4/11/2016

OPIM 5604 Predictive Modeling

Final Project – Group 5

Professor Cruz

NAVDEEP CHEEMA
RAVI KOLLA
JENNIFER NIELSEN
YIFEI TANG

Table of Contents

Executive Summary

The dating industry has grown rapidly over the past decade and is comprised of many players. Services such as telephone chat lines, singles bars, speed dating organizations, online dating, mobile dating, and others have shaped the industry into what it is now and faded the stigma associated with using dating services. The U.S. single population is estimated to be close to 100 million people (The Business of Dating). An increasing number of these singles are turning to dating services to find love, and it is working.

Satisfied customers are not pushing their friends to join the newest app in the market, but rather, to the one they had the most luck with. Therefore, reputation is one critical success factor. Dating services are rising their level of concentration to more effectively match customers to potential dates and attract new customers.

Our group project explored the dataset from a speed dating experiment facilitated by Columbia University. Specifically, we observed the perceived importance of 6 attributes and their impact on the subject's decision to go on a date. The purpose of comparison of these attributes can aid in the matching process for any dating service, not just speed dating. Additionally, there is significance in collecting useful data before matching to improve business performance and to provide insight into what men and women are looking for in a partner. For instance, it can be argued that customers who both equally value ambition in a partner are more suited for each other. If this data can be collected in advance, companies can more efficiently create matches. Dating service providers can strengthen their reputation using similar techniques to create successful matches.

Following the SEMMA approach, the speed dating data was first partitioned and variables were defined. We explored the data using various visualization techniques to discover patterns and abnormalities among the variables (e.g. correlations, outliers, missing data). Because the dataset was

already cleaned, no modification steps were performed. Logistical regression, decision tree, and neural network models were then created to predict the subject's likelihood of accepting or rejecting a date with a potential mate from the speed dating event. Finally, the models were measured and evaluated for usefulness. It was decided that the logistical regression model was the most optimal due to its simplicity and efficiency.

Objectives

- Design and implement a model that would most accurately predict a subject's decision to go on a date based on his/her attribute preference in a partner
- Examine and report on patterns of gender behavior differences when dating
- Explore trends to inform stakeholders how to market more effectively for client acquisition and retention
- Discover potential additional business opportunities for dating service providers
- Provide recommendations that could make the speed date process more effective

Data Introduction

Speed Dating Event

The experimental speed dating exercise took place for 2 years from 2002 to 2004 as part of a Columbia University study. Students were recruited to be subjects and completed a pre-event survey online. In this survey, they were asked to rate their preference for 6 attributes (intelligence, attractiveness, sincerity, fun, shared interests, and ambition) in a partner.

During the speed dating event, subjects met with a potential mate for 4 minutes and filled out a scorecard afterwards. This scorecard indicated each subject's decision to either accept or reject a date with who they just met with. This decision became our response variable in the modeling techniques. Subjects were also asked to rate their partner on those six attributes on the scorecard.

Dataset

The dataset contains 194 variables, which contain the data for the speed dating event. The data has the participant’s name, age, address, profession, race, preferences and the attributes they look for in partner. For this project, we’ve decided to not include all the variables, and work only on what attributes have the most significance for male and female participants. The team has decided to use the following six attributes that were most significant to choose the partner:

- Attractive (Attr)
- Sincere (Sinc)
- Intelligent (Intel)
- Fun (Fun)
- Ambitious (Amb)
- Shared Interests (Shar)

**please take note of the abbreviations which are referenced throughout the paper*

The decision variable is present in the dataset which tells us whether the participant is willing to go on another date with the partner.

iid	unique subject number, group (wave id gender)
dec	Participant's decision which tells us whether the participant is willing to go on another date with the current date. 1-Yes, 0-No
attr	Attractiveness of the partner. Rated on a scale of 1-10. (1 being the least, and 10 being the highest)
sinc	Describes how sincere the partner is. Rated on a scale of 1-10. (1 being the least, and 10 being the highest)
intel	Intelligence of the partner. Rated on a scale of 1-10. (1 being the least, and 10 being the highest)
fun	Fun factor of the partner. Rated on a scale of 1-10. (1 being the least, and 10 being the highest)
amb	ambitiousness of the partner. Rated on a scale of 1-10. (1 being the least, and 10 being the highest)
shar	shred interests of the partner. Rated on a scale of 1-10. (1 being the least, and 10 being the highest)
gender	Female=0 Male=1

Data Analysis

Sample

The dataset is a sample of 8378 rows of data from the speed dating event, where each row contains the data for individual decisions. We have selected 6 explanatory variables (attributes mentioned above) and 1 decision variable (see Figure 1). The data has been partitioned into training and validation sets, based on the “iid” column (uniquely identifies each participant in the experiment) and used the “stratified random” technique.

Explore

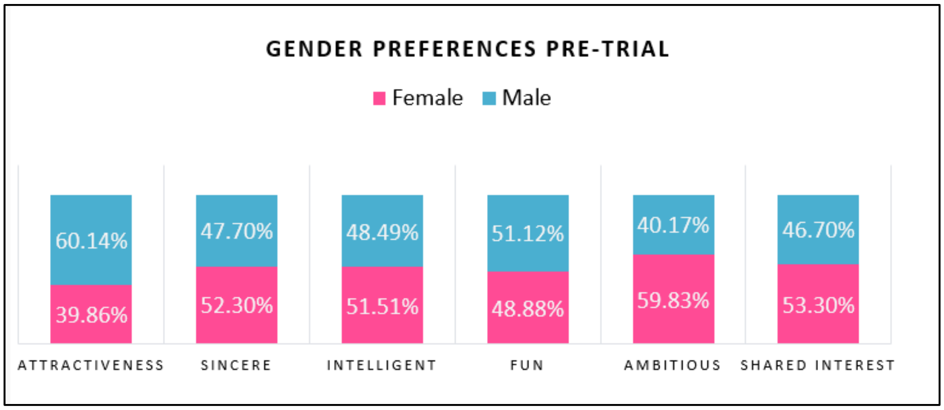
In our dataset, the attributes are all continuous variables with the values ranging from 1-10. The decision variable and the gender are ordinal values, with values 1 or 0. The following exploring steps were performed:

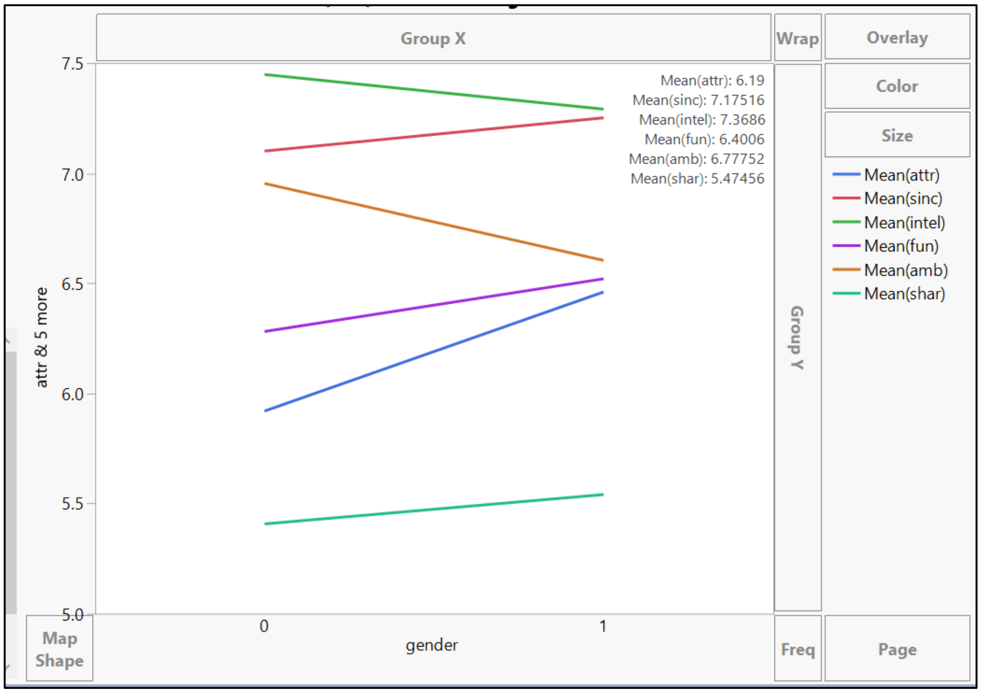
- Began exploring the data for missing values (see Figure 2). Dataset has 8378 rows, of which 1338 rows have missing values. The missing data accounts for 15.9% of the data, but the team decided to leave the missing values as they are. These missing values belong to participants and it only makes sense to leave them blank, as one can’t simply attribute one person’s behavior to another person.

