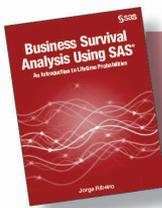


# **Business Survival Analysis Using SAS<sup>®</sup>**

**An Introduction to Lifetime Probabilities**



**Jorge Ribeiro**



From *Business Survival Analysis Using SAS*.  
Full book available for purchase [here](#).

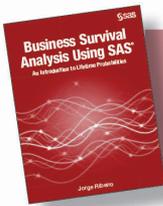
# Contents

<b>About The Book</b> .....	<b>vii</b>
<b>About the Author</b> .....	<b>xiii</b>
<b>Acknowledgments</b> .....	<b>xv</b>
<b>Contents</b> .....	<b>iii</b>
<b>Chapter 1: Data Preparation for Survival Models</b> .....	<b>1</b>
Introduction .....	1
Step 1: Collect Three Raw Variables .....	3
Step 2: Create Three Required Variables .....	3
Step 3: Collect and Process Many Predictor Variables .....	3
Data Set Structure .....	3
Characteristics of the Data Set Used in This Book .....	4
Special Considerations for Survival Analysis Data Input .....	4
Plan for the Collection of the Required Variables .....	4
Five Fundamental Questions for Time-to-Next-Purchase Survival Models .....	4
Data Set Creation .....	7
(Start, End) Calendar Timeline .....	7
Time and Censor Variables .....	9
Strata by Previous Purchases .....	12
Categorical Variables .....	13
Data Preparation with Base SAS, Macros, and SAS/STAT .....	14
Multiple Imputation .....	14
Winsorization .....	15
Binning Using Macros .....	15
Data Preparation with SAS Enterprise Miner .....	16
<b>Chapter 2: Exploratory Data Analysis and Basic Statistical Concepts</b> .....	<b>19</b>
Introduction .....	19
Overview of Continuous and Discrete Distributions .....	20
Deciding Whether to Use Continuous Time or Discrete Time .....	21
Understanding the Concept of a Customer Tie .....	22
Continuous Distribution .....	24
Survival Function for Continuous Time .....	24
Hazard Function for Continuous Time .....	27
Discrete Distribution .....	28

Data Formats .....	29
Customer-Level Data Set (Standard Data Format) .....	29
Calculation of the Discrete Hazard—An Example .....	29
Interpretation of the Results .....	31
Month Data Set (Expanded Data Format) .....	32
Creation of Dummy Variables .....	34
A Simple Example.....	34
A Time-to-Next-Purchase Example.....	35
PROC LOGISTIC to Compute Hazard.....	38
Odds Calculation (January) .....	39
Hazard Calculation (January) .....	40
Survival Function .....	41
Hazard and Survival Relationship.....	42
Summary of Distributions .....	43
Life Table Method .....	43
Step 1: PROC LIFETEST with METHOD = LIFE .....	44
Step 2: At-Risk Adjustment .....	49
Step 3: Empirical Hazard Calculation.....	49
Step 4: Graphics and Interpretation .....	50
Data Storage Formats .....	52
Customer-Level Data Set (Standard Format) .....	52
Month Data Set (Expanded Format) .....	52
<b>Chapter 3: Stratified Sampling Method for Survival Analysis .....</b>	<b>53</b>
Introduction .....	53
The Sampling Process with PROC SURVEYSELECT .....	54
Step 1: Analyze Sample Proportion for Strata.....	55
Step 2: Create the Model_Shopping Sample.....	56
Step 3: Create the Score_Shopping Sample .....	57
Step 4: Compare the Results from the Model_Shopping and Score_Shopping Data Sets .....	59
SAS Enterprise Miner Nodes .....	60
The Sample Node .....	60
The Data Partition Node .....	60
The Flowchart .....	61
<b>Chapter 4: SAS Enterprise Miner Use from Source to Customer-Level Model Output.....</b>	<b>63</b>
Introduction .....	63
Creating the Process Flow.....	64
Step 1: Create a New Project .....	65
Step 2: Create a New Library .....	66
Step 3: Create Start Code.....	67
Step 4: Create a New Data Source .....	69
Step 5: Create a Process Flow Diagram .....	74

Step 6: Insert the Data Partition Node .....	76
Step 7: Create the Survival Process Flow .....	80
Running a Model and Examining Results .....	81
Survival Node Properties Panel .....	81
Results .....	83
<b>Chapter 5: The Cubic Spline Regression and Model Interpretation .....</b>	<b>85</b>
Introduction .....	85
SAS Enterprise Miner Strategy for Standard Data Format .....	86
The Problem: The Nonlinearity of the Hazard Function .....	87
The Solution: The Cubic Spline Basis Functions .....	87
Two Challenges: Number of Knots to Choose and Placement of the Knots .....	89
The Mathematical Definition of a Knot.....	90
The Default Five-Knots Cubic Spline Basis Method without Stepwise Selection .....	92
The Default Five-Knots Cubic Splines Basis Method Using the Stepwise Selection.....	95
Example 1: Interpretation of Numeric Variables.....	97
Understanding the Stepwise Regression Spline Model in SAS Enterprise Miner .....	98
Writing the Model .....	100
Understanding the Odds Ratio.....	100
Example 2: Interpretation of Categorical Variables.....	105
Odds Calculation of Categorical Variables.....	105
Interpretations .....	107
Model Output.....	111
Model Information .....	111
Strata Analysis .....	116
Survival and Hazard Functions Saved in a Data Set.....	121
<b>Chapter 6: The Fully Expanded Data Set Model .....</b>	<b>123</b>
Introduction .....	123
Saving the Expanded Data Set .....	123
Include a SAS Code Node.....	123
Rename the SAS Code Node.....	124
Create a Permanent Library .....	125
Explore the SURV_EXPCENDATA Data Set .....	126
Running the Expanded Model.....	128
Create a Second Diagram and Assign Variable Roles .....	128
Set the Properties and Run .....	129
Comparing the Results.....	130
<b>Chapter 7: Statistics for Performance and Validation .....</b>	<b>131</b>
Introduction .....	131
Twofold Focus .....	132
SAS Enterprise Miner for Practical Solutions.....	132
Comparison of Model Validation Approaches .....	133
Validation: An Explanation.....	133
Traditional Logistic Regression .....	134
Survival Regression (Steps 1 through 4).....	136

Statistics for Performance and Validation.....	141
Step 5: Compute the Hazard for Only the First Month of the Period .....	141
Step 6: Create a Hit Variable Indicating Whether Repurchase Occurs in January.....	142
Step 7: Compute the Depth of the Hazard in Quantiles .....	143
Step 8: Compute the Sensitivity .....	146
Step 9: Create the Concentration Curve .....	147
Step 10: Compute Lift .....	149
Step 11: Create the Lift Curve .....	150
Step 12: Compute Benefit.....	152
Step 13: Create the Benefit Curve .....	152
Step 14: Select the Cutoff=Depth Where the Benefit Is Maximum .....	153
The Depth Measure and the Hazard Measure: Which to Choose.....	155
Two Business Situations.....	155
Business Situation 1: Use of Depth as Cutoff .....	156
Business Situation 2: Use of Average Hazard Function Value as Cutoff.....	158
Specificity and Suboptimal Cutoffs .....	159
Gini Concentration Ratio .....	163
Kolmogorov-Smirnov.....	165
Method 1—PROC NPAR1WAY .....	167
Method 2—PROC SQL .....	168
Density Analysis and PROC KDE .....	170
A Comprehensive Validation Approach .....	172
<b>Chapter 8: Scoring New and Old Customers.....</b>	<b>173</b>
Introduction .....	173
Business Situation 1: Scoring Customers .....	175
Preparing the Scoring Data Set .....	175
Scoring Using SAS Enterprise Miner.....	178
Analyzing the Optimized Score Code.....	190
Scoring Using SAS DATA Steps.....	197
Analyzing Customer ID = 4 .....	199
Using PROC COMPARE for Validation.....	203
Business Situation 2: Calibrating Scoring Results .....	205
Step 1: Compute Actual and Forecast of Part_TRAIN .....	207
Step 2: Plot an Actual and Forecast Graph—Part_TRAIN .....	208
Step 3: Compute the Calibration (Prior Proportion—Part_TRAIN) .....	209
Step 4: Compute the Full-Year Forecast—Part_VALIDATE .....	209
Step 5: Calibrate the Forecast of Part_VALIDATE .....	211
Step 6: Plot an Actual by Calibrated Forecast Graph .....	212
<b>References .....</b>	<b>213</b>
<b>Index .....</b>	<b>215</b>



From *Strategic Analytics and SAS*. Full book available for purchase [here](#).

