Will I Want You To Be My Date?

Introduction

Dating has been present in different shapes or forms over the course of human history. Even though traditions and dating channels have changed, people's interest in the science of dating has not decreased. People meet thousands of individuals over the course of their lifetime, yet they do not spend a vast amount of their time interacting with the majority of them. Brief interactions with a newly met person can shape the course of a potential relationship. Research has found that first impressions are formed incredibly quickly upon meeting someone new, and that accurate judgments of someone else's personality traits can be formed in interactions as short as four-minutes long (Bar et al., 2006). However, little is known as to how much different traits and interest compatibility matter when determining whether or not a person would be interested in going on a date with someone new. Therefore, the purpose of this study is to evaluate the importance of different demographic features and people's interests in predicting whether or not a person will want to date a potential romantic partner they have just met.

The data used in the report comes from an experimental study conducted at Columbia's Business School (Fisman et al., 2006). The researchers collected data from speed dating events that took place between 2002-2004 where participants had the chance to interact with potential romantic partners of the opposite sex. After a four-minute interaction, each participant had to answer a series of questions regarding their encounter and decide whether they would want to go on a date with the potential partner or not. In this study, we are interested in using demographic information and personal interests and preferences to predict whether a person would want to go on a date with the person they have just met.

Description of the Data

The final dataset contains information on 551 participants and 8,368 two-people interactions between them. Each row in the dataset consists of a distinct participant-partner relationship (for example, if participant 1 interacted with participants 2 and 3, the first row reflects the relationship between participants 1-2 and the second one of participants 1-3). The original data has 195 features; however, only 43 features were chosen to be analyzed in the study based on the question of interest. Following other studies that investigated the effects of different determinants of liking (Walster et al., 1966), we chose features that we believed may play a role in a person's decision to date someone, and we divided them into four big categories: demographic information such as religion, race, and age; interest in different leisure activities such as sport, exercise, dining, museums, art, hiking, and gaming; perception of the other person in terms of attractiveness, sincerity, intelligence, fun, ambition, and shared interests; and the importance that each participant assigns to each of those six factors when choosing a significant other.

Although we were interested in including other features such as wealth in our prediction, we opted out from doing so due to the large number of missing observations. Moreover, we created additional variables. We calculated the difference between each participant and their potential partner by subtracting one from another (if numerical) or comparing if they were the same (if categorical) for field of study; dating goal; frequency of dates and outings; age; importance of race and religion; and interest in sports,

tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv, theater, movies, concerts, music, shopping, and yoga; and race. The original 43 features chosen include 8 categorical variables (id and predictors: gender, race, field of study, dating goal, and frequency of dates and outings; and the response, whether or not they want to go on a date with a person). The other 35 variables are numerical.

Data Transformation & Cleaning

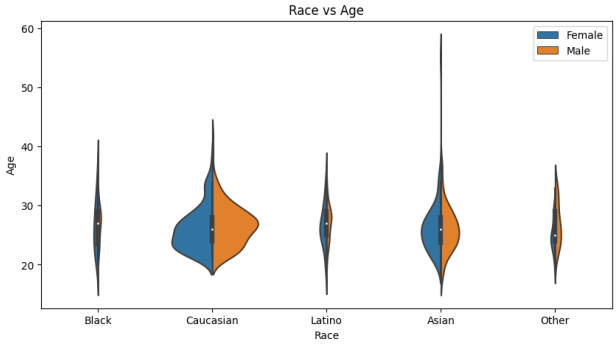
We started off the analysis by identifying missing observations in our dataframe. Out of the 43 features we were interested in, 40 had missing observations. Since not all individuals in the study had the same number of rows in the dataset (as different people interacted with a different number of potential partners), we were unable to input the missing values from the entire dataset (as the values of observations that appeared more often would be weighted more heavily when they should not be). Thus, we first created a smaller dataframe that only contained the rows with unique ids (in other words, we had one observation per participant). For the features that did not depend on the partner (demographics, interests, and importance of attributes), we imputed the missing values in the smaller dataset using kNN imputation (with 5 neighbors) for numerical variables and mode for categorical ones. We then merged the original dataset with the smaller one, and we replaced the values of features that originally had missing values with the newly created ones. We finalized the imputation process by imputing the numerical features that were dependent upon the partner (how much they like them, how likely they think the other person is to want to date them, and their rates on the six main attributes) with kNN imputation using 5 neighbors in the entire dataset.

