Addressing Class Imbalance for Logistic Regression

a thesis presented for the degree of

Doctor of Philosophy of Imperial College London

and the

Diploma of Imperial College

by

Yazhe Li

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2AZ

June 2020

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: _____

Copyright

Abstract

The challenge of class imbalance arises in classification problem when the minority class is observed much less than the majority class. This characteristic is endemic in many domains. Work by Owen [2007] has shown that, in a theoretical context related to infinite imbalance, logistic regression behaves such that all data in the rare class can be replaced by their mean vector to achieve the same coefficient estimates. Such results suggest that cluster structure among the minority class may be a specific problem in highly imbalanced logistic regression. In this thesis, we focus on highly imbalanced logistic regression and develop mitigation methods and diagnostic tools.

Theoretically, we extend the Owen [2007] results to show the phenomenon remains true for both weighted and penalized likelihood methods in the infinitely imbalanced regime, which suggests these alternative choices to logistic regression are not enough for highly imbalanced logistic regression.

For mitigation methods, we propose a novel relabeling solution based on relabeling the minority class to handle imbalance problem when using logistic regression, which essentially assigns new labels to the minority class observations. Two algorithms (the Genetic algorithm and the Expectation Maximization algorithm) are formalized to serve as tools for computing this relabeling. In simulation and real data experiments, we show that logistic regression is not able to provide the best out-of-sample predictive performance, and our relabeling approach that can capture underlying structure in the minority class is often superior.

For diagnostic tools to detect highly imbalanced logistic regression, different hypothesis testing methods, along with a graphical tool are proposed, based on the mathematical insights about highly imbalanced logistic regression. Simulation studies provide evidence that combining our diagnostic tools with mitigation methods as a systematic strategy has the potential to alleviate the class imbalance problem among logistic regression.

To my family and Yuxin.

Acknowledgments

First, I want to express my sincere gratitude to my supervisors Professor Niall Adams and Professor Tony Bellotti, for their expertise, motivation, and patience. Undertaking this Ph.D. has been a truly life-changing experience for me, and it would not have been possible to do without the support and guidance that I received from them during these past four years. I appreciate their more than the generous contribution of time, insightful suggestions, and continuous encouragement throughout my research. I would also like to thank Dr. Heather Battey and Professor Nick Heard for insightful discussions for my research.

Many thanks to my friends Dr. Can Gao, Xixi Yu, Pedro Bustamante Munguira, Dr. Papaioannou Georgiosmy, office mates in room 537, and many other friends for all the useful recommendations, wonderful talks, and encouragement throughout my life in London. My thanks also go out to my friends in California and Boston (too many to list here but you know who you are!), without your support and friendship, I could not start my journey in Statistics.

A very special thank you to Yuxin Fu for always believing in me and encouraging me to follow my dreams during this challenging period. Yuxin has been a true and great supporter and has unconditionally loved me during my good and bad times. There are no words to convey how much I love her.

In the end, I am grateful to my parents and family members for their unconditional support. I consider myself nothing without them.

Contents

List of Notations

N ..... the number of the majority class observations

n ..... the number of the minority class observations

M ..... refers to the size of a data set, usually $M = N + n$

β_0 ..... the intercept term of logistic regression

$\widehat{\beta}_0$ ..... the estimate of β_0

β ..... the slope vector of logistic regression

$\widehat{\beta}$ ..... the estimate of β

μ, $\boldsymbol{\mu}$ ..... the mean or the mean vector

σ, $\boldsymbol{\Sigma}$ ..... the variance or the covariance matrix

X ..... a random variable

x, $\mathbf{x}$ ..... a realization of X, scalar or vector

List of Acronyms

MLE ..... **M**aximum **L**ikelihood **E**stimation

EM ..... **E**xpectation **M**aximization algorithm

GA ..... **G**enetic **A**lgorithm

AUC ..... **A**rea **U**nder receiver operating characteristic **C**urve

1

Introduction

Classification problems can be defined as identifying the class which a new observation belongs to, based on learning from the data where the class information is known. High *class imbalance* refers to one or some classes that are extremely rare in the classification problem. Modeling imbalanced data is a challenging problem, which is the primary concern of this thesis. In the real world, the rare (minority) class usually links to a concept with higher interest; for example, the conflict between countries in political science [King and Zeng, 2001a], credit card transaction fraud [Brause et al., 1999] and default in the consumer credit risk industry (as we discuss below).

Credit providers use statistical models to evaluate the credit risk of lending to consumers, typically by constructing a classification rule to distinguish *good* and *bad* risk customers. Customers' classes are defined by their propensity to default, that is, to fail in satisfying their repayment obligations. This process is known as *credit scoring*. The most popular approach for consumer credit risk modeling is logistic regression [Thomas, 2009, p. 79], which is regarded as a benchmark in the financial industry. There are several unsolved problems in retail credit scoring, including reject inference [Hand, 1998], handling variation over time, and class imbalance. The latter is a common

problem in credit scoring in which one of the two credit-risk classes is much less frequent. Because logistic regression is the main tool for credit scoring; it motivates us to concentrate on the application of logistic regression and its related methods in highly imbalanced data sets.

Owen [2007] provides a striking asymptotic result, which suggests that, in cases of extreme class imbalance, the minority class only contributes to the logistic regression estimation via its sample mean vector. This deep mathematical insight raises concerns about the utility of such models, and potential unwanted consequences, especially when cluster structure emerges among the minority class. Throughout the thesis, we break down the highly imbalanced logistic regression problem into three sub-problems:

1. How widely used modifications to logistic regression (e.g. weighted or penalized logistic regression) perform in the highly imbalanced data?

2. Based on the theoretical results, can we propose some mitigation methods for highly imbalanced logistic regression?

3. How to identify that a problem exhibits high class imbalance with respect to logistic regression?

We propose a systematic approach (theory, detection, and alleviation) to handle highly imbalanced logistic regression by addressing these problems.

1.1 Research Contributions

This thesis aims to contribute towards the theory of highly imbalanced logistic regression and corresponding mitigation methodology. The main contributions of this thesis can be summarized as follows:

- Two natural choices to alleviate the class imbalance problem are penalizing and weighting the likelihood [Wang et al., 2015, King and Zeng, 2001b]. However, we show, by extending Owen [2007] result, that penalizing and weighting the likelihood are insufficient for handling the class imbalance problem. In fact, penalized logistic regression makes matters worse. This is part of our published paper [Li et al., 2019].

- We present two relabeling procedures that attempt to handle the class imbalance problem. Essentially, these procedures seek to partition the minority class into several new pseudo-classes and relabel them to improve the predictive performance of the model. They have different computational efficiency. A cross validation procedure is proposed for selecting the unobservable number of the pseudo-classes. These procedures are shown to be effective in a simulation study and real data. This material is collected in a paper under review [Li et al., 2020].

- Several diagnostic tools are proposed to detect highly imbalanced logistic regression problems. They focus on different aspects of the model, i.e. parameters, likelihood, and prediction. We explore their performance in different sample sizes by simulation and with real data.

1.2 Thesis Structure

Chapter 2 reviews the basics of some statistical concepts and methods, including probability theory, classification methods, performance assessment, and clustering methods. We also briefly review the high class imbalance problem with its mitigation methods.

In Chapter 3, for a binary classification problem, we explore the limit behavior of logistic regression as the number of the majority class cases tends to infinity while the number of minority class cases remains fixed (i.e. infinitely imbalanced logistic regression [Owen, 2007]). We provide the background to infinitely imbalanced logistic regression. Then we consider methods that extend logistic regression in the highly imbalanced data. New theorems are given for infinitely imbalanced weighted logistic regression and penalized logistic regression. These results explain why they are not attractive as mitigation methods for highly imbalanced logistic regression. We also give the theory of infinitely imbalanced multinomial logistic regression as a preparation for the mitigation methods proposed in the Chapter 4. Part of this chapter is from our paper [Li et al., 2019].

In Chapter 4, we introduce our relabeling approach. A brute force method

(Genetic Algorithm) and a computationally efficient method (Expectation Maximization Algorithm) are proposed, which seeks to relabel the minority class into several new distinct pseudo-classes. These algorithms are inspired by the theoretical results from the previous chapter. We demonstrate the performance of our methods with a simulation study and multiple real data sets. Some contents in this chapter are from a paper which is currently under review.

In Chapter 5, three hypothesis testing methods, along with a visualization tool, are proposed for detecting highly imbalanced logistic regression problems in light of the deeper mathematical insights from Owen [2007]. Again, we demonstrate our diagnostic tools through a simulation study and multiple real data sets. The results provide evidence that combing the diagnostic tools and our relabeling approach as a systematic strategy may alleviate the class imbalance for logistic regression.

We conclude our research and present the possible future work directions in Chapter 6.

2

Background

In this chapter, we outline some important statistical concepts and methods which are frequently used throughout the thesis. This background is essential for the high class imbalance theory we developed in Chapter 3 and the proposed relabeling approach in Chapter 4. In Section 2.1, we introduce the basic concepts of probability theory and address the Gaussian distribution, which will be used in Chapter 3. Four frequently used classification methods are discussed in Section 2.2. The high class imbalance problem and cutting edge mitigation methods are reviewed in Section 2.3. The topics in Sections 2.2 and 2.3 are the key concerns of this thesis, which will be further developed in Chapters 3, 4 and 5. Then, Section 2.4 defines some performance criteria for classification and introduces the relevant measurement methods. In addition, we introduce some findings of the area under receiver operating characteristic curve's (AUC) characteristics in the highly imbalanced data. In Section 2.5, a clustering method is discussed which will be used in Chapter 4.

2.1 Probability Theory

This section describes some basic concepts and definitions in probability theory. Section 2.1.1 describes the definition of the probability distribution function, expectation and variance, which are frequently used in the proofs in Chapter 3. The Gaussian distribution is introduced in Section 2.1.2, which is widely used in Section 3.3 for the illustration. The standard definitions in this section follow those in DeGroot and Schervish [2012].

2.1.1 Probability Distribution, Expectation, Variance and Covariance

For a continuous random variable X, if there is a non-negative function f, such that for every interval $\mathbf{I}$, the probability of X drawn from $\mathbf{I}$ equals to the integration of f over $\mathbf{I}$, i.e.

$$\Pr(X \in \mathbf{I}) = \int_{x \in \mathbf{I}} f(x)dx, \tag{2.1}$$

then the function f is called the probability density function (pdf).

The expectation and the variance of the distribution f are defined as follow:

- the expectation (mean) is

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx, \tag{2.2}$$

  which is usually denoted by μ,

- the variance is

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx, \tag{2.3}$$

  which is usually denoted by σ^2.

When we consider a joint distribution of multivariate variables $X_1, \ldots, X_p$,

the variance-covariance matrix is defined as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \cdots & \mathrm{Var}(X_p) \end{pmatrix}, \tag{2.4}$$

where $\mathrm{Cov}(X_i, X_j)$ represents the covariance between random variable X_i and X_j

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}\big([X_i - \mathrm{E}(X_i)][X_j - \mathrm{E}(X_j)]\big). \tag{2.5}$$

Note that $\mathbf{\Sigma}$ is a symmetric and positive-semidefinite matrix [DeGroot and Schervish, 2012, p. 741].

2.1.2 Gaussian Distribution

The Gaussian distribution, also known as the normal distribution, is by far the most widely used distribution in statistics. The density function for a univariate normal distribution with expectation μ and variance σ^2 is

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.6}$$

which is usually denoted by $X \sim N(\mu, \sigma^2)$.

For the p-dimensional multivariate Gaussian distribution with the expectation vector $\boldsymbol{\mu}$ and the variance covariance matrix $\mathbf{\Sigma}$, the density function is

$$f(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{|\mathbf{\Sigma}|(2\pi)^{p/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \tag{2.7}$$

where $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$ and $\mathbf{x}$ is a p-dimensional vector.

2.2 Classification Methods

In this section, we introduce several commonly used classification methods and the framework for modeling a classifier. Section 2.2.1 gives the con-

cepts of training, validation and test set as well as the misclassification cost in the binary classification problem. Logistic regression, introduced in Section 2.2.2, is a well-established classification algorithm, which remains a reference benchmark in many domains, like consumer credit risk, due to the regulatory requirement of interpretability. Deploying logistic regression in highly imbalanced data is the primary concern of this thesis. The limit behavior of logistic regression in highly imbalanced data and corresponding mitigation methods will be discussed in Chapters 3 and 4. Three classification methods will be introduced in Section 2.2.3 and Section 2.2.4, namely linear/quadratic discriminant analysis and k-nearest neighbors. They will be used later in Chapter 4 to expand our relabeling idea.

2.2.1 Framework

In this section, we briefly describe the concept of training, test, and validation sets, which are frequently used among the model training and test process. The concept of the misclassification cost for the binary classification task is also discussed in this section.

Training, Test, and Validation Sets

A training set is a data set used for fitting the model (e.g. fit the parameters of a classifier). A test set is independent from the training set but shares the same distribution with the training set, which is usually used to assess the model performance. The validation set is a sample set hold back from the training set, usually used to estimate the prediction error for model selection.

It is hard to provide a general rule on how to split a data set into training, test and validation set, but usually a three phrase process is frequently used for model evaluation [Hastie et al., 2009, p. 222]: 1. use the training set to train models; 2. use the validation set to estimate the prediction error for model selection; 3. use the test set to assess the performance of the selected model.

Misclassification Cost

Consider a binary classification task; denote the binary response as $Y \in \{0, 1\}$, $\Pr(Y = 1|X = \mathbf{x})$ denotes the conditional probability that the object belongs to class 1 given its feature vector $X = \mathbf{x}$, and $\Pr(Y = 0|X = \mathbf{x})$ has similar definition. Then the cost for misclassifying an particular observation $\mathbf{x}$ can be defined as [Domingos, 1999]

$$R(\mathbf{x}) = \Pr(Y = 1|X = \mathbf{x})C(0 \to 1) + \Pr(Y = 0|X = \mathbf{x})C(1 \to 0) \tag{2.8}$$

where $C(1 \to 0)$ represents the misclassification cost of misclassifying a class 1 observation to class 0 and $C(0 \to 1)$ vice versa. The optimal prediction for a particular $\mathbf{x}$ is the class 0 or 1 that minimize Equation (2.8) [Elkan, 2001]. This will be further discussed in Section 2.3 for high class imbalance problem.

2.2.2 Logistic Regression

Logistic regression is used to estimate the posterior probabilities of each class. We still denote the binary response as $Y \in \{0, 1\}$, then, for an observation $\mathbf{x}$, binary logistic regression has the form:

$$\Pr(Y = 1|X = \mathbf{x}) = \frac{e^{(\beta_0+\beta^T\mathbf{x})}}{1 + e^{(\beta_0+\beta^T\mathbf{x})}}, \tag{2.9}$$

where β_0 is an intercept, and $\beta^T = \{\beta_1, \cdots, \beta_p\}$ is a slope parameter vector, to be estimated. Suppose we have M observations, then the log-likelihood function for independent observations can be written as:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{M} \{y_i \log(\Pr(Y = 1|X = \mathbf{x}_i)) + (1 - y_i) \log(\Pr(Y = 0|X = \mathbf{x}_i))\} \\ &= \sum_{i=1}^{M} \{y_i(\beta_0 + \beta^T\mathbf{x}_i) - \log(1 + e^{\beta_0+\beta^T\mathbf{x}_i})\}. \end{aligned} \tag{2.10}$$

For convenience, we introduce another form of the log-likelihood function which is equivalent to Equation (2.10): consider n observations from class $Y = 1$, denoted by $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n}$, and N observations from class $Y = 0$, denoted by $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0N}$, thus $M = N + n$. Equation (2.10) can be transformed to

$$l(\beta) = \sum_{i=1}^{n} \log \frac{e^{\beta_0 + \beta^T \mathbf{x}_{1i}}}{1 + e^{\beta_0 + \beta^T \mathbf{x}_{1i}}} + \sum_{i=1}^{N} \log \frac{1}{1 + e^{\beta_0 + \beta^T \mathbf{x}_{0i}}}. \tag{2.11}$$

Equation (2.11) is frequently used in Chapters 3 and 4.

If the conditional distributions of X given $Y = y$ is $N(\boldsymbol{\mu_y}, \boldsymbol{\Sigma})$ (multivariate normal with equal covariance matrices in both classes), then the coefficient estimates of the logistic regression model are simply $\boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_0)$ [Anderson and Blair, 1982]. Otherwise, to maximize Equation (2.10), we set its derivatives to zero:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{M} \mathbf{x}_i (y_i - \frac{e^{(\beta_0 + \beta^T \mathbf{x}_i)}}{1 + e^{(\beta_0 + \beta^T \mathbf{x}_i)}}) = 0. \tag{2.12}$$

Equation (2.12) can be solved numerically (e.g. Newton methods) to get the maximum likelihood estimator (MLE) $\widehat{\beta_0}$ and $\widehat{\beta}$.

We can also calculate the variance of $\widehat{\beta}_j, j \in \{0, \cdots, p\}$ from the maximum likelihood estimator. We can consider

$$\frac{\partial l}{\partial \beta_u \partial \beta_v} = -\sum_{i=1}^{M} \frac{x_{iu} e^{\beta_0 + \beta^T \mathbf{x}_i} x_{iv}}{(1 + e^{\beta_0 + \beta^T \mathbf{x}_i})^2} = -\mathbf{X}_{\cdot u}^T W \mathbf{X}_{\cdot v}. \tag{2.13}$$

Here, x_{iu} and x_{iv} refer to the uth and vth elements of the vector $\mathbf{x}_i$, $\mathbf{X}_{\cdot u}$ and $\mathbf{X}_{\cdot v}$ refer to the uth and vth column in the design matrix $\mathbf{X}$ and W is a diagonal matrix

$$\begin{aligned} W &= \operatorname{diag}\left(\frac{e^{\beta_0 + \beta^T \mathbf{x}_1}}{(1 + e^{\beta_0 + \beta^T \mathbf{x}_1})^2}, \cdots, \frac{e^{\beta_0 + \beta^T \mathbf{x}_M}}{(1 + e^{\beta_0 + \beta^T \mathbf{x}_M})^2} \right)_{M \times M} \\ &= \operatorname{diag}(p_1(1 - p_1), \cdots, p_M(1 - p_M)), \end{aligned} \tag{2.14}$$

where p_i is the posterior probability $\Pr(Y = 1 | X = \mathbf{x}_i)$. Thus, the Fisher information $I_Y(\beta) = -E_\beta(\bigtriangledown^2 l(\beta)) = \mathbf{X}^T W \mathbf{X}$. Then, the central limit the-

orem shows that the distribution of $\widehat{\beta}$ converges to a multivariate normal distribution $N(\beta, (\mathbf{X}^T W \mathbf{X})^{-1})$ [Hastie et al., 2009, p. 125]. Here we can use a plugin approximation

$$\begin{aligned} W_{\widehat{\beta}} &= \text{diag}\left(\frac{e^{\widehat{\beta_0}+\widehat{\beta}^T \mathbf{x}_1}}{(1+e^{\widehat{\beta_0}+\widehat{\beta}^T \mathbf{x}_1})^2}, \cdots, \frac{e^{\widehat{\beta}_0+\widehat{\beta}^T \mathbf{x}_M}}{(1+e^{\widehat{\beta}_0+\widehat{\beta}^T \mathbf{x}_M})^2} \right)_{M \times M} \\ &= \text{diag}(\widehat{p}_1(1-\widehat{p}_1), \cdots, \widehat{p}_M(1-\widehat{p}_M)) \end{aligned}$$

as an estimate of W when M is large. The fact that the coefficient estimates $\widehat{\beta}$ follows a multivariate normal distribution will be used in Chapter 5.

Penalized Logistic Regression

Maximum likelihood estimation may become unstable if the dimension of data is high or several variables are highly correlated [Lessmann et al., 2015]. In order to perform parameter shrinkage and variable selection, penalized logistic regression is designed by adding penalty terms to the likelihood function (Equation 2.10). These penalties include l_1 (lasso [Tibshirani, 1996]), l_2 (ridge [Hoerl and Kennard, 1970]) and mixtures of the two (elastic-net [Zou and Hastie, 2005]). The general form for penalized logistic regression is

$$l(\beta) = \sum_{i=1}^{N} \left[y_i(\beta^T \mathbf{x}_i) - \log(1+e^{\beta^T \mathbf{x}_i}) \right] - \lambda \left[(1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \text{where } \lambda > 0. \tag{2.15}$$

The penalty term $\left[(1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right]$ uses two types of penalties, with $\|\beta\|_1$ and $\|\beta\|_2^2$ denoting the l_1 and the squared l_2 norms of β ($\alpha = 1$ is lasso, $\alpha = 0$ is ridge, $0 < \alpha < 1$ is elastic-net). When solving the penalized logistic regression, we usually do not penalize the intercept term and implement a standardization process before the penalty in order to make the penalty meaningful.

The effect of lasso is variable selection by exactly penalizing some parameters to zero as λ increases [Hastie et al., 2009, page 68]. However, the ridge penalty only makes parameters shrink toward to zero (but not equal to zero) with λ increasing [Hastie et al., 2009, page 61]. The parameter λ allows the

user to control the trade-off between the model complexity and the goodness of fit [Hastie et al., 2009], and we usually use cross-validation (discussed in Section 2.4.4) to choose the optimal penalty parameter λ. The highly imbalanced penalized logistic regression will be discussed in Section 3.3.

Multinomial Logistic Regression

Binary logistic regression can be easily extended to K class multinomial logistic regression by reorganize Equation (2.9) to

$$\begin{aligned}
\log\frac{\Pr(Y=1|X=\mathbf{x})}{\Pr(Y=K|X=\mathbf{x})} &= \beta_{10}+\beta_1^T\mathbf{x},\\
\log\frac{\Pr(Y=2|X=\mathbf{x})}{\Pr(Y=K|X=\mathbf{x})} &= \beta_{20}+\beta_2^T\mathbf{x},\\
&\vdots\\
\log\frac{\Pr(Y=K-1|X=\mathbf{x})}{\Pr(Y=K|X=\mathbf{x})} &= \beta_{(K-1)0}+\beta_{K-1}^T\mathbf{x},
\end{aligned} \tag{2.16}$$

where β_{k0} is the intercept term and β_k is the slope vector ($k \in \{1, 2, \ldots, K-1\}$). The highly imbalanced multinomial logistic regression will be discussed in Section 3.4.

2.2.3 Linear Discriminant Analysis and Quadratic Discriminant Analysis

Linear discriminant analysis (LDA [Lachenbruch and Goldstein, 1979]) and quadratic discriminant analysis (QDA [Klecka et al., 1980]) are two widely used classification methods. In a K class classification problem, assume $f_k(\mathbf{x})$ is the probability density distribution function of each class k and the prior class proportions are ϕ_k; then the posterior probability for an observation $\mathbf{x}$ comes from class k is

$$\Pr(Y=k|X=\mathbf{x}) = \frac{\phi_k f_k(\mathbf{x})}{\sum_{k=1}^K \phi_k f_k(\mathbf{x})}. \tag{2.17}$$

LDA considers a special case when $f_k(\mathbf{x}) \sim N(\boldsymbol{\mu_k}, \boldsymbol{\Sigma})$, i.e. each class is normally distributed with mean $\boldsymbol{\mu_k}$ and shares a common covariance matrix $\boldsymbol{\Sigma}$. Then the log-ratio between $\Pr(Y = k_1|X = x)$ and $\Pr(Y = k_2|X = x)$ is

$$\begin{aligned}\log\frac{\Pr(Y = k_1|X = \mathbf{x})}{\Pr(Y = k_2|X = \mathbf{x})} =& \log\frac{\phi_{k_1}}{\phi_{k_2}} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_{k_1}} - \boldsymbol{\mu_{k_2}}) \\ &- \frac{1}{2}(\boldsymbol{\mu_{k_1}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_{k_1}} + \boldsymbol{\mu_{k_2}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_{k_2}}),\end{aligned} \tag{2.18}$$

which is a linear function of $\mathbf{x}$. Thus we can define the linear discriminant function as

$$\mathrm{LDA}_k(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k} + \log\phi_k - \frac{1}{2}\boldsymbol{\mu_k}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k}, \text{where } k \in \{1, \ldots, K\}. \tag{2.19}$$

We find that the discriminant rule: " find an optimal k for particular $\mathbf{x}$ to minimize the classification error" is equivalent to "$k = \arg\max_k \mathrm{LDA}_k(\mathbf{x})$".

The "equal covariance" assumption leads to the cancellation of the quadratic part in the normal distribution (Equation 2.7) when we are calculating the log-ratio (Equation 2.18) for LDA. QDA further considers each class has different covariance matrix $\boldsymbol{\Sigma}_k$. Then the quadratic part will remain in the quadratic discriminant function:

$$\mathrm{QDA}_k(\mathbf{x}) = \log\phi_k - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_k})^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu_k}) - \frac{1}{2}|\boldsymbol{\Sigma}_k|, \text{where } k \in \{1, \ldots, K\}. \tag{2.20}$$

The discriminant rule for QDA is "$k = \arg\max_k \mathrm{QDA}_k(\mathbf{x})$".

2.2.4 k-Nearest Neighbors

Given labeled data $(\mathbf{x}_i, y_i)$, perhaps the simplest prediction rule is predicting an input $\mathbf{x}$ according its nearest neighbor:

$$\widehat{f}(\mathbf{x}) = y_i \text{ such that } \|\mathbf{x}_i - \mathbf{x}\| \text{ is the smallest.}$$

This is usually called the 1-nearest neighbor model. A natural extension is to consider the k-nearest neighbors (kNN [Cover and Hart, 1967]) of $\mathbf{x}$, call

them $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$, and then classify according to a majority vote:

$$\widehat{f}(\mathbf{x}) = j \text{ such that } \sum_{i=1}^{k} \mathbf{1}(\mathbf{x}_{(i)} = j) \text{ is the largest,}$$

where $\mathbf{1}$ is an identification function. Typically, we standardize each features to mean 0 and variance 1, because they may be measured in different units. The only parameter for kNN is the number of the nearest neighbors k, which can be selected by cross validation (introduced in Section 2.4.4). The concept of k-nearest neighbors is used in Section 2.3 for an oversampling algorithm.

2.3 Class Imbalance Problem

This section briefly describes the class imbalance problem, which is a common problem in the field of classification and is the main concern of our research. Strictly speaking, any data that has an unequal class proportion can be considered as imbalanced data. The common understanding in the community of **high class imbalance** usually refers to the situation when some classes are significantly under-populated among the data set like 20:1, 100:1 or 1000:1 [He and Garcia, 2008, Krawczyk, 2016]. Dealing with class imbalance is a challenging technical problem which is important in a variety of applications, including image classification [Buda et al., 2018], medical science [Mac Namee et al., 2002, Li et al., 2010], fraud detection [Guo et al., 2008], and political science [King and Zeng, 2001b]. In the real world, the minority class usually related to a concept of higher interest than the majority class like the loan default in the credit risk industry [Brown and Mues, 2012] and oil spills in satellite radar images [Kubat et al., 1998].

The performance of standard classification algorithms is restricted when learning from imbalanced data [Visa and Ralescu, 2005]. Addressing this challenge, various authors proposed different approaches to handle it. These approaches can be divided into three groups, based on their mechanisms [He and Garcia, 2008, García and Herrera, 2009]:

1. Data level: construct a balanced data set via sampling,

2. Algorithm level: modify algorithms for learning imbalanced data (e.g. cost-sensitive learning, active learning),

3. Combined data level and algorithm level approach.

The novel approach developed in Chapter 4, built around logistic regression, falls most naturally in the "algorithm level" category.

2.3.1 Data level

The data level approach seeks to produce a balanced data set by modifying the collection of the examples, e.g. "sampling" [Krawczyk, 2016]. Several studies have shown that some basic classification algorithms can benefit from learning on balanced data, which justifies the use of the sampling methods [Laurikkala, 2001, Estabrooks et al., 2004].

Random undersampling and random oversampling are two natural methods that randomly remove samples from the majority class or randomly replicate the minority class samples. Both of these methods can modify the class proportion to any level, providing the chance that modeling can benefit from a more balanced data set, but they bring their own problems, which may influence model performance. Random undersampling can lead to information loss in the majority class. For random oversampling, the duplicated observations in the minority class may change the data structure hence leading to overfitting [He and Garcia, 2008]. This is because the duplicated observations may result in a too specific decision boundary for a classifier; although performing well on the training set, the performance on the test set may undesirable [Mease et al., 2007].

Several informed undersampling methods are proposed to overcome the drawback of random undersampling. For example, Mani and Zhang [2003] proposed to use k nearest neighbor (kNN) approach for informed undersampling. Essentially, it will select a given number of the majority class observations which is nearest to each minority class observation. Thus it will guarantee the minority class is surrounded by the majority class. Similarly, Kubat et al. [1997] proposed a "one-sided selection procedure" aiming at removing

the "noisy" majority class observations to construct a representative subset of the majority class observations. Here, "noisy" majority class observation means those observations from the majority class occurring far away from the minority class. Liu et al. [2009] also proposed two informed undersampling methods to overcome the drawback of random undersampling, i.e., EasyEnsemble and BalanceCascade. EasyEnsemble method trains several classifiers on several subsets from the majority class, then ensemble the outputs of those classifiers. BalanceCascade method iteratively trains the classifiers; in each iteration, the correctly classified majority class observations are removed, hence forming an undersampled data set for the next iteration.

For informed oversampling, a widely used method is SMOTE (Synthetic Minority Over-Sampling technique), which shows good performance in application [Chawla et al., 2002]. Different to random oversampling, SMOTE adds artificially generated data that has the same distribution character as the minority class, rather than by oversampling with replacement like random oversampling. For continuous variables, the generated synthetic minority class samples are located on the line between one minority class observation and one of its k nearest neighbors, where the k nearest neighbors are considered within the minority class observations. We will compare the performance of our approach with SMOTE when handling the class imbalance problem.

SMOTE generates the same number of synthetic samples for each original minority class observation, without consideration of increasing the occurrence of overlapping between classes [López et al., 2013]. To this end, adaptive sampling further considers the overlap between the majority class and the minority class, like borderline-SMOTE method [Han et al., 2005] and adaptive synthetic sampling method [He et al., 2008], which increases the occurrence of overlapping between classes by selectively generating synthetic samples [Wang and Japkowicz, 2004]. The borderline-SMOTE method first selects "danger" minority class observations. Here, a "danger" minority class observation refers to a minority class observation which, in its k-nearest neighbors, the number of the majority class observations is greater than the

number of the minority class observations. Then, borderline-SMOTE deploy SMOTE method on "danger" minority class observations. The adaptive synthetic sampling method uses a density distribution Γ to automatically decide how many synthetic samples should be generated for each minority class observations. The distribution Γ is determined by the proportion of the majority class observations that appeared among the k-nearest neighbors of each minority class observation, thus, a higher proportion will lead to more synthetic samples for that minority class observation.

Cluster-based sampling [Jo and Japkowicz, 2004] also appears in the literature. Nickerson et al. [2001] argue that balancing the class proportion is not an effective approach when the small disjuncts appear among the minority class. Jo and Japkowicz [2004] proposed an oversampling method with the consideration of the small disjuncts. They suggest using unsupervised clustering methods, like K-means (see Section 2.5), to cluster both the majority class and the minority class into several small disjuncts; then randomly oversample all of these small disjuncts to the size of the largest disjunct.

2.3.2 Algorithm level

While the data level approach advocates editing data to alleviate the imbalance problem, the algorithm level approach seeks to modify the classifier to learn imbalanced data. One popular approach is cost-sensitive learning, which considering the misclassification cost between different classes. As we described in Section 2.2.1, let $C(1 \to 0)$ represent the misclassification cost of misclassifying a class 1 observation to class 0 and $C(0 \to 1)$ vice versa. In a binary classification problem, the cost for misclassifying $\mathbf{x}$ is defined as $R(\mathbf{x})$ (Equation 2.8). We usually assume that there is no cost when an observation is correctly classified i.e.

$$C(0 \to 0) = C(1 \to 1) = 0,$$

and

$$C(1 \to 0) > C(0 \to 1) > 0$$

when an observation is misclassified to the wrong class in the highly imbalanced scenario. The objective of cost-sensitive learning is to produce a classifier that minimizes the overall cost on the training set [Elkan, 2001].

For application, there are two approaches to conduct cost-sensitive learning. The first approach is introducing classifiers that are cost-sensitive. Because these techniques are specific to a particular classifier, there is no uniform framework for this type of cost-sensitive learning. We consider an example here for cost sensitive logistic regression [Bahnsen et al., 2014]. Let

$$p_i = \Pr(Y = 1|X = \mathbf{x}_i) = \frac{e^{(\beta_0+\beta^T\mathbf{x}_i)}}{1 + e^{(\beta_0+\beta^T\mathbf{x}_i)}},$$

where $\mathbf{x}_i$ is an observation, then the overall cost function for M observations is

$$\frac{1}{M}\sum_{i=1}^{M}\Big[y_i(1-p_i)C(1\to 0) + (1-y_i)p_iC(0\to 1)\Big]. \tag{2.21}$$

The objective of cost sensitive logistic regression is solving $(\hat{\beta}_0, \hat{\beta})$ to minimize the cost function (2.21), which can be achieved using numerical methods.

Another way to conduct cost sensitive learning is by designing a general method to modify a cost insensitive classifier to become cost sensitive. For example, for a given threshold t (see Section 2.4.1 for definition), the total misclassification cost for a set of observations can be calculated, and it is a function of t. The researcher can obtain this misclassification cost curve by calculating each possible threshold (i.e. the predicted probability of each observation). The threshold adjusting method [Sheng and Ling, 2006] simply chooses the best threshold, which minimizes this curve.

We can also achieve cost-sensitive learning by rebalancing the data set. Elkan [2001] proves that the number of class 0 observations in the training set should be rebalanced to

$$\text{the number of the class 0 observations} \times \frac{p\pi_1}{(1-p)\pi_0},$$

in order to make a classifier cost-sensitive, where π_0, π_1 are the prior propor-

tions of class 0 and class 1, and

$$p = \frac{C(0 \to 1)}{C(0 \to 1) + C(1 \to 0)}.$$

If a classification algorithm can use weights on training set, the weight for observations in class 0 can be directly set to $p\pi_1/((1-p)\pi_0)$, otherwise we can use undersampling method.

In addition to cost-sensitive learning, there are several modified classifiers targeting learning imbalanced data, which are algorithm level methods. Perhaps the most famous is balanced random forest [Chen et al., 2004]. Balanced random forest modifies the random forest as follows: for each tree in a random forest [Breiman, 2001], it will draw a bootstrap sample from the minority class and a bootstrap sample from the majority class with equal size to the size of the minority class. This technique may avoid the disappearance of the minority class when generating a bootstrap sample from the whole training set, hence provide a good predictive performance [Fitzpatrick and Mues, 2016].

2.3.3 Mitigation Methods for Highly Imbalanced Logistic Regression

Apart from general mitigation methods to handle the class imbalance problem (e.g. sampling and cost-sensitive learning introduced in the previous section), particular attention has been paid to mitigation methods for highly imbalanced logistic regression by several authors. As mentioned in Section 2.2.2, maximum likelihood estimation (MLE) is a widely used parameter estimation method for logistic regression. King and Zeng [2001b] investigate the bias of the intercept term in maximum likelihood estimates of logistic regression for highly imbalanced data. Assuming $\Pr(Y = 1|X = \mathbf{x}) = \frac{e^{\beta_0 + \beta^T \mathbf{x}}}{1 + e^{\beta_0 + \beta^T \mathbf{x}}}$, they find the bias of the MLE $\hat{\beta}_0$ is

$$\mathrm{E}(\hat{\beta}_0 - \beta_0) \approx \frac{\bar{p} - 0.5}{M\bar{p}(1-\bar{p})}, \tag{2.22}$$

where $\bar{p}$ is the proportion of the minority class ($Y = 1$) in the sample set and M is the sample size. For imbalanced data, $\bar{p}$ can be very small, resulting in large $\mathrm{E}(\hat{\beta}_0 - \beta_0)$. They proposed two correction methods based on prior information about the proportion of the minority class (which is denoted by τ):

- *prior correction*: the corrected intercept term is

$$\hat{\beta}_0 - \log\left[\frac{1-\tau}{\tau}\frac{\bar{p}}{1-\bar{p}}\right], \tag{2.23}$$

- *weighting correction*: find the MLE through a weighted log-likelihood function, where the weight for each minority class observation is $\tau/\bar{p}$ and for each majority class observation is $(1-\tau)/(1-\bar{p})$*.

Some literature also proposes "algorithm level" approaches to mitigate the class imbalance problem for logistic regression. For example, Dong et al. [2014] proposed a modified logistic regression by combining the recall metric. Consider the log-likelihood function (2.11) for logistic regression

$$l(\beta) = \sum_{i=1}^{n} \log \frac{e^{\beta_0+\beta^T \mathbf{x}_{1i}}}{1+e^{\beta_0+\beta^T \mathbf{x}_{1i}}} + \sum_{i=1}^{N} \log \frac{1}{1+e^{\beta_0+\beta^T \mathbf{x}_{1i}}},$$

where $\mathbf{x}_{1i}$ are the minority class observations and $\mathbf{x}_{0i}$ are the majority class observations; we can further define $R_1 = \sum_{i=1}^{n} \Pr(Y = 1|X = \mathbf{x}_{1i})/n$ and $R_0 = \sum_{i=1}^{N} \Pr(Y = 1|X = \mathbf{x}_{0i})/N$ as the recall ratio for class 1 and class 0 respectively. The objective function they propose is

$$l_{\text{modified}}(\beta) = l(\beta) + CM(R_1 + R_0), \tag{2.24}$$

where $C \in [0, 1]$ to control the trade-off between the $l(\beta)$ and the recall. Equation (2.24) attempts to enhance the recall of the majority class and the minority class simultaneously for better predictive performance on both classes.

*This method works for the bias of the intercept term in MLE; however, as we will discuss in Section 3.2, the slope vector obtained by weighted likelihood method still suffer from highly imbalanced data.

Table 2.1: Confusion Matrix for Binary Classification.

		Predicted	
		Class 1	Class 0
Actual	Class 1	True Positive (TP)	False Negative (FN)
	Class 0	False Positive (FP)	True Negative (TN)

2.4 Performance Assessment for Classification

For the purpose of assessing the prediction performance of classifiers in this thesis, in this section, we introduce several commonly used performance metrics and the corresponding measurement procedures. We will also derive some theoretical results for a widely used performance metric, namely area under the receiver operating characteristic curve (AUC), when applied in the high imbalance scenario. This is our original work.

2.4.1 Confusion Matrix

A binary classifier will predict the posterior probability or assign a class label (positive/negative) for each sample in the test set. Let the positive class be denoted by $Y = 1$ and the negative class denoted by $Y = 0$; in cases when the posterior probability is calculated, the threshold t, as a cutoff point, will be used for assigning an observation to positive class 1 when $\Pr(Y = 1|X = \mathbf{x}) > t$. Table 2.1 is a 2×2 confusion matrix for binary classifiers. Here, FP (false positive) is the fraction of negative samples that are classified as positive. FN (false negative), TP (true positive), and TN (true negative) have similar meaning.

For a binary classification problem, many performance metrics can be calculated from the confusion matrix. A list of common performance metrics, derived from this confusion matrix, is displayed in Table 2.2.

Table 2.2: Performance Metrics; $P = TP + FN$; $N = FP + TN$.

Performance Metric	Equation
Accuracy	$(TP + TN)/(P + N)$
Error rate	$1 - (TP + TN)/(P + N)$
False positive rate or False alarm rate	FP/N
Precision	$TP/(TP + FP)$
True positive rate or Recall or Sensitivity	TP/P
Specificity	TN/N
F measure	$2/\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}$

2.4.2 Area Under Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve [Fawcett, 2006] displays true positive rate (TPR on the vertical axis) against false positive rate (FPR on the horizontal axis) in a plot simultaneously as the threshold t varies. The area under the ROC curve (AUC) can be used to summarize the overall performance of a model. AUC can be interpreted in various ways, one interpretation is: the AUC is the probability of the event "when a randomly selected class 1 sample has a higher predicted probability than that for a randomly selected class 0 sample" [Thomas, 2009, p. 115]. Thus, higher AUC is preferred.

For the calculation of the AUC, we can define the "score of an observation $\mathbf{x}$" as $s(\mathbf{x}) = \Pr(Y = 1|X = \mathbf{x})$. Then, we can further define $f_0(s)$ and $f_1(s)$ as the probability density function of the scores for the negative class and the positive class respectively, with the cumulative distribution function $F_0(s)$ and $F_1(s)$. In this setting, for a certain threshold t, we have $\text{TPR} = 1 - F_1(t)$ and $\text{FPR} = 1 - F_0(t)$. Then the ROC curve is the plot of $1 - F_1(t)$ against $1 - F_0(t)$ and

$$\text{AUC} = \int_{-\infty}^{\infty} (1 - F_1(t)) f_0(t) dt. \tag{2.25}$$

In this section, for simplicity, let us use A to denote the AUC. In application, to obtain the estimate of A (denoted by $\widehat{A}$), we need the true label and the predicted probability of the observations in a test set. The simplest way is to produce a plot of the ROC curve, and calculate $\widehat{A}$ by using quadrature.

Hand and Till [2001] also give a way to calculate $\widehat{A}$ by ranking the scores $s(\mathbf{x}_i)$ on the test set to replace the theoretical function F_1 and f_0 by the observed values. We will discuss the variance of $\widehat{A}$ in the highly imbalanced data later.

The AUC is a commonly used model performance metric since the AUC transforms the ROC curve to a single numeric value, does not require error cost information, and summaries the performance across all thresholds [Wu et al., 2007, He and Garcia, 2009]. It is objective, and easy for calculation and interpretation. Researchers can get the same AUC given the same trained classifier and an identical test set, which let different classification models' performance can be easily compared by naturally ranking their AUC for a given data set [Kaymak et al., 2012]. As it can be seen in some benchmark comparison studies of classification algorithms performance for credit scoring [Baesens et al., 2003, Lessmann et al., 2015], for the consumer credit risk industry, using the AUC as a performance metric is a consistent standard practice and also explicitly mentioned in the Basel II capital accord [Lessmann et al., 2015]. Given the importance of the AUC, we use it as an important performance metric in our experiments.

However, the AUC has its pitfalls as well. The incoherency of the AUC has been investigated by Hand and Anagnostopoulos [2013], as we explain now. Let the *cost ratio* between the misclassification cost be defined by $C(0 \to 1)/C(1 \to 0)$, where $C(1 \to 0)$ represents the misclassification cost of misclassifying a class 1 observation to class 0 and $C(0 \to 1)$ vice versa. Once this cost ratio is determined, it will correspond to a optimal classification threshold t, which can minimize the overall misclassification cost on a given data set. Hand and Anagnostopoulos [2013] show that one can obtain the AUC for a classifier on a given data set by integrating the overall misclassification loss on a chosen distribution of threshold t. In particular, for calculating the AUC, this distribution of t is determined by the posterior probability calculated from the classifier itself. Putting it in different words, considering the link between the threshold t and the cost ratio between $C(0 \to 1)$ and $C(1 \to 0)$, using the AUC means assuming using different misclassification costs for different classifiers (i.e. misclassification cost depend on classifiers),

which is questionable since we usually believe these misclassification costs are problem domain dependent [Kaymak et al., 2012]. A new measure called the H-measure [Hand, 2009], which is a remedy to this problem, will be introduced in Section 2.4.3.

The variance of the AUC for High Class Imbalance Data

The AUC is calculated from the predicted probability $\Pr(Y = 1|X = \mathbf{x})$ and the true label on a test set, which does not require specifying a threshold. This makes it one of the most popular performance metrics when assessing classifiers. In this section, we focus on the variance of the estimated $\widehat{A}$ in highly imbalanced data. This section is original work.

Hanley and McNeil [1982] derived the variance of an empirical AUC as,

$$s^2(\widehat{A}) = \frac{1}{n_N n_P}\left[A(1-A) + (n_p - 1)(Q_1 - A^2) + (n_N - 1)(Q_2 - A^2)\right], \tag{2.26}$$

where $0.5 < A < 1$, n_N denotes the number of the negative class observations, n_P denotes the number of the positive observations, Q_1 is the probability that the predicted probability of two randomly selected positive observations exceeds the predicted probability of a randomly selected negative observation and Q_2 is the probability that the classification score of two randomly selected negative observations exceeds the predicted probability of a randomly selected positive observation. Q_1 and Q_2 are two complex functions based on the populations, but Hanley and McNeil [1982] give a good approximation of Q_1 and Q_2 which are $\frac{A}{2-A}$ and $\frac{2A^2}{1+A}$ respectively (see [Krzanowski and Hand, 2009, p. 79]). Without loss of generality, let $n_N > n_P$, which means the positive class is our minority class.

For simplicity, we use $x \in \mathbf{Z}^+$ to represent the number of the positive (minority) class observations n_P in this section. If M denotes the number of the

total observations, then we can rewrite Equation (2.26) as a function of x,

$$f(x) = \frac{1}{x(M-x)}\left[A(1-A) + (x-1)\left(\frac{A}{2-A} - A^2\right) + (M-x-1)\left(\frac{2A^2}{1+A} - A^2\right)\right], \tag{2.27}$$

and with a bit of foresight we require $2x < M+1$, which is obviously true in the high imbalance scenario. Here, we want to show:

Proposition 1. *Formula (2.27) $f(x)$ is a decreasing function for fixed M when $0.5 < A < 1$.*

The decreasing function (2.27) means the variance of the AUC, $s^2(\widehat{A})$, will increase when the data is more imbalanced.