## Chapter 1: Data Preparation for Survival Models

<b>Introduction</b> .....	<b>1</b>
<b>Step 1: Collect Three Raw Variables</b> .....	<b>3</b>
<b>Step 2: Create Three Required Variables</b> .....	<b>3</b>
<b>Step 3: Collect and Process Many Predictor Variables</b> .....	<b>3</b>
<b>Data Set Structure</b> .....	<b>3</b>
<b>Characteristics of the Data Set Used in This Book</b> .....	<b>4</b>
<b>Special Considerations for Survival Analysis Data Input</b> .....	<b>4</b>
<b>Plan for the Collection of the Required Variables</b> .....	<b>4</b>
<b>Five Fundamental Questions for Time-to-Next-Purchase Survival Models</b> .....	<b>4</b>
<b>Data Set Creation</b> .....	<b>7</b>
<b>(Start, End) Calendar Timeline</b> .....	<b>7</b>
<b>Time and Censor Variables</b> .....	<b>9</b>
<b>Strata by Previous Purchases</b> .....	<b>12</b>
<b>Categorical Variables</b> .....	<b>13</b>
<b>Data Preparation with Base SAS, Macros, and SAS/STAT</b> .....	<b>14</b>
<b>Multiple Imputation</b> .....	<b>14</b>
<b>Winsorization</b> .....	<b>15</b>
<b>Binning Using Macros</b> .....	<b>15</b>
<b>Data Preparation with SAS Enterprise Miner</b> .....	<b>16</b>

---

### Introduction

Development of any survival analysis model requires that input data be prepared in a specific way. This chapter introduces a Garden data set that contains real data from an anonymous company selling products across six main departments. This is the input data on which you will build the time-to-next-purchase survival model. In addition, this chapter explains general items of relevance to survival analysis, such as the following:

- Variables required (Start, End, Censor, Last\_day, and Time)
- Missing data
- Outliers or extreme values
- Non-normality
- Binning of categorical and continuous variables

If you have little experience in data mining or modeling, then for context and additional detail in data preparation, see Svolba (2006) and Refaat (2007).

## 2 Business Survival Analysis Using SAS: An Introduction to Lifetime Probabilities

The chapter highlights how SAS Enterprise Miner handles data preparation:

- Specific nodes designed for data preparation such as Filter and Interactive Binning are used.
- Advanced knowledge of Base SAS or the SAS Macro Language is less critical to implementing and validating models, making the process more transparent and accessible. Where this knowledge is required, SAS Enterprise Miner allows procedures and macros to be isolated in SAS code nodes.
- Whole end-to-end projects from source databases to final model output can be set up quickly with standardized folder names and processes, allowing for easy comparison of different data manipulation scenarios.

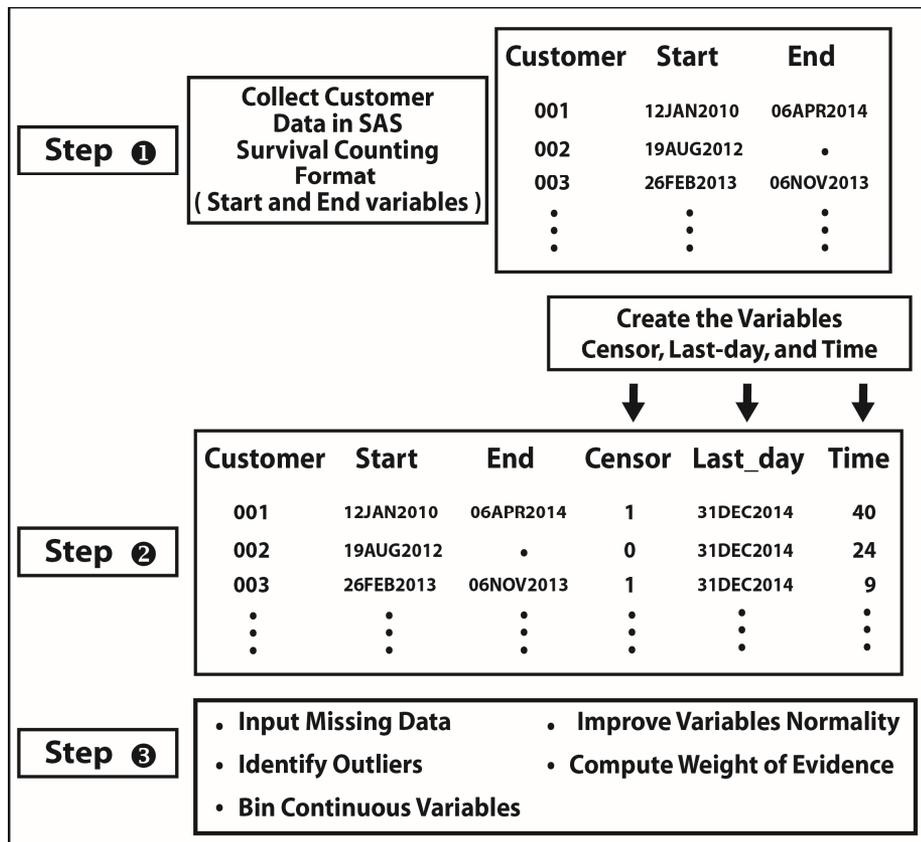
The Garden data set contains 676,635 customers who were observed over a four-year period. Their start time is when they make their first purchase. They are monitored until the end of the four-year period.

The aim of the model is to predict the time, in months, to the next purchase of a product. Analysis is centered on whether and when the second purchase (repurchase) occurs. To address this, the database must contain the relevant variables for the model.

In this chapter, the event called *Target* by the SAS Enterprise Miner node is repurchase or second purchase. To model the Target variable, SAS Enterprise Miner analyzes different input variables such as Age, Scorecard, Behavior, and channel (Internet, Telephone, and Catalog) to explain and predict the time elapsed until the customer makes a second purchase. Only variables available at the time of the first purchase are considered as input variables (predictors).

Figure 1.1 shows a step-by-step process to perform a systematic data preparation before model development.

**Figure 1.1: Data Preparation Flowchart for Time-to-Next-Purchase Models**



The information in the variables in steps 1 and 2 is universal to survival analysis. These variables are needed for the time to event and so that you can know whether an event occurs for an entity. These variables are also used in the input data for advanced models. After the discussion in this chapter, they are not discussed again. In contrast, the input variables are model-specific and selected based on the analysis required and stakeholder proposals—in this case, the variables to predict time-to-next-purchase. There are considerations in step 3 (such as missing data, outliers, and binning) that must be addressed when using these predictor variables to produce the right survival analysis model. SAS Enterprise Miner has a number of nodes for handling predictor variables.

---

## Step 1: Collect Three Raw Variables

Collect three raw variables:

- *Customer* is a numeric variable with a unique ID number for each customer.
- *Start* is a numeric variable that represents the SAS date when an event occurred in the study period. In this case, the event is a customer's first purchase in the study period 2010 to 2014. The variable is always populated (nonmissing) because customers making no purchases in the study period are out of the scope of the model.
- *End* is a numeric variable that represents the SAS date when a second event occurred in the study period. Its value can be greater than the value of *Start* or missing if no second event occurs. In this case, the event is a customer making a second purchase (repurchase) in the study period 2010 to 2014.

The information contained in these variables can be displayed in a standard (cross-sectional) data format or in an expanded data format. The Standard format shown in Figure 1.1 consists of one row per entity (in this case, one row per customer). The expanded format has one row per entity per time period. Both formats are covered in more detail in Chapter 2.

---

## Step 2: Create Three Required Variables

Create three required variables:

- *Censor* is a numeric variable equal either to 0 or 1. The value is 1 if the *End* variable is populated and 0 if it is missing.
- *Last\_day* is a numeric variable that represents the SAS date of the end of the study period. In this case, it is 31DEC2014 throughout the database.
- *Time* is a numeric variable that contains the number of interval time periods between *Start* and *End* if *End* is populated, or between *Start* and *Last\_day* if it is not.

---

## Step 3: Collect and Process Many Predictor Variables

Collect many raw variables containing customer information to predict time to event as denoted by the *Time* variable. Process these variables by imputing missing values, identifying outliers, and binning, as appropriate.

---

## Data Set Structure

The Garden data set is used throughout Chapter 1 to demonstrate survival analysis data preparation. It consists of data from an anonymous company, which sells across six main departments:

- Garden Tools
- Electrical Tools and Supplies
- Security and Safety
- Computer Accessories
- Car Accessories
- Art and Decoration

---

## Characteristics of the Data Set Used in This Book

The Garden data set consists of customer-level transaction data. Each of the 676,635 rows represents a customer who made their purchase between 1 January 2011 and 31 December 2014.

Some of these customers are new to the business. Others have made purchases prior to 1 January 2011. In the rest of the book, *first purchase* refers to the first purchase made by the customer during the study period 1 January 2011 to 31 December 2014. *Second purchase* refers to the subsequent purchase.

