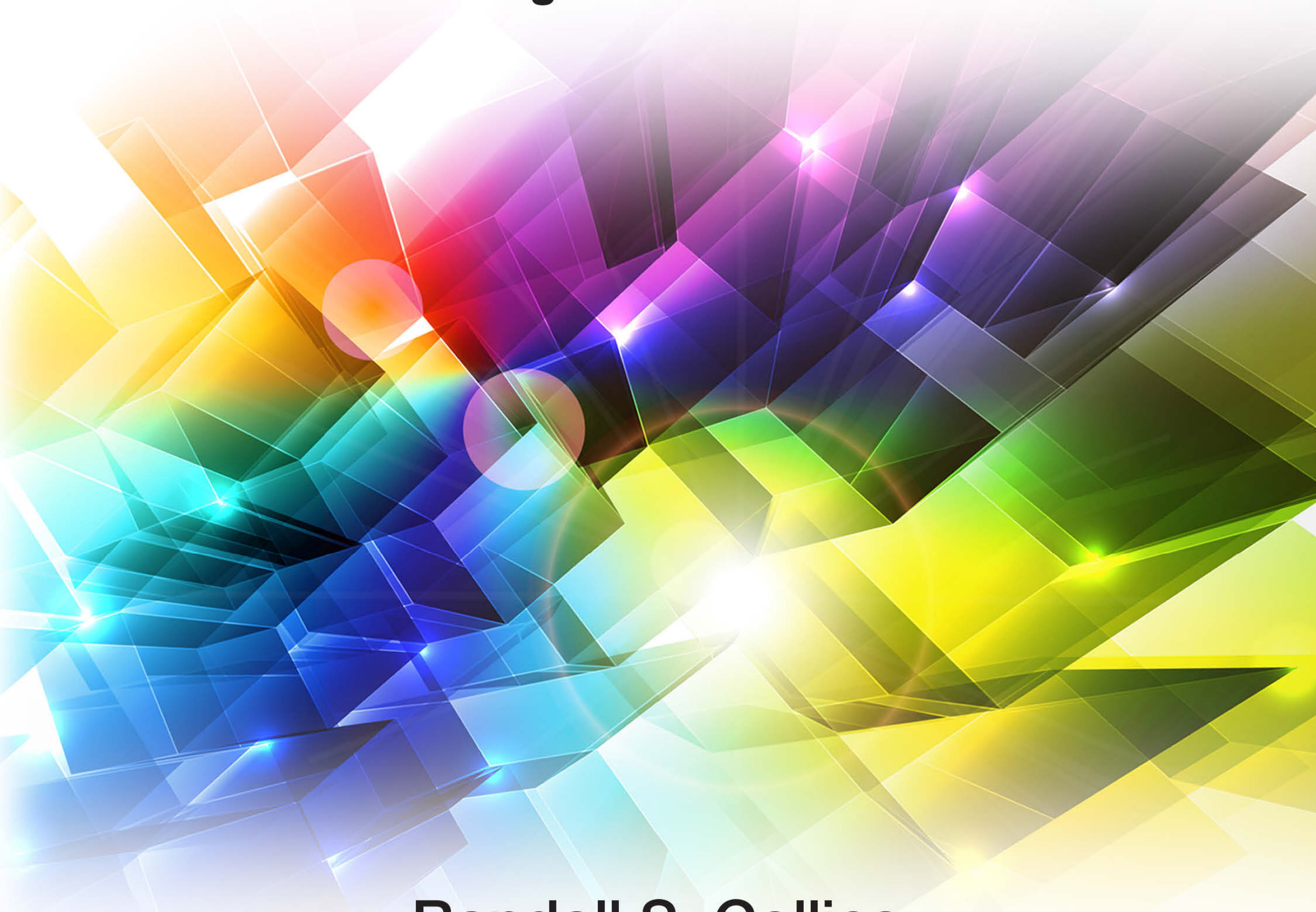
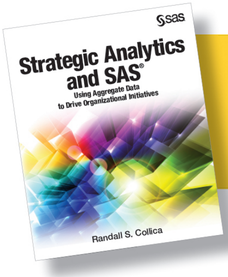
sas

