ANALYSIS OF SPEED DATING

GROUP 4

Aikun Tao, Dantian Liu, Shaka Lohardjo, Xuefan Han, Zhilin Zhang

RESEARCH QUESTIONS

Speed dating is one of the most popular ways for people to meet potential romantic partners

QUESTION 1

What factors will result in another date?

QUESTION 2

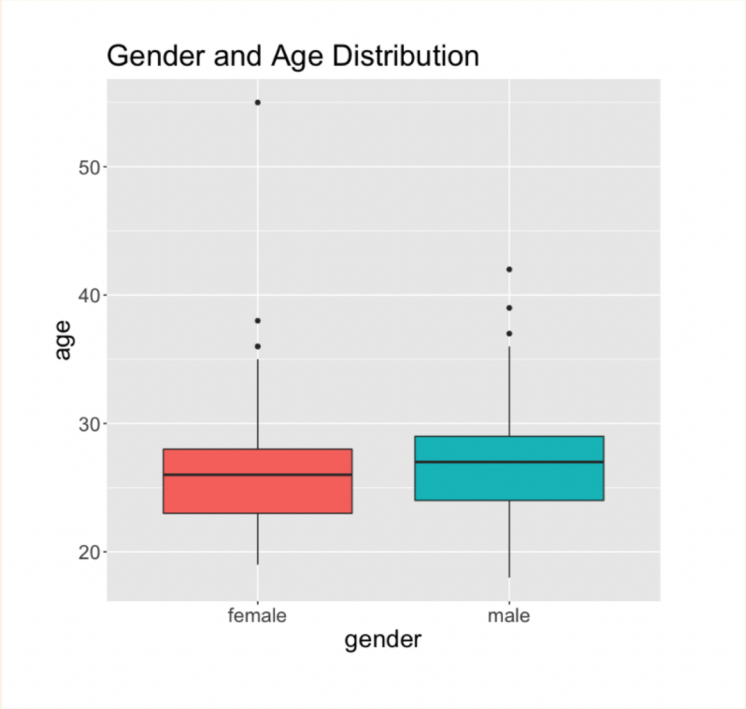
How to predict if two people will match after the first date?

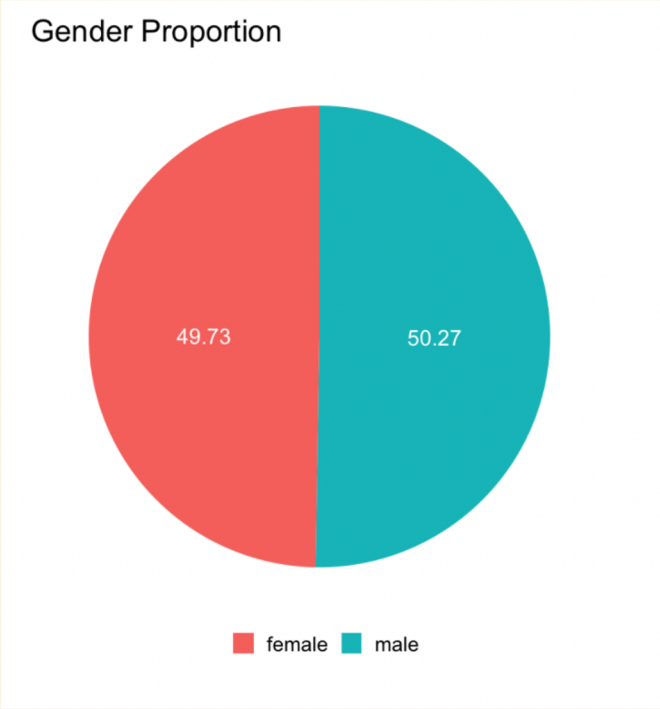
DATA RESOURCE

- Columbia Business School professors Ray Fisman and Sheena Iyengar's Speed Dating Experiment
- Data was collected from 552 participants in experimental speed dating events from 2002 to 2004
  - 21 waves during the events
- The original dataset contains 8,378 rows and 195 columns
  - 8 character variables and 187 numeric variables

