A Comparison of Data Mining Methods for Binary Response Variables in Direct Marketing

John Sparks
The University of Illinois, Chicago

Background & Introduction

Over my 15 years as an analyst in the direct- and database-marketing industry I was exposed to a number of data mining techniques for predicting response such as CHAID and Neural Networks. This in addition to some of the classic statistical techniques: logistic regression, linear regression and k-nearest-neighbor. Authors and presenters discussing these modeling techniques have been happy to show examples where their algorithm beat the competition, but the analyst could not know if the example file was selected precisely because it provided favorable results for that technique. This has been true both of conference presenters and authors publishing in academic journal. In short, the present state of the literature makes generalization of model performance comparisons difficult. This research aims to directly address this issue. <u>Analyzing nine files from the direct marketing industry using five different algorithms will provide more generalizable results than the comparatively narrower studies seen in past literature.</u>

The remainder of this document is organized as follows: A review of current research is presented next. Section 2 provides details of the experiment to-date, including a brief description of the modeling techniques and data files. Results are contained in Section 3. The document concludes with a discussion of next steps for the research.

Section 1: Literature Review

The major work in the field of comparative performance of modeling & data mining algorithms was completed by Michie, Spiegelhalter & Taylor (1994). This ambitious research analyzed 21 datasets using 23 algorithms. Due to the large number of algorithms and datasets, interpretation of comparative performance was complex. The authors noted that a hybrid regression-neural net algorithm called DIPOL92 had strong performance, as did a density estimating algorithm called ALLOC80. These newer algorithms—that are rather assumption free—were quickly followed, however, by a few variations of discriminant analysis. Disciminant analysis is the classic statistical technique with rather specific assumptions about the relationships in the data.

One draw back to this major comparative study was the wide range of sources and characteristics of the datasets used. The datasets ranged from one with a binary dependent variable, 7 independent variables and 20,000 rows, to another with a 91-level dependent variable, 56 independent variables and 7,080 rows. The datasets were also from a wide range of settings such as satellite imagery, credit analysis and gene identification. This makes it difficult to assess relative algorithm performance within one data origin, such as direct marketing.

A number of articles have focused on model performance within the direct-marketing area. These studies have been narrower, however, typically focusing on one or two datasets and a small number of algorithms. The articles below are representative of the publications in this area since 2000. Note that these comparative studies focused on predictive situations where the dependent variable was binary—response or non-response. Binary response models are the focus of this research as well.

McCarty & Hastak (2007) compared the performance of logistic regression, CHAID & RFM (Recency, Frequency & Monetary Value—a deductive selection approach) using two direct marketing files. They found little difference in performance but observed that logistic regression identified a smaller proportion of responders at small file depths (20% and 30%).

Levin & Zhavi (2001) executed two deductive selection methods, CHAID, a genetic algorithm, a modified AID algorithm and logistic regression on a single file from the collectibles industry. In contrast to McCarty & Hastak they generally found logistic regression superior to the other techniques.

Linder, Geier & Kolliker (2004) constructed 8 simulated datasets ranging in size from 1,100 to 11,000 and tested the performance of logistic regression, classification trees and neural networks. They found that neural networks had the best performance overall, but that logistic regression performed well when the complexity of the data was higher.

Diechman et al. (2002) analyzed one dataset using MARS basis functions as input to a logistic regression and linear regression model. They observed that using the basis functions improved lift performance vs. building a stepwise logistic regression model with the original independent variables.

In summary, the comparison of modeling techniques in the direct marketing industry has been a topic of interest in recent years. The publications to-date within the industry, however, have focused on a small number of techniques and files. The state of the literature would be improved by a study that compares a larger number of techniques and files in order to produce more generalizable results. Also, an understanding of how the characteristics of the independent variables co-vary with relative model performance would be helpful. This research will provide such a study, the details of which are discussed in the next section.

Section 2: Experimental Design and Data

This experiment compares the performance of five different data-mining algorithms for predicting binary response. Each of these algorithms is executed on nine different files from the direct marketing industry. All the files were available for academic study or in the public domain. Four of them were obtained from the Direct Marketing Education Foundation (DMEF). A detailed set of file characteristics is provided in the appendix.