Strategic Analytics and SAS®

Using Aggregate Data to Drive Organizational Initiatives



Randall S. Collica



Contents

About This Book . . . . . *v*

About the Author . . . . . *ix*

Acknowledgments . . . . . *xi*

Chapter 1 / Setting the Stage for Customer Strategic Analytics **1**

Introduction 1
Basis for Aggregating Customer Data and Predictive Models 2
The Difference between Tactical and Strategic Analytics 5
Use Cases Described in This Book 6

Chapter 2 / Use Case 1: Loyalty Analytics via Promoter Surveys **9**

What Are Promoter Scores? 9
Fusing Research Surveys with Customer Attributes 10
Text Mining Call Center Account Notes/Chats 11
Example: Turning Promoter Scores into an Analytic Predictive Model 13
Deploying a Promoter Score Model on a Larger Customer Base 27
Developing an Econometric Business Model from Promoter Score Aggregations 28
Strategic Analytics from Econometric Models 44
References for Further Reading 45

Chapter 3 / Use Case 2: Revenue Risk as a Customer Event **47**

Capturing Customer Revenue Patterns as Events 47
Concepts of Customer Event Histories 48
Survival – The Long and Short of It 50
How to Set Up Survival Data Prep for Survival Mining 52
Predicting Customer Probability of Revenue Risk 55
Aggregation and Further Analytics 65
Transform and Stationary 75
References and Further Reading 86

Chapter 4 / Use Case 3: Health Care Adverse Events **87**

Overview of the Adverse Events Problem 87
Exploratory Work with the VAERS Data 88
Predicting Adverse Events 92
Aggregation and Further Analysis 103

Defining Strategic Outcomes 112
References 116

Chapter 5 / Envisioning Strategic Analytics **117**
Understanding Data in a Visual Environment 117
Envisioning Analytic Results for Strategic Planning 118
Putting It All Together 125
Summing Up and Further Work 126
Reference 128

Index **129**

1

Setting the Stage for Customer Strategic Analytics

Introduction 1
Basis for Aggregating Customer Data and Predictive Models 2
The Difference between Tactical and Strategic Analytics 5
Use Cases Described in This Book 6

Introduction

This book will help bridge the gap between the detailed customer data records typically used for tactical projects and programs and the high-level, sweeping analytics that are typically used for understanding general trends, setting corporate directions, or planning the next stages for customer growth. What if a C-level executive or senior vice president asked you the following questions?

- Can we improve our customer satisfaction ratings, and, if so, how will such improvements grow our revenues and increase customer tenure?
- How likely are we to continue to hit our revenue targets over the next two quarters?
- Can physicians discern early warning signs of eminent life threats from sets of symptoms documented by medical professionals?

Questions such as these are strategic because they ask for high-level direction. However, the answers can be obtained from low-level customer or patient attributes. For example, in my previous position, a market research director had just conducted a survey that captured the attitudes toward information technology (IT). The market director asked me if there was a way we could use the results of a survey to improve customer offers and messaging in our tactical marketing programs. He also wondered if overarching marketing directions could be influenced by the attitudinal information obtained from the survey.

These were indeed intriguing questions, and I had to probe a bit to see if the survey answers could be linked to our customer data records. When I found out that they could, the task was to find a method or technique that would allow the linkage between detailed customer data and the high-level survey. What resulted was a set of analytics that enabled me to predict the five attitudinal segments with 85%accuracy on the customer base and also to combine the behavioral customer data attributes with the predicted attitudinal segments. This enabled strategic direction-setting for customers with certain behaviors and attitudes about IT to be combined in a way not ever before attempted. We could have a single segmentation that could provide both the product offer as well as the messaging, depending on the attitudinal segment. This capability was used to alter strategic planning for campaigns in several technology groups.

Basis for Aggregating Customer Data and Predictive Models

On what basis can we do these high-level strategic analytics by aggregating data and predictive model estimates and develop high-level models that provide answers to questions similar to those in the previous section? The answer is surprisingly simple.

But first let me tell you about an experience that changed my understanding of a subject many find very dry: *statistics.* As an engineering major, I was required to take at least one course in elementary statistics. I dreaded it greatly. I didn't believe I was learning anything, and the professor was drier than the Sahara desert! I just couldn't see the reason for all those bell-shaped curves and the like, and I didn't understand how it applied to me. This remained true until I worked in manufacturing, and we came across some problems that we desperately needed to solve, and quickly. We had a very likable consultant come in who was highly knowledgeable in practical experimentation, and he made practical statistics truly come alive. It wasn't dry at all. Engaging and intuitive, he held our attention, showing us some very simple techniques (really just simple addition and multiplication and very little algebra) that allowed us to find the root cause of the issues in some of our semiconductor manufacturing equipment. After designing some experimental manufacturing test runs, we aggregated each of them to form averages so that we could find the main effects of certain attributes such as temperature or voltage settings. These average estimates allowed us to make some general conclusions as to not only the main effects of our attributes, but also if one attribute was affecting another (called an interaction) like pressure and temperature. All of this from simple addition and multiplication!

