UNIVERSITY OF CALIFORNIA

Los Angeles

# The Secret of Love in Speed Dating

A thesis submitted in partial satisfaction of the requirements

for the degree

Master of Applied Statistics

by

Shi Yunfan

2019

ABSTRACT OF THE THESIS

The Secret of Love in Speed Dating

by

Shi Yunfan

Masters of Applied Statistics

University of California, Los Angeles, 2019

Professor Ying Nian Wu, Chair

Speed dating is a popular and fast way to meet new people and find life partner in nowadays society. Four professors from Columbia university did the speed dating experiment from 2002-2004 and I use their data-set in this paper to answer the research question: what are the gender differences on selecting opposite sex partner from speed dating event, and can we eventually predict people's decisions. In this speed dating experiment, every participant had a chance to meet a new person from opposite sex just through a 4-minute conversation. They collected everyone's basic information such as gender, race, age and so on. Before the speed dating event they also collected participants' hobbies, expectation about opposite sex, and what kind of person themselves are. During the event, each participants would also value how they think their partner is. In this paper, I did basic data analysis to explore gender difference and other useful information about different people's preference on opposite sex. I try to predict males' final decision, if they like the female who they just met, using either all information I have before or after the speed dating event. The base model I use is logistic regression model, and I improved the model by step-wise variable selection. The compared models are decision tree model, random forest model and XGBoost model. I separated the whole data set into 80% training data and 20% testing data to avoid over-fitting. The best model I finally have is XGBoost model. It has a 82.4% precision on testing data-set based on all the information we have after the speed dating event, and still a 70.2% precision on testing data-set even we only use all information before two people never actually met on the speed dating event. So we can believe that we have the ability to discover the secret of love with modern machine learning algorithms if we have enough information.

The thesis of Shi Yunfan is approved.

Frederic R Paik Schoenberg

Vivian Lew

Wu Ying Nian, Committee Chair

University of California, Los Angeles

2019

# Table of Contents

# Table of Figures

# Table of Tables

# CHAPTER 1

## 1 Introduction

To most of us, our wife or our husband is the people we will spend most of our time with. And not like our parents and children, our life partner is the only family member we can choose. So that makes the choice of our marriage partner the most important decision in our life.

But what make people fall in with someone? Some suggested childhood experience, some suggested DNA, some suggested personality. But very few of them came from carefully designed experiment and in a very scientific point of view. It is very hard to draw causal relationship for all social science questions because you can not manipulate human to suit your experiment requirement and control variables. So it is more logical to draw correlations.

In this paper, I will answer the research question what are the gender differences on selecting opposite sex from speed dating event, and can we eventually predict people's decisions.

The data-set I used was gathered from participants in experimental speed dating events from 2002-2004. I will explore and explain relationships between variables in this data-set and to see what makes gender difference in choosing opposite sex partners. I will predict people's decisions using logistic regression as the base model, and gradually improve the result by decision tree model, random forest model and XGBoost model.

# CHAPTER 2

## 2 Experiment design

This experiment and data set was compiled by Columbia Business School professors Ray Fisman and Sheena Iyengar for their paper Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment from 2002-2004.

The experimental was basically a four minute speed dating. Each participants had a four-minute conversations with every participants of opposite gender. They would decide whether or not they are interested in the person and want a second date. If both participants chose “Yes”, then we got a match, and they would be provided with the others contact information to set up the second date. If one of them chose “No”, or both of them chose “No”, it means someone was being rejected and there would be no information exchange. Before the meet-up session and after each of the four-minute conversation, participants would be asked to do questionnaires to answer some questions and how they feel about their partner, and we are interested in what made up their decisions.

Subjects—The experiment subjects came from graduate and professional schools at Columbia University and they were reached out through mass e-mail or fliers all over the campus. To sign up for this experiment, they had to register online and completed a pre-event questionnaire. So the research did not reach out all the population and the way they chose the subjects is biased. The subjects were not been random chosen from all population, so we could not draw any causal relationship of what make people like someone from neither the public or even Columbia University students. It was just an observational study in a more manageable way. But we could still use this data set, not because it is a perfectly designed, but to show we have the ability to analysis what make people like someone, or even predict what kind of person someone will fall in love with.

Setting—The Speed Dating were took place in an private room at a bar near the campus. The experiment designers tried to make table arrangement, lighting, and type and volume of music played constant across events, to have all other variables almost same except experiment subjects.

*Figure 1*

Procedure—One female and one male seated on each side of the table. Males rotated from table to table after each 4-minute conversation, meeting all of the females. After the four minutes, participants would take one minute to record their scores for their partners .

# CHAPTER 3

## 3 Data

The data set is relatively large with 552 individuals, 8378 rows and 71 useful variables. It has 4 types of variables:

1.basic personal information:

subject's information (gender,race,age...)

partner's information (gender,race, age...).

2.variables collected from consensus:

subject's self evaluation (attraction,intelligence,sincere,fun,ambition)

subject's hobbies (movie, dining, hiking...)

subject's demand for partner (attraction,intelligence,sincere,fun,ambition)

3.variables collected from speed dating:

subject's evaluation of partner (attraction,intelligence,sincere,fun,ambition)

partner's evaluation of subject (attraction,intelligence,sincere,fun,ambition)

4.variables created from original variables:

same race (if both subject and partner are from same race)

match (if both subject and partner choose "yes")

partners' self evaluation (attraction,intelligence,sincere,fun,ambition) age difference (age of male - age of female)

5. decision:

subject’s decision for partner (Yes/No)

partner’s decision for subject (Yes/No)

So group 5 is the most important variable and the y variable we want to predict. We want to explore the relationship between variables (group 1,2,3,4) with people’s decision (group 5) in speed dating process.

# CHAPTER 4

## 4 Data analysis

In this chapter, I will discover the general gender difference, and males' and female's different preference of race, age and partner's goal of the speed dating event.

### 4.1 General gender differences

Gender is the most important and basic information about speed dating as the goal of speed dating is to help us find the other half, a male or a female (for heterosexual relationship for this experiment and this paper). So from this point, every variable will be discussion separately by male and female. In this data set, we have almost same number of male and female participants.

Female are more selective in general towards the decision of whether they will go out for the date with the partner.

*Figure 2*

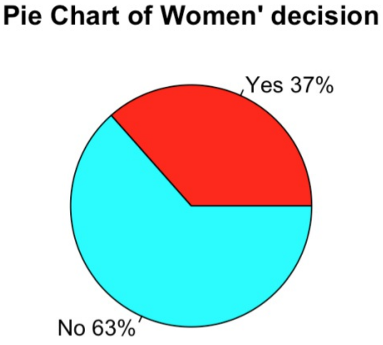
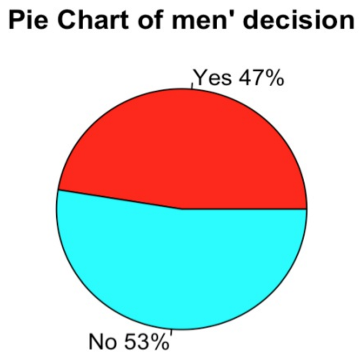


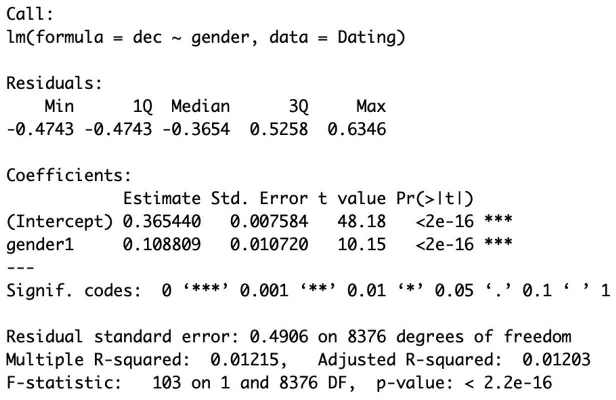
*Figure 3*



In general, female will say yes for 37% of the time and say no in 63% of the time. However, male in general say yes for 47% of the time and say no in 53% of the time, almost half-half for yes and no. So male have 10% chance more to be refused by a female, which correspond to the intuition that women are much more difficult to pursue and less open than male.

We will also confirm it by logistic regression - decision against gender:

*Figure 4*

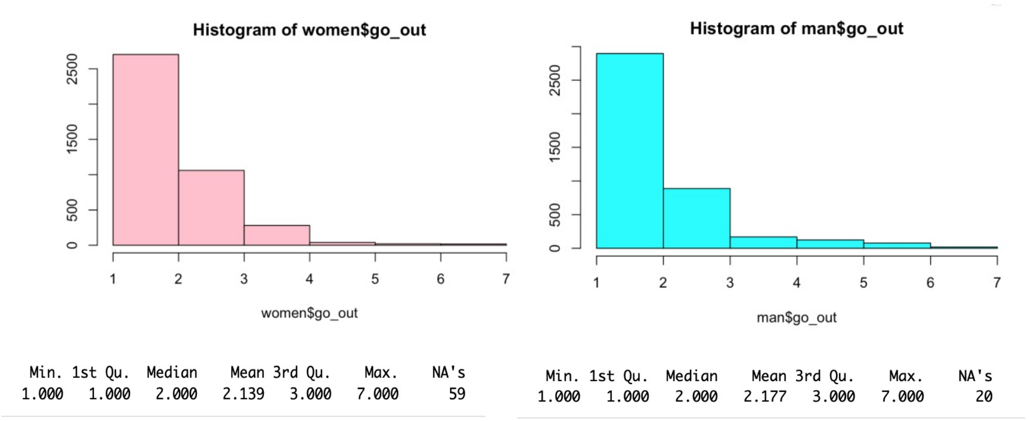


```
Call:
lm(formula = dec ~ gender, data = Dating)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4743 -0.4743 -0.3654  0.5258  0.6346 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.365440   0.007584   48.18   <2e-16 ***
gender1     0.108809   0.010720   10.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4906 on 8376 degrees of freedom
Multiple R-squared:  0.01215,	Adjusted R-squared:  0.01203 
F-statistic:   103 on 1 and 8376 DF,  p-value: < 2.2e-16
```

It confirms that gender difference is highly significant in Yes/No decision.

My first guess will be that women are less out-going than male and they go out less than men. In the experiment, we have typically 10 partners and if we accept 37% for female and 47% for male in general, we have about 4 dates for female and 5 dates for male. Will the reason be that female are less willing to hang out?

So I look into the variable - how often participants go out before the speed dating experiment. Here are the results:



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  1.000   1.000   2.000   2.139   3.000   7.000      59 
```

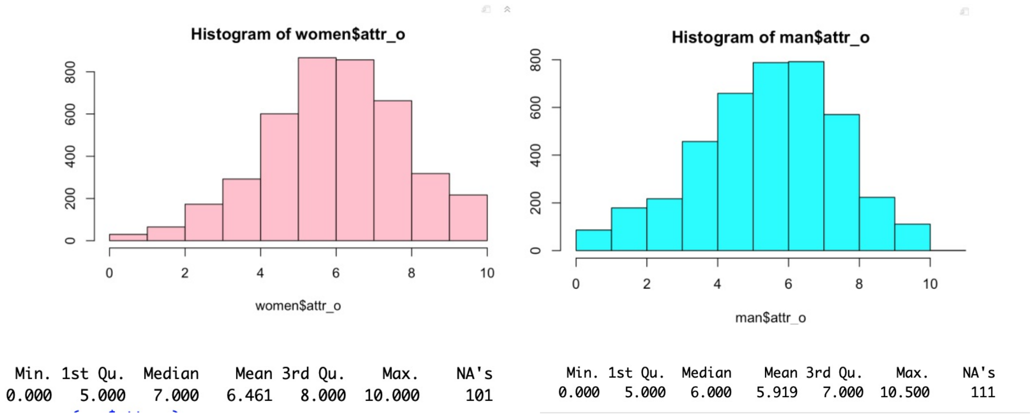
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  1.000   1.000   2.000   2.177   3.000   7.000      20 
```

*Figure 5*

As we can see from the histogram of go_out variable for both female and male, the median are both 2 and mean are very similar,female is 2.139 and male is 2.177, within 2% difference. Therefore, the difference of accept rate for the second date is not due to female are less willing to go out.

The second guess will be that women in this experiment are more desirable than men, so women refuse more male partners.