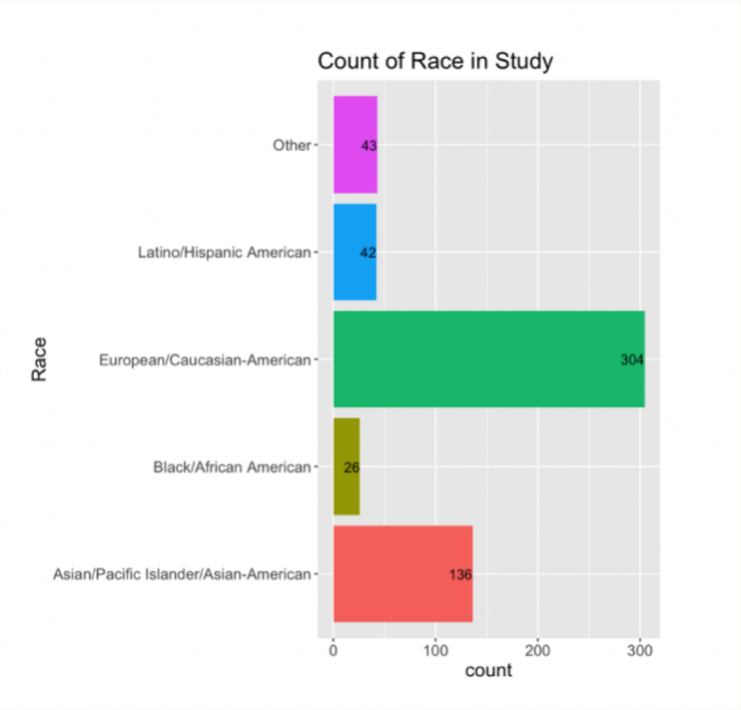
http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/

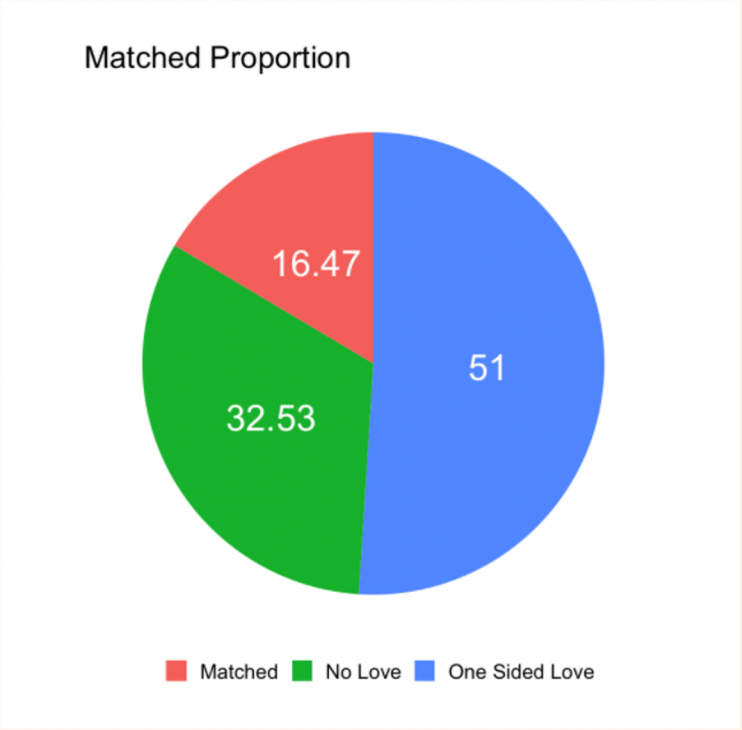
DATA EXPLORATION





DATA EXPLORATION





DATA CLEANING

- Removed duplicated fields
- Removed columns with more than 1000 null values
- Subset the data to only include relevant factors
- Imputed data entry mistakes
- Imputed remaining missing values with the mode of each column

FEATURE SELECTION

variables	tree1.variable.importance
like_o	189.62097
like	188.17105
attr	121.07047
fun	97.35299
attr_o	89.72151
fun_o	87.54267
prob_o	86.26018
sinc	62.96218
intel	61.53122
sinc_o	56.58380
prob	54.78527
intel_o	51.05774