The key was the design of the test runs and the levels of the attributes that made up those runs. In essence, the averages from certain runs of the designed experiment helped us to form conclusions from hypotheses that were based on our previous knowledge about how the process worked. This is a very simplistic form of a strategic analytic result that came from a design that allowed conclusions from a few runs, which in turn gave strategic insight into the root cause of the issue in our manufacturing process.

In this book, we'll discover ways to create a provisional design from the key attributes that are needed depending on the overall desired outcome. This type of design is not a factorial matrix like the designed experiment described earlier. Rather, it is an ad hoc design that will allow high-level analysis derived from low-level customer data records. The beauty behind this configuration is that once the high-level scenarios are fully explored, the plans that influence those strategic decisions can be implemented at the lower level. This is the case because the aggregation path used to produce those high-level analyses is linked to the lower level data..

Let's consider a simple pictorial example. Suppose you have a customer business-to-business data set that has the form shown in Figure 1.1 on page 3.

Figure 1.1 *Customer Data Example*

	Customer ID No.	CITY	STATE	Industry Segment Code	No of local employees	Corporate Revenue last fiscal yr.	0-No. 1=Yes	US Region Location of Business	Last Years Fiscal Revenue	This Years Fiscal Revenue YTD	Revenue for All Years	Year of 1st Purchase	Last Yr of Purchase
3	000045138	HARRISBURG	PA	PSV	9	940000	0	Northeast	0	0	3633.1	1998	1999
4	000055897	SAN FRANCISCO	CA	BKG	65	63000000	0	Western	-1892	0	383462.88	1998	1999
5	000056320	ARLINGTON	VA	PSV	5	63423000000	0	Southeast	114229.08	0	210767.08	1999	2000
6	000072249	NEW YORK	NY	MED	4500	225000000	0	Northeast	1461808.26	1385556.79	2905339.35	1996	2001
7	000084640	ATLANTA	GA	RTL	200	21600000	0	Southeast	21082.36	6854.28	98536.28	1998	2001
8	000091210	SEATTLE	WA	PSV	20	1300000	0	Western	0	0	728.5	1998	1998
9	000099101	TOCCOA	GA	AUT	211	8309000000	0	Southeast	0	0	73.32	1999	1999
10	000104463	AIKEN	SC	UTL	12300	2248228000	1	Southeast	396499.06	113593.01	9869176.47	1989	2001
11	000124750	FORT WORTH	TX	RTL	426	189900000	0	Central	0	0	38422.42	1994	1999
12	000137349	LAS VEGAS	NV	PSV	3	200000	0	Western	0	0	1138.34	1999	1999
13	000140400	MOUNT JULIET	TN	MED	90	8100000	0	Southeast	0	13256.32	13256.32	2001	2001
14	000146878	OKLAHOMA CITY	OK	RTL	3	190000	0	Central	0	0	29.14	1998	1998
15	000148908	LITTLE ROCK	AR	MED	11	23613000000	0	Central	0	0	3127.8	1994	1997
16	000152249	MIAMI	FL	BKG	0	26787000000	0	Southeast	35535.48	0	61345.34	1998	2000
17	000152660	SANTA MONICA	CA	CPG	5	360000	0	Western	2343	0	2343	2000	2000
18	000155010	PHARR	TX	MED	7	24322699	0	Central	300.96	0	3086.18	1999	2000

This data set contains more columns of data than what is shown here.[1] However, the main demographics are displayed. If a senior marketing executive asked you where the organization should next focus its efforts to drive increased customer revenue, perhaps you could start with the industries that are represented by their Industry Segment Code column and where the company is located in the US Region Location of Business column. These columns, along with the revenues received last year and the estimated spending column, might be used to help answer that question. The aggregated data of approximately 100,000 customer data records totals to 109 records shown in Figure 1.2 on page 4.

