Chapter 8
Low Default Portfolios

INTRODUCTION

In this chapter, we discuss low default portfolios (LDPs). We begin by discussing the modeling problem and provide some regulatory perspectives. We then elaborate on the development of predictive models for skewed data sets and cover the following approaches: varying the time window, undersampling and oversampling, and synthetic minority oversampling technique (SMOTE). The next subsections feature discussions on how to adjust the posterior probabilities to the original class distribution and perform cost-sensitive learning. We then focus on the shadow ratings approach, where the aim is to develop an internal credit risk model predicting external agency ratings. The confidence level approach is reviewed next, followed by some other methods to deal with LDPs. The chapter concludes by providing some thoughts on how to model loss given default (LGD) and exposure at default (EAD) for LDPs.

BASIC CONCEPTS

The Basel Accord provides no formal definition of a low default portfolio. The Bank of England earlier suggested 20 as the minimum number of required defaults to begin modeling (Prudential Regulation Authority 2013). Hence, if you have fewer than 20 defaults, you definitely have a low default portfolio. The definition of a low default portfolio strongly depends not only on the quantity, but also on the quality of the data. More specifically, to what extent is the data predictive for the given (limited) number of defaults? Put differently, if you have high-quality and highly predictive data, you don't need that many defaults in order to derive a meaningful default risk model.

When thinking about low default portfolios, a distinction needs to be made between a low number of defaults in an absolute and in a relative sense. For example, when you have a portfolio of 100,000 observations with a default rate of 1 percent (i.e., 1,000 defaulters), then the relative default rate is low, but the absolute number of defaulters is quite high. When the portfolio has 100 observations and 10 defaulters, then the relative default rate is high, but the absolute number of defaulters is low. Obviously, there can also be situations where both the absolute number and the relative number of defaulters are low.

Low default portfolios are quite common in a financial setting. A popular example is exposures to sovereigns; very few countries have gone into default in the past. Other examples are exposures to banks, insurance companies, and project finance, which is finance for large projects such as building highways or nuclear reactors. Exposures to large corporations and/or specialized lending are additional examples. When you bring new products to the market, it will also take some time before you have the necessary number of defaults to estimate standard credit risk models.

As already mentioned, for low default portfolios, typically you have a lack of modeling data, especially default data, which makes it very difficult to apply the advanced internal ratings based (IRB) approach, in which case you need to estimate the probability of default (PD), the LGD, and the EAD. Historical average default rates are not appropriate since they have been calculated on only a few observations. Because of data scarcity, the credit risk can thus be substantially underestimated or overestimated. This is a significant problem, especially given the fact that a substantial portion of a bank's assets might consist of low default portfolios.

Here you can see some statements made by the Basel Committee Accord Implementation Group's Validation Subgroup on the issue of low default portfolios (Basel Committee on Banking Supervision 2005):

- "LDPs should not, by their very nature, automatically be excluded from IRB treatment."
- "… an additional set of rules or principles specifically applying to LDPs is neither necessary nor desirable."
- "… relatively sparse data might require increased reliance on alternative data sources and data-enhancing tools for quantification and alternative techniques for validation."
- "… LDPs should not be considered or treated as conceptually different from other portfolios."

The Financial Services Authority (FSA), which was the predecessor of the Prudential Regulation Authority (PRA) in the United Kingdom, earlier also explicitly confirmed that it should be possible to include a firm's LDPs in the IRB approach (see Financial Services Authority 2006a, Section 7).

It needs to be noted, though, that given their intrinsic characteristics, LDPs do require special modeling and calibration approaches. We discuss these further in what follows. First, we take a look at the case where we have a skewed data set with a low number of defaults in a relative sense, but a sufficient number of defaults in an absolute sense.

DEVELOPING PREDICTIVE MODELS FOR SKEWED DATA SETS

Default risk data sets often have a very skewed target class distribution where typically only about 1 percent or even less of the transactions are defaulters. Obviously, this creates problems for the analytical techniques discussed earlier since they are being flooded by all the nondefault observations and will thus tend toward classifying every observation as nondefault. Think about decision trees, for example: If they start from a data set with 99 percent/1 percent nondefault/default observations, then the entropy is already very low and hence it is very likely that the decision tree does not find any useful split and classifies all observations as nondefault, thereby achieving a classification accuracy of 99 percent, but essentially detecting none of the defaulters. It is thus recommended to increase the number of default observations or their weight, such that the analytical techniques can pay better attention to them. Various procedures are possible to do this and will be outlined in what follows.

Varying the Sample Window

A first way to increase the number of defaulters is by increasing the time horizon for prediction. For example, instead of predicting default with a 12-month forward-looking time horizon, an 18- or 24-month time horizon can be adopted. This is likely to add more defaulters to the sample and thus enable the analytical techniques to find a meaningful discrimination. An ex post recalibration step may then be added to obtain Basel-compliant 12-month PDs. An example application may be the determination of default rates for investment-grade rating classes, which are often zero or close to zero for one year but become positive in cumulative terms over the longer term (say 20 to 30 years).

Another approach works by sampling every defaulter twice (or more), as depicted in Exhibit 8.1. Let's assume we predict default with a one-year forward-looking time horizon using information from a one-year backward-looking time horizon (e.g., as in behavioral scoring). By shifting the observation point earlier or later, the same default observation can be sampled twice. Obviously, the variables collected will be similar but not perfectly the same, since they are measured on a different (although overlapping) time frame. This added variability can then come in handy for the analytical techniques to better discriminate between the defaulters and nondefaulters. Note that depending on the skewness of the target, multiple observation points can be considered such that the number of defaulters is multiplied by 2, 3, 4, and so on. Finding the optimal number is subject to a trial-and-error exercise. Note that a disadvantage of this method is that the characteristics of the duplicated observations may be overstated.

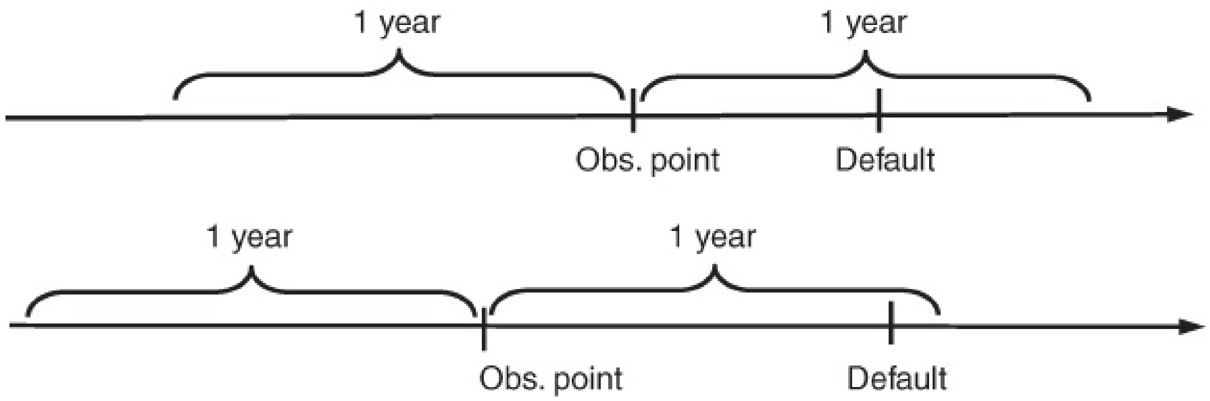


Exhibit 8.1 Varying the Time Window to Deal with Skewed Data Sets

Undersampling and Oversampling

Another way to increase the weight of the defaulters is by either oversampling them or by undersampling the nondefaulters. Oversampling is illustrated in Exhibit 8.2. Here, the idea is to replicate the defaulters two or more times so as to make the distribution less skewed. In our example, observations 1 and 4, both defaulters, have been replicated so as to create an equally balanced training sample having the same number of defaulters and nondefaulters.

