National College of Ireland

Machine Learning Approaches To Identify Preferences In An Ideal Match Using Speed Dating Data

MSc Research Project
Data Analytics

PRIYADARSHINI T S
x17110491

School of Computing
National College of Ireland

Supervisor: Dr.Pramod Pathak,Dr.Paul Stynes, Dympna O'Sullivan

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	PRIYADARSHINI T S
Student ID:	x17110491
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dr.Pramod Pathak, Dr.Paul Stynes, Dympna O'Sullivan
Submission Due Date:	13/08/2018
Project Title:	Machine Learning Approaches To Identify Preferences In An Ideal Match Using Speed Dating Data
Word Count:	4782

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Machine Learning Approaches To Identify Preferences In An Ideal Match Using Speed Dating Data

Priyadarshini T S
x17110491
MSc in Data Analytics
National College of Ireland

12th August 2018

Abstract

Speed dating data has various angles to look at. It is a 4-minute conversation between two people to find a desirable match. In this research, an analysis on attributes that influence a participant's decision on further contact is examined. The data is also explored in terms of gender preferences, hobbies and change in self-perception over time. Machine learning algorithms like Naïve Bayes, Random Forest, C5.0 and Gradient Boosted Decision Tree are used to determine the predominant attributes on decision making. Exploratory analysis was done on self-perception. Multiple regression was implemented to analyze a match with socio-cultural attributes. Gradient boosted decision tree performed better with an accuracy of 75.74 percent compared to other implemented classifiers. Attractiveness and fun stands out as desirable attributes in a prospective match.

Contents

1 Introduction

Making a choice is a difficult task. Understanding a personality will make it easy to bind like minded people. Speed dating gives a platform to meet a pool of people, make a conversation and then decide whether one want to accept or reject a person. As the interaction time during the event will be usually between 3-4 minutes, each person will be cautious and curios to know each other and give the best. It avoids awkward pauses when they meet and psychological effects if a person is rejected. Predicting a desirable person is a tough process. Match making companies are trying to provide a tailored list of matches based on their preferences. To provide a probable match and to click a first meeting, it is very important to analyze the most sought preferences. This can be achieved by interpreting a person's qualities and their expected qualities in a partner. The speed dating data is widely analyzed in psychological research fields. The approach is to analyze the attributes that play a part in decision making using machine learning algorithms and also to investigate self-perception. Having a potential partner and a long-term relationship is a universal desire. Finding a suitable partner is more than an opportunity as preferences play a major role. Speed dating gives an opportunity to see the data in various angles like gender preferences, behavioural analysis and effects of rejection Pe et al. (2016). Many factors contribute to complex relations which may dissolve a relationship. Speed dating gives a platform to quickly interact, share interests and be upright in their expectations in a partner and views and decide will they give a second chance to see will the relationship workout or not. Speed dating data gives divulge insights in the initial stage of a conversation. Technology has changed the way of finding a partner. It has definitely reshaped romance.It is an evolving culture in today's busy life style, they are usually arranged in social events, parties and pubs. The investment of time in long face to face meetings which can drain emotional energy can be avoided and there will be no obligations to talk about previous relationships. If they have nothing to share in common it is OK to converse and give a merciful ending.

Big data has showed a remarkable progress in the psychology field. Many applications are build to understand human behavior so the best can be delivered to the person based on their likes. Understanding what people are searching for is really important for both speed dating event organizers and online dating sites so the participants feelings are not hurt. By providing a tailored list of potential partners by understanding their idea of love will boost the success rates. It is a high responsibility to meet their expectations as study also says people who get rejected several times tend to lose interest in making new relationships and to come out and talk to people and go into depression which is a dangerous psychological disease Pe et al. (2016)

The purpose of this study is to cognizance the preferences of a person. It helps both the genders to know on what qualities they can groom and how to mark his/her presence in a short period of time.

The report is divided into different sections. Section 2 has a description of related work, section 3 has Methodology, section 4 about the design and implementation, section 5 has results and discussion and section 6 conclusion and future work.

Research Question – How the gender preferences differ in finding an ideal match? A study on speed dating data using machine learning algorithms

2 Related Work

Buss and Barnes (1986) explored the importance of preference and opportunities using the data from a speed dating agency in UK. Participants converse for 3 minutes. The data had 3600 participants who took part in the event between January 2004 and October 2005. Regression method is used to analyze the preferences. The results say that women like young and tall men and men like women who are slim and young. The linear probability model shows that the proposal percentage decreases for smokers for both the genders. Education and occupation have less priority. For male, education is positively correlated to age and height. The popularity attribute in the model says that women are more likely to propose a man who is less famous whereas it is vice versa for men. Belot and Francesconi (2006)

Social Relations model analysis is performed on 382 participants aged between 18 to 54 years from a community. The analysis involves both hypothesis and dyadic effects on attributes. An average 116 men and women got a match i.e. 60.7 percent participants resulted in a match. Both gender's interests are inclined towards physical attractiveness. Women are more drawn towards sociosexuality, income and education of men. Asendorpf et al. (2011)

Buss and Barnes (1986) did a study on consequences of mate preferences and sexual selection. The experiment was performed on 184 members in that 92 married couples were tested. They were made to fill the assessment questionnaire like confidential biographical questionnaire, marital preferences questionnaire, California Psychological Inventory (CPI), Eysenck Personality questionnaire (EPQ), Interpersonal Adjective Scales (IPA), Self and Spouse ratings, Interpersonal Dependency Scales (IDS), Personal Attributes Questionnaire, EASI Temperament Scales and Personal Attributes questionnaire (PAQ). t tests are used to determine sex differences, the tests say men demands in preferences are more than women. Varimax rotation is used for factor analysis. The mean and standard deviation of men and women are taken and the significance values and t values showed mate selection differs in men and women. Buss and Barnes (1986)

The study is conducted on undergraduate students from a Singaporean University. General linear model is applied on the data, the probability of men choosing women increases with attractiveness and women choosing men increases with social status. The participants were made to sit for another survey on short term and long-term relationship. A regression analysis shows men insisted on physical attractiveness for a long-term relationship whereas women had a moderate expectation on outlook. However, 80-90 percent of men and women choose physical attractiveness for a short-term relationship. Li et al. (2013)

Speed dating event was conducted for undergraduate students in 2005. Random forest was applied to predict actor and partner desire with gender as a predictor. The results show 5 to 18 percent of variance in actor desire and up to 18 to 27 percent variance in partner desire. People who tend to be more selective experienced less attraction. People with warmth personality experienced more attraction.Joel et al. (2017)

The subjects for the experiment were taken from Columbia University. A descriptive analysis was made on the field of career which showed that most of them were from business background. Both genders prefer partners from a populated area. A linear probability model shows female inclination towards intelligence and male inclination towards physical attractiveness. The effect of outlook is 18 percent higher in males saying 'Yes'. The

result also says men prefer women who are ambitious to his level and women prefer men who are more ambitious than her. Fisman et al. (2006)

