure 8. Recursive Feature Elimination and Bag of Trees Classification ROC Curve

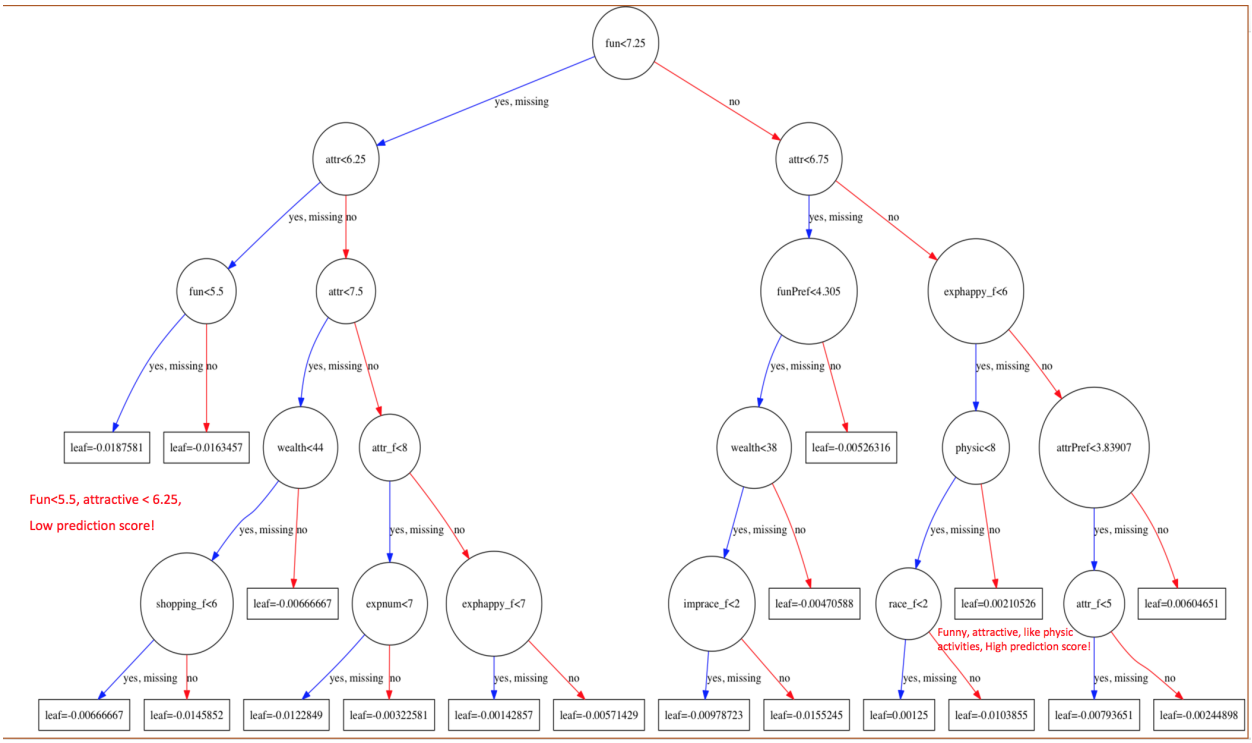

