Speed Dating: Becoming a Better Candidate

Laura Mathew, Andrew Nieto, and Alex Timperman

Department of Information and Decision Sciences, University of Illinois at Chicago

IDS 575: Machine Learning and Statistical Methods for Business Analytics

Dr. Moontae Lee

December 7, 2020

Abstract

The prominent purpose of the project is to classify which attributes are essential for matching with someone. Multiple different classification models were implemented to explore which attributes were key when matching with someone when participating in a speed dating event. Furthermore, the data of speed dating events between male and female participants were analyzed in order to understand whether these attributes differed between these participants. The classification models used include decision tree models as our baseline and Naive-Bayes, random forest, and logistic regression as our main models. The findings suggest that the random forest classifier had the highest accuracy, with the logistic regression models with the second highest. The significant variables determining whether or not two individuals will match are attractiveness and fun.

Introduction

Can you become a better candidate for the opposite sex through data science? Whether you are trying to find a partner online or through social events there is an opportunity for you to become a more appealing candidate. The goal is to determine if there are attributes that can increase the likelihood of you getting matched with someone else. Prior research shows that women in speed-dating settings are twice as choosy when picking a partner compared to men, and that education as well as professional status mattered to both men and women (McDonald, 2014). Further research also indicates that perceived similarity, not actual, made it much more likely for individuals to like the other person romantically (Tidwell, Eastwick, & Finkel, 2013).This will be explored as to what this means in the context of the research question. The prominent purpose of the project is to classify which attributes are essential for matching with someone. A subtopic that will also be explored to see if these attributes were consistent across both genders.

The dataset used for this project is part of a longitudinal study conducted by professors Ray Fisman and Sheena Iyengar at Columbia Business School. Data was gathered from 2003 - 2004, in speed dating event experiments, conducted over multiple waves(weeks).The project will be based on the speed dating events that recorded the background information of each candidate. The data has records of evaluation each recipient gave for each candidate, as well as how they evaluated themselves over the course of the study.

The paper published by the researchers of this dataset found female participants valued intelligence and race of the partner compared to male participants who valued attractiveness. (Fisman et al., 2006). Whether or not this holds true will be answered in the classification results and analysis of the data.

For the purposes of this project, various classification methods were used in order to classify predictions. A decision tree was first implemented as a baseline model and Naive-Bayes model, Random forest models, Logistic regression and Logistic regression models with Lasso and Ridge regularizer were applied to try to reach the goal. We were able to explore what makes someone more likely to receive matches, whether it was their interests or how others perceive them.

Data Manipulation

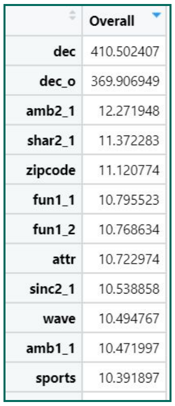
In order to do a proper analysis on this data set, we took a deep dive into the data to understand all variables, specifically, what each row of data represents. In the “Speed Dating Data” each participant has a unique identifier, however the data is not based on this identifier but based on each wave (week) of speed dating. Within each week there is an ID which identifies them for the event. For example, if there are 10 females and 10 males, then the ID numbers for that wave would range from 1-20. Each participant receives one number and these numbers are reused each week. One condition this study held was each participant of one gender must meet every single participant of the opposite gender in that wave. This held true regardless whether gender ratio was imbalanced for a wave.

Each row in the data represents one participant in a wave meeting with a participant of the opposite gender in the same wave. Information about each wave is provided as well as variations among them. During each meet, participants gave the score of whether or not they would be a match with the person they meet. They assigned scores of 0 and 1. This"match" variable will be used as the dependent variable to classify candidates. A substantial amount of detail was collected on the participants, ranging from background information and interests. Additionally, survey data on what they seek in a partner and evaluation scores they gave for each participant were also collected. Lastly, they recorded information on how participants changed how they see themselves and what they look for in a partner after participating in the speed dating event. These follow ups are some of the variables with the highest amount of missing information.

The variables with high missing information were omitted. The cutoff used to determine whether a variable was not worth including was if it consisted of 60% missing values. After checking for missing values in the variables that were classified as relevant, and found there was a large amount of columns with missing data. Omitting these values/participants would have caused the loss of data for future research questions. It is important to understand this was an experimental study and the investigators changed, added, as well as stopped recording for several items in the dataset as the study progressed.

One variable that was removed was *income* and was replaced with income2 in the dataset because of how it was listed in the csv file itself. A key was provided for the dataset and it was crucial to determine which variables we would need to omit. The missing values that were present in the dataset were all replaced by 0 since it was mostly due to participants not filling out the surveys as well as not answering correctly when the researchers gave them these surveys and questionnaires. We checked to see if there were more missing values and replaced all of the missing values until there were none. Replacing the missing values with a median was not possible because it would not be an accurate representation of the data that was collected and each row drastically differed depending on the answer each participant gave.

After accounting for all the missing values, the variable income2 was divided into quartiles to represent the levels of income. The dataset was then divided into male and female data subsets in order to complete further analysis as seen in studies conducted by the researchers themselves. Certain waves (weeks) were given a different scale and thus needed to be sure they were normalized. Since the goal is to identify which variables were the most relevant to finding a match and estimate how individuals rated their matches, a random forest analysis was conducted on the entire dataset in order to prioritize the most important variables. In the figure below you can see the variables *dec* & *dec_o* have an extremely high importance value, which means that these variables are too similar to the target variable and thus would contribute to overfit. We decided to omit these variables when implementing the models.



	Overall
dec	410.502407
dec_o	369.906949
amb2_1	12.271948
shar2_1	11.372283
zipcode	11.120774
fun1_1	10.795523
fun1_2	10.768634
attr	10.722974
sinc2_1	10.538858
wave	10.494767
amb1_1	10.471997
sports	10.391897

Models

Decision Tree / Random Forest

The decision tree model classifies the data by splitting it based on a set of predictors. In this model the goal is to classify the data on whether or not each participant receives a match. For the model the data was split based on information gain (entropy).

$$\boxed{\mathrm{Info}_{\mathrm{A}}(\mathrm{T}) = \sum \frac{|T_i|}{|T|}\ \mathrm{Info}(T_i)}$$

In this project the rpart package from R was used to run the decision tree. The pseudo code for decision tree can be listed as:

```
Model = rpart(dependent_variable ~ . , data, method , parms = (split etc.), control)
```

The model is making a split if there is a gain in information and will stop splitting as soon as a split does not result in a net gain for the model. After splitting the data based on these parameters what followed was pruning of the decision tree in order to make the model more robust. This is because oftentimes less complexity can lead to more stable models. Cost complexity pruning was used to reduce the number of splits in the Decision Tree Model. This was done by pruning to the lowest cost complexity value, where a comparison of the error values to the number of leaves was utilized to achieve a score.

$$\boxed{CC(T) = Err(T) + \alpha L(T)}$$

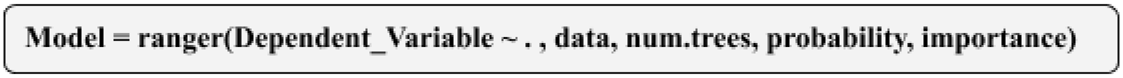
Lastly, an evaluation of the models was conducted based on the Accuracy and ROC curves. Accuracy was determined by calculating the correctly classified data points over the total predictions.

$$\boxed{(TP + TN) / (P + N)}$$

The ROC curve illustrates how the False Positive Rate increases as the True Positive Rate increases. It is necessary not to have too steep of a curve because that is a sign of overfitting. The optimization of a ROC curve is done by trying to achieve the largest area under the curve (AUC) while trying to minimize the probability of there being overfitting.

A random forest is a model that is a compilation of decision trees. It grows a set number of trees and each tree contributes to determining the variable's significance. The variable importance is determined by the average of all of the variable importance throughout the trees. Due to the assortment of Trees it can remove the majority of biased error from the model as well. This allows the Random Forest model to be one of the preferred classification methods. Comparing the accuracies of the model was made possible through the use of a confusion matrix to calculate accuracy and the ROC curves, similar to the method above.

In this project the ranger package from R was used to run the random forest. The pseudo code for the random forest can be listed as:



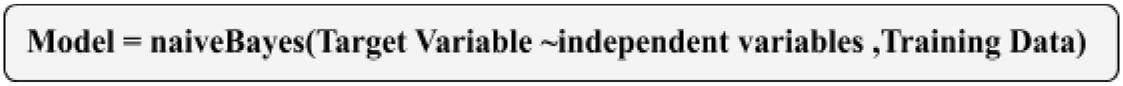
```
Model = ranger(Dependent_Variable ~ . , data, num.trees, probability, importance)
```

Naive-Bayes

The Naive-Bayes model is based on the naive-bayes theorem. This theorem assumes that each feature listed in the dataset is independent of each other and assumes that there is no relationship or interaction in between one another. This is useful when classifying binary results such as the target variable. It is a classification model through which we can find out if each participant matches or not. The naive-bayes classification uses probabilities in order to get the classification values. The Naive-Bayes model that was calculated for this section is the Bernoulli naive bayes, since we are classifying whether someone matched or not (1 or 0).

