On predicting speed dating decisions

Sook Yee Leung

May 13, 2015

Abstract

Speed dating is a condensed form of dating in which a group of individuals briefly meet each other, knowing that every met in the speed dating session is seeking a romantic relationship. Given the number of people each person can meet in a short period of time, the opportunity that participants have in completing survey forms between meetings, speed dating is a convenient context in which the precursors of romantic relationships can be studied. Using data from Fisman, Iyengar, Kamenica, and Simonson's (2006) speed dating study, this paper seeks to gain a better understanding of the factors that speed daters consider, which can lead speed daters to wish to meet the person she or he met again. Machine learning techniques and methodology are used to predict whether a particular speed dater decides to meet the person she or he met again. While this paper is a class project, the findings may be of interest to researchers studying the determinants of romantic relationships.

1. Introduction

Some may refer to it as love. Others may refer to it as romance. There are many metaphors for it: hearts and butterflies. Musicians have sung of it for centuries. It is also considered a precursor to marriage. The study of what leads to romantic relationships has interested researchers from many disciplines from fields such as psychology, sociology, and economics.

The search for a clear and decisive formula articulating what attracts one person to another remains elusive. Previous research suggest that multiple factors influence what attraction, including physical attractiveness, humor, similar backgrounds, and/or social status (Fisman et al., 2006; Fiore and Donath, 2009). Previous empirical research on successful romantic relationships has focused on various stages, such as marriage or dating (Wong, 2003; Fisman et al., 2006; Hitsch, Hortascu, and Ariely, 2005; Fiore and Donath, 2005). Researchers studying romantic relationship formation have conducted their studies within the online dating context or the speed dating context (Fiore and Donath, 2009; Fisman et al., 2006).

Such studies on relationship initiation is not only of interest in academia but is also of commercial interest since the aforementioned forms of dating are closely tied to online and speed dating industries, aiming to help users find their ideal romantic partner (Finkel and Sprecher, 2012; Tierney, 2013). This work will discuss romantic relationship initiation in the speed dating context.

Speed dating is a context in which a group of individuals briefly meet each other, aware that every met in the speed dating session is seeking a romantic relationship. The speed dating format is useful for social science research: data can be collected from between speed dates capturing participant opinions, such as her or his thoughts about the person she or he just met (Fisman et al., 2006; Luo and Zhang, 2009). Relevant demographic information may also be collected before the study. Follow-up information regarding whether the participant contacted their potential matches after the speed dating session. Moreover, lacking the limitations of time and event variability that can be encountered in traditional dates and the virtual environment in online dating, speed dating may mimic the actual dating context in a manner that better suits a research setting (Lou and Zhang, 2009).

Understanding the extent to which factors may contribute to initial attraction in the dating context may aid further understanding of romantic relationships. Factor types that have been studied include demographic, personality, contextual, perceptions of the ideal self, personality preferences in a partner, and similarity (Fisman et al., 2006; Luo and Zhang, 2009). This work utilizes data from Fisman and colleagues' speed dating study, to provide an understanding of how likely an individual is romantically interested in an individual she or he met (a.k.a. would like to see the other individual again).

There are ten sections in this paper. In addition to this first section, related work is discussed in the Section 2. The dataset, which was also used in prior research, is described in Section 3. The machine learning analysis procedure is described in Section 4. Sections 5 to 8 describes the analysis steps in detail. The final results are in section 9. A discussion in section 10 concludes this paper.

2. Related Work

The data that was used, as described in Section 3, was collected by researchers at Columbia University between 2002 and 2004 (Fisman et al., 2006, p.677). Study participants were undergraduate and graduate students, who provided email contact information to complete demographic information and follow-up surveys related to the speed dating sessions. Participants also rated and made decisions on whether they wanted to see their speed dating partner again after each four-minute speed date (Fisman et al., 2006, p.678). Fisman and colleagues (2006) were interested in how demographics, characteristic preferences in an ideal partner (e.g., ambition, attractiveness), perceived characteristics of a partner, and self-ratings on characteristics affected the decision of whether the participant wanted to see the other person again.