Original data

	ID	Variables	Class
Train	1		Defaulter
	2		Non-Defaulter
	3		Non-Defaulter
	4		Defaulter
	5		Non-Defaulter
	6		Non-Defaulter
Test	7		Non-Defaulter
	8		Non-Defaulter
	9		Defaulter
	10		Non-Defaulter

Oversampled data

	ID	Variables	Class
Train	1		Defaulter
	1		**Defaulter**
	2		Non-Defaulter
	3		Non-Defaulter
	4		Defaulter
	4		**Defaulter**
	5		Non-Defaulter
	6		Non-Defaulter
Test	7		Non-Defaulter
	8		Non-Defaulter
	9		Defaulter
	10		Non-Defaulter

Exhibit 8.2 Oversampling the Defaulters

Undersampling is illustrated in Exhibit 8.3. Here, observations 2 and 5, which are both nondefaulters, have been removed so as to create an equally balanced training sample. The undersampling can be done based on business experience where obviously legitimate observations are removed. Also, low-value transactions or inactive accounts can be considered for removal.

Original data

	ID	Variables	Class
Train	1		Defaulter
	2		Non-Defaulter
	3		Non-Defaulter
	4		Defaulter
	5		Non-Defaulter
	6		Non-Defaulter
Test	7		Non-Defaulter
	8		Non-Defaulter
	9		Defaulter
	10		Non-Defaulter

Undersampled data

	ID	Variables	Class
Train	1		Defaulter
	3		Non-Defaulter
	4		Defaulter
	6		Non-Defaulter
Test	7		Non-Defaulter
	8		Non-Defaulter
	9		Defaulter
	10		Non-Defaulter

Exhibit 8.3 Undersampling the Nondefaulters

Under- and oversampling can also be combined. In the literature, it has been shown that undersampling usually results in better classifiers than oversampling (Chawla et al. 2002).

Note that both undersampling and oversampling should be conducted on the training data and not on the test data. Remember, the latter should remain untouched during model development in order to give an unbiased view on model performance. A practical question concerns the optimal nondefaulter/defaulter odds that should be aimed for by doing under- or oversampling. Although working toward a balanced sample with the same number of defaulters and nondefaulters seems attractive, it severely biases the probabilities that will be output by the analytical technique. Hence, it is recommended to stay as close as possible to the original class distribution to avoid unnecessary bias. One practical approach to determining the optimal class distribution works as follows. In the first step, an analytical model is built on the original data set with the skew class distribution (e.g., 95 percent/5 percent nondefaulters/defaulters). The area under the curve (AUC) of this model is recorded (possibly on an independent validation data set). In a next step, over- or undersampling is used to change the class distribution by 5 percent (e.g., 90 percent/10 percent). Again, the AUC of the model is recorded. Subsequent models are built on samples of 85 percent/15 percent, 80 percent/20 percent, 75 percent/25 percent, and so on, each time recording their AUCs. Once the AUC starts to stagnate (or drop), the procedure stops and the optimal odds ratio has been found. Although it does depend on the data characteristics and quality, practical experience has shown that the ratio 80 percent/20 percent is quite commonly used in the industry.

In SAS, PROC SURVEYSELECT can be used to compose a sample with the required target distribution as follows:

```
PROC SORT DATA=data.mortgage out=mortgage;
BY default_time;
RUN;
PROC SURVEYSELECT DATA=mortgage
METHOD=SRS N=(1000,1000) SEED=12345 OUT=data.mySample;
STRATA default_time;
RUN;
```

The preceding statement will create a balanced sample of 2,000 observations by randomly selecting 1,000 defaulters and 1,000 nondefaulters. We can verify this as follows:

```
PROC FREQ DATA=data.mySample;
TABLE default_time;
RUN;
```

The output will be as shown in Exhibit 8.4.

The FREQ Procedure				
default_time	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	1000	50.00	1000	50.00
1	1000	50.00	2000	100.00

Exhibit 8.4 Creating a balanced sample using PROC FREQ

In SAS Enterprise Miner, the Sample node can be used to specify the desired properties of the sample (see Exhibit 8.5). In the stratification section, set the criterion property to level based and in the level based options section set the level section to rarest level (which is default in our case) and the sample proportion to 50 percent. This will create a sample with 50 percent defaulters and 50 percent nondefaulters. Running the sample node will give the output as displayed in Exhibit 8.6.

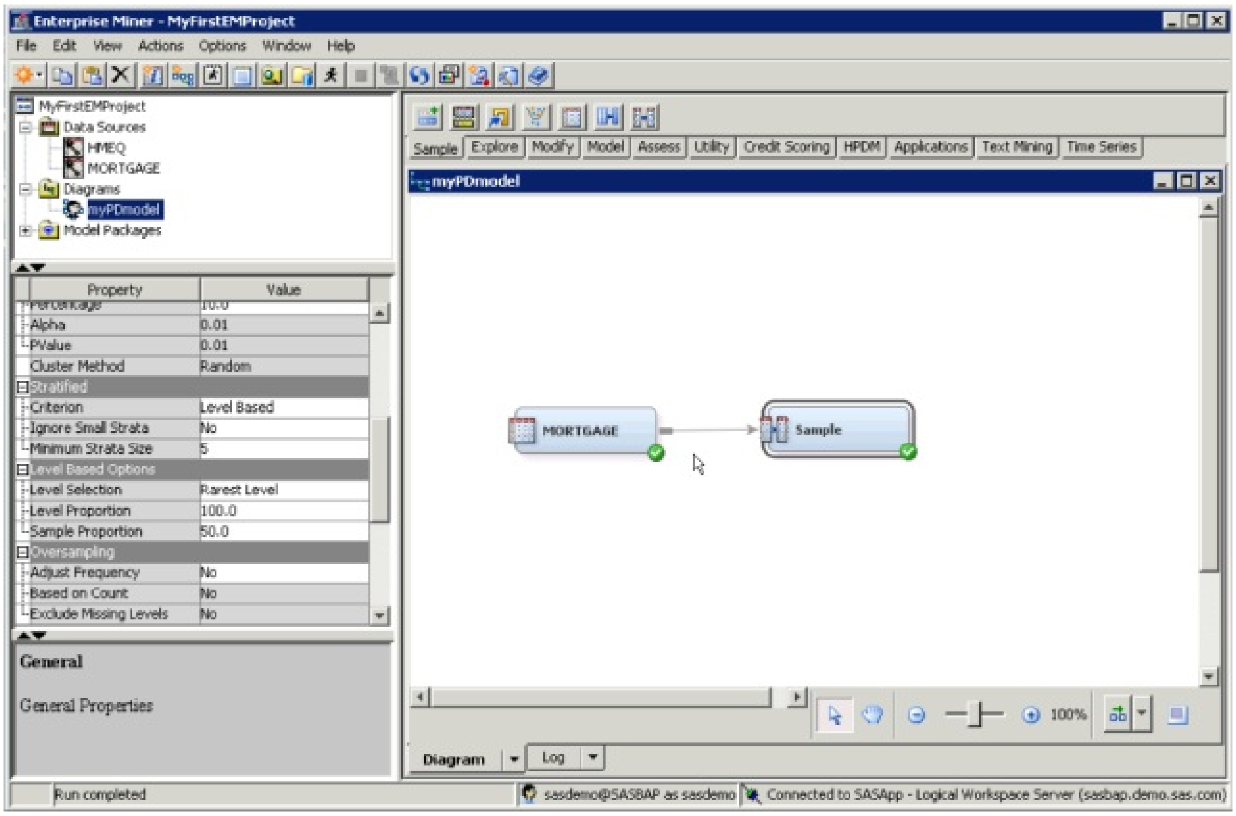
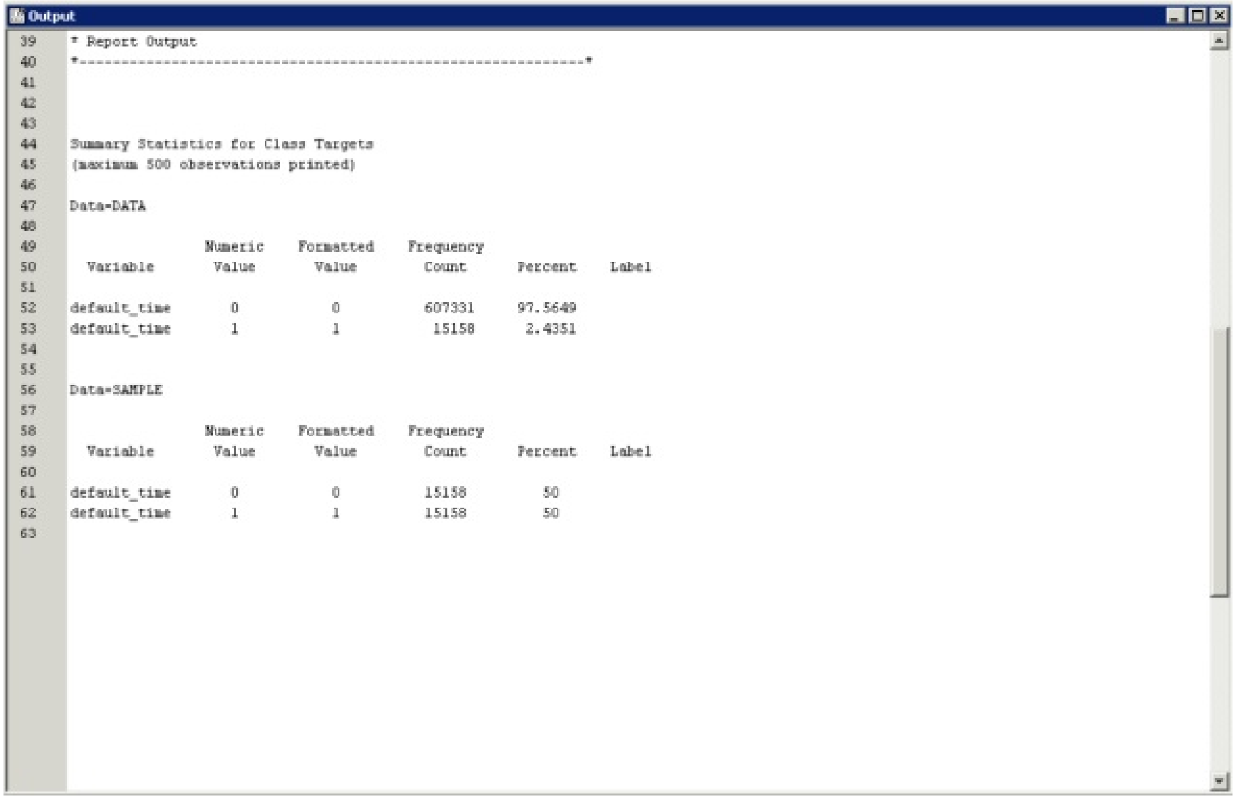