Figure 9. Classification and Regression tree: example 1

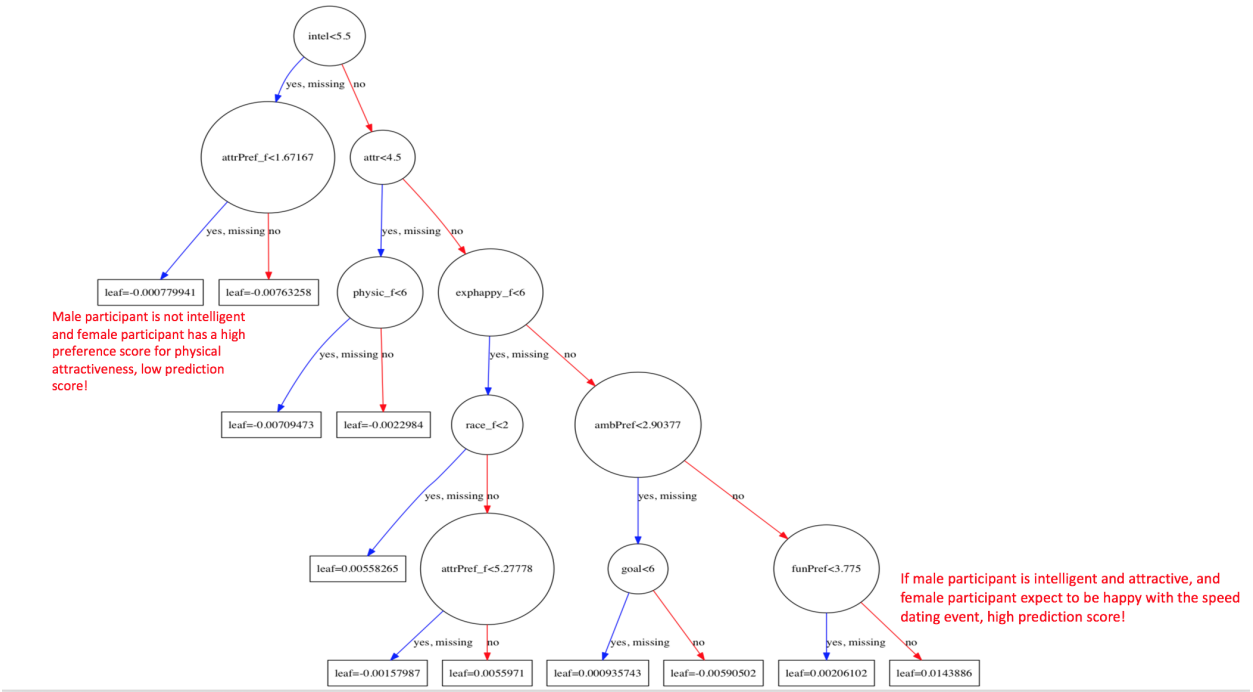

Figure 10. Classification and Regression tree: example 2

