**Final Report**

**Analyzing Factors that Influence the Success of a Date**

**Group 4: Aikun Tao, Dantian Liu, Shaka Lohardjo, Xuefan Han, Zhilin Zhang**

**APAN 5205**

**Instructor: Birol Emir**

**April 27, 2022**

**Statement of Purposes**

First organized in the late 90s, speed dating is still one of the most popular ways for people to meet potential romantic partners. While it is easy to participate in a speed dating event, the hard part is making it to a second date.

There are a lot of factors that influence the success rate of a first date, for example, one’s appearance, hobbies, and personality. Therefore our group decided to find out what factors are most important to a mutual match and how well we can predict if two people will match after their first date. In this project, we will be using data from Columbia Business professors Ray Fisman and Sheena Iyengar’s speed dating event. The data collected an individual’s experience at a speed dating event, including factors like race, career, and ratings on personality traits. Our team will analyze the data from participants in experimental speed dating events to explore the relationship between characteristics and attractiveness. This will be a supervised learning project. We are going to examine variable importance, use PCA to meltdown features and continue with less and more meaningful fields, and lastly use descriptive and predictive analysis to support our findings.

The research questions are

1. What factors will result in another date?
2. How to predict if two people will match after the first date?

**Data Discussion**

The data was collected from 552 participants in the experimental speed dating events from 2002 to 2004, and the process included signup, halfway through the meeting, follow-up after a few days, and follow-up after 3-4 weeks. The events were held in 21 waves with a varied number of participants, and each participant was required to fill out the same survey and questionnaire in each wave. The participants were asked to rate their date on multiple attributes: attractiveness, sincerity, intelligence, fun, ambition, and shared interests. The data gathered from the questionnaire included the fields of demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information.

The original dataset contains 8378 rows and 195 columns. There are 8 character variables with no missing values and 187 numeric variables with many null values. Variables can be categorized into the following groups: Unique identifiers, Study setting variables, Demographic variables, Psychographic variables, and Characteristic variables. Our team believes that this database is worth studying and exploring, as it contains a large amount of data and a large number of variables, which can reveal more secrets about dating than what’s already known.

Some variables are more worth noting in terms of understanding the basic demographic information of the experiment. For example, combining variables about the decision of the partner and matching results, we find out that only 16% can find mutual matches where both liked each other, 51% had a one-sided match while 33% didn't match at all. We also figure out that gender distribution is quite balanced with 49.73% female and 50.27% male. As for the age distribution, female participants tend to be younger than male participants in general, but the age range is similar. Among all participants, the European/Caucasian-American category takes up the largest part at about 55.1%, followed by Asian/Pacific Islander/Asian-American at about 24.6%.

**Data Preparation**

***Data cleaning***

We first removed the duplicates in the dataset. Then, we subset the data to only include relevant factors for our research problems. To ensure accuracy in our later analysis, we removed columns with more than 1000 null values. After doing so, we imputed the mistakes in our data. For character columns, we imputed the remaining missing values with the mode of each column. For numeric, we used the mice package to impute the missing values.