193 participants were taken from a university in northeastern part of United States. The participants interacted using a validated interface called ChatPlat. They were divided into 2 groups and had given instructions to ask at most 4 questions for one group and 9 questions for another group. The group which asked at most 9 questions received more responses and were interested to go on a second date compared to the group who asked 4 questions. Asking more questions increases liking's as it invites self-disclosure. They are more responsive to bind as.Huang et al. (2017)

The data set is taken from an American Business School. Random forest classifier is applied to the data which is divided into male and female. Men prefer partner who are outgoing. People who like theatre and hiking got more than 5 matches. Men got more matches who like gaming or clubbing. Women who likes reading got more matches. ColinLeverger (2016)

finkel2007speed conducted a study on participants from Northwestern University with a total of 163 participants. Social Relationship model on the data gave insights like, out of 163, 206 matches were formed based on the post-event form filled by the participants. Based on the answers filled in the questionnaire the matches were drawn. Finkel et al. (2007)

Social Relations model showed initial attraction and dyadic relationship between actors and partner. Meta-analysis showed the sexual differences and similarities. The similarity principle says that the individual is attracted to a similar personality. Reciprocating to others feelings will increase the liking between partners. Men are attracted to fun loving and socially active women. Women showed their interest to young and heavier men who are socially active, extroverted, cheerful and open minded.Luo and Zhang (2009)

When two people meet they usually feel in a happy or stressed. A social relationships model on 40 students confirms that a positive emotion prompts individual's reaction. They are more willing to meet again with whom they perceived responsiveness on romantic reaction. It correlates the affective presence. People usually connect who make a positive affective presence like agreeableness, portray emotions,expressiveness, good mood and how well they attach. Attributes like anxiety, disagreement,showing negative vibes are negative affective presence which can put off the spark. The study shows what kind of presence tickle a romantic spark between the two. People with greater positive effective presence showed likeliness to meet them again and it influence to initiate new potential relationship. Berrios et al. (2015)

Data from HurrayDate, a speed dating firm is used to find the preferences from the specifications made in the advertisement. A descriptive statistics and regression analysis shows the preferences of women interested in specific body types, whereas 56 percent of men and 74 percent of women did not insist on body type. Participants preferred European descent and women were more likely to express racial preference than men. Kurzban and Weeden (2007)

Aim of the match makers is to rank the candidates for a participant based on their liking's and retrieve them. An evaluation of their interests is made by collecting the data. A model is built to get deep understanding on preferences. The candidates are ranked using gradient boosted decision tree (GDBT). This algorithm is highly effective in ranking which helps to analyze and retrieve matching pairs. The relevance score of

the pair is computed. GDBT uses different subsets of features of runs. These runs are labelled as match only, candidate profile only, two-way match only and profile similarity only. A match weight is given according to the source of label given to profiles. The model performs well on the metrics given.Diaz et al. (2010)

Zang et al. (2017), conducted a study on online dating service named IBJ.INC in Japan. They extracted basic information and preferences from user profiles. The data set included 11,421 male and 8615 female users. Topological features like demographic features , user preferences, facial features, text features were created. Support vector machine (SVM), Naive Bayes, random forest and logistic regression were used fo find the matches based on preferences. SVM outperformed with 76.3 percent accuracy by getting user's with reciprocal interests leading to successful dates.

The above research findings are form social relationship model, linear probability and statistical models. In this research machine learning algorithms will be used to predict the attributes that sway the decision, hobbies that are of interests to both the genders which make them to reciprocate to feelings are examined and also to analyze the self-perception.

3 Methodology

The implementation of this project is based on CRISP-DM (Cross Industry Process for Data Mining) methodology. The steps are shown in figure 1:

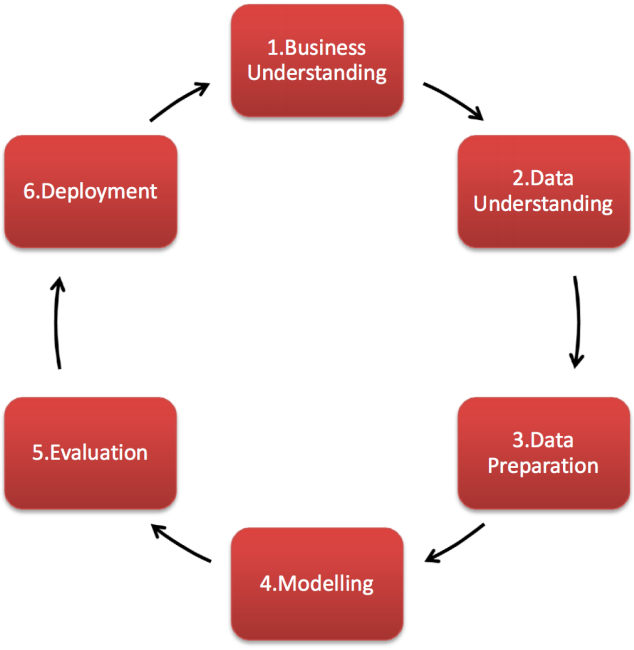


Figure 1: CRISP-DM Technique

3.1 Data description

The data was retrieved from a speed dating event conducted in Columbia University. The participants background data like race, religion, career, zip code, tuition, SAT score, age, how often they go out casually, how often they go out on dates and hobbies were

collected and a form was asked to fill up before, during and after the event, to record what they perceived and to find suitable matches and to know their willingness to see him/her in future. The participants were given 4 minutes to talk to one another. This record is used to find a probable match and the details were forwarded to the concerned participant for further meetings.

The attributes that influence to meet again, how self-perception change over time and attributes that influence a match in male and female are analyzed using machine learning algorithms like multiple regression, Naïve Bayes, decision trees like random forest, C50 and Gradient Boosted Decision Tree (GBM) and descriptive statistics. Naïve Bayes is a common approach for any classification and regression problem and it is used as a baseline classifier. As the analysis is majorly based on finding which attributes influence most in partner selection, random forest is implemented as it is a good indicator of feature importance. Decision tress are best suited to discover the interactions between variables and gives accurate predictions. After finding out the variable importance, boosted decision trees like C50 and GBM were implemented to achieve better accuracy. Metrics like accuracy – evaluates the model built, recall – actual positives identified, precision – accuracy of the positive identifications, f-measure – compares betterness and performance of the algorithm and AUC – gives the overall performance of the model, are used to evaluate the accuracy of the model built. The model is trained to overcome the effects of overfitting with k-fold cross validation (5 x 10 – fold cv) approach. The advantage of k-fold is it uses maximum possible data for training and test sets are mutually exclusive and covers almost entire data set. The model is tested against the test data and the results are evaluated based on the metrics mentioned above.

