Pertanika J. Sci. & Technol. 29 (1): 181 - 197 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Predictive Performance of Logistic Regression for Imbalanced Data with Categorical Covariate

Hezlin Aryani Abd Rahman[1]*, Yap Bee Wah[1,2] and Ong Seng Huat[3]

[1]Centre of Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]Advanced Analytics Engineering Centre, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[3]Department of Actuarial Science and Applied Statistics, UCSI University, 56000, Kuala Lumpur, Malaysia

ABSTRACT

Logistic regression is often used for the classification of a binary categorical dependent variable using various types of covariates (continuous or categorical). Imbalanced data will lead to biased parameter estimates and classification performance of the logistic regression model. Imbalanced data occurs when the number of cases in one category of the binary dependent variable is very much smaller than the other category. This simulation study investigates the effect of imbalanced data measured by imbalanced ratio on the parameter estimate of the binary logistic regression with a categorical covariate. Datasets were simulated with controlled different percentages of imbalance ratio (IR), from 1% to 50%, and for various sample sizes. The simulated datasets were then modeled using binary logistic regression. The bias in the estimates was measured using MSE (Mean Square Error). The simulation results provided evidence that the effect of imbalance ratio on the parameter estimate of the covariate decreased as sample size increased. The bias of the estimates depended on sample size whereby for sample size 100, 500, 1000 – 2000 and 2500 – 3500, the estimates were biased for IR below 30%, 10%, 5% and 2% respectively. Results also showed that parameter estimates were all biased at IR 1% for all sample size. An application using a real dataset supported the simulation results.

Keywords: Categorical covariate, imbalanced data, logistic regression, parameter estimates, predictive analytics, simulation

ARTICLE INFO

Article history:
Received: 13 February 2020
Accepted: 27 July 2020
Published: 22 January 2021

DOI: https://doi.org/10.47836/pjst.29.1.10

E-mail addresses:
hezlin@tmsk.uitm.edu.my (Hezlin Aryani Abd Rahman)
beewah@tmsk.uitm.edu.my (Yap Bee Wah)
ongsh@ucsiuniversity.edu.my (Ong Seng Huat)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Imbalanced data are a condition where the dependent variable contains one class which has more observations than the other. Imbalanced data will have prominent effect on the classification performance of classifiers such as logistic regression, decision trees, support vector machine (SVM) and artificial neural network (ANN). Imbalanced data also affects the classification "power" of various classifiers. The effect of imbalanced data has been reported by researchers through the application of real data sets (Blagus & Lusa, 2010; Longadge et al., 2013; Ramyachitra & Manikandan, 2014).

Logistic regression (LR) is frequently used in predictive modeling as a benchmark model when other classifiers' performances were evaluated. It is a conventional statistical model used widely in business, engineering, and social science research (Hamid, 2016 ; Hamid et al., 2018; Ahmad et al., 2011; Shariff et al., 2016; Yap et al., 2014), and medical and healthcare studies (Longadge et al., 2013; Mena & Gonzalez, 2006; Oztekin et al., 2009; Pourahmad et al., 2011; Rothstein, 2015; Roumani et al., 2013; Srinivasan & Arunasalam, 2013; Uyar et al., 2010). However, the presence of imbalanced data challenges LR's ability to classify, whereby majority of classifiers normally focus in the prediction without consideration on the relative distribution between the classes (Dong et al., 2014). Normally, when imbalance data are present, classification results from standard classifiers are biased towards the majority class. As a result, if the event of interest is the minority class, the sensitivity of the classifier will be zero and the specificity will be 100%. The real dataset in reality often suffers from some imbalance problem (Goel et al., 2013) and the minority class is often misclassified (Chawla et al., 2004; He & Garcia, 2009; Weiss & Provost, 2003). Thus, whenever imbalance problem is found in healthcare and medical datasets, the credibility of the models generated by the classifiers are often misleading.

Imbalanced problem affects standard classifiers (Chawla, 2003; Cohen et al., 2006; Galar et al., 2011) and logistic regression based on application to real datasets studies (Blagus & Lusa, 2010; Burez & Van den Poel, 2009; Mena & Gonzalez, 2006; Van Hulse et al., 2007). In our previous study, we performed simulation to study the impact of imbalanced ratio (IR) on LR parameter (β) estimates and the odds ratio (e^{β}) of the LR model using a continuous covariate (Rahman & Yap, 2016). The results provided enough evidence to conclude that extreme imbalanced ratio (IR = 1%, 2%, 5%) and small sample size have more serious effect on parameter estimates of LR model. Imbalanced ratio is the ratio of the number of cases in minority class to the majority class. For example, if the response variable is the presence of cancer and has two categories Cancer or No Cancer the imbalanced ratio is n_1/n_0, where n_1 is the number of patients diagnosed with cancer while n_0 is the number of patients who do not have cancer.

The effect of imbalanced data on the performance of the classifiers can be determined through simulation studies. In addition, the various types (categorical or continuous) of variables in a set of data might show different effects. In this simulation study, we focus on the logistic regression model, a useful statistical model for classification problem and investigate the imbalanced effects on the parameter estimate of the model with a single categorical covariate.

The aim of this study was to determine the effects of different IR on the logistic regression parameter estimate via simulation and an application to real dataset. The results of this study will guide practitioners on the severity of bias in estimates as a result imbalanced data.

MATERIALS AND METHOD

Review on Methods

Machine learning techniques i.e. LR, DT, ANN and SVM, may have great classification performance if it involves a balanced data. However, these techniques performs poorly when imbalanced problem arises (Anand et al., 2010).

Most studies concluded that there was an effect of IR towards the performance of standard classifiers (Rahman et al., 2012; Chawla, 2003; Lemnaru et al., 2012; Mena & Gonzalez, 2006; Prati et al., 2014; Van Hulse et al., 2007; Yap et al., 2014). A study by Mena and Gonzalez (2006) introduced a 3-step algorithm using simple LR called REMED (Rule Extraction Medical Diagnosis) which enabled users to select attributes for the model and improved the accuracy of the model by adjusting the percentage of the partition. Although REMED's algorithm claimed to improve the prediction accuracy, it is limited to medical diagnostics. Lemnaru et al. (2012) reported that IR, size and complexity of the dataset affects the predictive performance of different classifiers [(k-nearest neighbor (KNN), C4.5, SVM, multi-layered perceptron (MLP), Naïve Bayes (NB), and Adaboost (AB)]. In their extensive study, the IR was categorized into three categories (balance, small, large), four categories of dataset size (very small, small, medium, and large) and four categories of complexity of the dataset (small, medium, large and very large). They concluded that the performance of the classifiers was lower when the IR was high. Another extensive experiment performed by Van Hulse et al. (2007), using different sampling strategies (random oversampling (ROS), random undersampling (RUS), one-sided selection (OSS), cluster-based oversampling (CBOS), Wilson's editing (WE), SMOTE (SM), and borderline-SMOTE (BSM) on different classifiers (NB, DT C4.5, LR, random forest (RF), and SVM) on 35 real datasets with different ratio of imbalance (1.33% - 34.90%), concluded that sampling strategy improved the performance of the chosen classifiers. However, their study also concluded that there was no one universal sampling strategy that worked best for all classifiers. Chawla (2003) experimented on five real datasets using C4.5 as the classifier and reported that their synthetic sampling method, SMOTE, improved the performance of the classifier better than other sampling strategies. He also concluded that RUS was better than ROS with replication. Prati et al. (2014) also experimented on 22 real datasets with different IR on different classifiers (C4.5, C4.5Rules, CN2 and RIPPER, Back-propagation Neural Network, NB and SVM) and by using different sampling strategies (ROS, SMOTE, borderline-SMOTE, AdaSyn, and MetaCost). They concluded that in terms of accuracy (AUC), the rule-based algorithm (C4.5Rule, RIPPER) was the most affected while Support Vector Machine (SVM) was least affected by imbalanced data. However, the authors also stated that severe imbalanced class distributions would have a strong influence on SVM and any classifier for that matter.

