Quality Analysis of Graduates from Chiang Mai University Using Machine Learning Methods

Zihao Zhao[1] and Phisanu Chiawkhun [2]

[1] Data Science Consortium, Faculty of Engineering, Chiang Mai University, Thailand
[2] Statistics Department, Faculty of Science, Chiang Mai University, Thailand
Zhao_zihao@cmu.ac.th

Abstract. This research aims to measure the quality of Chiang Mai University graduates, construct the model which can predict income accurately and find the variables effected to Chiang Mai University graduates' quality. Through data collection and integration, we got the students' data in Chiang Mai University from academic year 2012-2014.Then we brought in data and used three machine learning models (artificial neural networks, logistic regression, and support vector machines) to perform multiple classifications. All three have relatively good prediction results with good accuracy. The results show us that the income of graduates of Chiang Mai University is normally distributed. Most of the graduates have a medium income, and a small number of people earn high and low incomes. The best way to increase income for students who have just entered university is to improve their English scores and choose medical-related majors. For senior students, choosing to study for a higher degree and maintaining a high GPA is a very effective way to increase their income.

Keywords: Machine learning, Explore data analysis Model prediction, Model evaluation.

1 Introduction

In Chiang Mai University, graduates' data will be entered into the files of the Registration Office (REG), Education Quality Development Office (EQD) and Information Technology Service Center (ITSC). We can apply for relevant data to the university for machine learning research through research. The first step of machine learning is data preprocessing, through a series of steps to get the data set we need. Next is data exploration and analysis, and a report on the initial data. Then we introduce the sorted data set into artificial neural network, logistic regression and support vector machine models. After the parameters are adjusted, we will evaluate the model using indicators such as accuracy. After getting a better model, you can use the model to evaluate and predict student data.

Data collected from academic year 2012-2014 graduates of Chiang Mai University. After screening, 9561 students' data were included in the dataset 1 for analysis of all students (Added degree level: bachelor, master and doctoral). 7306 students' data

were included in the dataset 2 for bachelor degree analysis. (Added the bachelor English score factor).

2 Literature Review

In this research, we read a lot of literature. According to different kinds of research methods and content, the study was conducted. First of all, in terms of data processing, we understand the data preprocessing methods of most data science researchers [1]. Use appropriate data preprocessing methods in different situations. Make appropriate adjustments according to the situation and content.

In the evaluation of graduate quality, related information related to data mining and big data. We can evaluate the data through the principles of data science and give corresponding evaluation indicators [2].

Artificial neural network is the most typical and widely used model in machine learning methods. We can use artificial neural networks to solve many types of prediction problems [3]. Its wide applicability and adjustability are the performance of artificial neural networks. Powerful reason. It can be used in many places such as evaluation [4].

Logistic regression is the basis of machine learning, which includes many basic statistical ideas. Logistic regression also has corresponding applications in machine learning. It is mostly used in traditional binary classification problems. Logistic regression focuses more on accuracy in machine learning [5].

Support vector machine is also one of the very typical machine learning methods. It has strong theoretical support, and its classification methods and functions are very suitable for binary classification problems. Support vector machines are more applicable than other machine learning methods in some situations [6].

3 Methodology

we evaluate and predict graduate income data through model building and parameter adjustment. The research method is divided into 5 parts. The relationship between the parts is as Fig.1 shows:

