Speed Dating Experiment - Capstone Final Report

Jose Reyes – Dec. 2017
Springboard Data Science Career Track

I. Introduction

The dating scene of recent years has seen the rise of a plethora of web-based dating apps such as Tinder, OkCupid, and Happn, which cater to the finding a romantic partner through the convenience of a mobile device. However, the weakness of such apps is that the human element (i.e. personality, chemistry… etc.) rarely comes across on an online profile, thus a potential connection can often either be dismissed or over-looked. It is for this reason that sites like match.com also advertise speed-dating events so that participants can meet new people and improve their chances of finding a romantic partner through their site. Introducing participants to each other and limiting dates to a four-minute introduction allows participants to gauge the human element and quantify various personality traits such as intelligence and humor and potentially discover a match that could have been otherwise overlooked on the online platform alone. This study will utilize data from speed dating events and will aim to identify the variables that predict a yes response from a speed-date partner, indicating that he/she would like to see the participant again. Though a predictive algorithm can be applied to both genders, this study will specifically focus on identifying the variables that will most likely yield men a yes from their female speed-date partner.

Utilizing the findings from this study, firms that host speed-dating events for their participants, either as a stand-alone service such as speedsanjosedating.com, or to supplement an online-dating platform, such as match.com, can apply machine learning algorithms to predict a yes/no response from their participants' speed-date partners and utilize the findings to recommend them to other in-event partners based on the preferred personality attributes. Coupled with participant permission, the firm can also aggregate and publish these personality-trait scores along with any other favorable key words to each participant's online profiles, thereby creating a more humanistic recommendation to other like-minded online users so as to better pique their interest.

II. Description of the Dataset and Data Wrangling

The speed-dating dataset utilized for this study was obtained from the Kaggle website and contains 195 variables and 8,378 rows and is stored in personal Google Drive so as to maintain the integrity of the original data.[1] This dataset is a collection of 4-minute dating encounters by 551 individuals (274 men and 277 women) participating in twenty-one speed dating events, referred to as "waves", held between October 16th, 2002 and April 7th, 2004. Transcripts of each speed-date, if available, can also be used for additional text mining to correlate against the given numeric ratings. The raw data contains each speed-date encounter, from both the male and

[1] https://www.kaggle.com/annavictoria/speed-dating-experiment

female perspective, as well as data collected during various times intervals, namely prior to the speed-date event and in the days and weeks after. Because this study is focused around the data collected at the time of each speed-date, the revised data frame will be comprised of basic demographics of each participant, namely age and race, their social and dating habits as well as the ratings received from each speed-date partner.

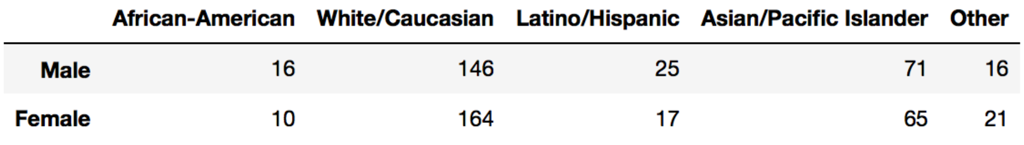
The Pandas python library was used to import the data from the Google Drive, inspect the each of 195 variables and identify the participant demographic data and in-event variables; effectively reducing the data frame to 35 variables with 8,378 rows. The Imputer from Scikit-Learn along with the most frequent value of each was column was used to populate missing data. Lastly, the finalized data frame was segmented into male and female data frames for data exploration and statistical analysis.

III. Data Story

Who are the participants?

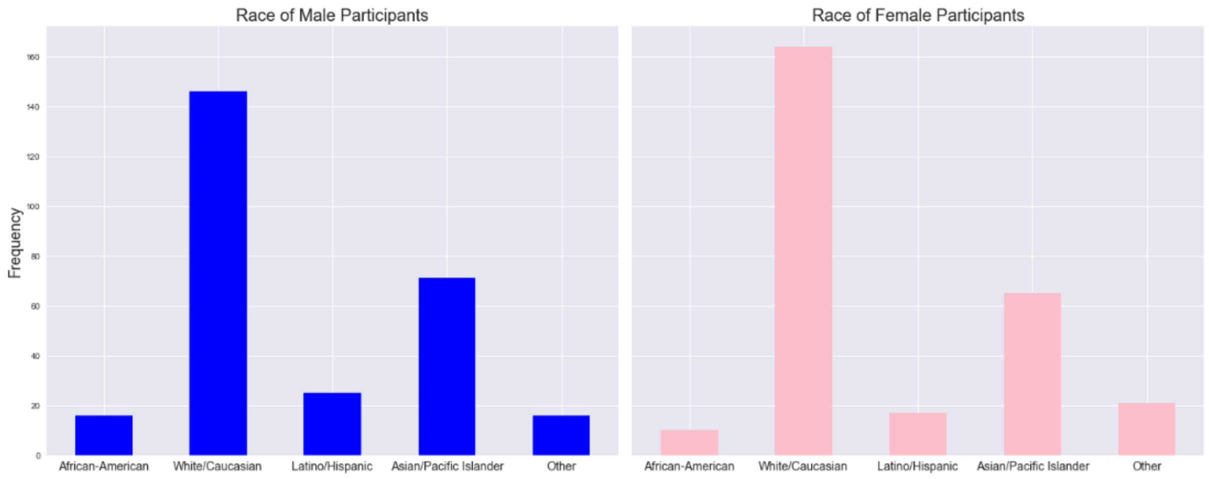
The participants of this study are primarily a mix of graduate and undergraduate students from Columbia University. The majority of the participants identify themselves as White/Caucasian with Asian/Pacific Islander being the next highest ethnic demographic among the participants. This break-down in ethnic and gender demographics reflects the current trends observed by the National Center for Education Statistics.[2]

