Speed Dating

01:31:59 AM 03 May, 2020

The purpose of the project is to analyze and identify the influential attributes in determining to meet someone again romantically after a speed dating event. In simple terms, "What influences love at first sight?" (Or , at least, love at the first four minutes?)

Speed Dating Experiment:

The dataset used for this experiment is gathered by Columbia Business school Professors Ray Fisman and Sheena Iyengar for their paper Gender Differences in Mate Selection: Evidence from a speed dating experiment. The dates lasted four minutes and a survey form is filled out to capture required personal information and dating experience. The participants are required to rate the importance of the following attributes in a potential data on a scale of 1-10((1=not at all important, 10=extremely important). Attractiveness, Sincerity, Intelligence, Fun, Ambitious, Shared interests/hobbies. The Questionnaire data is gathered at a different point in time during the experiment. Firstly at the time of Sign up,then the day after participating in the event, finally 3-4 weeks after they had been sent their matches.

Dataset & Variable Selection:

Each date will have 2 participants. p1 -participant 1, p2 participant2. There are 8378 observations and 195 variables in summary. Below are the key variables taken for study in this project.

attr : how attractive p1 thinks p2 is

sinc : how sincere p1 thinks p2 is

intel : how smart p1 thinks p2 is

fun : how fun p1 thinks p2 is

amb : how ambitious p1 thinks p2 is

shar : how much p1 believes they both (p1 and p2) share the same interests and hobbies

dec : whether p1 wants to meet p2 again given how the speed date went.

gender : gender of p1, 0 = woman

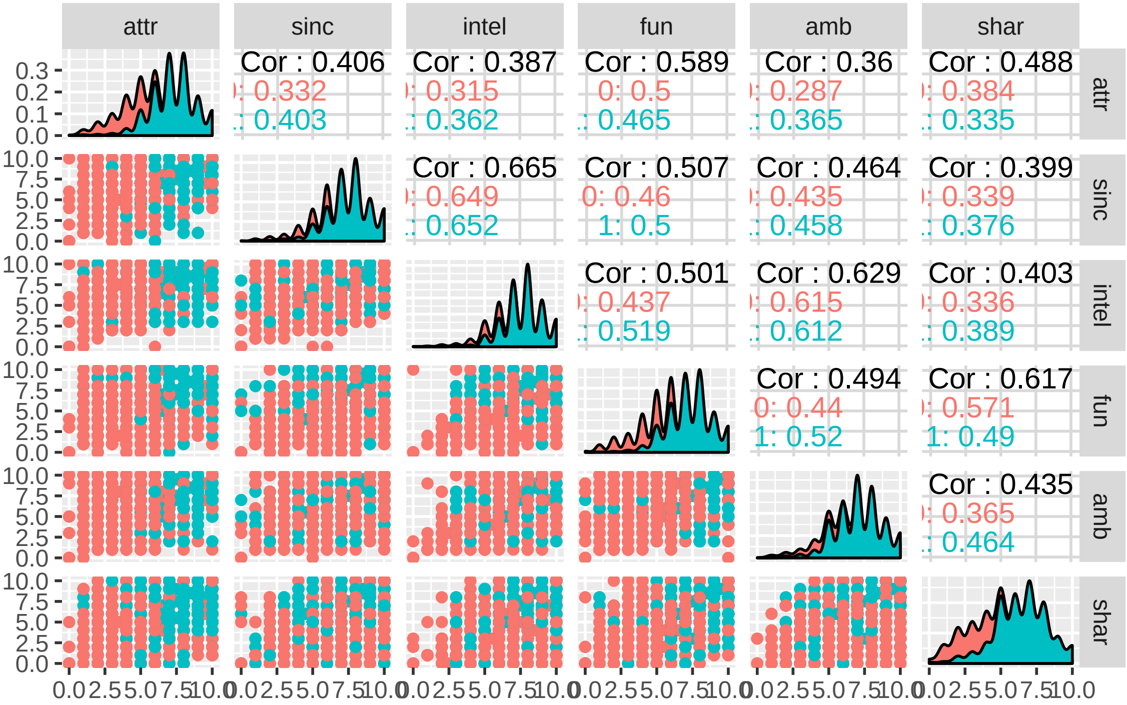
Data Slicing and Cleaning:

A subset of data with six influential attributes as predictor variables and decision attribute as a response variable is created from the main dataset. It is observed that some of the attribute values are NA and so rows having NA's are omitted.

```
req_columns_s <- c("dec", "attr", "sinc", "intel", "fun", "amb", "shar")
req_speed_dating_s <- speed_dating[,req_columns_s]
req_speed_dating_s <- na.omit(req_speed_dating_s)
req_speed_dating_s$dec <- as.factor(req_speed_dating_s$dec)
```

Collinearity Test: In this section, Ggpairs plot is plotted to check the relationship and existence of multi-collinearity among the predictor variables.

