INFS4203/7203 Data Mining
Speed Dating Experiment

Andi Nuruljihad 44159069
Dongnan Nie 44028053
Thanat Chokwijitkul 44522328
Xuanming Fu 43360774
Yong Jin Yeo 44063326

The University of Queensland
Brisbane, Australia

October 24, 2016

Contents

1 Background and Problem Definition

Speed dating experiment, a matchmaking process of dating system, encourages participants to meet a large number of potential partners of the opposite gender they are not familiar with and helps them get matched in a quick pace. Unmarried individuals always wonder how they can become successful in such dating event. Comparing several hyperbolic ways to prepare, one of the best methods may be to inform people what kind of human qualities are the most alluring and highly contribute to the possibility of getting matched in the speed dating environment. Therefore, the main aim of this project is using a large dataset to construct a classification model and analyse what is the most desirable attribute among all the personal attributes.

2 Objective

The primary purpose of this project is to compare six key attributes, including attractiveness, sincerity, intelligence, fun, ambition and shared interests rated by each of the participant's partners. This project will attempt to produce a classification model to determine whether the participants involving in the experiment would like to see their partners again in the future.

3 Dataset

The dataset used in this project was compiled by Columbia Business School Professors Ray Fisman and Sheena Iyengar for their paper *Gender Differences in Mate Selection: Evidence from a Speeding Dating Experiment* (Fisman, Iyengar, Kamenica, & Simonson, 2006).

3.1 Data Collection

All the data in the dataset was gathered from experimental speed dating events from 2002 to 2004. During the experiment, all the participants would have a four-minute dating experience with other attendees of the opposite gender and gave the feedback on their attributes and the likelihood whether they would like to see their date again.

3.2 Data Description

The input variables are basically personal attributes variables including attractiveness, sincerity, intelligence, fun, ambition and shared interests, which were chosen as the key variables. Other input variables such as basic personal information which the project did not rely on. For output variables, the most essential one is the likelihood whether attendees want to have a date again with their partners. Data attributes used in this project can be found in Table 1.

Table 1: Data attributes and descriptions

Attribute	Description
gender	0 = female, 1 = male
match	1 = yes, 0 = no
samerace	participant and partner are the same race, 1 = yes, 0 = no
age_o	age of partner
race_o	race of partner

Attribute	Description
pf_o_att	partner's stated preference for attractiveness
pf_o_sin	partner's stated preference for sincerity
pf_o_int	partner's stated preference for intelligence
pf_o_fun	partner's stated preference for fun
pf_o_amb	partner's stated preference for ambition
pf_o_sha	partner's stated preference for shared interests
dec_o	decision of partner the night of event
attr_o	rating by partner the night of the event for attractiveness
sinc_o	rating by partner the night of the event for sincerity
intel_o	rating by partner the night of the event for intelligence
fun_o	rating by partner the night of the event for fun
amb_o	rating by partner the night of the event for ambition
shar_o	rating by partner the night of the event for shared interests

3.3 Data Preparation

Originally, the dataset consists of 195 columns and 8,378 rows. At the start of the data cleaning process, almost 90 columns were eliminated since they represent irrelevant information regarding the participants' decisions. Due to the problem definition set before, this project only concentrates on the information which affects their decisions. Therefore, 6,378 rows were chosen as the training data and the rest of it as testing data.

Then several missing values were found in some rows. There is a large number of methods that can solve the issue of missing value, such as ignoring missing value, filling in with particular values and mean imputation.

Simply ignoring missing value does not have any advantage. By leading losing a large section of the original data, this is the first method which was negated, since deleting all the rows containing missing values means losing almost 1/10 portion of the dataset.

Since missing values cannot be ignored, the next possible approach is to modify the dataset by filling in special values. These special values can be represented as 0 or unknown, which also can lead to a situation of data deviation. In addition, suppose it is possible to use these special values, it may be considered that the special value is a new value of attributes, which make the model or regular pattern useless when the dataset is analysed by a data mining algorithm.

The missing data issue in this dataset was caused by skipping pattern in the survey, which was collected randomly. In fact, during the survey, most of the respondents in service were less likely to be missing. When using data analysis tool called IBM SPSS to analyse the missing value in the dataset, it yielded the result that the missing is completely at random, which means the type of missing value is MCAR. However, the test for MCAR could be not entirely accurate. It still gave an implication on how to deal with the missing value in a more reliable way (Humphries, 2010).