The Garden data set has one record per customer, with 29 columns containing internal data collected through the company's data warehouse. The variables include details of transactions and purchased products.

---

## Special Considerations for Survival Analysis Data Input

You first need to understand how the structure required for data input into a survival analysis model differs from that of other techniques such as linear, logistic regression, or time series. Blossfeld et al. (1989) comment that, because of the difficulty of data preparation and the complexity of dealing with censoring data, most inexperienced modelers avoid survival analysis. The complexities are related to the representativeness, quantity, and quality of the input variables.

It is important to consider what input variables are appropriate and necessary for the creation of a time-to-next-purchase survival analysis model. There are many possible variables that could be of interest. The most important fields for this particular model are as follows:

- Details about the customer at and previous to the first purchase
- Buyers' channels
- Internal family and income information about the customer.
- Products

With regard to buyers' channels, because segmenting the buyers is typically complex, this has been determined before starting the modeling process. You need to have as much information as you can to deliver a model with a high predictive ability, focused exclusively on a company's current customer database, supposing that new customers have the same profile as current customers.

---

## Plan for the Collection of the Required Variables

You should develop a detailed plan of how to collect the required variables. Remember that, no matter how thoroughly defined and logical your database, the results of survival analysis are not credible unless your variables perform well with the validation data set.

The collection of data and its availability for modeling purposes should be agreed on before the model-building process begins. It is good practice for a consensus to be reached among all stakeholders so that everyone can commit to the methodology from the beginning of the project. Creating and implementing models is somewhat of a political issue within some companies.

---

## Five Fundamental Questions for Time-to-Next-Purchase Survival Models

The model presented in this book was created to answer the five fundamental questions normally addressed by time-to-next-purchase survival models. In Figure 1.2, these questions are indicated next to the Variable column. In the following paragraphs, each question is reviewed in more detail.

Figure 1.2: Variables and the Questions That They Answer

	Variable	Type	Label
<b>When?</b> ①	1 Id_number	Numeric	Customer Unique ID Number
	2 Start	Numeric	Day of the First Purchase
	3 End	Numeric	Day of the Second Purchase or Missing if Censored
	4 Censor	Numeric	Second Purchase(Censor =1) -- Censored (Censor = 0)
	5 Last_day	Numeric	Last Day of Sample Selection 31/12/2014
	6 Time	Numeric	Time Elapsed after First Purchase
<b>How much?</b> ②	7 Garden	Numeric	Garden Tools
	8 Decorating	Numeric	Art, Craft and Decorating
	9 Car	Numeric	Car Accessories
	10 Electrical	Numeric	Electrical Tools
	11 Safety	Numeric	Security and Safety Products
	12 Computer	Numeric	Computer and Accessories
	13 Prev_Garden	Numeric	Previous Spent in Garden Tools
	14 Prev_Decorating	Numeric	Previous Spent in Art, Craft and Decorating
	15 Prev_Car	Numeric	Previous Spent in Car Accessories
	16 Prev_Electrical	Numeric	Previous Spent in Electrical Tools
<b>How many?</b> ③	17 Prev_Safety	Numeric	Previous Spent in Security and Safety Products
	18 Prev_Computer	Numeric	Previous Spent in Computer and Accessories
<b>How?</b> ④	19 Amount_CLV	Numeric	Total amount spent in Customer Lifetime Value
	20 Strata	Character	Strata Total Orders
	21 Account_Origin	Character	Account Origin
	22 Order	Character	Origin Purchase Order
<b>Who?</b> ⑤	23 Age	Character	Binned Age
	24 Credit_score	Character	Binned Credit Scorecard
	25 Behavior	Character	Binned Behavior
	26 Mosaic	Character	Mosaic Bureau Data
	27 Credicard	Character	Credicard Brand - Method of Payment
	28 Family	Numeric	Family Bureau Data
	29 Income	Numeric	Income Bureau Data

**Key Tip:** Include a label preceded by a number for any model created by SAS Enterprise Miner. The labels will be important later in the modeling process.

### Question 1: When Were the Purchases Made?

The When? ① variables address the question of when a customer (or group of customers) buys a specific product. The dates are required for survival analysis models created in SAS/STAT or SAS Enterprise Miner. The variables record the time of each purchase for each customer in the data set. The preparation of these variables for modeling purposes is discussed in detail in the section “Data Set Creation.”

### Question 2: How Much of What Was Purchased?

The primary goal of survival analysis is to model the Hazard function, assessing the relationship between the hazard and a set of variables (predictors) to determine whether they are statistically significant, controlling for other variables. The How much? ② variables numbered 7 to 12 represent the cost of the first purchase. The How much? ② variables numbered 13 to 18 represent the costs of previous purchases. They have the prefix (Prev\_) added to their original names.

You want to measure the impact of these characteristics on the time to the second purchase and to predict which customers are most likely to make a second purchase. In addition, you want to establish when the purchase is likely to occur.

The input variables 7 to 19 represent the amount of money spent on each product by each customer. A customer can buy multiple products from each department at one time. These variables represent the monetary relationship between the customer and the company. By including this information, you can quantify the relationship between past and current purchases in terms of profit per customer and customer lifetime value (CLV).

### Question 3: How Many Purchases?

For the How many? ⑤ variable, the categorical variable Strata segments customers according to the total number of purchases they made since their first purchase, which for some customers was 20 years ago. It is one of the most important input variables for time-to-next-purchase models and is discussed in greater depth in the section “Strata by Previous Purchases.”

### Question 4: How Were the Purchases Made?

The How? ④ variables record the channel through which the customer made their purchase.

#### Account\_Origin

The variable Account\_Origin tells you how the customers made their first purchase. It has the following categories:

- **Branch.** Customers who made their first purchase through a branch. A plausible hypothesis is that these customers are likely to purchase larger items from the Garden Tools department.
- **Internet.** Customers who made their first purchase on the company’s website. Another hypothesis is that customers who use this channel are younger than those using other channels.
- **Post.** Customers who made their first purchase through mail-order by using a product reference number from the company’s catalog and sending the order by post.
- **Telephone.** Customers who made their first purchase by telephone. A plausible hypothesis is that these customers tend to be older with no access to the Internet or who do not trust online shopping.

Output 1.1 shows that 46.64% of customers over the past made their first purchase using the Internet. The Account\_Origin variable could be useful if management wants to review channel strategy (for example, to assess how important branches are as a way of servicing customers). For example, management might be considering closing the least profitable stores (branches) and dispatching the products via other channels instead.

#### Output 1.1: Descriptive Statistics: Account\_Origin

Account_Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Branch	100830	14.90	100830	14.90
Internet	315575	46.64	416405	61.54
Post	108914	16.10	525319	77.64
Telephone	151316	22.36	676635	100.00

#### Order

Another important variable for marketing purposes is Order, which represents how a customer made his or her first purchase. Suppose that the marketing department is interested only in customers purchasing through one of the following three channels:

- Internet
- Post
- Telephone

This might be because the business is considering moving away from selling through the branch network and would like to assess the impact of this decision on total sales by creating a forecast without this channel of business.

Output 1.2 shows that 43.49% of customers made their first purchase through the Internet. Both Order and Account\_Origin could be used as a stratus, to categorize these customers.

**Output 1.2: Descriptive Statistics: Order**

Order	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Internet	294259	43.49	294259	43.49
Post	126287	18.66	420546	62.15
Telephone	256089	37.85	676635	100.00

**Question 5: Who Are the Purchasers?**

The Who? ⑤ variables numbered 23 to 29 were selected by the marketing department to describe a customer’s profile. The real problem is that there is no easy way to describe a perfect customer, and all available information (AAI) is limited to what is captured in a company’s database. Sometimes, customer-context variables that have an emphasis on income, gender, internal scorecard, behavior score, and type of credit card are enough to provide a reasonable prediction of time-to-next-purchase.

---

## Data Set Creation

In this section, the process of data set creation and preparation is discussed in depth. This topic is of paramount importance because data preparation for a typical survival analysis project represents up to 80% of the project timeline. A minor mistake in univariate analysis or absence of a critical part of univariate analysis can substantially reduce the predictive ability of the model.

For the moment, concentrate on only the first five variables:

- **Id\_number.** A unique numeric variable that allows the identification of the 676,635 customers. It is a mandatory variable in the modeling, sampling, and validation process.
- **Start.** The variable Start is, by definition, when the analysis time is 0—that is, the time at which the customer completes his or her first purchase.
- **End.** The variable End indicates the time of the second purchase. This variable is missing if the customer does not make a second purchase, so it is intrinsically linked to the Censor variable.
- **Censor.** This is a binary target that indicates whether the customer made a second purchase in the time period after the first purchase was made.
  - Censor = 1 if the outcome is New Purchase in the period of study.
  - Censor = 0 if the outcome is No Purchase in the period of study.