Proof. For simplicity, now we consider the log transform of $f(x)$,

$$\begin{aligned} g(x) &= \log(f(x)) \\ &= \log\left[A(1-A) + (x-1)\left(\frac{A}{2-A} - A^2\right) + (M-x-1)\left(\frac{2A^2}{1+A} - A^2\right)\right] \\ &\quad - \log(x) - \log(M-x). \end{aligned} \tag{2.28}$$

Here, we seek to show $g(x) - g(x-1) < 0$ when x is a positive integer between 1 and $\frac{M+1}{2}$:

$$\begin{aligned} g(x) - g(x-1) = {} & \log(x-1) + \log(M-x+1) - \log(x) - \log(M-x) \\ & + \log\left(\frac{(A(1-A) + (x-1)(\frac{A}{2-A} - A^2) + (M-x-1)(\frac{2A^2}{1+A} - A^2))}{(A(1-A) + (x-2)(\frac{A}{2-A} - A^2) + (M-x)(\frac{2A^2}{1+A} - A^2))}\right). \end{aligned} \tag{2.29}$$

We consider the first line of Equation (2.29)

$$\begin{aligned} & \log(x-1) + \log(M-x+1) - \log(x) - \log(M-x) \\ =& \log\left(\frac{-x^2 + (M+2)x - (M+1)}{-x^2 + Mx}\right). \end{aligned} \tag{2.30}$$

Because $2x < M + 1$, we know

$$\left(-x^2 + (M+2)x - (M+1)\right) - \left(-x^2 + Mx\right) = 2x - (M+1) < 0,$$

so we have

$$\frac{-x^2 + (M+2)x - (M+1)}{-x^2 + Mx} < 1;$$

The above results show that Equation (2.30) must be smaller than 0. Now, we consider the second line of Equation (2.29), let

$$q(x) = A(1-A) + (x-1)\left(\frac{A}{2-A} - A^2\right) + (M-x-1)\left(\frac{2A^2}{1+A} - A^2\right), \quad (2.31)$$

and with some simple calculation we can simplify $q(x)$ to

$$\begin{aligned} q(x) =& \left(\left(\frac{A}{2-A} - A^2\right) - \left(\frac{2A^2}{1+A} - A^2\right)\right) x \\ &+ \left(A(1-A) - \left(\frac{A}{2-A} - A^2\right) + (M-1)\left(\frac{2A^2}{1+A} - A^2\right)\right) \\ =& \frac{A(1-A)(1-2A)}{(2-A)(1+A)} x \\ &+ \left(A(1-A) - \left(\frac{A}{2-A} - A^2\right) + (M-1)\left(\frac{2A^2}{1+A} - A^2\right)\right). \end{aligned} \quad (2.32)$$

Since $0.5 < A < 1$, we can show that

$$q(x) - q(x-1) = \frac{A(1-A)(1-2A)}{(2-A)(1+A)} < 0,$$

which means that we have $q(x) < q(x-1)$ for the positive integer x. Consider the second line of Equation (2.29), we have

$$\log \frac{q(x)}{q(x-1)} < \log(1) = 0.$$

Thus we have the result $g(x) - g(x-1) < 0$. □

We have shown that $s^2(\widehat{A})$ is increasing when the imbalance level increases. This means that the variance of the empirical AUC is larger in the highly

imbalanced scenario, which makes the marginal AUC difference between different classifiers on the highly imbalanced data difficult to interpret.

Further considering the proportion of the minority class denoted by $\rho = x/M$, we can rewrite Equation (2.26) as

$$\begin{aligned} s^2(\widehat{A}) =& \frac{1}{M\rho(M-M\rho)}\left[A(1-A) + (M\rho-1)\left(\frac{A}{2-A} - A^2\right)\right. \\ & \left. + (M-M\rho-1)\left(\frac{2A^2}{1+A} - A^2\right)\right] \\ =& \frac{1}{M^2\rho(1-\rho)}\left[A(1-A) - \left(\frac{A}{2-A} - A^2\right) - \left(\frac{2A^2}{1+A} - A^2\right)\right. \\ & \left. + M\rho\left(\frac{A}{2-A} - A^2\right) + M(1-\rho)\left(\frac{2A^2}{1+A} - A^2\right)\right]. \end{aligned} \tag{2.33}$$

As the imbalance level increases, $\rho \to 1/M$, we will have

$$s^2(\widehat{A}) = \left(\frac{2}{1+A} - 1\right)A^2 + A\frac{1-A}{1+A}\frac{1}{M-1}. \tag{2.34}$$

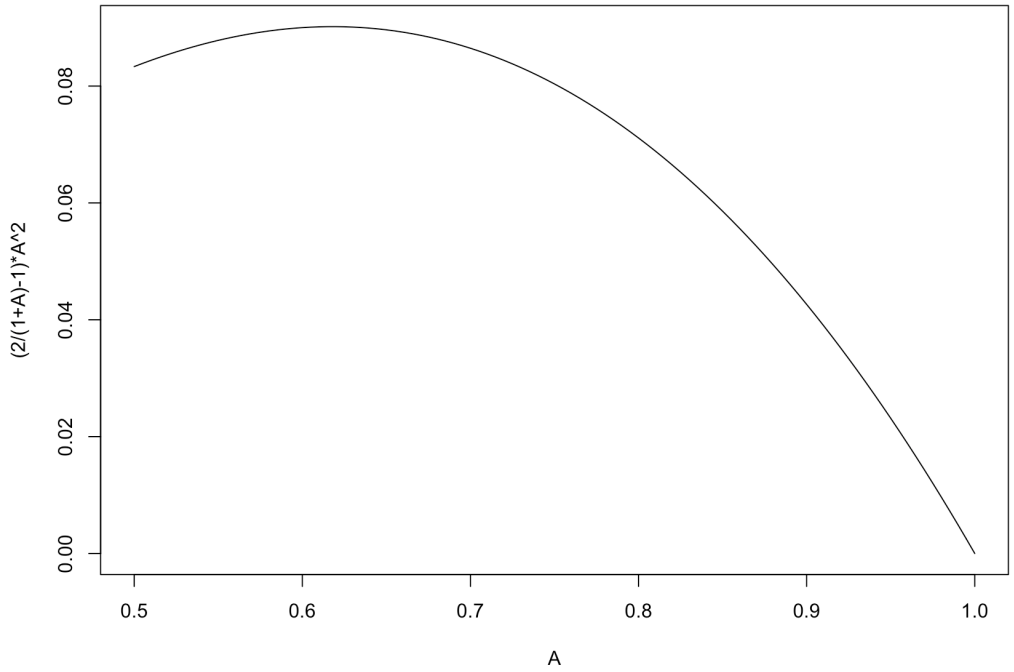
Equation (2.34) shows that the variance of the empirical AUC in extremely imbalanced scenario can be separated into two fractions: an unavoidable part $\left(\frac{2}{1+A} - 1\right)A^2$ (shows in Figure (2.1) with A varying between 0.5 and 1) and a term $\frac{A(1-A)}{(1+A)(M-1)}$ which can be reduced by increasing M.

2.4.3 H-measure

The H-measure is a coherent alternative to the AUC [Hand, 2009]. As discussed in [Hand, 2009], using the AUC assumes the misclassification cost is classifier dependent, thus, a remedy is incorporating a universal standard distribution to specify the relative severities of different misclassification errors. Hand [2009] proposes to use the Beta function

$$u_{\alpha,\beta}(c) = \frac{c^{\alpha-1}(1-c)^{\beta-1}}{\mathbf{B}(1;\alpha,\beta)}, \text{where } \mathbf{B}(1;\alpha,\beta) = \int_0^1 c^{(\alpha-1)}(1-c)^{(\beta-1)}dc,$$

Figure 2.1: the Curve of the Function $\left(\frac{2}{1+A}-1\right)A^2$ between 0.5 and 1



as a simple solution and the choice of parameters α, β depends on the cost of misclassifying each class (setting $\alpha = \beta$ leads to a symmetric Beta distribution). The general form of H-measure is

$$H = 1 - \frac{\int Q(T(c); b, c) u_{\alpha,\beta}(c) dc}{\pi_0 \int_0^{\pi_1} c u_{\alpha,\beta}(c) dc + \pi_1 \int_{\pi_1}^1 (1-c) u_{\alpha,\beta}(c) dc}, \tag{2.35}$$

where F_0 and F_1 are the cumulative distribution function of scores for the class 0 and 1 (the same definition in the previous section), b and c are the misclassification costs, π_0 and π_1 are the prior proportions of class 0 and 1, the loss function

$$Q(T(c); b, c) = \{c\pi_0(1 - F_0(T(c)) + (1-c)\pi_1 F_1(T(c))\}b$$

and

$$T(c) = \arg\min_t \{c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t)\}.$$

Without more information about the misclassification costs in the applications studied throughout this thesis, we use the default setting $\{\alpha = \pi_1 + 1, \beta = \pi_0 + 1, b = \pi_0, c = \pi_1\}$ proposed by its creators [Hand, 2009, Hand and Anagnostopoulos, 2014] to calculate the H-measure.

As an alternative to the AUC, the H-measure is not problem-free. The H-measure can face the risk of losing the fundamental requirement for a performance metric, which is objectiveness, because the choice of $\{\alpha, \beta\}$, and the assumption of the Beta function itself can be replaced by others. This may bring concern that different researchers will have different H-measure values under different assumptions.

2.4.4 Measurement Methods

In this section, we discuss two measurement methods: cross-validation and the bootstrap. These are resampling methods that iteratively fitting the model on the resampled data, which can give the additional information of the fitted model. For example, they provide the estimates of the test-set AUC and H-measure with their corresponding standard deviation. Both of them will be used in experiments in Chapter 4.

K Fold Cross Validation

The idea of K fold cross validation is to randomly divide the data into K equal sized folds. In each iteration, the kth fold is left out as a validation fold, where $k \in 1, \ldots, K$. We fit the model to the other $K - 1$ folds (combined) and then obtain the performance measure (e.g. error rate, the AUC, and area under the PR-curve) on the left out kth fold. This is done iteratively for each fold k and then the results are averaged. In each iteration, the validation fold is distinct from the other $K - 1$ folds for training, which is important since it generally results in a less biased and less optimistic estimate of the model performance [James et al., 2013, p. 181]. Setting K to the sample

size yields to the leave one out cross validation. Choosing $K = 5$ or $K = 10$ is a standard practice and generally be a good trade-off between bias and variance [Breiman and Spector, 1992].

Bootstrap

The bootstrap [Efron and Tibshirani, 1986] is a flexible statistical tool that can be used to quantify the uncertainty associated with a given estimator or a statistical learning method. The bootstrap obtains B distinct data sets by repeatedly sampling observations B times from the original data set with replacement. The sample size of these "bootstrap sample sets" are the same size as the original data (both sample size are M). As a result, some observations may appear multiple times in a given bootstrap data set and some may disappear. Those observations not appearing in a given bootstrap sample can form a "out of bootstrap sample set". From B different bootstrap data sets, we can refit the model of interest B times. This can be used to estimate the mean and the standard deviation of the parameters of the model.

The split between $K - 1$ training folds and kth validation fold is important for estimating the error rate, the AUC, or the area under the PR-curve in K fold cross validation. To estimate the performance measure by using the bootstrap method, we can think about using each "bootstrap sample set" as our training set, and the "out of bootstrap sample set" as our validation set. The B error rates, the AUCs, or the areas under the PR-curve on the out of bootstrap validation set can be used to estimate the mean and the standard deviation of those corresponding statistics. This process usually being called "out of bag (OOB) estimate". Actually, for a large enough data set, since

$$\begin{aligned} \Pr(\text{observation } i \in \text{a "boostrap sample set" }) &= 1 - (1 - \frac{1}{M})^M \\ &\approx 1 - e^{-1} = 0.632, \end{aligned} \tag{2.36}$$

where M is the number of the observations, roughly 63.2% observations will be used as training data and 36.8% observations will be used as validation data in each "OOB estimate" iteration.

2.5 Clustering Methods

Clustering methods are a type of learning algorithm used to discover data structures [Hastie et al., 2009, p. 485]. Clustering aims at learning the underlying structure of a data set based on the resemblance between observations, without a supervisor providing the correct answer. To be specific, clustering splits data into several clusters such that points in the same cluster are "more similar" than points in different clusters. In this section, we will briefly introduce two clustering methods: K-means and hierarchical clustering. They will be used to support relabeling idea in Chapter Four.

2.5.1 K-means

The K-means algorithm [Hartigan and Wong, 1979] classifies a given data set into a pre-specified number (K) of clusters. Given p-dimensional observations $\mathbf{x}_1, \cdots, \mathbf{x}_M$, and dissimilarity measure $d(\mathbf{x}_i, \mathbf{x}_j)$, we use $C(\mathbf{x}_i) = k$ to denote a clustering function C which assigns observation $\mathbf{x}_i$ to group $k \in \{1, \cdots, K\}$. When focusing on Euclidean space (dissimilarities are $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$), the within-point scatter can be written as

$$\begin{aligned} W &= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{C(\mathbf{x}_i)=k}\sum_{C(\mathbf{x}_i')=k}\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 \\ &= \sum_{k=1}^{K}\sum_{C(\mathbf{x}_i)=k}\|\mathbf{x}_i - \overline{\mathbf{x}}_k\|^2, \end{aligned} \tag{2.37}$$

where $\overline{\mathbf{x}}_k$ is the mean vector of group k, and n_k is the number of vectors in group k. Thus, we want to choose C to minimize the following equation

$$\min_{C}\sum_{k=1}^{K}\sum_{C(\mathbf{x}_i)=k}\|\mathbf{x}_i - \overline{\mathbf{x}}_k\|^2. \tag{2.38}$$

Another fact is that $\sum_{C(\mathbf{x}_i)=k} \|\mathbf{x}_i - c_k\|^2$ is minimized by $c_k = \overline{\mathbf{x}}_k$. Thus, the problem is the same as minimizing the following equation

$$\min_{C,c_1,\cdots,c_k} \sum_{k=1}^{K} \sum_{C(x_i)=k} \|\mathbf{x}_i - c_k\|^2. \tag{2.39}$$

Algorithm 1: K-means Clustering

1: Pick K points at random from $\mathbf{x}_1, \cdots, \mathbf{x}_M$ as the initial cluster centers $c_1, \cdots, c_k$, then repeat:

- 1 Minimize over C: for $i \in \{1, \cdots, M\}$, classify $\mathbf{x}_i$ based on the closest c_k (i.e. $C(\mathbf{x}_i) = k$).
- 2 Minimize over $c_1, \cdots, c_K$: for $k \in \{1, \cdots, K\}$, let the average vector of cluster k be the new cluster center ($c_k = \overline{\mathbf{x}}_k$).

2: End the iteration until within-cluster variation $\sum_{k=1}^{K} \sum_{C(\mathbf{x}_i)=k} \|\mathbf{x}_i - c_k\|^2$ doesn't change.

The K-means clustering algorithm minimizes Equation (2.38) by alternately minimizing over $C, c_1, \cdots, c_K$. In words, K-means classifies each vector based on the closest center, then calculates the new average vector in each cluster as each cluster's new center. One of the disadvantages of K-means is that there is a need to choose the number of clusters, which is usually an unknown prior.

2.6 Programming Language and Computational Package

Here, we list the programming language and the computational packages we continuously used through the thesis:

- logistic regression: we use `glm` function in `R`,
- penalized logistic regression: we use `glmnet` package in `R`,
- multinomial logistic regression: we use `mnlogit` package in `R`,

- AUC and H-measure: we use `hmeasure` package in `R` with the default setting for calculating the H-measure, i.e. $\{\alpha = \pi_1 + 1, \beta = \pi_0 + 1, b = \pi_0, c = \pi_1\}$ in Equation (2.35),
- the genetic algorithm: we use `GA` package in `R`,
- SMOTE oversampling: we use `SMOTE` function in `DMwR` package in `R`.

2.7 Summary

This chapter introduces the relevant background material on the probability theory, classification methods and the corresponding performance metric, clustering algorithms, and high class imbalance problem. These form the foundation of the main contents of the thesis.

As introduced in Section 2.3, modeling imbalanced data is a challenging problem, fraught with difficulties. The literature suggests that class imbalance is not well understood, and no single method is apparently superior to the others, in general. The exception is logistic regression, for which deeper mathematical insights are available, as we will discuss in the next chapter.

3

Highly Imbalanced Logistic Regression

Logistic regression is designed for modeling the posterior probability of each class (Section 2.2.2) and widely used in a wide range of fields by reason of its strong theoretical underpinning and high interpretability. As discussed in Section 2.3, the class imbalance is a common problem in the real world and becoming more widespread because of the increasing availability of data. In this chapter, we concentrate on logistic regression and related methods in highly imbalanced data sets.

Owen [2007] provides a striking asymptotic result which suggests that, in cases of extreme class imbalance, the minority class only contributes to the logistic regression estimation via its sample mean vector. This raises concerns about the utility of such models, and potential unwanted consequences. Two natural choices to alleviate these problems are penalizing and weighting the likelihood [Wang et al., 2015, King and Zeng, 2001b]. However, by extending Owen's result, we show that penalizing and weighting the likelihood are insufficient for handling the class imbalance problem. In fact, penalized logistic regression makes matters worse. A similar result for multinomial lo-

gistic regression with a specific defined highly imbalanced multi-class setting also being discussed. This result is a preliminary for our proposed relabeling approach in Chapter 4.

The outline of this chapter is as follows. The first section provides the background to highly imbalanced logistic regression. Section 3.2 and Section 3.3 consider methods that extend logistic regression. New theorems are given for infinitely imbalanced weighted logistic regression and penalized logistic regression. Infinitely imbalanced multinomial logistic regression theory is given in Section 3.4. Part of this chapter are from our published paper [Li et al., 2019].

3.1 Introduction to Infinitely Imbalanced Logistic Regression

In this section, we introduce the boundary behavior of logistic regression in the infinitely imbalanced data set [Owen, 2007] and a result about the existence of the maximum likelihood estimate (MLE) for logistic regression [Silvapulle, 1981]. They are the preparation needed for further investigation of highly imbalanced logistic regression in Sections 3.2 and 3.3.

3.1.1 Silvapulle's Results about Existence of MLE for Logistic Regression

Here, and when convenient in the sequel, we use the following notation: consider n p-dimensional feature vectors from class $Y = 1$ (the minority class), denoted by $\mathbf{x}_{11}, \cdots, \mathbf{x}_{1n}$, and N feature vectors from class $Y = 0$, $\mathbf{x}_{01}, \cdots, \mathbf{x}_{0N}$. In order to accommodate the intercept term in the regression parameters, let $\mathbf{z}_{0i} = (1, \mathbf{x}_{0i})$ for $i = (1, \cdots, N)$ and $\mathbf{z}_{1i} = (1, \mathbf{x}_{1i})$ for $i = (1, \cdots, n)$. Let S, F be the two relative interiors of the convex cones [Rockafellar, 2015, p. 10] generated by $\mathbf{x}_{11}, \cdots, \mathbf{x}_{1n}$ and $\mathbf{x}_{01}, \cdots, \mathbf{x}_{0N}$

respectively,

$$S = \left\{ \sum_{i=1}^{n} k_i \mathbf{z}_{1i} | k_i > 0 \right\} \text{ and } F = \left\{ \sum_{i=1}^{N} k_i \mathbf{z}_{0i} | k_i > 0 \right\}. \tag{3.1}$$

When $S \cap F \neq \emptyset$, then an unique MLE for logistic regression exists. However, if $S \cap F = \emptyset$ then no MLE exists [Silvapulle, 1981]. Put differently, the MLE for logistic regression only exists if the classes are not linearly separable. This result is required for the arguments of the following sections.

3.1.2 Owen's Results about Infinitely Imbalanced Logistic Regression

Here, we introduce the Owen [2007] result about the limit behavior of logistic regression in infinitely imbalanced problems. Following the notation in the previous section, we focus on the case when $n \ll N$. To demonstrate the result, Owen centers logistic regression around the minority class mean vector $\overline{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_{1i}/n$. Since in the infinitely imbalanced case $N \to \infty$, Owen also supposes that there is a good approximation for the conditional distribution of $\mathbf{x}$ given $Y = 0$ (majority class); denoted by F_0. Thus, it is shown that the log-likelihood function (2.10), can be written as

$$l(\beta_0, \beta) = n\beta_0 - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + (\mathbf{x}_i - \overline{\mathbf{x}})^T \beta}) - N \int \log(1 + e^{\beta_0 + (\mathbf{x} - \overline{\mathbf{x}})^T \beta}) dF_0(\mathbf{x}). \tag{3.2}$$

It is convenient to separate the intercept and slope terms in the parameter vector, thus β_0 denotes the intercept term, β denotes the slope terms, and $\mathbf{x}_i$ denotes minority class data only. Note that from here on, in this chapter, $\{\mathbf{x}_i, i \in [1, 2, \cdots, n]\}$ denotes only minority class data. Owen proposes the following definition to express the overlap condition (Section 3.1.1) in the case of infinitely imbalanced logistic regression.

Definition 1. *(Definition 3 in [Owen, 2007]) The distribution F on R^p has*

a point $\mathbf{x}_\star$ *surrounded if*

$$\int_{(\mathbf{x}-\mathbf{x}_\star)^T\psi\geq\epsilon} dF(\mathbf{x}) > \delta \tag{3.3}$$

holds for some $\epsilon > 0$*, some* $\delta > 0$ *and all* $\psi \in \Psi$*. Here* $\Psi = \{\psi \in R^d | \psi\psi^T = 1\}$*.*

An interesting example of the majority class distribution F_0 fails to satisfy this surrounded condition can be described with the following categorical variable example. Assume we have a univariate predictor variable $X \in \{0, 1\}$;

- for the minority class $Y = 1$ cases, X will always be 0 and never be 1,
- for the majority class $Y = 0$ cases, X will take 0 or 1.

Then the minority class mean $\bar{x} = 1$, stands on the boundary of the support of the distribution F_0. In this particular case, ψ should be either 1 or -1. When $\psi = -1$, for any $\epsilon > 0$, the integration (3.3) will either not exist or be 0, thus it will never greater than δ for some $\delta > 0$, which indicates F_0 fails to satisfy the surrounded condition.

Owen also assumes that

$$\int e^{\mathbf{x}^T\beta}(1 + \|\mathbf{x}\|)dF_0(\mathbf{x}) < \infty \tag{3.4}$$

for all $\beta \in R^p$, to ensure that the F_0 does not have tails that are so heavy that a degenerate logistic regression will arise. Then the main result is

Theorem 2. *(Theorem 8 in [Owen, 2007]) Let* $n \geq 1$*, and* $\mathbf{x}_1, \cdots, \mathbf{x}_n \in R^p$ *be fixed. Suppose that* F_0 *satisfies the tail condition (Equation 3.4) and surrounds the class mean vector* $\overline{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$ *as described in Definition 1. Then the maximizer* $\widehat{\beta}$ *of* $l(\beta_0, \beta)$ *given by Equation (3.2) satisfies*

$$\lim_{N\to\infty} \frac{\int e^{\mathbf{x}^T\widehat{\beta}}\mathbf{x}dF_0(\mathbf{x})}{\int e^{\mathbf{x}^T\widehat{\beta}}dF_0(\mathbf{x})} = \overline{\mathbf{x}}. \tag{3.5}$$

An immediate consequence of Theorem 2 is:

Corollary 3. *When $N \to \infty$, under the same conditions in Theorem 2, logistic regression only depends on the minority class data $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ through the minority class mean vector $\overline{\mathbf{x}}$*

Proof. As $N \to \infty$, the maximizer $\hat{\beta}$ of the log-likelihood function (3.2) satisfy

$$g(\widehat{\beta}) = \frac{\int e^{\mathbf{x}^T\widehat{\beta}}\mathbf{x}dF_0(\mathbf{x})}{\int e^{\mathbf{x}^T\widehat{\beta}}dF_0(\mathbf{x})} - \overline{\mathbf{x}} = 0.$$

Function $g(\widehat{\beta})$ is specified by $F_0(\mathbf{x})$ and $\bar{\mathbf{x}}$, thus the solution of $g(\widehat{\beta}) = 0$ only depends on $F_0(\mathbf{x})$ and $\bar{\mathbf{x}}$. □

This theorem can be further understand as we could replace $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ by one vector, the mean vector of the minority class, and obtain the same coefficient estimates of β in the limit $N \to \infty$. Theorem 2 is an asymptotic theoretical result and Owen's simulation [Owen, 2007, p.763 Table 1] shows that the convergence phenomena happens quickly when $N/n > 100$.

3.1.3 Does Owen's Results Really Matter?

Theorem 2 is a potentially worrying finding since it suggests the broader distributional structure in the minority class is not taken into account by highly imbalanced logistic regression. Owen [2007] also mentions the issues particularly related to the presence of cluster structure in the minority class, of which the overall minority class mean vector would be a poor representation. In this section, we demonstrate this problem with different simulations. To be specific, we want to compare the logistic regression performance with different number of the majority class observations N when **the minority class** has

1. **no** cluster structure, Gaussian distributed with the same variance as the majority class,

2. **no** cluster structure, Gaussian distributed with smaller variance than the majority class,

3. **two close** clusters, and each cluster is Gaussian distributed with the same variance as the majority class,

4. **two close** clusters, and each cluster is Gaussian distributed with smaller variance than the majority class,

5. **two well separated** clusters, and each cluster is Gaussian distributed with the same variance as the majority class,

6. **two well separated** clusters, and each cluster is Gaussian distributed with smaller variance than the majority class,

to answer the question "does Owen's results really matter". We fix the number of the minority class observations $n = 100$ and increase the number of the majority class observations N from 100 to 10000, for making the data more imbalanced; this follows the statement in Theorem 2. We simulate the above six scenarios in a two dimensional space; where

- for no cluster structure scenario: the minority class mean locates at $(1, 1)$,
- for two close clusters scenario: two cluster means locate at $(1.3, 0.7)$ and $(0.7, 1.3)$,
- for two well separated clusters scenario: two cluster means locate at $(0, 2)$ and $(2, 0)$.

We report the the average H-measure and AUC with its corresponding standard deviation in each simulation (each simulation is repeated 1000 times).

Simulation 1: no cluster structure, same variance between the majority and the minority class

In this simulation, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n = 100$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ as the minority class $Y = 1$. Here

$$\mu_0 = [0, 0], \mu_1 = [1, 1],$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The number of the majority class observations N varies in {100, 500, 1000, 5000, 10000}, and for each combination of (n, N), we repeat the simulation 1000 times by training a logistic regression on the training set and apply it on the test set. Table 3.1 gives the results.

Table 3.1: Simulation 1: the average H-measure and AUC with their corresponding standard deviation on the test set (1000 iterations).

N	H-measure	AUC
100	0.4259 (0.0558)	0.8524 (0.0251)
500	0.4012 (0.0442)	0.8526 (0.0199)
1000	0.3923 (0.0422)	0.8522 (0.0195)
5000	0.3841 (0.0400)	0.8527 (0.0185)
10000	0.3825 (0.0393)	0.8524 (0.0184)

Simulation 2: no cluster structure, the minority class has smaller variance than the majority class

In this simulation, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n = 100$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ as the minority class $Y = 0$. Here

$$\mu_0 = [0, 0], \mu_1 = [1, 1],$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}.$$

The number of the majority class observations N varies in {100, 500, 1000, 5000,10000}, and for each combination of (n, N), we repeat the simulation 1000 times by training a logistic regression on the training set and apply it on the test set.

Table 3.2: Simulation 2: the average H-measure and AUC with their corresponding standard deviation on the test set (1000 iterations).

N	H-measure	AUC
100	0.5983 (0.0548)	0.9049 (0.0207)
500	0.5362 (0.0376)	0.9043 (0.0117)
1000	0.5159 (0.0353)	0.9048 (0.0106)
5000	0.4981 (0.0316)	0.9041 (0.0087)
10000	0.4946 (0.0299)	0.9041 (0.0081)

Simulation 1 and 2 show that, in this contrived setting, as the number of the majority class N increases, there is no downward trend in the AUC value. However, we can observe a decline regarding the H-measure.

Simulation 3: the minority class has close cluster structure, same variance between each minority class cluster and the majority class

In this simulation, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n_1 = 50$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ and $n_2 = 50$ points are generated following $X \sim N(\mu_2, \Sigma_2)$. n_1 and n_2 are combined as the minority class $Y = 0$ (i.e. $n = n_1 + n_2 = 100$ minority class observations). Here

$$\mu_0 = [0, 0], \mu_1 = [1.3, 0.7], \mu_2 = [0.7, 1.3]$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The number of the majority class observations N varies in {100, 500, 1000, 5000, 10000}, and for each combination of (n, N), we repeat the simulation 1000 times by training a logistic regression on the training set and apply it on the test set.

Table 3.3: Simulation 3: the average H-measure and AUC with their corresponding standard deviation on the test set (1000 iterations).

N	H-measure	AUC
100	0.6566 (0.0536)	0.9432 (0.0154)
500	0.6326 (0.0406)	0.9396 (0.0120)
1000	0.6251 (0.0380)	0.9391 (0.0112)
5000	0.6189 (0.0373)	0.9392 (0.0111)
10000	0.6192 (0.0371)	0.9390 (0.0110)

Simulation 4: the minority class has close cluster structure, and each minority class cluster has smaller variance than the majority class

In this simulation, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n_1 = 50$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ and $n_2 = 50$ points are generated following $X \sim N(\mu_2, \Sigma_2)$. n_1 and n_2 are combined as the minority class $Y = 0$ (i.e. $n = n_1 + n_2 = 100$ minority class observations). Here

$$\mu_0 = [0, 0], \mu_1 = [1.3, 0.7], \mu_2 = [0.7, 1.3]$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}.$$

The number of the majority class observations N varies in {100, 500, 1000, 5000, 10000}, and for each combination of (n, N), we repeat the simulation 1000 times by training a logistic regression on the training set and apply it on the test set.

Table 3.4: Simulation 4: the average H-measure and AUC with their corresponding standard deviation on the test set (1000 iterations).

N	H-measure	AUC
100	0.8209 (0.0447)	0.9763 (0.0103)
500	0.7895 (0.0306)	0.9741 (0.0057)
1000	0.7836 (0.0266)	0.9740 (0.0045)
5000	0.7749 (0.0241)	0.9734 (0.0035)
10000	0.7716 (0.0230)	0.9731 (0.0033)

Simulation 3 and 4 show that, in this contrived setting, as the number of the majority class N increases, there is a downward trend in AUC. For example, the difference of the AUC value between $N = 100$ and $N = 1000$ is greater than one times their corresponding standard deviation in Table 3.4. The H-measure value also decreases when N increases.

Simulation 5: the minority class has well separated cluster structure, same variance between each minority class cluster and the majority class

In this simulation, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n_1 = 50$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ and $n_2 = 50$ points are generated following $X \sim N(\mu_2, \Sigma_2)$. n_1 and n_2 are combined as the minority class $Y = 0$ (i.e. $n = n_1 + n_2 = 100$ minority class observations). Here

$$\mu_0 = [0, 0], \mu_1 = [0, 2], \mu = [2, 0]$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The number of the majority class observations N varies in {100, 500, 1000, 5000, 10000}, and for each combination of (n, N), we repeat the simulation 1000 times by training a logistic regression on the training set and apply it on the test set.

Table 3.5: Simulation 5: the average H-measure and AUC with their corresponding standard deviation on the test set (1000 iterations).

N	H-measure	AUC
100	0.4290 (0.0575)	0.8578 (0.0257)
500	0.4025 (0.0458)	0.8534 (0.0207)
1000	0.3942 (0.0418)	0.8532 (0.0192)
5000	0.3810 (0.0399)	0.8515 (0.0187)
10000	0.3820 (0.0397)	0.8511 (0.0183)

Simulation 6: the minority class has has well separated clusters, and each minority class cluster has smaller variance than the majority class

In this simulation, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n_1 = 50$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ and $n_2 = 50$ points are generated following $X \sim N(\mu_2, \Sigma_2)$. n_1 and n_2 are combined as the minority class $Y = 0$ (i.e. $n = n_1 + n_2 = 100$ minority class observations). Here

$$\mu_0 = [0, 0], \mu_1 = [0, 2], \mu = [2, 0]$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}.$$

The number of the majority class observations N varies in {100, 500, 1000, 5000, 10000}, and for each combination of (n, N), we repeat the simulation 1000 times by training a logistic regression on the training set and apply it on the test set.

Table 3.6: Simulation 6: the average H-measure and AUC with their corresponding standard deviation on the test set (1000 iterations)

N	H-measure	AUC
100	0.6007 (0.0568)	0.9107 (0.0216)
500	0.5326 (0.0377)	0.9054 (0.0119)
1000	0.5156 (0.0332)	0.9048 (0.0099)
5000	0.4971 (0.0306)	0.9039 (0.0085)
10000	0.4934 (0.0305)	0.9035 (0.0084)

Simulation 5 and 6 show that, in this contrived setting, as the number of the majority class N increases, there is a downward trend in the AUC. The same as Simulations 3 and 4, the difference of the AUC value between $N = 100$ and $N = 1000$ is greater than one times their corresponding standard deviation. The H-measure also decreases when N increases. In fact, we see the H-measure always decreases as N increases from Simulation 1 to Simulation 6.

We conclude the results from Simulation 1 to 6 in Figure 3.1. In conclusion, by considering all simulations together, we can find that:

- when using the AUC as our performance metric, there is a downward trend when cluster structure is present in the minority class as N increases,
- when using the H-measure as our performance metric, a downward trend always exists, as N increases.

This evidence show that Theorem 2 does have the influence on the predictive performance of logistic regression. Actually, as pointed out in Owen [2007], it is reasonable to expect better results from logistic regression when we can detect pronounced cluster structure among the minority class, and further using a multi-class model. This point will be explored in the next chapter.

It is worth noting that the H-measure decreases in all simulations. This is suspicious. Without knowing more information about the misclassification costs, we use the default setting of $\{b = \pi_0, c = \pi_1\}$ (see Equation 2.35) proposed by Hand [2009], which are the prior proportion of both classes in a binary data set. Thus, as N increases, $\{\pi_0, \pi_1\}$ will be different, which brings the doubt on the utility of comparing the H-measure between balanced and imbalanced cases. Of course, we need to point out that, when comparing the model performance on the same highly imbalanced data, using the H-measure is still trustworthy.

3.2 Infinitely Imbalanced Weighted Logistic Regression

As mentioned in Section 2.3, a natural idea to address the imbalance problem would be to reweight the likelihood, leading to weighted logistic regression. Zeng [2017] proved that the MLE for *weighted* logistic regression exists and is unique when there is an overlap between the data points of the two classes, which extends the result in Silvapulle [1981], described in Section 3.1.1. Next, we identify the characteristics of infinitely imbalanced weighted logistic regression for a specific weighting strategy.

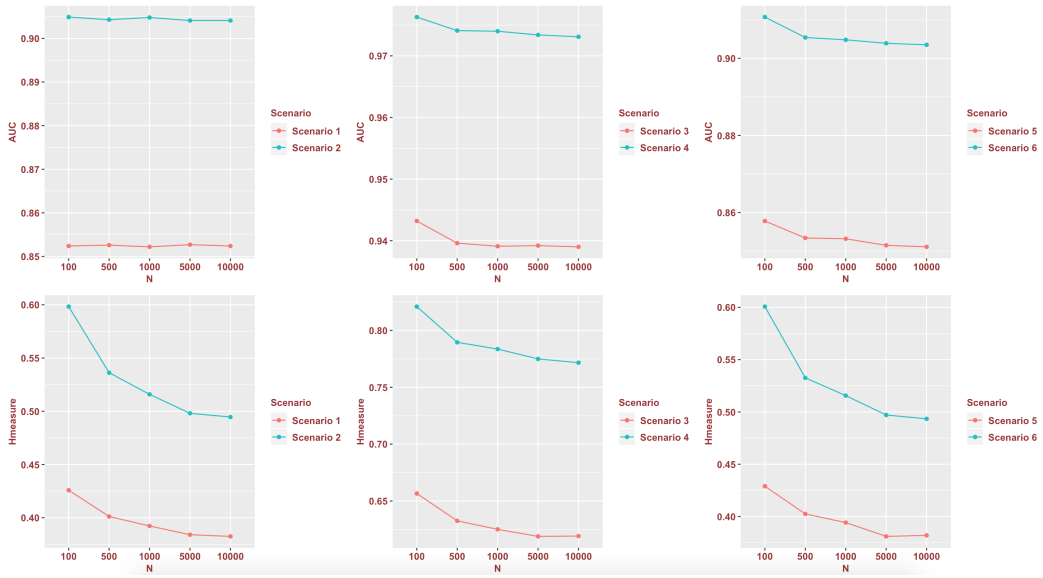


Figure 3.1: The results of the AUC and the H-measure in Similation 1 to Simulation 6.

Here, we consider a particular weighting scheme; retaining a weight of one for the majority class, as in Equation (3.2), and increasing the weight of observations in the minority class, with weights $\Omega = \{\omega_i > 1, i = 1, \cdots, n\}$. Thus, the log-likelihood function for weighted logistic regression is

$$\begin{aligned} l(\beta_0, \beta; \Omega) = &\sum_{i=1}^{n} \omega_i \beta_0 - \sum_{i=1}^{n} \omega_i \log(1 + e^{\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \beta}) \\ &- N \int \log(1 + e^{\beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \beta}) dF_0(\mathbf{x}), \end{aligned} \tag{3.6}$$

where $\bar{\mathbf{x}}$ is a weighted minority class mean vector $\bar{\mathbf{x}} = \sum_{i=1}^{n} \omega_i \mathbf{x}_i / \sum_{i=1}^{n} \omega_i$. We investigate the characteristics of infinite imbalance for this form of weighted logistic regression and sketch the whole proof before expanding on the relevant lemmas and theorems. We follow the lemmas and theorems given by Owen [2007]. Lemma 2 and Lemma 4 in Owen [2007] are numerical facts which also hold for this weighted logistic regression. Lemma 5 in Owen [2007] is used to prove the existence of a finite MLE when the surrounded condition (Definition 1) is satisfied. Lemma 6 and Lemma 7 in Owen [2007] are used to establish a boundary of $\left\|\hat{\beta}\right\|$ when $N \to \infty$. The following lemmas and theorem are our original work.

Lemma 4. *(Lemma 5 in [Owen, 2007]) Suppose that weights Ω are fixed, $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ are given and F_0 satisfies the surrounded condition (i.e. Equation 3.3) at the weighted minority class mean vector $\bar{\mathbf{x}}$. Then the log-likelihood function $l(\beta_0, \beta; \Omega)$ has a unique finite maximizer $(\hat{\beta}_0, \hat{\beta})$, when $0 < N < \infty$.*

Proof. The key point is to show that, for some direction vector, $\lambda > 0$, "$\frac{\partial}{\partial\lambda} l(\lambda\beta_0, \lambda\beta; \Omega)$ is always strictly negative" [Owen, 2007], which excludes the situation that $l(\beta_0, \beta; \Omega)$ will increase infinitely along some ray.

If $\beta = 0$, we assume that $0 < |\beta_0| < \epsilon/2$, ϵ comes from Definition 1. Thus,

$$\lim_{\lambda\to\infty} \frac{\partial}{\partial\lambda} l(\lambda\beta_0, \lambda\beta; \Omega) = \begin{cases} -N\beta_0, & \text{if } \beta_0 > 0 \\ \sum_{i=1}^n \omega_i \beta_0, & \text{if } \beta_0 < 0 \end{cases},$$

which means $\lim_{\lambda\to\infty} \frac{\partial}{\partial\lambda} l(\lambda\beta_0, \lambda\beta; \Omega) < 0$ when $\beta = 0$.

If $\beta \neq 0$, without loss of generality we may take $\beta^T\beta = 1$ and $0 \leq |\beta_0| < \epsilon/2$, ϵ comes from Definition 1, by following the ray back to the origin. Using $\sum_{i=1}^n \omega_i\beta_0 = \sum_{i=1}^n \omega_i[\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T\beta]$, we have

$$\begin{aligned} \lim_{\lambda\to\infty} \frac{\partial}{\partial\lambda} l(\lambda\beta_0, \lambda\beta; \Omega) = & \sum_{i:\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T\beta < 0} \omega_i[\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T\beta] \\ & - N \int_{\beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T\beta > 0} \beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T\beta dF_0(\mathbf{x}). \end{aligned} \tag{3.7}$$

The term involving a sum in this equation is smaller than or equal to 0, and the integral term is greater than 0 (equal to 0 is excluded by the surrounded condition). □

Lemma 5. *(Lemma 6 in [Owen, 2007]) Let $\hat{\beta}_0$ and $\hat{\beta}$ be the maximizer of the likelihood function, F_0 satisfies the surrounded condition at the weighted minority class mean vector $\bar{\mathbf{x}}$ and η is the infimum of δ. Then for any $N \geq 2\sum\omega_i/\eta$, we have $e^{\hat{\beta}_0} \leq 2\sum\omega_i/(N\eta)$.*

Proof. Let $e^{\beta_0} = A/N$, $0 < A < \infty$.

$$\begin{aligned}
\frac{\partial l}{\partial \beta_0} &= \sum_{i=1}^{n} \omega_i - \sum_{i=1}^{n} \omega_i \frac{e^{\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \beta}}{1 + e^{\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \beta}} - N \int \frac{e^{\beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \beta}}{1 + e^{\beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \beta}} dF_0(\mathbf{x}) \\
&= \sum_{i=1}^{n} \omega_i - \sum_{i=1}^{n} \omega_i \frac{AN^{-1} e^{(\mathbf{x}_i - \bar{\mathbf{x}})^T \beta}}{1 + AN^{-1} e^{(\mathbf{x}_i - \bar{\mathbf{x}})^T \beta}} - \int \frac{A e^{(\mathbf{x} - \bar{\mathbf{x}})^T \beta}}{1 + AN^{-1} e^{(\mathbf{x} - \bar{\mathbf{x}})^T \beta}} dF_0(\mathbf{x}) \\
&\leq \sum_{i=1}^{n} \omega_i - A \int_{(\mathbf{x} - \bar{\mathbf{x}})^T \beta > 0} \frac{e^{(\mathbf{x} - \bar{\mathbf{x}})^T \beta}}{1 + AN^{-1} e^{(\mathbf{x} - \bar{\mathbf{x}})^T \beta}} dF_0(\mathbf{x}) \\
&\leq \sum_{i=1}^{n} \omega_i - \frac{A\eta}{1 + A/N}.
\end{aligned} \tag{3.8}$$

Because $\eta \leq \delta$ and $\delta < \int_{(\mathbf{x} - \mathbf{x}_\star)^T \psi \geq \epsilon} dF(\mathbf{x}) \leq 1$ (see Definition 1), thus η is bounded in $0 < \eta \leq \delta < 1$. Since $N \geq 2 \sum \omega_i / \eta$, if we let $e^{\beta_0} > 2 \sum \omega_i / (N\eta)$, we will have $A > 2 \sum \omega_i / \eta$. Then, $\partial l / \partial \beta_0 < 0$. For the concave likelihood function, the negative partial derivative means that the maximizer $\hat{\beta}_0 < \log(2 \sum \omega_i / \eta N)$. □

Lemma 6. *(Lemma 7 in [Owen, 2007]) Under the same conditions as Lemma 5, we will have* $\limsup_{N \to \infty} \left\| \hat{\beta} \right\| < \infty$.

Proof. Under the surrounded condition (Equation 3.3), there exists a γ such that

$$\inf_{\psi \in \Psi} \int [(\mathbf{x} - \bar{\mathbf{x}})^T \psi]_+ dF_0(\mathbf{x}) \geq \gamma > 0, \tag{3.9}$$

where $\psi^T \psi = 1$ and $[(\mathbf{x} - \bar{\mathbf{x}})^T \psi]_+$ means the positive part of $[(\mathbf{x} - \bar{\mathbf{x}})^T \psi]$. We

still let $e^{\beta_0} = A/N$, then we have

$$\begin{aligned} l(\beta_0, 0; \Omega) - l(\beta_0, \beta; \Omega) &= -(\sum_{i=1}^{n} \omega_i + N)\log(1 + e^{\beta_0}) + \sum_{i=1}^{n} \omega_i \log(1 + e^{\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \beta}) \\ &\quad + N \int \log(1 + e^{\beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \beta}) dF_0(\mathbf{x}) \\ &> -(\sum_{i=1}^{n} \omega_i + N) e^{\beta_0} + \frac{e^{\beta_0} N}{1 + e^{\beta_0}} \int_{(\mathbf{x} - \bar{\mathbf{x}})^T \beta \geq 0} (\mathbf{x} - \bar{\mathbf{x}})^T \beta dF_0(\mathbf{x}) \\ &\geq -(\sum_{i=1}^{n} \omega_i + N) \frac{A}{N} + \frac{AN}{N + A} \|\beta\| \gamma. \end{aligned} \tag{3.10}$$

The first inequality above is taken from Lemma 4 in Owen [2007].