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)}$$

In this project, the e1071 package from R was used to implement naive bayes. The pseudo code for this section can be listed as:



```
Model = naiveBayes(Target Variable ~independent variables ,Training Data)
```

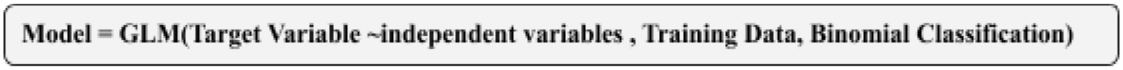
Logistic Regression

The logistic regression models measure the log-likelihood of a classification occurring between 0 and 1.This regressions is exclusively used for binary results since any predictions cannot be made outside of 0 and 1. Each individual coefficient will change the log-odds of a classification. For example, if a variable received a coefficient of 2, then for every one unit change in that vairale the log odds will increase by 2. The logistic regression will optimize the given parameter as follows:

$$\hat{\theta} = argmax_{\theta \in \mathbb{R}^N} L(\theta) = argmax_{\theta \in \mathbb{R}^N} \log(L(\theta))$$

The regression will optimize the parameters through a gradient descent. The algorithm it follows is:

$$\theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

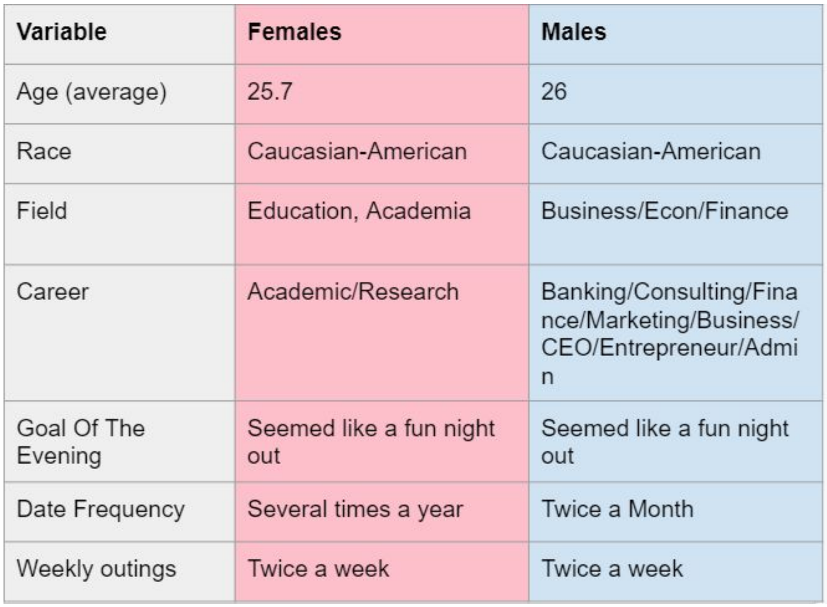


```
Model = GLM(Target Variable ~independent variables , Training Data, Binomial Classification)
```

Experimental Results

Dataset Explanation

The data set is the records of a speed dating experiment conducted between 2002-2004. The 195 features of the dataset can be broken down into 3 categories. The first set of variables are information about the participant. This includes information about the participant such as: age, race, education, field of work, zip code, and income. Also included here are the interests /hobbies as well as what they are looking for in the opposite sex. The second set and arguably the most essential, are the evaluations that participants filled out for other participants they met that night. It is at this point that participants decide whether or not to match with another participant. The last set of features are follow-up surveys given to participants weeks after the event. Looking below you can see the average statistics of the dataset. There is a clear difference between male and female participants in areas such as field, career, and their date frequency.



Variable	Females	Males
Age (average)	25.7	26
Race	Caucasian-American	Caucasian-American
Field	Education, Academia	Business/Econ/Finance
Career	Academic/Research	Banking/Consulting/Finance/Marketing/Business/CEO/Entrepreneur/Admin
Goal Of The Evening	Seemed like a fun night out	Seemed like a fun night out
Date Frequency	Several times a year	Twice a Month
Weekly outings	Twice a week	Twice a week

The experiment finished with a total of 8,378 matches. During each of the events participants had to meet with every participant of the opposite sex. The amount of females and males for each event was balanced, with an approximate equal number of each gender during every week. However, there was an imbalance in the classification feature. The “match” feature was the final say in whether participants matched or not during their encounter, yet 84% of the data resulted in participants not matching. It is here we see the foreshadowing of one of the issues, a large amount of false positives.

Problem Solving

After having an understanding of the dataset, the goal going forward would be to classify each pair as either match(1) or not (0). It was decided it could not be possible for two participants to be a partial match, the same way it was conducted in the experiment. From here the goal was to determine what were the most relevant variables in the study that increased the likelihood of two participants matching. To add on to this goal, we also wanted to do an analysis of the genders. As discussed previously there is reason to believe that male and females have different preferences when it comes to looking for a partner.

Experimental Processes

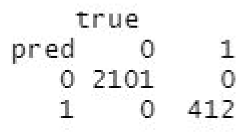
The very first step was taking time to understand the dataset. The original assumption on the data set was as follows; each row represented one participant and the results on how they performed. That was not the case, after discovering multiple attributes were staying the same across rows of data. This led to an understanding each row represented one participant meeting with another participant. So each participant would have X number of rows corresponding to them, where X is the number of participants of the opposite gender.

The experimental process that was followed was similar to a scientific method, but because the data had already been collected this step was omitted. The first step was developing a hypothesis based on the research conducted on speed-dating as well as the dataset that was selected. The next step was to pose research questions to investigate through analysis of the data. In order to understand the data better, data exploration and manipulation was necessary in order to account for missing values and impute missing values. A baseline model using a decision tree was implemented in order to understand the data and the hypothesis. After the baseline model, four different classification models were conducted- Naive-Bayes, Random Forest, Logistic Regression and Logistic Regression with Lasso and Ridge models. The experimental process concluded with determining the importance of the attributes that contributed to a match between both genders, as well as between male and female participants of the data. This allowed us to answer the research questions as well as how the hypothesis holds against the results of the classification models.

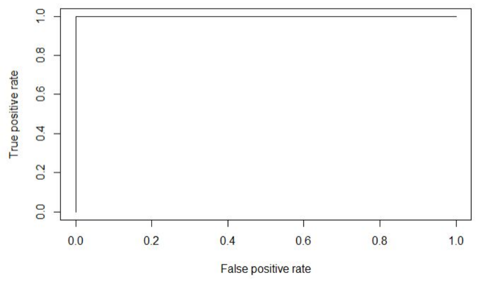
Model Results

Baseline Model: Decision Tree

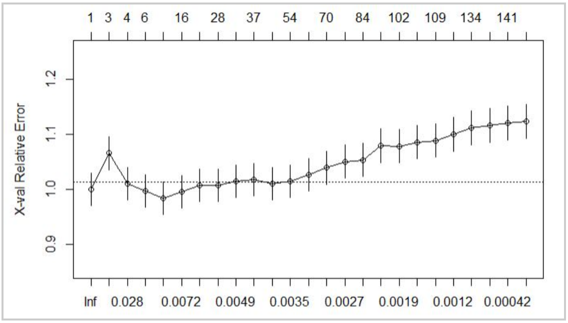
For the baseline model a decision tree was selected as the starting model because of its ability to effectively classify data of a binomial dependent variable. This allowed for a straightforward execution of the decision tree to determine whether a speed dating participant received a match (1) or no match (0). The starting point was to create a model by running it on all columns and with no complexity limits or split limits. This gave us a model that was incredibly overfitted and allowed us to begin crafting a better model. In the pictures below you can see there is an accuracy of exactly one and the roc curve is a perfect right angle. This gave us cause to believe there was serious data leakage and variables needed to be removed from the model in order to achieve a better baseline.

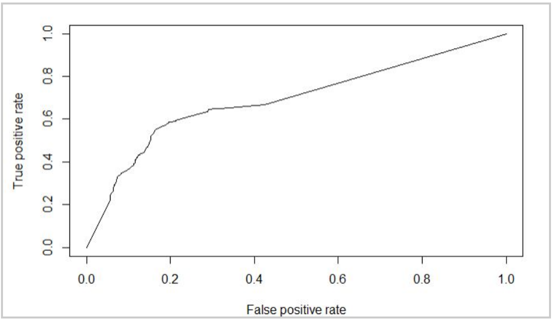


```
      true
pred    0    1
   0 2101    0
   1    0  412
```

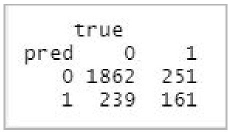


Once the decision tree was refined the results of the baseline model began to improve. Four variables from the model were removed because they were redundant and contributing heavily to data leakage. These variables were the decision variable, the from variable, and the like variable. The decision variable was binomial and the same as the target variable match. This was the biggest culprit of data leakage because they mirrored the target columns. The *from* variable was also removed because it was repetitive with *zip code* and from had much more variation, for example, some people said Chicago and some people said Chicago, Illinois. The *like* variable was also removed from the model because it duplicated another column that indicated whether the date liked the person they were matched with. Once these variables were removed the model saw significantly less overfit, but there were still some issues. In particular the model had a bumpy ROC, compared to being smooth, and the cross validation graph was ascending when it should be descending.