From an economics standpoint, Fisman and colleagues (2006) constructed several linear probability models to assess how the aforementioned characteristics affected the participant's decisions depending on the gender. Each linear probability model focused on a specific subset of factors. Models predicting decision were created for the following subsets: characteristic preferences in partner, demographics [e.g., SAT score, income], similarity [e.g., same race], availability of partners and characterstic preferences in partner considering ratings of own characteristics (Fisman et al., 2006). Specific features were highlighted as significant influences in each model and the goodness of fit for each model (R^2) was assessed for how well each subset explained participant decisions (Fisman et al., 2006; Frost , 2013). The goal of Fisman and colleagues' (2006) research was to better understand the extent of various factors contributed to "mate selection" (p.695).

The examination of specific factors that contribute to romantic partner selection in dating and/or marriage contexts is common across other academic works in this domain. The features studied tend to relate to theories about attraction and/or romantic partner selection. For instance, Houser, Horan, and Furler (2008) examined the "homophily or similarity principle" and a sense of closeness (i.e. "nonverbal immediacy") in the context of speed dating [p.754]. Variables of interest included survey items regarding background (e.g., "has a background similar to me") and observations about the other person (e.g., "uses hands while talking"), respectively [Houser et al., 2008. p.757]. Luo and Zhang (2009) assessed psychological personality characteristics (e.g., "neuroticism," "extroversion," "agreeableness," "avoidance" attachment style) from various empirically-validated psychology scales (e.g., "the Big Five Inventory", "Brennan and colleagues (1998) 36-item attachment measure") [p.943, 945]. Fiore and Donath (2009) have also assessed "homophily" in the online dating context (p.1371). Wong (2003) used economic modeling techniques to assess demographic factors that influence to the likelihood of marriage for white and black men (p.699). Wong (2003) used demographic variables from survey responses of the "Panel Study of Income Dynamics," regarding transitioning from being single to married (p.700).

This work seeks to utilize the data gathered from Fisman and colleague's (2006) study to predict the participant's decision per speed date. Unlike the aforementioned work, all of the factors was used to predict the participant's decision per speed date in this work. Much of the academic research on this subject in economics and psychology assess factor correlation and regression model influence on decisions, pertaining to partner selection. As a result, many previous academic works have not focused on directly predicting partner selection decision.

Nevertheless, prior work attempting to predict romantic partner selection is mainly within in the commercial realm. Companies, such as eHarmony and okCupid, use commercially licensed algorithms to aid customers in finding their ideal romantic partner (Bridle, 2014). Given the commercial nature of the algorithms, they have not been released for further study, but machine learning techniques are used (Bridle, 2014). For instance, the information collected can be used to classify users and recommend users of a particular classified group to users of another classified group (Bridle, 2014, para.17). These companies gather information about their customers and use such information to match their customers (Bridle, 2014). The features assessed may be similar to features assessed in academic studies, which include demographics, romantic partner preferences, personality, values, and so forth (Wong, 2003; Fisman et al., 2006; Luo aand Zhang, 2009). Still, academic researchers are skeptical about the matching accuracy of the algorithms used in dating sites, such as eHarmony (Tierney, 2013). The effectiveness of such algorithms is yet to be determined and is subject to further improvement.

3. Data Collection

Fisman and colleague's (2006) study used a speed dating script modeled after HurryDate's, "the largest speed dating company in New York" (p.676). Undergraduate and graduate students recruited through emails and fliers (Fisman et al., 2006, p.676). No participant partook in the study more than once. All forms of data were collected in survey form. There were 21 speed dating sessions in total with a varying number of participants (Fisman et al., 2006, p.676). Only heteronormative pairings were considered. Initial data collected online, which was part of the study registration process included the following: demographics, interests/hobbies, characteristic ratings from varying perspectives (i.e., the self, the self as perceived by others, the ideal romantic partner [of the opposite gender], the expected preferences of "fellow men/women," and the expected preferences of the opposite gender), study-specific variables, and speed date variables (*Figure 1*) [Gelman, 2006, p.5]. The only six characteristics assessed were: "attractive," "sincere," "intelligent," "fun," "ambitious," and "shared interests and hobbies" (Gelman, 2006, p.5). The variables of the initial survey were given the "_1" suffix (Gelman, 2006).

