Predicting Success on a Date

Shashi Roshan, Aastha Nargas, Kusum Vanwani, Pankaj Sharma

16 December, 2019

Project Members

- Aastha Nargas (anargas2)
- Kusum Vanwani (vanwani2)
- Pankaj Sharma (pankajs2)
- Shashi Roshan (sroshan2)

Abstract

Statistical learning models were applied to speed dating data set, in order to predict if the dating partner will decide to go on a second date with the subject. A variety of statistical learning techniques were explored and validated, and ultimately the predictive power of the statistical models appear to be good.

Introduction

Dating is a social meeting where two people assess each other's appropriateness for being a potential partner. If the subject and dating partner liked the first date, then they can decide to go for a second date. In a scenario where the subject has not heard back from their dating partner after the first date, and if the subject doesn't want to follow up, it can lead to waiting times for the subject. We use speed dating data to build statistical models, in order to predict if the dating partner will say Yes for a second date. We will mention this scenario as success on the first date (for the subject). Supervised statistical learning techniques are useful in approximating a function from the data, hence modelling a dependent variable in terms of single/multiple independent variables.

This data set is collected from Kaggle - "Speed Dating Experiment : What attributes influence the selection of a romantic partner?"[^1]. The data consists of details related to 8378 dating events, and whether the dating partner decided to go for a second date. The features include, but is not limited to, the importance of attractiveness, intelligence, sincerity, ambition, fun and shared interests of the subject and their dating partner. It also includes how they rated each other on these qualities, after the date was over. In this analysis, statistical learning techniques will be used on this data set to predict if the dating partner will decide to go on a second date, given a new set of feature variables.

Methods

Data

Data was collected in speed dating experiments among 552 participants during the period 2002-2004. The experiment consisted of multiple speed dating events which were about four minutes long with every other participant of the opposite sex. At the end of each event, the dating partner was asked if they would like a second date with the same person again (the subject). They were also asked to rate their dating mate on the attributes such as Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. Additional features like difference in the key attributes were created based on observation and intuition regarding the importance of these features.

The participants also answered a questionnaire at different time points in the process which included details on demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information. The data set consists of the subject's attributes, the dating partner's attributes, and the decision of the dating partner to go on a second date.

We will use supervised statistical learning techniques to predict success on a dating event - where success means that the dating partner will say Yes to a subject for a second date.

Assumption : The data was collected on speed dating experiments. Real life dates are rarely speed dates. In order to use this data for real life dating scenarios, we need to make the assumption that speed dating is similar to real life dating, and hence the **speed dating data can be applied to real life dating scenario**. We are proceeding with the belief that this assumption is true. However, this needs to be validated.

Modelling

In order to predict success on a date, following statistical learning techniques were used :

- Random Forests via the `RandomForest` package in `R` with OOB error rate as the validation metric was used.

- Gradient Boosting Classifier via the `gbm` package in `R`.
- Logistic Regression via the `glm` package in `R`.

The choice of hyperparameters was found using 10-fold cross validation.
For imputing missing values, we tried KNN imputation method and median imputation method. We considered medianImpute because of its computational efficiency. KnnImpute was also used since we assumed that people with similar attributes will give similar ratings. However, knnImpute is computationally heavy.
The response class is balanced in the training data, with 42% positive responses and 58% negative responses.

Evaluation

To evaluate the performance of different models, we split the data into train and test set. On the train set, we perform 10-fold cross validation to find the optimal set of hyperparameters. To select the best model while performing hyper parameter tuning, Accuracy metric was used. Once all the statistical models' performance were recorded using cross-validation results, we picked the best statistical model. Then, the best statistical model was tested on the test data set.

Results

The table shows the results of predictions on the test data using the statistical methods described in the methods section, with imputation methods like `knnImpute` and `medianImpute`. The best model was Gradient Boosting model (with median Imputation). The models were tuned on Accuracy.