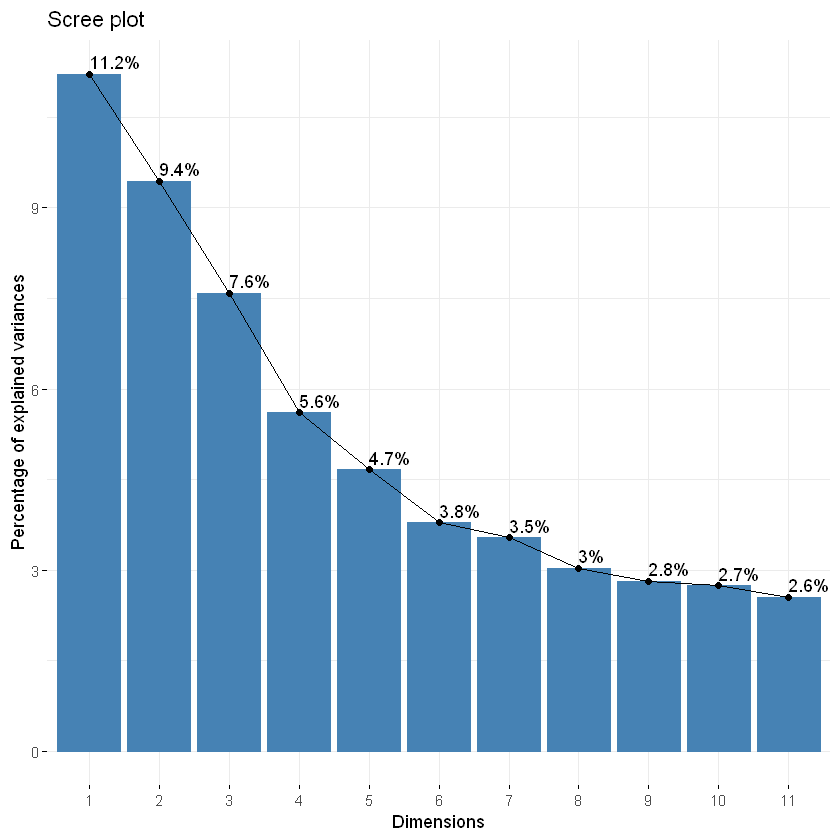
***Variable Importance***

This was the first of two techniques we used to attempt to reduce the number of variables and find the most important ones. Apart from some obvious factors like ‘like’ for ‘if one likes their date’ and ‘attr’ for attractiveness of their date, other most important factors include ‘fun’, ‘sinc’ for sincerity, and ‘intel’ for intelligence.

***Principal Components Analysis (PCA)***

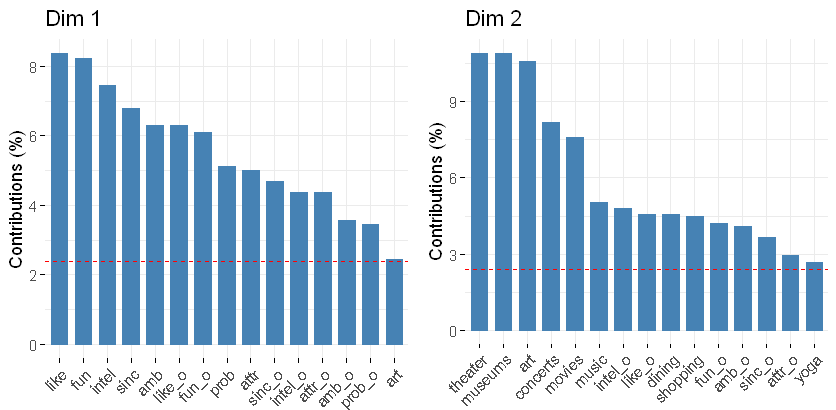
Since we had a very large number of features, we decided that PCA was suitable to reduce the variable dimensions into components that were related. This is the technique we used to get a deeper understanding of the relevancy of and correlation between factors.

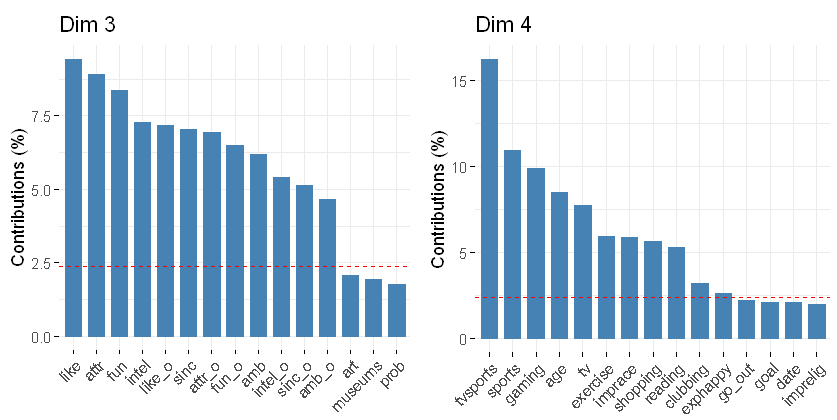
Based on the ‘elbow’ on a scree plot, 4 dimensions may be the appropriate number of dimensions to reduce to. Thus, we ran a principal components analysis with 4 dimensions.