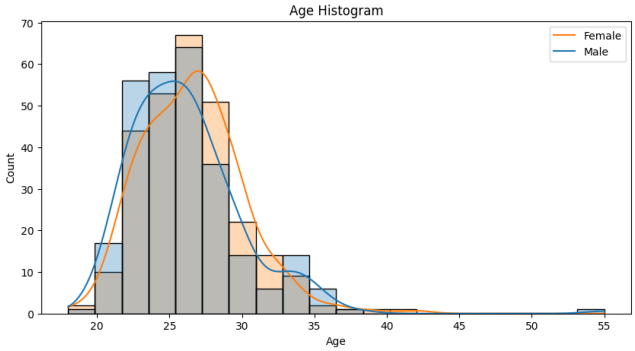
We also consolidated some categorical features like field of study, dating goal, and frequency of dates and outings into more comprehensive categories. Moreover, we noticed that some of our variables of interest had values that were outside of the permitted range. Even though the six features that measured the importance that each participant placed on the six main attributes of interest were supposed to add up to 100, there were nine observations that had a sum of less than or greater than 100. In those cases, we rescaled the variables so that the six attributes would add up to 100. Furthermore, a few features that were supposed to range from 1 to 10 (as participants had to determine how interested they were on certain activities on a scale from 1 to 10), had values lower than 1 or greater than 10. We tackled the issue by assigning values of 10 to any observation with a value greater than 10 and assigning values of 1 for any observation with values lower than 1.

We finalized the data cleaning and transformation process by deleting observations that had missing values for the id of their potential partners (pid), as our analysis is focused on such interaction. There were only 10 such observations. We then used the potential partner's pid to include the demographic and interest features from the potential partner in the same row as the individual, and we compared the participant and the partners' features to add them as the new "difference" columns described above. Nevertheless, as the new variables were created using a linear combination of the original predictors, we had to acknowledge the potential for high collinearity in our dataframe. Therefore, we used Variance Inflation Factor to identify features that were highly correlated with other predictors and decided to drop most of the original variables (both from the participant and the partner) that were used to create the new difference/match features, and kept only the new ones. We also removed the ids from the predictor matrix (as they might add noise to a model). Lastly, in order to run the models, we used One Hot Encoding to transform the 2 non-binary categorical variables (person's race and field of study) plus gender, which resulted in a grand total of 51 variables (50 independent and 1 response).

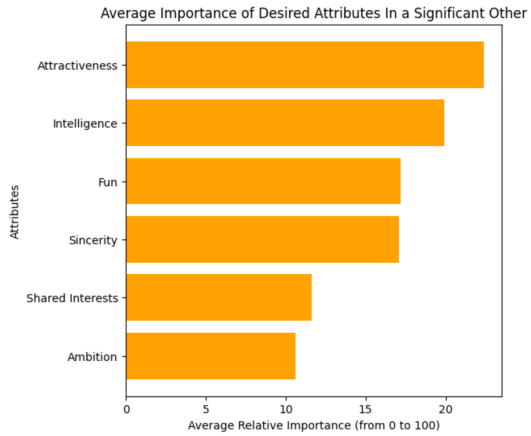
Data Visualization

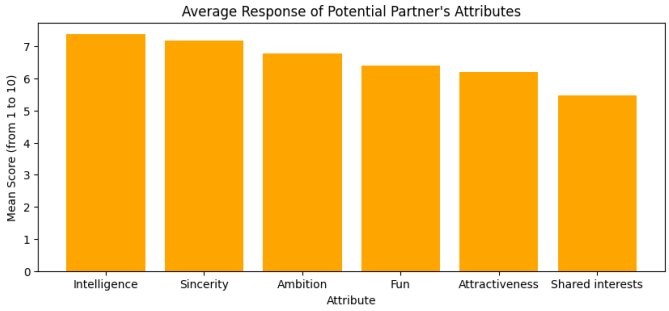
Once we cleaned the data, we started exploring the demographics of our participants to better understand our sample. As can be seen on the two plots below, most participants were Caucasian, followed by Asians, and were between 20 to 30 years old. Furthermore, we saw a similar proportion of female and male participants, which is expected as all interactions were between heterosexual individuals.





