A Comparative Study of Synthetic Over-sampling Method to Improve the Classification of Poor Households in Yogyakarta Province

To cite this article: B Santoso et al 2018 IOP Conf. Ser.: Earth Environ. Sci. 187 012048
A Comparative Study of Synthetic Over-sampling Method to Improve the Classification of Poor Households in Yogyakarta Province

B Santoso¹²*, H Wijayanto¹, K A Notodiputro¹ and B Sartono¹

¹Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia
²BPS-Statistics Indonesia, Jakarta, Indonesia

*email: budsant.bs@gmail.com

Abstract. The problems of class imbalance have attracted concerns from researchers in the last few years. Class imbalance problems occur when the data had unbalanced proportions between two or more groups of data which are usually called as minority and majority classes. These problems relate to creation of bias in parameter estimation as well as misclassification of the objects especially for the minority class. These will lead to incorrect prediction of the minority class, and eventually will risk the policy making. Several approaches have been proposed to correct misclassification such as data-based and algorithm-based approaches. As a data-based approach, over-sampling method is very popular nowadays. This approach is basically balancing the distribution of data through addition of synthetic data. This paper discusses the strategies of adding synthetic data in order to improve the accuracy of classification. Moreover, this paper also reviews several over sampling methods for class imbalanced problems. Specifically, the classification of poor households is illustrated by using the National Socio-Economic Survey (Susenas) data which has been stratified according to urban and rural areas. Finally, the K-Nearest Neighbor (KNN), Naïve Bayes, Support Vector Machine (SVM) and Generalized Linear Model (GLM) are employed to evaluate the classification performance by comparing the value of sensitivity and area under the ROC curve (AUC). The simulation result shows that there are bias on parameter estimation both on interception and on slope. The bias gets bigger as the data condition becomes more unbalanced and on small sample. Meanwhile, the classification accuracy will decrease with the decrement of probability (high imbalanced) value especially in the data with small sample. Decreased accuracy of classification mainly occurs in the minority class (sensitivity) and AUC. Based on the simulation result, it is clear that the synthetic over sampling approach can improve the accuracy of classification in minority class through increasing sensitivity value and AUC value. This occur at the small probability (unbalanced data). In line with the simulation results, the over sampling approach also shows the evident of improving the prediction of poor households in Yogyakarta Province. But on the other hand, it can also lead to decreased accuracy and specificity. However, further research is required to obtain a more accurate prediction result for all performance measures.

1. Introduction
Class imbalance commonly found in any real cases. The problems of class imbalance have attracted concerns from researchers in the last few years. Class imbalance problems occur when the data has unbalanced proportions between two or more groups of data which are usually called as minority and majority classes. These problems relate to creation of bias in parameter estimation as well as
misclassification of the objects especially for the minority class. These will lead to create an incorrect prediction of the minority class, and eventually will risk the policy making.

There are two approaches should be performed to solve imbalanced data problems, those are solution at data level and solution at algorithm level [1]. Solution at data level is applied by balancing the distribution of the majority and minority class through over sampling methods, under sampling methods or combination of both methods. Solution at algorithm level is applied by modifying in classifier methods or optimize the performance of learning algorithm [2]. The advantage of the data level is that can use independent of the classifier selected.

As a data-based approach, over-sampling method is very popular nowadays. This approach is basically balancing the distribution of data through addition of synthetic data. Over sampling approach or synthetic over sampling are used more frequently than under sampling since under sampling method will eliminate data in the majority class thus causing loss of important information of the data. According to Batista et al [3] in general, synthetic over sampling method gives better results than under sampling method. Over sampling method is more effective than under sampling, especially on data that has a high imbalance.

This paper discusses strategies of adding synthetic data in order to improve the accuracy of classification. This paper also reviews several synthetic over sampling methods for class imbalanced problems. These over sampling methods will be applied to classify poor households based on Susenas 2016 in Yogyakarta Province by using various classifier such as K-Nearest Neighbor (KNN), Naïve Bayes, Support Vector Machine (SVM) and Generalized Linear Model (GLM).

In addition, this paper also discusses data simulations with varying amounts of data and level of data imbalance. Furthermore, the results of parameter estimation and classification accuracy of simulation data will be explained. The SMOTE over sampling method is applied to the simulation data to compare the resulting output.

2. Material

2.1. Synthetic Over Sampling

Over sampling method is a data-based approach that is done by adding synthetic data in order to balance the distribution of data. Some approaches are made to obtain the right synthetic data that can characterize minority classes so as to improve the classification of minority data. The simple over sampling method is the Random Over Sampling (ROS), that is carried out by balancing the distribution of data through the application of minority data duplication at random. However, the method has the disadvantage that ROS cause problems overfitting [4]. Chawla et al [5] proposed a new technique, named SMOTE, to generate synthetic data based on the distance between the minority data and the closest minority data. Therefore, the new synthetic data will be formed between the two minority data.

Some researches are conducted to modify SMOTE in order to create more effective technique in improving the classification performance. Han et al [6] divided three location based on the amount of majority data in the nearest neighbors of minority data. If the majority data are all around the nearest neighbors, then the area is called noisy. If the majority data around the nearest neighbors are found higher than the minority or equal to the minority, the area is called borderline. Meanwhile if most of all data around the nearest neighbors are the minority data, then the area is called safe. Han et al [6] focused on the borderline area located on the boundary between the minority and the majority of data. Data generating through SMOTE is performed in the borderline areas, called the Borderline SMOTE (BR-SMOTE). Moreover, Bunkhumpornpat et al [7] paid attention to the safe area to perform over sampling based on the ratio between the number of minority data and the nearest neighbors. This method wants to make sure that the synthetic data that will be generated through the SMOTE are in areas that are completely safe. Therefore the method is called the Safe Level SMOTE (SL-SMOTE).
Other studies used an over-sampling approach based on SMOTE include Cluster-SMOTE [8], CURE-SMOTE [9], N-SMOTE [10] and Nuclear-SMOTE [11]. Other over sampling approaches include ADASYN [12], ADOMS [13], AHC [14] and SPIDER [15].

