Math Geosci
DOI 10.1007/s11004-010-9311-8

Sampling Bias and Class Imbalance in Maximum-likelihood Logistic Regression

Thomas Oommen · Laurie G. Baise · Richard M. Vogel

Received: 30 October 2009 / Accepted: 20 September 2010
© International Association for Mathematical Geosciences 2010

Abstract Logistic regression is a widely used statistical method to relate a binary response variable to a set of explanatory variables and maximum likelihood is the most commonly used method for parameter estimation. A maximum-likelihood logistic regression (MLLR) model predicts the probability of the event from binary data defining the event. Currently, MLLR models are used in a myriad of fields including geosciences, natural hazard evaluation, medical diagnosis, homeland security, finance, and many others. In such applications, the empirical sample data often exhibit class imbalance, where one class is represented by a large number of events while the other is represented by only a few. In addition, the data also exhibit sampling bias, which occurs when there is a difference between the class distribution in the sample compared to the actual class distribution in the population. Previous studies have evaluated how class imbalance and sampling bias affect the predictive capability of asymptotic classification algorithms such as MLLR, yet no definitive conclusions have been reached.

We hypothesize that the predictive capability of the model is related to the sampling bias associated with the data so that the MLLR model has perfect predictability when the data have no sampling bias. We test our hypotheses using two simulated datasets with class distributions that are 50:50 and 80:20, respectively. We construct a suite of controlled experiments by extracting multiple samples with varying class imbalance and sampling bias from the two simulated datasets and fitting MLLR models to each of these samples. The experiments suggest that it is important to develop a sample that has the same class distribution as the original population rather than

T. Oommen (✉) · L.G. Baise · R.M. Vogel
Department of Civil and Environmental Engineering, Tufts University, 113 Anderson Hall, Medford, MA 02155, USA
e-mail: toommen@mtu.edu

Present address:
T. Oommen
Dept. of Geological Engineering, Michigan Tech., Houghton, MI 49931, USA

Published online: 15 October 2010

ensuring that the classes are balanced. Furthermore, when sampling bias is reduced either by using over-sampling or under-sampling, both sampling techniques can improve the predictive capability of an MLLR model.

Keywords Probabilistic classification · Imbalance · Sampling · Binary class · Over-sampling · Under-sampling

1 Introduction

Linear regression is perhaps the most commonly used statistical method for predicting the value of a dependent variable from observed values of a set of predictor variables. However, linear regression requires the dependent variable to be continuous and the model residuals to be normally distributed. Logistic regression is a variation of linear regression for situations where the dependent variable is not a continuous parameter but rather a binary event (e.g., yes/no, 0/1, etc.). The value predicted using a logistic regression is a probability of the event, ranging from 0 to 1. A check of the ISI Web of Knowledge reveals that 11,725 papers were published in 2008 in which "logistic regression" appeared in either the title or among the key words. King and Zeng (2001) referred to the use of maximum likelihood as the nearly universal method for parameter estimation in logistic regression. Maximum-likelihood logistic regression (MLLR) has been applied to a wide range of applications, including, but not limited to, radar detection of hail (Lopez and Sanchez 2009), fingerprint matching (Cao et al. 2009), bank failure predictions (Boyacioglu et al. 2009), wildfire risk (Preisler et al. 2004), and diagnoses of rare medical conditions (Correia et al. 2009; Page et al. 2009).

MLLR is also widely used in engineering and geoscience applications such as evaluation of soil liquefaction potential (Juang et al. 2001, 2002; Lai et al. 2006), assessment of landslide susceptibility (Atkinson and Massari 1998; Carrara 1983; Chung and Fabbri 2003), classification of intermittent and perennial streams (Bent and Steeves 2006; Olson and Brouillette 2006), regionalization of low streamflows (Tasker 1989), mineral exploration (Agterberg 1974; Bonham-Carter and Chung 1989; Caumon et al. 2006), and for classifying wetlands (Toner and Keddy 1997). In these applications, the empirical data often have a large number of events/examples from one class while the other is represented by only a few instances. This imbalance in the data is referred to as class imbalance. For natural hazard applications, it is common for the hazard event (yes or one) to be sampled much more frequently than the non-hazard event (no or zero). For example, we see class imbalance in a soil liquefaction cone penetration test database where the class ratio of instances of liquefaction to non-liquefaction is 76:24 (Moss et al. 2006). In addition to class imbalance, the class ratio of the data/sample is often different from the class ratio in the population. This difference in the class ratio between the sample and the population is known as sampling bias. For example, if the true population of the data has a class ratio of 80:20 and a sample has a class ratio of 50:50, then the sample has no class imbalance but it exhibits sampling bias.

Since the 1980s and recently across a variety of disciplines, there have been several attempts by researchers to answer the following question: how does class imbalance affect the predictive capability of asymptotic classification algorithms such as MLLR? (Burez and Van den Poel 2008; Cosslett 1981a; Garcia et al. 2008; Gu et al. 2008; Liu et al. 2009; Seiffert et al. 2009; Sun et al. 2009; Tang et al. 2009; Williams et al. 2009). These studies have generally minimized or removed class imbalances using basic sampling methods. Two common methods to minimize class imbalance are under- and over-sampling. Under-sampling eliminates majority-class events and over-sampling duplicates minority-class events. However, none of the previously cited studies have drawn definitive conclusions regarding when and how class imbalance affects MLLR. The current consensus is that when both classes are equally easy to collect, an equal sampling is optimal in a few situations and near optimal in most situations. (Cosslett 1981b; Imbens 1992; King and Zeng 2001). Others have concluded that "there is no general answer to which class distribution will perform best, and the answer is surely method and case dependent," (Weiss and Provost 2003) or "there is no need to under-sample so that there are as many churners in your training set as non-churners" (Burez and Van den Poel 2008).

It is evident that no clear consensus has emerged on whether a dataset should have an equal number of classes or a unique class distribution (optimal imbalance) for optimal classifier performance. It is also not clear whether class imbalance is case dependent and, if so, how the optimal imbalance for each case can be determined a priori. From the literature, most of the discussion is on class imbalance without a clear discussion on how sampling bias and class imbalance interact.

We analyze the competing issues of sampling bias and class imbalance on the performance of MLLR models using controlled experiments based on simulated data. We test our hypotheses that the MLLR model has perfect predictability when the data have no sampling bias. We simulate two datasets with alternate class-imbalance ratios (50:50 and 80:20) and then sample these datasets to produce samples with the following class distributions: 50:50, 60:40, 70:30, 80:20, 90:10, 95:5, 98:2, and 99:1. Finally, we develop MLLR models on these samples and quantify their predictive capability using various statistical measures.

2 Simulated Data

We generate two samples from a logistic model with known model parameters to enable us to perform controlled MLLR experiments. Agresti (2002) and others document that for a logistic regression, the probability distribution of the dependent variable y follows the Bernoulli probability mass function

$$f(y,x,\alpha,\beta) = \bigl(1-\pi(x,\alpha,\beta)\bigr)\exp\biggl(y\cdot\ln\biggl(\frac{\pi(x,\alpha,\beta)}{1-\pi(x,\alpha,\beta)}\biggr)\biggr), \tag{1}$$

where

$$\pi(x,\alpha,\beta) = \frac{\exp(\alpha+\beta\cdot x)}{1+\exp(\alpha+\beta\cdot x)}, \tag{2}$$

and where y is a binary dependent variable of interest, x is a predictor variable, α and β are regression coefficients, and π is the continuous probability associated with the dependent variable of interest y. Taking logarithms, one can easily show that

$$\ln\bigl[\pi/(1-\pi)\bigr] = \alpha+\beta\cdot x, \tag{3}$$

where $\ln[\pi/(1-\pi)]$ is often termed the logit function.

The Bernoulli Probability Mass Function (PMF) yields

$$f_0(x,\alpha,\beta)=\bigl(1-\pi(x,\alpha,\beta)\bigr),\quad \text{and} \quad f_1(x,\alpha,\beta)=\pi(x,\alpha,\beta). \tag{4}$$

For a fixed probability p, one can generate the Bernoulli trial y using

$$y(p)=\left|\begin{array}{ll}\multicolumn{2}{l}{\text{dummy} \leftarrow \text{rnd}(1),}\\ 0 & \text{if dummy} < p,\\ 1 & \text{otherwise,}\end{array}\right. \tag{5}$$

where p is the probability of y. The conditional Bernoulli trials y are then generated by substitution of $\pi(x,\alpha,\beta)$ in place of $f(p)$ in (4) so that

$$y(\pi(x,\alpha,\beta))=\left|\begin{array}{ll}\multicolumn{2}{l}{\text{dummy} \leftarrow \text{rnd}(1),}\\ 0 & \text{if dummy} < \pi(x,\alpha,\beta),\\ 1 & \text{otherwise.}\end{array}\right. \tag{6}$$

We simulate a predictor variable x from a uniform distribution, which ranges from 0 to 10. For Case A, we use $\alpha=-10$ and $\beta=2$, and for Case B, we use $\alpha=-10$ and $\beta=3.85$. For each value of x, we calculate the true value of $\pi(x,\alpha,\beta)$ using the corresponding α and β values for each case. The true value of $\pi(x,\alpha,\beta)$ is substituted into (6) to generate the conditional Bernoulli trial y.

