Love is Blind

Group 9: Team Speed Date
Eustace Ednacot, Cindey Caine, Evan Azcuna, Momin Asadullah Khan

UC San Diego

Background

- This data comes from a study performed at Columbia Business School in the time between October 2002 and April 2004.
- Participants were paired across gender with all of their answers being confidential except when two people both agreed to match with each other, in which case they were informed.
- Data includes survey answers before, during, and after the experiments and includes:
  - demographic information
  - lifestyle information
  - lifestyle information
  - dating habits
  - self-rating across 5 attributes
  - ratings of 5 attributes of a subject's desired partner
  - ratings of 5 attributes in which a subject believes the opposite gender would rate theirs
- Perception of attributes in oneself and others is a key measurement used in this data set, with the attributes categorized and defined as follows:
  - Attractiveness: physical attractiveness
  - Sincerity: honesty in character
  - Intelligence: general intelligence and knowledge
  - Fun: pursuit of entertainment
  - Ambition: drive to succeed and accomplish things; and lastly
  - Shared Interests: the importance of shared interest among pairings

Research Question

- Using this dataset, we want to try and find a relationship between a general appeal factor versus the number of matches a person will receive and model the findings.
- Also, we want to see if there is a difference between how attraction is perceived in females versus males. What one gender is looking for versus what the opposite gender thinks the first gender is looking for in a partner.

Methodology

Data Visualization:

- Plotted several of the features against each other in an attempt to recognize any obvious trends.

Data Pre-Processing:

- Added an additional dimension representing the number of matches that a subject made. This was absent in the initial dataset.
- Separated data into subsets representing male and female participants.
- Removed data points of participants that did not answer relevant questions (some data points lacked data for specific features).
- Performed **PCA** on the five attributes, Attractiveness, Sincerity, Intelligence, Fun, and Ambition to reduce to a single dimension interpreted as the general appeal factor across three different perspectives asked about in the survey.
  - How would you rate your own attributes?
  - How would you rate the attributes of a potential partner?
  - How do you think a member of the opposite gender would rate the attributes of a member of your gender?
- Split the results of a survey question into categories of 'Active' Activities and 'Inactive' Activities where the participants rated their enjoyment of an activity.
- Performed PCA across the features in each activity group to reduce them into two singel-dimension3factors describing active interests (involved and participatory) and inactive interests (more passive and speculative).

Data Analysis

- Linear Regression was used to model a relationship between number of matches a subject makes and the appeal factor of a specific perspective.
- Sum Squared Error was then calculated to see which model was the best fit.
- The active and inactive factors were plotted against the appeal factor determined from a partner's rating of a subject to investigate data density and spread.

Results

Fig. 1

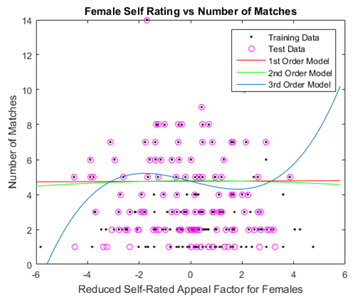


Fig. 2

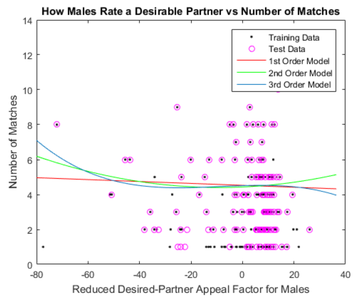


Fig. 3

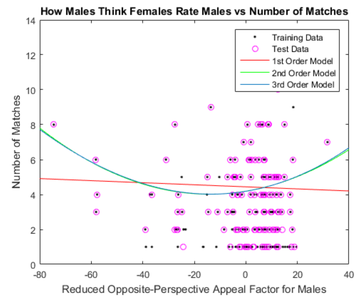


Fig. 4

Fig. 5

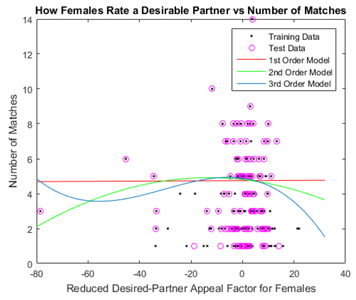


Fig. 6

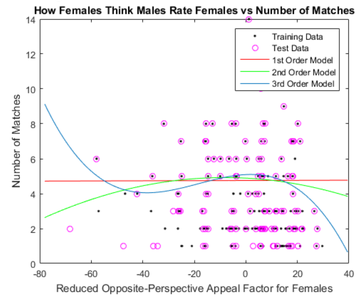


Fig. 1, Fig. 5, and Fig. 6:
The SSE for the 3rd order model was the smallest value. However, the curved tails are caused by outliers and may not be good predictors of model output.

Fig. 2, Fig. 3, and Fig. 4:
The training set SSE for the 3rd model is the smallest value. However, for Fig. 4, all SSEs for each model are very close. Therefore, for Fig. 4, any of the models can be used as predictors. For Fig. 2 and Fig. 3, the third order model over-fits the data due to the curved tails. Therefore, the 2nd order model should be analyzed instead.