Equation (3.10) implies that when $\|\beta\| > \frac{1}{\gamma}(1 + A/N)(1 + \sum \omega_i / N)$, we have $l(\beta_0, \beta; \Omega) < l(\beta_0, 0; \Omega)$. Thus, maximizing likelihood function $l(\beta_0, \beta; \Omega)$ will obviously let $\left\| \hat{\beta} \right\| \leq 2/\gamma$, with large enough N. □

Now, we prove the main theorem for infinitely weighted logistic regression:

Theorem 7. *Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ and the weights Ω be fixed. If F_0 satisfies the surrounding condition at $\bar{\mathbf{x}}$ and the tail conditions, then the maximizer $(\hat{\beta}_0, \hat{\beta})$ of $l(\beta_0, \beta; \Omega)$ satisfies*

$$\lim_{N \to \infty} \frac{\int \mathbf{x} e^{\mathbf{x}^T \hat{\beta}} dF_0(\mathbf{x})}{\int e^{\mathbf{x}^T \hat{\beta}} dF_0(\mathbf{x})} = \bar{\mathbf{x}}. \tag{3.11}$$

Proof. Set $\frac{\partial l}{\partial \beta} = 0$ we have,

$$-\sum_{i=1}^{n} \omega_i \frac{(\mathbf{x}_i - \bar{\mathbf{x}}) e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}} - N \int \frac{(\mathbf{x} - \bar{\mathbf{x}}) e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}} dF_0(\mathbf{x}) = 0. \tag{3.12}$$

Divide Equation (3.12) by $N e^{\hat{\beta}_0 - \bar{\mathbf{x}}^T \hat{\beta}}$ to yield

$$\int \frac{(\mathbf{x} - \bar{\mathbf{x}}) e^{\mathbf{x}^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}} dF_0(\mathbf{x}) = -\frac{1}{N} \sum_{i=1}^{n} \omega_i \frac{(\mathbf{x}_i - \bar{\mathbf{x}}) e^{\mathbf{x}_i^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}}. \tag{3.13}$$

Although the introduction of weights ω_i will lead to slower convergence than the unweighted logistic regression, the right side of Equation (3.13) still vanishes as $N \to \infty$ (because $\left\|\hat{\beta}\right\|$ is bounded by Lemma 6). In the left side of Equation (3.13), $\int e^{\mathbf{x}^T\hat{\beta}}[1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})\hat{\beta}}]^{-1}dF_0(\mathbf{x})$ is at most $\int e^{\mathbf{x}^T\hat{\beta}}dF_0(\mathbf{x})$ and is at least

$$\int e^{\mathbf{x}^T\hat{\beta}}[1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})\hat{\beta}}]^{-1}dF_0(\mathbf{x}) \to \int e^{\mathbf{x}^T\hat{\beta}}dF_0(\mathbf{x}),$$

as $N \to \infty$ because $\hat{\beta}_0 \to -\infty$ (see Lemma 5) and $\int e^{2\mathbf{x}^T\hat{\beta}}dF_0(\mathbf{x}) < \infty$ by the tail condition. Similarly we have

$$\int \mathbf{x}e^{\mathbf{x}^T\hat{\beta}}[1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})\hat{\beta}}]^{-1}dF_0(\mathbf{x}) \to \int \mathbf{x}e^{\mathbf{x}^T\hat{\beta}}dF_0(\mathbf{x}).$$

Thus, Equation (3.13) simplifies to

$$\bar{\mathbf{x}}\int e^{\mathbf{x}^T\hat{\beta}}dF_0(\mathbf{x}) - \int \mathbf{x}e^{\mathbf{x}^T\hat{\beta}}dF_0(\mathbf{x}) = 0, \text{ as } N \to \infty. \tag{3.14}$$

Thus, Equation (3.11) holds. □

Theorem 7 shows that this specific weighted logistic regression still only depends on the weighted minority mean vector in the class imbalance limit, $N \to \infty$. This finding has interesting implications for methods based on resampling the minority class.

3.3 Infinitely Imbalanced Penalized Logistic Regression

The previous section shows that resampling the minority class is insufficient for handling the class imbalance problem. Penalized logistic regression is another widely used method to handle imbalanced logistic regression problem [Wang et al., 2015, Ahmed et al., 2018]. It is known that maximum likelihood estimation may fail with high dimensional data or multiple highly-correlated variables [Efron and Hastie, 2016, p.303]. In order to perform parameter shrinkage and variable selection, penalized logistic regression is

designed to add penalty terms to the likelihood function of logistic regression. We might propose them as solutions to alleviate the imbalance problem. In this section, we consider the two common forms of penalty: the ridge and the lasso. For data as described in Section 3.1.2, we can give the objective function for ridge penalized logistic regression [Hoerl and Kennard, 1970] as:

$$\begin{aligned} l(\beta_0, \beta) =& \frac{1}{n+N}\left[\sum_{i=1}^{n} \log \frac{e^{\beta_0 + \mathbf{x}_{1i}^T \beta}}{1 + e^{\beta_0 + \mathbf{x}_{1i}^T \beta}} + \sum_{i=1}^{N} \log \frac{1}{1 + e^{\beta_0 + \mathbf{x}_{0i}^T \beta}}\right] \\ & - \frac{1}{2}\lambda \|\beta\|_2^2, \text{ where } \lambda > 0. \end{aligned} \tag{3.15}$$

Lasso penalized logistic regression has the same form except the penalty term at the end of the expression is given as $\lambda\|\beta\|_1$.

We first describe two simulations to explore the characteristics of penalized logistic regression in highly imbalanced data. In **Simulation A**, we consider samples of different size and different levels of imbalance, where $P(X|Y = 0) \sim N(0, 1)$ (the majority class), and we have 100 replicates from the $Y = 1$ class, all with $X = 1$. The role of the replicates is to handle computational issues*. As N (the number of majority class samples) increases, the problem becomes more imbalanced. The coefficient estimates of standard logistic regression and ridge penalized logistic regression (penalty parameter $\lambda = 0.1$) are given in Table 3.7. The table suggests that, as $N \to \infty$,

- Ne^{β_0} converges to n,
- β converges to 0,

in ridge penalized logistic regression, with this particular and arbitrary choice of penalty parameter. Actually, this is the behavior we want to prove in the next section.

However, in **Simulation B**, we consider $X \sim \text{Uniform}(0, 1)$ when $Y = 0$ (the majority class). We use $n = 100$ points for $Y = 1$; 50 points are $X = 0.5$

*We use the `glmnet` package in `R` programming language to implement penalized logistic regression. This package requires at least 8 points in each class. The replication of the single minority class observation can resolve this problem and not influence the coefficient estimates of the slope vector β.

Table 3.7: Simulation A for infinitely imbalanced penalized logistic regression. N observations in majority class ($Y = 0$) following $X \sim N(0, 1)$ and 100 observations in minority class with $Y = 1, X = 1$. As N (the number of the majority class samples) increases, the problem becomes more imbalanced, we can find Ne^{β_0} converges to n and β converges to 0 in ridge penalized logistic regression.

	Logistic Regression		Logistic Regression + Ridge	
N	β	Ne^{β_0}	β	Ne^{β_0}
100	1.1215	41.7805	0.6879	59.1750
1000	0.5656	65.3495	0.2454	85.5127
10000	0.5013	68.3830	0.0450	97.6581
100000	0.5007	68.6940	0.0049	99.7516
1000000	0.5001	68.7254	0.0005	99.9750
converge to	certain value k_1	certain value k_2	0	$n = 100$

and the others are $X = 2$. This setting fails the surrounded condition in Theorem 2, because the minority class mean $\bar{x} = 1.25$ is located outside the support of the majority class distribution. Table 3.8 shows the coefficient estimates of this simulation. We see that penalized logistic regression demonstrates shrinkage behavior, despite failing to satisfy the surrounded condition.[†]

3.3.1 Theoretical Results

In this section, we give results, following Owen, for penalized logistic regression in the infinitely imbalanced case. Considering the existence of a unique solution in penalized logistic regression, we follow the previous argument for logistic regression considered by Silvapulle [1981].

Theorem 8. *For data as described in Section 3.1.1, let* $\mathbf{x}_{1i}, i \in \{1, \cdots, n\}$ *denote p-dimensional feature vectors from class* $Y = 1$*, and* $\mathbf{x}_{0i}, i \in \{1, \cdots, N\}$ *denote p-dimensional feature vectors from class* $Y = 0$*. Assume that the* $n + N$ *by* $p + 1$ *data matrix (including the constant vector 1 to accommodate the intercept) has rank* $p + 1$*, then a unique finite ridge or lasso penalized logistic regression solution exists.*

[†]The convergence effect does not become apparent until $N/n > 10$ (see line 1 and line 2 in Table 3.8)

Table 3.8: Simulation B for infinitely imbalanced penalized logistic regression. N observations in majority class ($Y = 0$) following $X \sim \text{Uniform}(0, 1)$ and 100 observations in minority class (half of them with $Y = 1, X = 0.5$, the others with $Y = 1, X = 2$). As N (the number of the majority class samples) increases, the problem becomes more imbalanced, we can find Ne^{β_0} converges to n and β converges to 0 in ridge penalized logistic regression.

	Logistic Regression		Logistic Regression + Ridge	
N	β	Ne^{β_0}	β	Ne^{β_0}
100	2.2347	16.2756	1.2598	34.6374
1000	3.2033	8.4214	1.6478	31.6947
10000	4.6591	2.8035	0.7112	68.0441
100000	6.3475	0.7238	0.0878	95.6659
1000000	8.1866	0.1524	0.0090	99.5517
converge to	no converge	no converge	0	$n = 100$

Proof. We first consider the case of the ridge penalty, then give adjustments for lasso. In proof, we use B denotes $p + 1$ dimensional vector (β_0, β). In order to accommodate the intercept term in the regression parameters, let $\mathbf{z}_{1i} = (1, \mathbf{x}_{1i})$ and $\mathbf{z}_{0i} = (1, \mathbf{x}_{0i})$. We consider situations $S \cap F \neq \emptyset$ and $S \cap F = \emptyset$ separately.

If there is no separation between two convex cones S and F ($S \cap F \neq \emptyset$), we cannot find a hyperplane which separates S and F properly [Rockafellar, 2015, Theorem 11.3]. Therefore, there exists a unit $p+1$ dimensional vector $\mathbf{e}$, such that $\mathbf{z}_{1i}^T\mathbf{e}$ is negative for some $1 \leq i \leq n$ and $\mathbf{z}_{0i}^T\mathbf{e}$ is positive for some $1 \leq i \leq N$ ($\mathbf{z}$ is an observation in the design matrix). Let $A_1 = \{i|1 \leq i \leq n, \mathbf{z}_{1i}^T\mathbf{e} < 0\}$, $A_2 = \{i|1 \leq i \leq n, \mathbf{z}_{1i}^T\mathbf{e} \geq 0\}$, $A_3 = \{i|1 \leq i \leq N, \mathbf{z}_{0i}^T\mathbf{e} < 0\}$ and $A_4 = \{i|1 \leq i \leq N, \mathbf{z}_{0i}^T\mathbf{e} \geq 0\}$. Thus, considering $k \to \infty$,

$$\begin{aligned} l(B + (k+1)\mathbf{e}) - l(B + k\mathbf{e}) =& \frac{1}{n+N} \sum_{i \in A_1} \log \frac{e^{\mathbf{z}_{1i}^T(B+(k+1)\mathbf{e})}}{1 + e^{\mathbf{z}_{1i}^T(B+(k+1)\mathbf{e})}} \times \frac{1 + e^{\mathbf{z}_{1i}^T(B+k\mathbf{e})}}{e^{\mathbf{z}_{1i}^T(B+k\mathbf{e})}} \\ &+ \frac{1}{n+N} \sum_{i \in A_4} \log \frac{1 + e^{\mathbf{z}_{0i}^T(B+k\mathbf{e})}}{1 + e^{\mathbf{z}_{0i}^T(B+(k+1)\mathbf{e})}} \\ &- \frac{1}{2}\lambda(\|\beta + (k+1)\tilde{\mathbf{e}}\|_2^2 - \|\beta + k\tilde{\mathbf{e}}\|_2^2), \end{aligned} \tag{3.16}$$

(where $\tilde{\mathbf{e}}$ is $\mathbf{e}$ excluding the first component) since the summation parts of $i \in A_2$ and $i \in A_3$ will vanish as $k \to \infty$. We also notice that when k goes to infinity, the first term in Equation (3.16) will shrink to $\sum_{i \in A_1} \mathbf{z}_{1i}^T \mathbf{e} < 0$, the second term will shrink to $-\sum_{i \in A_4} \mathbf{z}_{0i}^T \mathbf{e} < 0$ and the third term is positive. Thus $l(B + (k+1)\mathbf{e}) - l(B + k\mathbf{e})$ will be negative for large k, so $l(B + k\mathbf{e})$ is a decreasing function in k. Therefore, $l(\beta_0, \beta)$ for fixed $\mathbf{x}$ does not have a direction of recession and the existence of the solution follows from Theorem 27.1(d) in Rockafellar [2015].

Assuming $S \cap F = \emptyset$, we could find a unit $p+1$ dimensional vector $\mathbf{e}$, such that $\{\mathbf{z}_{1i}^T \mathbf{e} \geq 0; 1 \geq i \geq n\}$ and $\{\mathbf{z}_{0i}^T \mathbf{e} < 0; 1 \geq i \geq N\}$ [Rockafellar, 2015, Theorem 11.3 and Theorem 11.7]. Rewrite the likelihood function (3.15) as:

$$l(B + k\mathbf{e}) = \frac{1}{n+N}\left[\sum_{i=1}^{n} \log \frac{e^{\mathbf{z}_{1i}^T (B+k\mathbf{e})}}{1 + e^{\mathbf{z}_{1i}^T (B+k\mathbf{e})}} + \sum_{i=1}^{N} \log \frac{1}{1 + e^{\mathbf{z}_{0i}^T (B+k\mathbf{e})}}\right] - \frac{1}{2}\lambda \|\beta + k\tilde{\mathbf{e}}\|_2^2. \tag{3.17}$$

Considering two numerical facts:

- $f_1(x) = \log \frac{e^x}{1+e^x}$ is a increasing function and $\lim_{x \to \infty} f_1(x) = 0$,
- $f_2(x) = \log \frac{1}{1+e^x}$ is a decreasing function and $\lim_{x \to -\infty} f_2(x) = 0$,

we know that the summation part in Equation (3.17) is increasing and approaching 0, as $k \to \infty$. Thus, term $-\frac{1}{2}\lambda \|\beta + k\tilde{\mathbf{e}}\|_2^2$ dominates Equation (3.17) as $k \to \infty$, leading to the result that $l(B + k\mathbf{e})$ is a decreasing function for large enough k. Therefore, $l(\beta_0, \beta)$ for fixed $\mathbf{x}$ does not have a direction of recession and the existence of the solution still follows from Theorem 27.1(d) in Rockafellar [2015]. This proof shows the existence of the solution for ridge penalized logistic regression. Following [Tibshirani, 2013, Lemma 5] and [van Wieringen, 2015, p.45], we know the extant solution for ridge penalized logistic regression is unique.

Through a similar proof process, Theorem 8 also holds with the lasso penalty.

Table 3.9: Infinitely imbalanced logistic regression shrinkage law.

Fixture	Logistic Regression	Ridge	Lasso
β_0	$-\infty$	$-\infty$	$-\infty$
β	certain value, k_1	0	0
Ne^{β_0}	certain value, k_2	n	n

The only change is for handling the penalty term in Equation (3.16):

$$\begin{aligned}\lambda\sum_{j=1}^{p}|\beta_j+k+1|-\lambda\sum_{j=1}^{p}|\beta_j+k| &= \lambda\sum_{j=1}^{p}\Big[|\beta_j+k+1|-|\beta_j+k|\Big]\\ &= \lambda p,\\ &\text{when } k\to\infty, \text{and } \beta_j \text{ is the } j\text{th element of } \beta,\end{aligned} \tag{3.18}$$

which is still a positive number. □

By repeating the previous numerical simulation (Table 3.7 and Table 3.8) in lasso penalized logistic regression, we obtain the shrinkage law (Table 3.9) when $N\to\infty$. The simulation results suggest that the intercept term β_0 shrinks to n/N and β shrinks to 0, when $N\to\infty$. We now prove this shrinkage law and further prove that the estimated parameter vector β only depends on the distribution F_0, the minority mean vector $\bar{\mathbf{x}}$ and the imbalance level N/n. The lasso penalty will be demonstrated in our proof process, because it is the more complex case than ridge penalty. The proof for ridge will be given in Appendix D.

We again use the notation of Section 3.1.2. In order to directly show the result, we center lasso penalized logistic regression around the minority class mean vector $\overline{\mathbf{x}}=\sum_{i=1}^{n}\mathbf{x}_i/n$. Since in the infinitely imbalanced case $N\to\infty$, we also suppose that there is a good approximation for the conditional distribution of $\mathbf{x}$ given $Y=0$; denoted by F_0. Thus, the objective function

for lasso penalized logistic regression [Tibshirani, 1996] is written as

$$
\begin{aligned}
l(\beta_0, \beta) =& \frac{1}{n+N}\Bigg[n\beta_0 - \sum_{i=1}^{n}\log(1+e^{\beta_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\beta}) \\
&- N\int \log(1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta})dF_0(\mathbf{x})\Bigg] - \lambda\sum_{j=1}^{p}|\beta_j|,\ \lambda>0,
\end{aligned}
\tag{3.19}
$$

where β_j is the jth element of β. We follow Owen's proof again: Lemma 4 and Lemma 5 in Owen [2007] still hold for penalized logistic regression. The three changes in the proof process are for Lemma 6, Lemma 7 and the main theorem in Owen [2007] (corresponding to our Lemma 9, Lemma 10 and Theorem 12 here).

Our Lemma 9 gives $e^{\hat{\beta}_0} \le n/(N-n)$ and Lemma 10 gives $\sum_{j=1}^{p}|\hat{\beta}_j| \le n/((N-n)\lambda)$. Note that in Lemma 9 and Lemma 10, we do **not** require the surrounded condition, which makes the proof **significantly different** from Owen's proof.

Lemma 9. *Let $\hat{\beta}_0$ and $\hat{\beta}$ be the maximizers of the objective function (3.19). Then $e^{\hat{\beta}_0} \le n/(N-n)$.*

Proof. Calculate the partial derivative with respect to β_0:

$$
\begin{aligned}
\frac{\partial l(\beta_0,\beta)}{\partial\beta_0} &= \frac{n}{n+N} - \frac{1}{n+N}\sum_{i=1}^{n}\frac{e^{\beta_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\beta}} \\
&\quad - \frac{N}{n+N}\int\frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}dF_0(\mathbf{x}) \\
&\le \frac{n}{n+N} - \frac{N}{n+N}\int\frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}dF_0(\mathbf{x}) \\
&\le \frac{n}{n+N} - \frac{N}{n+N}\int_{(\mathbf{x}-\bar{\mathbf{x}})^T\beta>0}\frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}dF_0(\mathbf{x}) \\
&\le \frac{n}{n+N} - \frac{N}{n+N}\frac{e^{\beta_0}}{1+e^{\beta_0}}\int_{(\mathbf{x}-\bar{\mathbf{x}})^T\beta>0}dF_0(\mathbf{x}) \\
&\le \frac{n}{n+N} - \frac{N}{n+N}\frac{e^{\beta_0}}{1+e^{\beta_0}}.
\end{aligned}
\tag{3.20}
$$

We applied the fact that $\frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}} \leq \frac{e^{\beta_0}}{1+e^{\beta_0}}$, when$(\mathbf{x}-\bar{\mathbf{x}})^T\beta > 0$, in the above inequality. Then, let $e^{\beta_0} > n/(N-n)$, the above equation leads to $\frac{\partial l(\beta_0,\beta)}{\partial \beta_0} < 0$. For the concave likelihood function, the negative derivative means that the maximizer $\hat{\beta}_0 \leq \log(\frac{n}{N-n})$. □

Lemma 10. *Let $\hat{\beta}_0$ and $\hat{\beta}$ be the maximizers of the log-likelihood function (3.19). Then* $\limsup_{N\to\infty} \left\|\hat{\beta}\right\| < \infty$.

Proof. Take arbitrary coefficient estimates $(\hat{\beta}_0, 0)$, we know $l(\hat{\beta}_0, \hat{\beta}) - l(\hat{\beta}_0, 0) \geq 0$. Then

$$l(\hat{\beta}_0, \hat{\beta}) - l(\hat{\beta}_0, 0) = \left[\frac{1}{n+N}\left[(n+N)\log(1+e^{\hat{\beta}_0}) - \sum_{i=1}^{n}\log(1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}) - N\int \log(1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}})dF_0(\mathbf{x})\right]\right] - \lambda\sum_{j=1}^{p}|\hat{\beta}_j| \geq 0. \tag{3.21}$$

Since $\log(1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}) \geq 0$ and $\int \log(1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}})dF_0(\mathbf{x}) \geq 0$, Equation (3.21) leads to:

$$\begin{aligned}\lambda\sum_{j=1}^{p}|\hat{\beta}_j| &\leq \frac{1}{n+N}\left[(n+N)\log(1+e^{\hat{\beta}_0}) - \sum_{i=1}^{n}\log(1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}) - N\int \log(1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}})dF_0(\mathbf{x})\right] \\ &\leq \log(1+e^{\hat{\beta}_0}) \leq e^{\hat{\beta}_0} \leq \frac{n}{N-n}.\end{aligned} \tag{3.22}$$

Thus, we have $\sum_{j=1}^{p}|\hat{\beta}_j| \leq \frac{n}{(N-n)\lambda}$, by considering the triangle inequality, we know $\left\|\hat{\beta}\right\|$ is bounded as $N \to \infty$. □

The following theorem demonstrates the behavior of $\hat{\beta}_0$ and $\hat{\beta}$ in infinitely imbalanced penalized logistic regression.

Theorem 11. *Let $n > 1$ and minority class vectors $\{\mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_n\}$ be fixed. Then the maximizer $(\hat{\beta}_0, \hat{\beta})$ of l given by Equation (3.19) have following shrinkage rules, $e^{\hat{\beta}_0} \to \frac{n}{N}$ and $\hat{\beta} \to 0$, when $N \to \infty$.*

Proof. From Lemma 9 and Lemma 10, we know $e^{\hat{\beta}_0} \leq \frac{n}{N-n}$ and $\left\|\hat{\beta}\right\|$ is bounded when $N \to \infty$.

Further considering Lemma 10, we have

$$\left\|\hat{\beta}\right\| \leq \sum_{j=1}^{p} |\hat{\beta}_j| \leq \frac{n}{(N-n)\lambda} \text{ when } N \to \infty. \tag{3.23}$$

Inequality (3.23) shows as $N \to \infty$, we have $\hat{\beta} = 0$. Thus $l(\hat{\beta}_0, \hat{\beta})$ simplifies to

$$l(\hat{\beta}_0, \hat{\beta}) = \frac{n}{n+N}\hat{\beta}_0 - \log(1 + e^{\hat{\beta}_0}). \tag{3.24}$$

Let the partial derivative of Equation (3.24) equal to 0 when $N \to \infty$,

$$\frac{\partial l(\hat{\beta}_0, \hat{\beta})}{\partial \hat{\beta}_0} = \frac{n}{n+N} - \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} = 0, \tag{3.25}$$

then we have $e^{\hat{\beta}_0} = n/N$. □

Since $Ne^{\hat{\beta}_0} \to n$ and $\hat{\beta} \to 0$ in penalized logistic regression when $N \to \infty$ (regardless of the data set), the shrinkage properties of penalized methods means that the data ceases to be involved. In this case, the estimated probability for the minority and the majority class will simply approach their marginal frequencies:

$$\Pr(\text{Y=1}|\mathbf{x} = \mathbf{x}) \xrightarrow{N\to\infty} \tfrac{n}{n+N},$$
$$\Pr(\text{Y=0}|\mathbf{x} = \mathbf{x}) \xrightarrow{N\to\infty} \tfrac{N}{n+N}.$$

Note that Theorem 11 is a strong result that demonstrates $\hat{\beta} = 0$ in infinitely imbalanced penalized logistic regression. We are also interested in how $\hat{\beta}$ shrinks to the limit when N approaches infinity. Therefore, we prove another theorem to demonstrate the behavior of $\hat{\beta}$ when approaching infinitely imbalanced penalized logistic regression. Here we will utilize $e^{\hat{\beta}_0} \to n/N$ as $N \to \infty$ from Theorem 11.

Since the lasso penalized log-likelihood function is nondifferentiable at $\beta = 0$, we alternatively use the subgradient [Nesterov, 2013, p. 126] of convex

Equation (3.19) with respect to β. Setting the subgradient of Equation (3.19) to 0, we have

$$-\frac{1}{n+N}\left[\sum_{i=1}^{n}\frac{(\mathbf{x}_i-\bar{\mathbf{x}})e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}}{1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}}+N\int\frac{(\mathbf{x}-\bar{\mathbf{x}})e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}{1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}dF_0(\mathbf{x})\right]-\lambda s=0,$$

where s is a p dimensional vector

$$s_j=\begin{cases}\text{Sign}(\hat{\beta}_j), & \text{if } \hat{\beta}_j\neq 0\\ \text{any number in } [-1,1], & \text{if } \hat{\beta}_j=0\end{cases}\quad\text{where } j\in\{1,\cdots,p\}.$$

We take advantage of a specific characteristic of the subgradient method; a convex function f attains its optimal value at a vector v if the zero vector is a subgradient of f at v [Nesterov, 2013, Theorem 3.1.15]. Thus, because Equation (3.19) is a convex function, we have the following main theorem:

Theorem 12. *Let $n\geq 1$ and minority class vectors $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ be fixed and suppose that F_0 surrounds $\bar{\mathbf{x}}=\sum_{i=1}^{n}\mathbf{x}_i/n$ as described. Then the maximizer $(\hat{\beta}_0,\hat{\beta})$ of l given by Equation (3.19) satisfies*

$$-\int(\mathbf{x}-\bar{\mathbf{x}})e^{(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}dF_0(\mathbf{x})=\frac{n+N}{n}\lambda s \tag{3.26}$$

as $N\to\infty$.

Proof. Setting the subgradient of Equation (3.19) to 0, we have

$$-\sum_{i=1}^{n}\frac{(\mathbf{x}_i-\bar{\mathbf{x}})e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}}{1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}}-N\int\frac{(\mathbf{x}-\bar{\mathbf{x}})e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}{1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}dF_0(\mathbf{x})-(n+N)\lambda s=0.$$

Dividing by N gives

$$-\int\frac{(\mathbf{x}-\bar{\mathbf{x}})e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}{1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}dF_0(\mathbf{x})-\frac{n+N}{N}\lambda s=\frac{1}{N}\sum_{i=1}^{n}\frac{(\mathbf{x}_i-\bar{\mathbf{x}})e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}}{1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}}. \tag{3.27}$$

As $N\to\infty$, the right side of Equation (3.27) vanishes because $\left\|\hat{\beta}\right\|$ is bounded as $N\to\infty$ by Lemma 10.

If we consider $N \to \infty$, we have $e^{\hat{\beta}_0} \to \frac{n}{N}$ and $\hat{\beta} \to 0$, yielding to $e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}} \to \frac{n}{N} e^0 \to 0$. Thus Equation (3.27) yields

$$-\int \frac{(\mathbf{x} - \bar{\mathbf{x}}) \frac{n}{N} e^{(\mathbf{x}-\bar{\mathbf{x}})^T \hat{\beta}}}{1} dF_0(\mathbf{x}) = \frac{n+N}{N} \lambda s \tag{3.28}$$

After simplification, Equation (3.26) holds. Notice that, s must converge to 0 due to the established Equation (3.26) when $N \to \infty$. □

Equation (3.26) shows the solution of β depends only on $\{\bar{\mathbf{x}}, F_0(\mathbf{x}), \frac{N}{n}\}$ when approaching infinite imbalance. We give some context and illustration of Theorem 12 with respect to the solution of penalized logistic regression with large finite N in the next section.

3.3.2 Numerical Explanations for Highly Imbalanced Lasso Penalized Logistic Regression

For application purposes, we are interested in investigating the solution of penalized logistic regression with large finite N. The following calculation and simulation provides some context and illustration of Theorem 12.

Assume the probability density function of the majority class, $f_0(\mathbf{x})$, is $N(0, I)$ where 0 is a p dimensional zero vector and I is the $p \times p$ identity covariance matrix. By taking advantage of independence between variables, we vectorise Equation (3.26) in **dimension** r (here $\mathbf{x}_{\cdot r}$ refers to the rth column

in the observation matrix):

$$\begin{aligned}
&\int (\mathbf{x}_{\cdot r} - \overline{\mathbf{x}}_{\cdot r}) e^{\sum_{k=1}^{p} (\mathbf{x}_{\cdot k} - \overline{\mathbf{x}}_{\cdot k})\beta_{\cdot k}} dF_0(\mathbf{x}) \\
&= \int (\mathbf{x}_{\cdot r} - \overline{\mathbf{x}}_{\cdot r}) e^{\sum_{k=1}^{p} (\mathbf{x}_{\cdot k} - \overline{\mathbf{x}}_{\cdot k})\beta_{\cdot k}} f_0(\mathbf{x}_{\cdot 1}) \cdots f_0(\mathbf{x}_{\cdot p}) d\mathbf{x}_{\cdot 1} \cdots d\mathbf{x}_{\cdot p} \\
&= \int (\mathbf{x}_{\cdot r} - \overline{\mathbf{x}}_{\cdot r}) e^{(\mathbf{x}_{\cdot r} - \overline{\mathbf{x}}_{\cdot r})\beta_{\cdot r}} f_0(\mathbf{x}_{\cdot r}) d\mathbf{x}_{\cdot r} \times \prod_{k \neq r} \int e^{(\mathbf{x}_{\cdot k} - \overline{\mathbf{x}}_{\cdot k})\beta_{\cdot k}} f_0(\mathbf{x}_{\cdot k}) d\mathbf{x}_{\cdot k} \\
&= \int \frac{(\mathbf{x}_{\cdot r} - \overline{\mathbf{x}}_{\cdot r}) e^{-\frac{\mathbf{x}_{\cdot r}^2}{\mathbf{x}}} e^{\mathbf{x}_{\cdot r}\beta_{\cdot r}}}{\sqrt{2\pi} e^{\overline{\mathbf{x}}_{\cdot r}\beta_{\cdot r}}} d\mathbf{x}_{\cdot r} \times \prod_{k \neq r} \int \frac{e^{-\frac{\mathbf{x}_{\cdot k}^2}{2}} e^{\mathbf{x}_{\cdot k}\beta_{\cdot k}}}{\sqrt{2\pi} e^{\overline{\mathbf{x}}_{\cdot k}\beta_{\cdot k}}} d\mathbf{x}_{\cdot k} \\
&= (\beta_{\cdot r} - \overline{\mathbf{x}}_{\cdot r}) \prod_{k=1}^{p} \frac{e^{\frac{\beta_{\cdot k}^2}{2}}}{e^{\overline{\mathbf{x}}_{\cdot k}\beta_{\cdot k}}}.
\end{aligned} \tag{3.29}$$

Consider that all $\beta_{\cdot r}, r \in \{1, \cdots, p\}$ are shrunk to 0, thus all $s_{\cdot r} \in [-1, 1]$. For all $r \in \{1, \cdots, p\}$, we have

$$(0 - \overline{\mathbf{x}}_{\cdot r}) = -\frac{n+N}{n} \lambda s_{\cdot r} \left(\prod_{k=1}^{p} \frac{e^{\frac{\beta_k^2}{2}}}{e^{\overline{\mathbf{x}}_k \beta_k}} \right)^{-1} = -\frac{n+N}{n} \lambda s_r, \tag{3.30}$$

so $\overline{\mathbf{x}}_{\cdot r}$ must be located in the interval $[-\lambda(1+\frac{N}{n}), \lambda(1+\frac{N}{n})]$. This interval will enlarge as the imbalance level $\frac{N}{n}$ increases, such that it will easily include all $\overline{\mathbf{x}}_{\cdot r}$.

Alternatively, if $\lambda > \frac{n}{n+N} \max_r \{\overline{\mathbf{x}}_{\cdot r}\}$, all $\beta_{\cdot r}$ will be shrunk to 0. When the imbalance level N/n is great, this condition is easily satisfied with respect to a fixed λ.

Following is a simple simulation to demonstrate the consequence of Equation (3.30). We generate $100,000$ five dimensional majority class data ($Y = 0$) following $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where

$$\boldsymbol{\mu}_0 = [0, 0, 0, 0, 0], \boldsymbol{\Sigma}_0 = I$$

and I is the identity matrix. $1,000$ minority class data ($Y = 1$) are generated

Table 3.10: Coefficient estimates of lasso penalized logistic regression with different penalty parameter λ.

λ	$\beta_{\cdot 1}$	$\beta_{\cdot 2}$	$\beta_{\cdot 3}$	$\beta_{\cdot 4}$	$\beta_{\cdot 5}$
0.0190	0	0	0	0	0
0.0168	0.1650	0	0	0	0
0.0153	0.3106	0.1148	0	0	0
0.0139	0.4388	0.2416	0.0377	0	0
0.0116	0.6435	0.4445	0.2392	0.0471	0
0.0087	0.8621	0.6581	0.4525	0.2547	0.0516

which follows $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where

$$\boldsymbol{\mu}_1 = [1.9, 1.7, 1.5, 1.3, 1.1], \boldsymbol{\Sigma}_1 = 0.01I.$$

The numbers in vector $\boldsymbol{\mu}_1$ are artificially set in a decreasing order, this is for later showing that how λ influences the coefficient estimates in different dimension. In this scenario, $\frac{n}{n+N} \max_r\{\overline{\mathbf{x}}_{\cdot r}\}$ equals 0.0188, and in our simulation, the simulated vector $\frac{n}{n+N}\bar{\mathbf{x}}_{\cdot r}$ is $[0.0188, 0.0168, 0.0149, 0.0129, 0.0109]$.

Table 3.10 shows the coefficient estimates of lasso penalized logistic regression for different values of λ. These λ are set in gaps between the numbers in vector $\frac{n}{n+N}\bar{\mathbf{x}}_{\cdot r}$. We find that the relationship between $\bar{\mathbf{x}}_{\cdot r}$ and λ determines whether the coefficient estimate $\beta_{\cdot r}$ will be shrunk to 0 (all coefficient estimates shrink to 0 when $\lambda > \frac{n}{n+N} \max_r\{\overline{\mathbf{x}}_{\cdot r}\}$, see Table 3.10 Line 1).

Without the assumption of the majority class distribution $F_0(\mathbf{x})$, if $\hat{\beta} = 0$, Equation (3.26) simplifies to

$$\overline{\mathbf{x}} - \int \mathbf{x} dF_0(\mathbf{x}) = \frac{n+N}{n}\lambda s, \text{ where } s \in [-1, 1]. \tag{3.31}$$

$\int \mathbf{x} dF_0(\mathbf{x})$ is the population mean $\mu_{\mathbf{x}}$ of the majority class distribution, which is determined by data set. Thus we need $\overline{\mathbf{x}} \in [\mu_{\mathbf{x}} - \frac{n+N}{n}\lambda, \mu_{\mathbf{x}} + \frac{n+N}{n}\lambda]$ to force all $\hat{\beta}$ shrink to 0, which is easy to satisfy with highly imbalanced data set.

3.4 Infinitely Imbalanced Multinomial Logistic Regression

We have briefly described infinitely imbalanced logistic regression. In this section, we give a similar result for multinomial logistic regression with a specific highly imbalanced multi-minority setting. This section is a preliminary for our proposed relabeling approach in Chapter 4; we are going to use this result for the calculation of our EM algorithm.

Multinomial logistic regression can be set up as a *one vs rest* model [Anderson, 1972], which we will consider a *base vs interest classes* model. Here, we use $\mathbf{x}_{0i}, i \in \{1, \cdots, N\}$ to denote the base class data. There are K interest classes, and each class has $n_k, k \in \{1, \cdots, K\}$ observations, denoted by $\mathbf{x}_{ki}, i \in \{1, \cdots, n_k\}$. We are particularly interested in the $n \ll N$ case as an extention of Owen [2007], where $n = \sum_{k=1}^{K} n_k$ (the number of the observations in rest classes). The main result we seek is Theorem 15: each minority class k only contribute to multinomial logistic regression via its mean vector $\bar{\mathbf{x}}_k = \sum_{i=1}^{n_k} \mathbf{x}_{ki}$, where $k \in \{1, \cdots, K\}$.

We still assume there is a good approximation of the base (majority) class distribution F, which does not have a heavy tail (Equation 3.4) and surrounds $\bar{x}_k$ where $k \in \{1, \ldots, K\}$. It is important to notice that Begg and Gray [1984] show that using a series of separate logistic regression asymptotically gives the same coefficient estimates as solving multinomial logistic regression. Howerer, directly combining [Owen, 2007] and [Begg and Gray, 1984] does not lead to the following Theorem 15, because the [Begg and Gray, 1984] result asymptotically requires $(N + \sum_{k=1}^{K} n_k) \to \infty$ but at the same time $n_k/N, k \in \{1, \ldots, K\}$ are nonnegligible, i.e. $\min(n_1, \ldots, n_k, N) \to \infty$, but Theorem 15 only requires $N \to \infty$.

Without loss of generality, we only investigate the limit behavior in class $k = 1$ when $N \to \infty$. We center multinomial logistic regression around $\bar{\mathbf{x}}_1$,

then the log-likelihood function is:

$$
\begin{aligned}
l(\beta_{01},\ldots,\beta_{0K},\beta_1,\ldots,\beta_K) = \; & n_1\beta_{01} - \sum_{i=1}^{n_1}\{\log(1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}_{1i}-\bar{\mathbf{x}}_1)^T\beta_k})\} \\
& + n_2\beta_{02} + \sum_{i=1}^{n_2}(\mathbf{x}_{2i}-\bar{\mathbf{x}}_1)^T\beta_2 - \sum_{i=1}^{n_2}\{\log(1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}_{2i}-\bar{\mathbf{x}}_1)^T\beta_k})\} \\
& + \cdots \\
& + n_K\beta_{0K} + \sum_{i=1}^{n_K}(\mathbf{x}_{Ki}-\bar{\mathbf{x}}_1)^T\beta_K - \sum_{i=1}^{n_K}\{\log(1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}_{Ki}-\bar{\mathbf{x}}_1)^T\beta_k})\} \\
& - N\int \log(1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k})dF(\mathbf{x}),
\end{aligned}
\tag{3.32}
$$

where β_{0k} is the intercept term, and β_k is the slope vector of each class $k \in \{1,\cdots,K\}$.

We sketch the proof of Theorem 15 first. We follow Owen [2007] proof: Lemma 4 and Lemma 5 in Owen [2007] still hold for multinomial logistic regression. Albert and Anderson [1984] which shows the surrounded condition is still essential for the existence and uniqueness of MLE in multinomial logistic regression. We illustrate three changes in Lemma 6, Lemma 7 and the main theorem of Owen [2007] (corresponding to Lemma 13, Lemma 14, and Theorem 15).

Lemma 13. *Let $\hat{\beta}_{01}$ and $\hat{\beta}_1$ be the maximizer of the likelihood function (3.32), F satisfies the surrounded condition at $\bar{\mathbf{x}}_k, k \in \{1,\cdots,K\}$ and η_1 is the infimum of δ_1. Then for any $N \geq \frac{2n_1}{\eta_1}$, we have $e^{\hat{\beta}_{01}} \leq \frac{2n_1}{N\eta_1}$.*

Proof. Let $e^{\beta_{0k}} = \frac{A_k}{N}$, where $0 < A_k < \infty$.

$$\begin{aligned}
\frac{\partial l}{\partial \beta_{01}} =& n_1 - \sum_{i=1}^{n} \frac{e^{\beta_{01}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_1}}{1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}_{ki}-\bar{\mathbf{x}}_1)^T\beta_k}} - N\int \frac{e^{\beta_{01}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}}{1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}} dF(\mathbf{x}) \\
\leq& n_1 - N\int \frac{e^{\beta_{01}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}}{1+\sum_{k=1}^{K} e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}} dF(\mathbf{x}) \\
=& n_1 - A_1\int \frac{e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}}{1+\sum_{k=1}^{K} A_k e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}/N} dF(\mathbf{x}) \\
=& n_1 - A_1\int \frac{e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}}{1+A_1 e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}/N+\sum_{k=2}^{K} A_k e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}/N} dF(\mathbf{x}) \\
\leq& n_1 - A_1\int_{\mathbf{x}\in\Omega} \frac{e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}}{1+A_1 N^{-1} e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1}+N^{-1}\sum_{k=2}^{K} A_k e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}} dF(\mathbf{x}) \\
\leq& n_1 - A_1\frac{1}{1+A_1N^{-1}+N^{-1}\sum_{k=2}^{K} A_k}\int_{\mathbf{x}\in\Omega} dF(\mathbf{x}) \\
\leq& n_1 - \frac{A_1\eta_1}{1+N^{-1}\sum_{k=1}^{K} A_k},
\end{aligned} \tag{3.33}$$

where, $\Omega = \{\mathbf{x}|(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1 \geq 0,\ (\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k \leq 0 \text{ for any } k\}$.

Because $\eta_1 \leq \delta_1$ and $\delta_1 < \int_{(\mathbf{x}-\mathbf{x}_\star)^T\psi\geq\epsilon} dF(\mathbf{x}) \leq 1$ (see Definition 1), thus η_1 is bounded in $0 < \eta_1 \leq \delta_1 < 1$. Since $N \geq 2n_1/\eta_1$, if we let $e^{\beta_{01}} > 2n_1/N\eta_1$, we will have $A_1 > 2n_1/\eta_1$. Then, we have

$$\frac{\partial l}{\partial \beta_{01}} < n_1 - \frac{\eta_1}{\frac{\eta_1}{n_1}+\frac{\sum_{k=2}^{K} e^{\beta_{0k}-\beta_{01}}}{N}}. \tag{3.34}$$

When N large enough, from above equation, we will have $\partial l/\partial\beta_{01} \leq 0$. For the concave likelihood function, the negative partial derivative means that the maximizer $e^{\hat{\alpha}_1} \leq 2n_1/(N\eta_1)$. □

Lemma 14. *Under the same condition as Lemma 13, we will have* $\limsup_{N\to\infty} \left\|\hat{\beta}_1\right\| < \infty$.

