Speed dating – partner compatibility analysis

Ivana Savin
University of Novi Sad, Faculty of Technical Sciences
Software engineering and information technologies
Novi Sad, Serbia
ivana.unitedforce@gmail.com

***Apstract*—The development of artificial intelligence has led to machine learning becoming intertwined with various branches of science, including humanistic and social studies. The idea of this paper is to try to analyze the case of speed dating using methods of logistic regression in order to examine whether it is possible to forecast the results of human interaction and predict possible relationships. The data set used is one collected at the Columbia Business School which includes various attributes significant for the analysis.**

Keywords — speed dating; machine learning; logistic regression; backward elimination; feature scaling

I. INTRODUCTION

This analysis is exploring the topic of speed dating. It aims to provide a basic understanding of people's behaviour in such events, including attributes most relevant for successful dating.

Speed dating is a formalized matchmaking process of dating system whose purpose is to encourage people to meet a large number of new people. Scholars have recently begun to harness the immense power of speed dating procedures to achieve important and novel insights into the dynamics of romantic attraction. Speed dating procedures allow researchers to study romantic dynamics dyadically, i.e. potentially meaningful relationships [4].

Chapter II describes the used data set and all of the significant attributes, chapter III uses graphic representation of the data for easier trend observation and analysis, chapter IV describes the logistic regression method and chapter V states the achieved results and also the conclusion of this paper, including ideas for further research.

II. DATA SET

The data set explored in this project is named „Speed Dating Experiment", as found on Kaggle.com. It was collected by professors Ray Fisman and Sheena Lyengar from Columbia Business School, originally used for their paper „Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment". It was generated from a series of experimental speed dating events from 2002 to 2004 and includes data related to demographics, dating habits, lifestyle information and an attribute evaluation questionnaire [5].

The data set includes 8379 rows and 195 columns, where each row represents an evaluation questionnaire for a specific partner and each column represents a different attribute. Table 1. Significant attributes displays the more significant attributes as well as those used in the regression algorithm.

Table 1. Significant attributes

Attribute name	Description
iid	Unique group number
id	Subject number within group
pid	Partner's iid number
partner	Partner's id number
attr	Grade for specific partner's attractiveness
sinc	Grade for specific partner's sincerity
intel	Grade for specific partner's intelligence
fun	Grade for specific partner's fun aspect
amb	Grade for specific partner's ambition
shar	Grade for shared interests with specific partner
attr1_1	Importance of attractiveness in potential date
sinc1_1	Importance of sincerity in potential date
intel1_1	Importance of intelligence in potential date
fun1_1	Importance of fun aspect in potential date
amb1_1	Importance of ambition in potential date
shar1_1	Importance of shared interests with potential date
match	Binary result of date
gender	Binary value of gender

A. *Data set cleaning*

Taking into account that the data set has incomplete rows in terms of contained attributes, some cleaning must be applied.

1) Inconclusive data

There are missing values in approximately one-third of the total data set, those values are considered inconclusive and are removed from the data set, resulting in a data set with 5850 rows.

2) *Normalizing attributes*

The *attr*, *sinc*, *intel*, *fun*, *amb* and *shar* attributes are all in the range of 0-10, based on their importance. But for the *attr1_1*, *sinc1_1*, *intel1_1*, *fun1_1*, *amb1_1* and *shar1_1* attributes a sum of 100 is divided between them all, also signifying importance.

Features scaling is very important in machine learning. It is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, including Logistic regression, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance [3].

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

The result of standardization (or Z-score normalization) is that the features will be rescaled so that they'll have the properties of a standard normal distribution with

$$\mu = 0 \text{ and } \sigma = 1$$

where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

III. Observations

Graphically representing the data leads to noticing trends clearer. In the following chapter several groups of data will be displayed: what participants are looking for in their matches, what participants think their same-sex colleagues are looking for, as well as what participants think the opposite sex is looking for.

A remark is that the charts in this chapter are constructed based on a survey conducted before the actual speed dating event.

A. What participants are looking for in their matches

Firstly, the examination whether there exists a difference for male and female participants. At this point in time, the participants have just signed up for the event and have not met anyone, as seen on Figure 1.

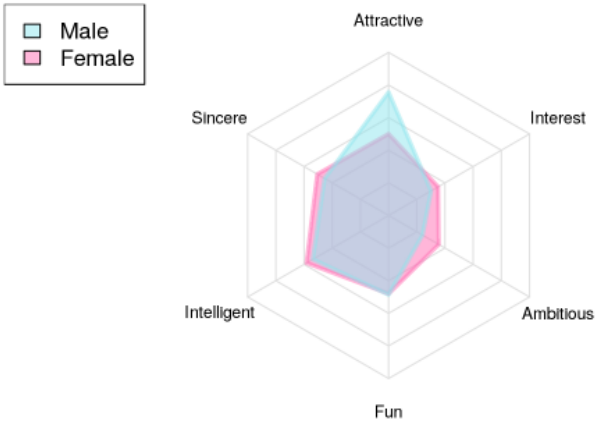

Figure 1. Radar chart of attributes participants are looking for in their matches

We can see that there is a great difference between what male and female participants are looking for:

- For male participants, the attractiveness of the female is given a lot more weight, and the ambitiousness or if they have any shared interest are not ranked as high
- For females, the points are more evenly ditributed across all of the attributes, with intelligence ranked slightly higher compared to others

Conclusion

Men are looking for attractive women, and are less concerned with a woman's ambition and shared interests. Women are, on the other hand, looking for a well-rounded, intelligent male.