So I look into the attraction variable, which indicates how attractive you score in your partner’s eye in the speed dating experiment. Here are the results:



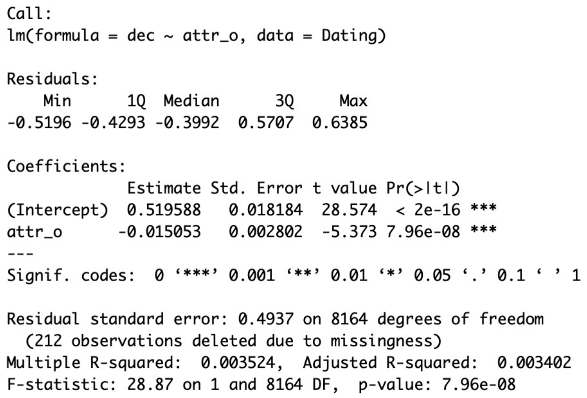
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.000   5.000   7.000   6.461   8.000  10.000     101 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.000   5.000   6.000   5.919   7.000  10.500     111 
```

*Figure 6*

We can see from histogram that more male scored very low(below 2), and more female above 8 (from table we can see there are 25% female score above 8 while only 25% of male score above 7). the median of female is 7 while median of male is only 6, mean of women receive 6.461 while men only receive 5.919 in general. So women receive 9.15% higher in general in attractiveness and perfectly match the 10% difference in Yes/No rate in decision for second date.

However, this is only my best guess for the gender difference in Yes/No rate for second date. We can not draw casual relationship here. It can be totally coincidence. We can do a very simple linear regression model on decision against attractiveness a person score. Here is the summary:



```
Call:
lm(formula = dec ~ attr_o, data = Dating)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5196 -0.4293 -0.3992  0.5707  0.6385 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.519588   0.018184  28.574  < 2e-16 ***
attr_o      -0.015053   0.002802  -5.373 7.96e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4937 on 8164 degrees of freedom
  (212 observations deleted due to missingness)
Multiple R-squared:  0.003524,  Adjusted R-squared:  0.003402 
F-statistic: 28.87 on 1 and 8164 DF,  p-value: 7.96e-08
```

*Figure 7*

We can see that although it indicates attractiveness is highly correlated to the decision making, it is not the only reason, as R square of the model is only about 0.003524. We can not draw the conclusion that difference of decision making is just because of the difference of attractiveness.

So at this point we need to understand that we can not draw any causal relationship between decision and any variables because we can not control the variable in the experiment design. We can not make 10 exactly same person, only have different out-looking score range from 1 to 10, and see what decision their partner will make, whether say yes or no for the second date. All participants are different, and different in various ways. Therefore, we can only draw correlation between decision making and all variables, to have our best guess that what factor affects people’s decision for the second date.

Let’s return to the question we facing right now. What cause the gender difference in decision making for this experiment?

It could be that female are more cautious and serious about dating in nature and thus more elective. Their opportunity cost for dating is higher because from evolution point of view, female need to be pregnant and raise one child at one time, so they need to be more cautious about their mate, make sure their mate will can are able to help them raise the next generation. It could be that female are more mature than male in mental and know what type of person they want, so they can can filter out more people unsuitable for them just for the 4-minute interaction with the male in this experiment. There are lots of reasons we can guess, but we can confirm none of them by the data itself. So from now on, this paper will focus on the result itself, more than the reasons behind these phenomenon.

## 4.2 Race

Race is another very important factor in speed dating decision process and we are very curious about how race matters in this process. “Data from the 5% sample of the 2000 Census reveal that among married blacks, 94% are married to other blacks. Members of other races are also unlikely to marry outside of their own group. While under random matching 44% of all marriages would be interracial, a mere 4% of marriages in the U.S. are between partners of different race”.

I will exam does race matter in decision making process in this experiment and how influential it is.

From calculating, among all matches, the percentage of different race match is 59.0% and percentage of same race match is 41.0%. If we random match, the percentage of different race match will be 53% and the experiment shows no large difference as random match, totally different from the Census result. Again, it has various reasons possible. It could be that Columbia students are highly educated so they have little or no racial bias and discrimination. It could be that rare interracial marriage is because of racial segregation, people have little chance to meet and date people from different race. It also could be that dating is just totally different from marriage while marriage is much more serious than dating. Still, we are not looking into the reasons, but we need to seek deep into racial preference in gender difference and each race.

### 4.2.1 Race & Gender

Who are more sensitive to race? Women or men? After calculating, we get this result:

If a male subject met a female partner from same race, they will say yes at 47.47% in general.

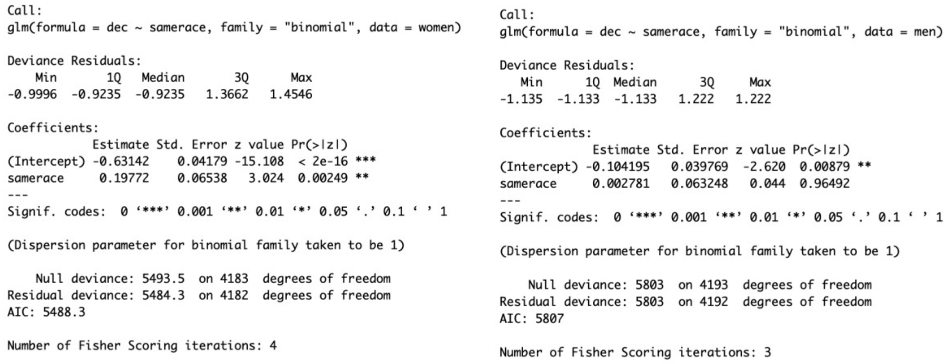
If a male subject met a female partner from different race, they will say yes at 47.39% in general.

If a female subject met a male partner from same race, they will say yes at 39.32% in general.

If a female subject met a male partner from different race, they will say yes at 34.72% in general.

It seems that male have no preference of race at all in general, while female prefer same race partner, yes rate increase about 5% from 34.72% to 39.32%.

I subset the data into women and men data set and see if same race factor matters:



```
Call:
glm(formula = dec ~ samerace, family = "binomial", data = women)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9996  -0.9235  -0.9235   1.3662   1.4546

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.63142    0.04179 -15.108  < 2e-16 ***
samerace     0.19772    0.06538   3.024  0.00249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5493.5  on 4183  degrees of freedom
Residual deviance: 5484.3  on 4182  degrees of freedom
AIC: 5488.3

Number of Fisher Scoring iterations: 4
```

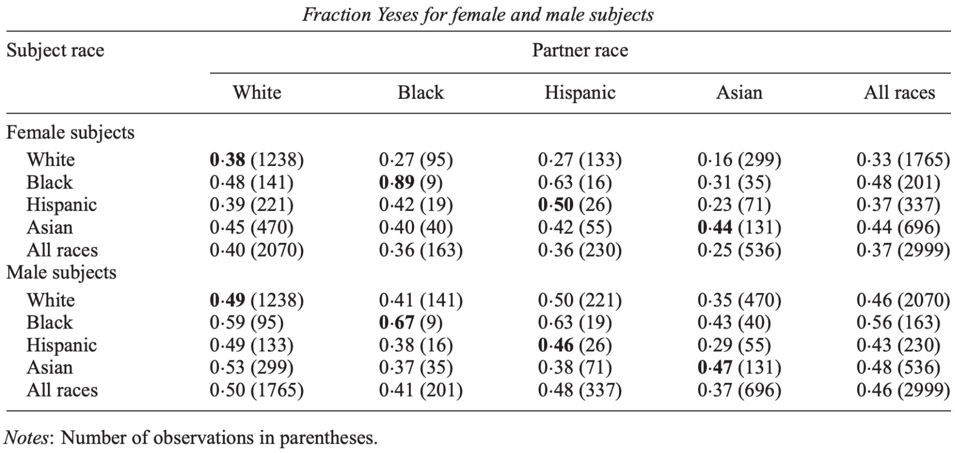
```
Call:
glm(formula = dec ~ samerace, family = "binomial", data = men)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.135  -1.133  -1.133   1.222   1.222

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.104195   0.039769  -2.620  0.00879 **
samerace     0.002781   0.063248   0.044  0.96492
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5803  on 4193  degrees of freedom
Residual deviance: 5803  on 4192  degrees of freedom
AIC: 5807

Number of Fisher Scoring iterations: 3
```

Figure 8

We can see that women have significant preference about same race while men have none racial preference at all.

### 4.2.2 Race in detail

I want to see detailed analysis about what race have preference towards what race. Here is the table from the original author’s paper from Columbia:



*Fraction Yeses for female and male subjects*

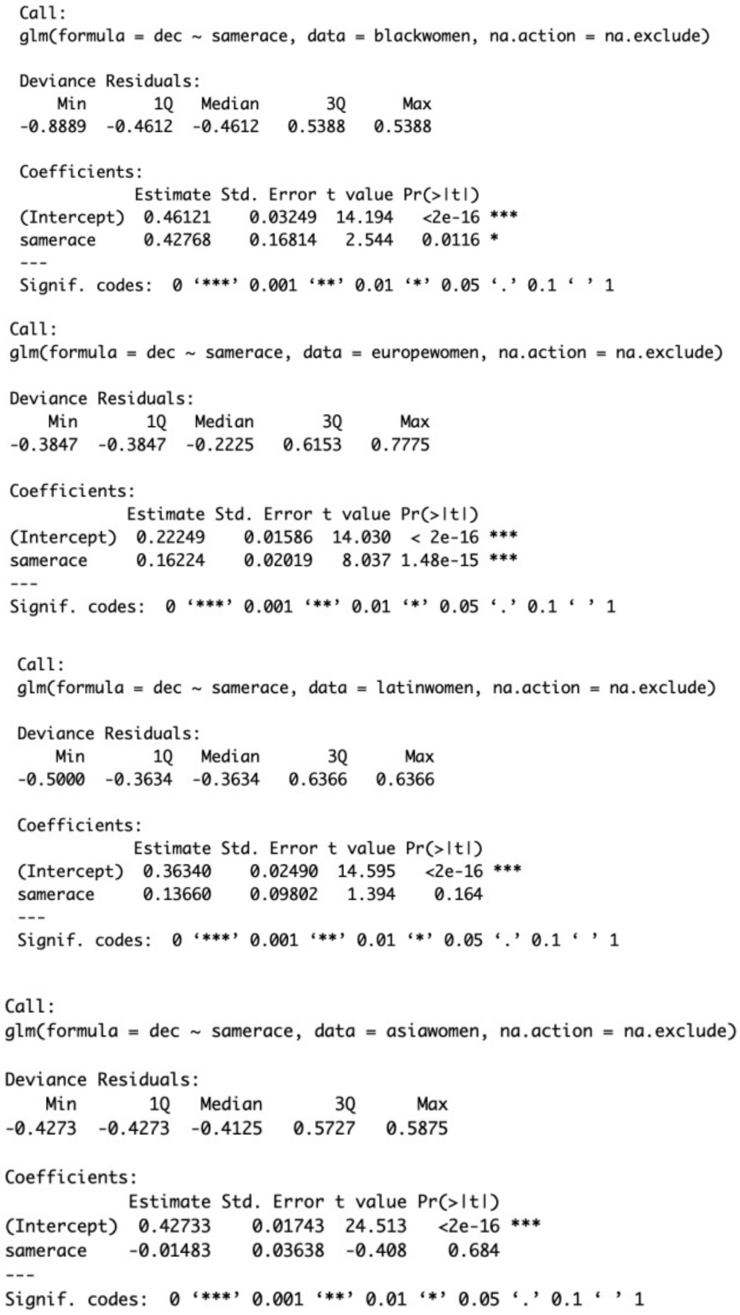
| Subject race | Partner race | | | | |
|---|---|---|---|---|---|
| | White | Black | Hispanic | Asian | All races |
| Female subjects | | | | | |
| White | **0·38** (1238) | 0·27 (95) | 0·27 (133) | 0·16 (299) | 0·33 (1765) |
| Black | 0·48 (141) | **0·89** (9) | 0·63 (16) | 0·31 (35) | 0·48 (201) |
| Hispanic | 0·39 (221) | 0·42 (19) | **0·50** (26) | 0·23 (71) | 0·37 (337) |
| Asian | 0·45 (470) | 0·40 (40) | 0·42 (55) | **0·44** (131) | 0·44 (696) |
| All races | 0·40 (2070) | 0·36 (163) | 0·36 (230) | 0·25 (536) | 0·37 (2999) |
| Male subjects | | | | | |
| White | **0·49** (1238) | 0·41 (141) | 0·50 (221) | 0·35 (470) | 0·46 (2070) |
| Black | 0·59 (95) | **0·67** (9) | 0·63 (19) | 0·43 (40) | 0·56 (163) |
| Hispanic | 0·49 (133) | 0·38 (16) | **0·46** (26) | 0·29 (55) | 0·43 (230) |
| Asian | 0·53 (299) | 0·37 (35) | 0·38 (71) | **0·47** (131) | 0·48 (536) |
| All races | 0·50 (1765) | 0·41 (201) | 0·48 (337) | 0·37 (696) | 0·46 (2999) |

*Notes*: Number of observations in parentheses.

Table 1

I will look into the significance of the samerace factor in each race and gender:

For female:



```
Call:
glm(formula = dec ~ samerace, data = blackwomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8889  -0.4612  -0.4612   0.5388   0.5388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46121    0.03249  14.194   <2e-16 ***
samerace     0.42768    0.16814   2.544   0.0116 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
glm(formula = dec ~ samerace, data = europewomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3847  -0.3847  -0.2225   0.6153   0.7775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22249    0.01586  14.030  < 2e-16 ***
samerace     0.16224    0.02019   8.037 1.48e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
glm(formula = dec ~ samerace, data = latinwomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5000  -0.3634  -0.3634   0.6366   0.6366

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.36340    0.02490  14.595   <2e-16 ***
samerace     0.13660    0.09802   1.394    0.164
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
glm(formula = dec ~ samerace, data = asiawomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4273  -0.4273  -0.4125   0.5727   0.5875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.42733    0.01743  24.513   <2e-16 ***
samerace    -0.01483    0.03638  -0.408    0.684
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 9*

