Andrew Bauman, Mark Miller
MSIS 645L
March 25, 2019

Project #1: Targeted Marketing

Companies routinely use surveys to estimate potential revenues in different markets. Survey responses are collected and aggregated in large databases where data analysis applications are used to generate reports to help in business decisions.

Suppose you are marketing manager attempting to optimize an advertising campaign. For years managers often send out bulk mailings (now bulk emails) to different prospect lists. Imagine you can measure cost for every mail sent. If every contact costs the company 5 dollars, sending a mail to 1000 people costs 5,000 dollars. This becomes very expensive. None of the prospects even guarantee a response, or even a response which leads to additional revenue. Companies employ data analysis techniques to select only those prospects most likely to respond to a campaign. By analyzing survey responses from prospects, we can segregate prospects into different groups of revenue probability.

Why are beer commercials shown during the Superbowl? Beer companies are showing targeted advertisements to a segment of the population they think are the most likely to go out and buy another 6 pack. Instead of showing commercials all year round and paying millions of dollars, they select the best potential sample of the population at the right time. How then, do they go about selecting the right audience? It seems it is becoming less common for people to make multi-million dollar decisions based on gut feeling and intuition.

In this case, you are a data analyst who works for the Land Rover automobile company. You desire to gain a better understanding of the lifestyle of potential SUV buyers. You commission a study of consumer attributes, interests, and opinions, and send out a questionnaire. The survey involves 30 questions, covering a variety of different consumer dimensions. The 31st question asked the consumer to rate themselves as to how likely they would be to purchase a SUV.

Respondents were asked to use a 9 point Likert scale to give their answer to each question. The value of "1" means the responder disagreed with the statement, and the value of "9" means the responder totally agrees. 400 consumers were surveyed. The profiles were obtained from the mailing lists of Car and Driver, Business Week, and Inc. magazines.

The survey questions (prefixed by their column name in the data file):

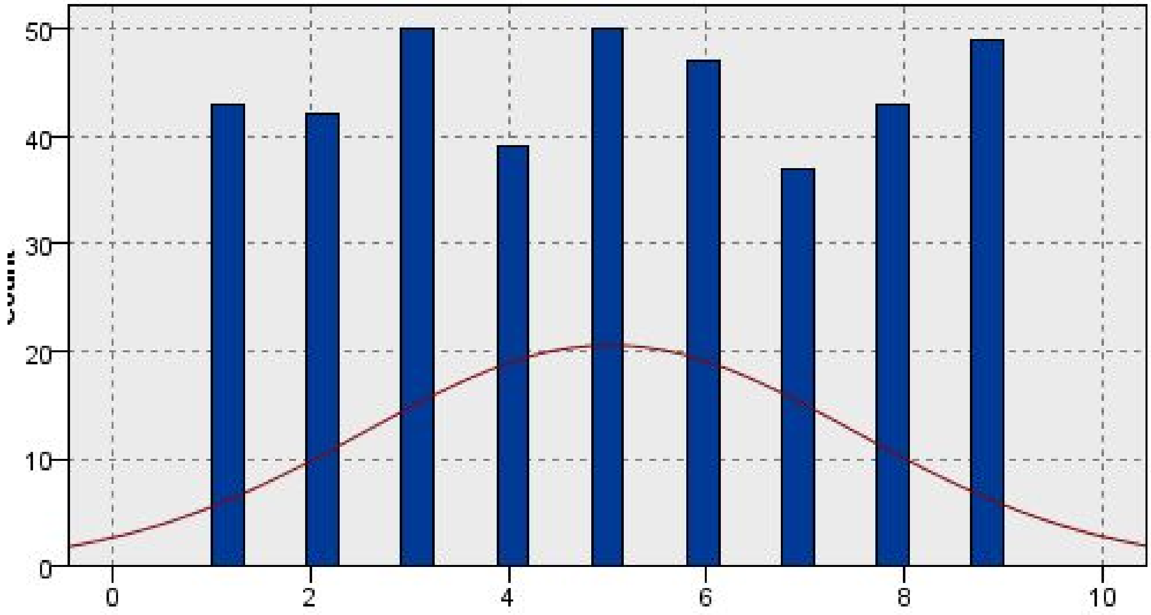
1. In Shape - I am in very good physical condition
2. Fashionable - When I must choose between the two, I dress for fashion, not comfort
3. Stylish - I have more stylish clothes than most of my friends
4. Individualistic - I want to look a little different from others
5. Risk Taker - Life is too short not to take some gambles
6. No Ozone Concern - I am not concerned about the ozone layer
7. Right To Pollute- I think the government is doing too much to control pollution
8. Society Fine - Basically, society today is fine
9. No Time For Charity - I don't have time to volunteer for charities
10. No Debt - Our family is not too heavily in debt today
11. Prefer Cash - I like to pay cash for everything I buy
12. Spendthrift - I pretty much spend for today and let tomorrow bring what it will
13. Prefer Credit - I use credit cards because I can pay the bill off slowly
14. No Coupons - I seldom use coupons when I shop
15. Low Interest Buyer - Interest rates are low enough to allow me to buy what I want
16. Confident - I have more self-confidence than most of my friends
17. Leader - I like to be considered a leader
18. Dependable - Others often ask me to help them out of a jam
19. Children Important - Children are the most important thing in a marriage
20. Introverted - I would rather spend a quiet evening at home than go out to a party
21. American Cars Rule - Foreign-made cars can't compare with American-made cars
22. Restrict Japan Imports - The government should restrict imports of products from Japan
23. Buy American - Americans should always try to buy American products
24. Adventurous - I would like to take a trip around the world
25. Midlife Crisis - I wish I could leave my present life and do something entirely different
26. Early Adopter - I am usually among the first to try new products
27. Active - I like to work hard and play hard
28. Skeptics Wrong - Skeptical predictions are usually wrong
29. Determined - I can do anything I set my mind to
30. Optimistic - Five years from now, my income will be a lot higher than it is now
31. Attitude - I would consider buying the Discovery made by Land Rover

The data is provided in an Excel file named SurveyData.xls. By this time you should be pretty familiar with loading data from an Excel file or other sources.