Proof. Under the surrounded condition, a γ exists such that

$$\inf_{\psi\in\Psi}\int[(\mathbf{x}-\bar{\mathbf{x}}_1)^T\psi]_+dF(\mathbf{x})\geq\gamma>0, \tag{3.35}$$

where $\psi^T\psi=1$ and $[(\mathbf{x}-\bar{\mathbf{x}}_1)^T\psi]_+$ means the positive part of $[(\mathbf{x}-\bar{\mathbf{x}}_1)^T\psi]$. We still let $e^{\beta_{01}}=A_1/N$, then we have

$$\begin{aligned}
l(\beta_{01},0)-l(\beta_{01},\beta_1)=&-\sum_{i=1}^{n}\log(1+e^{\beta_{01}}+\sum_{k=2}^{K}e^{\beta_{0k}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_k})\\
&-N\int\log(1+e^{\beta_{01}}+\sum_{k=2}^{K}e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k})dF(\mathbf{x})\\
&+\sum_{i=1}^{n}\log(1+\sum_{k=1}^{K}e^{\beta_{0k}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_k})\\
&+N\int\log(1+\sum_{k=1}^{K}e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k})dF(\mathbf{x}).
\end{aligned} \tag{3.36}$$

Because

$$\begin{aligned}
&\sum_{i=1}^{n}\log(1+\sum_{k=1}^{K}e^{\beta_{0k}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_k})+N\int\log(1+\sum_{k=1}^{K}e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k})dF(\mathbf{x})\\
\geq&\sum_{i=1}^{n}\log(1+e^{\beta_{01}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_1})+N\int\log(1+e^{\beta_{01}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_1})dF(\mathbf{x})\\
\geq&\sum_{i=1}^{n}\log(1+e^{\beta_{01}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_1}),
\end{aligned} \tag{3.37}$$

and when $\mathbf{x} > 0$, $\mathbf{x} > \log(\mathbf{x})$, we have following inequation when $N > 2n_1/\eta_1$,

$$\begin{aligned}
&-\sum_{i=1}^{n}\log(1+e^{\beta_{01}}+\sum_{k=2}^{K}e^{\beta_{0k}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_k})-N\int\log(1+e^{\beta_{01}}+\sum_{k=2}^{K}e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k})dF(\mathbf{x})\\
>&-(n+N)e^{\beta_{01}}-\sum_{i=1}^{n}\sum_{k=2}^{K}e^{\beta_{0k}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_k}-N\int\sum_{k=2}^{K}e^{\beta_{0k}+(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}dF(\mathbf{x})\\
>&-(n+N)e^{\beta_{01}}-\sum_{i=1}^{n}\sum_{k=2}^{K}\frac{2n_1}{N\eta_1}e^{(\mathbf{x}_i-\bar{\mathbf{x}}_1)\beta_k}-\int\sum_{k=2}^{K}\frac{2n_k}{\eta_k}e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}dF(\mathbf{x}).
\end{aligned}\tag{3.38}$$

From the tail condition we know

$$\int\sum_{k=2}^{K}\frac{2n_k}{\eta_k}e^{(\mathbf{x}-\bar{\mathbf{x}}_1)^T\beta_k}dF(\mathbf{x})$$

is a finite number (say q_1) and

$$\sum_{i=1}^{n}\sum_{k=2}^{K}\frac{2n_1}{N\eta_1}e^{(\mathbf{x}_i-\bar{\mathbf{x}}_1)\beta_k}$$

is smaller than a fixed number q_2 when N is big enough. Let $q = q_1 + q_2$, thus we have

$$\begin{aligned}
l(\beta_{01},0)-l(\beta_{01},\beta_1) &> -\sum_{i=1}^{n}\log(1+e^{\beta_{01}}+\sum_{i=1}^{n}\log(1+e^{\beta_{01}+(\mathbf{x}_i-\bar{\mathbf{x}}_1)^T\beta_1})-q\\
&\geq -(n+N)\frac{A}{N}+\frac{A+N}{N+A}\|\beta_1\|\gamma-q.
\end{aligned}\tag{3.39}$$

Equation (3.39) shows that when $\|\beta_1\| > \frac{1}{q\gamma}(1+A/N)(1+n/N)$, we have $l(\beta_{01},0) > l(\beta_{01},\beta_1)$. Thus, maximizing likelihood function will obviously let $\left\|\hat{\beta}_1\right\| < 2/q\gamma$, when N large enough. □

Theorem 15. *Let $n \geq 1$, and $\mathbf{x}_{ki}, k \in \{1,\cdots,K\}, i \in \{1,\cdots,n_k\}$ be fixed and suppose F satisfies the tail condition (Equation 3.4) and surrounded at $\bar{\mathbf{x}}_k$ as describe in Definition 1. Then the maximizer $(\hat{\beta}_{0k},\hat{\beta}_k)$ of Equation (3.32)*

satisfies

$$\bar{\mathbf{x}}_k = \lim_{N\to\infty} \frac{\int \mathbf{x} e^{\mathbf{x}^T\hat{\beta}_k} dF(\mathbf{x})}{\int e^{\mathbf{x}^T\hat{\beta}_k} dF(\mathbf{x})}, \text{ where } k \in \{1, \cdots, K\}. \tag{3.40}$$

Proof. Setting $\frac{\partial l}{\partial \beta_1} = 0$, we have

$$\begin{aligned} 0 = \frac{\partial l}{\partial \beta_1} = &- \sum_{i=1}^{n} \frac{(\mathbf{x}_i - \bar{\mathbf{x}}_1) e^{\beta_{01} + (\mathbf{x}_i - \bar{\mathbf{x}}_1)^T \beta_1}}{1 + \sum_{k=1}^{K} e^{\beta_{0k} + (\mathbf{x}_i - \bar{\mathbf{x}}_1)^T \beta_k}} \\ &- N \int \frac{(\mathbf{x} - \bar{\mathbf{x}}_1) e^{\beta_{01} + (\mathbf{x} - \bar{\mathbf{x}}_1)^T \beta_1}}{1 + \sum_{k=1}^{K} e^{\beta_{0k} + (\mathbf{x} - \bar{\mathbf{x}}_1)^T \beta_k}} dF(\mathbf{x}). \end{aligned} \tag{3.41}$$

Dividing by $Ne^{\beta_{01} + \bar{\mathbf{x}}_1^T \beta_1}$, we have

$$\int \frac{(\mathbf{x} - \bar{\mathbf{x}}_1) e^{\mathbf{x}^T \beta_1}}{1 + \sum_{k=1}^{K} e^{\beta_{0k} + (\mathbf{x} - \bar{\mathbf{x}}_1)^T \beta_k}} dF(\mathbf{x}) = -\frac{1}{N} \sum_{i=1}^{n} \frac{(\mathbf{x}_i - \bar{\mathbf{x}}_1) e^{\mathbf{x}_i^T \beta_1}}{1 + \sum_{k=1}^{K} e^{\beta_{0k} + (\mathbf{x}_i - \bar{\mathbf{x}}_1)^T \beta_k}}. \tag{3.42}$$

Because $\hat{\beta}_1$ is bounded, the right side will vanish as $N \to \infty$. Thus we have

$$\bar{\mathbf{x}}_1 \int \frac{e^{\mathbf{x}^T \beta_1}}{1 + \sum_{k=1}^{K} e^{\beta_{0k} + (\mathbf{x} - \bar{\mathbf{x}}_k)^T \beta_k}} dF(\mathbf{x}) = \int \frac{\mathbf{x} e^{\mathbf{x}^T \beta_1}}{1 + \sum_{k=1}^{K} e^{\beta_{0k} + (\mathbf{x} - \bar{\mathbf{x}}_k)^T \beta_k}} dF(\mathbf{x}) \tag{3.43}$$

Because $\hat{\beta}_{0k} \to -\infty$ and the tail condition, the results hold. □

This proof can be generalized to any $k \in \{1, \cdots, K\}$, which shows that each minority class only contribute to the multinomial logistic regression via their mean vector.

3.5 Summary

We have explored the impact of class imbalance for the logistic regression. Owen's results show that, in some limit, unwanted estimation artifacts arise when classes are highly imbalanced. We extend these results to show that some candidate approaches for mitigating these effects, namely weighted and penalized likelihood approaches, suffer from the same problem. This limiting

behavior is a characteristic of logistic regression, regardless of any particular imbalanced data. This tells us that the infinite imbalance problem is fundamental for logistic regression. We also extend Owen [2007] result to a multiclass scenario as a theoretical preparation for the next chapter.

Since logistic regression remains the workhorse of many applications [King and Zeng, 2001b, Zhu et al., 2006], the issue of imbalance certainly demands more attention, and the development of enhanced tools. In the following chapters, we sketch our mitigation methods and diagnostic tools to the highly imbalance logistic regression.

4

Relabeling Approach

Logistic regression is a well established classification algorithm, which remains a reference benchmark in many domains like consumer credit risk [Thomas, 2009], due to the regulatory requirement of interpretability [Basel II Accords, 2004]. Owen [2007] provides an asymptotic result which suggests that the minority class data contributes to logistic regression estimation only via its mean vector. In the previous chapter, we show that two alternative methods to logistic regression, namely weighted and penalized logistic regression, still suffer the same problem. These results suggest an obvious concern about unwanted consequences when cluster structure is present in the minority class. In this chapter, we propose a new approach to handle highly imbalanced classification problems when using logistic regression. Essentially, this approach seeks to *relabel* the minority class into several new pseudo-classes to circumvent the imbalance problem by exploiting cluster structure, then modeling on the new pseudo-classes for a multiclass model, hence improving predictive performance.

A binary classification problem involves training data labeled in the set $\{0, 1\}$. Assuming class 1 is the minority class, the relabeling concept seeks to construct a new pseudo-label in $\{1, 2, \ldots, K\}$ for the class 1 objects, which cap-

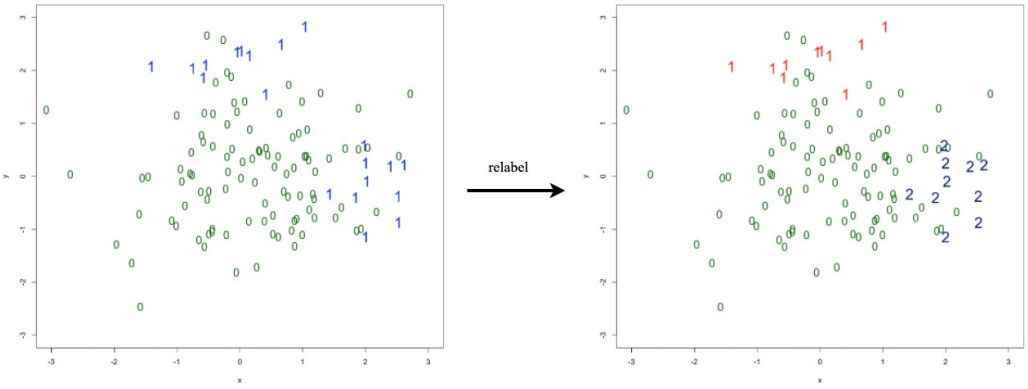


Figure 4.1: Illustration to the relabeling idea.

tures the cluster structure (i.e. generate some well separated pseudo-classes if cluster structure exists among the minority class). A classifier capable of handling multiple classes is then trained. Subsequently, in deployment on unlabeled data, the task is treated as a binary problem, by classifying to class 0 or otherwise (e.g. $\{1, 2, \ldots, K\}$). Figure 4.1 is an illustrative example of our relabeling idea: the left side is a binary classification problem, the relabeling idea seeks to grasp the unobserved cluster structure among the minority class by relabeling them into two new pseudo-classes (the right side in Figure 4.1). How to generate new pseudo-classes label will be introduced in this chapter. This relabeling concept is generic and pragmatic, but the known properties of logistic regression with imbalanced data will provide a clear demonstration; in this chapter, we will use logistic regression to derive an efficient algorithm to implement this relabeling idea.

There are two issues to resolve with the relabeling concept: selecting the number of pseudo-classes, K, and identifying the mapping of minority class instances to the K pseudo-classes. The latter problem is combinatorial, and hence computationally challenging even for small data sets. While brute-force attack method (i.e. checking all the possible relabeling solution, like using the genetic algorithm [Haupt and Ellen Haupt, 2004]) is possible, we have found it to be computationally demanding.

Expectation Maximization (EM algorithm) [Hastie et al., 2009, page 272] algorithm is an iterative optimization method for maximizing the posterior

likelihood with latent variables in the likelihood function. Generally, EM algorithm contains two steps in each iteration:

1. E-step: Given the maximum likelihood estimates or initial guess of the parameters, calculate the expectations of the latent variables.

2. M-step: Given the expectations of the latent variables, maximize the likelihood function to obtain new parameter estimates.

This chapter develops a new formulation of multiple logistic regression, which allows the identification of a given number of pseudo-classes and supports estimation via Expectation Maximization algorithm, which in turn reduces the computational burden significantly by efficiently searching the optimized relabeling solution among all possible solutions. The latter problem, selecting K, is resolved with a cross-validation approach.

To summarize the approach, we first define notation. Let $\mathbf{x}_{1i}, i \in \{1, \ldots, n\}$ denote the minority class observations, which are labeled as class 1, $Y = 1$, in the training data. Here we are seeking to selectively *relabel* the minority class into K pseudo-classes, which means assigning each minority observation $\mathbf{x}_{1i}$ a new tag $z_i = k$, where $i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$. Then, the core problem is estimating $\Pr(z_i = k|\mathbf{x}_{1i})$, the probability of a minority class observation belonging to pseudo-class k. If we think about the likelihood function of multinomial logistic regression when using $Y = 0$ as base class (for the observation $\mathbf{x}_{1i}$ part):

$$\text{(Likelihood for all } \mathbf{x}_{1i}) = \prod_{i=1}^{n} (\sum_{k=1}^{K} \mathbf{I}(z_i = k) \times \Pr(z_i = k|\mathbf{x}_{1i})),$$

where $\mathbf{I}(z_i = k) = 0$ or 1 is an indicator function. In our study, because $z_i = k$ are unobservable, we do not know the result of $\mathbf{I}(z_i = k)$. Thus, we assume z_i arises from a mixture of K proportions with prior weights $\Phi = \{\phi_1, \ldots, \phi_K\}$, i.e. $z_i \sim$ multinomial(Φ), where $\sum_{k=1}^{K} \phi_k = 1$, then, for any minority class observation $\mathbf{x}_{1i}$, we can write the case wise contributions

to the likelihood function with pseudo-classes as

$$\sum_{k=1}^{K} \phi_k \Pr(z_i = k|\mathbf{x}_{1i}). \tag{4.1}$$

By plugging Equation (4.1) into the log-likelihood function of multinomial logistic regression, we can use the EM algorithm to maximize the log-likelihood function and evaluate $\Pr(z_i = k|\mathbf{x}_{1i})$ simultaneously (details of this EM algorithm will be given in Section 4.3). On completion of this joint maximization, we relabel each minority class observation x_{1i} by assigning a label $z_i = k$, such that leads to the highest probability $\Pr(z_i = k|\mathbf{x}_{1i})$,

$$k = \arg\max_{k} \big(\Pr(z_i = k|\mathbf{x}_{1i})\big). \tag{4.2}$$

Constructing a multinomial logistic regression on this relabeled minority class data has the potential to alleviate the problem of highly imbalanced logistic regression, via utilizing each pseudo-classes' mean vector, especially when the cluster structure in the minority class do exist.

It is worth noting that our relabeling approach is motivated by finite mixture models, but it is different from finite mixture regression model, because the mixture model is designed to represent the unobserved population among the overall population, but we particularly interested in the minority class. Also, one can has a distributional assumption for the minority class data, then using EM algorithm purely on the minority class (e.g. solving a Gaussian mixture model in the minority class via using EM). This is different from our EM algorithm, because we deliberately let the majority class affects the relabeling of the minority class cases to reflect different aspects of discrimination of the new pseudo-class from the majority class. We will further show this point in Section 4.3.

We demonstrate that this relabeling method can bring two key advantages: improved prediction performance and very efficient computation compared to brute force method (i.e. checking all the possible relabeling solutions, like using genetic algorithm). Moreover, the method has the ability to find meaningful structure in the minority class. This point is demonstrated in

our credit scoring example later.

The outline of this chapter is: Section 4.1 gives the motivation to our approaches. Section 4.2 introduces a brute force method (genetic algorithm) to solve the relabel problem. This approach is time consuming but can always produces some relabeled minority classes. Our novel relabeling approach (EM algorithm) and corresponding algorithmic and deployment procedures are introduced in Section 4.3. We also provides the results of experiments with this relabeling approach on several imbalanced data sets to demonstrate its effectiveness.

4.1 Motivation

In this section, we illustrate the motivation to our relabeling approaches regarding the Owen [2007] results and the minority class's cluster structure.

As far as we know, the roughly analogous concept to the *relabeling* in the high class imbalance problem is the *small disjuncts* among the minority class. Nickerson et al. [2001] argue that balancing the class proportion is not an effective approach when the small disjuncts appear among the minority class. Jo and Japkowicz [2004] proposed an oversample method with the consideration of the small disjuncts. They suggest using unsupervised clustering methods, like K-means (see Section 2.5), to cluster both the majority class and the minority class into several small disjuncts; then random oversample all of these small disjuncts to the size of the largest disjunct.

In light of the results given in Chapter 3, if we suppose the minority class has some underlying structure, this could be problematic in the infinitely imbalanced regime, simply because the mean vector of clustered data is unlikely to be a good representation. Our relabeling approaches focus on the unwanted consequences when the cluster structure appears among the minority class. To explore this problem we first *relabel* the minority class into several new pseudo-classes using a clustering algorithm*, then use multinomial logistic regression to model the two pseudo-classes along with the majority class. In

*how to choose the number of the pseudo-classes are discussed in Section 4.3.3

the simplest case, this can be achieved by using standard clustering methods. This suggestion, to split the minority class into pseudo-classes, is familiar in credit scoring where different types of bad debt are considered[†], such as the "can't pay" and "won't pay" behavioural distinction [Bravo et al., 2015]. Here, it is still possible to reason about the good/bad dichotomy, simply by considering the predictions $\Pr(Y = 0|X = \mathbf{x})$.

We consider a contrived example here. To illustrate the issue of cluster structure in the minority class, we simulate a bivariate normal distribution example (see Figure 4.2). We generate 10,000 sample points following $X \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ from the majority class ($Y = 0$). Then 50 points following $X \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and 50 points following $X \sim N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ are generated and combined as the single minority class. Here

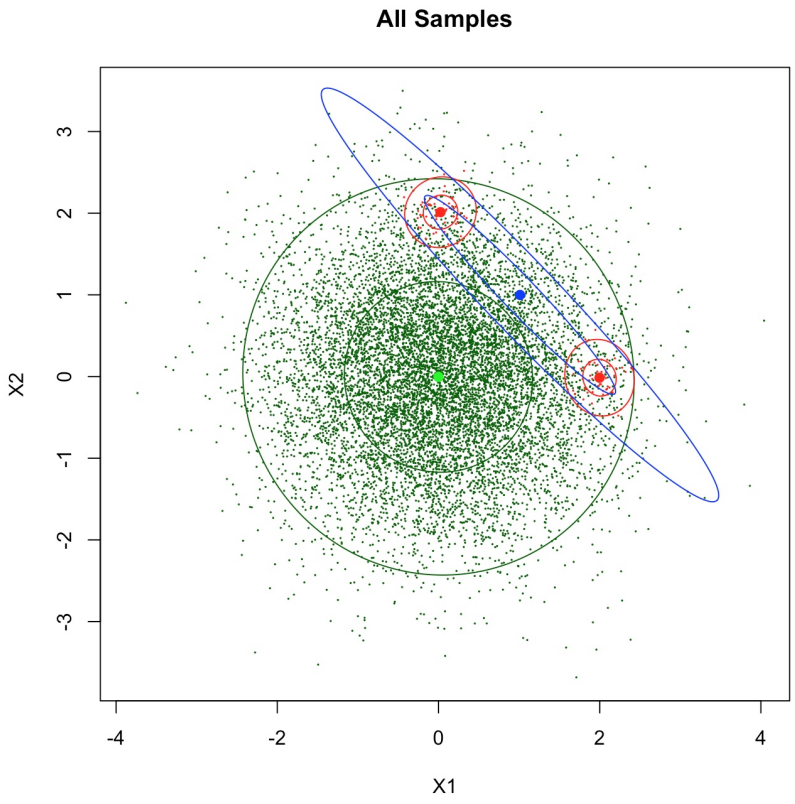
$$\boldsymbol{\mu}_1 = [0, 0], \boldsymbol{\mu}_2 = [0, 2], \boldsymbol{\mu}_3 = [2, 0], \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}.$$

In Figure 4.2, dark blue points represent the majority class, red points represent the minority class. The two red contour lines indicate that the minority class data are in two well-separated clusters.

We train two models on this data set. The first is a standard logistic regression model, the second is a multinomial logistic regression model which has one majority class ($c1, Y = 0$) and two separate minority pseudo-classes ($c2$ and $c3$, $Y = 1$) which are separated using K-means clustering (with $K = 2$). By construction, the minority class is well-separated. Finally, we generate test data following the distribution described above, to assess out-of-sample predictive performance. The prediction AUC of logistic regression is 0.918 which is lower than the AUC of multinomial logistic regression (0.954). These results suggest that logistic regression under-performs multinomial logistic regression when cluster structure is taken into account.

[†]Happy families are all alike; every unhappy family is unhappy in its own way. -*Leo Tolstoy*

Figure 4.2: Scatter plot of Simulations Samples. Green points represent the majority class and red points represent the minority class



4.2 Genetic Algorithm

Relabeling can be implemented in various ways. Unsupervised clustering methods (e.g. K-means [Hartigan and Wong, 1979] and hierarchical clustering [Johnson, 1967]) could be used to segment data into distinct pseudo-classes. However, the generated pseudo-classes may not lead to a better multinomial logistic regression performance because the objective functions of K-means and hierarchical clustering are not linked to the objective of optimizing discrimination between the majority and rare classes. In this section, we propose a brute force method, Genetic algorithm (GA) [Haupt and Ellen Haupt, 2004], to relabel the minority class data. The objective is searching over the possible mappings to the pseudo-classes for some well separated clusters among the minority class. Different from the unsupervised clustering methods, what GA is searching for is relabeled pseudo-classes which maximizes the objective of improving classification performance on the original $Y \in \{0, 1\}$ binary problem.

GA is an optimization method which is suitable for solving a variety of optimization problems. GA consist of five phases:

1. GA starts with a set of initial individuals (i.e. several guesses of the answer to the problem need to be solved), which is called the initial population. Each individual is made up with a set of character strings (e.g. 0/1 coding), named genes.

2. Fitness function is an objective function to be optimized. The output of the fitness function is a fitness score, which reflects how good the solution is. Fitness score decides the probability of each individual will be selected for next reproduction.

3. A portion of the fittest pairs of individuals will be selected as parents to pass their genes to the next generation.

4. Offspring are created by crossover genes between parents and replace current parents to form a new population.

5. In order to keep the diversity of the population, some genes in the offspring can be changed with a very low random probability. This process is known as mutation.

GA terminates when the population has converged, i.e. new offspring do not give a better fitness score. Although, GA is computationally slow and has no guaranteed convergence to an optimal solution [Kumar et al., 2010], in general, GA can provide solutions to a variety kinds of problems. We use the AUC on the training set as our objective function. By deploying GA, we attempt to split the minority class into several very separated pseudo-classes, which leads to a better multi-class logistic regression performance, i.e. higher AUC for classifying $Y \in \{0, 1\}$.

4.2.1 Algorithm Description

Next, we illustrate how to deploy GA to relabel the minority class. Figure 4.3 illustrates the relabeling process ($Y = 0$ represents the majority class and $Y = 1$ represents the minority class):

1. Classify the minority class data into K clusters $\{Z_1, \cdots, Z_K\}$ by K-means, as an initial population for GA to optimize.
2. Modify the minority clusters by deploying GA, the fitness function is the training set AUC.
   A. In each iteration, we train multi-class logistic regression based on the current clusters.
   B. Calculate the predicted probability of $(Y = 1)$ in training set, $\Pr(y_i = 1|\mathbf{x}) = \sum_{k=1}^{K} \Pr(z_i = k|\mathbf{x}); i \in \{1, \cdots, n\}$.
   C. Calculate the training set AUC, and modify minority clusters based on current clusters.
3. Relabel the minority class into several pseudo-classes based on the the final cluster provided by GA.

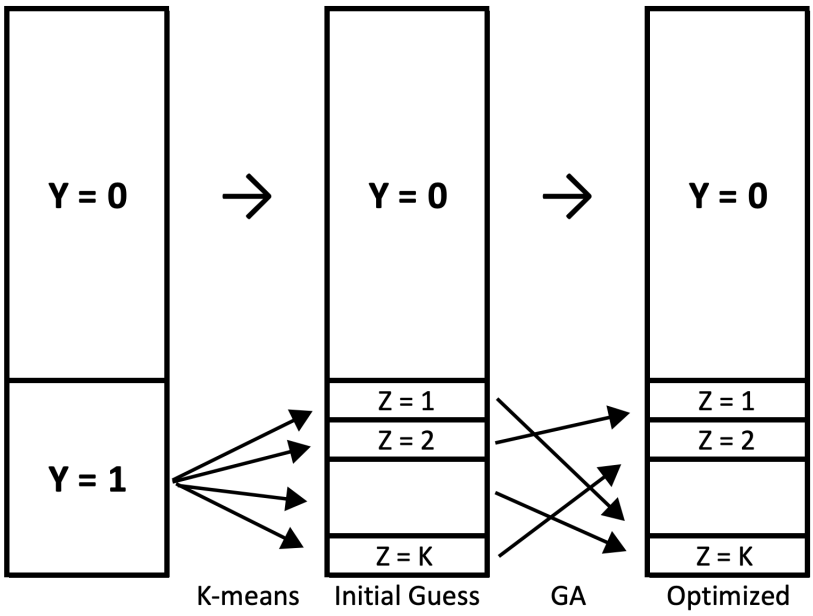


Figure 4.3: Illustration of the relabeling approach by GA.

We conduct a simple simulation still based on the simulation example (see Figure 4.2); GA successfully split the minority class into two designed pseudo-classes.

4.2.2 Experiment

In this section, we investigate this relabeling approach on the Freddie Mac mortgage credit data.

Data Description

Freddie Mac is a U.S. government-sponsored enterprise that purchases mortgage loans for later sale as part of mortgage-backed securities. Freddie Mac provides their "Single Family Loan-Level Data Set" including the fixed-rate mortgages they purchased from 1999 to 2015.[‡] Here, we define default as a

[‡]Data and detail available at:
http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html

mortgage that is 180 days or more overdue in making a repayment on the home loan. This is a standard definition of default in U.S. financial institutions [OCC, US Department of the Treasury, 2000, page. 1; US Federal Reserve Bank, 2007, page. 20]. The target variable we use is whether a mortgage moved to default status in the two years immediately following the first repayment date. The choice of two years was a balance between having a too long window and having too few defaults. In particular having a one year window would give some quarters with too few defaults for statistical modeling purposes (the default rate is very low, even in two years perspective). The upper plot of Figure 4.4 shows the number of new mortgages booked in each originating *quarter* from 2003 to 2013; the lower plot shows the default rate from 2003 to 2013. The number of applications fluctuates over this long time frame. We find a pronounced peak in default rate during the financial crisis period (2007-2008) with a peak of 6.8% in 2007 Q3; however, the default rate is extremely low in other quarters. Table 4.1 provides a description of the variables in this data set.

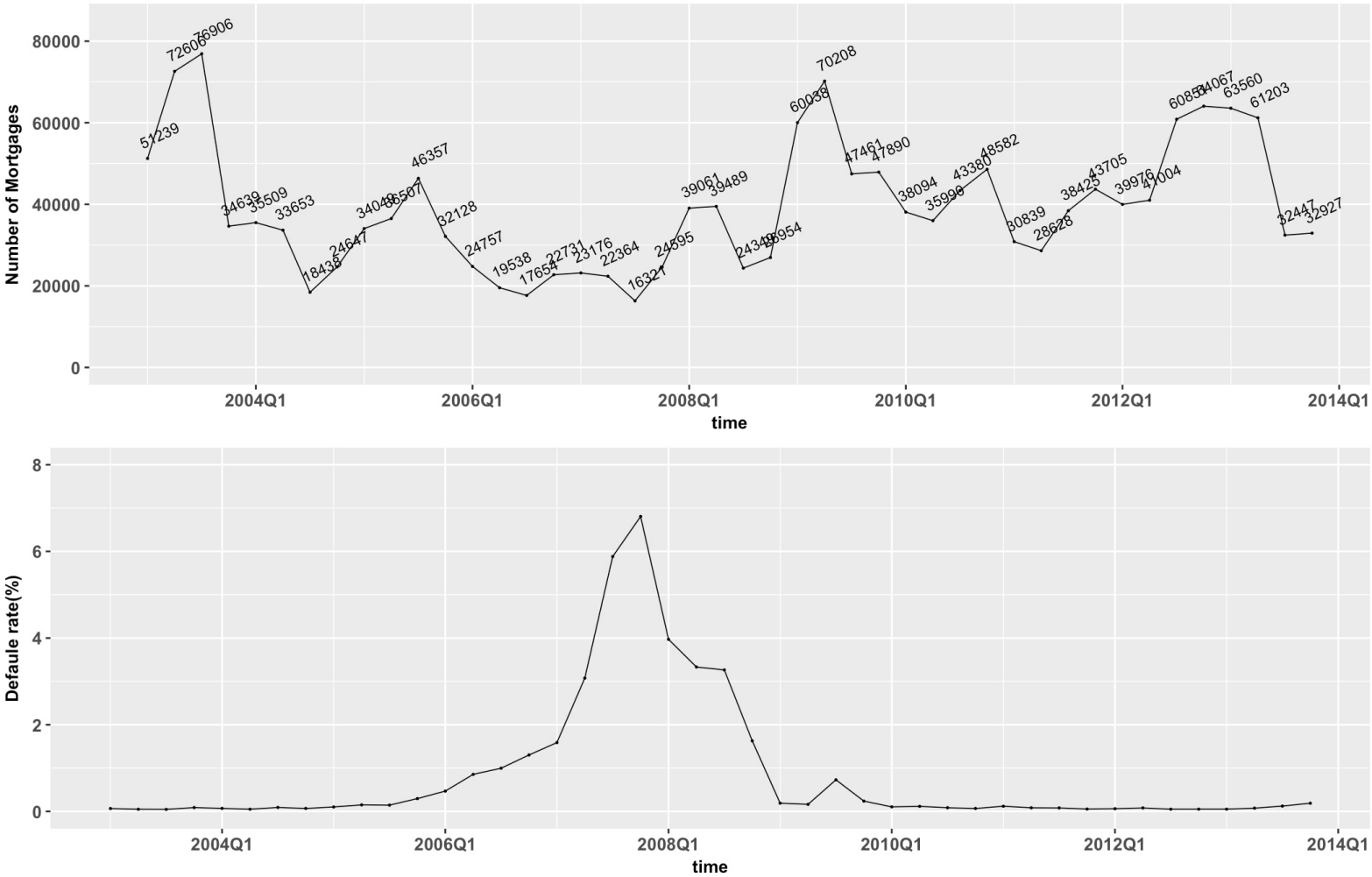


Figure 4.4: Sample size and default rate from 2003 to 2013 in the Freddie Mac data set.

Table 4.1: Description of variables in the Freddie Mac data set

Variable	Type	Description
Default	Categorical	Dependent variable: 1 if borrower greater than 180 days past due on monthly installments;0 otherwise.
Score	Continuous	A number, prepared by third parties, summarizing the borrower's creditworthiness, which may be indicative of the likelihood that the borrower will timely repay future obligations.
DTI	Continuous	Original Debt-To-Income Ratio.
UPB	Continuous	Unpaid Principal Balance.
LTV	Continuous	Original Loan-To-Value.
OIR	Continuous	Original Interest Rate.
Number of Borrowers	Categorical	The number of borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property. 1 = one borrower; 2 = more than one borrower.
Seller	Categorical	The entity acting in its capacity as a seller of mortgages to Freddie Mac at the time of acquisition.
Servicer	Categorical	The entity acting in its capacity as the servicer of mortgages to Freddie Mac as of the last period for which loan activity is reported in the Dataset.
First Time Homebuyer	Categorical	Y =yes; N = no.
Number of Units	Categorical	Denotes whether the mortgage is a one-, two-, three-, or four-unit property.
Occupancy Status	Categorical	O = Owner Occupied; I = Investment Property; S = Second Home; Space = Unknown.
Channel	Categorical	R = Retail; B = Broker; C = Correspondent; T = TPO Not Specified; Space = Unknown.
PPM	Categorical	Denotes whether the mortgage is a Prepayment Penalty Mortgage. Y = PPM; N = Not PPM.
Property Type	Categorical	CO = Condo; LH = Leasehold; PU = PUD; MH = Manufactured Housing; SF = 1-4 Fee Simple; CP = Co-op; Space = Unknown.
Channel	Categorical	R = Retail; B = Broker; C = Correspondent; T = TPO Not Specified; Space = Unknown.
Loan Purpose	Categorical	P = Purchase; C = Cash-out Refinance; N = No Cash-out Refinance; Space = Unknown.

Data Preparation

We use variables in Table 4.1, which give the information of each mortgage application (i.e. mortgage origination data). In the financial industry, the *log* transformation is frequently used to make highly skewed data less skewed and reduce the influence of outliers [Altman and Sabato, 2007]. We deploy the *log* transformation on the variable "UPB", since "UPB" is left skewed in the Freddie Mac data. "Seller" and "Servicer" are transformed from categorical variables to numerical variables using the weight of evidence approach [Thomas, 2009, p. 25], because "Seller" and "Servicer" contain many category levels. All other categorical variables are dummy coded to binary variables.

After dummy coding categorical variables into binary variables we delete the variables which are constant in the minority class, because linear separation appear in some dummy variables (this will make MLE not exist as we explained in Section 3.1.1). For example, the categorical variable "*number of units*" is dummy coded into several indicator variables ("*number.units2*", "*number.units3*", "*number.units4*") (e.g. "*number.units4*" $\in \{0, 1\}$ means whether "*number of units*" equal to 4 for an observation). However, in 2005 for example, some of these categories are barely populated. Thus, after dummy coding the variable"*number of units*" into binary variables, the new dummy variable "*number.units4*" has constant value 0 in the minority (default) class, which makes the coefficient estimate of "*number.units4*" in logistic regression fail to be finite.

Model Building Procedure Description

Table 4.2 explains the experimental procedure. After data preparation, we use data from an individual year (e.g. 2000) as a training set to train five different models: *"logistic"* and *"GA-K"* ($K \in \{2, 3, 4\}$). Here, *"logistic"* refers to modeling a logistic regression based on the prepared data. Model output is the estimated posterior probability of a mortgage account defaulting, given its feature vector.

Table 4.2: Experiment Procedure Times

Training set year	2000	2001	...
Default collection year	2001 2002	2002 2003	...
Testing set year	2003	2004	...

In *"GA-K"* procedure, we use GA to split the minority class into K pseudo-classes. These K pseudo-classes provide K new default classes, which, along with the non-default class data, made up a new *relabeled* data. Then this *relabeled* data are analyzed by multinomial logistic regression. The model output is still the posterior probability of defaulting by summing the posterior probability of K new default classes. Figure 4.5 is an example running GA-2 100 times in training set 2003. The green line represents the best fitness score (training set AUC) among populations in each iteration, and blue line represents the mean fitness score. We observe GA-2 enhance the best fitness score among population from 0.89 to 0.9, and converge after iterating 80 times.

A two-year gap (e.g. 2001-2002) is used for collecting default status information. Using contiguous windows for training is common practice in the credit risk industry. There are two main factors which influence the choice of the time gap:

- On one hand, we need to keep the observation time for default long enough to capture adequate default information.
- On the other hand, we also need to keep the model up-to-date, which requires short observation time. For example, if we increase the two-year gap to four years, the model is unlikely to be up-to-date. This means, for example, a model built on training set 2000 will be used to predict the data of each quarter in 2005 rather than 2003.

The "Two-year gap" is a reasonable choice to balance between having a too long window and having too few defaults. In particular having a one year window would give some quarters with too few defaults for statistical modeling purposes.

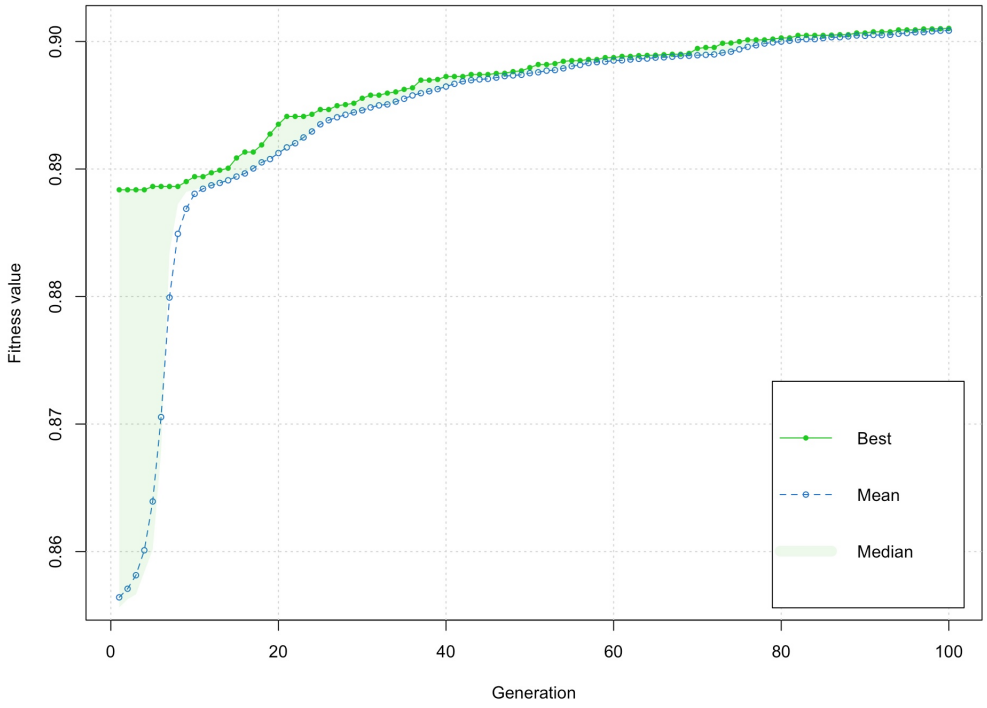


Figure 4.5: Training set AUC (vertical axis) in 2003. Horizontal axis is the number of the generation, green line represents the best fitness score among population in each iteration.

Finally, the obtained models (i.e. "logistic" and " GA-K") are used to forecast the data for the four quarters in the following third testing set year. The performance assessment metric is the AUC. To explore modeling and performance issues over a long time horizon, this procedure is repeated over a ten-year period.

Result and Discussion

The upper plot of Figure 4.6 displays the test-set AUC for each of the methods, over the observation period. We also plot the AUC difference with respect to the benchmark method logistic regression in the lower part of Figure 4.6. Overall, the results indicate that over a long time frame, different model's efficacy varies. We notice that relabeling methods (GA-K) do a good job to enhance the performance in most years, if we can correctly choose the

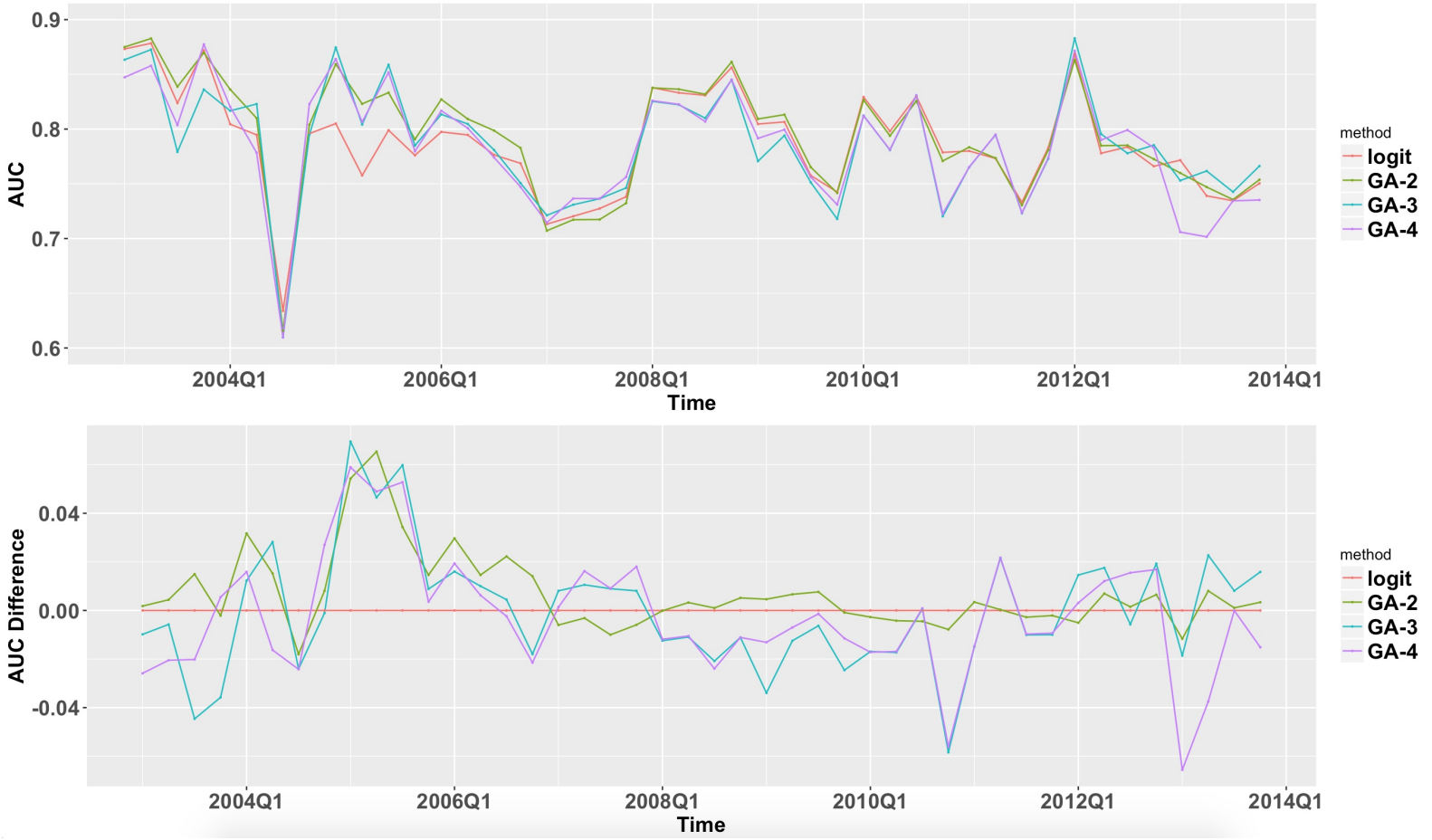


Figure 4.6: Prediction AUC from 2003 to 2013

number of clusters K in advance. We defer how to choose the right number by using cross validation procedure to Section 4.3.3.

In application, it is also meaningful to investigate the microeconomics meaning of different default group. We choose 2007 and 2009 for a further investigation.

- 2007: GA-4 > GA-3 > Logistic Regression > GA-2.
  Training set year 2004,
- 2009: GA-2 > Logistic Regression > GA-3 > GA-4
  Training set year 2006.

Figure 4.7-4.10 illustrate five important variables [Campbell and Cocco, 2015, Bagherpour, 2017] (score, DTI, UPB, LTV and OIR) with two or three minority classes in the year 2007 and 2009, which correspond to the training set in the year 2004 and 2006. We find in 2004, there is no significant difference between LTV and DTI when split minority class into two. However, the significant separation appears in three clusters. The "Default 2" group in three clusters of 2004 (Figure 4.8) shows high credit score, high DTI with high LTV group has a higher risk of default. This may explain why the three cluster relabeling solution works well in this case, by finding different kind of defaults that were not picked up in the basic logistic regression model. In 2006, three clusters produced by GA does not give a clear separation between clusters, but two clusters give a clear separation between the two groups may explain why $K = 2$ works well in 2006.

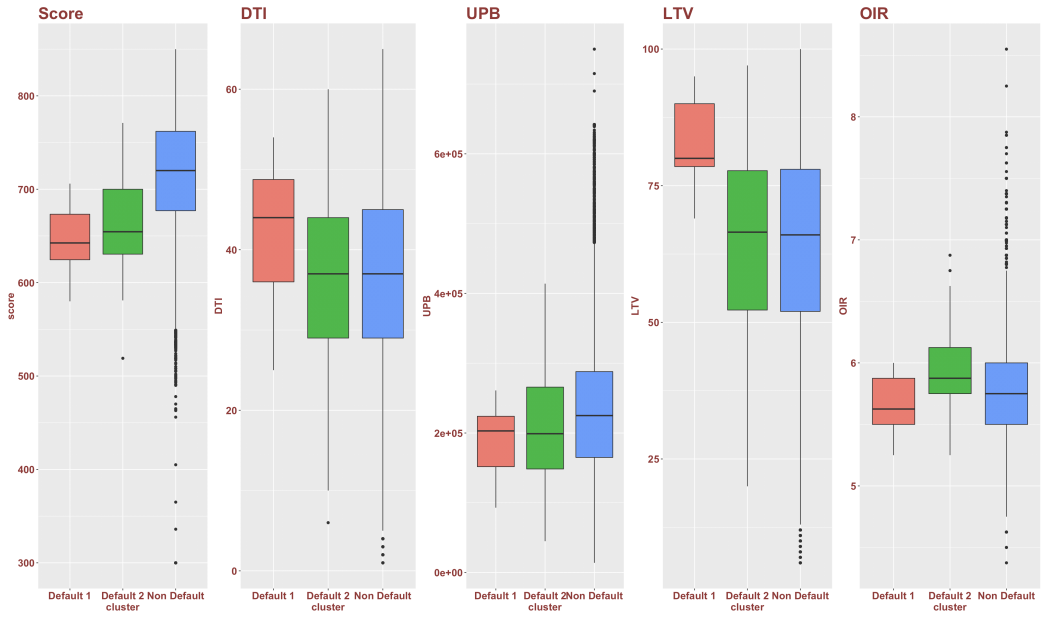


Figure 4.7: Five important variables in 2004, with two minority clusters.

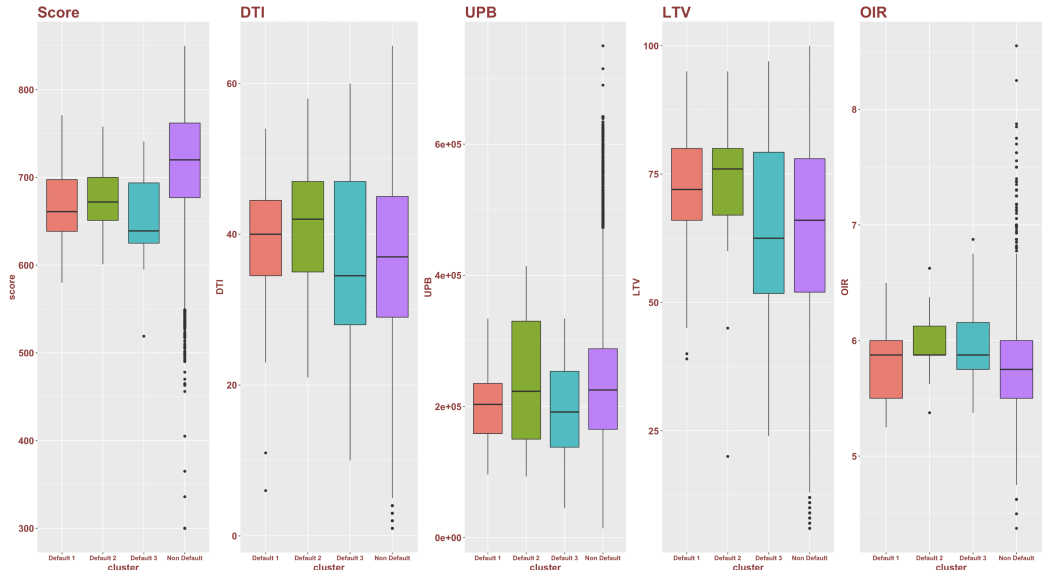


Figure 4.8: Five important variables in 2004, with three minority clusters.