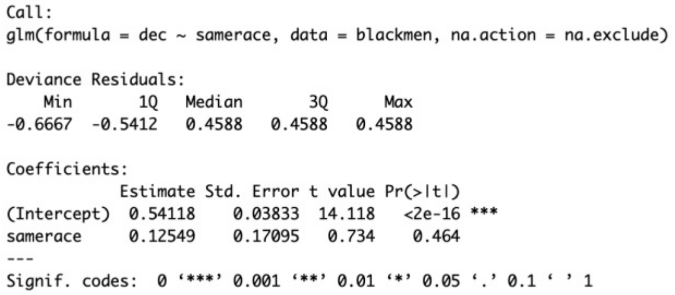
Black female: samerace is moderate significant

Europe female: samerace is highly significant,

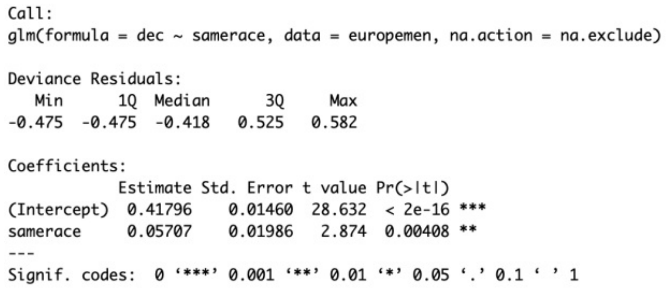
Latin female: samerace is not significant

Asian female: samerace is not significant

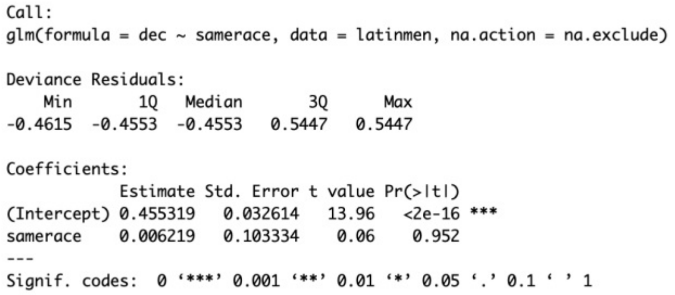
For male:



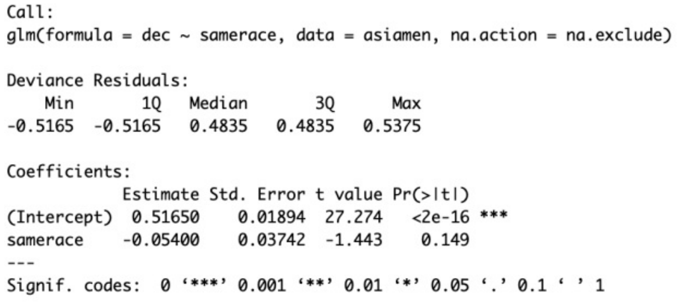
```
Call:
glm(formula = dec ~ samerace, data = blackmen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6667  -0.5412   0.4588   0.4588   0.4588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54118    0.03833  14.118   <2e-16 ***
samerace     0.12549    0.17095   0.734    0.464
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Call:
glm(formula = dec ~ samerace, data = europemen, na.action = na.exclude)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.475  -0.475  -0.418   0.525   0.582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.41796    0.01460  28.632  < 2e-16 ***
samerace     0.05707    0.01986   2.874  0.00408 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Call:
glm(formula = dec ~ samerace, data = latinmen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4615  -0.4553  -0.4553   0.5447   0.5447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.455319   0.032614   13.96   <2e-16 ***
samerace    0.006219   0.103334    0.06    0.952
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Call:
glm(formula = dec ~ samerace, data = asiamen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5165  -0.5165   0.4835   0.4835   0.5375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.51650    0.01894  27.274   <2e-16 ***
samerace    -0.05400    0.03742  -1.443    0.149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10

Black male: samerace is not significant

Europe male: samerace is significant

Latin male: samerace is not significant

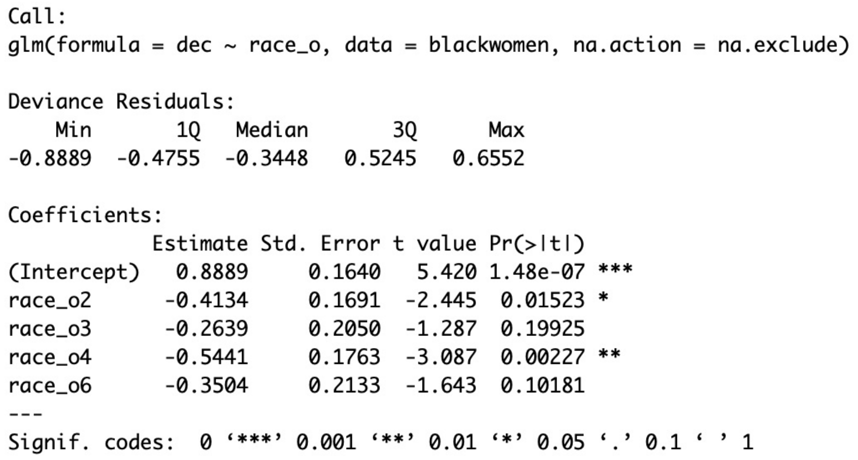
Asian male: samerace is not significant

We can see from these summaries that only black female, Europe female and Europe male have same race preference.

I want to dig deeper into what race each race prefer:

For black women:

We set black men as base model and apply logistic regression model:



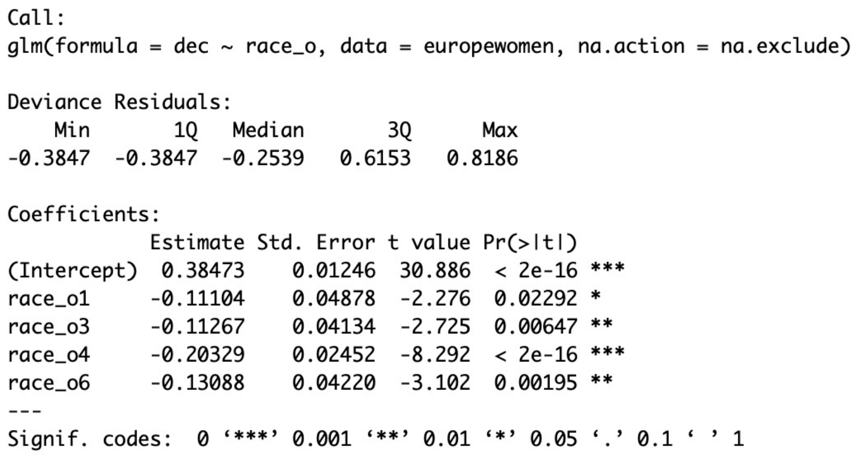
```
Call:
glm(formula = dec ~ race_o, data = blackwomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8889  -0.4755  -0.3448   0.5245   0.6552

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8889     0.1640   5.420 1.48e-07 ***
race_o2      -0.4134     0.1691  -2.445  0.01523 *
race_o3      -0.2639     0.2050  -1.287  0.19925
race_o4      -0.5441     0.1763  -3.087  0.00227 **
race_o6      -0.3504     0.2133  -1.643  0.10181
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11

Combining previous information, we can get the conclusion that black women have moderate significant preference on black men, moderate significantly dislike Europe men, no significant preference or dislike for Latin men and significantly dislike Asian men.

For Europe women:

We set Europe men as base model and apply logistic regression model:



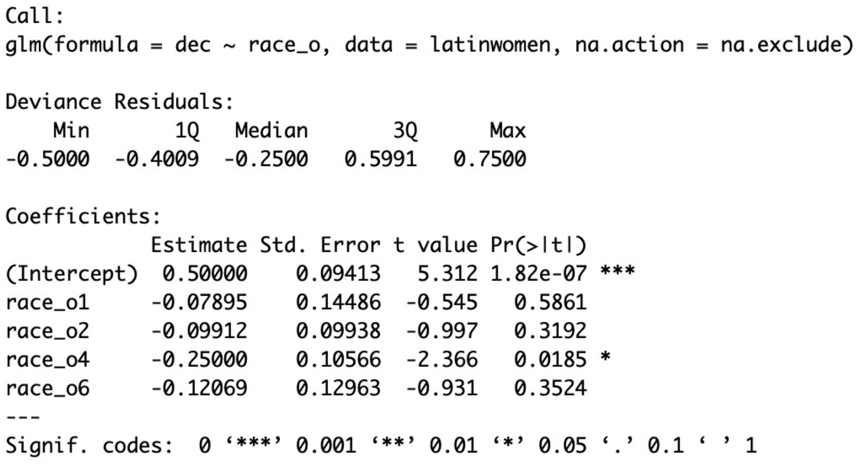
```
Call:
glm(formula = dec ~ race_o, data = europewomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3847  -0.3847  -0.2539   0.6153   0.8186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.38473    0.01246  30.886  < 2e-16 ***
race_o1     -0.11104    0.04878  -2.276  0.02292 *
race_o3     -0.11267    0.04134  -2.725  0.00647 **
race_o4     -0.20329    0.02452  -8.292  < 2e-16 ***
race_o6     -0.13088    0.04220  -3.102  0.00195 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12

Combining previous information, we can get the conclusion that Europe women have significant preference on Europe men, moderate significantly dislike black men, significant dislike for Latin men and highly significantly dislike Asian men.

For Latin women:

We set Latin men as base model and apply logistic regression model:



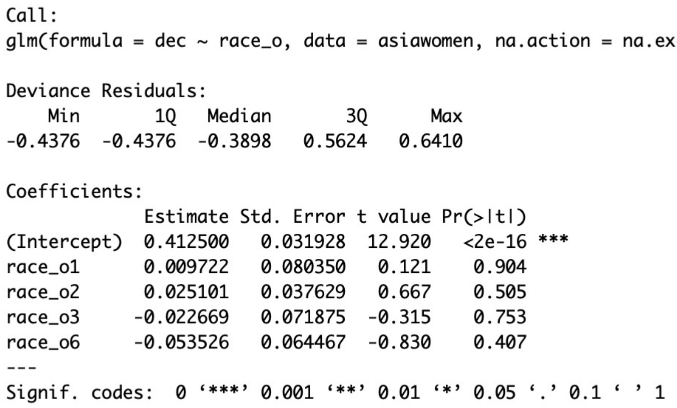
```
Call:
glm(formula = dec ~ race_o, data = latinwomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-0.5000  -0.4009  -0.2500   0.5991   0.7500  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.50000    0.09413   5.312 1.82e-07 ***
race_o1     -0.07895    0.14486  -0.545   0.5861    
race_o2     -0.09912    0.09938  -0.997   0.3192    
race_o4     -0.25000    0.10566  -2.366   0.0185 *  
race_o6     -0.12069    0.12963  -0.931   0.3524    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 13*

Combining previous information, we can get the conclusion that Latin women have no significant preference on Latin men or black men or Europe men and moderate significantly dislike Asian men.

For Asian women:

We set Asian men as base model and apply logistic regression model:



```
Call:
glm(formula = dec ~ race_o, data = asiawomen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-0.4376  -0.4376  -0.3898   0.5624   0.6410  

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.412500   0.031928  12.920   <2e-16 ***
race_o1      0.009722   0.080350   0.121    0.904    
race_o2      0.025101   0.037629   0.667    0.505    
race_o3     -0.022669   0.071875  -0.315    0.753    
race_o6     -0.053526   0.064467  -0.830    0.407    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 14*

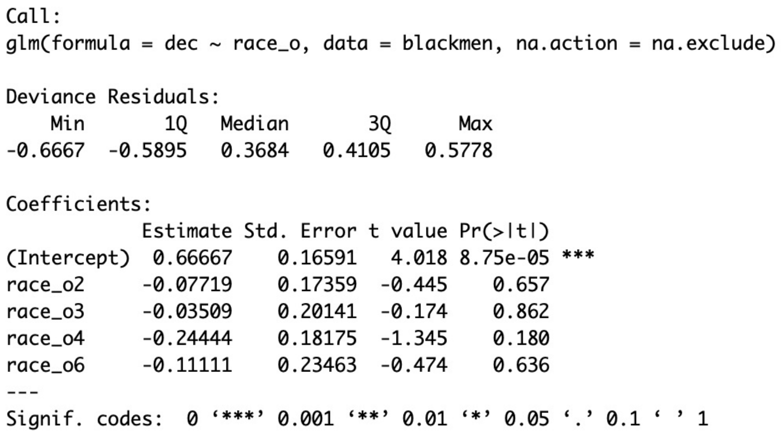
Combining previous information, we can get the conclusion that Asian women have no

preference on all races.

After the exploration for females' preference on each race, I still insist that we can only draw observational conclusions, but can not say that, for example, Europe women are most prejudice because they significantly dislike men from all other races, or Asian men are most unpopular because they are significantly disliked by female from all other races, even Asian women have no preference on them. All these kind of conclusions are not appropriate because the experiment subjects are not randomly choose at first place. All the conclusions may still be caused by the way the subjects approach the experiment designers or limitations on the population pool. So we will explore the male's preference but draw no conclusions on any deep reasons behind the phenomenon.

For black men:

We set black women as base model and apply logistic regression model:



```
Call:
glm(formula = dec ~ race_o, data = blackmen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6667  -0.5895   0.3684   0.4105   0.5778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.66667    0.16591   4.018 8.75e-05 ***
race_o2     -0.07719    0.17359  -0.445    0.657
race_o3     -0.03509    0.20141  -0.174    0.862
race_o4     -0.24444    0.18175  -1.345    0.180
race_o6     -0.11111    0.23463  -0.474    0.636
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

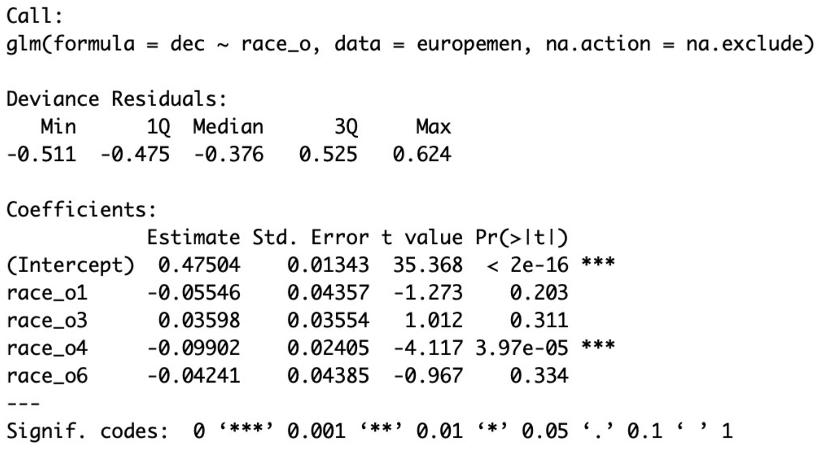
*Figure 15*

Combining previous information, we can get the conclusion that black men have no

preference on all races.