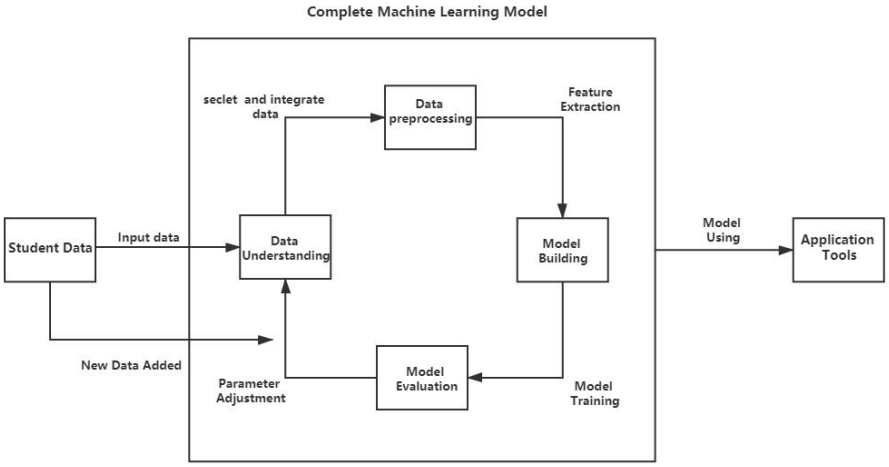

Fig. 1. Complete Machine Learning Model

3.1 Data Preprocessing

Data preprocessing refers to some kind of processing of data before the main processing. Due to the limitations of technology. It is difficult to collect complete data in life. Incomplete or erroneous data will cause the data to be unavailable for direct use. The research results are also unsatisfactory. In this case, we will use the most important data preprocessing. Data preprocessing is the first step in data analysis. There are many ways to do it. For example, data integration, data cleaning, data integration, data reduction, data conversion, etc. These data preprocessing methods greatly facilitate subsequent research and analysis.

3.2 Model Building and Training

The building and training of machine learning models are the core steps in all links. The three models we use are artificial neural networks, logistic regression and Support Vector machine. In the study, three models will be established separately and the parameters will be adjusted to train the models.

3.3 Model Evaluation and Prediction

After constructing the model, its parameters should be adjusted. We should combine the initial model with the data. Let the model be successfully applied to the data. There is no perfect model, we have to train as much as possible to try to improve the accuracy of the model. And with the addition of new data, the model also changes accordingly. Both the model and the data should fit together. After completing the construction of the two complete models, we will use them to evaluate student data. Compare the accuracy of the model. Finally, Models are used to predict random data.

4 Result

We used ‘explore data analysis’ to draw and analyze the correlation between variables. A lot of results have been obtained. As most people understand. GPA, English score, degree and other factors are all positive factors for high income. On the contrary, factors such as failed courses are negative factors. In addition, there are big differences between Faculty. The high-income probability of students engaged in medical and health care is far greater than that of students engaged in agriculture, agro-industry, and construction. Of course, no matter what faculty student, under the premise of high GPA and high degree, there is a greater probability of getting high income. The graduates working area is shown as Fig.2.

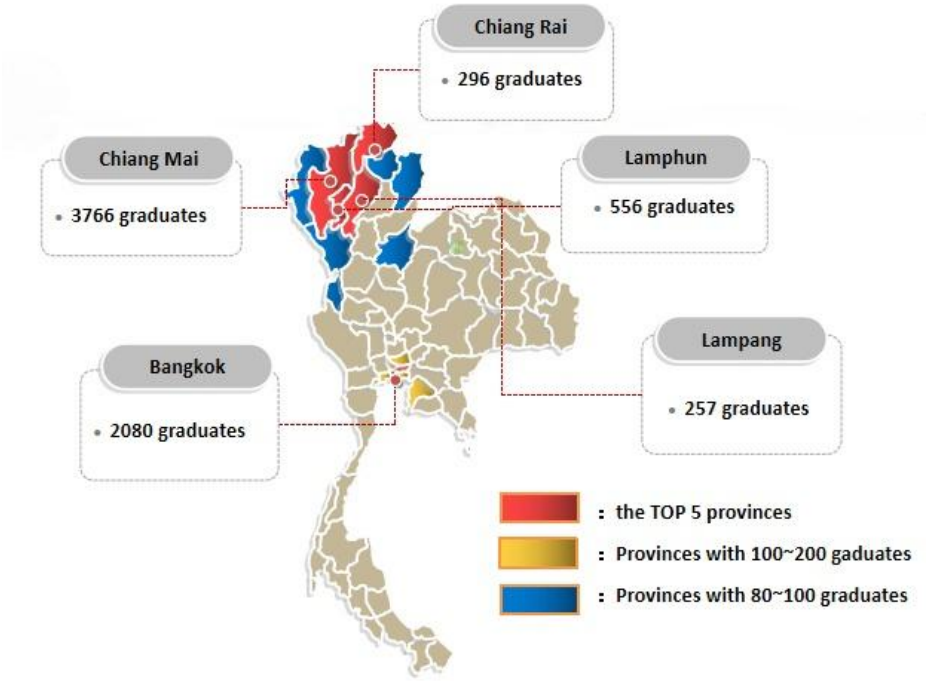

Fig. 2. Working area map

Although there are differences between the three research methods. We use models to adjust the parameters. Letting the model suit for the data. After some key parameters changed. The models gradually show us good output as Table 1. Finally we could use it to do some evaluation. The data itself is the standard that determines the upper limit of the model's accuracy. We use several standard indicators to measure the model. In this machine learning prediction, the three methods all showed the same level of prediction.

