Rattanakosin Journal of Science and Technology: RJST
2020, Volume 2 Issue 3: 73-88
Research Article
Online ISSN:2673-0553

Dealing with Imbalanced Data for Forecasting Model of Higher Education Selection of Grade 9 Students in Muang District Nakhonsawan Province

Chidchanok Sookwongtanon[1], Satriwit Chaleekorn[2] and Rungruttikarn Moungmai[3*]

Abstract

Nowadays big data has become more and more active because most organisations need to review their organisational performance in the past years and to seek impact factors both strengths and weaknesses for problem solving, operation plan and organisational development including the future prediction. A problem of big data is imbalanced data resulting in less predictive accuracy. Therefore, the study aims to create a proper model of multinomial logistic regression and to compare the predictive accuracy of the models between imbalanced data and balanced data using random over-sampling method and random under-sampling for correcting the data balancing. A real dataset of higher education selection of grade 9 students in Nakhonsawan, 2,692 students, was applied. It was found that main impact factor on students' decision was demographic. Moreover, the results revealed that the accuracy of forecasting models were more than 60% whereas the predictive accuracy of balanced data was better than another.

Keywords: forecasting, multinomial logistic regression, imbalanced data, predictive accuracy

[1,2,3] Department of Mathematics and Statistics, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Nakhonsawan, Thailand
Emails: [1]Chidchanok.S@nsru.ac.th, [2]Satrawit26@gmail.com
*Corresponding author, rungruttikarn.m@nsru.ac.th

1. INTRODUCTION

Big data is a huge dataset of information and is very important to all organisations [1]. For example, the government can analyse data, plan for management in various areas such as demography, education, economic development, formulation of strategies for national development, and natural resource management. For private entrepreneurs, getting customer information will help them gain insights into customer behaviour and make them better understand their customers. Also, this can be used to plan marketing, reach the distribution channels and help them expand their customer base quickly. Moreover, they can bring information to develop products or services to have more efficiency and can also be used to analyse future trends. Big data is very complex and diverse and may be classified into subgroups that each group might consist of a different amount of information.

A problem of big data analysis is imbalanced data which is not easy to handle. Since the dataset consists of variety numbers of each group that include both large number and small number of data. The large number of data groups might obstruct the properties of the small number of data groups hence the nature of the small group cannot be expressed [2]. For example, information on return treatment in hospitals of diabetes patients which were categorised into three groups, based on the nature of the recurrence at the hospital, that were (1) patients who did not return treatment or no disease, 54%, (2) patients who returned to treatment within 30 days of the last treatment, 11%, and (3) patients who returned treatment for more than 30 days from the last treatment, 35%. For analysing this data, it should give equal importance to all data groups. It is also less efficient than analysing data with imbalance correction. Therefore, the researchers were interested in how to correct the imbalanced data using the real dataset of the selection of higher education of grade 9 students in Muang district, Nakhonsawan province.

At the present, Thai education system is divided into four levels that are (1) grade 1 to 3 level, (2) grade 4 to 6 level, (3) junior high school or grade 7 to 9 level, and (4) senior high school or grade 10 to 12 level. In which the senior high school level, it is divided into two groups that are general education and vocational education. Level 1 to 3 of the vocational education is equivalent to the upper secondary or high school level [3]. Moreover, each school is divided into four groups based on the number of students in each school. These are small school, fewer than 499 students, medium school, 499 – 1,500 students, large school, 1,500 – 2,500 students, and extra-large school, 2,500 students or more. Not only the number of students in each group is different but factors and opportunities of each group are also different such as quality and number of teachers, school and building conditions, learning materials, and budget. These factors might affect student interest and opinions on their further education [4].

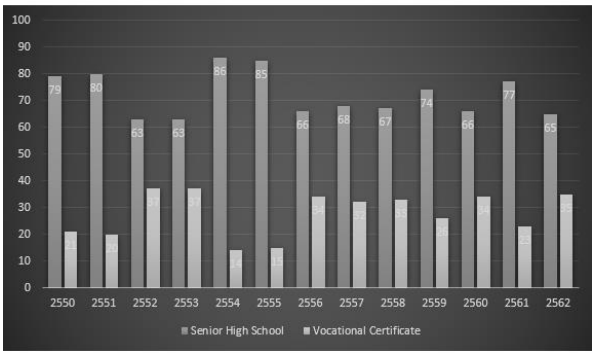

Figure 1 : The percentage of students who studied in senior high school and vocational education during the past 13 year.