Statistical Models	**Cross Validation Accuracy**	**Model Description**
Random Forest (with KNN imputation)	0.77	mtry = 63
Random Forest (with Median imputation)	0.77	mtry = 63
Random Forest (with no imputation)	0.77	mtry = 124
Gradient Boosting Model (with KNN imputation)	0.77	n.trees = 150, interaction.depth = 3, shrinkage = 0.1
Gradient Boosting Model (with Median imputation)	0.78	n.trees = 150, interaction.depth = 3, shrinkage = 0.1
Gradient Boosting Model (with no imputation)	0.78	n.trees = 100, interaction.depth = 3, shrinkage = 0.1
Logistic Regression (with KNN imputation)	0.76	Kappa = 0.49
Logistic Regression (with Median imputation)	0.77	Kappa = 0.51
Logistic Regression (with no imputation)	0.75	Kappa = 0.49

Discussion

Performance Metric	**Score (on test set)**
No Information Rate	0.59
Accuracy	0.77
Sensitivity	0.81
Specificity	0.71

We found that the statistical models performed really well in case of predicting success on a date. The best model was Gradient Boosting model (with median Imputation). The accuracy on test data set was 0.77. The model is better at predicting success on a date (Sensitivity 0.81), as compared to predicting failure on date (Specificity 0.71). Other statistical models like Random Forest Classifier and Logistic Regression also performed well. We can conclude that the statistical models used in this analysis are providing good results. Future work can include using this data to predict success even before a date has happened.

Also, the assumption that speed dating data is similar to real life dating data needs to be validated. As mentioned, real dates are different from speed dates but the variables used in this analysis will be relevant in a real date as well.

Another limitation of the dataset is that it only considers a very small sample of 552 participants going through multiple speed dates. This group of participants may not be considered the true representative of the larger population. Similar data on a much larger amount of participants which are randomly sampled from different segments of the communities will provide more robust predictions. However, the performance of the models used in this analysis and the choice of the parameters seem promising. This analysis can be treated as a first step towards a bigger research which can be used to revolutionize the dating world, and help people find **The One**.

Appendix

Speed Dating Experiment

[^1] : Speed Dating Experiment (https://www.kaggle.com/annavictoria/speed-dating-experiment)

Data Dictionary

The data set consists of following columns :

Subject's Attributes:

- `age` : Age of the subject
- `gender` : Gender of the subject
- `mn_sat` : The median SAT score of the undergraduate institution attended by the subject (taken from Barron's 25th Edition college profile book)
- `tuition` : Tuition fees of the undergraduate institution attended by the subject (taken from Barron's 25th Edition college profile book)
- `income` : Median household income based on the zipcode of the subject (When there is no income mentioned, it means that either the subject did not enter their zipcode, or they are from abroad)
- `exphappy` : How happy does the subject expect themselves to be with the people they will meet on the speed dating event? (on a scale of 1-10)
- `imprace` : How important is it to the subject that their dating partner is of the same racial/ethnic background (on a scale of 1-10)
- `imprelig` : How important is it to the subject that their dating partner is of the same religious background (on a scale of 1-10)

Subject's stated preference for following attributes (distribution of 100 points among all attributes. 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) :

- `attr1_1` : Importance of attractiveness in a dating partner
- `sinc1_1` : Importance of sincerity in a dating partner
- `intel1_1` : Importance of intelligence in a dating partner
- `fun1_1` : Importance of being funny in a dating partner
- `amb1_1` : Importance of being ambitious in a dating partner

Subject's opinion of their own qualities (on a scale of 1-10) :

- `attr3_1` : How the subject rates himself on attractiveness
- `sinc3_1` : How the subject rates himself on sincerity
- `intel3_1` : How the subject rates himself on intelligence
- `fun3_1` : How the subject rates himself on being funny
- `amb3_1` : How the subject rates himself on being ambitious

How interested is the subject in the following activities (on a scale of 1-10) :

- `sports`
- `tvsports`
- `excersice`
- `dining`
- `museums`
- `art`
- `hiking`
- `gaming`
- `clubbing`
- `reading`
- `tv`
- `theater`
- `movies`
- `concerts`
- `music`
- `shopping`
- `yoga`