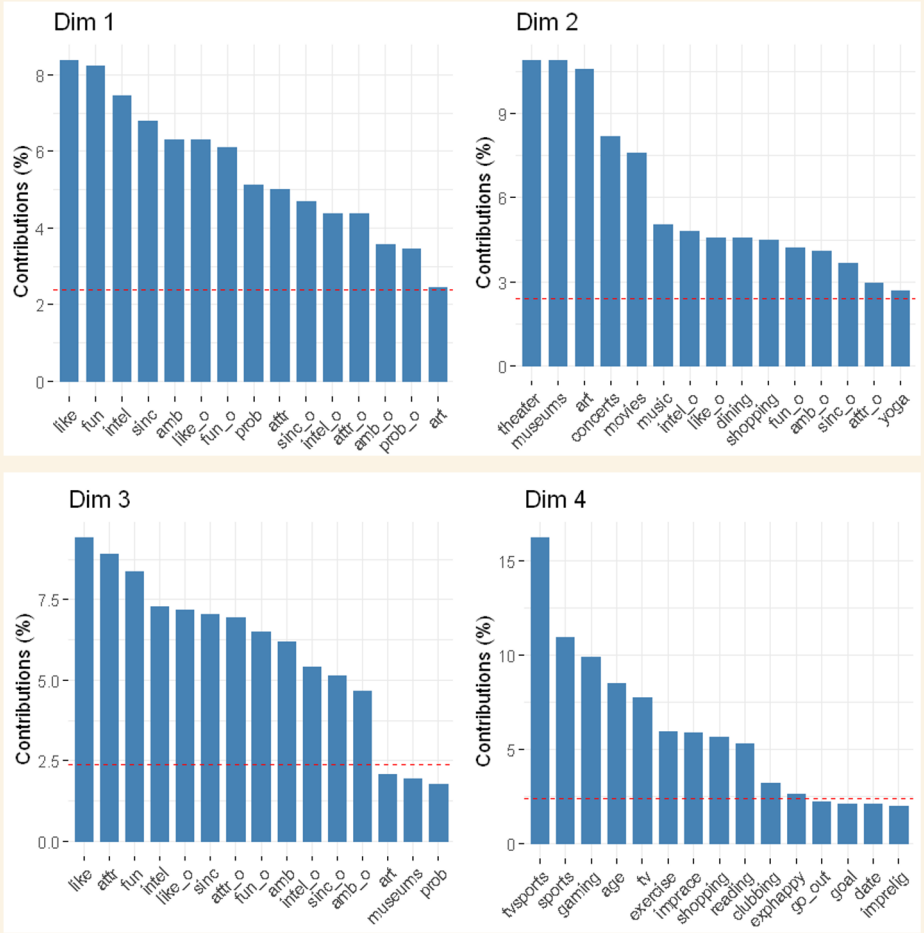
Variable Importance

- Apart from some obvious factors like ‘like’ for ‘if one likes their date’ and ‘attr’ for attractiveness of their date, other most important factors include ‘fun’, ‘sinc’ for sincerity, and ‘intel’ for intelligence

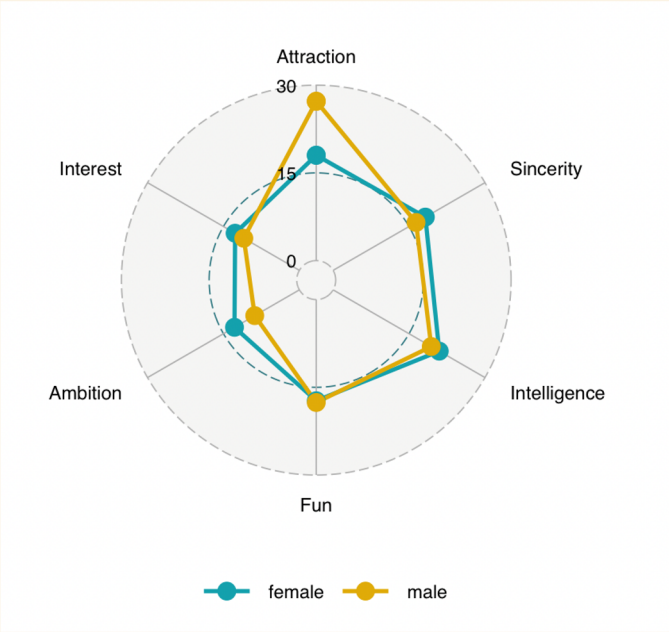
DIMENSION REDUCTION

Principal Components Analysis (PCA)

- Reduced to 4 dimensions
- Dimensions 1 and 2 are most relevant
- Personality factors like fun, sincerity, and intelligence are important in dates, and hobbies like going to the theater and museum and art are also contributing characteristics. We could also find that race and age were of relatively low importance in all dimensions