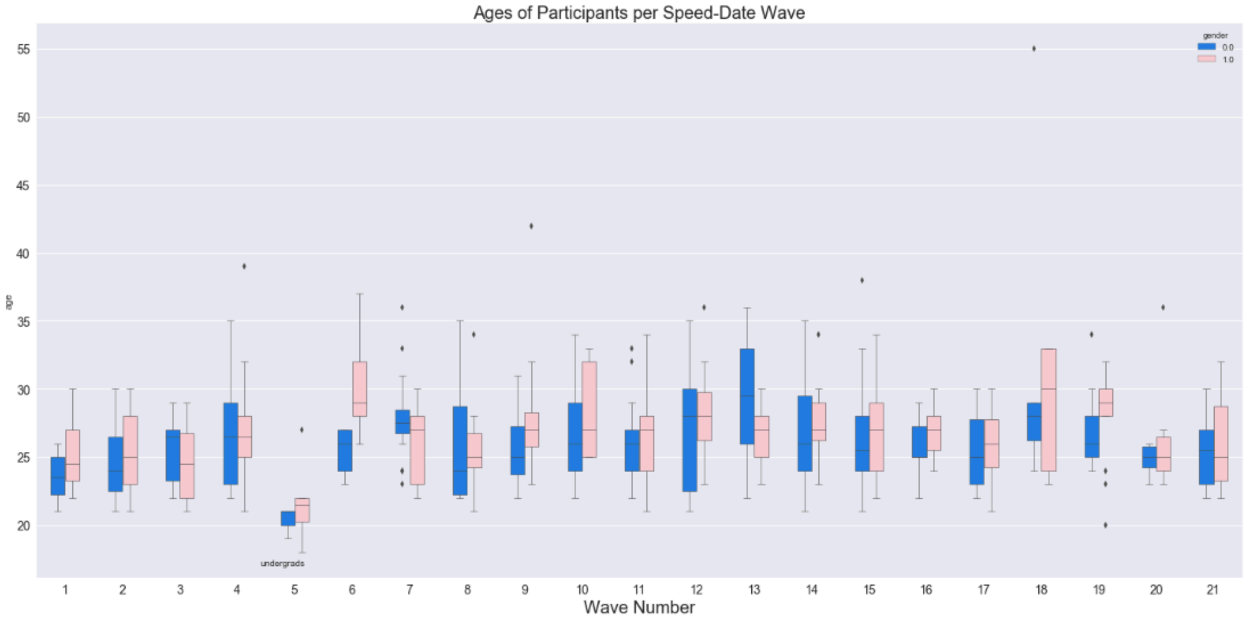
Initial analysis of the age for each gender shows that overall the female population was slightly older and had a higher age variance than the male population with the median age of the male participants being 26.0 and the average being being 26.14. Conversely, the median age of the female participants is 27.0 with the average being 26.59. A few participants were outside for the traditional college age with one male participant being 55 and one female participant being 42; these participants were not excluded from the analysis as they were assumed to be graduate college students. The box plots, shown in the next page, illustrate these findings. Wave 5 is important to note as it is comprised of only college undergrads that mainly belong to the traditional college age of 18-22, however this wave also follows the trend of an older and more age-variant female population.



	African-American	White/Caucasian	Latino/Hispanic	Asian/Pacific Islander	Other
Male	16	146	25	71	16
Female	10	164	17	65	21

[2] https://nces.ed.gov/pubs2010/2010015/indicator6_24.asp

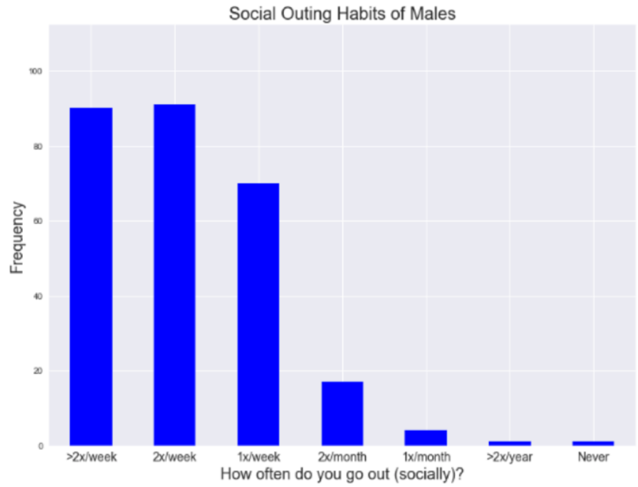


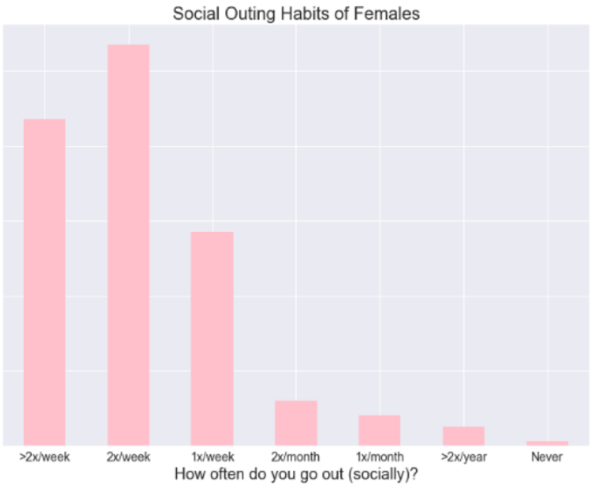


What are the social habits and why are they participating in these speed-dating events?