4 Design and Implementation

4.1 Data pre-processing and Feature Selection

The data is loaded into R Studio and analyzed for missing values. The missing values were removed. The data has 3 sessions: follow up time 1, follow up time 2 and follow up time 3. The ratings in few sessions are in a scale of 1 to 100, these values are normalized for the implementation of actor-partner model. In other sessions ratings are in a scale of 1 to 10. Features like SAT score, id, undergraduate school, zip code and tuition were omitted. However, features like hobbies, how often they go out, race and shared interests are retained which is used in personality analysis. The attributes like attractiveness, ambition, sincerity, fun and intelligence which gives insights on decision making and influence of these on participants were taken for analysis. After pre-processing, the data set has 5965 observations which is sufficient to fit a model.

4.2 Modelling

The following are the models that were implemented for this research:

a) Naïve Bayes

Naïve Bayes is the simple approach to a problem and it is used as a base classifier. It is fast, simple to build and non sensitive to irrelevant features. The pre-processed data with decision as a predictor and attributes like attractiveness, ambition, intelligence, fun and sincerity were taken and loaded into R studio. Set.seed function is used for random

sampling of data. The data is divided into training and testing set in the ratio of 75 and 25. e1071 library is used, the model was trained using a trained data set with a function called naiveBayes and tested against a test data set. To avoid overfitting k-fold cross validation (5x10 cv) was implemented and the model is trained and tested again. The model gave an accuracy of 69.30 percent and AUC value of 70.43. The results of Naïve Bayes model are shown in Table 1:

Metric	Value
Accuracy	69.30
AUC	70.43
Precision	79.77
Recall	63.92
F-Measure	70.97

Table 1:Results of Naive Bayes with k-fold cv

b) Random Forest

Random Forest helps to identify importance of attributes like attractiveness, ambition, intelligence, fun and sincerity over a dependent variable decision. Random Forest searches for the best features instead of the most important feature. Set.seed function is used for random sampling of data. The data is divided into training and testing set in the ratio of 72 and 25. randomForest library is used, the model is trained using a trained data set with a function called randomForest and tested against a test data set. From the variable importance plot in Figure 2, it can be said that attractiveness and fun are the major attributes that makes an impact on decision making. The model gave an accuracy of 72.99 percent and AUC value of 72.14. The results of random forest model are shown in Table 2:

Metric	Value
Accuracy	72.99
AUC	72.14
Precision	75.68
Recall	77.89
F-Measure	76.77

Table 2: Results of Random Forest model

To get a fair view of significance of variables, varImpPlot function is used. Gini measure (Mean Decrease Gini - Figure 2) shows how pure the leaf nodes are and a high score shows the variables are important. The most important variable from the plot is attractiveness and fun and the least important attribute is intelligence.

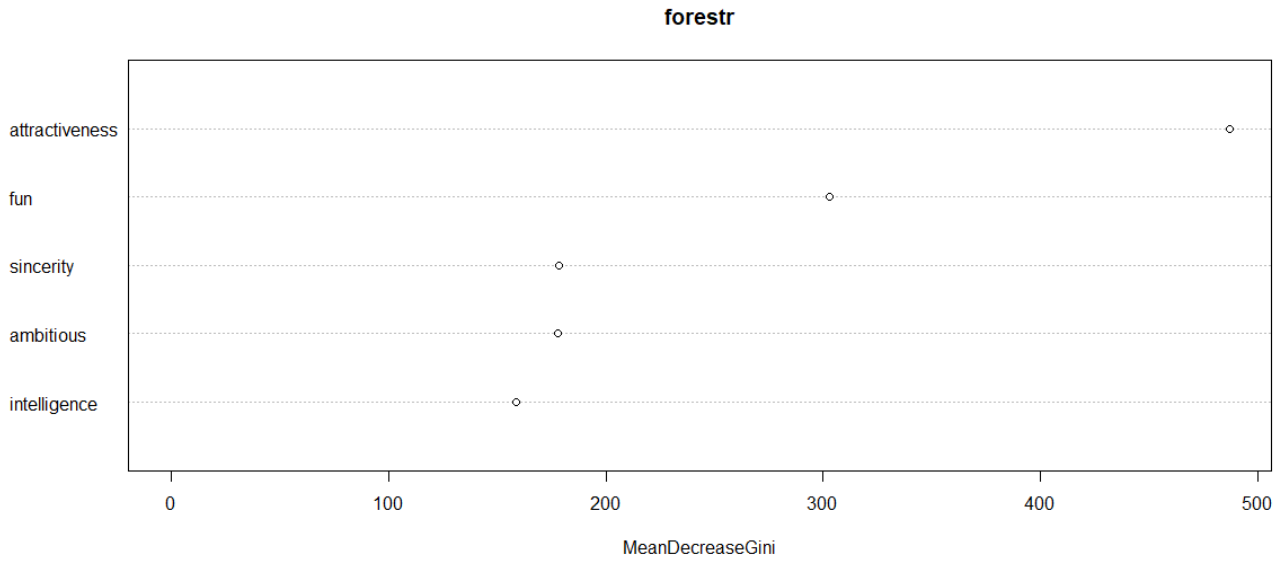


Figure 2: Variable Importance Plot using Random Forest

In this paper two boosted approaches of decision trees are implemented, they are, C50 and GBM to find better accuracy.

c) C5.0

C50 package is installed. C5.0 function is used to build a decision tree. Set.seed function is used for random sampling of data. The data is divided into training and testing set in the ratio of 75 and 25.The model is trained using a train data and the trials is set to 10, it is a boosting parameter and it will produce 10 trees and the best among them is voted. To reduce the impact of overfitting the model is again trained with k fold cross validation (5 x 10 cv) and the model is tested against the test data. The model gave an accuracy of 73.39 percent and AUC value of 72.33 percent. The results of C5.0 model is shown in Table 3:

Metric	Value
Accuracy	73.39
AUC	72.33
Precision	77.57
Recall	77.84
F-Measure	77.70

Table 3: Results of C5.0 with k-fold cv

From Figure 3 it can be seen that attractiveness (attr) and fun are the most sought attributes in a desirable partner.