This research will use over-sampling method of ROS, ROUS, SMOTE, BL-SMOTE and SL-SMOTE as an approach to obtain synthetic data. The over sampling method is applied to poor households data using urban and rural stratification. Stratification on various over sampling methods is performed in order to obtain more homogeneous synthetic data in minority class so it is expected to increase the accuracy of classification on minority data.

2.2. Classification Methods
Research on class imbalance has been done in many fields using various classification methods. Chawla [5] applied various data sets such as diabetes dataset, phoneme dataset, satimage dataset and oil dataset using C4.5, Ripper and Naïve Bayes classifier. Meanwhile, Bunkhumpornpat et al [7] used the classification method of C4.5, Naïve Bayes and Support Vector Machines (SVM) on haberman and satimage data. The C4.5 classification method is also used by Ramentol [1] in applying the ecoli dataset to complete class imbalanced. Kubat and Matwin [16] conducted research on oil slicks data using K-NN and C4.5 methods.

In Indonesia, class imbalanced research is mostly undertaken by using various classification methods such as C4.5, CART, K-NN or SVM. Komarek and Moore [17] conducted a study to improve the prediction accuracy of Logistic Regression (GLM) model. In some data mining implementation, GLM can outperform other algorithms such as Naïve Bayes, SVM and K-NN [18]. This research uses K-NN, Naïve Bayes, SVM and GLM classification methods to get prediction of classification accuracy.

3. Methods

3.1. Data Simulation
The simulation data used 2 predictor variables (X1 and X2) and response variable (Y). The predictor variables are generated by a normal distribution with mean 0 and varian 1 ($X_1 \sim N(0,1), X_2 \sim N(0,1)$). While response variables are generated by the Bernoulli distribution where the Y value follows the distribution:

$$Y = (1, \pi), \pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$$

(1)

The data generation process is performed on several types of data distributions, taking into account the level of imbalance of data such as 50:50 ($\pi = 0.5$), 80:20 ($\pi = 0.2$), 90:10 ($\pi = 0.1$), 95:5 ($\pi = 0.05$) and 99:1 ($\pi = 0.01$). The parameter quantity determined fix in accordance with the level of imbalance that is for the varying parameters of interception and the fixed fix slope parameter are $\beta_1 = 2$ and $\beta_2 = 3$. Table 1 shows different types of the level of imbalance and Linear predictor.

| Phi     | Linear predictor ($\beta_1 = 2$ and $\beta_2 = 2$) |
|---------|----------------------------------------------------|
| $\pi = 0.5$ (50:50) | $2X_1 + 3X_2$ |
| $\pi = 0.2$ (80:20) | $-3 + 2X_1 + 3X_2$ |
| $\pi = 0.1$ (90:10) | $-5 + 2X_1 + 3X_2$ |
| $\pi = 0.05$ (95:5) | $-7 + 2X_1 + 3X_2$ |
| $\pi = 0.01$ (99:1) | $-9 + 2X_1 + 3X_2$ |

For each value of X, the true $\pi (X, \beta)$ has been calculated by using the corresponding $\beta$ values for each case to generate the conditional Bernoulli trial Y. So there are five dataset containing $X_1$, $X_2$ and
Y with various class distribution. In addition, the data simulation process is applied by using different n values of n = 100, n = 500, n = 750, n = 1000, n = 2,500, n = 5,000, n = 7,500 and n = 10,000 with 1,000 replication. The next step is partition the data: 75 percent are training data and 25 percent are test data. Training data is use for modeling while the test data is used to test the accuracy of classification model.

3.2. Data Application
This study uses the poor households data in Yogyakarta (DIY) province in 2016 based on the results of the National Socio-Economic Survey (Susenas). The number of samples is 3,662 households which consist of 482 (13.2 percent) poor households and 3,180 (86.8 percent) non-poor households. This 13.2 percentage of poor households data in Yogyakarta province can be categorized as class imbalance. Meanwhile, stratified by urban-rural, the percentage of poor households in urban was 11.3 percent and in rural was 17.3 percent. Table 2 inform operational definition of variables. Poor households data (Y = 1) is minority class and non poor households (Y = 0) is majority class.

Table 2. Operational Definition of Variables

| Variable             | Symbol | Definition                                      |
|----------------------|--------|------------------------------------------------|
| Poor Status          | Y      | Y = 1 (poor households)                         |
|                      |        | Y = 0 (non poor households)                     |
| Education indicator  | X₁     | Head of Household’s Mean years schooling        |
| Housing indicator    | X₂     | Floor area Per capita                           |

3.3. Evaluation Measures
Table 3 shows confusion matrix used to determine the accuracy of classification in minority and majority classes. The class label of the minority class is positive and the class label of the majority class is negative. The first row of the table is the predicted class label and the first column present their actual class label. TP and TN denote the number of positive and negative examples that are classified correctly, while FP and FN denote the number of misclassified positive and negative examples respectively.

Table 3. Confusion matrix

| Predicted class | Actual class |               |               |
|-----------------|--------------|---------------|---------------|
|                 | No           | Yes           |               |
| No              | TN: True Negative | FN: False Negative |
| Yes             | FP: False Positive  | TP: True Positive  |

\[
\text{Accuracy} = \frac{TP + TN}{(TP + FP + TN + FN)}
\]

\[
\text{Sensitivity} = \frac{TP}{(TP + FN)}
\]

\[
\text{Specificity} = \frac{TN}{(TN + FP)}
\]

Using accuracy to evaluate the precision of classification on classed imbalance is not enough to show the goodness of classification. Accuracy only indicates the precision of classification on the composite of both majority and minority data classes. Performance measures such as sensitivity, specificity and under-curve area analysis Receiver Operator Characteristic (AUC-ROC) are more appropriate for class imbalance because they are more specific in predicting each class. Sensitivity is a measure of classification accuracy in minority data, while specificity is a measure of classification...
accuracy in majority data. Size that can accommodate the classification accuracy of both data groups is AUC-ROC. The ROC curve describes the classification performance in 2 dimensions: probability plot of false negative (1-specificity) and the correct prediction of true positive (sensitivity). AUC value is between 0.5 to 1 where the value close to 1 means the accuracy of classification is very good or can distinguish the class very well [19].