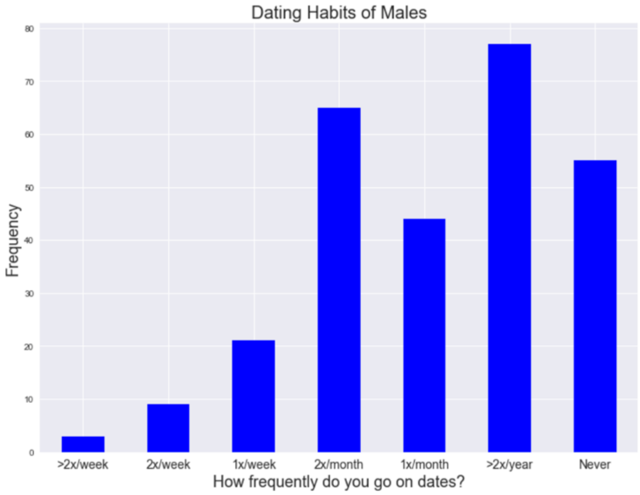
The large number of the participants enjoy an active social life with the majority of participants reporting as going out socially at least once a week or more. While women do report going out socially at least twice a week more so than the men, there is no major difference between social outings among the genders. Interestingly the trends flip when it comes to dating with the majority participants reporting as dating less frequently than going out socially.

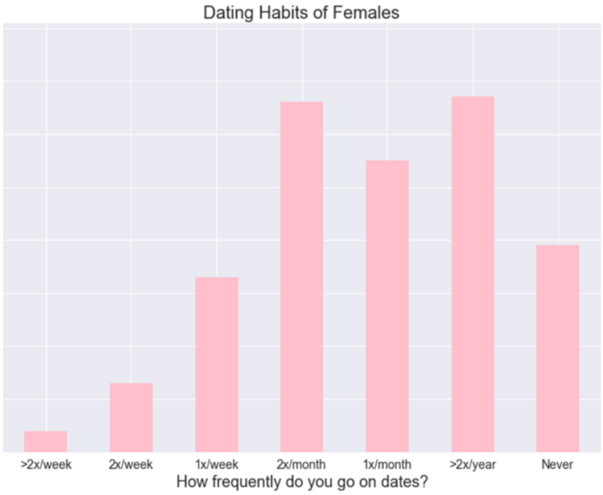
While some observable gender differences in dating frequency exist, overall the dating frequency for both genders is relatively equal. What is interesting is that a sizable portion of the participant pool reports as rarely dating which creates the speculation that participation in these events provides an opportunity to go out socially and meet new people. In fact, for both men and women, the primary reasons for participating in these speed-dating events was enjoyment and to meet new people. Very few participants had anterior motives for participation and even fewer were deliberately looking for a romantic connection in these events; this indicates that these event were mostly marketed as a fun way to meet new people or the participants perceived them as such.



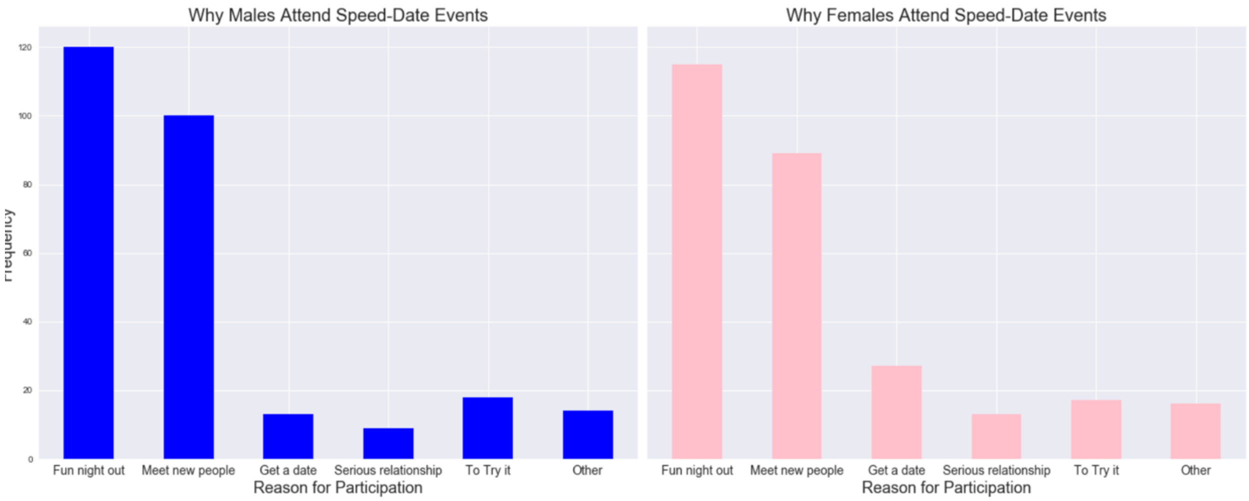


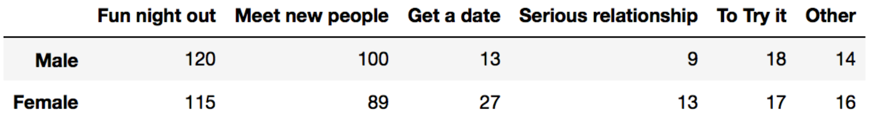
	>2x/week	2x/week	1x/week	2x/month	1x/month	>2x/year	Never
Male	90	91	70	17	4	1	1
Female	87	107	57	12	8	5	1





	>2x/week	2x/week	1x/week	2x/month	1x/month	>2x/year	Never
Male	3	9	21	65	44	77	55
Female	4	13	33	66	55	67	39



