WILLIAM & MARY

CHARTERED 1693

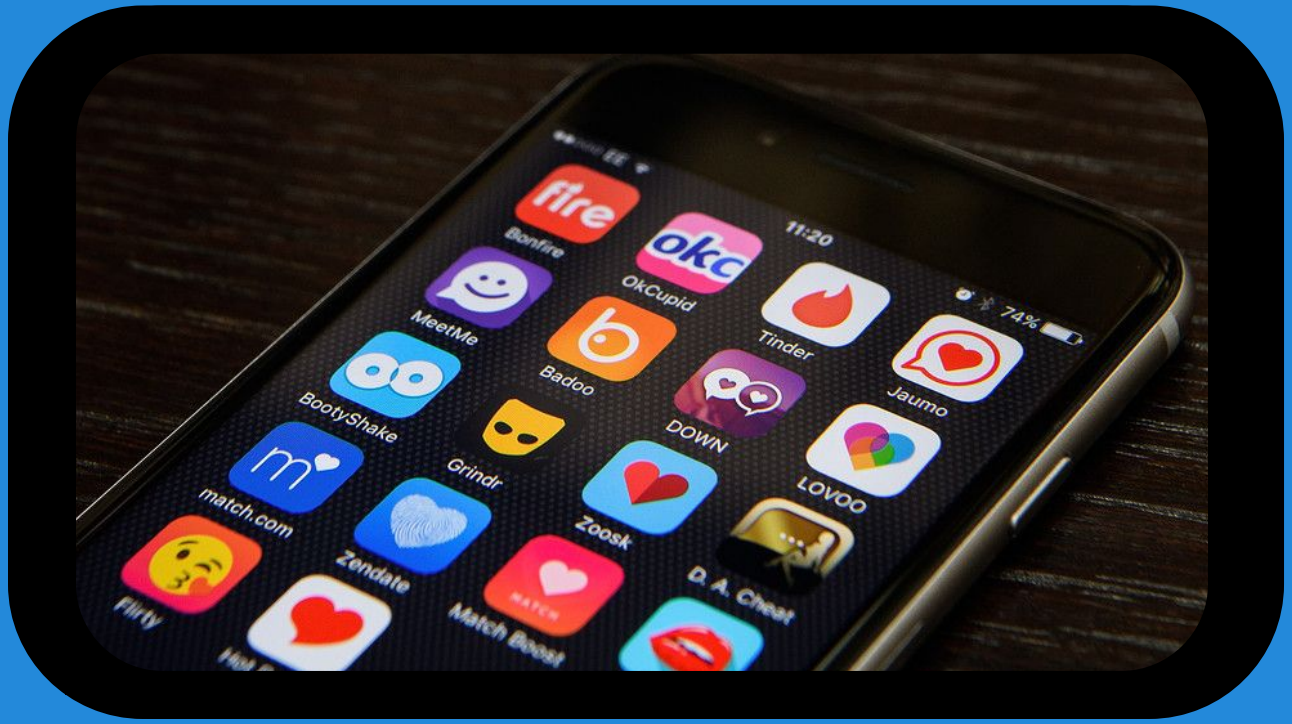
Speed Data-ing

Team 8

Isabel Hirama, Tammy Gong,

VJ Davey and Brian Knox

The Business Problem



Using data from a speed dating study, this project seeks to predict romantic interest, a question of major significance to the growing online dating industry.

The Business Problem

According to a 2015 Pew Research study, 15% of American adults have used online dating services.

The online dating market is crowded with many similar services in this $3 billion industry, so finding a way to distinguish a company from others is key to gaining market share.

The Business Problem

Our goal was to use machine learning techniques and a speed dating dataset to develop classification models to address the following:

Given a pair of speed dating participants:

A) Will the male be interested?

B) Will the female be interested?

The Business Problem

Reasoning Behind Research Question

- We first tried to predict matches based on pairs' ratings of each other's attractiveness, intelligence, etc.
- As a business case, this was not useful because online dating sites would not have this data (at least not until after a pair had already interacted)
- We chose to predict the decision of an individual side, male or female, based on the all available data from the speed dating experiment **except** their ratings of each other.
- This way a dating site can use previously collected information to determine matches before a couple meets and provides value to the users

Data Collection and Cleaning

Data Info

"Speed Dating Experiment" Dataset from Kaggle

- 8,378 observations (each observation is a different pair of speed-daters)
- 195 columns
- Data was collected between 2002 and 2004 from participants in speed dating events at Columbia University.
- Participants spent four minutes with a partner before moving on the next. Each participant was asked if they would like to see the other again, and rate them on attractiveness, sincerity, intelligence, fun, ambition, and shared interests.
- Participants were previously asked to rate how important each of the six attributes was to them in potential partners.

Data Cleaning

Excel

- Created unique identifier for each pair: **pair_id** by concatenating iid and pid
- Removed unwanted columns which either had too few observations or irrelevant information
- Removed 10 rows with missing partner id
- Removed 82 rows with missing partner-to-self ratings
- Renamed remaining columns with more descriptive variable names and updated the data dictionary
- Filled in missing values with information given in previous columns
- Corrected for different preference scales
- Created calculated fields:
    - Average rating for each person by their partners, for each of the 5 attributes
    - Proportion of each person's partners who were interested (interested partners/all partners)

Data Cleaning

SQL

- Imported the CSV into MySQL using Table Data Import Wizard
- Split data into two tables, male-first pairs and female-first pairs
- Deleted columns in the male table with the female data, and columns in the female table with male data
- Used an inner join to merge the two tables, each row became a pair with all data on both the male and female. There are no repeated pairs or duplicate columns
- Exported back to a CSV file

Variables

Y Variable

- **Female_dec, male_dec** (the decision of female and male)

X Variables (for females - all variables were also repeated for males)

f_rate_m_attr	f_age	f_career_num	f_like_reading	f_pref_sinc	f_genderpref_shar	f_rate_self_intel
f_rate_m_sinc	f_field	f_like_sports	f_like_tv	f_pref_intel	f_oppgender_pref_attr	f_rate_self_amb
f_rate_m_intel	f_field_num	f_like_tvsports	f_like_theater	f_pref_fun	f_oppgender_pref_sinc	f_avg_attr
f_rate_m_fun	f_race	f_like_exercise	f_like_movies	f_pref_amb	f_oppgender_pref_intel	f_avg_sinc
f_rate_m_amb	f_imp_race	f_like_dining	f_like_concerts	f_pref_shar	f_oppgender_pref_fun	f_avg_fun
f_rate_m_shar	f_imp_relig	f_like_museums	f_like_music	f_genderpref_attr	f_oppgender_pref_amb	f_avg_intel
f_rate_m_like	f_goal	f_like_art	f_like_shopping	f_genderpref_sinc	f_oppgender_pref_shar	f_avg_amb
f_rate_m_prob	f_dat_freq	f_like_hiking	f_like_yoga	f_genderpref_intel	f_rate_self_attr	f_avg_like
f_rate_m_met_before	f_out_freq	f_like_gaming	f_exp_happy	f_genderpref_fun	f_rate_self_sinc	
f_est_matches	f_career	f_like_clubbing	f_pref_attr	f_genderpref_amb	f_rate_self_fun	