In terms of modeling, the most difficult requirement is often precisely defining the Censor variable. In this model, it is a transition stage when the customer makes the second purchase.

---

### (Start, End) Calendar Timeline

Output 1.3 presents two required variables in survival analysis data—Start and End. These variables are in the calendar timeline format SAS DATE9. The variable Start marks the beginning of the contractual relationship between the supplier and the customer. This is also known as the starting point of the modeling process—the “beginning of the time” or the moment when the customer made the first purchase.

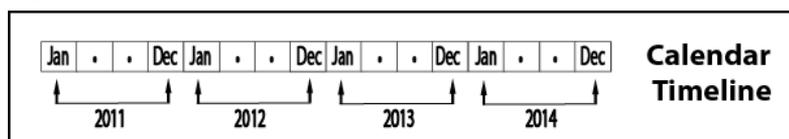
**Output 1.3: Variables Start and End**

Start	End
01JAN2011	04JAN2011
02JAN2011	05JAN2011
03JAN2011	17FEB2011
04JAN2011	06JAN2011
05JAN2011	18FEB2011
01JAN2011	08JAN2011
15JAN2011	25JAN2011
11JAN2011	19JAN2011
14JAN2011	16JAN2011
01JAN2011	27JAN2011

Modelers often encounter problems when defining the beginning of data collection. This can be attributed to a range of business-related issues such as merges between databases (when the company loses some customers), the implementation of new systems, or the acquisition of other businesses. The best solution for these types of problems is to design a timeline that coincides with how far back the oldest employees can give you accurate information about the historic problems faced by the company.

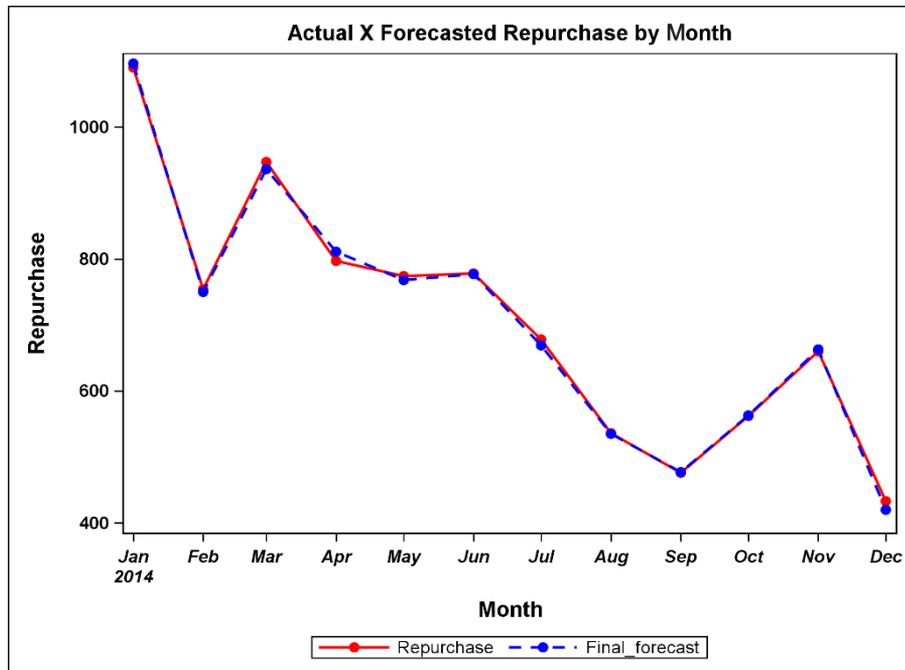
If there is an issue with the data prior to a certain month or period, modelers can choose an arbitrary Start time. This approach was adopted with the Garden data set. Following the completion of univariate analysis, it was decided that only customers purchasing after 1 January 2011 would be included in the final data set. Hence, the starting point of the study period is fixed to 01/JAN/2011, and the latest observation of the data set is fixed to 31/DEC/2014. SAS Enterprise Miner scans the database and creates a variable representing the censoring date based on the maximum date for the End variable and the interval time that you select (quarter, month, or day).

In Figure 1.3, the modeling period spans four years, which, in total, gives 48 observation points. This enables you to verify that the final model performs well throughout the course of the year, taking into account trends and seasonality such as purchases made for Christmas and Mother's Day.

**Figure 1.3: Calendar Time for Variables Start and End**

The variables Start and End are prerequisites for SAS Enterprise Miner and serve as a basis for the creation of the model's validation and performance graph in Figure 1.4.

Figure 1.4: Validation and Performance of Survival Analysis Model



## Time and Censor Variables

Time is the most important variable because it is the outcome variable of interest. Time is the number of months from the first purchase to the event (event being second purchase) or until the end of the period if the customer does not purchase again. Program 1.1 demonstrates how this variable is derived.

### Program 1.1: Computing the Time Variable

```

Data Time;
Set Garden (keep = Id_number Start End Censor);
Last_day = input ("31DEC2014" ①, anydtdte10.);
Time ③ = Intck ② ("MONTH", Start, Min(End, Last_day));
Format Last_day Date9.;
run;

```

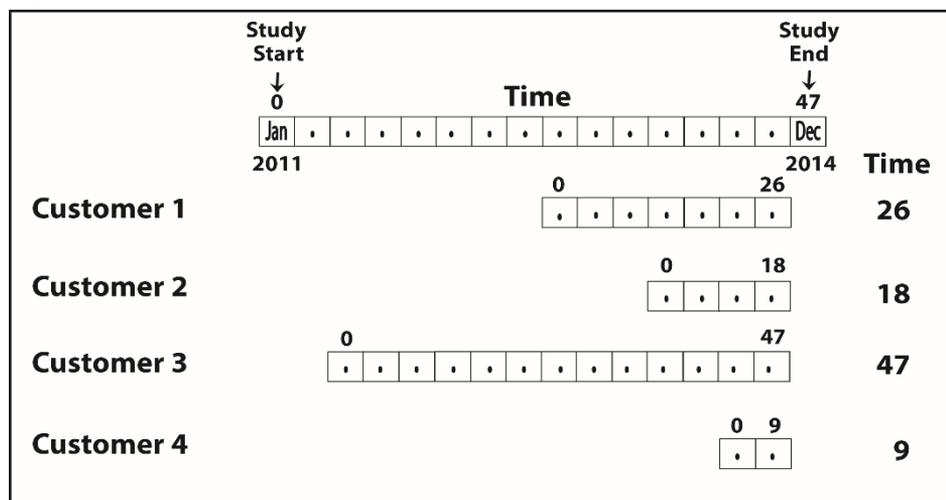
Before you create a predictive model using SAS Enterprise Miner, PROC LIFEREG, PROC PHREG, or any other procedure, you should first create the Time variable that represents the elapsed time between Start and the censoring date (Last\_day)—that is, the last day of data collection ①.

Although the data is collected daily, the Intck function ② creates the Time variable ③ by using a monthly interval. Therefore, the final model is created by month. The parameters of the Intck function can be summarized as follows: Intck ("INTERVAL", Start date, End date).

The Intck function returns the number of intervals (in months) between two dates, counting the number of months from the first value (Start) to the second (End).

One important feature of the Time variable is illustrated in Figure 1.5. Customers can start at any point during the data collection (study) period. Therefore, the Time variable needs to reflect the period when each customer is considered in the model.

Figure 1.5: Time Values for Different Customers



In Program 1.1, the function Min ensures that the results are correct for customers that are censored, which occurs when the End variable is missing (End = .). The function determines the smallest nonmissing value between the variables End and Last\_day. Last\_day represents the last day of data collection, also known as the censoring date. The Last\_day variable is used as a boundary in the modeling process, and SAS Enterprise Miner calculates it automatically for any survival model created.

Output 1.4 shows some selected customers from the data set named Time created by Program 1.1 to explain how they are processed in data preparation for modeling. The data was sorted by the variable Time starting in 01JAN2011.