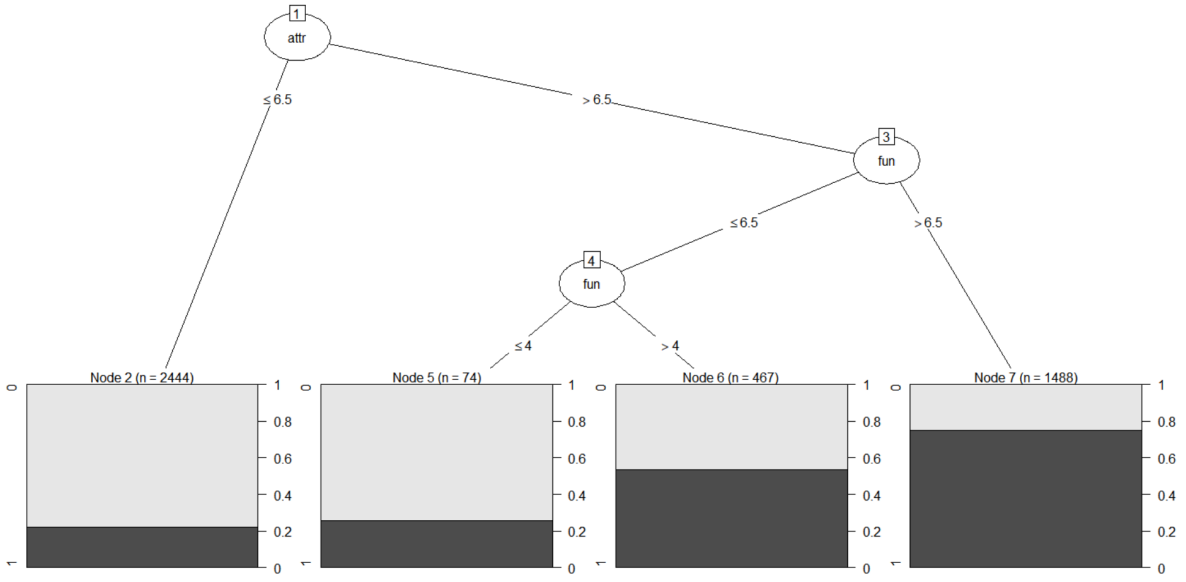


Figure 3: C5.0 Tree

d) Gradient Boosted Decision Tree (GBM)

GBM has a high prediction power as it combines weak learners and increases the robustness over a single classifier. It unifies weak and average predictors to generate a strong predictor. It minimizes the loss function and generates more accurate results Diaz et al. (2010). gbm package is installed and the model is fit using gbm function. Set.seed function is used for random sampling of data. The data is divided into training and testing set in the ratio of 75 and 25.To overcome the problem of overfitting the gbm model is trained using parameters like distribution - "Bernoulli" it is a type of loss function and it is default if the response value has 2 unique values, shrinkage – represents step size reduction and is set to 0.05, interaction depth is set to 1 as it is a additive model, minobsinnode is set to 1 – it is the minimum number of observations in trees terminal nodes, cross validation is (5 x 10). ntrees – it is the number of iterations, initially it is set to 1500. By performing gbm.perf function the best number of iteration is found i.e 185 which is shown in Figure 4. The model is tested against test data. It gave an accuracy of 75.74 percent. The results of GBM model are shown in Table 4:

Metric	Value
Accuracy	75.74
AUC	74.47
Precision	78.57
Recall	81.24
F-Measure	79.88

Table 4: Results of GBM with k-fold cv

Figure 4 shows best number of trees for prediction. The black line is the performance of training subset to fit a gbm model and the green line is the testing subset. As the number of trees goes up (i.e. the number of iteration in x-axis) the black line goes down and the training performance continues to improve. However, the test performance stops improving which is represented by green line and it is seen that the arc goes up, at this

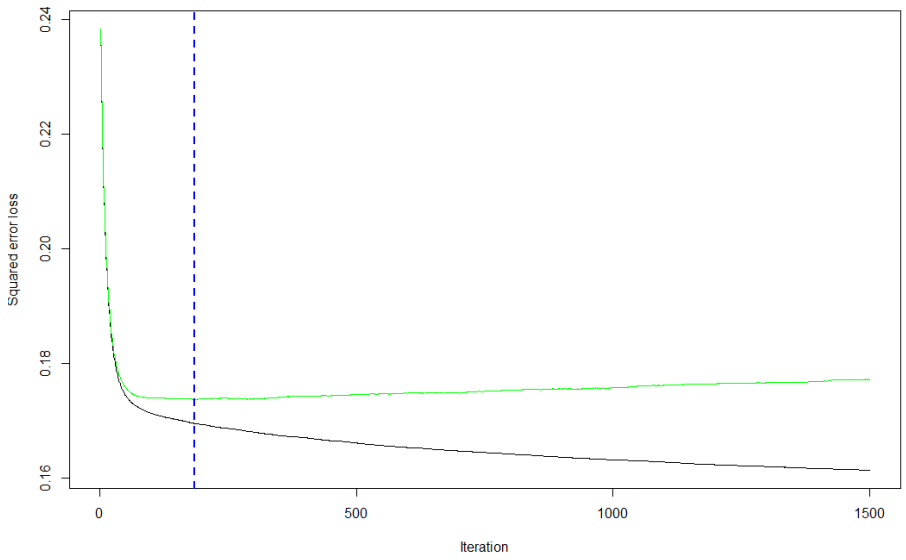


Figure 4: Best Number of Trees for Prediction

stage the gbm has picked the dotted line at 185 as the best iteration.The tree selected for prediction i.e. the vertical blue line, is the tree that minimizes the testing error on the cross-validation folds.

	var	rel.inf
attr	attr	51.852232
fun	fun	21.044544
amb	amb	9.531944
sinc	sinc	9.394599
intel	intel	8.176681

Figure 5: Summary of GBM model

Figure 5 shows the summary of the model. The summary of the model shows relative influence of a variable on decision making. The attribute in the top is the most important i.e. attractiveness(attr) and at the bottom is the least important i.e. intelligence(intel). The plot of this is shown in Figure 6.