General rule of AUC values:
- If AUC-ROC = 0.5 (diagonal line) this suggests no discrimination
- If 0.7 ≤ AUC-ROC < 0.8 this is considered acceptable discrimination
- If 0.8 ≤ AUC-ROC < 0.9 this is considered excellent discrimination
- If AUC-ROC ≥ 0.9 this is considered outstanding discrimination

4. Result and Discussion

4.1. Data Simulation
In the data simulation process, GLM model is used to generate parameter estimation in each class distribution. In addition, classification predictions are made on each different class distributions.

4.1.1. Parameter Estimation
Table 4 presents information of intercept parameters estimation based on different levels of data imbalance and amount of data. In the balanced data (π = 0.5) there is no bias on the intercept. At the moment the value of π = 0.2, bias occurs on the data with the sample under 1,000. Meanwhile, the value of π = 0.1 bias still occurs in the range sample under 5,000 data. Whereas in data with high imbalance (π = 0.05 and π = 0.01) occur bias at all different sample numbers. Therefore the bias on the intercept will enlarge as the higher the level of data imbalance and the smaller the number of sample.

Table 4. Mean Intercepts ($\beta_0$)

| n         | $\pi=0.5$ (0) | $\pi=0.2$ (-3) | $\pi=0.1$ (-5) | $\pi=0.05$ (-7) | $\pi=0.01$ (-9) |
|-----------|---------------|----------------|----------------|-----------------|----------------|
| n=100     | 0.001857      | -3.45584       | -18.5084       | -165.404        | -1568.93       |
| n=500     | -0.00370      | -3.08090       | -5.16498       | -7.47527        | -27.3375       |
| n=750     | 0.00195       | -3.03687       | -5.09400       | -7.27565        | -10.4883       |
| n=1,000   | -0.00064      | -3.01920       | -5.06422       | -7.20207        | -9.86111       |
| n=2,500   | -0.00157      | -3.00630       | -5.02762       | -7.11574        | -9.29462       |
| n=5,000   | 0.001851      | -3.00286       | -5.01874       | -7.03625        | -9.14672       |
| n=7,500   | 0.002457      | -3.00633       | -5.00725       | -7.02170        | -9.09407       |
| n=10,000  | 0.002267      | -3.00777       | -5.00746       | -7.02838        | -9.06599       |

Table 5 shows the average information for estimating the slope parameters $\beta_1$ and $\beta_2$ based on different levels of data imbalance and amount of data. In the slope $\beta_1$ and $\beta_2$ biases occur at all different values. In the slope $\beta_1$ when the value $\pi = 0.5$ and $\pi = 0.2$, the bias occurs in the data with the number sample below 1,000. Meanwhile, when the value of $\pi = 0.1$ bias is found in the data with the number sample below 2,500. While at the value of $\pi = 0.05$ bias occurs in the data with the number n under 5,000. In the data which value $\pi = 0.01$ occurs bias at all n values.

In the $\beta_2$ slope when the value $\pi = 0.5$ and $\pi = 0.2$, the bias occurs in the data with the number sample below 2,500. Meanwhile, when the value $\pi = 0.1$ and $\pi = 0.05$ is still found biased on the data with the number sampel under 5,000. In line with the expected value of parameter $\beta_1$, the bias occurs in the data with value $\pi = 0.01$ on all different sample values. So, in general, the bias on the slope will be more enlarged on the condition of the value of $\pi$ and samples getting smaller.
Table 5. Mean Slopes ($\beta_1$ and $\beta_2$)

| $\beta$ | n  | $\pi$  | $\pi=0.5$ | $\pi=0.2$ | $\pi=0.1$ | $\pi=0.05$ | $\pi=0.01$ |
|---------|----|--------|---------|---------|---------|---------|---------|
| $\beta_1=2$ | n=100 | 2.192841 | 2.297625 | 9.225770 | 47.80513 | 511.2142 |
|         | n=500 | 2.032058 | 2.058041 | 2.071574 | 2.149770 | 5.832223 |
|         | n=750 | 2.026163 | 2.023081 | 2.040823 | 2.087543 | 3.63052 |
|         | n=1,000 | 2.014577 | 2.013286 | 2.032178 | 2.069512 | 2.184043 |
|         | n=2,500 | 2.006580 | 2.005687 | 2.016784 | 2.032858 | 2.072915 |
|         | n=5,000 | 2.005232 | 2.003884 | 2.009619 | 2.011715 | 2.048653 |
|         | n=7,500 | 2.002807 | 2.005749 | 2.002683 | 2.007079 | 2.017387 |
|         | n=10,000 | 2.000138 | 2.005042 | 2.004971 | 2.007215 | 2.015082 |
| $\beta_2=3$ | n=100 | 3.324896 | 3.465245 | 10.66984 | 73.25643 | 660.5024 |
|         | n=500 | 3.051778 | 3.072096 | 3.107549 | 3.220332 | 9.677452 |
|         | n=750 | 3.037183 | 3.039288 | 3.068692 | 3.126216 | 3.500576 |
|         | n=1,000 | 3.024531 | 3.021123 | 3.036344 | 3.087298 | 3.317656 |
|         | n=2,500 | 3.014715 | 3.017095 | 3.013352 | 3.049676 | 3.108968 |
|         | n=5,000 | 3.008323 | 3.007012 | 3.010509 | 3.015634 | 3.044052 |
|         | n=7,500 | 3.003282 | 3.008546 | 3.004942 | 3.009466 | 3.040283 |
|         | n=10,000 | 3.000853 | 3.006682 | 3.004651 | 3.001304 | 3.022652 |

4.1.2. Classification Prediction

AUC is used as a measurement of classification accuracy evaluation that has accommodated the accuracy of classification in minority and majority class. Figure 1 describes the AUC based on different levels of data imbalance and amount of data. In general, AUC will decrease along with decreasing value of probability. So, it means that the more unbalanced the data will make the AUC smaller which decline the accuracy of classification. So that the minimum AUC can not distinguish the class accurately.

