ISSUES USING LOGISTIC REGRESSION WITH CLASS IMBALANCE, WITH A CASE STUDY FROM CREDIT RISK MODELLING

Yazhe Li*, Tony Bellotti and Niall Adams
Department of Mathematics, Imperial College London
London, SW7 2AZ, UK

Abstract. The class imbalance problem arises in two-class classification problems, when the less frequent (minority) class is observed much less than the majority class. This characteristic is endemic in many problems such as modeling default or fraud detection. Recent work by Owen [19] has shown that, in a theoretical context related to infinite imbalance, logistic regression behaves in such a way that all data in the rare class can be replaced by their mean vector to achieve the same coefficient estimates. We build on Owen’s results to show the phenomenon remains true for both weighted and penalized likelihood methods. Such results suggest that problems may occur if there is structure within the rare class that is not captured by the mean vector. We demonstrate this problem and suggest a relabelling solution based on clustering the minority class. In a simulation and a real mortgage dataset, we show that logistic regression is not able to provide the best out-of-sample predictive performance and that an approach that is able to model underlying structure in the minority class is often superior.

1. Introduction. Class imbalance is a common problem in the field of classification. For example, banks 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 whether or not they default, that is, fail to satisfy their repayment obligations. This process is known as credit scoring. The most popular approach for consumer credit risk modelling is logistic regression [25, p. 79], which is regarded as a benchmark in the financial industry. There are a number of statistical modelling challenges in credit scoring, such as reject inference [10], handling variation over time and class imbalance. The latter is a common problem in credit scoring in which the bad class is much less frequent than the good. As an example of high imbalance, we introduce the Freddie Mac mortgage data set for which the number of default samples is profoundly smaller than the number of non-default samples. 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 2015.¹ Here, we define default as a mortgage that is 180 days or more overdue in making a repayment on

¹Data and detail available at: http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html
their home loan. This is a standard definition of default in U.S. financial institutions [6]. The target variable we use is whether a mortgage moved to the default status in the two years immediately following the first repayment date. The upper plot of Figure 1 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.

Modelling such imbalanced data is a challenging problem, fraught with difficulties, and it is the main concern of this paper. Specifically, we concentrate on the application of logistic regression and related methods in highly imbalanced data sets. Owen [19] 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. In particular, if we suspect some underlying cluster structure within the minority class, then this will be lost when data is reduced to the sample mean vector. Two natural choices to alleviate these problems are penalizing and weighting the likelihood, which are used in the credit risk industry [28, 12]. 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 make matters worse.

We sketch a new procedure that attempts to handle a specific consequence of the highly imbalanced logistic regression, namely the minority class’s contribution in the estimation process only comes from its mean vector. Essentially, this procedure seeks to cluster the minority class into two (or more) latent subclasses to improve
the predictive performance of the model, with the underlying assumption that such useful subclasses exist within the minority class. This approach is shown to be effective in a simulation study and in modelling default for real mortgage data.

The outline of the paper is as follows. The next section provides background to logistic regression, highly imbalanced data, and the relationship between the two. Section 3 considers methods that extend logistic regression and are attractive proposals for credit risk modelling. New theorems are given for infinitely imbalanced weighted logistic regression and penalized logistic regression. We use a relabelling clustering approach to demonstrate problems arising from imbalanced classes with logistic regression, with a simulation study in Section 4 and using the Freddie Mac data in Section 5.

2. Logistic regression and class imbalance. In this section, we first introduce logistic regression and give a short review of the class imbalance problem, mentioning some sampling methods proposed in the literature to deal with this problem. Then we introduce the boundary behavior of logistic regression in the infinitely imbalanced data set [19] and a result about the existence of the maximum likelihood estimate (MLE) for logistic regression [23] as a preparation for further investigation of highly imbalanced logistic regression.

2.1. Logistic regression. Consumer credit risk classification usually uses binary logistic regression to predict the probability of a customer defaulting on a loan. Denote the binary response as $Y \in \{0, 1\}$ with $Y = 1$ indicating a default. $\Pr(Y = 1|X = x)$ refers to the conditional probability of a vector $x$ belonging to class 1 whenever $X = x$. The binary logistic regression model has the form:

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

(1)

where $\beta_0$ is an intercept and $\beta$ is a slope parameter vector, to be estimated. Suppose we have $M$ observations in a credit data set, then the log-likelihood function for independent observations can be written as:

$$l(\beta) = \sum_{i=1}^{M} \{y_i \log(\Pr(Y = 1|X = x)) + (1 - y_i) \log(\Pr(Y = 0|X = x))\}$$

$$= \sum_{i=1}^{M} \{y_i (\beta_0 + \beta^T x_i) - \log\left(1 + e^{\beta_0 + \beta^T x_i}\right)\}.$$  

(2)

To maximize Equation (2), we set its derivatives to zero and solve for $\beta_0$ and $\beta$. To provide solutions, and hence estimate parameters, various numerical optimization techniques are used.

2.2. Classification with imbalanced data. High imbalance means one of the two classes is extremely rare in the binary classification problem. Highly imbalanced credit data is common in the credit risk industry. For example, in mortgage data sets (such as that described in Section 1), the default rate is often less than 0.5%, which means the default class is extremely rare.

Various authors have investigated such imbalanced (especially highly imbalanced) problems. One type of proposal attempts to construct a balanced data set [14], and thus various sampling methods have been developed to address highly imbalanced problems. Random undersampling and random oversampling are two natural methods which randomly remove samples from the majority class or randomly replicate
the minority class. Random undersampling obviously leads to information loss in the majority class. Thus, Liu et al. [15] proposed two informed undersampling methods to overcome the drawback of random undersampling. Another widely used oversampling method is SMOTE (Synthetic Minority Over-Sampling technique), which shows good results in various applications [4]. In contrast to random oversampling, SMOTE adds artificially generated data which has the same character as the distribution of the minority class.

At the same time, an investigation also shows that models trained from the imbalanced data set could have the same predictive power as models trained from balanced data set [2].

Apart from sampling methods, another common technique to handle imbalance data is reweighting the likelihood function. Reweighting methods pass the weights to each observation directly to the likelihood function. Seiffet et al. [22] investigate the difference between reweighting and sampling methods for imbalanced data in boosting algorithms and find reweighting performs as well as sampling methods.

In summary, 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 now discuss.

2.3. 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 \( x_{11}, \ldots, x_{1n} \), and \( N \) feature vectors from class \( Y = 0 \), \( x_{01}, \ldots, x_{0N} \). In order to accommodate the intercept term in the regression parameters, let \( z_0 = (1, x_{0i}) \) for \( i = (1, \ldots, N) \) and \( z_1 = (1, x_{1i}) \) for \( i = (1, \ldots, n) \). Let \( S, F \) be the two relative interiors of the convex cones generated by \( x_{11}, \ldots, x_{1n} \) and \( x_{01}, \ldots, x_{0N} \) respectively,

\[
S = \left\{ \sum_{i=1}^{n} k_i z_{1i} | k_i > 0 \right\} \quad \text{and} \quad F = \left\{ \sum_{i=1}^{N} k_i z_{0i} | k_i > 0 \right\}.
\]

(3)

When \( S \cap F \neq \emptyset \), then a unique MLE for logistic regression exists. However, if \( S \cap F = \emptyset \) then no MLE exists [23]. 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.

2.4. Owen’s results about infinitely imbalanced logistic regression.

Here, we introduce the result about the limit behavior of logistic regression in infinitely imbalanced problems. Following the notation in [19], we focus on the case when \( n \ll N \). To demonstrate the result, Owen centers logistic regression around the minority class mean vector \( \bar{x} = \sum_{i=1}^{n} x_{1i} / n \). Since in the infinitely imbalanced case \( N \to \infty \), Owen also supposes that there is a good approximation for the conditional distribution of \( X \) given \( Y = 0 \) (majority class); denoted by \( F_0 \). Thus, it is shown that the log-likelihood function, Equation (2), can be written as

\[
l(\beta_0, \beta) = n \beta_0 - \sum_{i=1}^{n} \log \left( 1 + e^{\beta_0 + (x_i - \bar{x})^T \beta} \right) - N \int \log \left( 1 + e^{\beta_0 + (x - \bar{x})^T \beta} \right) dF_0(x),
\]

(4)

where \( F_0 \) is a distribution function for \( X \) given that \( Y = 0 \). It is convenient to separate the intercept and slope terms in the parameter vector, thus \( \beta_0 \) denotes the
intercept term, $\beta$ denotes the slope terms, and $x_i$ denotes minority class data only. Note that from here on, in this paper, $\{x_i, i \in [1, 2, \cdots, n]\}$ denotes only minority class data. Owen proposes the following definition to express the overlap condition (Section 2.3) in the case of infinitely imbalanced logistic regression.