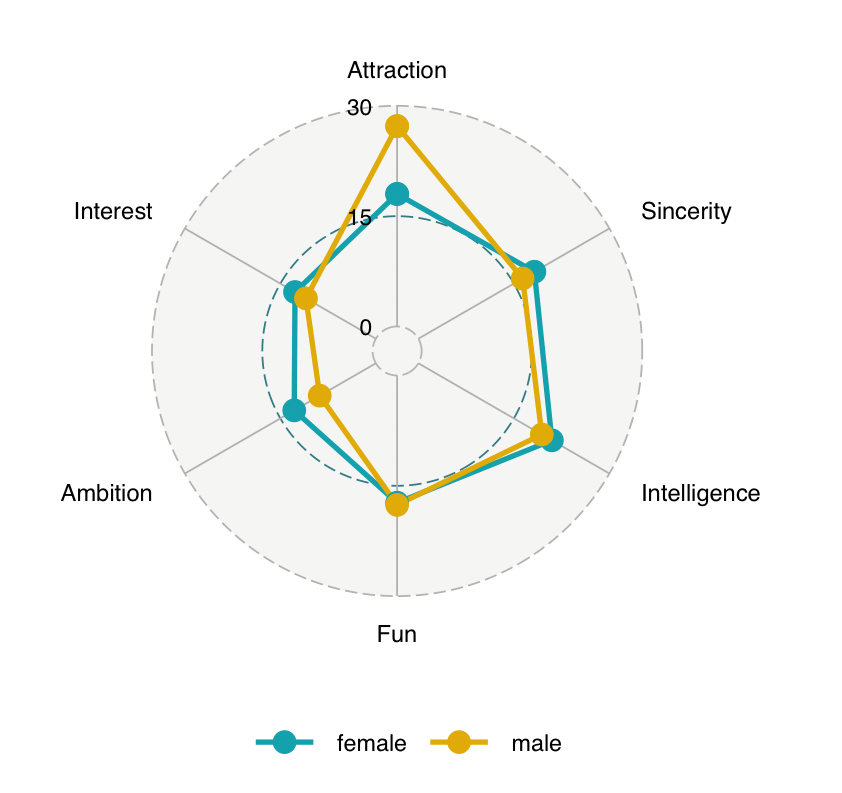
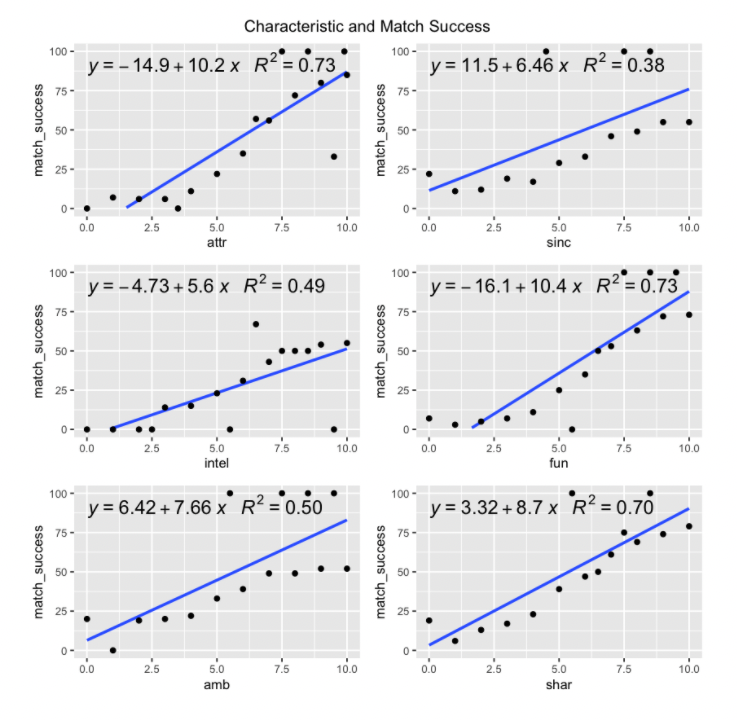


Next, we examined the elements comprising each component. Contributions of each variable to each component are charted out below. For dimensions 1 and 3, the factors that contributed the most are fun, intelligence, sincerity, and attractiveness. This was very much like what we found out in the variable importance analysis. For dimension 2, hobbies like going to the museum and theater and art are the most contributing elements. For dimension 4, the most important factor is ‘tvsports’ which reflects if the person likes to watch sports on television. This factor surpassed the second place by a large margin. Gaming and doing sports were of great importance in this dimension as well. Dimensions 1 and 2 are the most contributing overall.

Therefore, it seems like personality factors like fun, sincerity, and intelligence are important in dates, and hobbies like watching sports, doing sports, and gaming are also contributing characteristics. We could also find that race and age were of relatively low importance in all dimensions.