In the balanced data ($\pi=0.5$) it shows that the precision of classification is fine which is marked with high AUC. Meanwhile, the data with $\pi=0.2$ shows that the classification accuracy decreases compared with the balanced data. In the condition $\pi=0.1$, the AUC value decreases even the classification can not be distinguished at the small sample (n = 100). On data that has a high imbalance ($\pi=0.05$ and $\pi=0.01$) AUC value is decline. At the value of $\pi=0.05$, the classification can be distinguished on sample above 750. While in data with high imbalance rate (value $\pi=0.01$), the classification can be differentiated on sample above 5,000. Even in small size sample (n = 100 and 500) can not generated AUC value due to the absence of minority data (Y = 1) in the test data.

Based on the performance measure, there are different patterns between sizes. The value of AUC and sensitivity value will decrease along with unbalanced data (small value in $\pi$). The decrease in sensitivity values indicates that predictions in minority classes are increasingly inaccurate. However, there is an increasing value of accuracy and specificity. The increasing of specificity value indicates that prediction in the majority class is increasing dominantly. So that high accuracy value on unbalanced data is dominated by high specificity value (see supplementary table). In general, the simulation results show that the classification accuracy will be increase when the data is more balanced and in large data. (the more balanced the data, the more accurately the classification)

4.1.3. Classification Prediction After Over Sampling

Over sampling method use to prove and to overcome the problem of classification on unbalanced data. Figure 3 and figure 4 give evidence that over sampling method (SMOTE) can improve classification performance especially on data with small value in $\pi$. In another hand, the use of over sampling
method does not give an impact on classification accuracy when the data is balanced ($\pi = 0.5$). This can be seen from the performance measure that did not increase. Meanwhile, on unbalanced data (value $\pi \leq 0.2$), the use of over sampling method can increase the value of AUC and sensitivity value. Therefore The accuracy of classification in minority classes can be increased.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{AUC_value.png}
\caption{AUC value by $\Phi$}
\label{fig:AUC}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Sensitivity_value.png}
\caption{Sensitivity value by $\Phi$}
\label{fig:Sensitivity}
\end{figure}
However, over sampling method can reduce the value of specificity and accuracy value, because over sampling method is mainly focus to raise the classification performance on minority data that is through the addition of synthetic data in minority class. On the other hand, Figure 3 and figure 4 give
evidence that if the handling of the majority data is not done, so it can decrease the classification in the majority class, which is the specificity value and the accuracy degradation value.

4.2. Poor Households Data in Yogyakarta Province

Based on Susenas 2016 there are 3,662 households in Yogyakarta Province where 482 households are poor households (13.16%) and 3,180 are non-poor (86.84%). Then data is divided into training data (2,747 households) and test data (915 households).

The result of classification prediction on poor households in Yogyakarta Province shows that the accuracy of classification in poor households is very low. Poor households that are predicted to be poor only 13 households from 126 poor households. There are 113 poor households predicted to be non-poor households. Furthermore, some over sampling methods are used to improve accuracy, especially in predicting poor households such as ROS, ROUS, SMOTE, BL-SMOTE and SL-SMOTE.

In addition to the general approach, the process of adding synthetic data is distinguished by urban and rural stratification in order to increase homogeneity in the class. So there are additional 5 methods of Stratification-ROS, Stratification-ROUS, Stratification-SMOTE, Stratification-BLSMOTE and Stratification-SLSMOTE.

Table 6. Confusion matrix of Poor Households with Over Sampling

| Over sampling Methods | Actual class | Predicted class |  |  |
|-----------------------|--------------|-----------------|---|---|
|                       | Poor         | Non Poor        | Poor | Non Poor |
| Baseline              | 13           | 113             |     |         |
| ROS                   | 103          | 23              |     |         |
| Non Poor              | 385          | 404             |     |         |
| Stratified-ROS        | 103          | 23              |     |         |
| Non Poor              | 425          | 364             |     |         |
| ROUS                  | 97           | 29              |     |         |
| Non Poor              | 389          | 400             |     |         |
| Stratified-ROUS       | 93           | 33              |     |         |
| Non Poor              | 329          | 460             |     |         |
| SMOTE                 | 79           | 47              |     |         |
| Non Poor              | 283          | 506             |     |         |
| Stratified-SMOTE      | 79           | 47              |     |         |
| Non Poor              | 285          | 504             |     |         |
| BL-SMOTE              | 101          | 25              |     |         |
| Non Poor              | 255          | 534             |     |         |
| Stratified-BLSMOTE    | 90           | 36              |     |         |
| Non Poor              | 302          | 487             |     |         |
| SL-SMOTE              | 102          | 24              |     |         |
| Non Poor              | 303          | 486             |     |         |
| Stratified-SLSMOTE    | 85           | 41              |     |         |
| Non Poor              | 193          | 596             |     |         |

Table 6 illustrates the use of over sampling methods. Generally, all over sampling methods are able to increase predictions in poor households significantly. ROS and Stratified-ROS methods were able to predict poor households as poor households of 103 households from 126 households.
(sensitivity=0.82). Unfortunately, the method is not able to predict non-poor households optimally (specificity=0.4).

If it refers to the size that accommodates the prediction accuracy of both poor households and non-poor households, then the value of AUC is used. However, the value of AUC between over sampling method is inconsistent when viewed from the classification methods. In K-NN, the method with the highest AUC value is the BLSMOTE method. Meanwhile, Stratified-ROS and Stratified-SLSMOTE are the methods that have the highest AUC value in Naïve Bayes. In SVM, the highest AUC value is in the Stratified ROUS method, whereas in GLM, the highest AUC value is in the Stratified SMOTE method (see supplementary table). Each classification method has different procedures in determining the classification model to give different results. In the data of poor households in Yogyakarta Province, SVM is the classification method that produces the highest AUC value, however the most stable method is GLM.

