

# Developing Credit Risk Models Using SAS<sup>®</sup> Enterprise Miner<sup>™</sup> and SAS/STAT<sup>®</sup>

Theory and Applications



Iain L. J. Brown, PhD



From *Developing Credit Risk Models Using SAS® Enterprise Miner™ and SAS/STAT®*. Full book available for purchase [here](#).

## Contents

<b>About this Book</b> .....	<b>ix</b>
<b>About the Author</b> .....	<b>xiii</b>
<b>Acknowledgments</b> .....	<b>xv</b>
<b>Chapter 1 Introduction</b> .....	<b>1</b>
1.1 Book Overview.....	1
1.2 Overview of Credit Risk Modeling .....	2
1.3 Regulatory Environment.....	3
1.3.1 Minimum Capital Requirements.....	4
1.3.2 Expected Loss.....	5
1.3.3 Unexpected Loss .....	6
1.3.4 Risk Weighted Assets .....	6
1.4 SAS Software Utilized .....	7
1.5 Chapter Summary .....	11
1.6 References and Further Reading .....	11
<b>Chapter 2 Sampling and Data Pre-Processing</b> .....	<b>13</b>
2.1 Introduction .....	13
2.2 Sampling and Variable Selection.....	16
2.2.1 Sampling .....	17
2.2.2 Variable Selection .....	18
2.3 Missing Values and Outlier Treatment.....	19
2.3.1 Missing Values .....	19
2.3.2 Outlier Detection.....	21
2.4 Data Segmentation .....	22
2.4.1 Decision Trees for Segmentation .....	23
2.4.2 K-Means Clustering.....	24

2.5 Chapter Summary .....	25
2.6 References and Further Reading .....	25
<b>Chapter 3 Development of a Probability of Default (PD) Model .....</b>	<b>27</b>
3.1 Overview of Probability of Default.....	27
3.1.1 PD Models for Retail Credit.....	28
3.1.2 PD Models for Corporate Credit .....	28
3.1.3 PD Calibration .....	29
3.2 Classification Techniques for PD .....	29
3.2.1 Logistic Regression.....	29
3.2.2 Linear and Quadratic Discriminant Analysis .....	31
3.2.3 Neural Networks .....	32
3.2.4 Decision Trees .....	33
3.2.5 Memory Based Reasoning.....	34
3.2.6 Random Forests .....	34
3.2.7 Gradient Boosting.....	35
3.3 Model Development (Application Scorecards) .....	35
3.3.1 Motivation for Application Scorecards.....	36
3.3.2 Developing a PD Model for Application Scoring .....	36
3.4 Model Development (Behavioral Scoring) .....	47
3.4.1 Motivation for Behavioral Scorecards.....	48
3.4.2 Developing a PD Model for Behavioral Scoring .....	49
3.5 PD Model Reporting.....	52
3.5.1 Overview .....	52
3.5.2 Variable Worth Statistics .....	52
3.5.3 Scorecard Strength .....	54
3.5.4 Model Performance Measures .....	54
3.5.5 Tuning the Model.....	54
3.6 Model Deployment .....	55
3.6.1 Creating a Model Package .....	55
3.6.2 Registering a Model Package .....	56
3.7 Chapter Summary .....	57
3.8 References and Further Reading .....	58

<b>Chapter 4 Development of a Loss Given Default (LGD) Model.....</b>	<b>59</b>
4.1 Overview of Loss Given Default.....	59
4.1.1 LGD Models for Retail Credit .....	60
4.1.2 LGD Models for Corporate Credit.....	60
4.1.3 Economic Variables for LGD Estimation .....	61
4.1.4 Estimating Downturn LGD .....	61
4.2 Regression Techniques for LGD.....	62
4.2.1 Ordinary Least Squares – Linear Regression .....	64
4.2.2 Ordinary Least Squares with Beta Transformation .....	64
4.2.3 Beta Regression .....	65
4.2.4 Ordinary Least Squares with Box-Cox Transformation .....	66
4.2.5 Regression Trees.....	67
4.2.6 Artificial Neural Networks.....	67
4.2.7 Linear Regression and Non-linear Regression .....	68
4.2.8 Logistic Regression and Non-linear Regression.....	68
4.3 Performance Metrics for LGD.....	69
4.3.1 Root Mean Squared Error .....	69
4.3.2 Mean Absolute Error .....	70
4.3.3 Area Under the Receiver Operating Curve .....	70
4.3.4 Area Over the Regression Error Characteristic Curves .....	71
4.3.5 R-square .....	72
4.3.6 Pearson’s Correlation Coefficient.....	72
4.3.7 Spearman’s Correlation Coefficient .....	72
4.3.8 Kendall’s Correlation Coefficient.....	73
4.4 Model Development.....	73
4.4.1 Motivation for LGD models.....	73
4.4.2 Developing an LGD Model.....	73
4.5 Case Study: Benchmarking Regression Algorithms for LGD .....	77
4.5.1 Data Set Characteristics .....	77
4.5.2 Experimental Set-Up .....	78
4.5.3 Results and Discussion.....	79
4.6 Chapter Summary .....	83
4.7 References and Further Reading .....	84

<b>Chapter 5 Development of an Exposure at Default (EAD) Model .....</b>	<b>87</b>
5.1 Overview of Exposure at Default .....	87
5.2 Time Horizons for CCF .....	88
5.3 Data Preparation .....	90
5.4 CCF Distribution – Transformations.....	95
5.5 Model Development .....	97
5.5.1 Input Selection .....	97
5.5.2 Model Methodology.....	97
5.5.3 Performance Metrics.....	99
5.6 Model Validation and Reporting .....	103
5.6.1 Model Validation .....	103
5.6.2 Reports .....	104
5.7 Chapter Summary .....	106
5.8 References and Further Reading .....	107
<b>Chapter 6 Stress Testing .....</b>	<b>109</b>
6.1 Overview of Stress Testing .....	109
6.2 Purpose of Stress Testing .....	110
6.3 Stress Testing Methods.....	111
6.3.1 Sensitivity Testing.....	111
6.3.2 Scenario Testing .....	112
6.4 Regulatory Stress Testing .....	113
6.5 Chapter Summary .....	114
6.6 References and Further Reading .....	114
<b>Chapter 7 Producing Model Reports .....</b>	<b>115</b>
7.1 Surfacing Regulatory Reports .....	115
7.2 Model Validation.....	115
7.2.1 Model Performance .....	116
7.2.2 Model Stability .....	122
7.2.3 Model Calibration .....	125
7.3 SAS Model Manager Examples.....	127
7.3.1 Create a PD Report .....	127
7.3.2 Create a LGD Report.....	129
7.4 Chapter Summary .....	130

<b>Tutorial A – Getting Started with SAS Enterprise Miner.....</b>	<b>131</b>
A.1 Starting SAS Enterprise Miner .....	131
A.2 Assigning a Library Location .....	134
A.3 Defining a New Data Set.....	136
<b>Tutorial B – Developing an Application Scorecard Model in SAS Enterprise Miner.....</b>	<b>139</b>
B.1 Overview .....	139
B.1.1 Step 1 – Import the XML Diagram .....	140
B.1.2 Step 2 – Define the Data Source.....	140
B.1.3 Step 3 – Visualize the Data.....	141
B.1.4 Step 4 – Partition the Data .....	143
B.1.5 Step 5 –Perform Screening and Grouping with Interactive Grouping .....	143
B.1.6 Step 6 – Create a Scorecard and Fit a Logistic Regression Model .....	144
B.1.7 Step 7 – Create a Rejected Data Source .....	144
B.1.8 Step 8 – Perform Reject Inference and Create an Augmented Data Set .....	144
B.1.9 Step 9 – Partition the Augmented Data Set into Training, Test and Validation Samples .....	145
B.1.10 Step 10 – Perform Univariate Characteristic Screening and Grouping on the Augmented Data Set .....	145
B.1.11 Step 11 – Fit a Logistic Regression Model and Score the Augmented Data Set .....	145
B.2 Tutorial Summary .....	146
<b>Appendix A Data Used in This Book .....</b>	<b>147</b>
A.1 Data Used in This Book.....	147
Chapter 3: Known Good Bad Data.....	147
Chapter 3: Rejected Candidates Data .....	148
Chapter 4: LGD Data .....	148
Chapter 5: Exposure at Default Data .....	149
<b>Index .....</b>	<b>151</b>

From *Developing Credit Risk Models Using SAS® Enterprise Miner™ and SAS/STAT®: Theory and Application*, by Iain Brown. Copyright © 2014, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Developing Credit Risk Models Using SAS® Enterprise Miner™ and SAS/STAT®*. Full book available for purchase [here](#).