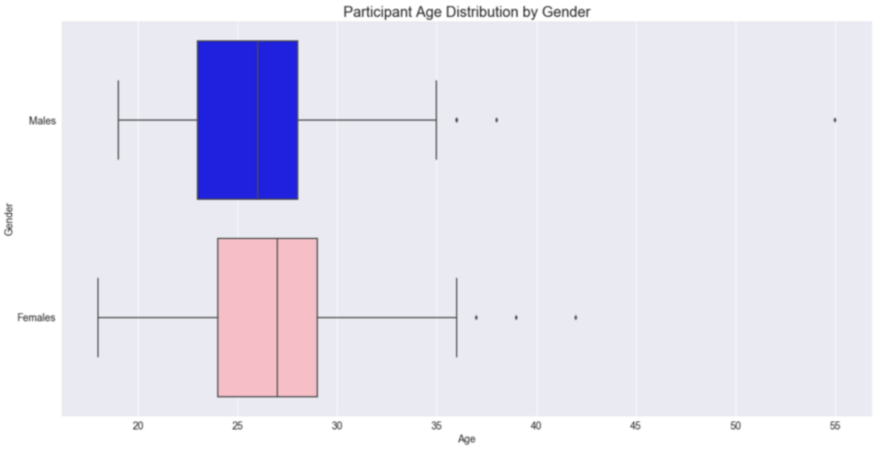


	Fun night out	Meet new people	Get a date	Serious relationship	To Try it	Other
Male	120	100	13	9	18	14
Female	115	89	27	13	17	16

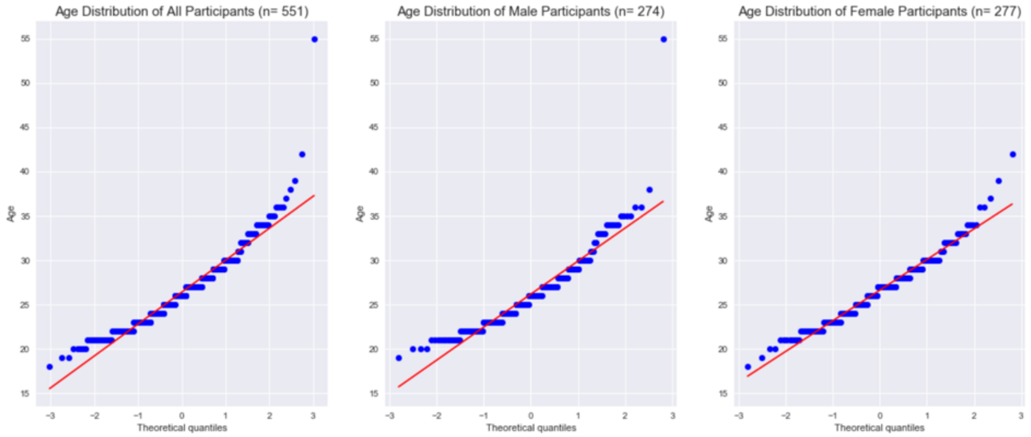
IV. Inferential Statistics

Age Demographics of the participants

Of the 21 speed-dating events analyzed, the median age of all participants was 26 and the average age of all participants being 26.37 having a standard deviation of 3.74. The oldest participants identified was a 55-year-old male, assumed to be a graduate student, and the youngest participant was an 18-year-old female. Specifically, males had a median age of 26 and an average age of 26.14 with a standard deviation of 3.96 while females had a median age of 27 and an average age of 26.59 with a standard deviation of 3.50.

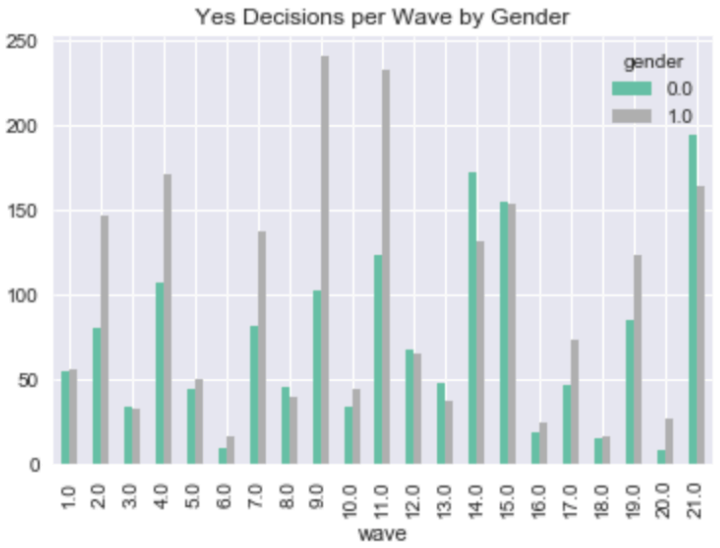


Collectively, the ages of the participants not normally distributed and are positively-skewed. The same was identified as true for males and females, respectively.



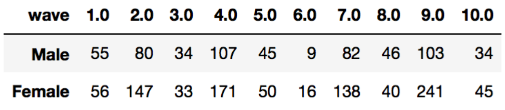
Speed-date outcomes

At the end of each 4-minute speed date participants must complete a 10-question scorecard rating their partner's various attributes such as level of attractiveness, sincerity and intelligence among others. The scorecard culminates with a yes/no decision of their date partner, indicating whether or not the participant would like to see them again. Interestingly males appear to be far more selective in their yes responses given that in 15 speed-dating events females made more yes decisions than their male counterparts. This goes against the traditional notion of evolutionary psychology which suggests that females are far more selective than their male counterparts as females pay a “greater reproductive cost by making a wrong choice.”[3] However because participants are attending these events primarily for enjoyment and to meet new people, mate selection pressures bare little influence in making a yes response.