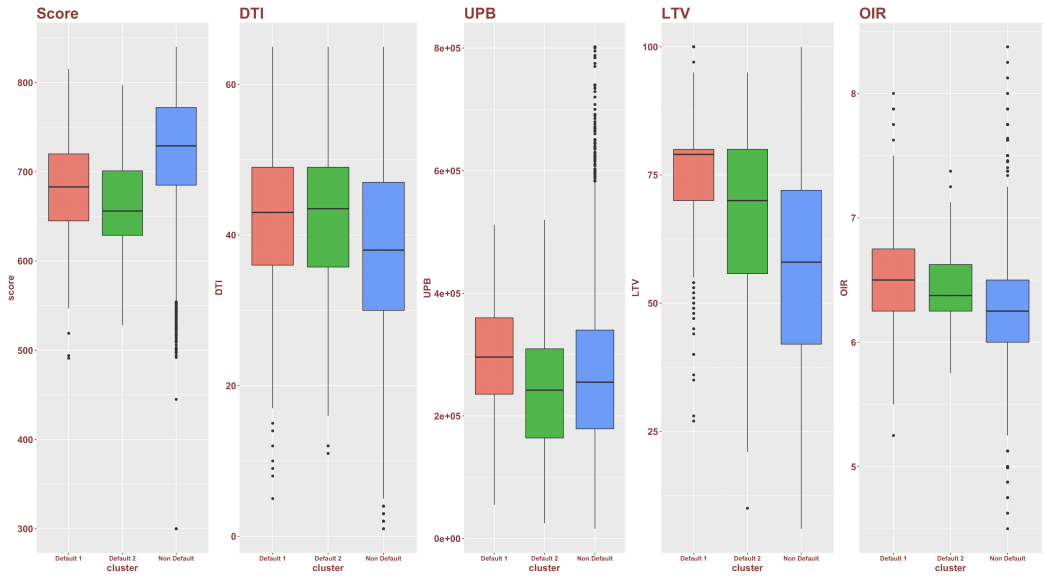


Figure 4.9: Five important variables in 2006, with two minority clusters.

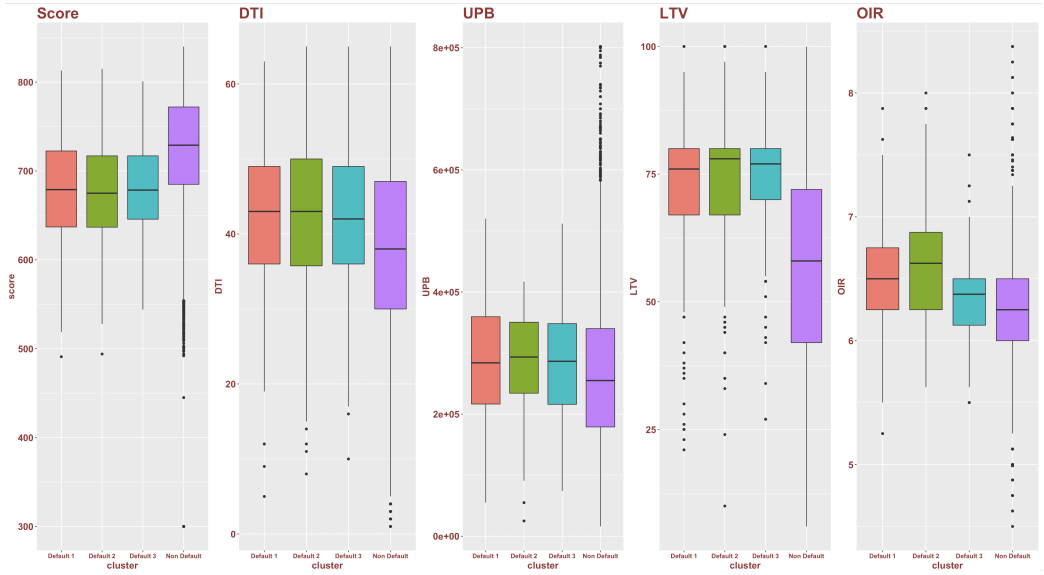


Figure 4.10: Five important variables in 2006, with three minority clusters.

4.3 Expectation Maximization Algorithm

Our GA approach provides promising potential to enhance the logistic regression performance in highly imbalanced data set by relabeling. However, we notice that using the GA to solve the relabeling problem is computationally expensive and time consuming; actually in our Freddie Mac experiment, it takes around two days to relabel a single year on a standard laptop in `R`. This is because it is an optimization over discrete subsets and hence faces a combinatorial explosion. For example, relabeling n minority observations to K pseudo-classes brings K^n potential solutions. Searching this space of relabeling using brute force method like GA is obviously computationally expensive.

To manage the computational burden and the objective of optimizing discrimination, we propose a novel relabeling procedure using the EM algorithm with multinomial logistic regression in this section.

4.3.1 Model Framework Description

The fundamental problem here is relabeling the minority class data into several new pseudo-classes. However, the pseudo-labels of the minority class data $\{\mathbf{x}_{1i}; i \in (1, \ldots, n)\}$ are unobserved. In order to demonstrate this problem as an incomplete-data problem, we introduce latent variable $\{z_i = k; i \in (1, \ldots, n), k \in (1, \ldots, K)\}$ to represent the pseudo-class label. For now, we assume K is known. If we assume z_i arises from a mixture of a finite number of subpopulations in proportions $\Phi = \{\phi_1, \ldots, \phi_K\}$, i.e. $z_i \sim$ mulitinomial(Φ), where $\sum_{k=1}^{K} \phi_k = 1$, then for a fixed observation i, we have case-wise contribution to the likelihood function with pseudo-classes as $\sum_{k=1}^{K} \phi_k \Pr(z_i = k|\mathbf{x}_{1i})$ (Equation 4.1). More specifically, the underlying population of the minority class is modeled as consisting of K distinct pseudo-classes with unknown populations Φ.

Writing down the log-likelihood function of the complete data:

$$L = \sum_{i=1}^{n} \left(\log \sum_{k=1}^{K} \phi_k \big(\Pr(z_i = k | \mathbf{x}_{1i}) \big) \right) + \sum_{i=1}^{N} \log(\Pr(y_i = 0 | \mathbf{x}_{0i})). \tag{4.3}$$

The probability function in Equation (4.3), $\Pr(z_i = k|\mathbf{x}_{1i})$ and $\Pr(y_i = 0|\mathbf{x}_{0i})$, can be modeled by various classification methods (Logistic Regression, Naive Bayes etc). If we consider the majority class as a base class (by simply relabeling $y_i = 0$ to $z_i = 0$), we can use multinomial logistic regression where

$$\textbf{logit}(\Pr(z_i = k|\mathbf{x}_{1i})) = \beta_{0k} + \mathbf{x}_{1i}^T \beta_k, \text{ where } k \in \{1, \ldots, K\}. \tag{4.4}$$

Henceforth, we use β_0 and β to denote the parameter set of scalar β_{0k} and vector β_k where $k \in \{1, \ldots, K\}$ respectively.

Finally, we make two remarks about this model. First, it is possible to reason about a binary classification output by considering $\Pr(Y = 0|X = \mathbf{x}) = 1 - \sum_{k=1}^{K} \Pr(Z = k|X = \mathbf{x})$. Second, this relabeling approach is useful to explore the latent subclasses in many application domains.

4.3.2 EM Algorithm

The parameters Φ, β_0, and β in Equations (4.3) and (4.4) are estimated in the proposed multinomial model. We use the EM algorithm to perform the optimization of this model in the highly imbalanced scenario $n \ll N$.

First, we write the log-likelihood function of the minority class, where $\hat{p}_{ik}$

denotes the estimated posterior probability $\Pr\{z_i = k\}$:

$$
\begin{aligned}
&\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \phi_k \Pr(z_i = k | \mathbf{x}_{1i}; \beta_{0k}, \beta_k) \right) \\
=&\sum_{i=1}^{n} \log \sum_{k=1}^{K} \left(\frac{\hat{p}_{ik} \phi_k \Pr(z_i = k | \mathbf{x}_{1i}; \beta_{0k}, \beta_k)}{\hat{p}_{ik}} \right) \\
\geq&\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik} \log \left(\frac{\phi_k \Pr(z_i = k | \mathbf{x}_{1i}; \beta_{0k}, \beta_k)}{\hat{p}_{ik}} \right) \text{ [\textbf{by Jensen's Inequality}]} \\
=&\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik} \log(\phi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik} \log \left(\frac{\Pr(z_i = k | \mathbf{x}_{1i}; \beta_{0k}, \beta_k)}{\hat{p}_{ik}} \right) \\
=&\underbrace{\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik} \log(\phi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik} \log \left(\Pr(z_i = k | \mathbf{x}_{1i}; \beta_{0k}, \beta_k) \right)}_{\text{to be optimized}} \\
&- \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik} \log(\hat{p}_{ik}).
\end{aligned}
\tag{4.5}
$$

Formula (4.5) provides a lower bound for the part in the log-likelihood function (4.3) related to the minority class.

In the EM algorithm, the E-step is used to construct a local lower-bound to the likelihood function, by conducting a soft assignment of each observation to each class; the lower-bound here is referred to as the Q function. Then, the M-step optimizes this lower bound by improving the coefficient estimates in the Q function. In general, the EM algorithm iteratively repeats E-step and M-step, to enhance the lower boundary of the likelihood, and ultimately reach the maximum likelihood solution. Notice that only the underlined part in Equation (4.5) involves the parameters we seek to estimate; thus, the Q

function for EM to optimize is

$$Q(\phi, \beta_0, \beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{p}_{ik}\log(\phi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{p}_{ik}\log\big(\Pr(z_i = k|\mathbf{x}_{1i}; \beta_{0k}, \beta_k)\big) + \sum_{i=1}^{N} \log(\Pr(y_i = 0|\mathbf{x}_{0i}; \beta_0, \beta)), \tag{4.6}$$

with the additional constraint that the ϕ_k's sum to one. For convenience, let θ denote all the model coefficients in the multinomial logistic regression and Θ denotes all parameters for optimization (model coefficients and latent class probabilities). Since the pseudo label $z_i = k$ is unknown, we start with randomly generated numbers between 0 and 1 as the initial guesses of $\hat{p}_{ik}^{(1)}, k \in \{1, \ldots, K\}, i \in \{1, \ldots, n\}$. Here, $\hat{p}_{ik}^{(1)}$ denotes $\hat{p}_{ik}$ in the first iteration and $\sum_{k=1}^{K} \hat{p}_{ik}^{(1)}$ is constrained to be one for each i. Then, we iterate following the **E - step** and the **M - step**, until convergence:

- **E - step**:

  In the $(u+1)$th iteration, given the current parameter estimates $\Theta^{(u)}$, the posterior probability of the minority observation i belonging to pseudo-class k is:

$$\hat{p}_{ik}^{(u+1)} = \frac{\phi_k^{(u)} \Pr\{z_i = k|\mathbf{x}_{1i}; \theta^{(u)}\}}{\sum_{k=1}^{K} \phi_k^{(u)} \Pr\{z_i = k|\mathbf{x}_{1i}; \theta^{(u)}\}}; i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}. \tag{4.7}$$

  Hence, we have new weighted pseudo-class 'center',

$$\mu_k^{(u+1)} = \sum_{i=1}^{n} \hat{p}_{ik}^{(u+1)} \mathbf{x}_{1i}; i \in \{1, 2, \ldots, n\}, k \in \{1, 2, \ldots, K\}. \tag{4.8}$$

- **M - step**:

  Given the estimated posterior probability $\hat{p}_{ik}^{(u+1)}$ in the E - step, calculate $\Theta^{(u+1)}$ by maximizing

$$Q(\Theta^{(u+1)}|\Theta^{(u)}) = Q_1(\theta^{(u+1)}|\Theta^{(u)}) + Q_2(\phi^{(u+1)}|\Theta^{(u)}), \tag{4.9}$$

  where

$$\begin{aligned} Q_1(\theta^{(u+1)}|\Theta^{(u)}) = &\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik}^{(u+1)} \log(\Pr(z_i = k|\mathbf{x}_{1i}; \theta_k^{(u+1)})) \\ &+ \sum_{i=1}^{N} \log(\Pr(y_i = 0|\mathbf{x}_{0i}; \theta^{(u+1)})), \end{aligned} \tag{4.10}$$

  and

$$Q_2(\phi^{(u+1)}|\Theta^{(u)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_{ik}^{(u+1)} \log(\phi_k^{(u+1)}). \tag{4.11}$$

In the **M - step**, Q_1 and Q_2 can be maximized separately, hence giving the maximizer of $\theta^{(u+1)}$ and $\phi^{(u+1)}$ respectively. Specifically, Q_1 is a weighted likelihood function for multinomial logistic regression, which means we could replace each group's minority class data with their weighted group centers

$\mu_k = \sum_{i=1}^{n} \hat{p}_{ik}\mathbf{x}_{1i}$ and get the same coefficient estimates. The maximizer of Q_2 with constraint $\sum_{k=1}^{K} \phi_k^{(u+1)} = 1$ is given by

$$\phi_k^{(u+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ik}^{(u+1)}. \tag{4.12}$$

The pseudo code for this EM algorithm has been provided in Appendix C.

The time complexity for our EM algorithm is $O(KnI)$ (see [Zhong and Ghosh, 2003]), where I is the number of the iterations needed to solve the multinomial logistic regression Q_1. Typically, the EM algorithm is not as popular as other clustering algorithms when the data set is large [Meilijson, 1989]. In practice, due to the nature of highly imbalanced data, n may be very small, making our algorithm viable for large problems (i.e. when N is large).

The algorithm terminates when relative changes in the likelihood function (4.6) are deemed sufficiently small. In our experience, terminating the process when the change is less the 0.01% is sufficient, usually resulting in convergence within 50 iterations. Upon convergence, we construct a relabeled data set by assigning each minority class observation $\mathbf{x}_{1i}$ a label $z_i = k$, where $k = \arg\max_k \big(\Pr(z_i = k|\mathbf{x}_{1i})\big)$. This relabeled data set is then amenable to standard multinomial logistic regression.

A general proof for the monotonicity of the EM algorithm is given by [Tanner, 2012, p. 34, Theorem 1], and we also provide a proof of monotonicity for the proposed algorithm.

Theorem 16. *Assume when* $Q(\Theta|\Theta^{(u)}) \geq Q(\Theta^{(u)}|\Theta^{(u)})$, *we have* $l(\Theta) \geq l(\Theta^{(u)})$.

Proof. Let Equation (4.5) be $Q(\Theta|\Theta^{(u)})+h(Z|X;\Theta^{(u)})$, where $h(Z|X;\Theta^{(u)}) = -\hat{p}_{ik}\log(\hat{p}_{ik})$. Now we have $l(\Theta) \geq Q(\Theta|\Theta^{(u)}) + h(Z|X;\Theta^{(u)})$.

Consider a special case when $\Theta = \Theta^{(u)}$, let $C_i^{(u)} = \sum_{k=1}^{K} \phi_k^{(u)}\Pr(z_i = k|\mathbf{x}_{1i};\theta^{(u)})$,

we notice that

$$\begin{aligned}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{p}_{ik}\log\left(\frac{\phi_k^{(u)}\Pr(Z_i=k|\mathbf{x}_{1i};\theta^{(u)})}{\hat{p}_{ik}}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\phi_k^{(u)}\Pr(z_i=k|\mathbf{x}_{1i};\theta^{(u)})}{C_i^{(u)}} \\ &\quad\times\log\left(\phi_k^{(u)}\Pr(z_i=k|\mathbf{x}_{1i};\theta^{(u)})\frac{C_i^{(u)}}{\phi_k^{(u)}\Pr(z_i=k|\mathbf{x}_{1i};\theta^{(u)})}\right) \\ &= \sum_{i=1}^{n}\frac{\sum_{k=1}^{K}\phi_k^{(u)}\Pr(z_i=k|\mathbf{x}_{1i};\theta_k^{(u)})}{C_i^{(u)}}\log(C_i^{(u)}) \\ &= \sum_{i=1}^{n}\log(C_i^{(u)}),\end{aligned} \tag{4.13}$$

thus, after simplification, we have $l(\Theta^{(u)}) = Q(\Theta^{(u)}|\Theta^{(u)}) + h(Z|X;\Theta^{(u)})$.

By combining the assumption in the theorem, we have

$$\begin{aligned}l(\Theta) &\geq Q(\Theta|\Theta^{(u)}) + h(Z|X;\Theta^{(u)}) \\ &\geq Q(\Theta^{(u)}|\Theta^{(u)}) + h(Z|X;\Theta^{(u)}) \\ &= l(\Theta^{(u)}),\end{aligned} \tag{4.14}$$

which states that improving the Q-function will at least not make the log-likelihood worse. □

We conclude our EM algorithm description by conducting a simple experiment on a contrived example which similar to the example we mentioned in Section 3.1.2, but with different pseudo-classes proportion. 2,000 sample points are generated following $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then 50 points following $X \sim N(\mu_1, \Sigma_1)$ and 100 points $X \sim N(\mu_2, \Sigma_2)$ as two minority class pseudo-classes. Here

$$\mu_0 = [0, 0], \mu_1 = [1, 1.8], \mu_2 = [1, -1.8],$$

$$\Sigma_0 = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \text{ and } \Sigma_1 = \Sigma_2 = \begin{bmatrix}0.2 & 0\\ 0 & 0.2\end{bmatrix}.$$

We deliberately make the number of the observations in two minority class

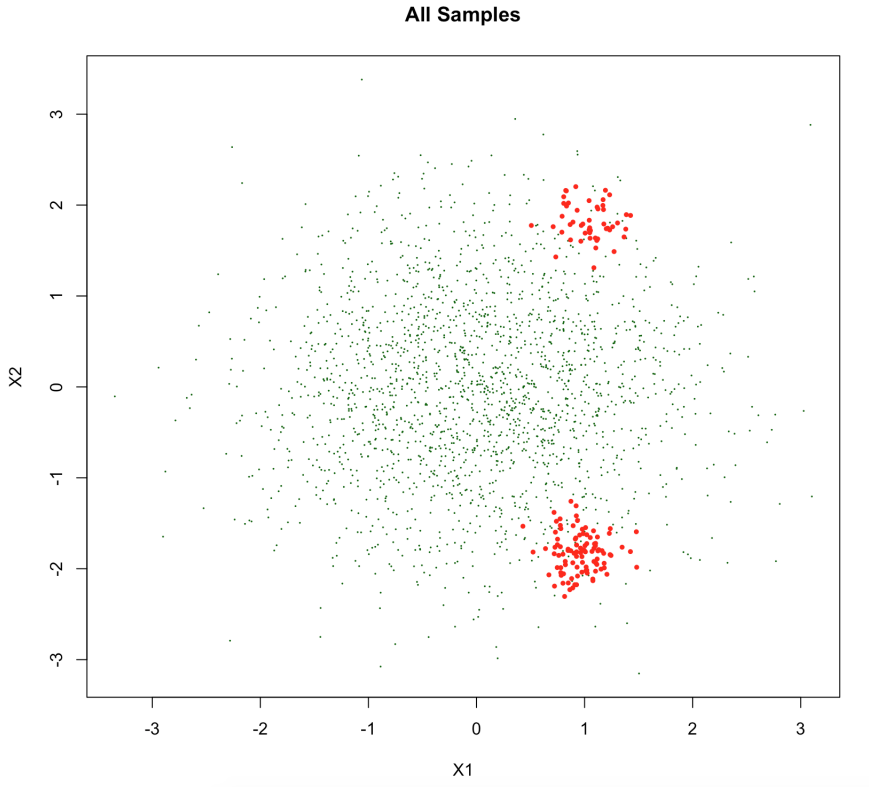


Figure 4.11: Scatter plots of simulation samples, including 2,000 majority observations and 150 minority observations.

pseudo-classes different, in order to check our EM algorithm performance (see Figure 4.11).

We iterate our EM algorithm 50 times with the setting $K = 2$ (how to select K will be discussed in Section 4.3.3). Figure 4.12 tracks $\hat{\phi}_1$ and $\hat{\phi}_2$ (proportions of the minority class) in each iteration. Figure 4.13 tracks the iteration of $\hat{p}_{i1}, i \in \{1, \cdots, n\}$ (the probability that minority class observation i belongs to pseudo-class 1). We find our EM algorithm successfully converges to the target proportions (50:100, which are the number of observations in two minority class clusters). Figure 4.14 gives the relabeling result by relabel each minority class observation into pseudo-class k based on

$$k = \arg\max_k \hat{p}_{ik}, k \in \{1, 2\}$$

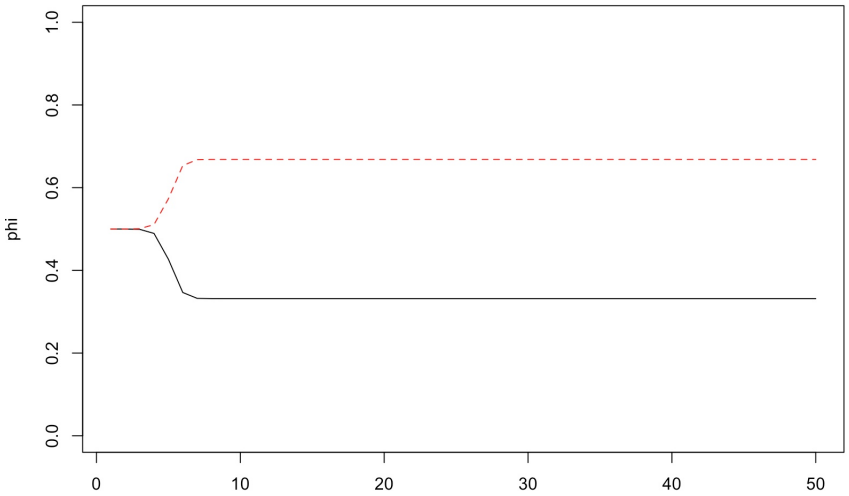


Figure 4.12: Proportions ($\hat{\phi}_1$ and $\hat{\phi}_2$) in 50 times iteration.

after 50 times iteration (blue and brown in Figure 4.14); we see our EM algorithm generate two well separated pseudo-classes.

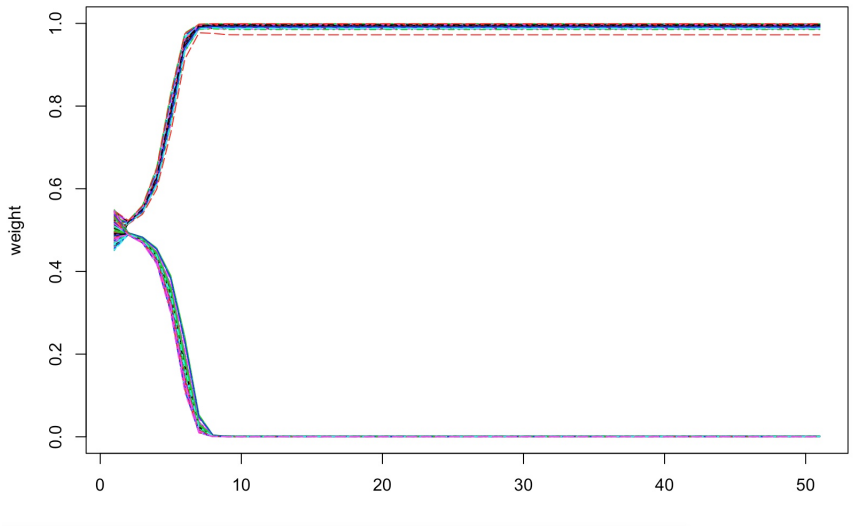


Figure 4.13: $\hat{p}_{i1}$ (the probability of observations belong to cluster 1) in 50 times iteration. $\hat{p}_{i1}$ converge to 1 or 0 means observation i belongs to cluster 1 or 2 respectively.

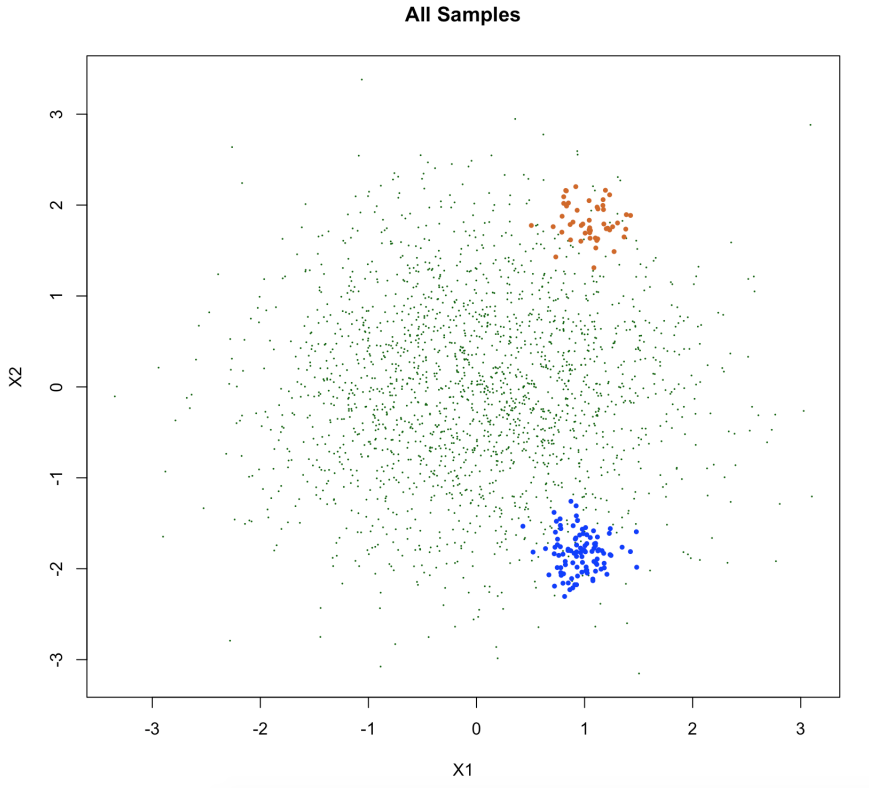


Figure 4.14: Scatter plots of simulation samples after relabeling, including one pseudo-class in blue and another one in brown.

Simulation Comparison between EM Relabeling Method and SMOTE

In this section, we use a simulation study to demonstrate the performance of our EM relabeling method and the SMOTE oversampling method (described in Section 2.3.1), when there are {two close clusters, two well separated clusters, or three clusters} in the minority class.

In our simulation, for the majority class $Y = 1$ (green points in Figure 4.15), we generate 2000 samples $X \sim N(\mu_0, \Sigma_0)$, where

$$\mu_0 = [0, 0], \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For the minority class (red points in Figure 4.15), we investigate three different scenarios:

- **Scenario 1, two close clusters structure** in the minority class $Y = 0$: we generate 75 samples for each cluster, which follows $X \sim N(\mu_1, \Sigma_1)$ and $X \sim N(\mu_2, \Sigma_2)$ respectively, where

$$\mu_1 = [1.5, 1], \mu_2 = [1.5, -1], \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

- **Scenario 2, two well separated clusters structure** in the minority class $Y = 0$: we generate 75 samples for each cluster, which follows $X \sim N(\mu_1, \Sigma_1)$ and $X \sim N(\mu_2, \Sigma_2)$ respectively, where

$$\mu_1 = [1.5, 2], \mu_2 = [1.5, -2], \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

- **Scenario 3, three clusters structure** in the minority class $Y = 0$: we generate 50 samples for each cluster, which follows $X \sim N(\mu_1, \Sigma_1)$,

$X \sim N(\mu_2, \Sigma_2)$, and $X \sim N(\mu_3, \Sigma_3)$ respectively, where

$$\mu_1 = [1.5, 2], \mu_2 = [1.5, 0], \mu_2 = [1.5, -2], \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

For the relabeling method, we use our EM algorithm to relabel the minority class ($Y = 0$) into two pseudo-classes in Scenarios 1 and 2, and three pseudo-classes in Scenario 3; these pseudo-classes together with the majority class constitute a relabeled training set. A multinomial logistic regression is trained on the relabeled data and the model output is the summation of the posterior probability of pseudo-classes. How to select the correct number of the pseudo-classes among the minority will be discussed in the next section.

For SMOTE, we oversample the minority class 10 times greater, i.e. generate 1350 synthetic minority class observations, which makes the size of the minority class increase to 1500. SMOTE has a hyperparameter k (the number indicating the number of nearest neighbors that are used to generate the synthetic examples of the minority class), where we will try $k = 5$ (the default setting from its creators [Chawla et al., 2002]) and a large $k = 50$. After SMOTE oversampling, we train a vanilla logistic regression on the oversampling data set. Figure 4.15 shows the scatter plots of the original data and the SMOTE oversampling data, where the left most plots in each row are the original data in each scenario, and the right two plots in each row are the scatter plots of the SMOTE oversampling data sets ($k = 5$ and $k = 50$).

In each scenario, we replicate our simulation 1000 times by modeling on the training set and deploy them on the corresponding test set which has the same distribution as the training set. Table 4.3 gives the average AUC and H-measure with their corresponding standard deviation on the test sets in different scenarios. We find that:

- regarding the AUC, the relabeling method gives a significant better performance in scenarios 2 and 3 (note the differences are higher than two times standard deviation). The relabeling method also provides some improvement in scenario 1.

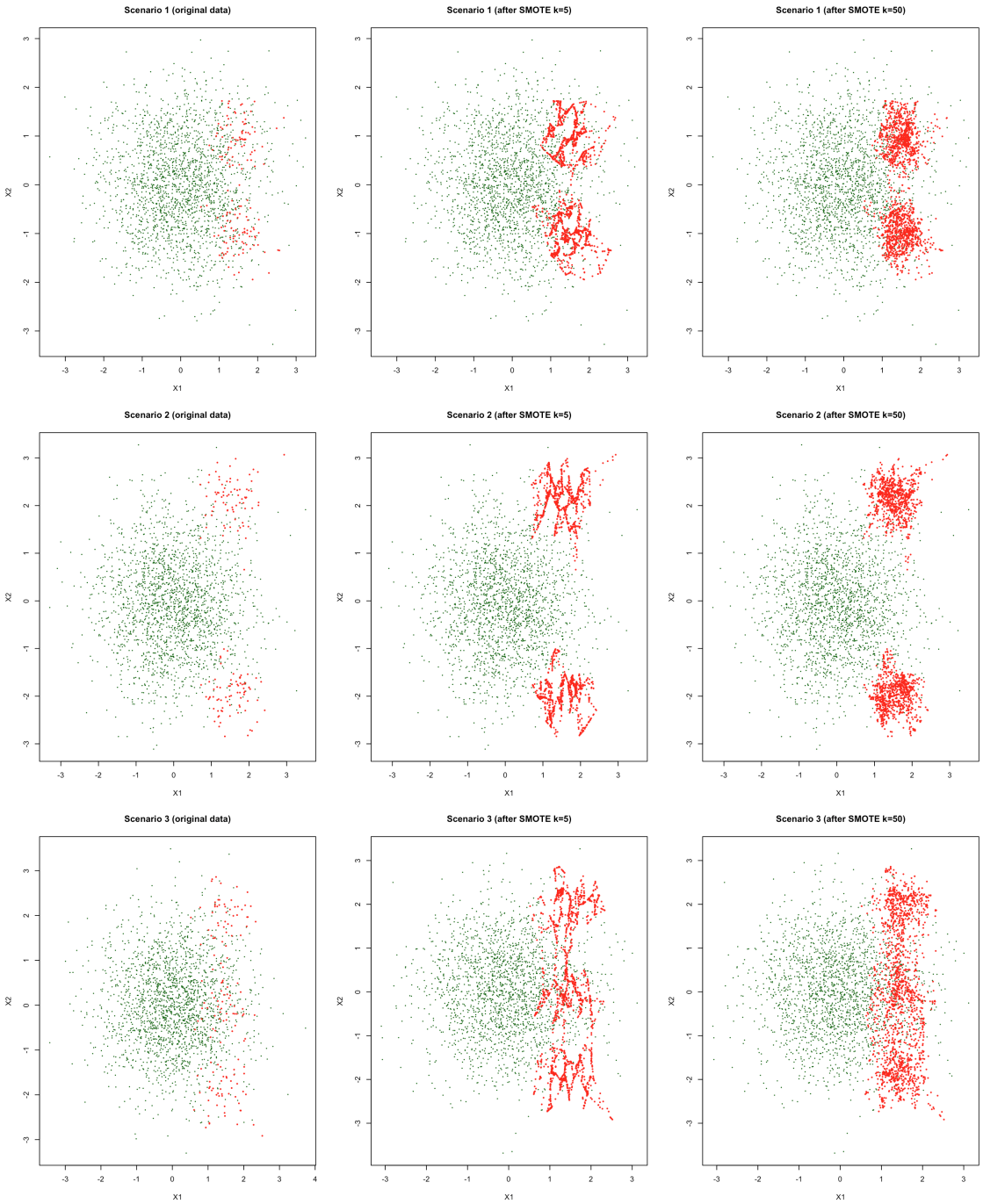


Figure 4.15: Scatter plots of Scenario 1 to 3, green points represent the majority class $Y = 1$ and red points represent the minority class $Y = 0$. The left most plots in each row are the original data in each scenario, and the right two plots in each row are the scatter plots of the SMOTE oversampling data sets ($k = 5$ and $k = 50$).

- regarding the H-measure, the relabeling method gives a better performance in scenario 2 (note the differences are higher than two times standard deviation). The relabeling method also provides some improvement in scenarios 1 and 3.

Overall, when the cluster structure is present among the minority class, our relabeling method will outperform SMOTE in this simulation study. We also try to deploy the EM relabeling method on the imbalanced data set where no cluster structure exists in the minority class. The EM algorithm will generate an empty pseudo-class hence still give a vanilla logistic after running EM. The reasons for this phenomenon along with how to choose the correct number of pseudo-classes K are given in the next section.

Table 4.3: Average AUC and H-measure with their corresponding standard deviation on the test sets in different scenarios.

Scenario 1: two close structures			
	Relabeling	SMOTE k=5	SMOTE k=50
AUC	0.9211 (0.0060)	0.9177 (0.0067)	0.9177 (0.0067)
H-measure	0.5923 (0.0257)	0.5899 (0.0252)	0.5900 (0.0251)
Scenario 2: two well separated structures			
	Relabeling	SMOTE k=5	SMOTE k=50
AUC	0.9430 (0.0054)	0.9171 (0.0069)	0.9172 (0.0069)
H-measure	0.6330 (0.0230)	0.5811 (0.0257)	0.5829 (0.0256)
Scenario 3: three clusters			
	Relabeling	SMOTE k=5	SMOTE k=50
AUC	0.9317 (0.0061)	0.9178 (0.0069)	0.9178 (0.0069)
H-measure	0.6279 (0.0252)	0.5906 (0.0263)	0.5906 (0.0254)

4.3.3 Identification of the Number of Clusters

In this section, we briefly touch the problem of identifying the number of pseudo-classes K. In general, selecting the unknown number of latent groups is a challenging [Ketchen and Shook, 1996]. Among various proposed methods to select the "right" number of groups, information criteria are widely used due to their simplicity. The two most popular information criteria are the Akaike information criterion (AIC) [Akaike, 1974] and the Bayesian information criterion (BIC) [Burnham and Anderson, 2004]. However, these are known to have problems; for example, they may overestimate or underestimate the number of groups for model based clustering [Zhong and Ghosh, 2003]. Besides the information criteria approach, statistical hypothesis testing is another popular procedure, often framed as testing the null hypothesis that there are K pseudo-classes in the minority class against the alternative hypothesis that there are $K+1$ pseudo-classes in the minority class. Unfortunately, the standard likelihood ratio test is not appropriate here, because the test statistic is not asymptotically chi-square distributed [Li et al., 1988, Titterington, 1990]. Bootstrap likelihood ratio test [McLachlan, 1987, Feng and McCulloch, 1996] and Monte Carlo methods [Smyth, 1997] are more precise, however they are computationally expensive. Compared to the above methods, cross validation is effective and fast when the sample size is adequate, and this approach is widely used in model based clustering [Smyth, 2000].

Here we use cross validation on the training set to estimate a performance measure for each choice of K. Motivated by problems in retail finance and its insensitivity to class prior distribution, we use the Area Under the ROC Curve (AUC) as a performance measure, though many other choices are valid and reasonable. In each iteration (fold) of cross validation, the procedure described above is used, namely: relabel the data to K pseudo-classes, fit a multinomial logistic regression, and estimate the posterior probability of each test set observation belonging to the base class. The latter stage provides the means to evaluate the AUC, which is then averaged over cross validation folds. The number of pseudo-classes, K, is searched in increasing order, and the process terminates when the estimated AUC decreases.

It is worth to note that the average AUC value from cross validation procedure when we relabel the minority class to K pseudo-classes may equal to the AUC when we relabel to $K+1$ pseudo-classes (and we will select K when this happen). There are two reasons for this phenomenon:

- the estimated mixture has an empty pseudo-class $\tilde{k}$ (i.e. $\phi_{\tilde{k}} \approx 0$ for a $\tilde{k} \in \{1, \ldots, K\}$), because $\hat{p}_{i\tilde{k}} \approx 0$ for all $i \in \{1, 2, \ldots, n\}$ may happen, and this leads to $\phi_{\tilde{k}} = \sum_{i=1}^{n} \hat{p}_{i\tilde{k}} \approx 0$,
- the estimated mixture has several identical pseudo-classes ($\{\beta_{0k}, \beta_k\} = \{\beta_{0k'}, \beta_{k'}\}$ for some $k, k' \in \{1, \ldots, K\}, k \neq k'$).

These could be investigated by comparing $\{\phi, \beta_0, \beta\}$ after running EM.

A Simulation Study for Cross Validation Procedure

In this section, we use a simulation study to show our cross validation method correctly selects K when there are {no clusters, close clusters, well separated clusters} in the minority class.

For the majority class $Y = 1$ (green points in Figure 4.16), we generate 10000 samples $X \sim N(\mu_0, \Sigma_0)$, where

$$\mu_0 = [0, 0], \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For the minority class (red points in Figure 4.16), we investigate four different scenarios:

- Scenario 1, **no cluster structure** in the minority class $Y = 0$: we generate 150 samples $X \sim N(\mu_1, \Sigma_1)$, where

$$\mu_1 = [1.5, 0], \Sigma_1 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

- Scenario 2, **two close clusters structure** in the minority class $Y = 0$: we generate 75 samples for each cluster, which follows $X \sim N(\mu_1, \Sigma_1)$ and $X \sim N(\mu_2, \Sigma_2)$, where

$$\mu_1 = [1.5, 1], \mu_2 = [1.5, -1], \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

- Scenario 3, **two well separated clusters structure** in the minority class $Y = 0$: we generate 75 samples for each cluster, which follows $X \sim N(\mu_1, \Sigma_1)$ and $X \sim N(\mu_2, \Sigma_2)$, where

$$\mu_1 = [1.5, 2], \mu_2 = [1.5, -2], \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

- Scenario 4, **three clusters structure** in the minority class $Y = 0$: we generate 50 samples for each cluster, which follows $X \sim N(\mu_1, \Sigma_1)$, $X \sim N(\mu_2, \Sigma_2)$ and $X \sim N(\mu_3, \Sigma_3)$, where

$$\mu_1 = [1.5, 2], \mu_2 = [1.5, 0], \mu_3 = [1.5, -2], \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

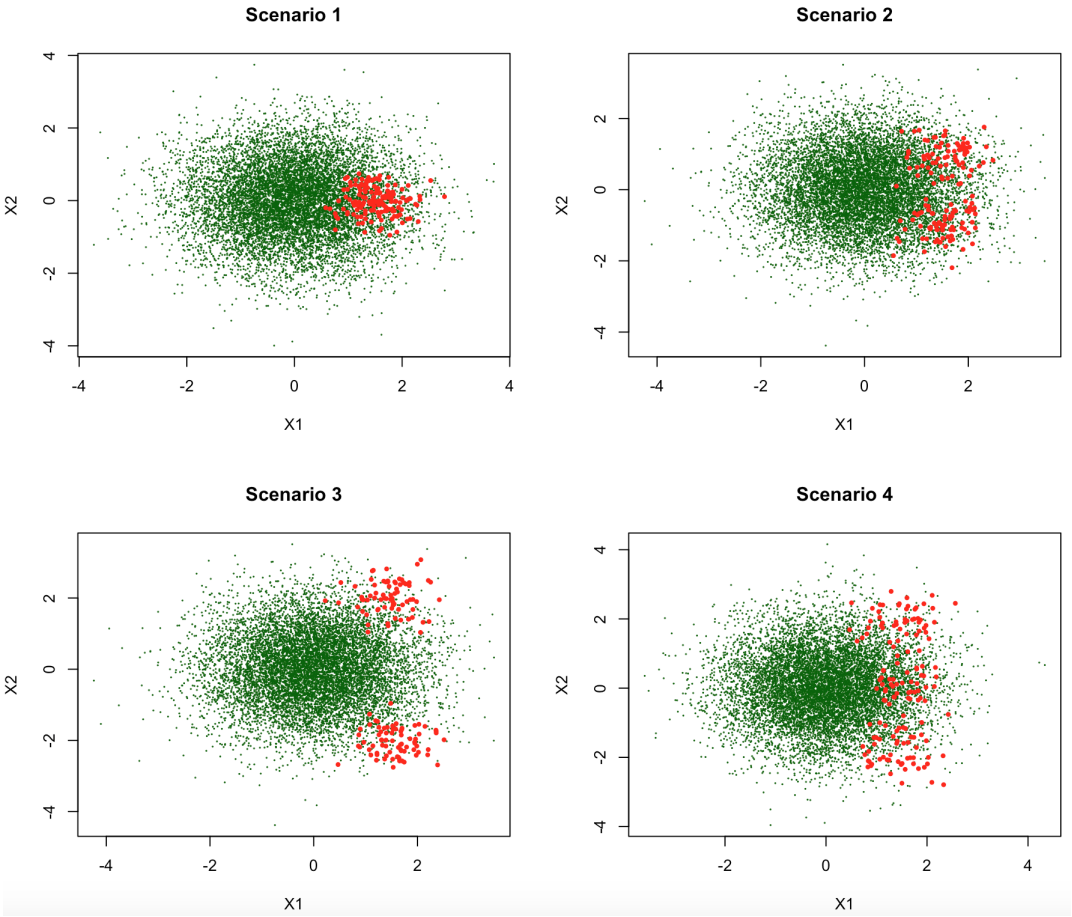


Figure 4.16: Scatter plot for Scenario 1 to 4, green points represent the majority class $Y = 1$, red points represent the minority class $Y = 0$.

We deploy our cross validation procedure with $K = 1, 2, 3$ in these scenarios; the average AUC are displayed in Figure 4.17. In Scenario 1, there is no difference regarding the AUC, and we select $K = 1$. In Scenarios 2 and 3, we select $K = 2$, and we can see that, when there well separated cluster structure among the minority class, $K = 2$ gives a more observable difference in AUC comparing to $K = 1$ and 3. In Scenarios 4, we select $K = 3$. Overall, the cross validation procedure selects the correct K which we set in different scenarios.

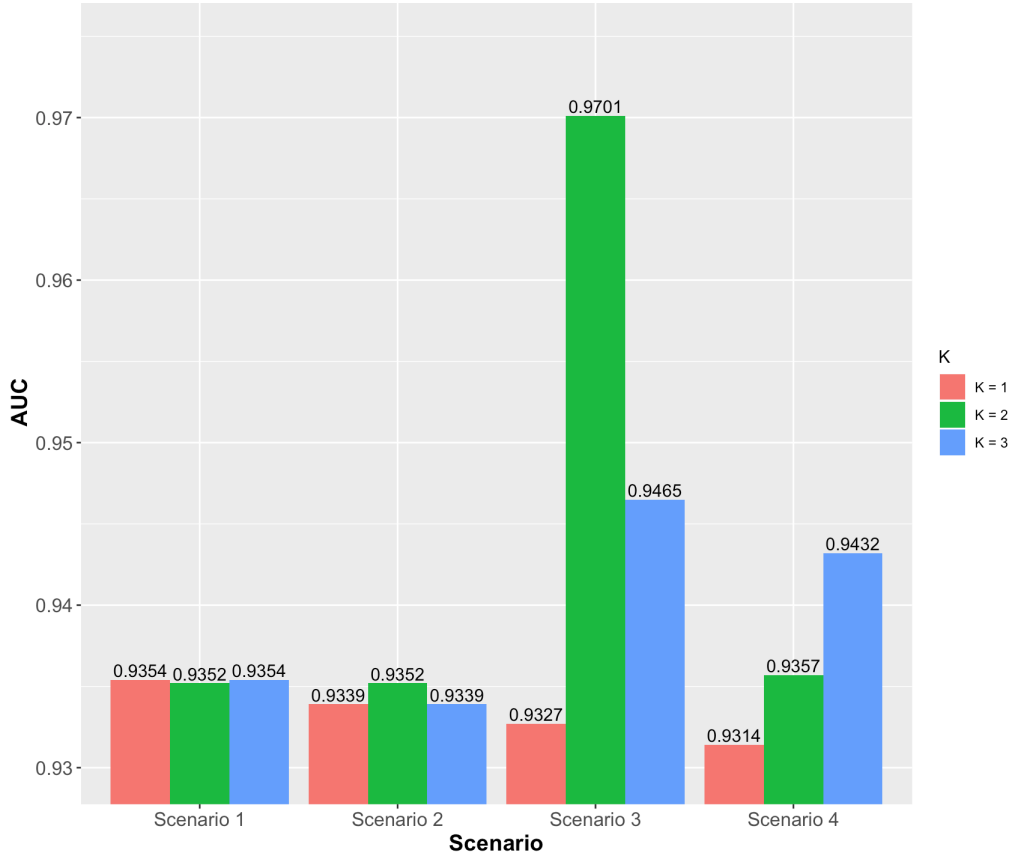


Figure 4.17: The average AUC obtained by cross validation procedure in Scenarios 1 to 4.