Tasks:

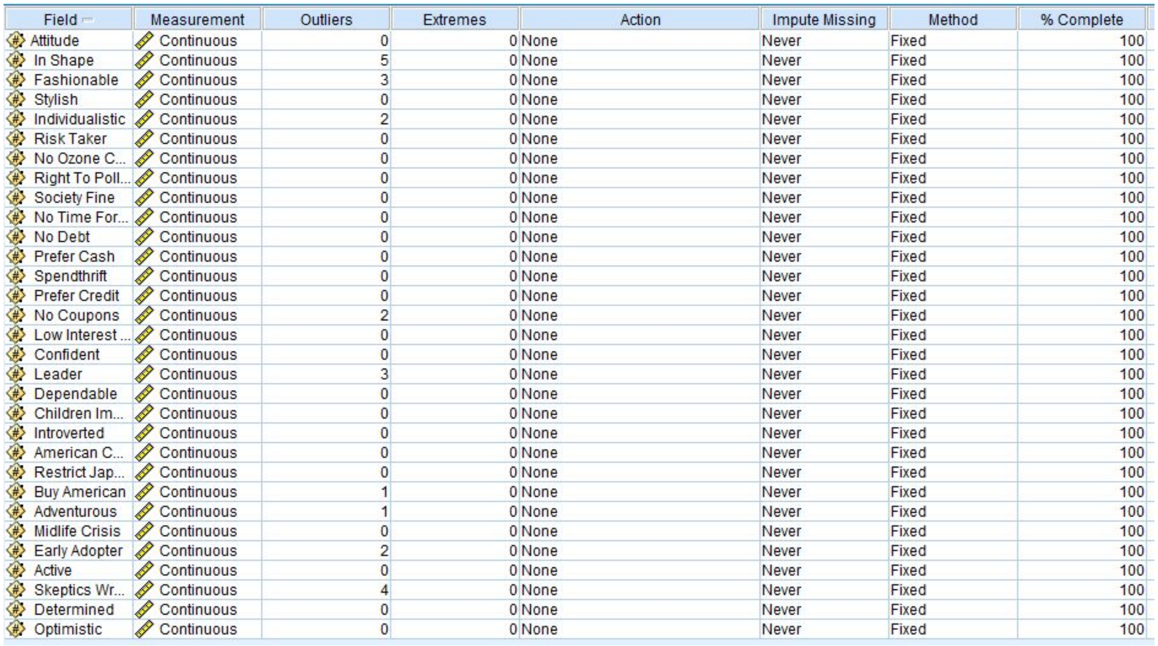
1. **Explore the data to get some initial insights. Add a few chart nodes and a statistics node to visualize the data. Report on your findings of the shape of the data. (Note: we assume that a Likert scale of 1-9 points can approximate a continuous variable)**

The shape of the data did not appear of be of a normal distribution, but we came to a conclusion that the shape of the data resembled a bimodal distribution. We used z-score normalization to even check further on the shape of the data but received no additional information leading us to our conclusion of a bimodal distribution.



2. Missing data can be a problem. Check for missing values (there may be some, there may none) and remediate the issue accordingly, if you think it is necessary.

In order to identify if we had missing data, we used a data audit node stemming from the SurveyData. From the picture below we can determine that there was no missing data found in the excel file.

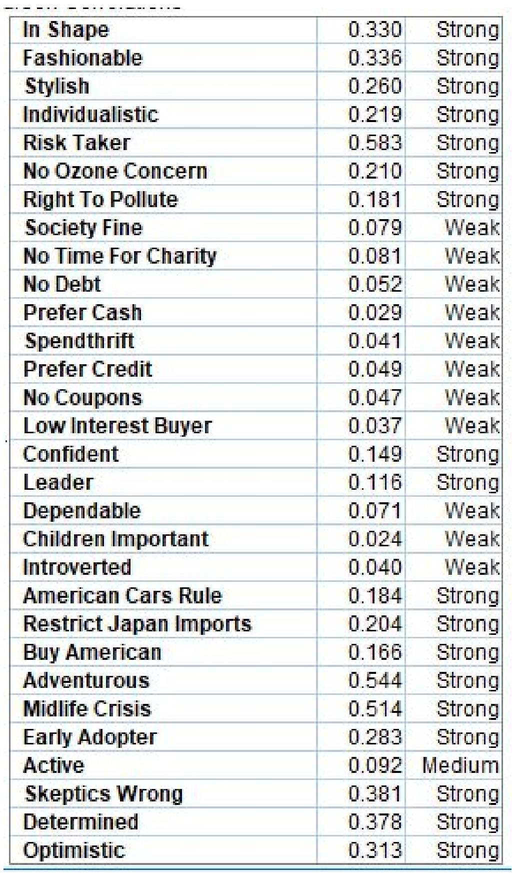


Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete
Attitude	Continuous	0	0	None	Never	Fixed	100
In Shape	Continuous	5	0	None	Never	Fixed	100
Fashionable	Continuous	3	0	None	Never	Fixed	100
Stylish	Continuous	0	0	None	Never	Fixed	100
Individualistic	Continuous	2	0	None	Never	Fixed	100
Risk Taker	Continuous	0	0	None	Never	Fixed	100
No Ozone C...	Continuous	0	0	None	Never	Fixed	100
Right To Poll...	Continuous	0	0	None	Never	Fixed	100
Society Fine	Continuous	0	0	None	Never	Fixed	100
No Time For...	Continuous	0	0	None	Never	Fixed	100
No Debt	Continuous	0	0	None	Never	Fixed	100
Prefer Cash	Continuous	0	0	None	Never	Fixed	100
Spendthrift	Continuous	0	0	None	Never	Fixed	100
Prefer Credit	Continuous	0	0	None	Never	Fixed	100
No Coupons	Continuous	2	0	None	Never	Fixed	100
Low Interest ...	Continuous	0	0	None	Never	Fixed	100
Confident	Continuous	0	0	None	Never	Fixed	100
Leader	Continuous	3	0	None	Never	Fixed	100
Dependable	Continuous	0	0	None	Never	Fixed	100
Children Im...	Continuous	0	0	None	Never	Fixed	100
Introverted	Continuous	0	0	None	Never	Fixed	100
American C...	Continuous	0	0	None	Never	Fixed	100
Restrict Jap...	Continuous	0	0	None	Never	Fixed	100
Buy American	Continuous	1	0	None	Never	Fixed	100
Adventurous	Continuous	1	0	None	Never	Fixed	100
Midlife Crisis	Continuous	0	0	None	Never	Fixed	100
Early Adopter	Continuous	2	0	None	Never	Fixed	100
Active	Continuous	0	0	None	Never	Fixed	100
Skeptics Wr...	Continuous	4	0	None	Never	Fixed	100
Determined	Continuous	0	0	None	Never	Fixed	100
Optimistic	Continuous	0	0	None	Never	Fixed	100