## Chapter 1 Introduction

<b>1.1 Book Overview .....</b>	<b>1</b>
<b>1.2 Overview of Credit Risk Modeling .....</b>	<b>2</b>
<b>1.3 Regulatory Environment .....</b>	<b>3</b>
1.3.1 Minimum Capital Requirements .....	4
1.3.2 Expected Loss .....	5
1.3.3 Unexpected Loss.....	6
1.3.4 Risk Weighted Assets .....	6
<b>1.4 SAS Software Utilized.....</b>	<b>7</b>
<b>1.5 Chapter Summary.....</b>	<b>11</b>
<b>1.6 References and Further Reading.....</b>	<b>11</b>

---

### 1.1 Book Overview

This book aims to define the concepts underpinning credit risk modeling and to show how these concepts can be formulated with practical examples using SAS software. Each chapter tackles a different problem encountered by practitioners working or looking to work in the field of credit risk and give a step-by-step approach to leverage the power of the SAS Analytics suite of software to solve these issues.

This chapter begins by giving an overview of what credit risk modeling entails, explaining the concepts and terms that one would typically come across working in this area. We then go on to scrutinize the current regulatory environment, highlighting the key reporting parameters that need to be estimated by financial institutions subject to the Basel capital requirements. Finally, we discuss the SAS analytics software used for the analysis part of this book.

The remaining chapters are structured as follows:

**Chapter 2** covers the area of sampling and data pre-processing. This chapter defines and contextualizes issues such as variable selection, missing values, and outlier detection within the area of credit risk modeling, and gives practical applications of how these issues can be solved.

**Chapter 3** details the theory and practical aspects behind the creation of Probability of Default (PD) models. This focuses on standard and novel modeling techniques, shows how each of these can be used in the estimation of PD, and demonstrates the full development of an application and behavioral scorecard using SAS Enterprise Miner.

**Chapter 4** focuses on the development of Loss Given Default (LGD) models and the considerations with regard to the distribution of LGD that have to be made for modeling this parameter. A variety of modeling approaches are discussed and compared in a case study in order to show how improvements over the traditional industry approach of linear regression can be made.

**Chapter 5** defines the concept of Exposure at Default (EAD) and how this parameter is formulated and estimated. A full model development process is shown through practical examples. The aim of this chapter is to fully explore the implications of model choice, input variables, and how best to estimate EAD.

**Chapter 6** defines and explains the concepts of stress testing under the three pillars of the Basel Capital Accord and what this entails for financial institutions.

**Chapter 7** focuses on how model reports can be generated from the procedures and methodologies created throughout this book. This chapter covers the key reporting outputs required within the regulatory framework and shows through SAS Model Manager and example code how these outputs can be created.

By the conclusion of this book, readers will have a comprehensive guide to developing credit risk models both from a theoretical and practical perspective. We also aim to show how analysts can create and implement credit risk models using example code and projects in SAS.

---

## 1.2 Overview of Credit Risk Modeling

With cyclical financial instabilities in the credit markets, the area of credit risk modeling has become ever more important, leading to the need for more accurate and robust models. Since the introduction of the Basel II Capital Accord (Basel Committee on Banking Supervision, 2004) over a decade ago, qualifying financial institutions have been able to derive their own internal credit risk models under the advanced internal ratings-based approach (A-IRB) without relying on regulator's fixed estimates.

The Basel II Capital Accord prescribes the minimum amount of regulatory capital an institution must hold so as to provide a safety cushion against unexpected losses. Under the advanced internal ratings-based approach (A-IRB), the accord allows financial institutions to build risk models for three key risk parameters: Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). PD is defined as the likelihood that a loan will not be repaid and will therefore fall into default. LGD is the estimated economic loss, expressed as a percentage of exposure, which will be incurred if an obligor goes into default. EAD is a measure of the monetary exposure should an obligor go into default. These topics will be explained in more detail in the next section.



With the arrival of Basel III and as a response to the latest financial crisis, the objective to strengthen global capital standards has been reinstated. A key focus here is the reduction in reliance on external ratings by the financial institutions, as well as a greater focus on stress testing. Although changes are inevitable, a key point worth noting is that with Basel III there is no major impact on underlying credit risk models. Hence the significance in creating these robust risk models continues to be of paramount importance.

In this book, we use theory and practical applications to show how these underlying credit risk models can be constructed and implemented through the use of SAS (in particular, SAS Enterprise Miner and SAS/STAT). To achieve this, we present a comprehensive guide to the classification and regression techniques needed to develop models for the prediction of all three components of expected loss: PD, LGD and EAD. The reason why these particular topics have been chosen is due in part to the increased scrutiny on the financial sector and the pressure placed on them by the financial regulators to move to the advanced internal ratings-based approach. The financial sector is therefore looking for the best possible models to determine their minimum capital requirements through the estimation of PD, LGD and EAD.

This introduction chapter is structured as follows. In the next section, we give an overview of the current regulatory environment, with emphasis on its implications to credit risk modeling. In this section, we explain the three key components of the minimum capital requirements: PD, LGD and EAD. Finally, we discuss the SAS software used in this book to support the practical applications of the concepts covered.

---

### 1.3 Regulatory Environment

The banking/financial sector is one of the most closely scrutinized and regulated industries and, as such, is subject to stringent controls. The reason for this is that banks can only lend out money in the form of loans if depositors trust that the bank and the banking system is stable enough and their money will be there when they require to withdraw it. However, in order for the banking sector to provide personal loans, credit cards, and mortgages, they must leverage depositors' savings, meaning that only with this trust can they continue to function. It is imperative, therefore, to prevent a loss of confidence and distrust in the banking sector from occurring, as it can have serious implications to the wider economy as a whole.

The job of the regulatory bodies is to contribute to ensuring the necessary trust and stability by limiting the level of risk that banks are allowed to take. In order for this to work effectively, the maximum risk level banks can take needs to be set in relation to the bank's own capital. From the bank's perspective, the high cost of acquiring and holding capital makes it prohibitive and unfeasible to have it fully cover all of a bank's risks. As a compromise, the major regulatory body of the banking industry, the Basel Committee on Banking Supervision, proposed guidelines in 1988 whereby a solvability coefficient of eight percent was introduced. In other words, the total assets, weighted for their risk, must not exceed eight percent of the bank's own capital (SAS Institute, 2002).

The figure of eight percent assigned by the Basel Committee was somewhat arbitrary, and as such, this has been subject to much debate since the conception of the idea. After the introduction of the Basel I Accord, more than one hundred countries worldwide adopted the guidelines, marking a major milestone in the history of global banking regulation. However, a number of the accord's inadequacies, in particular with regard to the way that credit risk was measured, became apparent over time (SAS Institute, 2002). To account for these issues, a revised accord, Basel II, was conceived. The aim of the Basel II Capital Accord was to further strengthen the financial sector through a three pillar approach. The following sections detail the current state of the regulatory environment and the constraints put upon financial institutions.