4.3.4 Experiment 1: Nested Cross Validation

We have proposed an EM algorithm for relabeling the minority class data into several distinct pseudo-classes. In this section and the next section, we illustrate some experiments to demonstrate our algorithm's effectiveness. Five binary classification data sets are used in this section. Nested cross validation experiments are conducted on five data sets.

Data Description

Five binary classification data sets are introduced in this section which will be used in our experiments. We report the number of variables, the proportion of the minority class and the source in Table 4.4. All data sets include some independent variables. Furthermore, each data set has a binary variable which indicates whether a target event happens. Details of these data are available at the source mentioned in Table 4.4.

Table 4.4: Summary of the experimental data sets

Data	*Variables*	*Proportion minority*	*Number of observation*	*Source*	*Target variable*
Taiwan credit card	23	22.12%	30000	[Yeh and Lien, 2009]	credit card default
European credit card transaction	29	0.17%	284807	[Dal Pozzolo et al., 2015]	fraudulent transaction
Bank telemarketing	20	11.26%	41188	[Moro et al., 2014]	successful telemarketing
Lending Club loan	33	11.95%	157085	[Wendy, 2004]	loan default
Loan recovery data	21	2.17%	8237	Confidential [Ye and Bellotti, 2019]	full recovery

Experimental Procedure

We randomly split each data set into ten folds. In each iteration, nine folds of data serve as the training set and one fold as the test set. Among the training data, we use the cross validation strategy described in Section 4.3.3, to identify the number of pseudo-classes K. To summarize, the inner cross validation loop is used for identifying the K, and the outer cross validation loop is used to measure the performance in the inner loop.

Results

Tables 4.5 to 4.8 give the mean of the AUC and the H-measure with their corresponding standard deviation in the ten-fold cross validation procedure and the test set respectively. The boldface text shows the number of pseudo-classes K selected by cross validation. For comparison, the relabeled model and standard logistic regression are deployed on the test set. The latter corresponds to $K = 1$ in the tables.

With respect to the AUC value, we find that relabeling the minority class into two pseudo-classes does enhance the performance on the Taiwan credit card data, European credit card fraud transaction data, and loan recovery data. In each case, $K = 2$ was the preferred choice. Especially for the Taiwan credit card data and loan full recovery data, we see the difference of the AUC from $K = 1$ to $K = 2$ are greater than twice of their corresponding standard deviation, which are significant improvements. Through the improvement in European credit card fraud transaction data are small; they are substantive in a big population of credit card transactions; a 2% increases in AUC may have substantial financial value. With the bank telemarketing data and Lending Club data, although relabeling the minority class data into two or three pseudo-classes did not enhance the predictive ability of the model, the cross validation procedure does correctly choose $K = 1$; hence the procedure does *not* reduce prediction performance. The reason of why $K = 1$ was the preferred choice in these two data set is that possibly there are no sharp cluster structure among the pseudo-classes. For an ad hoc analysis,

Figure 4.18 gives the boxplot and pie chart of some significant risk factors, which has been highlighted in [Janio, 2017] for Lending Club data, when we relabel the minority class into $K = 2$ new pseudo-classes. As we can seen in Figure 4.18, there are no significant difference between new pseudo-classes "Default 1" and "Default 2", and our cross validation procedure do choose $K = 1$ here. For the H-measure, we see similar results to the AUC; significant improvements still exist in Taiwan credit card data and loan full recovery data. Once pseudo-classes are estimated using the EM algorithm, we anticipate that they may express useful information regarding the minority class in the application domain. For some analysis of the characteristics between different pseudo-classes, see Appendix B.1.

We also provide indicative computation times in Tables 4.5 and 4.6. All of these experiments are conducted on the same computer[§]. It is interesting to notice that the computation time for $K = 2$ and $K = 3$ on the Taiwan credit card data and Lending Club data do not strictly follow the time complexity rule mentioned in Section 4.3.2. This is due to the EM algorithm terminating early when one of the $\phi_k = 0$ (a possibility desired above). This also explains why $K = 3$ gives identical AUC to $K = 1$ for the Taiwan credit card data and Lending Club data.

For the fraud detection problem, another important measure for model assessment is the fraud detection rate with a fixed alarm rate [Olszewski, 2014]. For example, if the bank pre-defines the alarm rate at 0.5%, we require to know how many fraud transactions could be detected. This performance metric is meaningful and vital in real application [Hand et al., 2008]. We provide the detected fraud rate (with alarm rate 0.5%) of the European credit card transaction data in Table 4.9. A notable enhancement from $K = 1$ to $K = 2$ is observed.

[§]Apple iMac with 4.2 GHz Intel Core i7 processor and 32 GB 2400 MHz DDR4 memory

Table 4.5: Mean and standard deviation of the **AUC** from **the cross validation** experiment. The computation times are presented in brackets. Note that K refers to the number of pseudo-classes in the minority class and $K = 1$ refers to standard logistic regression.

Taiwan Credit Card	*K (computation time)*	*Mean AUC*	*Standard deviation*
Cross validation AUC	1	0.7242	0.0022
	2 (270.4 mins)	**0.7545**	0.0030
	3 (351.7 mins)	0.7242	0.0022
Credit Fraud	*K (computation time)*	*Mean AUC*	*Standard deviation*
Cross validation AUC	1	0.9740	0.0027
	2 (570.3 mins)	**0.9747**	0.0024
	3 (834.2 mins)	0.9746	0.0025
Bank Telemarketing	*K (computation time)*	*Mean AUC*	*Standard deviation*
Cross validation AUC	1	**0.7913**	0.0018
	2 (26.9 mins)	0.7829	0.0018
	3 (28.8 mins)	0.7897	0.0020
Lending Club	*K (computation time)*	*Mean AUC*	*Standard deviation*
Cross validation AUC	1	**0.6879**	0.0005
	2 (2606.6 mins)	0.6730	0.0012
	3 (1116.7 mins)	0.6879	0.0005
Loan Full Recovery Data	*K (computation time)*	*Mean AUC*	*Standard deviation*
Cross validation AUC	1	0.8142	0.0070
	2 (12.5 mins)	**0.8497**	0.0074
	3 (23.8 mins)	0.8224	0.0068

Table 4.6: Mean and standard deviation of the **AUC** from the **test set** experiment. The computation times are presented in brackets. Note that K refers to the number of pseudo-classes in the minority class and $K = 1$ refers to standard logistic regression.

Taiwan Credit Card	*K (computation time)*	*Mean AUC*	*Standard deviation*
Test set AUC	1	0.7270	0.0062
	2 (27.4 mins)	**0.7562**	0.0096
	3 (36.8 mins)	0.7270	0.0062
Credit Fraud	*K (computation time)*	*Mean AUC*	*Standard deviation*
Test set AUC	1	0.9742	0.0151
	2 (54.0 mins)	**0.9753**	0.0148
	3 (88.4 mins)	0.9748	0.0142
Bank Telemarketing	*K (computation time)*	*Mean AUC*	*Standard deviation*
Test set AUC	**1**	**0.7913**	0.0163
	2 (3.6 mins)	0.7841	0.0169
	3 (3.8 mins)	0.7899	0.0161
Lending Club	*K (computation time)*	*Mean AUC*	*Standard deviation*
Test set AUC	**1**	**0.6882**	0.0046
	2 (272.5 mins)	0.6730	0.0049
	3 (139.9 mins)	0.6882	0.0046
Loan Full Recovery Data	*K (computation time)*	*Mean AUC*	*Standard deviation*
Test set AUC	1	0.8168	0.0101
	2 (1.5 mins)	**0.8576**	0.0049
	3 (2.1 mins)	0.8272	0.0044

Table 4.7: Mean and standard deviation of the **H-measure** from the **cross validation** experiment. The computation times are presented in brackets. Note that K refers to the number of pseudo-classes in the minority class and $K = 1$ refers to standard logistic regression.

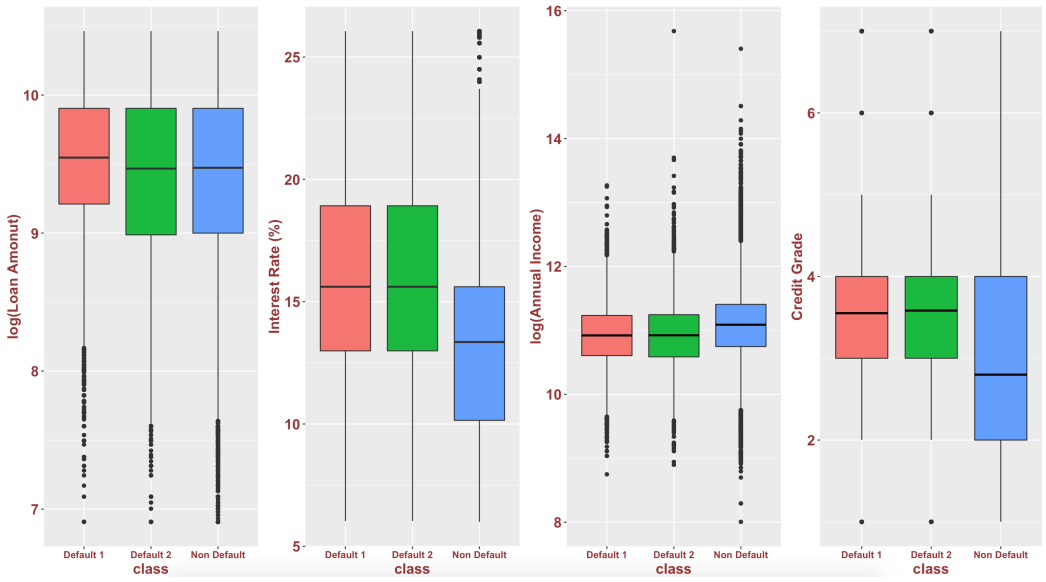
Taiwan Credit Card	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Cross validation	1	0.2488	0.0045
H-measure	2 (270.4 mins)	**0.2537**	0.0030
	3 (351.7 mins)	0.2488	0.0045
Credit Fraud	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Cross validation	1	0.8690	0.0050
H-measure	2 (570.3 mins)	**0.8698**	0.0045
	3 (834.2 mins)	0.8691	0.0051
Bank Telemarketing	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Cross validation	1	**0.3549**	0.0026
H-measure	2 (26.9 mins)	0.3389	0.0024
	3 (28.8 mins)	0.3438	0.0026
Lending Club	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Cross validation	1	**0.1115**	0.0058
H-measure	2 (2606.6 mins)	0.1112	0.0061
	3 (1116.7 mins)	0.1115	0.0058
Loan Full Recovery Data	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Cross validation	1	0.3830	0.0046
H-measure	2 (12.5 mins)	**0.4595**	0.0043
	3 (23.8 mins)	0.4034	0.0046

Table 4.8: Mean and standard deviation of the **H-measure** from the **test set** experiment. The computation times are presented in brackets. Note that K refers to the number of pseudo-classes in the minority class and $K = 1$ refers to standard logistic regression.

Taiwan Credit Card	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Test set	1	0.2434	0.0157
H-measure	**2 (27.4 mins)**	**0.2546**	0.0146
	3 (36.8 mins)	0.2434	0.0157
Credit Fraud	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Test set	1	0.8660	0.0198
H-measure	**2 (54.0 mins)**	**0.8691**	0.0160
	3 (88.4 mins)	0.8674	0.0230
Bank Telemarketing	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Test set	**1**	**0.3547**	0.0161
H-measure	2 (3.6 mins)	0.3406	0.0165
	3 (3.8 mins)	0.3507	0.0159
Lending Club	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Test set	**1**	**0.1115**	0.0059
H-measure	2 (272.5 mins)	0.1112	0.0061
	3 (139.9 mins)	0.1115	0.0059
Loan Full Recovery Data	*K (computation time)*	*Mean H-measure*	*Standard deviation*
Test set	1	0.3779	0.0390
H-measure	**2 (1.5 mins)**	**0.4603**	0.0303
	3 (2.1 mins)	0.4064	0.0351

Table 4.9: Fraud detection rate with fixed fraud alarm rate at 0.5% in Credit Fraud data

Number of pseudo-classes K	Cross validation detection rate	Test set detection rate
K=1	91.12%	88.20%
K=2	**91.20%**	**95.64%**
K=3	91.17%	95.58%



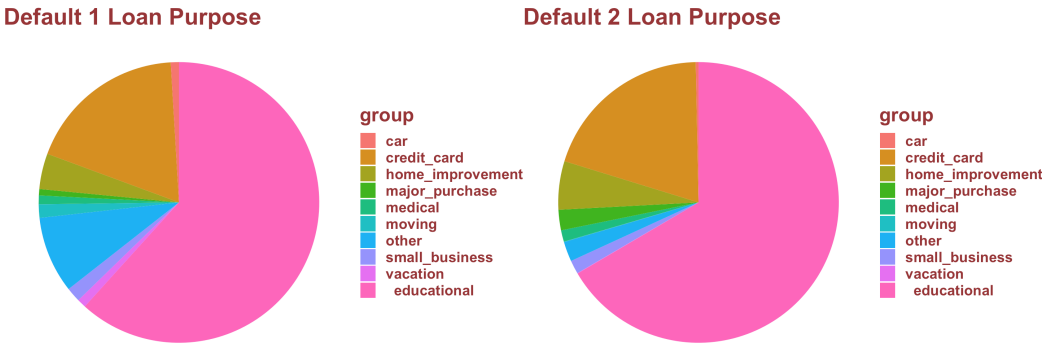


Figure 4.18: The boxplot of log(Loan Amount), Interest Rate (%), log(Annual Income), Credit Grade (1 = best, 7 = worst) and the pie chart of the proportion of different loan purpose for relabeled Lending Club data.

Additional Comparison to SMOTE Oversampling

As introduced in Section 2.3.1, the SMOTE method [Chawla et al., 2002] adds artificially generated data that has the same distribution character as the minority class, which is a widely used data level method to handle class imbalance problem. Here, we use SMOTE to oversample the minority class data among the training set, then build a logistic regression model and apply on the test set. The minority class is over-sampled at 100%, 200%, 300%, 400% and 500% of its original size for trails, which are the default setting by its creators [Chawla et al., 2002]. One of the drawbacks of SMOTE is that the user needs to predetermine k, the number of nearest neighbors that are used to generate the synthetic examples of the minority class. Here, we try $k = 5$ (default setting in [Chawla et al., 2002]) and a large $k = 50$. For comparison, the best SMOTE performance of each k along with relabeling procedure results are presented in Table 4.10.

We find that, when the relabeling procedure does relabel the minority class into new pseudo-classes, our relabeling method will outperform the SMOTE method in both the AUC value and the H-measure value (see Taiwan credit card and loan full recovery data). When the relabeling procedure choose not to relabel the minority class (i.e. a vanilla logistic regression is used), two procedure have similar performance. Actually, our results agree with a recent research from Maldonado et al. [2019], which shows that the performance of SMOTE may be restricted when incorporating with a linear model, like logistic regression, in high dimensional data.

Table 4.10: The AUC and the H-measure with their corresponding standard deviation by using relabeling method and SMOTE. Standard deviation are obtained by 10 fold cross validation. For relabeling procedure, capital K represents the number of the pseudo-classes selected by cross validation procedure; for SMOTE, small k represents the number of nearest neighbors that are used to generate the synthetic examples of the minority class.

Taiwan Credit Card	Relabeling $K = 2$	SMOTE $k = 5$	SMOTE $k = 50$
AUC	0.7562 (0.0062)	0.6993 (0.0073)	0.7205 (0.0095)
H-measure	0.2546 (0.0146)	0.1975 (0.0151)	0.2470 (0.0149)
Credit Fraud	Relabeling $K = 2$	SMOTE $k = 5$	SMOTE $k = 50$
AUC	0.9753 (0.0148)	0.9747 (0.0140)	0.9750 (0.0147)
H-measure	0.8691 (0.0160)	0.8618 (0.0148)	0.8670 (0.0172)
Bank Telemarketing	Relabeling $K = 1$	SMOTE $k = 5$	SMOTE $k = 50$
AUC	0.7913 (0.0163)	0.7913 (0.0203)	0.7895 (0.0191)
H-measure	0.3547 (0.0161)	0.3544 (0.0154)	0.3476 (0.0155)
Lending Club	Relabeling $K = 1$	SMOTE $k = 5$	SMOTE $k = 50$
AUC	0.6882 (0.0046)	0.6841 (0.0059)	0.6829 (0.0050)
H-measure	0.1115 (0.0059)	0.1111 (0.0068)	0.1097 (0.0059)
Loan Full Recovery Data	Relabeling $K = 2$	SMOTE $k = 5$	SMOTE $k = 50$
AUC	0.8576 (0.0049)	0.8248 (0.0066)	0.8262 (0.0662)
H-measure	0.4603 (0.0303)	0.4227 (0.0313)	0.4313 (0.0345)

4.3.5 Experiment 2: Mortgage Default Forecasting

In this section, we conduct a mortgage default forecasting experiment on the US mortgage data set (see Section 4.2.2 for detailed description to the data). The **same data preparation process**, as mentioned in Section 4.2.2, will be used again here.

Experiment Procedure

We use data from a single year for training two different models "with relabeling" and "without relabeling". Here, "without relabeling" refers to logistic regression. The "with relabeling" procedure splits the minority class into K pseudo-classes; K is obtained by ten fold cross validation on the training set (see Section 4.3.3). These K pseudo-classes together with the majority class constitute a relabeled training set. A multinomial logistic regression is trained based on the relabeled data (non-default as base class) and the model output is the probability of default by summing the posterior probability of each pseudo-class. In the cross validation process we substantially explored $K = 2$ or $K = 3$. The "with relabeling" and "without relabeling" models are deployed on four quarters in the following third year (see Table 4.2). The two year gap between training and testing is used to measure and collect default status.

Results

Figures 4.19 and 4.20 give the test set AUC and H-measure with their standard error bar for each method over the observation period. Here, standard error are estimated via bootstrapping each test set 1000 times. Points where the curves for the two methods, logistic regression and relabeling, coincide (year 2003, 2004, 2008, 2010, 2011, 2013) are due to the cross validation procedure selecting $K = 1$ on the corresponding training set. Note that this data is subject to appreciable concept or population drift [Krempl and Hofer, 2011], not least due to the 2008 financial crash. This makes detailed performance analysis more challenging.

For the AUC (Figure 4.19), over this decade, we observe that the relabeling procedure outperform logistic regression in 2005, 2006, 2007, 2009 and 2012, and the error bars in these years are not overlapping (i.e. the difference between the AUC for different methods is at least larger than one times their corresponding standard deviation). Even when the relabeling procedure does not suggest performance improvement, the cross validation on the training set will choose logistic regression ($K = 1$), hence preventing a drop in forecasting performance. The similar results can be observed in the H-measure plot (Figure 4.20); a better performance in 2005, 2006, 2007, 2009 and 2012. Table A.1 in Appendix A gives the mean AUC on the training set by cross validation procedure for choosing K. The relabeling approach also provides insights regarding different pseudo-classes of default; refer to Appendix B for a brief discussion.

Additional Comparison to SMOTE Oversampling

We conduct a comparison between the performance of the SMOTE method and our relabeling method on the Freddie Mac data. This comparison is similar to the comparison we conduct in the previous section. Here, in each training set, the minority class is over-sampled at 100%, 200%, 300%, 400% and 500% of its original size for trails, which are the default setting by its creators [Chawla et al., 2002]. For the number of nearest neighbors that are used to generate the synthetic examples of the minority class, we try $k = 5$ (default setting in [Chawla et al., 2002]) and a large $k = 50$. For comparison, the best SMOTE performance of each k along with relabeling procedure results are presented in Tables 4.11 and 4.12.

We find that, over this decade, *only in 2005 Q1*, SMOTE has significantly better performance than the relabeling method regarding both the AUC and the H-measure. In the other quarters, the relabeling procedure either performs better than SMOTE or has a similar performance to SMOTE ($k = 5$ or $k = 10$), regarding both the AUC and the H-measure. When relabeling outperforms SMOTE with more than one times the standard deviation, the corresponding quarters are labeled blue in Tables 4.11 and 4.12.

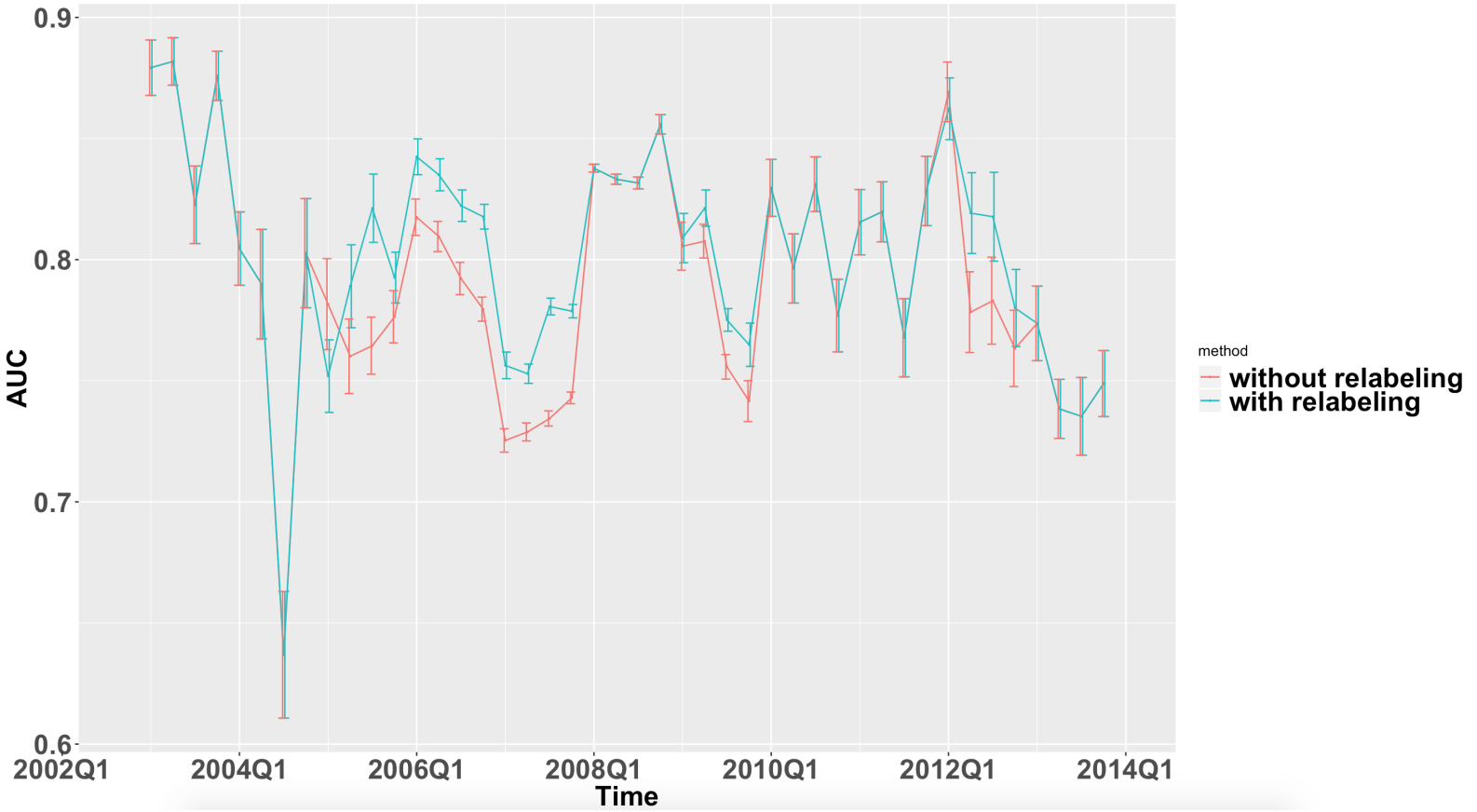


Figure 4.19: AUC and its corresponding error bar of different methods from 2003 to 2013.

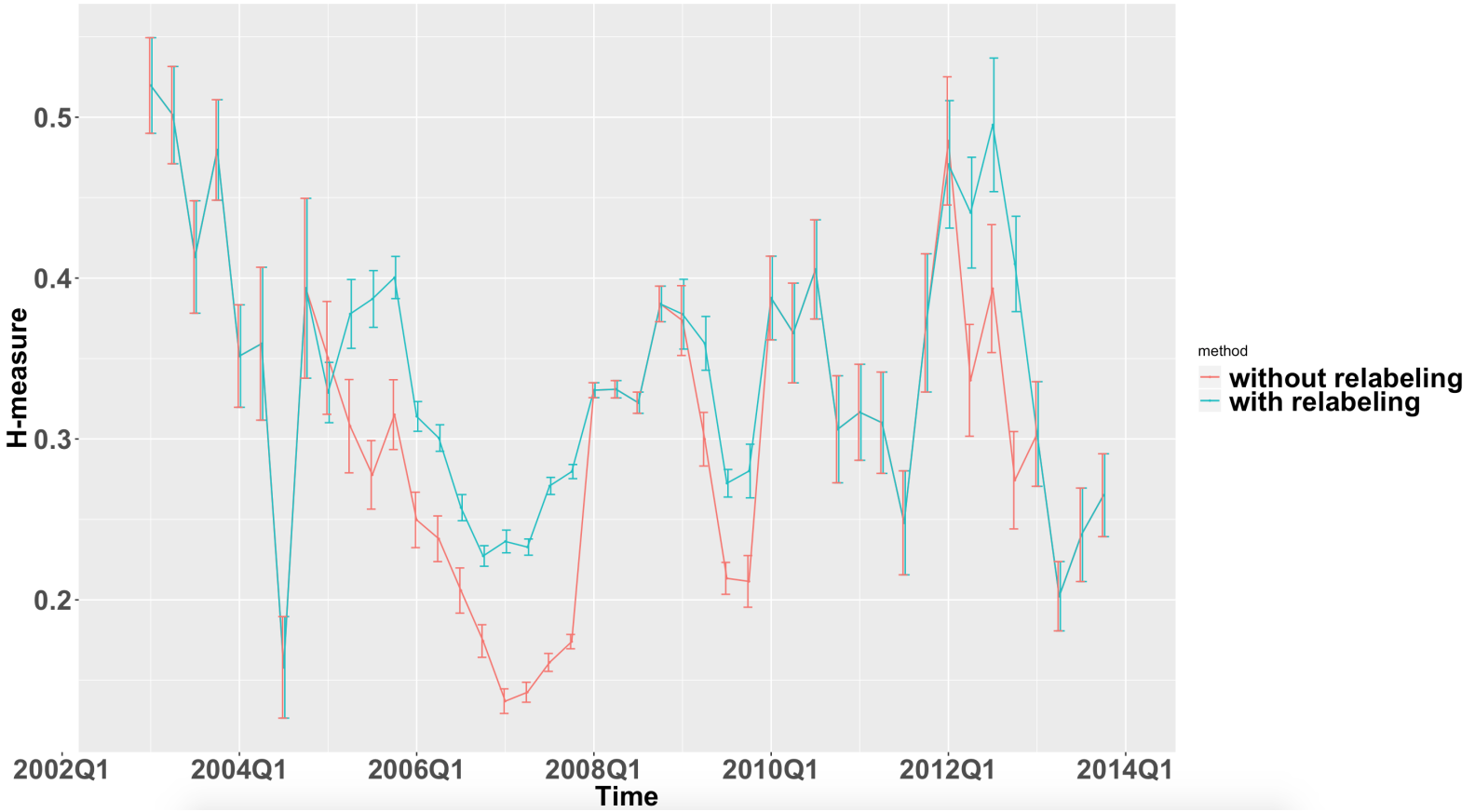


Figure 4.20: H-measure and its corresponding error bar of different methods from 2003 to 2013.

Table 4.11: The AUC with its corresponding standard deviation by using relabeling method and SMOTE from 2003 to 2013. Standard deviation are obtained by bootstrapping the test set 1000 times. When relabeling outperform SMOTE with more than one times standard deviation, it is blue labeled.

Test set	Relabelling		SMOTE $k = 5$		SMOTE $k = 50$	
time	AUC	SD	AUC	SD	AUC	SD
2003 Q1	0.8792	0.0114	0.8305	0.0171	0.8667	0.0140
2003 Q2	0.8818	0.0098	0.8305	0.0123	0.8718	0.0114
2003 Q3	0.8226	0.0160	0.7815	0.0159	0.7966	0.0182
2003 Q4	0.8759	0.0101	0.8296	0.0142	0.8578	0.0106
2004 Q1	0.8046	0.0151	0.7710	0.0156	0.7646	0.0165
2004 Q2	0.7899	0.0226	0.7838	0.0134	0.7711	0.0111
2004 Q3	0.6369	0.0261	0.6361	0.0155	0.6350	0.0156
2004 Q4	0.8027	0.0225	0.8077	0.0144	0.7461	0.0126
2005 Q1	0.7519	0.0150	0.7805	0.0115	0.7842	0.0135
2005 Q2	0.7890	0.0171	0.7899	0.0125	0.7600	0.0118
2005 Q3	0.8212	0.0141	0.7604	0.0108	0.7301	0.0085
2005 Q4	0.7926	0.0105	0.7735	0.0094	0.6736	0.0079
2006 Q1	0.8425	0.0074	0.7899	0.0079	0.8003	0.0095
2006 Q2	0.8350	0.0066	0.7654	0.0076	0.7885	0.0074
2006 Q3	0.8223	0.0065	0.7572	0.0061	0.7778	0.0078
2006 Q4	0.8177	0.0051	0.7445	0.0052	0.7664	0.0049
2007 Q1	0.7564	0.0055	0.6409	0.0056	0.6971	0.0044
2007 Q2	0.7529	0.0040	0.6481	0.0042	0.7135	0.0042
2007 Q3	0.7806	0.0035	0.6536	0.0039	0.7153	0.0035
2007 Q4	0.7787	0.0028	0.6655	0.0025	0.7329	0.0025
2008 Q1	0.8378	0.0016	0.8360	0.0016	0.8364	0.0018
2008 Q2	0.8332	0.0021	0.8321	0.0022	0.8328	0.0018
2008 Q3	0.8316	0.0024	0.8291	0.0027	0.8320	0.0027
2008 Q4	0.8559	0.0040	0.8572	0.0033	0.8571	0.0031
2009 Q1	0.8089	0.0102	0.8060	0.0114	0.8075	0.0089
2009 Q2	0.8213	0.0075	0.7987	0.0079	0.8013	0.0082
2009 Q3	0.7751	0.0047	0.7430	0.0056	0.7338	0.0062
2009 Q4	0.7649	0.0089	0.7349	0.0095	0.7310	0.0104
2010 Q1	0.8296	0.0118	0.8097	0.0126	0.7239	0.0139
2010 Q2	0.7964	0.0143	0.7979	0.0131	0.7952	0.0122
2010 Q3	0.8311	0.0113	0.8253	0.0149	0.8199	0.0119
2010 Q4	0.7769	0.0150	0.7778	0.0160	0.7814	0.0143
2011 Q1	0.8155	0.0135	0.7735	0.0166	0.7654	0.0152
2011 Q2	0.8197	0.0124	0.8089	0.0192	0.7582	0.0168
2011 Q3	0.7677	0.0161	0.7684	0.0202	0.7242	0.0204
2011 Q4	0.8284	0.0143	0.8271	0.0204	0.7825	0.0209
2012 Q1	0.8623	0.0127	0.8691	0.0137	0.8684	0.0132
2012 Q2	0.8192	0.0167	0.7836	0.0148	0.7895	0.0180
2012 Q3	0.8177	0.0183	0.7905	0.0179	0.7852	0.0163
2012 Q4	0.7800	0.0159	0.7778	0.0153	0.7735	0.0139
2013 Q1	0.7737	0.0154	0.7710	0.0156	0.7646	0.0165
2013 Q2	0.7384	0.0122	0.7138	0.0134	0.7211	0.0111
2013 Q3	0.7353	0.0160	0.7291	0.0155	0.7321	0.0156
2013 Q4	0.7488	0.0136	0.7377	0.0144	0.7461	0.0126

Table 4.12: The H-measure with its corresponding standard deviation by using relabeling method and SMOTE from 2003 to 2013. Standard deviation are obtained by bootstrapping the test set 1000 times. When relabeling outperform SMOTE with more than one times standard deviation, it is blue labeled.

Test set	Relabelling		SMOTE $k = 5$		SMOTE $k = 50$	
time	H-measure	SD	H-measure	SD	H-measure	SD
2003 Q1	0.5197	0.0297	0.4637	0.0348	0.4907	0.0291
2003 Q2	0.5013	0.0302	0.4735	0.0260	0.4604	0.0241
2003 Q3	0.4131	0.0349	0.3853	0.0266	0.3811	0.0322
2003 Q4	0.4796	0.0312	0.4112	0.0283	0.4577	0.0221
2004 Q1	0.3515	0.0318	0.2810	0.0232	0.2808	0.0252
2004 Q2	0.3592	0.0475	0.3550	0.0162	0.3520	0.0134
2004 Q3	0.1581	0.0315	0.1586	0.0246	0.1515	0.0233
2004 Q4	0.3937	0.0559	0.3918	0.0203	0.2655	0.0198
2005 Q1	0.3289	0.0187	0.3503	0.0227	0.3583	0.0153
2005 Q2	0.3777	0.0214	0.3138	0.0221	0.1360	0.0120
2005 Q3	0.3870	0.0176	0.2881	0.0210	0.2038	0.0129
2005 Q4	0.4004	0.0131	0.2821	0.0162	0.1246	0.0088
2006 Q1	0.3140	0.0092	0.2865	0.0120	0.2857	0.0165
2006 Q2	0.3006	0.0083	0.2518	0.0118	0.2679	0.0136
2006 Q3	0.2573	0.0081	0.2231	0.0095	0.2201	0.0129
2006 Q4	0.2273	0.0064	0.1925	0.0075	0.1918	0.0080
2007 Q1	0.2363	0.0071	0.0650	0.0043	0.1173	0.0051
2007 Q2	0.2328	0.0050	0.0675	0.0035	0.1315	0.0051
2007 Q3	0.2708	0.0053	0.0745	0.0034	0.1382	0.0042
2007 Q4	0.2797	0.0044	0.0860	0.0025	0.1598	0.0034
2008 Q1	0.3303	0.0046	0.3243	0.0035	0.3285	0.0040
2008 Q2	0.3309	0.0054	0.3247	0.0049	0.3322	0.0040
2008 Q3	0.3225	0.0066	0.3183	0.0058	0.3263	0.0059
2008 Q4	0.3839	0.0111	0.3874	0.0078	0.3992	0.0074
2009 Q1	0.3776	0.0217	0.3757	0.0191	0.3717	0.0159
2009 Q2	0.3595	0.0167	0.2812	0.0145	0.2901	0.0141
2009 Q3	0.2725	0.0086	0.2070	0.0077	0.2071	0.0079
2009 Q4	0.2801	0.0167	0.2072	0.0146	0.2053	0.0141
2010 Q1	0.3876	0.0260	0.3812	0.0212	0.3468	0.0221
2010 Q2	0.3659	0.0310	0.3545	0.0242	0.3386	0.0240
2010 Q3	0.4054	0.0308	0.4017	0.0271	0.3937	0.0226
2010 Q4	0.3160	0.0333	0.3133	0.0299	0.3152	0.0269
2011 Q1	0.3166	0.0298	0.3042	0.0256	0.2918	0.0235
2011 Q2	0.3101	0.0315	0.2667	0.0318	0.2913	0.0300
2011 Q3	0.2479	0.0323	0.2483	0.0314	0.2318	0.0252
2011 Q4	0.3722	0.0430	0.3751	0.0350	0.3784	0.0374
2012 Q1	0.4707	0.0396	0.5072	0.0356	0.5073	0.0307
2012 Q2	0.4407	0.0344	0.3709	0.0275	0.3762	0.0297
2012 Q3	0.4953	0.0415	0.3931	0.0360	0.3770	0.0316
2012 Q4	0.4088	0.0296	0.3271	0.0261	0.3365	0.0259
2013 Q1	0.3031	0.0325	0.2910	0.0232	0.2808	0.0252
2013 Q2	0.2023	0.0215	0.2050	0.0162	0.2020	0.0134
2013 Q3	0.2404	0.0291	0.2406	0.0246	0.2405	0.0233
2013 Q4	0.2650	0.0258	0.2618	0.0203	0.2605	0.0198

4.4 Summary

In this chapter, two relabeling procedure are proposed as mitigation methods for highly imbalanced logistic regression. Both of them use the likelihood function of logistic regression as the objective function to optimize, which is distinct from traditional unsupervised clustering methods (e.g. K-means and Hierarchical clustering).

Especially, we use the EM algorithm as the tool to obtain the underlying pseudo-class structure among the minority class in the presence of highly imbalanced data. Experiments show our EM algorithm can divide the minority class data into several distinct pseudo-classes on highly imbalanced data, and modeling on such relabeled data can enhance logistic regression performance when cluster structure is present among the minority class. Our cross validation procedure for selecting K, ensures that performance is not worse than using basic logistic regression in most cases.

As a generalized linear model, logistic regression is still widely used in many application and the theoretical limit behavior of logistic regression for highly imbalanced data is clear. In this thesis, our focus is a comprehensive study of highly imbalanced logistic regression. For a more balanced classification task, without the restriction to the choice of modeling methods, many more advanced nonlinear machine learning methods are available in the literature. Indeed, if we put aside computational details for a moment, the concept of the relabeling idea may also have value for balanced classification problems, and this is an area for further research.

Our relabeling approach also inspires us to think "when to use it?"; it is equivalent to ask whether our logistic regression has moved into the imbalanced regime? In the next chapter, we propose some diagnostic tools to this problem.

5

Diagnostic tools for Highly Imbalanced Logistic Regression

Being aware of an asymptotic result is one thing, but knowing that a given data set is moving into the asymptotic imbalance regime (with potential consequences) is another. As introduced in Section 2.3, technically, any data set has an imbalanced class proportion that can be recognized as an imbalanced data set [He and Garcia, 2009]. Researchers generally consider a data set as highly imbalanced if the class proportion displays a notable bias (e.g. 1:100, 1:1000) [Krawczyk, 2016]. We proposed a relabeling approach to mitigate the consequences of the highly imbalanced logistic regression in the previous chapter. With further consideration, from the user's point of view, any mitigation method to high class imbalance problem should be based on critically diagnostic to the high class imbalance problem itself (i.e. asking "have we moved into the imbalance regime?"). This kind of diagnostic is rarely considered in practice, due to the vague definition of high class imbalance and the absence of comprehensive diagnostic tools. In this chapter,

we propose some diagnostic tools for assessing whether a data set is in the high imbalance regime regarding logistic regression build upon the known asymptotic results of highly imbalanced logistic regression in Owen [2007]. On the one hand, the deeper mathematical insights from Owen [2007] present the characteristics of logistic regression in highly imbalanced data, which provides the opportunity to build diagnostic tools; on the other hand, trying to detect the evidence of an asymptotic behavior is hard. Thus, we propose different tools to investigate this problem from different aspects.

From Section 3.1.2, we know that the coefficient estimates $\hat{\beta}$ of infinitely imbalanced logistic regression converges to a limit, i.e. the coefficient estimates obtained by replacing the minority class with its mean. Generally speaking, comparing how similar or different the model is against one where the limit condition is artificially induced will indicate how near or far the model is from the imbalance regime. This idea leads to several potential tools:

- hypothesis tests between this limit and the real coefficient estimates,
- measuring the distance between this limit and the real coefficient estimates.

We now start with hypothesis testing tools.

5.1 Hypothesis Testing

In this section, we provide three different tools, focusing on the coefficient estimates, the likelihood, and the predictive probability of a logistic regression. We investigate the high class imbalance from different aspects and can be applied to different data sets. A simulation study will be used to investigate their behavior with different sample sizes.

5.1.1 Hotelling's T^2 Test

The two sample Hotelling's T^2 test [Hotelling, 1992] is a multivariate test (an equivalent to the Student's T test in the univariate case), which is used

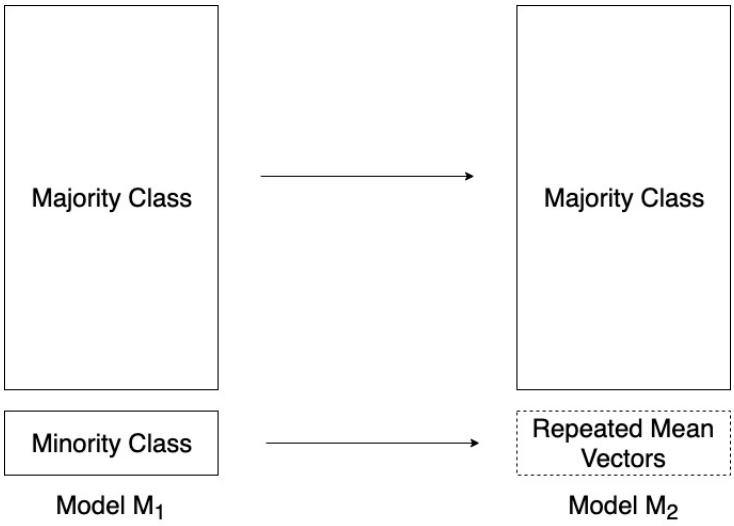


Figure 5.1: The notation in the hypothesis testing for highly imbalanced logistic regression, logistic regression M_1 and M_2 are modeled on the original data set and manipulated data set respectively, the coefficient estimates of Model M_1 and Model M_2 are denoted by $\hat{\beta}_{M_1}$ and $\hat{\beta}_{M_2}$.

to test whether the population means of two p-dimensional random vectors are equal or not. If we assume population one and population two are independently sampled from two multivariate normal distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ ($\boldsymbol{\Sigma}_1$ is not necessarily equal to $\boldsymbol{\Sigma}_2$), the null hypothesis of Hotelling's T^2 test [Hotelling, 1992] is $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, and the test statistic is

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \quad (5.1)$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the p-dimensional sample mean vectors of population one and two respectively, S_1 and S_2 are the sample variance-covariance matrices of population one and two respectively, n_1 and n_2 are the sample sizes of population one and two respectively. Under the null hypothesis, T^2 follows a F distribution with $(p, n_1 + n_2 - p - 1)$ degrees of freedom. We will reject H_0 at level α when $T^2 > F_{p,n_1+n_2-p-1,\alpha}$.

We know that replacing the minority class by its mean vector leads to the same coefficient estimates of the slope vector in the limit $N \to \infty$. Here we use $\hat{\beta}_{M_2}$ to denote the coefficient estimates after replacing the minority

class with its mean vector, and use $\hat{\beta}_{M_1}$ to denote the coefficient estimates estimated from the original data (see Figure 5.1). The replication of the minority class mean vector is to handle the computation issues of the intercept term. Equation (2.14) in Section 2.2.2 shows the coefficient estimates of logistic regression follows a multivariate normal distribution, thus we can use Hotelling T^2 test to test whether $\beta_{M_1} = \beta_{M_2}$. The diagnostic procedure as follow:

Diagnostic tool: Hotelling's T^2 test

1. Build a logistic regression (denoted by M_1) on the original data set, get the coefficient estimates $\hat{\beta}_{M_1}$ and corresponding covariance matrix $\widehat{\mathbf{\Sigma}}_{M_1}$

2. Replace the minority class data with its mean vector and repeat the mean vector n times; build a logistic regression (denoted by M_2) on the manipulated data set. Then, get the coefficient estimates $\hat{\beta}_{M_2}$ and corresponding covariance matrix $\widehat{\mathbf{\Sigma}}_{M_2}$. Here, n is the number of the minority class observations, and the role of the replicates is to handle the computational issues.

3. Calculate T^2 by plug $\bar{\mathbf{x}}_k = \hat{\beta}_{M_k}$, $S_k = \widehat{\mathbf{\Sigma}}_{M_k}$, $n_k = (n + N)$ into Equation (5.1), where $k \in \{1, 2\}$. Please note that we can omit the intercept term and only compare the slope vectors.

4. We will reject the null hypotheses $H_0 : \beta_{M_1} = \beta_{M_2}$ at level α (typically $\alpha = 0.05$) when $T^2 > F_{p,n_1+n_2-p-1,\alpha}$, where p is the dimension of the data.

Here, "reject the null hypotheses" means we have evidence to reject that logistic regression M_1 is in the highly imbalanced regime. Some issues remain for the Hotelling's T^2 test; here we use the assumption of the normality of β_{M_1} and β_{M_2}, however, the variable in the real data may not follow the normal distribution (for example, the categorical variable in the credit data).

5.1.2 Vuong's Non-nested Likelihood Ratio Test

Hotelling's T^2 test can be used to directly test whether $\beta_{M_1} = \beta_{M_2}$ or not, i.e. testing whether the coefficient estimates are "close enough" to the theoretical limits. In this section, we alternatively test the likelihood ratio between logistic regression models M_1 and M_2.

The likelihood ratio test compares the goodness-of-fit between the two models. The widely known likelihood ratio test introduced by Wilks [1938] requires one model is nested in another competing model on the same data set, i.e. the parameter space of one model should be a subspace of the competing model. However, in our case, due to the data manipulation process, we can not use this test, because $\hat{\beta}_{M_1}, \hat{\beta}_{M_2}$ are estimated on different data sets and they have the same parameter space. Vuong [1989] provides a likelihood ratio test by using the Kullback-Leibler information criterion to test the hypothesis that two models are equally close to the real data generating process, which does not require model nesting condition. Here, we give an introduction to Vuong's likelihood ratio test.