As the missing value was already known as a type of MCAR, one of the most common methods is the Expectation Maximisation (EM) algorithm, which is a way of maximising the latter iteratively and alternates between two steps including E and M steps , and use expectation max-

imisation filling in missing value. The E step calculates the expected complete data log-likelihood ratio, and then the M step maximises the ratio. After an E step and subsequent M step, the likelihood function will never decrease. Therefore, each column containing missing values were filled with the expectation maximisation values (Lauritzen, 2006).

4 Methodology

4.1 Expectation Maximisation (EM)

At the start of the project, the team decided to use C4.5 to analyse the dataset. However, there is too much noise in the dataset (i.e. missing data in attributes), resulting in the algorithm not performing well. Hence the first step was to reduce the noise in the dataset.

The dataset consists of a total of 195 columns. To narrow down the attributes that are useful for the project, only six key attributes were selected, including attractiveness, sincerity, intelligence, fun, ambition and shared interests as key attributes.

After choosing the attributes to be used for the project, the team found that there were some missing values in the dataset. One of the solutions is to utilise the Expectation Maximisation (EM) algorithm in filling up the missing data (Lauritzen, 2006). Suppose using special values like 'unknown' to treat missing attribute values, it may lead to a situation of data deviation. The easiest way is by using mean imputation to replace each missing value with the mean of the observed values. Unfortunately, this can severely distort the distribution for this variable. Therefore, as mentioned in the data preparation section, EM was the most suitable algorithm for filling up the missing data in the attributes.

4.2 Recursive Partitioning and Regression Tree (RPART)

Classification and Regression Tree (CART) in Recursive Partitioning and Regression Tree (RPART) package is the implementation of the idea found in the CART book and programs of Breiman, Friedman, Stone, and Olshen (1984). The main use of this algorithm is to create a decision tree to classify a member of the population by splitting into subpopulations based on several independent variables.

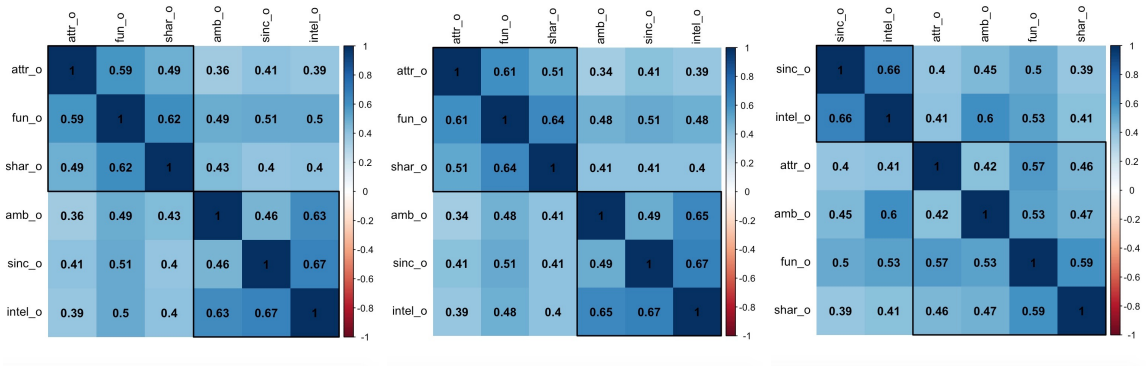
4.3 Random Forest

Random Forest is an ensemble learning method for classification and regression where a substantial number of decision trees are constructed. Every observation is fed into each of the decision trees. The most common outcome for each observation is then used as the final output. Afterwards, the new observation is fed into all those trees and taking a majority vote for each classification model.

Random Forest was used in the project because it runs efficiently on a large dataset, it can handle thousands of inputs without variable deletion, includes an effective method for estimating missing data while maintaining the accuracy even when a large proportion of data are missing. In this case, the dataset contains over 8,000 records, and almost every record had a missing value in some of its attributes.

5 Results and Discussion

5.1 Attribute Correlation



(a) Attributes (b) Male attributes (c) Female attributes

Figure 1: Attribute correlation

The figures above illustrate the attribute correlation regardless the participants gender, the male attribute correlation (rated by female participants) and the female attribute correlation (rated by male participants). As the scale on the right indicates, darker shades of blue correspond to stronger correlation. Three things to highlight are:

1. The correlations are positive: higher ratings on one direction are always associated with higher ratings on all dimensions. Even the squares that might initially appear to be white are in fact very faintly blue. Therefore, ratings of a given type contain information about ratings of other types.