### 1.3.1 Minimum Capital Requirements

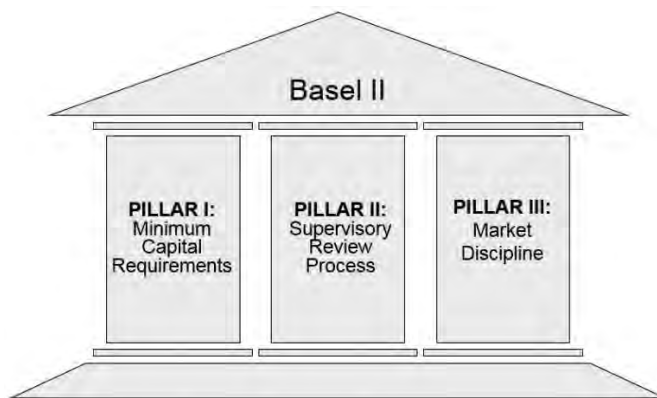
The Basel Capital Accord (Basel Committee on Banking Supervision, 2001a) prescribes the minimum amount of regulatory capital an institution must hold so as to provide a safety cushion against unexpected losses. The Accord is comprised of three pillars, as illustrated by Figure 1.1:

Pillar 1: Minimum Capital Requirements

Pillar 2: Supervisory Review Process

Pillar 3: Market Discipline (and Public Disclosure)

**Figure 1.1: Pillars of the Basel Capital Accord**



Pillar 1 aligns the minimum capital requirements to a bank's actual risk of economic loss. Various approaches to calculating this are prescribed in the Accord (including more risk-sensitive standardized and internal ratings-based approaches) which will be described in more detail and are of the main focus of this text. Pillar 2 refers to supervisors evaluating the activities and risk profiles of banks to determine whether they should hold higher levels of capital than those prescribed by Pillar 1, and offers guidelines for the supervisory review process, including the approval of internal rating systems. Pillar 3 leverages the ability of market discipline to motivate prudent management by enhancing the degree of transparency in banks' public disclosure (Basel, 2004).

Pillar 1 of the Basel II Capital Accord entitles banks to compute their credit risk capital in either of two ways:

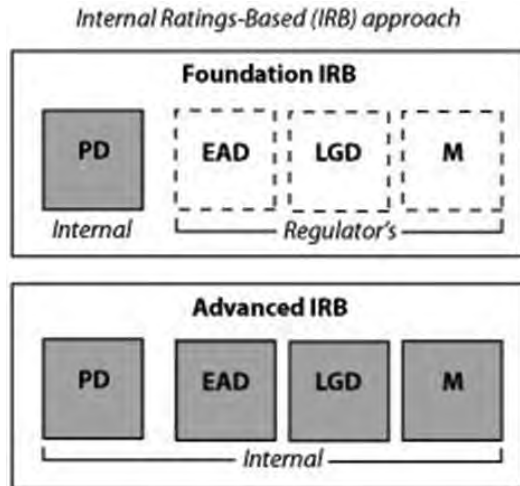
1. Standardized Approach
2. Internal Ratings-Based (IRB) Approach
  - a. Foundation Approach
  - b. Advanced Approach

Under the standardized approach, banks are required to use ratings from external credit rating agencies to quantify required capital. The main purpose and strategy of the Basel committee is to offer capital incentives to banks that move from a supervisory approach to a best-practice advanced internal ratings-based approach. The two versions of the internal ratings-based (IRB) approach permit banks to develop and use their own internal risk ratings, to varying degrees. The IRB approach is based on the following four key parameters:

1. Probability of Default (PD): the likelihood that a loan will not be repaid and will therefore fall into default in the next 12 months;
2. Loss Given Default (LGD): the estimated economic loss, expressed as a percentage of exposure, which will be incurred if an obligor goes into default - in other words, LGD equals: 1 minus the recovery rate;
3. Exposure At Default (EAD): a measure of the monetary exposure should an obligor go into default;
4. Maturity (M): is the length of time to the final payment date of a loan or other financial instrument.

The internal ratings-based approach requires financial institutions to estimate values for PD, LGD, and EAD for their various portfolios. Two IRB options are available to financial institutions: a foundation approach and an advanced approach (Figure 1.2) (Basel Committee on Banking Supervision, 2001a).

**Figure 1.2: Illustration of Foundation and Advanced Internal Ratings-Based (IRB) approach**



The difference between these two approaches is the degree to which the four parameters can be measured internally. For the foundation approach, only PD may be calculated internally, subject to supervisory review (Pillar 2). The values for LGD and EAD are fixed and based on supervisory values. For the final parameter, M, a single average maturity of 2.5 years is assumed for the portfolio. In the advanced IRB approach, all four parameters are to be calculated by the bank and are subject to supervisory review (Schuermann, 2004).

Under the A-IRB, financial institutions are also recommended to estimate a "Downturn LGD", which 'cannot be less than the long-run default-weighted average LGD calculated based on the average economic loss of all observed defaults with the data source for that type of facility' (Basel, 2004).

### 1.3.2 Expected Loss

Financial institutions expect a certain number of the loans they make to go into default; however they cannot identify in advance which loans will default. To account for this risk, a value for expected loss is priced into the products they offer. Expected Loss (EL) can be defined as the expected means loss over a 12 month period from which a basic premium rate is formulated. Regulatory controllers assume organizations will cover EL through loan loss provisions. Consumers experience this provisioning of expected loss in the form of the interest rates organizations charge on their loan products.

To calculate this value, the PD of an entity is multiplied by the estimated LGD and the current exposure if the entity were to go into default.

From the parameters, PD, LGD and EAD, expected loss (EL) can be derived as follows:

$$EL = PD \times LGD \times EAD \quad (1.1)$$

For example, if PD = 2%, LGD = 40% and EAD = \$10,000, then EL would equal \$80. Expected Loss can also be measured as a percentage of EAD:

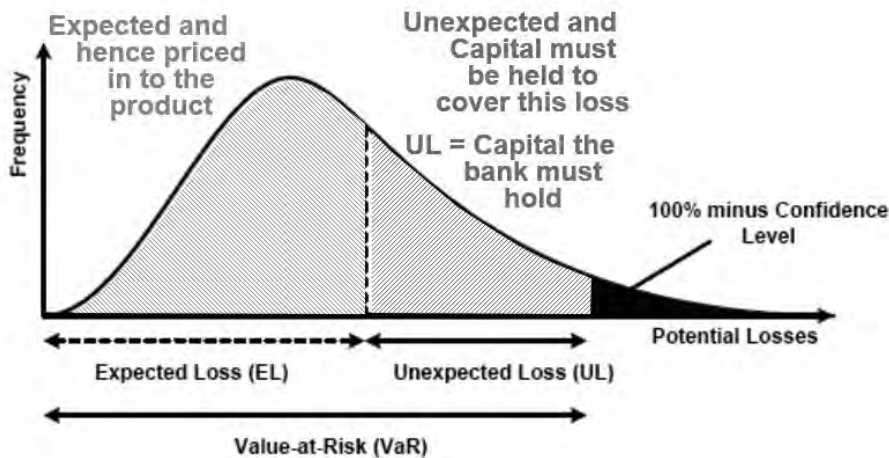
$$EL\% = PD \times LGD \quad (1.2)$$

In the previous example, expected loss as a percentage of EAD would be equal to  $EL\% = 0.8\%$ .

### 1.3.3 Unexpected Loss

Unexpected loss is defined as any loss on a financial product that was not expected by a financial organization and therefore not factored into the price of the product. The purpose of the Basel regulations is to force banks to retain capital to cover the entire amount of the Value-at-Risk (VaR), which is a combination of this unexpected loss plus the expected loss. Figure 1.3 highlights the Unexpected Loss, where UL is the difference between the Expected Loss and a 1 in 1000 chance level of loss.

