SENG 474 Final Project

Mining For Love:

A Statistical Analysis of Speed Dating

Stephen Scinocca V00809689
Austin Beauchamp V00696969
Kai Chao V00830178

1 Introduction

The modern era has seen an increase in online dating services, allowing individuals to search, analyze, and even talk to other singles. Many sites attempt to remove user labour by only bringing together algorithmically-determined pairs, or at the very minimum displaying a metric predicting compatibility. We investigate a speed dating dataset. A “speed date” consists of a short 2-5 minute meeting where candidates briefly talk and decide whether to pursue a future date or not. Each candidate provides numerical data about themselves and rates their date in a number of categories. This project aims to investigate trends amongst the speed dates, and most importantly, build a model with the ability to predict the probability of a successful date.

2 Dataset

The dataset for this project was found on Kaggle [1] from two Columbia University professors. There are 22 different speed dating events from 2002 to 2004, totalling over 8000 total dates with over 100 features per date. The number of dates per person ranged from 5 to 22. The applicants filled out various pieces of information before, during, and after the events. Information such as: their own interests, what they thought of themselves, what they thought of their current date, and their decision for each date. Ratings are typically on a 10-point scale. Candidates were contacted during weeks following the events for further feedback, but due to low response numbers, these attributes do not allow for any meaningful analysis.

3 Analysis

There are many different trends we examined from this dataset due to the large number of features per date. Since our focus was the predictive modelling, we only discuss a subset containing some of the most interesting results. All of the analysis and modelling we have done is on github [2].

3.1 Date Order

The Serial-position effect describes an individual's tendencies to remember the first and last items in a set. Although not based on memory, we were interested if date ordered had any impact on match rate, with a hypothesis that the first or last dates would be the most successful. Figure 1 shows a noticeable spike in the amount of matches from first dates compared to others. All dates following the first showed relatively stable numbers, suggesting that the first date was the most likely to result in a match. However, the last date did show any sort of significance, opposing the original hypothesis.

3.2 Interests

Candidates report their interests on 17 different areas (i.e. movies, gaming, exercise) using a 10 point scale. We attempt to predict a match between two people based purely off their interests. Figure 2 shows that different levels of interest for each area are correlated with different match rates. By using Jaccard Similarity, dates with a similarity greater than 0.5 are predicted to have a match. Similarity is calculated by the number of interests rated exactly the same divided by the total number of interests. Comparing the real matching from the data, the prediction is about 84% accurate. From this we can conclude that people sharings similar interests are more likely to match.

4 Predictive Modelling

With the end goal of this project being the ability to input data and output and expected result, we created various different models to compare their accuracy. Both a linear regression model and an artificial neural net are trained on the same parameters. For a single date between a person *A* and person *B*, the input data included the ratings of attraction, sincerity, intelligence, ambitiousness, and level of "fun" that person *A* gave person *B*. The final label was the decision that person *A* made, with a 1 meaning "yes" or a 0 for "no." We create binary classifiers to predict a person's decision based on the ratings they give their date. Due to most dates resulting in a "no" decision, a model that always outputted 0 would score an accuracy of 60%. The models created will use this as a baseline for determining effectiveness.

4.1 Linear Regression

Since the input consists of count data (nonnegative integers) capped at 10 with no outliers, we use multivariate linear regression due to its simplicity and effectiveness. We used two linear regression models: one with 10-fold cross validation, and the other training and testing on the same information. The predicted y-value was rounded to 0 or 1 to use 0-1 loss to try to compare the accuracy of the results. The loss for the 10 fold was 25.75%, meaning 74.25% accuracy. The loss for the other model was 25.83% meaning 74.17% accuracy.

4.2 Artificial Neural Net

We create another binary classifier using a 5-(10-50)-1 deep neural network. The network has the same five input parameters as described above, followed by two hidden layers of 10 and 50 fully connected nodes with rectified linear unit (ReLu) activation and normal kernel initializers. The final node with sigmoid activation outputs the percentage likelihood of a "yes" decision which is converted to a binary prediction with rounding. The input data is split and shuffled with 90% for training and 10% for testing. After training with 15 epochs, the model reports an accuracy of 74%.

4.3 Input

Both models would see improvements in accuracy with more parameters (such as "like," the overall "like" rating *A* gives *B*); however, the end goal of input restricts our training parameters because individuals only rate themselves in the five categories mentioned previously. To calculate how accurately individuals rate themselves, we take the difference between their self evaluation and the average of the ratings they got across all their dates. Plotting histograms of differences (figure 3), the mean values are taken as "knockdown" numbers. On average, each candidate overestimated each parameter by ~1. When a user inputs their own parameters into the model, we subtract the knockdown. This generates a more realistic representation of what an average date would rate. The output metric will be a percentage likelihood of an average date deciding "yes."