Variables

Key Variables Explained

- The f_rate or m_rate variables are the ratings of each individual on things like attractiveness, sincerity, intelligence, ambition and others
- The f_like or m_like variables are how much the individual likes certain topics such as movies, sports, books and more
- The f_genderpref or m_genderpref variables are how much the individual thinks their gender likes certain things in general
- The f_oppgender_pref or m_oppgender_pref variables are how much the individual thinks the opposite gender likes certain things in general
- The f_avg or m_avg variables calculated the average of the ratings given to the individual by their speed dating partners (between 10 and 20 people)

Unsupervised Learning

Exploration using hierarchical clustering

See Dendrogram section of R script

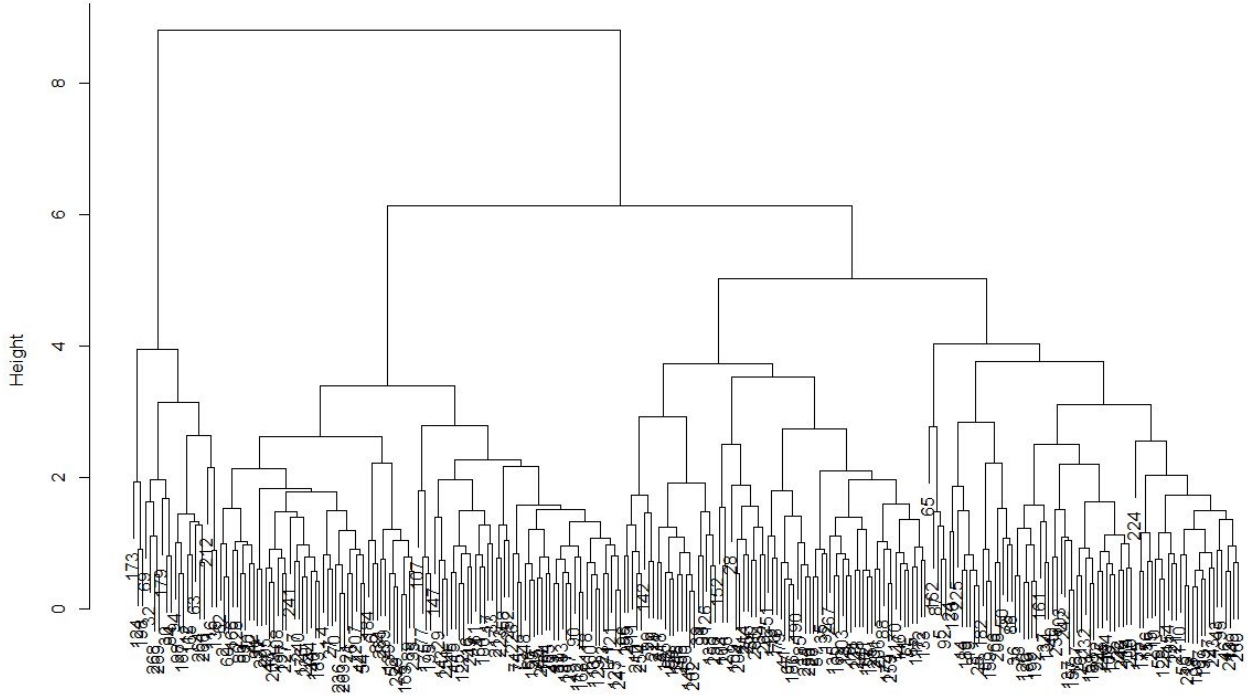
Exploration With Dendrograms

Using unsupervised learning, we conducted an initial exploration of what variables might be significant to partner decisions.

- For each gender, a dendrogram was created using the following attributes:
  - **Average Attractiveness Rating**
  - **Sincerity**
  - **Fun**
  - **Ambition**
  - **Intelligence**
  - **Likeability**
- Each node on the dendrogram represented a specific person and their data for each attribute was the previously calculated variable giving the average of the ratings given them by their partners (f_avg_attr, f_avg_sinc, etc)