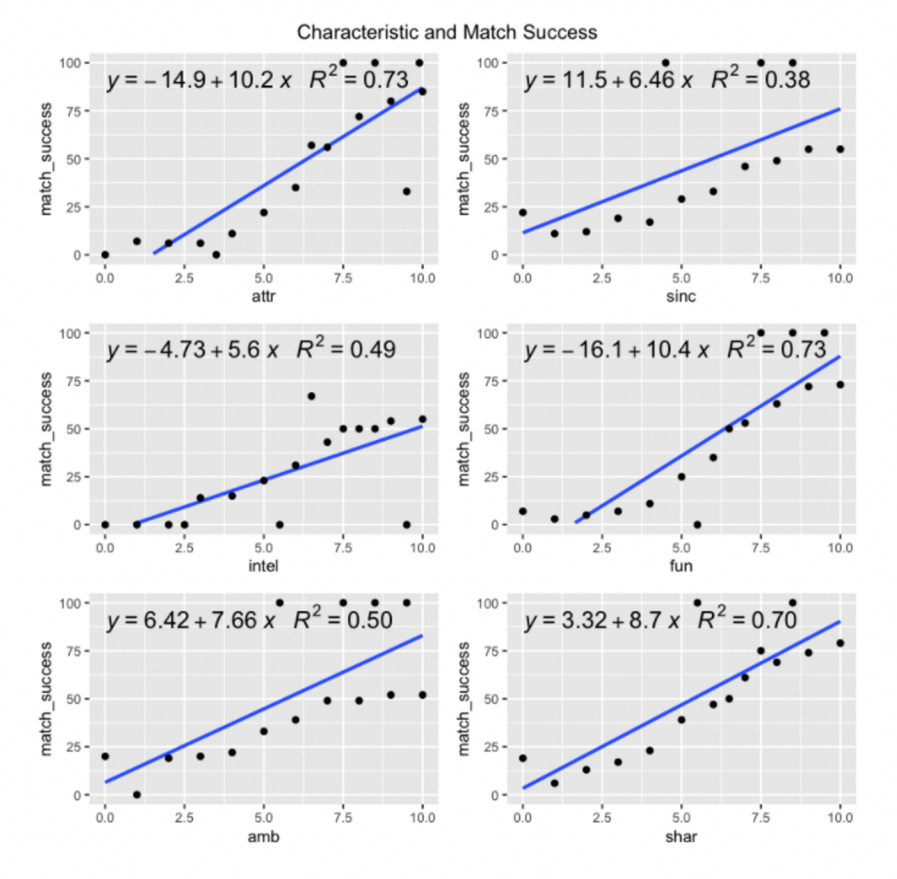


IMPORTANT CHARACTERISTICS BY GENDER



	Male	Female
Important	Attraction	Intelligence Attraction Sincerity
Unimportant	Ambition	Shared Interest Ambition

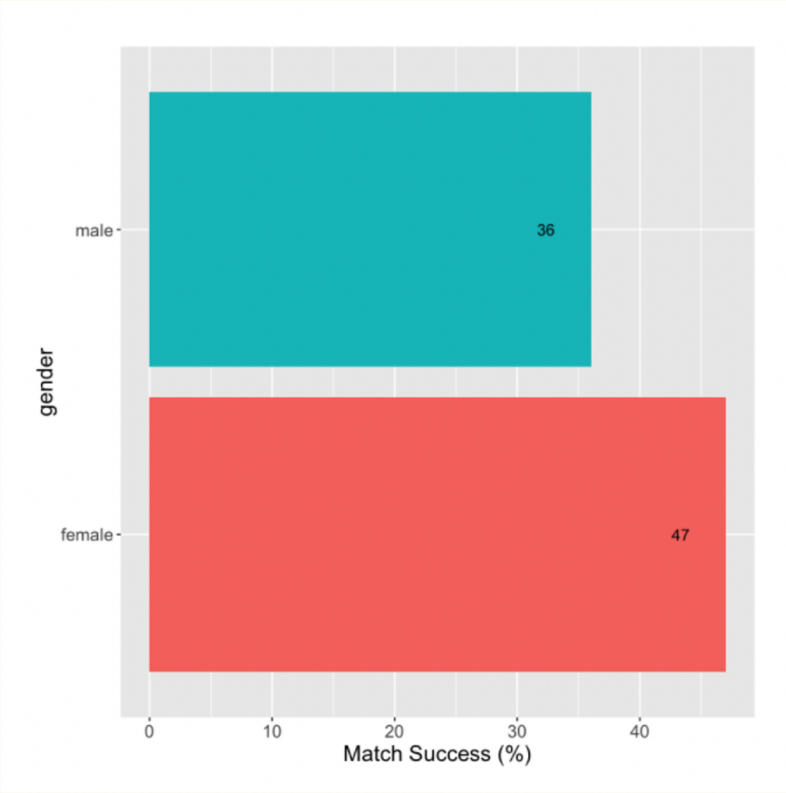
CHARACTERISTICS AND SUCCESSFUL ATES



Important Characteristics:

