Predicting matches using speed dating results

Gali, Geeta Madhav
gg6549@rit.edu

Kotha, Sai Venkat
sxk2606@rit.edu

ABSTRACT

The dataset selected is put together by Columbia Business school professors. Data was gathered from volunteers who participated in a speed dating experiment conducted between 2002-2004. During this experiment, each attendee have been on a date with every other participant of the opposite sex. After 4 minutes of each date, feedback is collected from the participant. Rating the partner based on Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests is part of collecting feedback. The main aim of this dataset, is to identify whether two individuals would like to go on a date or not based on the attributes described by both the individuals. We are trying to build various models for this data such as Zero rule, One rule, Decision tree, Naive based classifier and K Nearest Neighbors. Evaluating and comparing the results and understanding each model in detail is also part of this project. The classifier with the best performance is selected as the final classifier. We are planning to use Clustering algorithms such as K means clustering and Agglomeration to find the similar interest patterns and the state of thinking among the individuals. Also it helps us to find hidden patterns.[5]

1. OVERVIEW

Online dating is a billion dollar industry through out the world. They are using wide range of attributes such as mutual friends, race, gender, area, hobbies, interests, job, physical attributes, salary, what are you expecting from the other person. Using these many attributes will complicate the process of finding the correct match so, most of the companies are leveraging Machine learning algorithms to find the best match possible for their users. With this project, we are trying to create a similar model using the concepts we learnt as part of Big Data Analytics class. The dataset uses real data gathered from a speed dating experiment that was conducted on the students of Columbia University. The objective of this project is to build a classifier that predicts if there could be a match between two individuals. The dataset contains 195 attributes that provides background information of the individuals along with their likes/dislikes, interests, preferences, salary, job etc. This historical data available could be used to train the model. The section 'Data description' talks in detail about all the attributes present in the dataset. The 'Background' section gives an overview of previous work done on this topic. The 'Implementation' section provides information about the proposed models, work done so far and things that can be done further.

2. BACKGROUND

The speed dating experiment was conducted as part of a research done by Fisman, Iyengar, Kamenica and Simonson.[3] Their research titled 'Gender differences in mate selection: Evidence from a speed dating experiment' used Regression analysis to determine how different genders have different preferences during mate selection. Their research revealed that females give more importance to the partner's intelligence and race. Whereas, males give more importance to the partner's physical attractiveness. Males tend to reject a partner who is more intelligent or ambitious than them, while females tend to prefer a partner who grew up in wealthy neighborhoods.[3]

According to the research, the importance of attributes could be correlated to the reproduction capabilities. As women have limited time during which they are more reproductive, men tend to prefer women who look young and attractive. Women on the other hand tend to prefer men with good earning potential so that they can take proper care of their offsprings.[3]

Considering selectivity, the number of partners males accept is invariant with the group size whereas the number of partners females accept increases with the group size. When considering a particular group size, the number of partners females accept is less than that of the males.[3]

The Regression equation used by the research is mentioned in figure 1

.

$$\mathit{Decision}_{ij} = \alpha_i + \sum_{c \in C} \beta_{c0} * (\mathit{Rating}_{ijc} - \mathit{Self}_{ic}) + \sum_{c \in C} \beta_{c1} * (\mathit{Rating}_{ijc} - \mathit{Self}_{ic}) * (\mathit{Rating}_{ijc} > \mathit{Self}_{ic}) + \varepsilon_{ij}.$$

Figure 1:

$$\mathit{Decision}_{ij} = \beta_0 * \mathit{SameRace}_{ij} + \beta_1 * \mathit{SameField}_{ij} + \beta_2 * \mathit{SameRegion}_{ij} + \varepsilon_{ij}.$$

Figure 2:

While considering the importance of similarity, the regression equation is given in figure 2

The paper indicated that the attributes Fun, Shared Interests and Sincerity could be omitted from the analysis because it was determined that the almost all the participants gave equal rating of importance to these attributes and the removal of these attributes did not hinder the analysis or results.

