Department of Mathematics and Statistics

STA 401: Introduction to Data Mining

Spring 2024

Match Made in Algorithms: Predicting Love Compatibility

May 19, 2024

Instructor: Dr. Ayman Alzaatreh

Aysha B. Hashim	91558
Shahd Mahmoud	88887
Johnny Kortbawi	88581
Khalifa Almatrooshi	90847

Table of Contents

List of Tables

List of Figures

Abstract

The application of data science to predict love compatibility may have the potential to offer promising advancements. This study explores the efficacy of various statistical models and machine learning techniques in predicting match outcomes from a speed dating experiment, which gathers comprehensive behavioral and preference data from participants. The primary aim is to develop a predictive model that adequately forecasts the likelihood of compatibility between individuals to promote longer-lasting relationships. The dataset used is initially comprised of over 8,000 observations and is preprocessed to ensure the models appropriately fit the data. This preprocessing includes steps like imputing and removing missing values and ensuring the correct data format for each variable. Backward variable selection is applied to maintain the model fit while further minimizing the number of variables that add negligible value. Key machine learning models employed include Decision Trees, Random Forest, Logistic Regression, and Gradient Boosting, each tested for its prediction accuracy using a split of training and testing data. The models were evaluated based on their ability to predict mutual matches, which is crucial for recommending potential partners. Additionally, this study integrates a recommendation system that assesses the compatibility based on mutual preferences and characteristics to ensure that the values and attributes align, aiming to form quality matches. The blend of predictive modeling and personalized recommendation aims to refine the approach of digital matchmaking and better understand complex human preferences and behaviors.

Keywords: *Predictive Modeling, Machine Learning, Compatibility Prediction, Speed Dating Analysis, Recommendation Systems*

Match Made in Algorithms: Predicting Love Compatibility

In the realm of modern romance, the dynamics of attraction and mate selection have increasingly become integrated with the algorithms that fuel online dating sites. These algorithms promise to deliver a *match made in algorithms*, employing data to connect individuals based on compatibility and shared interests. Speed dating provides a platform for testing these algorithms, where the information of individuals interested in participating can be collected to determine their perfect match. This approach is especially crucial given the increasing difficulty in finding appropriate partners, as evidenced by the growing rates of divorce and relationship breakdowns over time [1]. Thus, individuals anticipate having long-term relationships with the one they have chosen after making a difficult decision [5]. Therefore, in efforts of reducing divorce rates, and contributing to long lasting, happy and healthy relationships, research efforts have invested time in understanding the factors that impact partners' compatibility.

Background

Speed dating provides a unique perspective on the complexity of romantic attraction. Individuals seeking for possible romantic partners attend a speed dating event, where they go on a series of short dates with other attendees. These dates last only a few minutes for each event. Following the event, attendees may say *yes* or *no* on whether they want to see each of their dates again. The option to get in touch for a future, probably more conventional date is granted to two attendees who say *yes* to one another [3]. This speed dating procedure enables researchers to examine romantic dynamics dyadically, considering the potential for meaningful relationships, and maintaining strong external validity [2]. Moreover, compared to conventional dating settings, speed dating has the potential of removing significant barriers to initiating a conversation with the desirable partner. For instance, one may presume with certainty that speed

daters are eager to meet prospective love partners and are romantically readily available and one is almost guaranteed to have a brief one-on-one meeting with each of the speed daters who are interested in the preferred sex. Thus, speed dating allows researchers to study romantic dynamics closely and facilitate connections by eliminating conversation barriers, offering insight into the complexities of romantic attraction in modern day romance.

Individuals are driven to say *yes* at the end of the speed dating event due to positive assortative mating. Positive assortative mating is explained in terms of similar preferences and interests of individuals who are choosing one another [6]. Matches are influenced by stronger preferences for socioeconomic similarity on factors such as age, education, and occupation [7]. Individuals may naturally gravitate towards certain preferences, such as desiring partners who share similar attributes, or they may frequently engage with individuals of comparable status or shared interests. Therefore, our study aims to delve deeper into this complex network of romantic compatibility by analyzing data from a speed dating experiment [4]. Through examining the different attributes collected during these encounters, including personality traits, interests, age, and income, we seek to develop a comprehensive understanding of the underlying factors that influence relationship compatibility. This project aims to apply various machine learning techniques discussed in class to generate a predictive model capable of forecasting whether individuals are compatible or not based on their respective attributes. Additionally, the paper will explore the use of recommendation systems in this context. Through quantifying romance in this way, machine learning may be a more successful matchmaker than humans, this project aims to explore its potential in contributing to successful pairings in the future.

Data

Variable Description

Variable Name	Description
gender	Gender of self
age	Age of self
age_o	Age of partner
d_age	Difference in age
race	Race of self
race_o	Race of partner
samerace	Whether the two persons have the same race or not.
importance_same_race	How important is it that partner is of same race?
importance_same_religion	How important is it that partner has same religion?
field	Field of study
pref_o_attractive	How important does partner rate attractiveness
pref_o_sinsere	How important does partner rate sincerity
pref_o_intelligence	How important does partner rate intelligence
pref_o_funny	How important does partner rate being funny
pref_o_ambitious	How important does partner rate ambition
pref_o_shared_interests	How important does partner rate having shared interests
attractive_o	Rating by partner (about me) at night of event on attractiveness
sincere_o	Rating by partner (about me) at night of event on sincerity

intelligence_o	Rating by partner (about me) at night of event on intelligence
funny_o	Rating by partner (about me) at night of event on being funny
ambitous_o	Rating by partner (about me) at night of event on being ambitious
shared_interests_o	Rating by partner (about me) at night of event on shared interest
attractive_important	What do you look for in a partner - attractiveness
sincere_important	What do you look for in a partner - sincerity
intellicence_important	What do you look for in a partner - intelligence
funny_important	What do you look for in a partner - being funny
ambtition_important	What do you look for in a partner - ambition
shared_interests_important	What do you look for in a partner - shared interests
attractive	Rate yourself - attractiveness
sincere	Rate yourself - sincerity
intelligence	Rate yourself - intelligence
funny	Rate yourself - being funny
ambition	Rate yourself - ambition
attractive_partner	Rate your partner - attractiveness
sincere_partner	Rate your partner - sincerity
intelligence_partner	Rate your partner - intelligence
funny_partner	Rate your partner - being funny
ambition_partner	Rate your partner - ambition

shared_interests_partner	Rate your partner - shared interests
sports	Your own interests [1-10]
tvsports	Your own interests [1-10]
exercise	Your own interests [1-10]
dining	Your own interests [1-10]
museums	Your own interests [1-10]
art	Your own interests [1-10]
hiking	Your own interests [1-10]
gaming	Your own interests [1-10]
clubbing	Your own interests [1-10]
reading	Your own interests [1-10]
tv	Your own interests [1-10]
theater	Your own interests [1-10]
movies	Your own interests [1-10]
concerts	Your own interests [1-10]
music	Your own interests [1-10]
shopping	Your own interests [1-10]
yoga	Your own interests [1-10]
interests_correlate	Correlation between participant's and partner's ratings of interests.
expected_happy_with_sd_people	How happy do you expect to be with the people you meet during the speed-dating event?

expected_num_interested_in_me	Out of the 20 people you will meet, how many do you expect will be interested in dating you?
expected_num_matches	How many matches do you expect to get?
like	Did you like your partner?
guess_prob_liked	How likely do you think it is that your partner likes you?
met	Have you met your partner before?
decision	Decision at night of event.
decision_o	Decision of partner at night of event.
match	Match (yes/no)

Data Preprocessing

As part of preprocessing the dataset used in this analysis, a series of cleaning steps are performed before applying the various machine learning models. Steps including deleting irrelevant features to our analysis, imputing or omitting missing values, ensuring that the variables are the correct data type, and balancing the dataset through oversampling and undersampling were explored in efforts of maximizing the accuracy of the model's match prediction.

