SAS® GLOBAL FORUM 2015

The Journey Is Yours

Increasing Profitability through Cluster Analysis and Simplicity:

Follow the Money

SAS GLOBAL FORUM

Increasing Profitability through Cluster Analysis and Simplicity: Follow the Money

Sherrie Rodriguez
Advising Faculty: Dr. Jennifer Priestley
Kennesaw State University

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Introduction

Developing a quality product or service, while at the same time improving cost management and maximizing profit, are challenging goals for any company. Finding the optimal balance between *efficiency* and *profitability* is not easy. The same can be said in regards to the development of a predictive statistical model. On the one hand, the model should predict as accurately as possible. But, on the other hand, having too many predictors can end up costing company money in the long run. One of the purposes of this project is to explore the cost of simplicity. When is it worth having a simpler model, and what are some of the costs of using a more complex one? The answer to that question leads us to another one: How can a predictive statistical model be maximized in order to increase a company's profitability?

Abstract

Using data from the consumer credit risk domain provided from CompuCredit, logistic regression was used to build binary classification models to predict the likelihood of default. This project compares two of these models. Although the original dataset had several hundred predictor variables and more than a million observations, I chose to use rather simple models. My goal was to develop a model with as few predictors as possible, while not going lower than a concordant level of 80%. Two models were evaluated and compared based on efficiency, simplicity, and profitability. Using the selected model, cluster analysis was then performed in order to maximize the estimated profitability. Finally, the analysis was taken one step further through a supervised segmentation process, in order to target the most profitable segment of the best cluster.

Objectives

- Develop an efficient statistical model that would predict the likelihood of default with as few predictors as possible
- Estimate the profitability that would be generated by such a model
- Maximize the estimated profitability by identifying the most profitable customers

Steps to Profitability

1. Data Preparation
   - Creation of binary response variable
   - Data cleansing and imputation
   - Multicollinearity assessment using regression and Variance Inflation Factor (VIF)
   - Variable transformations
   - Variable reduction
2. Modeling
   - Sampling
   - Model Development
   - Model testing and evaluation
   - Model comparison
3. Observation Segmentation
   - Unsupervised Cluster Analysis
   - Supervised Data Segmentation
   - Identifying the most profitable subpopulation to target

Increasing Profitability through Cluster Analysis and Simplicity: Follow the Money

Sherrie Rodriguez
Advising Faculty: Dr. Jennifer Priestley
Kennesaw State University

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Method 1: Logistic Regression

- Classification models were developed to predict the likelihood of default using the binary response variable GoodBad.
- Before building the models, random samples were taken from the dataset to create two independent files, a training file and a validation file.
- The goal was to build a model with as few predictors as possible, while not going below a concordant level of 80%.
- 131 predictors were initially placed into a model, using the backward selection option in the Proc Logistic procedure.
- Variables showing no effect were removed.
- Some remaining predictors included multiple versions of the same variable, due to variable transformations in the data preparation phase.
- Variables were selected based on the highest Chi-Square value.
- Several models were developed and tested.

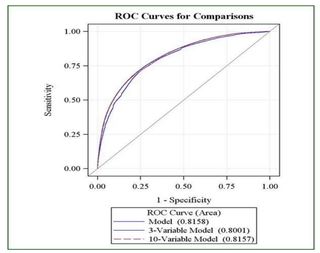
Model Comparison

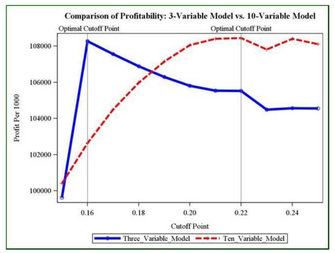
- Through the process of model development and validation, two models were selected for comparison.
- The two models were compared both on efficiency and profitability.
- Profitability reports were generated for each model using a profitability function.
- The cost of simplicity was an important factor in determining which model would be the best.