During the speed dating phase of the study (i.e., speed dating session), participants met with the study coordinators at a "popular bar/restaurant" near Columbia University (Fisman et al., 2006, p.677). The number of people in each session varied. Data collected during each speed date interaction included decisions about whether the participant wanted to see the other person she or he met again and characteristics ratings for each partner (p.677). During the middle of the event, participants were asked about characteristic ratings from the perspective of their ideal romantic partner and regarding themselves to assess any preference changes due to time or interactions (Fisman et al. 2006, p.677). The variables of the speed dating session were given the "_s" suffix (Gelman, 2006).

After the speed dating session, there were two follow-up surveys sent via email. The first was sent during the morning after the speed dating session. Completing the first follow-up allowed participants to be sent their "matches," which meant that the participant and another individual that she or he met during the speed dating session both indicated that they would like to see each other again on a survey (Fisman et al., 2006, p.678; Gelman, 2006). The follow-ups assessed participant satisfaction with the study session format and characteristic variables from varying perspectives (i.e., actual importance during the speed dating event, the ideal romantic partner, and the expected preferences of the opposite gender, and the self) [Gelman, 2006]. The variables of the first follow-up were given the "_2" suffix (Gelman, 2006).

The second follow-up was sent approximately three to four weeks after participants were sent their list of matches. There were questions related to outcomes, such as whether the participant contacted or dated one or more of her or his matches. Characteristic ratings from various perspectives (i.e., the self, the self as perceived by others, the ideal romantic partner [of the opposite gender], the expected preferences of "fellow men/women," the expected preferences of the opposite gender, and actual importance during the speed dating event) were asked (Gelman, 2006). The variables of the second follow-up were given the "_3" suffix (Gelman, 2006).

Each of the dataset instances is a four-minute speed date. Participants engaged in multiple speed dates in every session. While two instances can be of the same two individuals (identified by unique id numbers [iid] and [pid]), the two instances would be different in that one individual would be the target individual and the other individual would be the other person [Gelman, 2006, p.1]. Furthermore, if there were ten speed dates for each individual, there would be ten instances of a target individual paired with a different individual, which included accompanying initial survey data, first follow-up data, and second follow-up data. For the target participant identified by the unique id number, only the characteristic ratings and some background information about the other person would be vary in each instance (i.e., four minute speed date).

4. Procedure Outline

The objective of this paper is to present a trained classifier to predict decision outcomes of each speed date, considering features of interest shown in Figure 1. This section previews the aforementioned process. Section 5 presents the preparation of the feature space, which includes selecting and adding features relevant to this work. The *development, cross-validation*, and *final test* set creation process is also detailed. Section 6 discusses the initial analysis performed on the *development* dataset using models trained on the *cross-validation* data set. Section 7 details baseline and error analyses, focusing on the SMO algorithm with default settings that was selected based on the exploratory analysis. Section 8 presents parametric optimization of the SMO model after it is trained on the *cross-validation* set, comparing the model's baseline performance from Section 7 with the optimized performance on the *development* data set. Section 9 evaluates the performance of the optimized SMO model that was trained on the *cross-validation* data set on the *final test* set.

5. Data Preparation

The data set from Fisman and colleagues (2006) study was divided into three sets: the *development*, *cross-validation*, and *final test* set. The *cross-validation* set was used to train all models considered in this work. The *development* set was used to explore the data and as a test set to assess the trained model's performance during the data exploration, baseline analysis, and optimization phase. The *final test* set was used not used until the final test phase after the baseline model was optimized.

While there were 8379 instances of speed dates in total, only 6817 instances were used. A total of 1562 instances was removed because the original experimenters noted that participants of those phases used a different characteristic rating scale (e.g., "Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important") as opposed to the directions given for the following rating scale: "You have 100 points to distribute among the following attributes --give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100" (Gelman, 2006, p.5). Given the differing rating methods and the majority of the data utilizing the latter method, the minority of the data using the former rating scale were removed.