2. Consider the 6×6 square in the lower right, which corresponds to interrelations between attractiveness, fun, ambition, intelligence, share and sincerity. For each column, the square with the darkest shade of blue is the square on the diagonal. This corresponds to the fact that given a rating type R, the average rating thats most predictive of R is the average of R itself.

3. Consider the third matrix, which corresponds to individual mens decisions. The darkest shades of blue correspond to the average of other mens decisions on a woman, the average of their ratings of how much they like her overall primarily depend on ratings of her attractiveness.

5.2 Model Interpretation

The classification and regression tree generated showed that attractiveness was the dominant factor in determining the likelihood of someone matching with another individual. According to this model, if the physical attractiveness rated by a partner is below a threshold of 6.993, the classification result will suddenly fall into the class "no". However, physical attraction does not guarantee a match. With an attractiveness score above the 6.993 threshold, one would still need a "shared interests" score of 5.807 or greater in order to obtain the "yes" outcome. If attractiveness was rated 6.993 or higher, but the shared interest score was lower than 5.807, the correspond partner would have to give a "fun" rating of 6.953 or greater before the model will return a match. The results

showed that this model has an accuracy of 74.1% with a recall of 0.71 and precision of 0.68.A confusion matrix and ROC were also generated using a test sample size of 2,000 on this model to gauge its performance.

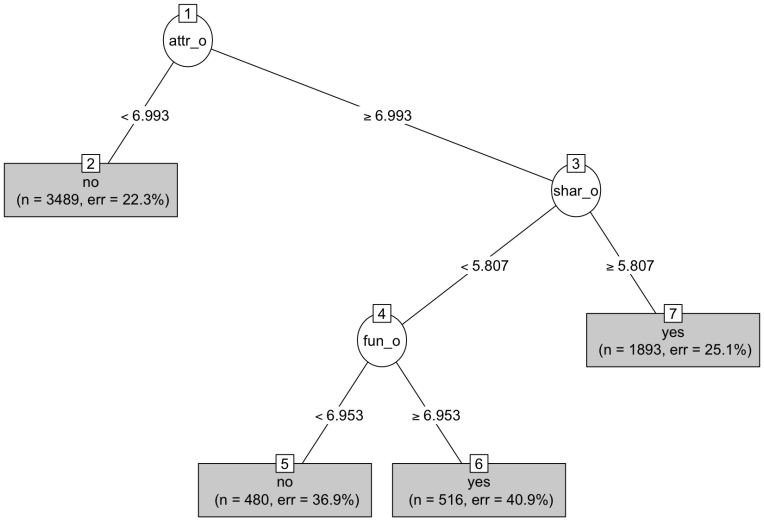
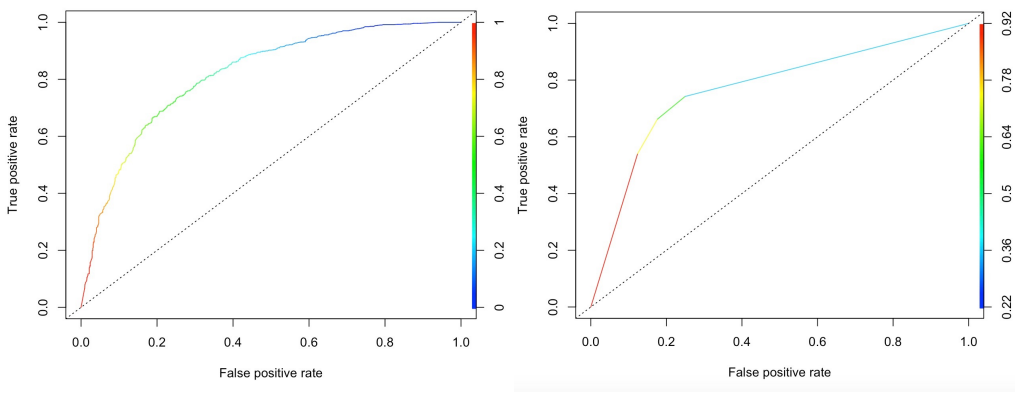


Figure 2: Classification tree generated by RPART

5.3 Model Performance



(a) Random Forest (b) RPART

Figure 3: Receiver Operating Characteristic (ROC)