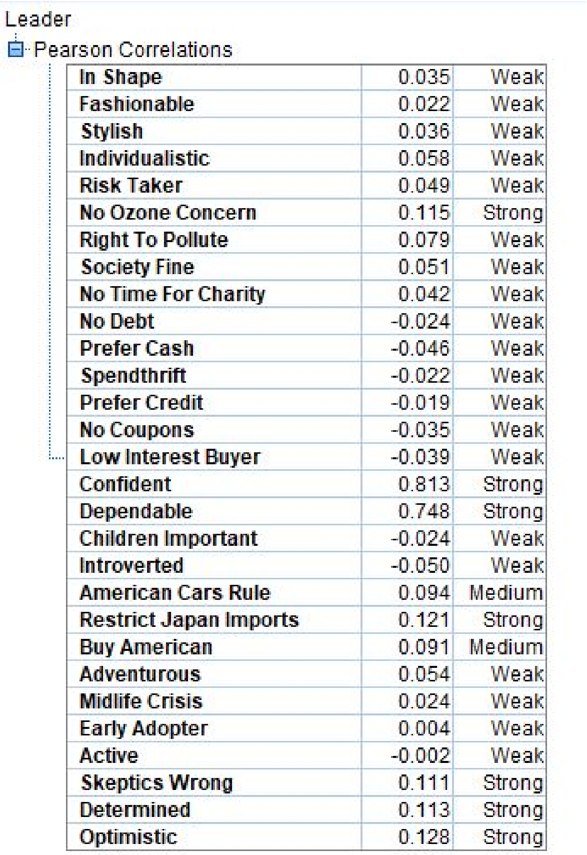
3. Provide a table describing the relationship of each survey question (questions 1-30) with Attitude (question 31). Also, investigate the correlation amongst the predictor variables (questions 1-30) [1]. Report your findings.

In order to gain information on correlation among the predictor values and the relationship of each survey question with attitude, we used a statistics node to generate the corresponding information. First, we used the Attitude against the 30 other survey questions and found the relationship varied among each variable.The relationship seemed to be spread even across the board between strong and weak ,but only found one relationship at medium.

Secondly, we had to investigate the correlation among the predictor values. For this we correlated each survey question against each other. The results of the statistics node gave us again, a varied result. This time we saw again a spread of strong and weak correlation with a few more mediums this time.



In Shape	0.330	Strong
Fashionable	0.336	Strong
Stylish	0.260	Strong
Individualistic	0.219	Strong
Risk Taker	0.583	Strong
No Ozone Concern	0.210	Strong
Right To Pollute	0.181	Strong
Society Fine	0.079	Weak
No Time For Charity	0.081	Weak
No Debt	0.052	Weak
Prefer Cash	0.029	Weak
Spendthrift	0.041	Weak
Prefer Credit	0.049	Weak
No Coupons	0.047	Weak
Low Interest Buyer	0.037	Weak
Confident	0.149	Strong
Leader	0.116	Strong
Dependable	0.071	Weak
Children Important	0.024	Weak
Introverted	0.040	Weak
American Cars Rule	0.184	Strong
Restrict Japan Imports	0.204	Strong
Buy American	0.166	Strong
Adventurous	0.544	Strong
Midlife Crisis	0.514	Strong
Early Adopter	0.283	Strong
Active	0.092	Medium
Skeptics Wrong	0.381	Strong
Determined	0.378	Strong
Optimistic	0.313	Strong



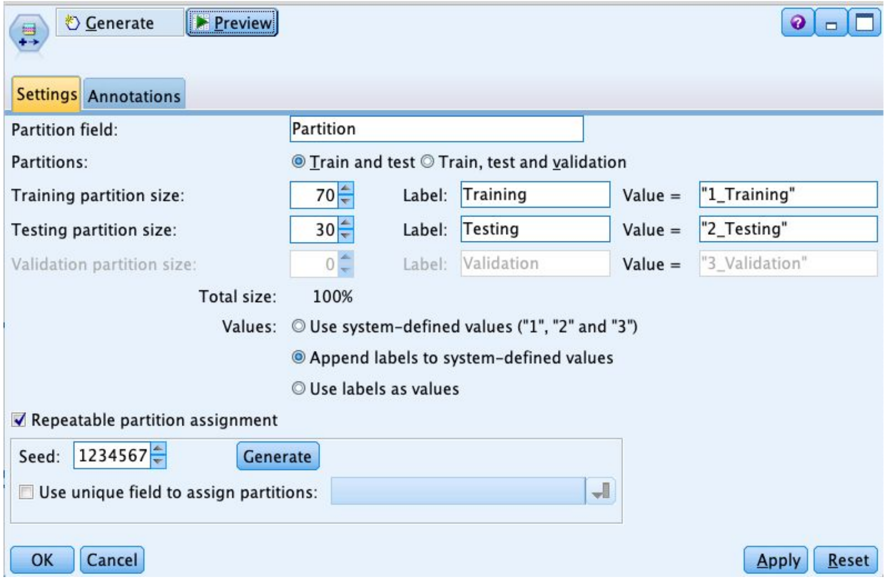
Leader

Pearson Correlations

In Shape	0.035	Weak
Fashionable	0.022	Weak
Stylish	0.036	Weak
Individualistic	0.058	Weak
Risk Taker	0.049	Weak
No Ozone Concern	0.115	Strong
Right To Pollute	0.079	Weak
Society Fine	0.051	Weak
No Time For Charity	0.042	Weak
No Debt	-0.024	Weak
Prefer Cash	-0.046	Weak
Spendthrift	-0.022	Weak
Prefer Credit	-0.019	Weak
No Coupons	-0.035	Weak
Low Interest Buyer	-0.039	Weak
Confident	0.813	Strong
Dependable	0.748	Strong
Children Important	-0.024	Weak
Introverted	-0.050	Weak
American Cars Rule	0.094	Medium
Restrict Japan Imports	0.121	Strong
Buy American	0.091	Medium
Adventurous	0.054	Weak
Midlife Crisis	0.024	Weak
Early Adopter	0.004	Weak
Active	-0.002	Weak
Skeptics Wrong	0.111	Strong
Determined	0.113	Strong
Optimistic	0.128	Strong

4. Partition the dataset: Use a 70% - 30% ratio for training and testing.

We partitioned the data by using a partition node and setting the training partition size and testing partition size to the correct values.