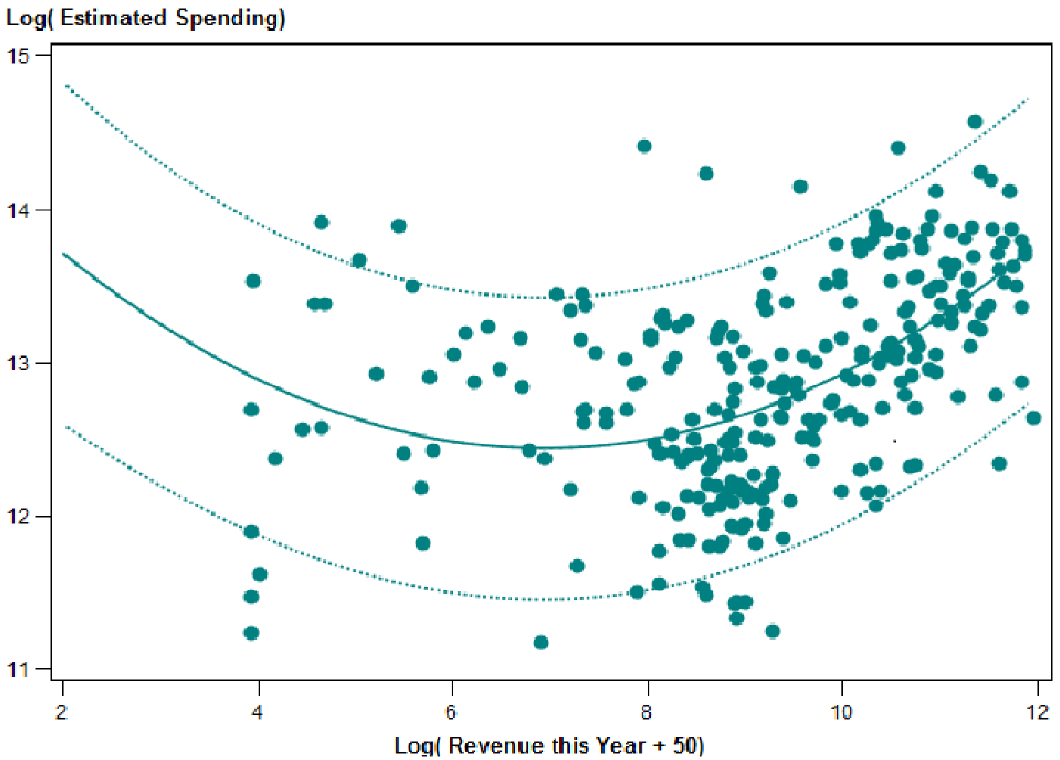
1 All data set attributes are described in the appendix given in the Example Code and Data found on the author's page (http://support.sas.com/publishing/authors/collica.html).

Figure 1.2 *Customer Data Aggregate*

	US Region Location of Business	Industry Segment Code	_FREQ_	This Years Fiscal Revenue YTD_Mean	Estimated Product-Service Spend_Mean
1	Central	AER	62	51187.575484	$1,516,277
2	Central	AUT	528	9463.9962121	$389,042
3	Central	BKG	1488	28165.274173	$387,647
4	Central	CHM	378	16406.060873	$390,582
5	Central	CPG	312	59781.952532	$458,983
6	Central	ELE	495	63254.987576	$690,520
7	Central	FMM	608	18586.859161	$307,320
8	Central	HCR	1404	17567.841731	$584,231
9	Central	INS	678	28770.540251	$557,578
10	Central	MED	951	20290.814017	$277,509
11	Central	NAT	118	26002.392119	$363,563
12	Central	OIL	412	75590.352379	$662,378
13	Central	PHM	228	24563.60886	$744,522
14	Central	PSV	5640	18234.566603	$305,620
15	Central	RTL	2442	17223.057932	$147,271
16	Central	SLE	2199	16148.374029	$697,297
17	Central	TEL	531	93204.316591	$682,345
18	Central	TRV	1078	20714.54744	$418,976
19	Central	UTL	509	81422.019096	$687,346
20	Great Lakes	AER	35	32388.161143	$989,954
21	Great Lakes	AUT	1492	17514.977621	$478,751
22	Great Lakes	BKG	1520	23519.590033	$384,536
23	Great Lakes	CHM	595	16393.203294	$328,096
24	Great Lakes	CPG	371	36250.088598	$469,752
25	Great Lakes	ELE	518	48841.226718	$378,421
26	Great Lakes	FMM	1329	16464.119789	$298,659
27	Great Lakes	HCR	1684	21514.483034	$468,491
28	Great Lakes	INS	749	41746.781682	$455,843
29	Great Lakes	MED	1084	15889.272878	$275,145
30	Great Lakes	NAT	63	11092.930159	$393,578