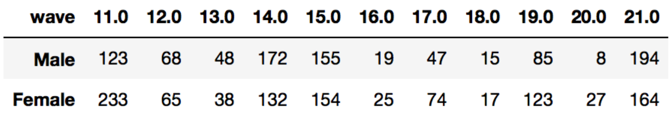


[3] https://www.psychologytoday.com/blog/the-scientific-fundamentalist/200908/are-women-always-more-selective-in-mate-choice-men-i

“Yes” responses by gender



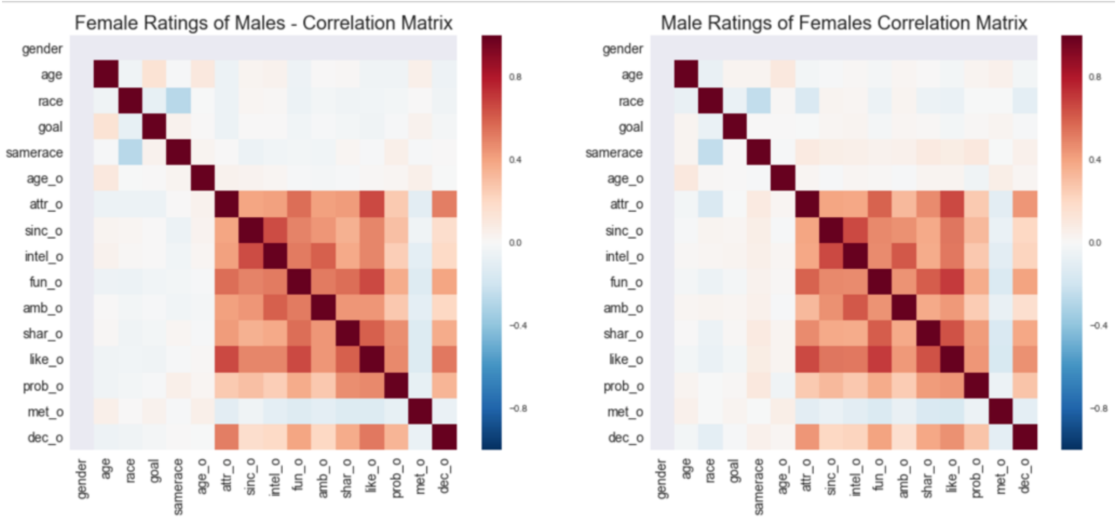
wave	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
Male	55	80	34	107	45	9	82	46	103	34
Female	56	147	33	171	50	16	138	40	241	45

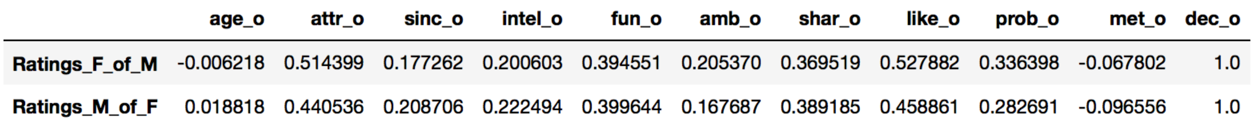


wave	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0
Male	123	68	48	172	155	19	47	15	85	8	194
Female	233	65	38	132	154	25	74	17	123	27	164

Correlation Studies

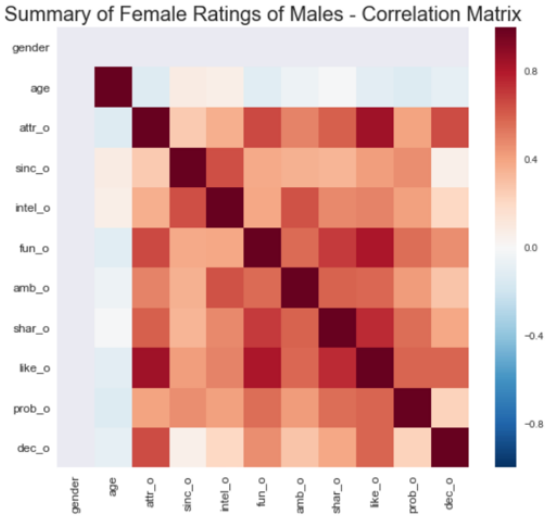
The male and female data frames were analyzed to identify correlations between the dec_o variable, indicating a partner’s yes/no response, and various other attributes. Surprisingly, participant demographics and motives for participation had no correlation with a yes/no response from their partner. Unsurprisingly however, at the speed-date encounter level, if a participant is deemed attractive and given a high “like” rating by their partner it correlates with a yes response. However, the correlations for attractiveness and like ratings of 0.44 and 0.45, respectively, from male participants and 0.51 and 0.52, respectively, from female participants are significant at best and not as strong as initially expected.