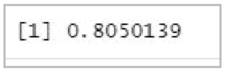




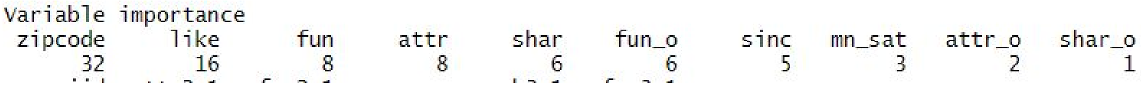
Even with these issues, this was still a much better predictive model because it had significantly less data leakage and overfitting. If you look at the confusion matrix and accuracy below, the model has an accuracy of around 80%, which can definitely be improved but is a realistic baseline score.



```
      true
pred    0    1
   0 1862  251
   1  239  161
```

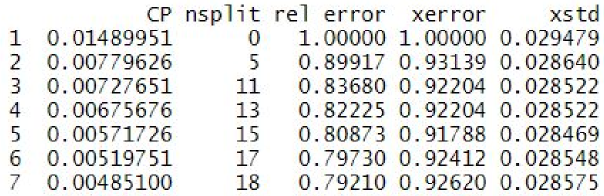


```
[1] 0.8050139
```

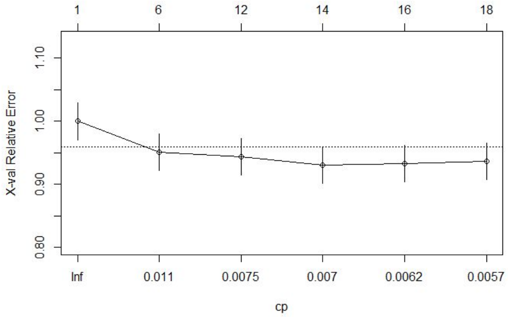


```
Variable importance
 zipcode     like      fun     attr     shar    fun_o     sinc   mn_sat   attr_o   shar_o
      32       16        8        8        6        6        5        3        2        1
```

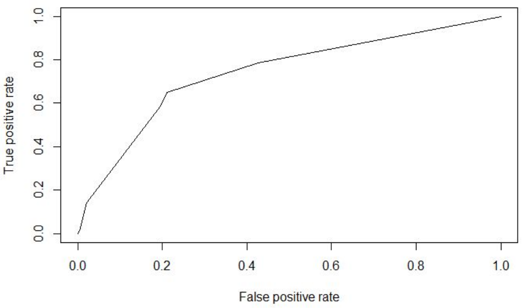
Continuing to remove variables that contributed to data leakage allowed the decision tree models to continue to improve Variables such as like and zip code were the biggest culprits because they had significant influence and were contributing to overfitting of the model. The model was then pruned to a complexity of 0.0057 by referencing the cross validation table.



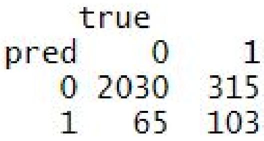
```
          CP nsplit rel error  xerror     xstd
1 0.01489951      0   1.00000 1.00000 0.029479
2 0.00779626      5   0.89917 0.93139 0.028640
3 0.00727651     11   0.83680 0.92204 0.028522
4 0.00675676     13   0.82225 0.92204 0.028522
5 0.00571726     15   0.80873 0.91788 0.028469
6 0.00519751     17   0.79730 0.92412 0.028548
7 0.00485100     18   0.79210 0.92620 0.028575
```



Using these different measures the model was significantly improved to 84% accuracy and an AUC of .74. Also the ROC curved compared to the baseline is much more rounded, displaying improvement in the model. Using pruning and variable removal methods significant improvement was made in the baseline model as seen above.



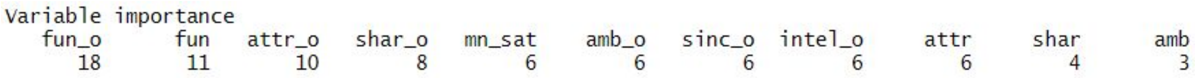
AUC: 0.742252



```
    true
pred    0    1
   0 2030  315
   1   65  103
```

Accuracy: 0.8487863

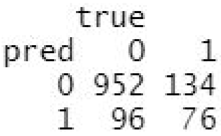
Using this improved model variable importance could now be used to gain insight into the variables which have the greatest impact on whether or not a contestant receives a match.



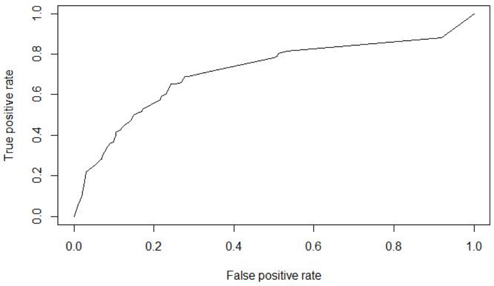
```
Variable importance
   fun_o       fun    attr_o    shar_o    mn_sat     amb_o    sinc_o   intel_o      attr      shar       amb 
      18        11        10         8         6         6         6         6         6         4         3 
```

Decision Tree illustrated whether the date found the contestant fun, though they were attractive, and if they shared common interests, were the biggest contributors to whether a contestant matched with their date. To gain greater insight and continue to improve the models it was then decided to split the dataset into males and females.

The next model run was a male-only subset. It decided to break up the data because it was hypothesised this would improve the model's accuracy and give important variables unique to each gender. The same variables that were in the previous model were removed in order to keep the columns consistent. The model ran and the result are as follows:



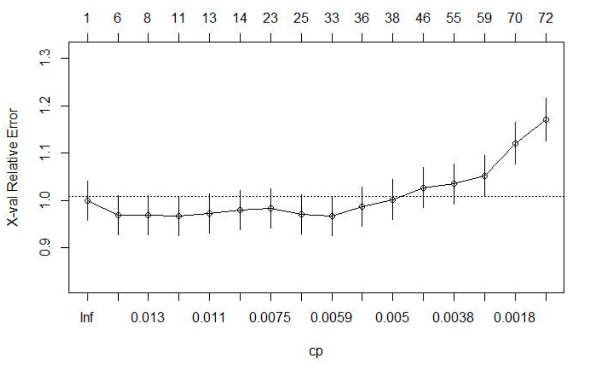
```
    true
pred   0   1
   0 952 134
   1  96  76
```

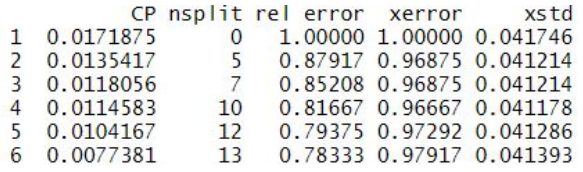


Accuracy: 0.8171701

AUC: 0.7133542

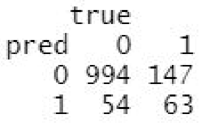
It was known from the previous model improving these numbers through pruning and identification would result in the preferred complexity value to be 0.014. This value would minimize the cross validation error.



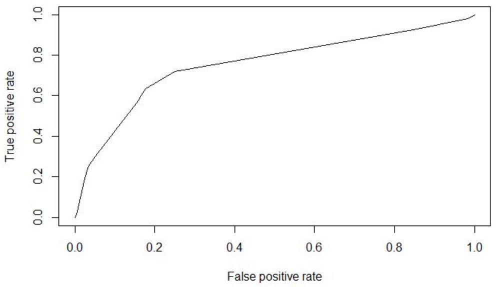


```
          CP nsplit rel error  xerror     xstd
1 0.0171875      0   1.00000 1.00000 0.041746
2 0.0135417      5   0.87917 0.96875 0.041214
3 0.0118056      7   0.85208 0.96875 0.041214
4 0.0114583     10   0.81667 0.96667 0.041178
5 0.0104167     12   0.79375 0.97292 0.041286
6 0.0077381     13   0.78333 0.97917 0.041393
```

This model was then rerun to see how pruning of 10 splits changed the models performance. Accuracy has improved from 81% to 84% while AUC improved from 0.71 to 0.75. What was surprising was that the accuracy of this submodel is nearly identical to the full dataset model and the AUC is similar as well. It was hypothesized the gender-specific dataset could have a better predicting model but that was not the case.

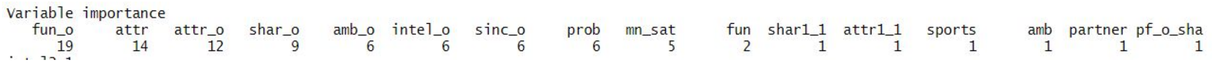


```
    true
pred   0   1
   0 994 147
   1  54  63
```



Accuracy: 0.8402226

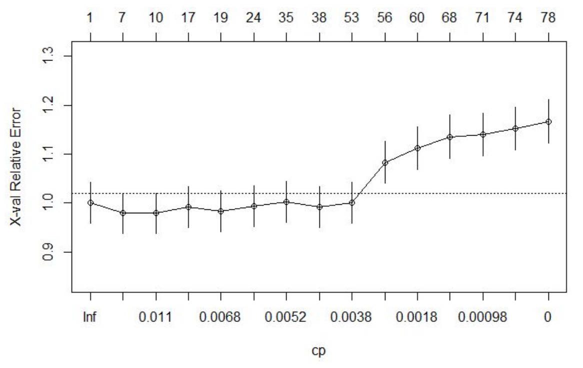
Accuracy: 0.7535305

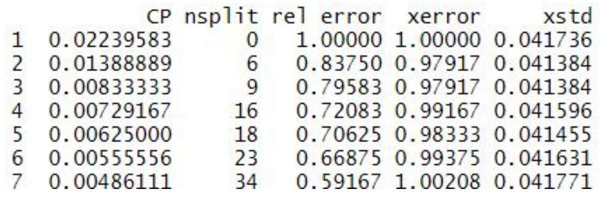


```
Variable importance
   fun_o     attr   attr_o   shar_o    amb_o  intel_o   sinc_o     prob   mn_sat      fun  shar1_1  attr1_1   sports      amb  partner pf_o_sha
      19       14       12        9        6        6        6        6        5        2        1        1        1        1        1        1
```

Variable importance for the male data set model allied to the full data set with the primary important factors of fun, attractiveness, and common interest playing an extensive role.