Attributes were removed and added to prepare the data for *Lightside* and *Weka*. Study specific variables that were not of domain interest (i.e., romantic relationship initiation) were removed, such as partner id and specific id numbers based on gender were removed as these variables added unnecessary information (Gelman, 2006). A textual “from” variable allowing the participant to describe where she or he was from was replaced with the “international” attribute identifying whether the target participant was from the United States (international = 0) or not (international = 1) [Gelman, 2006, p.2]. The field of study text variable was removed in favor of a nominal field code (e.g., 1 = Law, 2 = Math) variable that was already present (Gelman, 2006, p.2). Missing information for the field code variable was added manually based on the field of study text variable. The textual “career” variable was removed in favor of the “career code” variable (Gelman, 2006, p.4). Missing information for the career code variable was added manually based on the textual career variable. The recoding and feature creation was performed in R and Excel. *Table 1* presents a summary of the full set of attributes.

In total, 189 attributes were considered in this study. The attributes considered in the feature space included study context information, such as “condition” with 1 = “limited choice” or fewer speed dates and 2 = “extensive choice” or many speed dates, and the unique subject number associated [iid] with each speed date in order to allow *Lightside* to train models and assign instances to folds by unique ids to account for target participant related patterns (Gelman, 2006, p.1).

All other information from the initial survey, speed dating session, and follow-up surveys were kept in the dataset. While the second follow-up survey information variables had more missing data than other variables, these variables measured potentially informative study outcomes that can be aid decision prediction. For instance the numdat_3 variable, measuring how many dates the target participant has been in since the speed dating session (not necessarily with her and the matches variable, indicating speed date matches [both participants decided yes to see each other again]) may be useful [Gelman 2006, p.7]. For instance, if the aforementioned variable was particularly informative (e.g., higher weight given in a regression model or greater information value given in a decision tree model), then the variable may indicate that the target participant was keen (or not) on seeing one or more of her or his speed dates again or was keen on dating within three or four weeks after the study.

Of the 6817 instances, approximately 60% were assigned to the *cross-validation* set, 20% were assigned to the *development* set, and the remaining 20% were assigned the *final test* set. Given that groups of instances shared the same target participant’s unique id [iid], the data were randomly assigned to different machine learning sets based on the iid.

The random assignment of speed dates (instances) by target participant id was determined using R’s sampling and subsetting functions. For instance, data was allocated to the final test set by randomly selecting 20% of the 450 (excluding 102 of the participants using the different rating method) unique participant ids. Then, the instances with the common iids were extracted from the entire dataset to the final test set. Percentages of instance allocation may differ based on the varying number of speed dates a participant may have, considering her or his speed date session. As a result, 1365 (20.0%) instances were in the *final test* set, 1385 (20.3%) instances were in the *development* set, and 4067 (59.7%) instances were in the *cross-validation* set.

Example variable	Variable type	Variable explanation
Study-specific variables		
*Codes used to identify participants within entire study, sessions, etc		
iid	Nominal	Unique id number across the entire study
idg	Nominal	Subject number within gender, group(id gender)
id	Nominal	Id number within a session
condtn	Nominal	1 = limited choice; 0 = extensive choice
pid	Nominal	partner's iid number
gender	Nominal	Female = 0 and Male = 1
age	Numeric	Age of target participant
age_o	Numeric	Age of other person
field_code	Nominal	Categorization of field of study
Interest/hobbies variables		
*Activity variables rated on a scale of 1-10.		
museums	Numeric	Variable of interest rating
sports	Numeric	Variable of interest rating
Characteristic variables		
*The same characteristics are rated from different perspectives and study phases. *Suffixes provide indication of phase and perspective. *Includes the following: attractive, sincere, intelligent, fun, ambitious, and shared interests.		
attr1_1	Numeric	Variable of attractivness rating in terms of romantic partner preference (1) in the first survey(1)
attr3_1	Numeric	Variable of attractivness rating in terms of self-rating (3) in the first survey(1)
attr3_2	Numeric	Variable of attractivness rating in terms of self-rating (3) in the follow-up survey(2)
Speed date variables		
*Variables associated with the other person and other speed date related outcomes.		
dec	Nominal	Variable indicating the particpant's decision to meet again (1 = yes; 0 = no)
attr	Numeric	Variable of attractivness rating in terms of the other person
like	Numeric	Variable of liking rating in terms of the other person
dat_3	Nominal	Last follow-up variable indicating whether the participant dated on of her/his matches (1 = yes; 0 = no)
match	Nominal	Variable indicating whether the participant and the other person both chose to meet again (1 = yes; 0 = no)