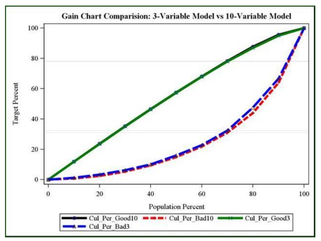
Results

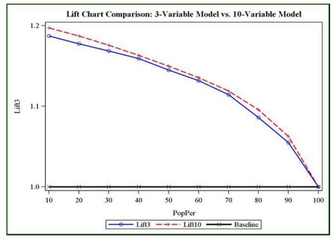
- The four charts provide comparisons of a 10-variable model versus a 3-variable model.
- The ROC curve shows that there is very little difference between the probabilities of the two models.
- The Gains and Lift charts show only a small advantage of the 10-variable model over the simpler one.
- The 10-variable model shows an expected profit of only $178.45 more (per 1000 customers) than the expected profit using the 3-variable model.

MODEL COMPARISON BY THE NUMBERS					
Model	Area Under Curve	KS Statistic	Max Lift	Optimal Cutoff Point	Profit per 1,000 Customers
10-Variable	.8157	47.34	1.197	0.22	$108,450.26
3-Variable	.8001	45.67	1.187	0.16	$108,271.81
Difference	.0156	1.67	0.01	0.06	$178.45



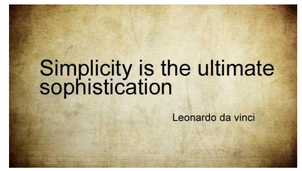






Discussion

At first glance, it may appear that the 10-variable model has a slight advantage over its competitor. However, there are other costs involved that the graphs do not show. First of all, there is the cost in terms of time. It takes time to collect data, so it is reasonable to assume that it would cost additional time if a more complex model was selected. Data collection also costs money, and the more variables there are in a model, the more data would need to be acquired. To gather the data needed for a 10-variable model would cost roughly three times more than the cost of collecting data for a model with only three predictors. Is the cost of simplicity worth it? In this specific case, I believe it is. In general, however, it all depends on whether or not the costs that are saved with a simpler model outweigh the benefits of a more complex one.



The model with three predictors was selected, and the profitability of $108,271.81 per 1000 customers served as a baseline for the remaining analyses.

Profitability Report Using 3-Variable Model				
outcometype	pct	n	profit	pper1000
ERROR1	0.0530854	39974	$-36,861,167.50	$-922,128.57
ERROR2	0.1953964	147136	$0.00	$0.00
VALID1	0.1226247	92338	$0.00	$0.00
VALID2	0.6288935	473565	$118,391,250.00	$250,000.00
	1	*753013*	*$81,530,082.50*	*$108,271.81*

Increasing Profitability through Cluster Analysis and Simplicity: Follow the Money

Sherrie Rodriguez
Advising Faculty: Dr. Jennifer Priestley
Kennesaw State University

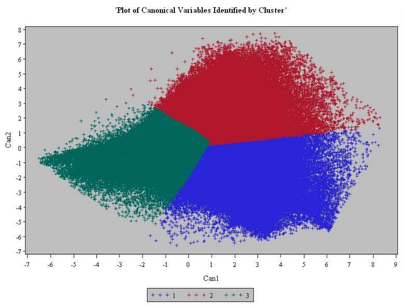
Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

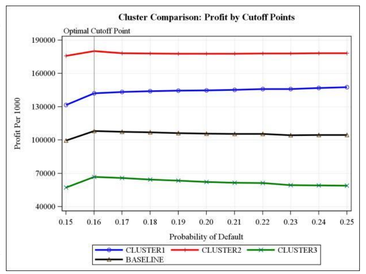
Method 2: Cluster Analysis

- This is an unsupervised method used to partition the data into groups based on similar and dissimilar characteristics.
- First the data was standardized.
- A hierarchical method was used to estimate the optimal number of clusters.
- Using K-means, the data was partitioned into three clusters.

