Speed Dating Analysis*

Statistical Analysis II 6337 Final Project

Léon Yuan

4/29/23

This report analyzes racial preferences, six primary attributes, and their relations to dating behaviors from graduate school at Columbia University in New York City from 2002 to 2004.

Table of contents

1 Introduction **2**

1.1 Outline experiment/data collection 2

1.2 What are you expecting to see (hypotheses)? 2

1.3 Importance/intrinsic interest of question addressed 2

2 Design & Data Collection **3**

2.1 How ambitious is the data collection and analysis required for your investigation 3

2.2 How appropriate/ideal are data for the question posed? 3

2.3 Clearly defined statistical hypotheses and correct description of how they link to question(s) of interest 3

2.4 Randomization (if relevant) & scientific rigor used in data collection and described in detail 4

3 Data Analysis **4**

3.1 Appropriate selection of statistical methodology 4

3.2 Analysis carried out correctly 4

3.3 Assumptions met? Appropriate remedies if not, or discussion of effects 5

3.4 Correct and complete interpretation of results 5

3.5 Proper and effective visualization of data, illuminating findings (See in Appendix) 6

4 Conclusion **6**

4.1 Concise and accurate summary of findings 6

4.2 Generalization / Scope of Inference 6

4.3 Thoughtful and realistic discussion of limitations and extensions/future questions 7

Appendix **7**

[*] Thank Dr. Stokes and other classmates for instructions and peer-helps through Spring 2023!

1 Introduction

From the past to now, human beings are a species in the world that most need to look for a mate or partner to live a life. In this context of human society, people name it dates. Then as more romantic emotions between mates become more robust, people name it relationships. After both mates and their affiliated relatives agree on this relationship, both mates get married and form a formal social bond protected by laws designed by humans.

1.1 Outline experiment/data collection

The data were male and female subjects or graduate and professional students who studied at Columbia University from 2002 to 2004 and volunteered to participate in this social science dating study. The whole distribution of participants in this sample is close to the distribution of the whole of Columbia University at that time.

The experimental design was each male participant would engage in conversations with each female participant within four minutes, as following Figure Figure 1. Then each subject of pairs would write down a survey to make decisions and rate scores to six attributes: Physical Attractiveness, Sincere, Intelligent, Fun, Ambitious, and Shared Interests of your partners. Next, each male rotates through all female subjects so that all people in one session(night) meet and talk to each other.

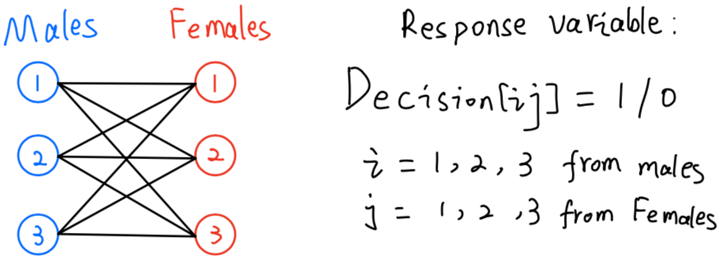

Figure 1: Dating Design

Before the dating experiments started, all students' demographic information would be collected, including race, age, major, intended career field, where they are from, income, and so on.

1.2 What are you expecting to see (hypotheses)?

I have four hypotheses to expect:

1. Asian males are the least popular in the dating market among all races and genders, while White females are the most popular.
2. Females of all races prefer same-race dating more than males do.
3. Both males and females of all races emphasize physical attractiveness at the top.
4. Males emphasize physical attractiveness much more than females do.

1.3 Importance/intrinsic interest of question addressed

This study and experiments are very useful and important for human beings to explicitly understand the **Biological Evolution** and promote a stable and harmonious society. In biology, as we all know, males tend to seek mates that are time-limited reproductive capacity signaled by females' appearance: physical attractiveness, while females tend to seek mates that are the ability of resource acquisition to raise their offspring: intelligence and ambition. The results of the study can further help males and females realize their shortcomings and how to improve their attributes to find their matched mates.