Thus, in a nutshell, we can conclude that the predictive performance of different standard classifiers compared by the mentioned studies arrived at different conclusions as to which classifier and sampling strategies performed better (Blagus & Lusa, 2010; Lemnaru et al., 2012; Mena & Gonzalez, 2006; Prati et al., 2014; Sarmanova & Albayrak, 2013).

In classification and predictive analytics, LR is normally considered a very informative classifier as it provides important information about the effect of an independent variable (IV) on the dependent variable (DV) through the odds ratio (Hosmer & Lemeshow, 2004). However, the presence of imbalanced problem hinders the predictive "power" of LR (Wallace & Dahabreh, 2012). Blagus & Lusa (2010) performed a simulation study to evaluate the performance of six types of classifiers (ANN, Linear Discriminant Analysis (LDA), RF, SVM and penalized logistic regression (PLR)) on highly imbalanced data. However, their results showed that the PLR with ROS method, failed to remove the biasness towards the majority class.

A simulation study by Hamid et al. (2015) discovered that when sample size was large (at least 500) the parameter estimates accuracy for LR improved. In addition, the estimation of LR parameters is severely affected by types of covariates; either continuous, categorical, or count data. Simulation studies, usually, enables us to provide a more conclusive evidence on the effect of IR, as the simulated datasets were mold perfectly to cater specific problem types. In our previous study (Rahman & Yap, 2016), our results were consistent with the study by Hamid et al., 2015, which reported that the performance of LR is affected by sample size. However, Hamid et al. (2015) did not consider imbalanced data. Simulation studies are important to obtain empirical evidence on the impacts of IR on the estimate of logistic regression parameter, β -value and the odds ratio of the LR model.

Simulation Methods

This study considered a simple binary logistic regression (LR). In the LR model, two unknown parameters, β_0 and β_1 , are estimated using the maximum likelihood method. Assuming observations to be independent, the likelihood function is given by the following Equation 1 (Hosmer & Lemeshow, 2004):

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{1}$$

To estimate β_0 and β_1, the maximization of the likelihood function is required. Therefore, the maximization of the natural logarithm of the likelihood function is denoted by the following Equation 2:

$$\log[L(\beta_{0,}\beta_1)] = \sum_{i=1}^{n} \{y_i \log[\pi(x_i)] + (1 - y_i)\log[1 - \pi(x_i)]\} \tag{2}$$

By referring to the simple LR Equation 1, the Equation 2 can also be expressed as Equation 3 (Hosmer & Lemeshow, 2004):

$$\log[L(\beta_{0,}\beta_1)] = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \log[1 + \exp(\beta_0 + \beta_1 x_i)] \tag{3}$$

By differentiating $\log[L(\beta_{0,}\beta_1)]$ with respect to β_0 and β_1 and setting the resulting Equation 4 to zero, we can obtain β that maximizes Equation 3.

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \tag{4}$$

The maximum likelihood estimates of β_0 and β_1, are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and is obtained using Newton -Raphson method. The probability that the event occurs, $\pi(x_i)$ for case *i* is then obtained as Equation 5:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} \tag{5}$$

In addition, $\hat{\pi}(x_i)$ is also known as fitted or predicted value and the sum of $\hat{\pi}(x_i)$ is equal to the sum of the observed values as in Equation 6:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{6}$$

The final estimated simple logistic regression model is written as Equation 7:

$$\log\left[\frac{\hat{\pi}(x_i)}{1 - \hat{\pi}(x_i)}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{7}$$

We assessed the effect of various percentages of IR and sample size on estimation of the parameter coefficient, β for binary LR model with one categorical independent variable. The estimate, $\hat{\beta}_1$ were compared with the true β_1 value. The simulations were performed using R-Studio. The value of the regression coefficient (β_0) for the logistic model was set at 2.08 which gave a significant odds ratio (*OR*) of 8.004 for *X* ($OR = e^{2.08} = 8.004$). The R code developed for this simulation is available at https://github.com/hezlin/simulationx1cat.git. It is also provided in the Appendix.

Odds-ratio provide important information of the effect of the covariate on the event (dependent variable). Given a binary Y(1=Died, 0=Survived) and a categorical covariate X(Hypertension-HPT) with two categories (1=Yes and 0=No), an odds-ratio of 1 will indicate both patients with or without HPT has equal chance of Y=1 (Died). Meanwhile, an odds-ratio greater than 1 will indicate that patients with HPT are more likely to die, and if odds-ratio is less than 1, patients with no HPT are more likely to die.

Eight imbalance ratios were considered for this simulation study: 1%, 2%, 5%, 10%, 20%, 30%, 40%, and 50%. Imbalance ratio (IR) is the percentage of occurrence of minority class between the two predictor classes. For example, in this simulation, if we generated a dataset N=100, if the IR = 1% means that 1 out of 100 has y=1 and the rest 99 out of 100 has y=0. The IR 5% or less represents high IR in the response variable. However, due to the complexity of generating the simulated dataset, especially for fixing definite percentages of IR, the simulation model required β_0 values to be flexible for different IR ratio. Thus, the full LR model used for this study is denoted as Equation 8:

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_{0k} + 2.08x_{ik} \qquad (8)$$

where β is determined by the IR and is not fixed at one value.

The data for the covariate (*X*) considered in this study were generated using a binomial distribution, *Bin (n= sample size, p=0.5).* We considered sample size of 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000. This simulation study involved 10,000 replications. The simulation algorithm is as follows:

Step 1: Generate random data for the categorical covariate X*, for* sample size, *n* and imbalance ratio, *IR.*

Step 2: Set β at 2.08 and obtain $f(x) = \beta_{0k} + 2.08x_{1k}$, where k = 1, 2, … 10,000. β_{0k} is not fixed to create a fix percentage of imbalance accordingly, whereby the confidence interval of β_{0k} is set within the range of (-2, 10).