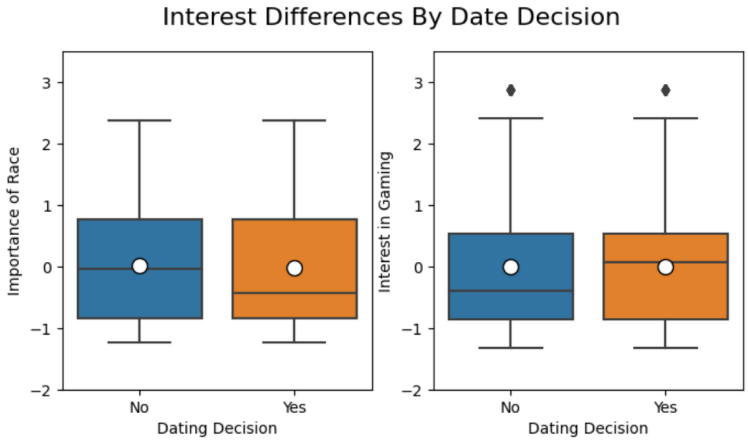
Since we were trying to better understand what attributes are associated with the willingness to go on a date with an individual, we also looked at what features our participants rate as most important in a potential romantic partner. Participants were asked to distribute 100 points across six different attributes according to the importance they place on each of them when dating. As we can see from the figure below, most individuals rated attractiveness as the most important attribute they look for in a partner, which is in accordance with the literature (Walster et al., 1966), although it is closely followed by intelligence. On the other hand, ambition and shared interest seem to be the attributes that people are least interested in when considering whether to date someone or not. We also looked at how well our participants were ranked in terms of those 6 attributes by one another. Although the averages of all six categories are quite similar, the average score for intelligence and sincerity was the highest in our sample, whereas attractiveness and shared interest were deemed the lowest.



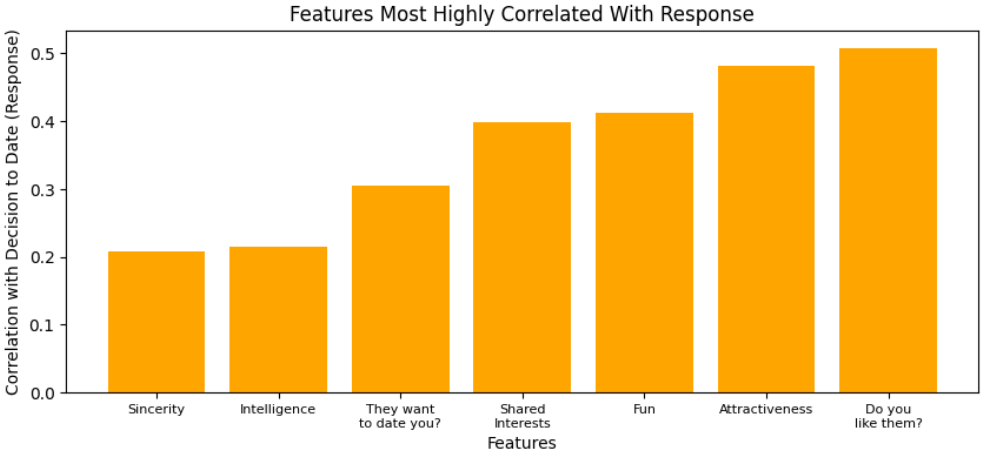


Furthermore, we were also interested in understanding whether there was any relationship between the magnitude of the difference between the person and his/her potential partner's interests and the decision to go on a date with them or not. Surprisingly, there were no clear patterns in terms of interest match and the decision to date the person or not. All of our variables of interest had quite similar

distributions regardless of the response, yet importance of race and interest in gaming are the ones with the highest difference in median between response outcome. The median difference in importance on race was smaller and the median difference in interest in gaming was greater among individuals who decided to date the potential partner.