**Figure 1.3: Illustration of the Difference between Expected/Unexpected Loss and a 1 in 1000 Chance Level of Loss**



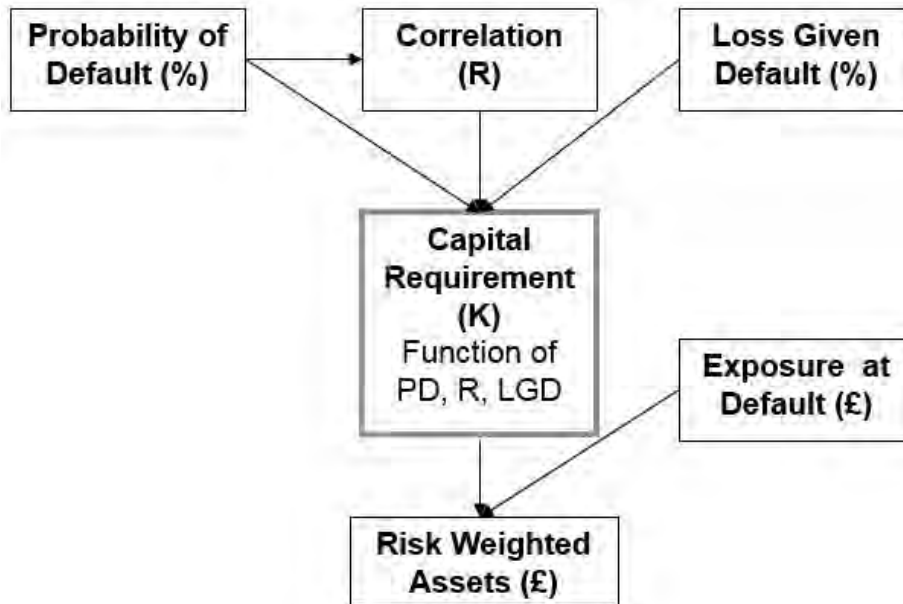
### 1.3.4 Risk Weighted Assets

Risk Weighted Assets (RWA) are the assets of the bank (money lent out to customers and businesses in the form of loans) accounted for by their riskiness. The RWA are a function of PD, LGD, EAD and M, where K is the capital requirement:

$$RWA = (12.5) \times K \times EAD \quad (1.3)$$

Under the Basel capital regulations, all banks must declare their RWA, hence the importance in estimating the three components, PD, LGD, and EAD, which go towards the formulation of RWA. The multiplication of the capital requirement (K) by 12.5  $\left(\frac{1}{12.5} = 0.08\right)$  is to ensure capital is no less than 8% of RWA. Figure 1.4 is a graphical representation of RWA and shows how each component feeds into the final RWA value.

Figure 1.4: Relationship between PD, LGD, EAD and RWA



The Capital Requirement (K) is defined as a function of PD, a correlation factor (R) and LGD

$$K = LGD \times \left( \phi \left( \sqrt{\frac{1}{1-R}} \phi^{-1}(PD) + \sqrt{\frac{R}{1-R}} \phi^{-1}(0.999) \right) - PD \right) \quad (1.4)$$

where  $\phi$  denotes the normal cumulative distribution function and  $\phi^{-1}$  denotes the inverse cumulative distribution function. The correlation factor (R) is determined based on the portfolio being assessed. For example, for revolving retail exposures (credit cards) not in default, the correlation factor is set to 4%. A full derivation of the capital requirement can be found in Basel Committee on Banking Supervision (2004).

In practice, how do estimations of PD, LGD and EAD impact the overall capital requirements? If we take PD as 0.03, LGD as 0.5, and EAD as \$10,000, then  $K(0.03, 0.5) \times (10000) = \$34.37$ . If an overestimate of 10% was made on PD, then the resulting capital required would then be  $K(0.033, 0.5) \times (10000) = \$36.73$ , requiring an increase of 6.9% in capital (\$2.36). However if an overestimate of 10% was made on LGD, then the resulting capital required would be  $K(0.03, 0.55) \times (10000) = \$37.80$ , requiring an increase of 10% in capital (\$3.43).

Because LGD and EAD enter the Risk Weight Function in a linear way, it is of crucial importance to have models that estimate LGD and EAD as accurately as possible, as LGD and EAD errors are more expensive than PD errors.

## 1.4 SAS Software Utilized

Throughout this book, examples and screenshots aid in the understanding and practical implementation of model development. The key tools used to achieve this are Base SAS programming with SAS/STAT procedures, as well as the point-and-click interfaces of SAS Enterprise Guide and SAS Enterprise Miner. For model report generation and performance monitoring, examples are drawn from SAS Model Manager. Base SAS is a comprehensive programming language used throughout multiple industries to manage and model data. SAS Enterprise Guide (Figure 1.5) is a powerful Microsoft Windows client application that provides a guided

mechanism to exploit the power of SAS and publish dynamic results throughout the organization through a point-and-click interface. SAS Enterprise Miner (Figure 1.6) is a powerful data mining tool for applying advanced modeling techniques to large volumes of data in order to achieve a greater understanding of the underlying data. SAS Model Manager (Figure 1.7) is a tool which encompasses the steps of creating, managing, deploying, monitoring, and operationalizing analytic models, ensuring the best model at the right time is in production.

Typically analysts utilize a variety of tools in their development and refinement of model building and data visualization. Through a step-by-step approach, we can identify which tool from the SAS toolbox is best suited for each task a modeler will encounter.