Figure 5 shows the performance measures based on classification methods and over sampling methods. The value of AUC and the sensitivity value in over sampling method increased compared to baseline data. However, there is a decreasing in accuracy and specificity in over sampling method compared to baseline data. It is necessary to do data cleaning technique after over sampling process so that the increase of sensitivity value does not decrease accuracy and specificity value. Meanwhile, it is clear that the application of stratification on over sampling method can improve the accuracy of classification in over sampling method.

**Figure 5.** Performance Measures by Classifiers and Over Sampling Methods
5. Conclusion
The simulation results show that in the unbalanced data caused a bias on the parameter estimation. The bias will get higher as the data imbalance gets higher and the data size is small. In addition to the large bias, the precision of classification will also decrease in the condition of unbalanced data, especially in small samples.

All over sampling methods are able to increase predictions in poor households in Yogyakarta province significantly and the use of urban-rural stratification can improve the prediction accuracy of poor households. Generally, over sampling method proved to improve the accuracy of classification, especially in minority classes, namely sensitivity and AUC. But on the other hand, it can also lead to decreased accuracy and specificity. In general, no dominant over-sampling method can provide the best results, so it is still open for further research to produce predictive improvement in the minority class without ignoring the accuracy of the majority class.

6. References
[1] Ramentol E, Caballero Y, Bello B and Herrera F 2011 SMOTE-RSB: A Hybrid Preprocessing Approach based on Oversampling and Undersampling for High Imbalanced Data-Sets using SMOTE and Rough Sets Theory. Knowledge and Information Systems (London: Springer) p 245
[2] Santoso B, Wijayanto H, Notodiputro KA and Sartono B 2017 Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : Review. IOP Conference Series : Earth and Environmental Science. 58.1-8.
[3] Batista G E A P A, Prati R C and Monard M C 2004 A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data, SIGKDD Exploration 6:1 p 20
[4] Tetko I, Livingstone D and Luik A 1995 Neural Network Studies : Comparison of Overfitting and Overtuning, Chemical Information & Computer Sciences p 826
[5] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE : Synthetic Minority Over Sampling Technique, Journal of Artificial Intelligence Research 16 p 321
[6] Han H, Wang W Y and Mao B H 2005 Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, ICIC05 Springer p 878
[7] Bunkhumpornpat C, Sinapiromsaran K and Lursinsap C 2009 Safe-level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem, PAKDD09 vol 5476 Springer p. 475
[8] Cieslak DA, Chawla NV, Striegel A 2006 Combating imbalance in network intrusion datasets. 732-737
[9] Ma L and Fan S 2017 CURE-SMOTE Algorithm and hybrid algorithm for feature selection and parameter optimization based on random forest. BMC Bioinformatics.
[10] Garcia V, Sanchez J S and Mollineda R A 2012 On the Effectiveness of Preprocessing Methods when dealing with different levels of class imbalance, Knowledge Based System vol 25 p 13
[11] Peng I, Wang X I, Yuan-chao 2007 A Classification method for imbalance data set based on hybrid strategy. Acta Electron Sin 35(11):2161-5
[12] He H, Bai Y, Garcia E A and Li S 2008 ADASYN: Adaptive Synthetic Sampling Approach for Imbalance Learning, IJCNN08 (Hongkong) p 1322
[13] Tang S and Chen S 2008 The Generation Mechanism of Synthetic Minority Class Examples, 5th Int. Conference on Information Technology and Applications in Biomedicine (ITAB) (China: Shenzhen) p 444-447
[14] 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 p 7
[15] Stefanowski J and Wilk S 2008 Selective pre-processing of imbalanced data for improving classification performance, 10th International Conference in Data Warehousing and Knowledge Discovery (Springer)
[16] Kubat M and Matwin S 1997 Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, *14th International Conference on Machine Learning (ICML97)* (USA: Tennessee) p 179

[17] Komarek P and Moore A W 2005 Making Logistic Regression A Core Data Mining Tool, School of Computer Science.

[18] Rianto H and Wahono R S 2015 Resampling Logistic Regression untuk penanganan ketidakseimbangan class prediksi cacat software. *Journal of Software Engineering*.

[19] Hosmer dan Lemeshow 2000 *Applied Logistic Regression 2nd Edition*, John Willey and Son, Inc., New York.
## Supplementary table

### Table 1 Mean Intercepts ($\beta_0$)

| R    | n   | $\pi=0.5$ (0) | $\pi=0.2$ (-3) | $\pi=0.1$ (-5) | $\pi=0.05$ (-7) | $P=0.01$ (-9) |
|------|-----|---------------|----------------|----------------|----------------|---------------|
| 500  | 100 | 0.005691      | -3.41194       | -19.0678       | -184.464       | -6.6E+12      |
| 500  | 500 | 0.003069      | -3.03422       | -5.17045       | -7.53668       | -28.3134      |
| 500  | 750 | 0.003154      | -3.03645       | -5.12709       | -7.28554       | -14.3023      |
| 1,000| 1,000| 0.000392     | -3.02675       | -5.08804       | -7.20594       | -9.65386      |
| 2,500| 2,500| -0.00126     | -3.00491       | -5.03568       | -7.10378       | -9.22308      |
| 5,000| 5,000| -0.00159     | -2.99886       | -5.02057       | -7.03235       | -9.13302      |
| 7,500| 7,500| 0.004863    | -2.99988       | -5.00431       | -7.02981       | -9.06468      |
| 10,000| 10,000| 0.000696     | -3.00365       | -5.01309       | -7.01465       | -9.05318      |
| 1,000| 100  | 0.01857      | -3.45584       | -18.5084       | -165.404       | -1568.93      |
| 1,000| 500  | -0.0037      | -3.0809        | -5.16498       | -7.47527       | -27.3375      |
| 1,000| 750  | 0.00195      | -3.03687       | -5.094         | -7.27565       | -10.4883      |
| 2,500| 1,000| -0.00064     | -3.0192        | -5.06422       | -7.20207       | -9.86111      |
| 2,500| 2,500| -0.00157     | -3.0063        | -5.02762       | -7.11574       | -9.29462      |
| 2,500| 5,000| 0.001851     | -3.00286       | -5.01874       | -7.03625       | -9.14672      |
| 2,500| 7,500| 0.002457     | -3.00633       | -5.00725       | -7.0217        | -9.09407      |
| 2,500| 10,000| 0.002267    | -3.00777       | -5.00746       | -7.02838       | -9.06599      |
| 2,500| 100  | 0.004366     | -3.41602       | -24.0463       | -235.564       | -1.2E+12      |
| 2,500| 500  | 0.006049     | -3.05944       | -5.16134       | -7.50539       | -38.391       |
| 2,500| 750  | -0.00111     | -3.03943       | -5.10736       | -7.3038        | -14.1865      |
| 2,500| 1,000| -0.00098     | -3.02769       | -5.08455       | -7.21631       | -9.74092      |
| 2,500| 2,500| 0.000807     | -3.01723       | -5.03116       | -7.09271       | -9.23352      |
| 2,500| 5,000| 0.000604     | -3.00831       | -5.01503       | -7.04477       | -9.13974      |
| 2,500| 7,500| 0.000145     | -3.00354       | -5.01039       | -7.02173       | -9.06768      |
| 2,500| 10,000| -0.00016    | -3.00399       | -5.01002       | -7.02129       | -9.07097      |
Table 2. Mean Slopes ($\beta_1$)