Subject's opinion of how other people will rate them on following attributes (on a scale of 1-10, 10 means highly rated) :

- `amb5_1` : Ambition
- `fun5_1` : Fun
- `intel5_1` : Intelligence
- `sinc5_1` : Sincerity
- `attr5_1` : Attractiveness

Subject's division of 100 points among following attributes, where more points are given to the attribute that they think are more important for the members of the opposite sex when they are deciding to date someone :

- `attr2_1` : Attractiveness
- `sinc2_1` : Sincerity
- `intel2_1` : Intelligence
- `fun2_1` : Fun
- `amb2_1` : Ambition
- `shar2_1` : Shared interests

Subject's division of 100 points among following attributes, where more points are given to the attribute that they think are more important for the members of their own sex when they are deciding to date someone :

- `attr4_1` : Attractiveness
- `sinc4_1` : Sincerity

- `intel4_1` : Intelligence
- `fun4_1` : Fun
- `amb4_1` : Ambition
- `shar4_1` : Shared interests

Subject's rating of their dating partner, after the date (on a scale of 1-10) :

- `attr` : How the subject rated their dating partner on attractiveness
- `sinc` : How the subject rated their dating partner on sincerity
- `intel` : How the subject rated their dating partner on intelligence
- `fun` : How the subject rated their dating partner on fun
- `amb` : How the subject rated their dating partner on ambition
- `like` : How much did the subject liked their dating partner? (on a scale of 1-10, where 10 means they liked them a lot)
- `prob` : How probable does the subject believe that their dating partner will say yes for a second date? (on a scale of 1-10, where 10 means highly probable)
- `met` : Has the subject met their dating partner before?

Dating Partner's Attributes :

- `age_o` : Age of the dating partner

Dating partner's stated preference for following attributes (distribution of 100 points among all attributes, where more points should be given to the attribute which they find more important in their dating partner) :

- `pf_o_att` : Importance of attractiveness in their dating partner
- `pf_o_sin` : Importance of sincerity in their dating partner
- `pf_o_int` : Importance of intelligence in their dating partner
- `pf_o_fun` : Importance of being funny in their dating partner
- `pf_o_amb` : Importance of being ambitious in their dating partner

How the dating partner rated the subject on following attributes, after the dating event (on a scale of 1-10) :

- `attr_o` : Attractiveness
- `sinc_o` : Sincerity
- `intel_o` : Intelligence
- `fun_o` : Fun
- `amb_o` : Ambition
- `shar_o` : Shared interests
- `prob_o` : How probable does the dating partner think that the subject will say Yes for a second date (10 = highly probable)
- `met_o` : Has the dating partner met the subject before?

Common Attributes :

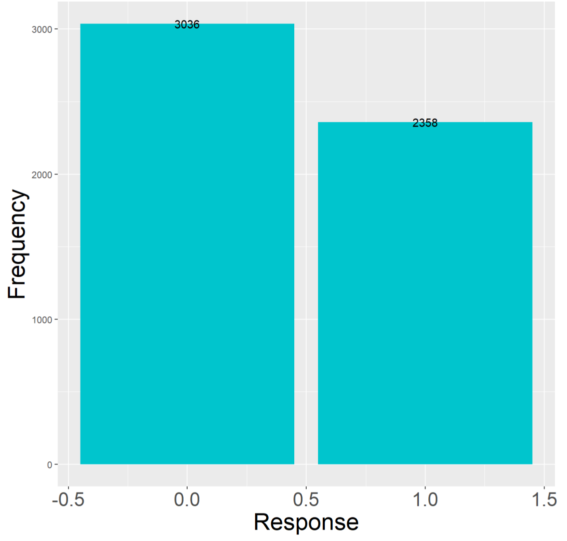
- `samerace` : Are the subject and their dating partner of the same race?
- `int_corr` : Correlation between the subject and the dating partner's interests

Response variable :