Output 1.4: Selected Customers' Time to Purchase

Id_number	Start	End	Ⓣ Censor	Last_day	Time
40	01JAN2011	.	0	31DEC2014	③ 47
41	01JAN2011	.	0	31DEC2014	47
42	01JAN2011	.	0	31DEC2014	47
43	01JAN2011	04JAN2011	1	31DEC2014	④ 0
44	01JAN2011	04JAN2011	1	31DEC2014	0
45	01JAN2011	05JAN2011	1	31DEC2014	0
46	01JAN2011	05JAN2011	1	31DEC2014	0
47	01JAN2011	06JAN2011	1	31DEC2014	0
48	01JAN2011	07JAN2011	1	31DEC2014	0
49	01JAN2011	08JAN2011	1	31DEC2014	0
50	01JAN2011	10JAN2011	1	31DEC2014	0
51	01JAN2011	12JAN2011	1	31DEC2014	0
52	01JAN2011	16JAN2011	1	31DEC2014	0
53	01JAN2011	27JAN2011	1	31DEC2014	0
54	01JAN2011	27JAN2011	1	31DEC2014	0
55	01JAN2011	16FEB2011	1	31DEC2014	⑤ 1
56	01JAN2011	17FEB2011	1	31DEC2014	1
57	01JAN2011	20FEB2011	1	31DEC2014	1

Id_number	Start	End	Ⓣ Censor	Last_day	Time
58	01JAN2011	03APR2011	1	31DEC2014	Ⓣ 3
59	01JAN2011	11APR2011	1	31DEC2014	3
60	01JAN2011	11APR2011	1	31DEC2014	3

### Customers 40 to 42

Time has a computed value of 47, which was created by the Intek function Ⓣ. You see that customers 40, 41, and 42 made their first purchase on 01JAN2011 as indicated by the Start variable. Because the End variable is set to missing (End = .), you conclude that by the last day of the study period, these customers did not make a second purchase. These customers are still present in the data set at the end of the period of data collection, but they have not made a second purchase. So, these customers are referred to as censored, and the variable Censor is set to 0 Ⓣ.

You do not know how long it will take these customers to make a second purchase. This is an important feature of this data set.

### Customers 43 to 54

These customers are significantly different when they are compared to the first group of customers. Here, Time is the difference between Start and End, and its calculated value is 0 Ⓣ because the customers made another purchase before the end of January 2011, the month in which they made their first purchase. These customers are not censored, and the variable Censor is set to 1. This value represents death in terms of survival analysis. Customers who make a second purchase do not come back and they are treated as new. The irreversibility of the event of death is an essential feature of survival analysis.

### Customers 55 to 57

These customers made a second purchase in February 2011. The Time variable is calculated as 1 Ⓣ.

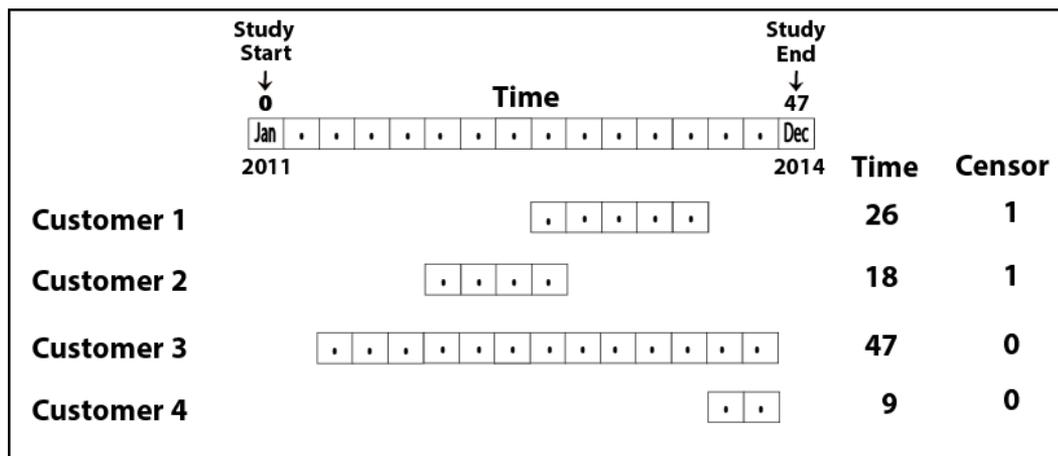
### Customers 58 to 60

These customers made a second purchase in April 2011. Therefore, the Time variable is calculated as 3 Ⓣ. Of the customers whose first purchase occurred in January, no one made a purchase in March 2011. Therefore, Time is not equal to 2 within this subset of customers.

### Censor and Time Variables for Selected Customers

To clarify the relationship between the Time and Censor variables, Figure 1.6 illustrates four customers. The graphic displays the experience of each customer even though you do not know when they made their first purchase.

Figure 1.6: Censor and Time for Four Customers



Customers 1 and 2, for example, both made a first purchase at different starting points. They are uncensored (Censor = 1) because they made a second purchase after 26 and 18 months, respectively.

Customer 3 made a first purchase in January 2011 (Time = 47). Customer 4 made a first purchase in March 2014 (Time = 9). They are censored (Censor = 0) because they did not make a second purchase after 47 and 9 months, respectively, until the end of the study in 12/31/2014.

The Time variable should be created using Program 1.1 at an early stage of the analysis. This program was created to match the SAS Enterprise Miner requirement. The Time variable is the basis of the construction of the Hazard function, which is discussed in detail in the next section.

---

## Strata by Previous Purchases

The Strata variable segments customers into categories based on specific criteria. For the Garden data set, the Strata variable is based on the number of historic purchases made by the customer. When a new customer makes their first purchase, they are assigned to the customer strata as “01”, “02”, “03”, and so on, depending on the number of purchases that they have made previously. Care must be taken when tracking each purchase made by the customer because each new purchase makes the initial state (first purchase) change to a second state (second purchase). Each period between purchases starts when the last purchase happens. In summary, each purchase is a new time (0). Every time that a customer repurchases, they start as a new customer.

**Key Tip:** Create a categorical variable based on the total number of previous purchases.

To stratify a continuous variable, you need to define the bins based on business experience. This can be challenging because stakeholders from different areas of the business might interpret the data differently. They might have conflicting opinions of how to bin the data. Business experience is a key factor to transform the distribution into a reduced number of bins.

Never create more than six bins. This ensures that there are robust volumes within each bin, as problems of convergence are likely to occur with inadequate volumes in each bin. If you do encounter this problem, PROC PHREG has the Firth’s method to fix it by adding a penalty term to the partial-log-likelihood. See PROC PHREG documentation in the *SAS/STAT User’s Guide* for more details.

A key finding from numerous research papers is that in some segments, customers with fewer purchases can be much more expensive to target when implementing marketing campaigns because they have a higher churn rate. Therefore, a key aspect to consider in the binning process is avoiding these segments of customers and placing emphasis on those customers who are more likely to increase the customer base.

Program 1.2 creates the Strata variable with six bins using a DATA step that can be customized for other models.

### Program 1.2: Binning Transformation

```
data strata;
  set total_orders;
  if total_orders in (1,2) and start < "01JAN2014"d then
    strata = "01-02-(2011-2013)"; ❶
  else if total_orders in (1,2) and start >= "01JAN2014"d then
    strata = "01-02-(2014)"; ❷
  else if 3 <= total_orders <= 5 then strata = "03-05";
  else if 5 < total_orders <= 10 then strata = "06-10";
  else if 10 < total_orders <= 20 then strata = "11-20";
  else if total_orders > 20 then strata = "21-High";
run;
```

The first strata, "01-02-(2011-2013)", contains customers that made their first purchase between 01JAN2011 and 31DEC2013, the first three years of the study period ❶. They made a maximum of two purchases (they either had 0, 1, or 2 purchases prior to 1 January 2011). You hypothesize that customers

who have not made an additional purchase after this time (more than three years) are the least likely to make any further purchases. You expect the model to allocate to customers in this segment (strata) the smallest probability of a second purchase.

In the second strata, "01-02-(2014)", customers have characteristics similar to those in the first strata ②. That is, they have less than two total purchases, but they made their first purchase in the last year of the study period, between 01JAN2014 and 31DEC2014. They are potential target customers for a marketing campaign because they are new to the database.

The other four strata represent the total number of purchases made without taking into account a time limit. Thus, the customer's full lifetime is considered.

One popular approach is to create a bin based on the start of each year. But, for time-to-next-purchase models, old customers are less likely to buy again. By creating a specific bin, you can compare old customers (using Survival and Hazard functions) with customers from recent years.

Another approach is one often used in econometrics, when some years perform better than others with respect to the economy. A hypothesis linked to the economy is that a marketing campaign performs better in a specific year. Then, to create a stratus for a specific year, the impact of the campaign on sales would be tracked. However, this approach is limited by the modeler's creativity, business knowledge, and understanding of their database.

Stratification can be created or considered from variables whose value is fixed at time equal to 0, such as gender or region.

---

## Categorical Variables

The variables numbered 23 to 29 in Figure 1.2 attempt to describe a customer's behavior. These variables help identify characteristics that are indicative of good customers. To identify good customers effectively, customers should be ranked by the number of purchases or by using a scorecard. In Output 1.5, consider the internal credit score at the time of application, segmenting the customers into 200-point bins with the best customers in the final bin 800-1000 as follows:

**Output 1.5: Credit Score from Bureau Data Binned**

Credit_score	Frequency	Cumulative Frequency
0 - 400	157793	157793
400 - 600	73994	231787
600 - 800	118060	349847
800 - 1000	200765	550612
Missing	126023	676635

An interesting approach is to bin the variables Age, Credit\_score, and Behavior. This is a typical framework derived from summarizing empirical results from other published models using categorical and binned variables.