In order to gain an in-depth understanding of these two models, the technique of receiver operating characteristic (ROC) curve has been introduced to evaluate the performance of both classifiers (Liaw & Wiener, 2002). In the ROC curve, perfect classification happens at 100% true positive rate and 0% false positive rate (Pencina, D'Agostino, & Vasan, 2008). Therefore, the perfect classification would indicate a curve at the upper left-hand corner of the graph, such that the closer our graph comes to that corner, the better we are at classification. The diagonal line depicts random guess so

that the distance of our graph over the diagonal line represents how much better we are predicting than a random guess. Our results indicate the similar outcome for both the RPART and Random Forest models. Notably, one significant difference lies in the shape of our graph. Since the Random Forest model consists of 500 trees, which provide a smoothed curve across the graph. Alternatively, with a single tree structure, the classification and regression tree tend to be composed of linearly distributed lines. In the case of RPART, the distribution of both true and false positive rate are skewed to the left due to the large size of the dataset. It is also evident that the predicted values of matched result are underestimated when the inflexion point are above the threshold of 0.4. Nevertheless, since Random Forest provides a smooth curve and avoid the presence of prominent inflexion point, the results from the Random Forest model are better than RPART in terms of reliability (Gromping, 2012).

5.4 Validation

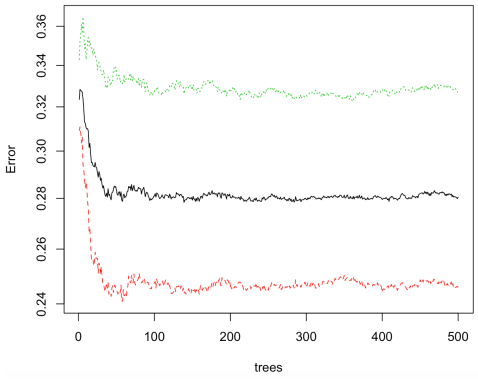


Figure 4: Random Forest error rate

With respect to the validity examination of the Random Forest model, there is no need for cross-validation or a separated test set to get an unbiased estimate of the test set error. It is estimated internally in the algorithms (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Each regression tree is constructed using a different bootstrap sample from the original data. Approximately one-third of the cases are left out of the bootstrap sample and not utilised in the construction of the tree. Thus, the method of out-of-bag error estimate has been employed to measure the internal prediction error.

5.5 Importance of Variables

In this study, Gini index has been employed to measure the relative importance of variables instead of Mean Squared Errors (MSEs) index. Gini importance estimates the average gain of purity by splits of a given variable. If the variable is meaningful, it tends to split mixed labelled nodes into pure single class nodes. Splitting by a permuted variables tends either to rise nor decrease node purities. Permuting a significant variable, it tends to give a relatively substantial decrease in mean Gini-gain. Gini importance is related to the local decision function that Random Forest uses to select the best possible split. Consequently, it does not exceed the desired limited time to compute. In contrast, mean Gini-gain in local splits, is not necessarily what is most meaningful to measure, in contrary to alteration of the overall model performance. Gini importance is inferior to (permutation-based) variable importance as it is relatively more unstable, more biased and tends to answer a more indirect question.

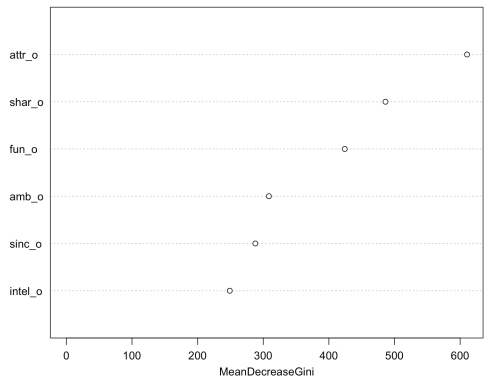


Figure 5: Mean Decrease Gini (MDG)

In this model, the most important variables according to the mean decrease Gini index were shown above, by descending order: attractiveness, shared interest, fun, ambition, sincerity and intelligence.

6 Conclusion

Two algorithms were used to generate classification models for this speed dating experiment dataset: 1) Classification and Regression Tree (CART/RPART), and 2) Random Forest. Gini index was used to calculate the relative importance of variables. The most important variables according to mean decrease Gini index were, in descending order: attractiveness, shared interests, fun, ambition, sincerity, and intelligence.

The RPART model had an accuracy of 75.6%, recall of 0.66, and precision of 0.82. This model showed that "attractiveness", "shared interest", and "fun" were the three most influential of the six variables used. Two individuals will be matched if both give the other an "attractiveness" rating greater than 6.993, "shared interests" rating greater than 5.807, and "fun" rating greater than 6.953. Any combination of scores that does not fulfil one or more of these requirements will not result in a match.