For Europe men:

We set Europe women as base model and apply logistic regression model:



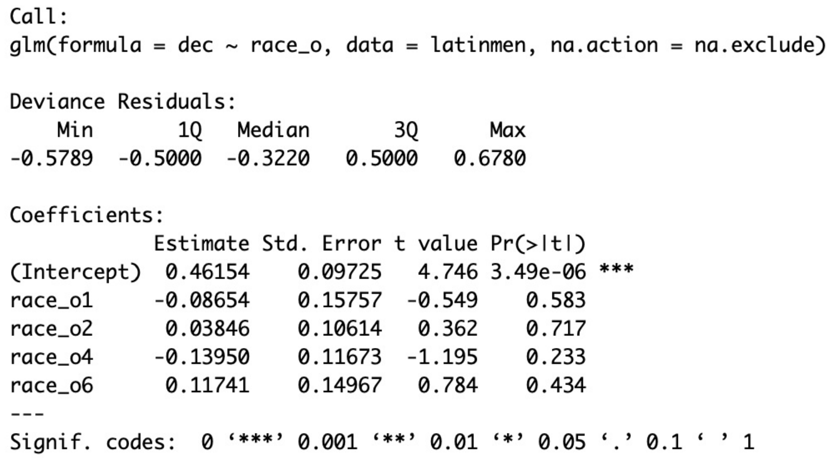
```
Call:
glm(formula = dec ~ race_o, data = europemen, na.action = na.exclude)

Deviance Residuals:
   Min      1Q  Median      3Q     Max  
-0.511  -0.475  -0.376   0.525   0.624  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.47504    0.01343  35.368  < 2e-16 ***
race_o1     -0.05546    0.04357  -1.273    0.203    
race_o3      0.03598    0.03554   1.012    0.311    
race_o4     -0.09902    0.02405  -4.117 3.97e-05 ***
race_o6     -0.04241    0.04385  -0.967    0.334    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 16*

Combining previous information, we can get the conclusion that Europe men have significant preference on Europe women, significantly dislike Asian women and have no preference on black or Latin women.

For Latin men:

We set Latin women as base model and apply logistic regression model:



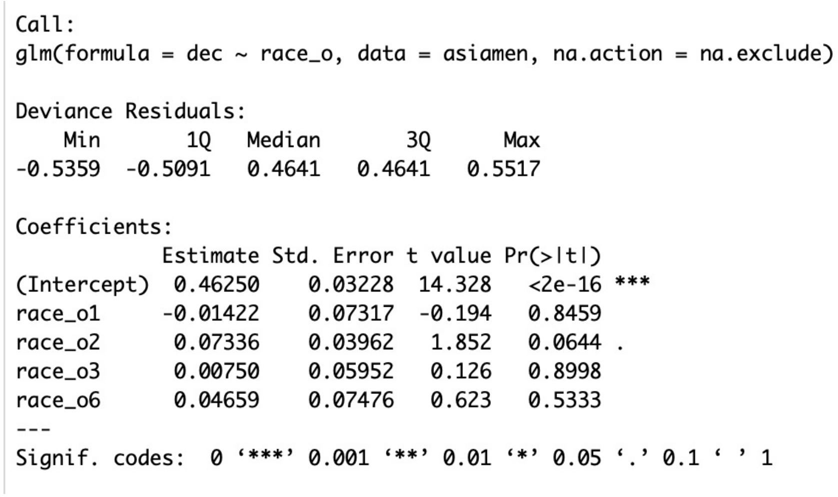
```
Call:
glm(formula = dec ~ race_o, data = latinmen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-0.5789  -0.5000  -0.3220   0.5000   0.6780  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.46154    0.09725   4.746 3.49e-06 ***
race_o1     -0.08654    0.15757  -0.549    0.583    
race_o2      0.03846    0.10614   0.362    0.717    
race_o4     -0.13950    0.11673  -1.195    0.233    
race_o6      0.11741    0.14967   0.784    0.434    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 17*

Combining previous information, we can get the conclusion that Latin men have no preference on all races.

For Asian men:

We set Asian women as base model and apply logistic regression model:



```
Call:
glm(formula = dec ~ race_o, data = asiamen, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-0.5359  -0.5091   0.4641   0.4641   0.5517  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.46250    0.03228  14.328   <2e-16 ***
race_o1     -0.01422    0.07317  -0.194   0.8459    
race_o2      0.07336    0.03962   1.852   0.0644 .  
race_o3      0.00750    0.05952   0.126   0.8998    
race_o6      0.04659    0.07476   0.623   0.5333    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 18*

Combining previous information, we can get the conclusion that Asian men have no

preference on all races.

## 4.3 Age

From 2018 U.S. Census, the average marriage age is 27.4. Do you feel the peer pressure or pressure from your parents when you approach 27.4 that you need to be married? Does age affect people's speed dating decisions?

Let us look at the age variable.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   24.00   26.00   26.36   28.00   55.00 
```

Median is 26 and average is 26.36, quite close but younger than 27.4, which is the exact population who need speed dating.

Let us look at the distribution of male and female:

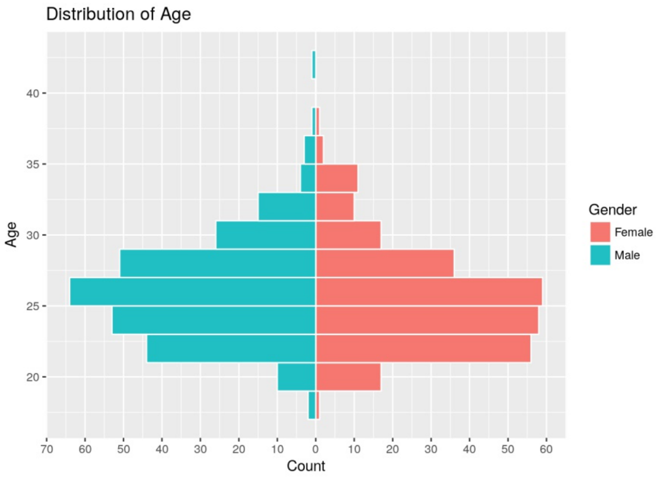
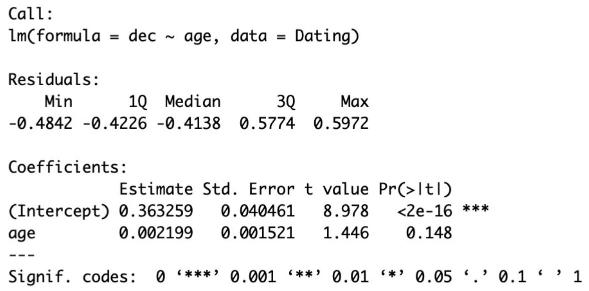


*Figure 19*

We can see the distribution are almost same for male and female, so hopefully each one will find a date.

Let us see if age matters the decision of male and female.

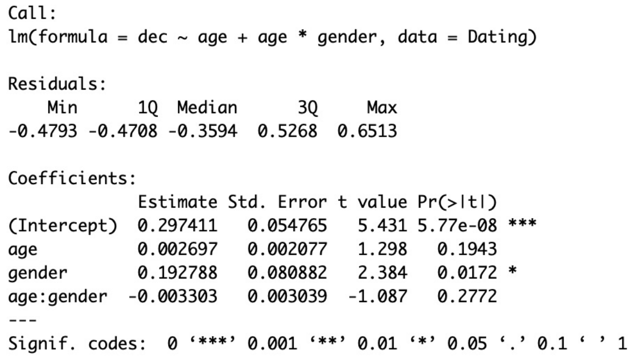


```
Call:
lm(formula = dec ~ age, data = Dating)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4842 -0.4226 -0.4138  0.5774  0.5972 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.363259   0.040461   8.978   <2e-16 ***
age         0.002199   0.001521   1.446    0.148    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 20*

It seems that age is not a influential factor to the final decision for both gender generally.

We will look into the interaction of gender and age to see if male and female react differently to the age.

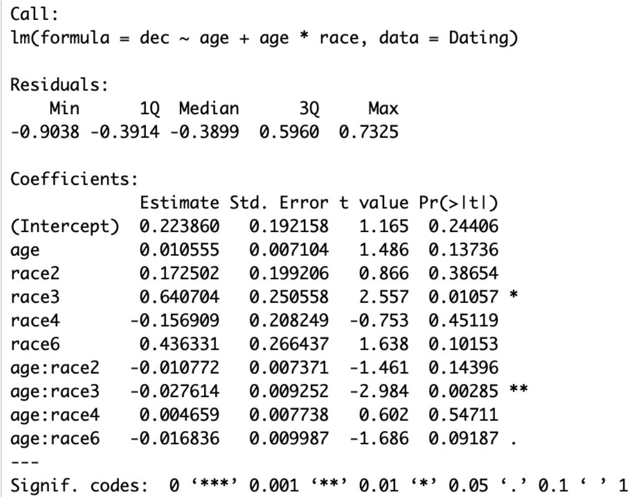


```
Call:
lm(formula = dec ~ age + age * gender, data = Dating)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4793 -0.4708 -0.3594  0.5268  0.6513 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.297411   0.054765   5.431 5.77e-08 ***
age          0.002697   0.002077   1.298   0.1943    
gender       0.192788   0.080882   2.384   0.0172 *  
age:gender  -0.003303   0.003039  -1.087   0.2772    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 21*

So both female and male have no preference on age in general.

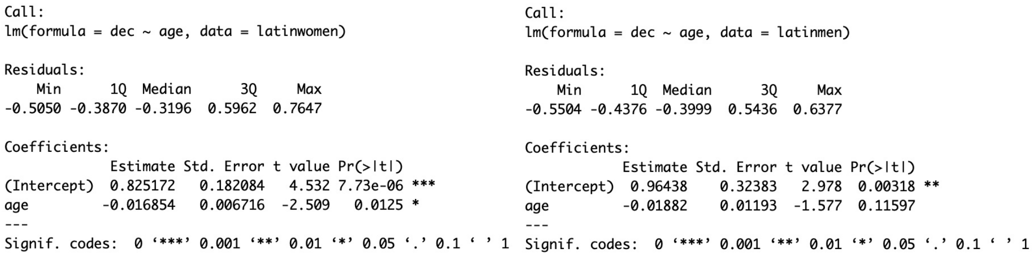
We will look into the interaction of race and age to see if different race react differently to the age.



```
Call:
lm(formula = dec ~ age + age * race, data = Dating)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9038 -0.3914 -0.3899  0.5960  0.7325 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.223860   0.192158   1.165  0.24406   
age          0.010555   0.007104   1.486  0.13736   
race2        0.172502   0.199206   0.866  0.38654   
race3        0.640704   0.250558   2.557  0.01057 * 
race4       -0.156909   0.208249  -0.753  0.45119   
race6        0.436331   0.266437   1.638  0.10153   
age:race2   -0.010772   0.007371  -1.461  0.14396   
age:race3   -0.027614   0.009252  -2.984  0.00285 **
age:race4    0.004659   0.007738   0.602  0.54711   
age:race6   -0.016836   0.009987  -1.686  0.09187 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 22*

So only age and Latin interaction is significant which means Latin have preference on age. Let us dig into it. Does Latin female or male have preference on age?