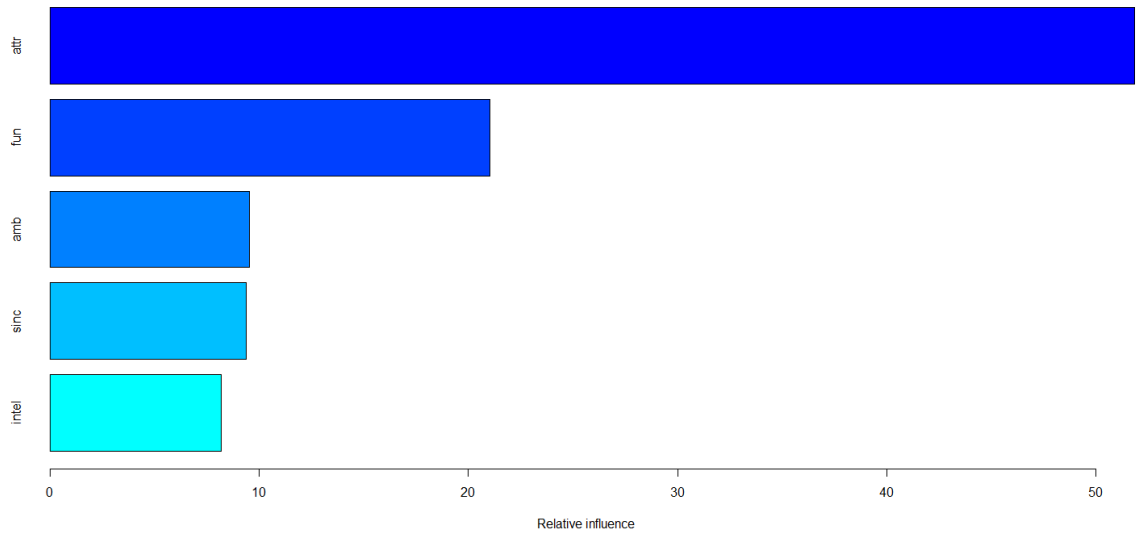


Figure 6: Relative Influence Plot using GBM model

e) Actor- partner model

Random forest algorithm is used to build an actor-partner model to analyze the participants on how they rated themselves and other participants. Actor model means self-perception and partner model means what attributes they look for in a partner. These ratings were taken after 3-4 weeks, after the matches were sent out. The model was built using all participants who got at least one match, so it can be analyzed on what attributes most of them found a match and their self-perception. The actor model shows how they perceived themselves after they found a match. From the variable importance plot as in Figure 7 it can be said that the participants rated themselves as more fun loving (fun) and attractive (attractiveness). The actor model gave 83.10 percent accuracy.

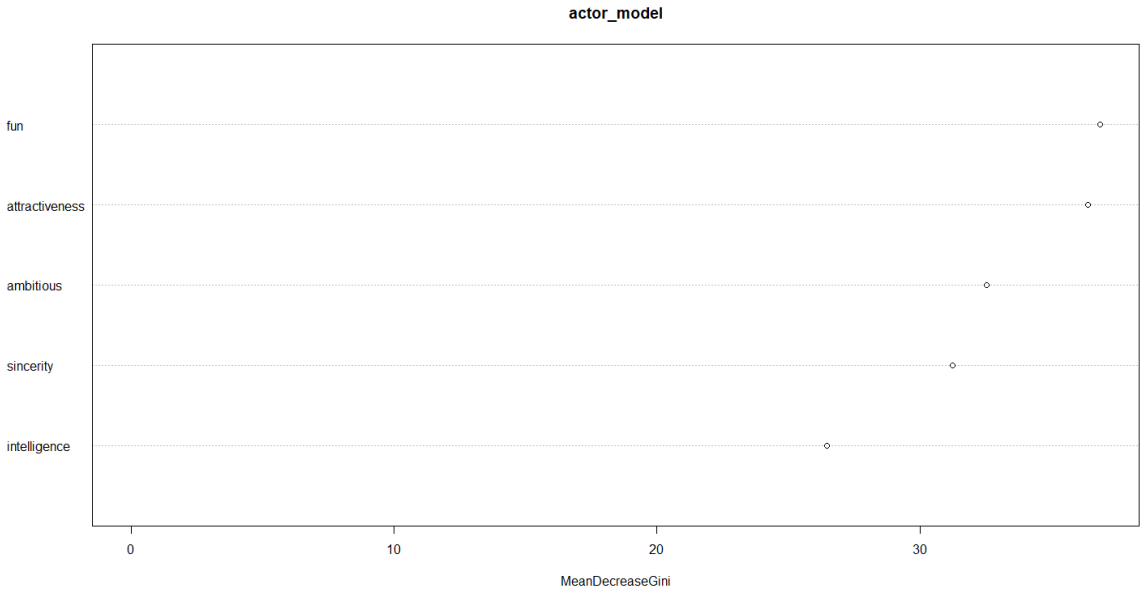


Figure 7: Actor model

As in Figure 8 the partner model shows the attribute ratings after they found at least one match. After 3-4 weeks of the event the participants preferred attractive (attractiveness) and sincere (sincerity) partners. The partner model gave 79.98 percent accuracy.

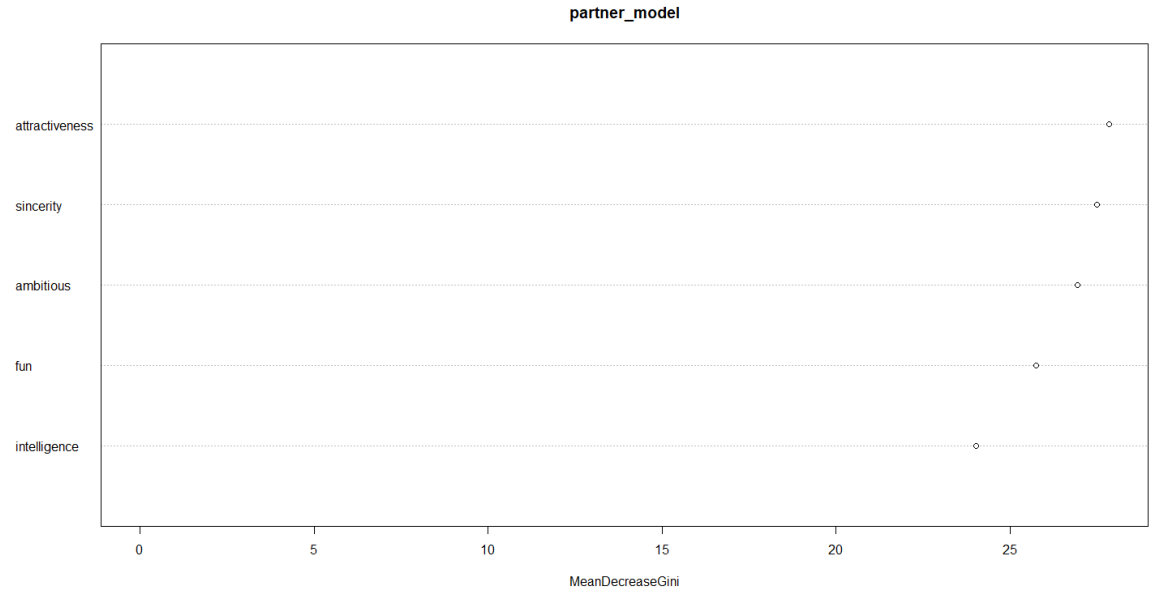


Figure 8: Partner model

f) Gender wise preferences

Gender wise preferences after they got a match were analyzed using randomForest function. To fit a model along with the above mentioned 5 attributes 'sharedinterests' attribute is also considered. The data was divided into male and female and their preferences are shown in a variable importance plot. From Figure 9 and Figure 10 it is seen that both men and women prefer partners who are attractive (attractiveness) and share common interests (sharedinterests).The male preference model gave an accuracy of 82.29 percent and female preference model performed with an accuracy of 80.94 percent