Table 1. Comparison of three methods.

Compari-son index in two dataset	Accuracy/Precision/Recall/F1-score	Auc
ANN1:	0.750	0.906
LR1:	0.750	0.905
SVM1:	0.760	0.894
ANN2:	0.748	0.905
LR2:	0.756	0.910
SVM2:	0.746	0.905

When discussing the influencing factors between variables, we abandon the machine learning method and pay more attention to accuracy. We used traditional logistic regression. Use 95% confidence intervals to determine the correlation of variables. Finally draw the conclusion: the biggest factor in determining high income is Health Science faculty as Fig.3. We can use the same method to conclude that all the students in the university, the Degree factor is the most critical factor in determining high income. This conclusion is consistent with the conclusion drawn by the correlation analysis of the heat map variables

.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.33554 -0.07422 -0.03061  0.01134  1.08305

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.0257669  0.0218340  -1.180 0.237988
GPA               -0.0286847  0.0080834  -3.549 0.000390 ***
Englishscore       0.0498147  0.0036472  13.658  < 2e-16 ***
preformancerating -0.0129045  0.0083623  -1.543 0.122831
Failedcourse      -0.0008673  0.0014423  -0.601 0.547615
workingarea       -0.0216503  0.0060727  -3.565 0.000366 ***
Feedback           0.0107625  0.0083247   1.293 0.196109
Healthscience      0.1742320  0.0073877  23.584  < 2e-16 ***
Sciencetechnology -0.0043979  0.0063360  -0.694 0.487632
Gender             0.0490568  0.0056902   8.621  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2226 on 7296 degrees of freedom
Multiple R-squared:  0.144,     Adjusted R-squared:  0.1429
F-statistic: 136.3 on 9 and 7296 DF,  p-value: < 2.2e-16
```

Fig. 3. Output of RStudio

5 Conclusion

In this study, we found that the income level of graduates satisfies a normal distribution, most people are at the middle income level and high-income and low-income are in the minority. We use data visualization to analyze the impact of faculty, GPA, English score, etc. on income through charts. Most of the results are consistent with common sense. That is, students who perform well (high GPA, high English Score, less failed courses and so on) in universities usually earn higher incomes. We can find the relationship between variables from a large number of visualization graphs. You can also draw more corresponding diagrams to visualize the data according to your needs. These explore data analysis give us a deep understanding of variables.

In terms of machine learning, this research has made progress in multiple categories on the basis of the traditional two-category classification. Each of the three learning methods has advantages and disadvantages. But in general, we maximize the functionality through model parameter adjustments. In this study, most of our parameters adopted the default parameters, but some core parameters were adjusted accordingly. As we expected, all three models have reached the ideal prediction effect. Through the model, we can infer the graduate income of students with a high probability. This is a good way for university students to assess their own level of employment.

As for the influencing variables on income. We used the traditional logistic regression method this time, because this method focuses on the correlation between variables, the accuracy of judgment is not high. And among the 7,361 data used for two classifications, only 450 are high-income ones. There is not enough data to support the influence of some variables. This is also the reason why the influencing factors of GPA and working area are negative. Besides, the adjusted r-squared of the goodness of fit of the model is 0.1429, which is caused by the large number of dummy variables. In conclusion. In addition to logistic regression analysis, we also used variable correlation analysis between every two variables. The conclusion of the two is roughly the same, that is faculty and degree are the two most obvious influencing factors. Choosing a faculty related to health science and improving a higher degree is the best way to get a high income. For some students who have their own faculty needs. Improving your GPA and English score is the best way to increase your income. This method is universal and suitable for all university students.

References

1. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), 111-117.
2. Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.
3. Elshorbagy, A., Simonovic, S. P., & Panu, U. S. (2000). Performance evaluation of artificial neural networks for runoff prediction. Journal of Hydrologic Engineering, 5(4), 424-427.

4. Adamowski, J., & Karapataki, C. (2010). Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN learning algorithms. Journal of Hydrologic Engineering, 15(10), 729-743.
5. Krishnapuram, B., Carin, L., Figueiredo, M. A., & Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE transactions on pattern analysis and machine intelligence, 27(6), 957-968.
6. Angulo, C., Ruiz, F. J., González, L., & Ortega, J. A. (2006). Multi-classification by using tri-class SVM. Neural Processing Letters, 23(1), 89-101.