Figure 1 presented Thailand's educational statistics from year 2007 to 2019, these number were from the Central Education Information System, Technology and Communication Center, Office of the Permanent Secretary for Education [5]. It was found that the percentage of high school students is more than half (7 years) with more than one third of the total number of students pursuing vocational education.

In addition, Figure 2 stated the number of students pursuing general and vocational educations. These numbers were from Strategy and Information division, Nakhonsawan provincial office [6]. It was found that the number of students studying in the ordinary field decreased while the number of vocational students has increased.

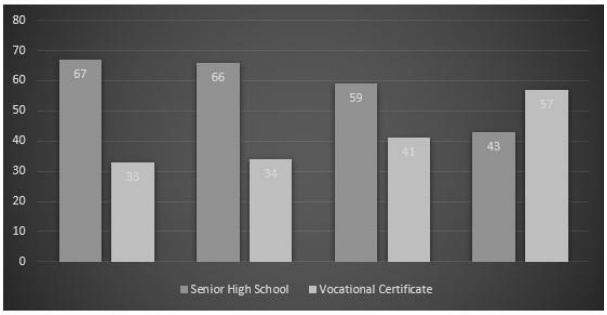

Figure 2 : The percentage of students who attended the general education senior high school and vocational education in Nakhonsawan province.

Figure 2 also showed that the trend of choosing to study in the ordinary high school decreased while choosing to study in vocational education increases. This affected the decreasing number of Science students. The researchers were therefore interested in studying trends and constructing a multinomial logistic regression model to predict the selection of higher education of grade 9 students in Muang district, Nakhonsawan province and to compare the predictive accuracy between imbalanced data model and balanced data models using two methods, Random Over-sampling (ROS) and Random Under-sampling (RUS).

2. IMBALANCED DATA

Sometimes, a dataset or variable value is classified into many subgroups where a number of each subgroup is not distributed the same or each group (class) does not have the same or different amount. For instance, customers were divided into two groups based on payment method that are cash and credit card. An accounting staff found that 70% of a total customers paid by cash while the rest customers paid by credit card. It can be seen that there was only two-third of the total customers paid by cash. This imbalanced dataset might affect some issues such as opinion survey and forecasting model. The finding might be biased into the majority data. Therefore, such problem cannot be ignored. Also, imbalanced data modification might gain more accuracy of the prediction. There are four popular methods to adjust the balance of data as follows [7].

1. Random Over-sampling (ROS) is a technique to balance the number of samples in a group(s) by adding some samples into the group with the smallest amount of data (minor group) until its amount is in balance with the group with the largest amount of data (major group).

2. Random Under-sampling (RUS) is a technique for reducing the number of samples by randomly deleting some samples in the major group until the distribution of the group is as balanced as the minor group.

3. Hybrid method is a combination method between over-sampling and under-sampling.

4. Synthetic minority over-sampling technique (SMOTE) is a technique used to solve the problem of classification of imbalanced data because each class has a different amount of data. Therefore, the SMOTE method will synthesize new data based on the old data and add them to the less informed group causing the distribution of the information in each group is more balanced. There are several methods of synthesis of new information, such as k-Nearest Neighbor and Genetic Algorithm (GA).

3. MULTINOMIAL LOGISTIC REGRESSION (MLR)

Logistic Regression is a forecasting technique to study the relationship between an independent variable(s) and one qualitative dependent variable or to study whether there is any independent variable(s) that can explain the variation of the dependent variable. In this study, the dependent variable was classified into three groups therefore multinomial logistic regression (MLR) was applied. MLR is also called Multi-equation model as there are many forecasting equations [8].

1.Modelling Construction

Let k be a number of groups of a dependent variable,
p be a number of independent variables (X),
n be a number of samples,
P_y = P(y = k) be a probability that the dependent variable is in group k (focus group), and
Q_y = Q(y = 0) be a probability of others that is not in a focus group.