Exhibit 8.5 Creating a Tailored Sample in SAS Enterprise Miner



```
Output

* Report Output
*------------------------------------------------------------*



Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

                     Numeric    Formatted    Frequency
  Variable            Value       Value        Count       Percent    Label

default_time            0           0          607331      97.5649
default_time            1           1           15158       2.4351


Data=SAMPLE

                     Numeric    Formatted    Frequency
  Variable            Value       Value        Count       Percent    Label

default_time            0           0           15158        50
default_time            1           1           15158        50
```

Exhibit 8.6 Creating a Tailored Sample in SAS Enterprise Miner: Results

Synthetic Minority Oversampling Technique (SMOTE)

Rather than replicating the minority observations (i.e., defaulters), synthetic minority oversampling works by creating synthetic observations based on the existing minority observations (Chawla et al. 2002). This is illustrated in Exhibit 8.7, where the circles represent the majority class (e.g., nondefaulters) and the squares the minority class (e.g., defaulters). For each minority class observation, SMOTE calculates the *k* nearest neighbors. Let's assume we consider the crossed square and pick the five nearest neighbors represented by the black squares. Depending on the amount of oversampling needed, one or more of the *k* nearest neighbors are selected to create the synthetic examples. Let's say our oversampling percentage is set at 200 percent. In this case, two of the five nearest neighbors are selected at random. The next step is then to randomly create two synthetic examples along the line connecting the observation under investigation (crossed square) with the two random nearest neighbors. These two synthetic examples are represented by dashed squares in Exhibit 8.7. As an example, consider an observation with characteristics (e.g., age and income) of 30 and 1,000, and its nearest neighbor with corresponding characteristics 62 and 3,200. We generate a random number between zero and one; let's say 0.75. The synthetic example then has age $30 + 0.75 * (62 - 30)$ or 54, and income $1{,}000 + 0.75 * (3{,}200 - 1{,}000) = 2{,}650$. SMOTE then combines the synthetic oversampling of the minority class with undersampling the majority class. Note that in their original paper, Chawla et al. (2002) also developed an extension of SMOTE to work with categorical variables. Empirical evidence has shown that SMOTE usually works better than either under- or oversampling. It has also proven to be very valuable for fraud detection (Van Vlasselaer et al. 2016).

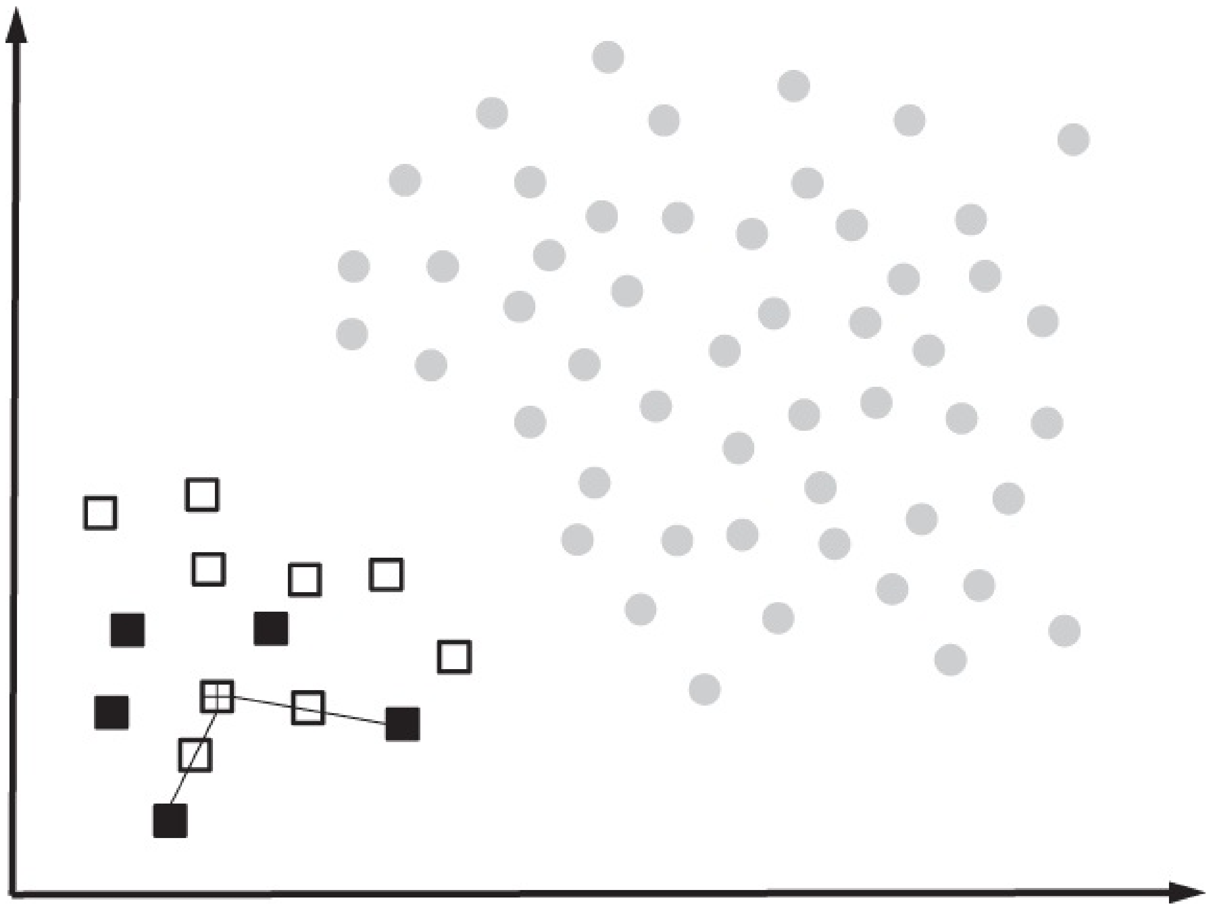


Exhibit 8.7 Synthetic Minority Oversampling Technique (SMOTE)

Adjusting Posterior Probabilities

The key idea of undersampling, oversampling, and SMOTE is to adjust the class priors to enable the analytical technique to come up with a meaningful model discriminating the defaulters from the nondefaulters. By doing so, the class posteriors will also become biased. This is not a problem in the case where the credit analyst is interested only in ranking the observations in terms of their default risk. However, if well-calibrated probabilities of default are needed (e.g., for Basel II/III), then the posterior probabilities need to be adjusted. One straightforward way to do this is by using the following formula (Saerens, Latinne, and Decaestecker 2002):

$$p(C_i|x) = \frac{\frac{p(C_i)}{p_r(C_i)} p_r(C_i|x)}{\sum_{j=1}^{2} \frac{p(C_j)}{p_r(C_j)} p_r(C_j|x)}$$

where C_i represents class *i* (e.g., class 1 for the defaulters and class 2 for the nondefaulters); $p(C_i)$ the prior probability (e.g., $p(C_1) = 1\%$ and $p(C_2) = 99\%$); $p_r(C_i)$ the resampled prior probability due to oversampling, undersampling, or other resampling procedures (e.g., $p_r(C_1) = 20\%$ and $p_r(C_2) = 80\%$); and $p_r(C_i|x)$ represents the posterior probability for observation *x* as calculated by the analytical technique using the resampled data. Note that the formula can be easily extended to more than two classes (e.g., in case of predicting ratings).