**Definition 1.** (Definition 3 in [19]) The distribution $F$ on $R^p$ has a point $x_*$ surrounded if

$$\int_{(x-x_*)^T \psi \geq \epsilon} dF(x) > \delta$$

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

Owen also assumes that

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

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

**Theorem 1.** (Theorem 8 in [19]) Let $n \geq 1$, and $x_1, \cdots, x_n \in R^p$ be fixed. Suppose that $F_0$ satisfies the tail condition (Equation 6) and surrounds the class mean vector $\overline{x} = \sum_{i=1}^{n} x_i/n$. Then the maximizer $(\hat{\beta}_0, \beta)$ of $l(\beta_0, \beta)$ given by Equation (4) satisfies

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

This theorem may be understood as: when $N \to \infty$, logistic regression only depends on the minority class data $\{x_1, x_2, \cdots, x_n\}$ through the minority class mean vector $\overline{x}$; we could replace $\{x_1, \cdots, x_n\}$ by one vector, the mean vector of the minority class, and obtain the same coefficient estimates of $\beta$ in the limit $N \to \infty$. Theorem 1 is an asymptotic theoretical result and Owen’s simulation [19, p.763 Table 1] shows that convergence happens rapidly from $N/n > 100$. In practice, we replace all the minority class with their mean vector in the Freddie Mac mortgage data set when we get the same coefficient estimates as using all the original minority class data.

Owen [19] notes other 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. These issues are explored later, in Sections 4 and 5.

### 3. Highly imbalanced weighted logistic regression and penalized logistic regression.

We have briefly described previous work about the existence of the maximum likelihood estimator for logistic regression and the limiting behavior of parameter estimates in highly imbalanced data sets. As mentioned in Section 1, a natural idea to address the imbalance problem would be to reweight the likelihood, leading to weighted logistic regression. Alternatively, penalized logistic regression adds a complexity penalty to the log-likelihood function (Equation 2) and, in the case of the lasso [26], is designed for coefficient shrinkage and variable selection, which is another commonly used alternative model to logistic regression in the credit risk industry [28, 8, p. 313]. We might propose them as solutions to alleviate the imbalance problem. However, we extend Owen’s results to weighted logistic regression and penalized logistic regression in Sections 3.1 and 3.2 and show that both these extensions still suffer from the limiting behavior with highly imbalanced data.
3.1. Highly imbalanced weighted logistic regression. Zeng [30] 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 [23], described in Section 2.3. Next, we identify the characteristics of infinitely imbalanced weighted logistic regression for a specific weighting strategy.

Consider retaining a weight of one for the majority class, as in Equation (4), 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

\[
l(\beta_0, \beta; \Omega) = \sum_{i=1}^{n} \omega_i \beta_0 - \sum_{i=1}^{n} \omega_i \log \left( 1 + e^{\beta_0 + (x_i - \bar{x})^T \beta} \right) \]

\[
- N \int \log \left( 1 + e^{\beta_0 + (x_i - \bar{x})^T \beta} \right) \ dF_0(x).
\]

where here \( \bar{x} \) is a weighted minority class mean vector \( \bar{x} = \sum_{i=1}^{n} \omega_i x_i / \sum_{i=1}^{n} \omega_i \).

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. The numbering of the lemmas corresponds to those in [19]. Lemma 2 and Lemma 4 in [19] are numerical facts which also hold for this weighted logistic regression. Lemma 5 is used to prove the existence of a finite MLE when the surrounded condition (Definition 1) is satisfied. Lemma 6a and Lemma 7a are used to establish a boundary of \( \| \beta \| \) when \( N \to \infty \).

**Lemma 5.** Suppose that weights \( \Omega \) are fixed, \( \{x_1, x_2, \cdots, x_n\} \) are given and \( F_0 \) satisfies the surrounded condition (i.e. Equation 5) at the weighted minority class mean vector \( \bar{x} \). 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” [19], which excludes the situation that \( l(\beta_0, \beta; \Omega) \) will increase infinitely along some ray.

If \( \beta = 0 \), we assume that \( 0 < |\beta_0| < \epsilon/2 \), \( \epsilon \) 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 \( 0 \leq |\beta_0| < \epsilon/2 \), \( \epsilon \) from Definition 1, by following the ray back to the origin. Using \( \sum_{i=1}^{n} \omega_i \beta_0 = \sum_{i=1}^{n} \omega_i \beta_0 + (x_i - \bar{x})^T \beta \), we have

\[
\lim_{\lambda \to \infty} \frac{\partial}{\partial \lambda} l(\lambda \beta_0, \lambda \beta; \Omega) = \sum_{i: \beta_0 + (x_i - \bar{x})^T \beta < 0} \omega_i \left[ \beta_0 + (x_i - \bar{x})^T \beta \right] \\
-N \int_{\beta_0 + (x - \bar{x})^T \beta > 0} \beta_0 + (x - \bar{x})^T \beta dF_0(x).
\]

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 6a. Let \( \hat{\beta}_0 \) and \( \hat{\beta} \) be the maximizer of the likelihood function, \( F_0 \) satisfy the surrounded condition at the weighted minority class mean vector \( \bar{x} \) and \( \eta \) is the infimum of \( \delta \). 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 \( e^{\beta_0} = A/N, 0 < A < \infty \).

\[
\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} \omega_i - \sum_{i=1}^{n} \omega_i \frac{e^{\beta_0 + (x_i - \bar{x})^T \beta}}{1 + e^{\beta_0 + (x_i - \bar{x})^T \beta}} - N \int \frac{e^{\beta_0 + (x - \bar{x})^T \beta}}{1 + e^{\beta_0 + (x - \bar{x})^T \beta}} dF_0(x)
\]

\[
= \sum_{i=1}^{n} \omega_i - \sum_{i=1}^{n} \omega_i \frac{AN^{-1} e^{(x_i - \bar{x})^T \beta}}{1 + AN^{-1} e^{(x_i - \bar{x})^T \beta}} - \int \frac{A e^{(x - \bar{x})^T \beta}}{1 + AN^{-1} e^{(x - \bar{x})^T \beta}} dF_0(x)
\]

\[
\leq \sum_{i=1}^{n} \omega_i - \frac{A \eta}{1 + A/N}.
\]

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

Lemma 7a. Under the same conditions, we will have \( \lim \sup_{N \to \infty} \|\hat{\beta}\| < \infty \).

Proof. Under the surrounded condition (Equation 5), there exists a \( \gamma \) such that

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

where \( \psi^T \psi = 1 \) and \( [(x - \bar{x})^T \psi]_+ \) means the positive part of \( [(x - \bar{x})^T \psi] \). We still let \( e^{\beta_0} = A/N \), then we have

\[
\begin{align*}
&l(\beta_0, 0; \Omega) - l(\beta_0, \beta; \Omega) \\
&\quad = - \left( \sum_{i=1}^{n} \omega_i + N \right) \log \left( 1 + e^{\beta_0} \right) + \sum_{i=1}^{n} \omega_i \log \left( 1 + e^{\beta_0 + (x_i - \bar{x})^T \beta} \right) \\
&\quad \quad + N \int \log \left( 1 + e^{\beta_0 + (x - \bar{x})^T \beta} \right) dF_0(x) \\
&\quad > - \left( \sum_{i=1}^{n} \omega_i + N \right) e^{\beta_0} + \frac{e^{\beta_0} N}{1 + e^{\beta_0}} \int_{(x - \bar{x})^T \beta \geq 0} (x - \bar{x})^T \beta dF_0(x) \\
&\quad \geq - \left( \sum_{i=1}^{n} \omega_i + N \right) \frac{A}{N} + \frac{AN}{N + A} \|\beta\| \gamma.
\end{align*}
\]

The first inequality above is taken from Lemma 4 in [19].

