SPRINGBOARD DATA SCIENCE CAPSTONE PROJECT:

PREDICTING A MATCH FOR SPEED DATING

SAMUEL BINENFELD

SEPTEMBER 21, 2017

TABLE OF CONTENTS

1 INTRODUCTION

With the growing number of single people in the world today, the dating market is larger than ever. Dating as an idea is great; bring two people together who are both looking for romance and connection. **The problem is that the dating process is inefficient.** More often than not, dates are unsuccessful and people leave having wasted their time.

In order to solve this issue, we will take a look at the Speed Dating Experiment dataset from Kaggle.com. Using this dataset, we will look at areas such as age, race, desires, and interests to answer the question, “What makes two people want to date each other?”. Once we’ve answered that, we will create a machine learning model that can predict whether or not two people are likely to be a match for each other. This model can be used to help streamline the dating process!

2 DATA CLEANING

2.1 Remove All Unnecessary Fields

The data set started out with 195 columns. This high number of columns (fields) was concerning, so we reduced this number down to 38. The fields that were removed were unnecessary, for a variety of reasons.

Some fields had significant missing data. If keeping the field was going to cause us to lose a large portion of the data (because of null values), we removed it. This included many of the survey questions such as “what do you think the opposite sex is most interested in?”. It also included rating variables, such as “rate your ambition” and “rate the level of shared interests between you and your date”. This step removed 108 fields.

Some fields were repetitive. The data asked the same questions multiple times throughout the experiment (before going on dates, halfway through going on dates, after the dates, etc.). Asking these questions so many times wasn't going to bring additional insight, so we removed these 16 fields.

Some fields were too varied to be able to gain insight from. These included things like "field of study" and the "zip code" you grew up in. The sample size was too small for these, so they had to be removed. This step removed 6 fields.

Some fields consisted of information gathered after the date. After the date, participants were asked to rate their date in a variety of areas such as attractiveness, sincerity, and intelligence. While this information would be useful for predicting a match, in the real world, we will not know any after-date information before match-making. Since we want our model to have practical use, we removed these 16 fields.

Some fields were irrelevant. We did data exploration on the remaining variables, but there were some fields where nothing of importance could be found. These included "position" of the dater and what "group" the dater was in. This step removed 11 fields.

2.2 Filter the rows

We began with 8,378 rows of data, but filtered this down to 8,038 rows, for a few reasons. We **removed any dates with null values**. We had already eliminated the fields with significant missing data, but we still lost 263 rows because of the nulls.

We **removed any dates that involved one person who was 55 years old**. This person was 13 years older than the next oldest person, and was an outlier in terms of age. Removing this person only lost 6 rows of data.

We **removed any dates that had a partner who was never a primary**. Every date consisted of two people, and to distinguish between the two, one person is labeled the "primary", and the other is labeled the "partner". The issue here was with the interest variables (where daters were asked to rate their level of interest in a variety of activities). These variables were only recorded for the primary dater, and not the partner.

Later in our exploration, we go on to create a variable that combines both the primary and partner's interests. For this reason, partners who were never a primary would create missing data and needed to be removed. This cost us 71 rows of data.

2.3 Fix "Date" and "Go Out" Variables

Participants were asked to rate how often they go on dates, and how often they go out in general. The rating scale for these variables was 1 through 7, with 1 being "very often", and 7 being "never". We found these variables to be easier to interpret when the scale was reversed, and so we subtracted each row from 8 to accomplish this.

2.4 Fix Race Variables

There was an error in many of the dates in recording the participants' race. To give an example, one participant may have been recorded as being "White/Caucasian" for 8 out of their 9 dates, and then on their last date would have been recorded as "Other". These errors were easily identifiable and we were able to correct them by inputting the race that the person was listed as most frequently.

Also, the race variable was an integer field, with each number corresponding to a race. In the original dataset, the race "Other" was listed as the number 6. Since there was no race with the number 5, we changed all race variables labeled "Other" from 6 to 5.

3 DATA EXPLORATION

3.1 Introduction to The Cleaned Data

Our dataset consists of 538 people, who went on a combined total of 4,019 dates. All of the dates were of heterosexual nature, and so each date involved one male and one female. Daters had four minutes to converse and get to know each other. At the end of the date, each participant was asked whether or not they would like to see their date again. Participants were also asked to answer questions about their demographics, interests, preferences, and desires. All of these questions are described in detail in *Appendix 1*.