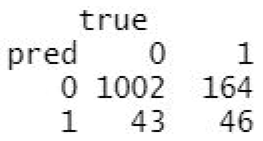
Next, a Decision Tree model was run on the female subset to compare the male model to. Following the same pruning methods as before , the model was pruned to a complexity of 0.00625 for 18 splits.



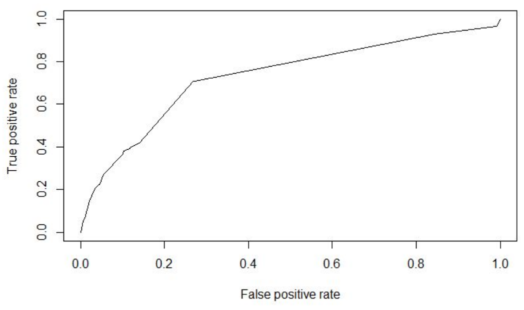


```
          CP nsplit rel error  xerror     xstd
1 0.02239583      0   1.00000 1.00000 0.041736
2 0.01388889      6   0.83750 0.97917 0.041384
3 0.00833333      9   0.79583 0.97917 0.041384
4 0.00729167     16   0.72083 0.99167 0.041596
5 0.00625000     18   0.70625 0.98333 0.041455
6 0.00555556     23   0.66875 0.99375 0.041631
7 0.00486111     34   0.59167 1.00208 0.041771
```

Taking a look at the result, the accuracy is very similar to that of the full data set and the male data set at 83.5% and the AUC is very similar as well at 0.73. The variable importance of the female dataset, again looked very similar to the previous models with the top three attributes being the same as the male dataset. With that being said, there is a slight difference as the median SAT score (mn_sat) was slightly higher than shared interest (shar) for the female dataset.

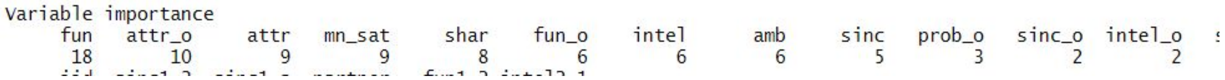


```
     true
pred    0    1
   0 1002  164
   1   43   46
```



Accuracy: 0.8350598

AUC: 0.7307268

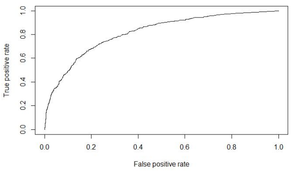
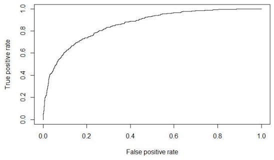
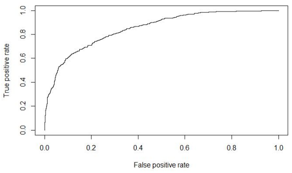
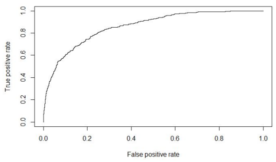


```
Variable importance
     fun   attr_o     attr   mn_sat     shar    fun_o    intel      amb     sinc   prob_o   sinc_o  intel_o
      18       10        9        9        8        6        6        6        5        3        2        2
```

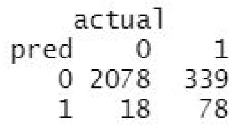
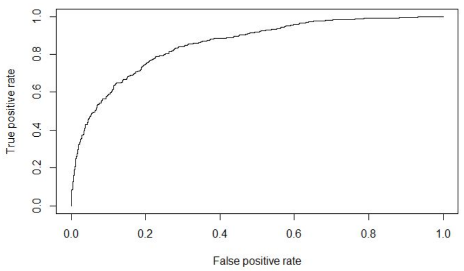
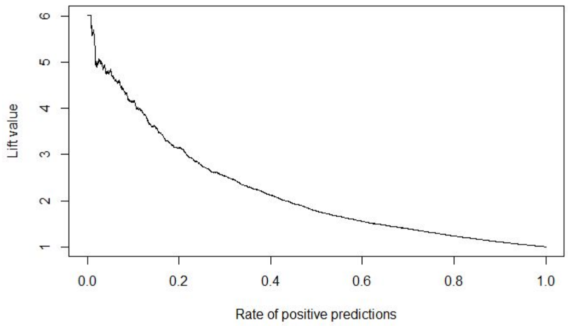
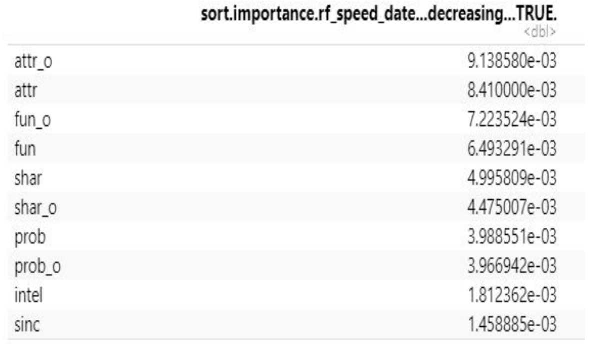
A large amount of information was gained from the decision trees. First, removing any overfitted variables such as the decision, zip code, and whether the person was liked or not. Also an initial draft of the variable importance, seeing attractiveness, and being fun were the highest ranked features, while other features such as like and sharing a common interest also play a role. However, it came as a surprise to see no significant difference between the male and female results of the models. It was hypothesized the subsets model could be more accurate and give contrasting important variables. Although decision trees are far from perfect predictors, the Random Forest model did serve the purpose of improving the Decision Tree models.

Random Forest Models

After working with the decision tree models a Random Forest model was used to enhance current models. Initially the random forest model was created on the entire dataset including both male and females. It was expected the model would perform significantly better categorizing the contestants into whether or not they received a match because it creates numerous decision trees and compiles them all into one predictive model. How many trees to create for the random forest was the first metric parameterized. Below you will see how the interactions were used to identify the appropriate amount of trees to use.

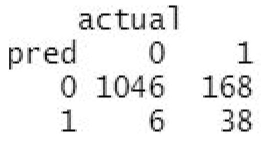
10 Trees	Accuracy:0.8547553 AUC:0.814794 	**200 Trees**	Accuracy:0.8595304 AUC:0.857698 
100 Trees	Accuracy:0.8591325 AUC:0.8496623 	**1000 Trees**	Accuracy:0.8595 AUC:0.8589 

All of the models above were similar, with the accuracies ranging from 84% to 85% and AUC's ranging from 0.81 to 0.85. Although the 1000 tree model performance was best, it was the 200 tree model that was selected. The 1000 trees simply took a drastically long time for a minimal improvement when compared to the 200 trees. The max depth size and min node size were experimented with, but neither made a significant impact on the performance of the model. The results of the Random Forest model for the full dataset is as follows:

 actual pred 0 1 0 2078 339 1 18 78	Accuracy: 0.8579387
	AUC: 0.8550539
	 sort.importance.rf_speed_date...decreasing...TRUE. <dbl> attr_o 9.138580e-03 attr 8.410000e-03 fun_o 7.223524e-03 fun 6.493291e-03 shar 4.995809e-03 shar_o 4.475007e-03 prob 3.988551e-03 prob_o 3.966942e-03 intel 1.812362e-03 sinc 1.458885e-03

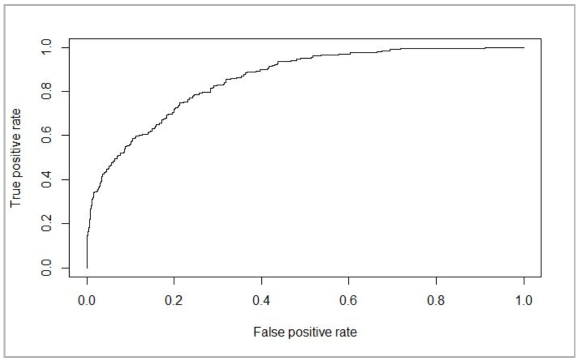
From the photos above it is clear the random forest is a significant improvement over the decision tree model because of its drastically increased accuracy of 86% and AUC of 0.86 as well. The ROC curve is also much more symmetrical using the random forest model thanks to the increase in the number of trees. In terms of variable importance, there is a very familiar trend, where attractiveness, being fun, and sharing common interests are the main contributors when determining the dependent variable. The prob column is significant in the model as well, which is whether the contestant thinks that their date will mark them a match or not. This result has reason to be there because if you believe someone will mark you as a match then you feel the date has gone well showing that your confidence after the date can be a good indicator of a match.