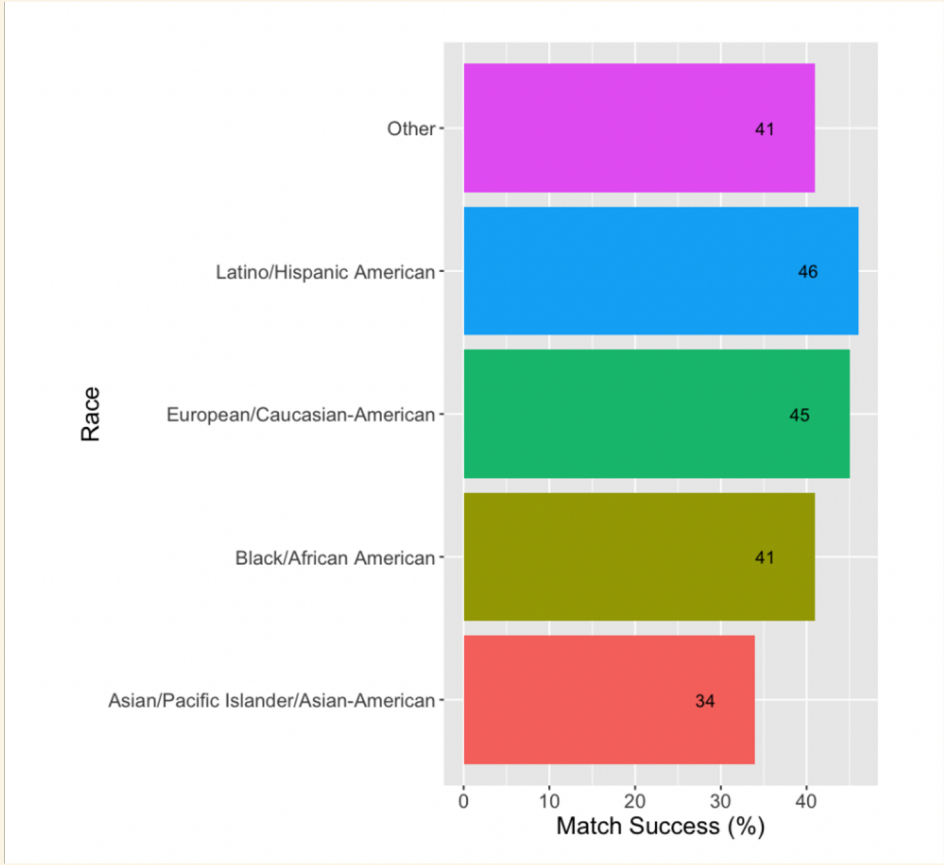
1. Attractiveness
2. Fun
3. Shared Interest