The theoretical properties of the simulated datasets are presented in Fig. 1 in terms of both the probabilities given by (2) and the logit function given in (3). Both the datasets have a total of 50,000 events, with the Case A dataset having a class distribution of about 50:50 (class 0: class 1:: 24970:25030) and Case B dataset having a class distribution of about 80:20 (class 0: class 1:: 39102:10757).

3 Methodology

To test the hypothesis that sampling bias controls the optimal class balance required for the best predictive performance of MLLR model, we extract samples from the Case A and Case B datasets with different class distributions. These samples have class distributions (that is, the ratio of class 0 to class 1) varying from 50:50, 60:40, 70:30, 80:20, 90:10, 95:5, 98:2, and 99:1, respectively. For each sample class distribution, we carry out 1000 Monte Carlo simulations and for each simulation we develop an MLLR model. The MLLR model is developed using the glm function in the stats package of the R programming language (R Development Core Team 2009). The predictive performance of each MLLR model is quantified using statistical model validation metrics such as area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AU-PRC), precision (Prec), recall (Rec), and F-measure (F-M)/F-score (van Rijsbergen 1979). These metrics are all computed from elements of the confusion matrix. A confusion matrix is a table used to evaluate the performance of a classifier. It is a matrix of the observed versus the predicted classes, with the observed classes in columns and the predicted classes in rows as shown in Table 1. The diagonal elements (where the row index equals

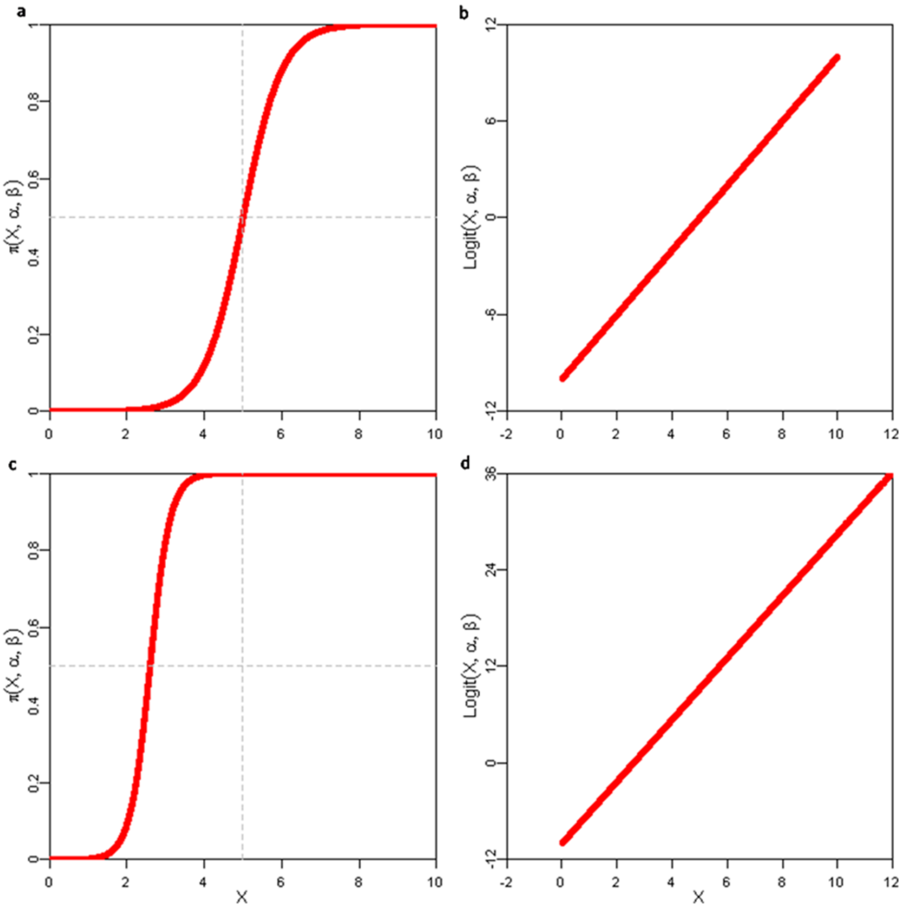


Fig. 1 Theoretical properties of the simulated datasets: (**a**) predictor variable x versus the probability of x for Case A, (**b**) predictor variable x versus the logit of x for Case A, (**c**) predictor variable x versus the probability of x for Case B, and (**d**) predictor variable x versus the logit of x for Case B

Table 1 Confusion matrix, presenting the observed classes in rows and the predicted classes in columns where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative

		Observed	
		Yes	No
Predicted	Yes	TP	FP
	No	FN	TN

the column index) include the frequencies of correctly classified instances and the non-diagonal elements include the frequencies of misclassifications.

To develop the confusion matrix for evaluating a probabilistic classifier, we must choose a probability threshold value that marks the class boundary between class 0 and class 1. The selection of a probability threshold to mark the class boundary is handled in two ways, either by choosing a single threshold value or by using the complete spectrum of thresholds (that is, operating conditions). In this study, we evaluated the MLLR models both ways.

When a single threshold value is chosen, we use the probability value that maximizes the accuracy as the optimal threshold. The accuracy is given by

$$\text{Accuracy} = (TP + TN)/(P + N), \tag{7}$$

where the true positive (*TP)* is the sum of instances of class 0 correctly predicted, true negative (*TN*) is the sum of instances of class 1 correctly predicted, P is all the instances of class 0, and N is all the instances of class 1. When a single threshold value is used, the predictive performance of the MLLR model is evaluated using metrics such as Prec, Rec, and F-M, applied separately to the different classes in the dataset.

Prec measures the accuracy of the predictions for a single class, whereas Rec measures accuracy of predictions only considering predicted values.

$$\text{Prec} = TP/(TP + FP), \tag{8}$$

and

$$\text{Rec} = TP/(TP + FN), \tag{9}$$

where the False Positive (*FP)* is the sum of instances of class 1 classified as class 0, and the where the false negative (*FN)* is the sum of instances of class 0 classified as class 1.

The F-M combines the Prec and Rec value to a single evaluation metric. The F-M is the weighted harmonic mean of the Prec and Rec

$$\text{F-M} = \left(1 + \beta^2\right)(\text{Prec} \cdot \text{Rec})/(\beta^2 \cdot \text{Prec} + \text{Rec}), \tag{10}$$

where β is a measure of the importance of Prec to Rec.

When the complete spectrum of probability thresholds is used to evaluate a probabilistic model, the evaluation metric is a two-dimensional curve. The commonly used two-dimensional evaluation curves for probabilistic classifiers are the Precision–Recall (P–R) curve and the receiver operating characteristic (ROC) curve (Fawcett 2006). P–R and ROC curves provide a measure of the classification performance for the complete spectrum of probability thresholds (that is, operating conditions). The P–R and ROC curves are developed by calculating the Prec, Rec, and the False Positive Rate (FPR) for each threshold from 0 to 1. The FPR is

$$\text{FPR} = \frac{FP}{FP + TN}. \tag{11}$$

Any point on either the P–R or ROC curve corresponds to a specific threshold. Figure 2 presents an idealized ROC curve, where the dashed line is the idealized best possible ROC curve. The AUC is a scalar measure that quantifies the ROC curve in terms of accuracy of the probabilistic classifier. The AUC varies from 1.0 (perfect accuracy) to 0. Randomly selecting a class produces the diagonal line connecting (0, 0)

Fig. 2 Receiver operating characteristic (ROC) curve illustrating its basic elements. The *dashed line* indicates a near perfect probability prediction, whereas the *dotted line* indicates predictions which result from random guessing