Equation (12) 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 \( \|\hat{\beta}\| \leq 2/\gamma \), with large enough \( N \).

Now, we prove the main theorem for infinitely weighted logistic regression:
Theorem 2. Let $x_1, \ldots, x_n$ and the weights $\Omega$ be fixed. If $F_0$ satisfies the surrounding condition at $\bar{x}$ and the tail conditions, then the maximizer $(\hat{\beta}_0, \hat{\beta})$ of $l(\beta_0, \beta; \Omega)$ satisfies

$$\lim_{N \to \infty} \frac{\int x e^{x^T \hat{\beta}} dF_0(x)}{\int e^{x^T \hat{\beta}} dF_0(x)} = \bar{x}. \quad (13)$$

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

$$- \sum_{i=1}^{n} \omega_i \frac{(x_i - \bar{x}) e^{\beta_0 + (x_i - \bar{x})^T \hat{\beta}}}{1 + e^{\beta_0 + (x_i - \bar{x})^T \hat{\beta}}} - N \int \frac{(x - \bar{x}) e^{\beta_0 + (x - \bar{x})^T \hat{\beta}}}{1 + e^{\beta_0 + (x - \bar{x})^T \hat{\beta}}} dF_0(x) = 0. \quad (14)$$

Divide Equation (14) by $N e^{\beta_0 - \bar{x}^T \hat{\beta}}$ to yield

$$\int \frac{(x - \bar{x}) e^{x^T \hat{\beta}}}{1 + e^{\beta_0 + (x - \bar{x})^T \hat{\beta}}} dF_0(x) = - \frac{1}{N} \sum_{i=1}^{n} \omega_i \frac{(x_i - \bar{x}) e^{x_i^T \hat{\beta}}}{1 + e^{\beta_0 + (x_i - \bar{x})^T \hat{\beta}}}. \quad (15)$$

Although the introduction of weights $\omega_i$ will lead to slower convergence than the unweighted logistic regression, the right side of Equation (15) still vanishes as $N \to \infty$ (because $\|\hat{\beta}\|$ is bounded by Lemma 7a). In the left side of Equation (15), $\int e^{x^T \hat{\beta}} \left[ 1 + e^{\beta_0 + (x - \bar{x})^T \hat{\beta}} \right]^{-1} dF_0(x)$ is at most $\int e^{x^T \hat{\beta}} dF_0(x)$ and is at least

$$\int e^{x^T \hat{\beta}} \left[ 1 + e^{\beta_0 + (x - \bar{x})^T \hat{\beta}} \right]^{-1} dF_0(x) \to \int e^{x^T \hat{\beta}} dF_0(x),$$

as $N \to \infty$ because $\hat{\beta}_0 \to -\infty$ (see Lemma 6a) and $\int e^{2x^T \hat{\beta}} dF_0(x) < \infty$ by the tail condition. Similarly we have

$$\int x e^{x^T \hat{\beta}} \left[ 1 + e^{\beta_0 + (x - \bar{x})^T \hat{\beta}} \right]^{-1} dF_0(x) \to \int x e^{x^T \hat{\beta}} dF_0(x).$$

Thus, Equation (15) simplifies to

$$\bar{x} \int e^{x^T \hat{\beta}} dF_0(x) - \int x e^{x^T \hat{\beta}} dF_0(x) = 0, \text{ as } N \to \infty. \quad (16)$$

Thus, Equation (13) holds. \hfill \square

Theorem 2 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.

3.2. Highly imbalanced penalized logistic regression. It is known that maximum likelihood estimation may fail with high-dimensional data or multiple highly-correlated variables [8, 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 consider the two common forms of penalty: the ridge and the lasso. For data as described in Section 2.3, we can give the objective function for ridge penalized logistic regression [11] as:

$$l(\beta_0, \beta) = \frac{1}{n + N} \left[ \sum_{i=1}^{n} \log \frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}} + \sum_{i=1}^{N} \log \frac{1}{1 + e^{\beta_0 + x_i^T \beta}} \right] - \frac{1}{2} \lambda \|\beta\|^2_2, \text{ where } \lambda > 0. \quad (17)$$
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 some simulations to explore the characteristics of penalized logistic regression in highly imbalanced data. We consider samples of different size and different levels of imbalance, where $P(X|Y=0) \sim N(0,1)$, 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 sample points) increases, the problem becomes more imbalanced. The coefficient estimates of standard logistic regression and ridge penalized logistic regression (penalty parameter is 0.1) are given in Table 1. The table suggests that, as $N \to \infty$, $N e^{\beta_0}$ converges to $n$ and $e^{\beta}$ converges to 1 in ridge penalized logistic regression, with this particular and arbitrary choice of penalty parameter. However, consider $X \sim \text{Uniform}(0,1)$ when $Y=0$. Again, we use $n=100$ points for $y=1$; 50 points are $x=0.5$ and the others are $x=2$. This setting fails the surrounded condition in Theorem 1. Table 2 shows the coefficient estimates of this simulation. We see that penalized logistic regression demonstrates shrinkage behavior, despite failing to satisfy the surrounded condition.$^2$

3.2.1. **Theoretical results.** In this section, we give results, following Owen, for penalized logistic regression in the infinitely imbalanced case. Considering the existence

# Table 1. Simulation A for infinitely imbalanced penalized logistic regression. $N$ observations in majority class ($Y=0$) following $N(0,1)$ and 100 observations in minority class with $Y=1$, $x=1$.

| $N$  | Logistic Regression | Ridge Penalized Logistic Regression |
|------|---------------------|-------------------------------------|
|      | $\beta$ $Ne^{\beta_0}$ | $\beta_0$ $e^{\beta_0}$ $Ne^{\beta_0}$ $\beta$ $e^\beta$ |
| 100  | 1.1215 41.7805       | -0.5247 0.5917 59.1750 0.6879 1.9896 |
| 1000 | 0.5656 65.3495       | -2.4591 0.0855 85.5127 0.2454 1.2782 |
| 10000| 0.5013 68.3830       | -4.6289 0.0098 97.6581 0.0450 1.0460 |
| 100000| 0.5007 68.7254      | -6.9102 0.0010 99.7516 0.0049 1.0050 |
| 1000000| 0.5001 68.7254   | -9.2106 0.0001 99.9750 0.0005 1.0005 |

# Table 2. Simulation B for infinitely imbalanced penalized logistic regression. $N$ observations in majority class ($Y=0$) following Uniform(0,1) and 100 observations in minority class (half of them with $Y=1$, $x=0.5$, the others with $Y=1$, $x=2$).

| $N$  | Logistic Regression | Ridge Penalized Logistic Regression |
|------|---------------------|-------------------------------------|
|      | $\beta$ $Ne^{\beta_0}$ | $\beta_0$ $e^{\beta_0}$ $Ne^{\beta_0}$ $\beta$ $e^\beta$ |
| 100  | 2.2347 16.2756       | -1.0602 0.3464 34.6374 1.2598 3.5246 |
| 1000 | 3.2033 8.4214        | -3.4516 0.0317 31.6947 1.6478 5.1958 |
| 10000| 4.6591 2.8035        | -4.9902 0.0068 68.0441 0.7112 2.0364 |
| 100000| 6.3475 0.7238        | -6.9521 0.0010 95.6659 0.0878 1.0918 |
| 1000000| 8.1866 0.1524   | -9.2148 0.0001 99.5517 0.0090 1.0090 |

$^2$The convergence effect does not become apparent until $N/n > 10$ (see line 1 and line 2 in Table 2)
of a unique solution in penalized logistic regression, we follow the previous argument for logistic regression considered in [23].