Partition field: Partition

Partitions: ● Train and test ○ Train, test and validation

Training partition size:	70	Label:	Training	Value =	"1_Training"
Testing partition size:	30	Label:	Testing	Value =	"2_Testing"
Validation partition size:	0	Label:	Validation	Value =	"3_Validation"

Total size: 100%

Values: ○ Use system-defined values ("1", "2" and "3")
● Append labels to system-defined values
○ Use labels as values

☑ Repeatable partition assignment

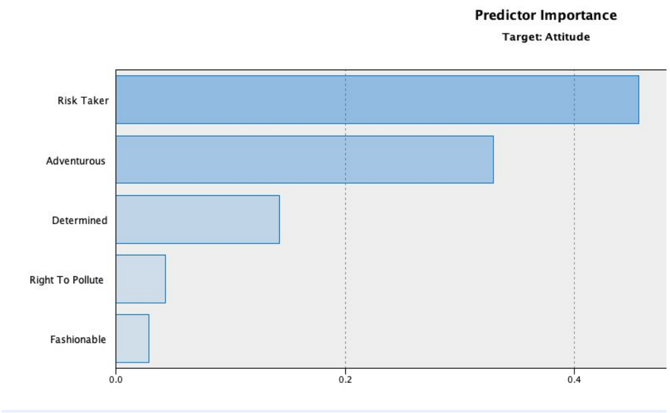
Seed: 1234567 Generate

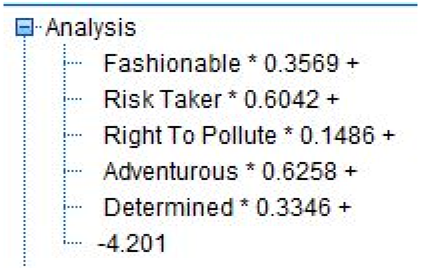
☐ Use unique field to assign partitions:

5. Use SPSS Modeler linear regression tool to investigate whether a linear relationship exists between attitude and the other variables. Use the method of your choice (all variables, a subset, stepwise / forwards / backwards) to build the multiple linear regression model, and justify your choice. For your model:

a. Write out the estimated regression equation and explain the meaning of the coefficients

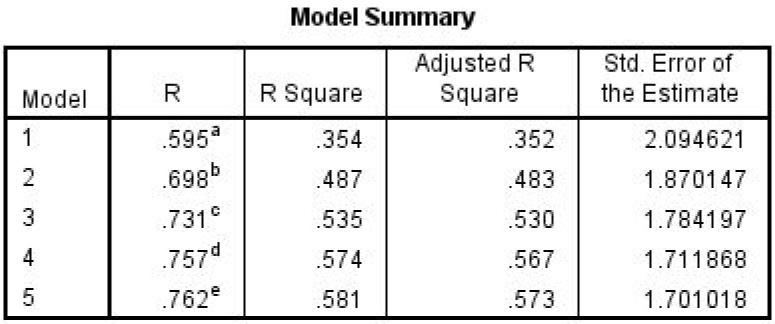
To investigate the linear relationship we opted to use the Stepwise method. SPSS identified the 5 predictors as being the most important. The estimated regression equation is show in the picture below.




Analysis
Fashionable * 0.3569 +
Risk Taker * 0.6042 +
Right To Pollute * 0.1486 +
Adventurous * 0.6258 +
Determined * 0.3346 +
-4.201

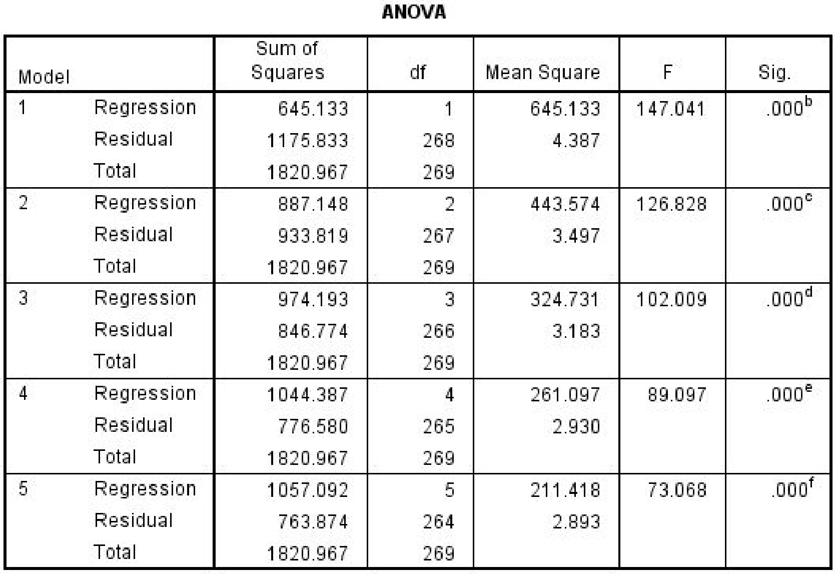
b. Provide a full report of the chosen regression model and report its metrics (goodness of fit, predictive performance) and statistics on training and test data.

The picture below shows both the R squared and adjusted R square. The R squared value tells us the goodness of fit metric. For our equation, model 5 is the goodness of fit for the entire equation, therefore our metric is 0.581 for the R squared and 0.573 for the adjusted R squared. We can also find the STD error of estimate which is the predicted error or,the typical error we see when we make a prediction. In our equation it was 1.70.