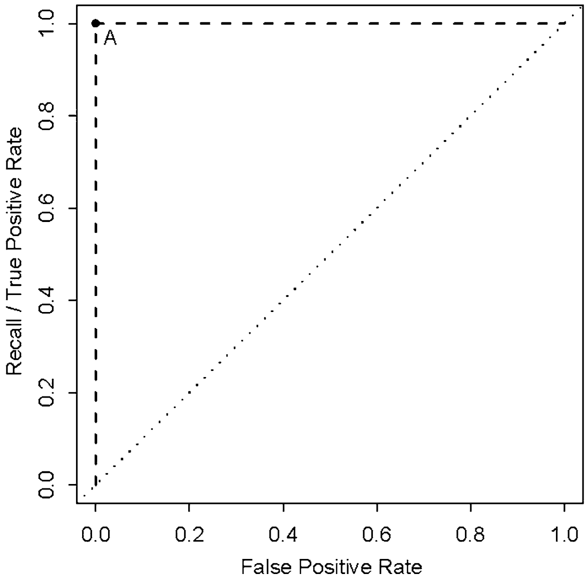
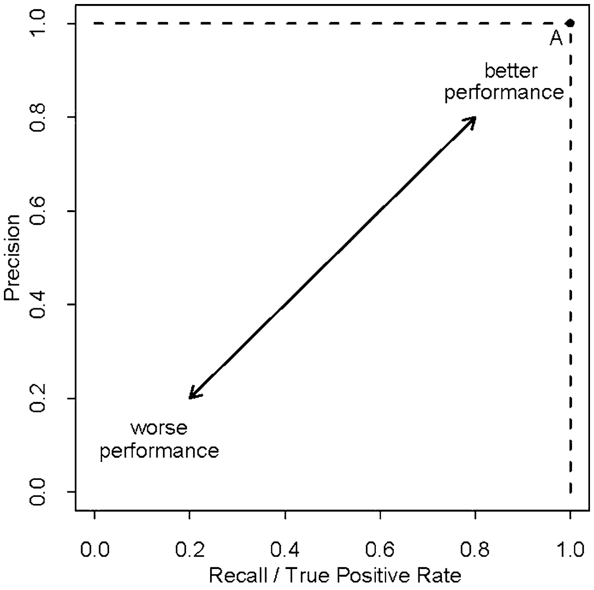


Fig. 3 Precision–recall (P–R) curve illustrating its basic elements. The *dashed line* represents the best P–R curve



and (1, 1) (shown as dotted diagonal line Fig. 2). This gives AUC $= 0.5$, thus it is unrealistic for a classifier to have an AUC less than 0.5.

Figure 3 presents an idealized P–R curve. The dashed line represents the best P–R curve with point A marking the best performance. AU-PRC is a scalar measure that quantifies the P–R curve signifying the predictive performance of the classifier. Unlike ROC curves, P–R curves are sensitive to the influence of class imbalance and sampling bias in a dataset (Oommen et al. 2010).

4 Logistic Regression and Maximum-likelihood Estimation

Logistic regression models a binary response variable. The logistic function is

$$\pi'(X) = \frac{1}{1+e^{-x}}, \tag{12}$$

where X is the input or the predictor variable and $\pi'(X)$ is the estimated output or the estimate of the response variable (Cox 1970). When there are multiple predictor variables, X can be expanded as the total contribution of all the predictor variables, given by

$$X = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{13}$$

where k is the number of predictor variables, α is the intercept, and $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients of the corresponding predictor variables.

The advantage of the logistic function is that it can take any value from $-\infty$ to ∞ as the input, whereas the output is confined to values between 0 and 1. This is achieved by a logit transformation (Cox 1970) of the logistic function, given earlier in (3)

$$\mathrm{loglt}\bigl(\pi'(X)\bigr) = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right). \tag{14}$$

Equations (12) and (14) form the bases of a logistic regression analysis, where $\pi'(x)$ represents a maximum-likelihood estimate of its true value $\pi(X)$, given in (2).

Although the above model used to find X in (13) looks like a simple linear regression, the underlying distribution of the dependent variable of interest y is binomial and the parameters α and $\beta_1, \beta_2, \ldots, \beta_k$ cannot be estimated in the same way as for simple linear regression because there is no normally distributed model error in logistic regression. Instead, the parameters are usually estimated using the method of maximum likelihood (ML). Nearly every software package which implements logistic regression uses the ML method, although other methods have been advanced, including exact logistic regression (Hirji et al. 1987).

5 Results and Discussion

5.1 Case A

In Case A, the two classes (class 0 and class 1) in the original population have a balanced distribution of 50:50. We extract eight random samples from the Case A population with varying class distribution of 50:50, 60:40, 70:30, 80:20, 90:10, 95:5, 98:2, and 99:1, respectively. Each sample is referred with the case number, followed by the sample class distribution in subscripts. For example, a sample with a class distribution 60:40 from Case A is referred as Case $\mathrm{A}_{60:40}$. The sample size for each of the eight samples is determined by fixing the length of the majority class (class 0) at 5000. An MLLR model is applied to each of the eight samples. A plot of the true $\pi(X)$ versus predicted $\pi'(x)$ probability is presented in Fig. 4. We observe from

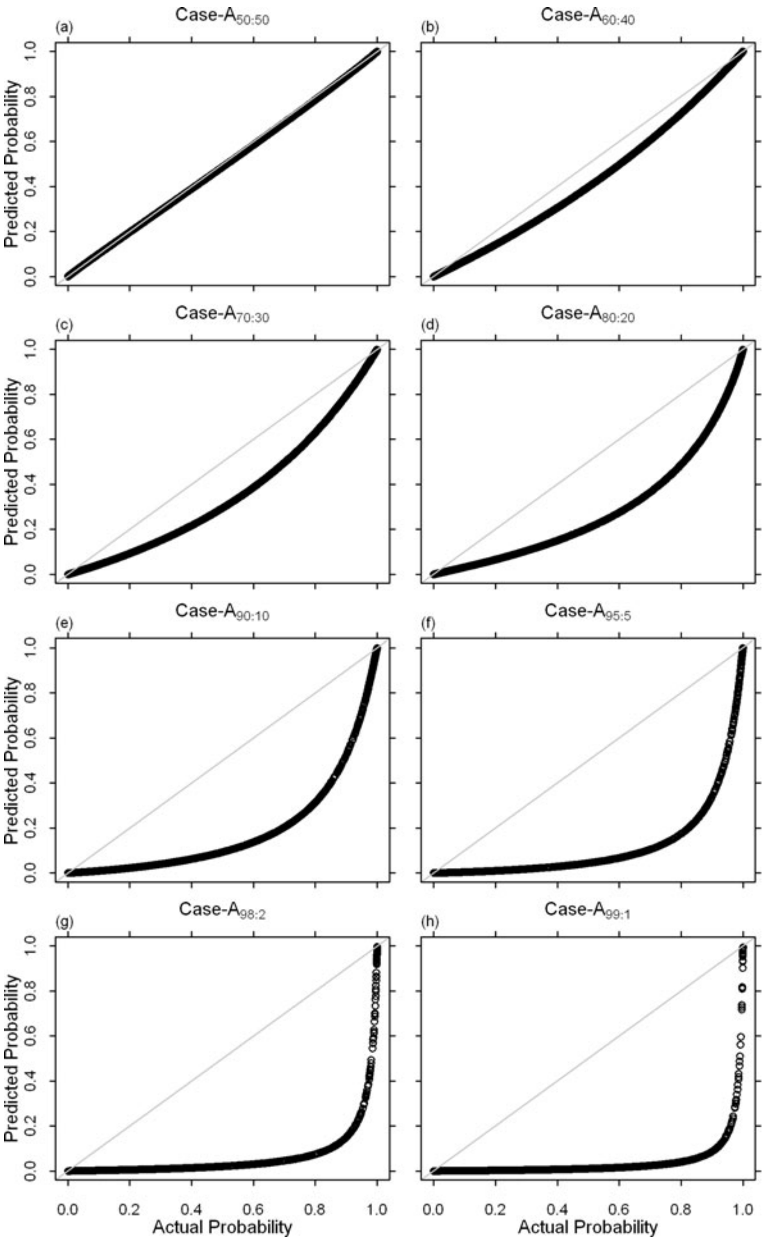


Fig. 4 Scatter plot of the true probability against the predicted probability using MLLR for Case A samples

Fig. 4 that the sample (Case A$_{50:50}$) with no class imbalance and no sampling bias (Fig. 4a) has the best overall predictive performance with the true probabilities being nearly equal to the predicted probabilities using MLLLR. We also observe that as the class imbalance increases from 60:40 to 99:1 resulting in increasing sampling bias and class imbalance (Figs. 4b to 4h); the MLLR model tends to under-predict the probability with the degree of under-prediction being proportional to the sampling bias and class imbalance. We observe that in Case A, the sample (Case $\mathrm{A}_{50:50}$) that has the best model performance (Fig. 4a) does not have a class imbalance (because both class 0 and class 1 have an approximately equal class distribution) or a sampling bias (because the distribution of classes in the sample is approximately equal to the distribution of classes in the population). Therefore, it is not evident from Case A whether both sampling bias and class imbalance are significant for the performance of the model or whether it is either one of them.