We also investigated how closely correlated our 43 features were to the response variable of whether or not they would want to go on a date with the potential partner. No variable had an absolute correlation value greater than 0.5. Nevertheless, using a threshold of 20%, we were able to identify the seven features that are most highly correlated to our response. As we can see in the plot below, the response to the question 'Do you like the person?' and attractiveness are the features more strongly associated with whether or not the person would want to go on a date. However, it is worth noting that the perception of whether the other person would want to go on a date with them is the fifth most highly correlated feature with our response. Such a finding not only shows that a person takes into account their chances with someone else when considering whether to go on a date with them or not, but also seems to be in accordance with the Matching Hypothesis in the Psychology field, which states that people select partners with whom they match in terms of social desirability (Shaw Taylor et al., 2011).



Project Question

The data exploration described above led us to develop our project question in more detail. The focus of this project is to predict whether a person will want to date a potential partner based on demographics and their interest matches.

Baseline model

Over the course of this project, we created different models to predict whether a person would want to go on a date with the potential partner or not after a short, 4-minute interaction (in our dataset, this variable is called “dec”). For this purpose, we first split our data into training and test datasets (70% of the data went to the training set). We ensured that the proportions of the categories of the response variable were equal across the train and test sets (around 42% of each set had a value of 1 for the response). We then implemented a Logistic Regression model with no cross-validation or penalty, and we included all 50 features in our design matrix X to predict our response variable “dec”. Since our data was well balanced, we decided to use accuracy to measure our models' performance across different algorithms. Surprisingly, the baseline model performed relatively well, as it had a training accuracy of 79.18% and a test accuracy of 76.16%, which are relatively high. Moreover, since the train and test accuracies were pretty close to each other, we concluded that the baseline model did not overfit the training dataset.

Model Selection

To improve our baseline model’s performance, we decided to test different algorithms that are suited for classification purposes. Since we have 50 predictors in our dataset, we considered that an improvement to the logistic regression baseline model would be to include regularization to identify the most important predictors to be kept. Thus, we first fit a logistic regression model with a Lasso penalty. Nevertheless, we also wanted to investigate a potential non-linear relationship between the features and the response. We first opted to fit a kNN model as it is one of the simplest machine learning algorithms that could describe such a relationship. However, since kNN tends to work best when there are fewer predictors (which is not necessarily our case), we also fit a decision tree as it is an easy and intuitive way to express the relationship between the predictors and response. Nonetheless, acknowledging that a single decision tree may not be flexible enough to capture the patterns in our data, we also increased the model’s complexity via XGBoost and, to decrease the high correlation between the trees, random forests. For all of the above models (except for the baseline), we performed cross-validation to find the optimal hyperparameters via a grid search.

The Lasso Regression model performed relatively similar to our baseline model, as it had a train accuracy of 79.2% and a test accuracy of 77.6%. The grid search identified that the best hyperparameter for our inverse regularization strength was 0.1, which, as it is the second-to-largest value for the hyperparameter we tested, is the second weaker regularization parameter. Nevertheless, the model still identified that 10 out of the 40 original predictors were not important, which decreased the complexity of the model. On the other hand, the grid search for kNN identified 30 neighbors and distance weights (through which closer neighbors become more influential than more distant ones) to be the optimal

combination to improve our model's performance. Nevertheless, this model had a high degree of overfitting, as the train accuracy was perfect (100%) whereas the test accuracy was 77.42%. Therefore, kNN did not improve our test accuracy significantly and, since the train accuracy was perfect, we can state that the kNN model would not generalize to unseen data as well as the Lasso Regression one.