2 Design & Data Collection

2.1 How ambitious is the data collection and analysis required for your investigation

The data is quite ambitious. The first step collection is finding volunteers of students on campus to participate in this experiment. The second step is collecting all participants' demographic information like, race, age, major, income, and career. The third collection is the most time-consuming and complicated because the experiment would make all participants date one on one from session 1 to session 14 with a four-minute conversation per each date, then let them make decisions and record their ratings and scores to each other. After all these steps are completed, the final data is 8378 rows and 195 variables, which shows this is a big data set. Since there are 195 variables there, there are too many different analyses and models which are reasonable to investigate. According to my hypothesis and questions, the analysis I would use involve both statistical modeling/inference and machine learning prediction/inference on a relatively big data set.

2.2 How appropriate/ideal are data for the question posed?

The subjects collected through this experiment were from Graduate and Professional Schools at Columbia University from 2002 to 2004. The samples were representative of the population of graduate students there. However, inference and prediction results may not be generalized or extended to larger or wider populations, such as New York City or even the whole country. However, within the scope of graduate school there, the data could be considered appropriate. This experimental design provided us with data that is very similar to a real-world dating setting because all dates happened in a real bar or restaurant, and all subjects gave their true decisions and attribute scores without under any public social pressure or political correctness as all feedback and survey were de-identified.

2.3 Clearly defined statistical hypotheses and correct description of how they link to question(s) of interest

1. Asian males are the least popular in the dating market among all races and genders, while White females are the most popular.

   $$H_0 : \beta_{Asian} = 0 \ ; H_a : \beta_{Asian} \neq 0$$

   where this parameter from a logistic regression that the response variable is the decision made by females.

   $$\beta_{White} = 0 \ ; H_a : \beta_{White} \neq 0$$

   where this parameter from a logistic regression that the response variable is the decision made by males to females.

2. Females of all races prefer same-race dating than males do. This is done by Cochran-Mantel-Haenszel Chi-Squared Test for Count Data.

$$H_0 : \text{true common odds ratio of interracial dating by gender is 1}$$

$$H_a : \text{true common odds ratio of interracial dating by gender is not 1}$$

Table 1: Same-Race and Interracial decisions by gender

	Same Race.Female	Different Race.Female	Same Race.Male	Different Race.Male
Yes	877	652	1199	787
No	1659	1006	1327	871

3. and 4. Both males and females of all races emphasize physical attractiveness at the top, and Males emphasize physical attractiveness much more than females do. These two hypotheses are verified by two separate random forest models. The similar statistical hypothesis would be:

$$H_0 : \text{Gini Decrease}_{attractive} > \text{all other variables' Gini decrease}; H_a : \text{Gini Decrease}_{attractive}\text{is not top}$$

$$H_0 : \text{Gini Decrease}_{attractive;males} \approx 2 \times \text{Gini Decrease}_{attractive;females}; H_a : \text{no twice relation}$$

2.4 Randomization (if relevant) & scientific rigor used in data collection and described in detail

The scientific rigor in data collection would be this dating experiment included all major races, two genders, most common majors/career fields, and domestic and international students across the world. Then all subjects of totally different backgrounds would meet at least once or more so that this experiment tried to incorporate all possible combinations of dating. As a result, statistical modeling and machine learning can have a completely valid input to make inferences and predictions.

3 Data Analysis

3.1 Appropriate selection of statistical methodology

For the second hypothesis that Females of all races prefer same-race dating than males do, I applied **Cochran-Mantel-Haenszel Chi-Squared Test** for Count Data. For the first hypothesis, I applied *ordinary logistics regression*. For the third and fourth hypotheses, I applied **random forest** and **Penalized Logistic Regression**.

3.2 Analysis carried out correctly

1. Carry Cochran-Mantel-Haenszel Chi-Squared Test Rayner and Rippon [1]

I did some data manipulation to aggregate the same race and interracial dating results by gender, as shown in Table 1.

2. Ordinary logistic regression

I conducted two logistic regressions separately for males and females. The reason I built two separate models for females and males is that there are some significant differences in dating behaviors between genders, and separate models are more accessible to interpret.

$$\log(\frac{Pr_i(Yes|X_i)}{1 - Pr_i(Yes|X_i)}) = \alpha + \beta \times \text{samerace}_i + \gamma \times \text{race}_i$$

3. Random Forest

I randomly split data into 80% training and 20% test set. I chose 500 trees and four random predictors at each split. The OOB estimate of error rate is 24.99% which has 75% accuracy on the training set, while on the test set, this RF has roughly 73% test accuracy, which is not bad on this dating data. The true "Yes" rate is $138/(138+118) = 53.90$, which is a little bit over 50% random guess rate. However, the true "No" rate is $347/(347+64) = 84.43$, which is a better prediction rate on the test set because the training set has more "No" classes than "Yes".

