Binary logistic regression classifying the gender of student towards Computer Learning in European schools

Chaman Verma, Veronika Stoffová, Zoltán Illés, Sanjay Dahiya

Abstract: The authors presented a gender prediction model of student based on answers provided into survey during academic year 2011 in Europe. This experimental study is performed in R-language by applying logistic regression on the large data-set available on the website of European Commission. More than 2500 schools, 27 countries, and more than 45000 students have participated in the survey held in 2011 and survey was conducted by European Commission on primary schools whose were studying at ISCED level 3 (upper secondary level of education). The dichotomous variable is gender and 6 predictors belong to attitude towards computer learning. The best cut-off and accuracy of the presented model is measured 0.499 and 0.628 respectively at 0.5 thresholds using Receiver Operating Characteristics (ROC) and Area under the curve (AUC) which signifies the model to predict more females with correctly as compared to males.

Keywords: Prediction, Confusion Matrix, ROC, AUC.

Introduction

Using the historical data on an attribute or event which can predict the future with specific probability, predictive modeling is required. Logistic regression plays a significant role in many fields which estimate the effects of independent variables on predictor variable as the probability. To modeling binary variables, it estimates relationships of other variables with the dependent variable. In case of various assumption distortions (such as normality, common variances etc.), logistic regression studies and practices are used as an alternative to discriminant analysis and crosstabs. If the dependent variable is binary like 0, 1 or discrete containing more than two levels, as the normality assumption is distorted, it also is an alternative to the linear regression analysis [1]. The term generalized linear models (GLM) usually refers to the large class of conventional linear regression models for a continuous response variable given continuous and/or categorical predictors following [2]. The authors focused on gender variables which are 2-level factor regressing on another 6 variable (related to student's attitude). Logistic regression is suitable in present problem in which gender is categorical and student attitude is numerical. Logistic regression is like discriminant analysis in terms of the aim of estimating a categorical dependent variable, and it necessitates less assumption. On the other hand, if the assumptions necessitated by the discriminant analysis are provided, the logistic regression may also be implemented [3]. Logistic regression was applied to develop the model for the early and reliable prediction of students pass or fail status of the undergraduate level and most found significant factors related to explore pass, fail or drop status of students based on their study [4] [5].The gender is one of the principal determinants of the probability of dropping out. In the binomial probity model, they used, males have a higher probability of dropping out relative to the reference group of females [7].

Experimental Design and Results

An experimental study is conducted in R- language to classify student gender based on their answers given to survey. The responses were belonging to European students of primary schools whose were studying at ISCED level 3 (upper secondary level of education). The dataset consists of more than approximately 150 attributes and 47000 instances. After self-reduction and liwise removal methods, we considered only 8 attributes, 45929 observations from dataset having 19771 male and 26158 female students [1]. An equation of logistic regression is defined to find the probability of accepting female is:

$$P = \frac{e^y}{1 + e^y} \tag{15}$$

where, y is gender.
The gender is considered as class or target variable and questions are predictors and the authors calculated the inter-rater reliability (IRR) of responses dividing the total no. of matched responses (25566)

by total no. of responses (45929).The predictors of model encoded from ST17Q01, ST17Q02 to ST17Q08 based on European Commission survey. To achieve the better performance of the model, dataset is trained and tested randomly at various training ratio of (test, train) such as (0.9,0.1), (0.8,0.2) up to (0.1,0.9). After applying regression model we found six significant variables which have provided meaningful contribution into the study such as ST17Q01, ST17Q02, ST17Q03, ST17Q04, ST17Q07 and ST17Q08 which contributed more than 99%. Hence, we considered only six significant variables into our classification model. To present a significant gender classification model we test and train the dataset using logistic model with 0.5 thresholds, The model equaiton of R- language is given below:

model<glm(GENDER~ST17Q01 + ST17Q02 + ST17Q03 + ST17Q04 + ST17Q07 + ST17Q08, data=train, family='binomial').

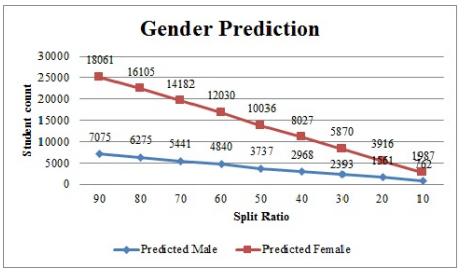

Figure 1: Prediction at Training Ratio (Source: Authors).

