Modeling Techniques in Predictive Analytics with Python and R

A Guide to Data Science

Thomas W. Miller

Associate Publisher: Amy Neidlinger
Executive Editor: Jeanne Glasser
Operations Specialist: Jodi Kemper
Cover Designer: Alan Clements
Managing Editor: Kristy Hart
Project Editor: Andy Beaster
Senior Compositor: Gloria Schurick
Manufacturing Buyer: Dan Uhrig

Published by Pearson Education, Inc.
Upper Saddle River, New Jersey 07458

Printed in the United States of America

First Printing October 2014

ISBN-10: 0-13-3892069
ISBN-13: 978-0-13-389206-2

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educacin de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.
Library of Congress Control Number: 2014948913

Contents

C.2 DriveTime Sedans

DriveTime in 2001 is an automobile dealership and financing firm with seventy-six dealerships in eight states. In a typical month the firm sells about four thousand used vehicles and processes about ten thousand credit applications. Virtually all sales are financed. The firm's stated mission is: "To be the auto dealership and finance company for people with less than perfect credit."

DriveTime generates traffic at its dealerships through television and radio advertising, referrals from other dealerships, and through its website. Customers who need financing to purchase vehicles are run through a custom credit risk scorecard, which uses both credit bureau and application information to determine credit worthiness. A generated risk score is used to determine the appropriate deal structure and credit policy.

DriveTime purchases most of its vehicles at auctions and from wholesalers. Vehicles include many makes and models of cars and trucks. The firm uses an information service known as Experian Autocheck to ensure that vehicles have correct odometer readings, have not been previously "totaled" (that is, evaluated as having no value after an accident), and have no other significant negative history. Vehicles that fail the Experian check are rejected and sent back to sellers. Those that pass are sent to a DriveTime reconditioning and inspection center, where they are put through additional checks and repaired as necessary. Vehicles are then delivered to the dealerships for sale.

Normal dealer sales occur within ninety days of delivery to the dealership. If a vehicle does not sell within ninety days, it is called an overage vehicle, meaning that it has been on the lot too long to generate normal dealer profits. Each overage vehicle has its sales price reduced in order to encourage a sale within the ensuing 91- to 119-day period. Profits on vehicles sold within the 91- to 119-day period are much lower than profits on vehicles sold within the normal 90-day period. Furthermore, if an overage vehicle fails to sell within 120 days, the vehicle is taken off the lot and sold at auction. DriveTime takes a loss on vehicles sold at auction.

Written by Thomas W. Miller and Steve Zemaitis. Data provided by DriveTime.

Table C.1 provides a hypothetical example, showing how normal and overage sales translate into business profits or losses for DriveTime. This example demonstrates the value of using a statistical model to select vehicles for sale. Profit contributions in the example represent gross rather than net profits. They do not account for operating costs, overhead costs, or taxes.

Table C.2 describes variables from the DriveTime vehicles database. The data, which represent 17,506 sedans sold and financed in the second half of 2001, are divided into three data sets for modeling work: 8,753 sedans comprise the training set, 4,377 the validation set, and 4,376 the test set.

Table C.3 shows how researchers use eight color categories to represent twenty-seven colors in the vehicles database. Color categories are defined so that each category has a sufficiently large frequency to warrant its use in modeling work. Gold becomes a catch-all or other color category, including gold, tan, cream, yellow, and brown tones.

Certain variables may be useful in developing vehicle selection models. Newer, lower mileage vehicles, for example, may be expected to sell faster than older, higher mileage vehicles. Sales prices are not included in the vehicles database, but we can assume that prices for vehicles sold within ninety days (normal dealer sales) are marked up, so that the firm recovers costs associated with purchasing, repairs, operations, and interest, and makes an appropriate profit.

DriveTime managers wonder whether it is possible to develop selection models for sedans using data from the vehicles database. Is a single model sufficient, or should separate models be built for the states in which DriveTime operated in 2001 (Arizona, California, Florida, Georgia, Nevada, New Mexico, Texas, and Virginia)? What would the models look like, and how much profit improvement would result from using the models?

Table C.1. *Hypothetical Profits from Model-guided Vehicle Selection*

The table below reflects hypothetical profits associated with DriveTime vehicle sales, given an average total cost per vehicle of $5,000, a 20 percent markup for normal dealer sales, 10 percent markup for overage dealer sales, and 20 percent loss for overage vehicles sold at auction. This example assumes that, of the approximately four thousand vehicles sold each month, about 85 percent are normal dealer sales, 10 percent overage dealer sales (within the 91- to 119-day period), and 5 percent overage auction sales.