4. Penalized Logistic Regression

I built a Penalized Logistic Regression because there are many predictors. Thus selecting and shrinking variables are necessary to build a good logistic regression. I encoded the train data frame into a form of dummy variables for all categorical variables, and I deleted the `from` column because it has 164 unique values, which produces a huge number of variables and also `field`. Then I use Elastic Net with logistics regression on this train matrix with alpha = 0.5. I build a final model with the lambda that gives the simplest model but also lies within one standard error of the optimal lambda selected by cross-validation measured by Binomial Deviance. Finally, I plot the cross-validation plot when selecting lambda.

$$\log\left(\frac{Pr_i(Yes|X_i)}{1-Pr_i(Yes|X_i)}\right) = \alpha + \beta \times \text{samerace}_i + \gamma \times \text{race}_i + \eta \times \text{age}_i + a \times \text{attractive}_i + b \times \text{sincere}_i$$

$$+c \times \text{intelligence}_i + d \times \text{fun}_i + e \times \text{shared interests} + f \times \text{interest correlation}$$

where

$$(1-0.5)/2||\text{coefficients}||_2^2 + 0.5||\text{coefficients}||_1 < \lambda$$

3.3 Assumptions met? Appropriate remedies if not, or discussion of effects

For the Mantel-Haenszel test, it assumes that the odds ratio is the same in different repeats. In our case, the data contains huge observations, which is close to a large sample size. We assume it is met. Random Forest and logistics regression do not make as many strict assumptions as linear regression and other tests do.

3.4 Correct and complete interpretation of results

Mantel-Haenszel: After conducting Mantel-Haenszel, then I found that the p-value is 0.0321, which is slightly less than 0.05; thus, I can tell that the odds ratio is not equal to 1 by gender. Females say more "Yes" to same-race dating compared with interracial dating than males do.

Ordinary logistic regression: First, I only care about how race affects females' decisions to males, only including the `samerace` and `race` columns in this logistic classification model. From the summary of the model, we can tell that all females are likely to reject the `Asian` males because `Asian` males have a 0.008 p-value which is the most significant in this model. The log odds of saying "yes" to Asian males by all females is -0.48977 given other variables fixed, and this is a significant negative coefficient meaning that Asian males are very unpopular when dating. Thus, the odd ratio of saying "yes" to Asian males is $e^{-0.48977} = 0.6127673$ which means when females date Asian males, they likely decrease 40% the probability of saying "yes" to Asian males. The opposite effect happens to white females, who only have a relatively significant p-value with a positive 0.25750 coefficient. This shows white females are the most popular in the dating market.

Random Forest: From the variable importance plots Figure 4 and Figure 6, I roughly classify the top 6 predictors into three classes. The first top class only has one predictor, physical attractiveness. This is consistent with my logistic regression. *Physical Attractiveness* has a 205.911 mean decrease of Gini, which is a measurement of building trees. This Gini decrease is almost twice that of other variables. Thus, *Physical attractiveness* is the most significant factor when females make decision to males. The second top class has *shared interests* and the correlation between participants' and partners' ratings of interests. These two variables are actually highly correlated however, random forest is robust to the highly correlated predictors because of its ability of randomly selecting a subset of variables at each split. *Shared Interests* has a 150.324 decrease Gini of mean, which is as three times as other less important variables. Females emphasize shared interests with males. The third top class has three variables: *fun, from, field*. They have a very close Gini decrease mean of about 125, which is twice as much as other less important variables. Females put equal emphasis on the fun, where males are from, and which field males' careers belong to. Overall, females are likely to date males who are very physically attractive and then have common/shared interests as they do, while males' career fields and where they're from play a secondary role in dating.

Penalized Logistic Regression: From the coefficients, only 6 six attribute variables are kept in the model. All other variables, such as race and age, are filtered out by elastic net variable selection. Age doesn't play any role in this analysis mainly because all participants are about from 20 to 30, which are considered young people. However, if participants of age greater than 40 or 50 are present in this study, age is expected to be significant.