The variable of interest throughout this report is *match*. A date is considered to be a match if both daters said yes to wanting to see their date again. The average match rate for all dates was 16.62%.

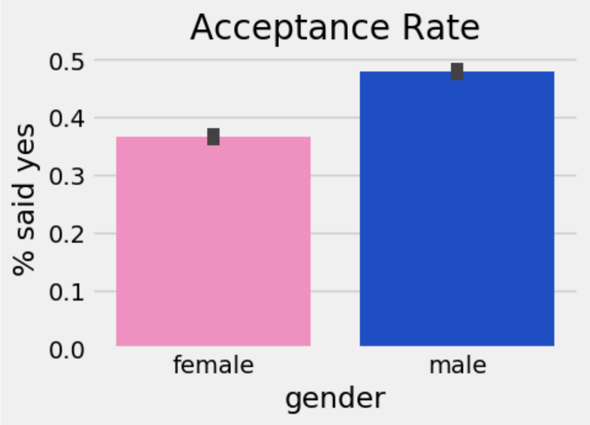
As a dater, there are two variables that make up your match rate: whether or not you say yes, and whether or not your date says yes to you. Therefore, your match rate will be influenced both by how accepting you are, and how desirable you are. We'll take a closer look at all the data, and see what factors are influential to the match rate. As we look at the different factors, the following terms will be used repeatedly (they are described in the first-person for interpretability).

- **Acceptance Rate -** % of the time you said yes
- **Desired Rate -** % of the time your date said yes
- **Match Rate -** % of the time both you and your date said yes

3.2 Gender

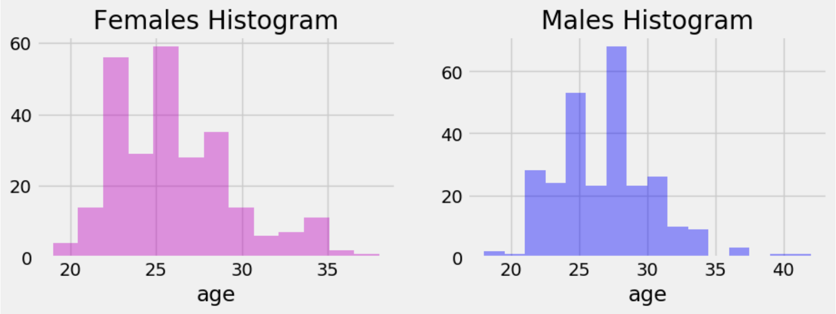
The distribution of gender was essentially equal. Of the 536 people, 265 were female and 271 were male (it's not exactly even because not everyone went on the same number of dates, and because we lost dates in our data cleaning process). In regards to acceptance rate, we see a difference between males and females. Females were less accepting, only saying yes to the man they dated 36.63% of the time. Males on the other hand, said yes to the woman 47.87% of the time.

```
Females: 36.63%
 Males: 47.87%
```