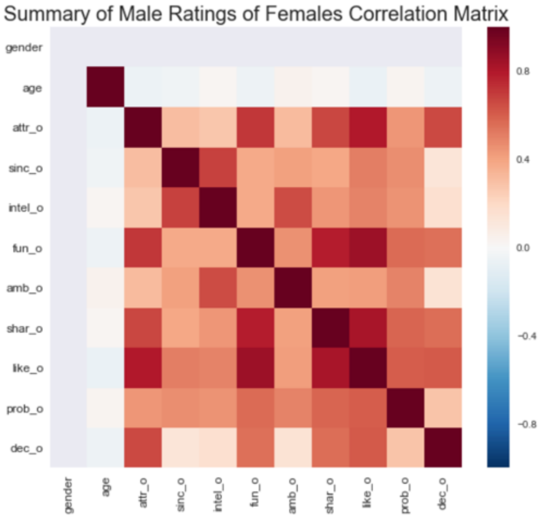


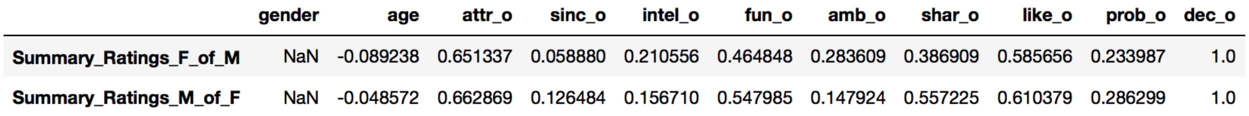


	age_o	attr_o	sinc_o	intel_o	fun_o	amb_o	shar_o	like_o	prob_o	met_o	dec_o
Ratings_F_of_M	-0.006218	0.514399	0.177262	0.200603	0.394551	0.205370	0.369519	0.527882	0.336398	-0.067802	1.0
Ratings_M_of_F	0.018818	0.440536	0.208706	0.222494	0.399644	0.167687	0.389185	0.458861	0.282691	-0.096556	1.0

Aggregating the average ratings for each participant and correlating it with the number of yes decisions received, a stronger correlation of the attractiveness and like ratings emerges. As expected, an additional attribute, fun_o (a partner’s rating of the participant’s level of fun) also emerges with a stronger correlation. The correlations for attractiveness, like, and fun ratings of 0.66, 0.61, and 0.55, respectively, from male participants and 0.65, 0.58, and 0.46 respectively, from female participants are stronger than the date-level correlations and will be considered as primary features in the predictive algorithm.







	gender	age	attr_o	sinc_o	intel_o	fun_o	amb_o	shar_o	like_o	prob_o	dec_o
Summary_Ratings_F_of_M	NaN	-0.089238	0.651337	0.058880	0.210556	0.464848	0.283609	0.386909	0.585656	0.233987	1.0
Summary_Ratings_M_of_F	NaN	-0.048572	0.662869	0.126484	0.156710	0.547985	0.147924	0.557225	0.610379	0.286299	1.0

Summary

The participants of these speed dating events are socially-active, mainly college-age individuals mostly looking for a fun night out and an opportunity and an opportunity to meet new people. The typical participant is an approximately 26 year-old Caucasian individual whose social and dating habits are mostly inversely proportional. Despite some variations, the overall ethnic and behavioral trends identified in this dataset are consistent with the clientele of college-targeted speed dating events.

The male data frame will be utilized for the machine learning component of this study and will be split into a striated training set comprising of 80% of the data and a testing set comprising of the remaining 20%. Utilizing the training set, a logistic regression analysis will initially be conducted, along with other algorithms, to identify the significant variables that best predict a positive result of the dependent variable "dec_o," denoting that a date partner would like to see the participant again. The independent variables will primarily be the personality attribute ratings the participant received from their partner, however with further exploration and analysis of the dataset, this list of independent variables may change depending on their level of significance in predicting a successful outcome. Utilizing the findings of the training set, the machine learning algorithm will be applied to the testing set and the predicted outcome will be compared against the actual outcome to measure the success and strength of the model.

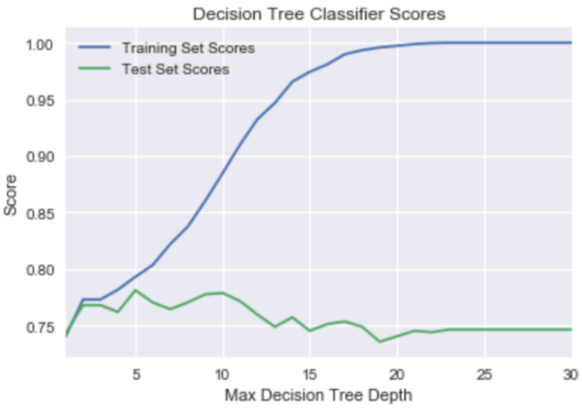
V. Machine Learning & Predictive Algorithm Results

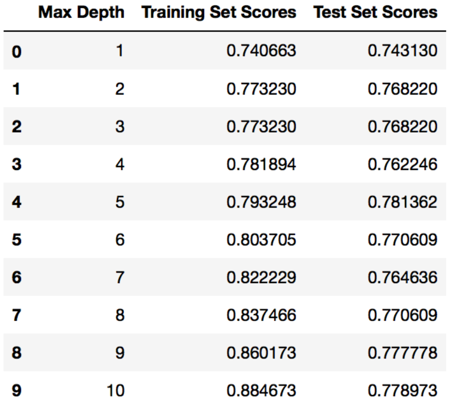
To predict whether or not a male's speed-date partner will make a "yes" response, three Scikit-Learn classification algorithms were used: Decision Tree, Random Forest, and Logistic Regression. Each of the algorithms was executed and scored, based on the performance results the algorithm was fine-tuned to maximize the performance. The dependent variable for this experience is the "dec_o" variable, while the independent variables are comprised of the variables listed on the correlation studies diagrams found in section IV.

To better asses the performance of each classification algorithm, a base model was created classifying all speed-date partner responses as "no" responses. Using test set data, the accuracy score of this mode is 0.5257, making this the accuracy score that the chosen classification algorithms had to beat.

Classifier 1: Decision Trees and Gradient-Boosted Trees

A decision tree was chosen because of its simplicity in dealing with noisy or incomplete data, its strong ability to identify the most discriminatory features, and its ease of classification without performing many calculations. A While Loop was generated to identify the training set and test set scores of decision trees with increasing tree depths no greater than 30, these are listed below.