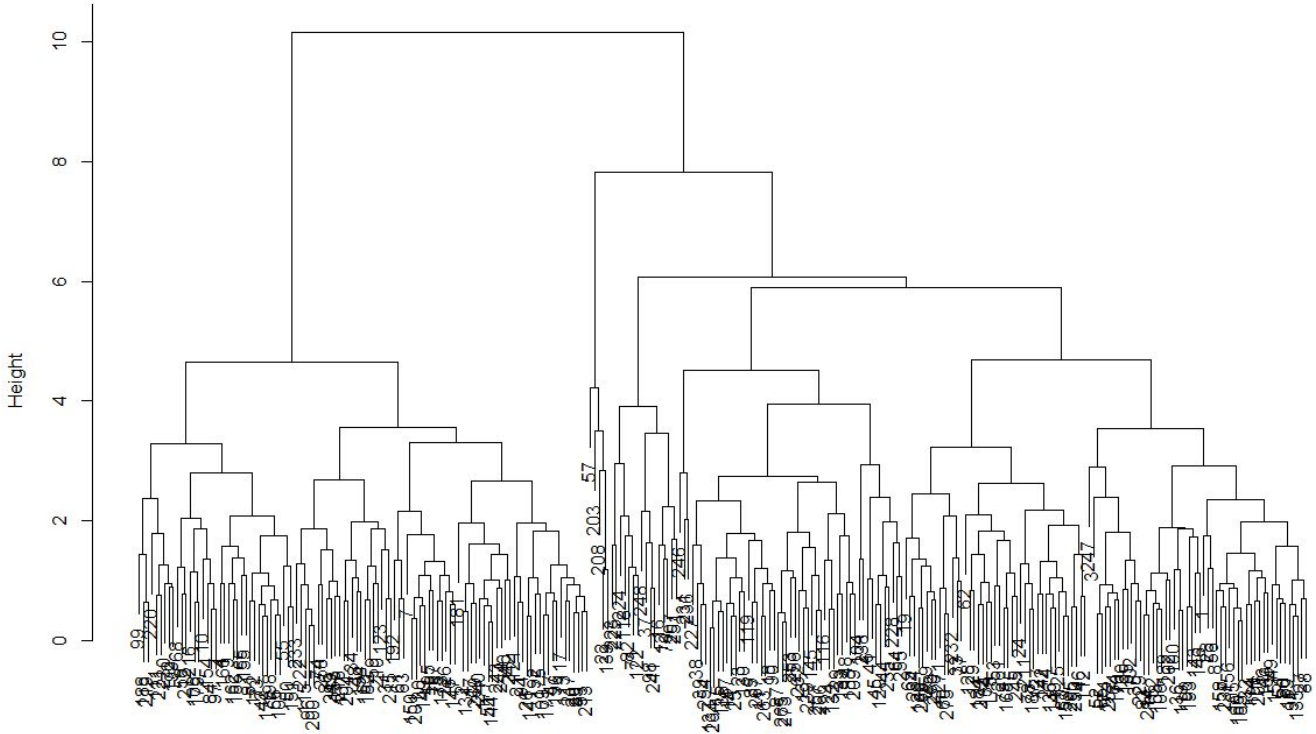
Exploration With Dendrograms

Dendrogram of Female Attribute Ratings



Exploration With Dendrograms

Dendrogram of Male Attribute Ratings



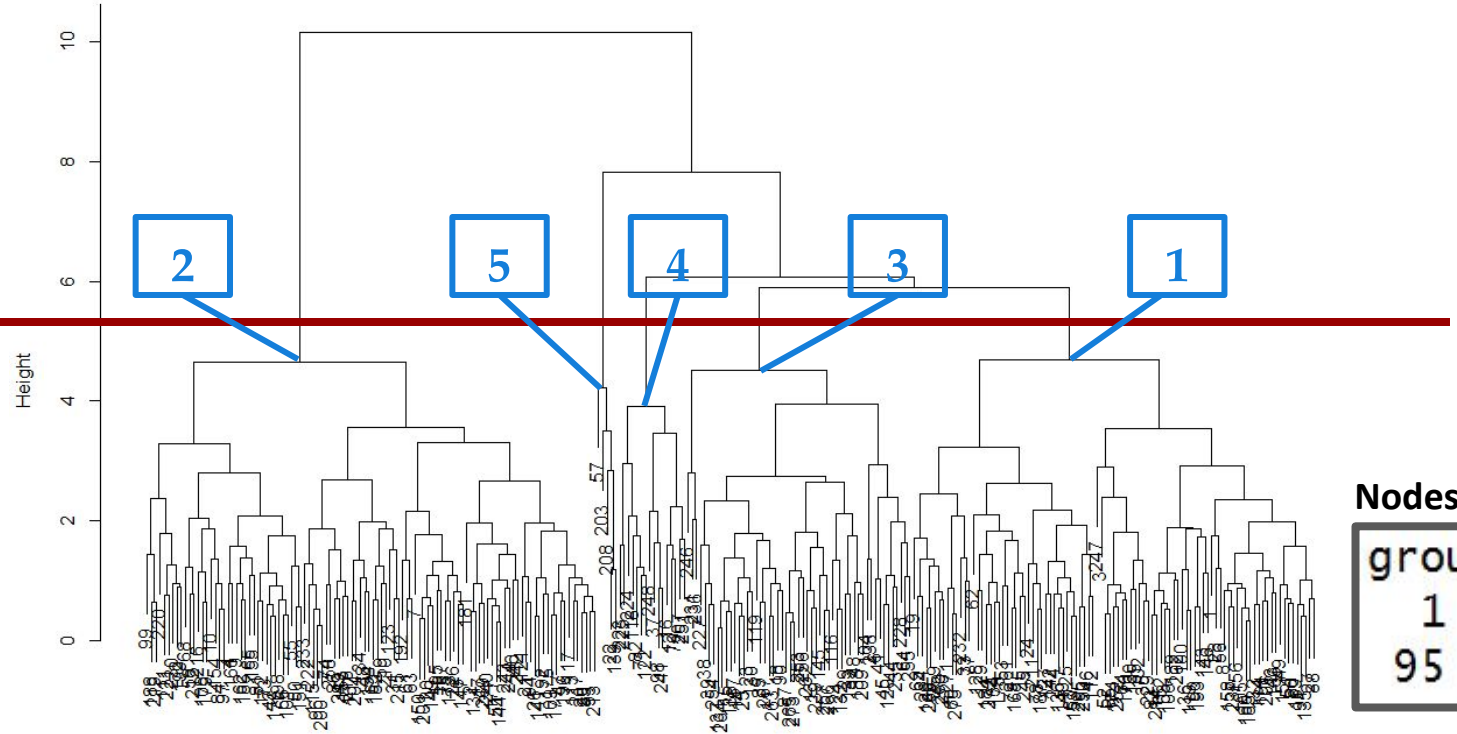
Exploration With Dendrograms

Next, we clipped the dendrograms to break users into clusters.

- For **males**, the dendrogram was clipped to form **five clusters.**
- The **female** dendrogram was clipped to form **four clusters**.
- Clipping decisions were made based on a visual assessment of the dendrograms.

Exploration With Dendrograms

Dendrogram of Male Attribute Ratings

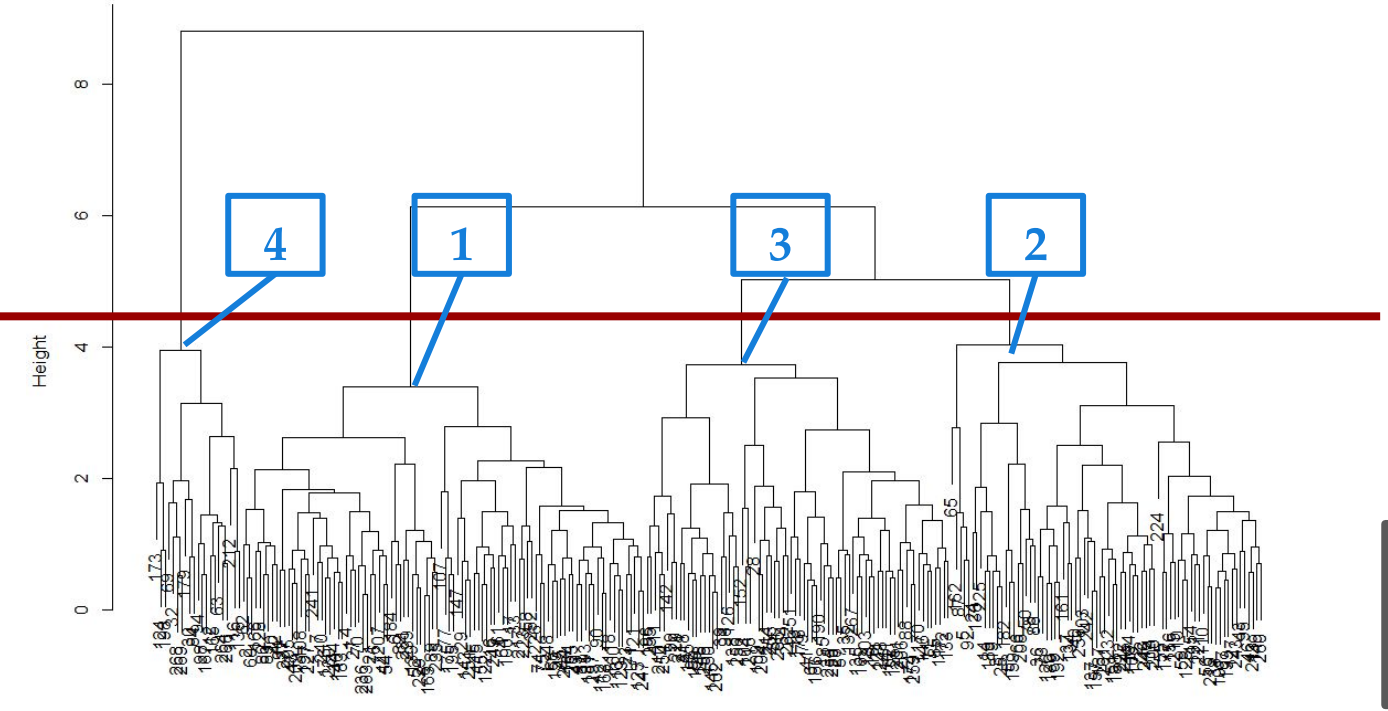


Nodes Per Cluster (Male)

```
groups.5
  1   2   3   4   5
 95 106  53  16   5
```

Exploration With Dendrograms

Dendrogram of Female Attribute Ratings



Nodes Per Cluster (Female)

```
groups.4
 1  2  3  4
98 76 74 21
```

Exploration With Dendrograms

Next, we viewed the average values for each of the attributes for each cluster

- We also calculated each cluster's average for a new variable: **percent_interest**
- This was the number of yes's (interested partners) divided by the total number of partners.
- E.g., for a participant where all partners said that would be interested in seeing them again, percent_interest would be 1.