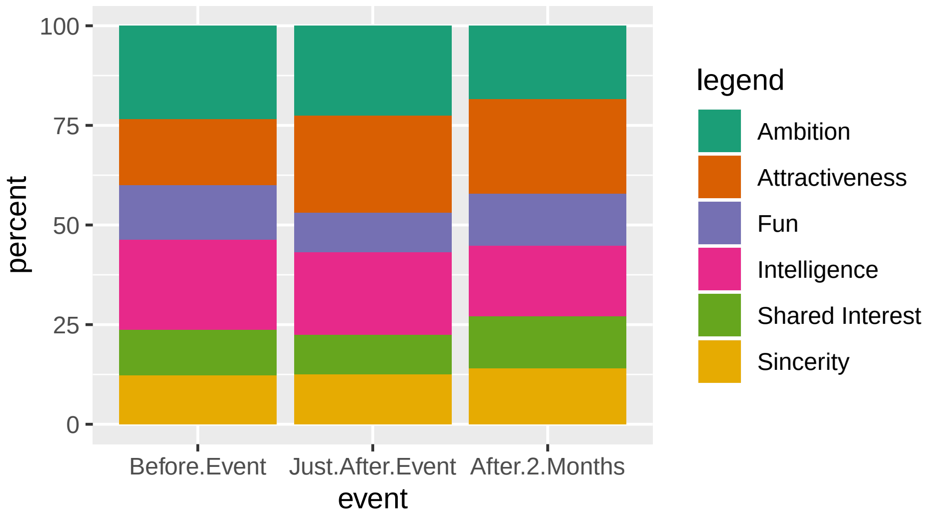
```
ggpairs(req_speed_dating_s,columns=2:7,aes(col=dec))
```



From the above plot there are no strong correlations among predictors is observed. Intelligence and sincerity has slightly high correlation ratio of around 66% but still not too high. Hence no action necessary to remove collinearity.

Overview of Data:

```r
## SDD<-read.csv("speed_dating_data.csv")
SDD_attributes<-speed_dating %>% select(attr1_1,sinc1_1,intel1_1,fun1_1,amb1_1,shar1_1,
                                attr7_2,sinc7_2,intel7_2,fun7_2,amb7_2,shar7_2,
                                attr7_3,sinc7_3,intel7_3,fun7_3,amb7_3,shar7_3)

SDD_attributes<-na.omit(SDD_attributes)
sum_data<-data.frame(value=apply(SDD_attributes,2,sum))

data <- matrix(sum_data$value,nrow=3)

colnames(data) <- c('Attractiveness','Sincerity','Intelligence','Fun','Ambition','Shared Interest')

rownames(data) <- factor(c("Before Event","Just After Event","After 2 Months"))

percent_data<-data.frame(apply(data, 1, function(x){x*100/sum(x,na.rm=T)}))

percent_data$attribute<-factor(rownames(percent_data))

percent_data<-gather(percent_data,Before.Event,Just.After.Event,After.2.Months,key=event,value=percent)

percent_data$event<-factor(percent_data$event,levels = c("Before.Event", "Just.After.Event", "After.2.Mo

ggplot(percent_data,mapping = aes(x=event,y=percent,fill=attribute))+
  geom_col()+
  scale_fill_manual("legend", values = brewer.pal(n = 6, name = "Dark2"))
```



The above graph indicates how participants perceived the 6 attributes at different timelines of the Dating event. From the graph we can see that there was a clear decrease in the importance of Intelligence. The participants gave a lower importance to Attractiveness before the event, compared to the later stages. As seen from the graph, after 2 months, the importance of Ambition reduced from the earlier phases.Overall, Attractiveness, Intelligence and Ambition dominated among all the factors that are perceived while going on a date

Analysis: As the problem Question is to select the best attribute has influence over decision making of the speed dating event, ML feature selection methods are referred. We have tried forward selection and stepwise regression. But the AIC scores are relatively close to each other and so we decided to analyse the give dataset with Logistic regression and Random Forest algorithms.