The Random Forest model consisted of 500 trees and had an accuracy of 74.1%, recall of 0.71, and precision of 0.68. The accuracy, precision and recall of both models were quite close, with the RPART model performing slightly better than the Random Forest model regarding the accuracy of each prediction but not the reliability in terms of variation of the values in a set of predictions.

In conclusion, apart from the analysis of the models, the main objective of this project is to model the process of speed dating and investigate the latent patterns behind the human interaction experiment. According to the results, is it true that humans nowadays are inevitably shallow and only base their decisions on superficial attributes such as attractiveness? The answer to this question may be unavoidably ambiguous. However, people should not despair because of the external validity of the experiment since it cannot precisely guarantee that people nowadays based their decisions solely on superficial attributes. During four minutes, the participants could not get to know their partners into depth and therefore they are likely to make their decisions based on external characteristics. This fact concludes why attractiveness ranks higher among other attributes.

References

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and Regression Trees.* Chapman and Hall.

Fisman, R., Iyengar, S., Kamenica, E., & Simonson, I. (2006). Gender Differences in Mate Selection: Evidence from a Speeding Dating Experiment. *The Quarterly Journal of Economics*, 673-697.

Gromping, U. (2012). *Variable Importance Assessment in Regression: Linear Regression Versus Random Forest.* The American Statistician.

Humphries, M. (2010). *Missing Data and How to Deal: An Overview of Missing Data.* https://liberalarts.utexas.edu/prc/files/cs/Missing-Data.pdf.

Lauritzen, S. (2006). *Missing Data and the Expectation Maximisation Algorithm.* http:// www.stats.ox.ac.uk/ steffen/teaching/fsmHT06/fsm406bw.pdf.

Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, *2*, 18-22.

Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC curve to Reclassification and Beyond. *Statistics in Medicine*, *27*, 157-172.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, *8*, 1.

Appendix A RPART Analysis

```
> rpart.analysis(rpart)
n= 6378

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 6378 2677 no (0.5802759 0.4197241)
   2) attr_o< 6.992849 3489  777 no (0.7773001 0.2226999) *
   3) attr_o>=6.992849 2889  989 yes (0.3423330 0.6576670)
     6) shar_o< 5.807483 996  482 no (0.5160643 0.4839357)
      12) fun_o< 6.953038 480  177 no (0.6312500 0.3687500) *
      13) fun_o>=6.953038 516  211 yes (0.4089147 0.5910853) *
     7) shar_o>=5.807483 1893  475 yes (0.2509245 0.7490755) *
> confusion.analysis(confusion)
           Pred:yes Pred:no
Actual:yes      555     283
Actual:no       205     957

Accuracy:                    0.756
Sensitivity/Recall:          0.662291169451074
Specificity:                 0.823580034423408
Fall-out:                    0.176419965576592
Miss rate:                   0.337708830548926
Precision:                   0.730263157894737
Negative predictive value:   0.771774193548387
False discovery rate:        0.269736842105263
F-measure:                   0.694618272841051
```

Appendix B Random Forest Analysis

```
> rf.analysis(rf)
Call:
 randomForest(formula = as.factor(dec_o) ~ attr_o + intel_o + sinc_o + fun_o +
 amb_o + shar_o, data = train, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 27.38%
Confusion matrix:
     0    1 class.error
0 2800  901   0.2434477
1  845 1832   0.3156518


Importance:
                0         1 MeanDecreaseAccuracy MeanDecreaseGini
attr_o  69.246698 139.52079           157.027493         618.2059
intel_o -2.866846  23.45674            18.361479         247.8271
sinc_o  -4.809755  45.29150            35.999591         293.3023
fun_o   32.829674  48.99429            66.838527         413.3848
amb_o    1.356246  10.06726             8.642003         308.4258
shar_o  28.140793  69.75619            74.036037         487.0278
> confusion.analysis(confusion)
           Pred:yes Pred:no
Actual:yes      595     243
Actual:no       275     887

Accuracy:                    0.741
Sensitivity/Recall:          0.710023866348449
Specificity:                 0.763339070567986
Fall-out:                    0.236660929432014
Miss rate:                   0.289976133651551
Precision:                   0.683908045977011
Negative predictive value:   0.784955752212389
False discovery rate:        0.316091954022989
F-measure:                   0.69672131147541
```