Due to the large number of new and established data mining techniques, one experiment cannot include all data mining algorithms. Therefore the selection of which data mining techniques to include is inherently somewhat arbitrary. The algorithms included here are based on the author's experience as an analyst in the direct marketing industry, availability in the major software packages used in private industry and the potential they represent to provide new insight on the research questions explored. Specifically, the five data mining algorithms used were logistic regression, linear regression, CHAID, neural network, and k-nearest neighbor (k-NN). Space does not permit a description of each of these algorithms. The interested reader may refer to Berry & Linoff (1997) for a business oriented discussion of most of these techniques or Hastie, Tibshirani & Friedman (2001) for a more detailed statistical description.

An unavoidable question when building a model for a binary dependent variable is, which independent variables should one submit to the model? For linear and logistic regression, the analyst can include all the variables and use a selection algorithm (forward, backward or stepwise) to select a set of statistically significant predictors. For a number of the other algorithms, using a large number of predictor variables may present too much noise to the algorithm and thus decrease predictive performance. The analyst may therefore choose to reduce the number of predictor variables by submitting the factors from the independent variables with an eigen value greater than one. Alternatively, these factors can be computed and the two variables most strongly associated with each factor can be used as model predictors. On a more heuristic basis, the analyst may determine how strongly each individual predictor is related to response and submit only a subset of the strongest variables. All four methods are tested here. Specifically the data forms used are referred to as:

1. All variables.
2. Factors.
3. Top two variables per factor.
4. Top 15 variables (The set of 15 variables with the strongest individual relationship with response as measured by the Schwarz Criterion or BIC).

Dependent Measure for Analysis

This study will focus on a business-driven measure for the performance of the algorithms previously discussed. In direct marketing one builds predictive models in order to increase the efficiency of the marketing expense. That is, the marketer seeks to maximize the number of sales for a given quantity of prospects contacted. One way to measure this is to examine the proportion of total responses captured in the top 50% of the individuals scored by the model. Absent any background regarding the particular marketing campaign, however, we cannot know if 50% is the appropriate percent of prospects at which to measure vs. 25%, 75%, etc. One way of circumventing the problem of selecting this percentage is to measure the total area under the gains curve.

Additional Methodological Notes

A 50/50 split for the analysis and holdout sample was used throughout this experiment. So, all dependent measures presented here are for the holdout sample only. Most of the data mining algorithms discussed here require that all the input variables to the model be quantitative. As a result all categorical variables from the original file were converted into a series of binary indicator variables so that they could be used by all the algorithms.

Ladder of powers was practiced to find the best fitting form between each of the predictor variables and the response variable. That is, the Schwarz Criteria was computed after building a simple logistic regression model between response and the following forms of each predictor variable: x, x^2 and x^3; x and x^2; x; log x; square root of x ;inverse of x. The form with the lowest Schwarz Criteria was used for logistic and linear regression.

Data Set Characteristics

While it will be informative to measure which data mining technique has the strongest overall performance, a logical extension is to examine how performance of the different algorithms relates to characteristics of the data sets. For example, does the proportion of binary variables in the data set favor neural networks vs. logistic regression? To that end, a number of characteristics were computed for each of the data sets used in this experiment[1]:

1. Response rate.
2. Number of variables.
3. Number of observations.
4. Ratio of observations to variables.
5. R-square from the analysis sample (see global R-square in Section 3).
6. Proportion of binary variables.
7. Number of factors with eigen values greater than one.
8. Ratio of variables to factors.
9. Mean and median r-squared (Pearson) between the binary dependent variable and the scored records from a simple model for each individual independent variable. For example, a logistic regression model was built for response using one of the independent variables. Each record was scored using the model. The correlation between the model scores and the binary dependent variable was calculated as a rough measure of association. This process was repeated for all the independent variables and the average and median correlation stored for each file. In order to account for the fact that the best relationship was not always linear, this was also done for each variable using the best model after ladder of powers adjustment.

To summarize, this experiment tests 6 different algorithms for predicting binary response using 9 different datasets and 4 data forms. This provides 216 data points for analysis. The primary

[1] The appendix contains a summary table of these and other measures for each data set.

dependent measure for analysis is the area under the gains curve from the holdout sample. The proportion of total responses contained in the top 25%, 50% and 75% of scored individuals is also available as a metric and will be used in the discussion of some results. A number of descriptive measures for each of the data sets will also be used to examine which of them co-varies with predictive ability of the data mining techniques.