Even though the KNN model did not perform as well as the Lasso Regression model, we still wanted to investigate a potential non-linear relationship between our predictors and response. Thus, we first fit a simple, single decision tree that performed relatively well. The model not only showed much lower levels of overfitting relative to the kNN model, but the test accuracy was not severely affected: the decision tree had a train accuracy of 85.36% and a test accuracy of 77%. The grid search identified a maximum depth of 8 and a splitter strategy of 'best', which always ensures that the algorithm chooses the best split, as the best hyperparameters for the model. Moreover, the algorithm deemed the predictors like and attr (how much they like the other person and how physically attractive they find them on a scale of 1 to 10, respectively) as the most important in predicting whether or not someone would want to go on a date with a newly met individual. We then tested fitting an Extreme Gradient Boosting algorithm to our data. The algorithm comes from the XGBoost Python package, and it includes more regularization than Gradient Boosting does, as the loss function is regularized using l1 and l2 and combines not just the residuals but also a penalty for model complexity (*XGBoost Documentation — Xgboost 1.7.1 Documentation*, n.d., (*How XGBoost Works - Amazon SageMaker*, n.d.)). The loss function utilized to perform the algorithm is the following:

$$\mathcal{L}^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t)$$

Where y_i is the real data, y-hat_$i^{(t-1)}$ is the $f(x + \Delta x)$ and $\Omega(f_t)$ is the regularization term (Leventis, 2022). Unlike gradient boosting, XGBoost fits the trees in a parallel way, and evaluates the performance of each possible split in the training dataset (*What Is XGBoost?*, n.d.). Even though the XGBoost includes more regularization than other models we tested, it still had quite a large degree of overfitting. The grid search determined that the best hyperparameters were a learning rate of 0.06, a maximum depth of 8, and 500 estimators, among others. Nevertheless, the test accuracy was 82.02%, but the train accuracy was 100%, showing, once more, that this model would not necessarily generalize well to unseen data. We also tested a Random Forest algorithm, as it would allow us to decrease the highly correlated trees from XGBoost. In this case, the grid search identified a maximum depth of 20 and 600 estimators to be the best hyperparameters to improve our model's accuracy. Nevertheless, this algorithm was also prone to overfitting, as it had a train accuracy of 100% and a test accuracy of 79.57%. Even though the Random Forest and the XGBoost models have the highest reported test accuracy, we are hesitant to state that such models outperformed the rest since the high degree of overfitting would lead to lower performance on newly seen data.

We finalized the report by fitting a Stacking Model, which is an ensemble method that takes different types of models to make intermediate predictions of the training set, and then uses such predictions to fit a new, finalized model ("Stacking in Machine Learning," 2019). In this case, we used the former models described (Logistic regression with Lasso Regularization, kNN, Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier) as base learners with the hyperparameters identified

using GridSearch. During the process of Stacking, the algorithm takes k-folds and uses the base learners to create predictions and saves them into a new feature matrix. This matrix along with the original predictions can be used to train the meta-learning algorithm, in our case Logistic Regression. Then we created an ensemble model with the trained meta-learning model and the original base learning models, which we used to generate our predictions ("Stacking in Machine Learning," 2019). Although the Stacking model combined the predictions from the previous models described, its performance did not improve relative to the original algorithms. The training accuracy was, once more, perfect, whereas the test accuracy was 82.08% (very similar to that of XGBoost). Thus, the Stacking model also presents a high degree of overfitting.