Categorization of the Field Column

In our dataset, the 'field' column that indicates the field of study contains 260 unique values. To streamline the analysis and enhance the interpretability of our results, we categorized these various fields into broader academic categories under a new column named 'field_category'. This transformation reduces the complexity of the data and allows for more generalized insights about group preferences. The categories we decided on are: Arts, Business,

Engineering, Health, Humanities, Sciences, Social Sciences, and Other. Other includes double majors and working fields. We developed a mapping function (Appendix) to assign each specific field to one of these categories automatically.

Omitting Selected Variables

Beginning with 8,378 observations and 124 features, the dataset initially contained 18,372 missing values. The following variables were deemed non-value adding to our analysis for varying reasons as seen in Table 1 and were dropped. After their omittance, the resultant dataset was reduced to 62 features.

Table 1: Dropped Variables

Variable	**Dropped Reason**
d_variables	*Binned ranges used for analysis. They group continuous data into discrete categories, making it easier to analyze patterns and trends. Will not be used for prediction*
decision & decision_o	*Leaking variables that the match response is based on*
expected_num_interested_in_me	*Accounts for over one third of the total missing values in the dataset*
Wave, has_null, met	*Irrelevant to this prediction*

Imputation of Missing values

After ensuring that all variables were set to the correct types—factor and numeric—median imputation was performed to address the missing values. Some variables, such as *expected_num_matches* and *shared_interests*, had over 1,000 missing values. Therefore, imputation was applied to variables with more than 100 missing values. For variables with fewer

than 100 missing values, the observations with missing data were omitted from the dataset. This minimal loss was deemed insignificant and did not negatively impact the models' predictive power.

Dataset Splitting into Training and Testing

Using random sampling, 70% of the data was reserved for training, and the remaining 30% was allocated to testing and validation of model performances. Hence, the resultant training dataset consisted of 5,865 observations while the testing dataset contained 2,513 observations.

Balancing the Dataset

Seeing as the original dataset is unbalanced, with approximately 83.5% no match and 16.5% match, the team looked into the options of oversampling the minority or undersampling the majority, made available through the *"ROSE"* library. Oversampling the minority was favored in this case due to the fact that undersampling yielded a severely smaller dataset, one that did not contain enough observations for the success of the prediction models. Hence, the minority was oversampled instead, creating a balanced 50% match, 50% no match dataset. Upon doing so, however, it was seen that the models performed better when the original, unbalanced dataset was used when compared to using the oversampled and balanced data. For example, logistic regression had a 77% accuracy with the balanced data, but an improved 83% using the unbalanced data. Other models also performed worse when using the balanced data. Hence, with the objective of enhancing the models' accuracy, it was decided to move forward with the unbalanced, original dataset.

EDA and Preliminary Insights

We have chosen to conduct Exploratory Data Analysis (EDA) solely on the unique rows. Excel was used to extract the columns relevant to an individual, leaving us with only duplicate

rows with iterating id. The remove duplicates function in Data was used, with id unchecked, to remove duplicate rows. Beginning with 8,378 observations and 124 features, we are left with 551 observations and 38 features, with 751 missing values. This ensures that our analysis accurately reflects the characteristics and distributions within the dataset.

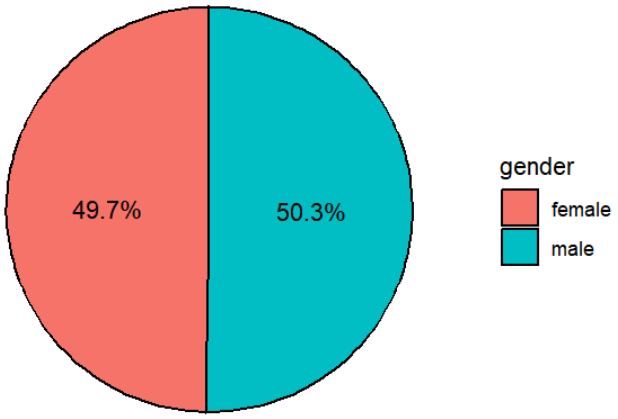


Figure 1: Distribution of Gender

Figure 1 shows a practically balanced demographic, with males constituting 50.3% and females 49.7% of the population. Therefore any insights from the report are not biased by a gender imbalance.

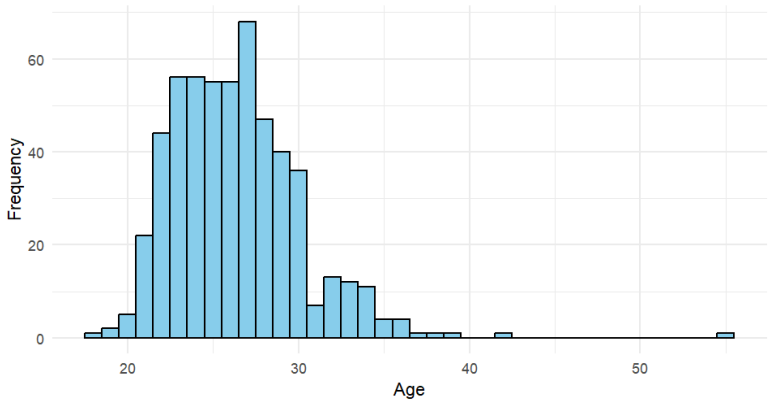


Figure 2: Distribution of Age

Figure 2 shows that a majority of participants are concentrated around the ages of 20 to 30 years, with a peak frequency around 27 years. This means that the dataset is primarily young adults, suggesting settings such as universities or business districts. The distribution is right-skewed, with fewer participants from 31 years and beyond. This could impact the generalizability of the models to older populations.

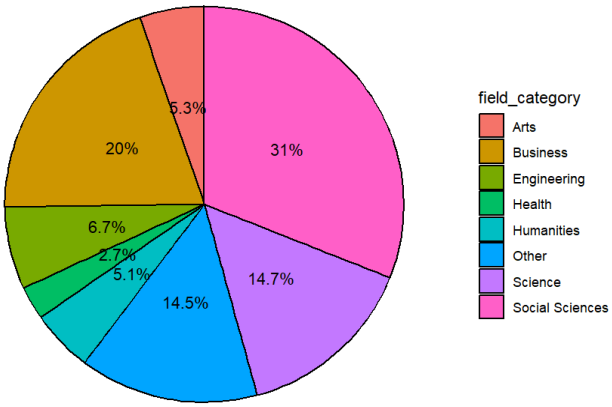


Figure 3: Distribution of Field Category

Figure 3 displays the distribution of participants across different academic fields. The largest segment is Social Sciences at 31%, followed by Business at 20%, and Science at 14.7%. The variety in fields suggests a diverse set of perspectives, but bias is expected as 51% of the participants are in Social Sciences or Business.