We carry out 1000 Monte Carlo simulations for each of the eight samples in Case A. For each simulation, an MLLR model is developed and the model predictive performance is quantified using AUC, AU-PRC, Prec, Rec, and F-M. Figure 5 presents the box plot of these predictive performance measures for each of the eight samples. We observe that as the class imbalance and sampling bias increase from Case $\mathrm{A}_{60:40}$ to Case $\mathrm{A}_{99:1}$ (Figs. 5b to 5h), the performance measures of the minority class (class 1) decrease considerably with the exception of the AUC. However, it is important to note that when there is no class imbalance and no sampling bias (Case $\mathrm{A}_{50:50}$) and when the predicted probability is close to the true probability (Fig. 5a), the mean Prec, Rec, and F-M were similar for both classes (class 0 and class 1), whereas these performance measures substantially differ for other samples (Figs. 5b to 5h).

AUC is similar to a rank sum test and measures the separability of the classes in a dataset. In Fig. 5, we observe that although all other performance measures indicate decrease in separability as class imbalance and sampling bias increase, AUC indicates that the separability is consistent. Therefore, to visualize the separability between the classes in each of the eight samples, we selected one simulation from each sample and plotted its estimated probability using the MLLR model against the number of instances for each class separately (as shown in Fig. 6). We observe from Fig. 6 that as the class imbalance and sampling bias increases from Case $\mathrm{A}_{60:40}$ to Case $\mathrm{A}_{99:1}$ (Figs. 5b to 5h), the separability in the dataset is not significantly different. This indicates that the measure of separability using AUC is more robust over other measures and is not influenced by class imbalance or sampling bias. This also indicates that although the predicted probability is significantly affected due to the increase in sampling bias and class imbalance in the data, the separability between the classes is unaffected. Therefore, the impact of class imbalance and sampling bias can be different for users depending on their objectives. For example, when the objective of the user using the MLLR model is to know whether a site is susceptible to landslide or not, the impact of class imbalance and sampling bias can be minimal; when the objective of the user is to compare between two sites that are susceptible to landslide, the user is interested in the probability of landslide susceptibility, in which case the impact of sampling bias and class imbalance can be significant.

Mehta and Patel (1995) recommend the use of exact logistic regression as an alternative for maximum-likelihood estimates when the data is sparse or unbalanced. However, we tried the exact logistic regression and the results did not improve the predicted probability compared to what was estimated using maximum likelihood.

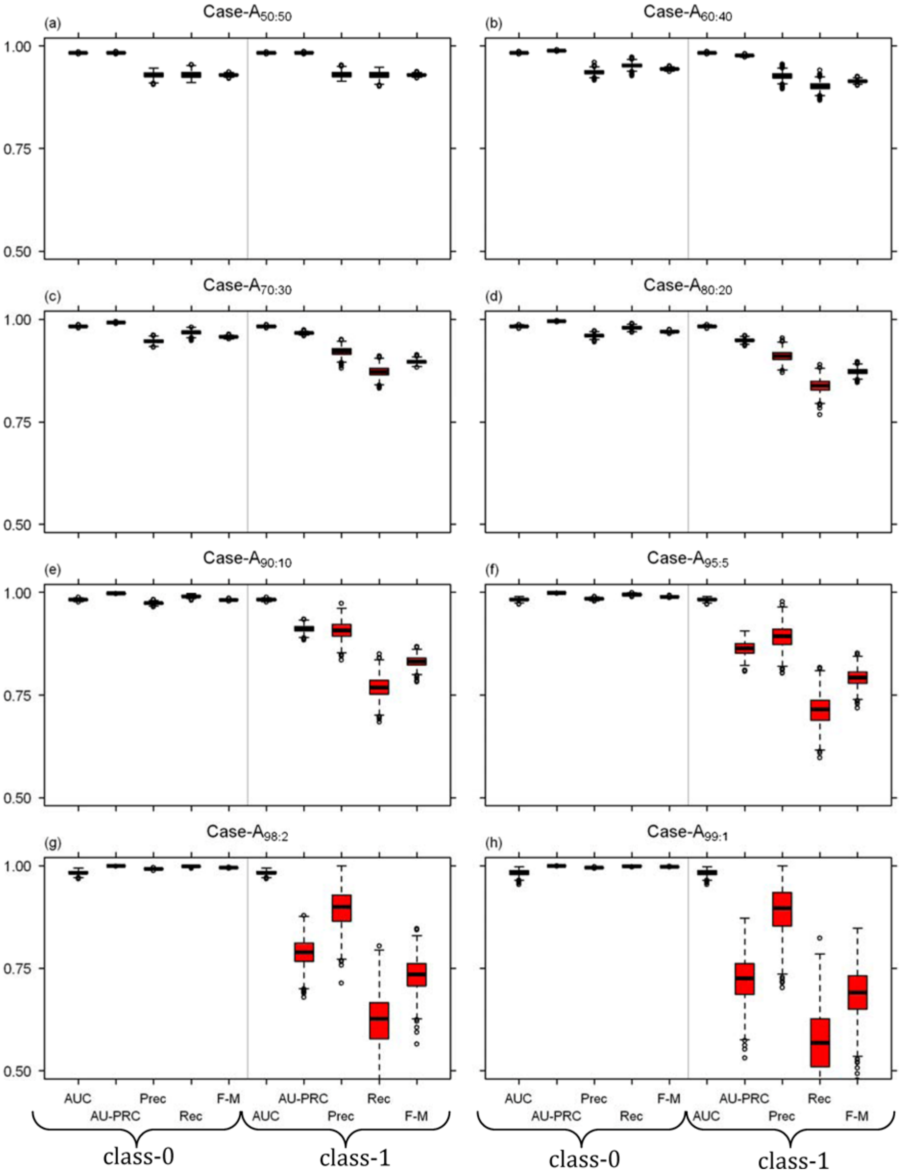


Fig. 5 Box plot of the predictive performance measures (AUC, AU-PRC, Prec, Rec, and F-M) of the MLLR model, computed separately for class 0 and class 1 for the 1000 Monte Carlo simulations of Case A samples

5.2 Case B

In Case B, the two classes (class 0 and class 1) in the original population have a class imbalance of 80:20. Similar to Case A, we extract eight random samples from the Case B population with the ratio of class 0 to class 1 varying from 50:50, 60:40,