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.595[a]	.354	.352	2.094621
2	.698[b]	.487	.483	1.870147
3	.731[c]	.535	.530	1.784197
4	.757[d]	.574	.567	1.711868
5	.762[e]	.581	.573	1.701018

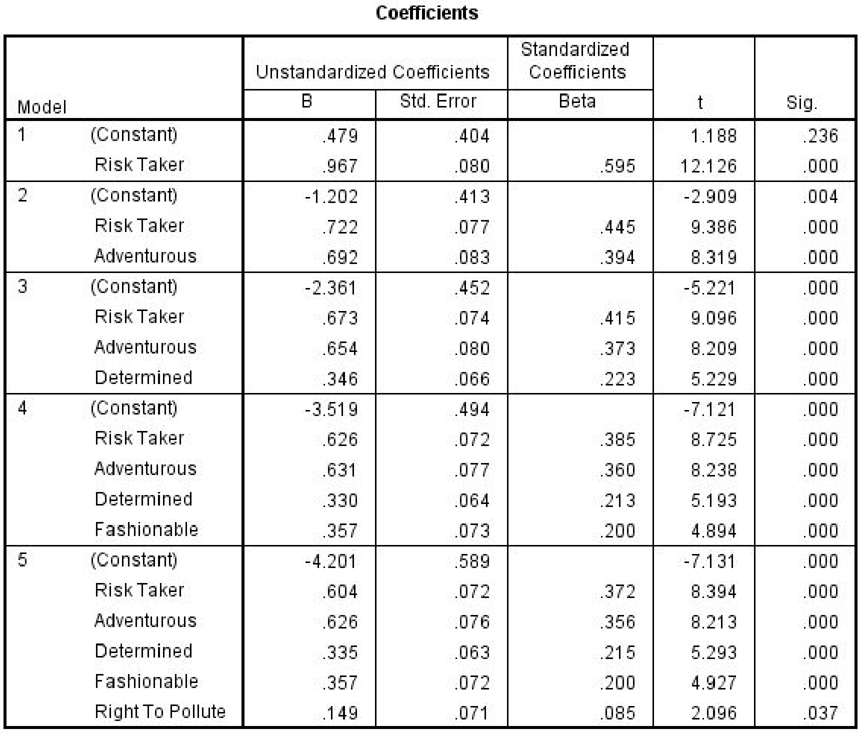
The picture below is our ANOVA summary of the regression model. In this summary we can find our F test which evaluates the relationship between response and a set of predictors.



ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	645.133	1	645.133	147.041	.000[b]
	Residual	1175.833	268	4.387		
	Total	1820.967	269			
2	Regression	887.148	2	443.574	126.828	.000[c]
	Residual	933.819	267	3.497		
	Total	1820.967	269			
3	Regression	974.193	3	324.731	102.009	.000[d]
	Residual	846.774	266	3.183		
	Total	1820.967	269			
4	Regression	1044.387	4	261.097	89.097	.000[e]
	Residual	776.580	265	2.930		
	Total	1820.967	269			
5	Regression	1057.092	5	211.418	73.068	.000[f]
	Residual	763.874	264	2.893		
	Total	1820.967	269			

This final summary gives us the majority of the important information. In here we can find the Y - Intercept and the slope. Our constant or the Y - intercept was -4.201. Next we had the slope for each predictor variable, Risk taker = 0.604, Adventurous = 0.626, Determined = 0.335, Fashionable = 0.357, and Right to Pollute = 0.149. These values are what make up our regression equation.

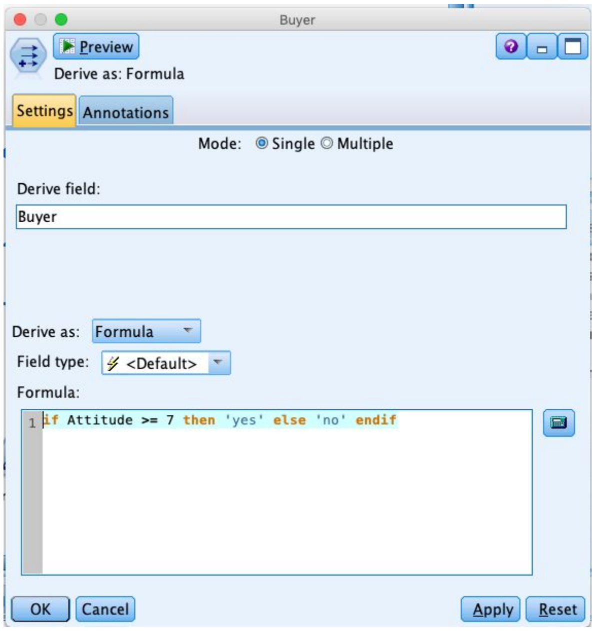


Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.479	.404		1.188	.236
	Risk Taker	.967	.080	.595	12.126	.000
2	(Constant)	-1.202	.413		-2.909	.004
	Risk Taker	.722	.077	.445	9.386	.000
	Adventurous	.692	.083	.394	8.319	.000
3	(Constant)	-2.361	.452		-5.221	.000
	Risk Taker	.673	.074	.415	9.096	.000
	Adventurous	.654	.080	.373	8.209	.000
	Determined	.346	.066	.223	5.229	.000
4	(Constant)	-3.519	.494		-7.121	.000
	Risk Taker	.626	.072	.385	8.725	.000
	Adventurous	.631	.077	.360	8.238	.000
	Determined	.330	.064	.213	5.193	.000
	Fashionable	.357	.073	.200	4.894	.000
5	(Constant)	-4.201	.589		-7.131	.000
	Risk Taker	.604	.072	.372	8.394	.000
	Adventurous	.626	.076	.356	8.213	.000
	Determined	.335	.063	.215	5.293	.000
	Fashionable	.357	.072	.200	4.927	.000
	Right To Pollute	.149	.071	.085	2.096	.037

6. Derive "Buyer" attribute: pick a threshold Attitude value (Attitude = 7), above which you can consider that that the responder is a highly likely buyer. Create a rule that tests the value of Attitude, and use this to derive a new attribute called Buyer (the rule is something like "Attitude" > 7)