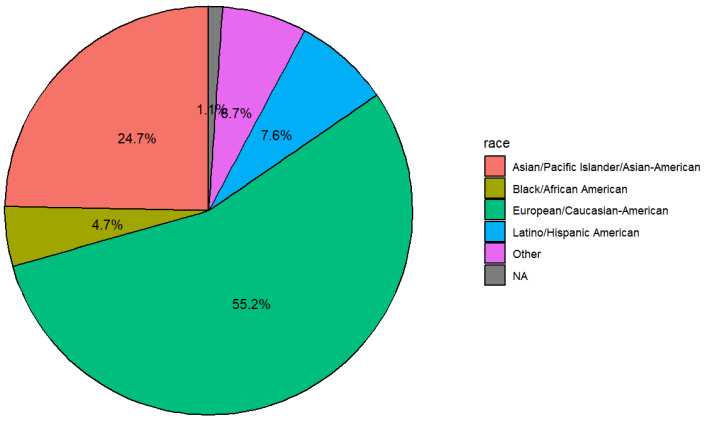


Figure 4: Distribution of Race

Figure 4 illustrates the racial composition of the participants, showing a dominant percentage of 55.2% European/Caucasian-American. The next largest group is Asian/Pacific Islander/Asian-American at 24.7%, followed by Latino/Hispanic American at 7.6%, and Other at 6.7%. A small fraction of 1.6% did not specify their race. Although the majority presence of European/Caucasian-Americans suggests potential biases, there is some racial diversity that can showcase the impact of race.

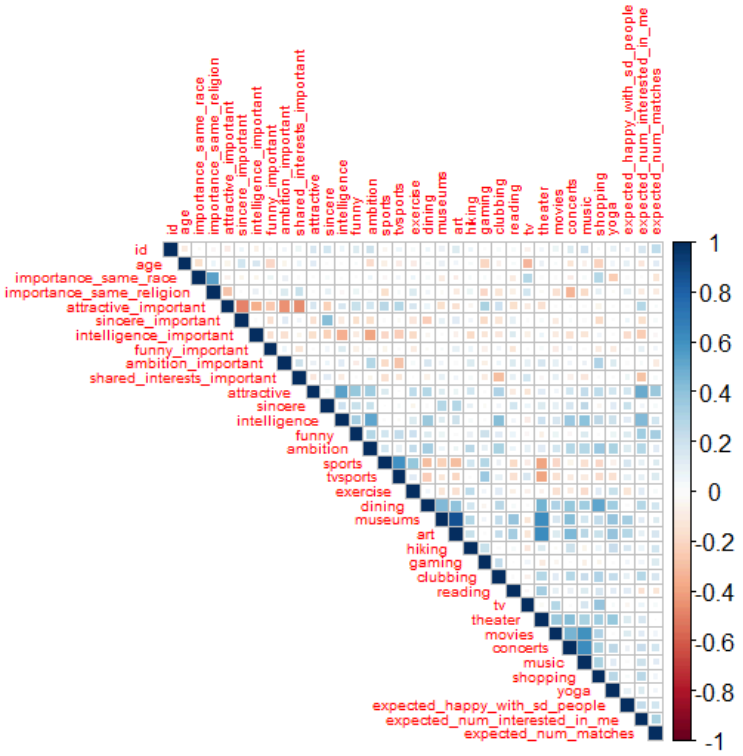


Figure 5: Correlation Matrix of Unique Individuals

Figure 5 shows a correlation matrix that visually explores the relationships between the numerical variables in the set. With this matrix we can form insights about the participants:

- A positive correlation between importance_same_race and importance_same_religion implies that participants that think one is important are likely to think the other is as important.
- For attractive_important, the more a participant thinks attractiveness is important, the less important the other attributes are. This reflects the idea that first impressions are often based on physical appearance, especially in the context of speed dating.
- For expected_num_interested_in_me, we see a person that rates themselves highly in attractiveness and intelligence expect more people to be interested in them.

The following graphs show the importance of traits by groups, specifically gender, field_category, and race. The ratings were done with a 100 points allocation. Since this is for EDA, we chose to remove outliers using the box plot method. This allows us to showcase the general trends between the participants, rather than being skewed by extreme values. From 551 to 341 observations.

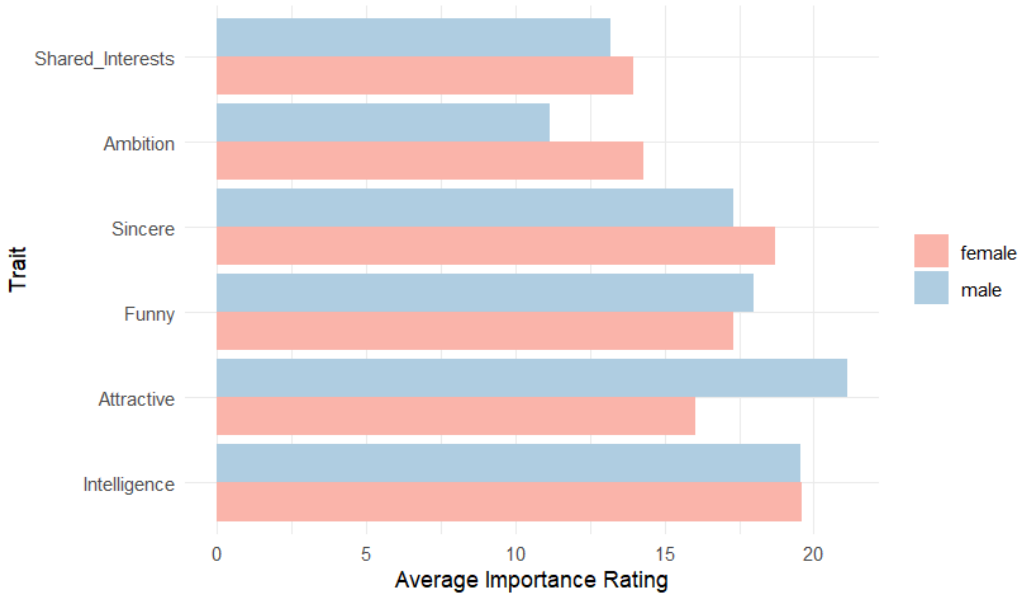


Figure 6: Importance of Traits by Gender

Figure 6 shows the average importance ratings assigned to traits by gender. Males value Funny slightly more than females and Attractive by a large margin. Females value Shared_Interests, Ambition, and Sincere. Both genders value intelligence the most and at a practically equal level.