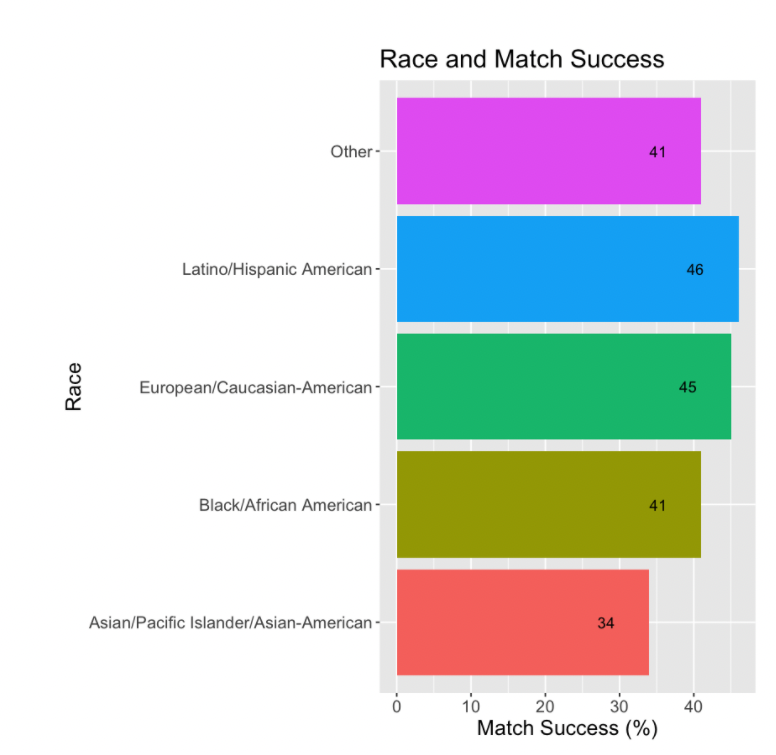
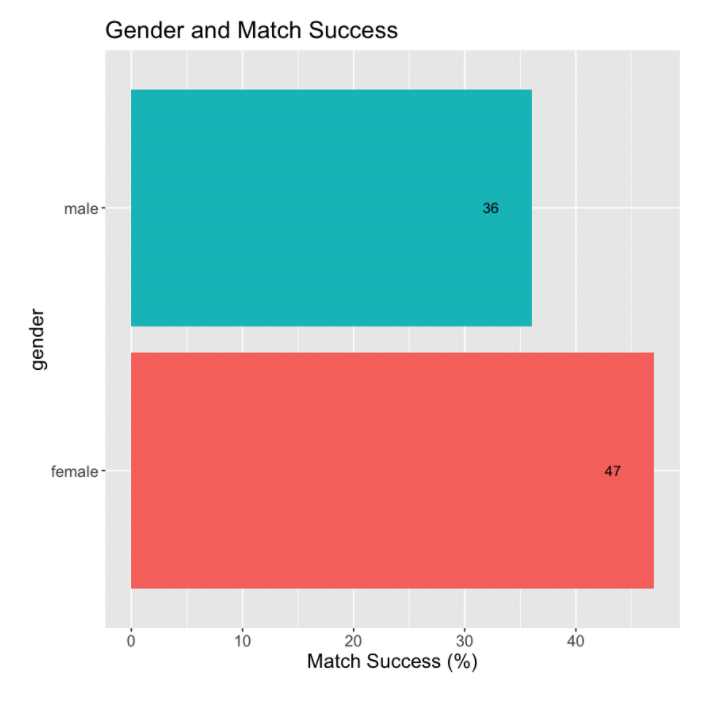
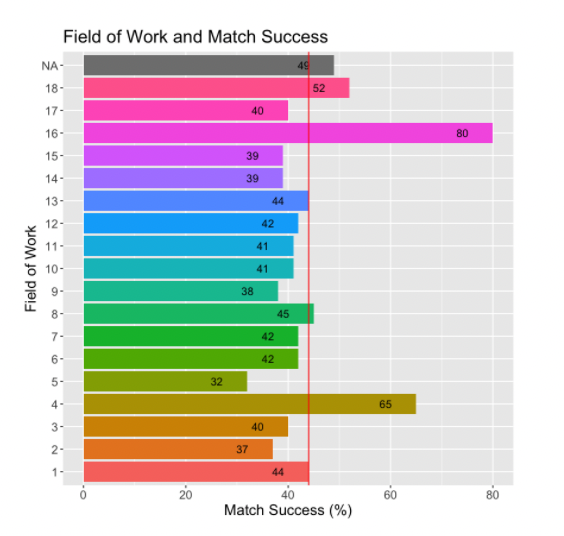


**Descriptive Analytics**

***Characteristics Importance***

Prior to the event, attendees were asked to rate the characteristics they look for in a significant other. Viewing the spider plot, for males, attractiveness is the most important and ambition is the least important. Meanwhile, for females, attractiveness, sincerity, fun, and intelligence are tied as the most important.