- Observed that there were no inconsistent values. Also, the dataset is fairly balanced, in terms of male and female participation (see Figure 3).

- Analyzed the data for outliers to see if there are any extreme values for the behavioral attributes. The boxplots in Figure 4 suggests that there might be some outliers in the dataset, but in reality, these values are bounded to a defined scale of 1-10.

- Performed “Mahalanobis outlier analysis” to see if the 2 methods tell a different story (see Figure 5).
- Searching for a potential reduction in dimensionality, the team decided it would be interesting to look for correlation between all the variables. When the correlation was performed (see Figure 6), we found that there is no negative correlation among the variables. It makes sense because these are behavioral attributes, and there wouldn’t be any scenario where one attribute would cause the other attribute to go down. (ex: A highly intelligent person is not so attractive). These attributes are unique for each person and we can’t predict a definite pattern. With not so strong correlation among these variables, performing principal component analysis, doesn’t add much value to model and it doesn’t call for the reduction of dimensionality.
- A quick review of the data provides some interesting insights. First, we report the valuation of attributes by men and women. Women put greater weight on intelligence than men do, while men place more value on physical appearance. We observe that a man’s demand for intelligence and ambition does not extend to women who are more intelligent or ambitious than he is. In fact, a man is significantly less likely to accept a woman who is more ambitious than he. Also, women put more emphasis on their partner’s ambitiousness, while men are more concerned about the sincerity and the fun factor when choosing the partner. The common attribute where men and women both put similar value is on the similar interests. The visual on the left and below illustrates these differences.





Modify

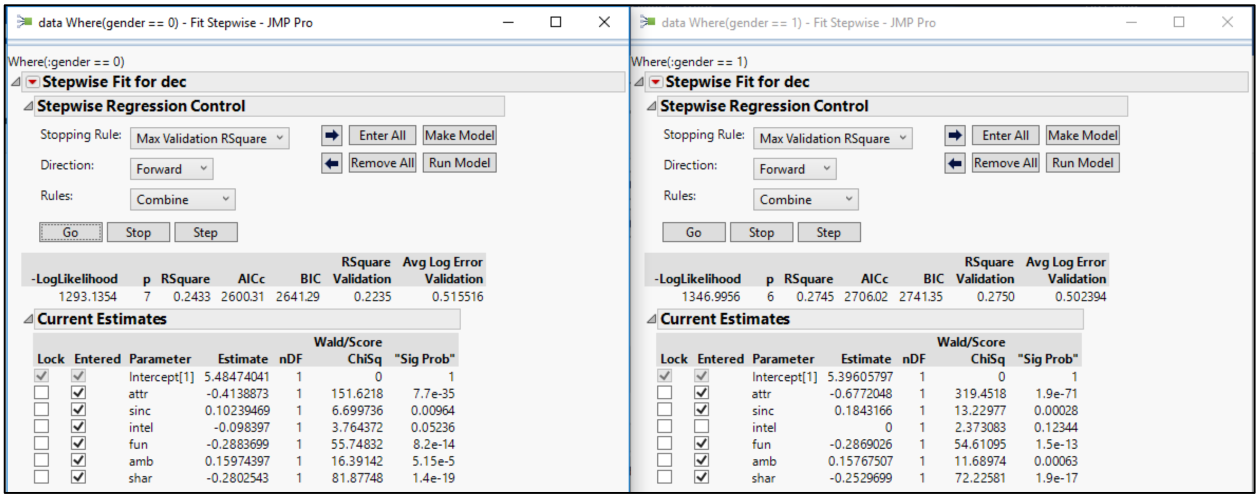
There was not an opportunity to modify this data in the project, although for future exploration and recommendations one suggestion is to combine data set with additional age groups to provide a more comprehensive case comparison.

Model

In the model analysis, we established three models: logistic regression model, decision tree and neural network, and selected the most favorable by comparing simplicity and efficiency of these three models.

Logistical Regression

Ran the Stepwise for female and male respectively, refer to images below.



data Where(gender == 0) - Fit Stepwise - JMP Pro

Where(:gender == 0)

Stepwise Fit for dec

Stepwise Regression Control

Stopping Rule: Max Validation RSquare

Direction: Forward

Rules: Combine

-LogLikelihood	p	RSquare	AICc	BIC	RSquare Validation	Avg Log Error Validation
1293.1354	7	0.2433	2600.31	2641.29	0.2235	0.515516

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
☑	☑	Intercept[1]	5.48474041	1	0	1
□	☑	attr	-0.4138873	1	151.6218	7.7e-35
□	☑	sinc	0.10239469	1	6.699736	0.00964
□	☑	intel	-0.098397	1	3.764372	0.05236
□	☑	fun	-0.2883699	1	55.74832	8.2e-14
□	☑	amb	0.15974397	1	16.39142	5.15e-5
□	☑	shar	-0.2802543	1	81.87748	1.4e-19

data Where(gender == 1) - Fit Stepwise - JMP Pro

Where(:gender == 1)

Stepwise Fit for dec

Stepwise Regression Control

Stopping Rule: Max Validation RSquare

Direction: Forward

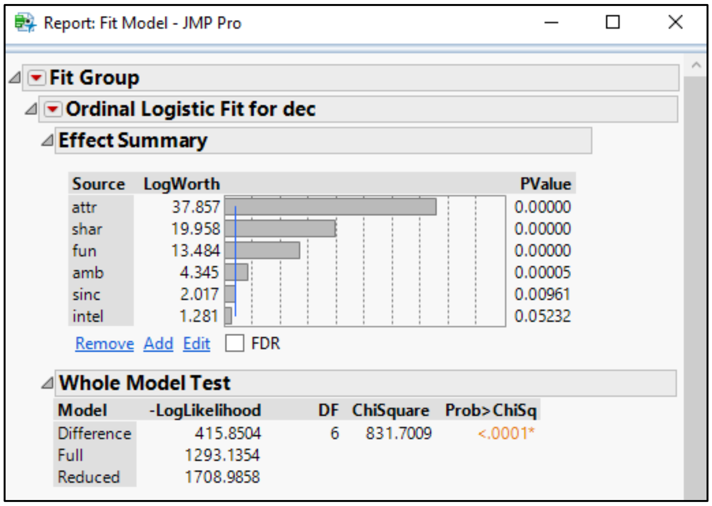
Rules: Combine

-LogLikelihood	p	RSquare	AICc	BIC	RSquare Validation	Avg Log Error Validation
1346.9956	6	0.2745	2706.02	2741.35	0.2750	0.502394

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
☑	☑	Intercept[1]	5.39605797	1	0	1
□	☑	attr	-0.6772048	1	319.4518	1.9e-71
□	☑	sinc	0.1843166	1	13.22977	0.00028
□	□	intel	0	1	2.373083	0.12344
□	☑	fun	-0.2869026	1	54.61095	1.5e-13
□	☑	amb	0.15767507	1	11.68974	0.00063
□	☑	shar	-0.2529699	1	72.22581	1.9e-17

For female, all the six attributes are chosen, and for male, intelligence is excluded by Stepwise. We ran the logistic regression model for female to see the contribution of each attribute.