Another paper titled 'Racial preferences in dating' published by Fisman, Iyengar, Kamenica and Simonson talked about the racial preferences of different individuals during mate selection. This paper described that women of all races tend to prefer partners of the same race while, men's preference for the same race is statistically insignificant. Older subjects of both the genders do not give more preference to the partner's race and attractiveness is not correlated to the race. Also, women do not find partners of the same race more attractive than the partners of other race.[4]

The paper suggests that the background of a subject plays a vital role in the subject's racial preferences. This background includes state or country of origin. Subjects that grew up in intolerant countries have strong same-race preferences. Surprisingly, if a subject grew up in a neighborhood where the majority of the inhabitants belong to a particular race, the subject is less willing to date a partner belonging to that race. People who are more physically attractive do not care about the partner's race.[4]

While determining the racial preferences the regression equation used is given in figure 3

$$\text{Decision}_{ij} = \alpha_i + Race_j + \beta_1 \text{Attractiveness}_j + \beta_2 \text{SameRace}_{ij} + \beta_3 \text{SameRace}_{ij} \times \text{Male}_i + \beta_4 \text{SameRace}_{ij} \times X_{ij} + \varepsilon_{ij},$$

Figure 3:

The paper indicates that even though the experiment was performed on a group of progressive individuals, the results show that these individuals have strong racial preferences. This could be considered as a reason for low rate of interracial marriages in the United States.[4]

To learn about the influence of personal resources and environmental resource pressures on mate preferences, a research study was performed by Rindy C. Anderson and Casey A. Klofstad. The objective of their research titled 'For love or money? The influence of personal resources and environmental resource pressures on human mate preferences' was to determine two things. First, whether women are more likely than men to seek resources from their partner. Second, to determine whether people living in areas with high cost of living are still seeking resources from their partners and if this behavior is more frequent in men or women.[2]

The data used for this research was collected from 2944 profiles from an American Dating Website. The daters reported their income, the preferred income of their potential partner, their address etc. along with other general information. The daters reported their income as one of the 7 categories that range from less than $25,000 to $150,000 or more. The daters could select the desired income of their partners as one or more of the options provided. From the selected options the lowest category of income selected is used for analysis. These income categories are encoded as numbers ranging from 1 to 7. The cost of living index for the zip code of the dater is also considered and is rated more than 100 if it is greater than the national average and is rated less than 100 if it is less than the national average.[2]

The analysis of the data indicated that women are more concerned with resource seeking than men. From the sample used for this research, women wanted a date with a minimum income of $50,000, whereas men wanted a date with a minimum income of $35,000. It was also discovered that women are more likely to list their desired income of the date in the dating profile than men. Men are more likely than women to list their personal income in their dating profile. Linear regression analysis was used to determine the relationship between individual resources, resource seeking and resource pressures. The analysis indicated that as the resource pressure increases, both men and women seek more resources. The relationship between personal income and income seeking is the same for both men and women.[2]

Another extensive research was done by Gunter J. Hitsch, Ali Hortacsu and Dan Ariely on the users of online dating. The data for this research consisted of 3,702 men and 2,783 women. This research included people who were only looking for long term relationships and did not consider any married or engaged individuals. The dataset gathered contained 597,167 observations of user actions for men and 196,363 user actions for women. It was discovered that men contacted 12.5% of all the women whose profiles they viewed and women contacted 9% of all the men whose profiles they viewed. This data suggests that women are more selective while choosing their mate than men.[6]

From the sample of men and women collected for this data, 51% of the individuals uploaded their pictures on their online dating profile. The pictures of all the individuals were rated for attractiveness on a scale of 1 - 10 by 100 students from University of Chicago. Regression analysis was performed on the attractiveness rating and annual income of the individuals. It was determined that for every one standard deviation increase in the attractiveness of men, there is a 10% increase in the annual income and this rate is at 12% for men.[6]

Based on the attractiveness rating all the individuals of the sample were split into deciles. The probability that a man contacts a woman belonging to 81-90th percentile range of attractiveness is 10.7% more than for a woman belonging to lower decile. Correspondingly, this probability for women is 7.9%. This indicates that men have more preference for attractiveness than woman.[6]

When considering the BMI of individuals, men prefer women who have a similar BMI as them and dislike women who have higher BMI. Whereas, women prefer men who have similar BMI as them and dislike men who have a lower BMI. When it comes to education, men and women prefer a partner with similar education level. Both men and women prefer a partner with higher income over a partner with lower income. Both men and women have same-race preference. The probability of a white man contacting a black woman is 10% lower than a white man contacting a white woman. The probability of a white woman contacting an Asian man is 12% less than that of a white woman contacting a white man.[6]

A research done by Joshua Akehurst, Irena Koprinska, Kalina Yacef, Luiz Pizzato, Judy Kay and Tomek Rej on the users of online dating studies the explicit and implicit preferences stated by the users to determine whether the explicit preferences are a good predictor for successful interactions or are the implicit preferences a good predictor for successful interactions.[1]

The data used was gathered from an Australian dating website and represented information and interactions of 8,012 users. The explicit preferences are the desired attributes in the partner stated by the user while signing up on the website. The implicit preferences are the attributes derived by analyzing the successful and unsuccessful interactions of users.[1]