Exploration With Dendrograms

Females

Cluster	f_avg_attr	f_avg_sinc	f_avg_fun	f_avg_intel	f_avg_amb	f_avg_like	f_percent_interest
1	7.500000	7.555556	7.542484	7.200000	7.054412	6.968750	0.6846591
2	6.418831	7.420168	7.394444	6.581250	6.645833	6.368056	0.4583333
3	5.750000	6.950000	7.000000	6.000000	6.285948	5.690972	0.2817460
4	4.666667	6.625000	6.750000	4.857143	5.700000	4.952381	0.1500000

Males

Cluster	m_avg_attr	m_avg_sinc	m_avg_fun	m_avg_intel	m_avg_amb	m_avg_like	m_percent_interest
2	7.000000	7.444444	7.900000	7.224790	7.452083	6.804762	0.57142857
1	5.833333	7.375000	7.500000	6.058824	6.882353	5.857143	0.33333333
3	4.700000	6.647059	6.904762	5.142857	6.300000	5.150000	0.18181818
5	3.937500	4.850000	6.500000	4.200000	6.111111	3.833333	0.05555556
4	3.951389	6.525000	7.666667	5.083333	7.431373	4.964286	0.02500000

Insights:

For both genders, the cluster with the highest **interest** scores had high ratings for all attributes.

As **interest** decreased, there was a steady decrease in ratings for **attractiveness**, **intelligence**, and **likeability**, suggesting that these attributes are valuable predictors of interest.

The other three attributes did not vary as greatly or steadily, suggesting that they were less impactful on interest.

Modeling

Steps taken during our modeling process

Modeling

Models Used

- Principle Component Analysis
- Decision Trees
- Boosting
- Random Forest
- Support Vector Machines

Principal Component Analysis

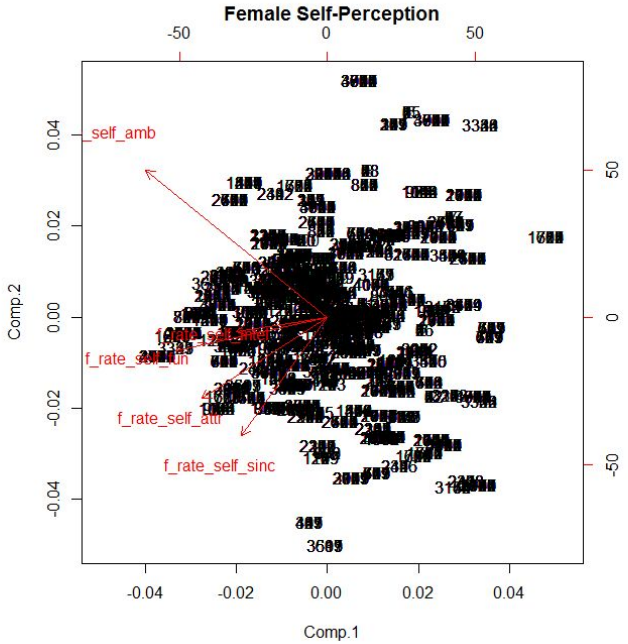
See Principal Component creating R script

Principal Component Analysis

Principle Components created included the following for both males and females. Male categories listed below. Female categories included the same variables but with the prefix "m_" swapped out for "f_"

- Self Perception (a combination of variables with the "m_rate_self" prefix for males)
- Opposite Gender Perception (a combination of variables with the "m_oppgender_pref_" prefix for males)
- Preferences (a combination of variables with the "m_pref_" prefix for males)
- Interests (a combination of variables with the "m_like" prefix for males)

Principal Component Analysis



Pictured above is a biplot of the first two Principal Components made based around the 5 variables which measured female self-rating. In all, we would end up using four Principal Components in our analysis, each contributing to a new group of variables we would call "female self perception"

Assessment of Using Principal Components

Pros

- Compresses a set of multiple vectors into a set with fewer vectors that still captures a certain amount of variance of the data
- May lead to better predictions when faced with many variables when unsure of how they all relate to each other
- The importance of a given principal component is easily interpretable through models

Cons

- Understanding relationships suggested by models may be difficult to interpret without explicit knowledge of how each variable for a principal component relates to that component

Decision Trees

See Decision Tree model in R script

Assessment of Categorical Decision Trees

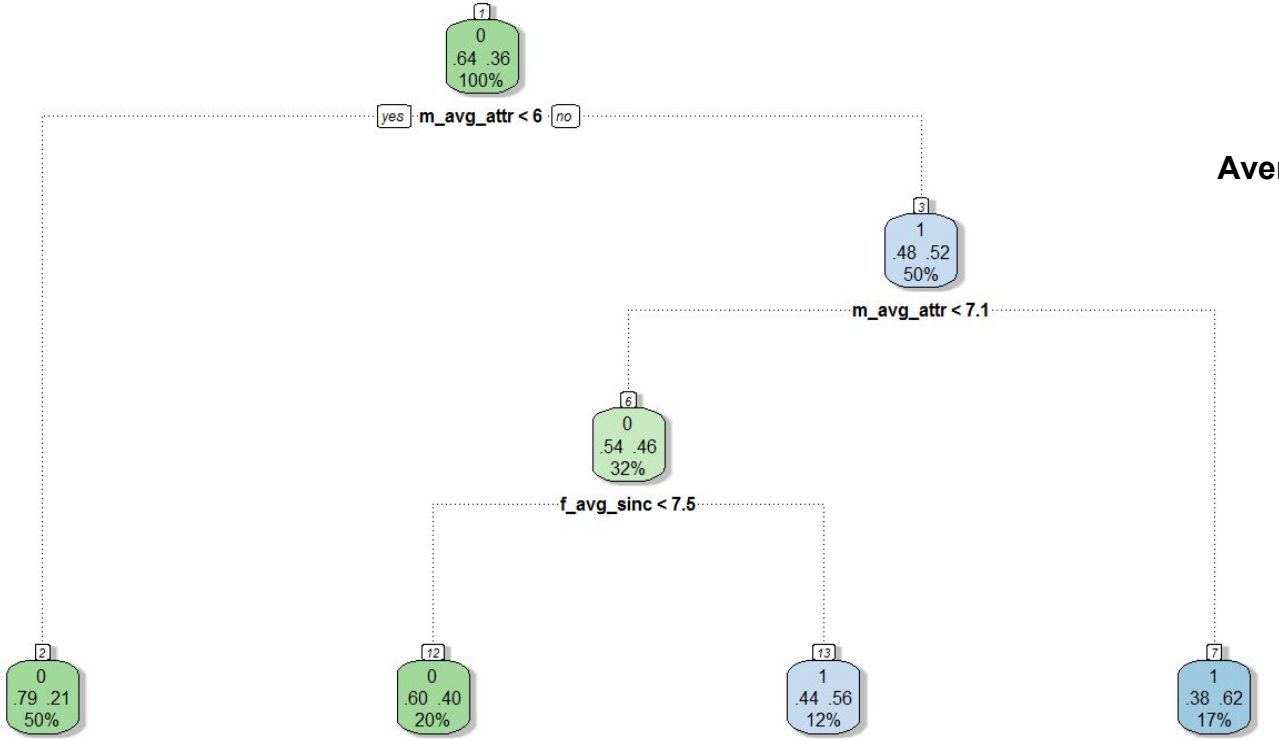
Pros

- Interpretability
- Ability to customize complexity

Cons

- Higher error rates than other models
- Error rates do not reliably improve with increased complexity compared to base error rates