Section 3: Results and Analysis

Results of the experiment to-date are discussed here. The principle findings are:

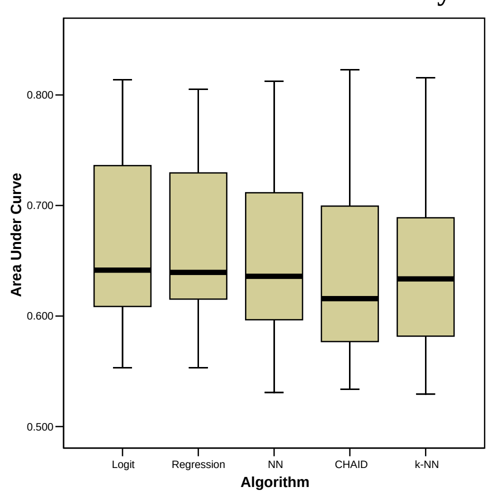
1. The statistical tests for algorithm and form are significant at alpha=0.05. Comparisons of the area under the gains curve indicate that linear methods using the top 15 variables with the strongest relationship to response are the best choice.
2. The area under the gains curve for the holdout sample can generally be predicted using the global R-squared (defined below) and the response rate for the file. The proportion of total responses from the top 25%, 50% and 75% of the scored holdout sample are also predictable using this model.

Algorithm and Form

Distribution of Area Under The Gains Curve. The two principle effects that are examined in this experiment are algorithm and data form. An analysis of variance statistical test for these factors will be examined in the next subsection. However, we can first examine the distributions using graphical summaries.

The box-plots in the figure below show the distributions for the area under the gains curve for the five data mining algorithms[2]. The distributions are sorted from left to right by descending mean area.

Distributions of Area Under Gains Curve by Algorithm.

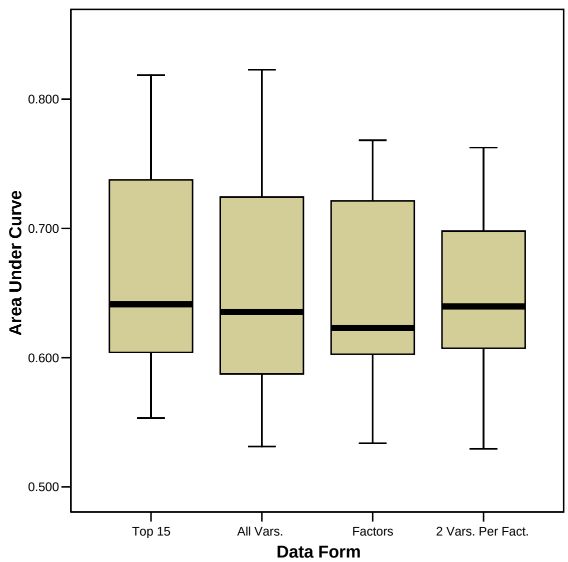


[2] There are 9 files and 4 data forms, so each box-plot is based on 36 data points.

Inspection of the graph shows that no one algorithm is obviously superior. The distributions are generally not that different, with fairly similar spreads and medians that appear roughly equivalent. Interestingly, the two linear methods (logistic regression and linear regression) have the best overall performance, with very little difference between the two apparent from the graph.

These findings are remarkable because they suggest, based only on an inspection of these box-plots, that the different algorithms do not produce remarkably different distributions of results for these nine direct marketing files. Thus a direct marketing modeler could choose almost any of these algorithms with little reason to believe that using a different algorithm would provide remarkably superior results. However, given the widespread availability of logistic and linear regression and their superior performance here, either of these would be the preferred choice. The distributions of results by data form are similarly shown on the box-plot below.

Distributions of Area Under Gains Curve by Data Form



Again, no remarkable difference in the distributions is obvious from the plot. The inter-quartile range of the top 15 form, however, appears to have a higher location than the others. The same general but tentative conclusion can be drawn from this plot: although the top 15 form has the best apparent performance, the direct marketing modeler can use any of these data forms with an expectation that using one form vs. another would not have a dramatic impact on the amount of area under the gains curve.

The performance of the top 15 form has an interesting additional implication. It appears that using the 15 strongest independent variables provides the same if not superior predictive power vs. submitting all the variables to the algorithm, or a linear representation of all the variables. This is important because submitting only 15 variables has the potential to remarkably reduce processing time for a number of the algorithms and to reduce the complexity of the final model.