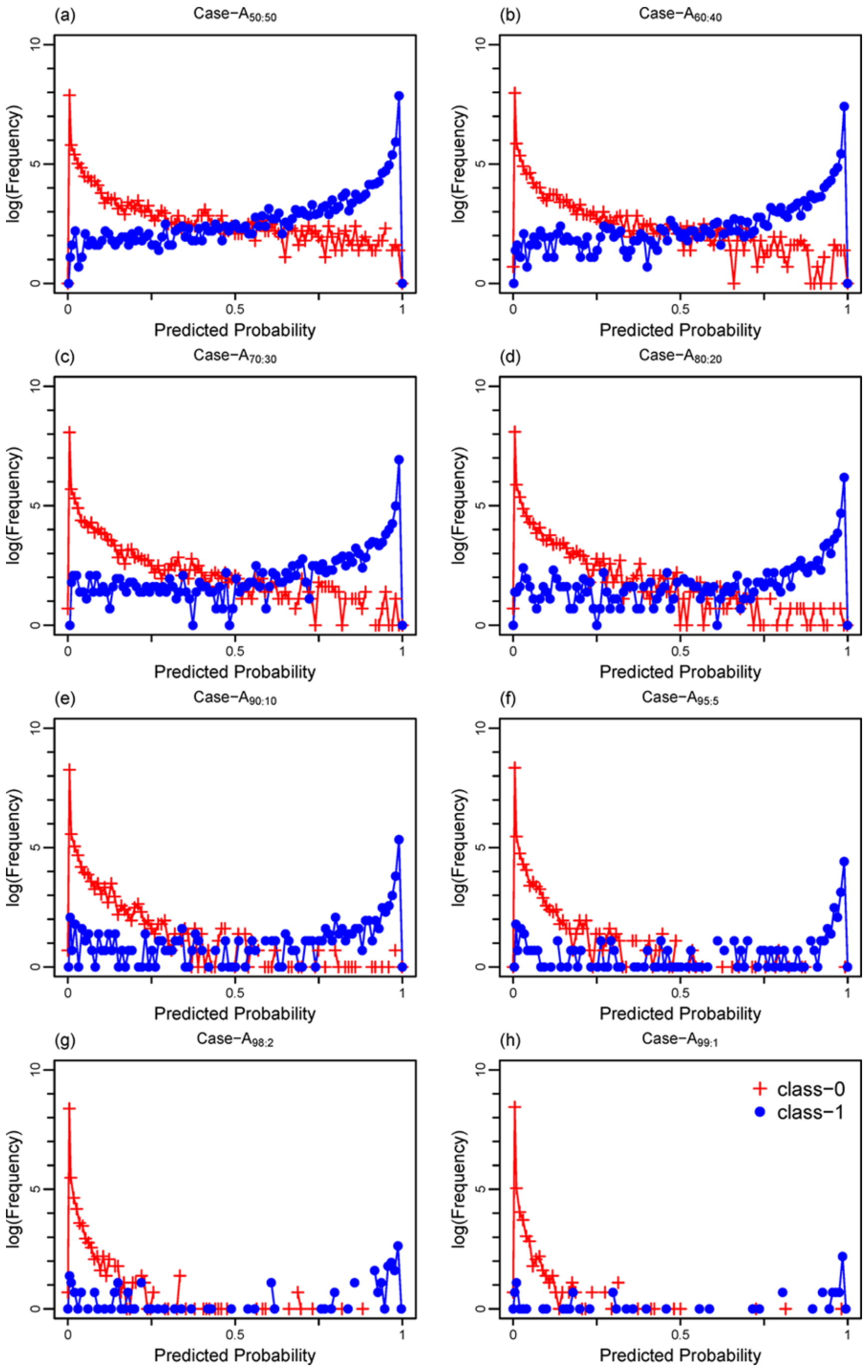


Fig. 6 Plot of the predicted probability against the log of the frequency showing the separability between class 0 and class 1 for the different samples of Case A

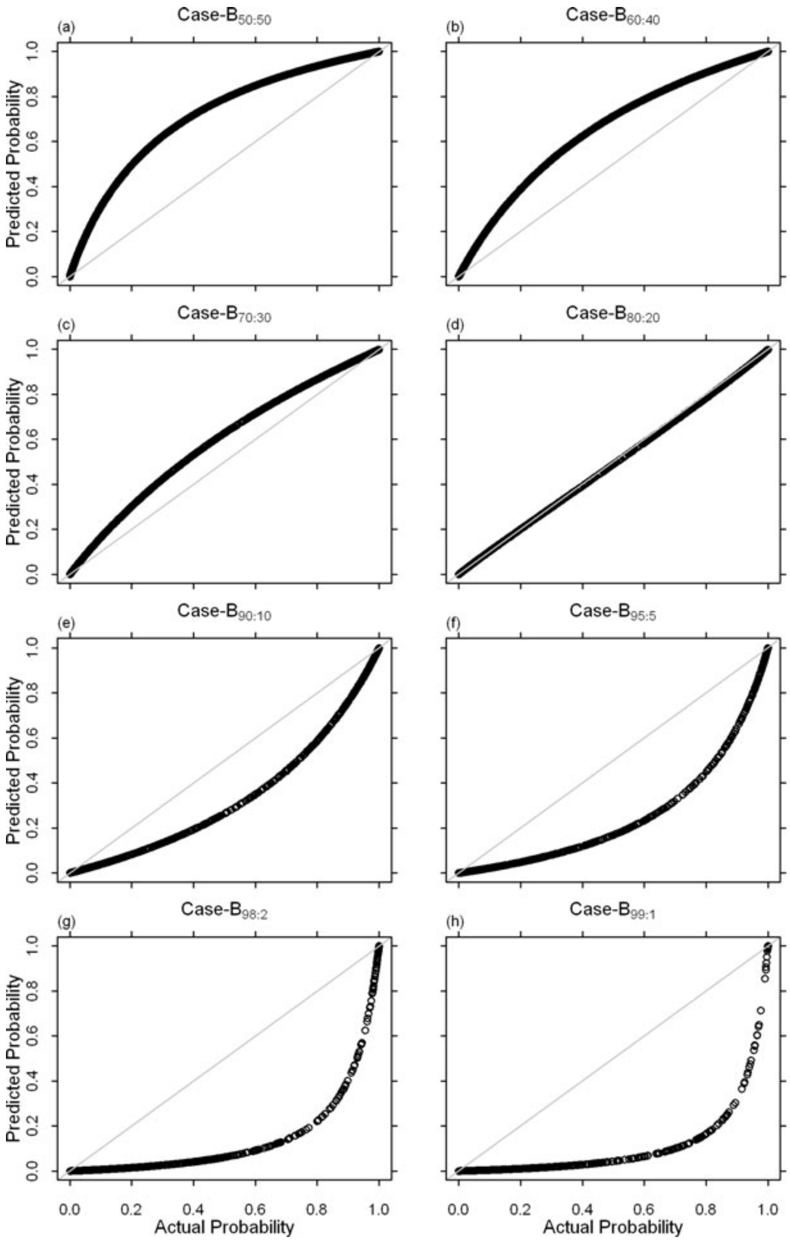


Fig. 7 Scatter plot of the true probability against the predicted probability using MLLR for Case B samples

70:30, 80:20, 90:10, 95:5, 98:2, and 99:1, respectively. A plot of the true probability against the probability that is predicted using the MLLR model developed for each of the eight samples is presented in Fig. 7. Here, it is evident from Fig. 7 that the sample (Case $B_{50:50}$) that has a balanced class (Fig. 7a) no longer performs best, as observed in Case A. Instead, the sample (Case $B_{80:20}$) that has no sampling bias (Fig. 7d) performs the best.

The sample (Case $B_{80:20}$) that performs the best (Fig. 7d) does not have a sampling bias because in that case, the class distribution of the sample (80:20) is equal to the class distribution of the population. We observe once again, that as the sampling bias increases (Figs. 7a to 7c and 7e to 7h), the MLLR model highly under- or over-predicts the probability. When the distribution of the minority class in the sample is more than the distribution in the population, the MLLR model over-predicts the probability, and if the distribution of the minority class in the sample is less than the distribution in the population, the MLLR model under-predicts the probability.

We also carry out 1000 Monte Carlo simulations for each of the eight samples in Case B. Figure 8 presents the box plot of the predictive performance for each of the eight samples using the same evaluation measures reported for Case A. For the case when the percentage of the minority class in the sample is less than its occurrence in the population, and as the sampling bias increases from samples Case $B_{90:10}$ to Case $B_{99:1}$ (Figs. 8e to 8h), the mean Prec, Rec, and F-M for the minority class (class 1) decrease considerably; when the distribution of the minority class in the sample is greater than the distribution in the population (Figs. 8a to 8c), both classes had high mean Prec, Rec, and F-M values, even though the predicted probability in these cases (Case $B_{50:50}$ to Case $B_{70:30}$) are greater than the true probability. Similar to Case A, we observe from Fig. 8 that there is no significant change in the AUC value due to class imbalance or sampling bias. Also similar to Fig. 6, Fig. 9 verifies that in Case B the separability between the classes is unaffected by class imbalance or sampling bias.

Next, we compare the performance measures from the samples that have the best MLLR model from Case A (Fig. 4a) and Case B (Fig. 7d). In Case A, the sample (Case $A_{50:50}$) that performs best has no sampling bias or class imbalance; in Case B, the sample (Case $B_{80:20}$) that performs the best has no sampling bias, but does exhibit some class imbalance. From these comparisons, we conclude that the performance of an MLLR model is more dependent upon the sampling bias than upon the class imbalance. In Case A, for the best performing sample (Case $A_{50:50}$), the mean Prec, Rec, and F-M are similar within and between the classes, whereas in Case B, for the best performing sample (Case $B_{80:20}$), the mean Prec, Rec, and F-M are similar within the classes. This indicates that the difference in the mean Prec, Rec, and F-M within the class and between the classes can be used as an indicator of sampling bias and the resulting inaccuracy of the predicted probabilities using the MLLR model.

In order to analyze the impact of sample size on the performance of the MLLR model, we carried out a sensitivity analysis by varying the majority-class size of the best sample from Case $B_{80:20}$. The majority-class size varied from 5000 to 100 instances. Results indicated that the sample size did not impact the predictive performance significantly. However, if the sample size is considerably small, it could impact the predictive performance of the MLLR model.