Rule: if Attitude >= 7 then 'yes' else 'no' endif



Buyer

Derive as: Formula

Mode: ● Single ○ Multiple

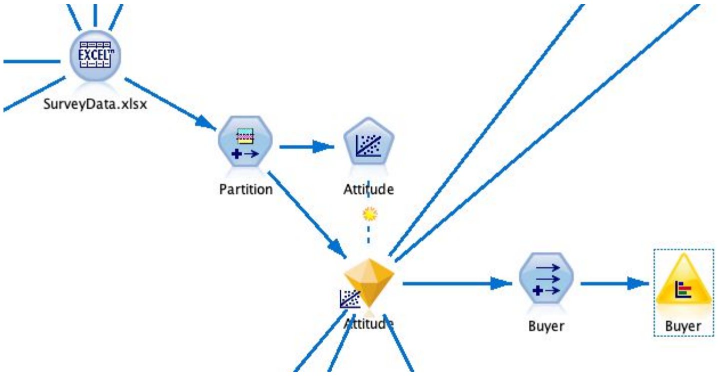
Derive field: Buyer

Derive as: Formula

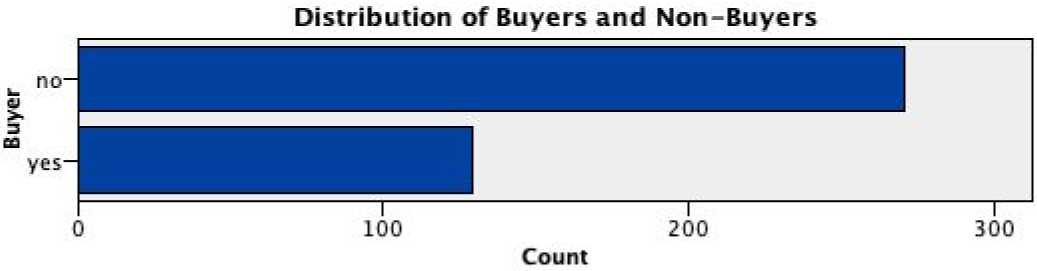
Field type: <Default>

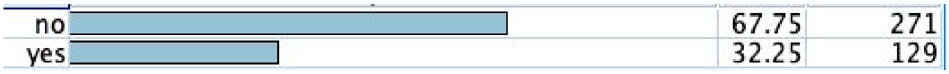
Formula:

```
if Attitude >= 7 then 'yes' else 'no' endif
```



7. Check the distribution of Buyers and non-Buyers in the dataset to see if the data set is unbalanced (you can plot a graph, or create a table).

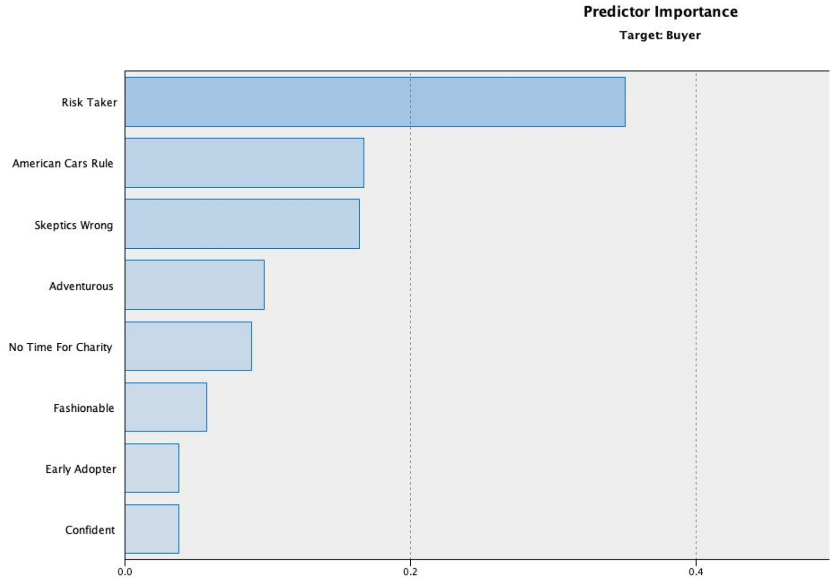




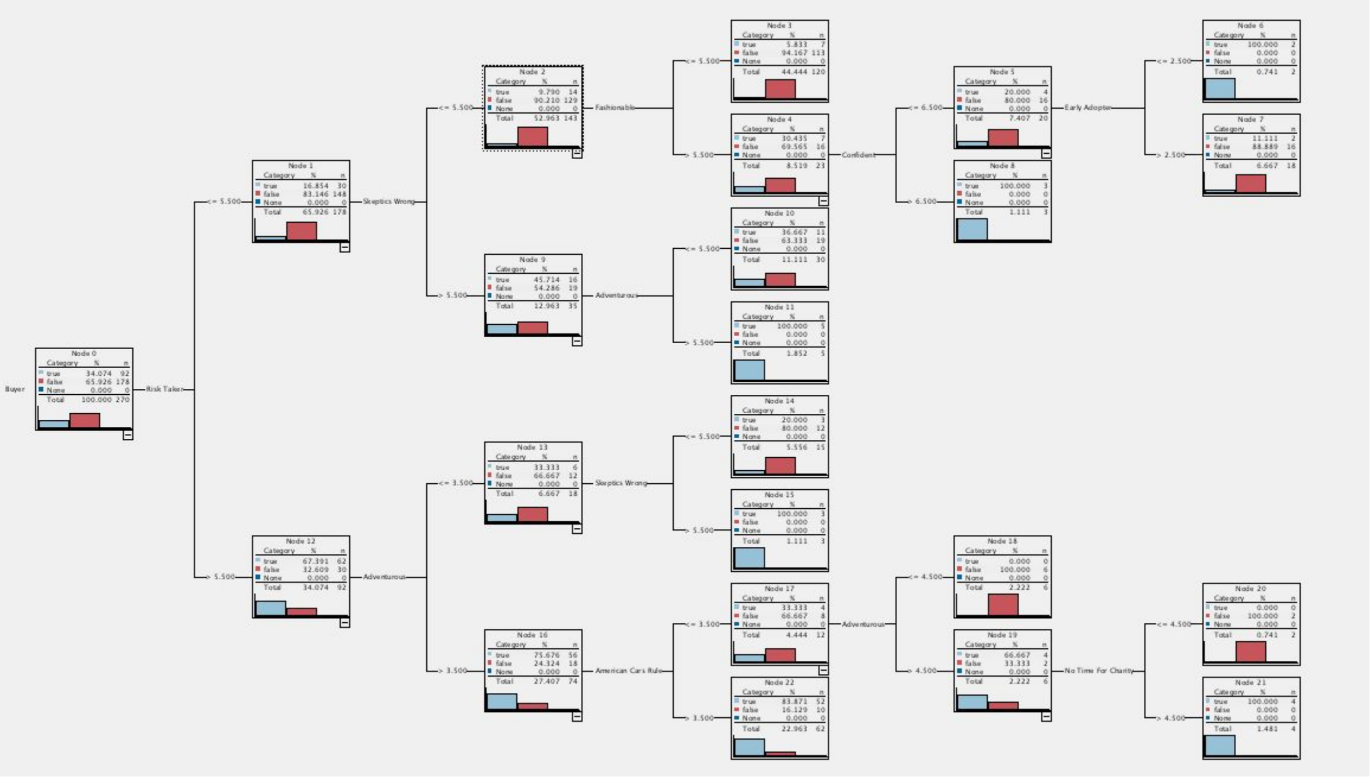
no		67.75	271
yes		32.25	129

Upon a brief analysis of the data, we can see that approximately 67.75% of the given sample are non-buyers, meaning the dataset is unbalanced, with recognized buyers representing only 32.25% of the overall dataset.