Exhibit 8.8 shows an example of adjusting the posterior probability, where $p(C_1) = 0.01, p(C_2) = 0.99, p_r(C_1) = 0.20$, and $p_r(C_2) = 0.80$. It can be easily verified that the rank ordering of the customers in terms of their default risk remains preserved after the adjustment.

	Posteriors Using Resampled Data		**Posteriors Recalibrated to Original Data**	
	***P*(Default)**	***P*(Nondefault)**	***P*(Default)**	***P*(Nondefault)**
Customer 1	0.1	0.9	0.004	0.996
Customer 2	0.3	0.7	0.017	0.983
Customer 3	0.5	0.5	0.039	0.961
Customer 4	0.6	0.4	0.057	0.943
Customer 5	0.85	0.15	0.186	0.814
Customer 6	0.9	0.1	0.267	0.733

Exhibit 8.8 Adjusting the Posterior Probability

The preceding calculation can be implemented in Base SAS as follows:

```
DATA posteriors;
INPUT probdef;
DATALINES;
0.1
0.3
0.5
0.6
0.85
0.9
;
DATA posteriors2;
SET posteriors;
probnondef=1-probdef;
oldprior=0.20;
newprior=0.01;
temp1=probdef/oldprior*newprior;
temp2=(1-probdef)/(1-oldprior)*(1-newprior);
newprobdef=temp1/(temp1+temp2);
newprobnondef=temp2/(temp1+temp2);
DROP temp1 temp2 oldprior newprior;
RUN;
/*Print data set*/
Proc PRINT DATA= posteriors2;
RUN;
```

The output will be as shown in Exhibit 8.9.

Obs	probdef	probnondef	newprobdef	newprobnondef
1	0.10	0.90	0.00447	0.99553
2	0.30	0.70	0.01702	0.98298
3	0.50	0.50	0.03883	0.96117
4	0.60	0.40	0.05714	0.94286
5	0.85	0.15	0.18630	0.81370
6	0.90	0.10	0.26667	0.73333

Exhibit 8.9 Adjusting the Posterior Probability in Base SAS

Cost-Sensitive Learning

Cost-sensitive learning is another alternative for dealing with highly skewed data sets. The idea is to assign higher misclassification costs to the minority class, which in our case is the defaulters. These costs are then taken into account during classifier estimation or evaluation. Exhibit 8.10 gives the overview of the costs in a binary classification setting where $C(i,j)$ represents the cost of misclassifying an example from class j into class i.

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	$C(+,+)$	$C(-,+)$
	Negative	$C(+,-)$	$C(-,-)$

Exhibit 8.10 Misclassification Costs

Note that usually $C(+,+) = C(-,-) = 0$, and $C(-,+) > C(+,-)$, with + referring to the defaulters and – to the nondefaulters. The costs are typically also determined on an aggregated basis, rather than on an observation-by-observation basis.

A first straightforward way to make a classifier cost-sensitive is by adopting a cost-sensitive cutoff to map the posterior class probabilities to class labels. In other words, an observation x will be assigned to the class that minimizes the expected misclassification cost:

$$argmin_i\left(\sum_{j\in\{-,+\}} P(j|x)\times C(i,j)\right)$$

where $P(j|x)$ is the posterior probability of observation x to belong to class j. As an example, consider a default risk setting where class 1 consists of the defaulters and class 2 the nondefaulters. An observation x will be classified as a defaulter (class 1) if:

$$P(1|x) \times C(1,1) + P(2|x) \times C(1,2) < P(1|x) \times C(2,1) + P(2|x) \times C(2,2)$$
$$P(1|x) \times C(2,1) > P(2|x) \times C(1,2)$$
$$P(1|x) \times C(2,1) > (1 - P(1|x)) \times C(1,2)$$
$$P(1|x) > \frac{C(1,2)}{C(1,2) + C(2,1)}$$
$$P(1|x) > \frac{1}{1 + \frac{C(2,1)}{C(1,2)}}$$

As a result, the cutoff depends only on the ratio of the misclassification costs, which may be easier to determine than the individual misclassification costs themselves.

Another approach to cost-sensitive learning works by directly minimizing the misclassification cost during classifier learning. Again assuming there is no cost for correct classifications, the total misclassification cost is then:

$$\text{Total cost} = C(-,+) \times FN + C(+,-) \times FP$$

where *FN* represents the number of false negatives, and *FP* the positives. Various cost-sensitive versions of existing classification techniques have been introduced in the literature. Ting (2002) introduced a cost-sensitive version of the C4.5 decision tree algorithm where the splitting and stopping decisions are based on the misclassification cost. Veropolous et al. (1999) developed a cost-sensitive version of support vector machines (SVMs) where the misclassification costs are taken into account in the objective function of the SVM. Domingos (1999) introduced MetaCost, which is a meta-algorithm capable of turning any classifier into a cost-sensitive classifier by first relabeling observations with their estimated minimal-cost classes and then estimating a new classifier on the relabeled data set. Fan et al. (1999) developed AdaCost, a cost-sensitive variant of AdaBoost that uses the misclassification costs to update the weights in successive boosting runs.

To summarize, cost-sensitive learning approaches are usually more complex to work with than the sampling approaches discussed earlier. López et al. (2012) conducted a comparison of sampling versus cost-sensitive learning approaches for imbalanced data sets and found that both methods are good and equivalent. From a pragmatic viewpoint, it is recommended to use sampling approaches.

MAPPING TO AN EXTERNAL RATING AGENCY

Another approach is to purchase external ratings from a rating agency, such as Moody's, Standard & Poor's, or Fitch. These rating agencies generate issuer ratings (corporate and sovereign), ratings of issues (loan and bond), and structured finance ratings (e.g., asset-backed security [ABS], collateralized debt obligation [CDO], and residential mortgage-backed security [RMBS]). Let's assume we want to build a sovereign rating system and obtain country ratings. A next step is then to collect predictors that could potentially have an influence on the rating. In our country rating example, one could think here of predictors such as gross domestic product (GDP), inflation, unemployment, imports, and exports. These predictors can then be put into a cumulative logistic regression model with the rating as the target variable. For each rating, the default rates reported by the rating agencies can then be used to determine the final calibrated PD. Earlier we referred to this approach as the shadow rating approach.