```
Call:
lm(formula = dec ~ age, data = latinwomen)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5050 -0.3870 -0.3196  0.5962  0.7647 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.825172   0.182084   4.532 7.73e-06 ***
age         -0.016854   0.006716  -2.509   0.0125 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

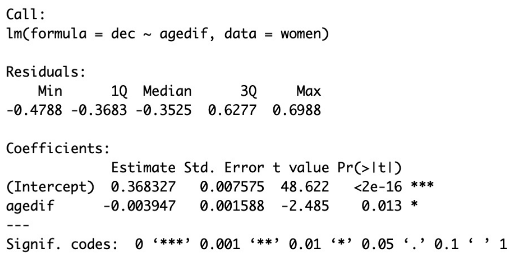
```
Call:
lm(formula = dec ~ age, data = latinmen)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5504 -0.4376 -0.3999  0.5436  0.6377 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.96438    0.32383   2.978  0.00318 **
age         -0.01882    0.01193  -1.577  0.11597   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 23*

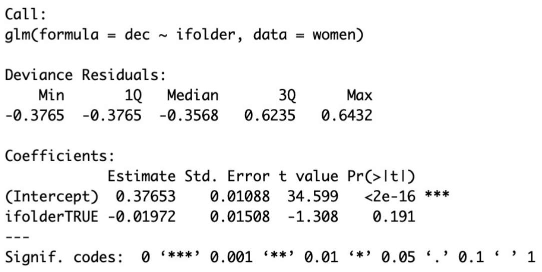
So we get the conclusion that only Latin female have preference on age. They prefer young male.

However, from the 2018 U.S. Census we know that male are generally older than female when they get married, so does the age difference matters, or does it necessary that male prefer younger female or female prefer older male?

Thus we create 2 new variable age-difference(age of male - age of female) and if-older(if male is older than female) for each pair and see are these two factors significant.

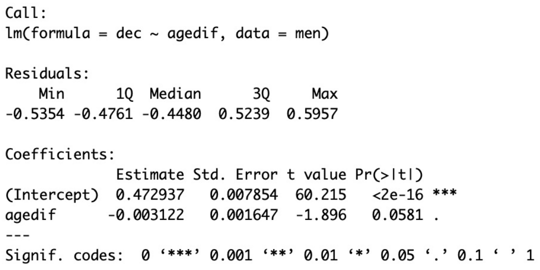


```
Call:
lm(formula = dec ~ agedif, data = women)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4788 -0.3683 -0.3525  0.6277  0.6988 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.368327   0.007575  48.622   <2e-16 ***
agedif      -0.003947   0.001588  -2.485    0.013 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

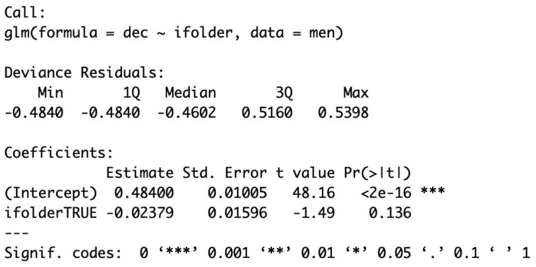


```
Call:
glm(formula = dec ~ ifolder, data = women)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.3765  -0.3765  -0.3568   0.6235   0.6432  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.37653    0.01088  34.599   <2e-16 ***
ifolderTRUE -0.01972    0.01508  -1.308    0.191    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 24*



```
Call:
lm(formula = dec ~ agedif, data = men)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5354 -0.4761 -0.4480  0.5239  0.5957 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.472937   0.007854  60.215   <2e-16 ***
agedif      -0.003122   0.001647  -1.896   0.0581 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Call:
glm(formula = dec ~ ifolder, data = men)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.4840  -0.4840  -0.4602   0.5160   0.5398  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.48400    0.01005   48.16   <2e-16 ***
ifolderTRUE -0.02379    0.01596   -1.49    0.136    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 25*

So we can see that only female care about the age difference. They prefer male partner to be close to their own age, not too younger or older. However, both female and male does not care if their partner is younger or older than themselves.

In general age is not such a big influential variable so we will not dig into much deeper.

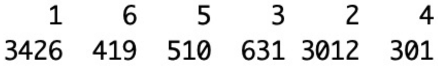
## 4.4 Goal

We always say that attitude matters in our life. Does attitude matters in speed dating?

Let us study the variable goal which represents subject’s primary goal in participating in this speed dating event. Each score (1-6) represent different goal for the event.

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

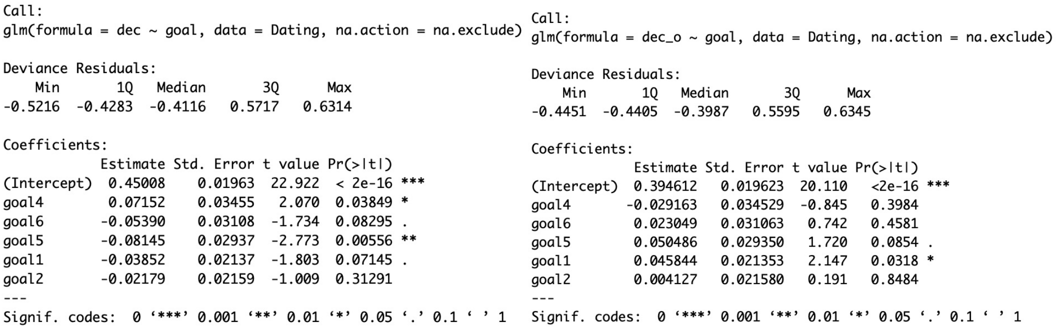
Let us see the summary of the goal.



```
   1    6    5    3    2    4 
3426  419  510  631 3012  301 
```

So most people have the goal of 1 and 2, which is seemed like a fun night out and to meet new people. So most people don't seem to be desperate to have a date or want to be married.

Since the event is speed dating event, so we set goal 3 - to get a date as base model and to see if goal matters in decision making for speed dating.



```
Call:
glm(formula = dec ~ goal, data = Dating, na.action = na.exclude)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.5216  -0.4283  -0.4116   0.5717   0.6314  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.45008    0.01963  22.922  < 2e-16 ***
goal4        0.07152    0.03455   2.070  0.03849 *  
goal6       -0.05390    0.03108  -1.734  0.08295 .  
goal5       -0.08145    0.02937  -2.773  0.00556 ** 
goal1       -0.03852    0.02137  -1.803  0.07145 .  
goal2       -0.02179    0.02159  -1.009  0.31291    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

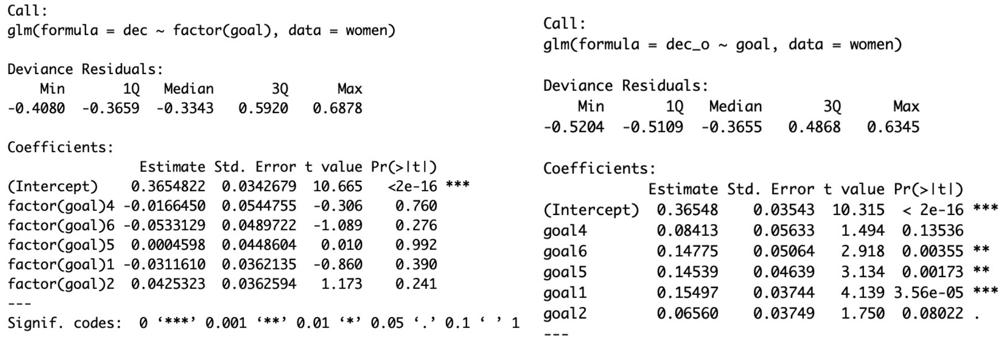
```
Call:
glm(formula = dec_o ~ goal, data = Dating, na.action = na.exclude)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.4451  -0.4405  -0.3987   0.5595   0.6345  

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.394612   0.019623  20.110   <2e-16 ***
goal4       -0.029163   0.034529  -0.845   0.3984    
goal6        0.023049   0.031063   0.742   0.4581    
goal5        0.050486   0.029350   1.720   0.0854 .  
goal1        0.045844   0.021353   2.147   0.0318 *  
goal2        0.004127   0.021580   0.191   0.8484    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 26*

So generally, if subject has goal 4 which is looking for a serious relationship, he or she will more likely to say yes to the partner compare to people just want a date, and subjects who just want a fun night or meet new people or say I did it will turn down more people, which is perfectly justified the result. And for the partner, subject will moderate significantly more likely to say yes for partner who has the goal of just want a fun night. The reason behind it is I think maybe people choose to have a fun night are more easy going and less nervous during the event.

We also want to see different result on both female and male.

For female:



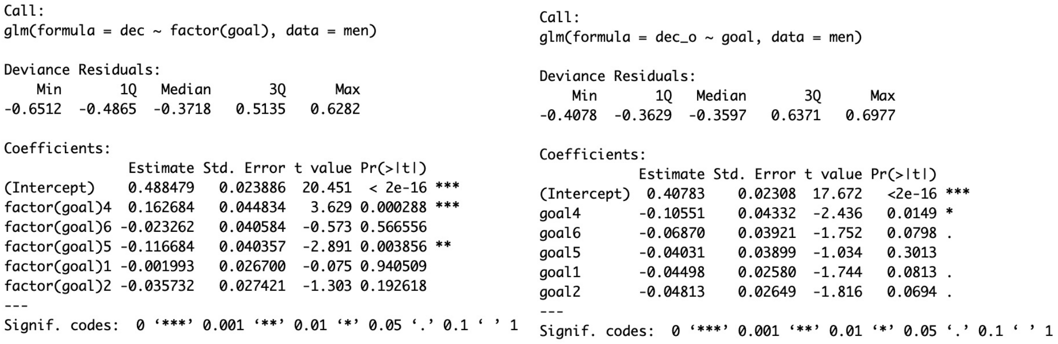
```
Call:
glm(formula = dec ~ factor(goal), data = women)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4080  -0.3659  -0.3343   0.5920   0.6878

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.3654822  0.0342679  10.665   <2e-16 ***
factor(goal)4 -0.0166450  0.0544755  -0.306    0.760
factor(goal)6 -0.0533129  0.0489722  -1.089    0.276
factor(goal)5  0.0004598  0.0448604   0.010    0.992
factor(goal)1 -0.0311610  0.0362135  -0.860    0.390
factor(goal)2  0.0425323  0.0362594   1.173    0.241
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
glm(formula = dec_o ~ goal, data = women)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5204  -0.5109  -0.3655   0.4868   0.6345

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.36548    0.03543  10.315  < 2e-16 ***
goal4        0.08413    0.05633   1.494  0.13536
goal6        0.14775    0.05064   2.918  0.00355 **
goal5        0.14539    0.04639   3.134  0.00173 **
goal1        0.15497    0.03744   4.139 3.56e-05 ***
goal2        0.06560    0.03749   1.750  0.08022 .
---
```

*Figure 27*

Female perform almost same no matter what goal they have for the event. Male highly significantly prefer female who just want a fun night or meet new people or say I did it. So male generally don’t like serious female.