These plots show that not much difference in the distributions of the area under the gains curve is obvious by visual inspection. It remains to determine, however, whether the differences are statistically significant. This will be analyzed next.

Repeated Measures Analysis of Variance. Because all five algorithms were run on each of the nine files a repeated measures ANOVA is appropriate. Similarly, data form is treated as a fixed effect that was repeated on each of the nine 'subjects', in this case, data files. The ANOVA table is included in the appendix. The p-values from the ANOVA were 0.007 for algorithm and 0.03 for form[3]. So, statistically speaking, both of these effects are related to average area under the gains curve, with algorithm demonstrating a stronger statistical relationship than data form.

The next obvious question is, which algorithms and forms are statistically different from the others? The table below shows the average area by algorithm. Also shown is the p-value for the difference between the algorithm and the average area for the algorithms listed below it. For example, the 0.015 on the first row is the p-value for the contrast between the area for logistic regression and the average area for all four other algorithms. The .012 is the p-value for the difference between linear regression and the remaining three algorithms listed below it.

Algorithm	Avgerage Area	P-Value For Avg Below
Logistic Regression	0.6641	0.015
Linear Regression	0.6633	0.012
Neural Network	0.6480	0.100
CHAID	0.6430	0.530
k-Nearest Neighbor	0.6376	

The average area for logistic regression is only eight basis points greater than that for linear regression; there is a decrease of 153 basis points, however, between linear regression and neural network. The p-values indicate that there is a strong difference between the average performance of logistic regression vs. the other algorithms and for linear regression vs. the others. The p-value for the difference between neural network and the remaining two algorithms is borderline significant. These findings suggest that logistic regression, linear regression and neural net are statically better choices than CHAID & k-NN. Given the widespread availability of the linear methods and their strong statistical performance, the direct marketing analyst has every incentive to use them, albeit after practicing ladder of powers.

The average areas by data form are shown below.

[3] P-values in the ANOVA are based on the Greenhouse & Geisser adjustment to the degrees of freedom for lack of symmetry in the covariance matrix.

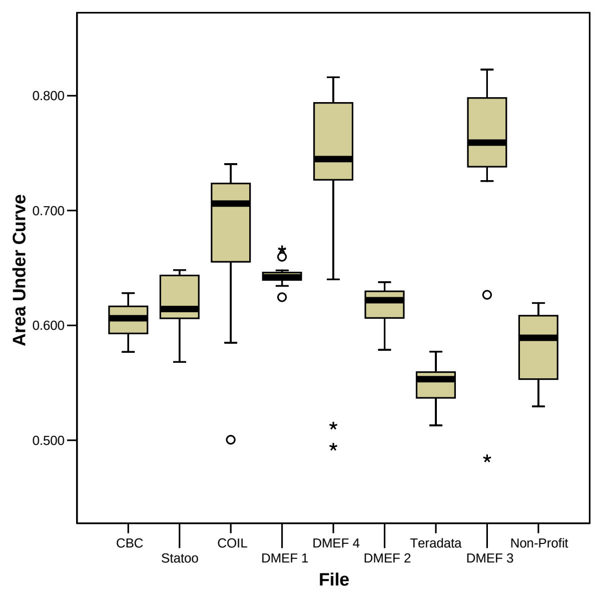
Form	Average Area	P-Value For Avg Below
Top 15 Variables	0.6576	0.022
All Variables	0.6476	0.151
Factors	0.6357	0.934
2 Vars. Per Factor	0.6334	

There is a 100 basis point drop from the top 15 form to the all variables form and a 119 basis point drop from all variables to factors. The only remaining decrease in average area is the much smaller 23 basis point difference between factors and two variables per factor. The p-values indicate that the top 15 form is significantly different from the average of all the remaining forms. As mentioned previously, limiting the number of variables decreases computation time and potential complications of the model. Therefore, these results and those practical considerations provide a strong argument for limiting the number of variables to the initial best 15 rather than attempt to reduce the number of variables based on some variation of principle components.