Report: Fit Model - JMP Pro

Fit Group

Ordinal Logistic Fit for dec

Effect Summary

Source	LogWorth	PValue
attr	37.857	0.00000
shar	19.958	0.00000
fun	13.484	0.00000
amb	4.345	0.00005
sinc	2.017	0.00961
intel	1.281	0.05232

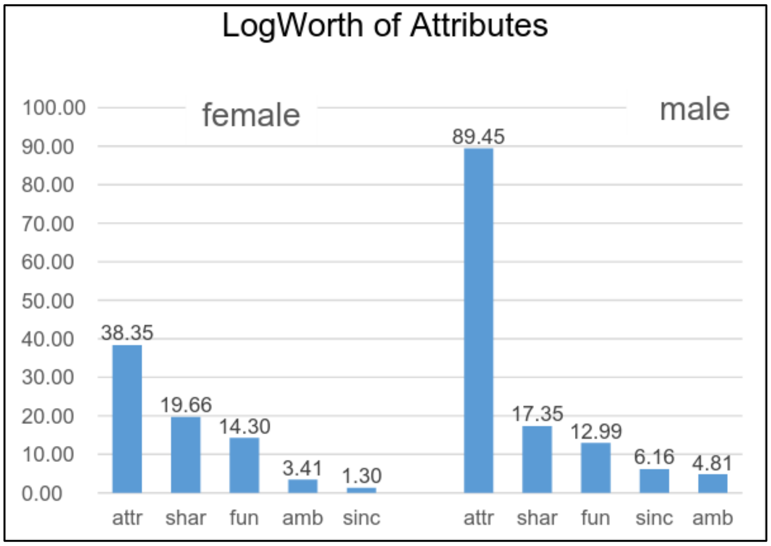
Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	415.8504	6	831.7009	<.0001*
Full	1293.1354			
Reduced	1708.9858			

As can be seen from the report, among all the variables intelligence is the least significant one and is insignificant at 0.01 level.

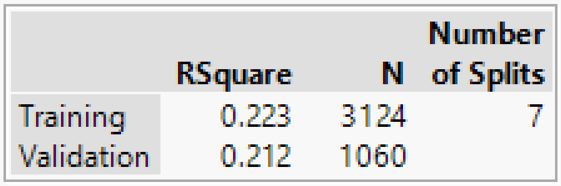
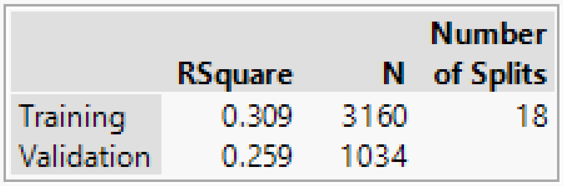
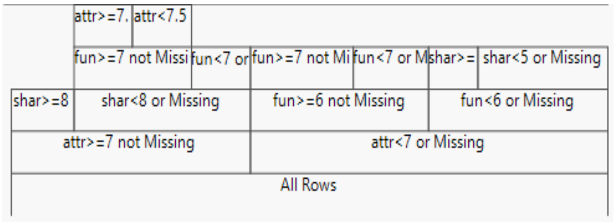
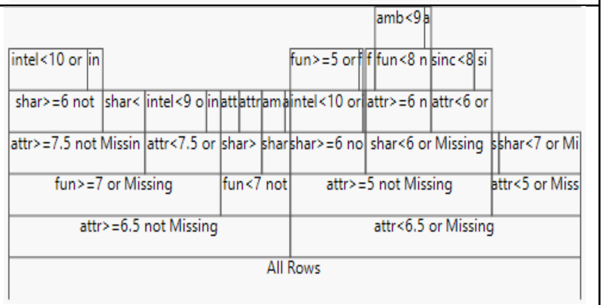
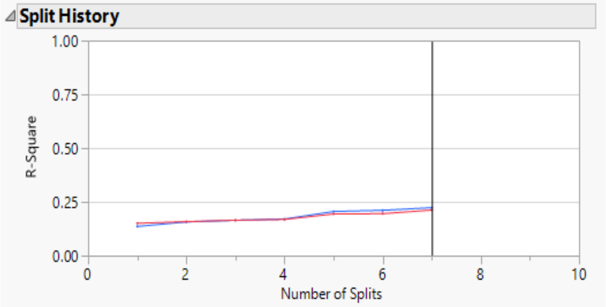
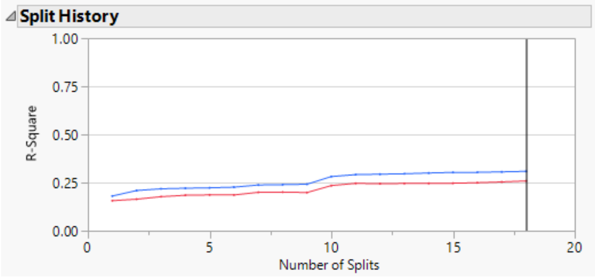
As a result, we decide to choose five attributes to fit the model for both male and female: attractiveness, shared interests, fun, ambition and sincerity.

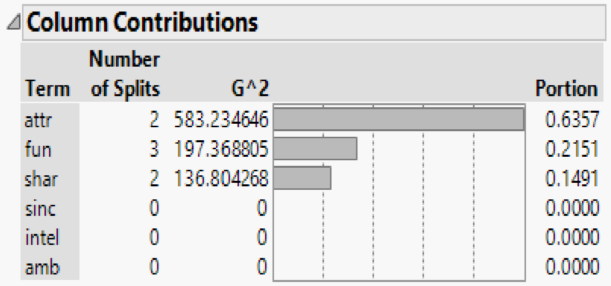
Both female and male care more about attractive while they decide to accept or decline the date. One noteworthy thing is that males pay much more attention to attractiveness than females (see chart on right).



Decision Tree

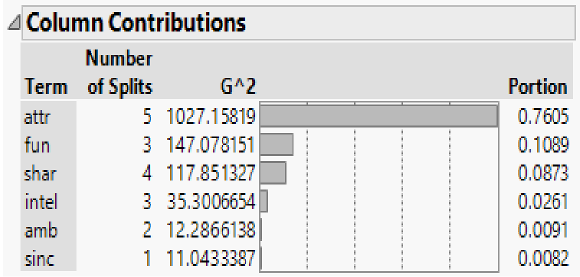
Following the logistical regression, a decision tree model was built. The below table summarizes the key measurements from this model:

<table>
<tr><th>Female</th><th>Male</th></tr>
<tr><td>
<table>
<tr><th></th><th>RSquare</th><th>N</th><th>Number of Splits</th></tr>
<tr><td>Training</td><td>0.223</td><td>3124</td><td>7</td></tr>
<tr><td>Validation</td><td>0.212</td><td>1060</td><td></td></tr>
</table></td>
<td>
<table>
<tr><th></th><th>RSquare</th><th>N</th><th>Number of Splits</th></tr>
<tr><td>Training</td><td>0.309</td><td>3160</td><td>18</td></tr>
<tr><td>Validation</td><td>0.259</td><td>1034</td><td></td></tr>
</table></td></tr>
<tr><td>
</td>
<td>
</td></tr>
<tr><td>
</td>
<td>
</td></tr>
</table>



Column Contributions

Term	Number of Splits	G^2	Portion
attr	2	583.234646	0.6357
fun	3	197.368805	0.2151
shar	2	136.804268	0.1491
sinc	0	0	0.0000
intel	0	0	0.0000
amb	0	0	0.0000