**Lemma 1.** For data as described in Section 2.3, 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.

**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 $(\beta_0, \beta)$. Other notations are same as those in Section 2.3 in our paper. 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 [21, Theorem 11.3]. Therefore, there exists a unit $p + 1$ dimensional vector $e$, such that $z_i^T e$ is negative for some $1 \leq i \leq n$ and $z_0^T e$ is positive for some $1 \leq i \leq N$ (see a observation in the design matrix). Let $A_1 = \{i| 1 \leq i \leq n, z_i^T e < 0\}$, $A_2 = \{i| 1 \leq i \leq n, z_i^T e \geq 0\}$, $A_3 = \{i| 1 \leq i \leq N, z_0^T e < 0\}$ and $A_4 = \{i| 1 \leq i \leq N, z_0^T e \geq 0\}$. Thus, considering $k \rightarrow \infty$,

$$
\begin{align*}
  l(B + (k + 1)e) - l(B + ke) &= \frac{1}{n + N} \sum_{i \in A_1} \log \left( \frac{e^{z_i^T(B + (k + 1)e)}}{1 + e^{z_i^T(B + (k + 1)e)}} \right) + \frac{1}{n + N} \sum_{i \in A_4} \log \left( \frac{1 + e^{z_i^T(B + ke)}}{1 + e^{z_i^T(B + (k + 1)e)}} \right) \\
  &\quad + \frac{1}{n + N} \sum_{i \in A_3} \log \left( \frac{1 + e^{z_i^T(B + ke)}}{1 + e^{z_i^T(B + (k + 1)e)}} \right) \\
  &\quad - \frac{1}{2} \lambda \parallel \beta + (k + 1)e \parallel_2^2 - \parallel \beta + ke \parallel_2^2.
\end{align*}
$$

(18)

(where $\bar{e}$ is $e$ excluding the first component) since the summation parts of $i \in A_2$ and $i \in A_3$ will vanish as $k \rightarrow \infty$. We also notice that when $k$ goes to infinity, the first term in Equation (18) will shrink to $\sum_{i \in A_1} e^{z_i^T e} < 0$, the second term will shrink to $-\sum_{i \in A_4} e^{z_i^T e} < 0$ and the third term is positive. Thus $l(B + (k + 1)e) - l(B + ke)$ will be negative for large $k$, so $l(B + ke)$ is a decreasing function in $k$. Therefore, $l(\beta_0, \beta)$ for fixed $x$ does not have a direction of recession and the existence of the solution follows from Theorem 27.1(d) in [21].

Assuming $S \cap F = \emptyset$, we could find a unit $p + 1$ dimensional vector $e$, such that \{\begin{align*}
  &z_i^T e \geq 0; 1 \geq i \geq n \} \text{ and } \{z_0^T e < 0; 1 \geq i \geq N \} \text{[21, Theorem 11.3 and Theorem 11.7].}
\end{align*}

Rewrite the likelihood function (17) as:

$$
\begin{align*}
l(B + ke) = \frac{1}{n + N} \left[ \sum_{i=1}^n \log \frac{e^{z_i^T(B + ke)}}{1 + e^{z_i^T(B + ke)}} + \sum_{i=1}^N \log \frac{1}{1 + e^{z_i^T(B + ke)}} \right] \\
- \frac{1}{2} \lambda \|\beta + ke\|_2^2.
\end{align*}
$$

(19)

Considering two numerical facts:

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

we know that the summation part in Equation (19) is increasing and approaching 0, as $k \rightarrow \infty$. Thus, term $-\frac{1}{2} \lambda \|\beta + ke\|_2^2$ dominates Equation (19) as $k \rightarrow \infty$, leads to the result that $l(B + ke)$ is a decreasing function for large enough $k$. Therefore,
Table 3. Infinitely imbalanced logistic regression shrinkage law.

| Fixture          | Logistic Regression | Ridge | Lasso |
|------------------|---------------------|-------|-------|
| $\beta_0$        | $-\infty$          | $-\infty$ | $-\infty$ |
| $N e^{\beta_0}$  | certain value, $k_1$ | $n$    | $n$    |
| $\beta$          | certain value, $k_2$ | $0$    | $0$    |

$l(\beta_0, \beta)$ for fixed $x$ does not have a direction of recession and the existence of the solution still follows from Theorem 27.1(d) in [21]. This proof shows the existence of the solution for ridge penalized logistic regression. Following lemmas in [27, Lemma 5] and [29, p.45], we know the extant solution for ridge penalized logistic regression is unique.

Through a similar proof process, Lemma 1 also holds with the lasso penalty. The only change is for handling the penalty term in Equation (18):

$$\lambda \sum_{j=1}^{p} |\beta_j + k + 1| - \lambda \sum_{j=1}^{p} |\beta_j + k|$$

$$= \lambda \sum_{j=1}^{p} \left( |\beta_j + k + 1| - |\beta_j + k| \right)$$

$$= \lambda p, \text{ when } k \to \infty, \text{ and } \beta_j \text{ is the } j\text{th element of } \beta,$$

which is still a positive number.

By repeating the previous numerical simulation (Table 1 and Table 2) in lasso penalized logistic regression, we obtain the shrinkage law (Table 3) when $N \to \infty$. The simulation results suggest that the intercept term $\beta_0$ shrinks to $n/N$ and $\beta$ shrinks to 0, when $N \to \infty$. We now prove this shrinkage law and further prove that the estimated parameter vector $\hat{\beta}$ only depends on the distribution $F_0$, the minority mean vector $\bar{x}$ 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 difference in the proof process between lasso and ridge will be pointed out in Section 3.2.2.

We again use the notation of Section 2.4. In order to directly show the result, we center lasso penalized logistic regression around the minority class mean vector $\bar{x} = \sum_{i=1}^{n} x_i/n$. Since in the infinitely imbalanced case $N \to \infty$, we also suppose that there is a good approximation for the conditional distribution of $X$ given $Y = 0$; denoted by $F_0$. Thus, the objective function for lasso penalized logistic regression [26] is written as

$$l(\beta_0, \beta) = \frac{1}{n+N} \left[ n \beta_0 - \sum_{i=1}^{n} \log \left( 1 + e^{\beta_0 + (x_i - \bar{x})^T \beta} \right) \right]$$

$$- N \int \log \left( 1 + e^{\beta_0 + (x - \bar{x})^T \beta} \right) dF_0(x) \right] - \lambda \sum_{j=1}^{p} |\beta_j|, \lambda > 0,$$

where $\beta_j$ is the $j$th element of $\beta$. We follow Owen’s proof again: Lemma 4 and Lemma 5 still hold for penalized logistic regression. The three changes in the proof process are for Lemma 6, Lemma 7 and the main theorem. Lemma 6b gives $e^{\hat{\beta}_0} \leq n/(N - n)$ and Lemma 7b gives $\sum_{j=1}^{p} |\hat{\beta}_j| \leq n/(N - n)\lambda$. Note that in
Lemma 6b and Lemma 7b, we do not require the surrounded condition, which makes the proof significantly different from Owen’s proof.

**Lemma 6b.** Let \( \hat{\beta}_0 \) and \( \hat{\beta} \) be the maximizers of the objective function (21). Then \( e^{\hat{\beta}_0} \leq n/(N - n) \).

**Proof.** Calculate the partial derivative with respect to \( \beta_0 \):

\[
\frac{\partial l(\hat{\beta}_0, \beta)}{\partial \beta_0} = \frac{n}{n + N} - \frac{1}{n + N} \sum_{i=1}^{n} e^{\hat{\beta}_0 + (x_i - \bar{x})^T \beta} \beta \leq \frac{N}{n + N} - \frac{N}{n + N} \int \frac{e^{\hat{\beta}_0 + (x - \bar{x})^T \beta}}{1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \beta}} dF_0(x)
\]

\[\leq \frac{n}{n + N} - \frac{N}{n + N} \int_{(x - \bar{x})^T \beta > 0} \frac{e^{\hat{\beta}_0 + (x - \bar{x})^T \beta}}{1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \beta}} dF_0(x)
\]

\[\leq \frac{n}{n + N} - \frac{N}{n + N} 1 + e^{\hat{\beta}_0} \int_{(x - \bar{x})^T \beta > 0} dF_0(x)
\]

\[\leq \frac{n}{n + N} - \frac{N}{n + N} 1 + e^{\hat{\beta}_0}.
\]

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

\( \Box \)

**Lemma 7b.** Let \( \hat{\beta}_0 \) and \( \hat{\beta} \) be the maximizers of the log-likelihood function (21). Then \( \lim_{n \to \infty} \| \hat{\beta} \| < \infty \).