The experiment on the gender specific datasets to determine any deeper insights was expanded upon using the random forest model. The Random Forest model was first run on the male dataset using the same parameter of 200 trees. Below, the results of this model are displayed:

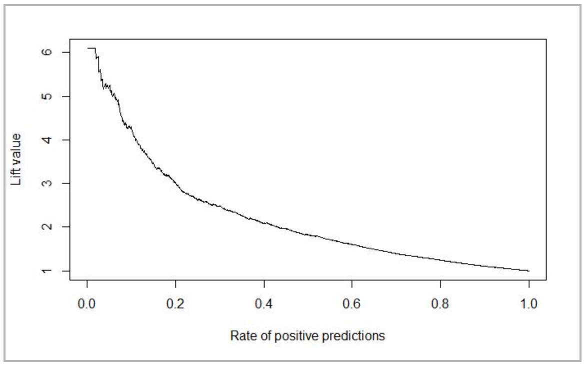


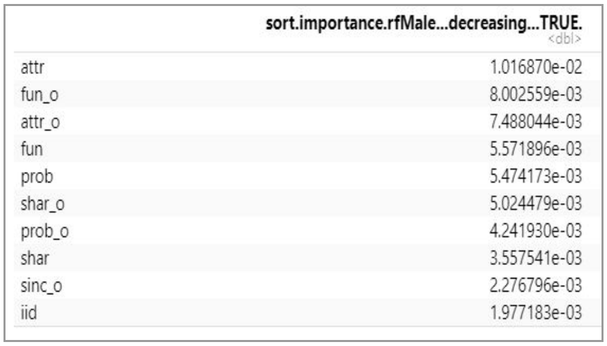
```
    actual
pred    0    1
   0 1046  168
   1    6   38
```

Accuracy: 0.8616852



AUC: 0.8569484

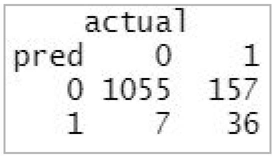
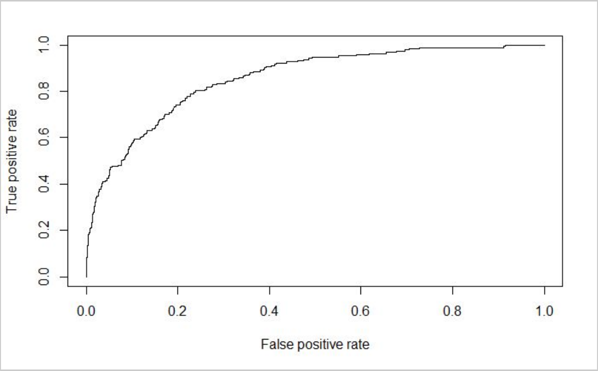
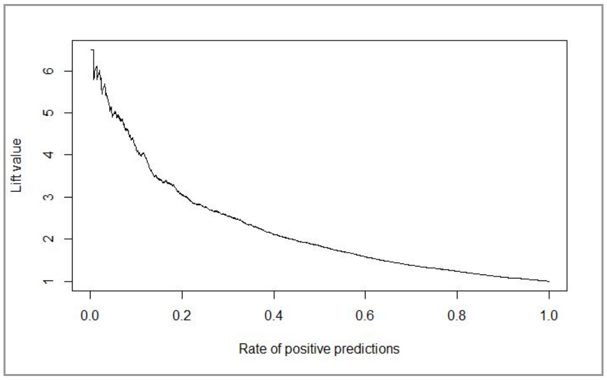
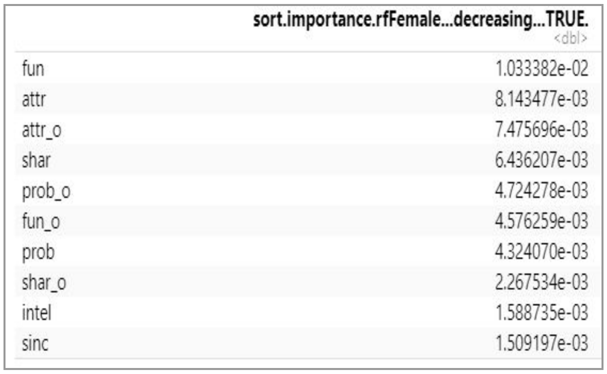




	sort.importance.rfMale...decreasing...TRUE. <dbl>
attr	1.016870e-02
fun_o	8.002559e-03
attr_o	7.488044e-03
fun	5.571896e-03
prob	5.474173e-03
shar_o	5.024479e-03
prob_o	4.241930e-03
shar	3.557541e-03
sinc_o	2.276796e-03
iid	1.977183e-03

The male dataset shows very similar results to the full dataset model. Overall, the hypothesis that the accuracy of the model would improve on gender specific datasets continues to be false. As the male dataset has a very similar AUC and Accuracy to the full dataset model. Variable importance is very similar to the previous model with attractive and fun being the main contributors. The probability of a match is slighter higher though on this dataset as it is slightly more important than shared interest for males.

Next, the same model was performed, but this time on the female dataset. Again, the parameters used were 200 trees and found the results below.

 actual pred 0 1 0 1055 157 1 7 36	Accuracy: 0.8693227
	AUC: 0.8557346
	 sort.importance.rfFemale...decreasing...TRUE. <dbl> fun 1.033382e-02 attr 8.143477e-03 attr_o 7.475696e-03 shar 6.436207e-03 prob_o 4.724278e-03 fun_o 4.576259e-03 prob 4.324070e-03 shar_o 2.267534e-03 intel 1.588735e-03 sinc 1.509197e-03

In the female dataset again there is no improvement in the accuracy and means, again displaying a key takeaway that model performance does not improve when the data is split by gender. Variable importance shows a change in the most influential variable to the model. The model indicates being fun is the most influential variable, which is different compared to the random forest on the male dataset where attractive was the most influential feature. This shows that the female dataset cared more about their date being fun than attractive which is the opposite of the male dataset.

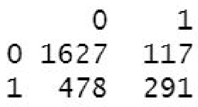
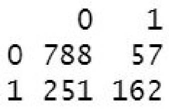
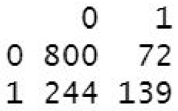
Overall the random forest gave us significantly better results compared to the decision tree models. It also gave a deeper insight into gender specific datasets than the decision tree models did. Even though the most influential features are the same across all the datasets the order of these features varies from female to male.

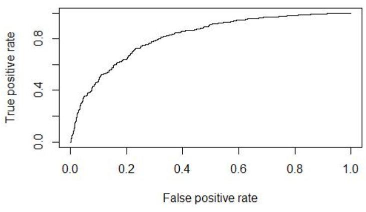
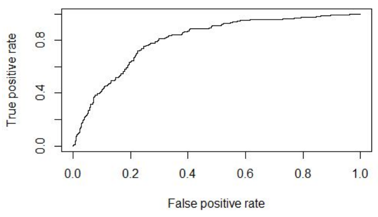
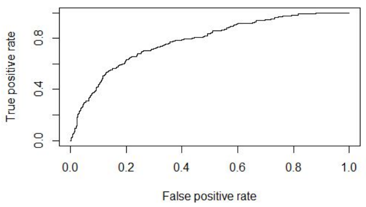
Naive-Bayes Models

The next model that was used is the Naive-Bayes for classification because of its simplicity. The model structure and code is similar to how the random forest model was written and conducted. All the data used for the models were split 70% to training and 30 % test previously. For this model, there were three different iterations of the mode run on three different datasets: *speed_date* dataset, male dataset and the female dataset. In all the models implemented a subset was made from the training data used in each model excluding these variables: *zip code,like,dec_o,dec, from, like_o*. These variables were excluded because they were too similar to the target variable and in the case of *zip code*, it was causing a lot of noise in the models as well as errors.

In the models which were implemented, the e1071 library was used to call the naiveBayes function with *dummymatch* as the target variable. The method used for each model was class. When predicted for the test data, "type=raw" had to be used in order for us to be able to calculate the accuracy. For all the confusion matrices "max.col" function must be used in order to calculate it because the levels were not the same for predicted value and target variable. Below are each of the confusion matrices, accuracies, and AUC curves.