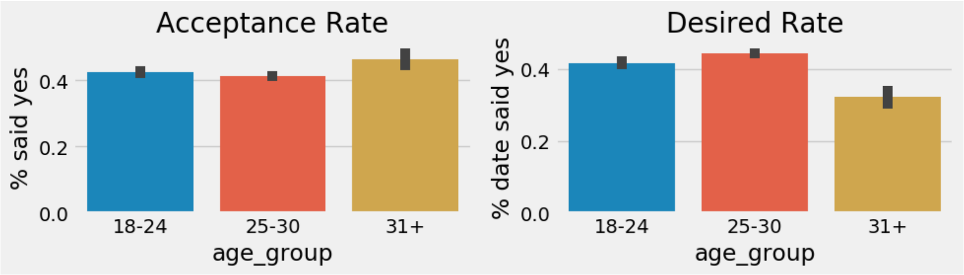


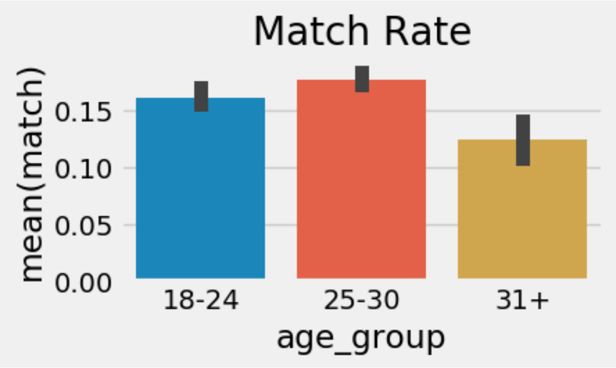
3.3 Age

Ages ranged from 18-42, with an average age of about 26.25. As this low average suggests, the age distribution was slightly skewed to the right. In other words, the majority of the daters were in their twenties, but we still had a fair number of people older than that (roughly 10% of the daters were over age 30). The age distribution in regards to gender was similar to that of the overall; the average ages for females and males were 26 and 26.6 respectively.

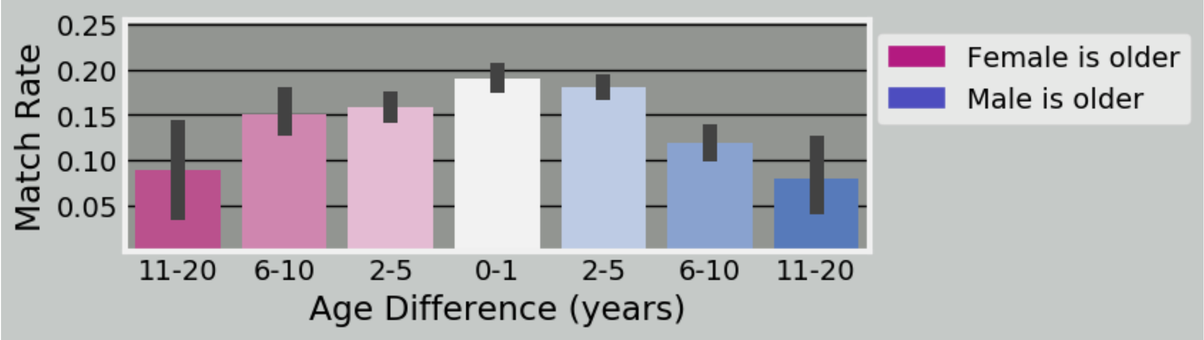


Because the distribution was skewed to the right, we want to see if the older daters struggled to find matches. To do this, we separate the daters into three different age groups, and look at each group's match rate. It turns out that those over the age of 30 did have a lower match rate, despite the fact that they were more accepting.





This leads us to consider the possibility that daters preferred those who were closer to their own age. To investigate this, we look at the age difference in each date, and compare the match rates. What we see is that there was a decrease in match rate as the age gap got larger. The following chart shows this, and is sorted by gender. To the right of 0-1 in blue is when the male is older, and to the left of 0-1 in pink is when the female was older.

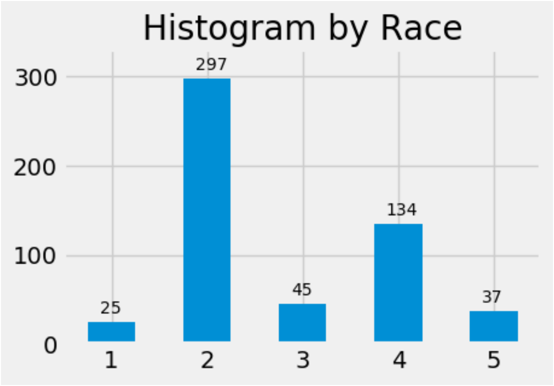


The match rate decline is clear for both males and females. We see that for males, the effect is not that strong until the age gap exceeds 5 years, but once it reaches this point, it does drop significantly. For females, we see an immediate drop for an age gap of greater than 1 year, but it doesn't become significant until the gap is over 10 years. The age difference is good to note, but we still need to investigate other areas.