**Figure 1.5: Enterprise Guide Interface**

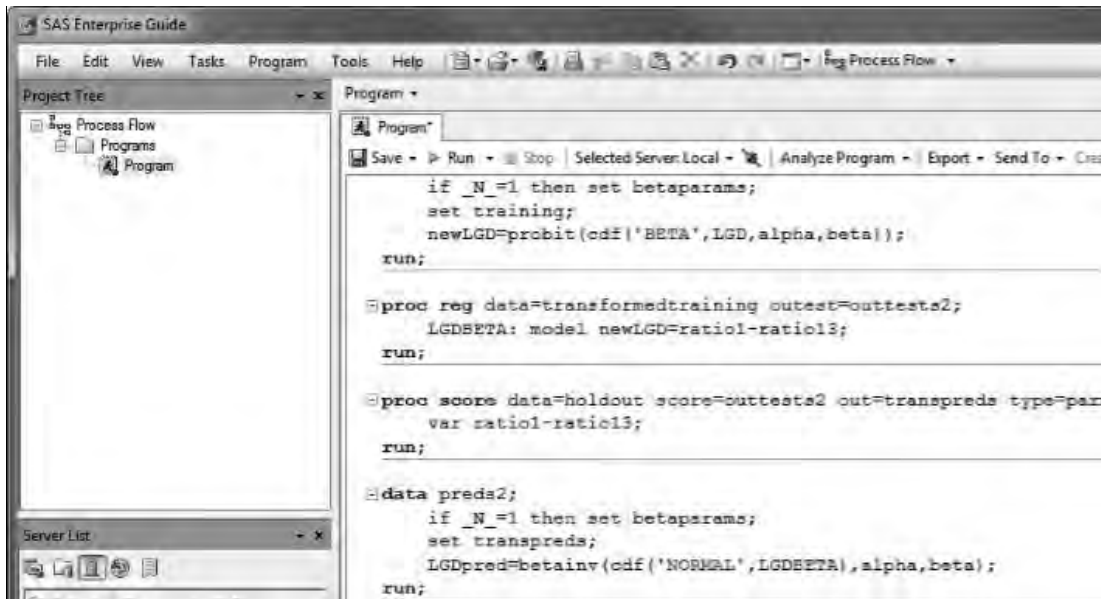
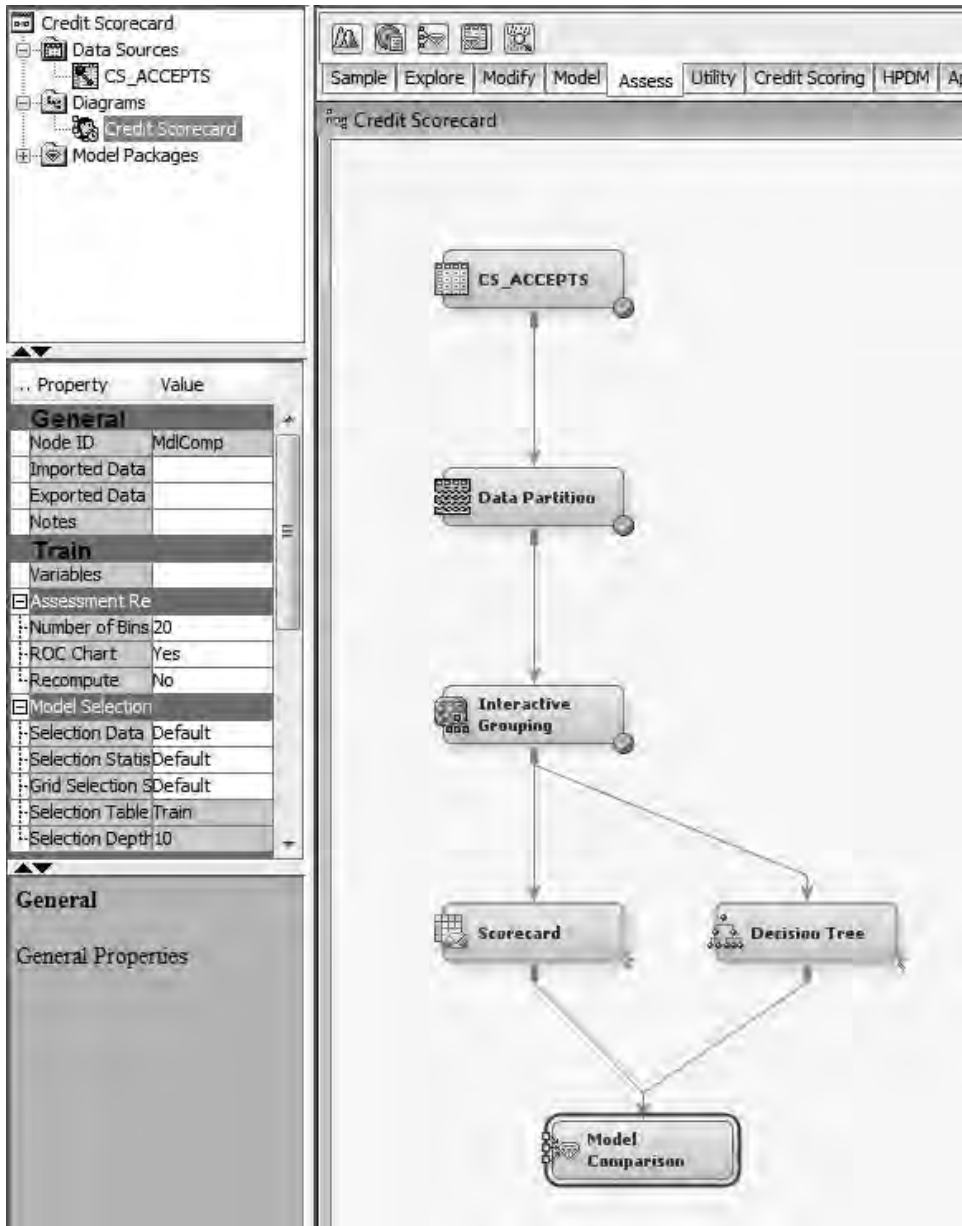
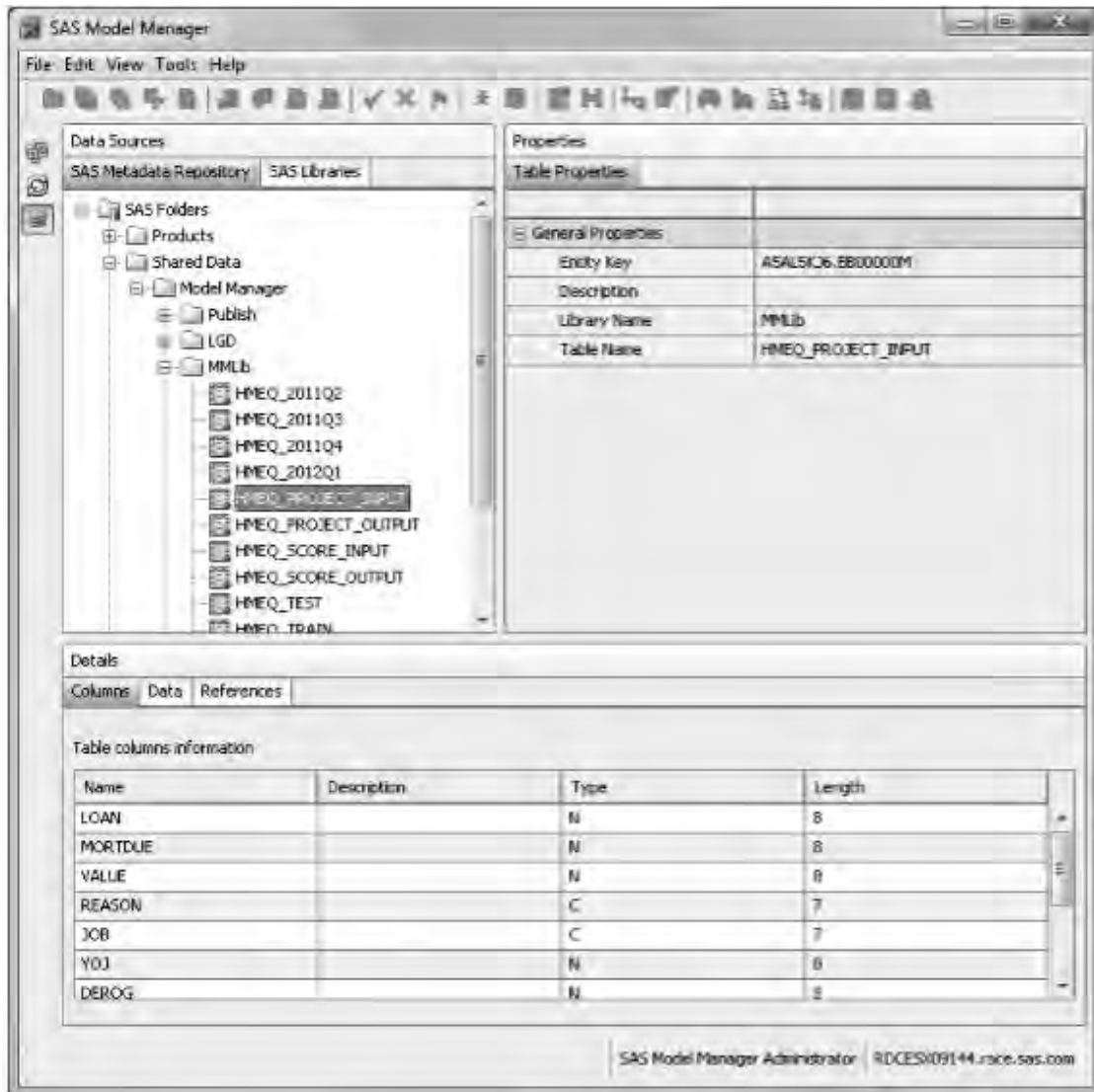