For male:



```
Call:
glm(formula = dec ~ factor(goal), data = men)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6512  -0.4865  -0.3718   0.5135   0.6282

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.488479   0.023886  20.451  < 2e-16 ***
factor(goal)4  0.162684   0.044834   3.629 0.000288 ***
factor(goal)6 -0.023262   0.040584  -0.573 0.566556
factor(goal)5 -0.116684   0.040357  -2.891 0.003856 **
factor(goal)1 -0.001993   0.026700  -0.075 0.940509
factor(goal)2 -0.035732   0.027421  -1.303 0.192618
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
glm(formula = dec_o ~ goal, data = men)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4078  -0.3629  -0.3597   0.6371   0.6977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40783    0.02308  17.672   <2e-16 ***
goal4       -0.10551    0.04332  -2.436   0.0149 *
goal6       -0.06870    0.03921  -1.752   0.0798 .
goal5       -0.04031    0.03899  -1.034   0.3013
goal1       -0.04498    0.02580  -1.744   0.0813 .
goal2       -0.04813    0.02649  -1.816   0.0694 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 28*

Male will highly significantly say more yes if they have the goal for serious relationship and significantly more selective if they just want to say they did speed dating. And also, female don’t like serious men.

## 4.5 Conclusion

1. Female tend to be more selective than male.

2. Female appear to score higher in attraction.

3. Female and male go out in similar rate.

4. Only black women, Europe women and Europe men appear to have same race preference.

5. Black women appear to have moderate significant preference on black men, moderate significantly dislike Europe men, no significant preference or dislike for Latin men and significantly dislike Asian men.

6. Europe women tend to have significant preference on Europe men, moderate significantly dislike black men, significant dislike for Latin men and highly significantly dislike Asian men.

7. Latin women appear to have no significant preference on Latin men or black men or Europe men and moderate significantly dislike Asian men.

8. Asian women appear to have no preference on all races.

9. Black men, Latin men and Asian men appear to have no preference on all races.

10. Europe men tend to have significant preference on Europe women, significantly dislike Asian women and have no preference on black or Latin women.

11. Both female and male have no preference on age in general. Only Latin female appear to have preference on age. They prefer young male.

12. Female tend to prefer male partner to be close to their own age, not too younger or older. However, both female and male does not care if their partner is younger or older than themselves.

13. Female perform almost same no matter what goal they have for the event. Male appear to highly significantly prefer female who just want a fun night or meet new people or say I did it. So most likely male generally don’t like serious female.

14. Male highly significantly appear to say more yes if they have the goal for serious relationship and significantly more selective if they just want to say they did speed dating. And also, female most likely don’t like serious men.

# CHAPTER 5

## 5 Models

I will predict people's decisions using logistic regression as the base model, and gradually improve the result by decision tree model, random forest model and XGBoost model.

### 5.1 Predict result after speed dating

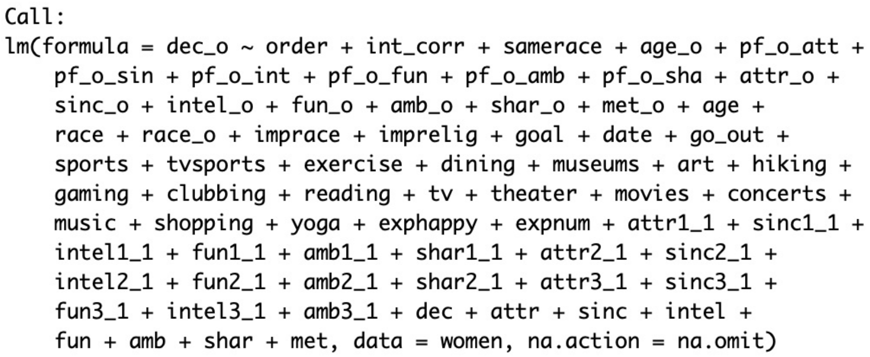
Before we build models, we need to consider what question we want to answer and what data we want to use. So first we want to predict people's decision based on all variables we have, including basic personal information or information collected from consensus before and during the speed dating event.

I choose data set women to build models and dec_o as y variable which is the variable we want to predict. I will also divide the data set into 2 parts, 80% of training data set and 20% of testing data set. The reasons are that as we mention in chapter 4 that in general women say yes at 37% of the time and say no at 63% of the time while men say yes at 47% of the time and say no at 53% of the time, so dec_o which is the decision of men to the women will have distribution of 47% of 1(yes) and 53% of 0(no). So if even the model fails and random choose 1 or 0, the prediction power should be almost same. So if the model has prediction power close to 50%, means the model predicts nothing, because even we random choose 0 or 1 for all men, we still get a model which has prediction power of 50%. We will consider the model is useful if the model has prediction power above 50%.

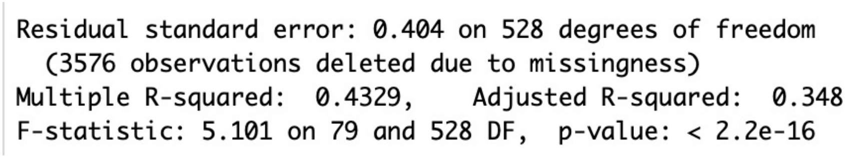
### 5.1.1 Model 1 - Logistic regression

Logistic regression, also known as logit regression, is the most commonly used method to predict a event existing only as two states, like the decision in our data set. We will use logistic function to have a function, and the decision of either 0 (means no) and 1(means yes) based on all other variables, such as gender, race, attraction, intelligence, fun and so on. Each independent variable will have a parameter to determine how important the variable is in the function to affect the probability of the dependent variable - the partner's decision. We will use the logistic regression model as base model for following reasons. First, logistic regression came out around 1830s and been used widely. It is easy to build the model. Secondly, it is the only model we can interpret the results, comparing to other machine learning methods which can not been understand its logic. So we will use logistic regression model as base model.

Now, let us put all variables into the logistic regression model to see its summary:



```
Call:
lm(formula = dec_o ~ order + int_corr + samerace + age_o + pf_o_att +
    pf_o_sin + pf_o_int + pf_o_fun + pf_o_amb + pf_o_sha + attr_o +
    sinc_o + intel_o + fun_o + amb_o + shar_o + met_o + age +
    race + race_o + imprace + imprelig + goal + date + go_out +
    sports + tvsports + exercise + dining + museums + art + hiking +
    gaming + clubbing + reading + tv + theater + movies + concerts +
    music + shopping + yoga + exphappy + expnum + attr1_1 + sinc1_1 +
    intel1_1 + fun1_1 + amb1_1 + shar1_1 + attr2_1 + sinc2_1 +
    intel2_1 + fun2_1 + amb2_1 + shar2_1 + attr3_1 + sinc3_1 +
    fun3_1 + intel3_1 + amb3_1 + dec + attr + sinc + intel +
    fun + amb + shar + met, data = women, na.action = na.omit)
```

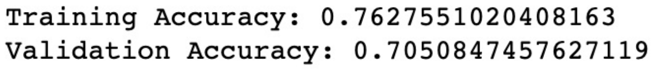


```
Residual standard error: 0.404 on 528 degrees of freedom
  (3576 observations deleted due to missingness)
Multiple R-squared:  0.4329,    Adjusted R-squared:  0.348
F-statistic: 5.101 on 79 and 528 DF,  p-value: < 2.2e-16
```

Since model's p-value is very low, we believe the model is useful. However, the multiple

R-square is only 43.3%, which means only less than half the data can be explained by the model. The adjusted R-square is only 34.8% which is much lower than the multiple R-square, which means we have too many variables which causes over-fitting . There is much more improvement of the model should be done.

We will also check its prediction power:



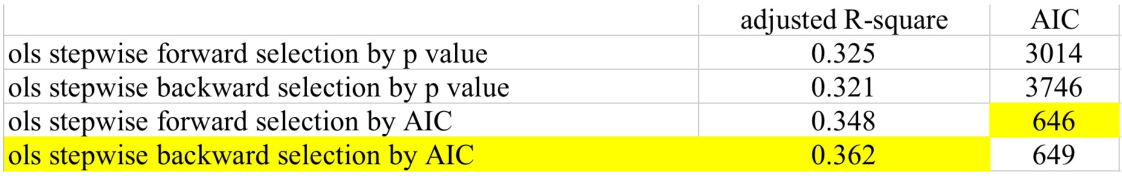
```
Training Accuracy: 0.7627551020408163
Validation Accuracy: 0.7050847457627119
```

We get 76.3% accuracy for 80% of the training data set and 70.5% accuracy for testing data set, which means if we conduct another round of speed dating event, we should be able to predict male's decision correctly after the speed dating event for 70.5% of the time. So the model is reasonable useful. But let us improving the model by variable selection first.

The reason to use variable selection is because there may be correlation between multiple variables, which means two or more variables are too similar in the model and to cause over-fitting. Which means the model perform well only in the training data-set but perform poorly in the testing data-set.

There are two general types of variable selection method: forward selection and backward selection. Forward selection will start from no variables in the model, and add one variable at a time, to improve certain criterion of the model. It will add the most important and significant variable first, until no more variable is significant to the function and improve the certain criterion of the model. The backward selection method is opposite. Backward selection will start from full model with all variables in the model, and delete one variable at a time, to improve certain criterion of the model. It will delete the most unimportant and insignificant variable first, until every variable is significant to the function and delete any of the variable will decrease the certain criterion of the model. AIC and p-value is the two criterion we will use to evaluate the performance of the variables and thus to choose variables. We will use AIC and adjusted R-square to evaluate the performance of the model.

So there are total four methods for variable selection: ols stepwise forward selection by p value, ols stepwise backward selection by p value, ols stepwise forward selection by AIC, ols stepwise backward selection by AIC.

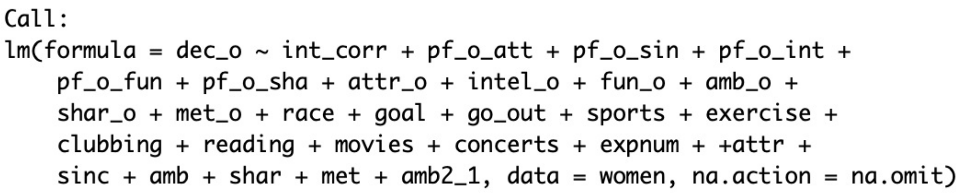


| | adjusted R-square | AIC |
|---|---|---|
| ols stepwise forward selection by p value | 0.325 | 3014 |
| ols stepwise backward selection by p value | 0.321 | 3746 |
| ols stepwise forward selection by AIC | 0.348 | 646 |
| ols stepwise backward selection by AIC | 0.362 | 649 |

*Table 2*

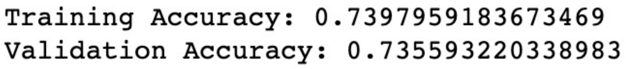
We can see that ols stepwise forward selection by AIC has lowest AIC and ols stepwise backward selection by AIC has highest adjusted R-square. I will choose one of these method. I choose ols stepwise backward selection by AIC because it is the only method improve adjusted R-square compare to 0.348 for the full model and its AIC is also close to the lowest. I will try explain this method here.

Here is the final model for logistic regression model after variable selection:



```
Call:
lm(formula = dec_o ~ int_corr + pf_o_att + pf_o_sin + pf_o_int +
    pf_o_fun + pf_o_sha + attr_o + intel_o + fun_o + amb_o +
    shar_o + met_o + race + goal + go_out + sports + exercise +
    clubbing + reading + movies + concerts + expnum + +attr +
    sinc + amb + shar + met + amb2_1, data = women, na.action = na.omit)
```

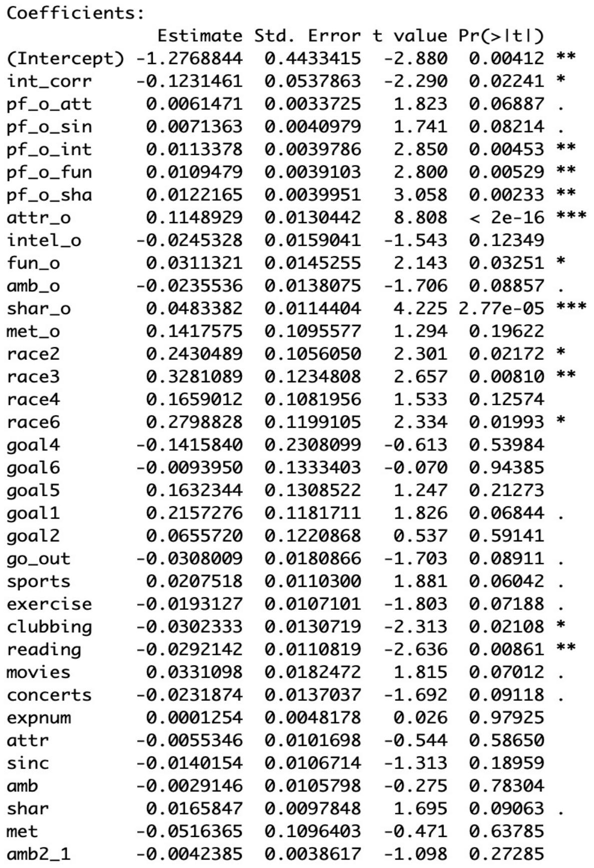
The final model has only 35 variables compare to 79 variables for the original model and has even higher adjusted R-square. Which means these 35 variables are capable of predicting final result - decision, as much as all 79 variables, even in a greater extent. We will check its prediction power:



```
Training Accuracy: 0.7397959183673469
Validation Accuracy: 0.735593220338983
```

We get 74.0% accuracy for 80% of the training data set and 73.6% accuracy for testing data set, which shows the model is improved by decreasing accuracy of training data and over-fitting to achieve a better performance for testing data set. This means if we conduct another round of speed dating event, we should be able to predict male's decision correctly after the speed dating event for 73.6% of the time. We can see the accuracy for testing data-set is almost same as the accuracy for training data-set, which means we eliminate variable over-fitting to the largest extent.

We can also exam the importance of each variables in this final model. We can see what makes a man say yes or no to the female:



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.2768844  0.4433415  -2.880  0.00412 ** 
int_corr    -0.1231461  0.0537863  -2.290  0.02241 *  
pf_o_att     0.0061471  0.0033725   1.823  0.06887 .  
pf_o_sin     0.0071363  0.0040979   1.741  0.08214 .  
pf_o_int     0.0113378  0.0039786   2.850  0.00453 ** 
pf_o_fun     0.0109479  0.0039103   2.800  0.00529 ** 
pf_o_sha     0.0122165  0.0039951   3.058  0.00233 ** 
attr_o       0.1148929  0.0130442   8.808  < 2e-16 ***
intel_o     -0.0245328  0.0159041  -1.543  0.12349    
fun_o        0.0311321  0.0145255   2.143  0.03251 *  
amb_o       -0.0235536  0.0138075  -1.706  0.08857 .  
shar_o       0.0483382  0.0114404   4.225 2.77e-05 ***
met_o        0.1417575  0.1095577   1.294  0.19622    
race2        0.2430489  0.1056050   2.301  0.02172 *  
race3        0.3281089  0.1234808   2.657  0.00810 ** 
race4        0.1659012  0.1081956   1.533  0.12574    
race6        0.2798828  0.1199105   2.334  0.01993 *  
goal4       -0.1415840  0.2308099  -0.613  0.53984    
goal6       -0.0093950  0.1333403  -0.070  0.94385    
goal5        0.1632344  0.1308522   1.247  0.21273    
goal1        0.2157276  0.1181711   1.826  0.06844 .  
goal2        0.0655720  0.1220868   0.537  0.59141    
go_out      -0.0308009  0.0180866  -1.703  0.08911 .  
sports       0.0207518  0.0110300   1.881  0.06042 .  
exercise    -0.0193127  0.0107101  -1.803  0.07188 .  
clubbing    -0.0302333  0.0130719  -2.313  0.02108 *  
reading     -0.0292142  0.0110819  -2.636  0.00861 ** 
movies       0.0331098  0.0182472   1.815  0.07012 .  
concerts    -0.0231874  0.0137037  -1.692  0.09118 .  
expnum       0.0001254  0.0048178   0.026  0.97925    
attr        -0.0055346  0.0101698  -0.544  0.58650    
sinc        -0.0140154  0.0106714  -1.313  0.18959    
amb         -0.0029146  0.0105798  -0.275  0.78304    
shar         0.0165847  0.0097848   1.695  0.09063 .  
met         -0.0516365  0.1096403  -0.471  0.63785    
amb2_1      -0.0042385  0.0038617  -1.098  0.27285    
```

*Figure 29*

We can see that attraction is the most important thing a men is looking. If a women score one point higher for the attraction, she has 11.5% more chance to have a second date with the men. The second most important thing a men looking for is shared interest. If a women score one point higher for the shared interest, she has 4.8% more chance to have a second date with the men.