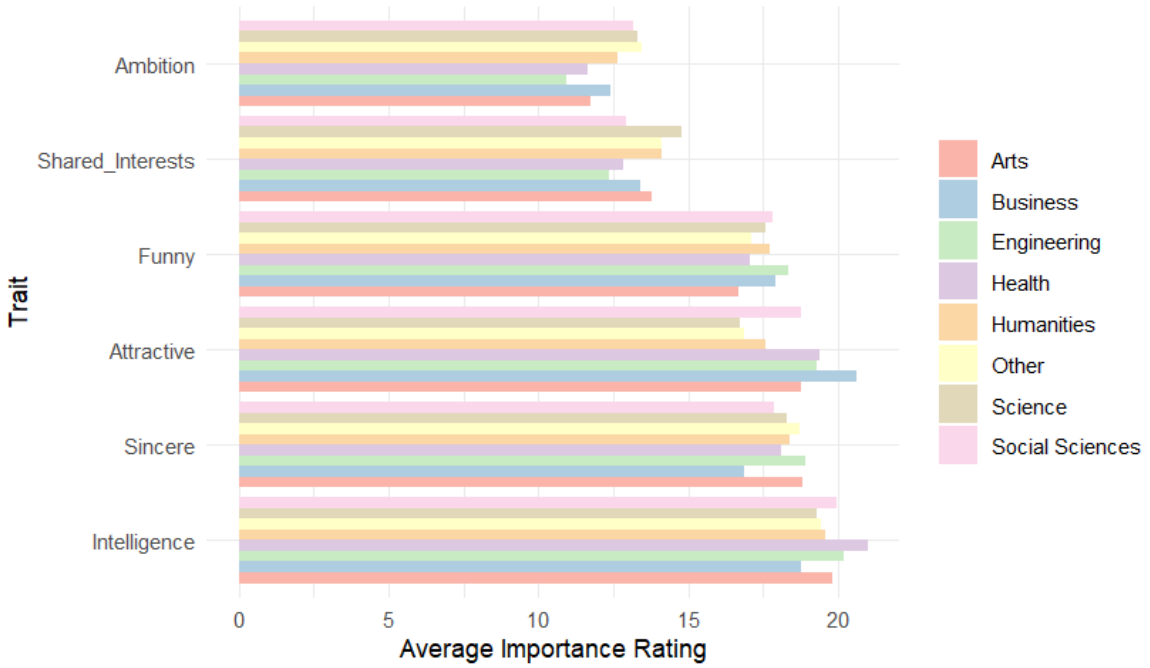


Figure 7: Importance of Traits by Field Category

Figure 7 shows the average importance ratings assigned to traits by field category. Ambition is valued the least and Intelligence is valued the highest by all fields. Ambition is valued the most by Science and the least by Engineering. Attractiveness is valued the most by Business and the least by Science. Intelligence is valued the most by Health and the least by Business.

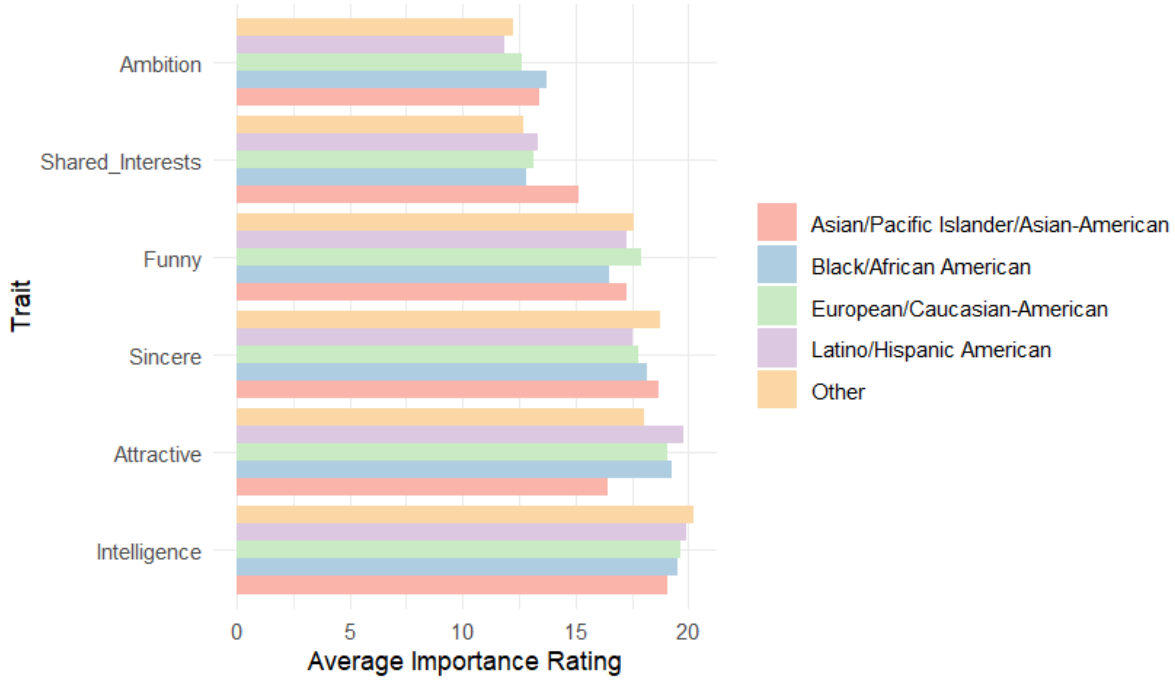


Figure 8: Importance of Traits by Race

Figure 8 shows the average importance ratings assigned to traits by race. Intelligence is valued the most by all races. Attractiveness is valued similarly by all except Asians. Funny has less variability across races. Sincere and Shared Interests are valued the highest by Asians.

The following graphs show the average self-rating of traits by groups, specifically gender, field_category, and race. The ratings were done with a 1-10 scale. Since this is for EDA, we chose to remove outliers using the box plot method. This allows us to showcase the general trends between the participants, rather than being skewed by extreme values. From 551 to 455 observations.