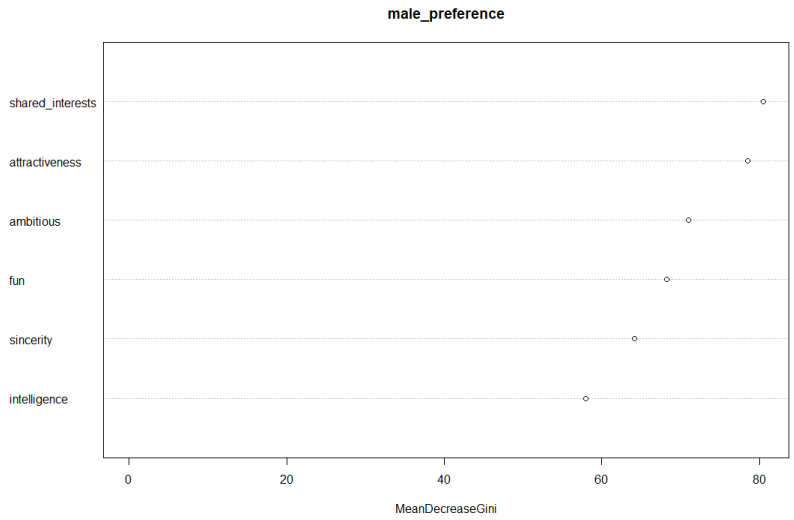


Figure 9: Male Preferences

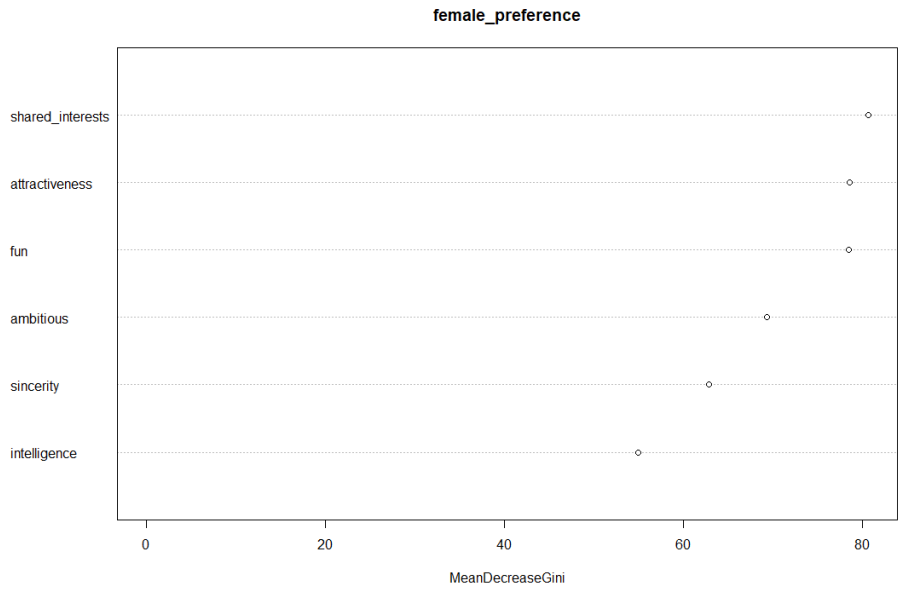


Figure 10: Female Preferences

g) Gender wise analysis of hobbies who got more matches

The data is divided into male and female as it helps to analyze other attributes like hobbies. A variable importance plot is implemented using random forest. The plot shows the hobbies who got at least or more than 3 matches in the event.

The Figure 11 shows that women whose hobbies were tv sports, concerts and clubbing were more likely to find a good number match. The best splits chosen to form a tree are shown in IncNodePurity. The most useful variable gets a high score. In male hobbies model tv, reading and theatre hobbies got a high score similarly in female hobbies model tv sports, concerts and clubbing hobbies got a high score.

Figure 12 shows, men whose hobbies were tv, reading and theatre were more likely to find a good number match.