Both Gender Naive Bayes	Male Data Naive Bayes	Female Data Naive Bayes

 `      0    1`
`0 1627  117`
`1  478  291`
	 `     0   1`
	`0 788  57`
	`1 251 162`
		 `    0   1`
		`0 800  72`
		`1 244 139`

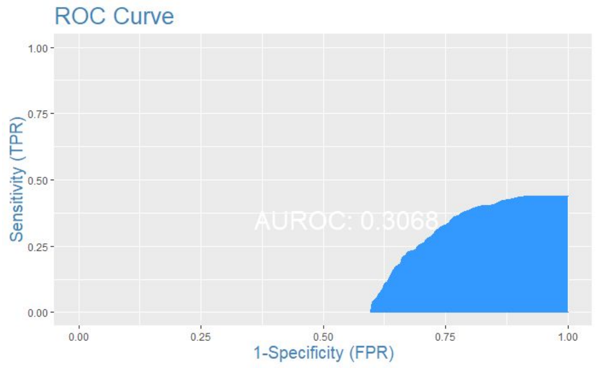
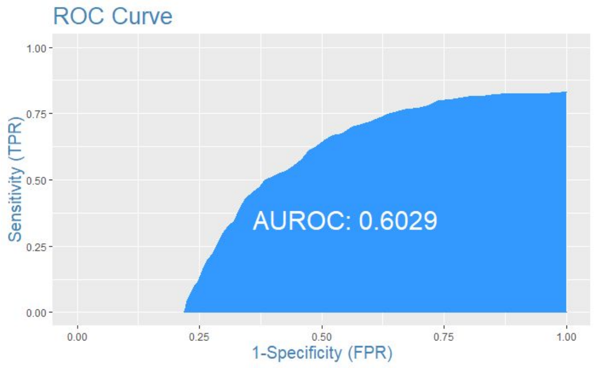
Accuracy : 76.3%	Accuracy : 75.5%	Accuracy : 74.8%
AUC = 81.6 	AUC = 80.6 	AUC= 82.8 

As seen in the table below, the accuracy for the Naive-Bayes models are much lower compared to the baseline model, decision tree. The confusion matrices show for the male and female models that the true positive values for the female dataset is much higher than male one. It is also apparent that the true negative value for male dataset is higher than the female one. The accuracy of the naive-bayes model ranges from 74% - 76%. With the female dataset reporting the lowest accuracy and there was no significant difference in accuracy between all these three iterations of the model. The accuracy for the naive-bayes models is so low because it is a very simple classifier, because the scope is only whether more individuals would match or not this makes sense for the speed date data which contained quite a large amount of missing values. When looking at AUC values and the graphs, the AUC graphs are not as smooth compared to random forest models. The model done with the female dataset has the highest AUC value compared to the others and the male dataset model has the lowest. It is possible to see that the false positive value is much higher for the female dataset compared to male dataset in the confusion matrix.

Logistic Regression

Logistic Regression is a general linear model which deals primarily with binomial data. By developing a logistic regression model the goal is to classify the candidate matches, coded in binary with 0 = no and 1 = yes. For this dataset, it was imperative to understand how match or no match occurred in conjunction with the independent variables. Next, some variables were removed because they were not useful in the context of the research question.While the rest were selected by the results of previous models and variables that were hypothesized some significance. A logistic regression on the entire dataset was not possible due to the enormous amount and types of missing data that was present in the dataset.

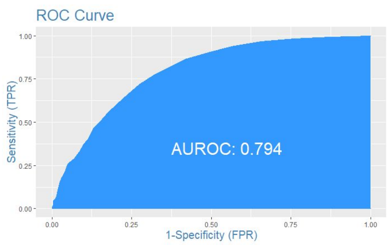
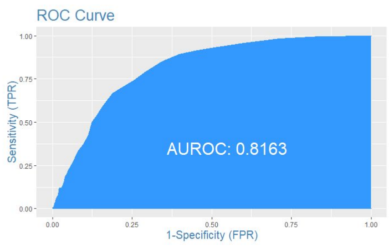
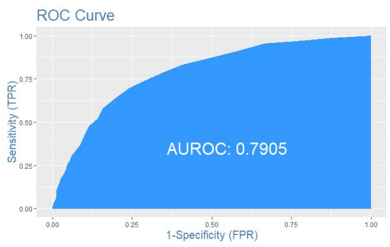
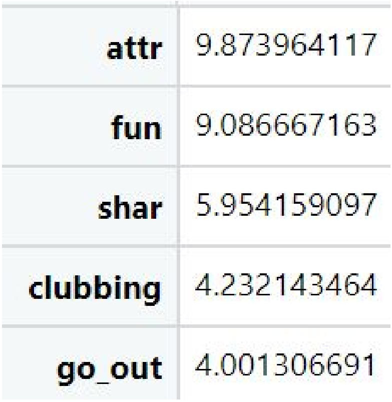
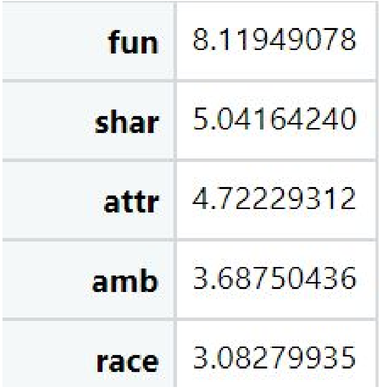
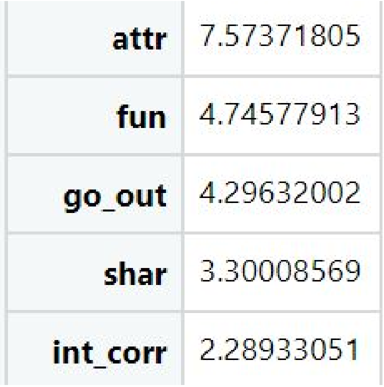
Since, there was no way to remove or account for all the missing data present in the dataset, two different logistic regressions were run, with two different sets of predictors. In Log Reg 1, income was included along with all other variables, and Log Reg 2 did not include income. As seen below, the ROC curve for Log Reg 2 is much higher than Log Reg 1, due to how income was classified within the model.

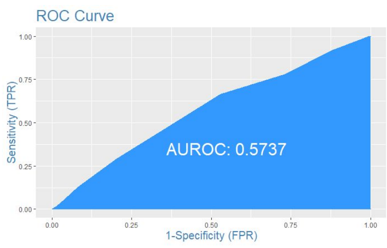
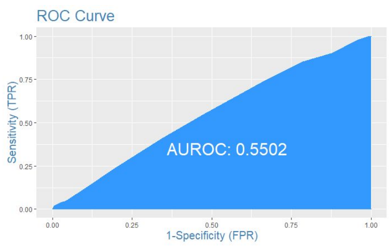
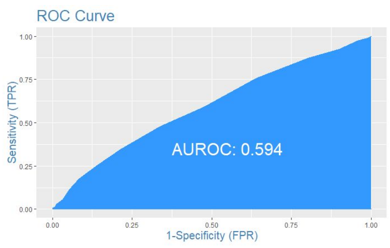
Base Models	
LOG REG 1	**LOG REG 2**
dummymatch ~ iid + wave +position + pid + int_corr +samerace + field_cd +career_c + imprace +imprelig + race + goal + go_out + sports+ tvsports +exercise+ dining +museums+ art+ hiking+ gaming+ clubbing +reading+ tv+ theater+ movies+ concerts+ music+ shopping+ yoga + attr1_1 + sinc1_1+ intel1_1 + fun1_1 + amb1_1 +shar1_1 +attr+ sinc +intel +fun +amb +shar+like +prob+met +income	dummymatch ~ iid + wave +position + pid + int_corr +samerace + field_cd +career_c + imprace +imprelig + race + goal + go_out + sports+ tvsports +exercise+ dining +museums+ art+ hiking+ gaming+ clubbing +reading+ tv+ theater+ movies+ concerts+ music+ shopping+ yoga + attr1_1 + sinc1_1+ intel1_1 + fun1_1 + amb1_1 +shar1_1 +attr+ sinc +intel +fun +amb +shar+like +prob+met
	

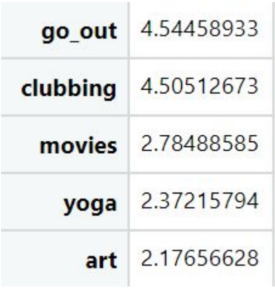
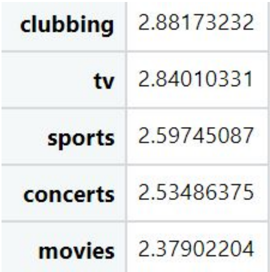
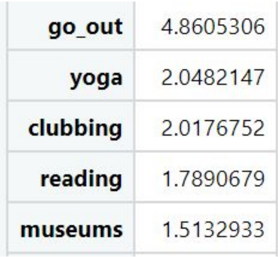
This first iteration of a logistic model was improved by finding out which variables were more important to the match variable by conducting a random forest model and then using the varImp function. Much like the decision tree baseline model, data leakage had to be decreased by eliminating and adding variables which gives us a much more accurate representation of the matches. Improvement was also achieved t imputing for missing values or replacing the missing values. This caused the amount of data leakage to be lower as well as help boost the AUC and accuracy of the models created.