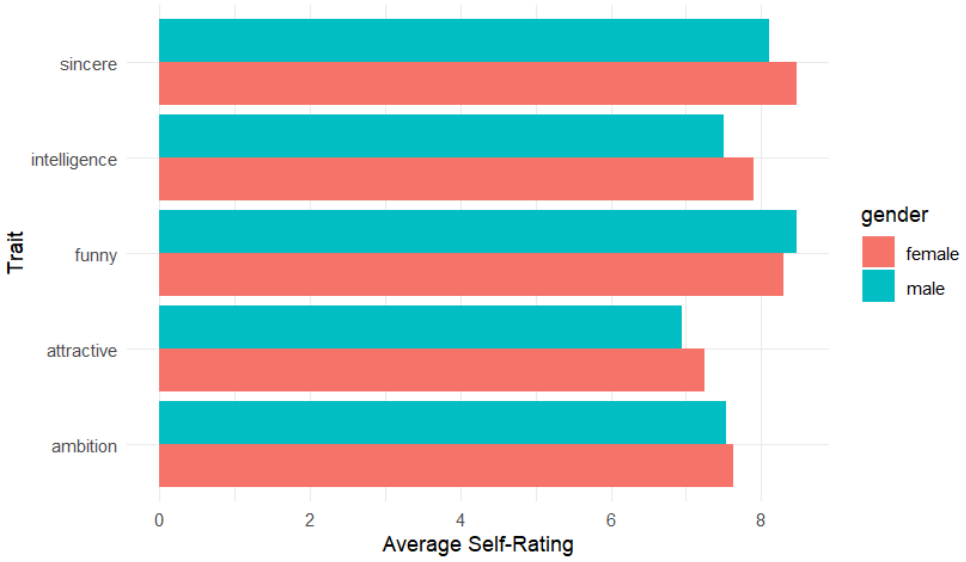


Figure 9: Self-ratings of Traits by Gender

Figure 9 shows the self-ratings of traits by gender. Funny is rated the most by males while sincere is rated the most by females. Attractiveness is rated the least by males and females. Females rate themselves higher in all traits except Funny.

Figure 10: Self-ratings of Traits by Field Category

Figure 10 shows the self-ratings of traits by field category. Sincere and Funny is rated the highest similarly by all fields, while Attractive is rated the least.

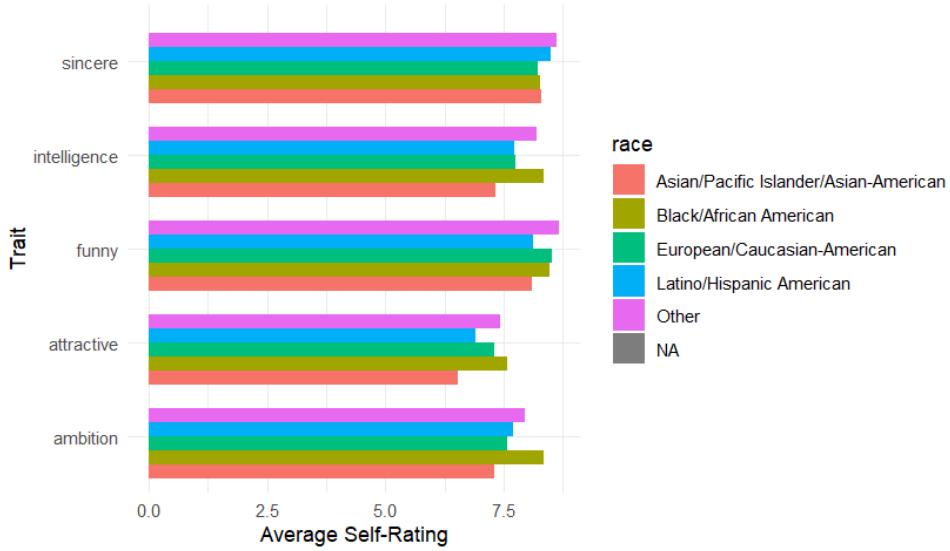


Figure 11: Self-ratings of Traits by Race

Figure 11 shows the self-ratings of traits by race. Sincere and Funny is rated the highest similarly by all races, while Attractive is rated the least.

The following correlation matrix is created based on the entire dataset that contains one pair of people in each observation to explore the relationship between how the rating of two people may be conceptualized.

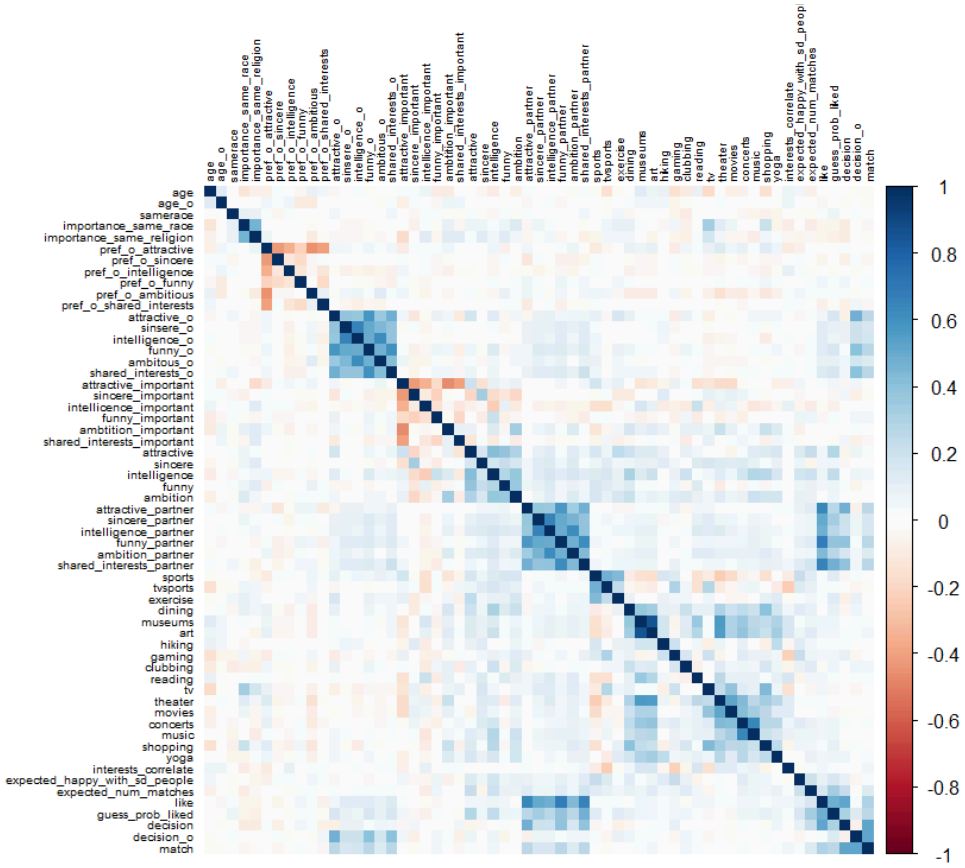


Figure 12: Correlation Matrix of Paired Observations

From the correlation matrix in Figure 12 above, it can be seen that:

- Blue blocks suggest high correlations among related variables, such as partner preferences (e.g., attractive_partner, sincere_partner, intelligence_partner) or personal traits rated by the partner (attractive_o, sincere_o, intelligence_o). High correlations within these blocks suggest that participants who rate high on one attribute (like attractiveness) also rate high on other similar attributes (like sincerity or intelligence), either for themselves or in their preferences for partners.
- The positive correlation between what people find important in others and how they perceive those traits in their partners may suggest that preferences influence perceptions, or vice versa. Additionally, the negative correlations between what the person finds

important suggests that participants may lean to valuing one specific while not prioritzing other values.

- The positive correlations between some of the interests suggest that these enjoying some interests inherently implies that other alligned interests are also enjoyed, while the negative correlations between interests suggest that enjoying one in particular may influence a lower rating on other unassociated interests.
- The blue boxes on the last row or last column indicate the variables that are most positively correlated with the response of match. These variables include the person's rating attractiveness, intelligence, sincerity, and funniness of their partner, as well as the partner's perception of those attributes on the person. Also the "like" attribute shows that the more you like your partner the more likely you are to match.