Step 3: Fit binary logistic regression to the generated data in Step 2.

Step 4: Obtain the parameter estimate, $\hat{\beta}$.

Step 5: Repeat Steps 1-4 for 10,000 replications.

Step 6: Calculate the MSE where $MSE = \frac{\sum_{i=1}^{10000}(\beta - \hat{\beta})^2}{10,000}$

Repeat Steps 1 – 6 for different sample size and imbalanced ratio.

RESULTS AND DISCUSSION

Simulation Results

Table 1 presents the simulation results for the LR parameter estimates for various sample sizes and IR. %.). The effect of IR was reduced when sample size increased. The results showed that the estimates for β_0 *and* β_1 were very far from the true parameter values for smaller sample size (n=100) and for IR 1%, 2%, 5%, 10%, 20% and 30%. The bias in estimate was clearly seen for IR 20% or less for n=500. Meanwhile, for n=1000, the bias was seen for IR 10% or less. However, the effect of IR was less for sample size more than 3000 and above was only affected by IR of 1% and 2%. Table 2 summarizes the findings.

Figure 1 presents the effect of sample size and IR on the parameter estimate values. It clearly shows the parameter estimates was biased for IR 30% and below for n=100. The Figure 1 also shows that for all sample sizes, the estimate was close to the true parameter values at IR=30% and above. In Figure 2, we focused on high IR, 1% to 10% and omitting 20% to 50% so that visualization of the effect is clearer. The Figure 2 shows threshold of effect of IR decreases as sample size increases. For example, estimates are biased for n=500 for IR 5% and below, while for n=1000, estimates are biased at IR % and below. When estimates are biased the MSE will be larger. Figure 3 illustrates the effect of IR and sample size through the MSE and Figure 4 further emphasizes results in Figure 3 by focusing on the IR of 1% to 10%, by omitting the 20% to 50% ratios. In Figure 3, the effect of imbalance is less (lower MSE) at IR=30%, similar to the illustration in Figure 1. Further focusing on highly imbalanced ratios, Figure 4 illustrates that the MSE values are the largest for small sample sizes (n=100 and n=500).

Figures 5 and 6 illustrate the effects of imbalanced using a clustered boxplot. As shown in Figure 5, the effect of imbalanced data is obvious for sample size n=500 (IR=1% and 2%) and n=1000 (IR=1%). In Figure 6, we omit the imbalanced ratio 1% and 2%, and now there are no huge spikes in the boxplots. Figures 5 and 6 clearly showed the effect of IR for various sample sizes, whereby the patterns show that the effect of IR on the bias of parameter estimates depend on sample size. The estimates get closer to the true value when the sample size and IR increases. The dispersion (standard deviation) of $\hat{\beta}_1$ also improves as sample size and IR increases.

Table 1

Parameter estimates for categorical covariate (β=2.08) for model with different n and IR

Size	IR	$\hat{\beta}_1$	C.I (lower)	C.I (upper)	β_0	$\hat{\beta}_1$	Size	IR		C.I (lower)	C.I (upper)	β_0	$\hat{\beta}_1$	Size	IR	$\hat{\beta}_1$	C.I (lower)	C.I (upper)	β_0	$\hat{\beta}_0$
100	**1**	**13.8056**	13.5891	14.0222	-6.5778	-19.6347	**2000**	**1**	**3.5815**	3.4914	3.6717	-6.1199	-7.5977	**4000**	**1**	**2.3202**	2.2895	2.3508	-6.1038	-6.3349
	2	**14.1244**	13.9647	14.2841	-5.6621	-17.6550		**2**	**2.3260**	2.2945	2.3574	-5.3962	-5.6309		**2**	2.1316	2.1241	2.1391	-5.3910	-5.4365
	5	**10.3026**	10.1344	10.4707	-4.5099	-12.6521		**5**	2.1204	2.1140	2.1269	-4.4212	-4.4580		**5**	2.0982	2.0938	2.1026	-4.4184	-4.4357
	10	**6.0332**	5.8896	6.1768	-3.6729	-7.5887		**10**	2.0970	2.0927	2.1013	-3.6369	-3.6524		**10**	2.0874	2.0844	2.0905	-3.6363	-3.6427
	20	**2.8190**	2.7535	2.8845	-2.7598	-3.4784		**20**	2.0864	2.0835	2.0894	-2.7453	-2.7514		**20**	2.0850	2.0830	2.0871	-2.7445	-2.7494
	30	**2.2406**	2.2164	2.2648	-2.1098	-2.2586		**30**	2.0836	2.0812	2.0860	-2.1130	-2.1151		**30**	2.0823	2.0806	2.0839	-2.1116	-2.1138
	40	2.1501	2.1394	2.1607	-1.5605	-1.6111		**40**	2.0837	2.0816	2.0858	-1.5612	-1.5645		**40**	2.0810	2.0796	2.0825	-1.5623	-1.5631
	50	2.1370	2.1275	2.1464	-1.0423	-1.0704		**50**	2.0822	2.0802	2.0842	-1.0402	-1.0411		**50**	2.0801	2.0787	2.0815	-1.0402	-1.0399
500	**1**	**10.2400**	10.0788	10.4011	-6.1980	-14.2681	**2500**	**1**	**3.0220**	2.9515	3.0924	-6.1132	-7.0367	**4500**	**1**	**2.2362**	2.2144	2.2579	-6.1051	-6.2509
	2	**6.7500**	6.6006	6.8993	-5.4361	-10.0587		**2**	**2.1933**	2.1766	2.2100	-5.3953	-5.4988		**2**	2.1231	2.1161	2.1301	-5.3879	-5.4281
	5	**2.8327**	2.7680	2.8975	-4.4340	-5.1721		**5**	2.1088	2.1029	2.1146	-4.4198	-4.4469		**5**	2.0964	2.0922	2.1006	-4.4192	-4.4340
	10	**2.1722**	2.1581	2.1863	-3.6419	-3.7273		**10**	2.0933	2.0895	2.0972	-3.6371	-3.6486		**10**	2.0878	2.0850	2.0906	-3.6362	-3.6431
	20	2.1168	2.1107	2.1229	-2.7487	-2.7787		**20**	2.0842	2.0816	2.0869	-2.7455	-2.7489		**20**	2.0824	2.0805	2.0843	-2.7454	-2.7474
	30	2.0902	2.0854	2.0950	-2.1131	-2.1219		**30**	2.0821	2.0800	2.0842	-2.1111	-2.1137		**30**	2.0825	2.0809	2.0841	-2.1112	-2.1139
	40	2.0912	2.0870	2.0954	-1.5615	-1.5692		**40**	2.0838	2.0819	2.0856	-1.5616	-1.5647		**40**	2.0812	2.0798	2.0826	-1.5619	-1.5630
	50	2.0928	2.0887	2.0968	-1.0390	-1.0458		**50**	2.0817	2.0799	2.0835	-1.0393	-1.0409		**50**	2.0804	2.0791	2.0817	-1.0400	-1.0405
1000	**1**	**6.4870**	6.3443	6.6298	-6.1395	-10.5075	**3000**	**1**	**2.6591**	2.6051	2.7132	-6.1100	-6.6739	**5000**	**1**	**2.2074**	2.1893	2.2254	-6.1037	-6.2219
	2	**3.5189**	3.4285	3.6094	-5.4078	-6.8257		**2**	**2.1619**	2.1493	2.1746	-5.3902	-5.4671		**2**	2.1176	2.1111	2.1242	-5.3905	-5.4227
	5	**2.1924**	2.1767	2.2081	-4.4262	-4.5303		**5**	2.1020	2.0968	2.1072	-4.4196	-4.4398		**5**	2.0928	2.0888	2.0968	-4.4193	-4.4305
	10	2.1163	2.1099	2.1227	-3.6365	-3.6712		**10**	2.0920	2.0885	2.0955	-3.6365	-3.6474		**10**	2.0874	2.0847	2.0901	-3.6372	-3.6429
	20	2.0976	2.0933	2.1018	-2.7460	-2.7619		**20**	2.0867	2.0843	2.0891	-2.7457	-2.7511		**20**	2.0841	2.0823	2.0860	-2.7450	-2.7487
	30	2.0883	2.0850	2.0917	-2.1118	-2.1192		**30**	2.0833	2.0814	2.0852	-2.1121	-2.1146		**30**	2.0825	2.0809	2.0841	-2.1112	-2.1139
	40	2.0861	2.0832	2.0891	-1.5620	-1.5665		**40**	2.0814	2.0797	2.0831	-1.5626	-1.5634		**40**	2.0817	2.0803	2.0832	-2.1118	-2.1133
	50	2.0844	2.0816	2.0872	-1.0401	-1.0418		**50**	2.0809	2.0793	2.0825	-1.0399	-1.0402		**50**	2.0804	2.0791	2.0817	-1.0400	-1.0402
1500	**1**	**4.6278**	4.5124	4.7431	-6.1250	-8.6452	**3500**	**1**	**2.4233**	2.3839	2.4628	-6.1072	-6.4387							
	2	**2.5863**	2.5351	2.6376	-5.3983	-5.8921		**2**	2.1460	2.1361	2.1559	-5.3877	-5.4511							
	5	2.1322	2.1246	2.1322	-4.4223	-4.4704		**5**	2.1011	2.0963	2.1059	-4.4195	-4.4386							
	10	2.1034	2.0984	2.1085	-3.6382	-3.6587		**10**	2.0913	2.0881	2.0946	-3.6369	-3.6466							
	20	2.0894	2.0860	2.0929	-2.7467	-2.7541		**20**	2.0858	2.0835	2.0880	-2.7457	-2.7501							
	30	2.0855	2.0827	2.0883	-2.1102	-2.1162		**30**	2.0817	2.0799	2.0834	-2.1118	-2.1134							
	40	2.0835	2.0810	2.0859	-1.5618	-1.5649		**40**	2.0811	2.0796	2.0827	-1.5612	-1.5626							
	50	2.0841	2.0818	2.0864	-1.0405	-1.0424		**50**	2.0808	2.0793	2.0823	-1.0400	-1.0404							