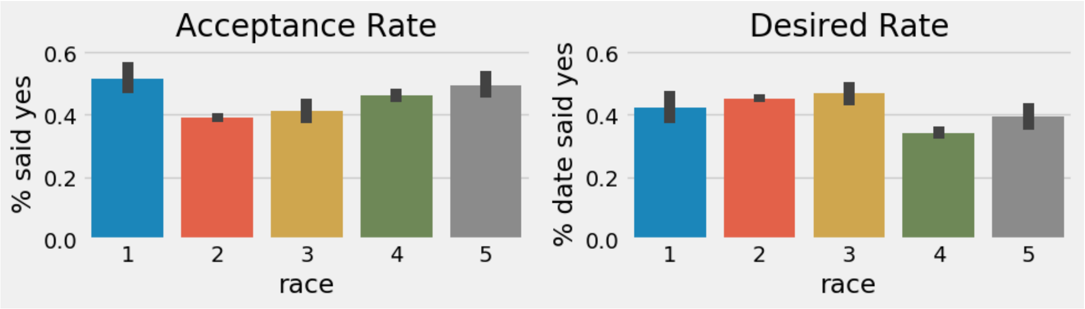
3.4 Race

The race distribution was not balanced in this dating experiment, and this made it somewhat difficult to draw conclusions based on race. Over half of the daters were Caucasian/European and about one fourth were Asian/Asian-American. The full distribution can be seen by this histogram.

```
1- African American
2- Caucasian/European
3- Latino/Hispanic
4- Asian/Pacific Islander/Asian-American
5- Other
```



The first step for us is to look at each race in terms of the accepting, desirable, and match rate characteristics. The following chart shows this.



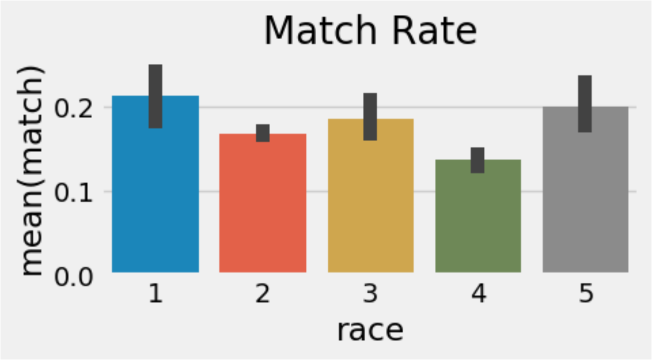
1- African American

2- Caucasian/European

3- Latino/Hispanic

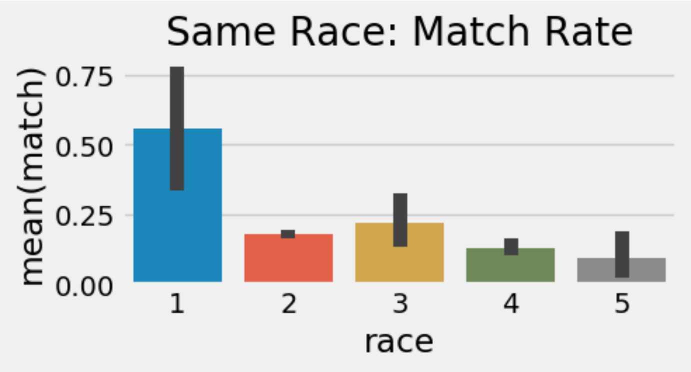
4- Asian/Pacific Islander/Asian-American

5- Other



There doesn't appear to be an overwhelming difference between races. We see that African Americans had the highest match rate, but this may be because they were more accepting overall. We also see a noticeably lower desirability and match rate for Asians/Pacific Islanders/Asian-Americans.

What we also want to look at, is if there was a preference to date others who were of the same race. We look at dates where both participants were the same race and compare the match rates. What we find is that the most dramatic increase was with African Americans, as match rate shoots up to 50%.