### 5.1.2 Model 2 - Decision tree

We will use decision tree as the compared model. Decision tree is a tree-like model and each node represent a criterion. If the person fulfill the criterion, male will go the left to the next level, otherwise will go to the right side of the tree. The decision tree model will also give us the most significant variables. It is also a very clear and explainable model we can use. Here is the result:

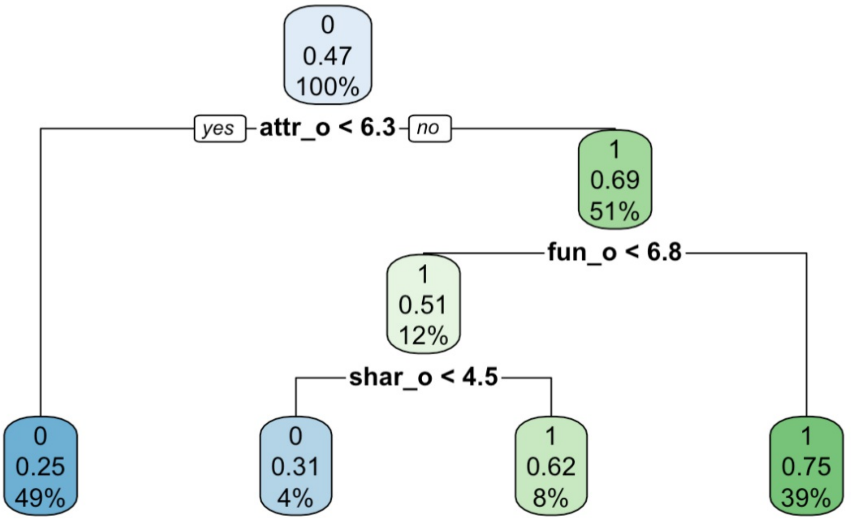


*Figure 30*

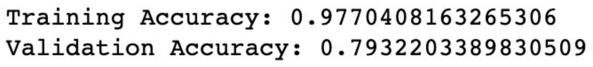
The model indicates that if a women score less than 6.3 for attraction, she has 25% of chance to be rejected by men, and if a women score more than 6.3 for attraction and more than 6.8 for fun, she has 75% of chance that the male partner want to have a second date with she. We can see that attraction, fun and shared interest are 3 most significant variables in decision tree model to predict male's decision, which also correspond to the logistic regression model to have similar result. The prediction power of decision tree model is 73.4%, which is very close to logistic regression model.

### 5.1.3 Model 3 - Random forest

Random forest is an ensemble learning method for regression by constructing multitude decision trees to avoid over fitting for individual decision tree. It takes random observations and random variables to form decision trees, and the maximum votes will give us a better prediction. It is very useful in predicting results, but it sacrifice the explainability as we can not explain the exact algorithms behind the model. It is also a much more latest algorithm which was first published on 1995.

The parameter we need to choose is the number and maximum depth of trees for the model. As number and maximum depth of trees increased, we will sacrifice computing speed and may cause over-fitting. But too few trees will also not reach its maximum prediction power.

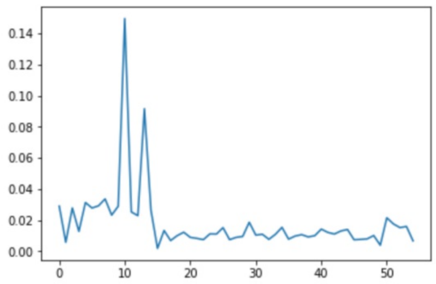
So the number of trees we will try is from 10 to 1000 and maximum depth of trees is from 3 to 10. When we choose 410 trees and maximum depth of trees as 9, the model reach its best prediction power. Now let us see the result for random forest model:

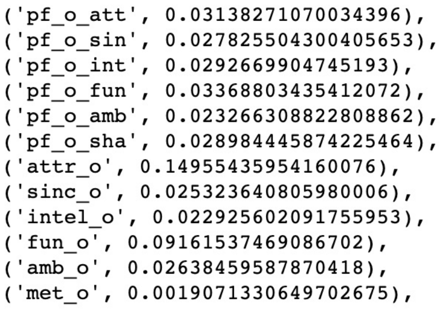


```
Training Accuracy: 0.9770408163265306
Validation Accuracy: 0.7932203389830509
```

We get 97.7% accuracy for 80% of the training data set and 79.3% accuracy for testing data set, which means if we conduct another round of speed dating event, we should be able to predict male’s decision correctly after the speed dating event for 79.3% of the time. So random forest model improved 6% accuracy compare to logistic regression models.

We can also have a view of what variables contributing most to the model:





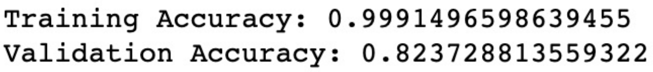
```
('pf_o_att', 0.03138271070034396),
('pf_o_sin', 0.027825504300405653),
('pf_o_int', 0.0292669904745193),
('pf_o_fun', 0.03368803435412072),
('pf_o_amb', 0.023266308822808862),
('pf_o_sha', 0.028984445874225464),
('attr_o', 0.14955435954160076),
('sinc_o', 0.025323640805980006),
('intel_o', 0.022925602091755953),
('fun_o', 0.09161537469086702),
('amb_o', 0.02638459587870418),
('met_o', 0.0019071330649702675),
```

*Figure 31*

The highest two points are attraction of the female and fun of the female. Attraction of the female contributes 15% to the model and fun of female contributes 9.2% to the model.

### 5.1.4 Model 3 - XGBoost

XGBoost is a open-source deep machine learning method. It also sacrificed the explainability of the algorithm for the prediction power. The most crucial parameters are maximum depth of the tree(normally from 3 to 10), the number of trees(normally from 10 to 1000), and learning rate (normally from 0.01 to 0.3). The final model we have for the parameters are: number of trees is 450, max depth is 3 and learning rate is 0.1, which reaches its best prediction power. Now let us see the result for XGBoost model:

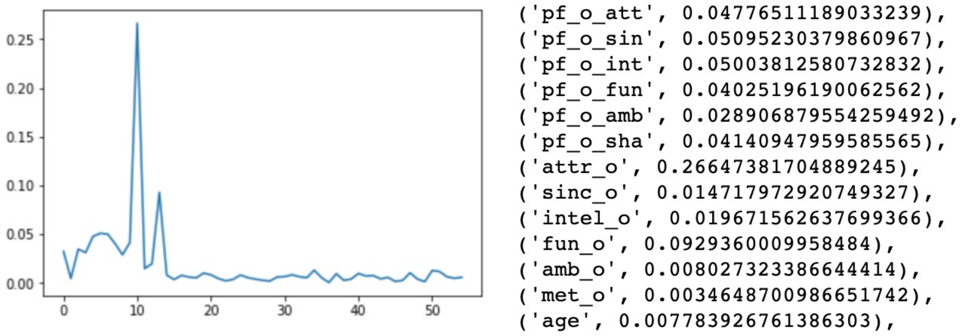


```
Training Accuracy: 0.9991496598639455
Validation Accuracy: 0.823728813559322
```

We get 99.9% accuracy for 80% of the training data set and 82.4% accuracy for testing data set, which means if we conduct another round of speed dating event, we should be able to predict male’s decision correctly after the speed dating event for 82.4% of the time.

XGBoost model improved 9% accuracy compare to logistic regression models.

We can also have a view of what variables contributing most to the model:



```
('pf_o_att', 0.04776511189033239),
('pf_o_sin', 0.05095230379860967),
('pf_o_int', 0.05003812580732832),
('pf_o_fun', 0.04025196190062562),
('pf_o_amb', 0.028906879554259492),
('pf_o_sha', 0.04140947959585565),
('attr_o', 0.26647381704889245),
('sinc_o', 0.014717972920749327),
('intel_o', 0.019671562637699366),
('fun_o', 0.0929360009958484),
('amb_o', 0.008027323386644414),
('met_o', 0.0034648700986651742),
('age', 0.007783926761386303),
```

*Figure 32*

The highest two points are attraction of the female and fun of the female. Attraction of the female contributes 26.6% to the model and fun of female contributes 9.3% to the model.

### 5.1.5 Final model:

| | training | testing |
|---|---|---|
| Logistic regression model | 76.28% | 70.51% |
| Logistic regression model - varibale selection | 73.80% | 73.56% |
| Decision tree model | 73.77% | 73.44% |
| Random forest model | 97.70% | 79.32% |
| XGBoost model | 99.91% | 82.37% |

*Figure 33*

So the final model we use to predict men's decision is XGBoost model, which means if we conduct another round of speed dating event, we should be able to predict male's decision correctly after the speed dating event for 82.4% of the time.

## 5.2 Predict result before speed dating

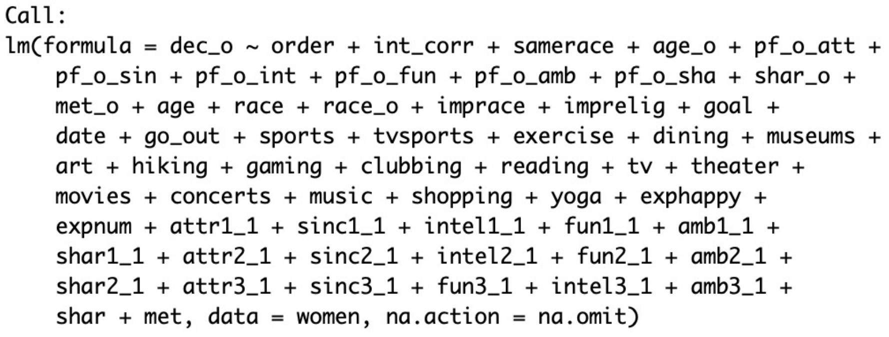
Speed dating event is try to help people find their dates, or hopefully, their life partner in

4 minutes. However, is that possible to find your true love before the event? We want to rule out people that are definitely not your type and save your time. We want to increase the efficiency before people went for the speed dating event. So is that possible?

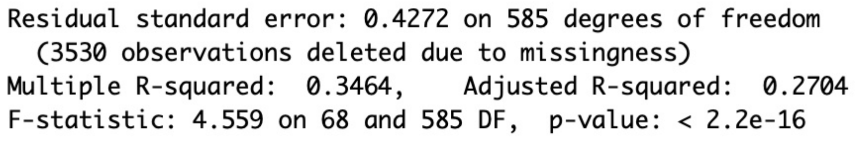
To find out the result, we need to eliminate all variables that we get after they met, which are attr, fun, amb, sinc, intel, attr_o, fun_o, amb_o, sinc_o,intel_o, dec, total 11 variables. Basically all these 11 variables matters for previous models, especially the most significant variables - attraction and fun of the women. Now we only have basic personal information of each participants. However, although we know nothing about how male will judge their female partner, we know what each male subject is looking for in the opposite sex (attr1_1 and so on) and how female think they measure up (attr3_1 and so on). So hopefully we can still have some useful models which has prediction power above 50%.

### 5.2.1 Model 1 - Logistic regression

Now, let us put all variables into the logistic regression model to see its summary:



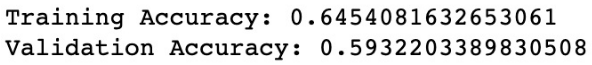
```
Call:
lm(formula = dec_o ~ order + int_corr + samerace + age_o + pf_o_att +
    pf_o_sin + pf_o_int + pf_o_fun + pf_o_amb + pf_o_sha + shar_o +
    met_o + age + race + race_o + imprace + imprelig + goal +
    date + go_out + sports + tvsports + exercise + dining + museums +
    art + hiking + gaming + clubbing + reading + tv + theater +
    movies + concerts + music + shopping + yoga + exphappy +
    expnum + attr1_1 + sinc1_1 + intel1_1 + fun1_1 + amb1_1 +
    shar1_1 + attr2_1 + sinc2_1 + intel2_1 + fun2_1 + amb2_1 +
    shar2_1 + attr3_1 + sinc3_1 + fun3_1 + intel3_1 + amb3_1 +
    shar + met, data = women, na.action = na.omit)
```



```
Residual standard error: 0.4272 on 585 degrees of freedom
  (3530 observations deleted due to missingness)
Multiple R-squared:  0.3464,    Adjusted R-squared:  0.2704
F-statistic: 4.559 on 68 and 585 DF,  p-value: < 2.2e-16
```

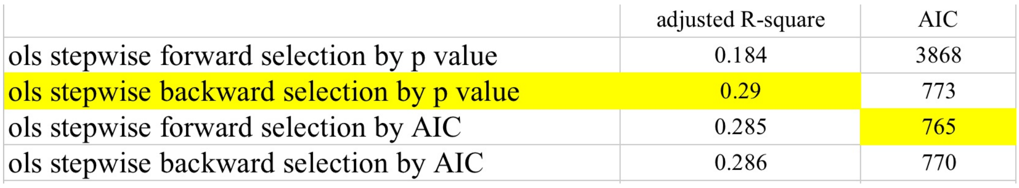
Since model’s p-value is very low, we believe the model is useful. However, the multiple R-square is only 34.6%, which means only 34.6% of the data can be explained by the model. The adjusted R-square is even lower, only 27.0% which is much lower than the multiple R-square, which means we have too many variables which causes over-fitting . There is much more improvement of the model should be done.