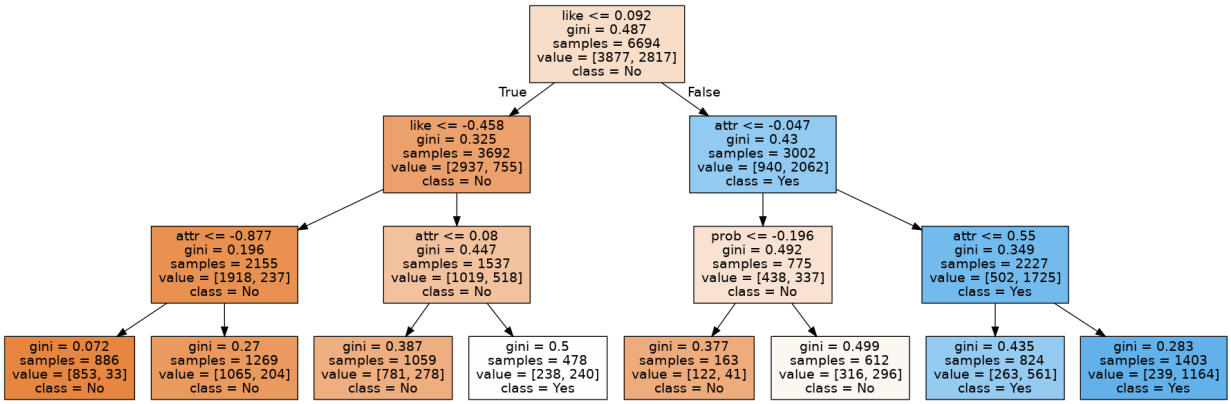
Chosen Model

Even though none of the models had a particularly high test accuracy, they all performed decently well. However, some models had quite high degrees of overfitting, as is shown by the difference between training and test accuracies and the perfect train accuracy.

Model	Train Accuracy	Test Accuracy
Baseline	79.18%	76.16%
Logistic Regression	79.24%	76.7%
KNN	100%	77.42%
Decision Tree	85.36%	77%
Random Forest	100%	79.57%
XGB	100%	82.02%
Stacking	100%	82.02%

We consider that a good model for our data is one that not only has high accuracy in newly seen data (and a low degree of overfitting), but also one that is parsimonious. In our case, the models with the highest accuracy and lowest degree of overfitting were the logistic regression (with and without regularization) and the single decision tree. Between the two logistic regression models, the one with Lasso regularization is preferred, as the non-important predictors make it simpler than the baseline model. The decision between the single decision tree and the logistic regression with regularization is slightly more complicated. While both models had a similar performance (79.24% train accuracy, 76.70% test accuracy for Regularized Logistic Regression, and 85.36% train accuracy, 77% test accuracy for Simple Decision Tree), the decision tree's performance was slightly better for both train and test set. Additionally, decision trees have the great advantage of being an intuitive algorithm that is easy to explain and interpret. Therefore, given the accuracy reported in both train and test sets and the interpretability and simplicity of the model, we believe that the decision tree is the best one out of all the models we tested. Nevertheless, decision trees do suffer from the disadvantage of having an unstable nature compared to other models. Given that we are not looking at a big sample of data, a small trend in the data results in a big change in the tree. Still, the model performs relatively well and is deemed as the best of all the ones

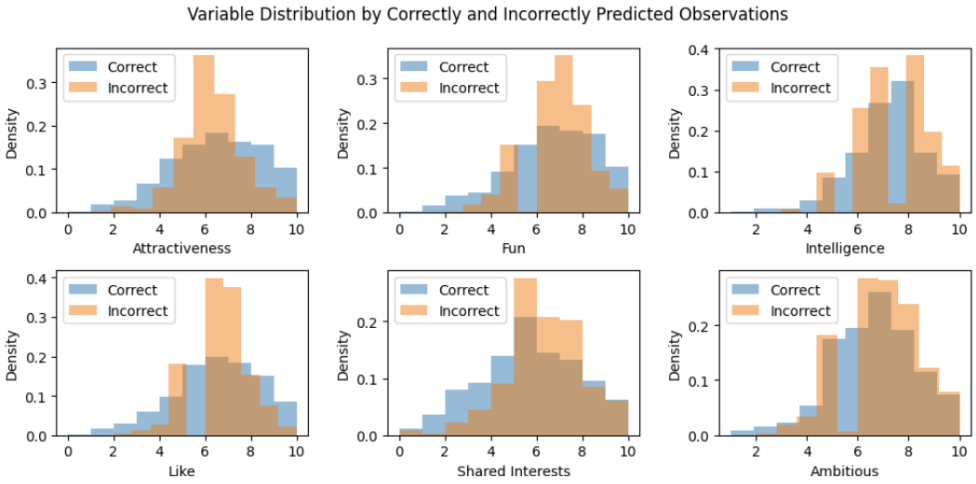
tested. The plot of a shorter version of the chosen model (with less depth for visualization purposes but with the same hyperparameters) is seen below. The tree has liking the potential partner in the first split which is, thus, considered to be the most important predictor.