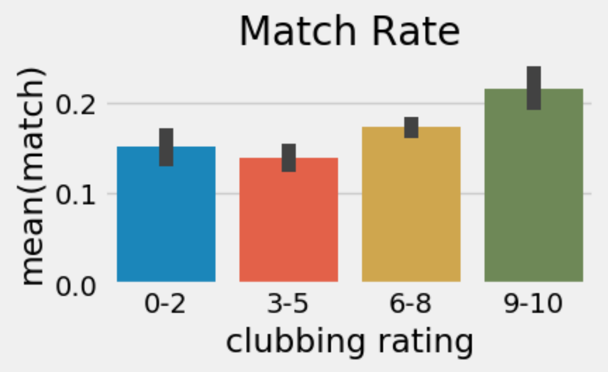
At first glance, this looks like it will be very valuable to us. While it is good information to have, we need to put it in perspective. The problem is that the number of African Americans in our dataset was low, and the match rate we see was actually only based on 8 dates, which is too small of a sample size. As well, because the occurrence of both daters being African American was so rare in this dataset, even if this higher match rate were true, it wouldn't have major significance in our model. For future study and modeling, we should look to get a more racially diverse dataset.

3.5 Interests

Participants were asked to rate their level of interest on a variety of activities such as sports, movies, art, etc. (the full list can be seen in *Appendix 1*). To get a look at all of the interests, and their relationship with match rate, we perform a correlation test. Here are the top 5 and bottom 5 interests, along with their correlation coefficients (rounded to thousandth place).

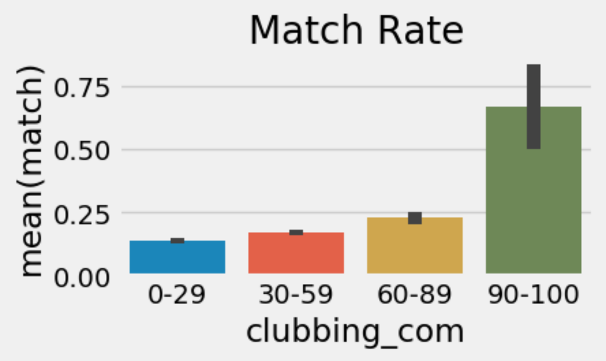
Top 5		Bottom 5	
clubbing	0.052	**movies**	-0.030
yoga	0.033	**tv**	-0.016
art	0.029	**tvsports**	-0.008
dining	0.026	**theater**	-0.008
concerts	0.026	**shopping**	-0.006

Clubbing and yoga were the most positively correlated, and movies and television were the most negatively correlated. Therefore, those who rate clubbing highly, should have a higher match rate. When we graph this out, we see this is true, especially for those who rated clubbing a 9 or 10 (the scale is 1-10 with 10 being the highest).



It's clear that a person with a high interest in clubbing will, on average, have a higher match rate. The next question is, what would the match rate be if **both** daters had a high interest in clubbing? It seems logical to think that people with common interests are more likely to be a match, so we think it would be higher.

To test this theory, we need a variable that reflects both participants' interests. Since we don't have one in the dataset, we create a new variable for each interest, which is the product of the primary and partner's rating for that interest. For example: If John and Katie went on a date, and John rated clubbing an 8, and Katie rated clubbing a 6, then our new variable would equal 48 (8 * 6). We create these and take a look at the new variable we made for clubbing, labeled *clubbing_com*.



Wow! We see a major jump in match rate to over 60% when we have a combined clubbing rating of 90 or above. Now to put this in perspective, we do suffer from the problem of small sample size, just as we did with the same race, African American dates. There were only 11 dates where the *clubbing_com* is 90 or higher.

However, the good news is that in addition to clubbing, there were 16 other interests that we have to work with. In fact, there were 1,682 dates (about 42% of the total) that have a combined rating of 90 or higher in at least one of the interests. If the other interests (when combined) also show us a higher match rate, then this will definitely help our model.

3.6 Desires and Preferences

Desires were ratings of attributes in response to the question "what do you look for in a date?". *Preferences* were ratings of a diverse, mixed-bag of questions. These categories are both described in detail in *Appendix 1*.

We'll analyze each of these categories the same way we analyzed our interest variables. First, we perform correlation tests to see the relationship between each variable and *match*.

Desires		Preferences	
fun1_1	0.047	**go_out**	0.063
intel1_1	0.022	**date**	0.061
attr1_1	0.015	**exphappy**	0.032
sinc1_1	-0.043	**imprelig**	-0.024
shar1_1	-0.047	**imprace**	-0.049

For *desires*, we see that those who desired their partner to be fun, were more likely to get a match. Conversely, those who wanted their partner to have shared interests, and those who wanted their partner to be sincere, were less likely to get a match.