Since the dependent variable is a category with k levels hence to conduct a forecasting model, k – 1 logit equations will be created and, then, these will be compared with a baseline. Firstly, P_y can be written as

$$P(y = k) = \frac{e^Z}{1+e^Z} \text{ or } \frac{1}{1+e^{-Z}} \qquad (1)$$

and

$$Q(y = 0) = 1 - P(y = k)$$

hence $Q(y = 0) = \frac{1}{1+e^Z}$.

Next, a ratio between P_y and Q_y, called odds, will be presented as

$$\text{Odds}\left(\frac{P(y=1)}{Q(y=0)}\right) = \frac{\frac{e^Z}{1+e^Z}}{\frac{1}{1+e^Z}} = e^Z \qquad (2)$$

After that, logit or log odds can be found from taking log into the equation (2) so

$$\text{logit} = \text{log odds} = Z$$

where

$$Z = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \qquad (3)$$

It can be seen that the logit in equation (3) is similar to the multiple linear regression model hence parameters (β_j) can be estimated by maximum likelihood estimation.

In this study, the dependent variable, choosing to study high school, was classified into three groups as follows:

y = 1 means choosing to study in Science,
y = 2 means choosing to study in Arts,
y = 3 means choosing to study in vocational education.

Therefore, to construct the MLR model, a group of y = 1 was set as the baseline and two logit equations were created as

$$\log\left(\frac{P(y=2)}{Q(y=0)}\right) = \hat{\beta}_0 + \hat{\beta}_{12}X_1 + \hat{\beta}_{22}X_2 + \cdots + \hat{\beta}_{p2}X_p \qquad (4)$$

and

$$\log\left(\frac{P(y=3)}{Q(y=0)}\right) = \hat{\beta}_0 + \hat{\beta}_{13}X_1 + \hat{\beta}_{23}X_2 + \cdots + \hat{\beta}_{p3}X_p \qquad (5)$$

2. Forecasting Interpretation

Due to there are two logit models, therefore the predicted value ($\hat{Y}_i$) has to be interpreted together as follows.

$\hat{Y}_I = 2$ if the logit or equation (4) is greater than 0.5.

$\hat{Y}_I = 3$ if the logit or equation (4) is less than or equal to 0.5 and the logit or equation (5) is greater than 0.5.

$\hat{Y}_I = 1$ if the logit or equation (4) is less than or equal to 0.5 and logit or equation (5) is less than or equal to 0.5.

3. Parameters Estimation

The logit in equation (3) or multiple linear regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{11} + \cdots + \beta_p X_{pi} + \varepsilon_i \qquad (6)$$

The parameters can be estimated from samples, so

$$\hat{Y}_i = b_0 + b_1 X_{11} + \cdots + b_p X_{pi} \qquad (7)$$

Because the data used in this study is a sample group hence the approximation of parameters start with setting Y_i, X_{ji}, and β_j metrics as follows:

$$Y_i = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}$$

where i = 1, 2, …, n,

$$X_{ji} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{p1} \\ 1 & X_{21} & \ldots & X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & \ldots & X_{pn} \end{bmatrix}_{n\times(p+1)}$$

where j = 1, 2, …, p, and

$$\hat{\beta}_j = b_j = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}_{(p+1)\times 1} .$$

Then, the parameters can be estimated from

$$b_j = (X'X)^{-1}X'Y \qquad (8)$$

After approximating the parameter estimators, the test of Wald statistic will be used to evaluate whether the parameter is appropriate in the model. The statistic can be calculated from

$$W = \left[\frac{b_j}{SE(b_j)}\right]^2 \qquad (9)$$

4. Suitability Investigation of the Model

To investigate the proposed model, -2LL and Hosmer and Lemeshow test will be used to assess whether such model is appropriate.

4.1 Considering the value of -2LL

The value of – 2LL or -2 log likelihood is firstly assessed. Considering the -2LL value of each model, if such of the final model is the lowest value, the logistic equation is most appropriate. Moreover, this test can be considered by using the Chi-square model test (χ^2 – test) at df = p. If the test is statistically significant or accepts H_1, then the ith independent variable of the nth data set can be used to predict the likelihood of occurrence of an event of interest (y = 1) by belief (1 – α)100%.

