Team Project: Speed Dating

**Team 31** Section 10B

Team Members:   
Lau, Eric  
Lu, Wen  
Oteng-Nyame, Jason  
Phatraprasit, Sun

Business Understanding

Speed dating and online matchmaking is much more common place in the world today and is no longer a taboo subject. Most dating services use a questionnaire to hone in on your characteristics as well as desired characteristics in an attempt to match you with a person you may like to date. This billion-dollar industry relies on the accuracy of matches so that users can feel confident that the dates they go on can potentially lead to a lifelong partner. We need to build a speed dating business that is reputable and has a history of successful matches so that we can increase participant enrollment. If our services have a higher success rate, we can charge more per event and add additional services to increase per participant revenue.

The current business model is to assign people to random waves/group per night since people are limited in the number of people they meet per night. This randomized placement does not account for any features of the participant, who the participants are, or their preferences. By utilizing data mining, we can potentially create combinations of waves/ group of people that have a greater chance of matching. The data mining would predict a yes or no of two individuals matching based on a predetermined questionnaire before the event. Each individual would be rated against all other individuals of that night based on a match outcome. Then, we would seek to maximize the number of match connections within a specific wave, which would increase the number of matches at the end of the event.

Data Understanding

We acquired the data from Kaggle: Speed Dating Experiment Dataset. The dataset was gathered from 552 participants in an experimental speed-dating event from 2002-2004. To support our data mining process and business problem goals, we only used features that were recorded before the dating event started. Data includes attributes about the participants such as:

* Demographics
* Lifestyle information
* Beliefs on what others find valuable in partner
* Objective of participating
* Dating habits
* Self-perception of the participants
* Likes and Dislikes of activities

At the end of the event, participants were asked to rate their date on attributes such as:

* Attractiveness
* Sincerity
* Intelligence
* Fun
* Ambition
* Shared Interests/Hobbies

To understand how match occur, we used association rules to identify the factors that contributed to the decision-making process of men and women. We decided to look at these two groups separately by dividing the dataset by gender, and then converted the data type from ordinal to nominal. We then split each data set further into two different parts, where the decision is yes (dec\_o = 1) and the decision is no (dec\_o = 0). Association rules were applied to these four parts to determine what mattered the most when it comes to decision-making. We have included short summaries of the association rules outcome and the top 24 lifts are in the appendix.

1. Men’s Decision

This part of the association rule only includes the male segment of the data, the LHS is the attribute combination that determined the decision, and the RHS is constrained to either dec\_o = 1 or dec\_o = 0.  (note: for rating scales, we converted them from 0-100 to low, medium and high for easier interpretation)

1. For example, for the first association when men said yes, men were looking for higher/ a lot of fun in their partner, women with medium theater experience, and medium movie interest, and men who are okay that their partner is a different racial background. These aspects are combined and contributed to men’s decision of saying yes.
2. For the highest association for when men said no, women had a high interest in sports, low interest in watching TV sports, medium interest in going to the museums and low yoga interest.
3. Women’s Decision

This part of the association rule only includes the female segment of the data, the LHS is the attribute combination that determined the decision, and the RHS is constrained to either dec\_o = 1 or dec\_o = 0.  (note: for rating scales, we converted them from 0-100 to low, medium and high for easier interpretation)

1. For the highest association for when women said yes,  women are looking for men that do not necessarily have to be attractive, with medium sincerity, medium intelligence, doesn’t have to be super fun, but with medium ambition.
2. For the highest association when women said no, women did not want men with low sports interest and low interest in museums.

Data Preparation

We encountered many blank fields and  inconsistencies with responses in the dataset. We discovered that half of the dataset had more than 20% of the data missing in their corresponding fields. We assumed that this could be due to the nature of survey questions and the participants’ laziness. The data was cleaned in the following way:

* Populated the missing id using the iid of the person.
* Changed 0’s in zip code to NAs.
* Changed the male and female attribute to M and F.
* Input the missing pids using the partner and wave features.
* Separated data to ensure female is always iid and male is always pid.

There are other variables that have missing information, but we are unable to input anything because they are personalized information specific to the participant. For our purposes, we decided to exclude blank or NA fields for analysis and our final dataset contained.