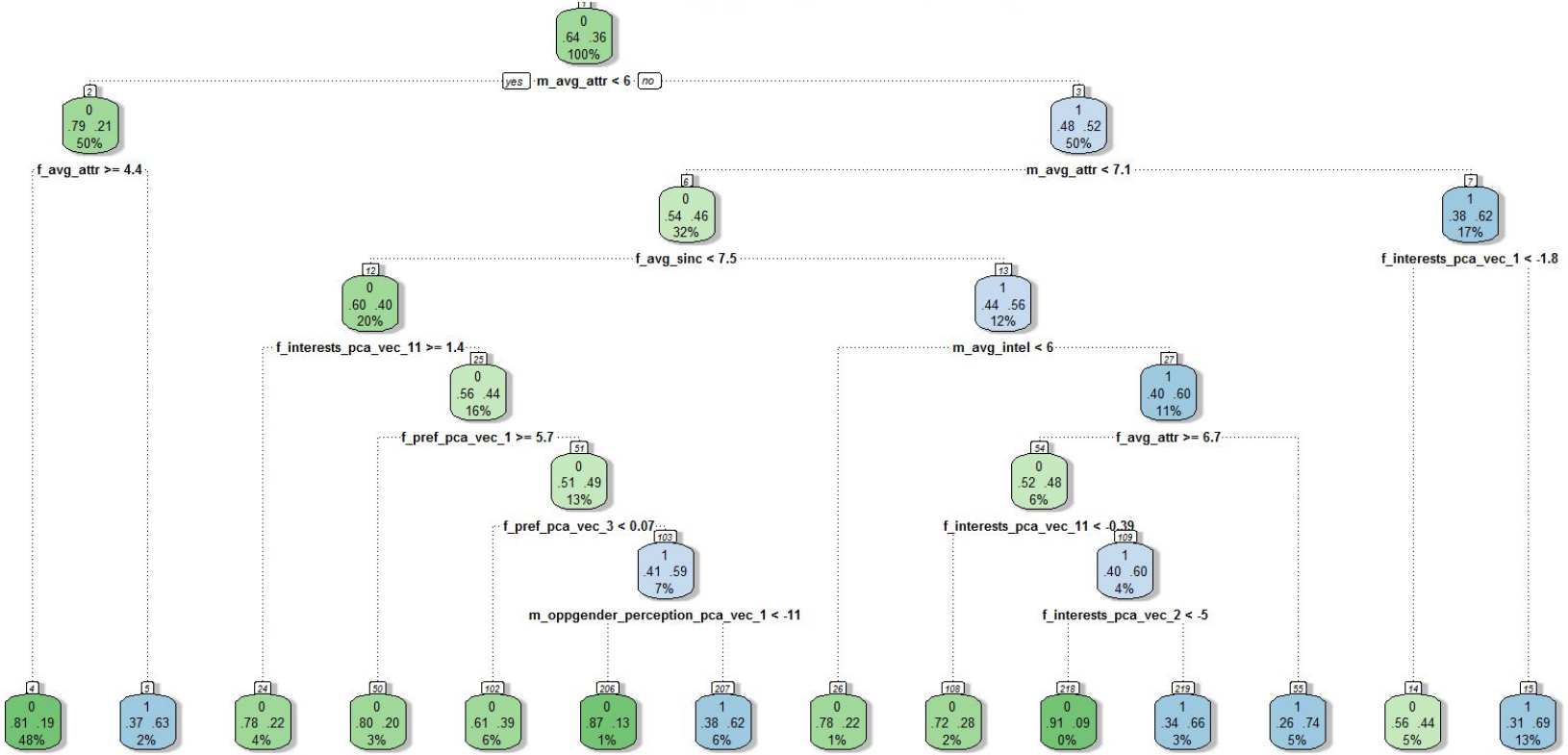
Female Decision - Simple



Complexity: .025
Overall error: 35%
Averaged class error: 39%

Interpretation: Generally, if a male's attractiveness is under 6, females will say no. If the male is over 6 but under 7.1, then females with sincerity ratings over 7.5 will say yes.

Female Decision - Complex



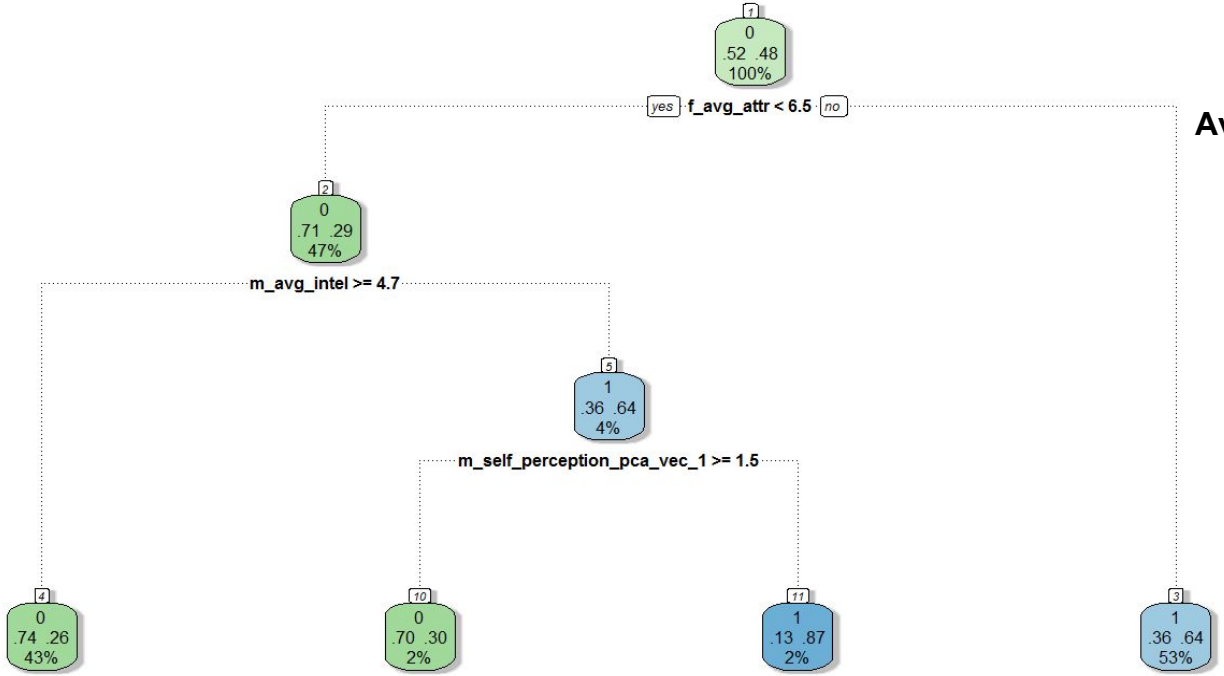
Complexity: .008 **Overall error:** 31%, **Averaged class error:** 36%