| $R$  | $n$ | $\pi=0.5$ | $\pi=0.2$ | $\pi=0.1$ | $\pi=0.05$ | $P=0.01$ |
|------|-----|-----------|-----------|-----------|-----------|----------|
| 500  | 100 | 2.07493  | 2.28293  | 7.377008  | 51.18712  | 3.23E+12 |
| 500  | 500 | 2.028379 | 2.010666 | 2.080134  | 2.155301  | 6.023661 |
| 500  | 750 | 2.022147 | 2.02416  | 2.047327  | 2.100374  | 3.277496 |
| 1.000| 100 | 2.013097 | 2.027413 | 2.03144   | 2.08102   | 2.155915 |
| 1.000| 500 | 2.007763 | 2.005047 | 2.014375  | 2.037401  | 2.051062 |
| 1.000| 500 | 2.004441 | 2.000036 | 2.012333  | 2.02414   | 2.02448  |
| 1.000| 750 | 2.004849 | 1.999954 | 2.004363  | 2.011618  | 2.012459 |
| 1.000| 1000| 2.000528 | 2.00193  | 2.003134  | 2.00349   | 2.017965 |
| 1.000| 100 | 2.192841 | 2.297625 | 9.2257    | 47.80513  | 511.2142 |
| 1.000| 500 | 2.032058 | 2.058041 | 2.071574  | 2.14977   | 5.832223 |
| 1.000| 750 | 2.026163 | 2.023081 | 2.040823  | 2.087543  | 2.363052 |
| 1.000| 100 | 2.014577 | 2.013286 | 2.032178  | 2.069512  | 2.184043 |
| 1.000| 500 | 2.00658  | 2.005687 | 2.016784  | 2.032858  | 2.072915 |
| 1.000| 750 | 2.005232 | 2.003884 | 2.009619  | 2.011715  | 2.048652 |
| 1.000| 1000| 2.002807 | 2.005749 | 2.002683  | 2.007079  | 2.017387 |
| 1.000| 100 | 2.192841 | 2.297625 | 9.2257    | 47.80513  | 511.2142 |
| 1.000| 500 | 2.032058 | 2.058041 | 2.071574  | 2.14977   | 5.832223 |
| 1.000| 750 | 2.026163 | 2.023081 | 2.040823  | 2.087543  | 2.363052 |
| 1.000| 100 | 2.014577 | 2.013286 | 2.032178  | 2.069512  | 2.184043 |
| 1.000| 500 | 2.00658  | 2.005687 | 2.016784  | 2.032858  | 2.072915 |
| 1.000| 750 | 2.005232 | 2.003884 | 2.009619  | 2.011715  | 2.048652 |
| 1.000| 1000| 2.002807 | 2.005749 | 2.002683  | 2.007079  | 2.017387 |
| 2.500| 100 | 2.185362 | 2.308378 | 10.1219   | 48.43391  | 7.1E+11  |
| 2.500| 500 | 2.03891  | 2.044352 | 2.067146  | 2.150713  | 8.470388 |
| 2.500| 750 | 2.025478 | 2.029084 | 2.041994  | 2.088894  | 3.067979 |
| 2.500| 100 | 2.009674 | 2.015507 | 2.036533  | 2.063955  | 2.170248 |
| 2.500| 500 | 2.002627 | 2.013826 | 2.014692  | 2.025444  | 2.060076 |
| 2.500| 750 | 2.002969 | 2.005296 | 2.00628   | 2.01364   | 2.03265  |
| 2.500| 1000| 2.001537 | 2.003498 | 2.00544   | 2.008027  | 2.013663 |
| 2.500| 100 | 2.001983 | 2.002922 | 2.003497  | 2.005433  | 2.016726 |
Table 3. Mean Slopes ($\beta_2$)