Over the past 10 years, there has been an expansion in customer research, specifically marked by some trends such as credit card behavior, postal code for geo-localization, and text mining. Tracking these trends enables you to collect more data about a customer's experience. The collection of this additional data enables you to create internal variables, which could be considered as input variables to predict a customer's behavior.

Although income has previously been considered as a principal variable for predicting customer behavior, recent internal research for web analysis investigates the effect of discount prices on customer behavior. Early results show that this is a more powerful characteristic when considering a customer's behavior.

Binning variables is important for a variety of reasons, but its primary purpose is to provide a framework (structure) for the Hazard function. Once you have established a clear definition of the binning process, you can assess the impact of the binned variable on the model by checking the sensitivity of the model with and without the binned variable.

In Output 1.6, data is binned by the Age variable.

**Output 1.6: Age Variable Binned**

Age	Frequency	Cumulative Frequency
18 - 31	58902	58902
32 - 43	72487	131389
44 - 56	125824	257213
57 - 68	164404	421617
69 - 81	155015	576632
82 - high	46324	622956
Missing	53679	676635

The goals and output of the model depend on your company's needs. You could have any of the following variables:

- Geographic
- Products
- Gender
- Seasonality
- Price

Without a deep understanding of how the customers are segmented, you will have more difficulty interpreting the coefficients and hazard results.

In summary, segmenting continuous variables and regrouping discrete ones into smaller categories are the key aspects of data preparation in survival analysis.

---

## Data Preparation with Base SAS, Macros, and SAS/STAT

This section presents the different approaches using Base SAS, SAS macros, and SAS/STAT for data preparation. Techniques and strategies are presented.

---

### Multiple Imputation

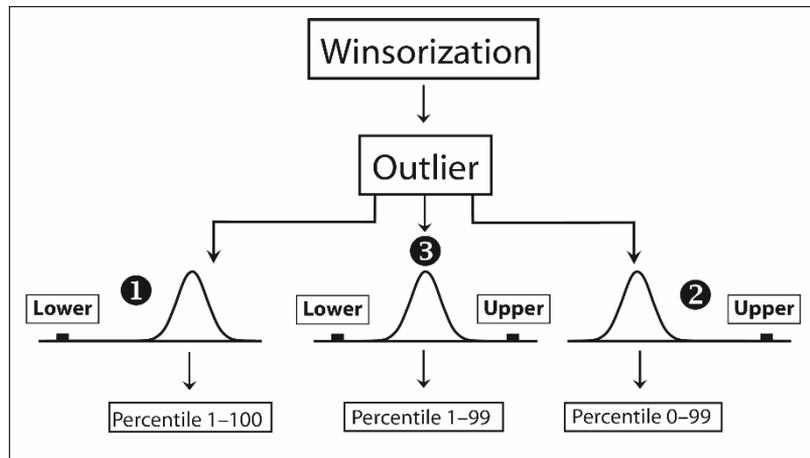
*Multiple imputation* for missing data is such a complex technique that SAS published the book *Multiple Imputation of Missing Data Using SAS* by Patricia Berglund and Steven Heeringa. How should missing data values be treated? There is no easy answer. Choosing the “best” missing value replacement technique inherently requires the modeler to make assumptions about the true (missing) data. It is beyond the scope of this book to extend the discussion about these methods.

For further details, read about the MI procedure and the MIANALYZE procedure in the *SAS/STAT User's Guide*. There, you will find a detailed explanation of missing methods. The documentation shows how the MI procedure can create multiple imputed data sets for incomplete multivariate data and how the MIANALYZE procedure can generate valid statistical inferences about parameters by combining results from the imputations.

## Winsorization

*Winsorization* is a technique named after the engineer-biostatistician Charles P. Winsor. This process is sometimes called *trimming* or *truncation*. Winsorization is a method of censoring. It involves replacing the original data by an arbitrary percentile that you believe is an acceptable value. For example, a 95% Winsorization would replace all data above the 95th percentile by the value of the 95th percentile. This process is standard in modeling in financial services and insurance companies. It can be implemented using PROC UNIVARIATE or macros (Löeffler and Poach 2011). Figure 1.7 provides an example of Winsorization.

**Figure 1.7: Winsorization**



The maximum percentile distribution ranges from 0 to 100 and covers the full distribution of a variable. Winsorization can be used to customize the range of the distribution so that the range of the variable falls between a limited range of percentiles.

In Figure 1.7, you see that the outlier is located in the extreme negative area, called the lower area of the distribution ❶. A 1% Winsorization replaces all data below the 1st percentile by the value of the 1st percentile.

An outlier is also located in the extreme positive area, called the upper area of the distribution ❷. A 99% Winsorization replaces all data above the 99th percentile by the value of the 99th percentile. This technique with the upper and lower outliers is called one-sided Winsorization.

In the first distribution graph in Figure 1.7, you see the percentile values 1 to 100. In other words, you are replacing the outlier from the lower area by the 1<sup>st</sup> percentile, but you are not replacing the right side. The 100th percentile stays the same value because it is not an outlier. The same reasoning is valid in the third distribution graph, where you see percentile values 0 to 99. You replace the values above the 99<sup>th</sup> percentile, but the left side stays unchanged. It is represented by 0, the minimum value of the variable.

In the second distribution graph, you apply simultaneously both the 1<sup>st</sup> percentile and the 99<sup>th</sup> percentile because the variable has Lower and Upper outliers. This technique is called two-sided Winsorization.

One reason that Winsorization is popular within the financial industry is because when you exclude an observation, you also exclude all the information present in the other variables of that observation, whether they have outliers or not. The main advantage of Winsorization compared to the outright exclusion of an observation is that this information remains available during the modeling process.

## Binning Using Macros

Binning continuous variables needs someone who has an expert level of Base SAS programming and who has used the SAS Macro Language. Mamdouh Refaat (2007) implemented these techniques very well using advanced macros:

**Equal Width**

creates bounds by dividing the variable into an equal number of ranges of equal width.

**Equal Height**

creates upper and lower limits for each bin.

**Optimal Binning**

attempts to minimize the frequencies of the observations in the new bins.

---

## Data Preparation with SAS Enterprise Miner

The preparation process with SAS Enterprise Miner can modify the data by creating new variables. Variables can be selected and transformed to improve the model-selection process by techniques such as the stepwise selection. The Modify step, which is one of the SEMMA steps, includes the use of advanced tools and awards algorithms for defining transformations, missing-value handling, values recoding, and interactive binning:

- **S**—Sample
- **E**—Explore
- **M**—Modify
- **M**—Model
- **A**—Assess

Furthermore, advanced visualization tools enable you to quickly and easily examine large amounts of data in multidimensional histograms and to graphically compare different models and transformations. After you have completed the Modify step, you can apply the scoring modeling formula to a new data set without any Base SAS manipulation or macros. You can then select the champion model with the best binning transformation.

The time to select the final model can be reduced from three weeks for a standard survival analysis project to one or two days.

Figure 1.8 shows an example of data preparation using a SAS Enterprise Miner process flow, which will be presented in detail in Chapter 4. Table 1.1 provides details.

**Figure 1.8: Data Preparation Process Flow**

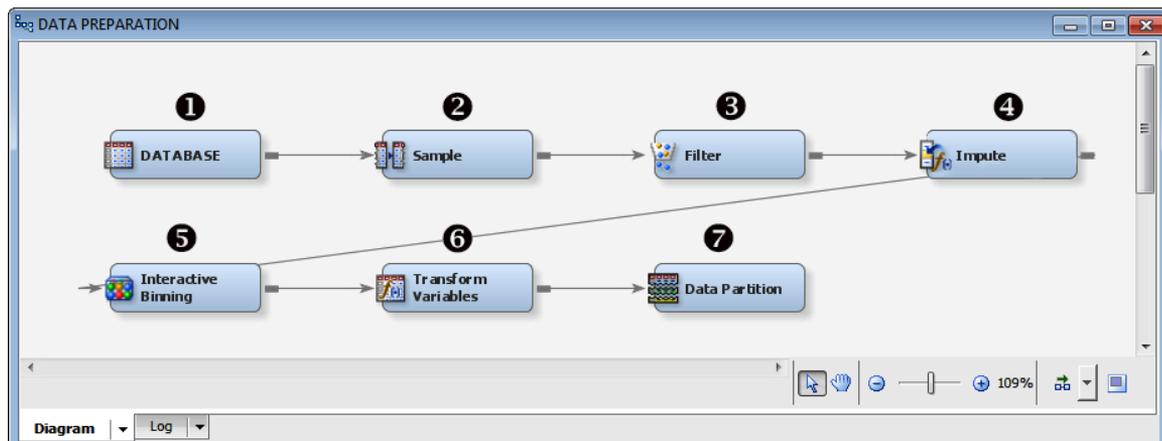
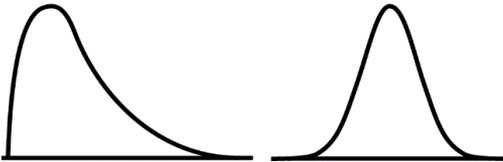