Model Strength and File Characteristics

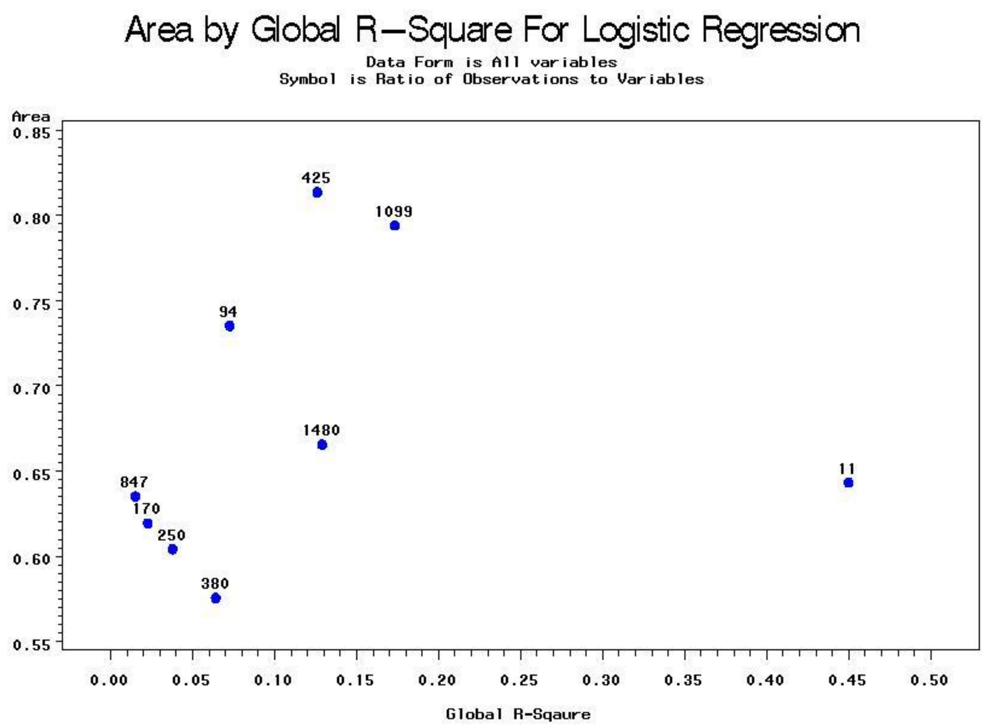
Distribution of Area Under Gains Curve, Average Correlation & Global R-Squared. The box plot below shows the distribution of the area under the gains curve for each file. Some files have stronger overall performance (DMEF 3) than others (Teradata). This leads us to question whether we can use a characteristic of the file to predict how well the algorithms will be able to perform.

Box-Plots for Area Under Gains Curve by File.
24 Observations Per File (6 Algorithms by 4 Data Forms)



One would expect that a file that has independent variables that are more strongly related to the dependent variable will produce a model with greater area under the gains curve. To examine this, the relationship between the area under the gains curve and the average correlation of the binary dependent variable with a simple model of each of the independent variables was examined. The results did not show a particularly strong relationship. The area under the gains curve, however, is a result of building a model using a combination of multiple input variables. The average correlation is a measure of center for the entire set of individual input variables. A measure of the combined correlation of the input variables to the dependent variable should provide a better predictive measure for the area under the gains curve.

One rough measure of the combined strength of this relationship is the R-square calculated by building a regression model using all the input variables available in the file. This model was built for the analysis sample of each file[4]. This measure is referred to hereafter as the global R-square. The figure below shows the relationship between the global R-square from the analysis sample and the area under the gains curve from the holdout sample for logistic regression using the all variables form. The number listed next to each data point is the ratio of observations to variables for the file.



Two observations are apparent from the figure:

- The file with 11 observations per predictor variable (file two) is an outlier.

[4] After conversion of categorical to binary variables and elimination of linearly redundant variables but before practicing ladder of powers.

- Ignoring this outlier, a relationship is apparent between the global R-square of the analysis sample and the area under the gains curve for the holdout sample.

In fact, ignoring file two gives an R-square using the remaining eight files of 0.56 for a regression of global R-square on the area under the gains curve.

Further, for all data forms and algorithms a significant linear relationship was found between these two measures. The r-square varied from 0.44 to 0.70. A table containing summaries of the regression relationships is available in the appendix. A significant relationship was also found between the global R-square and the proportion of total responses from the top 25%, 50%, and 75% of the scored holdout sample. When the response rate of the file was added to the relationship, the R-squares increased to around 0.85 to 0.90. The models including the response rate are not discussed further, however, due to the possibility of over-fitting given the small number of files used in the regression relationship.