Logistic Regression:

Here, the dataset is partioned into training dataset with 50% of the dataset and testing dataset with the rest of the data for the purpose of cross validation. The data in the training and testing dataset are randomly selected.

```
set.seed(1)
s <- sample(nrow(req_speed_dating_s), round(.5*nrow(req_speed_dating_s)))
train_req_speed_dating_s <- req_speed_dating_s[s,]
test_req_speed_dating_s <- req_speed_dating_s[-s,]
```

Now a logistic regression model is created using the training dataset.

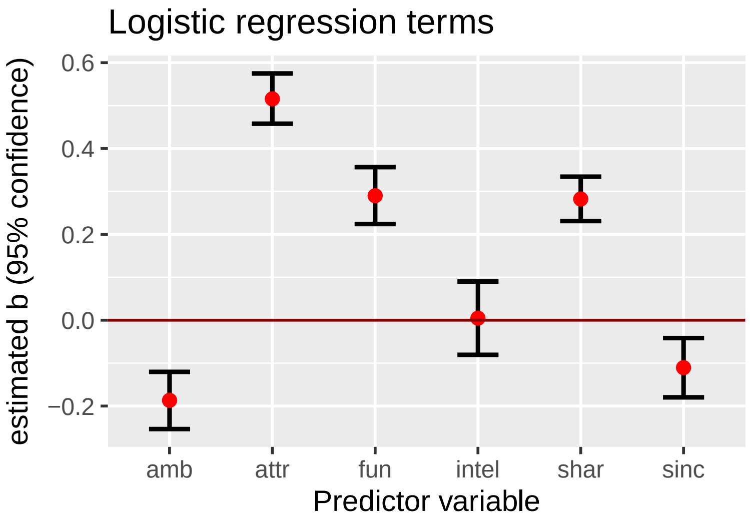
```
log_reg_model_s <- glm(dec ~ ., data = train_req_speed_dating_s, family = binomial("logit"),maxit = 100
summary(log_reg_model_s)$coefficients
```

```
##                 Estimate Std. Error     z value     Pr(>|z|)
## (Intercept) -5.033048247 0.25921600 -19.4164258 5.605825e-84
## attr         0.515675246 0.02984577  17.2779996 6.889637e-67
## sinc        -0.110541328 0.03520751  -3.1397088 1.691158e-03
## intel        0.004759121 0.04358059   0.1092028 9.130417e-01
## fun          0.290004886 0.03381793   8.5754768 9.867714e-18
## amb         -0.186700631 0.03398938  -5.4929107 3.953627e-08
## shar         0.282388417 0.02636882  10.7091796 9.217506e-27
```

From the model(full model) fit we can see that attr, fun, amb and shar is very significant parameters for making a decision followed by sinc whereas intel is not significant at all. From the model fit we can see that sinc and amb has negative impact on our response i.e dec whereas attr,fun and shar are having positive impact on dec.

The most influencing predictor can be identified by calculating the confidence interval. The calculated confidence intervals for all the predictors are visulaized below.

```
broom::tidy(log_reg_model_s,conf.int = TRUE,conf.level = 0.95) %>%
  filter(term != "(Intercept)") %>%
  ggplot(aes(term, estimate,ymin = conf.low,ymax = conf.high)) +
  geom_errorbar(size = 0.8, width= 0.4) +
  geom_point(color = "red", size = 2) +
  geom_hline(yintercept = 0, colour = "darkred") +
  labs(x = "Predictor variable",title = "Logistic regression terms",y = expression(paste("estimated ",
```



From the above plot, we can clearly say that the predictor attr is having the highest positive confidence interval than other predictors.

Model Predictions Using Logistic Regression: Using the logistic regression model which was created in the previous section, it is possible to predict a match between two individals with these six predictor variable values.

Prediction with Training Dataset:

```
pred_s <- predict(log_reg_model_s, train_req_speed_dating_s, type="response")
train_bin_preds_s <- as.numeric(pred_s >= 0.5)
train_accuracy_s <- mean(train_bin_preds_s == train_req_speed_dating_s["dec"])
train_accuracy_s
```

```
## [1] 0.7488636
```

Prediction with Test Dataset:

```
test_preds_s <- predict(log_reg_model_s, test_req_speed_dating_s, type="response")
test_bin_preds_s <- as.numeric(test_preds_s >= 0.5)
test_accuracy_s <- mean(test_bin_preds_s==test_req_speed_dating_s["dec"])
test_accuracy_s
```