8. Decision Tree Classification: Model the dependence of how likely someone will purchase a Land Rover using a C5.0 decision tree. With the resulting tree, create a profile of the potential Buyer (Note: when creating the tree remember to drop Attitude from the list of predictors). This means explaining the rules created by the tree (or walking down the branches of the tree, same thing)

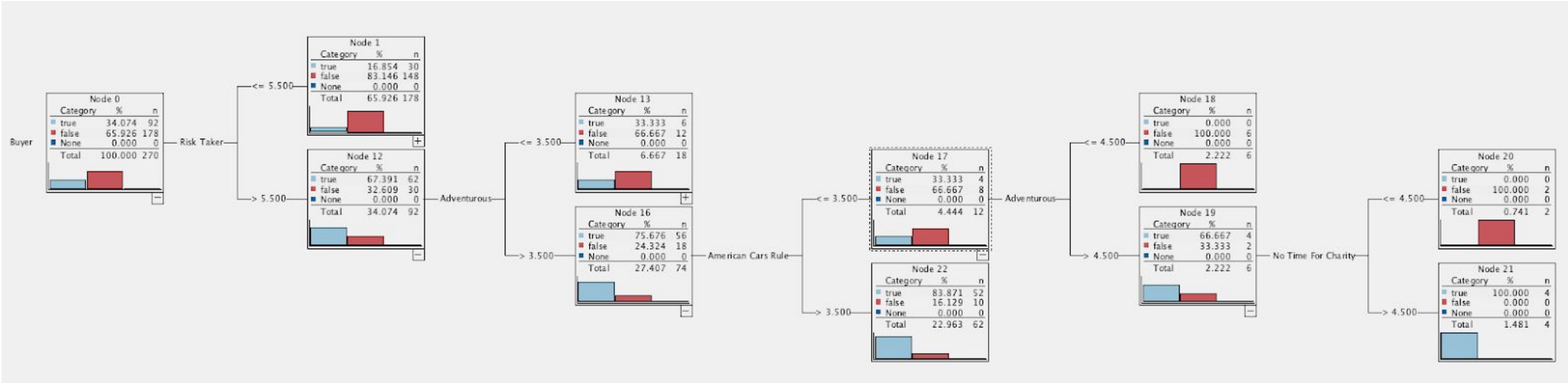


The C 5.0 Tree Algorithm identified the eight variables shown here (left) as the most important predictors of the derived target variable, *Buyer.*


Buyer
Node 0 Category % n; true 34.074 92; false 65.926 178; None 0.000 0; Total 100.000 270
Risk Taker
<= 5.500
Node 1 Category % n; true 16.854 30; false 83.146 148; None 0.000 0; Total 65.926 178
Skeptics Wrong
<= 5.500
Node 2 Category % n; true 9.790 14; false 90.210 129; None 0.000 0; Total 52.963 143
Fashionable
<= 5.500
Node 3 Category % n; true 5.833 7; false 94.167 113; None 0.000 0; Total 44.444 120
> 5.500
Node 4 Category % n; true 30.435 7; false 69.565 16; None 0.000 0; Total 8.519 23
Confident
<= 6.500
Node 5 Category % n; true 20.000 4; false 80.000 16; None 0.000 0; Total 7.407 20
Early Adopter
<= 2.500
Node 6 Category % n; true 100.000 2; false 0.000 0; None 0.000 0; Total 0.741 2
> 2.500
Node 7 Category % n; true 11.111 2; false 88.889 16; None 0.000 0; Total 6.667 18
> 6.500
Node 8 Category % n; true 100.000 3; false 0.000 0; None 0.000 0; Total 1.111 3
> 5.500
Node 9 Category % n; true 45.714 16; false 54.286 19; None 0.000 0; Total 12.963 35
Adventurous
<= 5.500
Node 10 Category % n; true 36.667 11; false 63.333 19; None 0.000 0; Total 11.111 30
> 5.500
Node 11 Category % n; true 100.000 5; false 0.000 0; None 0.000 0; Total 1.852 5
> 5.500
Node 12 Category % n; true 67.391 62; false 32.609 30; None 0.000 0; Total 34.074 92
Adventurous
<= 3.500
Node 13 Category % n; true 33.333 6; false 66.667 12; None 0.000 0; Total 6.667 18
Skeptics Wrong
<= 5.500
Node 14 Category % n; true 20.000 3; false 80.000 12; None 0.000 0; Total 5.556 15
> 5.500
Node 15 Category % n; true 100.000 3; false 0.000 0; None 0.000 0; Total 1.111 3
> 3.500
Node 16 Category % n; true 75.676 56; false 24.324 18; None 0.000 0; Total 27.407 74
American Cars Rule
<= 3.500
Node 17 Category % n; true 33.333 4; false 66.667 8; None 0.000 0; Total 4.444 12
Adventurous
<= 4.500
Node 18 Category % n; true 0.000 0; false 100.000 6; None 0.000 0; Total 2.222 6
> 4.500
Node 19 Category % n; true 66.667 4; false 33.333 2; None 0.000 0; Total 2.222 6
No Time For Charity
<= 4.500
Node 20 Category % n; true 0.000 0; false 100.000 2; None 0.000 0; Total 0.741 2
> 4.500
Node 21 Category % n; true 100.000 4; false 0.000 0; None 0.000 0; Total 1.481 4
> 3.500
Node 22 Category % n; true 83.871 52; false 16.129 10; None 0.000 0; Total 22.963 62