Cumulative logistic regression, also called ordinal logistic regression, is an extension of logistic regression to deal with ordinal multiclass targets such as credit ratings. Remember, as discussed previously, when adopting the shadow rating approach, you want to build a classification model predicting ratings. Ratings are ordinal and have an ordering between them: AAA is better than AA, AA is better than A, and so on. In other words, when thinking about cumulative probabilities, we then have:

$$P(C \leq AAA) \geq P(C \leq AA) \geq P(C \leq A) \geq P(C \leq BBB)\ldots$$

In cumulative logistic regression, the cumulative probabilities are modeled using a logit type of transformation. The formulation goes as follows: the probability that a firm with characteristics $x_1,\ldots, x_N$ has a particular rating AAA, AA,…is calculated as (Allison 2001):

$$P(Y \leq AAA) = \frac{1}{1 + e^{-\theta_{AAA}+\beta_1 x_1+\ldots+\beta_N x_N}}$$
$$P(Y \leq AA) = \frac{1}{1 + e^{-\theta_{AA}+\beta_1 x_1+\ldots+\beta_N x_N}}$$

…

$$P(Y \leq D) = \frac{1}{1 + e^{-\theta_{D}+\beta_1 x_1+\ldots+\beta_N x_N}}$$

In a corporate setting, the *x* variables can represent firm characteristics such as accounting ratios, profitability information, and stock price information. This can then also be reformulated in terms of the odds as follows:

$$\frac{P(Y \leq AAA)}{1 - P(Y \leq AAA)} = e^{-\theta_{AAA}+\beta_1 x_1+\ldots+\beta_N x_N}$$
$$\frac{P(Y \leq AA)}{1 - P(Y \leq AA)} = e^{-\theta_{AA}+\beta_1 x_1+\ldots+\beta_N x_N}$$

…

$$\frac{P(Y \leq D)}{1 - P(Y \leq D)} = e^{-\theta_{D}+\beta_1 x_1+\ldots+\beta_N x_N}$$

Note that since $P(Y \leq AAA) = 1$, $\theta_{AAA} = +\infty$. Taking the logarithm gives linear expressions as follows:

$$\frac{P(Y \leq AAA)}{1 - P(Y \leq AAA)} = -\theta_{AAA} + \beta_1 x_1 + \ldots + \beta_N x_N$$

$$\frac{P(Y \leq AA)}{1 - P(Y \leq AA)} = -\theta_{AA} + \beta_1 x_1 + \ldots + \beta_N x_N$$

…

$$\frac{P(Y \leq D)}{1 - P(Y \leq D)} = -\theta_D + \beta_1 x_1 + \ldots + \beta_N x_N$$

The logit functions for all ratings are parallel since they differ only in the intercept. Hence, this model is also referred to as the proportional odds model (see Allison, 2001).

Just as with binary logistic regression, the β parameters of the cumulative logistic regression model can be estimated using the maximum likelihood procedure. In SAS, this is implemented by PROC LOGISTIC. The individual rating probabilities can then be obtained as follows:

$$P(Y = AA) = P(Y \leq AAA) - P(Y \leq AA)$$

$$P(Y = 1) = P(Y \leq AA) - P(Y \leq A)$$

…

$$P(Y = D) = P(D \leq 1)$$

When assigning a rating to a new observation using the cumulative logistic regression model, you can calculate all the rating probabilities and assign the observation to the rating with the highest probability (winner-take-all learning).

Exhibit 8.11 shows some examples of rating probability distributions for a model predicting insurance ratings in terms of the z-score, which is (see Van Gestel et al. 2007):

$$z = -\theta_R + \beta_1 x_1 + \ldots + \beta_N x_N$$

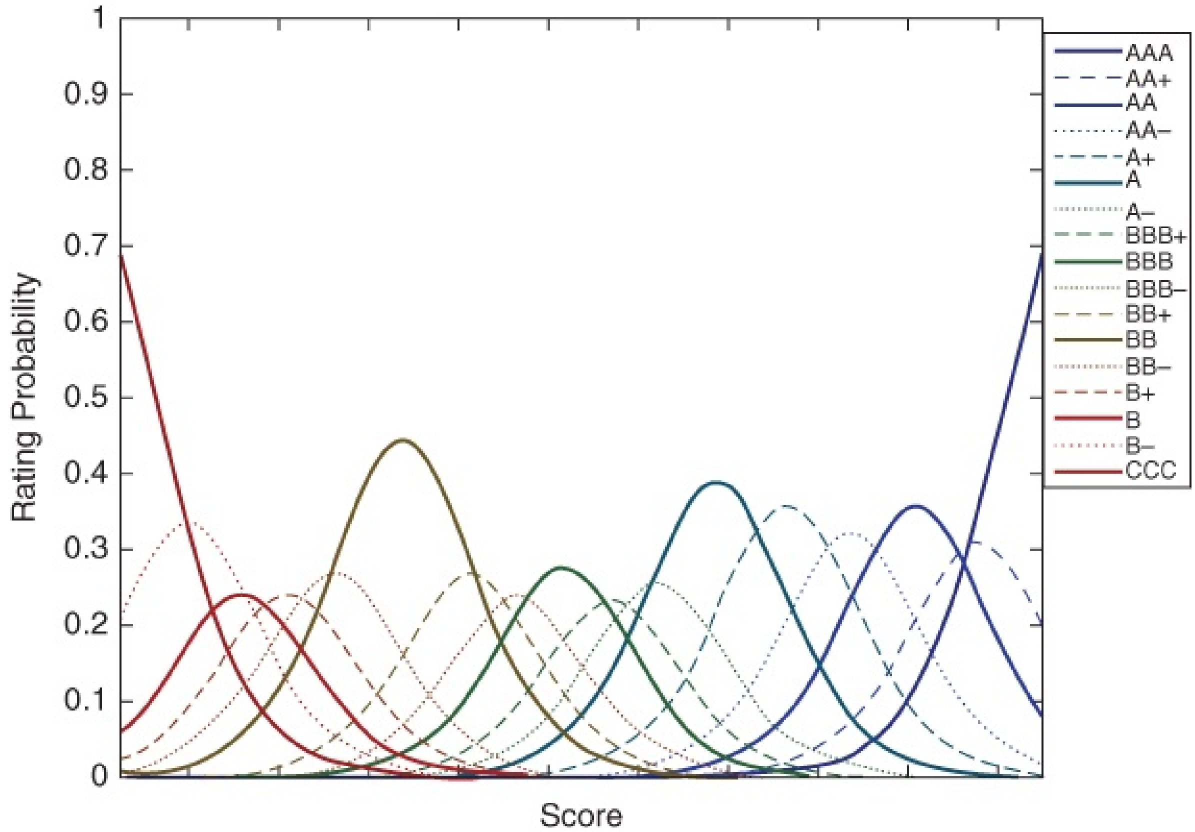


Exhibit 8.11 Rating Probability Distribution (Van Gestel et al. 2007)

It can be clearly seen that, depending on the z-score, every rating pops up at least once as the winner.

In SAS, cumulative logistic regression can be implemented using PROC LOGISTIC. The ratings data set contains ratings of 197 corporates where the target rating is numerically coded as 1 $(= \text{AAA}), 2\ (= \text{AA}), 3\ (= \text{A}), \ldots 9\ (= \text{C})$, and $10\ (= \text{Default})$. The predictors are the well-known CAMELS variables, encoded as follows:

- COMMEQTA = common equity to total assets (capital adequacy)
- LLPLOANS = loan loss provision to total loans (asset quality)
- COSTTOINCOME = operating costs to operating income (management quality)
- ROE = return on equity (earnings performance)
- LIQASSTA = liquid assets to total assets (liquidity)
- SIZE = natural logarithm of total assets (size)

We invite the reader to explore this data set using PROC MEANS and PROC UNIVARIATE.

We can now estimate a cumulative logistic regression model as follows:

```
PROC LOGISTIC DATA=data.ratings;
MODEL rating= COMMEQTA LLPLOANS COSTTOINCOME ROE LIQASSTA SIZE;
OUTPUT OUT =ratingsout PREDPROBS=INDIVIDUAL;
RUN;
```

The output statement specifies that the output should be written to the data set ratingsout. The PREDPROBS=INDIVIDUAL option indicates that we want to include estimated probabilities for each of the ratings individually.

Running the preceding PROC LOGISTIC gives the maximum likelihood estimates shown in Exhibit 8.12.