B. What participants think their same-sex colleagues are looking for in a partner

Next, the examination of what people think men/women of their same sex are looking for, determening if they separate their own views from the majority, as seen on Figure 2.

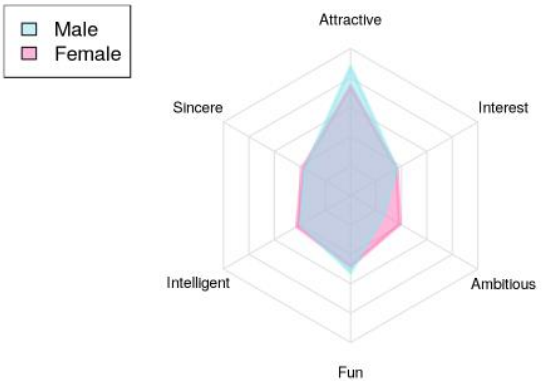

Figure 2. Radar chart of attributes participants think their same-sex colleagues are looking for in a partner

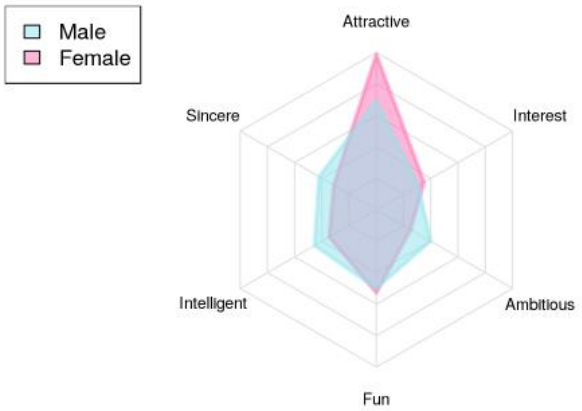

Figure 3. Radar chart of attributes participants think their opposite sex is looking for in a partner

Analyzing the chart, both men and women think people of their same gender are most concerned with finding an attractive partner:

- As in the previous analysis, men think their fellow males highly value attractiveness and are less concerned with a woman's ambition
- As for women, there is a significant difference in comparison to the previous analysis. Women say that they themselves are looking for a well-rounded man, however they think that other women are mainly looking for attractive men

Conclusion

What women say they are looking for is drastically different from what they think other women value.

C. What participants think the opposite sex is looking for

Finally, the examination of what participants think their opposite sex is looking for. We can see on Figure 3 if there is any difference in the expectations of men and women.

Analyzing the chart, we can see that:

- Women strongly feel that men are most concerned with a woman's attractiveness and that other attributes are not as important
- Comparing female and male answers in the first graph, we can see that there are not that much difference between the two. However, women almost accurately predicted what men are looking for in their partners, i.e. attractiveness
- Nevertheless, men's predictions were not far off either. What men think women are looking for also closely resembles what women say they are looking fore. The main differences, though, are a higher attractiveness score and a lower shared interest score.

Conclusion

Both men and women can predict what the opposite sex are looking for in their partners to a certain degree.

IV. LOGISTIC REGRESSION

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes) [1].

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, etc.) or 0 (FALSE, failure, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

In the case of this project, the dependent variable is *match* and the independent variables are *attr, sinc, intel, fun, amb, shar, attr1_1, sinc1_1, intel1_1, fun1_1, amb1_1, shar1_1*.

Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Among these independent variables there are some that are highly statistically significant, that have great impact or great effect on the dependent variable and some that are not significant at all. So the goal is to find a set of optimal independent variables so that each independent variable has great impact on the dependent variable. To achieve this, backward elimination algorithm will be used.

A. Backward elimination

Backward elimination is an approach which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.

1. Select a significance level to stay in the model (e.g. SL = 0.05)
2. Fit the full model with all possible predictors
3. Consider the predictor with the **highest** P-value. If P > SL, go to step 4, otherwise go to END
4. Remove the predictor
5. Fit model without this variable

END. The model is ready

Figure 4. Backward elimination algorithm

V. RESULTS AND CONCLUSION

The dataset is divided to train and test sets, where the test set contains 25% of the whole dataset, which is 1463 rows. Using logistic regression and forecasting using its results, an accuracy of 85.3% has been achieved, which shows that it is feasible to predict the outcome based on the Speed Dating data set. One possible way to increase the accuracy can be combining different models after carefully studying their own performance.

Further research would include implementing other classification methods such as Support Vector Machine (SVM), Random Forest and K-Nearest Neighbors (k-NN), PCA for dimensionality reduction. It is important to compare the performance of multiple different machine learning algorithms consistently, in order to choose the one that gives the best result.

REFERENCES

[1] Menard, Scott. *Applied logistic regression analysis*. Vol. 106. Sage, 2002.

[2] Sutter, Jon M., and John H. Kalivas. "Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection." Microchemical journal 47.1-2 (1993): 60-66.Hart, P.E., N.J. Nilsson, and B. Raphael, „A formal basis for the heuristic determination of minimum cost paths“, *IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-4*, No. 2, July 1968, pp. 100-107.

[3] Blum, Avrim L., and Pat Langley. "Selection of relevant features and examples in machine learning." Artificial intelligence 97.1 (1997): 245-271.

[4] Finkel, Eli J., and Paul W. Eastwick. "Speed-dating." Current Directions in Psychological Science 17.3 (2008): 193-197.

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