GENDER AND MATCHES



Females have more success than males

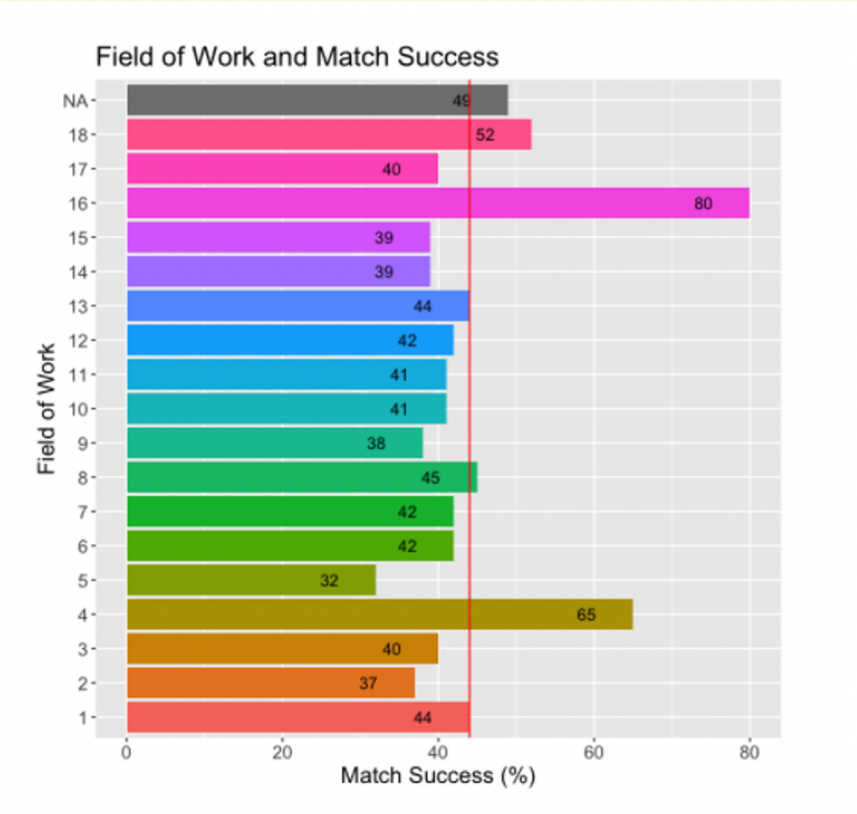
RACE AND MATCHES



Highest:
Latino/Hispanic American - 46%

Lowest:
Asian/Pacific Islander/Asian-American - 34%

FIELD OF WORK AND MATCHES



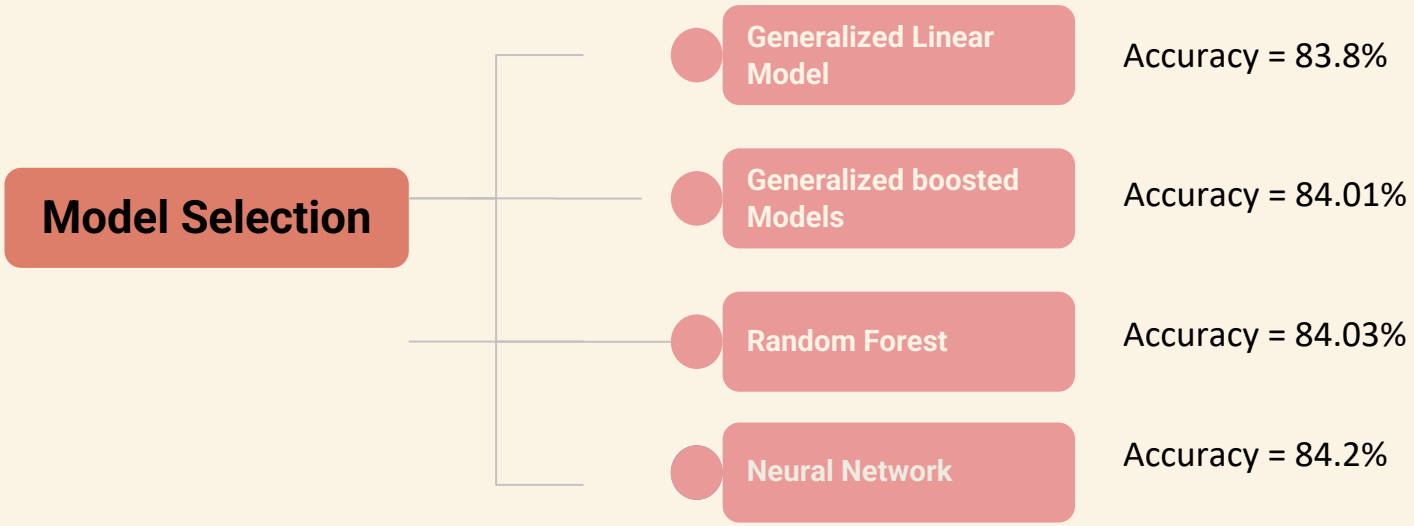
Highest:

- Language - 80%
- Medical Science/Pharmaceuticals - 65%

Lowest:

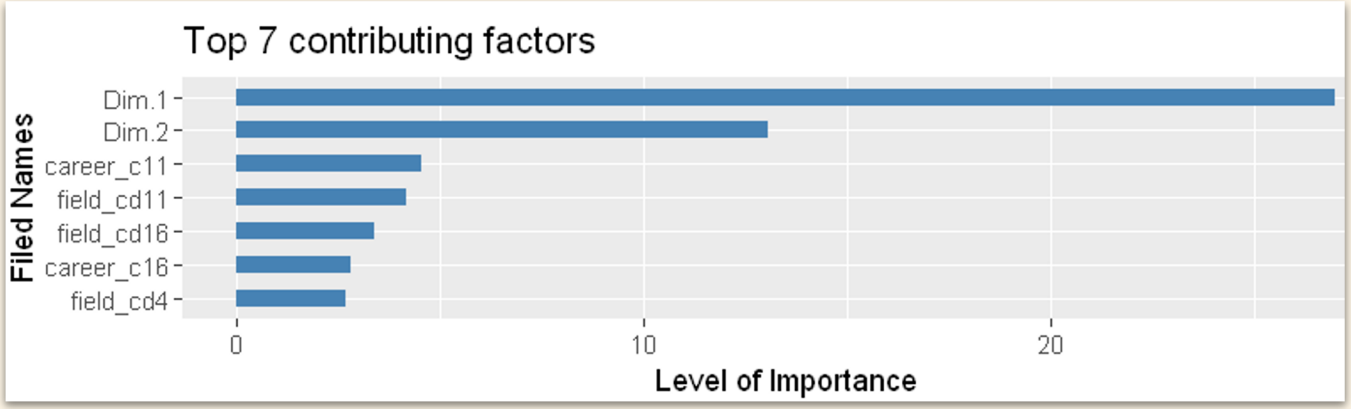
- Engineering - 32%
- Math - 37%

MODELS


Model Selection
Generalized Linear Model
Accuracy = 83.8%
Generalized boosted Models
Accuracy = 84.01%
Random Forest
Accuracy = 84.03%
Neural Network
Accuracy = 84.2%