4.2 Considering Hosmer and Lemeshow test

To evaluate whether the proposed model is appropriate, a null hypothesis (H_0) of suitable model is tested using Hosmer and Lemeshow test.

Let G be the number of groups,

n_g be the number of observations for the **g**th group, **O**g be the observed events,

Eg be the expected events.

The test statistic follows a Chi-squared distribution with (G-2) degrees of freedom and the test statistics can be written as

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)} \qquad (10)$$

In the test, if χ^2 is not statistically significant or accepts H_0, then the proposed model is suitable.

5. Accuracy Calculation

An accuracy value presents the accuracy of the prediction of the model. There are many statistics such as ROC, MAPE and RMSE. In this study, an accuracy rate will be used.

Let TP be a number of events that are positive of both observed and predicted values,

FP be a number of events that observed values are negative and predicted values are positive,

FN be a number of events that observed values are positive and predicted values are negative,

TN be a number of events that are negative of both observed and predicted values.

These numbers can be presented as shown in Table 1

The accuracy rate can be calculated as

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (11)$$

Table 1. Amount of students classified by observed and predicted values.

		Observed value	
		positive	**negative**
Predicted value	**positive**	True positive (TP)	False positive (FP)
	negative	False negative (FN)	True negative (TN)

III. METHOD

To conduct this research, there were five steps as shown in Figure 3:

Step 1: Data Cleaning.
Evaluating the integrity of the raw dataset.

Step2: Descriptive Statistics.
Conducting an initial statistical analysis

Step 3: Model 1 Creation
Creating model of imbalanced data.

Step 4: Model 2 and 3 Creation
Modifying the data balancing by two methods.
(1) Random Over-sampling (Model 2).
(2) Random Under-sampling (Model 3).

Step 5: Accuracy rate
Comparing the predictive accuracy of all models.

4. RESULTS

1.Dataset

The population used in this study was grade 9 students who were studying in academic year 2019 in Muang district, Nakhonsawan, 3,062 students. They were classified into four groups based on their school size including small, medium, large, and extra-large schools. For the reliability and accuracy of the data analysis, students' data with missing information were ignored from the dataset. Hence, the remaining dataset consisted of 2,692 students.

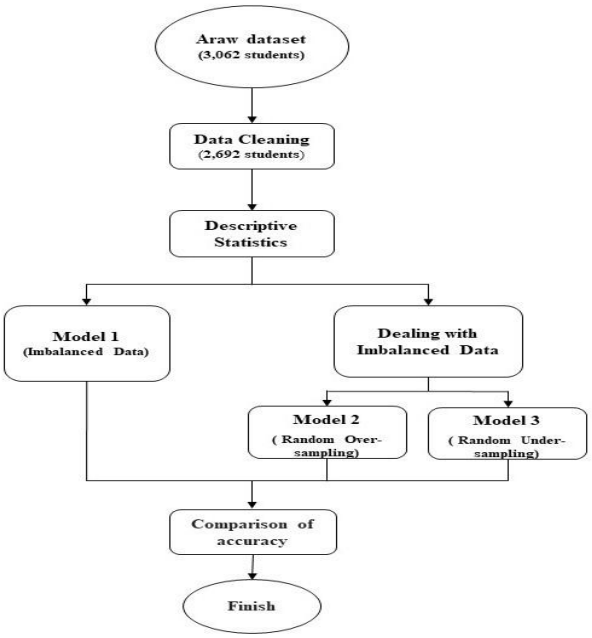

Figure 3 : The process of conducting research

In this study, there were twenty-three independent variables (p = 23) including school size, gender, GPA, parental status, family income, parents' occupations, parents' education, and student opinions that were rating scale with five levels.

2.General Information

Table 2. GPA, amount and percentage of samples classified by school size and choosing to study (Science, Arts and Vocational education).