The LOGISTIC Procedure						
Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**1**	1	−18.5764	2.1870	72.1484	<.0001
Intercept	**2**	1	−17.5111	2.1331	67.3889	<.0001
Intercept	**3**	1	−16.3356	2.0981	60.6218	<.0001
Intercept	**4**	1	−15.0194	2.0588	53.2212	<.0001
Intercept	**5**	1	−13.8925	2.0231	47.1553	<.0001
Intercept	**6**	1	−12.3558	1.9753	39.1261	<.0001
Intercept	**7**	1	−11.4118	1.9525	34.1625	<.0001
Intercept	**8**	1	−10.3187	1.9381	28.3453	<.0001
Intercept	**9**	1	−8.5034	1.9856	18.3402	<.0001
COMMEQTA		1	10.1167	3.5110	8.3024	0.0040
LLPLOANS		1	−29.4204	11.6879	6.3362	0.0118
COSTTOINCOME		1	−3.3613	0.7982	17.7350	<.0001
ROE		1	3.2776	1.2446	6.9345	0.0085
LIQASSTA		1	3.8190	0.9057	17.7802	<.0001
SIZE		1	0.7807	0.1078	52.4427	<.0001

Exhibit 8.12 Maximum Likelihood Estimates from PROC LOGISTIC

From Exhibit 8.12, it can be seen that all variables are significant. The performance can be evaluated by measuring the association between the predicted probabilities and the observed responses as indicated in Exhibit 8.13.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	75.2	**Somers' D**	0.508
Percent Discordant	24.3	**Gamma**	0.511
Percent Tied	0.5	**Tau-a**	0.432
Pairs	16386	**c**	0.754

Exhibit 8.13 Association Statistics from PROC LOGISTIC

The table in Exhibit 8.13 shows that the model has good performance. As discussed earlier, the ratingsout data set contains the predicted probabilities for each of the ratings, combined with the predicted rating where the latter is decided based on the largest probability (winner-take-all learning). We can now evaluate the performance of this model by creating a notch difference table and corresponding graph. We first create the following temporary SAS data set:

```
DATA temp;
SET ratingsout;
notchdiff= ABS(_FROM_ - _INTO_);
RUN;
```

The _FROM_ and _INTO_ variables contain the original rating and the predicted rating, respectively. The notchdiff variable then calculates the absolute value of the difference between them. We can now run PROC FREQ as follows:

```
ODS GRAPHICS ON;
PROC FREQ DATA=temp;
TABLES notchdiff / PLOTS=CUMFREQPLOT(SCALE=PERCENT);
RUN;
ODS GRAPHICS OFF;
```

This will give the output shown in Exhibit 8.14.

The FREQ Procedure				
notchdiff	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	63	31.98	63	31.98
1	77	39.09	140	71.07
2	34	17.26	174	88.32
3	17	8.63	191	96.95
4	3	1.52	194	98.48
5	2	1.02	196	99.49
6	1	0.51	197	100.00

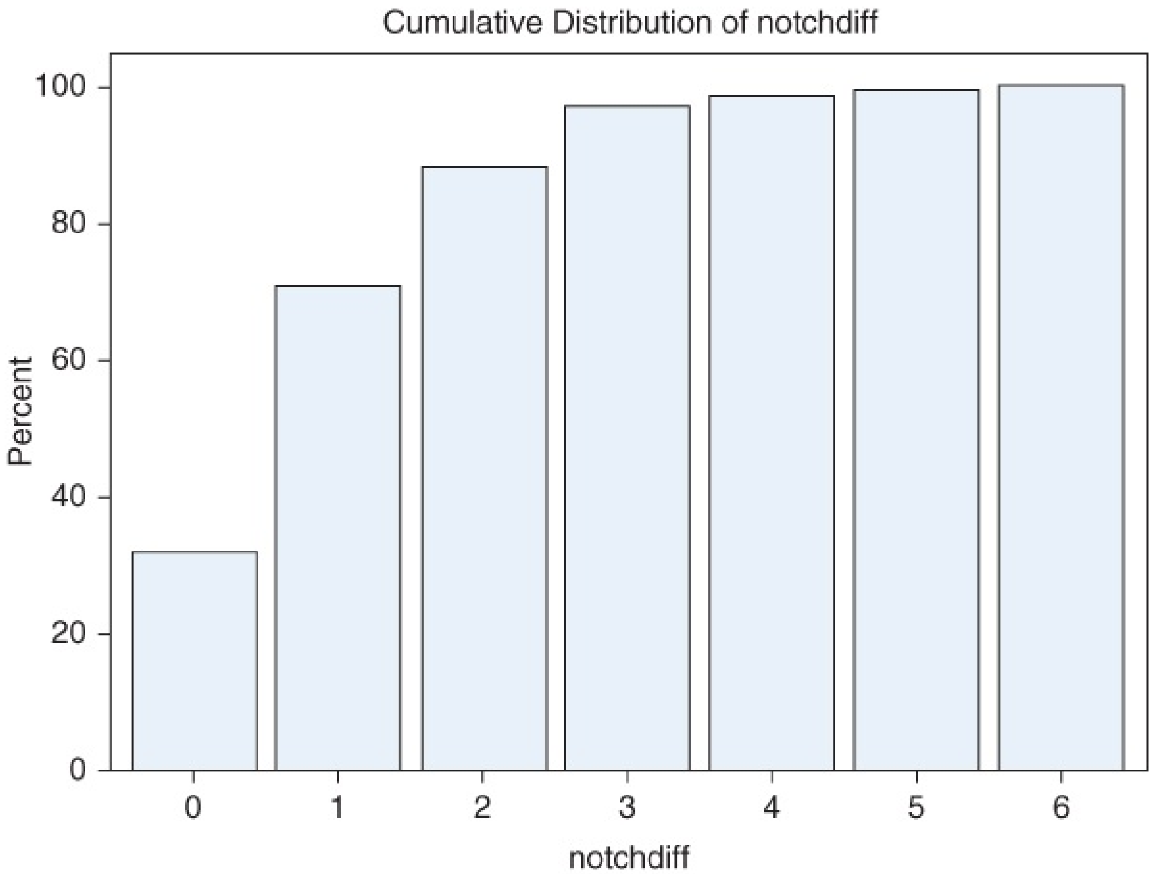


Exhibit 8.14 Notch Difference Graph for the Ratings Data Set

From Exhibit 8.14 we can see that at a 0-notch difference level, the model obtains a classification accuracy of 31.98 percent. By allowing a 1-notch difference between the predicted and target ratings, the cumulative accuracy increases to 71.07 percent.

CONFIDENCE LEVEL BASED APPROACH

Another interesting approach for working with a low number of defaulters is the confidence level based approach developed by Pluto and Tasche (2005).

Let's start with the most extreme example of a skewed data set, which is a data set with no defaulters at all. Obviously, none of the sampling approaches discussed so far will work for this. Assume now that we have an expert-based credit risk model that can discriminate the observations into default risk classes A, B, and C using a set of predefined business rules. Although these three classes allow us to discriminate among the observations in terms of their default risk, it would also be handy to accompany each of these classes with default probability estimates. These probabilities can then be used to calculate both the expected and the unexpected credit losses.

In a first step, we calculate the probability of default (PD) for class A, PD_A . A key assumption we make is that default occurs independently. Although this assumption might seem naive at first sight, it allows us to derive probability estimates in a fairly straightforward way given this complex setting with no data about defaulters. More specifically, we first assume that the ranking of the observations across the three default risk classes is correct, or, in other words, $PD_A \leq PD_B \leq PD_C$. The most prudent estimate (sometimes also referred to as the most conservative estimate) is then obtained under the temporary assumption that $PD_A = PD_B = PD_C$. Hence, the probability of default equals PD_A for every observation. Given that we have n_A observations in rating A, n_B observations in rating B, and n_C observations in rating C, and that default occurs independently, the probability of not observing any defaulter in the total data set equals:

$$(1 - PD_A)^{n_A + n_B + n_C}$$

We can now specify a confidence region for PD_A, which is the region of all values of PD_A such that the probability of not observing any defaulter is higher than 1 – α, or, in other words:

$$1 - \alpha \leq (1 - PD_A)^{n_A + n_B + n_C}$$

or

$$PD_A \leq 1 - (1 - \alpha)^{1/(n_A + n_B + n_C)}$$

Assume we have 100 observations in rating A, 200 in rating B, and 50 in rating C. Exhibit 8.15 illustrates the values obtained for PD_A by varying the confidence level from 50 percent to 99.9 percent. As can be observed, PD_A increases as the confidence level increases.

α	**50%**	**75%**	**90%**	**95%**	**99%**	**99.9%**
PD_A	0.20%	0.39%	0.65%	0.85%	1.31%	1.95%

Exhibit 8.15 Values for PD_A for a Data Set with No Defaulters

We can now continue this same procedure to compute PD_B. We have $n_B + n_C$ observations left. The most prudent estimate of PD_B is obtained by again assuming $PD_B = PD_C$. Hence, we have:

$$1 - \alpha \leq (1 - PD_B)^{n_B + n_C}$$

or

$$PD_B \leq 1 - (1 - \alpha)^{1/(n_B + n_C)}$$

For our data set, this gives the values reported in Exhibit 8.16.