We will also check its prediction power:



```
Training Accuracy: 0.6454081632653061
Validation Accuracy: 0.5932203389830508
```

We get 64.5% accuracy for 80% of the training data set and 59.3% accuracy for testing data set, which means if we conduct another round of speed dating event, we should be able to predict male’s decision correctly after the speed dating event for 59.3% of the time. Since the prediction power is very close to 53%, we can not even confidently state that the logistic regression model is useful anymore. So Let us improve the model by variable selection first.

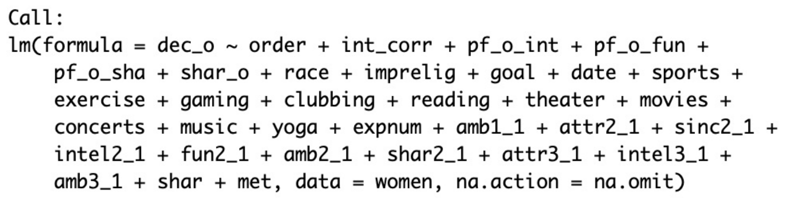


| | adjusted R-square | AIC |
|---|---|---|
| ols stepwise forward selection by p value | 0.184 | 3868 |
| ols stepwise backward selection by p value | 0.29 | 773 |
| ols stepwise forward selection by AIC | 0.285 | 765 |
| ols stepwise backward selection by AIC | 0.286 | 770 |

*Table 3*

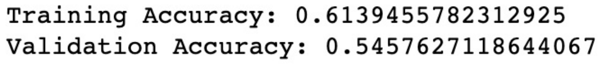
We can see that ols stepwise forward selection by AIC has lowest AIC and ols stepwise backward selection by p-value has highest adjusted R-square. Both methods have very similar adjusted R-square and AIC, so either one is fine. So I choose ols stepwise backward selection by p-value. Its adjusted R-square improved 2% compare to full model.

Here is the final model for logistic regression model after variable selection:



```
Call:
lm(formula = dec_o ~ order + int_corr + pf_o_int + pf_o_fun +
    pf_o_sha + shar_o + race + imprelig + goal + date + sports +
    exercise + gaming + clubbing + reading + theater + movies +
    concerts + music + yoga + expnum + amb1_1 + attr2_1 + sinc2_1 +
    intel2_1 + fun2_1 + amb2_1 + shar2_1 + attr3_1 + intel3_1 +
    amb3_1 + shar + met, data = women, na.action = na.omit)
```

We will check its prediction power:



```
Training Accuracy: 0.6139455782312925
Validation Accuracy: 0.5457627118644067
```

We get 61.4% accuracy for 80% of the training data set and 54.6% accuracy for testing data set, which means although variable selection method remain the same adjusted R-square, it loses its prediction power. So we will keep the full model here as base model, instead of using variable selection method.

### 5.2.2 Model 2 - Decision tree

We will use decision tree as the compared model. Here is the result:

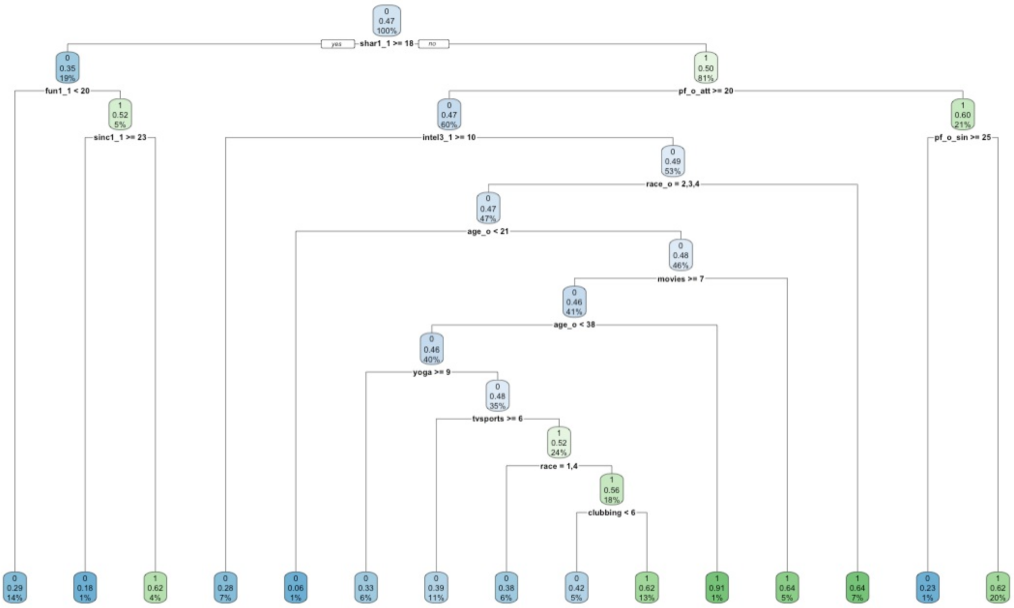
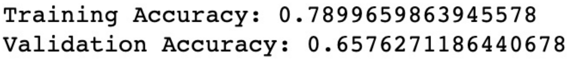


*Figure 34*

We can see the model is much more complicated than decision tree model to predict males' decision after the speed dating event, even with fewer independent variables. It means since we lose our most significant variables such as attraction and fun score from male to the female, all other variables matter in the model. The prediction power of decision tree model is 63.3%, which is 4% better than logistic regression model.

### 5.2.3 Model 3 - Random forest

For random forest model this time, we choose 20 trees and maximum depth of trees as 6, the model reach its best prediction power. Now let us see the result for random forest model:



```
Training Accuracy: 0.7899659863945578
Validation Accuracy: 0.6576271186440678
```

We get 79.0% accuracy for 80% of the training data set and 65.8% accuracy for testing data set, which means if we conduct another round of speed dating event, we should be able to predict male's decision correctly even before the speed dating event for 65.8% of the time. So random forest model improved 6.4% accuracy compare to logistic regression models.

We can also have a view of what variables contributing most to the model:

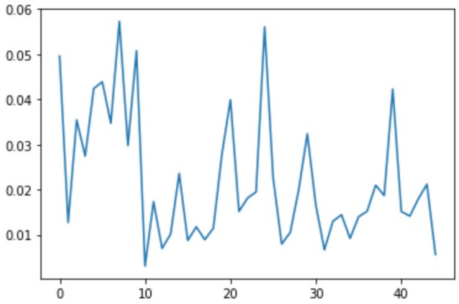
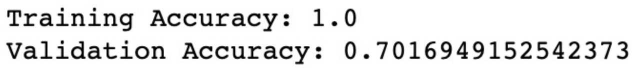


*Figure 35*

So this time, a lot of variables are all contributing significantly to the model. The the model algorithms become even more vague.

### 5.2.4 Model 3 - XGBoost

XGBoost model here we use parameters : number of trees is 85, max depth is 7 and learning rate is 0.15, which reaches its best prediction power. Now let us see the result for XGBoost model:

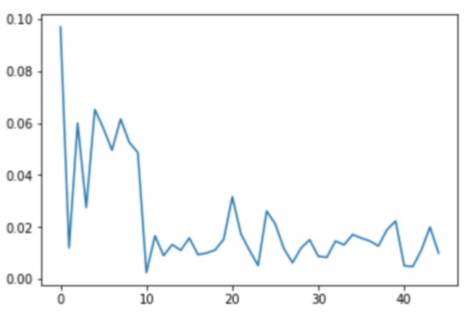


```
Training Accuracy: 1.0
Validation Accuracy: 0.7016949152542373
```

We get 100% accuracy for 80% of the training data set and 70.2% accuracy for testing data set, which means if we conduct another round of speed dating event, we should be able to predict male's decision correctly after the speed dating event for 70.2% of the time.

XGBoost model improved 11% accuracy compare to logistic regression models.

We can also have a view of what variables contributing most to the model:



```
[('int_corr', 0.09694354146666823),
 ('samerace', 0.011993280984169975),
 ('age_o', 0.05994800425998787),
 ('race_o', 0.027515866026236714),
 ('pf_o_att', 0.06515873311761673),
 ('pf_o_sin', 0.05792221392681177),
 ('pf_o_int', 0.049565459895385904),
 ('pf_o_fun', 0.06150609060435874),
 ('pf_o_amb', 0.052587794646043594),
 ('pf_o_sha', 0.048616863133196714),
 ('met_o', 0.002450022236359496),
 ('age', 0.01661383593962953),
 ('race', 0.008980596778975041),
 ('imprace', 0.013229520666776663),
 ('imprelig', 0.011030653814524375),
```

*Figure 36*

The highest point is shared interest which contributes 9.7% to the model. Age of female, expectation of the male are also significantly contributes to the model around 6% each.

### 5.2.5 Final model:

| | training | testing |
|---|---|---|
| Logistic regression model | 64.54% | 59.32% |
| Logistic regression model - varibale selection | 61.39% | 54.58% |
| Decision tree model | 66.90% | 63.28% |
| Random forest model | 79.00% | 65.76% |
| XGBoost model | 100.00% | 70.17% |

*Table 4*

We can see that XGBoost is still the best model to predict male's decision before the speed dating event. Since 53% of the time men will say yes to the women they met in speed dating, logistic regression model with 59.3% prediction power is not a significant improvement. XGBoost has prediction power of 70.2%, which is about 11% higher than logistic regression and 17% higher than random choose “no” for all the time. This means XGBoost is significantly a useful model to predict male’s decision before he even met the women than random guess. This gives us the hope that we may have ability to discover the secret of love - what makes you like somebody, with the fast development of machine learning algorithms nowadays.

## 5.3 Conclusion

The best model I finally have is XGBoost model. It has a 82.4% precision on testing data-set based on all the information we have after the speed dating event, and still a 70.2% precision on testing data-set even we only use all information before two people never actually met on the speed dating event. So we can believe that we have the ability to discover the secret of love with modern machine learning algorithms if we have enough information.

# CHAPTER 6

## 6 Future Research Directions

I think there are many possible applications for this thesis and numerous future research directions. There are three possible ways to improve- data, variables and methods.

### 6.1 Data

The data is only collected from University of Columbia and it is biased in multiple ways. We should choose participants in a random way from all population. However, it is very much difficult in real life situation and even not necessary in practical. The most practical way is to acknowledge the limitation, the conclusion only apply to the limitation of the data, for example, only apply to New York area if we expand the speed dating events from University of Columbia to New York area or only apply to all US collage students if we expand the speed dating events to all US collages students. The other most practical way is to build an app and store all participants information, update our results after each round of speed dating event. Although the conclusion may still be biased, but it is less and less biased as the data become bigger and bigger, and it will always provide useful suggestion for particular population. Take this data set as an example, the conclusion will be the most useful information for students from University of Columbia who are willing to take part in the speed dating event.

### 6.2 Variables

We could also expand our variables from very basic personal survey information and consensus after the speed dating event, to more personal life information if we could work with large website like Facebook, Amazon, Tinder and so on. We could gather who are their friends, what kind of stuff they prefer to buy, what kind of person they think are attractive. With all kinds of information from previous life and different area of the life, we can get a very clear portfolio of the person and may know what kind of person are the perfect match for them.

## 6.3 Methods

In this paper, I only use linear regression, random forest and XGBoost methods. However, as data get larger and more dimension, we could even build a better algorithm to better perform the relationship between love and person.

# 7 References

1.RAYMOND FISMAN, SHEENA S. IYENGAR, EMIR KAMENICA, ITAMAR SIMONSON, “GENDER DIFFERENCES IN MATE SELECTION: EVIDENCE FROM A SPEED DATING EXPERIMENT”,2008 The Review of Economic Studies Limited

2.RAYMOND FISMAN, SHEENA S. IYENGAR, EMIR KAMENICA, ITAMAR SIMONSON, “Racial Preferences in Dating”,2008 The Review of Economic Studies Limited

3.Ho, Tin Kam, “ Random Decision Forests ”. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 5 June 2016.

4.H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

5. M. Kubat and S. Matwin, “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection,” in Fourteenth International Conference on Machine Learning, Morgan Kaufmann, 1997, pp. 179- 186.

6.C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, “Conditional variable importance for random forests,” BMC Bioinformatics, vol. 9, no. 1, p. 307, 2008.

7. C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution,” BMC Bioinformatics, vol. 8, no. 1, p. 25, 2007.

8. Liaw A . "Documentation for R package randomForest". Retrieved 15 March 2013.

9.Breiman L . "Random Forests". Machine

Learning. 5–32. doi:10.1023/A:1010933404324.