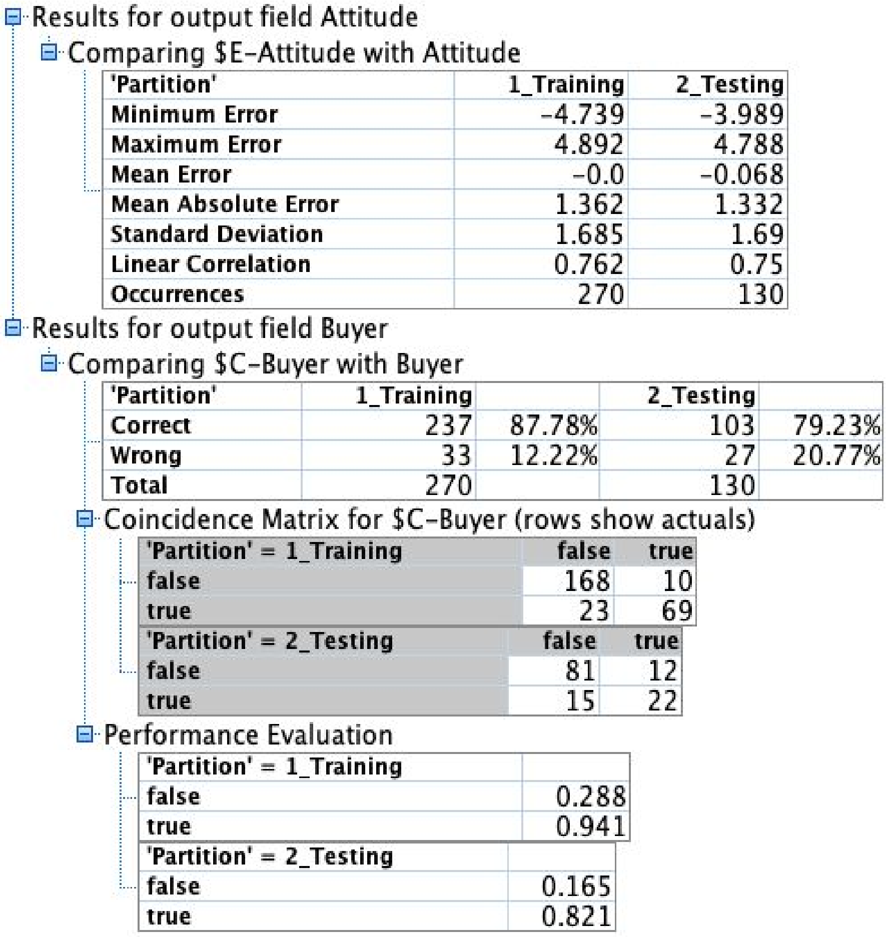
The full decision tree, generated by the C 5.0 Tree Node (Above)

A reduced decision tree, generated by the C 5.0 Tree Node (Below) depicting the likely characteristics of those who are more likely to be identified as a prospective buyer.



Based on this decision tree, we are able to conclude that a good, prospective buyer is someone who considers themselves to be an adventurous risk taker, who favor American cars and feel they don't have time for charity.

9. Performance Evaluation (1) – Confusion Matrix and Derived Metrics. Derive proper performance metrics considering the (balanced / unbalanced) nature of the data set.



Results for output field Attitude

Comparing $E-Attitude with Attitude

'Partition'	1_Training	2_Testing
Minimum Error	-4.739	-3.989
Maximum Error	4.892	4.788
Mean Error	-0.0	-0.068
Mean Absolute Error	1.362	1.332
Standard Deviation	1.685	1.69
Linear Correlation	0.762	0.75
Occurrences	270	130

Results for output field Buyer

Comparing $C-Buyer with Buyer

'Partition'	1_Training		2_Testing	
Correct	237	87.78%	103	79.23%
Wrong	33	12.22%	27	20.77%
Total	270		130	

Coincidence Matrix for $C-Buyer (rows show actuals)

'Partition' = 1_Training	false	true
false	168	10
true	23	69
'Partition' = 2_Testing	false	true
false	81	12
true	15	22

Performance Evaluation

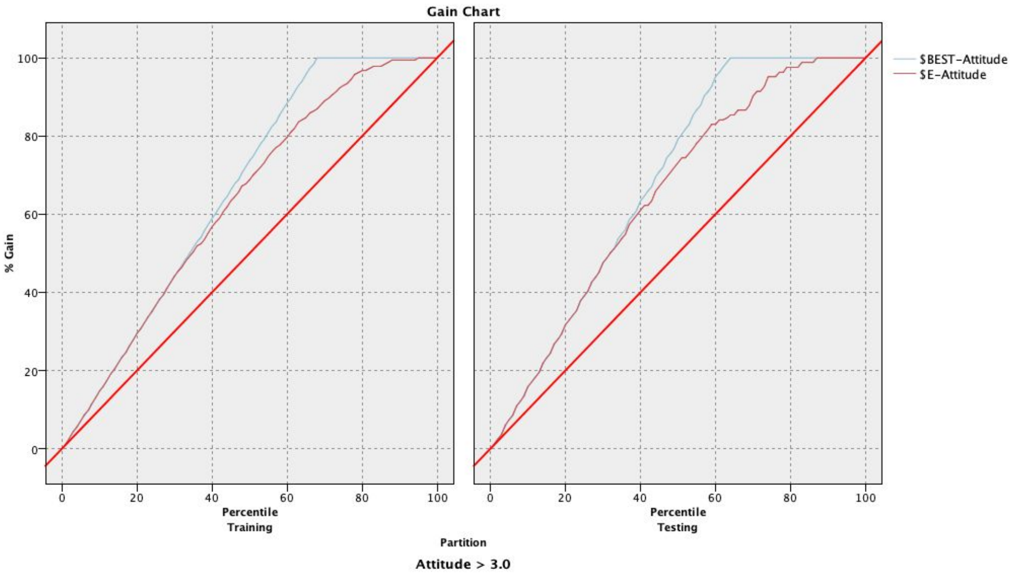
'Partition' = 1_Training	
false	0.288
true	0.941
'Partition' = 2_Testing	
false	0.165
true	0.821

Based on this included coincidence Matrix, the model has the following derived metrics:

Accuracy: [(81+22)/(81+15+12+22)] = 79.23%
Recall: [(81) / (81+12)] = 87.1%
Specificity: [(22) / (22+15)] = 59.46%
Precision: [(81) / (81+15)] = 84.38%
F-Statistic: 1- [(22) / (22+15)] = 40.54%

10. Performance Evaluation (2) – Gain Chart. Plot a Gain Chart and explain it

The cumulative gains chart shows the percentage of the overall number of cases in a given category gained by targeting a percentage of total cases. The diagonal line shown is the baseline. The other two lines are showing the lift curve which uses the predictions of the response model to calculate the percentage of positive responses. Below we show two charts, one for testing and one for training. Each chart has a $BEST - Attitude and $E - Attitude lift curve. We can determine what to expect from each category by looking closely at the points on the curve.



[1]*You can use the statistics node in SPSS, but you may profit from using the Data Analysis add-in in Excel. In there, choose the correlation option, and it will build for you a nice two-way table with all the correlations (aka correlation matrix)*