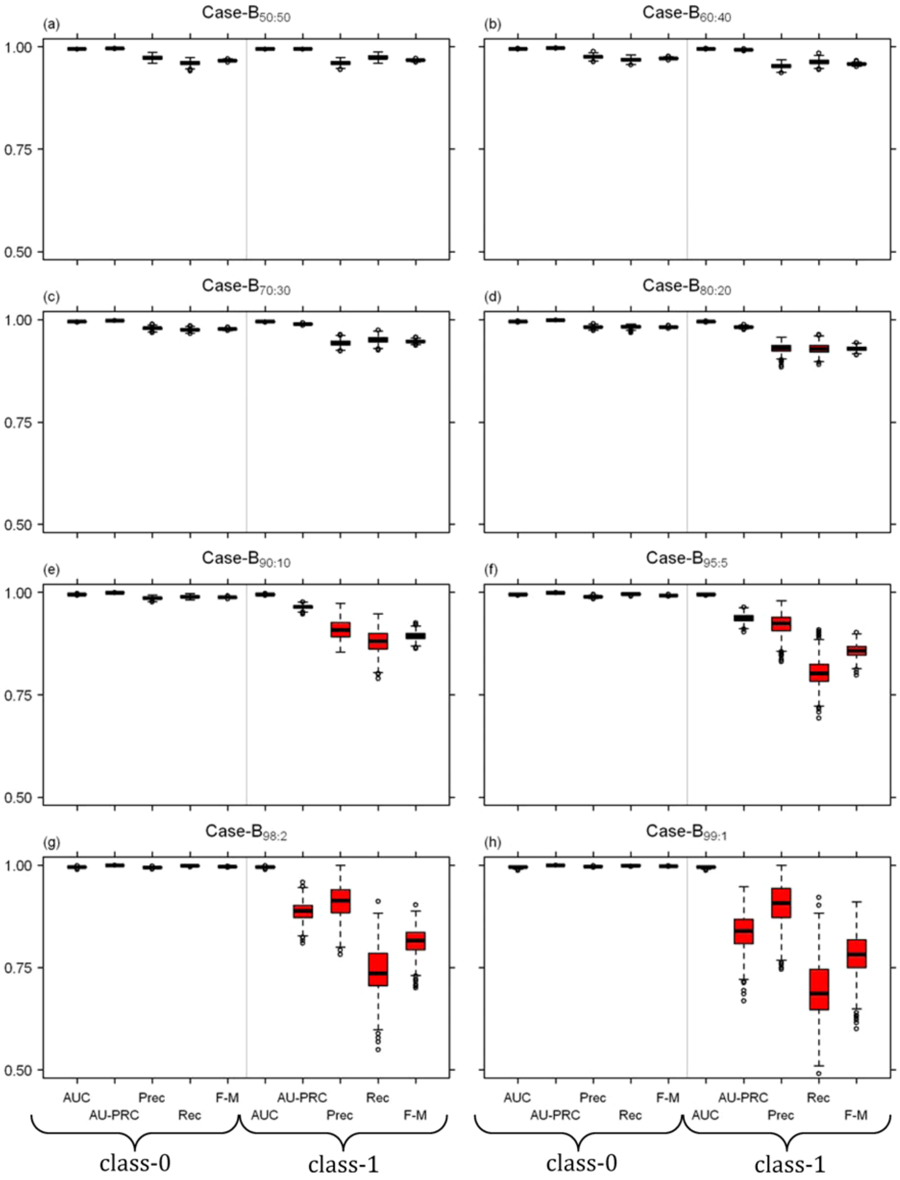


Fig. 8 Box plot of the predictive performance measures (AUC, AU-PRC, Prec, Rec, and F-M) of the MLLR model computed separately for class 0 and class 1 for the 1000 Monte Carlo simulations of Case B samples

6 Improving MLLR Models Using Sampling Techniques

We have tried to improve the performance of the worst performing samples of Case A (Case A$_{99:1}$) and Case B (Case B$_{99:1}$) by using common sampling techniques: under- and over-sampling. The worst performing sample of Case A (Fig. 4h) is further

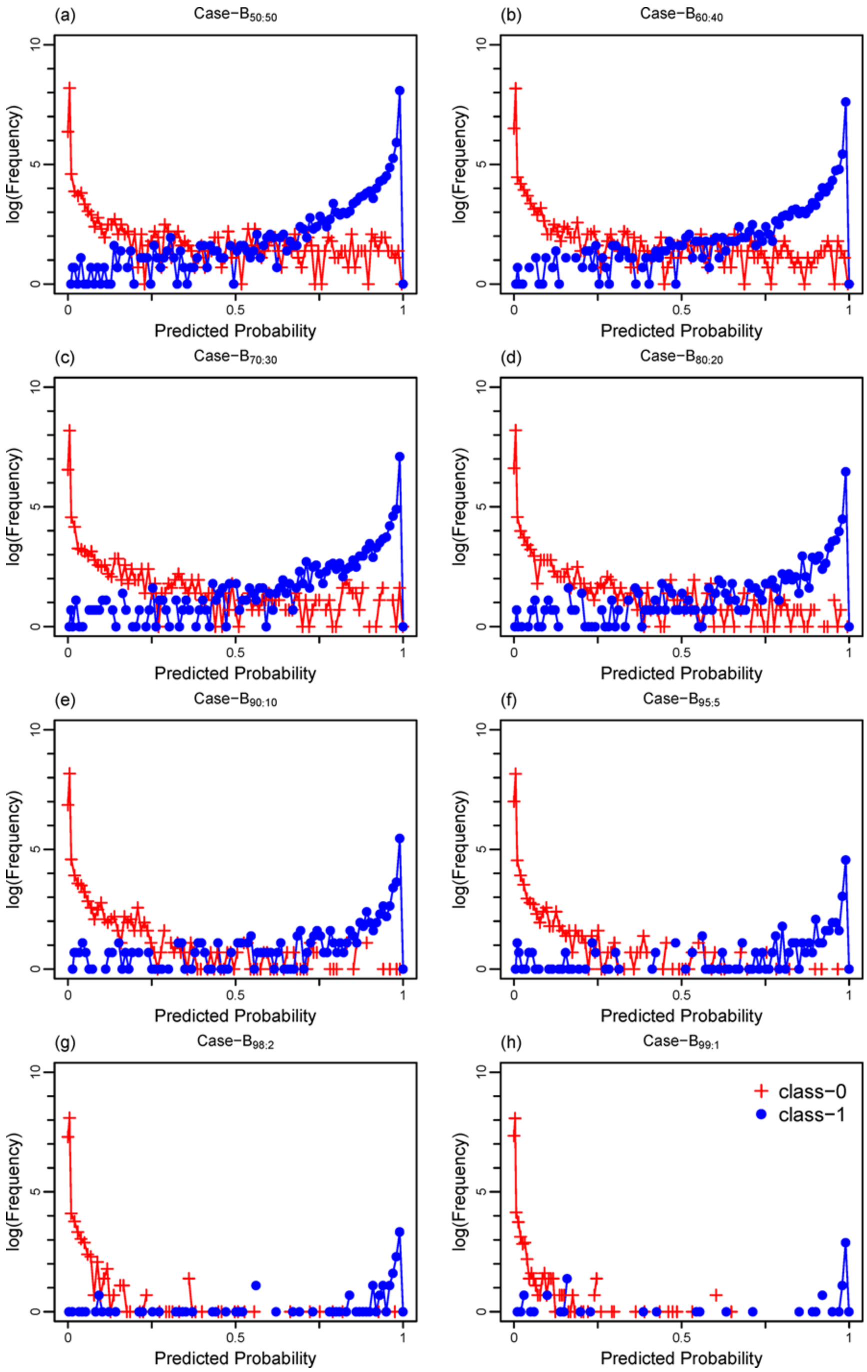


Fig. 9 Plot of the predicted probability against the log of the frequency showing the separability between class 0 and class 1 for the different samples of Case B

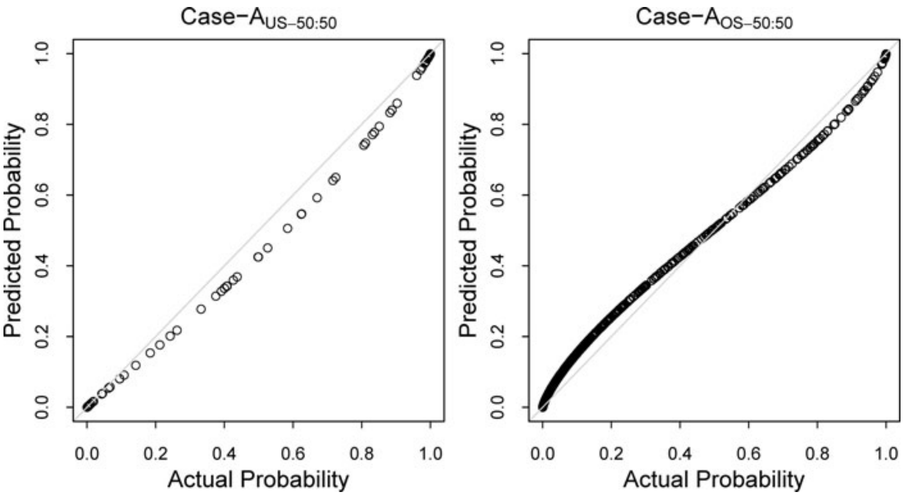


Fig. 10 Scatter plot of the true probability against the predicted probability using MLLR for the over-sampled (Case $A_{OS\text{-}50:50}$) and under-sampled (Case $A_{US\text{-}50:50}$) samples, compared to the original sample (Case $A_{99:1}$) (Fig. 4h, class distribution of 99:1)