Female Decision - Complex

Insights

- Instead of being a terminal node, for males given an attractiveness rating less than 6, then females who are rated as less attractive are more likely to say yes
- For males with an attractiveness rating over 7.1, the female's decision varied based on their own interests
- Males whose attractiveness rating is between 6 and 7.1 the female's decision varied based on the female's average sincerity, the male's average intelligence, and other factors

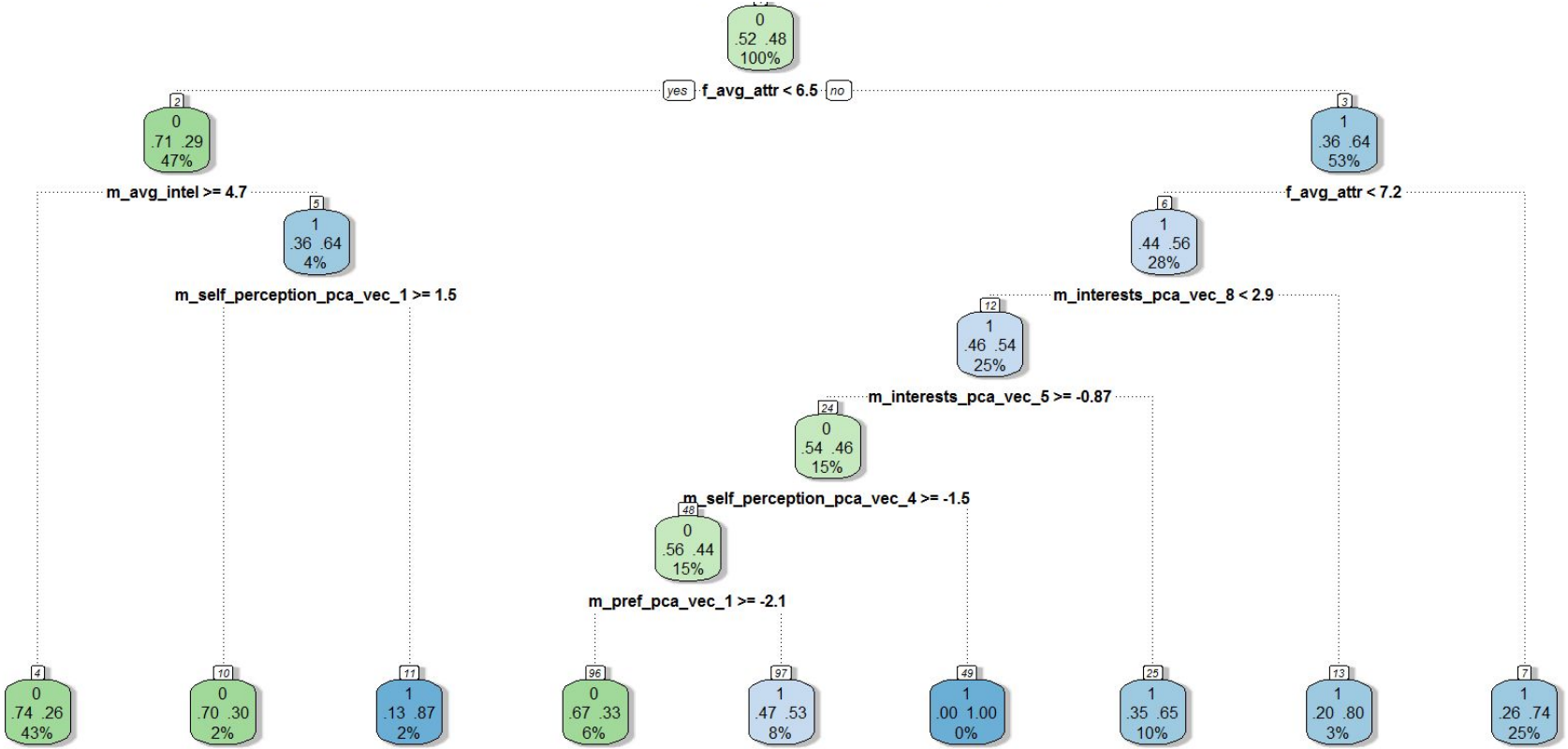
Male Decision - Simple



Complexity: .01
Overall error: 36%
Averaged class error: 36%

Interpretation: Generally, if a female's attractiveness is over 6.5, males will say yes. If it is under 6.5, males who were rated under 4.7 for intelligence and had a self-perception PCA score of under 1.5 will say yes.

Male Decision - Complex



Complexity: .008 **Overall error:** 37%, **Averaged class error:** 37%

Male Decision - Complex

Insights

- For females with attractiveness ratings between 6.5 and 7.2, the male decision varies based on his interests, self-perception, and preferences
- 68% of the decision can be calculated with 74% accuracy based on only two different variables: female attractiveness rating and male intelligence rating

Boosting Models

See AdaBoosting model in R script

Results : Male Decision

Training: 28% error

Test: 29% error

Type 1 Test: 25% error

Type 2 Test: 32% error

	Predicted No's	Predicted Yes's
Actual No's	239	81
Actual Yes's	96	201

Number of trees: 420 **Complexity:** .001 **Max Depth:** 30 **Min Split:** 20

Results : Female Decision

Training: 24% error

Test: 24% error

Type 1 Test: 14% error

Type 2 Test: 40% error

	Predicted No's	Predicted Yes's
Actual No's	331	56
Actual Yes's	92	138

Number of trees: 420 **Complexity:** .0005 **Max Depth:** 30 **Min Split:** 20

Random Forest Models

See Random Forest model in R script

Results : Male Decision

Training: 26% error

Test: 29% error

Type 1 Test: 27% error

Type 2 Test: 32% error

	Predicted No's	Predicted Yes's
Actual No's	235	85
Actual Yes's	96	201

Number of trees: 800 **Number of Variables (mtry):** 30

Results : Female Decision

Training: 25% error

Test: 24% error

Type 1 Test: 15% error

Type 2 Test: 39% error

	Predicted No's	Predicted Yes's
Actual No's	328	59
Actual Yes's	89	141

Number of trees: 420 **Number of Variables (mtry):** 30

Assessment for Boosting & Random Forest Models

Pros

- Fairly row error rates for all types of error

Cons

- Slower than other types
- Less useful for interpretation (compared to standard Decision Trees)