α	**50%**	**75%**	**90%**	**95%**	**99%**	**99.9%**
PD_B	0.28%	0.55%	0.92%	1.19%	1.82%	2.72%

Exhibit 8.16 Values for PD_B for a Data Set with No Defaulters

Finally, we can calculate PD_C as follows:

$$1 - \alpha \leq (1 - PD_C)^{n_C}$$

or

$$PD_C \leq 1-(1-\alpha)^{1/n_C}$$

This gives the values reported in Exhibit 8.17.

α	50%	75%	90%	95%	99%	99.9%
PD_C	1.38%	2.73%	4.50%	5.81%	8.80%	12.90%

Exhibit 8.17 Values for PD_C for a Data Set with No Defaulters

Note that despite having no defaulters in the data, PD_C at the 99.9 percent confidence level equals 12.90 percent, which is quite high. Also observe that for a given confidence level, $PD_A \leq PD_B \leq PD_C$ as required at the outset. An obvious question is what confidence level should be adopted. Before answering this question, we illustrate how this approach can be implemented in SAS.

Assume we have a portfolio with three ratings as follows: A (100 obligors), B (150 obligors), and C (80 obligors):

```
%LET nA=100;
%LET nB=150;
%LET nC=80;
/*Set the confidence level to 99%*/
%LET sig=0.99;
/*Create a data set LDP with the calculations*/
DATA LDP;
PDA=1-(1-&sig)**(1/(&nA+&nB+&nC));
PDB=1-(1-&sig)**(1/(&nB+&nC));
PDC=1-(1-&sig)**(1/(&nC));
RUN;
/*Print data set*/
Proc PRINT DATA=LDP;
RUN;
```

Running and inspecting this data set then yields the results reported in Exhibit 8.18.

Obs	PDA	PDB	PDC
1	0.013858	0.019823	0.055939

Exhibit 8.18 Example of Confidence Level Based Approach in Base SAS

Let's now assume we have one defaulter in rating A, two in rating B, and four in rating C (seven in all). We first determine PD_A using again the most prudent estimate principle: $PD_A = PD_B = PD_C$. By using the binomial distribution to calculate the probability of observing less than or equal to seven defaulters, PD_A can be found as follows:

$$1-\alpha \leq \sum_{i=0}^{7} \binom{n_A+n_B+n_C}{i} PD_A^i (1-PD_A)^{n_A+n_B+n_C-i}$$

Likewise, PD_B and PD_C can be found as follows:

$$1-\alpha \leq \sum_{i=0}^{6} \binom{n_B+n_C}{i} PD_B^i (1-PD_B)^{n_B+n_C-i}$$

$$1-\alpha \leq \sum_{i=0}^{4} \binom{n_C}{i} PD_C^i (1-PD_C)^{n_C-i}$$

Exhibit 8.19 displays the values obtained depending upon the confidence levels. Again, note that the probabilities increase for increasing confidence levels. Just as in the previous examples, also observe that $PD_A \leq PD_B \leq PD_C$ as required at the outset.

α	**50%**	**75%**	**90%**	**95%**	**99%**	**99.9%**
PD_A	2.19%	2.76%	3.34%	3.72%	4.51%	5.51%
PD_B	2.66%	3.41%	4.17%	4.68%	5.73%	7.05%
PD_C	9.28%	12.26%	15.35%	17.38%	21.50%	26.56%

Exhibit 8.19 Values for PD_A, PD_B, and PD_C for a Data Set with Defaulters

As already mentioned, a key question to answer when adopting this approach is the setting of the confidence level. Obviously, this depends on how conservative the estimates should be. As illustrated, a higher confidence level results in a higher probability of default estimate. In their original paper, Pluto and Tasche (2005) suggest not to exceed 95 percent. Benjamin, Cathcart, and Ryan (2006) suggest adopting confidence levels between 50 percent and 75 percent.

Various extensions of this standard approach have been developed. A first extension takes into account correlated default events by using the Basel single-factor model and the asset correlations mentioned in the Basel Accord. More specifically, in the case of our earlier example (one defaulter in rating A, two in rating B, and four in rating C), the confidence region at level α for PD_A corresponds to the set of values of PD_A that satisfy the following inequality:

$$1-\alpha \leq \int_{-\infty}^{+\infty} \phi(x) \sum_{i=0}^{7} \binom{n_A+n_B+n_C}{i} CDF(PD_A,\rho,x)^i (1-CDF(PD_A,\rho,x))^{n_A+n_B+n_C-i} dx$$

where $\phi(x)$ is the standard normal density and $CDF(PD_A,\rho,x)$ represents the conditional PD based on the unconditional average PD_A, asset correlation ρ, and systematic factor x, and is calculated as follows:

$$CDF(PD_A,\rho,x) = \Phi\left(\frac{\Phi^{-1}(PD_A)-\sqrt{\rho}x}{\sqrt{1-\rho}}\right)$$

with Φ the cumulative standard normal distribution and Φ^{-1} its inverse. Note that the right-hand side of the inequality represents the one-period probability of not observing any default among the $n_A + n_B + n_C$ obligors with default probability PD_A. Following the Basel Accord, the asset correlation ρ for corporate exposures can be set to 12 percent. The preceding inequality can then be solved numerically (Pluto and Tasche 2005) or by using a simulation approach (Clifford, Marianski, and Sebestyen 2013). Similar expressions can then be derived for PD_B and PD_C as follows:

$$1-\alpha \le \int_{-\infty}^{+\infty} \phi(\mathrm{x}) \sum_{i=0}^{6} \binom{n_B+n_C}{i} CDF(PD_B,\rho,x)^i (1-CDF(PD_B,\rho,x))^{n_B+n_C-i} dx$$
$$1-\alpha \le \int_{-\infty}^{+\infty} \phi(\mathrm{x}) \sum_{i=0}^{4} \binom{n_C}{i} CDF(PD_C,\rho,\mathrm{x})^i (1-CDF(PD_C,\rho,\mathrm{x}))^{n_C-i} dx$$

As illustrated in Pluto and Tasche (2005), the effect of including correlations is that the probabilities PD_A, PD_B, and PD_C will increase. Given the increased complexity of taking into account default correlation, one might consider using the uncorrelated case with a higher confidence level instead.

Another extension proposed in Pluto and Tasche (2005) ensures that the average estimated portfolio PD equals the observed portfolio PD by introducing a scaling factor K as follows:

$$\frac{\widehat{PD}_A n_A + \widehat{PD}_B n_B + \widehat{PD}_C n_C}{n_A+n_B+n_C} K = PD_{Portfolio}$$

where $PD_{Portfolio}$ represents the observed portfolio PD. The scaled PD estimates then become:

$$\widehat{PD}_{A,scaled} = K \cdot \widehat{PD}_A; \widehat{PD}_{B,scaled} = K \cdot \widehat{PD}_B; \widehat{PD}_{C,scaled} = K \cdot \widehat{PD}_C$$

Note that this scaling will obviously make the estimates less conservative. One way to deal with this is by scaling according to the upper bound of the overall portfolio PD, which is also determined using the most prudent estimation principle.

This approach can also be extended to a multiperiod setting. See Pluto and Tasche (2005) for more details.

OTHER METHODS

Other methods can also be adopted for modeling low default portfolios. A first popular approach is to pool your low default portfolio data with other banks or market participants, such as credit bureaus, for example. Another approach is to aggregate subportfolios with similar risk characteristics to increase the default history. Suppose you have a portfolio with U.S.-based small and medium-sized enterprises (SMEs), which would be a low default portfolio. If you also happen to have a portfolio with Canadian SMEs, which also are low default, you could consider combining the two portfolios, thereby increasing the number of defaulters and facilitating the modeling. You could also consider combining rating categories and analyzing the PDs of the combined category in a manner consistent with the Basel Accord, for corporates, sovereigns, and banks. Alternatively, you could use an upper bound on the PD estimate as input to the capital requirement formulas, or infer the PD estimates with a horizon of more than one year and then annualize the resulting figure. Finally, you can also use the lowest nondefault rating as a proxy for default. Note, however, that here it is still necessary to calibrate the ratings to a PD in a manner that is consistent with the Basel definition.

Another statistical approach for LDPs can be applied via Bayesian statistics. We will describe this technique in the chapter on Bayesian methods for credit risk modeling and will revisit computing PDs for LDPs there.