For *preferences*, we see that those who said they go out often, and those who said they date often, were more likely to get a match. On the other side, those who said that their partner's race was very important to them, were less likely to get a match.

Like interests, we feel there is a chance that participants with similar desires, or similar preferences, are more likely to be a match than those without. Our next step is to create new variables that reflect the product of the primary and partner's responses for each of the questions (just like we did with the interest variables). After completing this, we feel that we have a good understanding of our variables, have created some valuable interaction variables, and are ready to move on to modeling.

4 MODELING

4.1 Introduction to Modeling

We will be using a number of supervised learning, classification algorithms to make prediction models for our variable *match*. The algorithms we will use are: Logistic Regression, Support Vector Classifier, Random Forest Classifier, and Gradient-Boosted Machine. We will create models for each of these, test them for accuracy, and see which ones perform well. We will then take the most successful models, and tune the parameters to ensure we get optimal results.

It's a good idea for us to get a benchmark goal for the accuracy of our models. Given that the average match rate for all dates was 16.62%, we realize that a model that predicts "no match" for every date, would be 83.38% accurate. This means that at a minimum, we want our models to strive for higher than 83.38% test set accuracy.

In addition to accuracy, we'll take a look at a couple of other metrics. Since our dataset is imbalanced (as in there are few matches compared to lots of non-matches), we want to examine the AUC of the PR Curve, and this will be the primary metric used to rate the models. Though this is the primary metric, we will also look at the AUC of the ROC curve.

4.2 Feature Engineering

The following is a list of the created features that will be used in our models. Some of these have already been mentioned in the exploration section.

- Create interaction terms (primary rating * partner rating) for all variables in the interests, desires, and preferences categories
- Create *age difference* variable (male age – female age)
- Create *age group* variable that separates age into bins of [18-24, 25-30, 31-42]
- Create *combined age* variable (primary age + partner age)
- Create category variable that contains both the male and female's race

4.3 Data Pre-Processing

The first thing we need to do is split the dataset in two parts, *X* and *y*. *X* is what we will plug into the model, and *y* is what we want the model to predict. **For *y*, we use only the field *match***. For *X*, we want to use all of our data, except what we're trying to predict. *Match* is what we want to predict, but the variables *dec* and *dec_o* make up the *match* variable, so we can't have those either. ***X* will be our data set with the fields *dec, dec_o,* and *match* removed**.

Now that we have our *X* and *y*, we need to split each of these up into training and testing tests. We complete this, putting 75% of the data in our training sets, and 25% of the data in our test sets. With this done, our data is now ready for modeling.

4.4 Model the Data

Finally, we're ready to see some results on how well we can predict matches. We plug in our data to each of the four algorithms, keeping in mind that we want a test accuracy rate higher than 83.38%. The results were as follows.

Model	Train Accuracy	Test Accuracy	Precision	Recall	ROC AUC	PR AUC
Random Forest	98.47%	84.03%	74.51%	10.98%	75.32%	47.60%

Logistic Regression	83.69%	82.74%	40.00%	0.58%	63.44%	29.22%
SVC	100.00%	82.79%	0.00%	0.00%	93.54%	88.45%
Gradient Boosting	85.95%	84.13%	93.55%	8.38%	77.17%	47.58%

These results are disappointing. The highest test accuracy we received was 84.13%, which is barely over random guessing (83.38%). Even though we're not happy with the results, we still want to get some insight into what our models were doing. Random Forest has a method that allows us to look at the level of importance for each feature, and so we'll take a look at the top 10 most important features.

Features	Importance
clubbing_com	0.035848
yoga_com	0.027607
date_com	0.026963
exercise_com	0.02595
shopping_com	0.02567
concerts_com	0.025075
hiking_com	0.025022
sinc1_com	0.02416
pid	0.023971
exphappy_com	0.023931

What we see is that the top 7 features are all combined interest features. After that, we see 1 combined desire (*sinc1_com*) and 1 combined preference (*exphappy_com*). The 9th most important feature is *pid*, which is a variable used only to identify daters, and so we want to disregard this as it will not help performance on unseen data.