Modeling

Below, we model the events that could occur and create a model framework. For customer with features and customer with features the decision would be to Match , if both say yes to each other, or Not Match , if at least one of them says no. The goal is to maximize the total number of matches and the decision variable is to create the waves of groups for the speed dating. We assume that the current placement into different waves is random and that we will have the relevant feature information about each participant before we create the waves. Let’s assume there are number of participants in each wave and number of waves.

Therefore we want to decide placement into different waves in order to maximize:

If participants are assigned into waves randomly then

And since we know that the for our baseline model we assume which means :

Therefore to maximize we can place participants into waves such that:

Or

That is we create waves such that the expected number of matches in a given wave is greater than the expected number of matches in a wave if waves were assigned randomly. We select who goes into each wave based on their probability of matching as opposed to doing it randomly.

To do this we must first estimate the probability of matching based on the features for each pair irrespective of the wave . We simply need to be able to beat the null baseline model in order to achieve a higher number of matches after creating our waves. Then we use these predicted probability values to create waves such that the expected value of total matches given features for both participants and each wave is greater than the expected value of total matches if waves were randomly assigned.

Speed dating is built on the number of matches and the historical success rate of the company. The speed-dating world thrives on word of mouth and individuals willing to give it a try. We can improve speed-dating outcomes and increase our match rate by grouping our participants based on our algorithm match/non-match outcome. As our match rate increases, the company’s reputation is built up and more and more individuals are willing to participate in a speed dating night with our company. We can analyze how our match rates affect our overall business through the number of speed dating nights or the number of participants. If we see an overall positive correlation between the number of matches and participation rate, we can determine if more matches brings increased revenue.

Our return on investment can be calculated with overall metrics of company performance. We can record number of matches, number of participants per night, new participant rate, number of speed dating nights per week/ per month and revenue per participant.

Our goal is to identify the probability that two participants will match given the features we have about them. As we want to classify people into two categories, match and no match, our core data-mining task is classification.  Predictive modeling will give us the probability of a match to help us determine our results. We considered several classification models to solve our problem: Decision Tree, SVM, Random Forest and Neural Networks. Before running the more complicated models (SVM etc), we balanced our data using the smote function, which artificially creates new entries of the minority class (in this case the minority class was a 1 for a match) using the nearest neighbor of the observations.

Decision Trees: We first considered the data mining method of decision trees because it is the most simplistic method of classification and easy to interpret. Decision trees can handle categorical and continuous variables, which was a positive because our dataset contain vast information of person preferences and ratings. In addition, decision trees implicitly perform variable/ feature selection, which is useful to narrow down our variables.  However, we encountered that the decision tree would suffer with other aspects of our data. The decision tree suffered from too much information because of the inclusion over 15 variables in our model and lost the advantage of being easily understood. Also, the decision tree performed poorly because most of the data we had were continuous variables such preference ratings(1-10) and personal ratings (1-100). For these reasons, we do not implement decision trees in our final model selection.

After exploring decision trees, we explored Random Forest as an option to negate the disadvantages of decision trees. We utilized Random forests because it could handle our categorical and continuous numerical variables easily without interpretation, it automatically performs variable/feature selection, and implementation time and optimization is easily achieved. In addition, random forests can procedure highly accurate classifiers and can handle our input variables with ease. As a disadvantage, random forests are known to overfit the data, however, we will be using a training and out of sample data set (test set) to ensure we do not overfit the data.

Support vector machine(SVM) was the third data mining method considered because of the improvement over logistic regression. SVMs provide more optimal boundaries, leading to better performance, and can theoretically classify our data better. Another positive would be that SVMs rely on a subset of points for an approximation of a boundary, making it useful in our very convoluted data. SVM was tested as a model to utilize and was optimized to increased match and non-match rates.

Lastly, we decided to implement neural networks as a black box method because of the complexities of human behavior.  Neural networks are able to implicitly detect complex nonlinear relationships between dependent and independent variables, detect all possible interactions between our outcome variable, and relative ease of use. The disadvantages include its "black box" nature and proneness to overfitting. As a result, the neural network provides ability for the modeler to maximize predictive capabilities of complex human characteristics with high accuracy, but unknown reasons.