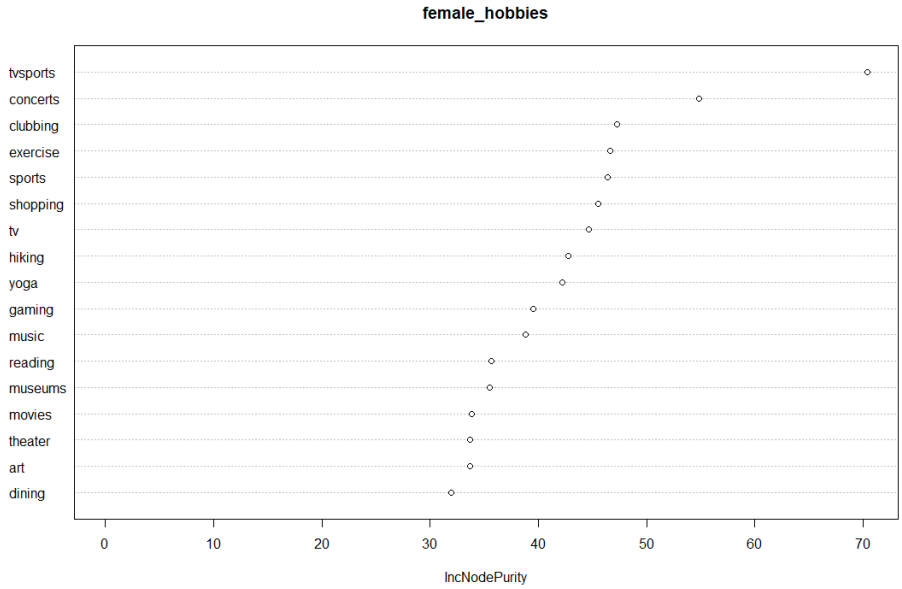


Figure 11: Female Hobbies

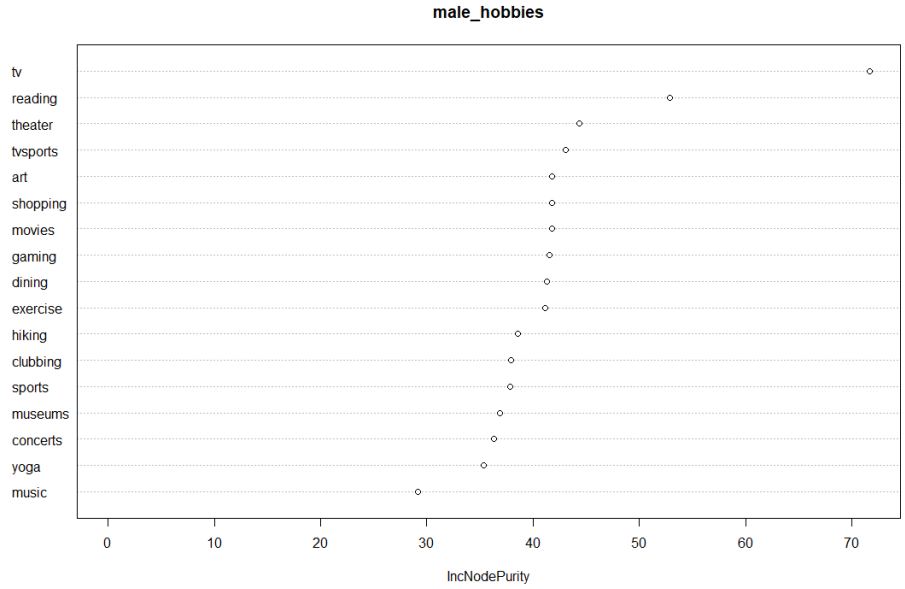
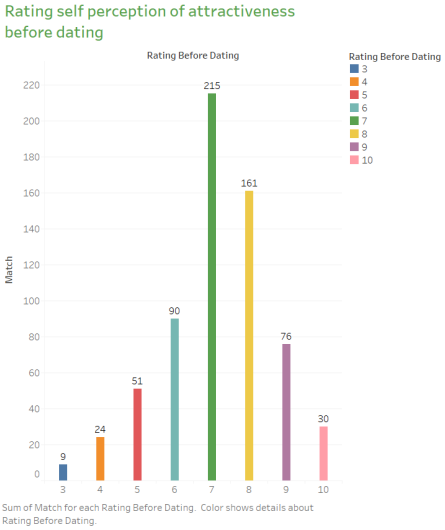


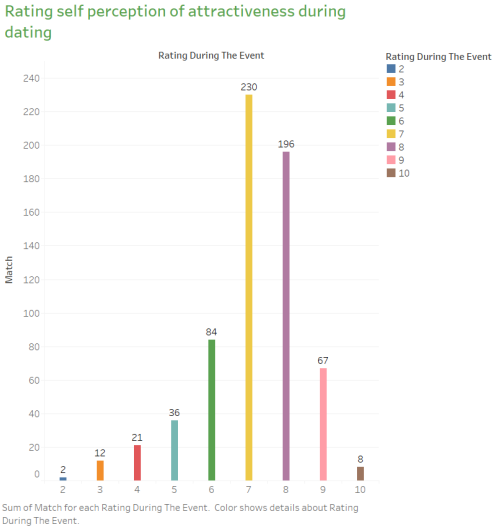
Figure 12: Male Hobbies

h) Variation of ratings in self-perception over time

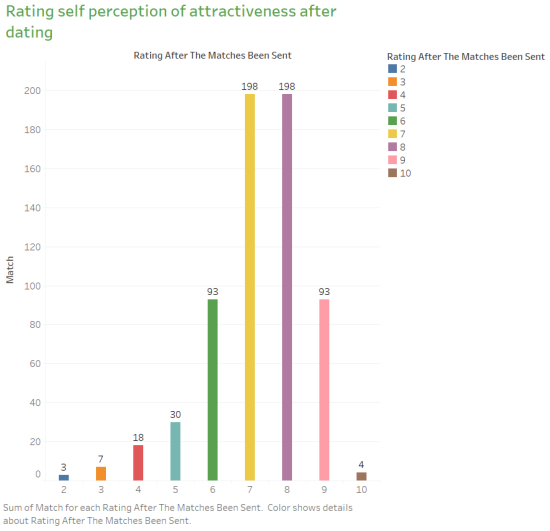
From the results of the implemented models it can be said that attractiveness is the most sought attribute. So, a descriptive analysis is done on how people rated themselves on attractiveness over a period of time i.e. before the event, during the event and after the event. The Figure 13 (a) shows the ratings before they got a match i.e. before the event, most of them rated between 6 to 8 upon 10 points. Figure 13 (b) shows the how they rated during the event, they rated between 7-8 and Figure 13 (c) shows the ratings of after the event, they rated between 6-9.



(a) Self perception before the event who found a match

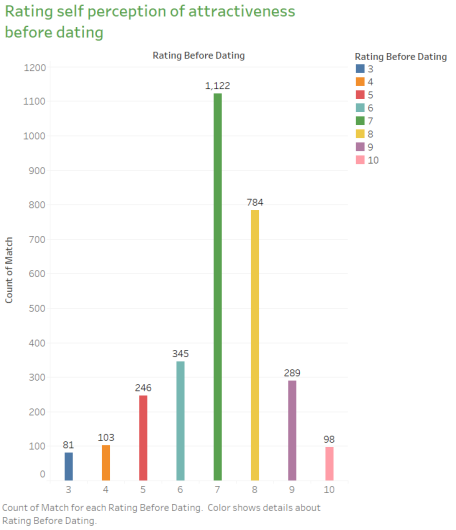


(b) Self perception during the event who found a match

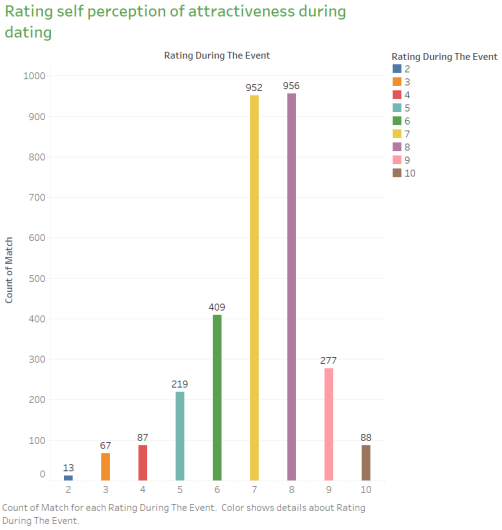


(c) Self perception after the event who found a match

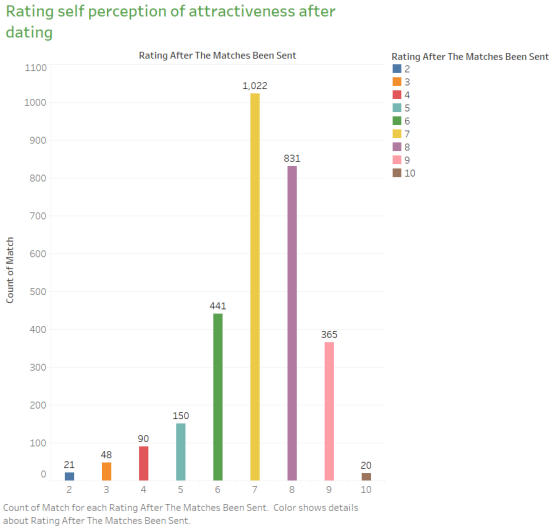
Figure 13: Self Perception who found a match