Variable Selection

For this project, the backward selection method is used for variable selection. Backward selection is particularly effective in reducing the risk of omitting relevant variables that may have significant combined effects with others, despite appearing insignificant when considered in isolation. This methodical reduction helps to keep only the variables that contribute to meaningful predictions of matches, therefore maintaining accuracy and interpretability despite reducing the number of variables included. The metric used to assess the number of variables to be selected by the backward selection algorithm is the Bayesian Information Criterion (BIC). The penalty term in BIC helps to provide a balance between the fitting capabilities and complexity of the model, which may further improve the model's generalizability. The BIC is chosen for 8 variables as evident from Figure 13 below:

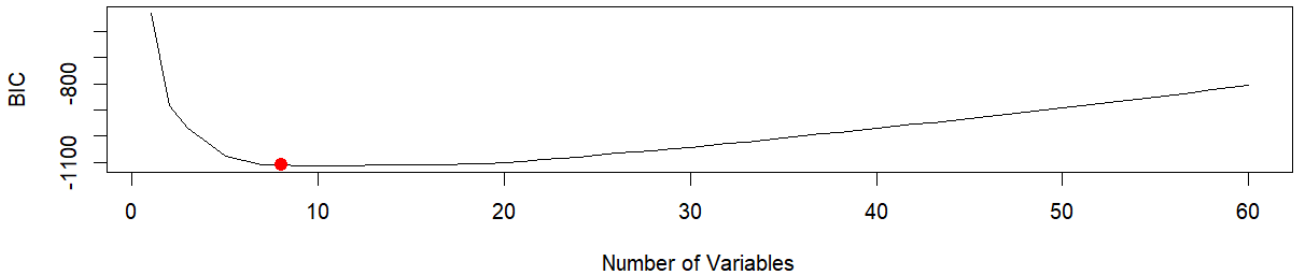


Figure 13: BIC vs. Number of Variables Plot

Predictive Models

Decision Tree with Pruning

Decision trees are a powerful tool in data analytics due to their simplicity and ease of interpretation. They produce a visual representation of the decision-making process that is involved in determining the final response. In this project, the pruned decision tree is modeled on the speed-dating dataset to predict successful matches between two people based on various relevant participant attributes.

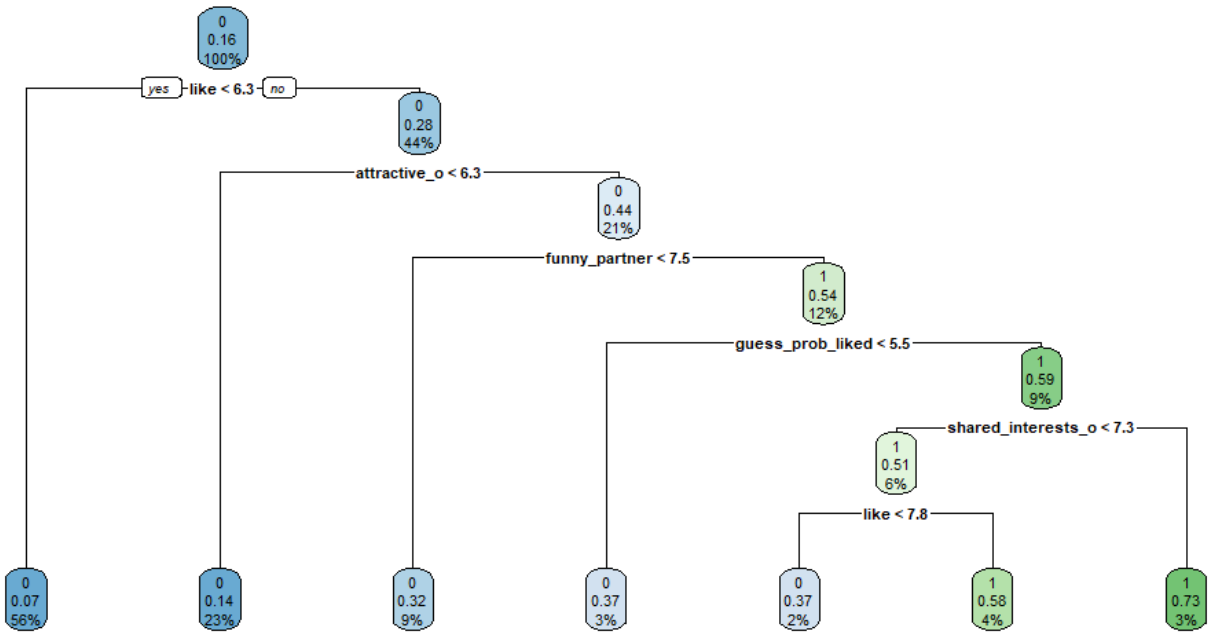


Figure 14: Pruned Decision Tree

The first split in the tree is based on whether the “like” rating, which is based on liking your partner, is less than 6.3. If it is less than the tree predicts no match which is the case for 56% of the training data. The second split is according to the “attractive_o” attribute and suggests that if your partner does not rate your attractiveness above 6.3 then the two would likely not match, which is the case for 23% of the training data. As the tree branches out, it incorporates decisions based on relevant factors such as “funny_partner”, “guess_prob_liked”, and “shared_interests_o”. The tree suggests that attaining high scores in these factors improves the likelihood of obtaining a match. The decision tree algorithm was implemented using all post-data-processing predictors and then with only the variables selected from the backward selection. Both produced almost identical results with the first achieving 84.8% accuracy, while the significantly reduced model attained 84.76% accuracy. The confusion matrix created by testing the decision tree of the latter is displayed below along with relevant metrics:

Table 2: Decision Tree Confusion Matrix

	Actual	
Predicted	**0**	**1**
0	2024	312
1	71	106

Accuracy: 84.76%

Misclassification Rate: 15.24%

Random Forest

Well-known for its promising accuracy and prediction power, the random forest algorithm reduces the risk of overfitting by combining several decision trees’ predictions and

basing the output on the majority vote of predictions [8]. Its retention of accuracy when a significant amount of data is missing, coupled with its efficient and significant accuracy–even with larger datasets– make it an attractive model for both regression and classification applications.

The model was run with the original training dataset containing all the variables, and another time including only the 8 significant variables recommended through the backward subset selection. Using the full training dataset, the algorithm yielded an accuracy of 84.7%, and a misclassification rate of 15.3%. However, the model performance was further enhanced when using the training subset that only includes the most significant 8 variables, yielding an improved accuracy of 85.6% and a reduced misclassification rate of 14.4% as seen in Table 3.

Table 3: Random Forest Confusion Matrix

	Actual	
Predicted	**0**	**1**
0	1895	264
1	78	143

Accuracy: 85.6%

Misclassification Rate: 14.4%

The Mean Decrease in Accuracy plot, shown in Figure 15, measures the importance of each feature based on the relative decrease of the model's accuracy when the feature is excluded. Hence, features that cause a larger decrease in accuracy are considered more important. From the included variables, it is evident that attractive_o is the most important variable as its exclusion results in the highest decrease in accuracy. The next most important feature is like, followed by

attractive_partner and 3 other significant variables. In that order, their exclusion affects the model performance to lesser extents.