Table 1.1: SAS Enterprise Miner Nodes for Data Preparation

Data Preparation	Node	Description
Input data	 DATABASE	The input data node specifies the data set (Garden) used in this project. It is the first node in the process flow diagram.
Sample	 Sample	The Sample node enables you to extract a sample from the input data source (Garden). The node and the sampling technique are covered in detail in Chapter 4.
Exclude outliers	 Filter	The Filter node excludes outliers and enables the user to choose cutoffs and different methods.
Replace missing data	 Impute	The Impute node enables you to replace missing values in the data set before you start the modeling process. The node has more than 12 different imputation techniques to replace missing values.
Bin continuous variables	 Interactive Binning	<p>One of the more important tools in data preparation in SAS Enterprise Miner is the Interactive Binning node, which can create bins, buckets, or classes of all input variables. These include both class and interval input variables. You can create bins to reduce the number of unique levels as well as to attempt to improve the predictive power of each input variable.</p> <p>In the Interactive Binning node, the predictive power of a variable (its ability to separate high-risk survival customers from low-risk ones) is assessed by its Gini Concentration Ratio by grouping the selected characteristics based on business considerations. The node is helpful in shaping the data to reduce overfitting. In most cases, binning attributes leads to stronger predictive power and improvement in the survival model without any Base SAS preparation. The code generated by the Interactive Binning node can be saved, exported, and implemented using Base SAS.</p>
Transform variables	 Transform Variables	<p>The Transform Variables node transforms variables to improve normality and the fit of the model as presented below:</p> <div style="text-align: center;"> <p><b>Before Transformation</b>                      <b>After Transformation</b></p>  </div>
Partition data	 Data Partition	A Data Partition node can statistically split a data set into three data sets: one for training, one for validation, and one for scoring. The one for scoring is not used in building the model, but is used as a holdout sample. These data sets are carried along through the process flow and used during the model assessment.

A detailed explanation of each node can be found in SAS Enterprise Miner documentation.

# About This Book

---

## What Does This Book Cover?

The primary purpose of this book, beyond introducing underlying survival analysis theories, methods, and requirements, is to illustrate how to build a time-to-next-purchase survival model in SAS Enterprise Miner. It explains each step with regard to statistics and to Base SAS and SAS/STAT (with which the reader might be more familiar with than SAS Enterprise Miner). It addresses the development and application of survival analysis techniques to common scenarios faced by financial services, insurance companies, marketing, and the telecommunications industry. These scenarios include the following:

- Time-to-next-purchase for marketing
- Employer turnover for human resources
- Small business portfolio macroeconomic stress test for banks
- Mortgage International Financial Reporting Standard (IFRS 9) lifetime probability of default for building societies
- Churn or attrition models for mobile and insurance industries

---

## Business Concepts

This book emphasizes business problems involving time-to-next-purchase and churn.

### Time-to-Next-Purchase

The objective of a time-to-next-purchase model is to predict the time until the next purchase for customers of a retail outlet. Thereby, the model addresses common business concerns, including the following:

- When and for how long to undertake an advertising campaign
- Resource planning for staff distribution, call centers, and the like, on the basis of anticipated demand
- Required stock of products subject to seasonality effects
- Customer churn (change and turn)—the proportion of customers who leave each month

Two primary objectives in marketing are to know which products to offer next to a customer and when to send targeted advertisements to customers who are likely to be responsive during certain months of the year. By targeting an appropriate subset of customers, a marketing department can reduce the amount of wasted time, effort, and costs of marketing on customers who are unlikely to respond.

The emphasis is on *which customers* make a second purchase through a web page, catalog, or branch, and the *duration* between the first and second purchase. Accordingly, the focus of the model is on *time to event*, which, in this case, is the time to repurchase. Survival analysis differs from logistic regression in that beyond looking at a binary outcome (such as whether a customer makes another purchase in a fixed period), it looks at the time to event: namely, when a second purchase will occur.

### Churn

In this progressively more digital age, a retail outlet has to have an efficient marketing strategy as competition intensifies from new entrants in catalogs and online. The market is saturated and aggressive, and new customers are both hard and expensive to find. To be competitive, the company requires a strategy for long-term retention of customers and to avoid churn. Equally, it needs to track the impact of initiatives to persuade new customers to repurchase.

---

## Overview of Chapters

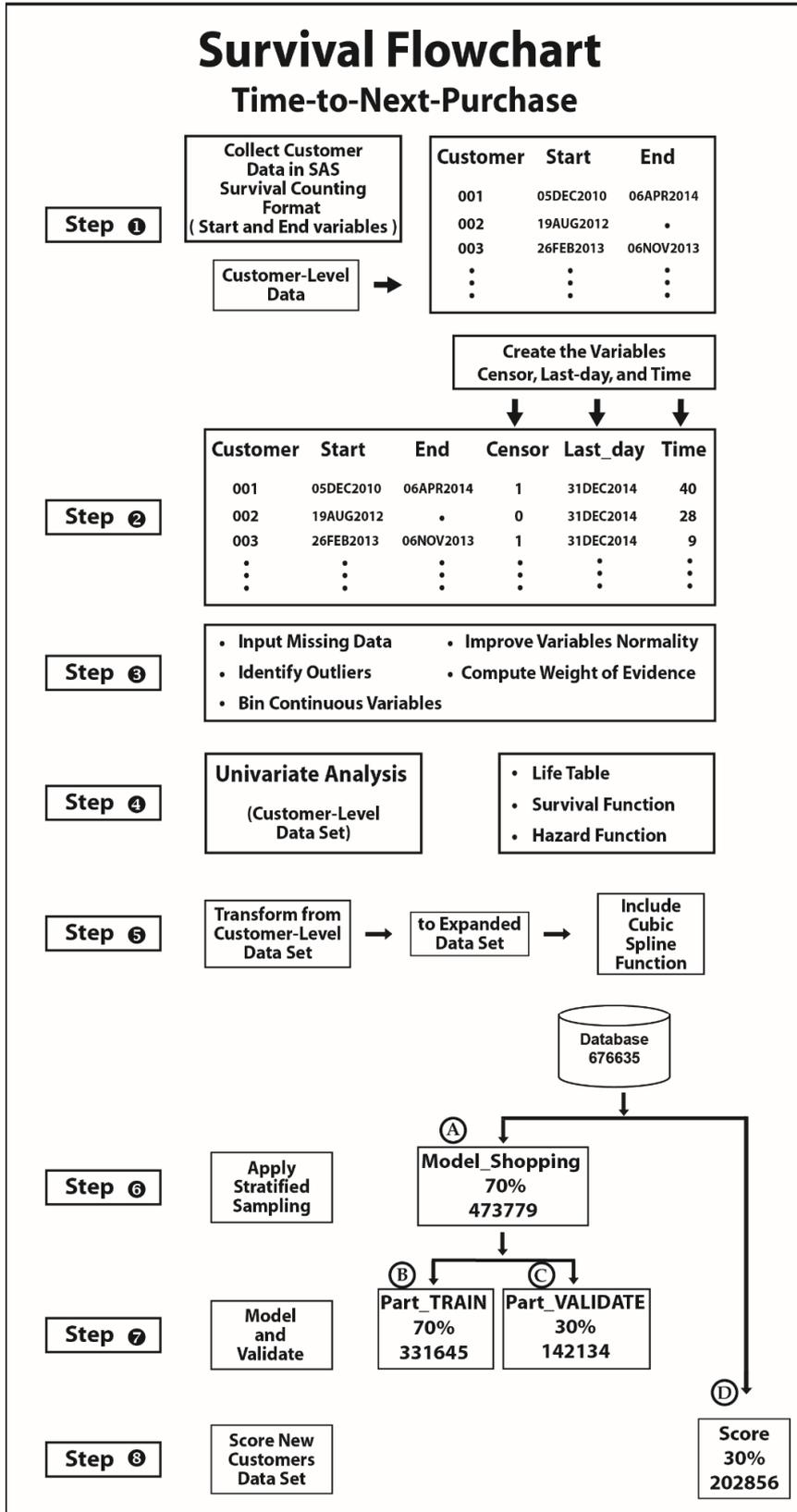
This book teaches survival analysis using its first principles. It highlights its relevance to real-world business cases. It leads the reader through an actual example to contextualize the statistical methods and coding while understanding that this is an introduction to survival analysis models.