Ultimately, we want to maximize the number of correct matches and correct non-matches we can predict in order to create waves/ groups of people that can match. Random forest, SVM, and neural networks provides predictive matches and non-matches in which we can use to create wave/groups of participants that are matches based on communities or clusters of participants.

Evaluation

We evaluated the models using a confusion matrices because it provides a standard evaluation method to view performance of our binary classification prediction. The confusion matrix provides insight into the errors being made by our algorithms and, more importantly, the types of errors that are being made. With the matrix, each algorithm’s outcome will have a true positive, false positive, false negative, and true negative results. These results can be aggregated to create our key performance indicators (KPIs). We propose to use the following KPIs:

1. **Area Under the Curve (AUC):** The AUC is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.
2. **Accuracy:** Determines the overall effectiveness of the algorithm based on true positive and true negative outcomes.
3. **Sensitivity:** Determines how many matches were correctly identified. (True Positive Rate)
4. **Specificity:** Determines how many non-matches were correctly identified. (True Negative Rate)

We used an out of sample data set to generate the following performance statistic.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AUC | Accuracy | Sensitivity | Specificity |
| Random Forest | 97% | 91.4% | 99% | 85% |
| Support Vector Machine | 94% | 85.0% | 90% | 84% |
| Neural Network | 90% | 85.8% | 87% | 80% |
| Baseline | N/A | 83.16% | 0 | 1 |

The baseline model predicts all zeros for matches (based on the average value of 0.165) and this is worse than any of our other models because it doesn’t add any extra information to our decision on how to create the waves and thus the waves will still be created randomly using the null model. All our models estimate a probability of matching and then based on a probability threshold which maximizes the AUC, we predict matches. We settled on the Random Forest as our final model because it has the highest value for all our metrics of performance. More importantly it provides us with a high sensitivity which means we can make more correct predictions of matches which we can use to decide which participants should be in each wave.

Deployment

The key to our data mining strategy is to predict the match probabilities between a couple. First, our algorithm runs on the information initially collected from the participants when they signed up. With information about themselves and what type of person they are seeking, we can predict the probability of a couple matching and based on the probability threshold to maximize AUC, we predict matches for each couple. Since the current business solution is to randomize the waves of people meeting, a group is not maximized based on potential matches. In our proposed speed dating setup, we can tailor the waves to maximize the number of potential matches.  We will utilize network optimizations strategies to create waves of participants that are maximized by modularity. Our goal is to maximize number of matches (edges) within groups in our network and have a good division of participants (nodes) into different waves (communities).

We deployed our data mining matching algorithm on an out of sample list of participants matched to all participants to identify all combinations of potential matches. The list of matches of participants was analyzed to ensure we did not have any participants without any matches. If we had participants without matches, we would include them in to clusters with lower number of participants then randomly in different clusters. Based on 516 participants in out of sample set, we selected 26 clusters to divide each wave/ cluster into roughly 20 participants. We considered two options fastgreedy.community and walktrap.community, but chose fastgreedy.community because it tries to optimize modularity through maximizing separate communities. The fastgreedy.community method is fast and does not require an parameter turning. We did not utilize walktrap.community because of a natural tendency to segregate nodes with single matches and group large clusters with many more matches.

Results

This improvement centers on the main purpose of speed dating, which is to obtain matches. If more participants obtain matches, our speed dating company’s reputation increases, thereby causing an increase in referrals and new participants.

1. Overall total of match percent
2. Average matches per cluster

Risks and Considerations

There are many aspects of the business that could be affected if we implement our data mining algorithms, ranging from logistical to satisfaction. By grouping the waves, we would have to propose a date for the participants to come instead of him choosing a date himself. This could cause a scheduling conflict for the participants and could be placed in a suboptimal wave. Also, participants may want to meet a wide variety of people, do not know what their preferences in a partner, and do not want to restrict their options to just those with a “high probable” match with them. In this situation, our data mining would not correctly match participants, which could lead to bad reviews if the participants feel that they are only meeting certain types of people. In addition, if a participant does not know what their preferences are or incorrectly fills out the survey information, they can be placed in a suboptimal group and cause the same reputational issues. Also, we don’t foresee any serious ethical concerns with our deployment as we are increasing the chance of people getting their match, which is their true objective for using this application.