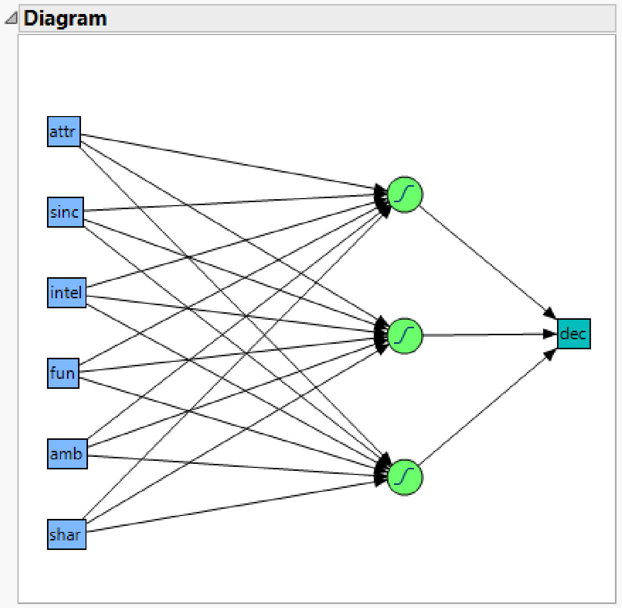
Column Contributions

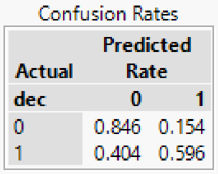
Term	Number of Splits	G^2	Portion
attr	5	1027.15819	0.7605
fun	3	147.078151	0.1089
shar	4	117.851327	0.0873
intel	3	35.3006654	0.0261
amb	2	12.2866138	0.0091
sinc	1	11.0433387	0.0082

Neural Network

A neural network model was also built, with the following results:

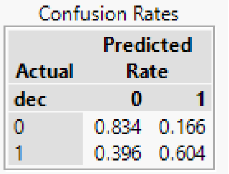
Female	Male





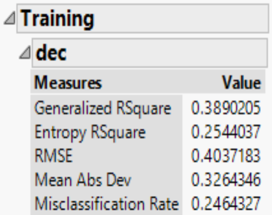
Confusion Rates

Actual	Predicted Rate	
dec	0	1
0	0.846	0.154
1	0.404	0.596



Confusion Rates

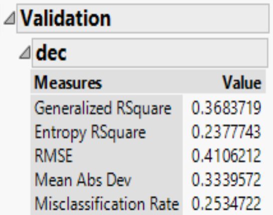
Actual	Predicted Rate	
dec	0	1
0	0.834	0.166
1	0.396	0.604



Training

dec

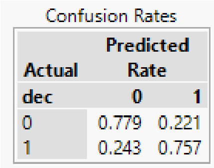
Measures	Value
Generalized RSquare	0.3890205
Entropy RSquare	0.2544037
RMSE	0.4037183
Mean Abs Dev	0.3264346
Misclassification Rate	0.2464327



Validation

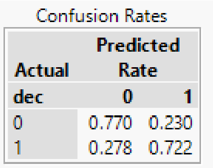
dec

Measures	Value
Generalized RSquare	0.3683719
Entropy RSquare	0.2377743
RMSE	0.4106212
Mean Abs Dev	0.3339572
Misclassification Rate	0.2534722



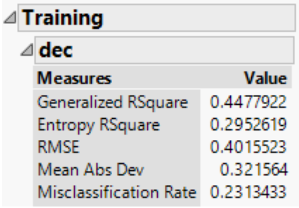
Confusion Rates

Actual	Predicted Rate	
dec	0	1
0	0.779	0.221
1	0.243	0.757



Confusion Rates

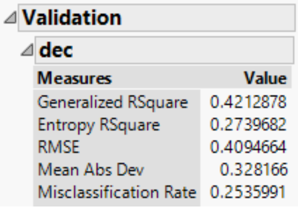
Actual	Predicted Rate	
dec	0	1
0	0.770	0.230
1	0.278	0.722



Training

dec

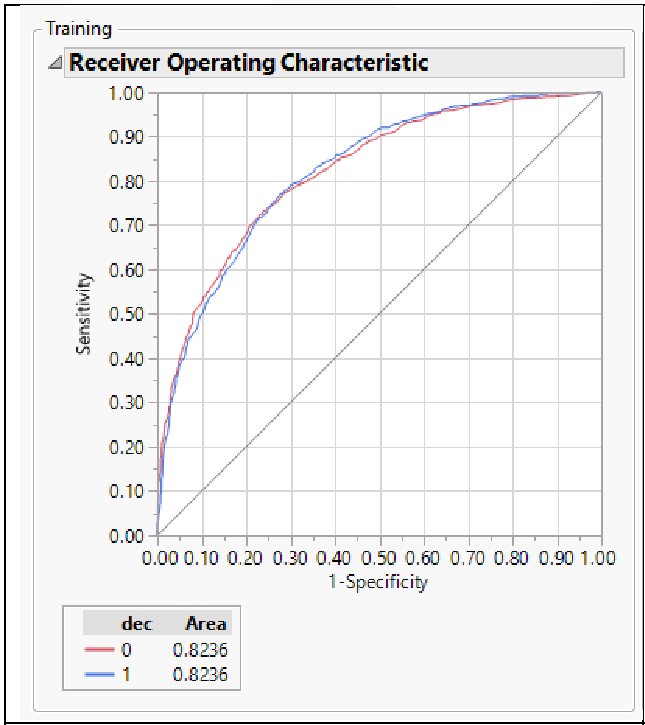
Measures	Value
Generalized RSquare	0.4477922
Entropy RSquare	0.2952619
RMSE	0.4015523
Mean Abs Dev	0.321564
Misclassification Rate	0.2313433



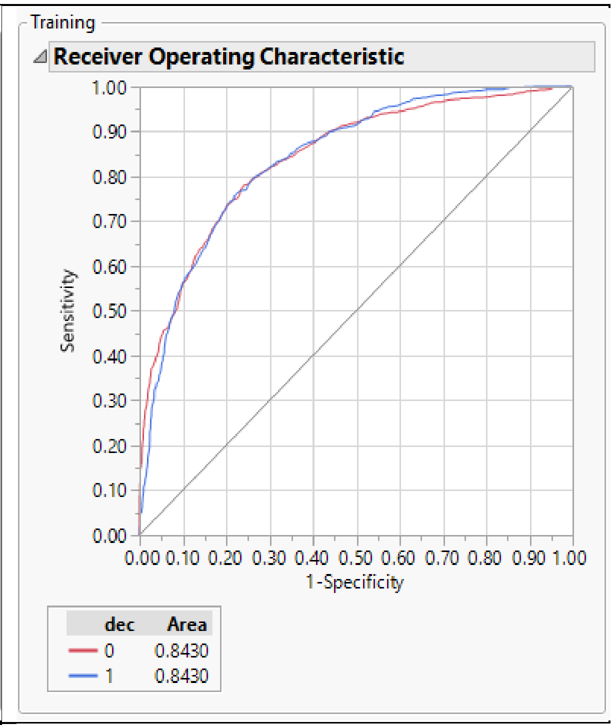
Validation

dec

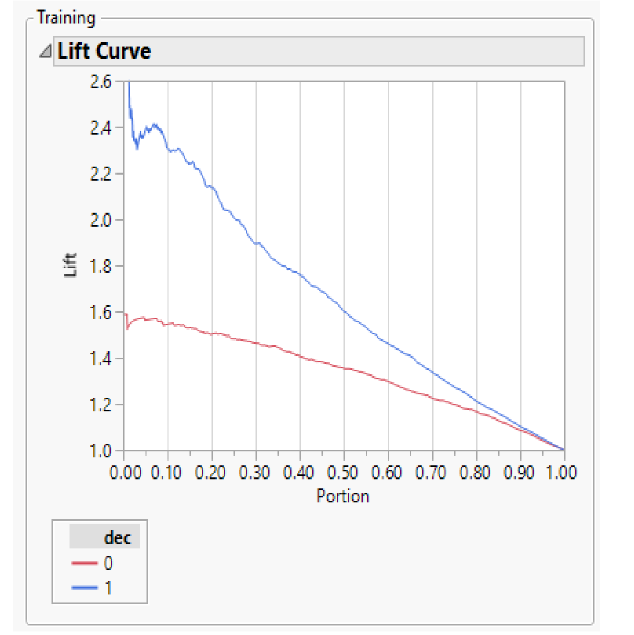
Measures	Value
Generalized RSquare	0.4212878
Entropy RSquare	0.2739682
RMSE	0.4094664
Mean Abs Dev	0.328166
Misclassification Rate	0.2535991