5 Concluding Results

After training both models and analyzing their outputs, the most influential factors for receiving a positive reaction from a date are attractiveness and "fun." The linear regression model weighs attractiveness 12 times more heavily than intelligence. Also, the weights for both sincerity and ambitiousness are negative, meaning higher sincerity and ambitious values result in a lower likelihood of a second date. Through manual testing, the neural net has similar results to the regression model: showing more importance in attractiveness and fun, with negative correlations in sincerity and ambitiousness. The most successful candidate at a speed dating event would be attractive and fun with little sincerity and ambitiousness.

6 Future Work

For future work with this data, we could expand the predictive models by assigning a similarity score to each training sample. When training on new information, the model can weigh more heavily on samples that are more similar to the testing sample. This should make the model more accurate. For analysis, we could also look at the differences in results between men and women, and factor that into the model as well.

Appendix A

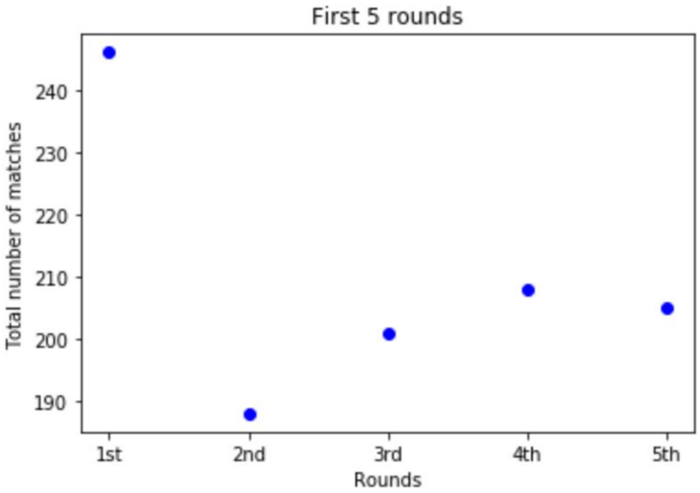

Figure 1: Number of matches versus round number. First dates resulted in more matches compared to other rounds.

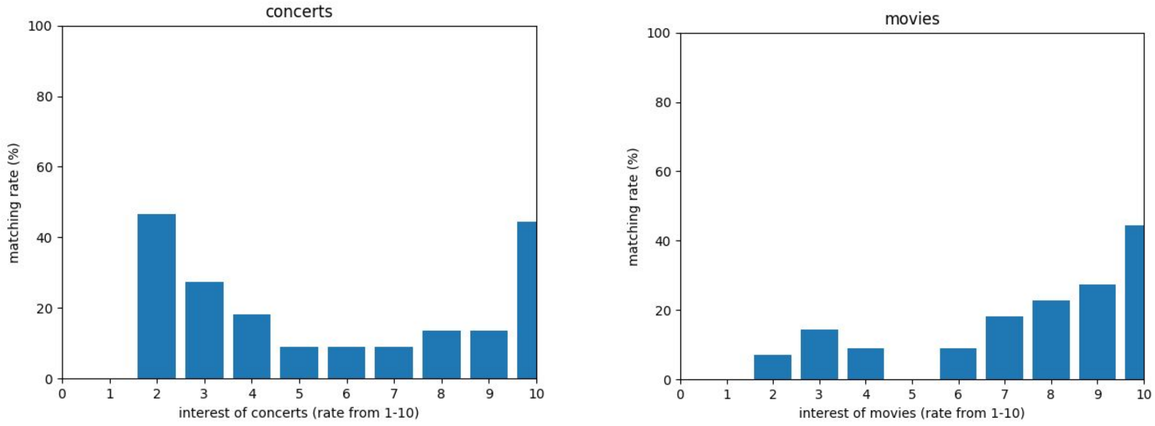

Figure 2: Example graphs showing different interest levels relate to difference match rates.

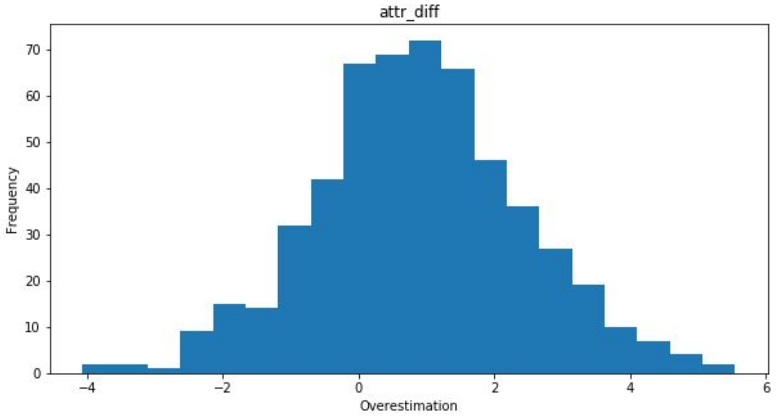

Figure 3: Example histogram in self-rated vs date-rated attractiveness values, showing that majority of people rated themselves ~1 higher in attraction.

References

(1) https://www.kaggle.com/annavictoria/speed-dating-experiment
(2) https://github.com/austinbeauch/speeddating