It is not surprising that the area under the gains curve from the holdout sample is related to the global R-squared from the analysis sample: both are measures of the predictive ability of a model. The significance of this finding, however, is the quantification of the relationship. Using this quantification a direct marketing analyst can calculate the global R-squared from an analysis sample and use it to estimate of the proportion of total responses from the top 75% of the scored holdout sample, before beginning any actual modeling.

Section 4: Summary & Conclusions

In conclusion, this study analyzed the performance of five data mining algorithms using four data forms and nine files from the direct marketing industry. Results showed that linear methods and the top 15 variables had the strongest results. The area under the gains curve (a key metrics used in the direct marketing industry) is highly predictable using the global R-squared of the file and the response rate.

The direct marketing analyst can use the results to follow the following process. First, convert all variables to a quantitative representation. Then divide the file into analysis & holdout samples and calculate the global R-squared for the analysis sample. Practice ladder of powers using the analysis sample and build a regression model using the top 15 variables. The actual area under the gains curve can then be compared to the predicted value using the regression equation available in the appendix to this document. If there is an unacceptable difference between the actual and predicted values, then one of the other algorithms can potentially be used to build a model and the comparison repeated for confirmation.

Some next steps building from this base of research include a comparison to new data mining algorithms and use of monte carlo simulation to increase the number of data sets.

A number of new data mining algorithms have become available in the last few years, but are not widely available in the major software packages used in the direct marketing industry (SAS &

SPSS). Clearly it would be interesting to know if any of them increase model fit, the area under the gains curve, and ultimately return from marketing spending. These techniques include MARS, Random Forests, Latent Class Modeling and General Additive Models.

To address the fact that only nine data sets were available for analysis, an alternative next step would be to build a larger number of data sets using monte carlo methods. The nine data files used here would serve as guides for the general composition of the data sets so that the simulated data would still be representative of data seen in the direct marketing industry. Building a set of simulated data sets would also allow for an experiment where a number of the key data set characteristics (such as the number of variables, observations, response rate, ratio of observations to variables & global R-squared) would vary in a more controlled fashion. Additionally, basing the experiment on a larger number of data files would make the results more generalizable.

Appendix

Detailed Data Set Characteristics.

File Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Description	CBC--Book Club	Statoo Case Study--Retail Bank	COIL Tournament (Specialty Insurance)	DMEF 1 (Non-Profit)	DMEF 4 (Specialty Cataloger)	DMEF 2 (Multi-Division Cataloger)	Teradata Cell Phone Churn	DMEF 3 (Gift Cataloger)	Non-Profit
Number of Observations	4,000	2,158	9,822	99,180	101,094	96,550	100,000	106,284	94,001
Number of Variables	16	195	104	67	92	114	263	250	554
Ratio of Observations to Variables	250	11	94	1,480	1,099	847	380	425	170
Response Rate	9.2%	49.4%	5.9%	27.4%	9.1%	2.4%	49.5%	5.4%	5.2%
Number of Binary Variables	4	26	59	22	44	58	161	133	203
Percent Binary Variables to Total Variables	25%	13%	57%	33%	48%	51%	61%	53%	37%
Number of Factors (Egien>1)	5	37	44	15	28	30	93	55	144
Ratio of Variables to Factors	3.2	5.3	2.4	4.5	3.3	3.8	2.8	4.5	3.8
Mean Correlation of Independent to Dependent Variables: Linear Only	0.05	0.12	0.03	0.09	0.07	0.03	0.02	0.03	0.01
Median Correlation of Independent to Dependent Variables: Linear Only	0.04	0.11	0.03	0.06	0.05	0.02	0.01	0.01	0.01
Mean Correlation of Independent to Dependent Variables: After Ladder of Powers	0.06	0.13	0.03	0.09	0.09	0.03	0.03	0.03	0.01
Median Correlation of Independent to Dependent Variables: After Ladder of Powers	0.05	0.13	0.03	0.06	0.05	0.03	0.02	0.02	0.01
Global R-Square	0.04	0.45	0.07	0.13	0.17	0.02	0.06	0.13	0.02

The ANOVA Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Anova SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
Algorithm	4	0.02070634	0.00517659	5.88	0.0012	0.0071	0.0018
Error(Algorithm)	32	0.02819175	0.00088099				