The implicit preferences are represented by a binary classifier that uses the likes and dislikes of a user. The classifier used is a NBTree classifier which is a hybrid of Decision Tree and a Naive Bayes classifier. This classifier is trained on the users' previous successful and unsuccessful interactions.[1]

Method used in the research for rating the explicit preferences:
If only the explicit preferences of the users are used to train a classifier that predicts if an interaction of a user with a potential partner could be successful or not, the accuracy would be 49.43% which is less than the baseline accuracy of always predicting the majority class(ZeroR Rule)which is 59.78%. This indicates that the explicit preferences are not a good predictor for the success of interactions between the users.[1]

Method used in the research for rating the implicit preferences:
To measure the predictive power of implicit preferences, rather than using the explicit preferences as attributes, the users' interactions are used to train the NBTree classifier. Each user has a set of successful and unsuccessful interactions with other users. These interactions are used to build the classifier. The NBTree classifier trained had an accuracy of 82.29% which is greater than the ZeroR baseline of 63.5% and the accuracy of explicit preferences classifier.[1]

The researchers then built a hybrid content-collaborative reciprocal recommender which uses both implicit preferences and explicit preferences of users. The users recommended by the recommender are then ranked and the shortlist of candidates are suggested to a user. The content part of the recommender finds the users that are similar to each other based on their preferences. The collaborative part of the recommender uses the implicit preferences and interactions to determine the users that match. This recommender is reciprocal because it considers the likes/dislikes of both the users while evaluating a match.[1]

3. DATA DESCRIPTION

The data is collected from a speed dating experiment conducted on the students of Columbia University over the week nights during 2002-2004. Before the experiment all the participants filled in a survey that asked them questions about their background information, preferences in their partners, interests etc. Usually, two sessions were conducted during a night and this data is gathered from 14 of those sessions. The participants were randomly distributed among each other. During a session every participant has a 4 minute data with all the participants of opposite gender. After the date, both the participants rate each other on 6 attributes namely Attractiveness, Sincerity, Intelligence, Ambitious, Fun, Shared Interests and indicate whether he/she would go out again with the partner for a second date. If both the participants of a date indicated that they would like to go out again, it is considered a match.

The description of attributes
match - This is the target variable. It is a boolean binary variable and 1 indicates that it is a match and 0 otherwise
gender: Gender of the participant. Male or female
samerace: This is a boolean binary variable. 1 if both the participant and the partner belong to the same race and 0 otherwise
age_o: age of the partner
race_o: race of the partner
pf_o_att: partner's preference for attractiveness while looking for a date
pf_o_int: partner's preference for intelligence while looking for a date
pf_o_amb: partner's preference for ambition while looking for a date
pf_o_sha: partner's preference for shared interests while looking for a date
pf_o_fun: partner's preference for fun while looking for a date

dec_o: decision of the partner
attr_o: rating of attractiveness in the participant by the partner
sinc_o: rating of sincerity in the participant by the partner
intel_o: rating of intelligence in the participant by the partner
shar_o: rating of shared interests in the participant by the partner
amb_o: rating of ambition in the participant by the partner
fun_o: rating of fun in the participant by the partner
age: age of the participant
field: participant's field of study
race: participant's race
from: participant's country of origin
career: participant's intended career

Below are the attributes which describe the interest of the participant in the respective activities, on a scale of 1-10?
sports, tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv, theater, movies, concerts, music, shopping, yoga
exphappy: how happy does the participant is expecting during the speed dating event

attr1_*, sinc1_*, intel1_*, fun1_*, amb1_*, shar1_*: what does a participant look for in opposite sex. where attr - attractiveness, sinc - sincere, intel - intelligence, fun - fun, amb - ambitious, shar - has shared his/ her interests or hobbies

attr2_*, sinc2_*, intel2_*, fun2_*, amb2_*, shar2_*: what does participant think the opposite sex looks for in a date. where attr - attractiveness, sinc - sincere, intel - intelligence, fun - fun, amb - ambitious, shar - has shared his/ her interests or hobbies

attr3_*, sinc3_*, intel3_*, fun3_*, amb3_*, shar3_*: what does a participant think of himself/herself. where attr - attractiveness, sinc - sincere, intel - intelligence, fun - fun, amb - ambitious, shar - has shared his/ her interests or hobbies

attr4_*, sinc4_*, intel4_*, fun4_*, amb4_*, shar4_*: what participant think most of his/her fellow men/women look for in the opposite sex. where attr - attractiveness, sinc - sincere, intel - intelligence, fun - fun, amb - ambitious, shar - has shared his/ her interests or hobbies

attr5_*, sinc5_*, intel5_*, fun5_*, amb5_*, shar5_*: how does participant think others perceive himself/herself. where attr - attractiveness, sinc - sincere, intel - intelligence, fun - fun, amb - ambitious, shar - has shared his/ her interests or hobbies

dec: decision (1 - yes, 0 - no)
like: how much participant likes this person
prob: how probable does the participant think that the other person will say yes
match_es: how many matches the participant is expecting