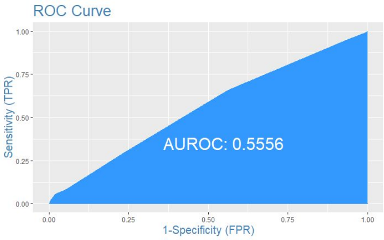
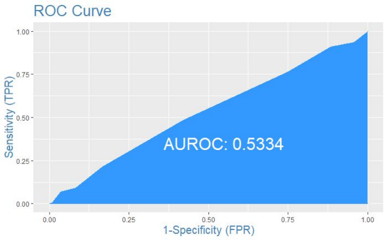
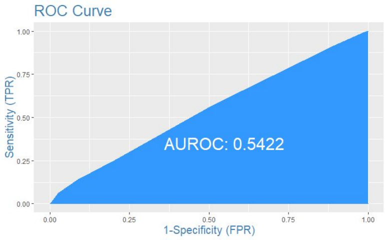
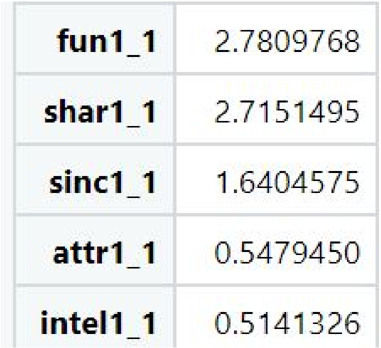
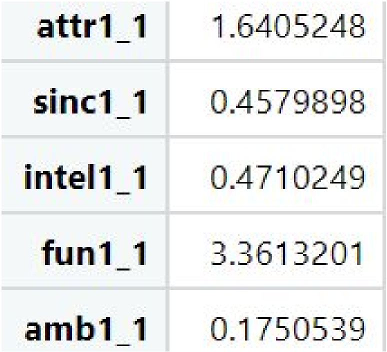
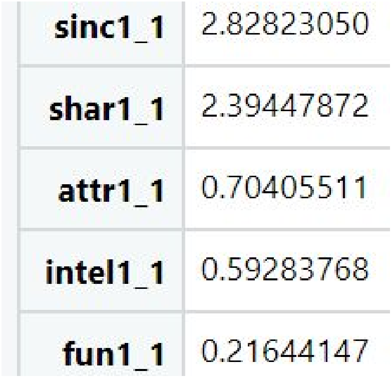
GLM Results

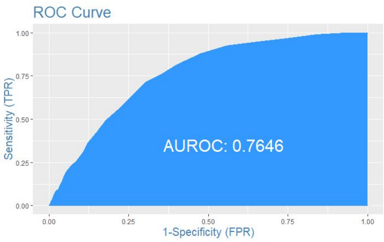
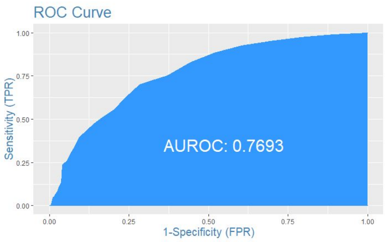
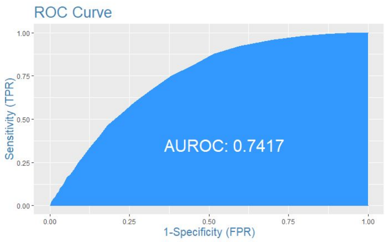
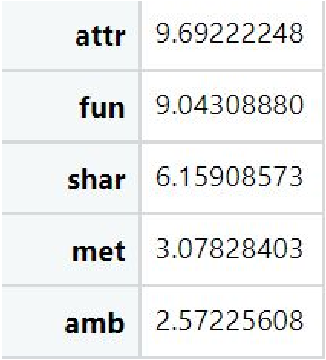
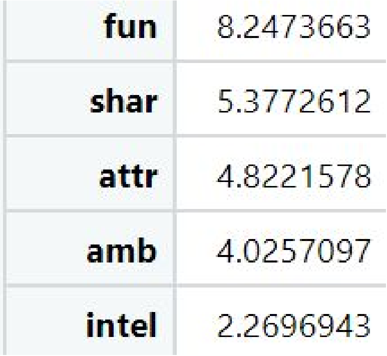
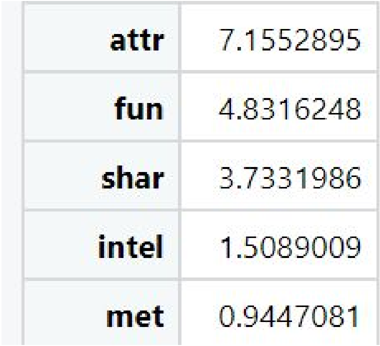
After all of the missing values had been properly taken care of, there were dramatic changes in the logistic regressions. It is here that it was decided to explore predicting the dependent variable with different subsets. The first GLM was a combination of participant background information, what they look for in a partner, hobbies/interests, and scorecard result from each individual they met. While the others were focused on the four sets of characteristics mentioned. Lastly for each GLM conducted on the entire population, the exact GLM was conducted on each gender. This was done in order to see if any relevant difference would appear between male and females. The following are the results of out models:

Logistic Regression on all relevant variables.		
Both Genders	**Female**	**Male**
		
Accuracy: 0.8432	**Accuracy**: 0.8271	**Accuracy**: 0.8235
Sensitivity: 0.1542289	**Sensitivity**: 0.1572052	**Sensitivity**: 0.1287554
Specificity: 0.9744197	**Specificity**: 0.9766082	**Specificity**:0.9814634
 attr 9.873964117 fun 9.086667163 shar 5.954159097 clubbing 4.232143464 go_out 4.001306691	 fun 8.11949078 shar 5.04164240 attr 4.72229312 amb 3.68750436 race 3.08279935	 attr 7.57371805 fun 4.74577913 go_out 4.29632002 shar 3.30008569 int_corr 2.28933051

Logistic Regression based on Hobbies/Interests		
Both Genders	**Female**	**Male**
		
Accuracy: 0.84	**Accuracy**: 0.8175	**Accuracy**: 0.8148

Sensitivity: 0	**Sensitivity**: 0	**Sensitivity**: 0
Specificity: 1	**Specificity**: 1	**Specificity**: 1
 go_out 4.54458933 clubbing 4.50512673 movies 2.78488585 yoga 2.37215794 art 2.17656628	 clubbing 2.88173232 tv 2.84010331 sports 2.59745087 concerts 2.53486375 movies 2.37902204	 go_out 4.8605306 yoga 2.0482147 clubbing 2.0176752 reading 1.7890679 museums 1.5132933

Logistic Regression on Based what you look for in the opposite sex		
Both Genders	**Female**	**Male**
		
Accuracy: 0.84	**Accuracy**: 0.8175	**Accuracy**: 0.8148
Sensitivity: 0	**Sensitivity**: 0	**Sensitivity**: 0
Specificity: 1	**Specificity**: 1	**Specificity**: 1
 fun1_1 2.7809768 shar1_1 2.7151495 sinc1_1 1.6404575 attr1_1 0.5479450 intel1_1 0.5141326	 attr1_1 1.6405248 sinc1_1 0.4579898 intel1_1 0.4710249 fun1_1 3.3613201 amb1_1 0.1750539	 sinc1_1 2.82823050 shar1_1 2.39447872 attr1_1 0.70405511 intel1_1 0.59283768 fun1_1 0.21644147

<table>
<tr><th colspan="3">Logistic Regression on Based on Scorecard results</th></tr>
<tr><th>Both Genders</th><th>Female</th><th>Male</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr><td><b>Accuracy</b>: 0.8368</td><td><b>Accuracy</b>: 0.8215</td><td><b>Accuracy</b>: 0.814</td></tr>
<tr><td><b>Sensitivity</b>: 0.05970149</td><td><b>Sensitivity</b>: 0.1048035</td><td><b>Sensitivity</b>: 0.05150215</td></tr>
<tr><td><b>Specificity</b>: 0.9848413</td><td><b>Specificity</b>: 0.9814815</td><td><b>Specificity</b>: 0.9873171</td></tr>
<tr>
<td>
<table>
<tr><td>attr</td><td>9.69222248</td></tr>
<tr><td>fun</td><td>9.04308880</td></tr>
<tr><td>shar</td><td>6.15908573</td></tr>
<tr><td>met</td><td>3.07828403</td></tr>
<tr><td>amb</td><td>2.57225608</td></tr>
</table></td>
<td>
<table>
<tr><td>fun</td><td>8.2473663</td></tr>
<tr><td>shar</td><td>5.3772612</td></tr>
<tr><td>attr</td><td>4.8221578</td></tr>
<tr><td>amb</td><td>4.0257097</td></tr>
<tr><td>intel</td><td>2.2696943</td></tr>
</table></td>
<td>
<table>
<tr><td>attr</td><td>7.1552895</td></tr>
<tr><td>fun</td><td>4.8316248</td></tr>
<tr><td>shar</td><td>3.7331986</td></tr>
<tr><td>intel</td><td>1.5089009</td></tr>
<tr><td>met</td><td>0.9447081</td></tr>
</table></td>
</tr>
</table>

First, it is necessary to discuss the models which did not give us conclusive results. The regression on Hobbies/Interests and the survey on what you look for in the opposite sex provided us with ROC curves that look no better than a coin flip. One hypothesis is that variables were not relevant enough to create a prediction. What an individual wants may be different from what they choose when it comes to matching with someone. While no specific interests can provide you with more matches, what could be of value is finding out for future use would be a value assigned to how many interests they have in common. Perhaps two individuals are more likely to match if they have more in common rather than being interested in a particular hobby.

The logistic regression on all relevant variables led to three common variables showing up in the coefficients. Attractiveness, fun, and sharing common interests were all relevant when it came to predicting whether or not two individuals would match. For a woman how “fun” a participant was was the highest coefficient in determining whether or not those two individuals would match. However, for men, how attractive a participant was was the determining factor in whether or not two individuals would match. Some noticeable differences were women preferred participants who were ambitious and had the same race as them. While men were interested in participants that showed an interest in going out and shared the same hobbies as them.