From two plots Figure 3 and Figure 5, I can tell that Random Forest and Penalized Logistic Regression perform almost equally, while Penalized Logistic Regression performs slightly better than Random Forest.

3.5 Proper and effective visualization of data, illuminating findings (See in Appendix)

Asian males are the least popular in the dating market among all races and genders, while White females are the most popular. This finding can be further validated by the data visualization shown in Figure 2 in the appendix.

4 Conclusion

4.1 Concise and accurate summary of findings

Although Asian males are the least popular in dating and females of all races prefer the same race dating, improving physical attractiveness and having common interests can significantly dominate racial preference in dating.

4.2 Generalization / Scope of Inference

All the participants in this data were from graduate and professional schools at Columbia University. The prediction and inference results may only be valid to that school and surrounding area, such as New York City. These experiments and subjects happened in a very international and diverse society. Thus inference may be generalized to other similar social structures like the “Big Blue State”: New York, California. This can not be extended to “Big red state” rural areas in the US.

4.3 Thoughtful and realistic discussion of limitations and extensions/future questions

Limitations: When applying Mantel-Haenszel Chi-Squared Test to racial preference, there may be other confounding variables to affect the result except for gender. There are maybe other interaction terms between race and attributes.

Extension: Given more time, I would like to build a generalized linear mixed model for this data to consider cross-nested effects when all subjects date each other.