4. IMPLEMENTATION

The Cross Industry Standard Process for Data Mining is used while implementing this project. The online dating industry in the United States has been improving at a faster growth and there has been approximately 36% increase in membership over the past two years. During the recent years the industry has been adopting to more machine learning techniques to improve the quality of their business.

The data being used for this project has been gathered from the students of Columbia University who volunteered to participate in the speed dating experiment. The data contains personal information about the individuals and does not defy any ethical considerations as this information is relevant and required so that the other participants can know about their potential mates.

The dataset comes with 195 attributes and all of these attributes are not relevant as most of these have redundant information or many missing values. The data has to be cleaned where the missing values and inconsistent values have to be handled. There are certain attributes which have to be normalized. After the data is cleaned and prepped for modeling, 5 classifiers will be trained on the data.

Before training the classifiers, feature selection will be performed to determine the best features that help in building a better classifier. Principle Component Analysis would be suitable for this data as the dataset has numerous attributes and PCA reduces the dimensionality of the data which results in simple models.

One of such classifiers is a OneR rule which uses the prior probability and determines the best attribute with minimum misclassification error and uses the attribute to predict the target variable. A Decision Tree is also a good classifier which is more complex than a OneR rule and uses multiple attributes to form a tree like model where the values of attributes determine the decision and consequences and predicts the target variable.

A Naive Bayes classifier also works well for this kind of data. The Naive Bayes classifier uses the historical data and computes the probabilistic inference to predict the target variable. The other classifier that will be used is 'k nearest neighbors' which uses the specified k number of neighbors and determines the majority class among the neighbors and predicts this majority class as the label for the target variable.

Another classifier that could be used is a Support Vector Machine. SVM also works well for this data as this problem is a binary classification problem and SVMs create a separation that classifies the instances into either of the two classes.

K fold cross validation technique will be used for validating the models while training. After the models are built, they are evaluated by computing the performance metrics such as accuracy, precision and recall. Accuracy gives us the percentage of correctly classified instances, precision gives us the fraction of instances that are actually a match from all the cases that were classified as a match and recall gives us the fraction of instances reported from the instances that were actually a match. The performances of all the classifiers are compared and the best classifier is chosen as the final model.

5. WORK DONE SO FAR

5.1 Data Cleaning

The dataset contains many unique identifiers which identify the participant, partner, the group of the participants, the number of the group etc. To avoid overfitting of the classifier all these unique attributes have been removed. Most of the participants did not mention their undergraduate school and the attribute 'undergra' has many missing values. To resolve this, this attribute has been removed. The attributes 'mn_sat' and 'tuition' have many missing values and these attributes have also been removed.

The attributes 'attr2_1, 'sinc2_1', 'int2_1', 'amb2_1', 'sha2_1' and 'fun2_1' represent what the participant thinks the opposite sex looks for in a date. These attributes can be removed because it does not matter what the participant thinks about other person's view while looking for a match.

The attributes 'attr3_1', 'sinc3_1', 'int3_1', 'amb3_1', 'sha3_1' and 'fun3_1' represent how the participant rates himself on the qualities. We have consolidated these 6 attributes and built a new feature named 'partnerSelfConfidence' which is the average rating of all the 6 attributes the participant gave himself. If the value for this attribute is a decimal, if the fractional part is greater than 0.5 then the ceiling of the value is considered otherwise the floor of the value is considered.

The attributes 'attr4_1', 'sinc4_1', 'int4_1', 'amb4_1', 'sha4_1' and 'fun4_1' represent what the participant thinks his/her fellow men/women look for in the opposite sex. These attributes might not be relevant for the classifier as these attributes provide redundant information that has already been stated by the individual participants.

The attributes 'attr5_1', 'sinc5_1', 'int5_1', 'amb5_1', 'sha5_1' and 'fun5_1' represent what the participant thinks about how others perceive him. These attributes have been removed because we think these attributes represent redundant information and also these values wouldn't be helpful in predicting a match as it is the participant's own perspective.