The average spending can be plotted against the average revenue received per year. If we were to plot these average revenues in a scatter plot, we might better understand the high-level relationship that might exist. As you can see in Figure 1.2 on page 4, a column labeled as _FREQ_ contains the counts of how many observations were aggregated from the original data set. The scatter plot in Figure 1.3 on page 5 gives the Log of Estimated Spending on the vertical axis and the Log of Revenues (added to 50) on the horizontal axis. The addition of 50 will is explained below.

Figure 1.3 *Customer Data Aggregate Scatter Plot*



The scatter plot from the aggregated customer detail data indicates that there is some curvature in the relationship between the Log of Estimated Spending attribute and the Log of Revenues, which was generated from customers this year. We add 50 to move the data to the right along the horizontal axis by 50 units in order to avoid negative values, since you cannot take the log of a negative number. This addition doesn't change the general shape of the relationship of the two attributes. It just scales the data so that the relationship can be more easily observed. This simple exercise shows how detailed data can be aggregated and how the general high-level relationship between two customer attributes can be observed graphically.

The Difference between Tactical and Strategic Analytics

Data and analytics at the individual customer level imply that an estimate, prediction, or classification takes place at a customer detail level, and, therefore, any business action taken will also be at that level; and this is a tactical example. For example, if customers are classified into segments that each represent customers with relatively similar attributes, then any action within those segments will generally be at the customer level. An example might be a marketing campaign that is targeting the segment that the customer is associated with. In strategic analytics, the model is at higher levels than the detail data. You wouldn't apply or score data records at the customer level that was developed from a model aggregated at a high level. This principle applies to data mining, machine learning, and in time-series forecasting. Therefore, the questions about business activities are referenced at the aggregate level rather than at the detail level. While this seems rather straightforward, the additional elements needed for best-in-class strategic analytics will typically involve some model scores or estimates of the following:

- Market share estimates obtained from syndicated data sources
- Regional or demographic econometric data elements or estimates
- Overlaid model scores that are obtained from third-party data providers at a high level, such as geography or econometric measures

These types of data, combined with aggregate customer or organizational data, will help form the basis to answer business questions such as the following:

1. Where should our organization focus its sales efforts next year by each industry market segment?
2. Is our share-of-wallet market too low in some region, industry, or country where we desperately need to improve?
3. Has our customer loyalty improved in some regions but not in others? What are the main factors driving this behavior? Can something be done to improve this situation?

Use Cases Described in This Book

These questions are directed at organizational shift areas. Data at the detail customer or transactional level needs to be aggregated in order to answer such questions. In Chapter 2, we'll look at a case study of customer loyalty behavior, attempt to answer the third question above, and offer some estimates of the driving mechanisms and the rates of potential improvements that might be obtained.

In Chapter 3, we will look at taking customer revenue transactions and aggregating them to understand risk of revenue decline and how to estimate customer risk at an organizational level.

Chapter 4 uses health care adverse event data to explore the potential mechanisms and explain the driving forces behind these adverse events, answering questions such as, Do hospital locations make a difference in the types of events observed?

And, finally, Chapter 5 takes a look at how we can envision these strategic analytics and develop methods for communicating the results of these high-level strategic analytics for consumption by a broader audience.

Note: In each use case example, the analyses presented are reflective of the ad hoc nature of the data. The actions that should be taken after such analyses are to design carefully controlled experiments to test and verify the alleged mechanisms to observe if they are real or just happenstance occurrences in the data.