Appendix

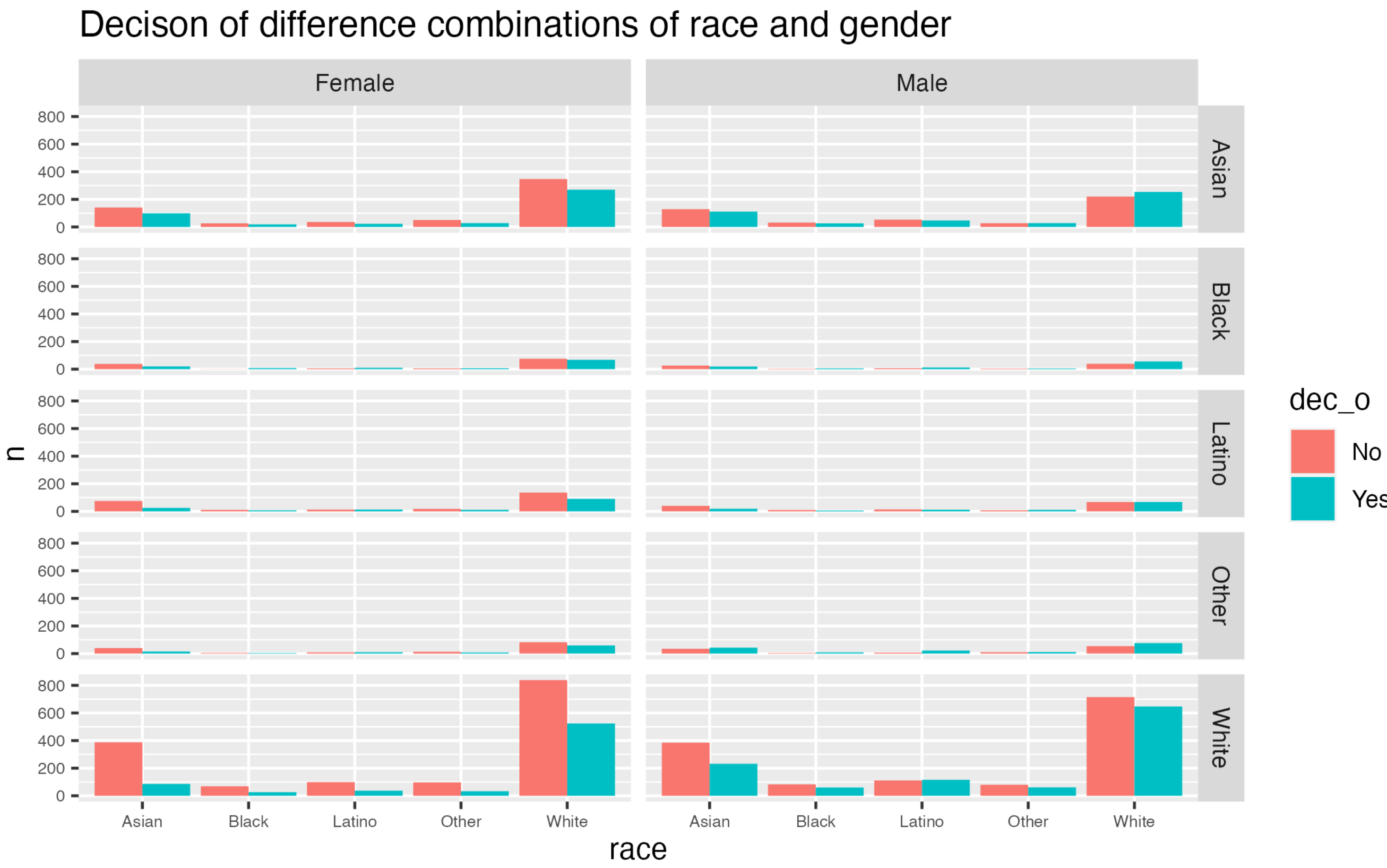

Figure 2: racial preference

References

[1] J. C. W. Rayner and Paul Rippon. "Recent Extensions to the Cochran–Mantel–Haenszel Tests". In: *Stats* 1.1 (2018), pp. 98–111. ISSN: 2571-905X. DOI: 10.3390/stats1010008. URL: https://www.mdpi.com/2571-905X/1/1/8.