Obs.	GPA	School size				Total
		XL	**L**	**M**	**S**	
Sci.	3.33	1,021 (37.93%)	91 (3.38%)	10 (0.37%)	34 (1.26%)	1,156 (42.94%)
Arts	2.94	568 (21.1%)	62 (2.3%)	11 (0.41%)	10 (0.36%)	651 (24.17%)

Obs.	GPA	School size				Total
		XL	**L**	**M**	**S**	
Voc.	2.62	478 (17.76%)	232 (8.62%)	82 (3.05%)	93 (3.46%)	885 (32.89%)
Total		2,067 (76.79%)	385 (14.3%)	103 (3.38%)	137 (5.08%)	2,692 (100%)

As shown in Table 2, the choosing to study of a sample group was selected in order of Science (42.94%), vocational education (32.89%), and Arts (24.17%) respectively. GPA of each group was 3.33, 2.62, and 2.94 respectively. In addition, it was found that the majority samples of students from extra large schools chose to study Science (37.93%) while the majority samples of students from large, medium and small schools chose to study in vocational education at 8.62%, 3.05% and 3.46% respectively.

The finding of student opinions found that there were four aspects of influences affecting the decision to choose to continue studying. These concluded personal reasons, persuasion from others, internal educational guidance, and the cost of education. The most influencing factors on decision-making were personal reasons, educational expenses, and internal educational guidance and persuasion from others. However, these influences had no different effect on choosing to study in Science, Arts and vocational education groups.

Before creating a model, relationship between the dependent variable and each independent variable was verified using Chi-square test. It was found that such correlation was statistically significant. Additionally, all qualitative independent variables were converted to dummy variables.

3.Imbalanced Data Model

To construct a multinomial logistic regression model, a total of 2,692 samples were analysed to created two logit models using a group of students who chose to study in Science (y = 1) as the baseline, the results were as follows. Model 3.1 can be written as

$$\mathrm{P(y=2)} = \frac{1}{1+e^{-(z)}}$$

where

$$\begin{aligned} Z = {} & 0.932x_1d_1 - 0.372x_1d_2 + 0.517x_1d_3 - 0.302x_2 - 1.591x_3 - 4.784x_{11}d_1 \\ & -4.613x_{11}d_2 - 5.065x_{11}d_3 - 2.153x_{11}d_4 - 0.897x_{17}d_1 - 0.805x_{17}d_2 \\ & -0.829x_{17}d_3 - 0.889x_{17}d_4 - 2.426x_{22}d_1 - 2.685x_{22}d_2 - 2.199x_{22}d_3 \\ & -1.803x_{22}d_4 \end{aligned} \quad (12).$$

Model 3.2 can be written as

$$\mathrm{P(y=3)} = \frac{1}{1+e^{-(z)}}$$

where

$$Z = -1.056x_1d_1 - 2.075x_1d_2 - 0.443x_1d_3 + 0.909x_2 - 2.019x_3 \quad (13).$$

Table 3. Model fitting information of imbalanced data model

Model	Model Fitting Criteria -2 Log Likelihood	Likelihood Ratio Tests			Pseudo R-Square	
		Chi-Square	df	Sig.	Cox and Snell	Nagel-kerke
1.Intercept Only	4040.09					
2.Final	2867.067	1173	196	0	0.46	0.525

As shown in Table 3, there were two models where model 1 consisted of only a constant in the model and model 2, the final model, included both constant and independent variables in the model. It can be seen that the value -2LL of model 2 was less than the other indicating that model 2 was the best model. It also stated Cox and Snell's R^2 and Nagelkerke R^2 values that were 46.35% and 52.5%, respectively. These can be used to describe the relationship between dependent and independent variables. These can be also said that independent variables can describe the dependent variable 46.35% and 52.5%, respectively.

Table 4 : Predictive accuracy of imbalanced data model.

Observed	Predicted			Percent Correct
	Science	Arts	Vocational education	
Science	651	70	90	80.30%
Arts	196	148	111	32.50%
Vocational education	105	61	453	73.20%
Overall Percentage	50.50%	14.80%	34.70%	66.40%

It was found from the Table 4 that an overall predictive accuracy of the proposed model was 66.4%. Considering each group of the dependent variable, it was found that the correct prediction of the choosing to study in Science, Arts, and vocation education were 80.30%, 32.50%, and 73.20% respectively.