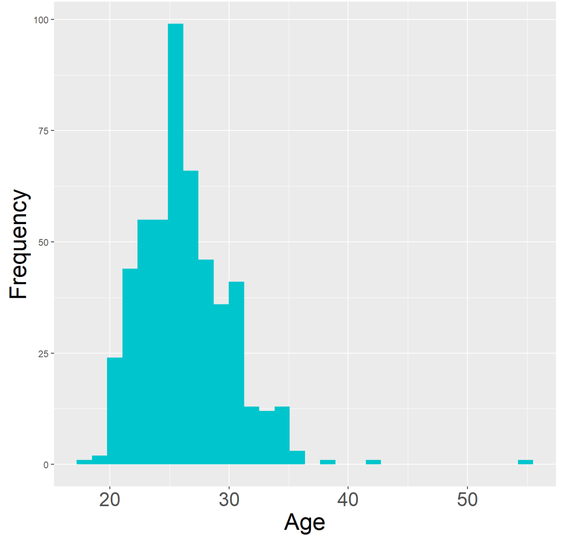
- `dec_o` : Decision of the dating partner to go on a second date with the subject. It can take two possible values : 0 and 1. 1 denotes that the dating partner decided to go on a second date with the subject, and 0 means that they did not.

Exploratory Data Analysis

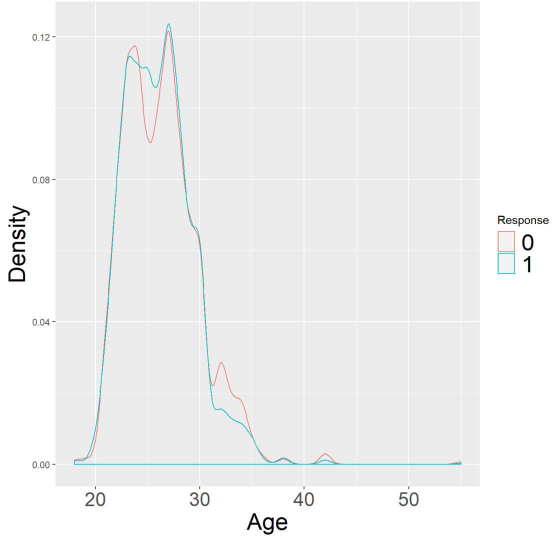
Distribution of Response Variable



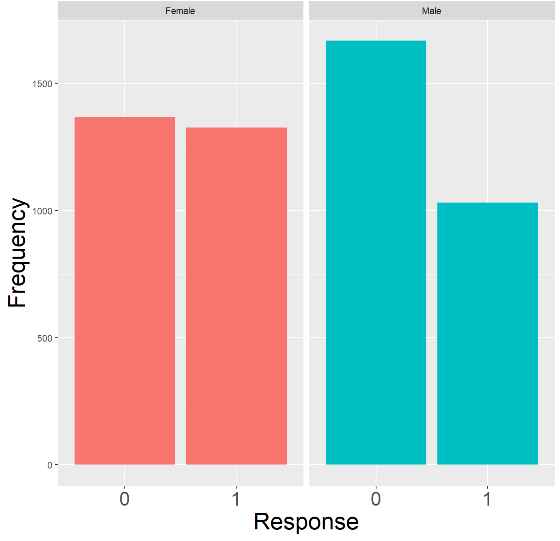