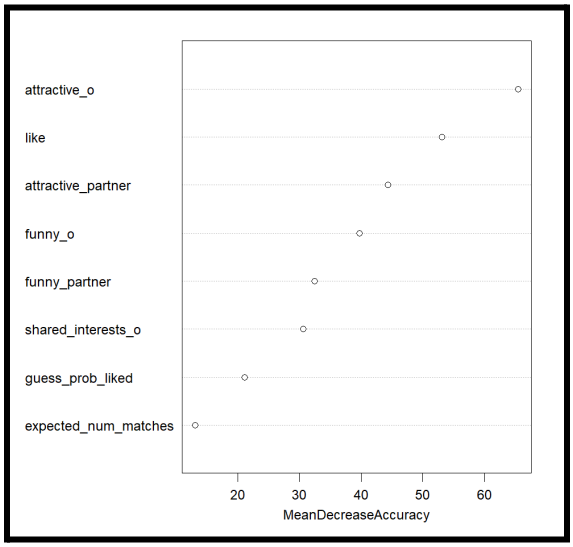


Figure 15: Mean Decrease in Accuracy plot

Logistic Regression

Logistic regression is a powerful statistical method used for binary classification problems, where the outcome variable is categorical with two possible outcomes. In the context of this report, logistic regression is highly relevant as it allows us to model and predict whether a speed dating match will occur based on various participant attributes and ratings. We first fitted a model using all available predictors to understand their impact on the likelihood of a match. The predictors with a p-value less than 0.05 are age_o, raceEuropean/Caucasian-American, attractive_o, intelligence_o, funny_o, ambitious_o, shared_interests_o, attractive_partner, sincere_partner, funny_partner, ambition_partner, art, tv, shopping, expected_num_matches, like, guess_prob_liked, field_categoryOther.

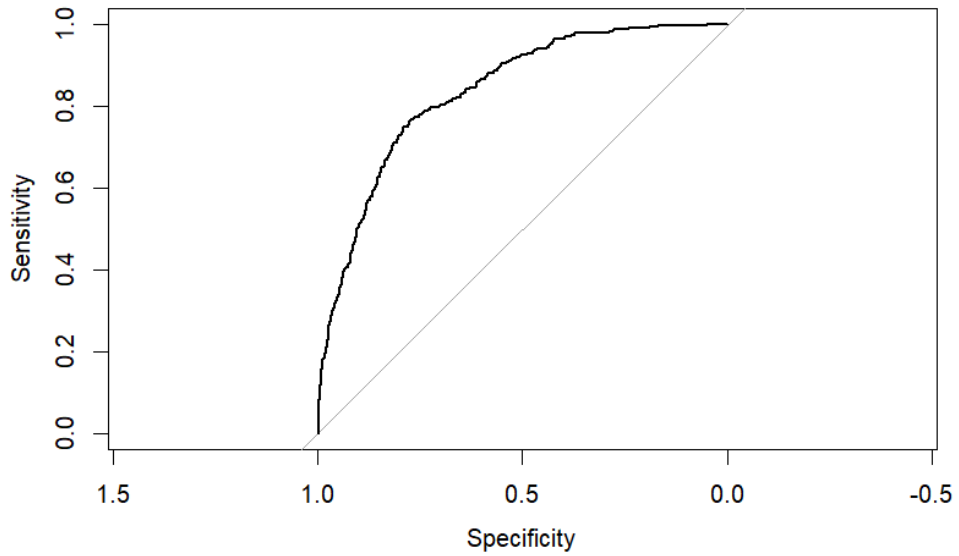
Table 4: Confusion Matrix for Logistic Regression With All Predictors

	Actual	
Predicted	**0**	**1**
0	1778	288
1	72	129

Accuracy: 84.11%

Misclassification Rate: 15.89%

We also plotted the Receiver Operating Characteristic (ROC) curve and calculated the Area Under the Curve (AUC) to assess the model's discriminatory power. The ROC curve bends towards the top-left corner, indicating that the model has a good separability between the positive and negative classes. A perfect model would have a curve that passes through the top-left corner. An AUC of 83.58% signifies that the model performs well in distinguishing between matched and non-matches.



Area under the curve: 83.58%

Figure 16: ROC Curve and AUC for Logistic Regression with all Predictors

To improve the model's performance, we identified the best variables using backwards selection. The same prediction and evaluation steps are followed.

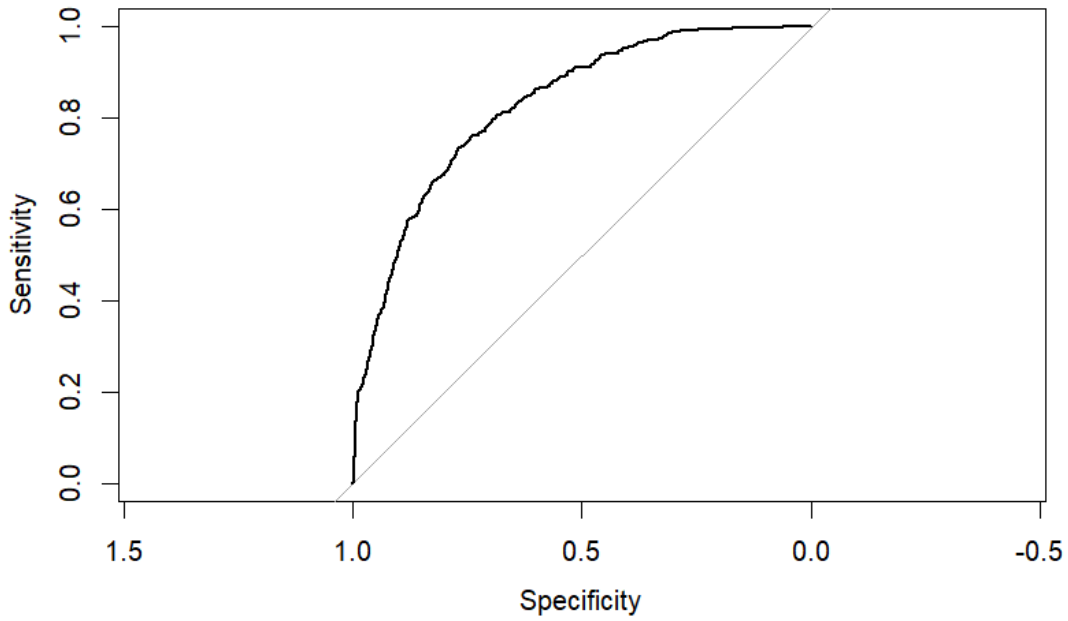
Table 5: Confusion Matrix for Logistic Regression With Best Predictors

	Actual	
Predicted	**0**	**1**
0	1792	311
1	58	106

Accuracy: 83.7%

Misclassification Rate: 16.3%

This model performs slightly worse, meaning that we lost some information in model selection, or the full model was overfitting the training data.