Table 2
Summary of findings on the effect of IR and associated sample size

Sample Size	Estimates biased if IR is
100	30% and below
500	10% and below
1000 – 2000	5% and below
2500 – 3500	2% and below
4000 and above	1% and below

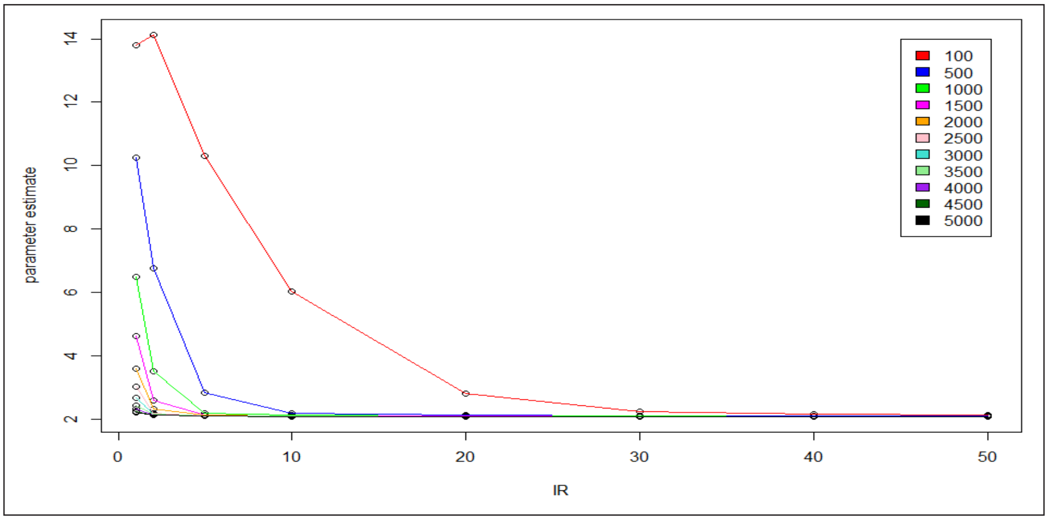

Figure 1. Categorical covariate's parameter estimates, $\hat{\beta}_1$, for different sample size and imbalance ratio (Imbalance Ratio (IR): 1% to 50%).