**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

\[
l(\hat{\beta}_0, \hat{\beta}) - l(\hat{\beta}_0, 0) = \left[ \frac{1}{n + N} \left[ (n + N) \log \left( 1 + e^{\hat{\beta}_0} \right) - \sum_{i=1}^{n} \log \left( 1 + e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}} \right) \right.ight.
\]

\[\left. - \left. N \int \log \left( 1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}} \right) dF_0(x) \right] - \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \geq 0.
\]

(23)

Since \( \log \left( 1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}} \right) \geq 0 \) and

\[
\int \log \left( 1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}} \right) dF_0(x) \geq 0,
\]

Equation (23) leads to:

\[
\lambda \sum_{j=1}^{p} |\hat{\beta}_j| \leq \frac{1}{n + N} \left[ (n + N) \log \left( 1 + e^{\hat{\beta}_0} \right) - \sum_{i=1}^{n} \log \left( 1 + e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}} \right) \right.ight.
\]

\[\left. - \left. N \int \log \left( 1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}} \right) dF_0(x) \right].
\]
Then we have
\[ -N \int \log \left( 1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}} \right) dF_0(x) \leq \log \left( 1 + e^{\hat{\beta}_0} \right) \leq e^{\hat{\beta}_0} \leq \frac{n}{N - n}. \]

Thus, we have \( \sum_{j=1}^{p} |\hat{\beta}_j| \leq \frac{n}{(N - n)\lambda} \), by considering the triangle inequality, we know \( ||\hat{\beta}|| \) 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 3.** Let \( n > 1 \) and minority class vectors \( \{x_1, x_2, \ldots, x_n\} \) be fixed. Then the maximizer \( (\hat{\beta}_0, \hat{\beta}) \) of \( l \) given by Equation (21) have following shrinkage rules, \( e^{\hat{\beta}_0} \to \frac{n}{N} \) and \( \hat{\beta} \to 0 \), when \( N \to \infty \).

**Proof.** From Lemma 6b and Lemma 7b, we know \( e^{\hat{\beta}_0} \leq \frac{n}{N - n} \) and \( ||\hat{\beta}|| \) is bounded when \( N \to \infty \).

Further considering Lemma 7b, we have
\[ ||\hat{\beta}|| \leq \sum_{j=1}^{p} |\hat{\beta}_j| \leq \frac{n}{(N - n)\lambda} \text{ when } N \to \infty. \]  

Inequality (25) shows as \( N \to \infty \), we have \( \hat{\beta} = 0 \). Thus \( l(\hat{\beta}_0, \hat{\beta}) \) simplifies to
\[ l(\hat{\beta}_0, \hat{\beta}) = \frac{n}{n + \lambda} \hat{\beta}_0 - \log \left( 1 + e^{\hat{\beta}_0} \right). \]

Let the partial derivative of Equation (26) equal to 0 when \( N \to \infty \),
\[ \frac{\partial l(\hat{\beta}_0, \hat{\beta})}{\partial \hat{\beta}_0} = \frac{n}{n + N} - \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} = 0, \]
then we have \( e^{\hat{\beta}_0} = n/N \).

Since \( N e^{\hat{\beta}_0} \to n \) 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:
\[ \Pr(Y=1|X = x) \xrightarrow{N \to \infty} \frac{n}{n + N}, \]
\[ \Pr(Y=0|X = x) \xrightarrow{N \to \infty} \frac{N}{n + N}. \]

Note that Theorem 3 is a strong result that demonstrates \( \hat{\beta} = 0 \) in infinitely imbalanced penalized logistic regression. We are also interested in how \( \hat{\beta} \) shrinks 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 3.

Since the lasso penalized log-likelihood function is nondifferentiable at \( \beta = 0 \), we alternatively use the subgradient [18, p. 126] of convex Equation (21) with respect to \( \beta \). Setting the subgradient of Equation (21) to 0, we have
\[ -\frac{1}{n + \lambda} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}} + N \int \frac{(x - \bar{x}) e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}} dF_0(x) - \lambda s = 0, \]
where \( s \) is a \( p \) dimensional vector

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

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 \( v \) [18, Theorem 3.1.15]. Thus, because Equation (21) is a convex function, we have the following main theorem:

**Theorem 4.** Let \( n \geq 1 \) and minority class vectors \( \{x_1, \ldots, x_n\} \) be fixed and suppose that \( F_0 \) surrounds \( \bar{x} = \sum_{i=1}^n x_i/n \) as described. Then the maximizer \((\hat{\beta}_0, \hat{\beta})\) of \( t \) given by Equation (21) satisfies

\[
- \int (x - \bar{x}) e^{(x - \bar{x})^T \hat{\beta}} dF_0(x) = \frac{n + N}{n} \lambda s
\]

as \( N \to \infty \).

**Proof.** Setting the subgradient of Equation (21) to 0, we have

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

Dividing by \( N \) gives

\[
- \int \frac{(x - \bar{x}) e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}} dF_0(x) - \frac{n + N}{N} \lambda s = \frac{1}{N} \sum_{i=1}^n \frac{(x_i - \bar{x}) e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}}. \tag{29}
\]

As \( N \to \infty \), the right side of Equation (29) vanishes because \( \|\hat{\beta}\| \) is bounded as \( N \to \infty \) by Lemma 7b.

If we consider \( N \to \infty \), we have \( e^{\hat{\beta}_0} \to \frac{n}{N} \) and \( \hat{\beta} \to 0 \), yielding to \( e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}} \to \frac{n}{N} e^0 \to 0 \).

Thus Equation (29) yields

\[
- \int \frac{(x - \bar{x}) \frac{n}{N} e^{(x - \bar{x})^T \hat{\beta}}}{1} dF_0(x) = \frac{n + N}{N} \lambda s \tag{30}
\]

After simplification, Equation (28) holds. Notice that, \( s \) must converge to 0 due to the established Equation (28) when \( N \to \infty \).

The notional limiting operation \( N \to \infty \) in Equation (24) and Equation (29) is used to obtain analytic approximations. However the expression so obtained is an asymptotic form rather than a limit and gives useful results even for finite but large \( N \). Equation (28) shows the asymptotic solution of \( \beta \) depends only on \( \{\bar{x}, F_0(x), \frac{n}{N}\} \) when approaching infinite imbalance. We give some context and illustration of Theorem 4 with respect to the solution of penalized logistic regression with large finite \( N \) in the next section.

### 3.2.2. Changes in the proof for infinitely imbalanced ridge penalized logistic regression

The proof given above is specifically for the lasso penalty. Here, we briefly demonstrate several changes in the proof process required for infinitely imbalanced ridge penalized logistic regression.
In the proof of Lemma 7b, Equation (23) is changed to
\[
\ell(\hat{\beta}_0, 0) - \ell(\hat{\beta}_0, \hat{\beta}) = \frac{1}{n + N} \left[ -(n + N) \log \left(1 + e^{\hat{\beta}_0}\right) + \sum_{i=1}^{n} \log \left(1 + e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}\right) \right.
\]
\[
+ N \int \log \left(1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}\right) dF_0(x) \left. \right] + \frac{1}{2} \lambda \|\hat{\beta}\|^2
\]
\[
> \frac{1}{n + N} \left[ -(n + N) \frac{A}{N} + A \frac{\|\hat{\beta}\| \delta}{1 + A/N} \right] + \frac{1}{2} \lambda \|\hat{\beta}\|^2
\]
\[
> \frac{1}{n + N} \left[ -(n + N) \frac{A}{N} + A \frac{\|\hat{\beta}\| \delta}{1 + A/N} \right].
\]
(31)

The bounded result of \(\|\hat{\beta}\|\) could still be derived, by the way letting \(\ell(\hat{\beta}_0, 0) - \ell(\hat{\beta}_0, \hat{\beta}) > 0\).

In the proof of Theorem 3, Equation (25) is changed to
\[
\ell(\hat{\beta}_0, \hat{\beta}) \leq 0 - \frac{1}{2} \lambda \|\hat{\beta}\|^2, \; \lambda > 0 \text{ when } N \to \infty,
\]
and maximize \(\ell(\hat{\beta}_0, \hat{\beta})\) leading to \(\hat{\beta} = 0\). Then the results still hold.