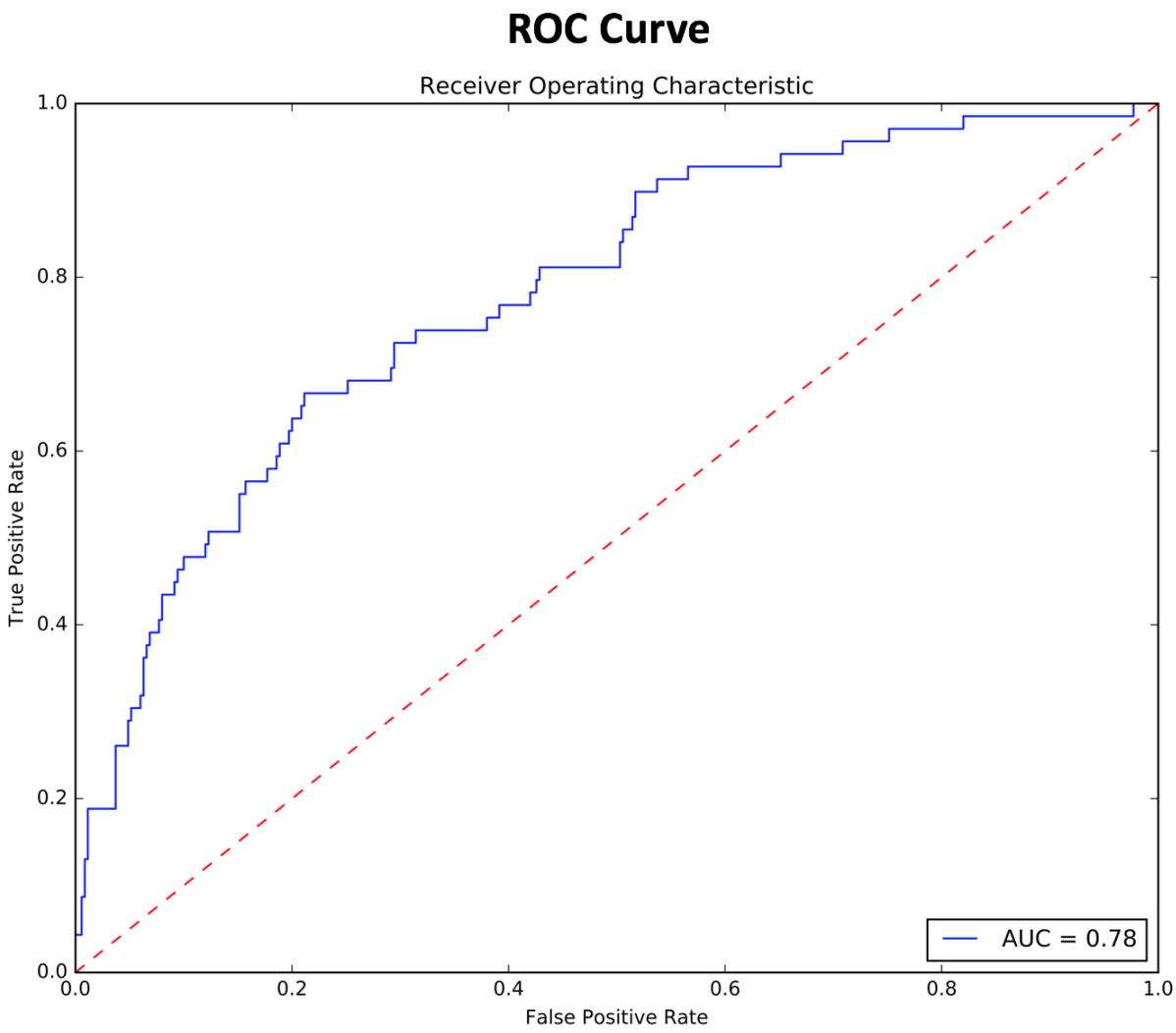

Figure 11. XGboost: ROC curve

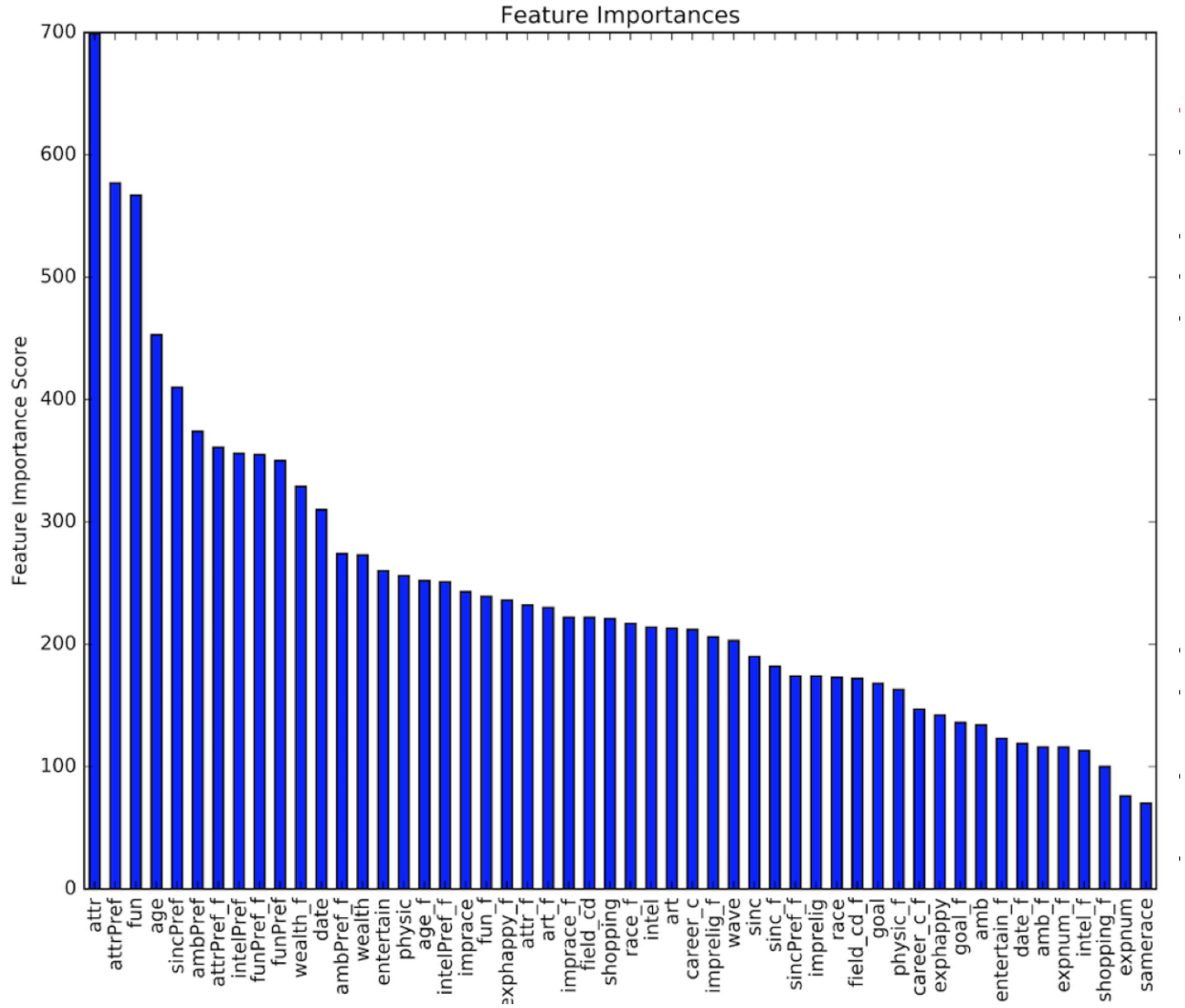