The attributes 'attr3_s', 'sinc3_s', 'int3_s', 'amb3_s', 'sha3_s' and 'fun3_s' represent how the subject would rate himself on the 6 qualities mentioned. These attributes have been condensed to create a new attribute named 'subjectSelfConfidence' which is the average rating of all the 6 attributes the subject gave himself. If the value for this attribute is a decimal, if the fractional part is greater than 0.5 then the ceiling of the value is considered otherwise the floor of the value is considered.

The attribute 'satis_2' describes how satisfied the participant was with the people they met. This attribute is not important for the classifier because while predicting a match between two individuals, the perception of an individual on the overall quality of all the individuals is not important. So, this attributed has been removed from the dataset. The attribute 'numdat_2', 'numdat_3' indicates how the participant feels about the number of speed dates he has had i.e, if the participant felt that the dates were too many or too less or right. This attribute can be removed because the participant's feeling about the number of decisions would not affect his/her decision about his partner.

The attribute 'match_es' represent how many matches the subject expects that he/she would get. The number of expected matches does not affect the decision of the participant. So, this attribute can be removed. The attributes that represent the information collected using a survey that was filled by the participants the next day after the date and 3 weeks after the date have been removed because they provide redundant information also these attributes have many missing values.

The attributes 'you_call' and 'them_call' represent the number of matches the subject has called to set up a date and the number of calls from matches the subject has received. These attributes have been removed because this information is not important as we are only worried about predicting a match between two people.

The attribute 'met' represents if the subject has met the partner before or not. Even if the subject has met the partner before, the subject's rating of the partner would not change because of it and thus, this attribute can be deleted.

The attribute 'dec' represents the decision of the subject about his/her partner. This attribute might cause bias and overfit the classifier. Hence, this attribute has been removed.

The attribute 'like' represents how much the subject likes his/her partner overall on a scale of 1 - 10. This attribute could also overfit the classifier and thus it has been removed.

The attribute 'expnum' represents how many people the participant expects would like him/her. This attribute does not provide any relevant information and so can be deleted.

Among the other attributes present, all the attributes for which more than 40% of the values are missing, such attributes have been removed. The values of the attribute 'income' are not specific to a participant. These values are determined based on the zip code and the median income of each zip code is used. The attribute 'int_corr' represents the correlation between participant's and partner's rating of interests. This attribute could overfit the classifier and has been removed.

5.2 Data Preparation

The attributes 'pf_o_att', 'pf_o_sin', 'pf_o_int', 'pf_o_amb', 'pf_o_sha' and 'pf_o_fun' are used for specifying the preferences for qualities while looking for a mate. Some of the participants rated the preference for each of these qualities on a scale of 10 while the other participants distributed a total of 100 points among these qualities. The values for these attributes have been normalized to a scale of 0 - 1 to maintain consistency. This normalization is done by dividing the value of each attribute with the total of all the 6 attributes mentioned.

The attributes 'attr1_s', 'sinc1_s', 'int1_s', 'amb1_s', 'sha1_s', 'fun1_s' , 'attr3_s', 'sinc3_s', 'int3_s', 'amb3_s', 'sha3_s', amd 'fun3_s' also have the same problem where some participants used the scale 1 - 10 and some distributed 100 points over the attributes. These values are normalized in the same way by dividing the value of each attribute with the total of all the 6 attributes mentioned.

After normalizing the values, if the values are decimals, if the fractional part is greater than 0.5 then the ceiling of the value is considered otherwise the floor of the value is considered.

The attribute 'field' which represents the participant's current field of study has some inconsistent values. For example, some participants whose filed of study is 'Law', it is represented as 'Law' in some cases and as 'law' in other cases. To avoid inconsistency, all the values have been changed to 'Law'.

The attribute 'career' that represents the intended career of participants also has inconsistent values. One such case is where the intended career of participants is 'Law', it is represented as both 'Law' and 'Lawyer' in multiple cases. This has been handled by using the value 'Lawyer' for all such instances.

The remaining attributes for which there are missing values, the missing values are filled withe mode of the values for that attribute.

6. WORK TO BE DONE

The data is cleaned and prepped for further tasks. Principle Component Analysis has to be performed to reduce the dimensionality of the data and to find the best features that help in building a better classifier. All the five classifiers have to be trained and parameter tuning should be done to build good classifiers. The performance of all the classifiers have to be evaluated and the best classifier will be selected.

7. REFERENCES

[1] Explicit and implicit user preferences in online dating.

[2] For love or money? the influence of personal resources and environmental resource pressures on human mate preferences.

[3] Gender differences in mate selection: Evidence from a speed dating experiment.

[4] Racial preferences in dating.

[5] Speed dating experiment.

[6] What makes you click?-mate preferences in online dating.