Results

- Profitability reports were generated for each cluster.
- The clusters were given names that reflected their characteristics according to the amount of profitability per 1000 customers:
- Prompt Payers (red) ~ $180,239.98
- Common Consumers (blue) ~ $142,139.51
- Despairing Deadbeats (green) ~ $66,889.49
- Targeting the Prompt Payers, the estimated profitability was increased by 66.47% above baseline.





The Model

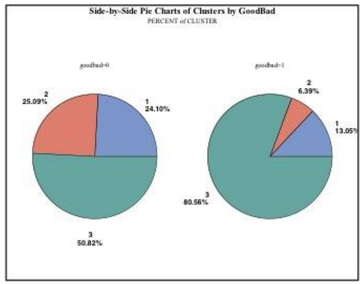
- A simple model with only 3 predictors allows the cluster analysis to go one step further.
- The binary response variable GoodBad was used to predict the probability of default.
- The model has three predictors: one continuous variable and two discrete.
- The model equation is shown below, along with a table containing the variable descriptions.

$$\text{GoodBad} = e^{(-3.86+2.18(BRADBM)+.83(BRR4524)+.60(TRR23))}$$

3-Variable Model: Variable Descriptions		
Variable	**Type**	**Description**
GoodBad	Binary	0 = "Good" customer, 1 = "Bad" customer
Bradbm	Continuous	Maximum open bank revolving utilization ratio
Brr4524	Discrete	Number of bank revolving accounts over 90 days past-due in the past 24 months
Trr23	Discrete	Number of trades in 30 or 60 days overdue

Clusters by GoodBad

Table of CLUSTERS by GoodBad			
CLUSTER	**GoodBad**		
Frequency Percent Row Pct Col Pct	**0**	**1**	**Total**
Common Consumers (Blue)	149561 19.86 89.65 24.10	17264 2.29 10.35 13.05	166825 22.15
Prompt Payers (Red)	155704 20.68 94.85 25.09	8458 1.12 5.15 6.39	164162 21.80
Despairing Deadbeats (Green)	315436 41.89 74.74 50.82	106590 14.16 25.26 80.56	422026 56.04
Total	620701 82.43	132312 17.57	753013 100.00



Method 3: Supervised Segmentation

- Using the two discrete predictors, binary variables were created to form four independent groups.
- Profitability reports were generated for each group.
- The model was tested again multiple times, using different validation files, each time increasing the profitability more than 50% above baseline.

Results

- The following amounts are per 1000 observations scored:
- Cream of the Crop ~ $190,795.07
- Wafflers 1 ~ $133,937.22
- Wafflers 2 ~ $129,259.49
- Tardy Traders~ $180,239.98
- Those who were consistent in paying their bills, whether on time or late, generated the most profit.
- The Cream of the Crop was found to be the most profitable, increasing the profitability an additional 9.75% bringing the total to 76.22% above baseline.
- Gains and Lift charts were of no use with this method since they depend on ratios of the good and the bad.

Conclusions

- There were two things that set the Cream of the Crop apart from the rest of the sample: 1) Over a two year period, they were never more than 90 days late in paying their credit card bills. 2) They had zero trade payments that were 30 or 90 days overdue.
- Through cluster analysis and simplicity, the profitability of a company can be increased significantly by targeting the most profitable subpopulation.

Increasing Profitability through Cluster Analysis and Simplicity: Follow the Money

Sherrie Rodriguez
Advising Faculty: Dr. Jennifer Priestley
Kennesaw State University

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Following the Money

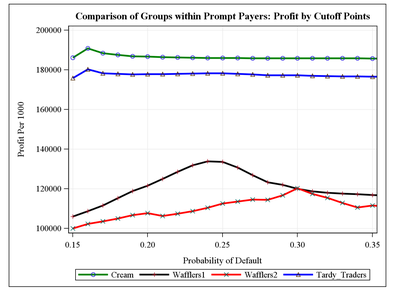
The Cream of the Crop

The frequency table below shows the distribution of the groups within the Prompt Payers. Approximately 88% consists of the Cream of the Crop.