Distribution of Age



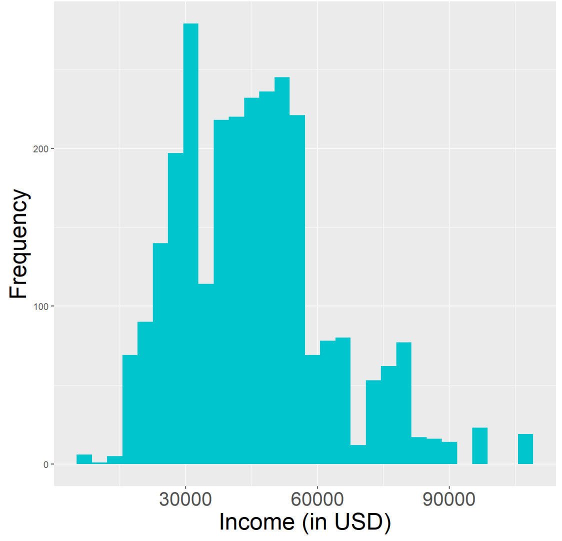
Density of Age



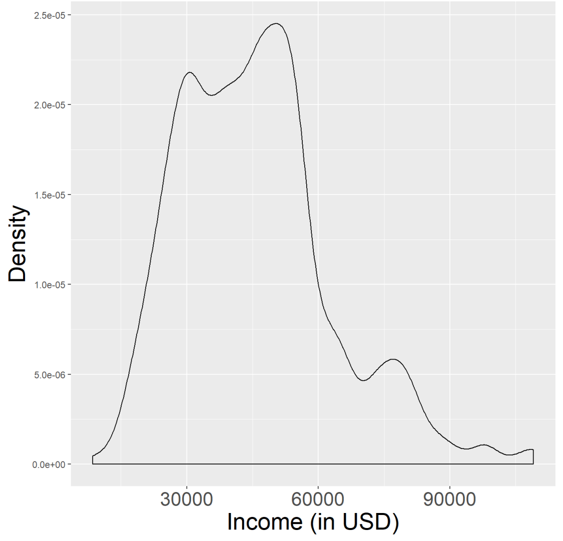
Distribution of Response by Gender



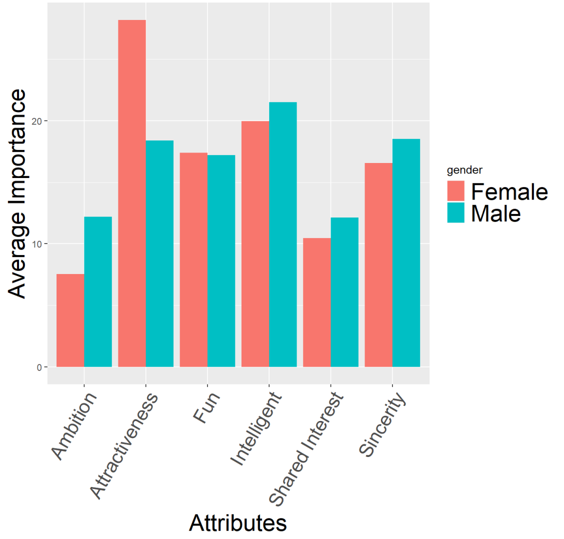
Distribution of Income



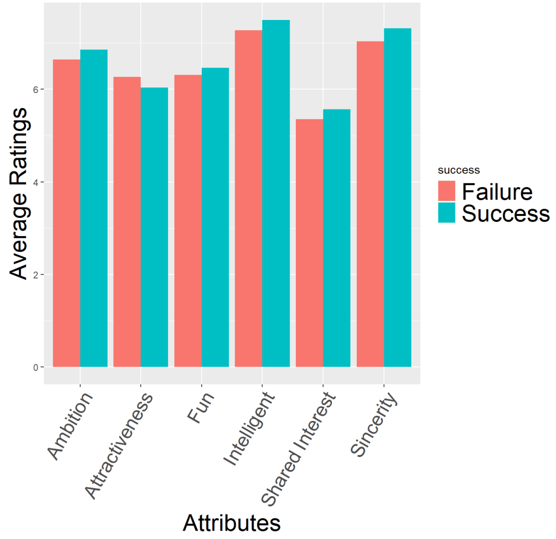
Density of Income



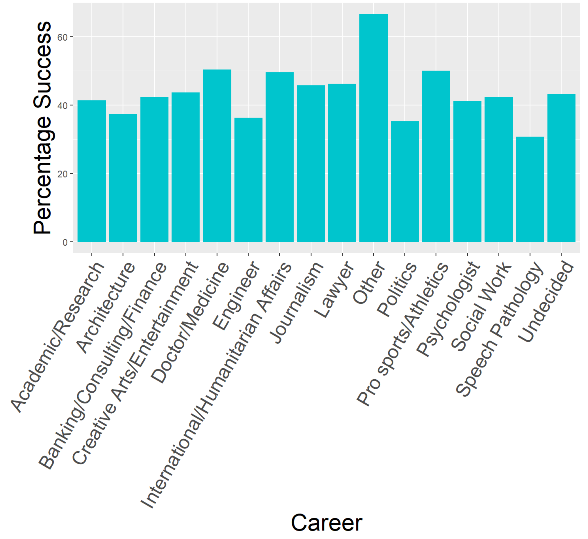
Importance of Attributes by Gender



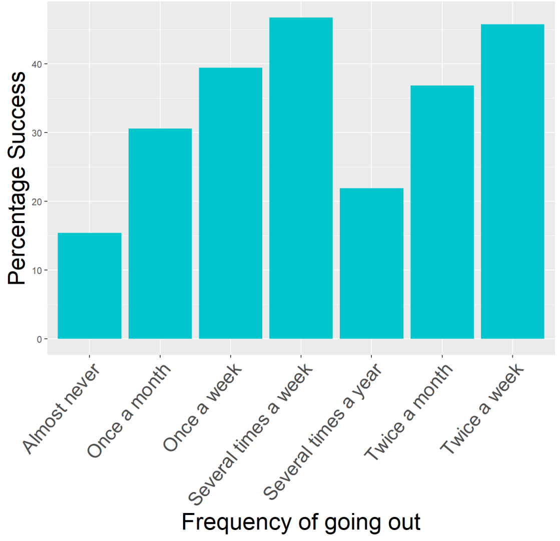