Figure 1. Brief summary of attributes. The complete list is available vis-à-vis Gelman (2006).

6. Data Exploration

Data exploration was conducted on the *development* set. Ascertaining the feature space was related to this stage. Each variable's contribution towards predicting whether the speed date resulted in the target participant wanting to see the other person again was considered against domain knowledge discussed in Section 2. Variables that were study specific (e.g., pid, idg) were removed. Further descriptive statistics on each variable was obtained from Minitab and Weka.

Algorithm Type (Weka algorithm name)	**% Correct (All features)**	**Kappa (All features)**	**% Correct (Missing removed)**	**Kappa (Missing removed)**
Naïve Bayes	73.48%	0.4679	72.98%	0.4605
Regression (SMO)	82.88%	0.6487	80.78%	0.6040
Rule based classifier (JRip)	84.47%	0.6781	76.62%	0.5808
Bayesian Network (BayesNet)	77.46%	0.5467	77.17%	0.5407
Decision Tree (J48)	83.74%	0.6655	78.31%	0.5575

Figure 2. Model performance results

Five types of algorithms were used to gain initial understanding of how speed date decisions were predicted and of the type of performance to be expected on the initial feature space of 189 variables. The models from the algorithms shown in *Figure 2* (and discussed throughout this work) were trained on the *cross-validation* set, using ten-fold cross validation. The results shown in *Figure 2* on the models that considered all of the features of the original feature space were obtained from the trained model's performance on the *development set*. Default settings for each algorithm were used.

The model results using all selected features, provided a baseline comparison to models created from alterations of the feature space. Upon closer examination of the feature space through Minitab and Weka, 64 variables were found to have nearly fifty percent or more missing data. Descriptive statistics and Weka was used to examine the variables with high amounts of missing data. Missing data can harm the generalizability of the trained model to unseen data, which can include making overly optimistic predictions about the models' performance on unseen data. The variables with high rates of missing data included all variables from the final follow-up phase that was three to four weeks after the speed dating session and all variables pertaining to characteristic ratings halfway through the speed dating session. It is possible that participants lost interest in the study after nearly a month and may not have had enough time to input her or his ratings, respectively.

Missing data may harm model generalizability if the non-missing data was not missing at random. For instance, of the non-missing data from the final follow-up variables, participants, who have found an ideal romantic partner via the study, may be eager to report her or his outcome and characteristic rating perceptions in the final follow-up. The responses from the particular group of participants may cause the model to overfit the training data and overpredict a particular decision (e.g., more yeses).

Thus, the feature space was adjusted to exclude features with nearly fifty percent (45% or above) or more variables missing, according to the *development* set. The adjusted feature space utilized 125 features. A total of 64 features were removed. The results shown in *Figure 2* on the models under the "missing removed" columns demonstrates the trained model's performance on the *development set*. Default settings for each algorithm were used. As was demonstrated throughout the course, it is clear that the Naïve Bayes and Bayesian Network algorithms were minimally impacted by missing data.

However, the rule-based and decision tree algorithms were greatly impacted by the feature space adjustment: both algorithms performed notably worse. Given that decision trees generated classifications starting with the most highly informative variable (based on information gain) and then branched based the information gain value of subsequent variables in a greedy manner, more influence might have been given to responses to the variables with a great missing data percentage. Likewise, given that the rule-based algorithm uses a "separate and conquer" method to determine rules that classify some instances correctly while ignoring others, the variables with a great missing data percentage may have been incorporated in many of the rules that predict decisions ("Rule Learning", 1998).