Segmentation of Prompt Payers using Binary Variables Created from Discrete Variables			
	N	Binary_Brr4524	Binary_Trr23
Cream of the Crop	144996	0	0
Wafflers 1	13611	0	1
Wafflers 2	4006	1	0
Tardy Traders	1549	1	1

The profitability report generated by the Cream of the Crop is shown below, along with a graph comparing the four groups.

Profitability Report of the Cream of the Crop				
outcometype	pct	n	profit	pper1000
ERROR1	0.0329526	4778	$-6,157,228.50	$-1,288,662.31
ERROR2	0.0258904	3754	$0.00	$0.00
VALID1	0.0081175	1177	$0.00	$0.00
VALID2	0.9330395	135287	$33,821,750.00	$250,000.00
	1	*144996*	*$27,664,521.50*	*$190,795.07*



Reference Links

Department of Statistics and Analytical Sciences at Kennesaw State University
https://analytics.kennesaw.edu/

Dr. Jennifer Lewis Priestley, Kennesaw State University
Home Page

STAT 4330/8330, Kennesaw State University
https://analytics.kennesaw.edu/~jpriestl/STAT4490/Schedule4490.htm

Increasing Profitability through Cluster Analysis and Simplicity: Follow the Money

Sherrie Rodriguez
Advising Faculty: Dr. Jennifer Priestley
Kennesaw State University