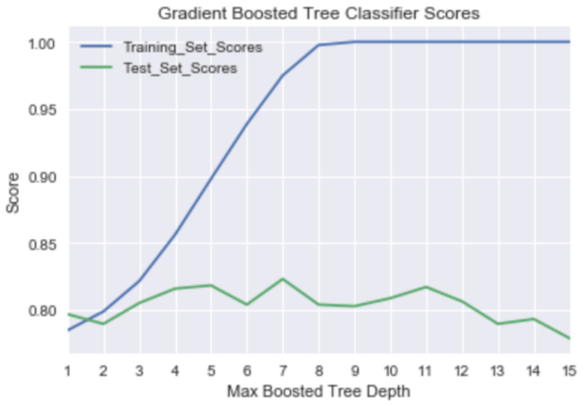


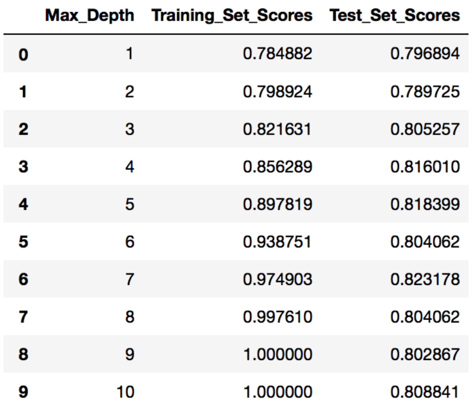


	Max Depth	Training Set Scores	Test Set Scores
0	1	0.740663	0.743130
1	2	0.773230	0.768220
2	3	0.773230	0.768220
3	4	0.781894	0.762246
4	5	0.793248	0.781362
5	6	0.803705	0.770609
6	7	0.822229	0.764636
7	8	0.837466	0.770609
8	9	0.860173	0.777778
9	10	0.884673	0.778973

The best score measured using the decision tree algorithm using test set data was 0.7814 with the decision tree having a maximum depth of 5 levels. Overfitting starts to occur beyond this point as the accuracy scores using test set data begin to worsen.

A Gradient-Boosted Tree algorithm was also selected for this experiment with a similar While Loop was created to capture the training and test set scores at various depths. This classifier has a best test data set score of 0.8232 with a boosted tree depth of 7. The test score accuracy of this model begins to decrease after a max depth of 7, with some intermittent spikes in performance thereafter, indicating some possible overfitting beyond this point.

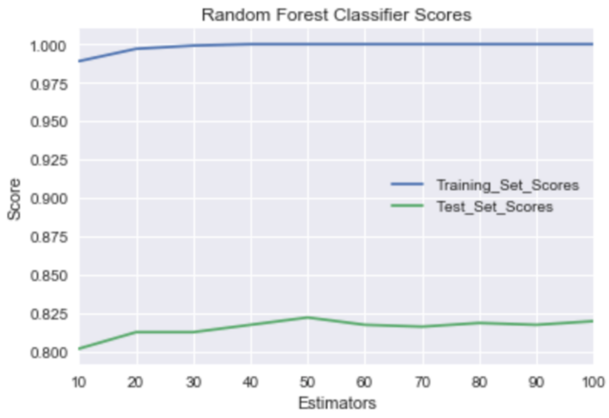


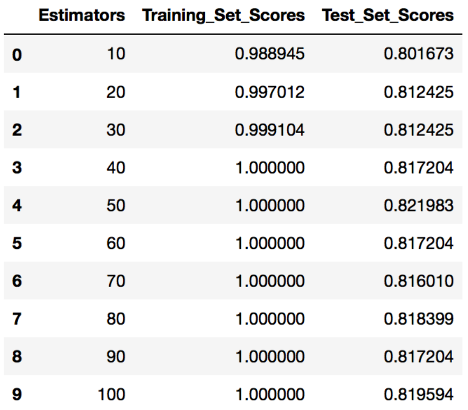


	Max_Depth	Training_Set_Scores	Test_Set_Scores
0	1	0.784882	0.796894
1	2	0.798924	0.789725
2	3	0.821631	0.805257
3	4	0.856289	0.816010
4	5	0.897819	0.818399
5	6	0.938751	0.804062
6	7	0.974903	0.823178
7	8	0.997610	0.804062
8	9	1.000000	0.802867
9	10	1.000000	0.808841

Classifier 2: Random Forest.

A basic Random Forest algorithm was also utilized to predict a "yes" response from a speed-date partner due to its reputation as a highly accurate classifier. To better record and visualize the number of estimators providing the highest scores on both the training and test data sets, a While Loop was also generated, scoring each forest and increasing the estimator by a 10 trees, culminating in a total of 100 trees per forest, these are listed below:





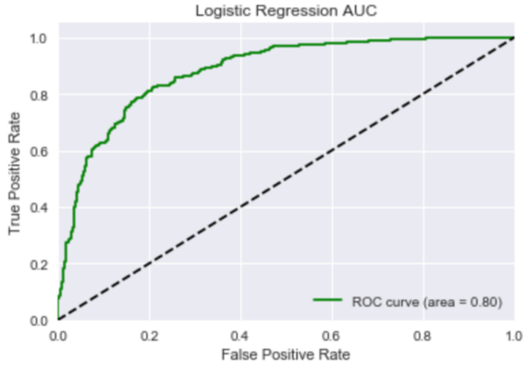
	Estimators	Training_Set_Scores	Test_Set_Scores
0	10	0.988945	0.801673
1	20	0.997012	0.812425
2	30	0.999104	0.812425
3	40	1.000000	0.817204
4	50	1.000000	0.821983
5	60	1.000000	0.817204
6	70	1.000000	0.816010
7	80	1.000000	0.818399
8	90	1.000000	0.817204
9	100	1.000000	0.819594

The score achieved was 0.8220 with each forest having a total of 50 trees each.