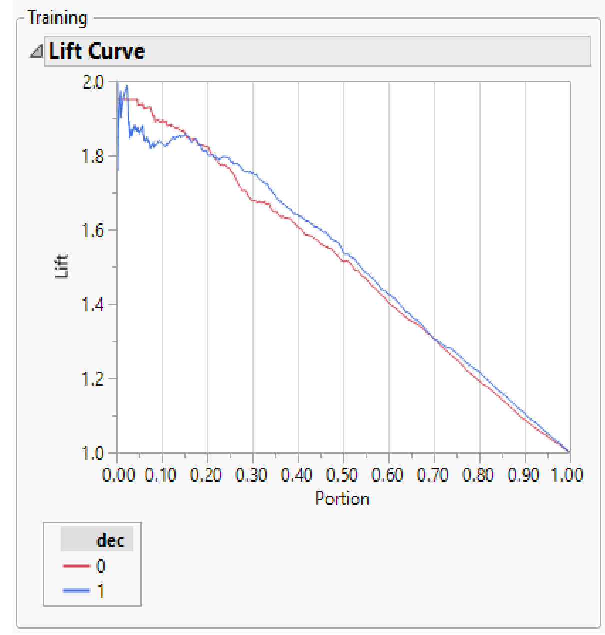


dec	Area
0	0.8236
1	0.8236



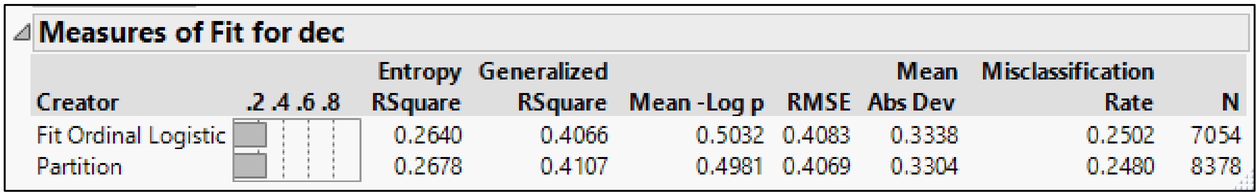
dec	Area
0	0.8430
1	0.8430





Assess

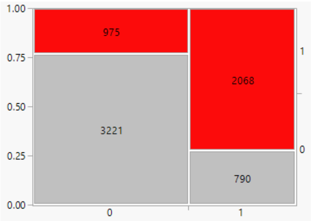
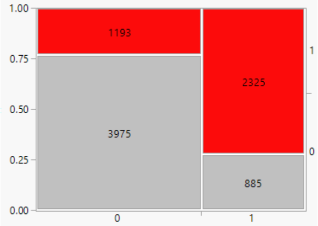
The 3 models were compared side-by-side. The value of R^2 is important because it measures how close the data is to the fitted regression line. This statistic provides some information about the goodness of fit of a model.



Measures of Fit for dec

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Ordinal Logistic		0.2640	0.4066	0.5032	0.4083	0.3338	0.2502	7054
Partition		0.2678	0.4107	0.4981	0.4069	0.3304	0.2480	8378

The confusion matrix table visually described the performance of the model. The results are summarized in the below chart:

	Logistic Regression		Decision Tree		Neural Network[1]	
	1	0	1	0	1	0
Dec 1	29.32%	13.82%	27.75%	14.24%	59.59%	13.54%
Dec 0	11.20%	45.66%	10.56%	47.45%	10.85%	46.02%
Misclassification Rates	0.2502		0.2480		0.2439	
					*	

All the three models have relative low Entropy RSquare due to ordinal responses. By comparing the misclassification rates, we find that all the models have similar performance. However, the simpler one is

[1] Unable to save probability formula for Neural Network model

the better one. Logistic regression model is the simplest model and easier to understand even though its misclassification rate is a little bit higher than those of any other two.

Model Interpretation

The value in the table below is the mean rating of each attribute by gender. Given the mean rating of a partner, a man is 46.86% likely to choose to go on a date, and woman, 30.75%.

	Male	Female
attr	6.46	5.92
sinc	7.25	7.10
fun	6.52	6.28
amb	6.60	6.95
share	5.54	5.41
linear	-5.27130	-4.44932
fail	0.531356	0.692544
succeed	0.468644	0.307456

If we add 1 to the rating of each attribute, some changes may happen in the probability of a successful date. As can be seen below:

	Prob [1] - Male	Change	Prob [1] - Female	Change
attr+1	0.6347960	35.45%	0.402210	30.82%

sinc+1	0.4227500	-9.79%	0.292927	-4.73%
fun+1	0.5402740	15.28%	0.373844	21.59%
amb+1	0.4296930	-8.31%	0.281075	-8.58%
share+1	0.5319400	13.51%	0.369480	20.17%

1. The probability of a successful date will increase with the rise rating of attractiveness, fun and shared interests for both male and female. However, increased sincerity and ambition will have a negative effect on the probability of a successful date.
2. For both male and female, attractiveness makes the most difference. One more score in attractiveness will bring more than 35% probability for men to say YES, which is 5% higher than for women. That is to say, men care more about attractiveness than women when they choose speed dating partner.

3. The second most important factor is fun both for male and for female. Women pay more attention to fun than men.

(Note: refer to formulas in Appendix B for further analysis.)

Business Insights

Speed dating was trademarked by Rabbi Yaacov Deyo in 1998 for singles specifically looking for a serious relationship and potential life partner. Maximizing the efficiency of identifying a marriage partner did not compromise quality of match-making results or quality of the clientele.

As the idea integrated into the dating industry, organizations tried implementing this process into areas that ended up weakening the results and cheapening the quality. While industry followers were initially successful in cashing in on this trend, it was apparent by 2003 that the market had lost control of this idea. This forum was initially so successful because it was about quality for a group of people looking for a serious relationship. In order to revive this idea back to its maximum earning potential developers and investors had to redesign their approach.

In this speed dating study specifically we found that planning of the *pre dating* questions is valuable. Framing some of the pre dating questions to get insight on attributes and experiences of *prior* dating partners and experiences would be a recommendation for data analysts to explore. These insights on prior experiences could potentially help offer events and services that would increase match potential more than *desired* traits.

Attributes is an important area to explore when marketing for online dating – specifically speed dating. For example, in pre-trial ratings subjects anticipated intelligence would be something that would be a significant influence on the success rate of their match/preference. However, when actual results post -dating were examined, the model we used eliminated this attribute as a significant factor for males immediately during the match process with relevance to females quickly following in the 2nd round. Women weighted the attributes more evenly than the men did, with intelligence on top and ambition on the bottom.

Recommendations

Top Recommendations:

- Speed dating organizations should invest more effort into attractiveness and fun, since these were the top attributes in the model, as opposed to the insignificant attributes (e.g. ambition).
- Redefine the rating system to include most meaningful attributes and goals of target demographic to most effectively match participants to event.

- Offer separate events to keep participants involved most well matched. For example, speed dating for those looking for a serious relationship and separate events for those looking more for companionship/friendship.
- To increase participation and success in this niche market, focus speed dating events on traditional relationship seeking audience.

- Offer events to financial planning associations to pitch their products and funds to potential investors. (Puko)

Other Recommendations:

- Offer speed dating events to specific demographics. Studies show 80% of those dating prefer to date someone of their ethnic background.
- To expand customer base customize speed dating forum to offer services to business market. Employers look for low cost venues to help in the pre-screening process when replacing or filling a position. Employment agency fees range from 10%-25% of filled position salary, while use of headhunters for upper management /executive level positions can climb as high as 213% of annual salary. (Employment Agency Cost)
- Expand services to students, specifically foreign students for networking opportunities. It is often difficult to acclimate to a new culture. This type of service could help bridge the gap students and job seekers find when entering the workplace as an international candidate.