In practice, when people went on the actual date we can see the correlation between the characteristics and second date success. The results show that there are only a few characteristics that have a strong positive relationship with date success. The important characteristics for the entire population were attractiveness, fun, and shared interest with an R^2 of 0.7 or higher.

***Success of Match by Race, Gender, and Field of Work***

We wanted to compare demographic variables such as race, the field of work, and gender. If any demographic group experiences higher success in getting a date.

For race, we can see that “Latino/Hispanic American” experience the highest success of matches with 46% success in attracting their date. While “Asian/Pacific Islander/Asian-American,” experienced the lowest success at 34%.

For the field of work, the most successful field is Language (16) at 80% and Medical Sciences (4) at 65%. While the field of work that encounters the least success in Engineering (5) at 32% and Math (2) at 37%. For gender, in general, females experience more success at 47% than males at 36.

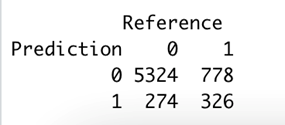
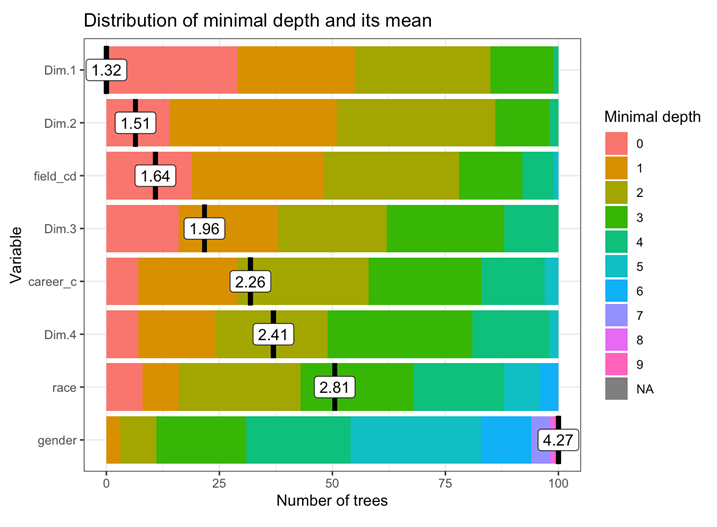
**Predictive Analytics**

After PCA, the number of variables was reduced to 4 dimensions combined with factor variables ‘field\_cd’, ‘career\_c’, ‘race’, and ‘gender’. The data was then split into a train sample with 80% of the data and a test sample with the rest based on the ‘match’ column for prediction. We tried 4 models in total and finally chose Neural Network.

***Random Forest***

We used Random Forest to find the relationships between the factors that may influence the matching result of people after a first date. The number of trees is set to 100 and the number of variables tried at each split is 2. The classification error in ‘matched’ is 4.89%, and ‘not matched’ is 70%. According to the confusion matrix, out of 5598 not matched people, 5324 of them were correctly classified as ‘not matched’, and 274 were mistakenly classified as being matched. The classification error implies that the model predicts well on people who are not being matched, and is poor at predicting people who get matched. The overall accuracy on the train set is 84.0%.

The plot below shows the distribution of minimal depth among the trees of the Random Forest model. Usually, the smaller the mean minimal depth, the more important the variable is. According to the plot, Dim.1 has the lowest mean minimal depth.