Consider the equivalence between model M_1 and model M_2 (Figure 5.1) with respect to their likelihood on the p dimensional original data set with n observations in the minority class and N observations in the majority class. The likelihood function for model M_1 and M_2 in the original data set is

$$l(M_1) = \prod_{i=1}^{n+N} f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i) \text{ and } l(M_2) = \prod_{i=1}^{n+N} f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i),$$

where $f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)$ and $f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i)$ are the case wise contributions to the likelihood function of the logistic regression, with parameters $\hat{\beta}_{M_1}$ and $\hat{\beta}_{M_2}$ respectively:

$$f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i) = y_i \frac{e^{\hat{\beta}_{M_1}^T \mathbf{x}_i}}{1 + e^{\hat{\beta}_{M_1}^T \mathbf{x}_i}} + (1 - y_i) \frac{1}{1 + e^{\hat{\beta}_{M_1}^T \mathbf{x}_i}},$$

$$f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i) = y_i \frac{e^{\hat{\beta}_{M_2}^T \mathbf{x}_i}}{1 + e^{\hat{\beta}_{M_2}^T \mathbf{x}_i}} + (1 - y_i) \frac{1}{1 + e^{\hat{\beta}_{M_2}^T \mathbf{x}_i}}.$$

Then the test statistics of Vuong's likelihood ratio test is the variance of the case wise log-likelihood ratio,

$$\omega_\star^2 = \text{Var}\left(\log\left(\frac{f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i)}\right)\right), i \in \{1, \cdots, n+N\}. \tag{5.2}$$

A natural estimate of $\omega_\star^2$ is the sample analog:

$$\hat{\omega}_\star^2 = \frac{1}{n+N}\sum_{i=1}^{n+N}\left[\log\frac{f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i)}\right]^2 - \left[\frac{1}{n+N}\sum_{i=1}^{n+N}\log\frac{f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i)}\right]^2. \tag{5.3}$$

Vuong [1989] proves that model M_1 and model M_2 are equivalent, if and only if $\omega_\star^2 = 0$. Thus, the hypotheses for Vuong's test is $H_0 : \omega_\star^2 = 0$. Vuong [1989] also proves that under several conditions*, $(n+N)\hat{\omega}_\star^2$ asymptotically follows a weighted sum of χ^2 distributions (denoted by $\sum \lambda_\star^2 \chi^2$, where $\lambda_\star^2$ is a weights vector), when the null hypotheses H_0 is true. The weights of this summation can be calculated via the squared eigenvalues of a particular matrix $\mathbf{W}$.

To obtain $\mathbf{W}$, we need to calculate the following matrices

$$A_{M_1}(\hat{\beta}_{M_1}) = \text{E}\left[\frac{\partial \log f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{\partial \hat{\beta}_{M_1} \hat{\beta}'_{M_1}}\right], \tag{5.4}$$

$$B_{M_1}(\hat{\beta}_{M_1}) = \text{E}\left[\frac{\partial \log f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{\partial \hat{\beta}_{M_1}} \frac{\partial \log f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{\partial \hat{\beta}'_{M_1}}\right]. \tag{5.5}$$

The matrices A_{M_2} and B_{M_2} can be calculated similarly. Further, we also need to calculate

$$B_{M_1 M_2}(\hat{\beta}_{M_1}\hat{\beta}_{M_2}) = \text{E}\left[\frac{\partial \log f(\hat{\beta}_{M_1}, \mathbf{x}_i, y_i)}{\partial \hat{\beta}_{M_1}} \frac{\partial \log f(\hat{\beta}_{M_2}, \mathbf{x}_i, y_i)}{\partial \hat{\beta}'_{M_2}}\right]. \tag{5.6}$$

*observations are i.i.d, first and second derivative of the likelihood function exist and maximum likelihood estimates are unique.

Then, the matrix $\mathbf{W}$ is defined as:

$$\mathbf{W} = \begin{pmatrix} -B_{M_1} A_{M_1}^{-1} & -B_{M_1 M_2} A_{M_2}^{-1} \\ B_{M_1 M_2}^T A_{M_1}^{-1} & B_{M_2} A_{M_2}^{-1} \end{pmatrix}, \tag{5.7}$$

and the squared eigenvalues of $\mathbf{W}$ are $\lambda_\star^2$. For a detailed review to Vuong's likelihood ratio test, please refer to Golden [2000]. Here we use Vuong's likelihood ratio test as a diagnostic tool:

Diagnostic tool: non nested likelihood ratio test

1. Build logistic regression model M_1 on the original data set.
2. Replace the minority class with its mean vector and repeat the mean vector n times, build a logistic regression M_2 on the manipulated data set.
3. Deploy both M_1 and M_2 on the original data set, calculate $\hat{\omega}_\star^2$ and $\lambda_\star^2$.
4. We will reject the null hypotheses $H_0 : \omega_\star^2 = 0$ at level α (typically $\alpha = 0.05$), when $(n+N)\hat{\omega}_\star^2 > \sum \lambda_\star^2 {\chi^2}_\alpha$.

Here, "reject the null hypotheses" means we have the evidence to reject that logistic regression M_1 is in the highly imbalanced regime. It is very important to deploy model M_2 on the original full data set to obtain $\hat{\omega}_\star^2$; in order to ensure the same data set is used for calculating case wise log-likelihood ratio (see Equation 5.2). The implementation of Vuong's likelihood ratio test with `R` is given in Appendix C.

5.1.3 Brier Score z Test

Separate from the coefficient estimates and the likelihood, model predictions provide another view to assess whether two models are different or not. The Brier score (BS) [Brier, 1950] is an accuracy measure of the posterior prob-

ability prediction, which is defined as:

$$BS = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{p}_i)^2, \tag{5.8}$$

where $M = n + N$ is the total number of the observations, $y_i \in \{0, 1\}$ is the true binary label of the data, and $\hat{p}_i$ is the predicted probability of $Y = 1$ given observation $\mathbf{x}_i$ (estimates of $\Pr(Y = 1|X = \mathbf{x}_i)$).

For simplicity, let π_i denote the true but unknown probability of $\Pr(Y = 1|X = \mathbf{x}_i)$. Then, Spiegelhalter [1986] proves, by central limit theorem, if $\pi_i = \hat{p}_i$, then, BS follows a normal distribution with the expectation ($E(BS)$)

$$E(BS) = \frac{1}{M}\sum_{i=1}^{M}(\hat{p}_i(1 - \hat{p}_i)), \tag{5.9}$$

and the variance of the Brier score ($\sigma^2(BS)$) is

$$\sigma^2(BS) = \frac{1}{M^2}\hat{p}_i(1 - 2\hat{p}_i)^2(1 - \hat{p}_i). \tag{5.10}$$

Following Spiegelhalter [1986], Redelmeier et al. [1991] provide a z statistics test to compare two Brier scores of model M_1 and model M_2. Here, we use

$$d = \frac{1}{M}((y_i - \hat{p}_i)^2 - (\pi_i - \hat{p}_i)^2) = \frac{1}{M}\sum_{i=1}^{M}(y_i - \pi_i^2 - 2\hat{p}_i(y_i - \pi_i))$$

to denote the difference between the BS and the correspondent expectation. Assuming models M_1 and M_2 assign estimated probability $\hat{p}_{M_1 i}$ and $\hat{p}_{M_2 i}$ to each observations i respectively, then for model M_1, the estimation of d_{M_1} can be write as:

$$\hat{d}_{M_1} = \frac{1}{M}\sum_{i=1}^{M}(y_i - \pi_i^2 - 2\hat{p}_{M_1 i}(y_i - \pi_i))$$

and similar $\hat{d}_{M_2}$ for model M_2. We want to test whether d_{M_1} is significantly different from d_{M_2}, i.e. the null hypothesis is $H_0 : d_{M_1} = d_{M_2}$.

The difference between $\hat{d}_{M_1}$ and $\hat{d}_{M_2}$ is

$$\begin{aligned}\hat{d}_{M_1-M_2} &= \hat{d}_{M_1} - \hat{d}_{M_2} \\ &= \frac{2}{M}\sum_{i=1}^{M}(\hat{p}_{M_1 i} - \hat{p}_{M_2 i})(\pi_i - y_i),\end{aligned} \tag{5.11}$$

and the variance of $\hat{d}_{M_1-M_2}$ is

$$\sigma^2(\hat{d}_{M_1-M_2}) = \frac{4}{M}\sum_{i=1}^{M}\pi_i(1-\pi_i)(\hat{p}_{M_1 i} - \hat{p}_{M_2 i})^2. \tag{5.12}$$

Hence the test statistics z is

$$z = \frac{\hat{d}_{M_1-M_2}}{\sqrt{\sigma^2(\hat{d}_{M_1-M_2})}}, \tag{5.13}$$

and it follows a standard normal distribution when the null hypothesis is true. Equations (5.11, 5.12) depend on π_i which are some unknown true probability. Redelmeier et al. [1991] propose two options, one is set π_i equal to the class proportion among the observations, another one is $\pi_i = (\hat{p}_{M_1 i} + \hat{p}_{M_2 i})/2$. Here, we choose to use $\pi_i = (\hat{p}_{M_1 i} + \hat{p}_{M_2 i})/2$. Because our target is comparing the difference between two Brier scores, essentially, we seek to compare the difference between $\hat{p}_{M_1 i}$ and $\hat{p}_{M_2 i}$, thus it can be more reasonable to set π_i equal to the average between $\hat{p}_{M_1 i}$ and $\hat{p}_{M_2 i}$ than using the prior class proportion.

Here we use the Brier score z test as a diagnostic tool:

Diagnostic tool: Brier score z test

1. Build a logistic regression (M_1) on the original data set.
2. Replace the minority class with its mean vector and repeat the mean vector n times, build a logistic regression M_2 on the manipulated data set.
3. Deploy both M_1 and M_2 on the original data set, calculate $\hat{p}_{M_1 i}$ and $\hat{p}_{M_2 i}$, $i \in \{1, \cdots, M\}$.
4. Assume $\pi_i = (\hat{p}_{M_1 i} + \hat{p}_{M_2 i})/2$, calculate $d_{M_1 - M_2}$ and the z statistics by Equations (5.11, 5.12).
5. We will not reject the null hypotheses $H_0 : d_{M_1 - M_2} = 0$ at level α (typically $\alpha = 0.05$), when $z < z_\alpha$.

5.1.4 Simulation Results

In this section, we simulate a bivariate normal distribution example to assess the performance of our diagnostic tools. We generate M binary observations where

$$M \in \{500, 1000, 5000, 10000, 50000, 100000\},$$

and vary the imbalance level γ in

$$\gamma \in \{0.1\%, 0.5\%, 1\%, 5\%, 10\%, 20\%, 30\%, 40\%\}.$$

The majority class ($Y = 0$) follows $X \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ which has $(1 - \gamma)M$ observations and the minority class ($Y = 1$) follows $X \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ which has γM observations. Here

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \text{and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.64 & 0 \\ 0 & 0.64 \end{pmatrix}.$$

In each generated simulation data, we deploy our diagnostic tools. Monte Carlo replication are conducted 1000 times.

Tables (5.1, 5.2, 5.3) give the results. We could claim that our logistic regression is in the high imbalance regime when the p-value is greater than 0.05[†]. Highlighted in Tables (5.1, 5.2, 5.3) are p-value higher than 0.05. We find that:

- When the sample size $M \geq 50000$, the Hotelling T^2 test's p-value will always be smaller than 0.05, no matter how the imbalance level varies. Particularly, we find that when $M \in \{5000, 10000\}$ and the imbalance level $\gamma = 1\%$, Hotelling T^2 test does not indicate that our logistic regression has moved into the imbalanced regime; but Vuong's likelihood ratio test and Brier score z test do provide this indication in the same setting.
- Most of the test results from Vuong's likelihood ratio test are consistent with the Brier score z test; the only exception occurs when $M = 500$ and imbalance rate γ at 10%.

In summary, all of these diagnostic tests can effectively detect highly imbalanced logistic regression on a small sample set. However, in large samples, we find that the Hotelling's T^2 test does not provide the indication of highly imbalanced logistic regression, when comparing to the results obtained from the other two tests in the same setting; this is reasonable by considering the sample size terms in Equation (5.1). We recommend to use Vuong's likelihood ratio test and the Brier score z test when the sample size is large (for example when $M > 10000$ in our simulation experiment).

[†]This is because "p-value > 0.05" means that there is no evidence to reject the null hypothesis: "logistic regression is in the high imbalance regime".

Table 5.1: p-value table - Hotellling T^2 test, NA comes from the low proportion of the minority class.

Imbalance γ M	0.1%	0.5%	1%	5%	10%	20%	30%	40%
500	NA	**0.5718**	**0.2159**	0.0440	0.0374	0.0228	0.0200	0.0234
1000	NA	**0.2037**	**0.0857**	0.0259	0.0257	0.0167	0.0144	0.0110
5000	**0.3137**	**0.0576**	0.0373	0.0094	0.0094	0.0030	0.0031	0.0039
10000	**0.1544**	0.0315	0.0182	0.0033	0.0017	0.0003	0.0010	0.0007
50000	0.0282	0.0023	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
100000	0.0119	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 5.2: p-value table - Vuong's likelihood ratio test, NA comes from the low proportion of the minority class.

Imbalance γ M	0.1%	0.5%	1%	5%	10%	20%	30%	40%
500	NA	**0.7106**	**0.6650**	**0.2308**	0.0300	0.0009	0.0017	0.0028
1000	NA	**0.6847**	**0.5684**	**0.0888**	0.0020	0.0000	0.0000	0.0000
5000	**0.6660**	**0.3595**	**0.1825**	0.0002	0.0000	0.0000	0.0000	0.0000
10000	**0.5503**	**0.1900**	**0.0574**	0.0000	0.0000	0.0000	0.0000	0.0000
50000	**0.2242**	0.0031	0.0000	0.0000	0.0000	0.0008	0.0064	0.0093
100000	**0.0974**	0.0000	0.0000	0.0000	0.0006	0.0087	0.0277	0.0364

Table 5.3: p-value table - Brier Score z test, NA comes from the low proportion of the minority class.

Imbalance γ M	0.1%	0.5%	1%	5%	10%	20%	30%	40%
500	NA	**0.6573**	**0.6067**	**0.2285**	**0.0893**	0.0045	0.0001	0.0000
1000	NA	**0.6286**	**0.5181**	**0.1101**	0.0150	0.0000	0.0000	0.0000
5000	**0.6088**	**0.3925**	**0.1724**	0.0003	0.0000	0.0000	0.0000	0.0000
10000	**0.5535**	**0.2369**	**0.0553**	0.0000	0.0000	0.0000	0.0000	0.0000
50000	**0.3012**	0.0081	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100000	**0.1500**	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 5.4: Diagnostic tool table. D is the original data set, D_s is a data set for which all the minority class data are replaced by their mean vector, D_i is a resampled data set with $i \times n$ observations in the majority class.

Data set	D (original data set)	D_s	D_1	D_2	$\cdots$
Number of observations in majority class	N	N	n	$2n$	$\cdots$
Number of observations in minority class	n	1	n	n	$\cdots$
Logistic regression model	M	M_s	M_1	M_2	$\cdots$

5.2 Mahalanobis Distance

Our simulation results show that the extremely large sample size always results in a statistical significance in hypothesis testing. Essentially, the p-value can be as small as you want by increasing the sample size [Demidenko, 2016]. In this section, we propose a visualization method to alleviate the influence of the extremely large sample size. The core idea is to artificially tune the imbalance rate by undersampling the majority class and use the Mahalanobis distance [McLachlan, 1999] to measure the distance between the coefficient estimates of these logistic regression and the limit coefficient estimates of the logistic regression when replace the whole minority class with its mean.

Mahalanobis distance is a measure of the distance between a single point and a distribution. If we use $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to denote the mean vector and the covariance matrix of a population, then the Mahalanobis distance between a point $\mathbf{x}$ and this population is defined as

$$\mathrm{MD}(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}. \tag{5.14}$$

Considering the limit coefficient estimate $\widehat{\beta}$ and its corresponding covariance matrix $\widehat{\boldsymbol{\Sigma}}$ (obtained by replacing the minority class with its mean vector), we can calculate the Mahalanobis distance between this limit and any other coefficient estimates via Equation (5.14).

For a clear illustration, a sequence of data sets is constructed as in Table 5.4.

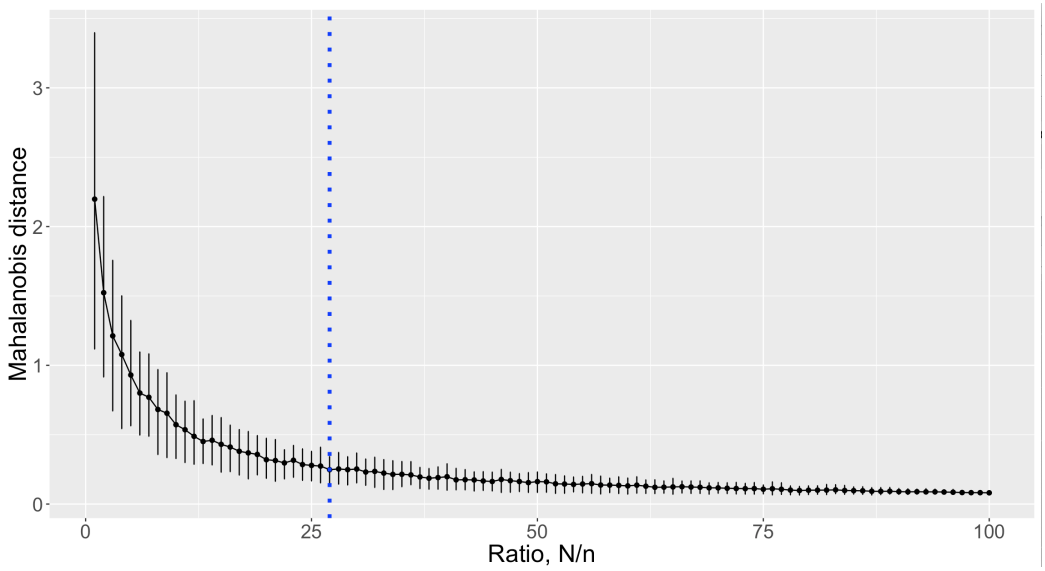


Figure 5.2: The Mahalanobis distance between β_{M_s} and $\beta_{M_k}, k \in \{1, \cdots, 100\}$, the 5% and 95% quantile bar and mean from Monte Carlo replication are presented.

Here D_s is obtained by replacing the minority class with its mean vector, and $\{D_1, D_2, \ldots\}$ are obtained by random undersampling the majority class to $\{1n, 2n, \ldots\}$ observations. For each constructed data set, a logistic regression model is computed, and the parameter estimates retained. Then, the Mahalanobis distance between the parameter vector for M_S and $\{M, M_1, M_2, \cdots\}$ are computed. The corresponding covariance matrices are computed from the logistic regression procedure to support the construction of the distance metric. We expect the Mahalanobis distance will approach 0 as class imbalance increases, consistent with the Owen [2007] theoretical result. This diagnostic method is illustrated in the following simulation, where Monte Carlo replicates over the D_i are considered.

We generate 10000 sample points following $X \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ from the majority class ($Y = 0$). The 100 points following $X \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are generated as the minority class. Here

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \text{and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.16 & 0.08 \\ 0.08 & 0.16 \end{pmatrix}.$$

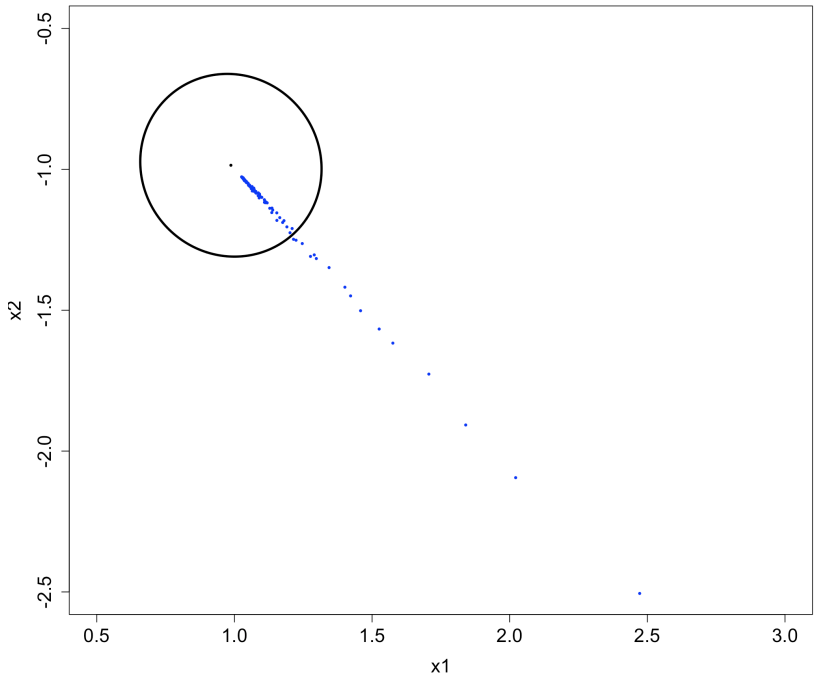


Figure 5.3: The scatter plot of $\beta_{M_k}, k \in \{1, \cdots, 100\}$ (blue) and β_{M_s} (black) and a two-dimensional ellipse that traces the bivariate normal density contour for β_{M_s} at 95%.

The imbalance ratio between the majority class and the minority class is 100. We construct D_s, $D_1, \cdots, D_{100}$ as illustrated in Table 5.4. Figure 5.2 gives the results. The 5% and 95% quantile bar and mean in the plot results from undersampling the majority class data 100 times in each ratio level N/n. The plots show that when $N/n > 10$, the Mahalanobis distance drops towards 0 very quickly.

For this bivariate simulation, we can present the coefficient estimates $\hat{\beta}_{M_k}, k \in \{1, \cdots, 100\}$ and $\hat{\beta}_{M_s}$ in a scatter plot. The covariance matrix of $\hat{\beta}_{M_s}$ can also let us draw a 95% contour, which traces this bivariate normal distribution. Figure 5.3 gives the result, the black contour is the 95% ellipse of the bivariate normal distribution of $\hat{\beta}_{M_s}$. It shows that as the imbalance ratio N/n increasing, the coefficient estimates $\hat{\beta}_{M_k}$ will approach $\hat{\beta}_{M_s}$ very quickly. We note the link between this visualization tool and the Hotelling's T^2 test, since $\text{MD}(x)$ is χ^2 distributed when the population is drawn from a multi-

variate normal distribution and Hotelling's T^2 statistic is asymptotically χ^2 distributed. However, this visualization method can serve as a supplement tool, especially when the sample size $N + n$ is extremely large.

Figure 5.2 also shows that the curve of the Mahalanobis distance exhibits a "L" shape or "Knee" shape, which brings a famous problem in many domains: "searching a knee point (also named as the indication point), based on recent trends" [Salvador and Chan, 2004]. We seek to use some automatic indication point searching algorithm, which could help us find an indication point in the Mahalanobis distance plot. This method can make users aware of the highly imbalanced logistic regression. Satopaa et al. [2011] provide a knee point detection algorithm named "Kneedle"[‡]. We use it as a tool to search a knee point in the Mahalanobis distance plot. This algorithm will select an imbalance ratio $N/n = 27$ as knee the point, which is displayed as the vertical blue line in Figure 5.2.

[‡]Generally, we prefer an algorithm which does not require parameter tuning for different "Knee" shape curves and suitable for both online and offline operating. The algorithm Satopaa et al. [2011] proposed satisfy these requirements.

Diagnostic tool: Mahalanobis distance plot

1. Generate a sequence of data sets $(D_s, D_1, D_2, \cdots)$, which artificially varied the imbalance ratio k.
2. For each constructed data set, build a logistic regression $(M_s, M_1, M_2, \cdots)$.
3. Calculate the Mahalanobis distance between the coefficient estimates of M_s and M_k, $k \in \{1, 2, \cdots\}$.
4. Repeat step 1, 2, and 3 multiple times to have the average of the Mahalanobis distance between coefficient estimates of M_s and M_1, M_s and M_2, $\cdots$.
5. Plot the average Mahalanobis distance with the corresponding imbalance ratio k.
6. We could further consider "Kneedle" algorithm to detect a indication point. If the true imbalance ratio is higher than the indication point, we aware that logistic regression already moved into imbalanced regime.

5.3 Discussion of the Diagnostic Tools and Relabeling

We have proposed several diagnostic tools for detecting highly imbalanced logistic regression. In this section, we deploy our diagnostic tools on two simulation data sets where cluster structure is present the minority class.

Simulation A, two clusters among the minority class: Here, we generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$. Then $n_1 = 25$ points are generated following $X \sim N(\mu_1, \Sigma_1)$ and $n_2 = 25$ points are generated following $X \sim N(\mu_2, \Sigma_2)$. n_1 and n_2 are combined as the minority class $Y = 0$ (i.e. $n = n_1 + n_2 = 50$ minority class observations). Here

$$\mu_0 = [0, 0], \mu_1 = [1.5, 1], \mu_2 = [1.5, -1]$$

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

The number of the majority class observations N varies in $\{50, 100, 200, \ldots, 1000\}$. As N increases, data become more imbalanced; this follows the approach in Owen's Theorem to construct imbalance. For each combination of (n, N), we repeat the simulation 1000 times by deploying the relabeling approach and a vanilla logistic regression on the training set then apply them on the test set. We also apply two diagnostic tools (Vuong's likelihood ratio test and Brier score z test) on the training set. Figure 5.4 gives the average AUC and H-measure with their corresponding error bars (standard errors) on the test set. The vertical blue lines represent the moment when diagnostic tools indicate us that logistic regression has moved into the imbalance regime (i.e. p-value > 0.05) in all 1000 replications.

For the AUC, we can find that the overlap part of the error bars between vanilla logistic regression and the relabeling approach shrinks as N increases; and a relatively noticeable difference of the AUC between two methods can be found when $N > 800$. At the same time, the Brier score z test and Vuong's likelihood ratio test will give indications of "has moved into imbalance regime" at $N = 500$ and $N = 600$ respectively. This means that, in this contrived scenario, we can get a significant performance improvement by using the relabeling approach when these diagnostic tools indicate us. For the H-measure, we see a similar phenomenon to the AUC, which also justifies the effectiveness of diagnostic tools and its collaboration with the relabeling approach.

Simulation B, three clusters among the minority class: Here, the minority class is modified to have a three clusters structure. We generate N points $X \sim N(\mu_0, \Sigma_0)$ as the majority class $Y = 0$ (the same as Simulation 1). Then $n_1 = 16$ points are generated following $X \sim N(\mu_1, \Sigma_1)$, $n_2 = 16$ points are generated following $X \sim N(\mu_2, \Sigma_2)$ and $n_3 = 16$ points are generated following $X \sim N(\mu_3, \Sigma_3)$. n_1, n_2 and n_3 are combined as the minority class $Y = 0$ (i.e. $n = n_1 + n_2 + n_3 = 48$ minority class observations). Here

$$\mu_0 = [0, 0], \mu_1 = [1.5, 1], \mu_2 = [1.5, 0], \mu_3 = [1.5, -1]$$

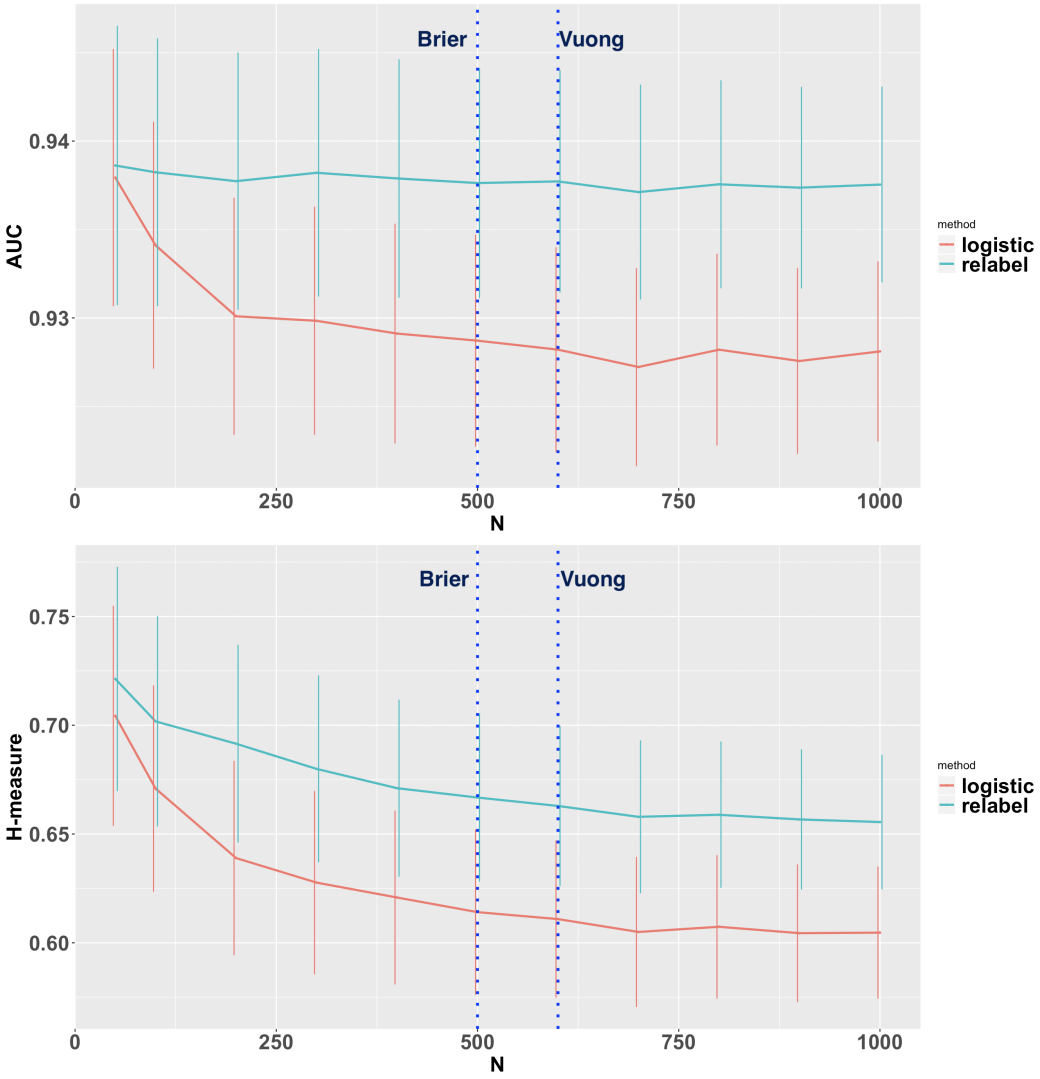


Figure 5.4: Simulation A: The average AUC and H-measure with their corresponding error bars (standard errors) on the test set. The vertical blue lines represent the moment when diagnostic tools indicate us that logistic regression has moved into the imbalance regime (i.e. p-value > 0.05) in all 1000 replications.

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{and } \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.16 \end{bmatrix}.$$

The number of the majority class observations N varies in $\{50, 100, 200, \ldots, 1000\}$. For each combination of (n, N), we repeat the simulation 1000 times by deploying the relabeling approach and a vanilla logistic regression on the training set then apply them on the test set. We also apply two diagnostic tools (Vuong's likelihood ratio test and Brier score z test) on the training set. Figure 5.5 gives the average AUC and H-measure with their corresponding error bars (standard errors) on the test set. The vertical blue lines represent the moment when diagnostic tools indicate us that logistic regression has moved into the imbalance regime (i.e. p-value > 0.05) in all 1000 replications.

In this simulation, for the AUC, a relatively noticeable difference of the AUC between two methods can be found when $N > 500$. At the same time, the Brier score z test and Vuong's likelihood ratio test will give indications of "has moved into imbalance regime" at $N = 400$ and $N = 600$ respectively. This means that, in this contrived scenario, the Brier score z test does a better job than Vuong's likelihood ratio test to give us an indication to use relabeling approach. We also applied Hotelling T^2 test in Simulation A and Simulation B, however, it will not give us any indication of the "imbalanced regime". The results from Simulations A and B show that we should consider the results from all proposed diagnostic tools carefully. It is worth to try relabeling methods mentioned in the previous chapter when any hint of highly imbalanced logistic regression emerges. The diagnostic methods do not provide definitive answers except for somewhat trivial cases, because we are trying to detect an asymptotic phenomenon. While a diagnostic check is worthwhile, as a matter of standard practice, the EM relabeling method is recommended as generally preferable.

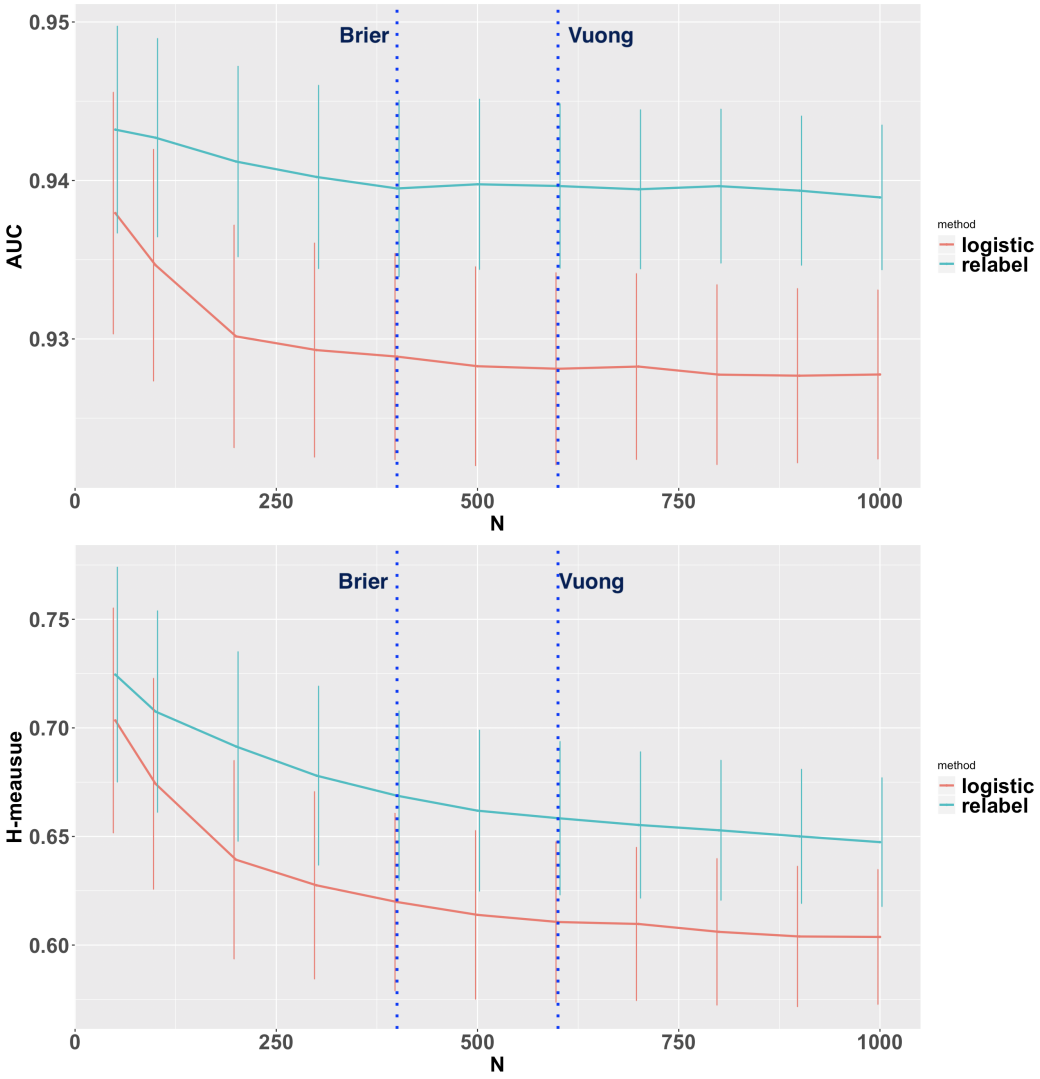


Figure 5.5: Simulation B: The average AUC and H-measure with their corresponding error bars (standard errors) on the test set. The vertical blue lines represent the moment when diagnostic tools indicate us that logistic regression has moved into the imbalance regime (i.e. p-value > 0.05) in all 1000 replications.

5.4 Real Data Application

In this section, we deploy our diagnostic tools on two real data sets. These data are from the credit risk industry, which mentioned in the previous chapter.

5.4.1 Loan Recovery Data

For Loan Recovery Data, the proportion of the minority class data is 2.17% and the total number of the observations is 8237. The Hotelling T^2 test, the Vuong's likelihood ratio test and the Brier score z test give p-value 2.3675×10^{-4}, 0.1479, and 0.8692 respectively, which means Vuong's likelihood ratio test and Brier score z test will suggest logistic regression is in the highly imbalanced regime. The Mahalanobis distance plot is given in Figure 5.6. The "Kneedle" algorithm will select an imbalance ratio $N/n = 9$ as the indication point.

It is clear that Hotelling T^2 test suffers from the large sample size here, and considering the outputs from these diagnostic tools, we suggest that this data set is in the highly imbalanced regime.

5.4.2 Freddie Mac Mortgage Data

The target variable in Freddie Mac mortgage data is whether a mortgage moves into default status in the following two years after the first payment. In the previous chapter, we use the data from a single year (between 2000 and 2010) as a training set. Figure 4.4 shows the number of applications fluctuates over this long time frame. We find a pronounced peak in default rate during the financial crisis period (2007-2008) with a peak of 6.8% in 2007 Q3; however, the default rate is extremely low in other quarters. For our diagnostic tools, Table 5.5 gives the p-value of the Vuong's likelihood ratio test and Brier score z test in each year. Due to the large sample size, the Hotelling's T^2 test will return a p-value uniformly equal to 0 in all years.

We find both these tests will indicate highly imbalanced logistic regression

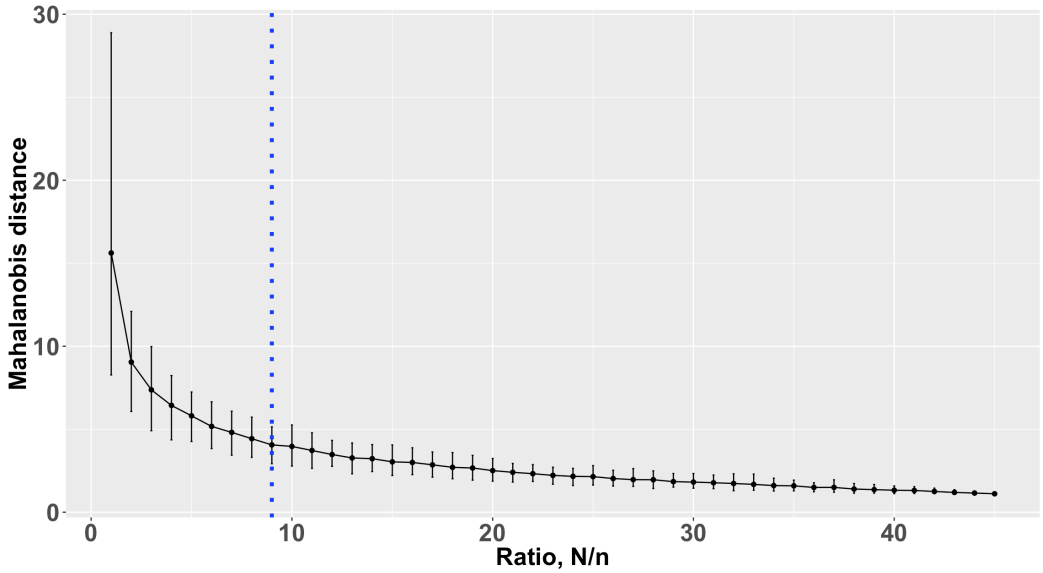


Figure 5.6: Mahalanobis Distance Plot for Recovery Data, the 5% and 95% quantile bar and mean from Monte Carlo replication are presented, and the vertical blue line is the knee point detected by "Kneedle" algorithm.

in the years 2000, 2001, 2002, 2003, 2004, 2005, 2006, and 2010. Figure 5.7 gives the Mahalanobis distance plot; the red dot line is the indication points detected by the "Kneedle" algorithm. The visualization plots tend to show that logistic regression moves into a highly imbalanced regime in all years; actually, all indication points in Figure 5.7 are smaller than 10. The conflict results in the years 2007, 2008, and 2009 are reasonable considering the large sample size of the Freddie Mac data. We can also find the confidence bands in Figure 5.7 usually shrink with the imbalance ratio increase; this is caused by the effect of the sample size.

Table 5.5: p-value Table for the diagnostic to the Freddie Mac data in each year, the fifth and sixth columns are the p-value of the Vuong's likelihood ratio test and the Brier score z test respectively.

Year	Proportion of the Minority Class	Number of the Minority Class	Total Number of the Observations	p-value Vuong	p-value Brier
2000	0.4112%	308	74902	0.3857	0.0747
2001	0.2014%	465	230840	0.2190	0.0606
2002	0.1015%	234	230330	0.0640	0.0638
2003	0.0582%	137	235391	0.3196	0.5715
2004	0.0677%	76	112247	0.6663	0.1158
2005	0.1690%	252	149041	0.2576	0.0712
2006	0.8915%	755	84680	0.4001	0.3905
2007	4.2679%	3690	86458	0.0000	0.0000
2008	3.1580%	4101	129857	0.0000	0.0000
2009	0.3049%	688	225600	0.0000	0.0000
2010	0.0915%	152	166046	0.6380	0.2980

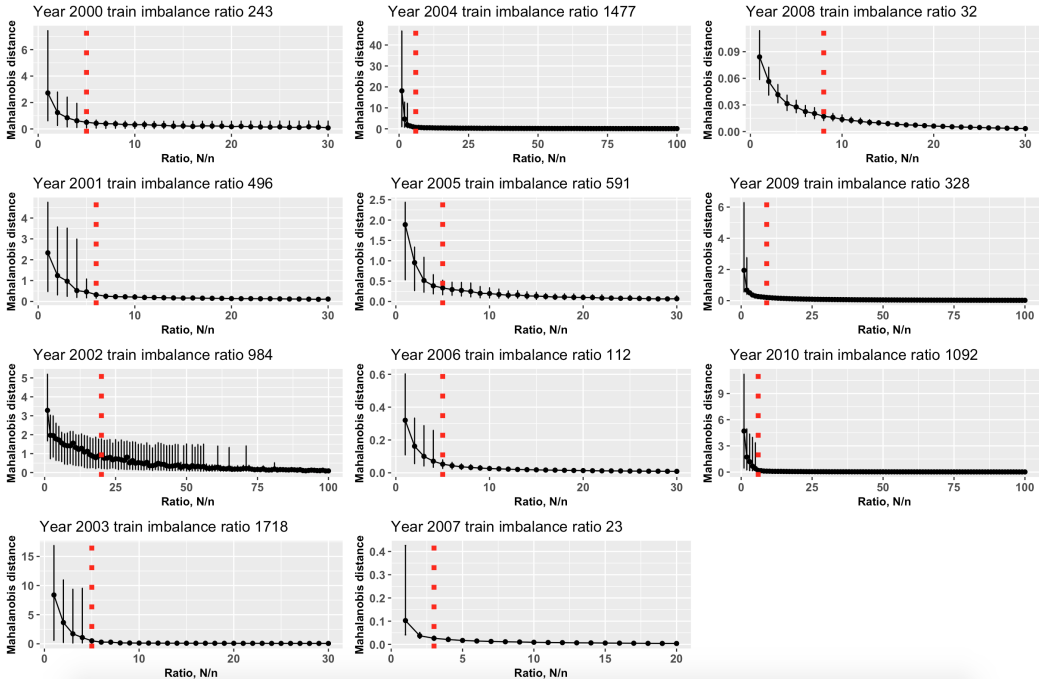


Figure 5.7: Mahalanobis Distance Plot for Freddie Mac Data, the 5% and 95% quantile bar and mean from Monte Carlo replication are presented, and the vertical red line is the knee point detected by "Kneedle" algorithm.

5.5 Summary and Recommendations

As stated at the beginning of this chapter, knowing that a given data set is moving into the asymptotic imbalance regime is vital from the user's point of view. The diagnostic tools proposed in this chapter are focusing on detecting highly imbalanced logistic regression in light of Owen [2007] theoretical results. From different angles of view, we propose to measure the difference of coefficient estimates, likelihood, or predicted probability between the theoretical limit and the real model as diagnostic tools. We also proposed a visualization tool to observe the imbalance.

It is important to note that the calibration between these diagnostic tools is essential since we find the conflict diagnostic results on the real data. The simulation results show that the sample size becomes a key concern when deploying these tools. Hotelling T^2 test tends to be more sensitive than Vuong's likelihood ratio test and Brier score z test when the sample size is large, which resulting in the omitting of the highly imbalanced logistic regression. Thus, we recommend that the Hotelling T^2 test is only suitable for the small sample set. Vuong's likelihood ratio test and Brier score z test are more stable, but still suffer from the extremely large sample size, which suggests that the user should not state any probabilistic conclusion, especially when diagnosing on the big data. The visualization of the Mahalanobis distance for more massive data sets move away from statistical tests, that may alleviate the unwanted effect of the large sample size. We suggest that users consider the results from all proposed diagnostic tools carefully. It is worth to try relabeling methods mentioned in the previous chapter when any hint of highly imbalanced logistic regression emerges.