In the proof of Theorem 4, we do not use subgradient method because the derivative of ridge penalty exist. Setting the partial derivative with respect to \(\beta\) of Equation (17) to 0, we have:
\[
- \sum_{i=1}^{n} \frac{(x_i - \bar{x}) e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (x_i - \bar{x})^T \hat{\beta}}} - N \int \frac{(x - \bar{x}) e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}}{1 + e^{\hat{\beta}_0 + (x - \bar{x})^T \hat{\beta}}} dF_0(x) - (n + N) \lambda \hat{\beta} = 0.
\]
(33)

After similarly simplify process we have the maximizer \((\hat{\beta}_0, \hat{\beta})\) of \(l\) given by Equation (17) satisfies
\[
- \int (x - \bar{x}) e^{(x - \bar{x})^T \hat{\beta}} dF_0(x) = \frac{n + N}{n} \lambda \hat{\beta}
\]
as \(N \to \infty\).

3.3. Numerical explanations for highly imbalanced 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 4.

Assume the probability density function of the majority class, \(f_0(x)\), 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 (28) in dimension \(r\) (here \(x_r\) refers to the \(r\)th column in the observation matrix):
\[
\int (x_r - \bar{x}_r) e^{\sum_{k=1}^{p} (x_k - \bar{x}_k) \beta_k} dF_0(x)
\]
\[
= \int (x_r - \bar{x}_r) e^{\sum_{k=1}^{p} (x_k - \bar{x}_k) \beta_k} f_0(x_1) \cdots f_0(x_p) dx_1 \cdots dx_p
\]
\[
= \int (x_r - \bar{x}_r) e^{(x_r - \bar{x}_r) \beta_r} f_0(x_r) dx_r \times \prod_{k \neq r} \int e^{(x_k - \bar{x}_k) \beta_k} f_0(x_k) dx_k
Consider that all \( \beta_r, r \in \{1, \cdots, p\} \) are shrunk to 0, thus all \( s_r \in [-1, 1] \). For all \( r \in \{1, \cdots, p\} \), we have

\[
(0 - \bar{x}_r) = -\frac{n + N}{n} \lambda s_r \left( \prod_{k=1}^p e^{\frac{\beta_k}{\sigma_k}} \right)^{-1} = -\frac{n + N}{n} \lambda s_r, \tag{36}
\]

so \( \bar{x}_r \) 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 \( \bar{x}_r \).

Alternatively, if \( \lambda > \frac{n}{n + N} \max \{\bar{x}_r\} \), all \( \beta_r \) will be shrunk to 0. When the imbalance level \( N/n \) is great, this condition is easily satisfied with respect to a fixed \( \lambda \).

Following is a simple simulation to demonstrate the consequence of Equation (36). We generate 100,000 five dimensional majority class data \((Y = 0)\) following \(N(\mu_0, \Sigma_0)\), where \(\mu_0 = 0\), \(\Sigma_0 = I\) and \(I\) is the identity matrix. 1,000 minority class data \((Y = 1)\) are generated which follows \(N(\mu_1, \Sigma_1)\), where \(\mu_1 = [1.9, 1.7, 1.5, 1.3, 1.1]\) and \(\Sigma_1 = 0.01I\). In this scenario, \(\frac{n}{n + N} \max \{\bar{x}_r\} \) equals 0.0188. In our simulation, the simulated vector \(\frac{n}{n + N} \bar{x}_r\) is \([0.0188, 0.0168, 0.0149, 0.0129, 0.0109]\).

Table 4 shows the coefficient estimates of lasso penalized logistic regression for different values of \( \lambda \). We find that the relationship between \( \bar{x}_r \) and \( \lambda \) determines whether the coefficient estimate \( \beta_r \) will be shrunk to 0 (all coefficient estimates shrink to 0 when \( \lambda > \frac{n}{n + N} \max \{\bar{x}_r\} \), see Table 4 Line 1).

Without the assumption of the majority class distribution \(F_0(x)\), if \( \hat{\beta} = 0 \), Equation (28) simplifies to

\[
\bar{x} - \int xdF_0(x) = \frac{n + N}{n} \lambda s, \text{ where } s \in [-1, 1]. \tag{37}
\]

\( \int xdF_0(x) \) is the population mean \( \mu_x \) of the majority class distribution, which is determined by data set. Thus we need \( \bar{x} \in [\mu_x - \frac{n+N}{n} \lambda, \mu_x + \frac{n+N}{n} \lambda] \) to force all \( \hat{\beta} \) shrink to 0, which is easy to satisfy with highly imbalanced data set.
Table 5. Coefficient of logistic regression and two clusters multinomial logistic regression. The left three columns are logistic regression and right four columns are multinomial logistic regression.

| Logistic Regression | Multinomial Logistic Regression |
|---------------------|---------------------------------|
| Coefficients Estimate Pr(>|z|) | Cluster Coefficients Estimate Pr(>|t|) |
| Intercept -5.7705 $< 2 \times 10^{-16}$ | c2 Intercept -7.6828 $< 2 \times 10^{-16}$ |
| $x_1$ 1.1384 $< 2 \times 10^{-16}$ | c3 Intercept -7.6828 $< 2 \times 10^{-16}$ |
| $x_2$ 1.1287 $< 2 \times 10^{-16}$ | c2 $x_1$ 0.0818 0.6045 |
|                              | c3 $x_1$ 2.2775 $< 2 \times 10^{-16}$ |
|                              | c2 $x_2$ 2.3532 $< 2 \times 10^{-16}$ |
|                              | c3 $x_2$ 0.0953 0.5412 |

4. Example: Cluster structure in the minority class and relabelling. In light of the results given so far, 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. To explore this problem we first relabel the minority class into two new (both “bad”) subclasses using a clustering algorithm, then use multinomial logistic regression to model the two bad classes along with the majority good class. This can be achieved by using standard clustering methods. Here, it is still possible to reason about the good/bad dichotomy, simply by considering the predictions $Pr(Y = 0 | X = x)$. This suggestion, to split the “bads” into subclasses, 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 [3].

To illustrate the issue of cluster structure in the minority class, we simulate a bivariate normal distribution example. We generate 10,000 sample points following $X \sim N(\mu_1, \Sigma_1)$ from the majority class ($Y = 0$). Then 50 points following $X \sim N(\mu_2, \Sigma_2)$ and 50 points following $X \sim N(\mu_3, \Sigma_3)$ are generated and combined as the single minority class. Here $\mu_1 = [0, 0]$, $\mu_2 = [0, 2]$, $\mu_3 = [2, 0]$, $\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_2 = \Sigma_3 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$. Clearly, this is rather contrived, but serves for the purposes of illustration. In Figure 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 classes ($c2, c3, Y = 1$) which are separated using $K$-means clustering (with $K = 2$). Table 5 shows the coefficients for the two different models. The first three columns are the logistic regression coefficients. By construction, the minority class is well-separated. The right four columns are multinomial logistic regression coefficients. Finally, we generate test data following the distribution described above, to assess out-of-sample predictive performance. The prediction AUC (area under the ROC curve [9]) of logistic regression is 0.918 which is lower than the AUC of multinomial logistic regression (0.954). Although the coefficient table shows that both intercept and slope terms are significant in logistic regression, all of these results suggest that logistic regression under-perform multinomial logistic regression when cluster structure is taken into account.
In this section, we conduct an analysis of the Freddie Mac data set introduced in Section 1. Table 8 in the Appendix provide a description of the variables in this data set. The theory in previous sections show that penalized or weighted logistic regression cannot avoid the consequence of highly imbalanced data. We try our relabelling approach on Freddie Mac data to test if and when a problem emerges in highly imbalanced logistic regression.

5.1. **Data preparation.** In the financial industry, the log transformation is frequently used to make highly skewed data less skewed and reduce the influence of outliers [1]. 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 [25, 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. For example, the variable “number of units” is dummy
Table 6. Experiment Procedure Time Table

| Training set year | 2000 | 2001 | ⋮  |
|-------------------|------|------|----|
| Default collection year | 2001 2002 | 2002 2003 | ⋮  |
| Testing set year | 2003 | 2004 | ⋮  |

coded into several indicator variables ("number.units2", "number.units3", "number.units4"). 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.