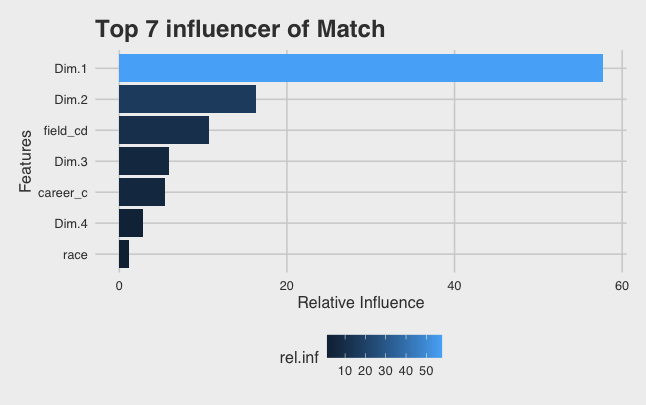
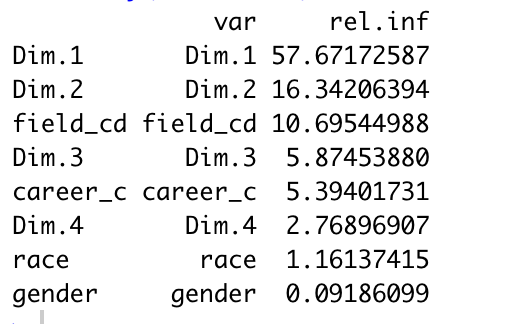


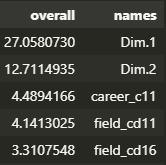
***Gradient Boosting Machines***

Next, we used the GBM model. Two models were tried to fit the train set, and the one with the highest accuracy (84.01%) was built with a total number of trees equaling 1000, and the depth of tree equaling 3. The model also includes cv.folds to perform a 5-fold cross-validation.

The results show that the model has a minimum CV RMSE of 0.841. This means on average the model is about 0.841 off from the actual matching result. Since the value of the dependent variable ‘match’ only has two levels (0 or 1), the model does not predict well based on the minimum CV RMSE.

An important feature in the GBM modeling is the variable importance. The result table below ranks the individual variables based on their relative influence, which is a measure indicating the relative importance of each variable in training the model. The result of the model shows that Dim.1, Dim.2, and field\_cd are the top 3 variables that have the largest influence on the match result.

***Logistic Modeling***

We then used a logistic regression model to predict matches. The accuracy of a GLM model is about 83.8%. The chart on the right shows the top five important variables for this model. Dimensions 1 and 2 contributed the most, and factors ‘career’ and ‘field’ were relatively more important than the rest. However, overfitting is a problem with this model as we have more parameters than our data set has observations. Therefore, the prediction may be misleading due to a rank-deficient fit.

***Neural Networks***

We chose the neural network to make predictions of the match result. We attempted by using basic neural network models with nnet, using h2o deep learning framework, using h2o deep learning framework with manual-adjusted hyperparameters, and using h2o deep learning framework with random-searched hyperparameters. We also approached each method by using all the variables and using 4 component variables that were derived from PCA.

First of all, we used a basic neural network model with nnet to predict the match result based on all the variables including 4 PCA components, field\_cd (field of study encoded), career\_c (job categories encoded), race, and gender. We predicted on the test sample and acquired a 0.840 accuracy. Then we teased out encoded factors and used the 4 PCA components to predict the result, acquiring a 0.842 accuracy.

After, we used h2o deep learning framework and used all variables to predict the match result, obtaining a 0.786 accuracy rate. We also teased out factor variables and predicted again, obtaining a 0.788 accuracy rate.

Furthermore, we manually adjusted the hyperparameters and used a different number of nodes and hidden layers, obtaining 0.810 and 0.814 accuracy rates respectively when using all variables and using only PCA components.

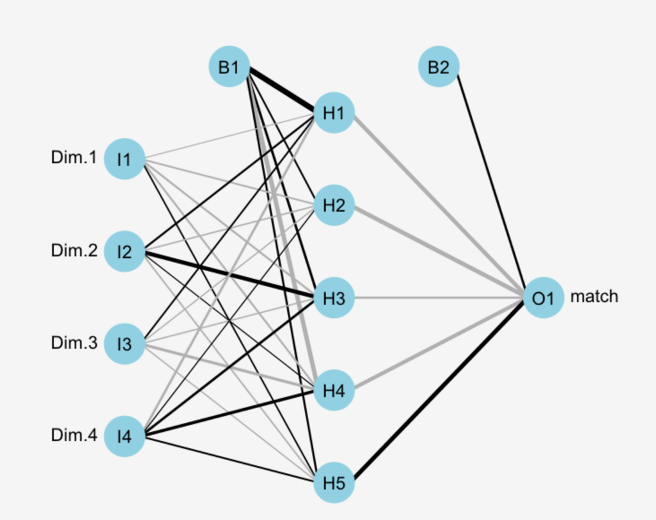
Last, we used random search to tune the hyperparameters, obtaining 0.802 and 0.837 accuracy rates respectively when using all variables and using only PCA components.