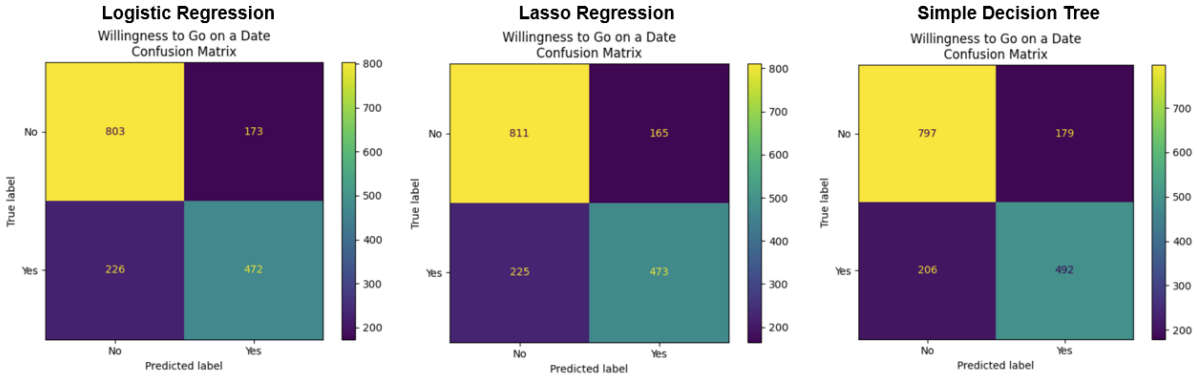
We also calculated the feature importance for the baseline, Lasso logistic regression, and ensemble models (decision tree, random forest and XGBoost), with similar results obtained across all of them. As we can observe from the table below (based on the single decision tree), by far the most important feature is how much the participant likes the other person (0.5), and in second place, the attractiveness of the potential date (0.12). To a much lower extent (0.03), the perceived probability of the other person wanting to date them, the importance of shared interest/hobbies, and the importance of how fun the other person is also play a role in predicting the decision of wanting or not to date a newly met person.

Most Important Features		Feature Importance
1.	Overall, how much do you like this person?	0.50
2.	Rate attractiveness of potential date	0.12
3.	How probable do you think it is that this person will say “yes” for you?	0.03
4.	How important is it for you that a potential date has shared interests / hobbies?	0.03
5.	How important is it for you that a potential date is fun?	0.03

We then plotted the variable distribution by correctly and incorrectly predicting observations for every variable. We noticed that most of the variables had a similar distribution, with the exception of the six variables displayed: Attractiveness, Fun, Intelligence, Like, Shared Interest, and Ambitious, which had a considerable difference as shown in the graphs above.



In our analysis, we also computed the confusion matrix for both the baseline model (simple logistic regression, the lasso logistic regression, and our chosen model which was the simple decision tree)



Metric	Baseline Logistic Regression	Lasso Logistic Regression	Single Decision Tree
True Negative Rate	82.27%	83.09%	81.66%
True Positive Rate	67.62%	67.66%	70.49%
Precision	73.18%	74.14%	73.32%
F1 Score	70.29%	70.81%	71.88%

If we compare the metrics, the regularized Logistic Regression (Lasso) outperforms the simple logistic regression in every single one: The True Negative Rate (TNR), True Positive Rate (TPR), Precision (which measures the proportion of positive identifications that were actually correct), and F1 Score (which combines precision and TPR). We can also identify that the TNR (specificity - 83.09%) is

substantially larger than the TPR (recall - 67.77%), meaning that the model is better at predicting the decision of 'not' wanting to go on a date than predicting the decision of wanting to go on a date with the newly met individual. In our problem, sensitivity is a more important metric as we would like to predict positive decisions.
In this sense, the results of the decision tree seem better as the sensitivity is larger (70.49%), as well as the F1 score (71.88% vs 70.81% of the Lasso Regression). Additionally, the gap between the TNR and TPR is smaller, meaning that the model is more parsimonious and robust for predicting correctly the decision on whether to date or not.