5.2. Model building procedure. Table 6 illustrates the experimental procedure. After data preparation, we use data from an individual year (e.g. 2000) as a training set to train two different models: “without relabelling” and “with relabelling”. Here, “without relabelling” refers to modelling a logistic regression based on the prepared data. Model output is the estimated posterior probability of a mortgage account defaulting, given its feature vector.

In “with relabelling” procedure, we use hierarchical clustering to split the minority class into two clusters. These two clusters provide two new default classes, which, along with the non-default class data, made up a new relabeled data. Then multinomial logistic regression is used to model the three classes in this relabeled data. The model output is still the posterior probability of defaulting by summarizing the posterior probability of two new default classes.

Although many kinds of unsupervised clustering methods are available, we choose hierarchical clustering here because it is fast and also because the algorithm is deterministic [17]. We manually set a threshold, which the minority class is clustered into two groups. We also considered K-means and K-medians clustering as well, since we used these in the simulation (Section 4) and also because they are a common choice for clustering. However, hierarchical clustering gave marginally better results.

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 in 2000 will be used to predict the data of each quarter in 2005 rather than 2003.

A two-year gap is a reasonable choice to balance this trade-off.

Finally, these models (i.e. “without relabelling” and “with relabelling”) are used to forecast default events in the four quarters in the test year. The performance assessment metric is AUC as this is a common industry measure of performance [25]. To explore modelling and performance issues over a long time horizon, this procedure is repeated over an eleven year period.
5.3. **Result and discussion.** Figure 3 shows the test set AUC for each of the methods, over the observation period and Table 9 in the Appendix shows AUC on the test sets and corresponding standard deviation calculated by DeLong’s method [7], repeated for 200 random bootstrap test sets and 200 stratified bootstrap [20] test sets in each quarter separately. Notice that the relabelling method does well to enhance performance in certain years, such as 2005, 2006 and 2012, but fails to achieve this in other years, 2007 and 2010. However, where relabelling’s performance is not as good as logistic regression, it is not a large difference. On the other hand, we observe a large improvement in certain periods.

Given the context of the data, there are a number of other issues (except imbalance) that make a simple performance comparison difficult. Any model’s efficacy may vary over time for a number of reasons and population drift (the distribution changes over time [13]) is a key concern, especially when considering a long time frame experiment. We want to provide some simple evidence in light of population drift, to enable us to better understand the behavior of logistic regression with imbalance.

Detecting evidence of population drift can be challenging. Here, we use a simple heuristic procedure to suggest which years (training set and corresponding test set) exhibit evidence of drift. Essentially, via the bootstrap, we compare a bias-corrected estimate of the distribution of the “resubstitution” AUC (the measure computed on the training data) with the corresponding distribution of the test set AUC. Differences in these distributions provide a crude indication for the presence of population drift. Specifically, bootstrapped training sets are used for model training and out of bootstrap training samples are used to calculate the debiased AUC of the training set. We repeat this process 100 times for calculating the distributions of the training set AUC. We also bootstrap test sets 100 times in each test set quarter to calculate the distributions of the test set AUC. Figure 7 in the Appendix shows the results. Different lines are the distributions of the training set AUC and the corresponding test set AUC in four test quarters respectively. This figure suggests that population drift occurs in a number of years.
We seek to quantify the population drift between these distributions. The K-S statistic [16] can calculate the distance \(D\) between two empirical distributions. Here, we use 100 debiased training set AUCs and all four quarters’ test set AUCs \((4 \times 100)\) as the samples of the two empirical distributions. Table 7 gives \(D\) for each year. Generally, we find that population drift happens very often in our experiment (unsurprising since the observation period spans the financial crash of 2008) but the performance of the “without relabelling” and “with relabelling” method with respect to population drift are different. The \(D\) statistics of “with relabelling” are consistently less than “without relabelling”, which may suggest that our approach helps alleviate the influence of population drift.

Table 10 in Appendix provides information about the default rate and the mean AUC difference (“with relabelling” - “without relabelling”) in each year. Figure 4 plots this mean AUC difference in each year with respect to different default rate in training year and population drift scenario. For illustration, we arbitrarily define the absence of population drift when \(D < 0.5\) for the “with relabelling” approach. First, note that higher default rates are associated with population drift. Second, for these high default scenario, there is no compelling evidence to prefer either model. However, in the low default rate scenario, ie with class imbalance, “with relabelling” performs better than “without relabelling” when population drift does not happen, while logistic regression is preferred when population drift is present. These post-hoc diagnostic results suggest the shortcomings of logistic regression with imbalanced data.

For this application, it is also meaningful to investigate the difference in characteristics between different default groups arising from relabelling. We look particularly at years 2002 and 2004 since those training years correspond to test years 2005 and 2007 respectively: 2005 was a year logistic regression performed poorly, relative to the relabelling approach, and in 2007, logistic regression was acceptable in performance, as can be read from Figure 3. The distribution of key variables are shown in Figures 5 and 6. For 2002, we see clear differences in distribution of DTI and LTV for the two default clusters: in particular cluster 1 represents a group with much higher LTV but interestingly a similar range of FICO score. For 2004, we also see clear differences in LTV. However one of the default groups has generally lower LTV than the non-default group which is contrary to conventional credit-scoring wisdom: expert knowledge [5] in mortgage markets shows higher LTV brings higher risk of default. This may therefore not express a genuinely useful clustering for modelling default risk.
**Figure 4.** Scatter plot of mean AUC difference in test year v.s. default rate in training year.

**Figure 5.** Boxplots for Score, DTI, UPB, LTV and OIR in the year 2002. The \( p \)-values in the plot are calculated through Student’s \( t \)-test \([24]\) between “Default 1” and “Default 2” in each variable.
Figure 6. Boxplots for Score, DTI, UPB, LTV and OIR in the year 2004. The p-values in the plot are calculated through Student’s t-test [24] between “Default 1” and “Default 2” in each variable.

6. Conclusion. We have explored the impact of class imbalance for the logistic regression methods used for credit scoring. 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 dataset. This tells us that the infinite imbalance problem is fundamental for logistic regression. The challenge of severe model imbalance is profound. Since logistic regression remains the workhorse of credit scoring and many other applications [12, 31], the issue of imbalance certainly demands more attention, and the development of enhanced tools.

To mitigate the problem of highly imbalanced logistic regression, we made use of a relabelling approach. Results for the credit scoring example are promising and indicate that the relabelling method works well in conditions of low population drift and high imbalance. This suggests proposing the relabelling approach itself as a practical solution to the problems discussed here. Further work is required to determine the correct clustering methodology and to identify more precisely the circumstances when the clustering of the rare class gives benefit for the classification problem. This will be the subject of future research.

Appendix. This section provides figures and tables which mentioned in our paper.

Acknowledgments. We are grateful to Dr Heather Battey for insightful discussions during the development of this paper.
# Table 8. 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.                                                             |
Table 9. AUC and standard deviation in Freddie Mac experiment