4. Balanced Data Models

From previous section, the amount of each group of choosing to study was different. The largest number of samples or major group was choosing to study in Science, almost 50%, and the smallest number of samples or minor group was choosing to study in Arts, 24%. It can be noticed that the amount of the major group was twice of the minor group, resulting in the finding. The predicted values might be biased towards the majority group, choosing to study in Science. Also, the major group might obscure characteristics of other sample groups. Therefore, before creating a forecasting model, it was necessary to sort out the imbalanced data problem by adjusting the balance of the dataset. That was the amount of samples of both groups, Science and Arts, had to be the most similar amount of samples by using Random Over-sampling and Random Under-sampling methods. To modify the data balancing some data have to adding or cut off from the dataset. However, we cannot know that there is a better way to deal with unbalanced data than random increase and random reduction. Each randomized data can be changed. So if random increase and random decrease in the next round May cause the information to be changed because our sample selection is very random.

4.1 Random Over-sampling Model

This method will randomly select some data using simple sampling technique by increasing the number of data in the minor group to have the same number of data as the major group. In this study, the number of data of the existing minor group was 651 students while the number of data of the existing major group was 1,156 students. Therefore, the data of 505 students using simple random sampling technique from Arts had to be added into the minor group, resulting in the total number of samples equal to 3,197 students. A new dataset was divided into 3 groups as follows: 1,156 students who chose to continue in Science (36.15%), 1,156 students who chose to continue in Arts (36.15%) and 885 students who chose to continue in vocational education (27.68%). After that the new dataset was used to conduct the model using a group of students who chose to study in Science (y = 1) as the baseline, the results were as follows. Model 4.1.1 can be written as

$$P(y = 2) = \frac{1}{1+e^{-(z)}}$$

where

$$Z = -0.42x_2 - 1.685x_3 - 3.091x_{11}d_1 - 3.085x_{11}d_2 - 3.364x_{11}d_3 - 0.055x_{11}d_4 \qquad (14)$$

Model 4.1.2 can be written as

$$P(y = 3) = \frac{1}{1+e^{-(z)}}$$

where

$$Z = -1.137x_1d_1 - 2.191x_1d_2 - 0.579x_1d_3 + 0.745x_2 - 2.123x_3 + 1.323x_{18}d_1 + 1.201x_{18}d_2 + 1.254x_{18}d_3 + 1.027x_{18}d_4 - 1.698x_{23}d_1 - 1.699x_{23}d_2 - 1.610x_{23}d_3 - 1.756x_{23}d_4 \qquad (15).$$

Table 5 : Model fitting information of Random Over-sampling model.

Model	Model Fitting Criteria -2 Log Likelihood	Likelihood Ratio Tests			Pseudo R-Square	
		Chi-Square	df	Sig.	Cox and Snell	Nagel-kerke
1.Intercept Only	6974.535					
2.Final	5265.591	1708.943	198	0	0.414	0.467

As shown in Table 5, there were two models where model 1 consisted of only a constant in the model and model 2, the final model, included both constant and independent variables in the model. It can be seen that the value -2LL of model 2 was less than the other indicating that model 2 was the best model. It also stated Cox and Snell's R^2 and Nagelkerke R^2 values that were 41.4% and 46.7%, respectively. These can be used to describe the relationship between dependent and independent variables. These can be also said that independent variables can describe the dependent variable 41.4% and 46.7%, respectively.

It was found from the Table 6 that an overall predictive accuracy of the proposed model was 63.80%. Considering each group of the dependent variable, it was found that the correct prediction of the choosing to study in Science, Arts, and vocation education were 70.10%, 57.84%, and 63.50% respectively.

4.2 Under-sampling method

This method will randomly select some data using simple sampling technique by reducing the number of data in the major group to have the same number of data as the minor group. In this study, the number of data of the existing major group was 1,156 students while the number of data of the existing minor group was 651 students.