Conclusion

Even though individuals look for different attributes in their potential partners, different trends can be analyzed to better predict if a person would want to date another individual. After cleaning the data, performing feature engineering, and fitting a baseline logistic regression model, we fitted a regularized logistic regression, kNN, decision tree, random forest, XGBoost, and stacking models. Building these statistical models allowed us to answer the main question and to predict whether a person would be willing to date another one after a short 4-minute interaction. Out of all the models we fitted, we believed the single decision tree to be the best one, as it is a parsimonious, easily interpretable model that had one of the best test accuracies and the lowest degree of overfitting. Moreover, such a model deemed the most important predictors of our response to be the degree to which they like the other person and the importance they place on attractiveness in a significant other, and, to a much lower extent, the perceived probability of the other person wanting to date them, the importance of shared interest/hobbies, and the importance of how fun the other person is. This report shows that being someone that other people enjoy being around (being a likable person) can boost one's chances of success in dating. Thus, the implications of this work for individuals who are single is that focusing on being likable is their one best shot at improving their chances of going on a date with someone new. However, being likable is quite subjective, and it can range from being outgoing and friendly to being kind and empathetic, and more work should be performed to decrease the subjectivity of the quality.

Even though we ran into some difficulties with regard to data cleaning and transformation, as the original dataset required quite a lot of preprocessing work for it to be ready, we were able to overcome such obstacles relatively easily and fit models that behaved reasonably well. Nevertheless, although we have identified the single decision tree to outperform the other models, this model is far from perfect, and more work is needed to improve the results of this report. Firstly, additional data should be gathered to improve the generalizability of our model. Not only did our data omit certain important variables that play a role in dating environments, such as income or family goals, but the experiment also excluded homosexual couples from the study. We strongly believe that including data on homosexual couples would help us better understand the dating paradigm, and could help us identify interesting differences in terms of feature importance and predictive power between genders. Moreover, the dataset only included information on around 8,000 interactions, so data on more interactions should be included to improve our predictive power. Furthermore, we believe that other variables of interest in the dataset should be evaluated. In this report, we predicted whether a person would want to go on a date with another individual they had just met. However, it would be interesting to study the features that predict not only the decision to go on a date but also the compatibility between individuals, as measured by a match between them. Lastly, we believe that testing out other machine learning algorithms and engaging in more

cross-validation to reduce overfitting would help us fit better models that more accurately predict our response.

Dating is as much of a science as it is an art. Determining which factors drive the compatibility between individuals, however, can help us reduce the uncertainty of dating and can help individuals focus on the right features to improve to boost their chances of finding a suitable romantic partner. Even though more work is needed, the present report adds to the existing literature by providing machine learning models and by identifying the most important features to predict whether a person would want to go on a date with an individual they have just met.

References

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. https://doi.org/10.1037/1528-3542.6.2.269

Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2006). Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment*. *The Quarterly Journal of Economics*, *121*(2), 673–697. https://doi.org/10.1162/qjec.2006.121.2.673

How XGBoost Works—Amazon SageMaker. (n.d.). Retrieved December 1, 2022, from https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

Leventis, D. (2022, January 2). XGBoost Mathematics Explained. *Medium*. https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a

Stacking in Machine Learning. (2019, May 20). *GeeksforGeeks*. https://www.geeksforgeeks.org/stacking-in-machine-learning/

Shaw Taylor, L., Fiore, A. T., Mendelsohn, G. A., & Cheshire, C. (2011). "Out of My League": A Real-World Test of the Matching Hypothesis. *Personality and Social Psychology Bulletin*, *37*(7), 942–954. https://doi.org/10.1177/0146167211409947

Walster, E., Aronson, V., Abrahams, D., & Rottman, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of Personality and Social Psychology*, *4*, 508–516. https://doi.org/10.1037/h0021188

What is XGBoost? (n.d.). NVIDIA Data Science Glossary. Retrieved December 11, 2022, from https://www.nvidia.com/en-us/glossary/data-science/xgboost/

XGBoost Documentation—Xgboost 1.7.1 documentation. (n.d.). Retrieved December 1, 2022, from https://xgboost.readthedocs.io/en/stable/index.html