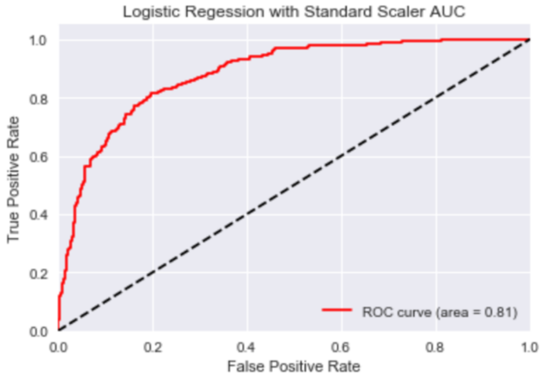
Classifier 3: Logistic Regression

A basic regression calculation was performed and fine-tuned with Scikit-Learn's Standard Scaler. The model was then refined to include only the ratings received and lastly the top 3 coefficients were identified and the model was refined once more. The top 3 coefficients were the attractiveness rating, the like rating, and met which denotes whether or not the speed-date partner as met the participant. The basic model alone had and AUC score of 0.8039, the standardized model did slight better with an AUC score of 0.8053. The first refined model performed the best out of the four regressions with an AUC score of 0.8065 and the top-3 coefficient regression model scored the lowest with an AUC score of 0.7670.

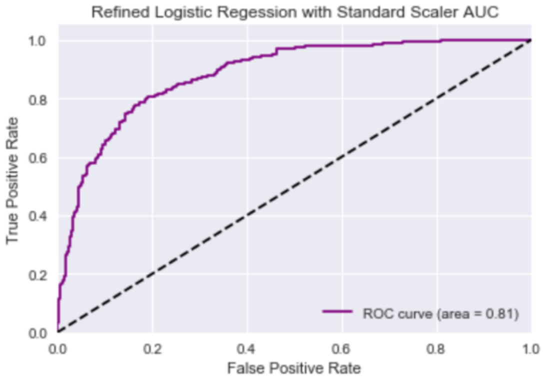
Logistic Regression Results:



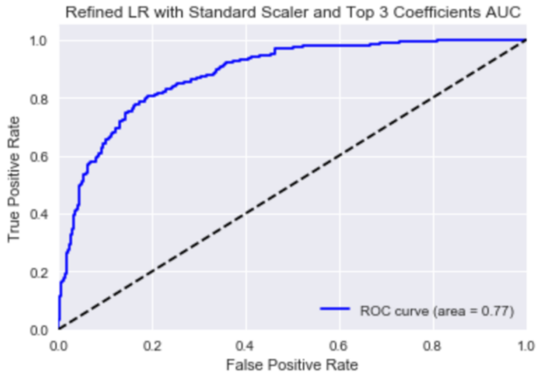
```
---Logistic Regression Model---

Accuracy of Logistic Regression Model: 0.8041

Logistic Regression AUC: 0.8039

             precision    recall  f1-score   support

        0.0       0.82      0.81      0.81       440
        1.0       0.79      0.80      0.80       397

avg / total       0.80      0.80      0.80       837
```



```
---Logistic Regression with StandardScaler Model---

Accuracy of LR Model with StandardScaler: 0.8053

Logistic Regression with StandardScaler AUC: 0.8053

             precision    recall  f1-score   support

        0.0       0.82      0.80      0.81       440
        1.0       0.79      0.81      0.80       397

avg / total       0.81      0.81      0.81       837
```



```
---Refined LR with StandardScaler Model---

Accuracy of LR Model with StandardScaler: 0.8065

Logistic Regression with StandardScaler AUC: 0.8064

             precision    recall  f1-score   support

        0.0       0.82      0.81      0.81       440
        1.0       0.79      0.81      0.80       397

avg / total       0.81      0.81      0.81       837
```



```
---Refined LR Model with Top 3 Coefficients ---

Accuracy of LR Model Top 3 Coefficients: 0.7670

Logistic Regression with StandardScaler AUC: 0.7675

             precision    recall  f1-score   support

        0.0       0.79      0.76      0.77       440
        1.0       0.74      0.78      0.76       397

avg / total       0.77      0.77      0.77       837
```

VI. Conclusion

The features utilized to predict whether or not a speed-date partner would like to see the participant again were the speed-partner's ratings of the participant along with basic demographic information about the participant, namely age and race, and other variables such as whether or not the participant and the speed-date partner are of the same race or if they had met previously. In predicting a "yes" response from a female speed-date partner, denoting specifically that she would like to see the male participant again, the Gradient Boosted Tree algorithm proved to be the most accurate when testing on unseen data, performing slightly better than the random forest algorithm. The summary of the scores can be found below.

Predictive Model Final Results	*Test Data Set Score*
Base Model	0.5257
Decision Tree	0.7814
Gradient Boosted Trees	0.8232
Random Forest	0.8220
Logistic Regression	0.8041 (AUC = 0.8039)
Logistic Regression with Standard Scaler	0.8053 (AUC = 0.8053)
Logistic Regression with Standard Scaler and less features	0.8065 (AUC = 0 .8064)
Logistic Regression with Standard Scaler and top 3 features	0.7670 (AUC = 0.7675)

Though a score of 0.8232, achieved by the Gradient-Boosted tree algorithm is descent, it would be advisable to further refine these algorithms until a minimum score of 0.90 or above can be achieved. To further refine this predictive algorithm, it is recommended that the collection of conversational data in the form of text transcript be collected so as to determine whether or not other variables that exist during the date, other than the physical ones presented, had any impact on a "yes" response. Lastly, it is recommended that non-invasive thermal scanners be incorporated to asses the genuineness of a "yes" response. This would allow for better and more accurate predictions and better marketing of speed-date services and it would allow for better feed-back from participants, improved customer satisfaction, and a better bottom line for the company.