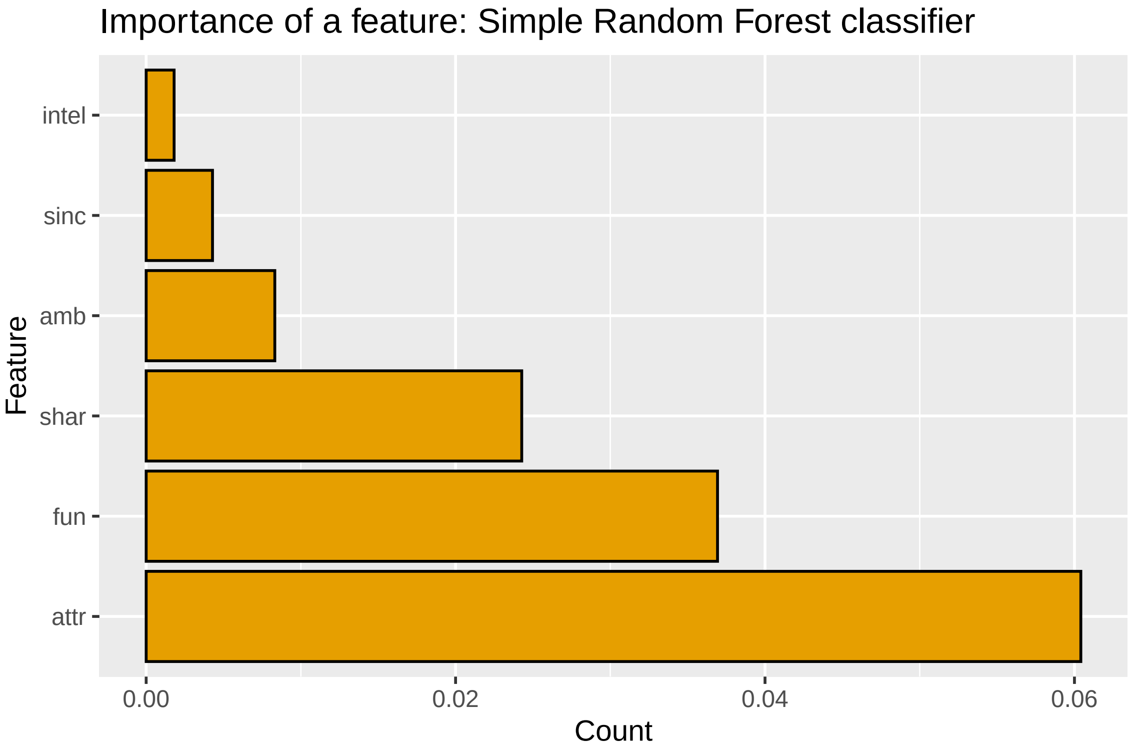
```
## [1] 0.7480114
```

This Logistic regression model achieved a 74.88% and 74.80% with training and test dataset respectively.

Random Forest: Random Forest with built in option in feature selection produces a clear output which is plotted using ggplot as below.

```
fit <- randomForest(dec ~ attr+sinc+intel+fun+amb+shar,
          data = req_speed_dating_s,importance=TRUE,ntree=600)
```

```r
importance.features <- tibble::rownames_to_column(data.frame(fit$importance[,c(1)]))
colnames(importance.features) <- c("rowname", "value")

ggplot(importance.features, aes(x = reorder(rowname, -value), y = value)) +
  geom_bar(stat = "identity", position = "dodge", fill="#E69F00", colour="black") +
  xlab("Feature") + ylab("Count") + ggtitle("Importance of a feature: Simple Random Forest classifier")
  coord_flip()
```



From the above graph, it is observed that the predictor attraction followed by fun are having the most influence over the response variable and the predictors Intelligence is having least influence than other variables.

Conclusion: Contradicting the famous saying “No beauty shines brighter than that of a good heart”, Attractiveness drove participants’ decision to select / reject their partner the most. Both Logistics and Random Forest showed us that Attractiveness was the most influential factor, whereas Intelligence and Sincerity were rated the lowest. From the initial Perception graph, which was taken during 3 timelines, Intelligence was in the top 3, but when it comes to making actual decision for selecting a partner, Intelligence was the least important.

References : RAYMOND, FISMAN. SHEENA, S. IYENGAR.,EMIR, KAMENICA.,ITAMAR. SIMONSON.(2008) “Racial Preferences in Dating”, Review of Economic Studies (2008) 75, 117–132 COLIN. LEVERGER.(2016) “Exploring Speed Dating”, https://colinleverger.github.io/speed-dating-experiment-r/ ANNA. MONTOYA.(2016) “Speed Dating Experiment”, https://data.world/annavmontoya/speed-dating-experiment