Only the regression algorithm created models that maintained the highest and a fairly consistent performance regardless of the feature space adjustment. Unlike the rule-based and decision tree algorithms, regression algorithms do not divide the feature space in order to predict the classifier. Instead, a decision boundary is applied to all of the instances, considering the weight of all of the features. Furthermore, the SMO algorithm used normalizes all variable values and fills in missing values in variables with a default value ("SMO", n.d.).

Considering the consistency of the SMO algorithm's performance in both feature space versions, it is likely that the SMO algorithm is assigning weights in a manner that does not emphasize the features with the missing data, aiding in model generalizability to unseen data that may or may not have data for the variables with the high missing data percentage rate. The SMO model utilizing all of the features performed did not perform significantly better than the SMO model utilizing less features (Fisher's exact x^2 two-tailed test, p-value = 0.1675). Still, in order to leverage the information from having more variables to make predictions, the SMO model utilizing all features was chosen for further analysis.

7. Baseline Performance

Baseline analysis was conducted with default Weka settings. The initial feature space consisting of 189 variables was used. Models in this section were trained using 10-fold cross-validation on the *cross-validation* set. The performance results were based on the trained model's performance on the *development* set.

In order to better understand the SMO model's performance in comparison to simpler models. A ZeroR model, which makes predictions based on the majority class value, was evaluated. Based on the cross-validation set, the trained model predicted that decision = no. Given that the ZeroR algorithm yields a model that makes predictions in manner that is similar to chance by predicting the majority class the kappa value was 0.0. The model's accuracy was 57.66%, correctly classifying 799 instances in the *development* set.

A OneR model, which makes predictions based on one rule, was also evaluated. Based on the cross validation set, predictions were made based on consideration of whether the like variable had a value above (decision = yes) or below 6.25 (decision = no), which was a rating from a scale of one to ten of how much the target participant liked the other person (Gelman, 2006). The trained OneR model had an accuracy of 77.57%, correctly classifying 1074 instances in the development set, and a kappa of 0.4836.

It is clear that the SMO model performed better than the simpler models, demonstrating that the baseline model selected from the data exploration phase performed better than chance and better than just considering one variable. The SMO model had an accuracy of 82.88%, correctly classifying 1149 instances, and a kappa of 0.6487. The performance difference between the SMO and OneR model is statistically significant (Fisher's exact x^2 two-tailed test, p-value = 0.001).

An error analysis was conducted on the SMO model classification. According to the error matrix, the features that had the highest horizontal and vertical absolute differences (i.e., at least ten times more than subsequent variables) within the false positives and false negatives predictions were the zipcode, tuition, income, and mean sat score variables. Upon further analysis of the descriptive statistics results from my development dataset, my development dataset, and study documents, these variables were only relevant to the participants from the United States, which according was 32% of the participants according to the international variable based on the "from" variable of the original dataset (Gelman, 2006, p.3). These variables were eliminated from the feature space.

Considering the similarity principle from previous research, two new variables that were correlations generated in Excel of ratings were generated (Luo and Zhang, 2009, p.934). The principle is based previous work as discussed in Section 2. Variable creation was also based on the SMO model variable weights: the match variable [a binary variable that indicated whether the speed dating pair both decided yes (match = 1) or otherwise (match = 0)], an indicator of decision similarity was the most influential in the model with a coefficient of 7.05. A correlation of self-ratings on all six characteristics and other person's preferences on all six characteristics: a correlation of the participant's personal characteristic ratings and other person's characteristic preference ratings (selfper_o_corr). Though not as straightforward as the other variables, this variable may indicate whether the target participant's decision was influenced by a romantic characteristic preference match with the other person aligning with self-perceptions. The other variable assessed a correlation of self-ratings based on how "others perceive [the target participant]" and the other person's preferences on all six characteristics: outper_o_corr (Gelman, 2006, p.7). Though not as straightforward as the other variables, this variable may indicate whether the target participant's decision was influenced by a romantic characteristic preference match with the other person based on more superficial characteristics. These new variables were added to the feature space.