At the end of a speed dating session each participant had to evaluate the participants they met as well as determining whether or not they would decide to match with that individual. It was hypothesized that at this moment the evaluations can properly classify whether or not two individuals would match. The more relevant variables for both genders both included Fun, Sharing Common Interests, and attractiveness, and intelligence. A noteworthy point is from variables listed intelligence was lowest ranking one among them. An exclusive variable for the females was ambition, women were more likely to match with a participant if they presented themselves that way. Males were more likely to match with someone if they had met the person previously.

Ridge and Lasso Models:

To improve the logistic regression model Ridge and Lasso models were used to regularize the coefficients. The Ridge and Lasso model are regularizing models that will shrink coefficients that become too large based on a penalty defined by lambda. The difference between these two models is that the ridge model does not remove insignificant variables but leaves them in the model, with very small coefficients. However, the Lasso model will penalize coefficients to zero, effectively removing these coefficients from the model. Both approaches were executed in order to minimize the error found in the model. Multiple Ridge and Lasso models were run using different values of lambda, in particular minimum lambda and lambda.1se. Below you can see the ridge and lasso model results using these different penalty values.

RIdge Model	
Lambda Value	Mean Squared Error
Lambda.min	11.20285
Lambda.1se	9.507508

Lasso Model	
Lambda Value	Mean Squared Error
Lambda.min	11.71545
Lambda.1se	9.9597

Overall from running these different models it is apparent that the Ridge Models tend to have lower MSE's regardless of the Lambda value. Also the preferred Lambda is the, Lambda.1se, value because it has the lowest error. Using this information the coefficients for the Lasso and Ridge models were calculated based on the preferred Lambda values.

Ridge Model Coefficients	**Lasso Model** Coefficients
(Intercept) -4.334011e+00	(Intercept) -4.542865189
iid -2.195161e-05	iid .
wave -2.153056e-04	wave .
position -3.394332e-04	position .
pid 4.327780e-05	pid .
int_corr 1.219458e-01	int_corr .
samerace 5.333656e-03	samerace .
field_cd -7.301337e-03	field_cd .
career_c -2.585687e-04	career_c .
imprace -2.115834e-02	imprace -0.001934979
imprelig -1.916403e-03	imprelig .
race -4.356839e-03	race .
goal -1.809810e-02	goal .
go_out -8.154797e-02	go_out -0.069549460
sports 2.180597e-03	sports .
tvsports -8.787118e-03	tvsports .
exercise 2.089232e-03	exercise .
dining 1.038310e-02	dining .
museums -6.193179e-03	museums .
art 1.004573e-02	art .
hiking 8.566895e-03	hiking .
gaming -1.709204e-03	gaming .
clubbing 2.982120e-02	clubbing 0.009715697
reading -3.548345e-04	reading .
tv -3.264881e-03	tv .
theater -9.564172e-03	theater .
movies -2.627138e-02	movies .
concerts 1.688862e-02	concerts .
music 4.605636e-03	music .
shopping -1.271549e-02	shopping .
yoga 1.559084e-02	yoga .
attr1_1 -6.650975e-04	attr1_1 .
sinc1_1 -5.568140e-03	sinc1_1 .
intel1_1 6.512281e-03	intel1_1 .
fun1_1 7.902105e-03	fun1_1 .
amb1_1 -2.847770e-03	amb1_1 .
shar1_1 -8.169726e-03	shar1_1 .
attr 1.367453e-01	attr 0.199587703
sinc 4.554127e-02	sinc .
intel 4.967896e-02	intel .
fun 1.229872e-01	fun 0.198047224
amb 2.810467e-03	amb .
shar 7.627502e-02	shar 0.074636791
met 4.019764e-02	met .

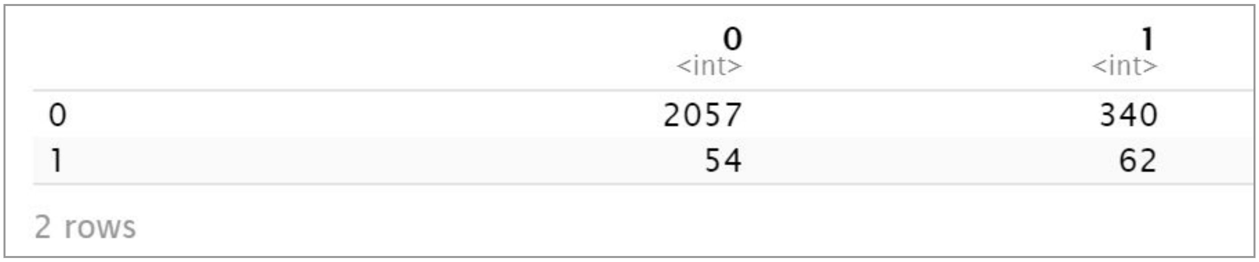
Sum square of Coefficients: 0.07122332	0.08956387

Overall between the two models the important variables are similar, with attractiveness and fun being the two most significant variables across all of the models. The major difference that this illustrates is how the Lasso model removes variables from the model, while the ridge model keeps all values in the model and just reduces the coefficient significantly. Also some variables are interpreted differently between the two models, for example clubbing is a positive coefficient in the Lasso model while it is a negative coefficient in the Ridge model. Overall the ridge model is preferred for the dataset because it returns the lowest MSE for the data.

Explanation about dataset (origins, features, size, label imbalance, ...) possibly with descriptive analysis.

Outcomes

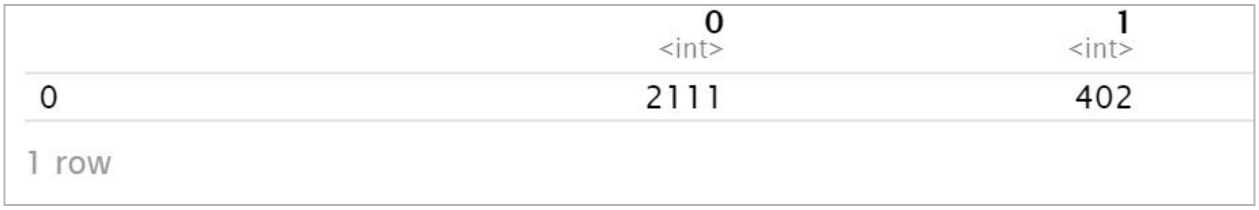
The ability to predict whether two individuals will match is in fact possible, however this does not mean the experiment was perfect. After taking a look back we see that the following issues may be impacting the current results. The first of which is the unbalanced data set. The number of matches in the data set were few, and thus made classifying matching participants harder. This leads to the fact that the data models did not take into account weights for the data sets. Perhaps increased the values of the true positives would have resulted in better results. In one of the confusion matrices, only 62 matches were correctly predicted.



	0 <int>	1 <int>
0	2057	340
1	54	62

2 rows

When tackling the models, some of the logistic regression were not able to predict any true positives that lead to misleading confusion matrices. Below you will see results of a confusion matrix. That was not able to predict anything as a true positive. Lastly was the inability to have income as a relevant variable in any models. Countless issues occurred when trying to adapt income into any of the models. Even after dividing the incomes into quartiles, they only hurt the models and did not improve them. Every time income was included in any model: AUC was lowered, ROC curved worsened, and accuracy dropped. Even as the end of the experiment approaches a solution to this issue was not able to be discovered. Thus it is entirely possible that important information is missing all of the models.



	0 <int>	1 <int>
0	2111	402

1 row

Conclusion

What an individual looks for in a mate and what an individual wants can be two different things. After multiple iterations, the conclusion of the experiment were that the most important factors when trying to find a match is how attractive and fun you can present yourself. Although it may not come as a surprise that attractiveness is a major factor of when two should match. It can be a simple truth that can remind that two people who find one another attractive is really all that is necessary to be a match with someone. Attractiveness is a variable that is hard to improve on, it is out of your control whether the opposite gender will find you attractive. The relevancy of the fun variable was surprising. It makes sense that a participant would want to meet up with someone if they had an enjoyable experience with them. It creates the impression that this person is someone you would like to spend time with outside the event. So the final conclusion when trying to find your perfect match is to try to come across as someone entertaining and exciting.

One of the key steps missed during the analysis is normalizing the data. An opportunity was missed when removing zip code, it could have influence on how people choose their partners, we also were unable to analyze how income played a part in how people matched with others due to how the data was organized in the source file.Some of the analysis that we missed are comparing the attributes and expectations with both genders themselves. Such as, what do male participants think other males in the study value in their matches, and the same for female participants. In the future we could explore more of how male and female participants thought about the opposite gender but in the context of answering the survey questions. We would also like to try to explore the attributes distributed within each gender more as well as find out what attributes were important to each gender. Analyze whether a candidate over values themselves compared to how the opposing gender rates them and is overlooked. It would be also beneficial in the future to try and develop a SVM classifier to figure out whether the participants match in regards with the attributes provided. Another thing that we would like to try in the future is to classify how men and women describe attributes of themselves and each other, and how accurate this classification could be.

References

Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, *121*(2), 673-697

McDonald, C. (2014, February 15). Speed dating: Why are women more choosy? Retrieved November 11, 2020, from https://www.bbc.com/news/magazine-26172314

Tidwell, N. D., Eastwick, P. W., & Finkel, E. J. (2013). Perceived, not actual, similarity predicts initial attraction in a live romantic context: Evidence from the speed‐dating paradigm. *Personal Relationships*, *20*(2), 199-215.