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Sample SAS Code

```
*Logistic Regression Code;
Proc Logistic data = train des outest = sr.betas3
outmodel=sr.scoringdata3
PLOTS(MAXPOINTS=NONE);
model GOODBAD = BRADBM BRR4524 TRR23 /selection = backward
outroc=sr.roc3 CTABLE pprob=(0 to 1 by .1)
LACKFIT RISKLIMITs;
output out = sr.output3 p = predicted;
score data=sr.valid out=sr.score3;
Run;

*Optimal cutoff point was found to be .16;;
data probs;
set score3;
if p_1 ge .16 then preds = 1;
else preds = 0;
run;

*Cluster analysis using K-means method;
%let inputs = BRADBM BRAGE BRHIC;

Proc stdize data = sr.score3 method = range out = temp;
var &inputs;
run;

proc fastclus data=temp maxc=3 outstat =stats out=clus;
var &inputs;
run;

DATA CLUSTER1;
SET CLUS;
WHERE CLUSTER=1;
RUN;

DATA CLUSTER2;
SET CLUS;
WHERE CLUSTER=2;
RUN;

DATA CLUSTER3;
SET CLUS;
WHERE CLUSTER=3;
RUN;
```

```
*Code to create profitability report using values from CompuCredit;
data probs;
set score3;
if p_1 ge .16 then preds = 1;
else preds = 0;
run;

data probs1;
set probs (keep=preds GOODBAD crelim);
crelim2 = crelim/2;
if preds = 1 AND GOODBAD = 0 then outcometype = "ERROR2";
else if preds = 0 AND GOODBAD = 1 then outcometype = "ERROR1";
else if preds =1 AND GOODBAD=1 then outcometype = "VALID1";
else outcometype = "VALID2";
if outcometype = "ERROR1" then profit = -crelim2;
else if outcometype = "ERROR2" then profit = 0;
else if outcometype = "VALID2" then profit = 250;
else if outcometype = "VALID1" then profit = 0;
run;

Proc means data=probs1 sum mean;
var profit;
class outcometype;
Run;

Proc sort data=probs1;
by outcometype;
run;

PROC REPORT DATA= probs1 nowd;
COLUMN outcometype pct n profit pper1000;
DEFINE outcometype /group width  = 8 ;
DEFINE profit /format=dollar15.2;
define pper1000 / computed format=dollar15.2;
/*get the overall number of obs*/
compute before;
  overall=n;
endcomp;

compute pper1000;
pper1000 = (profit.sum/n)*1000;
endcomp;
compute before outcometype;
totaln=n;
endcomp;
compute pct;
pct = (totaln/overall);
if _break_ = '_RBREAK_' then pct= (overall/overall);
endcomp;
rbreak after/summarize dol;
RUN;
```

```
*ROC curve comparison of the two models;
ods graphics on  labelmax=54800;
proc logistic data = sr.train plots=effect
plots=ROC (id=prob)plots(maxpoints=none);
model goodbad(event='1') = BRADBM BRAGE BRCRATE4 BRHIC
BRMINB BRR4524 DCWCRATE TRR23 LODDSBRWCRATE
LODDSTR224
ODDSFFWCRATE ODSEQBRMAXB WCRATE;
roc '3-Variable Model' bradbm brr4524 trr23;
roc '10-Variable Model' BRADBM BRAGE BRCRATE4 BRHIC
BRR4524 DCWCRATE TRR23 LODDSBRWCRATE LODDSTR224
ODDSFFWCRATE;
roccontrast reference('3-Variable Model') /estimate;
run;
ods graphics off;

*Graph comparing gains charts of the two variables;
proc sgplot data=gains;
series x=PopPer y= cul_per_good10 /
markers lineattrs=(color=GREEN)
markers lineattrs=(pattern=solid)
markers lineattrs= (thickness =5);
series x=PopPer y=cul_per_bad10 /
 markers lineattrs= (color=red)
MARKERS LINEATTRS = (THICKNESS = 5);
xaxis type=discrete;
xaxis label="Population Percent";
yaxis label="Target Percent";
series x=PopPer y= cul_per_good3 /
markers lineattrs=(color=green)
markers lineattrs=(pattern=solid)
markers lineattrs=(thickness=5);
series x=PopPer y=cul_per_bad3 /
markers lineattrs=(color=blue)
markers lineattrs = (thickness=5);
refline 32.34 78.01 / transparency =.8 axis=Y;
refline 30.98 78.32 / transparency=.8  axis=Y;
title"Gain Chart: 3-Variable Model";
run;
```

```
*Graph comparing the lift charts of the two variables;
proc sgplot data=lift;
series x=Population_Percent y= lift3 /
markers lineattrs=(color=blue)
markers lineattrs=(pattern=solid)
markers lineattrs= (thickness =2);
xaxis type =discrete min=10 max=100;
series x=Population_Percent y= lift10/
markers lineattrs=(color=red)
markers lineattrs=(thickness=2);
series x=Population_Percent y=Baseline /
markers lineattrs=(color=black)
markers lineattrs=(pattern=solid)
markers lineattrs= (thickness =3);
run;

*Graph comparing the two profitability curves;
proc sgplot data=cutoff;
series x=cutoffpoint y= Three_Variable_Model /
markers lineattrs=(color=blue)
markers lineattrs=(pattern=solid)
markers lineattrs= (thickness =5);
series x=cutoffpoint y=Ten_Variable_Model /
 markers lineattrs= (color=red)
MARKERS LINEATTRS = (THICKNESS = 4);
xaxis type=discrete;
xaxis label="Cutoff Point";
yaxis label="Profit Per 1000";
refline .16 .22 / axis=x
label=('Optimal Cutoff Point' 'Optimal Cutoff Point');
run;

* Side-by-Side pie chart of Clusters by GoodBad;
PROC GCHART DATA = CLUS;
TITLE 'Side-by-Side Pie Charts of Clusters by GoodBad';
PIE cluster / TYPE = PCT  DISCRETE GROUP=GOODBAD
ACROSS=2
slice = OUTSIDE
plabel =(font='Albany AMT/bold' h=1.3);
LEGEND;
RUN; QUIT;
```

SAS® GLOBAL FORUM

April 26-29
Dallas, TX