Table 6 : Predictive accuracy of Random Over-sampling model

Observed	Predicted			Percent Correct
	Science	Arts	Vocational education	
Science	810	239	107	70.10%
Arts	309	668	179	57.80%
Vocational education	119	204	561	63.50%
Overall Percentage	38.70%	34.80%	26.50%	63.80%

Therefore, the data of 505 students using simple random sampling technique from Science had to be randomly reduced from the major group, resulting in the total number of samples equal to 2,187 students. A new dataset was divided into 3 groups as follows: 651 students who chose to continue in Science (29.76%), 651 students who chose to continue in Arts (29.76%) and 885 people who chose to continue in vocational education (40.46%). As the previous method, the new dataset was used to conduct the model using a group of students who chose to study in Science (y = 1) as the baseline, the results were as follows. Model 4.2.1 can be written as

$$P(y = 2) = \frac{1}{1+e^{-(z)}}$$

where

$$Z = -0.401x_2 - 1.646x_3 - 4.367x_{11}d_1 - 4.333x_{11}d_2 - 4.664x_{11}d_3 - 1.043x_{11}d_4 + 1.054x_{18}d_1 + 1.220x_{18}d_2 + 1.068x_{18}d_3 + 0.932x_{18}d_4 \quad (16).$$

Model 4.2.2 can be written as

$$P(y = 3) = \frac{1}{1+e^{-(z)}}$$

where

$$Z = -1.272x_1d_1 - 2.578x_1d_2 - 0.622x_1d_3 + 0.070x_2 - 2.180x_3 + 1.472x_{18}d_1 + 1.436x_{18}d_2 + 1.384x_{18}d_3 + 1.280x_{18}d_4 \quad (17).$$

Table 7. Model fitting information of Random Under-sampling model.

Model	**Model Fitting Criteria**	**Likelihood Ratio Tests**			**Pseudo R-Square**	
	-2 Log Likeli-hood	**Chi-Square**	**df**	**Sig.**	**Cox and Snell**	**Nagel-kerke**
1.Intercept Only	4759.816					
2.Final	3498.707	1261.109	200	0	0.438	0.494

As shown in Table 7, there were two models where model 1 consisted of only a constant in the model and model 2, the final model, included both constant and independent variables in the model. It can be seen that the value -2LL of model 2 was less than the other indicating that model 2 was the best model. It also stated Cox and Snell's R^2 and Nagelkerke R^2 values that were 43.8% and 49.4%, respectively. These can be used to describe the relationship between dependent and independent variables. These can be also said that independent variables can describe the dependent variable 43.8% and 49.4%, respectively.

Table 8 : Predictive accuracy of Random Under-sampling model

Observed	Predicted			Percent Correct
	Science	Arts	Vocational education	
Science	447	107	98	68.60%
Arts	163	293	196	44.90%
Vocational education	81	120	683	77.30%
Overall Percentage	31.60%	23.80%	44.70%	65.00%

It was found from the Table 8 that an overall predictive accuracy of the proposed model was 65.00%. Considering each group of the dependent variable, it was found that the correct prediction of the choosing to study in Science, Arts, and vocation education were 68.60%, 44.90%, and 77.30% respectively.

4.3 comparison of imbalanced and balanced data models

Table 9 : Predictive accuracy of imbalanced data and balanced data models

Data management	Total	Choosing to study		
		Science	Arts	Vocational education
No data modifying	66.40%	80.30%	32.50%	73.20%
Modifying Using ROS	63.80%	70.10%	57.80%	63.20%
Modifying Using RUS	65.00%	68.60%	44.90%	77.30%

As displayed in the Table 9, the overall predictive accuracies of imbalanced data, ROS, and RUS models were 66.40%, 63.80%, and 65.00%, respectively. It can be noticed that the accuracy of each model was slightly different. Considering the correct prediction of each group of the dependent variable, it was found that some models were improved to increase the accuracy while some developed model caused a slight decrease of the accuracy.

The correct prediction of choosing to study in Science was found that after modifying the balance of the data, the correction of these models were slightly less than the correction of imbalanced data model. On the other hand, among the choosing to study in Arts group, the correction of balanced data models were greater than the correction of imbalanced data model almost two times. However, among the choosing to study in vocational education group, the correction of balanced data model using ROS method was reduced while the correction of the other balanced data model was slightly increase.