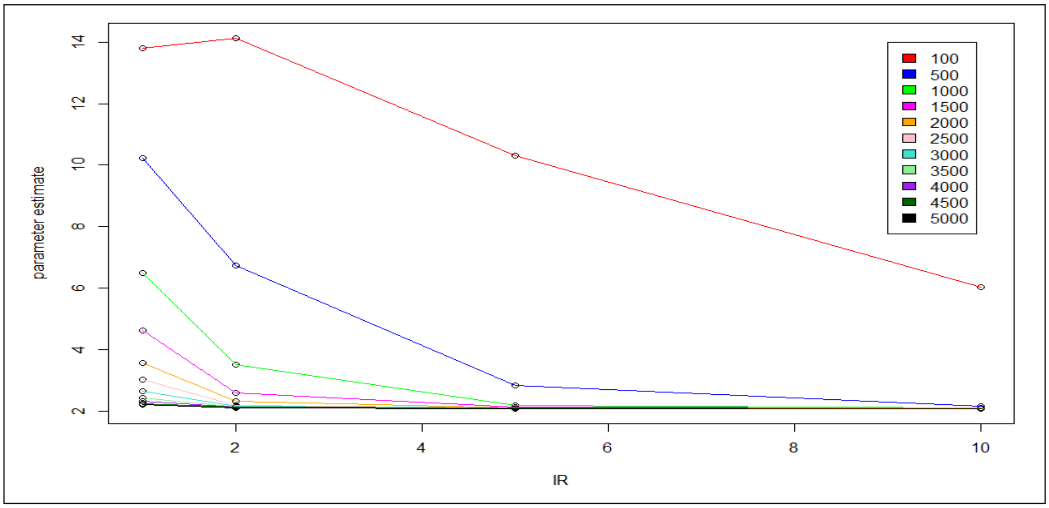

Figure 2. Categorical covariate's parameter estimates, $\hat{\beta}_1$, for different sample size and highly imbalance ratio (IR : 1-10%).

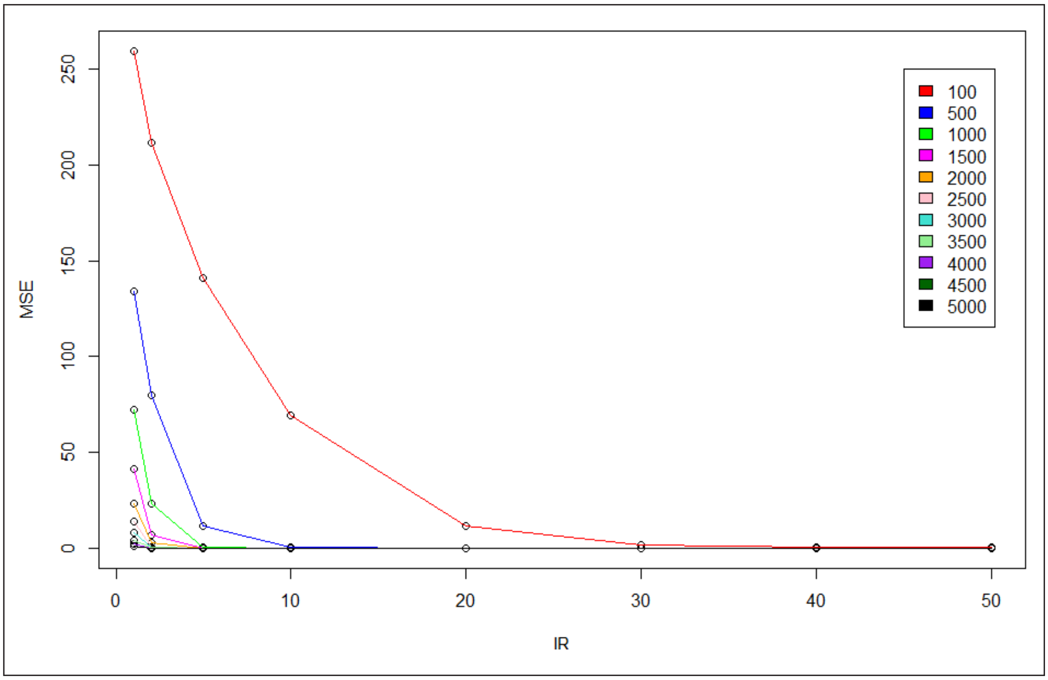

Figure 3. Mean square error (MSE) of categorical covariate's parameter estimates, $\hat{\beta}_1$, for different sample sizes and imbalance ratio

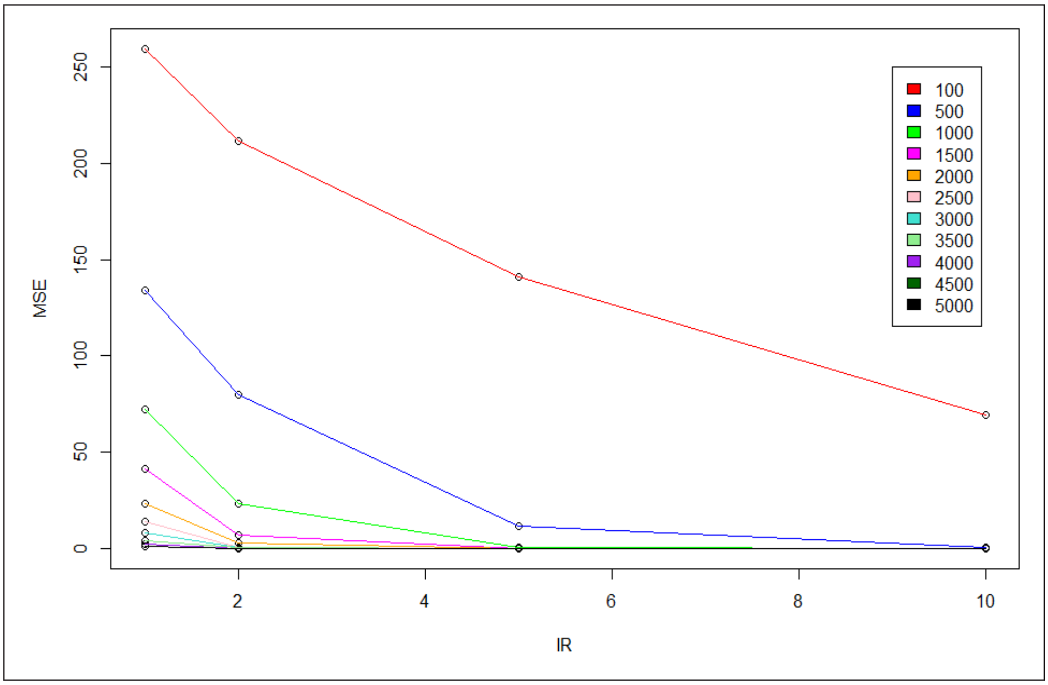

Figure 4. Mean square error (MSE) categorical covariate's parameter estimates, $\hat{\beta}_1$, for different sample size and highly imbalance ratio (IR : 1-10%).