6

Conclusion

This chapter concludes this thesis by summarizing our research contributions and possible future work directions. The target of this thesis towards a solid theoretical understanding of highly imbalanced logistic regression and its corresponding mitigation and diagnostic methods. With these in mind, we conduct our research focusing on answering the questions we proposed in Chapter 1:

In the **theory part**, we extend Owen [2007] results of the infinitely imbalanced logistic regression to its two natural alternative choices (weighted and penalized logistic regression). We prove that:

- the minority class only contributes to the infinitely imbalanced weighted logistic regression via its weighted mean vector,
- the slope vector of the infinitely imbalanced penalized logistic always converges to zero, which makes the predicted probability uniformly equal to the prior class proportion without considering the input $\mathbf{x}$.

The main message from these theories is that weighting or penalizing the likelihood function is not enough for highly imbalanced logistic regression.

Especially in the presence of cluster structure among the minority class; the mean vector of the minority class can be a poor representation.

For **mitigation methods**, we propose to relabel the minority class into several new pseudo-classes to circumvent the imbalance problem by exploiting the cluster structure. This relabeling idea is achieved by solving two subproblems:

- the Genetic algorithm and the Expectation Maximization algorithm are proposed to establish the mapping between the minority class and the K pseudo-classes.
- the cross validation procedure is proposed to identify the number of pseudo-classes K.

Simulation studies and real data experiments show that modeling multinomial logistic regression on the relabeled data can enhance logistic regression performance, especially when cluster structure is present.

For **diagnostic tools**, we propose to use different hypothesis testing methods for detecting highly imbalanced logistic regression. These methods, essentially build on the more in-depth mathematical insight of the infinitely imbalanced logistic regression, emphasize different ways to detect highly imbalanced logistic regression, i.e. the coefficient estimates, the likelihood, or the predicted probability. Considering the huge data sets we often deal with in real life, a graphical method is designed as a supplementary tool to hypothesis testing methods. From our simulation studies and real data experiments, we recommend the user to try all diagnostic tools before reaching any conclusion regarding class imbalance.

As mentioned in Section 2.3, class imbalance is a challenging and important problem, not only because class imbalance is common in many real-life data but also due to the concept linked to the rare class can being critical in application. The thesis presents a systematic approach for handling the class imbalance problem with logistic regression, which is tested on multiple simulation and real data set and has shown improvement regarding prediction.

Future Work

Some open questions remain to be answered beyond those considered in this thesis. Some possible directions for future work are listed below:

Understanding the Fundamental Theory:

Currently, most research on the class imbalance problem are focused on case studies or specific algorithms, without the theoretical understanding. We see the immediate impact from the consequences of Theorems 2, 7, and 11, i.e. the cluster structure among the minority class is problematic for highly imbalanced logistic regression, and weighted or penalized likelihood function can not alleviate the problem. A deeper understanding of questions like "How do imbalanced data affect widely used classification algorithms (e.g. tree based methods, discriminant analysis)?" can help us to propose learning algorithms for imbalanced data with more targeted approach.

Relabeling with Cluster Structure in Balanced Binary Classification Problem:

The relabeling approaches we proposed in Chapter 4 are designed for imbalanced data. A natural question is whether the cluster structure will affect the *balanced* binary classification problem. An investigation of this problem, when using logistic regression with a binary data set, is a subject for further research. The mixture of experts model [Masoudnia and Ebrahimpour, 2014] hierarchically divides the problem space into several subspaces, and assign different inputs to different learners. We cannot use the mixture of experts model directly, because it does not consider the binary data structure (we are trying to cluster the data within one class). However, it gives a good example to capture cluster structure in data.

Relabeling with Other Classifiers:

Considering the proposed EM algorithm in Chapter 4, we iteratively improve the lower bound of the likelihood function with the latent variable (pseudo-classes). However, the concept of relabeling may have more general applicability. Indeed, if we put aside computational concerns for a moment, any classifiers with a clear likelihood expression can fit into this EM framework. An example is linear or quadratic discriminant analysis, which immediately

leads to an EM algorithm solving Gaussian mixtures model among the minority class. Whether this EM procedure can be an effective framework for relabeling with other classifiers is an interesting problem.

Application:

We have extensively conducted experiments on the credit risk data, due to its importance in the consumer credit risk industry. The results show that the cluster structure among the minority class has a direct impact on the prediction performance of logistic regression. More application experiments on other domains are also very interesting.

A

Tables

Table A.1: Mean AUC and its corresponding standard deviation on training set by **ten fold cross validation**, boldface indicates the choosen K.

Train year	2000	2001	2002
$K = 1$	**0.8569** (0.0101)	**0.8508** (0.0157)	0.8615 (0.0270)
$K = 2$	0.8475 (0.0135)	0.8450 (0.0147)	0.8629 (0.0230)
$K = 3$	0.8213 (0.0138)	0.8396 (0.0150)	**0.8642** (0.0185)
Train year	2003	2004	2005
$K = 1$	0.8509 (0.0185)	0.7819 (0.0183)	**0.8469** (0.0175)
$K = 2$	0.8549 (0.0138)	**0.7853** (0.0201)	0.8408 (0.0159)
$K = 3$	**0.8697** (0.0141)	0.7631 (0.0196)	0.8356 (0.0180)
Train year	2006	2007	2008
$K = 1$	0.8484 (0.0075)	**0.8660** (0.0026)	**0.8708** (0.0031)
$K = 2$	**0.8487** (0.0069)	0.8613 (0.0027)	0.8682 (0.0036)
$K = 3$	0.7599 (0.0065)	0.8598 (0.0036)	0.8664 (0.0038)
Train year	2009	2010	
$K = 1$	0.8882 (0.0120)	**0.8322** (0.0115)	
$K = 2$	**0.9217** (0.0118)	0.8253 (0.0126)	
$K = 3$	0.9204 (0.0113)	0.8297 (0.0177)	

B

Analysis to the Pseudo-classes among the Minority Class

B.1 Recovery Data-Full Recovery

This section gives a brief analysis to the pseudo-classes among the recovery data set (Section 4.3.4). Table B.1 shows the descriptive statistics for the two pseudo-classes for full recovery (i.e. recovery rate = 100%). They show two clearly different groups based on loan amount: C2 are high value loans, whilst C3 are low value loans, in general. Note that interest payment due, credit limit and OB (outstanding amount at default) all reflect loan amount. Amongst other variables we see that the high loan group, C2, includes generally older people and proportionally less female borrowers than C3, which is a reflection of lending demographics for this cohort. The C2 pseudo-class has significantly lower average credit score, which is surprising since lenders would typically want to take less risk with higher value loans. As we might expect, the pre-purchase recovery rate (RRpre) is higher for C3

since it is easier to recover smaller outstanding values. Marital status=Other is strongly associated with C2 (note: in whole data, 14.5% of observation have Married=O). It is difficult to know exactly what a borrower intends when recording this status.

Table B.2 shows coefficient estimates on boundaries for the two separate recovery pseudo-classes using multinomial logistic regression for the recoveries data. There are a few variables that stand out. We find that Age behaves differently for each pseudo-class. So for the large loan value C2 pseudo-class, older people are more likely to repay, whereas among the low value C3 pseudo-class it is the other way round. Also, although it is less likely for a female borrower to be in C2, it turns out that if they are, they are less likely to repay the full amount, whereas they are more likely to repay a small loan, on average. The total number of outstanding loans (Total.number) is a risk factor for large loans, but not so much for smaller loans. The total_num_calls and total_num_contacts both relate to how frequently the customer was contacted prior to the debt being sold to the debt collection company. The context of the contact is unknown, but the coefficients demonstrate that the more previous contacts, the more probable a larger loan is paid off, whereas the contrary is true for a smaller loan. We would imagine that the type of recovery contact associated with a large loan would be quite different than for a small loan and this may explain the discrepancy. Similar reasoning can be given for the difference with pre-sale payment frequency. Looking at the relative difference of coefficient estimates amongst the marital statuses for each group, just one is striking: the coefficient for widows in C2 is very low; however, this is because there are no widows in C2 for the training data, hence this is just a poor estimate.

Table B.1: Descriptive statistics for in the two clusters of defaults in recover data, the second and third columns are the means of each variable the first and third shows the significant test and correspond p-value.

Variable	Test	C2 (N=67)	C3 (N=97)	p-value (less than 0.01)
Principal	T	2710 (1940)	385 (417)	<0.0001
Interest payments due	T	649 (407)	179 (177)	<0.0001
Credit limit	T	5490 (2820)	2460 (1960)	<0.0001
Outstanding balance at default	T	3660 (2220)	824 (465)	<0.0001
Age	T	36.3 (12.4)	26.9 (11.5)	<0.0001
Delphi Score	T	149 (183)	251 (179)	<0.0001
Total number	T	0.851 (1.41)	1.60 (1.78)	0.0031
RRpre	T	0.385 (0.351)	0.733 (0.387)	<0.0001
Sex=F	Fe	16.4%	54.1%	<0.0001
Married=O	Fe	65.7%	3.1%	<0.0001
Bureau information exists	Fe	61.1%	34.0%	0.0007753

Principal: original loan amount borrowed.

RRpre: recovery rates prior to debt being purchased.

Total number: total number of loan accounts.

Married: Marital status, O for other.

Table B.2: Coefficient estimates of multinomial logistic regression for recovery data

Variable	C2	C3
(Intercept)	-6.21962	1.72086
Principal	0.00213	0.00461
Interest	0.00261	0.00367
Insurance	0.07348	0.11017
Late charges	0.00176	0.00306
Over limit fees	-0.00699	-0.00757
Credit limit	0.00022	0.00020
Outstanding balance	-0.00248	-0.00723
Age	0.03263	-0.03963
Delphi Score	-0.00152	-0.00243
Total number	-0.22769	-0.03836
Total net paid amount	0.00016	0.00066
Total numbel of calls	0.00031	-0.00392
Total number of contacts	0.00546	-0.01014
RRpre	-0.67422	-0.19672
Product R	-16.48621	-0.33173
Product C	This is the base category	
Sex F	-0.72327	0.29184
Sex M	This is the base category	
Married S	2.19710	-0.26762
Married Blank	3.34404	-1.56225
Married D	0.67555	-0.32375
Married O	2.57717	0.37771
Married W	-15.20640	0.41571
Married M	This is the base category	
Employer No Information	-0.61682	-0.33218
Employer Employer Provided	This is the base category	
Number of Files 2	0.90711	-16.68695
Number of Files 1	This is the base category	
Bureau information exists	0.56645	-1.37245

Principal: Original loan amount borrowed.

Insurance: Insurance fees.

Late charges: Fees for late repayments.

Total number: total number of loan accounts.

Total net paid amount: Total loan amount for an individual in all accounts.

RRpre: recovery rates prior to debt being purchased.

Product: Type of loan, C for credit card, R for mortgage.

Married: Marital status, M for married, S for single, D for divorced, O for other, W for widow, Blank for unknown.

Employer: Data of the employer, categorical with two levels No Information/Employer Provided.

Number of Files: Not sure the meaning, with very little variation.

B.2 Freddie Mac 2009 Data

Table B.3 shows descriptive statistics for observations found in the two pseudo-classes of defaults, which are named C2 (119 observations) and C3 (18 observations). Only those variables that exhibit a difference between pseudo-classes that is statistically significant using either a t-test for continuous variables or a Fisher exact test for categorical variables at a 0.0001 level are shown. Comparing the smaller pseudo-class C3 with C2, we find it is characterized broadly as individuals with lower FICO scores, higher principal balance (UPB) and higher loan-to-value (LTV): all indicators of higher risk. The exception to this characteristic is debt-to-income (DTI) which is much lower in C3 than C2, on average. Hence C3 accounts for defaulters who are generally higher risk, except for a lower DTI. Amongst the other mortgage characteristics, C3 is less likely to include investment mortgages (note, this is typically the case when the first time homebuyer indicator is set to blank as well as when Occupancy.status=I) and less likely to have originated from a correspondent lender. Indeed, the last line of the table shows that these variables, together, are strongly associated with pseudo-class C2 rather than C3. Finally, we see that C3 is more likely to include condominium property purchases.

Table B.4 shows coefficient estimates on boundaries for the two separate default pseudo-classes using multinomial logistic regression. Focussing on the five main risk factors, we see that C2 has relatively high magnitude of coefficient estimates for DTI, UPB and OIR, whereas C3 has high coefficient estimates for FICO score and LTV. The differences in coefficient estimates are quite high. Indeed, surprisingly, FICO score has a negligible association with default for C2 and also UPB has negative association with default, which is counter-intuitive, unless we consider UPB a proxy for wealth, when taken along with DTI. OIR is often a proxy for credit risk, since it would have been set partly as a function of credit risk, hence in C2 it seems to be supplanting the FICO score in this role. Given that C2 contains more investment properties or originate from correspondent lenders, perhaps the OIR contains more credit risk signal from underwriting decisions than just

the FICO score, alone. The emphasis of C2 on DTI as a main risk factor reflects that DTI is much higher in C2, which suggests some threshold effect of DTI on default risk: ie it only really begins to have an effect above some level. The emphasis of C3 on FICO score and LTV risk factors is reflective of C3 being a slightly higher risk group than C2.

Table B.3: Descriptive statistics for the two pseudo-classes of defaults for US mrotgages in 2009, the second and third columns are the means of each variable the first and forth shows the significant test and correspond p-value.

Variable	Test	C2 (N=416)	C3 (N=272)	p-value (less than 0.01)
Score	T	748 (46)	723 (53)	<0.0001
DTI	T	59.0 (3.9)	38.3 (10.1)	<0.0001
UPB (log)	T	12.3 (0.6)	12.5 (0.5)	<0.0001
LTV	T	65.3 (16.2)	72.2 (14.3)	<0.0001
Original loan term	T	344 (51)	353 (34)	0.00027
Number of borrowers (1,2) = 2	Fe	47.1%	30.1%	<0.0001
First time homebuyer=blank (not applicable)	Fe	43.0%	18.4%	<0.0001
Insurance = yes	Fe	0%	13.6%	<0.0001
Occupancy status=I (investment)	Fe	33.9%	4.4%	<0.0001
Channel=C (correspondent)	Fe	71.6%	29.4%	<0.0001
Property type=CO (condo)	Fe	1.2%	9.6%	<0.0001
Loan purpose=P (purchase)	Fe	54.8%	39.3%	<0.0001
First time homebuyer=blank & Occupancy.status=I & Channel=C	Fe	22.8%	0.7%	<0.0001

Table B.4: Coefficient estimates for logistic regression between C1 and C2 (single model) and multinomial logistic regression

Variable	Single model	C2 coefficients	C3 coefficients
(Intercept)	-11.70993	-24.99017	1.06121
Score	-0.00650	-0.00101	-5.36259
DTI	0.15057	0.34276	-0.01665
UPB (log)	0.00126	-0.09806	0.00522
LTV	0.01007	0.00603	0.27384
OIR	0.92498	0.64968	0.03392
Original loan term	-0.00232	-0.00475	1.36464
Number of borrowers	-0.55512	-0.24292	0.00116
Seller	-0.73125	-1.35303	-1.12875
Servicer	-0.19476	-0.25508	0.55009
First time homebuyer=blank	-0.01674	0.37475	-0.65135
First time homebuyer=Y	-0.02910	0.38030	-0.73761
First time homebuyer=N	This is the base category		
Insurance	0.92890	-15.42656	-0.20138
Number of units=2	-0.05732	0.53622	0.91431
Number of units=4	0.83219	0.45550	-1.32299
Number of units=3	0.53823	0.41463	1.84850
Number of units=1	This is the base category		
Occupancy status=I	0.24243	0.58407	0.68492
Occupancy status=S	0.51913	0.28594	-1.05676
Occupancy status=O	This is the base category		
Channel=B	-0.10472	-0.02674	0.85781
Channel=C	0.58863	0.93971	-0.35180
Channel=R	This is the base category		
PPM	1.14164	-0.05567	-0.12578
Property type=PU	0.16950	0.34265	0.98686
Property type=CO	-0.26160	-0.72814	-0.03873
Property type=SF	This is the base category		
Loan purpose=C	0.39416	1.15580	-0.07411
Loan purpose=P	0.25857	1.06121	0.03092
Loan purpose=N	This is the base category		

First time homebuyer=blank means property is not eligible for first time home buyer status, which is the case of "Investment Properties, Second Homes and Refinance".

Number of units denotes whether the mortgage is a 1-, 2-, 3-, or 4-unit property.

Occupancy status=I, O, S means Investment, Owner Occupied, and Second Home respectively

Channel=R, B, C denotes the channel involved in origination of the mortgage loan, R = Retail, B = Broker, C = Correspondent.

PPM denotes whether the mortgage is a Prepayment Penalty Mortgage.

Property type=PU, CO, SF means Planned Unit development, COndominium, and Single Family home.

Loan.purpose=C, P, N means Cash-out Refinance, Purchase, and No Cash-out refinance.

C

Codes

C.1 Vuong's Likelihood Ratio Test in R

In this section, we give the R code for implementing Vuong's likelihood ratio test [Vuong, 1989] in R. In this process, we need the following function: the `imhof` function in the `CompQuadForm` package for calculating the p-value of a weighted sum of χ^2 distribution and the `estfun` function in the `sandwich` package for calculating the derivative of the likelihood function with respect to the parameter vector. Some functions are borrowed from the `nonnset2` [Merkle et al., 2016] package with modifications; the modifications are aiming at ensuring two models are deployed on the original full data set. We start with building two logistic regression on a simple simulation example: the majority class ($Y = 0$) follows $X \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and the minority class ($Y = 1$) follows $X \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.64 & 0 \\ 0 & 0.64 \end{pmatrix}$.

```
library(MASS)
# number of the minority class observations
```

```
n = 50
# imbalance ratio
r = 100
# number of the majority class observations
N =  n*r
# correlation coefficient
rho <- 0
mu1 <- 0; s1 <- 1
mu2 <- 0; s2 <- 1
mu <- c(mu1,mu2)
sigma <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2)
bvn1 <- mvrnorm(N, mu = mu, Sigma = sigma )

rho <- 0
mu1 <- 2; s1 <- 0.8
mu2 <- 2; s2 <- 0.8
mu <- c(mu1,mu2)
sigma <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2)
bvn2 <- mvrnorm(n, mu = mu, Sigma = sigma )

data = cbind(rbind(bvn1,bvn2),c(rep(0,N),rep(1,n)))
data = as.data.frame(data)
colnames(data) = c("x1","x2","y")

df = data
default = df[which(df$y == TRUE),]
non_default = df[which(df$y == FALSE),]

temp = as.data.frame(matrix(rep(colMeans(default),
                      each=nrow(default)),nrow=nrow(default)))
names(temp) = names(non_default)

# this is the manipulated data set
drop_data= rbind( non_default, temp)
# model M_1 in Section 5.1.2
glm.fit = glm(y ~. , data=data,
```

```
                family=binomial(link = "logit"))
# model M_2 in Section 5.1.2
glm.fit.drop = glm(y ~. , data=drop_data,
                family=binomial(link = "logit"))
```

Now, we have two logistic regression **glm.fit** and **glm.fit.drop**. The following three functions are used to calculate Equations (5.4, 5.5, 5.6, 5.7).

```
# Equation (5.7)
calcLambda <- function(object1, object2, n,
                       score1, score2, vc1, vc2){
  AB1 <- calcAB(object1, n, score1, vc1)
  AB2 <- calcAB(object2, n, score2, vc2)
  Bc <- calcBcross(AB1$sc, AB2$sc, n)
  W <- cbind(rbind(-AB1$B %*% chol2inv(chol(AB1$A)),
                   t(Bc) %*% chol2inv(chol(AB1$A))),
             rbind(-Bc %*% chol2inv(chol(AB2$A)),
                   AB2$B %*% chol2inv(chol(AB2$A))))

  lamstar <- eigen(W, only.values=TRUE)$values
  Re(lamstar)
}

# A, B as defined in Equations (5.4) and (5.5)

calcAB <- function(object, n, scfun, vc){
  scaling <- summary(object)$sigma
  if(is.null(scaling)){
    scaling <- 1
  } else {
    scaling <- scaling^2
  }
  tmpvc <- n * vc(object)
  A <- chol2inv(chol(tmpvc))
  sc <- (1/scaling) * sandwich::estfun(object)
  sc.cp <- crossprod(sc)/n
  B <- matrix(sc.cp, nrow(A), nrow(A))
```

```
  list(A=A, B=B, sc=sc)
}

# Equation (5.6)
calcBcross <- function(sc1, sc2, n){
  crossprod(sc1, sc2)/n
}
```

We now start to calculate the case wise contributions to the likelihood function of models **glm.fit** and **glm.fit.drop**.

```
llA = data$y*log(glm.fit$fitted.values)
        +(1-data$y)*log(1-glm.fit$fitted.values)
llB = data$y*log(glm.fit.drop$fitted.values)
        +(1-data$y)*log(1-glm.fit.drop$fitted.values)
```

Then we calculate $\hat{\omega}_\star^2$ (Equation 5.3) and the p-value of Vuong's likelihood ratio test.

```
omega.hat.2 <- (n+N-1)/(n+N) * var(llA - llB, na.rm = TRUE)
lr <- sum(llA - llB, na.rm = TRUE)
teststat <- (1/sqrt(n+N)) * lr/sqrt(omega.hat.2)
vcl <- function(obj) vcov(obj, full=TRUE)
lamstar <- calcLambda(glm.fit, glm.fit_drop, n=n+N,
       score1=Null, score2=Null, vc1=vcl, vc2=vcl)
pOmega <- CompQuadForm::imhof((n+N) * omega.hat.2, lamstar^2)$Qq
pOmega
```

The final result `pOmega` is the p-value of Vuong's likelihood ratio test, which should be around 9.2791×10^{-9} in our simulation.

C.2 Hotelling T^2 test in R

```
diffhotelling = function(y1, y2, se1, se2, df1, df2){
  # y1, se1, df1 are coefficients estimates, the corresponding
  # covariance matrix and number of observations used in model 1
  # y2, se2, df2 are coefficients estimates, the corresponding
```

```
  # covariance matrix and number of observations used in model 2
  n1 = df1
  n2 = df2

  p<-length(y1)

  S.pooled = ((n1-1)*se1 + (n2-1)*se2) / (n1+n2-2)

  test.statistic<-as.numeric(n1)*as.numeric(n2)/(n1+n2)*t(y1-y2)%*
  %solve(S.pooled)%*%(y1-y2)*((n1+n2-p-1)/(p*(n1+n2-2)))

  df.1<-p
  df.2<-n1+n2-p-1

  p.value<-1-pf(test.statistic,df.1,df.2)

  return(list(pvalue=p.value, t2=test.statistic, diff=y1-y2))
}
```

C.3 Brier Score z test in R

```
brier_score_test = function(p_ai,p_bi,pi_i,y_i){
# p_ai, p_bi: posterior probability from two model
# pi_i "true" posterior probability, can use (p_ai+p_bi)/2
# y_i: true label

  num = length(p_ai)

  d_ab =  sum(((p_ai-p_bi)*pi_i-(p_ai-p_bi)*y_i)*2/num)

  v_ab =  sum((p_ai-p_bi)^2*pi_i*(1-pi_i)*4/(num^2))

  z = d_ab/sqrt(v_ab)

  p_value = 2*pnorm(abs(z), lower.tail = F)
```

```
  return(c(z,p_value))
}
```

C.4 EM Algorithm Pseudo Code

The following are the pseudo codes for our EM algorithm:

Algorithm 2: Expectation Maximization Algorithm for Relabeling

Data: minority class $x_{1i}, i \in \{1, \cdots, n\}$ and majority class $x_{0i}, i \in \{1, \cdots, N\}$.

Result: $\widehat{p}_{ik}$, where $i \in \{1, \cdots, n\}$ and $k \in \{1, \cdots, K\}$, the posterior probability of a minority class observation i arises from pseudo-class k.

Initialization: iteration number $(u) = 1$;

$\widehat{p}_{ik}^{(u)}$ = a random number between 0 and 1;

$\phi_k^{(u)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{p}_{ik}^{(u)}$;

optimize $\alpha_k^{(u)}$ and $\beta_k^{(u)}$ in multinomial logistic regression

$Q_1^{(u)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{p}_{ik}^{(u)} \log(\Pr(z_i = k \mid x_{1i})) + \sum_{i=1}^{N} \log(\Pr(y_i = 0 \mid x_{0i}))$;

and $Q_2^{(u)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{p}_{ik}^{(u)} \log(\phi_k^{(u)})$;

then assign $Q^{(u)} = Q_1^{(u)} + Q_2^{(u)}$ and $Q^{(u-1)} = 0.5 \times Q^{(u)}$;

while $(Q^{(u)} - Q^{(u-1)})/Q^{(u-1)} > 0.01\%$ **do**

update $\widehat{p}_{ik}^{(u+1)} = \frac{\phi_k^{(u)} \Pr\{z_i = k | x_{1i}\}}{\sum_{k=1}^{K} \phi_k^{(u)} \Pr\{z_i = k | x_{1i}\}}; i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$;

// E-step

optimize $\alpha_k^{(u+1)}$ and $\beta_k^{(u+1)}$ in multinomial logistic regression

$Q_1^{(u+1)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{p}_{ik}^{(u+1)} \log(\Pr(z_i = k \mid x_{1i})) + \sum_{i=1}^{N} \log(\Pr(y_i = 0 \mid x_{0i}))$; // M-step

update $Q_2^{(u+1)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{p}_{ik}^{(u+1)} \log(\phi_k^{(u+1)})$ by letting

$\phi_k^{(u+1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{p}_{ik}^{(u+1)}$; // M-step

$Q^{(u+1)} = Q_1^{(u+1)} + Q_2^{(u+1)}$; // new log-likelihood

$u = u + 1$;

end

Here (u) is the number of current iteration. In each iteration (u), Q_1 is a weighted multinomial logistic regression with

$$\Pr(z_i = k \mid x_{1i}) = \frac{e^{\alpha_k^{(u)} + x_{1i}^T \beta_k^{(u)}}}{1 + \sum_{j=1}^{K} e^{\alpha_j^{(u)} + x_{1i}^T \beta_j^{(u)}}}, i \in \{1, \cdots, n\}.$$

and

$$\Pr(y_i = 0 \mid x_{0i}) = \frac{1}{1 + \sum_{j=1}^{K} e^{\alpha_j^{(u)} + x_{0i}^T \beta_j^{(u)}}}, i \in \{1, \cdots, N\}.$$

Demo R code can be found here: https://github.com/yazheli/code-and-data.

D

Infinitely Imbalanced Ridge Penalized Logistic Regression

We again use the notation of Section 3.1.2. In order to directly show the result, we center ridge penalized logistic regression around the minority class mean vector $\overline{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$. Since in the infinitely imbalanced case $N \to \infty$, we also suppose that there is a good approximation for the conditional distribution of $\mathbf{x}$ given $Y = 0$; denoted by F_0. Thus, the objective function for ridge penalized logistic regression [Hoerl and Kennard, 1970] is written as

$$\begin{aligned} l(\beta_0, \beta) = &\frac{1}{n+N}\Bigg[n\beta_0 - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T\beta}) \\ &- N \int \log(1 + e^{\beta_0 + (\mathbf{x} - \bar{\mathbf{x}})^T\beta}) dF_0(\mathbf{x})\Bigg] - \frac{1}{2}\lambda\|\beta\|^2, \ \lambda > 0, \end{aligned} \tag{D.1}$$

We follow Owen's proof again: Lemma 4 and Lemma 5 in Owen [2007] still hold for penalized logistic regression. The three changes in the proof process are for Lemma 6, Lemma 7 and the main theorem in Owen [2007] (corresponding to our Lemma 17, Lemma 18 and Theorem 20 here). Our Lemma 17

gives $e^{\hat{\beta}_0} \leq n/(N-n)$ and Lemma 18 gives $\left\|\hat{\beta}\right\| \leq \sqrt{2n/((N-n)\lambda)}$ as $N \to \infty$. Note that in Lemma 17 and Lemma 18, we do **not** require the surrounded condition, which makes the proof **significantly different** from Owen's proof.

Lemma 17. *Let $\hat{\beta}_0$ and $\hat{\beta}$ be the maximizers of the objective function (D.1). Then $e^{\hat{\beta}_0} \leq n/(N-n)$.*

Proof. Calculate the partial derivative with respect to β_0:

$$\begin{aligned}
\frac{\partial l(\beta_0,\beta)}{\partial \beta_0} &= \frac{n}{n+N} - \frac{1}{n+N}\sum_{i=1}^{n}\frac{e^{\beta_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\beta}} \\
&\quad - \frac{N}{n+N}\int \frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}} dF_0(\mathbf{x}) \\
&\leq \frac{n}{n+N} - \frac{N}{n+N}\int \frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}} dF_0(\mathbf{x}) \\
&\leq \frac{n}{n+N} - \frac{N}{n+N}\int_{(\mathbf{x}-\bar{\mathbf{x}})^T\beta>0} \frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}} dF_0(\mathbf{x}) \\
&\leq \frac{n}{n+N} - \frac{N}{n+N}\frac{e^{\beta_0}}{1+e^{\beta_0}}\int_{(\mathbf{x}-\bar{\mathbf{x}})^T\beta>0} dF_0(\mathbf{x}) \\
&\leq \frac{n}{n+N} - \frac{N}{n+N}\frac{e^{\beta_0}}{1+e^{\beta_0}}.
\end{aligned} \tag{D.2}$$

We applied the fact that $\frac{e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}}{1+e^{\beta_0+(\mathbf{x}-\bar{\mathbf{x}})^T\beta}} \leq \frac{e^{\beta_0}}{1+e^{\beta_0}}$, when$(\mathbf{x}-\bar{\mathbf{x}})^T\beta > 0$, in the above inequality. Then, let $e^{\beta_0} > n/(N-n)$, the above equation leads to $\frac{\partial l(\beta_0,\beta)}{\partial \beta_0} < 0$. For the concave likelihood function, the negative derivative means that the maximizer $\hat{\beta}_0 \leq \log(\frac{n}{N-n})$. □

Lemma 18. *Let $\hat{\beta}_0$ and $\hat{\beta}$ be the maximizers of the log-likelihood function (D.1). Then $\limsup_{N\to\infty}\left\|\hat{\beta}\right\| < \infty$.*

Proof. Take arbitrary coefficient estimates $(\hat{\beta}_0, 0)$, we know $l(\hat{\beta}_0,\hat{\beta}) - l(\hat{\beta}_0, 0) \geq$

0. Then

$$l(\hat{\beta}_0, 0) - l(\hat{\beta}_0, \hat{\beta}) = \frac{1}{n+N}\Bigg[- (n+N)\log(1+e^{\hat{\beta}_0}) + \sum_{i=1}^{n} \log(1+e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T\hat{\beta}}) + N\int \log(1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}})dF_0(\mathbf{x})\Bigg] + \frac{1}{2}\lambda\left\|\hat{\beta}\right\|^2 \geq 0. \tag{D.3}$$

Since $\log(1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}) \geq 0$ and $\int \log(1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}})dF_0(\mathbf{x}) \geq 0$, Equation (D.3) leads to:

$$\begin{aligned}\frac{1}{2}\lambda\|\beta\|^2 \leq \frac{1}{n+N}\Bigg[&(n+N)\log(1+e^{\hat{\beta}_0}) - \sum_{i=1}^{n}\log(1+e^{\hat{\beta}_0+(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\beta}}) \\ &- N\int \log(1+e^{\hat{\beta}_0+(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}})dF_0(\mathbf{x})\Bigg] \\ \leq \log(1+e^{\hat{\beta}_0}) \leq e^{\hat{\beta}_0} &\leq \frac{n}{N-n}.\end{aligned} \tag{D.4}$$

Thus, we know $\left\|\hat{\beta}\right\|$ is bounded as $N \to \infty$. □

The following theorem demonstrates the behavior of $\hat{\beta}_0$ and $\hat{\beta}$ in infinitely imbalanced ridge penalized logistic regression.

Theorem 19. *Let $n > 1$ and minority class vectors $\{\mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_n\}$ be fixed. Then the maximizer $(\hat{\beta}_0, \hat{\beta})$ of l given by Equation (D.1) have following shrinkage rules, $e^{\hat{\beta}_0} \to \frac{n}{N}$ and $\hat{\beta} \to 0$, when $N \to \infty$.*

Proof. From Lemma 17 and Lemma 18, we know $e^{\hat{\beta}_0} \leq \frac{n}{N-n}$ and $\left\|\hat{\beta}\right\|$ is bounded when $N \to \infty$.

Further considering Lemma 18, we have

$$\left\|\hat{\beta}\right\| \leq \sqrt{\frac{2n}{(N-n)\lambda}}, \text{ when } N \to \infty. \tag{D.5}$$

Inequality (D.5) shows as $N \to \infty$, we have $\hat{\beta} = 0$. Thus $l(\hat{\beta}_0, \hat{\beta})$ simplifies

to

$$l(\hat{\beta}_0, \hat{\beta}) = \frac{n}{n+N}\hat{\beta}_0 - \log(1 + e^{\hat{\beta}_0}). \tag{D.6}$$

Let the partial derivative of Equation (D.6) equal to 0 when $N \to \infty$,

$$\frac{\partial l(\hat{\beta}_0, \hat{\beta})}{\partial \hat{\beta}_0} = \frac{n}{n+N} - \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} = 0, \tag{D.7}$$

then we have $e^{\hat{\beta}_0} = n/N$. □

In the next theorem, we do not use subgradient method for lasso again because the derivative of ridge penalty exists.

Theorem 20. *Let $n \geq 1$ and minority class vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be fixed and suppose that F_0 surrounds $\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$ as described. Then the maximizer $(\hat{\beta}_0, \hat{\beta})$ of l given by Equation (D.1) satisfies*

$$-\int (\mathbf{x} - \bar{\mathbf{x}}) e^{(\mathbf{x}-\bar{\mathbf{x}})^T \hat{\beta}} dF_0(\mathbf{x}) = \frac{n+N}{n}\lambda\hat{\beta} \tag{D.8}$$

as $N \to \infty$.

Proof. Setting the partial derivative with respect to β of Equation (D.1) to 0, we have:

$$-\sum_{i=1}^{n} \frac{(\mathbf{x}_i - \bar{\mathbf{x}}) e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}} - N \int \frac{(\mathbf{x} - \bar{\mathbf{x}}) e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}} dF_0(\mathbf{x}) - (n+N)\lambda\hat{\beta} = 0. \tag{D.9}$$

Dividing by N gives

$$-\int \frac{(\mathbf{x} - \bar{\mathbf{x}}) e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\beta}}} dF_0(\mathbf{x}) - \frac{n+N}{N}\lambda\hat{\beta} = \frac{1}{N}\sum_{i=1}^{n} \frac{(\mathbf{x}_i - \bar{\mathbf{x}}) e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}}}. \tag{D.10}$$

As $N \to \infty$, the right side of Equation (D.10) vanishes because $\left\|\hat{\beta}\right\|$ is bounded as $N \to \infty$ by Lemma 18.

If we consider $N \to \infty$, we have $e^{\hat{\beta}_0} \to \frac{n}{N}$ and $\hat{\beta} \to 0$, yielding to $e^{\hat{\beta}_0 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\beta}} \to$

$\frac{n}{N}e^0 \to 0$. Thus Equation (3.27) yields

$$-\int \frac{(\mathbf{x}-\bar{\mathbf{x}})\frac{n}{N}e^{(\mathbf{x}-\bar{\mathbf{x}})^T\hat{\beta}}}{1}dF_0(\mathbf{x}) = \frac{n+N}{N}\lambda\hat{\beta} \tag{D.11}$$

After simplification, Equation (D.8) holds. □

Equation (D.8) shows the solution of β depends only on $\{\bar{\mathbf{x}}, F_0(\mathbf{x}), \frac{N}{n}\}$ when approaching infinite imbalance.

References

I. Ahmed, A. Pariente, and P. Tubert-Bitter. Class imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research*, 27(3):785–797, 2018.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

E. I. Altman and G. Sabato. Modelling credit risk for SMEs: evidence from the US market. *Journal of Accounting Finance and Business Studies*, 43 (3):332–357, 2007.

J. Anderson and V. Blair. Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, 69(1):123–136, 1982.

J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59(1): 19–35, 1972.

B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

A. Bagherpour. Predicting mortgage loan default with machine learning methods. Technical report, University of California, Riverside, 2017. URL https://pdfs.semanticscholar.org/a4e5/3d7255dd397da78242c4ad41213a404cb51e.pdf.

A. C. Bahnsen, D. Aouada, and B. Ottersten. Example-dependent cost-sensitive logistic regression for credit scoring. In *13th International Conference on Machine Learning and Applications*, pages 263–269. IEEE, 2014.

Basel II Accords. International convergence of capital measurement and capital standards: a revised framework. *Bank for International Settlements*, 2004.

C. B. Begg and R. Gray. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71(1):11–18, 1984.

R. Brause, T. Langsdorf, and M. Hepp. Neural data mining for credit card fraud detection. In *Proceedings 11th International Conference on Tools with Artificial Intelligence*, pages 103–106. IEEE, 1999.

C. Bravo, L. C. Thomas, and R. Weber. Improving credit scoring by differentiating defaulter behaviour. *Journal of the Operational Research Society*, 66(5):771–781, 2015.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman and P. Spector. Submodel selection and evaluation in regression, the X-random case. *International Statistical Review*, 60(3):291–319, 1992.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

I. Brown and C. Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

K. P. Burnham and D. R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33 (2):261–304, 2004.

J. Y. Campbell and J. F. Cocco. A model of mortgage default. *The Journal of Finance*, 70(4):1495–1554, 2015.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. Technical report, University of California, Berkeley, 2004. URL https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166. IEEE, 2015.

M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Pearson Education, 2012.

E. Demidenko. The p-value you can not buy. *The American Statistician*, 70 (1):33–38, 2016.

P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 99, pages 155–164, 1999.

Y. Dong, H. Guo, W. Zhi, and M. Fan. Class imbalance oriented logistic regression. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 187–192. IEEE, 2014.

B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2016.

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1): 18–36, 2004.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Z. D. Feng and C. E. McCulloch. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3):609–617, 1996.

T. Fitzpatrick and C. Mues. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249(2):427–439, 2016.

S. García and F. Herrera. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. *Evolutionary Computation*, 17(3):275–306, 2009.

R. M. Golden. Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, 44(1):153–170, 2000.

X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. On the class imbalance problem. In *ICNC Fourth International Natural Computation Conference*, volume 4, pages 192–201. IEEE, 2008.

H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-SMOTE: a new oversampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887. Springer, 2005.

D. J. Hand. Reject inference in credit operations. In *Credit Risk Modeling: Design and Application*, chapter 11, pages 181–190. Fitzroy Dearborn, 1998.

D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, 2009.

D. J. Hand and C. Anagnostopoulos. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34(5):492–495, 2013.

D. J. Hand and C. Anagnostopoulos. A better Beta for the H measure of classification performance. *Pattern Recognition Letters*, 40:41–46, 2014.

D. J. Hand and R. J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

D. J. Hand, C. Whitrow, N. M. Adams, P. Juszczak, and D. Weston. Performance criteria for plastic card fraud detection tools. *Journal of the Operational Research Society*, 59(7):956–962, 2008.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108, 1979.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science and Business Media, 2009.

R. L. Haupt and S. Ellen Haupt. *Practical Genetic Algorithms*. Wiley Online Library, 2004.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, (9):1263–1284, 2008.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

H. Hotelling. The generalization of Student ratio. In *Breakthroughs in Statistics*, pages 54–65. Springer, 1992.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

M. Janio. Lending club loan analysis. *Kaggle*, 2017. URL https://www.kaggle.com/janiobachmann/lending-club-risk-analysis-and-metrics.

T. Jo and N. Japkowicz. Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1):40–49, 2004.

S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

U. Kaymak, A. Ben-David, and R. Potharst. The auk: A simple alternative to the auc. *Engineering Applications of Artificial Intelligence*, 25(5):1082–1089, 2012.

D. J. Ketchen and C. L. Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17(6):441–458, 1996.

G. King and L. Zeng. Explaining rare events in international relations. *International Organization*, 55(3):693–715, 2001a.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001b.

W. R. Klecka, G. R. Iversen, and W. R. Klecka. *Discriminant Analysis*. Sage, 1980.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

G. Krempl and V. Hofer. Classification in presence of drift and latency. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 596–603. IEEE, 2011.

W. J. Krzanowski and D. J. Hand. *ROC Curves for Continuous Data*. Chapman and Hall, 2009.

M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *International Conference on Machine Learning*, volume 97, pages 179–186. Citeseer, 1997.

M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3):195–215, 1998.

M. Kumar, M. Husian, N. Upreti, and D. Gupta. Genetic algorithm: review and application. *International Journal of Information Technology and Knowledge Management*, 2(2):451–454, 2010.

P. A. Lachenbruch and M. Goldstein. Discriminant analysis. *Biometrics*, 35 (1):69–85, 1979.

J. Laurikkala. Improving identification of difficult small classes by balancing class distribution. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 63–66. Springer, 2001.

S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: a ten-year update. *European Journal of Operational Research*, 247(1):124–136, 2015.

D.-C. Li, C.-W. Liu, and S. C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*, 40 (5):509–518, 2010.

L. Li, N. Sedransk, et al. Mixtures of distributions: a topological approach. *The Annals of Statistics*, 16(4):1623–1634, 1988.

Y. Li, T. Bellotti, and N. Adams. Issues using logistic regression with class imbalance, with a case study from credit risk modelling. *Foundations of Data Science*, 1(4):389–417, 2019.

Y. Li, N. Adams, and T. Bellotti. A relabeling approach to handling the class imbalance problem for logistic regression. *Journal of Computational and Graphical Statistics*, under review, 2020.

X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

V. López, A. Fernández, S. García, V. Palade, and F. Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

B. Mac Namee, P. Cunningham, S. Byrne, and O. I. Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, 24(1):51–70, 2002.

S. Maldonado, J. López, and C. Vairetti. An alternative smote oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76:380–389, 2019.

I. Mani and I. Zhang. kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*, volume 126, 2003.

S. Masoudnia and R. Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

G. J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324, 1987.

G. J. McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999.

D. Mease, A. J. Wyner, and A. Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8 (3):409–439, 2007.

I. Meilijson. A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1): 127–138, 1989.

E. C. Merkle, D. You, and K. J. Preacher. Testing nonnested structural equation models. *Psychological Methods*, 21(2):151–163, 2016.

S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science and Business Media, 2013.

A. Nickerson, N. Japkowicz, and E. E. Milios. Using unsupervised learning to guide resampling in imbalanced data sets. In *International Conference on Artificial Intelligence and Statistics*, 2001.

OCC, US Department of the Treasury. Uniform retail credit classification and account management policy. 2000. URL https://ithandbook.ffiec.gov/media/resources/3677/occ-bl2000-20_ffiec_uniform_retail_credit_class.pdf.

D. Olszewski. Fraud detection using self-organizing map visualizing the user profiles. *Knowledge-Based Systems*, 70:324–334, 2014.

A. B. Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(4):761–773, 2007.

D. A. Redelmeier, D. A. Bloch, and D. H. Hickam. Assessing predictive accuracy: how to compare Brier scores. *Journal of Clinical Epidemiology*, 44(11):1141–1146, 1991.

R. Rockafellar. *Convex Analysis*. Princeton University Press, 2015.

S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 576–584. IEEE, 2004.

V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE, 2011.

V. S. Sheng and C. X. Ling. Thresholding for making classifiers cost-sensitive. In *AAAI 21st National Conference on Artificial Intelligence*, pages 476–481, 2006.

M. J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(3):310–313, 1981.

P. Smyth. Clustering sequences with hidden Markov models. In *9th International Conference on Neural Information Processing System*, pages 648–654, 1997.

P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

D. J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5(5):421–433, 1986.

M. A. Tanner. *Tools for Statistical Inference*. Springer, 2012.

L. C. Thomas. *Consumer Credit Models: Pricing, Profit and Portfolios*. OUP Oxford, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

R. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

D. Titterington. Some recent research in the analysis of mixture distributions. *Statistics*, 21(4):619–641, 1990.

US Federal Reserve Bank. Joint final rule: Risk-based capital standards: Advanced capital adequacy framework Basel II. 2007. URL https://www.govinfo.gov/content/pkg/FR-2007-12-07/pdf/07-5729.pdf.

W. N. van Wieringen. Lecture notes on ridge regression. *preprint arXiv:1509.09169*, 2015.

S. Visa and A. Ralescu. Issues in mining imbalanced data sets, a review paper. In *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*, pages 67–73. MAICS, 2005.

Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

B. X. Wang and N. Japkowicz. Imbalanced data set learning with synthetic samples. In *IRIS Machine Learning Workshop*, 2004.

H. Wang, Q. Xu, and L. Zhou. Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLoS One*, 10(2):e0117844, 2015.

K. Wendy. Lending club loan data analyze. *Kaggle*, 2004. URL https://www.kaggle.com/wendykan/lending-club-loan-data.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

S. Wu, P. Flach, and C. Ferri. An improved model selection heuristic for auc. In *European Conference on Machine Learning*, pages 478–489. Springer, 2007.

H. Ye and A. Bellotti. Modelling recovery rates for non-performing loans. *Risks*, 7(1):19, 2019.

I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

G. Zeng. On the existence of maximum likelihood estimates for weighted logistic regression. *Communications in Statistics-Theory and Methods*, pages 1–10, 2017.

S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4(11):1001–1037, 2003.

M. Zhu, W. Su, and H. A. Chipman. Lago: A computationally efficient approach for statistical detection. *Technometrics*, 48(2):193–205, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodology)*, 67 (2):301–320, 2005.