Figure 1.6: Enterprise Miner Interface



**Figure 1.7: Model Manager Interface**





## 1.5 Chapter Summary

This introductory chapter explores the key concepts that comprise credit risk modeling, and how this impacts financial institutions in the form of the regulatory environment. We have also looked at how regulations have evolved over time to better account for global risks and to fundamentally prevent financial institutions from over exposing themselves to difficult market factors. To summarize, Basel defines how financial institutions calculate:

- **Expected Loss (EL)** - the means loss over 12 months
- **Unexpected Loss (UL)** - the difference between the Expected Loss and a 1 in 1000 chance level of loss
- **Risk-Weighted Assets (RWA)** - the assets of the financial institution (money lent out to customers & businesses) accounted for by their riskiness
- How much **Capital** financial institutions hold to cover these losses

Three key parameters underpin the calculation of expected loss and risk weighted assets:

- **Probability of Default (PD)** - the likelihood that a loan will not be repaid and will therefore fall into default in the next 12 months
- **Loss Given Default (LGD)** - the estimated economic loss, expressed as a percentage of exposure, which will be incurred if an obligor goes into default - in other words, LGD equals: 1 minus the recovery rate
- **Exposure At Default (EAD)** - a measure of the monetary exposure should an obligor go into default

The purpose of these regulatory requirements is to strengthen the stability of the banking system by ensuring adequate provisions for loss are made.

We have also outlined the SAS technology which will be used through a step-by-step approach to apply the theoretical information given into practical examples.

In order for financial institutions to estimate these three key parameters that underpin the calculation of EL and RWA, they must begin by utilizing the correct data. Chapter 2 covers the area of sampling and data pre-processing. In this chapter, issues such as variable selection, missing values, and outlier detection are defined and contextualized within the area of credit risk modeling. Practical applications of how these issues can be solved are also given.

## 1.6 References and Further Reading

Basel Committee on Banking Supervision. 2001a. *The New Basel Capital Accord*. Jan. Available at: <http://www.bis.org/publ/bcbsca03.pdf>.

Basel Committee on Banking Supervision. 2004. *International Convergence of Capital Measurement and Capital Standards: a Revised Framework*. Bank for International Settlements.

SAS Institute. 2002. "Comply and Exceed: Credit Risk Management for Basel II and Beyond." A SAS White Paper.

Schuermann T. 2004. "What do we know about loss given default?" Working Paper No. 04-01, Wharton Financial Institutions Center, Feb.

From *Developing Credit Risk Models Using SAS® Enterprise Miner™ and SAS/STAT®: Theory and Application*, by Iain Brown. Copyright © 2014, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Developing Credit Risk Models Using SAS® Enterprise Miner™ and SAS/STAT®*. Full book available for purchase [here](#).

# Index

## A

Accuracy performance measure 117  
Accuracy Ratio (AR) performance measure 54, 117  
Accuracy Ratio Trend, graphically representing in SAS Enterprise Guide 121–122  
advanced internal ratings-based approach (A-IRB) 2  
Analytical Base Table (ABT) format 50  
application scorecards  
  about 35  
  creating 144  
  data partitioning for 40  
  data preparation for 37–38  
  data sampling for 39–40  
  developing models in SAS Enterprise Miner 139–145  
  developing PD model for 36–47  
  filtering for 40  
  input variables for 37–38  
  for Known Good Bad Data (KGB) 39  
  model creation process flow for 38  
  model validation for 46–47  
  modeling for 41–45  
  motivation for 36–37  
  outlier detection for 40  
  reject inference for 45–46  
  scaling for 41–45  
  strength of 54  
  transforming input variables for 40–41  
  variable classing and selection for 41  
application scoring 16  
Area Over the Curve (AOC) 71  
Area Over the Regression Error Characteristic (REC) Curves 71–72  
Area Under Curve (AUC) 54, 70–72, 117  
ARIMA procedure 113  
Artificial Neural Networks (ANN) 63, 67, 79  
assigning library locations 134–136  
augmented data sets  
  creating 144–145  
  grouping 145  
  partitioning into training, test and validation 145  
  scoring 145  
augmented good bad (AGB) data set 46  
AUOTREG procedure 113

## B

Basel Committee on Banking Supervision 4, 8  
Basel II Capital Accord 2, 4  
Basel III 3  
Bayesian Error Rate (BER), as performance measure 117  
behavioral scoring  
  about 17, 47

  data preparation for 49–50  
  developing PD model for 49–52  
  input variables for 49  
  model creation process flow for 50–52  
  motivation for 48  
benchmarking algorithms for LGD 77–82  
Beta Regression (BR) 63, 65–67  
beta transformation, linear regression nodes combined with 65  
Binary Logit models 98–99  
binary variables 15  
Binomial Test 125  
"black-box" techniques 44  
Box-Cox transformation, linear regression nodes combined with 63  
Brier Skill Score (BSS) 125

## C

calibration, of Probability of Default (PD) models 29  
capital requirement (K) 6  
Captured Event Plot 54  
case study: benchmarking algorithms for LGD 77–82  
classification techniques, for Probability of Default (PD) models 29–35  
Cluster node (SAS Enterprise Miner) 24–25  
Cohort Approach 89  
Confidence Interval (CI) 125  
corporate credit  
  Loss Given Default (LGD) models for 60–61  
  Probability of Default (PD) models for 28  
Correlation Analysis 125  
correlation factor (R) 7  
correlation scenario analysis 112  
creating  
  application scorecards 144  
  augmented data sets 144–145  
  Fit Logistic Regression Model 145–146  
  Loss Given Default (LGD) reports 129–130  
  Probability of Default (PD) reports 127–129  
  rejected data source 144  
creation process flow  
  application scorecards 39  
  for behavioral scoring 50–52  
  for Loss Given Default (LGD) 74–75  
credit conversion factor (CCF)  
  about 92  
  distribution 93–94  
  time horizons for 88–90  
credit risk modeling 2–3  
Cumulative Logit models 30, 98–99  
cumulative probability 30

**D**

- D Statistic, as performance measure 117
- data
  - Loss Given Default (LGD) 75
  - partitioning 40, 143
  - preparation for behavioral scoring 49–50
  - preparation for Exposure at Default (EAD) model 90–95
  - preparation of application scorecards 37–38
  - pre-processing 13–18
  - used in this book 147–150
  - visualizing 141–143
- Data Partition node (SAS Enterprise Miner) 18, 40, 45, 75, 96, 143
- data pooling phase 37
- data sampling
  - See* sampling
- data segmentation
  - about 22–23
  - decision trees 23–24, 28, 33–34
  - K-Means clustering 24–25
- data sets
  - See also* augmented data sets
  - characteristics for Loss Given Default (LGD) case study 77–78
  - defining 136–138
- data sources, defining 140
- data values 14
- Decision Tree node (SAS Enterprise Miner) 33
- decision trees 23–24, 28, 33–34
- defining
  - data sets 136–138
  - data sources 140
- discrete variables 14, 22
- discrim procedure 31–32
- discussion, for LGD case study 79–82

**E**

- economic variables, for LGD models 61
- End Group Processing node (SAS Enterprise Miner) 46–47
- Enterprise Miner Data Source Wizard 15–16
- Error Rate, as performance measure 117
- estimating downturn LGD 61–62
- examples (SAS Model Manager) 127–130
- Expected Loss (EL) 5–6, 11
- experimental set-up, for LGD case study 78–79
- expert judgment scenario analysis 112
- Exposure at Default (EAD)
  - about 2-3, 4, 11, 87–91
  - CCF distribution - transformations 94–96
  - data preparation 90–95
  - data used in this book 149
  - model development 97–103
  - model methodology 90–95
  - model performance measures 105–106

- model validation 103–106
  - performance metrics 99–103
  - reporting 103–106
  - time horizons for CCF 88–90
- extreme outliers 14