References

Employment Agency Cost. Ed. INC CostHelper. n.d. 10 April 2016. <http://personalfinance.costhelper.com/employment-agencies.html>.

Merhar, Christina. *Small Business Employee Benefits and HR Blog*. February 4 2016. 10 April 2016. <http://www.zanebenefits.com/blog/bid/312123/Employee-Retention-The-Real-Cost-of-Losing-an-Employee>.

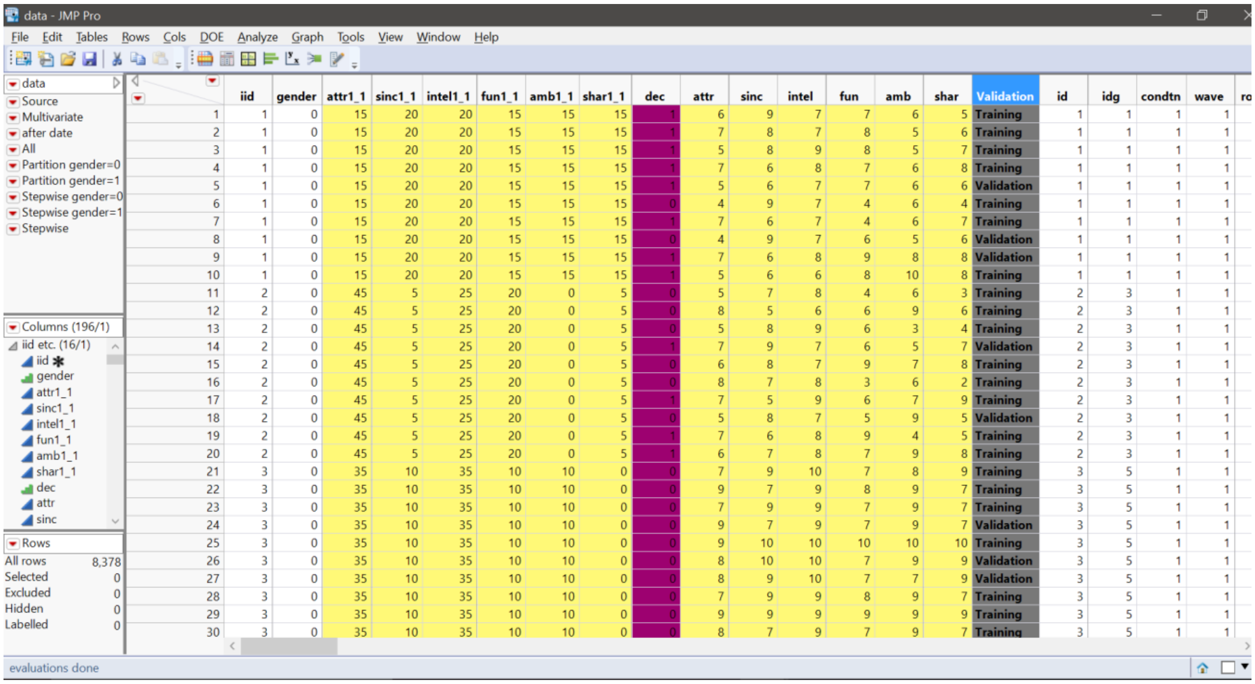
Puko, Tim. *The Wall Street Journal*. 5 February 2016. 10 April 2016. <http://blogs.wsj.com/moneybeat/2016/02/05/in-search-for-investors-hedge-funds-try-speed-dating/>.

Random Facts. 8 January 2010. April 2016. <http://facts.randomhistory.com/dating-and-relationship-facts.html>.

The Business of Dating. n.d. WikiDot. 10 April 2016. <http://thebusinessofdating.wikidot.com/the-dating-industry>.

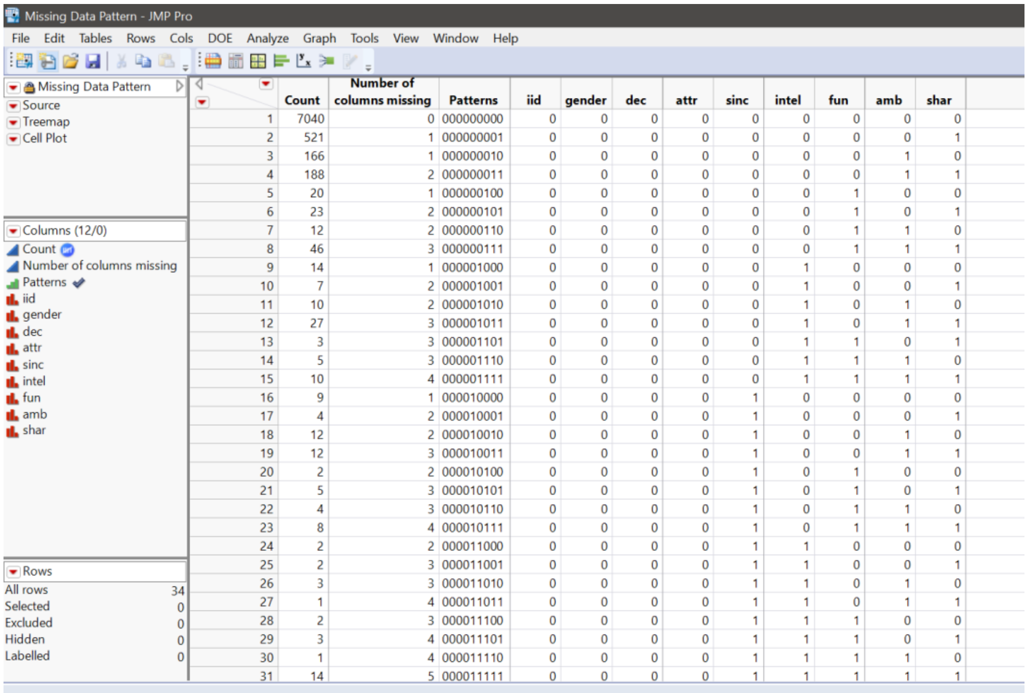
Appendix A: Exploration Screenshots

Figure 1



	iid	gender	attr1_1	sinc1_1	intel1_1	fun1_1	amb1_1	shar1_1	dec	attr	sinc	intel	fun	amb	shar	Validation	id	idg	condtn	wave	ro
1	1	0	15	20	20	15	15	15	1	6	9	7	7	6	5	Training	1	1	1	1	
2	1	0	15	20	20	15	15	15	1	7	8	7	8	5	6	Training	1	1	1	1	
3	1	0	15	20	20	15	15	15	1	5	8	9	8	5	7	Training	1	1	1	1	
4	1	0	15	20	20	15	15	15	1	7	6	8	7	6	8	Training	1	1	1	1	
5	1	0	15	20	20	15	15	15	1	5	6	7	7	6	6	Validation	1	1	1	1	
6	1	0	15	20	20	15	15	15	0	4	9	7	4	6	4	Training	1	1	1	1	
7	1	0	15	20	20	15	15	15	1	7	6	7	4	6	7	Training	1	1	1	1	
8	1	0	15	20	20	15	15	15	0	4	9	7	6	5	6	Validation	1	1	1	1	
9	1	0	15	20	20	15	15	15	1	7	6	8	9	8	8	Validation	1	1	1	1	
10	1	0	15	20	20	15	15	15	1	5	6	6	8	10	8	Training	1	1	1	1	
11	2	0	45	5	25	20	0	5	0	5	7	8	4	6	3	Training	2	3	1	1	
12	2	0	45	5	25	20	0	5	0	8	5	6	6	9	6	Training	2	3	1	1	
13	2	0	45	5	25	20	0	5	0	5	8	9	6	3	4	Training	2	3	1	1	
14	2	0	45	5	25	20	0	5	1	7	9	7	6	5	7	Validation	2	3	1	1	
15	2	0	45	5	25	20	0	5	0	6	8	7	9	7	8	Training	2	3	1	1	
16	2	0	45	5	25	20	0	5	0	8	7	8	3	6	2	Training	2	3	1	1	
17	2	0	45	5	25	20	0	5	1	7	5	9	6	7	9	Training	2	3	1	1	
18	2	0	45	5	25	20	0	5	0	5	8	7	5	9	5	Validation	2	3	1	1	
19	2	0	45	5	25	20	0	5	1	7	6	8	9	4	5	Training	2	3	1	1	
20	2	0	45	5	25	20	0	5	1	6	7	8	7	9	8	Training	2	3	1	1	
21	3	0	35	10	35	10	10	0	0	7	9	10	7	8	9	Training	3	5	1	1	
22	3	0	35	10	35	10	10	0	0	9	7	9	8	9	7	Training	3	5	1	1	
23	3	0	35	10	35	10	10	0	0	7	9	9	7	9	7	Training	3	5	1	1	
24	3	0	35	10	35	10	10	0	0	9	7	9	7	9	7	Validation	3	5	1	1	
25	3	0	35	10	35	10	10	0	0	9	10	10	10	10	10	Training	3	5	1	1	
26	3	0	35	10	35	10	10	0	0	8	10	10	7	9	9	Validation	3	5	1	1	
27	3	0	35	10	35	10	10	0	0	8	9	10	7	7	9	Validation	3	5	1	1	
28	3	0	35	10	35	10	10	0	0	7	9	9	8	9	7	Training	3	5	1	1	
29	3	0	35	10	35	10	10	0	0	9	9	9	9	9	9	Training	3	5	1	1	
30	3	0	35	10	35	10	10	0	0	8	7	9	7	9	7	Training	3	5	1	1	