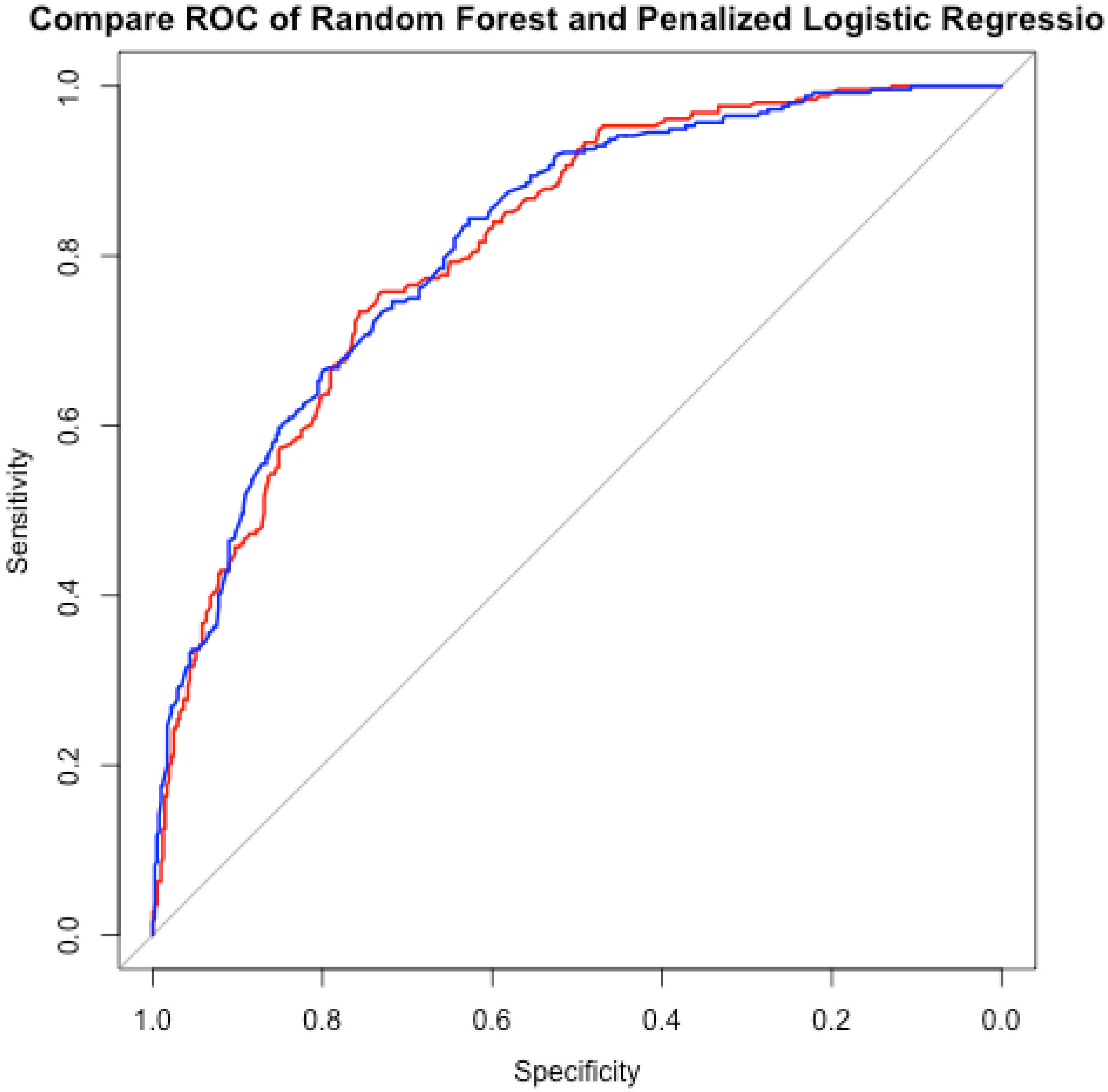

Figure 3: Compare ROC for decisions made by females to males

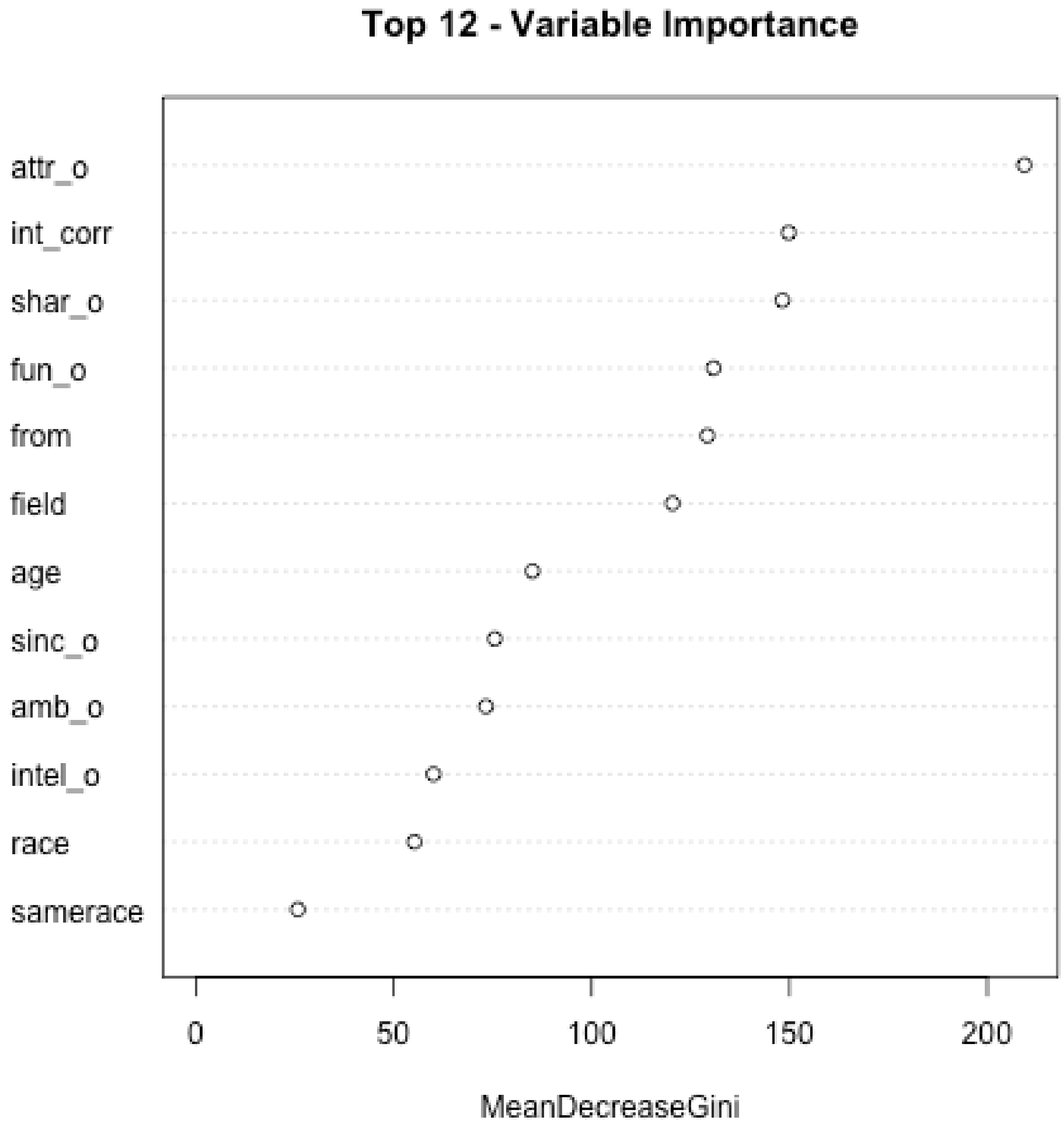