**F**

- Filter node (SAS Enterprise Miner) 21, 40, 95
- filtering
  - for application scorecards 40
  - methods for 21, 40
- Fit Logistic Regression model, creating 144
- Fit Statistics window 54
- fitting logistic regression model 145
- Fixed-Horizon Approach 90
- Friedman test 78
- FSA Stress Testing Thematic review (website) 113
- Fuzzy Augmentation 45
- fuzzy reject inference 145

**G**

- "garbage in, garbage out" 14
- Gini Statistic 52–54, 71
- gradient boosting, for Probability of Default (PD) models 35
- Gradient Boosting node (SAS Enterprise Miner) 35
- graphical Key performance indicator (KPI) charts 123
- grouping
  - augmented data set 145
  - performing with interactive grouping 145

**H**

- Hard Cutoff Method 45, 145
- historical scenarios 112
- Hosmer-Lemeshow Test (p-value) 125
- HP Forest node (SAS Enterprise Miner) 34
- hypothetical scenarios 112

**I**

- importing XML diagrams 140
- Impute node (SAS Enterprise Miner) 20–21
- Information Statistic (I), as performance measure 117
- information value (IV) 52–54
- input variables
  - application scorecards 37, 40–41
  - behavioral scoring 49
- Interactive Grouping node (SAS Enterprise Miner) 33, 41, 46, 53, 93, 143, 145
- interval variables 14, 21

**K**

- Kendall's Correlation Coefficient 73
- Kendall's Tau-b, as performance measure 117
- K-Means clustering 24–25
- Known Good Bad (KGB) data
  - about 23, 139

- application scorecards 39
  - sample 37
  - used in this book 147–148
- Kolmogorov-Smirnov Plot 42–43, 54, 117
- K-S Statistic 54
- Kullback-Leibler Statistic (KL), as performance measure 117

**L**

- Least Square Support Vector Machines 28
- library locations, assigning 134–136
- lift charts 105
- linear discriminant analysis (LDA), for Probability of Default (PD) 31–32
- linear probability models 28
- linear regression
  - non-linear regression and 63, 68–69
  - Ordinary Least Squares (OLS) and 63
  - techniques for 63
- linear regression nodes
  - combined with beta transformation 64
  - combined with Box-Cox transformation 66
- Loan Equivalency Factor (LEQ) 87
- logistic procedure 41, 113
- logistic regression
  - fitting 145
  - non-linear regression and 68–69
  - for Probability of Default (PD) 29–30
- Logistic Regression node 75–76
- logit models 28
- Log+(non-) linear regression techniques 63
- loss, predicting amount of 76
- Loss Given Default (LGD)
  - about 2–3, 4, 11, 59
  - benchmarking algorithms for 77–82
  - case study: benchmarking algorithms for LGD 77–82
  - for corporate credit 60–61
  - creating reports 129–130
  - creation process flow for 74–75
  - data 75
  - data used in this book 148
  - economic variables for 61
  - estimating downturn 61–62
  - model development 73–77
  - models for retail credit 60
  - motivation for 73
  - performance metrics for 69–73
  - regression techniques for 62–69

**M**

- macroeconomic approaches, stress testing using 113
- market downturn, as a hypothetical scenario 112
- market position, as a hypothetical scenario 112
- market reputation, as a hypothetical scenario 112
- Maturity (M) 4
- Mean Absolute Deviation (MAD) 117, 125

- Mean Absolute Error (MAE) 60
- Mean Absolute Percent Error (MAPE) 117, 126
- Mean Square Error (MSE) 117, 126
- memory based reasoning, for Probability of Default (PD) models 34
- Metadata node 96
- minimum capital requirements 4–5
- missing values 16, 19–22
- model calibration 116, 125–126
- Model Comparison node 77, 103, 119
- model development
  - Exposure at Default (EAD) 97–103
  - Loss Given Default (LGD) 73–77
  - Probability of Default (PD) 36–47
  - in SAS Enterprise Miner 139–140
- model reports
  - producing 115–130
  - regulatory reports 115
  - SAS Model Manager examples 127–130
  - validation 115–127
- model stability 122–125
- model validation
  - about 77
  - application scorecards 46–47
  - Exposure at Default (EAD) 97–103
  - for reports 115–127
- modeling, for application scorecards 41–44
- models
  - deployment for Probability of Default (PD) 55–57
  - performance measures for 54, 116–122
  - registering package 56–57
  - tuning 54
- Multilayer Perceptron (MLP) 32
- multiple discriminant analysis models 28

**N**

- Nemenyi's post hoc test 62
- Neural Network node (SAS Enterprise Miner) 33
- Neural Networks (NN) 32
- nlmixed procedure 66
- nominal variables 14–15
- non-defaults, scoring 76
- non-linear regression
  - linear regression and 63, 68–69
  - logistic regression and 68–69
  - techniques for 63
- Normal Test 126

**O**

- Observed Versus Estimated Index 126
- 1-PH Statistic (1-PH), as performance measure 117
- ordinal variables 14
- Ordinary Least Squares (OLS)
  - about 63, 97–98
  - linear regression and 64
- Ordinary Least Squares + Neural Networks (OLS + ANN) 63

Ordinary Least Squares + Regression Trees (OLS + RT) 63  
 Ordinary Least Squares with Beta Transformation (B-OLS) 63, 64, 65  
 Ordinary Least Squares with Box-Cox Transformation (BC-OLS) 63, 66–67, 79  
 outlier detection 21–22, 40

**P**

parameters, setting and tuning for LGD case study 79  
 Parceling Method 45, 145  
 partitioning  
   augmented data set into training, test and validation 145  
   data 40, 143  
 Pearson's Correlation Coefficient 72, 99  
 performance measures  
   Exposure at Default (EAD) model 105–106  
   SAS Model Manager 117–118  
 performance metrics  
   Exposure at Default (EAD) 99–103  
   for Loss Given Default (LGD) 69–73  
 performing  
   reject inference 144–145  
   screening and grouping with interactive grouping 143–144  
   univariate characteristic screening 145  
 Pietra Index, as performance measure 118  
 Pillar 1/2/3 4  
 Precision, as performance measure 118  
 predicting amount of loss 76–77  
 pre-processing data 13–16  
 Probability of Default (PD)  
   about 2–3, 4, 11, 24  
   behavioral scoring 47–52  
   calibration 29  
   classification techniques for 29–35  
   creating reports 127–129  
   decision trees for 33–34  
   gradient boosting for 35  
   linear discriminant analysis (LDA) for 31–32  
   logistic regression for 29–30  
   memory based reasoning for 34  
   model deployment 55–57  
   model development 35–47  
   models for corporate credit 28  
   models for retail credit 28  
   Neural Networks (NN) for 32–33  
   quadratic discriminant analysis (QDA) for 31–32  
   random forests for 34–35  
   reporting 52–55  
 probit models 28  
 "pseudo residuals" 35

**Q**

quadratic discriminant analysis (QDA), for Probability of Default (PD) models 31–32

**R**

random forests, for Probability of Default (PD) models 34–35  
 reg procedure 64  
 registering model package 56–57  
 Regression node (SAS Enterprise Miner) 30, 41, 44, 64, 77, 113  
 regression techniques, for Loss Given Default (LGD) models 62–69  
 Regression Trees (RT) 63, 67, 79  
 regulatory environment  
   about 3–4  
   Expected Loss (EL) 5–6  
   minimum capital requirements 4–5  
   Risk Weighted Assets (RWA) 6–7  
   Unexpected Loss (UL) 6  
 regulatory reports 115  
 regulatory stress testing 113  
 reject inference  
   for application scorecards 45–46  
   performing 144–145  
 Reject Inference node 45, 144  
 rejected candidates data, used in this book 148  
 rejected data source, creating 144  
 reporting  
   Exposure at Default (EAD) 103–106  
   Probability of Default (PD) 52–54  
 results, for LGD case study 79–82  
 retail credit  
   Loss Given Default (LGD) models for 60  
   Probability of Default (PD) models for 28–29  
 Risk Weighted Assets (RWA) 6–7, 11  
 ROC Plot 54  
 Root Mean Squared Error (RMSE) 69–70, 99  
 root node 23–24  
 R-Square 72, 99