Figure 2



	Count	Number of columns missing	Patterns	iid	gender	dec	attr	sinc	intel	fun	amb	shar
1	7040	0	000000000	0	0	0	0	0	0	0	0	0
2	521	1	000000001	0	0	0	0	0	0	0	0	1
3	166	1	000000010	0	0	0	0	0	0	0	1	0
4	188	2	000000011	0	0	0	0	0	0	0	1	1
5	20	1	000000100	0	0	0	0	0	0	1	0	0
6	23	2	000000101	0	0	0	0	0	0	1	0	1
7	12	2	000000110	0	0	0	0	0	0	1	1	0
8	46	3	000000111	0	0	0	0	0	0	1	1	1
9	14	1	000001000	0	0	0	0	0	1	0	0	0
10	7	2	000001001	0	0	0	0	0	1	0	0	1
11	10	2	000001010	0	0	0	0	0	1	0	1	0
12	27	3	000001011	0	0	0	0	0	1	0	1	1
13	3	3	000001101	0	0	0	0	0	1	1	0	1
14	5	3	000001110	0	0	0	0	0	1	1	1	0
15	10	4	000001111	0	0	0	0	0	1	1	1	1
16	9	1	000010000	0	0	0	0	1	0	0	0	0
17	4	2	000010001	0	0	0	0	1	0	0	0	1
18	12	2	000010010	0	0	0	0	1	0	0	1	0
19	12	3	000010011	0	0	0	0	1	0	0	1	1
20	2	2	000010100	0	0	0	0	1	0	1	0	0
21	5	3	000010101	0	0	0	0	1	0	1	0	1
22	4	3	000010110	0	0	0	0	1	0	1	1	0
23	8	4	000010111	0	0	0	0	1	0	1	1	1
24	2	2	000011000	0	0	0	0	1	1	0	0	0
25	2	3	000011001	0	0	0	0	1	1	0	0	1
26	3	3	000011010	0	0	0	0	1	1	0	1	0
27	1	4	000011011	0	0	0	0	1	1	0	1	1
28	2	3	000011100	0	0	0	0	1	1	1	0	0
29	3	4	000011101	0	0	0	0	1	1	1	0	1
30	1	4	000011110	0	0	0	0	1	1	1	1	0
31	14	5	000011111	0	0	0	0	1	1	1	1	1