4.4 comparison with other work

Comparing the proposed model with the model of Poosumpa et al. [9], it was found that the proposed model accuracy was lower than the accuracy of such model. Although, see Table 9, the accuracy of each total model was lower than the model of Poosumpa et al., the accuracies of some groups were more than 70%. However, the dependent variable of the proposed model was divided into three groups, choosing to study in Science, Arts and Vocational education, whereas the dependent variable of the model of Poosumpa et al. was divided into 2 groups, choosing to study in Science and not choosing to study in Science. For further work, the model of Poosumpa et al. should be modified using ROS and RUS. Also, other methods to adjust the balance of data should be applied.

5. CONCLUSION

In this study, the researchers were interested in conducting a multinomial logistic regression model to predict the selection of higher education of grade 9 students in Muang district, Nakhonsawan province and to compare the predictive accuracy between imbalanced data and balanced data. To create such model, choosing to study in high school was set as the dependent variable which was divided into three 3 groups, choosing to study in Science, Arts, and vocational education.

It was found that factors affecting the selection of further study were gender, academic performance, and school size. There were also other factors that were (1) personal reasons, to have the knowledge and educational background to be used in their careers, (2) influence of persuasion from others, the persuasion of friends to choose further education, (3) the educational guidance within the institute, the recommendation of the homeroom teacher, and (4) the educational expenses and living expenses such as food, accommodation and fares, and book expenses.

In addition, the results revealed that the predictive accuracy of the balanced data models was not different from the imbalanced data models. However, ROS and RUS methods improved the correct prediction especially choosing to study in Arts group, yields up to 28%. Also, the predictive accuracy of the other groups were more than two-third of the total. Therefore, modifying the balance of data before constructing the forecasting model will give better predictive accuracy and will be more efficient than imbalanced data model. However, the accuracies of both imbalanced and balanced data models are not different. There might be other factors that influence the students' decision. Also, an appropriate number of data to be randomly added or reduced into the minor or major groups should be considered.

6. ACKNOWLEDGEMENT

Thank you Ms. Sarocha Poosumpa, Ms. Photchanat Sathongkhao, Ms.Tasanee Supprasert and Dr. Rungruttikarn Moungmai who provided information used in this research.

7. REFERENCES

[1] Chomboon K. (2015). Classication Techique Minority Class On Imbalanced Dataset With Data Partitioning Method. Master thesis, B.Eng., Suranaree University, Nakhon Ratchasima.

[2] Kesornsit W., Lorchirachoonkul V. and Jitthavech J.(2018).Imbalanced Data Problem Solving in Classification of Diabetes Patients.Master thesis, M.S., Khon Kaen University, Khon Kaen.

[3] Chomnok P. and Yanwiset C. Secondary education. Retrieved May 28, 2020, from https://sites.google.com/site/fihakhwamru/kar-suksa-radab-mathymsuksa

[4] Institute for the Promotion of Teching Science and Technology.Different school sizes. Retrieved January 11, 2021, from https://pisathailand.ipst.ac.th/issue-2019-41/?fbclid=IwAR0-Q-XWo8GA83wUPJXvEWHr7DQVtpEJwq2A6XzI0rgAWAxiEg7-3d4KVJg

[5] Bureau of information and communication technology and Office of the permanent secretary ministry of education. (2007-2019). Statistics on further study selection of Mathayomsuksa 3 students. Retrieved May 28, 2020, from http://www.bict.moe.go.th/2020/index.php

[6] Nakhonsawan Provincial Statistical Office. (2018-2019). Nakhon Sawan Provincial Statistical Report. Retrieved May 30, 2020, from http://nksawan.nso.go.th/index.php?option=com_content&view=article&id=335&Itemid=507

[7] Tongpool P., Jamrueng P., Boonrit R. and Sinsomboonthong S. Performance Comparison in Predietion of Imbalanced Data in Data Mining Classification. Master thesis, King Mongkut's Institute of Technology Ladkrabang, Bangkok.

[8] Woraphongsathorn T. (4/2/2018). Multiple Logistic Regression Analysis. Retrieved May 30, 2020, from http://oec.anamai.moph.go.th/download/OEC_2016/MEETTING2561/APRIL2561/2_5April2561/6-Multiple%20Logistic%20Regression%20Analysis.pdf

[9] Poosumpa S., Sathongkhao P. and Supprasert T. Forecasting Model for the Number of Grade 9 Students Who Intend to Study the Upper Secondary Level in Science, Muang District, Nakhonsawan Province