	Type of Sale			
	Normal Dealer	*Overage Dealer*	*Overage Auction*	*Monthly Totals*
Unit total cost	$5,000	$5,000	$5,000	
Unit price	$6,000	$5,500	$4,000	
Unit margin profit (loss)	$1,000	$ 500	($1,000)	
Units sold	3,400	400	200	4,000
	(85%)	(10%)	(5%)	
Profit (loss)	$3,400,000	$200,000	($200,000)	$3,400,000

Suppose that researchers are able to develop a model that is reasonably accurate in predicting how long it takes to sell a vehicle. Suppose further that, using this time-to-sale model to guide inventory decisions, DriveTime is able to increase normal dealer sales from 85 to 90 percent, with corresponding declines in overage vehicle sales. Assuming no change in vehicle costs or prices, what would be the effect upon profits? The following table suggests that monthly profits would increase by $220,000. Twelve months of sales of this type would contribute more than $2.6 million in profit a year. This demonstrates the value of using statistical models to guide business decisions.

	Type of Sale			
	Normal Dealer	*Overage Dealer*	*Overage Auction*	*Monthly Totals*
Unit total cost	$5,000	$5,000	$5,000	
Unit price	$6,000	$5,500	$4,000	
Unit margin profit (loss)	$1,000	$ 500	($1,000)	
Units sold	3,600	280	120	4,000
	(90%)	(7%)	(3%)	
Profit (loss)	$3,600,000	$140,000	($120,000)	$3,620,000

Table C.2. *DriveTime Data for Sedans*

Variable Name	*Description*
data.set	Data set for modeling (TRAIN, VALIDATE, or TEST)
total.cost	Total cost of vehicle (purchase cost + repair cost + other costs)
lot.sale.days	Days from vehicle delivery to dealership to sale (days on lot)
overage	Overage vehicle (NO = 0–90 days on lot, YES = 91+ days on lot)
vehicle.type	Type of sedan (ECONOMY, FAMILY.SMALL, FAMILY.MEDIUM, FAMILY.LARGE, or LUXURY)
domestic.import	Type of manufacturer (Domestic or Import)
vehicle.age	Age of vehicle in years (year of sale minus model year)
vehicle.age.group	Age group of vehicle (ONE-THREE, FOUR, FIVE, SIX, or SEVEN+)
color.set	Color category (BLACK, WHITE, . . . , GOLD)
makex	Make/manufacturer of vehicle (BUICK, CADILLAC, . . . , TOYOTA)
state	State location of dealership where vehicle sold (AZ = Arizona, CA = California, . . . , VA = Virginia)
make.model	Make and model of sedan (ACURA.INTEGRA, BUICK.CENTURY, . . . , TOYOTA.TERCEL)

Table C.3. *DriveTime Sedan Color Map with Frequency Counts*

Color in	*Color Category Defined by Researchers*								
Database	Black	White	Blue	Green	Red	Purple	Silver	Gold	*Count*
Aluminum/ Silver	0	0	0	0	0	0	1234	0	1234
Beige	0	0	0	0	0	0	0	123	123
Black	1216	0	0	0	0	0	0	0	1216
Blue	0	0	2149	0	0	0	0	0	2149
Blue - Dark	0	0	16	0	0	0	0	0	16
Blue - Light	0	0	53	0	0	0	0	0	53
Bronze	0	0	0	0	0	0	0	15	15
Brown	0	0	0	0	0	0	0	64	64
Burgundy/ Maroon	0	0	0	0	0	1410	0	0	1410
Cream	0	0	0	0	0	0	0	76	76
Chrome/ Stainless Steel	0	0	0	0	0	0	1	0	1
Copper	0	0	0	0	0	0	0	9	9
Gray	0	0	0	0	0	0	618	0	618
Gold	0	0	0	0	0	0	0	1003	1003
Green	0	0	0	3309	0	0	0	0	3309
Green - Dark	0	0	0	59	0	0	0	0	59
Green - Light	0	0	0	20	0	0	0	0	20
Lavender	0	0	0	0	0	8	0	0	8
Mauve	0	0	0	0	12	0	0	0	12
Orange	0	0	0	0	9	0	0	0	9
Pink	0	0	0	0	7	0	0	0	7
Purple	0	0	0	0	0	366	0	0	366
Red	0	0	0	0	1406	0	0	0	1406
Tan	0	0	0	0	0	0	0	414	414
Taupe	0	0	0	0	0	0	0	11	11
Teal	0	0	289	0	0	0	0	0	289
Turquoise	0	0	2	0	0	0	0	0	2
White	0	3603	0	0	0	0	0	0	3603
Yellow	0	0	0	0	0	0	0	4	4
Count	1216	3603	2509	3388	1434	1784	1853	1719	17506