SVM Models

See SVM model in R script

Results : Male Decision

Training: 29% error

Test: 31% error

Type 1 Test: 26% error

Type 2 Test: 37% error

	Predicted No's	Predicted Yes's
Actual No's	238	82
Actual Yes's	109	188

Kernel: Radial **Cost:** 1

Results : Female Decision

Training: 28% error

Test: 29% error

Type 1 Test: 25% error

Type 2 Test: 32% error

	Predicted No's	Predicted Yes's
Actual No's	239	81
Actual Yes's	96	201

Kernel: Radial **Cost:** 1

Assessment for SVM Models

Pros

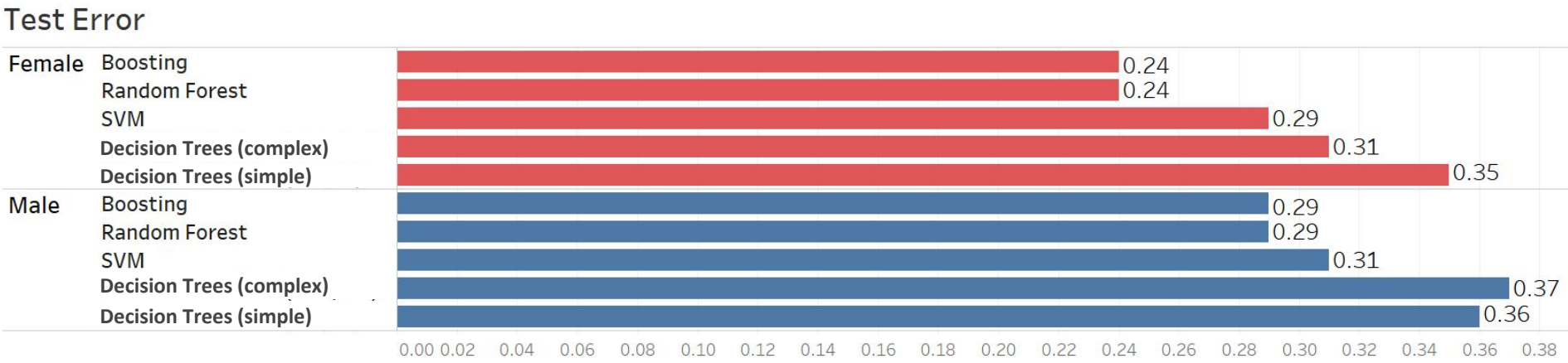
- Low type 2 error for females
- Lower error rates than standard classification trees

Cons

- Higher error rates than Boosting and Random Forest Models
- Less useful for interpretation (compared to Decision Trees)

Model Comparison & Discussion

Test Error



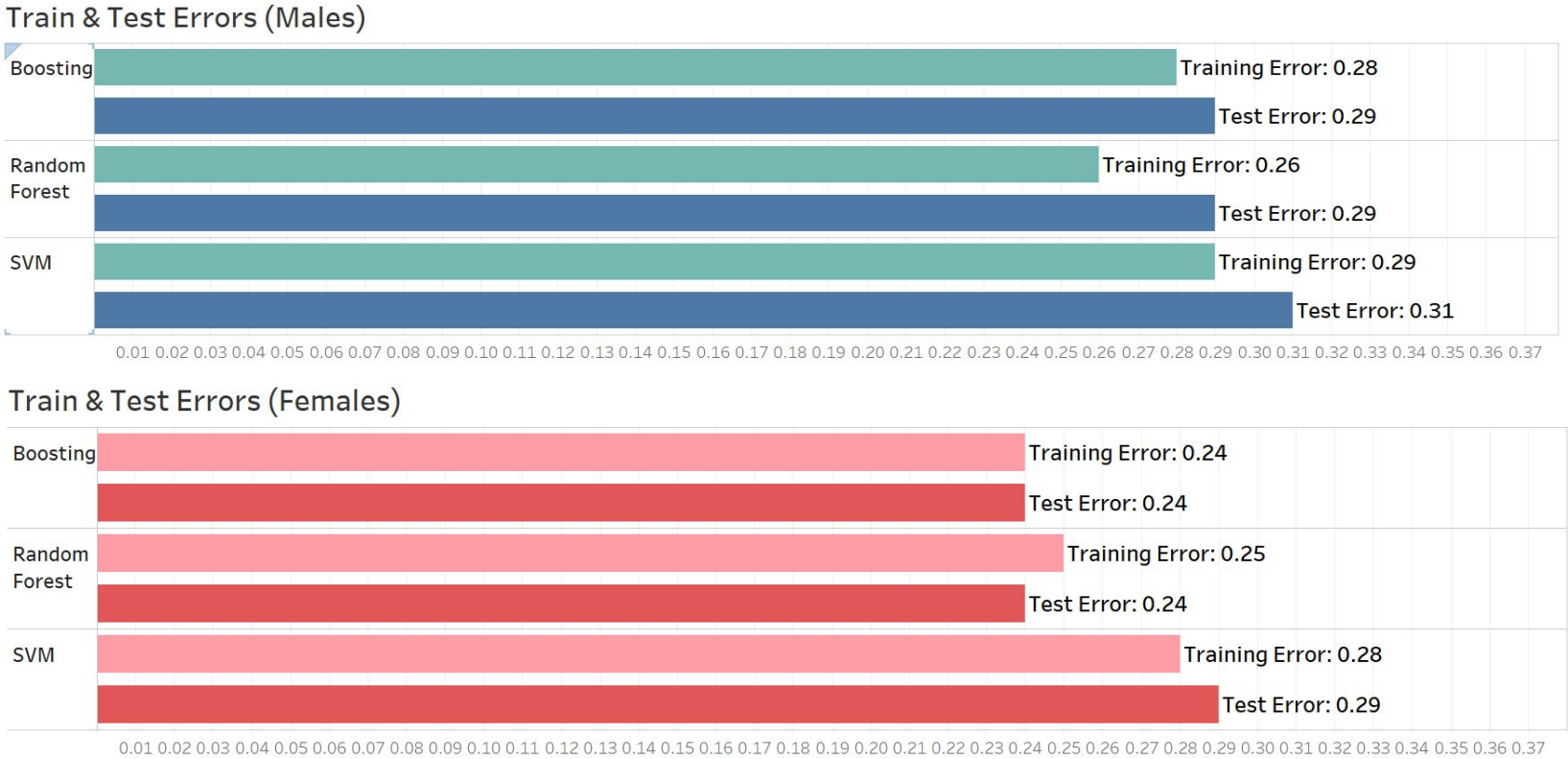
Insights:

Classification (decision) trees are the weakest in terms of accuracy.

Boosting and Random Forest models performed equally well.

Lower error rates overall for females.