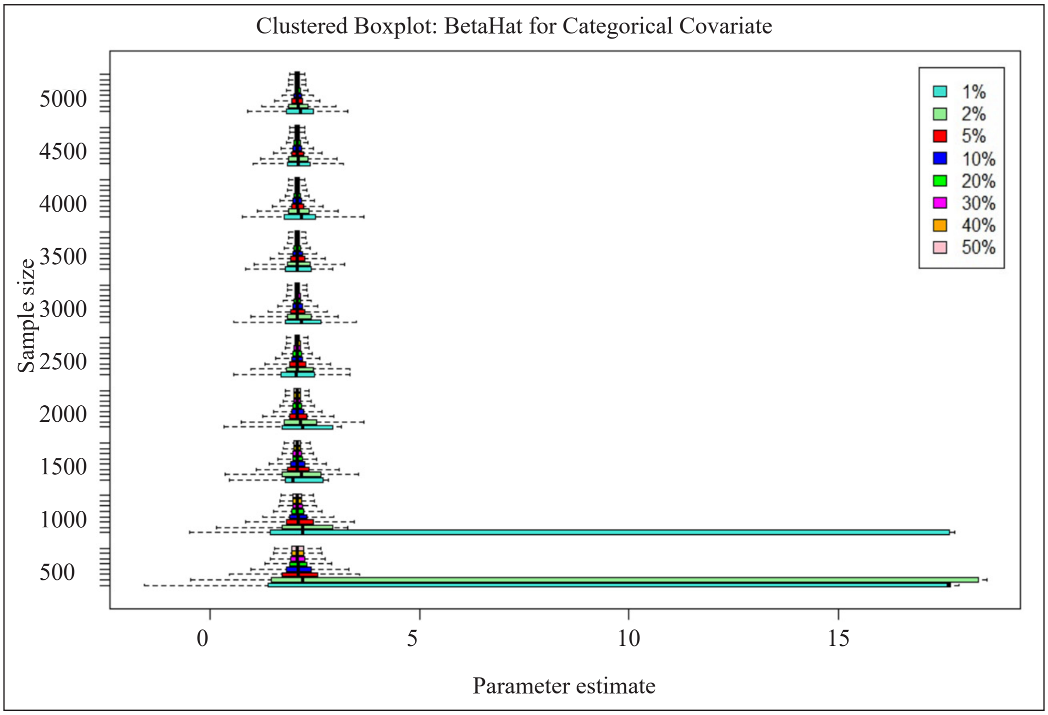

Figure 5. Clustered boxplots of $\hat{\beta}_1$ for a categorical covariate

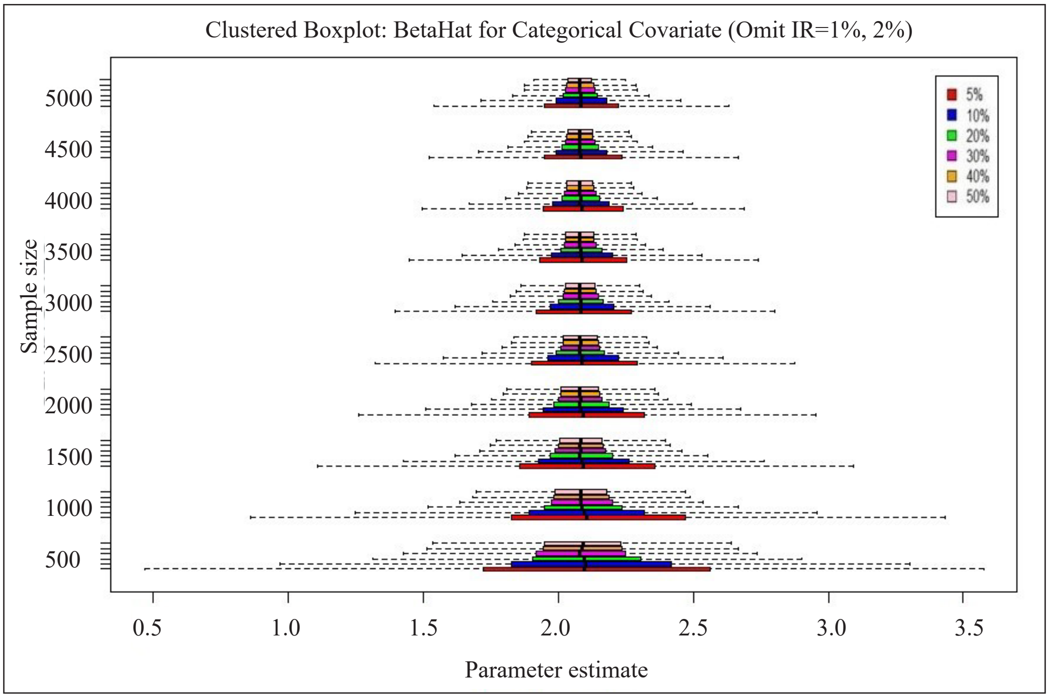

Figure 6. Clustered boxplots for $\hat{\beta}_1$ for a categorical covariate (omit IR=1%, 2%)

Hence, by referring to all the figures (Figures 1 to 6), it can be concluded that the effect of the imbalanced problem on the categorical covariate's parameter estimation was most severe for smaller sample sizes ($n \leq 500$) and for highly imbalanced ratios ($IR \leq 5\%$). The severity of the imbalanced problem was identified by the difference between the parameter estimated values and the fixed true beta value (β_1=2.08), as well as larger value of MSE. MSE is a good indicator of the bias in parameter estimates of the model. A larger MSE will indicate estimates are biased.

From this simulation results, the effect of IR for small sample size was very prominent with large MSE. Even for larger sample size (n=1000 and 1500), the effect of imbalance towards the parameter estimation was still apparent. For small sample size, n=100, only at IR = 30% onwards the value of the estimates became closer to the actual parameter value. Sample size n=500, the estimates improve at IR = 10% onwards. For other sample sizes $1000 \leq n \leq 2000$, $2500 \leq n \leq 3500$ and $n \geq 4000$, the parameter estimation improved at IR = 5%, 2% and 1% onwards. The summary of these findings is shown in Table 2.

Application to Real Data Results

This section illustrates and application using a real medical dataset (Diabetes Messidor dataset) from the UCI repository which has 16 covariates and known as "The Diabetes Messidor" dataset (Antal & Hajdu, 2014), consists of 1151 observations. This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not (DR status). All features represent either a detected lesion, a descriptive feature of an anatomical part or an image-level descriptor. The two categorical covariates selected for this illustration are the *retinal abnormality* and *AMFM status*. We modeled the binary dependent variable, *DR status* (1=with DR (53%) and 0=without DR (47%)). We used retinal abnormality (1 = yes, 0 = no) and AMFM status (0 = AM, 1 = FM) as the independent variable in Model 1 and Model 2 respectively. Using stratified sampling on the original dataset, we obtained the IR percentage as shown in Table 3.

Results in Table 3 show that the estimate $\hat{\beta}_1$ in Model 1 was affected for IR 5% and below. The p-values for $\hat{\beta}_1$ increases (leading to independent variable becoming insignificant) as imbalance becomes more severe thus leading to misleading results. Results of Model 2 shows the effect of imbalance on odds-ratio. The odds-ratios were extremely large at IR 1% and 2%. This application to real dataset confirmed the results of the simulation study, which strengthened the conclusion that imbalanced problem will be misleading on the effect of the independent variable on the response variable.