Area under the curve: 82.85%

Figure 17: ROC Curve and AUC for Logistic Regression with Best Predictors

Gradient Boosting

Gradient boosting is another powerful machine learning algorithm that can be applied to this context. Being an ensemble learning algorithm, this method combines weaker performing learners, which typically consist of shallow decision trees, creating a stronger predictive model. Applying it to the speed dating dataset, the mode was run with incorporating the best variables selected earlier by the backwards selection. The threshold was set as 0.45, and not the default 0.5 as it was fine tuned using a function that yields the threshold that provides the highest accuracy. This threshold can be further tuned to cater to varying concerns and priorities. WIthin the context of this problem, the team believes that accuracy is most important as a metric, when compared to precision, or recall. False positives and false negatives are almost equally unfavorable in this problem. Matching with someone that is incompatible with you can likely result in a failed and

unhealthy relationship. Similarly, failing to match with someone who in reality is compatible with you will result in an opportunity loss, and sacrifices the loss of a love that could have flourished, resulting in a successful relationship.

Table 6: Confusion Matrix of Gradient Boosting

	Actual	
Predicted	**0**	**1**
0	1976	288
1	94	155

Accuracy: *84.8%*

Misclassification Rate: *15.2%*

Recommendation System

Recommendation systems, particularly relevant in the context of speed dating, are designed to help individuals discover potential partners whose preferences and characteristics align with theirs, thus potentially enhancing the possibility of mutual attraction and compatibility. It goes through data to find patterns and similarities among users. For this project we implemented a recommendation system through the following:

- **Grouping and Identification**: Firstly, the participants are grouped based on personal and partner preferences. These variables include intrinsic attributes like age, race, and personal interests, as well as extrinsic preferences such as what they look for in a partner. Then we generate identifiers (personID and partnerID) for each unique combination of these attributes and identify each individual participant.

- **Content-based Filtering:** We employ the cosine function to assess similarities across different attributes of different participants to capture similarities in higher dimensions. A similarity matrix is then created that shows how similar each participant is to every other participant.
- **Collaborative-based Filtering:** The implementation also involves creating a sparse matrix from historical match data, indicating which participants have chosen each other in past speed dating rounds. This matrix helps to add data about the participants' past decisions to influence the prediction of future matches.
- **Scores and Recommendations:** The scores can be based on the combination of similarities and past decisions, or only on similarities. Additionally, any scores given to same-gender matches are set to zero to ensure that only opposite-gender matches are recommended. Finally, the top recommendations for each participant are extracted from the recommendations based on the scores.

Results Summary

Table 7: Accuracy of Models

Model	Accuracy
Decision Tree	84.76%
Random Forest	85.6%
Logistic Regression	83.7%
Gradient Boosting	84.8%

The table above summarizes the accuracy of our models used to predict the likelihood of a match based on the dataset. Based on the results, random forest performed the best, and logistic

regression had the lowest accuracy. However, all the models yielded accuracies within a considerably narrow range of values between 83% and 85%.

- Decision Trees: Accuracy of 84.76%. The model is easy to interpret and handles non-linear relationships between variables effectively, however it is prone to overfitting.
- Random Forest: Accuracy of 85.6%. This model mitigates the overfitting from decision trees, and its superior performance indicates that it can capture complex interactions between the variables better than a single decision tree.
- Logistic Regression: Accuracy of 83.7%. This model is useful for binary classification problems as is the case here and useful for its simplicity. It does have the lowest accuracy among the models but it does remain interpretable and beneficial through the p-value from using all predictors.
- Gradient Boosting: Accuracy of 84.8%. This model builds an ensemble of trees sequentially with each tree correcting the errors of the previous one. Its accuracy being close to that of Decision Trees suggests that it is not effective in capturing the complex interactions between variables.

Conclusion and Future Work

The applications of machine learning algorithms are fascinatingly diverse, and may extend to highly personal and complex phenomena in our lives including the possibilities of finding the perfect love match for individuals. This study proves that algorithmic matchmaking is indeed promising with model accuracies as high as 85.6% as seen in random forest. This project lays a strong foundation for future research and work on this topic with the objective of developing models that can predict couples' compatibility more accurately, further contributing to reduced rates of divorce and failed relationships. This project provides valuable insights on

what each of the genders values most. Additionally it predicts couples' compatibility based on the most significant attributes through various model types and algorithms. Furthermore, it uses a recommendation system to find potential matches for an individual based on their own attributes through calculating a tailored compatibility score. This research can be further developed by running a similar speed dating experiment in different regions of the world so that the dataset is more representative. Hence, including different individuals from a broader set of geographical regions such as the Middle East, for instance, adds an interesting layer to the research as different cultural contexts often influence priorities and findings. Furthermore, the experiment's questions can be modified to include deeper topics, ones that are more fundamental to judging a person's character to make strong conclusions such as their compatibility with the other participants. Finally, variations of the recommendation system can be explored. For instance, instead of matching candidates based on similarities and closeness in answers, one could consider the potential attraction to varying interests, and differing passions, linking back to the famous ideology of opposites attracting one another. Hence, it would be interesting to build a recommendation system that considers factors that are different from closeness and similarity.

References

[1] M. Coleman, L. Ganong, and K. Leon, “Divorce and postdivorce relationships,” *The Cambridge Handbook of Personal Relationships*, pp. 157–173, Jun. 2006. doi:10.1017/cbo9780511606632.010

[2] E. J. Finkel and P. W. Eastwick, “Speed-dating,” *Current Directions in Psychological Science*, vol. 17, no. 3, pp. 193–197, Jun. 2008. doi:10.1111/j.1467-8721.2008.00573.x

[3] E. J. Finkel, P. W. Eastwick, and J. Matthews, “Speed‐dating as an invaluable tool for studying Romantic attraction: A methodological primer,” *Personal Relationships*, vol. 14, no. 1, pp. 149–166, Mar. 2007. doi:10.1111/j.1475-6811.2006.00146.x

[4] R. Fisman and S. Iyengar, 2004, “SpeedDating”, Columbia Business School. [Online] Available: https://www.openml.org/search?type=data&sort=runs&id=40536&status=active

[5] A. Paul and S. Ahmed, “Computed compatibility: Examining user perceptions of AI and matchmaking algorithms,” *Behaviour & Information Technology*, vol. 43, no. 5, pp. 1002–1015, Apr. 2023. doi:10.1080/0144929x.2023.2196579

[6] G. S. Becker, “A theory of marriage: Part I,” *Journal of Political Economy*, vol. 81, no. 4, pp. 813–846, Jul. 1973. doi:10.1086/260084

[7] M. V. Belot and M. Francesconi, “Can anyone be ‘the’ one? evidence on mate selection from speed dating,” *SSRN Electronic Journal*, 2006. doi:10.2139/ssrn.941111

[8] “Random Forest algorithm in machine learning,” AlmaBetter, https://www.almabetter.com/bytes/tutorials/data-science/random-forest (accessed May 18, 2024).