Here is a results comparison chart of the models constructed:

| Neural Network Models | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Package | nnet | nnet | h2o | h2o | h2o | h2o | h2o | h2o |
| Variable | all | 4 | all | 4 | all | 4 | all | 4 |
| Tuned (Y/N) | N | N | N | N | Manual | Manual | Grid  Search | Grid  Search |
| Accuracy | 0.840 | 0.842 | 0.786 | 0.788 | 0.810 | 0.814 | 0.802 | 0.837 |

We concluded that the model with the best accuracy is the one using the nnet package with default parameters and 4 PCA components as input variables. (Plot is as shown below)

Neural network models are flexible and inclusive, which are reliable to deal with problems with many features. It can split the task into layers and nodes of simple elements and can run fast after it has been trained. On the other hand, they are computationally expensive and time-consuming to run. They also heavily rely on training data and are prone to overfitting. On top of that, they are more like black boxes that cannot be easily interpreted. In this sense, we can use neural networks to yield accurate prediction results but should choose other models to have better interpretability regarding the importance of the chosen variables.



**Conclusion/Recommendations/Limitations**

After using various methods of analysis and experimenting with different models to predict the match result, we have some interesting findings regarding the success of the first date and significant factors influencing another date.

By calculating the variable importance, we found out that personality traits play crucial roles. Apart from obvious factors such as ‘if one likes his/her date’ and general attractiveness of the date, attributes such as fun, sincerity, and intelligence weigh high among other factors.

While performing dimension reduction using PCA, we found out that Dimension 1 was an important factor in other models. Within this dimension, it weighed that the most critical component as fun, intelligence, and sincerity. On top of that, Dimension 2 was important too with high weights in activities like museums, theater, and arts.

Preferences of different genders vary. While the most important attribute is attraction and the least important is an ambition for males, the most important is intelligence, attraction, and sincerity, and the least is shared interests and ambitions for females.

As for demographic factors, we noticed that Latino/Hispanic Americans generally have the highest success rate, however, Asian/Pacific Islander/Asian-Americans have the lowest. When partitioned by field of work, we figured that people in the field of work in Language, have a significantly higher success match rate, followed by Medical Science/Pharmaceuticals. For language, we can infer that it is due to the ability to communicate whether in multiple languages or conversational. The lowest-performing fields are engineering and math.

We also found out that females have a higher rate in terms of match success when compared to males.

For our target audience, we have the following suggestions:

Traits like attractiveness, fun, sincerity, intelligence, ambition and shared interest are important in regard to a successful match. Among the traits, an increase in fun attributes and attractiveness will increase match success most significantly. For instance, a one-point increase in fun results in an increase of 10.4% in matching success, while a one-point increase in attractiveness results in an increase of 10.2% in matching success. Other traits like shared interest, ambition, sincerity, and intelligence give 8.7%, 7.66%, 6.46%, and 5.4% respectively.

For characteristic suggestions, we would recommend:

* Having excellent communication skills might help boost the relationship.
* Caring about your appearance can help increase your attractiveness.
* Having a wide variety of interests to share experiences with others can help boost your attractiveness.
* Adopting a lively lifestyle and having an interest in activities like going to the museum, theater, and interest arts are helpful.
* If your area of study is in Language or Medical Sciences, you have a big advantage.

There are certain areas that we can work on to improve our project outcome:

* Neural Network Model
  + Though our Neural Network Model can obtain high accuracy, it has a black-box nature, meaning that it is hard for us to interpret the result. We can further improve by trying out other models and finding the one that has the best trade-off between accuracy and interpretability.
* Untidy but meaningful data
  + Data in “From” is really messy with different types of data like city or state or country inputted. Similarly “Zipcode” is also inconsistent and mostly empty. This makes it very difficult to clean since we have to make interpretations.
* Loss of track and follow-up
  + Our scope only looks at during the event, we did not consider long-term impact after the event. The data for the follow-up contact had a significant drop in responses.
* Loss of timeliness
  + Some survey data are obtained after 3 - 4weeks, meaning that they might have changed their perceptions after the event.