Fig. 7

Sum Squared Error for the Training Set						
Gender	Female			Male		
Appeal Factor Type	Self-Rating	Desired-Partner	Opposite-Perspective	Self-Rating	Desired-Partner	Opposite-Perspective
1st Order	3743.9004	3698.2181	3844.2177	2636.3583	2939.004	2753.059
2nd Order	3743.2979	3684.4915	3827.3862	2635.9052	2928.7123	2662.4659
3rd Order	**3624.3918**	**3667.9243**	**3787.2126**	**2635.5843**	**2924.1269**	**2662.4095**

Fig. 8

Sum Squared Error for the Test Set						
Gender	Female			Male		
Appeal Factor Type	Self-Rating	Desired-Partner	Opposite-Perspective	Self-Rating	Desired-Partner	Opposite-Perspective
1st Order	1502.3331	1556.6508	1411.2612	**1128.1129**	1099.7206	1303.9404
2nd Order	1506.8472	**1553.5385**	**1388.5663**	1129.341	1089.2386	1279.5148
3rd Order	**1435.0738**	1558.149	1406.2011	1129.2074	**1067.2195**	**1280.0918**

Fig. 9

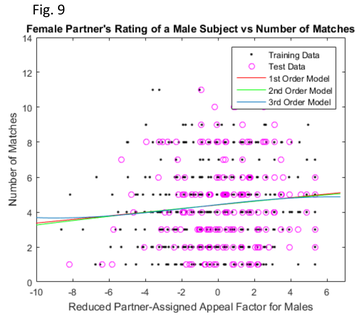


Fig. 10

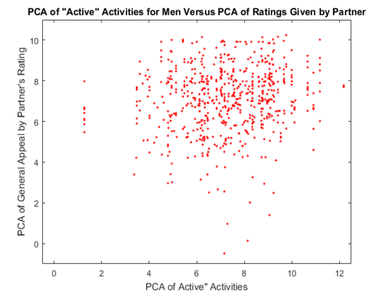


Fig. 11

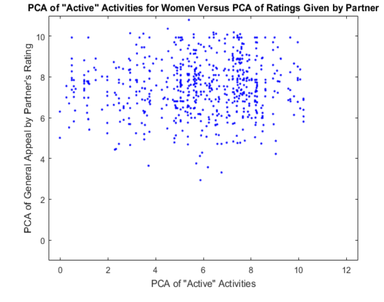


Fig. 12

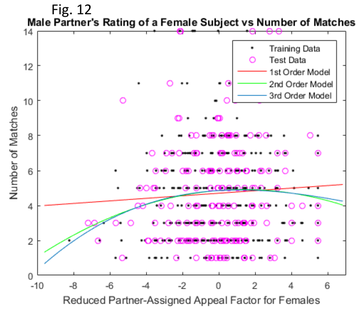


Fig. 13

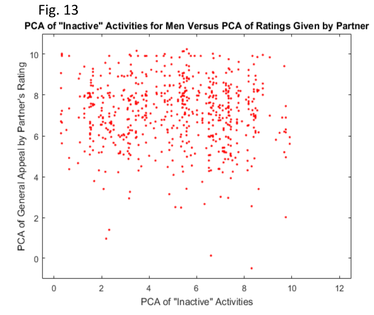
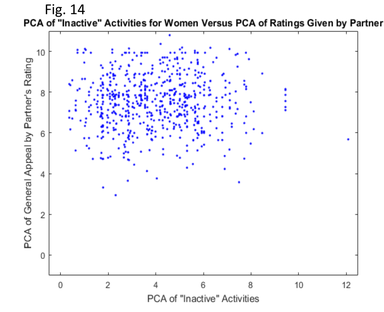


Fig. 14



Interpretation

- Fig. 2 versus Fig. 6: Based on the similar shapes of these two models, it can be inferred that females have a generally good understanding of how a male would rate a female.
- Fig. 1, 2, 3, 5, 6: Generally, no matter whether the appeal factor across all three perceptions increase or decrease, there is little to no influence on the predicted number of matches a subject makes.
- Fig. 4: There is a very noticeable positive correlation between how a male subject rates themselves and how many matches he makes. Generally, the higher that a male subject rates themselves, the higher the number of matches he makes. This can be interpreted as a simple display of self-confidence.
- Fig. 10 and Fig. 13: Among males, the data shows more variance in levels of interest with matches that valued inactive activities. The denser active activities leaned towards the higher values, indicating that males that valued active activities were more successful in getting a match.
- Fig. 11 and Fig. 14: Females seemed to follow a similar trend with interests in active activities, however inactive activities skews to the lower end, suggesting that women who valued inactive activities less made up the bulk of women who got matches.
- Fig. 9 and Fig. 11: Female Subjects who were given a lower rating by a Male Partner were able to find more matches, as compared to a Male Subject who were given a lower rating by a Female Partner.

Conclusion

- Generally, the appeal factor across all three perspectives did not have as heavy an influence on the number of matches a subject made.
- For future analysis, we would like to further process the initial data set so that we can find the mean ratings of each person given by their partner. Also finding the overall mean of each factor.
- We would like to train and utilize a perceptron to model and predict whether a match happens or not based on high dimensional input.
- Rather than use PCA, we may also take all of the key attributes and active/inactive activities and cluster them into groups of similar personalities and interests.
  - We can run the perceptron on each of these clusters to determine which group would be most successful in matching.

References

https://www.kaggle.com/annavictoria/speed-dating-experiment