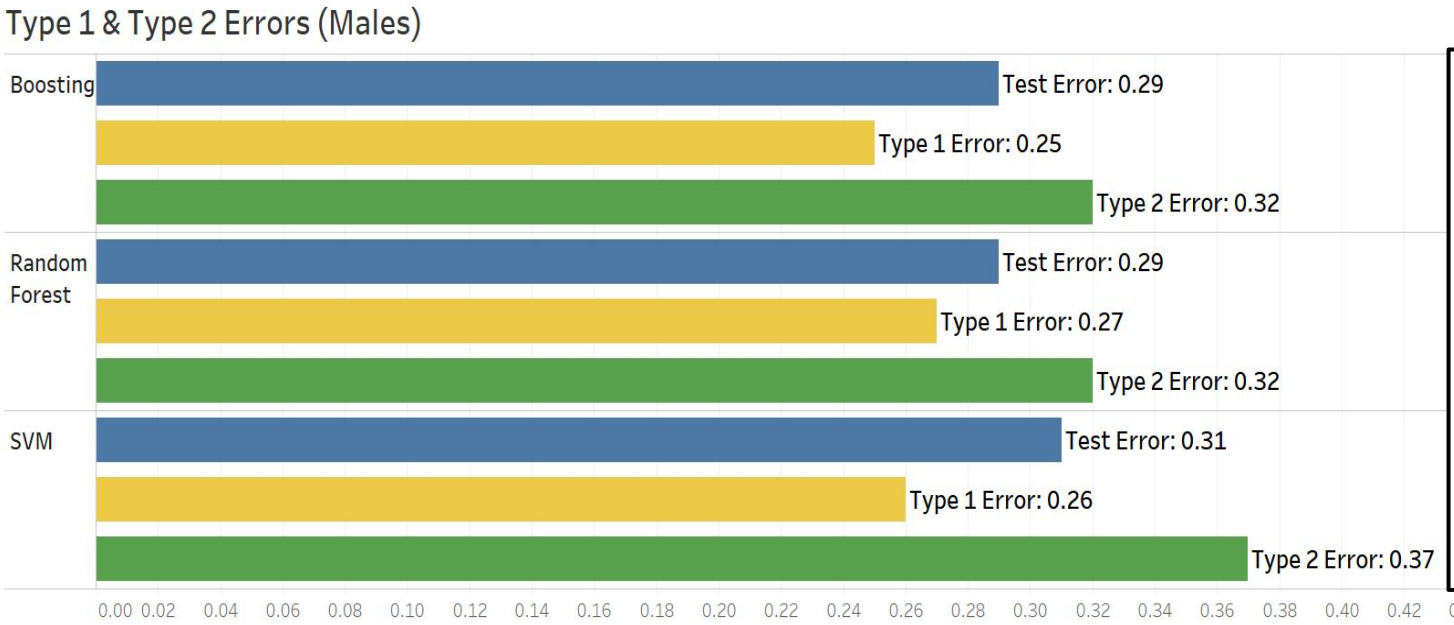
Training & Test Errors



Insights:

Test error rates are similar to training error rates, suggesting that models were not overtrained.

Type 1 & Type 2 Errors (Males)



Type 1: Predicted interest in partner, when the person was **not** interested.

Type 2: Predicted not interested in partner, when the person **was** interested

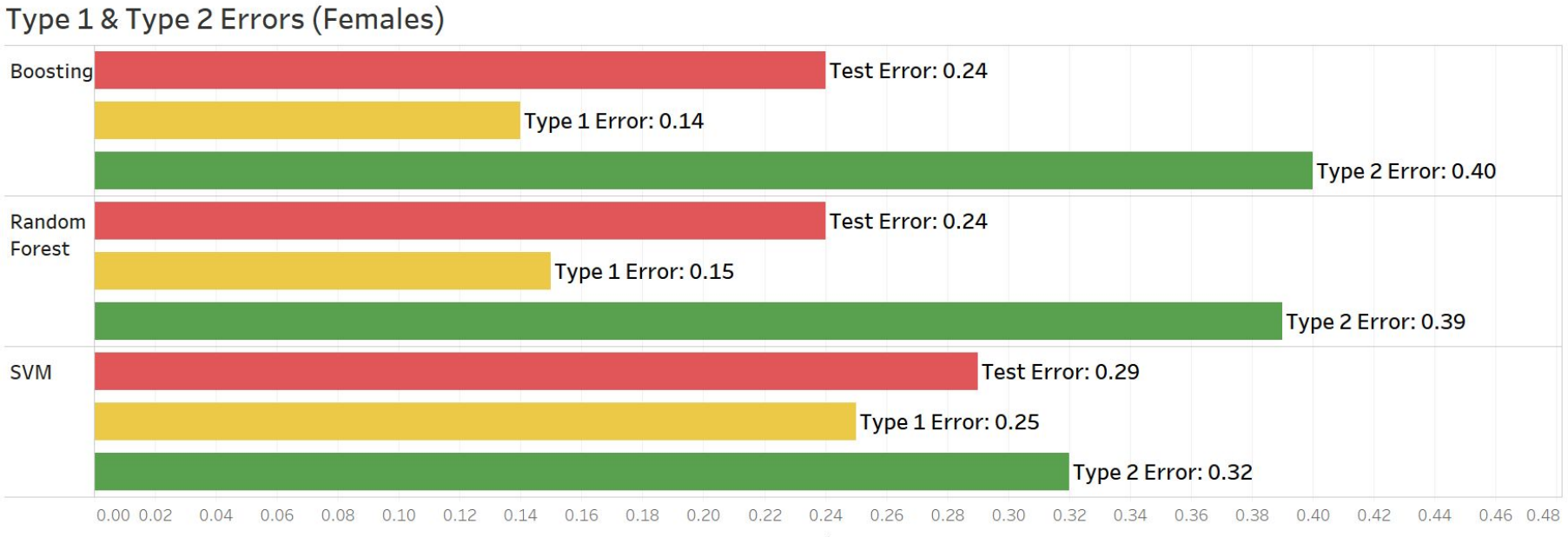
Insights:

For all models, type 2 errors were greater than type 1.

Boosting and Random Forest had the same type 2 error, but Boosting had lower type 1.

SVM performed worst for both error rates.

Type 1 & Type 2 Errors (Females)



Insights:

As with males, type 2 error was greater than type 2 for all models.

SVM had the lowest type 2 error, despite having the highest type 1 and overall error.

Discussion

Type 1 vs Type 2 Errors and Gender

- Type 1 error rates were lower for females, probably because females said yes less often (36.5%), whereas males said yes more often (47.5%).
- Therefore, female yes's were the hardest to predict - which aligns with the idea that the biggest challenge for dating services (and men in general) is finding interested females

Discussion

Type 1 vs Type 2 Errors: Which is worse?

- Depends on format of dating service and customer expectations:
  - **Tinder**: users are shown many other nearby users in quick succession, and choose to show interest or not. Keeping **type 2 error** low is more important to prevent users from missing out on potential matches.
  - **Coffee Meets Bagel**: users are shown a small number of potential matches each day, based on user data. Keeping **type 1 error** low is important, because users expect high quality over quantity.

Discussion

Considerations for Choosing a Model

- **Type of dating service:** as stated previously, some may focus on minimizing type 1 error, others type 2 error
- **Gender:** different models should be used for males and females
- **Our recommendations:**
  - For predicting **female interest**, we recommend using the **SVM model**. Despite a slightly lower overall error rate, its low type 2 error rates make it preferable to the other two.
  - For predicting **male interest**, we recommend using the **Boosting model**, as its overall error rate was identical to the Random Forest model, and type 1 error was very slightly lower.
  - To determine the **most important information to collect from users**, we recommend using **Classification Trees**, due to their high level of interpretability.

Discussion

Future Improvements

- Strategically identifying:
  - New questions to ask users in order to better predict matches.
  - Questions that are unnecessary to ask (questions that do not yield valuable predictors).
  - Ways to collect data without burdening users with questions (analyzing their behavior on the site, how long they spend viewing profiles, etc.)
- Explore other types of models
- Experiment with using different models for different age groups

Thank You!