Table 3
Effect of imbalanced with application to real dataset (Diabetes Messidor)

Independent Variable	**Data/ IR**	$\hat{\beta}_0$ **, [p-value]** **C.I (lower, upper)**	$\hat{\beta}_1$ **, [p-value]** **C.I (lower, upper)**	***Odds-Ratio (OR)*** **C.I (lower, upper)**
Retinal Abnormality (1 = yes, 0 = no)	**Original** **(540:611)**	0.6614, [0.002] (0.6613, 0.6614)	-0.5838, [0.010] (-0.5837, -0.5838)	0.5578 (0.5577, 0.5578)
	40% **(407:611)**	0.9477, [0.000] (0.9414, 0.9539)	-0.5872, [0.026] (-0.5938, -0.5805)	0.5591 (0.5554, 0.5626)
	30% **(261:611)**	1.4151,[0.000] (1.4033, 1.427)	--0.6110, [0.075] (-0.6236, -0.5983)	0.5537 (0.5470, 0.5604)
	20% **(152:611)**	1.9717,[0.000] (1.9527, 1.9906)	-0.6262, [0.178] (-0.6462, -0.6062)	0.5609 (0.5506, 0.5712)
	10% **(68:611)**	2.9467,[0.007] (2.8593, 3.0342)	-0.7962, [0.345] -0.8846, -0.7078)	0.5700 (0.5528, 0.5871)
	5% **(35:611)**	4.6431,[0.080] (4.3845, 4.9016)	-1.8264, [0.502] -2.0863, -1.5664)	0.5920 (0.5680, 0.6159)
	2% **(13:611)**	10.5126, [0.418] (10.0310, 10.9941)	-6.7028, [0.806] (-7.1881, -6.2176)	0.6347 (0.5902, 0.6792)
	1% **(7:611)**	13.9315, [0.631] (13.4669, 14.3960)	-9.5009, [0.856] (-9.9711, -9.0306))	0.6799 (0.6137, 0.7461)
AMFM status (1 = FM, 0 = AM)	**Original** **(540:611)**	0.1837, [0.011] (0.1836, 0.1837)	-0.1785, [0.153] (-0.1784, -0.1785)	0.8364 (0.8364, 0.8365)
	40% **(407:611)**	0.4669, [0.000] (0.4657, 0.4680)	-0.1787, [0.219] (-0.1820, -0.1754)	0.8375 (0.8348, 0.8403)
	30% **(261:611)**	0.9124, [0.000] (0.9104, 0.9143)	-0.1804, [0.307] (-0.1860, -0.1750)	0.8381 (0.8335, 0.8428)
	20% **(152:611)**	1.4515, [0.000] (1.4483, 1.4548)	-0.1721, [0.404] (-0.1811, -0.1630)	0.8510 (0.8432, 0.8587)
	10% **(68:611)**	2.2623, [0.000] (2.2570, 2.2677)	-0.1785, [0.451] (-0.1934, -0.1635)	0.8615 (0.8480, 0.8749)
	5% **(35:611)**	2.9268, [0.000] (2.9193, 2.9343)	-0.1601, [0.480] (-0.1816, -0.1387)	0.9061 (0.8850, 0.9272)
	2% (13:611)	**3.9382, [0.000]** **(3.9240, 3.9525)**	**-0.0850, [0.500]** **(-0.1557, -0.0143)**	**79921.66** **(-10534.34, 170377.67)**
	1% (7:611)	**4.5989, [2.0000e-03]** **(4.5452, 4.6526)**	**0.5738, [5.4100e-01]** **(0.3403, 0.8074)**	**1676700** **(1185601, 2167798)**

CONCLUSIONS

Imbalanced data has effect on the parameter estimates and classification performance of binary logistic regression model with a categorical covariate. The optimal IR for different sample size for less biased estimates was determined via a simulation study. It was concluded that all samples are affected by imbalanced even for larger sample sizes. The effect of imbalanced data on parameter estimates reduces as sample size increases. The imbalanced ratio in the response variable will not only affect the parameter estimates, but the p-value and odds- ratio for the covariate as well. Hence, imbalanced data can lead to inaccurate findings. There are approaches recommended for handling imbalanced problem such as resampling strategies (ROS (Random Oversampling), RUS, (Random Undersampling) and SMOTE (Synthetic Minority Oversampling Technique). Future simulation studies can investigate which sampling techniques can improve the parameter estimates and predictive performance of the binary logistic regression when data is highly imbalanced.

ACKNOWLEDGEMENT

Our gratitude goes to Universiti Teknologi MARA for the indirect contribution of this publication. We also thank Prof. Dr. Haibo He (Rhodes Island University), Prof. Dr. Ronaldo Prati (Universidade Federal do ABC), Dr. Pam Davey and Dr. Carolle Birrell (University of Wollongong) for sharing their knowledge and providing valuable comments for this study. Not to forget Dr. Hamzah Abdul Hamid (Univeristi Malaysia Perlis) for his indirect contribution to the simulation studies.

REFERENCES

Ahmad, S., Midi, H., & Ramli, N. M. (2011). Diagnostics for residual outliers using deviance component in binary logistic regression. *World Applied Sciences Journal, 14*(8), 1125-1130.

Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*, *39*(5), 1385-1391. doi: https://doi.org/10.1007/s00726-010-0595-2

Antal, B., & Hajdu, A. (2014). An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems*, *60*, 20-27. doi: https://doi.org/10.1016/j.knosys.2013.12.023

Blagus, R., & Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *11*(1), 1-17. doi: https://doi.org/10.1186/1471-2105-11-523

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626-4636. doi: https://doi.org/10.1016/j.eswa.2008.05.027

Chawla, N. V. (2003, August 21). C4. 5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Set II* (Vol. 3, p. 66). Washington, DC.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1-6. doi: https://doi.org/10.1145/1007730.1007733

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., & Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, *37*(1), 7-18. doi: https://doi.org/10.1016/j.artmed.2005.03.002

Dong, Y., Guo, H., Zhi, W., & Fan, M. (2014, October 13-15). Class imbalance oriented logistic regression. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (pp. 187-192). Shanghai, China. doi: https://doi.org/10.1109/CyberC.2014.42

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463-484. doi: 10.1109/TSMCC.2011.2161285

Goel, G., Maguire, L., Li, Y., & McLoone, S. (2013). Evaluation of sampling methods for learning from imbalanced data. *Intelligent Computing Theories*, *7995*, 392-401. doi: https://doi.org/10.1007/978-3-642-39479-9_47

Hamid, H. A. (2016). Effects of different type of covariates and sample size on parameter estimation for multinomial logistic regression model. *Jurnal Teknologi*, *78*(12-3), 155-161. doi: https://doi.org/10.11113/jt.v78.10036