Rating of Attributes after a date



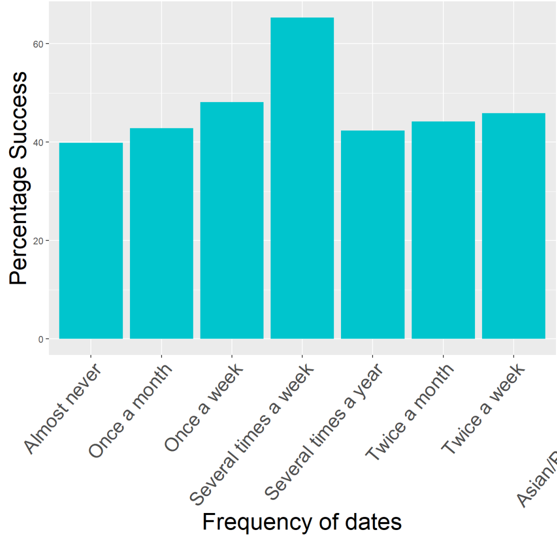
Success of Date based on career



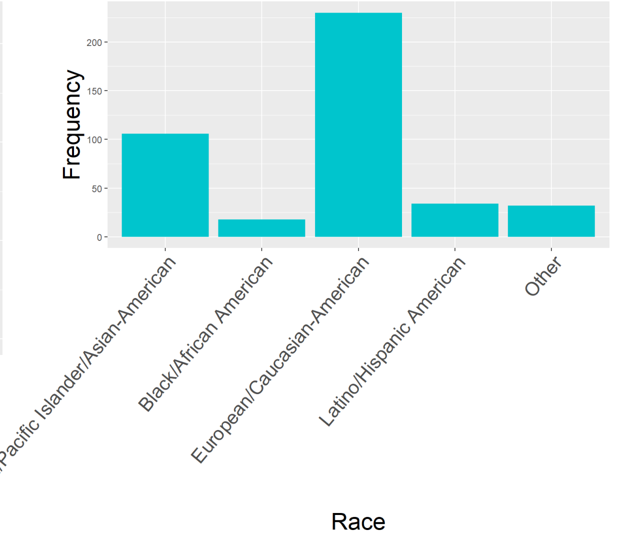
Success of Date based on frequency of going



Success of Date based on frequency of dates



Distribution of Race


Frequency
Asian/Pacific Islander/Asian-American
Black/African American
European/Caucasian-American
Latino/Hispanic American
Other
Race