To facilitate learning, a single example is discussed throughout the book. The example illustrates, using minimum statistical theory, how to create a survival analysis model to predict time-to-next-purchase (or repurchase for customers who have already made a first purchase in the past). The book emphasizes survival analysis techniques that can be easily implemented, rather than delving into the heavy detail of the underlying mathematics. As such, it serves as a fundamentals course for understanding problems and techniques that are usually described only in more advanced books and papers.

This book is an introduction to survival analysis. But, unlike other books, this one is focused on business applications using data mining and regression, not on pharmaceutical or medical research. The following elements are presented in detail:

- Continuous and discrete time-to-event definitions
- Censor variables
- Survival function
- Hazard function
- Time-dependent variables and expanded data sets
- Cubic spline function and logistic regression models
- Specific sample methods for survival analysis
- Statistics for validation, such as the following:
  - Depth
  - Lift
  - Benefit
  - Concentration curve
  - Gini Concentration Ratio
  - Kolmogorov-Smirnov Statistic
- Customer scoring

The chapters in this book lead you through the steps to build a time-to-next-purchase survival model, specifically, and survival models, more generally. By following the steps shown in the following flowchart, you can develop any survival analysis model:



## **Chapter 1: Data Preparation for Survival Models**

Chapter 1 focuses on the problems addressed by survival analysis: objectives, notation, and terminology. The Garden data set and creation of the required variables Censor and Time are presented. Issues such as defining the target (or dependent) variable are explained. The process starts with data sample collection. Next is the Time definition, which includes discussion of important considerations for univariate analysis.

## **Chapter 2: Exploratory Data Analysis and Basic Statistical Concepts**

Chapter 2 introduces the logistic regression model and teaches basic statistical theory concerning the Survival and Hazard functions. This crucial chapter demonstrates the link between the Hazard and Survival functions and the logistic regression model.

## **Chapter 3: Stratified Sampling Method for Survival Analysis**

Chapter 3 shows you how to create the unbiased stratified samples required for developing a survival analysis model.

## **Chapter 4: SAS Enterprise Miner Use from Source to Customer-Level Model Output**

Chapter 4 shows you how to navigate, start a new project, create a data source, and complete other tasks in SAS Enterprise Miner. The focus is on creating the relevant process flow in SAS Enterprise Miner to develop, in detail, a model using the Survival node.

## **Chapter 5: The Cubic Spline Regression and Model Interpretation**

Chapter 5 presents the cubic spline basis variables, which consist of a function joined at points called *knots*. The chapter shows you how the incorporation of the splines improves the survival model by adding flexibility to the Hazard function. The focus is on the interpretation of the coefficients and the cubic spline basis variables within the logistic survival model. You learn how to interpret SAS Enterprise Miner survival model output for continuous and categorical variables and about methods of checking the adequacy of a fitted model. By the end of the chapter, you will understand the meaning of the Hazard function and splines, as well as their relationships in the modeling process of survival analysis.

## **Chapter 6: The Fully Expanded Data Set Model**

Chapter 6 shows you how to create a model using a fully expanded data set format. This model is compared to the customer-level model created in Chapter 4.

## **Chapter 7: Statistics for Performance and Validation**

Chapter 7 is devoted to all the statistical measures necessary to assess the performance and validation of survival models:

- Depth
- Lift
- Benefit
- Concentration curve
- Gini Concentration Ratio
- Kolmogorov-Smirnov Statistic

## **Chapter 8: The Scoring of New and Old Customers**

In Chapter 8, you develop an algorithm to score and validate new customers for a different date or time period.

---

## Is This Book for You?

This book is intended for the following:

- Graduates of economics, business, and marketing programs
- Analysts who want to create statistical models and are working in areas such as credit risk

The examples in this book show people with limited modeling experience how to apply models. They present techniques in a way that avoids high-level theoretical considerations and detailed advanced topics.

---

## What Are the Prerequisites for This Book?

Ideally, the reader should be familiar with linear regression, but statistics is kept at a minimum level to focus on the interpretation of SAS output at each stage of the modeling process.

Familiarity with statistical modeling is beneficial, but it is not a *must-have* condition for the reader. Basic knowledge of SAS 9.4 is recommended, but not of SAS macros. Minimum knowledge of SAS Enterprise Miner enables the reader to understand each of the steps presented in the flowchart.

It is assumed that the reader is familiar with Base SAS, including the following:

- Assigning a SAS library
- Basic DATA steps such as creating SAS data sets and SAS variables
- SAS functions
- Procedures (statements and options)

---

## What Should You Know about the Examples?

This book includes tutorials for you to follow to gain hands-on experience with SAS.

---

### Software Used to Develop the Book's Content

The book covers the use of SAS statistical programming (Base SAS, SAS 9.4, SAS/STAT 14.1, and SAS Enterprise Miner 14.1).

The large volume and granularity of data, combined with the complexity of the survival model calculations applied to the data to score customers' probabilities to repurchase, create substantial pressure on the delivery of output (in essence, a list of customers likely to repurchase in the next month). Marketing managers simply will not be able to wait 10 days, for example, to run a full recalculation based on scoring codes that change every month. As a result, marketing departments and banks need new, high-performance technology that can scale to meet business needs.

SAS Enterprise Miner can achieve optimal results with minimal human intervention to develop and implement survival data mining models. It provides an environment that supports efficient documentation, a strong sample process, flexibility to change time periods, model management, traceability, workflow, and audit trails. The modeler needs only to understand how to rerun calculations and analyses.

For this reason, it is important that the underlying end-to-end Survival Flowchart be traceable (for finance and audit purposes). The Survival Flowchart needs to be usable by marketing analysts so that they can perform this work themselves, rather than having to rely on data miners or consultants to implement a new monthly advertising campaign.

---

### Example Code and Data

The depth and guidance of this book are sufficient to begin developing any business- or finance-related survival analysis model. The book is designed to be read linearly, from Chapter 1 through to Chapter 8, gradually demystifying the complexity of survival analysis. However, this book should be *applied* rather than just *read*. The same can be said about cookbooks, whose practical requirements educate you through

trial and error. Consequently, to succeed, run the SAS code of each example. To understand the necessary modification at each step, open the respective SAS data set. Think about any mistakes, and try again.

The Garden data set and all variables are explained in detail in Chapter 1.

You can access the example code and data for this book by linking to its author page at <https://support.sas.com/authors>.

---

## **SAS University Edition**

If you are using SAS University Edition to access data and run programs, then check the SAS University Edition page to ensure that the software contains the product or products that you need to run the code: [www.sas.com/universityedition](http://www.sas.com/universityedition).

---

## **Output and Graphics**

All the output and graphics in this book were created by SAS Enterprise Miner 14.1 or by the SAS 9.4 SGPLOT procedure. Some graphics were modified using Adobe Photoshop and Adobe Illustrator for pedagogical reasons. The respective Adobe Photoshop (PSD) and Adobe Illustrator (AI) files are available from the author by request. These graphics can be useful for model validation documentation, validation reports, or PowerPoint presentations.

---

## **We Want to Hear from You**

SAS Press books are written *by* SAS users *for* SAS users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit <https://support.sas.com/publishing> to do the following:

- Sign up to review a book
- Recommend a topic
- Request authoring information
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through [saspress@sas.com](mailto:saspress@sas.com) or [https://support.sas.com/author\\_feedback](https://support.sas.com/author_feedback).

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: <https://support.sas.com/publishing>.

## About The Author



Jorge Ribeiro currently works in the area of econometric model management and development in the retail credit risk industry. He has worked previously as a stress test methodology manager at Yorkshire Bank, head of modeling at Direct Line Group Insurance, Vice President of Barclays Bank in the internal validation unit, head of modeling at HML Mortgages (IFRS 9), and principal data mining and modeler consultant at JD Williams & Co. in the United Kingdom. As a former professor of mathematics, he has more than 20 years of academic experience in advanced econometric techniques, such as vector autoregressive and Bayesian analyses, as well as rational expectations and brand awareness with latent variables using factor analysis and constraint optimization for call center management. He has used SAS since 1986 and has attended more than 50 SAS training courses and has presented at conferences worldwide over the past 25 years. Jorge holds a master's degree in economics from the Fluminense Federal University in Rio de Janeiro. He completed postgraduate work in financial modeling using SAS, and he attended the doctoral program in financial econometrics at the Université de Nantes.

Learn more about this author by visiting his author page at [http://support.sas.com/ribeiro\\_j](http://support.sas.com/ribeiro_j). There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

# Ready to take your SAS<sup>®</sup> and JMP<sup>®</sup> skills up a notch?



Be among the first to know about new books,  
special events, and exclusive discounts.

**[support.sas.com/newbooks](https://support.sas.com/newbooks)**

Share your expertise. Write a book with SAS.

**[support.sas.com/publish](https://support.sas.com/publish)**

 [sas.com/books](https://sas.com/books)  
for additional books and resources.

  
THE POWER TO KNOW.®