LGD AND EAD FOR LOW DEFAULT PORTFOLIOS

In this chapter, we have discussed various methods for modeling PD in the case of low default portfolios. An obvious question is: What about LGD and EAD? In fact, for LGD and EAD, the problem becomes more challenging as you are restricted to the use of the defaulters only, whereas in the case of PD you can at least use both the defaulters and the nondefaulters to build the model. Here you can see a statement from the FSA in the United Kingdom (Financial Services Authority 2006b):

> It is accepted that the U.K. and international community are not well advanced in their thinking on EADs, and as a consequence estimation for LDPs adds further difficulty. As a result the EG fully endorses implementation to be pragmatic, flexible, and principle based.

However, this quote doesn't give precise input on how to deal with LGD and EAD in the presence of low default portfolios.

One approach might be to make use of a relationship between PD and LGD, as suggested by Frye (2013). In line with the capital requirements formulas in the Basel Accord, the latter assumes that the conditional default rate follows a Vasicek distribution. Suppose we now also assume that the conditional expected loss rate (cLoss) has a Vasicek distribution with the same value of ρ:

$$\text{cLoss} = CDF^{-1}_{cLoss}(q) = \Phi\left(\frac{\Phi^{-1}(EL) + \sqrt{\rho}\,\Phi^{-1}(q)}{\sqrt{1-\rho}}\right)$$
$$= \Phi\left(\Phi^{-1}(cDR) - \frac{\Phi^{-1}(PD) - \Phi^{-1}(EL)}{\sqrt{1-\rho}}\right)$$

with CDF the cumulative distribution and CDF^{-1} its inverse, Φ the cumulative standard normal distribution and Φ^{-1} its inverse, q the quantile considered, EL the unconditional expected loss, PD the unconditional PD, and cDR the conditional default rate. The conditional LGD can then be obtained by dividing this equation by *cDR* as follows:

$$cLGD = \frac{\Phi(\Phi^{-1}(cDR) - k)}{cDR} \text{ with } k = \frac{\Phi^{-1}(PD) - \Phi^{-1}(EL)}{\sqrt{1-\rho}}$$

This relationship allows us to calculate the *cLGD* based on a loan's PD, ρ, and EL. The PD can be determined using any of the methods discussed before, the asset correlation ρ can be set as specified in the Basel Accord (e.g., between 12 percent and 24 percent for corporate loans), and the EL can be determined based on the spread of the loan. Other relationships between PD and LGD have been developed in Frye and Jacobs (2012), Frye (2000), Pykhtin (2003), Tasche (2004), Giese (2005), Hillebrand (2006), and Rösch and Scheule (2012). Note that each of these approaches starts from making some assumptions and introduces additional parameters that need to be set in either a quantitative way (i.e., based on statistical analysis) or an expert-based way.

Another option to determine LGD and EAD for LDPs is to use the reference values set in the foundation internal ratings based (IRB) approach. Remember, in the foundation IRB approach the LGD is set to 45 percent for senior claims and 75 percent for subordinated claims. This approach is also suggested by the European Banking Authority (EBA) in its 2015 discussion paper, as the following quote illustrates (European Banking Authority 2015):

> It proved particularly difficult to calculate LGD for LDP. As a result in such cases the LGD values for unsecured senior exposures are often close to the regulatory value under the FIRB Approach (45%) for credit institutions and large corporate portfolios.

For EAD, the foundation IRB credit conversion factor (CCF) is set to 20 percent for commitments less than one year, and 50 percent for commitments more than one year.

PRACTICE QUESTIONS

1. What is the difference between SMOTE and undersampling/oversampling?
2. Consider the following table with logistic regression probabilities:

	Posteriors Using Resampled Data	
	***P*(Default)**	***P*(Nondefault)**
Customer 1	0.1	0.9
Customer 2	0.3	0.7
Customer 3	0.5	0.5
Customer 4	0.6	0.4
Customer 5	0.85	0.15
Customer 6	0.9	0.1

where $p(C_1) = 0.10, p(C_2) = 0.90, p_r(C_1) = 0.20$, and $p_r(C_2) = 0.80$. Calculate the posteriors recalibrated to the original data.

Build a decision tree for the ratings data set in Enterprise Miner and compare it to a cumulative logistic regression model in terms of model representation, significant variables, and performance.

3. Let's assume we have an LDP with three ratings: A (200 obligors), B (100 obligors), and C (50 obligors). Use the confidence level based approach to calculate PD_A, PD_B, and PD_C for confidence levels 90 percent, 95 percent, and 99 percent.
4. Discuss some approaches that could be used to determine LGD and EAD for LDPs.

References

Allison, P. D. 2001. *Logistic Regression Using the SAS System: Theory and Application*. New York: John Wiley & Sons–SAS.

Basel Committee on Banking Supervision. 2005. “Validation of Low-Default Portfolios in the Basel II Framework.” *Basel Committee Newsletter no. 6*, September.

Benjamin, N., A. Cathcart, and K. Ryan. 2006. “Low Default Portfolios: A Proposal for Conservative Estimation of Default Probabilities.” Discussion paper, Financial Services Authority.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16:321–357.

Clifford, T., A. Marianski, and K. Sebestyen. 2013. “Low Default Portfolio Modeling–Probability of Default–Calibration Conundrum.” Deloitte.

Domingos, P. 1999. “MetaCost: A General Method for Making Classifiers Cost-Sensitive.” In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, 155–164. New York: ACM Press.

European Banking Authority. 2015. “Future of the IRB Approach.” Discussion Paper 2015/01.

Fan, W., S. J. Stolfo, J. Zhang, and P. K. Chan. 1999. “AdaCost: Misclassification Cost-Sensitive Boosting.” In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, 97–105. Waltham, MA: Morgan Kaufmann.

Financial Services Authority (FSA). 2006a. “CP06/3 Strengthening Capital Standards 2.”

Financial Services Authority (FSA). 2006b. “Expert Group Paper on Exposure at Default.”

Frye, J. 2000. “Depressing Recoveries.” *Risk*, 108–111.

Frye, J., 2013. “Loss Given Default as a Function of the Default Rate,” Federal Reserve Bank of Chicago.

Frye, J., and M. Jacobs. 2012. “Credit Loss and Systematic Loss Given Default.” *Journal of Credit Risk* 8 (1): 1–32.

Giese, G. 2005. “The Impact of PD/LGD Correlations on Credit Risk Capital.” *Risk*, 79–84.

Hillebrand, M. 2006. “Modeling and Estimating Dependent Loss Given Default.” *Risk*, 120–125.

López, V., A. Fernández, J. Moreno-Torres, and F. Herrera. 2012. “Analysis of Preprocessing vs. Cost-Sensitive Learning for Imbalanced Classification: Open Problems on Intrinsic Data Characteristics.” *Expert Systems with Applications* 39:6585–6608.

Pluto, K., and D. Tasche. 2005. “Estimating Probabilities of Default for Low Default Portfolios.” In *The Basel II Risk Parameters*, edited by B. Engelmann and R. Rauhmeier. New York: Springer.

Prudential Regulation Authority. 2013. “Internal Ratings Based Approaches.” SS 11/13, December.

Pykhtin, M. 2003. “Unexpected Recovery Risk.” *Risk*, 74–78.

Rösch, D., and H. Scheule. 2012. “Forecasting Probabilities of Default and Loss Rates Given Default in the Presence of Selection.” *Journal of the Operational Research Society* 65 (3): 393–407.

Saerens, M., P. Latinne, and C. Decaestecker. 2002. “Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure.” *Neural Computation* 14 (1): 21–41.

Tasche, D. 2004. “The Single Risk Factor Approach to Capital Charges in the Case of Correlated Loss Given Default Rates.” Working paper.

Ting, K.M. 2002. “An Instance-Weighting Method to Induce Cost-Sensitive Decision Trees,” IEEE Transactions on Knowledge and Data Engineering, 14(3):659–665, 2002.

Van Gestel, T., D. Martens, B. Baesens, D. Feremans, J. Huysmans, and J. Vanthienen. 2007. “Forecasting and Analyzing Insurance Companies' Ratings.” *International Journal of Forecasting* 23 (3): 513–529.

Van Vlasselaer, V., T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens. 2016. GOTCHA! Network-Based Fraud Detection for Social Security Fraud.” *Management Science*, forthcoming.

Veropoulos, K., C. Campbell, and N. Cristianini. 1999. “Controlling the Sensitivity of Support Vector Machines,” *In Proceedings of the International Joint Conference on AI*, 55–60.