Index

A

aggregation 65, 103
Allison, Paul 51
analytics
  customer strategic 1
  descriptive 127
  loyalty 9
  predictive 127
  prescriptive 127
  strategic 5, 44, 117, 118
  tactical 5
AUTOREG procedure 75

B

binning 127

C

call center account notes/chats, text mining 11
charts 128
Chi-square test 66
Clustering 125
confidence band 125
continuous time 50
customer attributes, fusing research surveys with 10
customer data, predictive models and 2
customer events
  capturing customer revenue patterns as 47
  concepts of histories 48
  revenue risk as a 47
customer probability, predicting for revenue risk 55
customer revenue patterns, capturing as events 47
customer strategic analytics 1

D

data
  in a visual environment 117
  multidimensional 127
  predictive models and customer 2
  querying 127
  survival 52
data acquisition wizard 128
data mining
  about 48
  in a visual environment 117
Data Partition node 17, 59, 96
decision trees 100, 125, 127
descriptive analytics 127
descriptive statistics 127
designed experimental study 65
Detractor category 13, 27
discrete time 50
display rules 127
dynamic filtering 127

E

econometric models
  developing from promoter score aggregations 28
  strategic analytics from 44

ellipsis (...) icon 19, 60
ESRI mapping technology 128
estimate ststament 39
Expectation-Maximization algorithm 96

F

F-test 66
filters 128
forecasting
  about 127
  performing basic 122

G

General Linear Models 125
GLM procedure 79
Gradient Boosting node 22, 24
graph templates 128
grouping 127

H

hazard 50
hazard curve 51
Hazard Rate histograms 62
health care adverse events
  about 87
  aggregation 103
  defining strategic outcomes 112
  exploratory work 88
  predicting 92
hierarchies
  about 128
  drillable 127

I

in-memory analysis 127
instrument statement 39
interface, for reports 127
interval censoring 49

L

label statement 39
layout capabilities, for reports 128
LOGISTIC procedure 107
logistic regression 125
Logistic Regression model 101
loyalty analytics, via promoter surveys 9
LSMeans statement 115

M

Machine Learning (ML) technique 12
mashups 128
Merge node 98, 99
Metadata node 20, 96, 100
Model Comparison node 21, 24, 100
MODEL procedure 38
model statement 39
models
  comparing 125
  deploying promoter score 27
multidimensional data 127
Mutual Information algorithm 96

N

network diagrams
  about 127
  analyzing 121
nodes
  Data Partition 17, 59, 96
  Gradient Boosting 22, 24
  Merge 98, 99
  Metadata 20, 96, 100
  Model Comparison 21, 24, 100
  Program 74, 78

Regression 20, 21, 25, 99, 100, 101
Replacement 17, 95
Reporter 64
Rule induction 100
Score 25, 64, 101
Survival 59
Text Cluster 21, 22, 96, 98
Text Filter 18, 19, 21, 96, 98
Text Parsing 17, 18, 96
Text Rule Builder 98, 101
Text Topic 19, 96, 98, 100, 102
Variable Selection 98, 100
null hypothesis 65

O

odds ratios 83
one-click ability 127
orthogonal variables 65

P

Passive category 13, 27
pop-up sections 127
predictive analytics 127
predictive models
customer data and 2
turning promoter scores into an analytic 13
prescriptive analytics 127
procedures
AUTOREG 75
GLM 79
LOGISTIC 107
MODEL 38
SGPLOT 79
Program node 74, 78
Project Library 30
Promoter category 13, 27
promoter scores
about 9
deploying models 27
developing econometric business models from aggregations of 28
turning into an analytic predictive model 13
promoter surveys, loyalty analytics via 9

Q

querying data 127

R

range-based alerts 128
ranking 127
receiver operating characteristics (ROC) 100
Regression node 20, 21, 25, 99, 100, 101
relationships, analyzing 119
Replacement node 17, 95
Reporter node 64
reports
creating 125
features of 127
research surveys, fusing with customer attributes 10
revenue risk
as a customer event 47
predicting customer probability of 55
right censoring 49
Rule induction node 100

S

Sales Trend Chart, visualizing 122
SAS LASR Analytic Server 128
SAS Visual Analytics 125, 127
SAS Visual Statistics 125
scenario analysis 127
Score node 25, 64, 101
SGPLOT procedure 79
Singular Value Decomposition (SVD) 12

stationary time series 75
statistical hypothesis testing 65
strategic analytics
  compared with tactical analytics 5
  data in a visual environment 117
  envisioning results for strategic planning 118
  from econometric models 44
strategic outcomes, defining 112
strategic planning, envisioning results for 118
survival analysis
  about 48
  types of 50
survival data, setting up prep of for survival mining 52
survival mining, setting up survival data prep for 52
Survival node 59