Considering the adjusted feature space, new models were trained on the *cross-validation* set. The results for the ZeroR and OneR models were the same. The SMO model yielded an accuracy of 88.17%, correctly classifying 1221 instances in the development set, and a kappa of 0.7343. The SMO model utilizing the adjusted feature space performed significantly better than the SMO model utilizing the entire feature space (Fisher's exact x^2 two-tailed test, p-value = 0.0001). The improved model performance was not surprising considering that features, which confused the prediction model were eliminated and more information was added to the feature space.

8. Optimization

The SMO model based on the adjusted feature space from Section 7 was optimized in order to use the best parameter value for better performance on my dataset. The complexity parameter, C, was adjusted. Given the number of features in my feature space and utilization of features with a large percentage of missing values, it is worth considering a lower C values to avoid overfitting my model.

The *CVParameterSelection* technique from Weka was utilized. An initial evaluation of the above assumption was performed. Five folds were selected to evaluate the SMO algorithm, while specifying that the parameter selection adjust C values between 1 and 3, in 3 steps. A C value of 3 was selected. After training an SMO model with a C=3 setting, the resultant model performance on unseen data an accuracy of 89.67%, correctly classifying 1242 instances in the development set, and a kappa of 0.7520.

Complexity parameter values considering C = 1 and below were then evaluated. Five folds were selected to evaluate the SMO algorithm, while specifying that the parameter selection adjust C values between 0.1 and 1, in 9 steps. Development set Cross validation set from 0.1 to 1 in 9 steps. The selected C value was 0.325. Performance on unseen data was an accuracy of 91.17%, correctly classifying 1262 instances in the development set, and a kappa of 0.7746.

While the more complex SMO model may better fit the training dataset, given the number of features considered, the less complex model may perform better on unseen data. Further statistical analysis indicates that the difference between the baseline SMO model with a C= 1 and the optimized SMO model with C=0.325 is statistically significant (Fisher's exact x^2 two-tailed test, p-value = 0.0104).

9. Final Result

A final test of the optimized SMO model with the adjusted feature space from Section 7 was performed on the *final test* set, which was not used until this step. An SMO model with C=0.325 was trained using the cross-validation set. The resulting performance on the final test set was a 90.87% accuracy, correctly classifying 1365 instances, and a 0.7857 kappa. The OneR model trained on the cross validation dataset, correctly classified 1028 instances, yielding and accuracy rate of 75.31% and a kappa of 0.4976. The performance difference between the two models was statistically significant (Fisher's exact x^2 two-tailed test, p-value = 0.0001).

10. Discussion

Various machine learning methodology and algorithms were explored in predicting speed dating decisions based on demographic, characteristic, interest, and contextual factors. The process described in this work involved data preparation, feature engineering and selection, baseline performance analysis, error analysis, and optimization. The optimized SMO model successfully predicted over 90% of the final test set data.

Unlike prior academic work focusing on specific variables associated with specific theories mentioned in Section 2, this work considered a large variety of features made available via Fisman and colleagues' (2006) study data set. Similar to commercial classification algorithms used to determine potential online dating matches, this research has used machine learning methods to determine speed date decisions regarding whether the target speed dater wanted to see the other person again.

This study may contribute to further exploration of how the dissonance between superficially projected personality characteristics and personal perception of personality characteristics affect speed date decisions and/or other relationship outcomes. While the final optimized model coefficients for the variables representing the correlation between the other person's characteristic preference and target participant's self-rating of characteristics (selfper_o_corr) and the correlation between the other person's characteristic preference and target participant's characteristic self-ratings based on assumptions of others' perceptions (outper_o_corr) were rather small (i.e., 0.175 and -0.154, respectively), the polarity of the coefficients indicate that the dissonance between an outward and personal perception of characteristics affects speed dating outcomes. In other words, it is likely that individuals with a higher dissonance between the outwardly projected and the inwardly acknowledged self affects dating interaction decisions. Future research may explore other features and interactions to provide more insight on this potential interaction.