Figure 12. Feature Importance Score

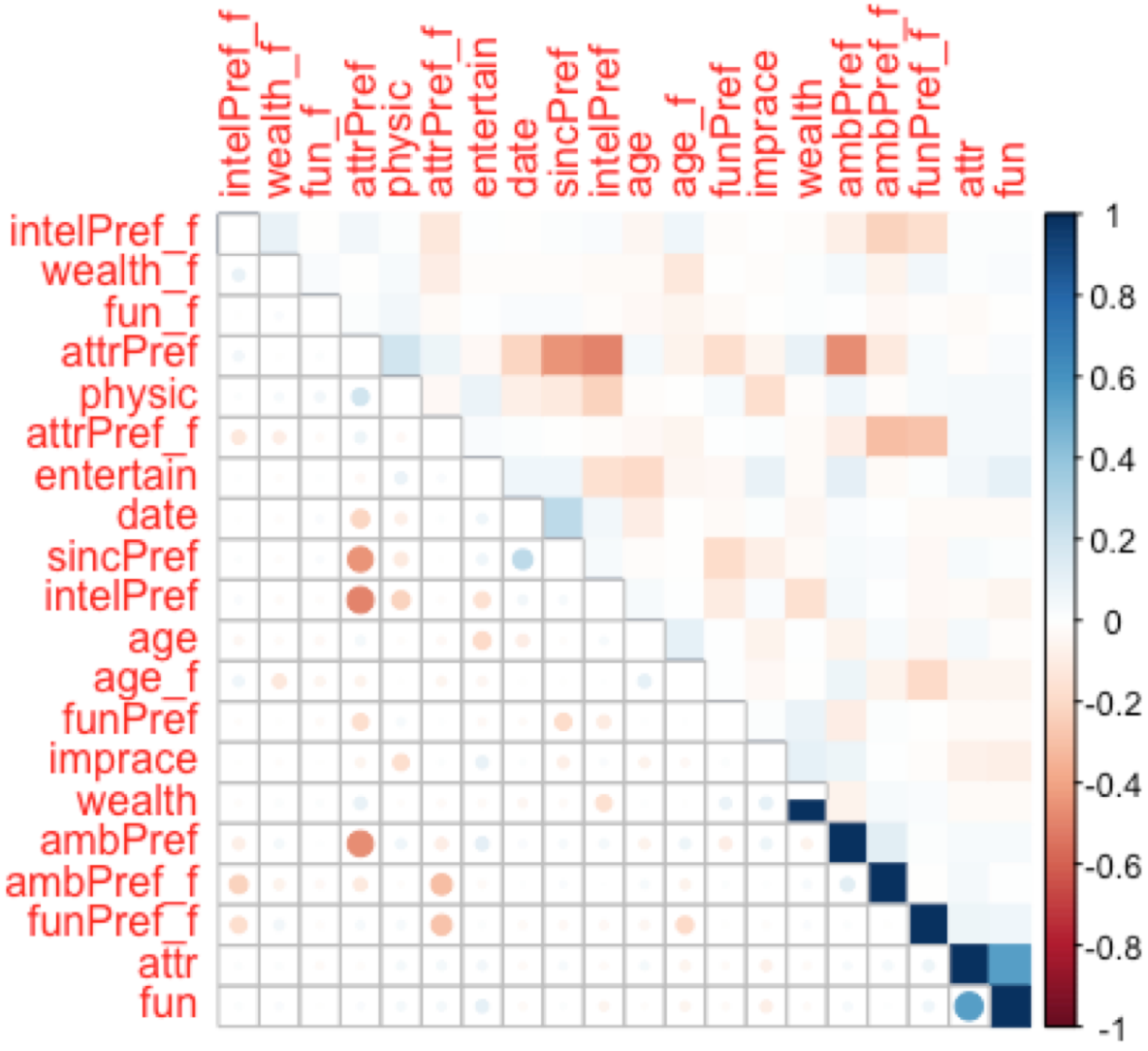

Figure 13. Correlation matrix for the 20 most important features