Greenhouse-Geisser Epsilon	0.6135
Huynh-Feldt Epsilon	0.9054

Source	DF	Anova SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
Preparation	3	0.01711755	0.00570585	4.39	0.0135	0.0300	0.0173
Error(Preparation)	24	0.03120907	0.00130038				

Greenhouse-Geisser Epsilon	0.6703
Huynh-Feldt Epsilon	0.8960

Source	DF	Anova SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
Algorithm*Preparatio	12	0.00995402	0.00082950	4.72	<.0001	0.0151	0.0049
Error(Algorithm*Preparatio)	96	0.01686781	0.00017571				

Greenhouse-Geisser Epsilon	0.2109
Huynh-Feldt Epsilon	0.3166

Regression of Area Under Gains Curve on Global R-Squared

Algorithm	Form	Intercept	Intercept P-Value	Slope	Slope P-Value	R-Squared
Chaid	All Variables	0.55	<.0001	1.45	0.017	0.64
Knn	All Variables	0.54	<.0001	1.21	0.012	0.68
Logit	All Variables	0.59	<.0001	1.18	0.032	0.56
NN	All Variables	0.53	<.0001	1.35	0.013	0.67
Regression	All Variables	0.59	<.0001	1.15	0.026	0.59
Mean		**0.56**		**1.27**		
Chaid	Top 15 Variables	0.56	<.0001	1.35	0.025	0.59
Knn	Top 15 Variables	0.56	<.0001	1.37	0.018	0.64
Logit	Top 15 Variables	0.58	<.0001	1.14	0.048	0.50
NN	Top 15 Variables	0.57	<.0001	1.24	0.040	0.53
Regression	Top 15 Variables	0.59	<.0001	1.06	0.056	0.48
Mean		**0.57**		**1.23**		
Chaid	Factors	0.54	<.0001	1.06	0.020	0.62
Knn	Factors	0.55	<.0001	1.03	0.022	0.61
Logit	Factors	0.59	<.0001	0.90	0.074	0.44
NN	Factors	0.58	<.0001	0.92	0.067	0.46
Regression	Factors	0.59	<.0001	0.85	0.075	0.43
Mean		**0.57**		**0.95**		
Chaid	2 Vars. Per Factor	0.56	<.0001	0.98	0.021	0.61
Knn	2 Vars. Per Factor	0.55	<.0001	0.96	0.036	0.55
Logit	2 Vars. Per Factor	0.59	<.0001	0.87	0.070	0.45
NN	2 Vars. Per Factor	0.55	<.0001	1.00	0.030	0.57
Regression	2 Vars. Per Factor	0.59	<.0001	0.82	0.069	0.45
Mean		**0.57**		**0.93**		

Cited Literature

Berry, M., Linoff, G., (1997) *Data Mining Techniques*. New York: John Wiley & Sons, Inc.

Clementine 10.1 Algorithms Guide. Integral Solutions Limited.

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., Teebagy, N. (2002) 'Application of Multiple Adaptive Regression Splines (MARS) in Direct Response Modeling', *Journal of Interactive Marketing*, 16 (4), 15-27.

Freidman, J (1991) 'Multivariate Adaptive Regression Splines," Annals of Statistics, 19 (1), 1-67.

Hatsie, T., Tibshirani, R., Friedman, J. (2001) *The Elements of Statistical Learning*. New York: Springer.

Levin, N., and Zahavi, J. (2001) 'Predictive Modeling Using Segmentation', *Journal of Interactive Marketing*, 15 (2), 2-22.

Linder, R., Geier, J., Kolliker, M. (2004) 'Artificial Neural Networks, Classification Trees and Regression: Which Method for Which Customer Base', *Database Marketing & Customer Strategy Management*, 11(4), 344-356.

McCarty, J., and Hastak, M. (2007) 'Segmentation Approaches in Data-Mining: A Comparison of RFM, CHAID, and Logistic Regression', *Journal of Business Research*, 60 (6), 656-662.

Michie, D., Spiegelhalter, D., Taylor, C. (1994). *Machine Learning, Neural and Statistical Classification*. Hertfordshire, England: Ellis Horwood,

Neslin, S., Gupta, S., Kamakura, W., Lu, J., Mason, C. (2006) 'Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models', *Journal of Marketing Research*, 43 (2), 204-211.

Witten, H., Frank, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufman.

Weiss, S., Indurkhya, N., (1998). *Predictive Data Mining: A Practical Guide*. San Francisco: Morgan Kaufman.