Table 2 Mean predictive performance of the MLLR model for the simulations of under-sampled (Case $A_{US\text{-}50:50}$) and over-sampled (Case $A_{OS\text{-}50:50}$) samples, compared to the worst performing sample (Case $A_{99:1}$) of Case A

Case	Class 0					Class 1				
	AUC	AU-PRC	Prec	Rec	F-M	AUC	AU-PRC	Prec	Rec	F-M
$A_{99:1}$	0.99	0.99	0.99	0.99	0.99	0.99	0.83	0.90	0.69	0.77
$A_{US\text{-}50:50}$	0.98	0.98	0.93	0.95	0.94	0.98	0.98	0.95	0.93	0.94
$A_{OS\text{-}50:50}$	0.99	0.99	0.98	0.98	0.98	0.99	0.97	0.92	0.93	0.93

sampled to derive two different samples of which one is over-sampled and the other is under-sampled. The resulting samples have a class distribution of 50:50. The sample that is over-sampled is referred to as Case $A_{OS\text{-}50:50}$ and the sample that is under-sampled is referred to as Case $A_{US\text{-}50:50}$, respectively. The class distribution of 50:50 for the sample from Case A represents a balanced class with no sampling bias. A plot of the true probability against the probability predicted using an MLLR model for both the over- and under-sampled samples is presented in Fig. 10. It is evident from Fig. 10 that both over- and under-sampling improved the predictive capability of the sample compared to the original sample (Case $A_{99:1}$) (Fig. 4h, class distribution of 99:1).

We carried out 1000 Monte Carlo simulations of over- and under-sampling, and developed an MLLR model for each simulation. In Table 2, we present the comparison of the mean predictive performance of the MLLR model for the simulation compared to the predictive performance of the MLLR model in the original sample. The mean predictive performance measures for the under- and over-sampled samples have improved considerably over the original (biased and imbalanced) sample, es-

Table 3 Mean predictive performance of the MLLR model for the simulations of under-sampled (Case $B_{US\text{-}50:50}$ and Case $B_{US\text{-}80:20}$) and over-sampled (Case $B_{OS\text{-}50:50}$ and Case $B_{OS\text{-}80:20}$) samples, compared to the worst performing sample (Case $B_{99:1}$) of Case B

Case	Class 0					Class 1				
	AUC	AU-PRC	Prec	Rec	F-M	AUC	AU-PRC	Prec	Rec	F-M
$B_{99:1}$	0.99	0.99	0.99	0.99	0.99	0.99	0.83	0.90	0.69	0.77
$B_{US\text{-}50:50}$	0.99	0.99	0.97	0.97	0.97	0.99	0.99	0.97	0.97	0.97
$B_{OS\text{-}50:50}$	0.99	0.99	0.98	0.96	0.97	0.99	0.99	0.96	0.98	0.97
$B_{US\text{-}80:20}$	0.99	0.99	0.98	0.98	0.98	0.99	0.98	0.95	0.93	0.94
$B_{OS\text{-}80:20}$	0.99	0.99	0.98	0.98	0.98	0.99	0.97	0.92	0.93	0.93

pecially for class 1. The comparison of the mean predictive performance measures for the under- and over-sampled cases reveals that they both have similar predictive performance.

In Case B, the worst performing sample (Case $B_{99:1}$) is further sampled to derive four different samples of which two are over-sampled and the other two are under-sampled. The resulting samples have class distributions of 50:50 and 80:20. The samples that are over-sampled are referred to as Case $B_{OS\text{-}50:50}$ and Case $B_{OS\text{-}80:20}$, respectively, and the samples that are under-sampled are referred to as Case $B_{US\text{-}50:50}$ and Case $B_{US\text{-}80:20}$, respectively. Figure 11 compares the true probability with the predicted probability for both the over-sampled and under-sampled cases for class distributions of 50:50 and 80:20. Here we observe for the case of over- and under-sampling that the sample with no sampling bias (Case $B_{OS\text{-}80:20}$ and Case $B_{US\text{-}80:20}$) results in predictions which outperform the model based on the sample with no class imbalance (Case $B_{OS\text{-}50:50}$ and Case $B_{US\text{-}50:50}$). As in Case A, a comparison of the model performance based on over- with under-sampling indicates that the true and predicted probability are in better agreement for the case of over-sampling than for the case of under-sampling.

Similar to Case A, we have carried out 1000 Monte Carlo simulations of over- and under-sampling for both class distributions in Case B and developed an MLLR model for each simulation. In Table 3, we present the comparison of the mean predictive performance of the MLLR model compared to the predictive performance of the MLLR model based on the original sample. The mean predictive performance for both class distributions (50:50 and 80:20) and both sampling techniques (under- and over-sampling) demonstrates considerable improvement over the original sample. It is observed from Table 3 that the difference between the mean Prec, Rec, and F-M is minimal within and between the classes in the case of balanced samples (Case $B_{OS\text{-}50:50}$ and Case $B_{US\text{-}50:50}$), whereas it is minimal within the classes when there is no sampling bias (Case $B_{OS\text{-}80:20}$ and Case $B_{US\text{-}80:20}$). This indicates that the difference in mean Prec, Rec, and F-M can be used as an indicator of sampling bias even when the samples are re-sampled using under- or over-sampling. Our results, which indicate that over- and under-sampling can improve predictive performance, are consistent with previous results in the literature but provide more guidance on when over- or under-sampling should be used and what the sampling ratio should be. Cosslett (1981b), Imbens (1992), and King and Zeng (2001) have recommended

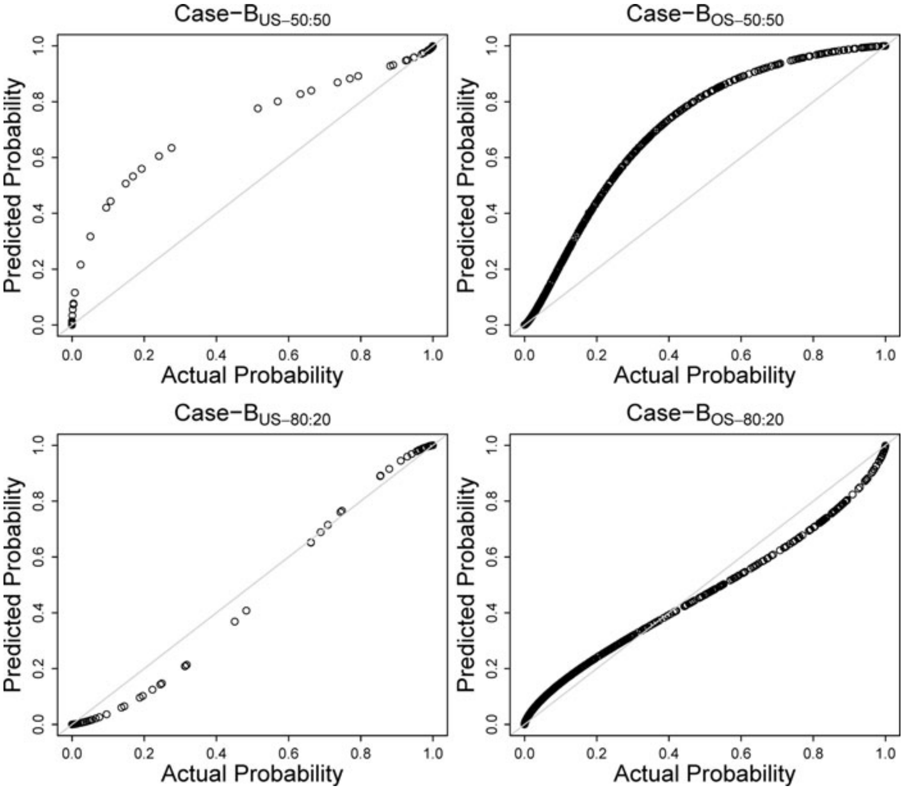


Fig. 11 Scatter plot of the true probability against the predicted probability using MLLR for the over-sampled (Case $\mathrm{B}_{\text{OS-50:50}}$ and Case $\mathrm{B}_{\text{OS-80:20}}$) and under-sampled (Case $\mathrm{B}_{\text{US-50:50}}$ and Case $\mathrm{B}_{\text{US-80:20}}$) samples from the worst performing sample (Case $\mathrm{B}_{99:1}$) of Case B

correction and weighting methods for the estimation of the maximum-likelihood estimates to correct for the sampling bias. However, the proposed methods require the a priori knowledge of class distribution of the population and are not always available (King and Zeng 2001).