Vis-à-vis Section 8, the better performance on unseen data yielded from setting a smaller complexity value (0.325) for the SMO algorithm as opposed to a larger complexity value (3.0), indicates that the optimal value is small. Since a smaller complexity setting was preferred, it is clear that considering the large feature space, which includes demographic variables, characteristic variables across time and from different perspectives, and contextual variables, may lead to variable weight distributions patterns that can mislead the SMO model.

Unsurprisingly, the OneR model performed rather well, correctly classifying the decisions of over two-thirds of instances in the *development* set and the *final test* set, considering only whether the target participant rated liking the other person with a score at or above 6.5. In the context of the work discussed in Section 2, the concept of liking is complex and can be explained by theories such as similarity, characteristics such as attractiveness, and so forth (Luo and Zhang, 2009; Fiore and Donath, 2009; Houser et al., 2008). However, it is interesting that liking ratings did not nearly completely predict speed date decisions. This indicates that other factors instead of the subjective feeling of liking influences future dating decisions. Previous research supports this. For instance, Fisman and colleagues (2006) found that in sessions with less participants, female participants were more likely to choose to see their partner again (i.e., decision = yes).

A number of limitations are to be considered in this work. Technical limitations prevented the use of models that may have better guarded against overfitting, such as Random Forest models. From a domain-based perspective, due to the wide variety of factors that may influence speed date decisions, more decisive factors may not have been considered in the original study's data collection phase.

Foremost, this research is a class project that seeks to evaluate machine learning methods in the context of an actual data set. Future work may integrate more recent theories on speed dating decisions since Fisman and colleagues' (2006) almost a decade ago. An example includes research from Luo and Zhang (2009), which indicates that assessing factors related to reciprocated liking may useful predicting future dating decisions (p. 954). Considering additional theories to drive the data collection and feature engineering phases of machine learning allows a more informative set of factors to be considered.

References

Bridle, J. (2014). The algorithm method: How internet dating became everyone's route to a perfect love match. Retrieved May 13, 2015, from http://www.theguardian.com/lifeandstyle/2014/feb/09/match-eharmony-algorithm-internet-dating

Finkel, E., & Sprecher, S. (2012). The Scientific Flaws of Online Dating Sites. Retrieved May 13, 2015, from http://www.scientificamerican.com/article/scientific-flaws-online-dating-sites/

Fiore, A. T., & Donath, J. S. (2005, April). Homophily in online dating: when do you like someone like yourself?. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems* (pp. 1371-1374). ACM.

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*, 673-697.

Frost, J. (2013). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Retrieved May 13, 2015, from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Gelman, A. (2006). Speed Dating Key. Retrieved May 13, 2015, from http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/Speed Dating Data Key.doc

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2005, January). What makes you click: An empirical analysis of online dating. In *2005 Meeting Papers* (Vol. 207). Society for Economic Dynamics.

Houser, M. L., Horan, S. M., & Furler, L. A. (2008). Dating in the fast lane: How communication predicts speed-dating success. *Journal of Social and Personal Relationships*, *25*(5), 749-768.

Luo, S., & Zhang, G. (2009). What Leads to Romantic Attraction: Similarity, Reciprocity, Security, or Beauty? Evidence From a Speed-Dating Study.*Journal of personality*, *77*(4), 933-964.

Tierney, J. (2013, February 11). A Match Made in the Code. Retrieved May 13, 2015, from http://www.nytimes.com/2013/02/12/science/skepticism-as-eharmony-defends-its-matchmaking-algorithm.html?pagewanted=all&_r=0

Wong, L. Y. (2003). Structural estimation of marriage models. *Journal of Labor*

Economics, *21*(3), 699-727.

Rule Learning versus Decision Tree Learning. (1998). Retrieved May 13, 2015, from http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume8/fuernkranz98a-html/node6.html

SMO. (n.d.). Retrieved May 13, 2015, from http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html