Figure 3

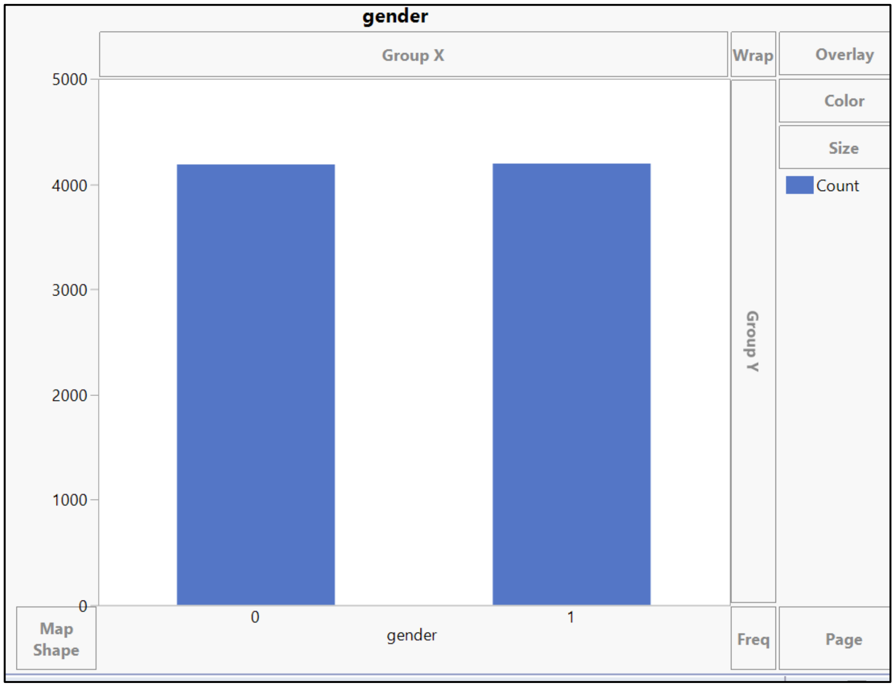


Figure 4

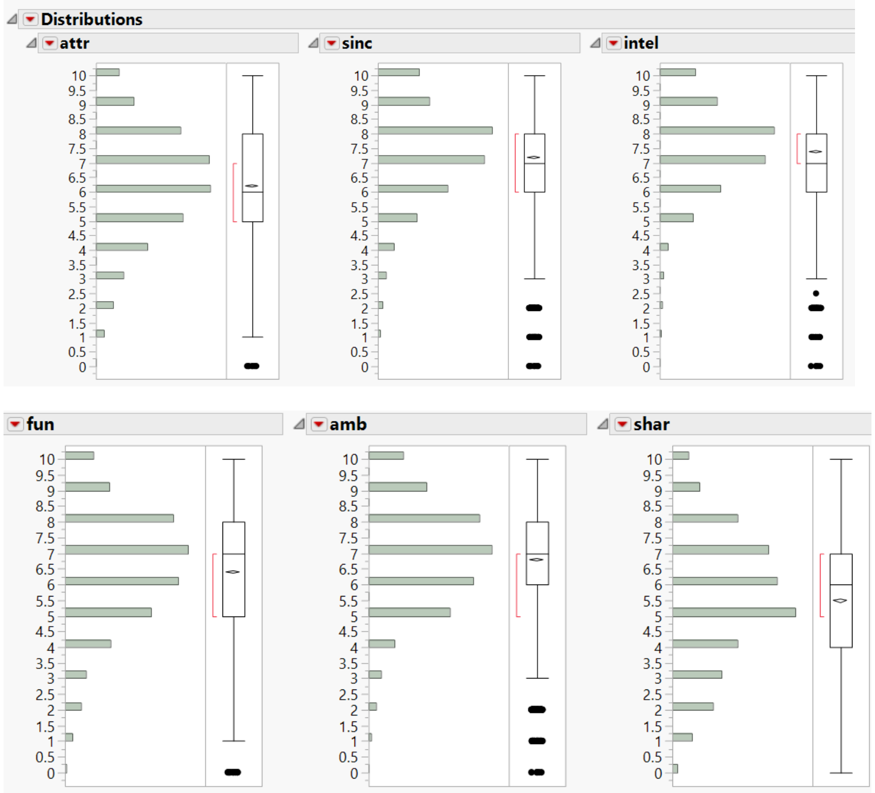


Figure 5

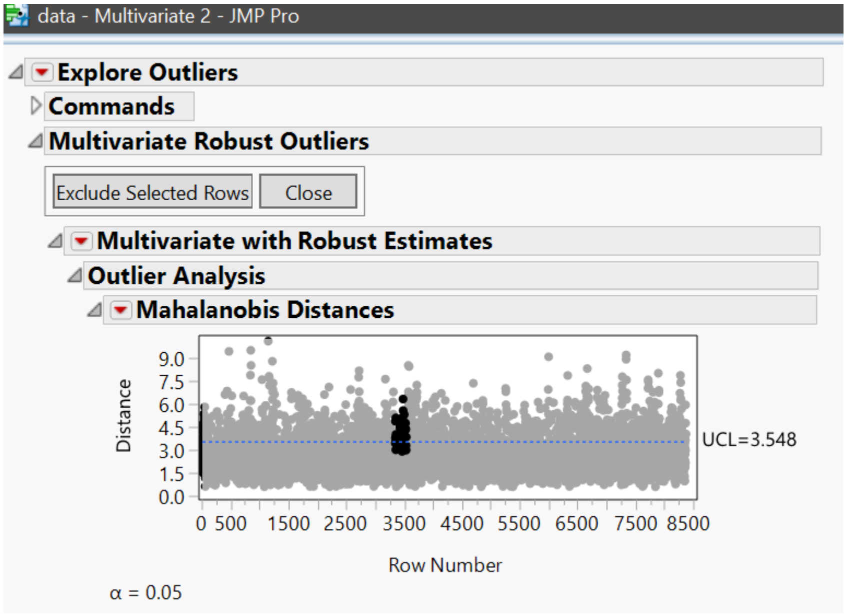
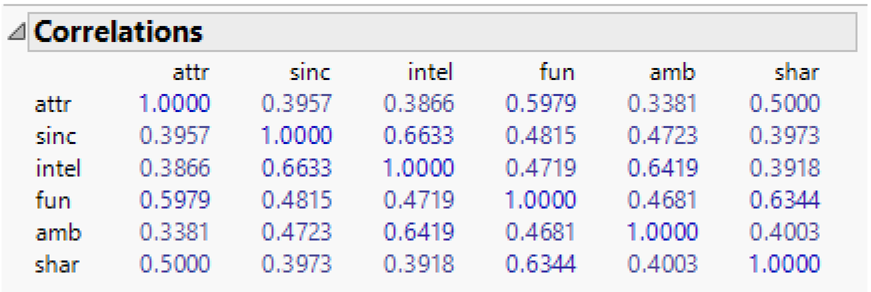


Figure 6

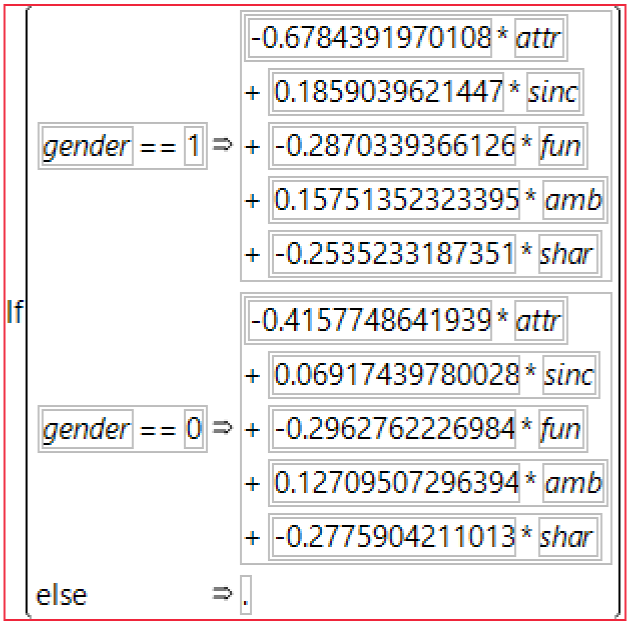


Correlations

	attr	sinc	intel	fun	amb	shar
attr	1.0000	0.3957	0.3866	0.5979	0.3381	0.5000
sinc	0.3957	1.0000	0.6633	0.4815	0.4723	0.3973
intel	0.3866	0.6633	1.0000	0.4719	0.6419	0.3918
fun	0.5979	0.4815	0.4719	1.0000	0.4681	0.6344
amb	0.3381	0.4723	0.6419	0.4681	1.0000	0.4003
shar	0.5000	0.3973	0.3918	0.6344	0.4003	1.0000

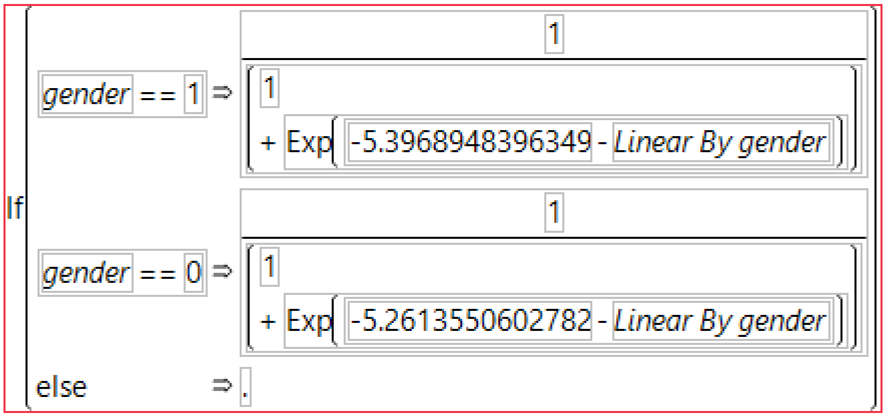
Appendix B: Formulas

Leaner by gender:



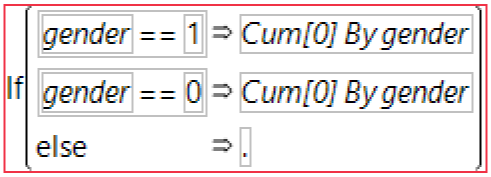
$$\text{If}\begin{cases} gender == 1 \Rightarrow -0.6784391970108 * attr + 0.1859039621447 * sinc + -0.2870339366126 * fun + 0.15751352323395 * amb + -0.2535233187351 * shar \\ gender == 0 \Rightarrow -0.4157748641939 * attr + 0.06917439780028 * sinc + -0.2962762226984 * fun + 0.12709507296394 * amb + -0.2775904211013 * shar \\ \text{else} \Rightarrow . \end{cases}$$

Cum [0] By gender:



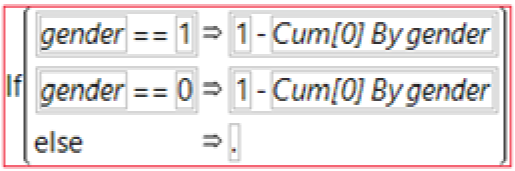
$$\text{If}\begin{cases} gender == 1 \Rightarrow \dfrac{1}{\left(1 + \text{Exp}\left(-5.3968948396349 - \textit{Linear By gender}\right)\right)} \\ gender == 0 \Rightarrow \dfrac{1}{\left(1 + \text{Exp}\left(-5.2613550602782 - \textit{Linear By gender}\right)\right)} \\ \text{else} \Rightarrow . \end{cases}$$

Probability [0] By gender:



$$\text{If}\begin{cases} gender == 1 \Rightarrow \textit{Cum[0] By gender} \\ gender == 0 \Rightarrow \textit{Cum[0] By gender} \\ \text{else} \Rightarrow . \end{cases}$$

Probability [1] By gender:



```
If  gender == 1 ⇒ 1 - Cum[0] By gender
    gender == 0 ⇒ 1 - Cum[0] By gender
    else        ⇒ .
```