| R    | n   | $\pi=0.5$ | $\pi=0.2$ | $\pi=0.1$ | $\pi=0.05$ | $P=0.01$ |
|------|-----|-----------|-----------|-----------|------------|----------|
| 500  | 100 | 3.317464  | 3.450494  | 11.30377  | 85.94024   | 1.9E+12  |
|      | 500 | 3.059715  | 3.027923  | 3.117048  | 3.257927   | 9.312972 |
|      | 750 | 3.041977  | 3.040837  | 3.085866  | 3.113785   | 4.935434 |
| 1.000| 100 | 3.023825  | 3.036296  | 3.046237  | 3.082961   | 3.235466 |
|      | 500 | 3.015471  | 3.001738  | 3.021854  | 3.050553   | 3.08138  |
|      | 750 | 3.00634   | 3.002722  | 3.014624  | 3.011797   | 3.05283  |
| 1.000| 100 | 3.007345  | 2.998268  | 3.003602  | 3.013912   | 3.02247  |
|      | 500 | 2.990077  | 3.001353  | 3.008486  | 3.004722   | 3.009408 |
| 2.500| 100 | 3.324896  | 3.465245  | 10.66984  | 73.25643   | 660.5024 |
|      | 500 | 3.051778  | 3.072096  | 3.107549  | 3.220332   | 9.677452 |
|      | 750 | 3.037183  | 3.039288  | 3.068692  | 3.126216   | 3.500576 |
| 1.000| 100 | 3.024531  | 3.021123  | 3.036344  | 3.087298   | 3.317656 |
|      | 500 | 3.014715  | 3.007095  | 3.013352  | 3.049676   | 3.108698 |
|      | 750 | 3.008323  | 3.007012  | 3.010509  | 3.015634   | 3.044052 |
|      | 1000| 3.003282  | 3.008466  | 3.008492  | 3.008466   | 3.040283 |
| 2.500| 100 | 3.281134  | 3.444309  | 14.91598  | 91.62312   | 5.94E+11 |
|      | 500 | 3.05346   | 3.05877   | 3.104231  | 3.238292   | 13.54139 |
|      | 750 | 3.033853  | 3.045361  | 3.076257  | 3.142492   | 4.818249 |
| 1.000| 100 | 3.017863  | 3.021949  | 3.053823  | 3.10095    | 3.258656 |
|      | 2.500| 3.004788  | 3.015933  | 3.018574  | 3.044723   | 3.084399 |
|      | 5.000| 3.00341   | 3.008065  | 3.008742  | 3.022458   | 3.050966 |
|      | 7.500| 3.004337  | 3.004177  | 3.006654  | 3.012442   | 3.025519 |
|      | 10.000| 3.002272  | 3.003143  | 3.006575  | 3.010295   | 3.028062 |
Table 4. Performance Measures of Baseline and Synthetic Over Sampling (simulated data)

| Phi | N   | Baseline |          |          |          |          |          |          |          |
|-----|-----|----------|----------|----------|----------|----------|----------|----------|----------|
|     |     | Accuracy | Sensitivity | Specificity | AUC | Accuracy | Sensitivity | Specificity | AUC |
| 0.5 | 100 | 0.84     | 0.7       | 0.93      | 0.817   | 0.84     | 0.7        | 0.93      | 0.817   |
|     | 500 | 0.82     | 0.8       | 0.83      | 0.815   | 0.82     | 0.8        | 0.83      | 0.815   |
|     | 750 | 0.84     | 0.8       | 0.83      | 0.84    | 0.83     | 0.8        | 0.84      | 0.834   |
|     | 1000| 0.84     | 0.86      | 0.82      | 0.837   | 0.84     | 0.85       | 0.82      | 0.837   |
|     | 2500| 0.85     | 0.86      | 0.85      | 0.852   | 0.85     | 0.85       | 0.85      | 0.85    |
|     | 5000| 0.86     | 0.86      | 0.86      | 0.865   | 0.87     | 0.86       | 0.87      | 0.867   |
|     | 7500| 0.87     | 0.87      | 0.87      | 0.872   | 0.87     | 0.88       | 0.86      | 0.873   |
|     | 10000| 0.86    | 0.86      | 0.86      | 0.861   | 0.86     | 0.86       | 0.86      | 0.862   |
| 0.2 | 100 | 0.97     | 0.43      | 0.97      | 0.702   | 0.76     | 0.74       | 0.87      | 0.87    |
|     | 500 | 0.88     | 0.67      | 0.94      | 0.803   | 0.88     | 0.85       | 0.88      | 0.87    |
|     | 750 | 0.94     | 0.8       | 0.94      | 0.87    | 0.86     | 0.85       | 0.85      | 0.856   |
|     | 1000| 0.88     | 0.67      | 0.94      | 0.808   | 0.88     | 0.82       | 0.89      | 0.855   |
|     | 2500| 0.9      | 0.7       | 0.94      | 0.823   | 0.88     | 0.87       | 0.88      | 0.877   |
|     | 5000| 0.89     | 0.69      | 0.95      | 0.822   | 0.88     | 0.84       | 0.89      | 0.866   |
|     | 7500| 0.88     | 0.69      | 0.95      | 0.818   | 0.86     | 0.83       | 0.87      | 0.852   |
|     | 10000| 0.89   | 0.69      | 0.95      | 0.819   | 0.87     | 0.84       | 0.88      | 0.859   |
| 0.1 | 100 | 0.96     | 0        | 1         | 0.5     | 0.96     | 1         | 0.95      | 0.979   |
|     | 500 | 0.94     | 0.69      | 0.97      | 0.833   | 0.9      | 0.77       | 0.92      | 0.844   |
|     | 750 | 0.94     | 0.55      | 0.99      | 0.769   | 0.88     | 0.7        | 0.9       | 0.802   |
|     | 1000| 0.98     | 0.74      | 0.99      | 0.866   | 0.95     | 0.89       | 0.95      | 0.924   |
|     | 2500| 0.95     | 0.61      | 0.98      | 0.797   | 0.91     | 0.82       | 0.92      | 0.871   |
|     | 5000| 0.94     | 0.54      | 0.98      | 0.76    | 0.88     | 0.77       | 0.89      | 0.834   |
|     | 7500| 0.93     | 0.61      | 0.97      | 0.793   | 0.89     | 0.87       | 0.89      | 0.88    |
|     | 10000| 0.94   | 0.6       | 0.98      | 0.788   | 0.89     | 0.84       | 0.9       | 0.871   |
| 0.05| 500 | 0.94     | 0.69      | 0.97      | 0.833   | 0.94     | 1         | 0.93      | 0.966   |
|     | 750 | 0.97     | 0.6       | 0.98      | 0.792   | 0.96     | 1         | 0.96      | 0.981   |
|     | 1000| 0.98     | 0.5       | 1         | 0.866   | 0.96     | 1         | 0.96      | 0.982   |
|     | 2500| 0.96     | 0.45      | 0.99      | 0.718   | 0.92     | 0.79       | 0.92      | 0.858   |
|     | 5000| 0.97     | 0.38      | 0.99      | 0.691   | 0.94     | 0.83       | 0.94      | 0.883   |
|     | 7500| 0.97     | 0.48      | 0.99      | 0.734   | 0.92     | 0.8        | 0.92      | 0.863   |
|     | 10000| 0.98   | 0.39      | 0.99      | 0.696   | 0.93     | 0.88       | 0.94      | 0.91    |
| 0.01| 750 | 0.98     | 0.25      | 1         | 0.625   | 0.95     | 1         | 0.95      | 0.975   |
|     | 1000| 0.99     | 0        | 1         | 0.5     | 0.98     | 0.5       | 0.99      | 0.744   |
|     | 2500| 0.99     | 0        | 0.99      | 0.502   | 0.94     | 0.75       | 0.94      | 0.845   |
|     | 5000| 0.99     | 0.23      | 0.99      | 0.615   | 0.95     | 0.85       | 0.96      | 0.901   |
|     | 7500| 0.99     | 0.35      | 0.99      | 0.674   | 0.93     | 0.95       | 0.93      | 0.942   |
|     | 10000| 0.99    | 0.39      | 0.99      | 0.696   | 0.96     | 0.92       | 0.96      | 0.939   |
Table 5. Comparison of Accuracy, Sensitivity, Specificity and AUC with KNN Classifier (Poor Household Data)