T

t-test 66
tactical analytics, compared with strategic analytics 5
text analysis 127
Text Cluster node 21, 22, 96, 98
Text Filter node 18, 19, 21, 96, 98
text mining call center account notes/chats 11
Text Parsing node 17, 18, 96
Text Rule Builder node 98, 101
Text Topic node 19, 96, 98, 100, 102
threshold-based alerts 128
Transform 75

U

Unit Root Tests 75

V

Variable Selection node 98, 100
views 127
visual environment, data in a 117

About This Book

Purpose

This manuscript is a culmination of my desire to help organizations that are struggling to connect the dots between detail data on their customers, patients, partners, and so on, and the high-level business objectives that support key goals and initiatives of these organizations. The shift from answering tactical business issues to answering both tactical and strategic business issues allows an organization to have a bigger picture view on how the analytics that they perform can affect their business.

Is This Book for You?

If you are a data miner, data scientist, or senior business analyst who desires to improve business outcomes, then this book will assist you in achieving that objective. If you are a director of analytics or customer experience, then this book can help you and your team see how your own detail data and model results can be used in more strategic initiatives.

Prerequisites

Users should be familiar with (but not necessarily expert in) regression techniques and methods, logistic analysis, confidence intervals, and a basic understanding of hypothesis tests. In addition, familiarity with machine learning and data mining techniques will also be helpful but not mandatory.

What's New?

What is different or new in this document when compared to the last version (optional)?

Scope of This Book

his book will help bridge the gap between the detailed customer data records typically used for tactical projects and programs and the high-level sweeping analytics used typically for understanding general trends, setting corporate directions, or planning the next stages for customer growth. What if a C-level executive or senior vice president asked you the following questions?

- Can we improve our customer satisfaction ratings, and, if so, how will such improvements grow our revenues and increase customer tenure?
- How likely are we to continue to hit our revenue targets over the next two quarters?
- Can physicians discern early warning signs of eminent life threats from sets of symptoms documented by medical professionals?

About the Examples

Software Used to Develop the Book's Content

SAS Enterprise Miner, SAS Text Miner, SAS Enterprise Guide, SAS/STAT, SAS/ETS, SAS/GRAPH, and SAS Visual Analytics.

Example Code and Data

My Author page will contain a ZIP file that has all the SAS code examples as well as the SAS Enterprise Miner and SAS Text Miner XML diagrams and data sets that complete each example.

Output and Graphics Used in This Book

For most of the exercises, SAS/GRAPH, SAS Statistical Graphics, and ODS were used to generate the graphics within SAS Enterprise Guide. Some of the graphics were within SAS Enterprise Miner and SAS Text Miner. Chapter 5's visualizations are all SAS Visual Analytics.

Exercise Solutions

All of the solutions are given in the data and code examples obtained on my Author page in a ZIP file that contains chapter folders. All of the SAS Enterprise Guide projects, SAS code, data sets, and completed XML diagrams are given as well.

Additional Resources

At the end of each chapter, there are additional resources and reference materials that pertain to each chapter's subject matter.

Keep in Touch

We look forward to hearing from you. We invite questions, comments, and concerns. If you want to contact us about a specific book, please include the book title in your correspondence.

To Contact the Author through SAS Press

By e-mail: saspress@sas.com

Via the Web: http://support.sas.com/author_feedback

SAS Books

For a complete list of books available through SAS, visit http://support.sas.com/bookstore.

Phone: 1-800-727-3228

Fax: 1-919-677-8166

E-mail: sasbook@sas.com

SAS Book Report

Receive up-to-date information about all new SAS publications via e-mail by subscribing to the SAS Book Report monthly eNewsletter. Visit http://support.sas.com/sbr.

Publish with SAS

SAS is recruiting authors! Are you interested in writing a book? Visit http://support.sas.com/saspress for more information.

About the Author

Randall S. Collica is a Principal Solutions Architect at SAS supporting the retail, communications, consumer, and media industries. His research interests include segmentation, clustering, ensemble models, missing data and imputation, Bayesian techniques, and text mining for use in business and customer intelligence. He has authored and coauthored 11 articles and a book, *Customer Segmentation and Clustering Using SAS® Enterprise Miner™, Second Edition*. He holds a US patent titled "System and Method of Combining Segmentation Data." He received a BS degree in electronic engineering from Northern Arizona University.