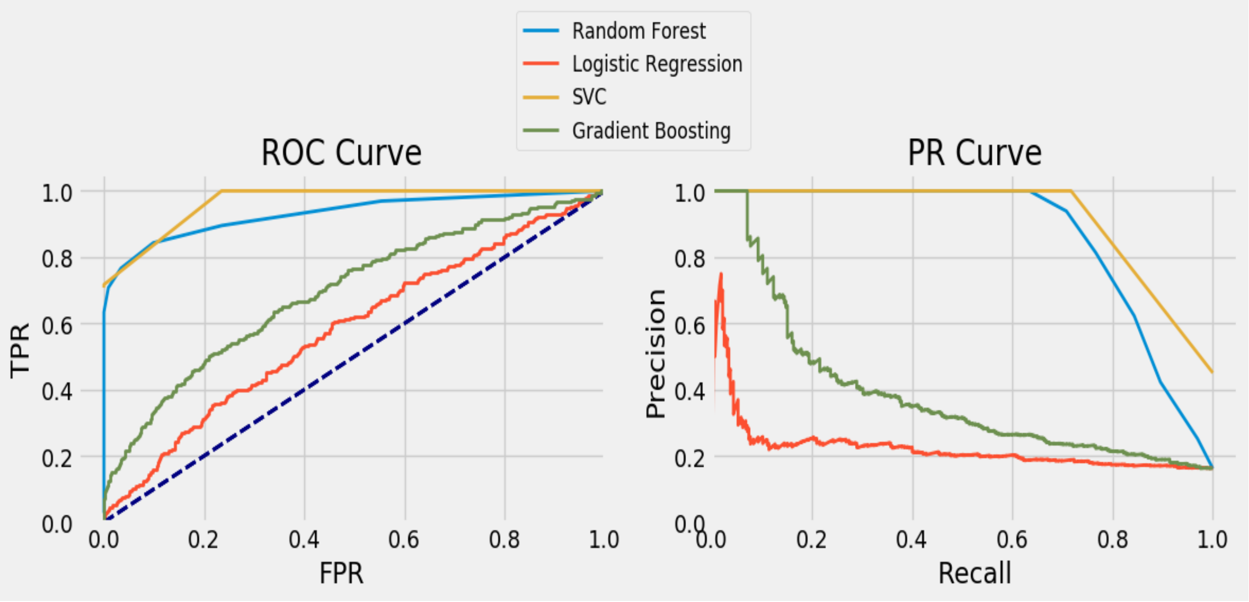
A concern of ours is that our model has too many features in it. If a model has too many features, it may over-fit to the training data, and we see this may be happening for both the Random Forest and SVC (we can tell because the Training Accuracy ratings were at or close to

100%). To attempt to improve our models' results, we want to reduce the number of features that we put into it.

Since the combined interests' features were the most important, we'll try re-modeling the data with only those fields. This will help us reduce noise, not over-fit to training data, and decrease computation requirements. Here are the results.

Model	Train Accuracy	Test Accuracy	Precision	Recall	ROC AUC	PR AUC
Random Forest	99.45%	94.08%	100.00%	63.38%	93.15%	87.06%
Logistic Regression	83.23%	83.83%	0.00%	0.00%	58.40%	22.17%
SVC	100.00%	95.42%	100.00%	71.69%	96.67%	92.23%
Gradient Boosting	85.50%	84.88%	86.21%	7.69%	69.65%	38.54%

This is a great improvement! As highlighted in green, the Random Forest and SVC performed very well on the test set. We also see strong performance from these algorithms in the area under curve percentages. We can take a look at both the ROC and PR curves and compare each of the models.



We tried modeling the other groups, such as combined desires, combined preferences, our other created variables, and various combinations of all of these. Unfortunately, nothing was able to outperform combined interests, and even adding the others to the combined interest variables didn't show noticeable improvement. With this in mind, we decide to keep the combined interest variables as our model's features, and we choose to only go forward with the Random Forest and SVC algorithms. We move on to tuning the parameters of these two algorithms.

4.5 Parameter Tuning

Random Forest

For the Random Forest Classifier, the parameters we will tune are: n_estimators, min_samples_leaf, and max_features. The range of values we will try for each parameter are as follows.