Hamid, H. A., Yap, B. W., Xie, X. J., & Rahman, H. A. A. (2015). Assessing the effects of different types of covariates for binary logistic regression. In *AIP Conference Proceedings 1643* (Vol. 425, pp. 425-430). New York, USA: American Institute of Physics. doi: https://doi.org/10.1063/1.4907476

Hamid, H. A., Yap, B. W., Xie, X. J., & Ong, S. H. (2018). Investigating the power of goodness-of-fit tests for multinomial logistic regression. *Communications in Statistics: Simulation and Computation*, *47*(4), 1039-1055. doi: https://doi.org/10.1080/03610918.2017.1303727

He, H., & Garcia, E. E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: https://doi.org/10.1109/TKDE.2008.239

Hosmer, D. W., & Lemeshow, S. (2004). *Applied logistic regression, second edition*. New York, NY: John Wiley & Sons, Inc. doi: https://doi.org/10.1002/0471722146

Lemnaru, C., Potolea, R., Lenmaru, C., & Potolea, R. (2012). Imbalanced classification problems: Systematic study, issues and best practices. *Enterprise Information Systems: Lecture Notes in Business Information Processing*, *102*, 35-50. doi: https://doi.org/10.1007/978-3-642-29958-2

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*, *2*(1), 83-87. doi: https://doi.org/10.1109/SIU.2013.6531574

Mena, L., & Gonzalez, J. A. (2006, May 11-13). Machine learning for imbalanced datasets: Application in medical diagnostic. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2006)* (pp. 574-579). Florida, USA.

Oztekin, A., Delen, D., & Kong, Z. J. (2009). Predicting the graft survival for heart-lung transplantation patients: An integrated data mining methodology. *International Journal of Medical Informatics*, *78*(12), e84-e96. doi: https://doi.org/10.1016/j.ijmedinf.2009.04.007

Pourahmad, S., Ayatollahi, S. M. T., & Taheri, S. M. (2011). Fuzzy logistic regression: A new possibilistic model and its application in clinical vague status. *Iranian Journal of Fuzzy Systems*, *8*(1), 1-17.

Prati, R. C., Batista, G. E. A. P. A., & Silva, D. F. (2014). Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, *45*(1), 247-270. doi: https://doi.org/10.1007/s10115-014-0794-3

Rahman, H. A. A., & Yap, B. W. (2016). Imbalance effects on classification using binary logistic regression. In *International Conference on Soft Computing in Data Science* (pp. 136-147). Singapore: Springer. doi: https://doi.org/https://doi.org/10.1007/978-981-10-2777-2_12

Rahman, H. A. A., Yap, B. W., Khairudin, Z., & Abdullah, N. N. (2012, September 10-12). Comparison of predictive models to predict survival of cardiac surgery patients. In *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)* (pp. 1-5). doi: https://doi.org/10.1109/ICSSBE.2012.6396534

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: A review. *International Journal of Computing and Business Research*, *5*(4), 1-29.

Rothstein, M. A. (2015). Ethical issues in big data health research: Currents in contemporary bioethics. *The Journal of Law, Medicine and Ethics*, *43*(2), 425-429. doi: https://doi.org/10.1111/jlme.12258

Roumani, Y. F., May, J. H., Strum, D. P., & Vargas, L. G. (2013). Classifying highly imbalanced ICU data. *Health Care Management Science*, *16*(2), 119-128. doi: https://doi.org/10.1007/s10729-012-9216-9

Sarmanova, A., & Albayrak, S. (2013, April 24-26). Alleviating class imbalance problem in data mining. In *2013 21st Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). Haspolat, Turkey. doi: 10.1109/SIU.2013.6531574

Shariff, S. S. R., Rodzi, N. A. M., Rahman, K. A., Zahari, S. M., & Deni, S. M. (2016). Predicting the "graduate on time (GOT)" of PhD students using binary logistics regression model. In *AIP Conference Proceedings* (Vol. 1782, No. 1, p. 050015). New York, USA: AIP Publishing LLC. doi: https://doi.org/10.1063/1.4966105

Srinivasan, U., & Arunasalam, B. (2013). Leveraging big data analytics to reduce healthcare costs. *IT Professional*, *15*(6), 21-28. doi: https://doi.org/10.1109/MITP.2013.55

Uyar, A., Bener, A., Ciracy, H. N., & Bahceci, M. (2010). Handling the imbalance problem of IVF implantation prediction. *IAENG International Journal of Computer Science, 37*(2), 164-170.

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning* (pp. 935-942). New York, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/1273496.1273614

Wallace, B. C., & Dahabreh, I. J. (2012, December 10-13). Class probability estimates are unreliable for imbalanced data (and how to fix them). In 2*012 IEEE 12th International Conference on Data Mining* (pp. 695-704). Brussels, Belgium. doi: 10.1109/ICDM.2012.115

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315-354. doi: https://doi.org/10.1613/jair.1199

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)* (pp. 13-22). Singapore: Springer. doi: https://doi.org/10.1007/978-981-4585-18-7

APPENDIX

```
#fitting the model
set.seed(54321)
ndata <- 100
nrep <- 10000 #set the number of replications
start <- -10 #set initial value of bnot
end <- 10 #set end value for bnot
n <- 1 #set initial value for the loop
perc <- 40 #set the percentage of imbalance

#replication setup
beta0Hat<-rep(NA,nrep)
beta1Hat<-rep(NA,nrep)
betanot<-rep(NA,nrep)
betaone <- 2.08

while(n<=nrep)
{
  #set bnot value
  for(i in seq(start,end,0.001))
  {
    x <- rbinom(ndata,1,1/2)
    rx <-chartr(“01”, “AB”, x)
    dummy(x)
    k <-dummy(x)
    linpred <- cbind(1,dummy(x)[,-1])%*% c(i,betaone) #(b)
    pi<-exp(linpred)/(1+exp(linpred))
    ru <- runif(ndata,0,1)
    u<-as.vector(ru)
    ry <- ifelse((u<=pi),1,0)
    m_y <- (mean(ry)*100)
    if(m_y == perc && n <=nrep)
    {
      dt <-data.frame(x=rx, y=ry) #fit the logistic model
      #print(dt)
      betanot[n]<-i
      mod <- glm(y~x, family=”binomial”, data=dt)
      beta0Hat[n]<-mod$coef[1]
      beta1Hat[n]<-mod$coef[2]
      n <- n + 1
    }
  }
}

Round1<-round(c(beta0=mean(beta0Hat),beta1Hat=mean(beta1Hat)),3)
mean(beta1Hat)
ci.b1 <- CI(beta1Hat,ci=0.95)
MSEbeta1Hat <- round(sum((beta1Hat-2.08)^2/nrep),3)
meanb0 <- mean(betanot)
mean(beta0Hat)
ci.b0 <- CI(beta0Hat,ci=0.95)
```