In order to improve the performance of an MLLR model when a priori knowledge of class distribution is not known, King and Zeng (2001) recommended a sequential practice involving the equal collection of both classes, that is, class 0 (frequent event) and class 1 (rare event), and verify if the standard errors and confidence intervals are narrow enough to stop. Otherwise, they suggest continuing sampling class 0 randomly until the confidence intervals get sufficiently small for the substantive purpose at hand. In effect, by this sequential practice, the user is reducing the sampling bias between the sample and the population. The results from our study support the recommendation of King and Zeng (2001) when dealing with samples for which the class distribution in the population is unknown. In fields such as geosciences, geohazards, and medical diagnosis, the distribution of the classes in the population is unknown a priori. However, we recommend that in fields where the distribution of the classes in the population is known a priori, the user should choose a sample that has the same distribution as the population to ensure optimal performance of the MLLR model.

7 Conclusions and Future Work

We have analyzed the influence of class imbalance and sampling bias on the performance of the asymptotic MLLR classifier. We have generated two synthetic datasets with class distributions of 50:50 and 80:20. Here we have generated datasets with completely known properties from an underlying logistic regression model for which we know the true model parameters. We have performed a set of simulations that extract several samples from these datasets that have different class distributions and have compared the resulting MLLR models, which were fit to each of the samples in terms of various commonly used statistical performance measures which are all based on the confusion matrix.

The following specific conclusions arise from our performance evaluation of the resulting fitted MLLR models:

- The predicted probability using an MLLR model is closest to the true probability when the sample has the same class distribution as the original population. Therefore, in probabilistic modeling using MLLR, it is important to develop a sample that has the same class distribution as the original population rather than ensuring that the classes are equally sampled. In summary, it is critical that the user of MLLR attempts to limit sampling bias.
- When the objective of the MLLR model is only to separate between the classes in the dataset, neither sampling bias nor class imbalance are significant, that is, the separability of the data is not affected by the difference in the true and predicted probabilities. However, when the user is interested in comparing predictions within a class, it is important that the sample has minimal sampling bias to ensure that the difference in the true and predicted probabilities will be minimal.
- AUC is a widely used measure of the predictive performance of probabilistic classifiers, which ranges from 0.5 (random performance) to 1 (perfect performance). Independent of the class imbalance and sampling bias in the data, AUC measures the separability between the classes.
- For the evaluation of probabilistic models, we recommend the use of AUC to evaluate the separability between the classes. The difference between the Prec, Rec, and F-M within a class can be used as an indicator of the sampling bias as well as the difference in the true and predicted probabilities.
- When the sampling bias is reduced using basic sampling techniques, both over- and under-sampling will tend to improve the predictive capability of the model with over-sampling providing slightly improved estimates of the probability of interest. While the use of over- and under-sampling is not new, our results provide guidance in using these sampling methods. The goal of sampling should be to mimic the population class ratio in the sample.

In this study, using two synthetic datasets from a model with known parameters, we demonstrate that sampling bias limits the predictive performance of MLLR, an asymptotic classifier. Our future work will focus on how to identify the correct class distribution of the original population from a sample. This information is needed to correct for sampling bias and thus derive better estimates of probability using MLLR classifiers.

References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley series in probability and statistics. Wiley, New York

Agterberg FP (1974) Automatic contouring of geological maps to detect target areas for mineral exploration. Math Geol 6:373–395

Atkinson PM, Massari R (1998) Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy. Comput Geosci 24:373–385

Bent GC, Steeves PA (2006) A revised logistic regression equation and an automated procedure for mapping the probability of a stream flowing perennially in Massachusetts. US Geological Survey Scientific Investigations Report 2006-5031, 1 CD-ROM

Bonham-Carter GF, Chung CF (1989) Integration of mineral resource data for Kasmere lake area, Northwest Manitoba, with emphasis on uranium. Comput Geosci 15(1):25–45

Boyacioglu MA, Kara Y, Baykan OK (2009) Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. Expert Syst Appl 36:3355–3366

Burez J, Van den Poel D (2008) Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. Expert Syst Appl 35:497–514

Cao K, Yang X, Tian J, Zhang YY, Li P, Tao XQ (2009) Fingerprint matching based on neighboring information and penalized logistic regression. Adv Biom 5558:617–626

Carrara (1983) Multivariate models for landslide hazard evaluation. Math Geol 15(3):403–426

Caumon G, Ortiz JM, Rabeau O (2006) Comparative study of three data-driven mineral potential mapping techniques. In: Int assoc for mathematical geology, XI$^{\text{th}}$ international congress, Belgium, S13-05

Chung CF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30:451–472

Correia LCL, Rocha MS, Esteves JP (2009) HDL-cholesterol level provides additional prognosis in acute coronary syndromes. Int J Cardiol 136:307–14

Cosslett SR (1981a) Maximum-likelihood estimator for choice-based samples. Econometrica 49:1289–1316

Cosslett SR (1981b) Efficient estimation of discrete-choice models. MIT Press, Cambridge

Cox DR (1970) Analysis of binary data. Methuen, London

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874

Garcia V, Mollineda RA, Sanchez JS (2008) On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Appl 11:269–280

Gu Q, Cai ZH, Zhu L, Huang B (2008) Data mining on imbalanced data sets. In: International conference on advanced computer theory and engineering, pp 1020–1024

Hirji KF, Mehta CR, Patel NR (1987) Computing distributions for exact logistic regression. J Am Stat Assoc 82:1110–1117

Imbens GW (1992) An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60:1187–1214

Juang CH, Chen CJ, Jiang T (2001) Probabilistic framework for liquefaction potential by shear wave velocity. J Geotech Geoenviron Eng 127:670–678

Juang CH, Jiang T, Andrus RD (2002) Assessing probability-based methods for liquefaction potential evaluation. J Geotech Geoenviron Eng 128:580–589

King G, Zeng L (2001) Explaining rare events in international relations. Int Organ 55:693–715

Lai SY, Chang WJ, Lin PS (2006) Logistic regression model for evaluating soil liquefaction probability using CPT data. J Geotech Geoenviron Eng 132:694–704

Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B, Cybern 39:539–50

Lopez L, Sanchez JL (2009) Discriminant methods for radar detection of hail. In: 4th European conference on severe storms, vol 93, pp 358–368

Mehta CR, Patel NR (1995) Exact logistic regression: Theory and examples. Stat Med 14:2143–2160

Moss RES, Seed RB, Kayen RE, Stewart JP, Kiureghian AD, Cetin KO (2006) CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. J Geotech Geoenviron Eng 132(8):1032–1051

Olson SA, Brouillette MC (2006) A logistic regression equation for estimating the probability of a stream in Vermont having intermittent flow: US Geological Survey Scientific Investigations Report 2006–5217

Oommen T, Baise LG, Vogel R (2010) Validation and application of empirical liquefaction models. J Geotech Geoenviron Eng. doi:10.1061/(ASCE)GT.1943-5606.0000395

Page RL, Ellison CG, Lee J (2009) Does religiosity affect health risk behaviors in pregnant and postpartum women? Matern Child Health J 13:621–632

Preisler HK, Brillinger DR, Burgan RE, Benoit JW (2004) Probability based models for estimation of wildfire risk. Int J Wildland Fire 13:133–142

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna

Seiffert C, Khoshgoftaar TM, Van Hulse J (2009) Hybrid sampling for imbalanced data. Integr Comput -Aided Eng 16:193–210

Sun YM, Wong AKC, Kamel MS (2009) Classification of imbalanced data: A review. Int J Pattern Recognit Artif Intell 23:687–719

Tang YC, Zhang YQ, Chawla NV, Krasser S (2009) SVMs modeling for highly imbalanced cClassification. IEEE Trans Syst Man Cybern Part B, Cybern 39:281–288

Tasker GD (1989) Regionalization of low flow characteristics using logistic and GLS regression. In: Kavvas ML (ed) New directions for surface water modeling. IAHS Publication, vol 181, pp 323–331

Toner M, Keddy P (1997) River hydrology and riparian wetlands: A predictive model for ecological assembly. Ecol Appl 7:236–246

van Rijsbergen C (1979) Information retrieval. Butterworths, London

Weiss GM, Provost F (2003) Learning when training data are costly: The effect of class distribution on tree induction. J Artif Intell Res 19:315–354

Williams DP, Myers V, Silvious MS (2009) Mine classification with imbalanced data. IEEE Geosci Remote Sens Lett 6:528–532