RESULTS OF LOGISTIC MODEL

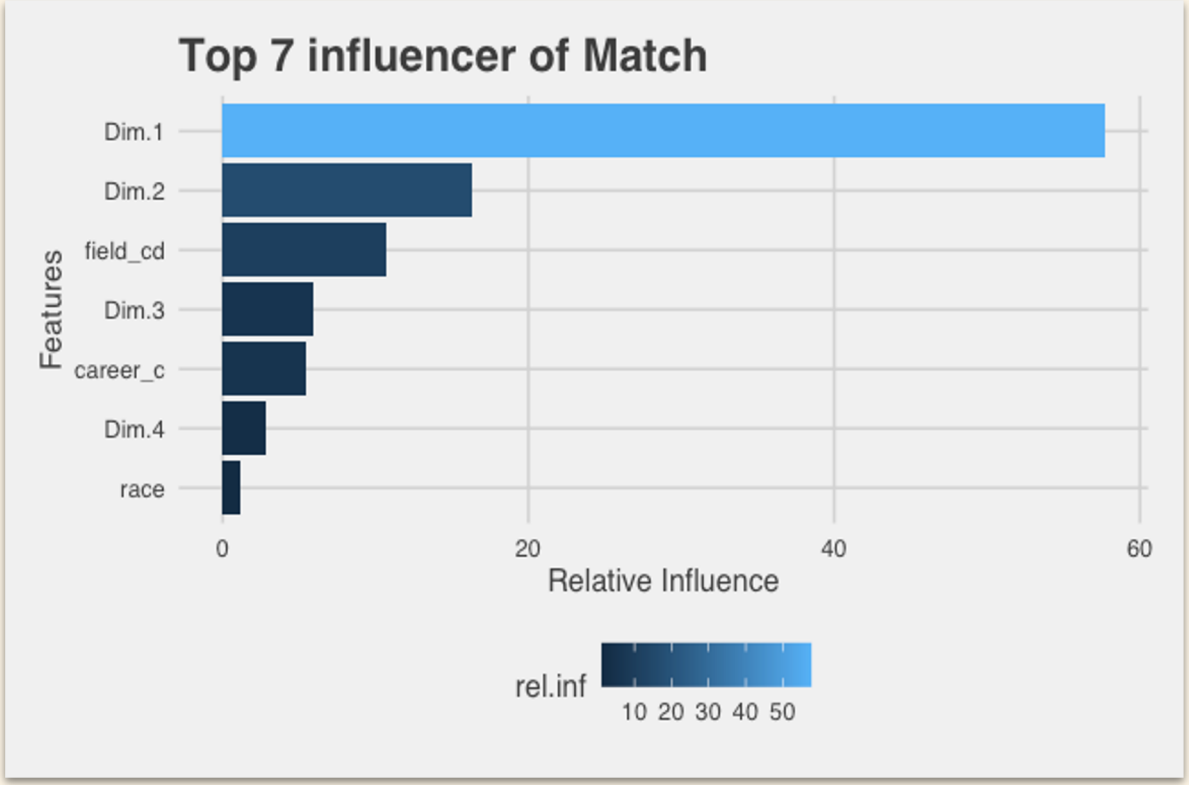
overall	names
26.9699708	Dim.1
13.0327609	Dim.2
4.5431878	career_c11
4.1439115	field_cd11
3.3769800	field_cd16
2.7878949	career_c16
2.6747610	field_cd4



- Both the chart and the bar plot show the top seven important variables for this model.
- Dimensions 1 and 2 contributed the most, and factors ‘career’ and ‘field’ were relatively more important than the rest.