Figure 4: Variable Importance made by females to males

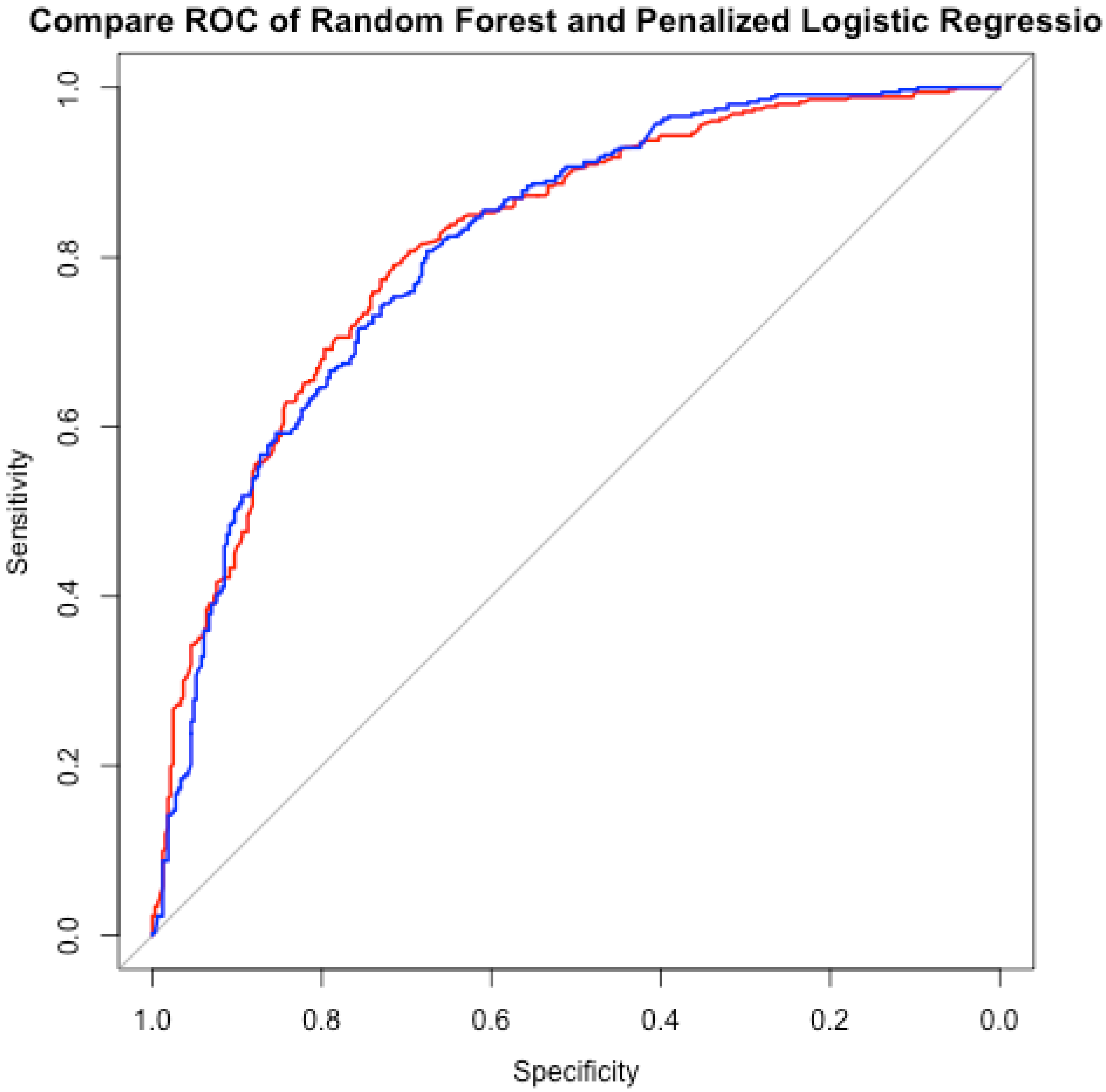

Figure 5: Compare ROC