**S**

Sample node (SAS Enterprise Miner) 17, 39–40  
 sampling  
   about 13–16  
   for application scorecards 39–40  
   variable selection and 16–19  
 SAS  
   software 7–10  
   website 35  
 SAS Code node 32, 65, 66, 67, 75, 94, 95, 96  
 SAS Enterprise Guide  
   about 7  
   graphically representing Accuracy Ratio Trend in 121

- SAS Enterprise Miner
    - about 7
    - developing application scorecard models in 139–146
    - getting started with 131–138
    - starting 131–134
  - SAS Model Manager
    - about 7
    - documentation 127
    - examples 127–130
    - performance measures 117–118
    - website 116
  - scenario testing 112–113
  - Score node (SAS Enterprise Miner) 51, 55
  - Scorecard node (SAS Enterprise Miner) 41, 44, 46, 54
  - scoring
    - See also* behavioral scoring
    - augmented data set 145–146
    - non-defaults 76
  - screening, performing with interactive grouping 143–144
  - Segment Profile node (SAS Enterprise Miner) 24
  - segmentation
    - See* data segmentation
  - SEMMA (Sample, Explore, Modify, Model, and Assess tabs) methodology 38
  - sensitivity measurement 54, 118
  - sensitivity testing 111
  - simulation scenario analysis 112–113
  - software (SAS) 7–10
  - Somers' D (p-value), as performance measure 118
  - Spearman's Correlation Coefficient 72, 99
  - specificity measurement 69, 118
  - standard procedure 66
  - Start Group Processing node (SAS Enterprise Miner) 46–47
  - starting SAS Enterprise Miner 131–134
  - stress testing
    - about 109–110
    - methods of 111–113
    - purpose of 110
    - regulatory 113
    - using macroeconomic approaches 113
  - surveyselect procedure (SAS/STAT) 17
  - System Stability Index (SS) 122
- T**
- "through-the-door" population 45, 144
  - Traffic Lights Test 126
  - Transform Variables node (SAS Enterprise Miner) 33, 40–41, 46–47, 65, 76
  - transformations 40–41, 95–96
  - transreg procedure 67
  - tuning models 54
  - tutorials
    - developing application scorecard models in SAS Enterprise Miner 139–146
    - getting started with SAS Enterprise Miner 131–138
- U**
- Unexpected Loss (UL) 6, 11
  - univariate characteristic screening, performing 145
- V**
- validation
    - See* model validation
  - Validation Score, as performance measure 118
  - Value-at-Risk (VaR) 110
  - varclus procedure (SAS/STAT) 50
  - Variable Clustering node (SAS Enterprise Miner) 18, 50
  - Variable Selection node (SAS Enterprise Miner) 18
  - Variable Time Horizon Approach 90
  - variable worth statistics 52–53
  - variables
    - binary 14
    - discrete 14, 22
    - economic 60
    - interval 14, 22
    - nominal 14–15
    - ordinal 14
    - sampling 16–19
    - selecting 16–19
  - visualizing data 141–142
- W**
- websites
    - FSA Stress Testing Thematic review 113
    - SAS 35
    - SAS Model Manager 116, 127
  - Weight of Evidence (WOE) 33, 41
  - worst-case scenario analysis 112
- X**
- XML diagrams, importing 140

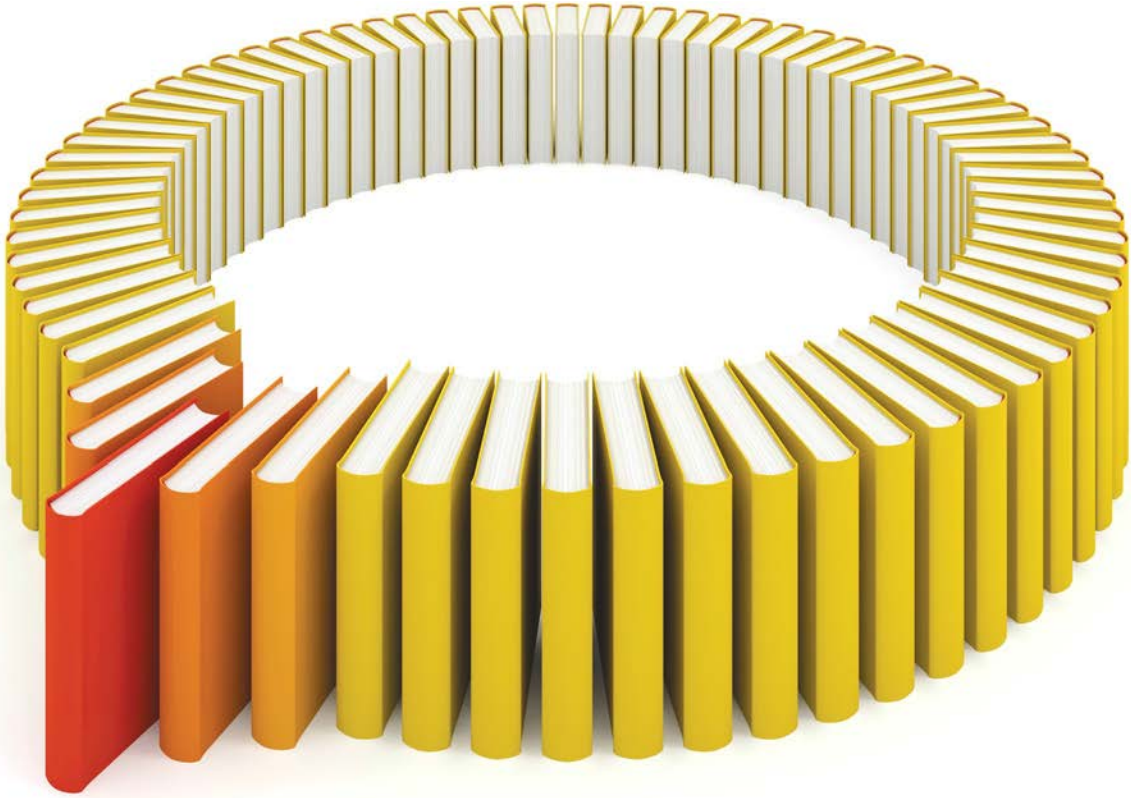
## About The Author



Dr. Iain Brown is an Analytics Specialist Consultant at SAS, specializing in Credit Risk. Prior to joining SAS in 2011, he worked as a Credit Risk Analyst at a major UK retail bank where he built and validated PD, LGD, and EAD models using SAS software. He has spoken at a number of internationally renowned conferences and conventions and has published papers on the topic of credit risk modeling in the International Journal of Forecasting and the Journal of Expert Systems with Applications. In 2011, he won the SAS Student Ambassador award for his doctoral research, which recognizes and supports students who use SAS technologies in innovative ways to benefit their respective industries and fields of study.

Iain has a BBA in Business from the University of Kent, an MSc in Operational Research from the London School of Economics and Political Science (LSE), and a PhD in Credit Risk from the University of Southampton. Iain is also an active member of the Operational Research (OR) Society; in July 2014, he was awarded the title of Associate Fellow of the OR Society (AFORS) for his contribution to the field of OR. His research interests include data mining, credit scoring, credit risk modeling, and Basel compliancy.

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



# Gain Greater Insight into Your SAS<sup>®</sup> Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

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

  
THE POWER TO KNOW.®