- **n_estimators** – [10, 25, 50, 75, 100]
- **min_samples_leaf** – [1, 25, 50, 75, 100]
- **max_features** – [.1, .25, .5, .75]

Often when tuning the N_estimators parameter, it is worthwhile to consider the trade-off between accuracy and computational speed. However, in our case the data set we're analyzing isn't exceptionally large, so computational speed is not an issue. Our final step here is to use sci-kit learn's GridSearchCV tool to find our optimal parameters. Here are the results.

- **n_estimators** = 100
- **min_samples_leaf** = 1
- **max_features** = .1

Now that we have our optimal parameters, we can compare the Random Forest with default parameters vs. with the new parameters. We see a definite improvement, especially in recall.

Model	Train Accuracy	Test Accuracy	Precision	Recall	ROC AUC	PR AUC
RF (default)	99.45%	94.08%	100.00%	63.38%	93.15%	87.06%
RF (new params)	100.00%	95.52%	100.00%	73.13%	96.15%	90.30%
Improvement	0.55%	1.44%	0.00%	9.75%	3.00%	3.24%

Support Vector Classifier

For the SVC, the parameters we will tune are: C, kernel, and gamma. The range of values we will try in our grid search are as follows.

- **C** – [.0001,.001,.01,.1,1]
- **kernel** – ['linear', 'rbf', 'sigmoid', 'poly']
- **gamma** - [0.01,0.02,0.03,0.04,0.05,0.10,0.2,0.3,0.4,0.5]

After getting the results, we found that changing the gamma parameter had no effect on accuracy, and so we no longer need to be concerned with it. For the remaining two parameters, the optimal choices are as follows.

- **C** – 1
- **kernel** – 'rbf'

Ironically, these are the default values for the SVC classifier. This means our previous model was already optimal, and there is no need for a comparison. We now have our best performing models with optimal parameters.

5 CONCLUSION

Our original goal was to see if we could find a way to predict the likelihood that two speed daters would be a match for each other. The benchmark for our test was to beat random guessing, which would be an accuracy rate of 83.38%. We were able to exceed this, with our highest accuracy score reaching 95.52%.

We found the best way to predict whether or not two people will be a match for each other, is to look at their combined interests. The interests we found most helpful in prediction were clubbing and yoga. For further analysis, we should look at incorporating a wider range of interests to find other useful variables that are missing from this dataset.

The best performing models were the Random Forest with tuned parameters, and the SVC. Both performed well in test set accuracy, and in the AUC of the PR Curve. The performance was nearly identical in both, and so we are comfortable using either one of these. Overall, we are excited about our results, and believe this model has the potential to be helpful in the dating market.

APPENDICES

Appendix 1

Since each date involves two people, we often have two fields for each variable (one for each person). To distinguish between each person, we refer to them as either the “primary” or the “partner”.

- **IDs:** Each participant in the study is given a unique identification number.

iid	Primary’s ID #	**pid**	Partner’s ID #

- **Facts:** These are self-explanatory.

gender	Primary’s Gender	**samerace**	Yes or No
age	Primary’s Age	**age_o**	Partner’s Age
race	Primary’s Race	**race_o**	Partner’s Race

- **Interests:** These are the primary’s ratings on the scale of 1-10 for a collection of interests. (1- not at all interested, 10- very interested)

sports	**tvsports**	**exercise**	**dining**	**museums**
art	**hiking**	**gaming**	**clubbing**	**reading**
tv	**theatre**	**movies**	**concerts**	**music**
shopping	**yoga**			

- **Desires:** The primary was given 100 points to allocate among these 5 categories in terms of what they look for in a partner.

attr1_1	Attractiveness	**fun1_1**	Fun

sinc1_1	Sincerity	**shar1_1**	Shared Interests
intel1_1	Intelligence		

- **Preferences:** These are a collection of the primary's preferences

Rated 1-10 (10 highest)		**Rated 1-7 (1 - most often)**	
imprace	How important is race to you?	**go_out**	How often do you go out?
imprelig	How important is religion to you?	**date**	How often do you date?
exphappy	How happy to expect to be from these dates?		

- **Decisions:** This is the decision of whether or not you want to see the other person again. Values are 0 or 1. If both are 1, then it is a match.

dec	Primary's Decision
dec_o	Partner's Decision
match	1- If both *dec* and *dec_o* are 1 0- If either *dec* or *dec_o* are 0