for decisions made by males to females

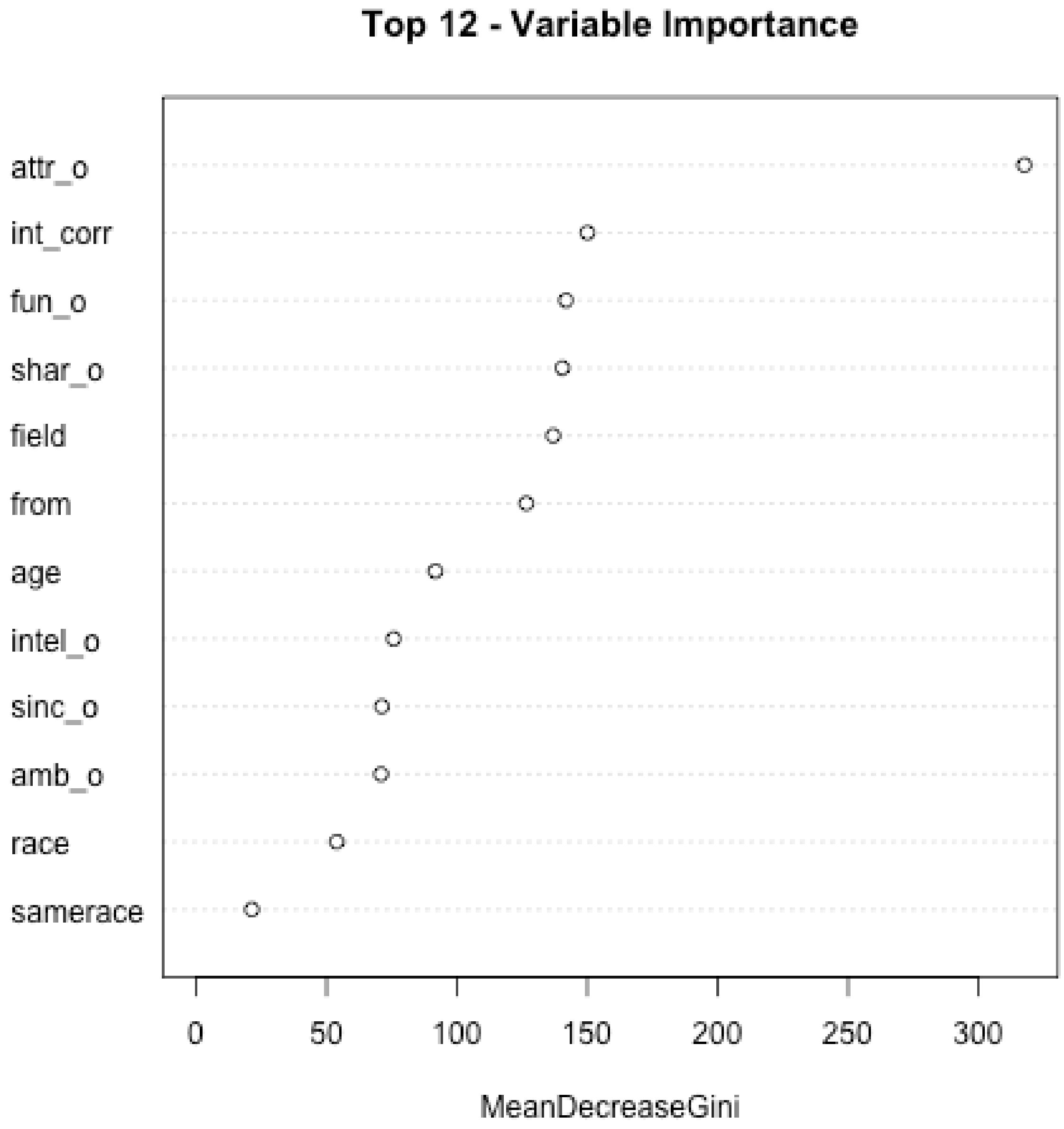

Figure 6: Variable Importance made by males to females