(a) Self perception before the event who did not find a match



(b) Self perception during the event who did not find a match



(c) Self perception after the event who did not find a match

Figure 14: Self Perception who did not found a match

The above graphs in Figure 14 shows the ratings of self-perception on attractiveness who did not get a match over a period of time as mentioned above. Before the event most of them rated between 7-8 as shown in Figure 14 (a), during the event most of them they rated between 6-8 as shown in Figure 14 (b) and after the event most people rated between 7-8 as shown in Figure 14 (c). Usually, we expect to see a decrease or increase in the ratings if someone is rejected or accepted but from the above graphs it can be said that there is no much change in self-perception throughout the process. This may be due to conversing with so many different people with different expectations, influenced the participants to decide on how they were perceived by others.

i) Multiple Regression

Multiple regression is done to analyze what kind of personalities get along and found a match. To do this analysis the data set was split into male and female. The model was

built with match as a predictor and attributes like same race (samerace), how often they go out (goout), how often they date someone (date) and shared interests (sharedinterests) were considered. For male model the probability of getting a match increased for the participant who is from same race (samerace) and shares similar interests (sharedinterests) whereas outgoing female (goout) and female who go out on a date frequently (date) has negative influence on the predictor. For female model the probability of finding a match increases with shared interests (sharedinterests) and all other attributes like same race, goout and date has a negative influence on the predictor. This is shown in the below equations:

P male ((getting a match)/(not getting a match)) = 0.50 + 0.044 x samerace - 0.026 x date – 0.040 x goout + 0.012 x sharedinterests - equation (1)

P female (getting a match) /)/(not getting a match) = 0.53 – 0.018 x samerace - 0.006 x date – 0.026 x goout + 0.007 x sharedinterests - equation (2)

5 Results and Discussion

First Naive Bayes,Random Forest, C50 and GBM models were implemented with decision as a predictor. These models were built with an objective to analyze which attribute is most influential in decision of meeting a person further or not. The results of these models are summarized in Table 5 and compared under metrics like accuracy and AUC curve. The results are shown in Table 5

Method	Accuracy	AUV value
Naive Bayes	69.30	70.43
Random Forest	72.99	72.14
C50	73.39	72.33
GBM	75.74	74.47

Table 5: Summary of the Results

From the results it can be seen that GBM which is an ensemble method performed best with 75.74 percent accuracy and AUC value of 74.47 percent. From the previous related work, it has been seen that attractiveness is the most desirable variable Asendorpf et al. (2011). This research also predicted 'attractiveness' as the most desirable attribute in a potential partner.

The other models like actor-partner model, gender wise preferences, gender wise analysis of hobbies were implemented after they got at least one match. These results are summarized in Table 6

Model	Accuracy
Actor model	83.10
Partner model	79.98
Male Preference model	82.29
Female Preference model	80.94

Table 6: Results

The actor-partner model and male and female preference model examined self-perception, attributes desired in a partner and male and female preferences which helps to see the data in a different dimension and it gave reasonable results.

Multiple regression was implemented using attributes which describes the other qualities of a participant like shared interests, same race, how often they go out and how often they go out on a date, the results of multiple regression male model shows that same race and shared interests are positively correlated with the predictor match and for female regression model only shared interests is positively correlated with the predictor match. A descriptive analysis of self-perception over time was also done, which did not show much changes in ratings of self-attractiveness though they found a match or not. On the whole, the key findings of the analysis are attractiveness and fun, these attributes are perceived as a major attribute that influenced the participants decision.

6 Conclusion and Future Work

A set of models were implemented to analyse the attributes that makes an impact on decision on how they perceived the other participant and are they interested to take it forward or not. The C5.0 and GBM models gave more accurate results when trained with 5-fold cross validation but have to compromise with time as the execution time is bit more when more trials are run. The models can be used on other data like online dating, matching jobs and any data related to ratings such as products, chocolate and many more.

The present research is based on 5 attributes and the impact of those attributes on decision making. In this research the gender preferences were also looked at as it is important to know the desires of men and women to make an event more successful. A better understanding of self-perception over time is also seen in a descriptive analysis. In future work different combination of attributes can be considered which may give better accuracy.

References

Asendorpf, J. B., Penke, L. and Back, M. D. (2011). From dating to mating and relating: Predictors of initial and long-term outcomes of speed-dating in a community sample, *European Journal of Personality* **25**(1): 16–30.

Belot, M. and Francesconi, M. (2006). Can anyone be'the'one? evidence on mate selection from speed dating.

Berrios, R., Totterdell, P. and Niven, K. (2015). Why do you make us feel good? correlates and interpersonal consequences of affective presence in speed-dating, *European journal of personality* **29**(1): 72–82.

Buss, D. M. and Barnes, M. (1986). Preferences in human mate selection., *Journal of personality and social psychology* **50**(3): 559.

ColinLeverger (2016). Analysis of speed dating exp dataset with r,gephi.

Diaz, F., Metzler, D. and Amer-Yahia, S. (2010). Relevance and ranking in online dating systems, *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 66–73.

Finkel, E. J., Eastwick, P. W. and Matthews, J. (2007). Speed-dating as an invaluable tool for studying romantic attraction: A methodological primer, *Personal Relationships* **14**(1): 149–166.

Fisman, R., Iyengar, S. S., Kamenica, E. and Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment, *The Quarterly Journal of Economics* **121**(2): 673–697.

Huang, K., Yeomans, M., Brooks, A. W., Minson, J. and Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking., *Journal of personality and social psychology* **113**(3): 430.

Joel, S., Eastwick, P. W. and Finkel, E. J. (2017). Is romantic desire predictable? machine learning applied to initial romantic attraction, *Psychological science* **28**(10): 1478–1489.

Kurzban, R. and Weeden, J. (2007). Do advertised preferences predict the behavior of speed daters?, *Personal Relationships* **14**(4): 623–632.

Li, N. P., Yong, J. C., Tov, W., Sng, O., Fletcher, G. J., Valentine, K. A., Jiang, Y. F. and Balliet, D. (2013). Mate preferences do predict attraction and choices in the early stages of mate selection., *Journal of Personality and Social Psychology* **105**(5): 757.

Luo, S. and Zhang, G. (2009). What leads to romantic attraction: Similarity, reciprocity, security, or beauty? evidence from a speed-dating study, *Journal of Personality* **77**(4): 933–964.

Pe, M. L., Gotlib, I. H., Van Den Noortgate, W. and Kuppens, P. (2016). Revisiting depression contagion as a mediator of the relation between depression and rejection: A speed-dating study, *Clinical Psychological Science* **4**(4): 675–682.

Zang, X., Yamasaki, T., Aizawa, K., Nakamoto, T., Kuwabara, E., Egami, S. and Fuchida, Y. (2017). You will succeed or not? matching prediction in a marriage consulting service, *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*, IEEE, pp. 109–116.