We could mitigate the problems mentioned above by being as transparent about our data mining process and allow the participant the option to opt-out of the prediction matching process. If a participant opts-out, we can place them randomly in a cluster or let the participant choose a day if they insist on picking a day themselves. Also, we can ensure the participant’s privacy by using de-identified data in all data mining activities, ensuring that their data safe and protected.

Appendix

Data understanding

I.               Men’s Decision

This part of the association rule only includes the male segment of the data, the LHS is the attribute combination that determined the decision, and the RHS is constrained to either dec\_o = 1 or dec\_o = 0.  (Note: for rating scales, we converted them from 0-100 to low, medium and high for easier interpretation)

a.     Men Said Yes

This is when RHS is equal to dec\_o = 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LHS | RHS | Support | Confidence | Lift | Count |
| Fun3\_1=H, theater=M, movies=M, samerace=0 | {dec\_o=1} | 0.01039861 | 0.9333333 | 1.973682 | 42 |
| fun3\_1=H, sports=H,  theater=M, movies=M,  samerace=0 | {dec\_o=1} | 0.01039861 | 0.9333333 | 1.973682 | 42 |
| hiking=H, theater=M,  movies=M | {dec\_o=1} | 0.01015103 | 0.9318182 | 1.970478 | 41 |

The way to interpret this table would be for rule 1, men looking for higher/ a lot of fun in their partner, women with medium theater experience, medium movie interest, and men who’s okay with their partner has a different racial background than they’re combined contributed to men’s decision of saying yes.

b.     Men Said No

This is when RHS is equal to dec\_o = 0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LHS | RHS | Support | Confidence | Lift | Count |
| Sports=H, Tvsports=L,  Museums=M, Yoga=L | {dec\_o=0} | 0.01039861 | 0.9545455 | 1.810901 | 42 |
| Sports=H, Tvsports=L,  Art=M, Yoga=L | {dec\_o=0} | 0.01039861 | 0.9545455 | 1.810901 | 42 |
| Sports=H, TVsports=L,  Art=M, Concerts=H | {dec\_o=0} | 0.01039861 | 0.9545455 | 1.810901 | 42 |

The way to interpret this table would be for rule 1, women with a high interest in sports, low interest in watching TV sports, medium interest in going to the museums and low yoga interest combined contributed to men’s decision of saying no.

II.              Women’s Decision

This part of the association rule only includes the female segment of the data, the LHS is the attribute combination that determined the decision, and the RHS is constrained to either dec\_o = 1 or dec\_o = 0.  (note: for rating scales, we converted them from 0-100 to low, medium and high for easier interpretation)

a.     Women Said Yes

This is when RHS is equal to dec\_o = 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LHS | RHS | Support | Confidence | Lift | Count |
| Attr1\_1.1=L, Sinc1\_1.1=M,  Intel1\_1.1=M, Fun1\_1.1=L,  Amb1\_1.1=M | {dec\_o=1} | 0.005942065 | 0.9600000 | 2.630556 | 24 |
| Reading=M, Movies=H,  Yoga=M, Sinc1\_1.1=M,  Amb1\_1.1=H | {dec\_o=1} | 0.005446893 | 0.9166667 | 2.511816 | 22 |
| Since3\_1=H, Exercise=M,  Gaming=H, Yoga=H,  Amb1\_1.1=M | {dec\_o=1} | 0.005199307 | 0.9130435 | 2.501888 | 21 |

The way to interpret this table would be for rule 1, women are looking for men that doesn’t necessarily have to be attractive, with medium sincerity, medium intelligence, doesn’t have to be super fun, but with medium ambition combined contributed to the decision of women saying yes.

b. Women Said No

This is when RHS is equal to dec\_o = 0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LHS | RHS | Support | Confidence | Lift | Count |
| sports=L, museums=L | {dec\_o=0} | 0.005199307 | 1 | 1.574659 | 21 |
| museums=L, yoga=H | {dec\_o=0} | 0.005199307 | 1 | 1.574659 | 21 |
| museums=L, gaming=H | {dec\_o=0} | 0.005199307 | 1 | 1.574659 | 21 |

The way to interpret this table would be for rule 1, men with low sports interest and low interest in museums combined contributed to women’s decision of saying no.

Modeling