| Method               | Accuracy | Sensitivity | Specificity | AUC   |
|----------------------|----------|-------------|-------------|-------|
| Baseline             | 0.82     | 0.10        | 0.94        | 0.519 |
| ROS                  | 0.87     | 0.25        | **0.96**    | 0.605 |
| Stratified ROS       | 0.84     | 0.36        | 0.91        | 0.639 |
| ROUS                 | 0.86     | 0.27        | 0.95        | 0.611 |
| Stratified ROUS      | 0.81     | 0.43        | 0.87        | 0.648 |
| SMOTE                | 0.85     | 0.22        | **0.96**    | 0.589 |
| Stratified SMOTE     | 0.85     | 0.25        | 0.95        | 0.600 |
| BLSMOTE              | 0.69     | 0.8         | 0.68        | **0.739** |
| Stratified BLSMOTE   | 0.77     | 0.59        | 0.79        | 0.691 |
| SLSMOTE              | 0.64     | **0.81**    | 0.62        | 0.713 |
| Stratified SLSMOTE   | 0.74     | 0.67        | 0.76        | 0.715 |

Table 6. Performance Measures by NaïveBayes Classifier (Poor Household Data)

| Method               | Accuracy | Sensitivity | Specificity | AUC   |
|----------------------|----------|-------------|-------------|-------|
| Baseline             | **0.86** | 0           | 1           | 0.5   |
| ROS                  | 0.49     | **0.82**    | 0.44        | 0.629 |
| Stratified ROS       | 0.51     | **0.82**    | 0.46        | **0.639** |
| ROUS                 | 0.54     | 0.77        | 0.51        | 0.638 |
| Stratified ROUS      | 0.61     | 0.66        | 0.60        | 0.632 |
| SMOTE                | 0.64     | 0.63        | 0.64        | 0.634 |
| Stratified SMOTE     | 0.64     | 0.63        | 0.64        | 0.634 |
| BLSMOTE              | 0.61     | 0.66        | 0.61        | 0.633 |
| Stratified BLSMOTE   | 0.61     | 0.65        | 0.61        | 0.629 |
| SLSMOTE              | 0.65     | 0.62        | **0.65**    | 0.635 |
| Stratified SLSMOTE   | 0.64     | 0.63        | 0.64        | **0.639** |

Table 7. Performance Measures by SVM Classifier (Poor Household Data)

| Method               | Accuracy | Sensitivity | Specificity | AUC   |
|----------------------|----------|-------------|-------------|-------|
| Baseline             | **0.86** | 0           | 1           | 0.538 |
| ROS                  | 0.55     | **0.82**    | 0.51        | 0.711 |
| Stratified ROS       | 0.58     | 0.79        | 0.55        | 0.713 |
| ROUS                 | 0.59     | 0.76        | 0.57        | 0.714 |
| Stratified ROUS      | 0.6      | 0.74        | 0.58        | **0.718** |
| SMOTE                | 0.74     | 0.51        | **0.78**    | 0.702 |
| Stratified SMOTE     | 0.73     | 0.59        | 0.76        | 0.712 |
| BLSMOTE              | 0.63     | 0.69        | 0.62        | 0.715 |
| Stratified BLSMOTE   | 0.63     | 0.71        | 0.62        | 0.705 |
| SLSMOTE              | 0.72     | 0.56        | 0.75        | 0.707 |
| Stratified SLSMOTE   | 0.73     | 0.6         | 0.76        | 0.708 |
Table 8. Performance Measures by GLM Classifier
(Poor Household Data)

| Method              | Accuracy | Sensitivity | Specificity | AUC   |
|---------------------|----------|-------------|-------------|-------|
| Baseline            | 0.86     | 0           | 1           | 0.5   |
| ROS                 | 0.64     | 0.67        | 0.64        | 0.651 |
| Stratified ROS      | 0.64     | 0.67        | 0.63        | 0.652 |
| ROUS                | 0.64     | 0.67        | 0.64        | 0.653 |
| Stratified ROUS     | 0.66     | 0.64        | 0.66        | 0.652 |
| SMOTE               | 0.75     | 0.48        | 0.80        | 0.637 |
| Stratified SMOTE    | 0.74     | 0.56        | 0.77        | 0.662 |
| BLSMOTE             | 0.66     | 0.63        | 0.67        | 0.649 |
| Stratified BLSMOTE  | 0.67     | 0.63        | 0.67        | 0.65  |
| SLSMOTE             | 0.75     | 0.48        | 0.79        | 0.637 |
| Stratified SLSMOTE  | 0.75     | 0.51        | 0.79        | 0.648 |