| Time       | With Relabelling | Without Relabelling |
|------------|------------------|---------------------|
|            | AUC              |                     |
|            | DeLong           | Bootstrap           |
|            | Stratified       |                     |
|            | AUC              | DeLong             |
|            |                  | Bootstrap          |
|            |                  | Stratified         |
| 2003 Q1    | 0.879            | 0.033              |
|            | 0.035            | 0.032              |
| 2003 Q2    | 0.880            | 0.025              |
|            | 0.024            | 0.024              |
| 2003 Q3    | 0.839            | 0.035              |
|            | 0.033            | 0.031              |
| 2003 Q4    | 0.872            | 0.025              |
|            | 0.025            | 0.025              |
| 2004 Q1    | 0.808            | 0.042              |
|            | 0.041            | 0.041              |
| 2004 Q2    | 0.804            | 0.053              |
|            | 0.056            | 0.053              |
| 2004 Q3    | 0.636            | 0.067              |
|            | 0.063            | 0.067              |
| 2004 Q4    | 0.806            | 0.046              |
|            | 0.045            | 0.046              |
| 2005 Q1    | 0.865            | 0.025              |
|            | 0.024            | 0.027              |
| 2005 Q2    | 0.841            | 0.026              |
|            | 0.025            | 0.026              |
| 2005 Q3    | 0.849            | 0.021              |
|            | 0.020            | 0.022              |
| 2005 Q4    | 0.814            | 0.022              |
|            | 0.022            | 0.021              |
| 2006 Q1    | 0.817            | 0.017              |
|            | 0.016            | 0.016              |
| 2006 Q2    | 0.803            | 0.015              |
|            | 0.016            | 0.016              |
| 2006 Q3    | 0.789            | 0.016              |
|            | 0.015            | 0.015              |
| 2006 Q4    | 0.776            | 0.012              |
|            | 0.012            | 0.012              |
| 2007 Q1    | 0.697            | 0.013              |
|            | 0.013            | 0.014              |
| 2007 Q2    | 0.704            | 0.010              |
|            | 0.010            | 0.010              |
| 2007 Q3    | 0.725            | 0.008              |
|            | 0.008            | 0.008              |
| 2007 Q4    | 0.720            | 0.006              |
|            | 0.006            | 0.007              |
| 2008 Q1    | 0.837            | 0.004              |
|            | 0.004            | 0.004              |
| 2008 Q2    | 0.832            | 0.005              |
|            | 0.005            | 0.005              |
| 2008 Q3    | 0.830            | 0.006              |
|            | 0.006            | 0.007              |
| 2008 Q4    | 0.857            | 0.008              |
|            | 0.008            | 0.008              |
| 2009 Q1    | 0.804            | 0.024              |
|            | 0.023            | 0.022              |
| 2009 Q2    | 0.811            | 0.018              |
|            | 0.019            | 0.017              |
| 2009 Q3    | 0.757            | 0.013              |
|            | 0.013            | 0.013              |
| 2009 Q4    | 0.738            | 0.023              |
|            | 0.025            | 0.022              |
| 2010 Q1    | 0.825            | 0.033              |
|            | 0.034            | 0.032              |
| 2010 Q2    | 0.793            | 0.038              |
|            | 0.039            | 0.037              |
| 2010 Q3    | 0.826            | 0.034              |
|            | 0.031            | 0.034              |
| 2010 Q4    | 0.769            | 0.036              |
|            | 0.038            | 0.034              |
| 2011 Q1    | 0.789            | 0.039              |
|            | 0.037            | 0.035              |
| 2011 Q2    | 0.780            | 0.042              |
|            | 0.041            | 0.039              |
| 2011 Q3    | 0.740            | 0.048              |
|            | 0.048            | 0.044              |
| 2011 Q4    | 0.782            | 0.050              |
|            | 0.043            | 0.047              |
| 2012 Q1    | 0.861            | 0.034              |
|            | 0.032            | 0.033              |
| 2012 Q2    | 0.776            | 0.043              |
|            | 0.045            | 0.038              |
| 2012 Q3    | 0.771            | 0.045              |
|            | 0.043            | 0.045              |
| 2012 Q4    | 0.771            | 0.038              |
|            | 0.036            | 0.034              |
| 2013 Q1    | 0.769            | 0.039              |
|            | 0.037            | 0.039              |
| 2013 Q2    | 0.738            | 0.029              |
|            | 0.028            | 0.029              |
| 2013 Q3    | 0.730            | 0.040              |
|            | 0.039            | 0.041              |
| 2013 Q4    | 0.754            | 0.033              |
|            | 0.031            | 0.032              |
Table 10. Mean AUC difference (Hierarchical - Logistic) in each year.

| Train year | Training Default rate | Test year | Test Default rate | AUC difference |
|------------|------------------------|-----------|-------------------|----------------|
| 2000       | 0.41%                  | 2003      | 0.06%             | 0.0057         |
| 2001       | 0.20%                  | 2004      | 0.07%             | 0.0063         |
| 2002       | 0.10%                  | 2005      | 0.18%             | 0.0578         |
| 2003       | 0.06%                  | 2006      | 0.89%             | 0.0119         |
| 2004       | 0.07%                  | 2007      | 4.26%             | -0.0133        |
| 2005       | 0.18%                  | 2008      | 3.15%             | -0.0005        |
| 2006       | 0.89%                  | 2009      | 0.30%             | -0.0003        |
| 2007       | 4.26%                  | 2010      | 0.00%             | -0.0055        |
| 2008       | 3.15%                  | 2011      | 0.08%             | 0.0055         |
| 2009       | 0.30%                  | 2012      | 0.06%             | -0.0041        |
| 2010       | 0.09%                  | 2013      | 0.10%             | -0.0011        |

Figure 7. Density plot of AUC on training year and four test quarters respectively, the left side is “with relabelling” method and the right side is “without relabelling”.

![Density plots for different test years with and without relabelling](image-url)
REFERENCES

[1] E. I. Altman and G. Sabato, Modelling credit risk for smes: Evidence from the US market, *Abacus*, 43 (2007), 322–357.

[2] G. E. Batista, R. C. Prati and M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, *ACM SIGKDD Explorations Newsletter*, 6 (2004), 20–29.

[3] C. Bravo, L. C. Thomas and R. Weber, Improving credit scoring by differentiating defaulter behaviour, *Journal of the Operational Research Society*, 66 (2015), 771–781.

[4] 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 (2002), 321–357.

[5] T. M. Clauretie, A note on mortgage risk: Default vs. loss rates, *Real Estate Economics*, 18 (1990), 202–206.

[6] Cornell Law School, Definition of default, date of default, and requirement of notice of default, URL https://www.law.cornell.edu/cfr/text/24/203.467.

[7] E. R. DeLong and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics*, 44 (1988), 837–845.

[8] B. Efron and T. Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Institute of Mathematical Statistics (IMS) Monographs, 5. Cambridge University Press, New York, 2016.

[9] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters*, 27 (2006), 861–874.

[10] D. J. Hand, Reject inference in credit operations, *Credit Risk Modeling: Design and Application*, 181–190.

[11] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12 (1970), 55–67.

[12] G. King and L. Zeng, Logistic regression in rare events data, *Political analysis*, 9 (2001), 137–163.

[13] G. Krempl and V. Hofer, Classification in presence of drift and latency, in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, IEEE, 2011, 596–603.

[14] J. Laurikkala, Improving identification of difficult small classes by balancing class distribution, *Artificial Intelligence in Medicine*, 2101 (2001), 63–66.

[15] 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 (2009), 539–550.

[16] F. J. Massey Jr, The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American Statistical Association*, 46 (1951), 68–78.

[17] F. Murtagh and P. Contreras, Algorithms for hierarchical clustering: An overview, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2 (2012), 86–97.

[18] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Applied Optimization, 87. Kluwer Academic Publishers, Boston, MA, 2004.

[19] A. B. Owen, Infinitely imbalanced logistic regression, *Journal of Machine Learning Research*, 8 (2007), 761–773.

[20] O. Pons, Bootstrap of means under stratified sampling, *Electronic Journal of Statistics*, 1 (2007), 381–391.

[21] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J. 1970.

[22] C. Seiffert, T. M. Khoshgolftaar, J. Van Hulse and A. Napolitano, Resampling or reweighting: A comparison of boosting implementations, in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, 1 (2008), 445–451.

[23] 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 (1981), 310–313.

[24] Student, The probable error of a mean, *Biometrika*, 6 (1908), 1–25.

[25] L. C. Thomas, *Consumer Credit Models: Pricing, Profit and Portfolios*, Oxford, 2009.

[26] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, 58 (1996), 267–288.

[27] R. Tibshirani, The lasso problem and uniqueness, *Electronic Journal of Statistics*, 7 (2013), 1456–1490.
[28] H. Wang, Q. Xu and L. Zhou, Large unbalanced credit scoring using lasso-logistic regression ensemble, *PLoS ONE*, **10** (2015), e0117844.

[29] V. Wieringen and Wessel, Lecture notes on ridge regression, arXiv preprint, arXiv:1509.09169.

[30] G. Zeng, On the existence of maximum likelihood estimates for weighted logistic regression, *Communications in Statistics-Theory and Methods*, **46** (2017), 11194–11203.

[31] M. Zhu, W. Su and H. A. Chipman, Lago: A computationally efficient approach for statistical detection, *Technometrics*, **48** (2006), 193–205.

E-mail address: y.li16@imperial.ac.uk
E-mail address: anthony-graham.bellotti@nottingham.edu.cn
E-mail address: n.adams@imperial.ac.uk