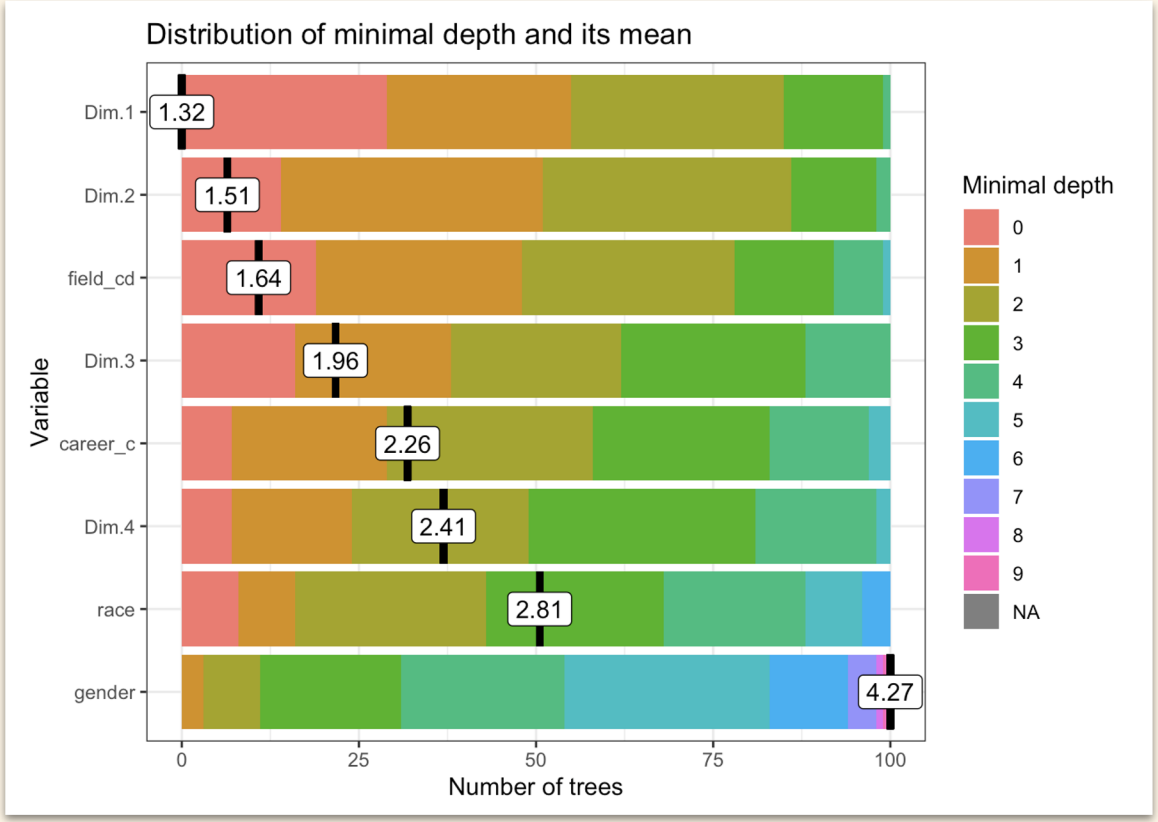
RESULTS OF GBM MODEL

- The result of the GBM model shows that **Dim.1**, **Dim.2** and **field_cd** are the top 3 variables that have the largest influence on the match result.



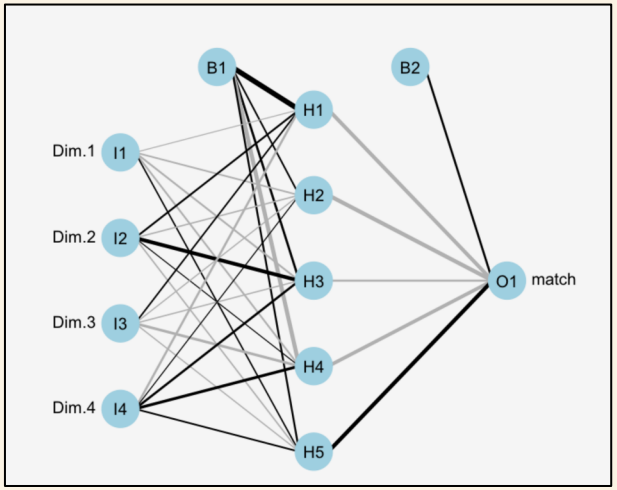
RESULTS OF RANDOM FOREST MODEL

- The distribution of the **mean minimal depth** reveals the variable's role in the random forest's structure and prediction.
- Smaller the mean minimal depth, the more important the variable is;
- **Dim.1** has a lowest mean minimal depth



FINAL MODEL: NEURAL NETWORKS

Neural Network Models								
Package	nnet	nnet	h2o	h2o	h2o	h2o	h2o	h2o
Variable	all	4	all	4	all	4	all	4
Tuned (Y/N)	N	N	N	N	Manual	Manual	Grid Search	Grid Search
Accuracy	0.840	0.842	0.786	0.788	0.810	0.814	0.802	0.837



CONCLUSIONS

QUESTION 1

What factors will result in another date?

Characteristics	Attractiveness, fun, shared interests, sincerity, intelligence, ambition
Study Field	Language, Medical Science/Pharmaceuticals
Hobbies	Theater and museum and art
Others	Race: Latino/Hispanic Americans; Gender: Female

QUESTION 2

How to predict if two people will match after the first date?

```
library(nnet)
set.seed(1031)
model_test = nnet(match~ Dim.1 + Dim.2+ Dim.3 + Dim.4,
                  data = train, size=5, decay=0.1, MaxNWts=10000, maxit=100)
```

RECOMMENDATIONS

- Be more extroverted when interacting with your date.
- Caring about your appearance can help increase your attractiveness.
- Having a wide variety of interests to share experiences with others can help boost your attractiveness.
- Adopting a lively lifestyle and having interests in sports,theater and museum and art are helpful.
- Having excellent communication skills might help boost the relationship.
- If your area of study is in Language and have certain knowledge in related fields, leverage your advantage.

LIMITATIONS

- Neural Network Model
  - Black-box nature, trade-off between accuracy and interpretability
- Untidy but meaningful data
  - “From” and “Zip code”
- Loss of track and follow-up
  - Whether the match was successful or not.
- Loss of timeliness
  - Might not reflect participants’ perceptions accurately.

THANK YOU