Fig.1 reflects gender-wise prediction count of students at the various splitting ratio. It can be seen at the ratio (90, 10), the model predicts more female as compared to male and found to be much significant at this level. The decreasing level of splitting ratio of training data causes poor prediction of gender. At ratio (10, 90), only 762 male and 1987 female are predicted. The gender classification model gives better accuracy at (0.9, 0.1) split. Table 1 shows the total number of the predicted male is 12562 and the total

Table 1: Confusion Matrix (Source: Authors)

Gender	Male (1)	Female (2)
Male (1)	7075	5487
Female (2)	10749	18061

number of predicted female is 28810. Out of total 12562 predicted male, 7075 are predicted accurately and 10749 went to the wrong class named female. It can also see that 18061 students who are predicted as female category correctly with the 6 predictors, while 5487 are predicted into male category incorrectly. The sensitivity or true positive rate (Tpr) of the model is calculated by dividing the total number correct predicted female by the total number of an actual female which is 0.766. The false positive rate (Fpr) is 0.603 which is calculated by dividing the total number of incorrectly predicted female by the total number of an actual male. Further, specificity (1-Fpr) of the model is 0.396 (total number incorrect predicted male/actual male). The precision value is estimated as 0.626 (total no. of correct predicted female/ total no. of an actual female. Hence, at 0.50 thresholds, the accuracy of predicting gender is found 62%. The overall accuracy of the model is calculated as 0.607((total no. of corrected female |+| total no. of corrected male) |/| total no. of students). The prevalence of model is 0.569 which is calculated by dividing the total no. of an actual female by total no. of students which specifies the correct prediction of the female student.

Fig. 2 shows probability is varying in between the range of 0.2 to 0.8. It can be observed that most of the probability to predict the gender of the student against their responses is lies in between 20% to 80%. At the point 0.62 maximum student are found to belongs to female category due to the threshold equation pred1<-ifelse(p1>0.5,2,1) specifying prediction of the female student at 0.5 cutoffs.

The performance of the model is evaluated by ROC and AUC in R-Language. In Fig. 3 the 45 degree reference line (in black) is the line of non-discrimination or benchmark of the model and area under

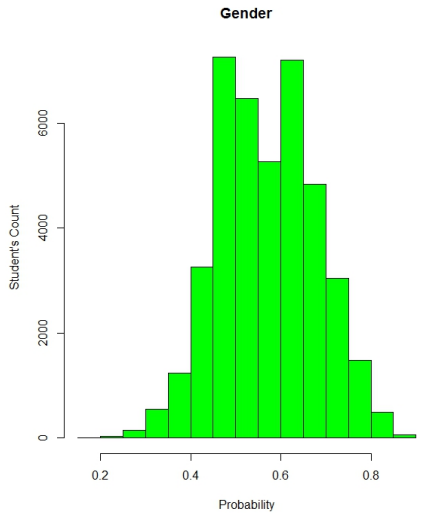

Figure 2: Probability Count (Source: Authors).

the curve (AUC) is 0.6287 which is less than a straight line and ROC curve of the model is above than benchmark or reference line. As the calculated true positive rate (Tpr) or sensitivity is 0.766 and false positive rate (Fpr) is 0.603. The sensitivity of the model measures the ability correct prediction of female and specificity measures the ability to correctly predict the male.

Conclusion

The classification model predicts the student's gender according to their given responses against 6 signficant questions. This model predicts student's gender with more than 60% accurately based on their response. The present study explores the higher count of prediction of female students as compared to male students according to responses and the precision value of the model is measured 0.626. The true positive rate (sensitivity) is 0.766 of model specifies better prediction of the female towards their responses. More than 50% positive predictive value concludes the male and female classes are perfectly balanced. The presented model is efficient to predict student's gender with supporting by ROC curve and confusion matrix.response to computer learning. In future, the sensitivity of model can be enhancing by exploring other remaining variables/features in the survey using other classifiers.

Acknowledgments

The present study is supported by Eotvos Lorand University and Tempus public Foundation, Hungary.

References

[1] European Commission: "https://ec.europa.eu/digital-single-market/news/ict-education-essie-survey-smart-20100039", Accessed on 14 Feb 2018.

[2] McCullagh P., et.al. (1989), "Generalized Linear Models", Second Edition, London: CRC Press Publishers.

[3] Korkmaz1 M., et.al. (2012), "The importance of logistic regression implementations in The Turkish Livestock Sector and logistic regression implementations/fields", J.Agric. Fac. HR.U., 2012, 16(2): 25-36.

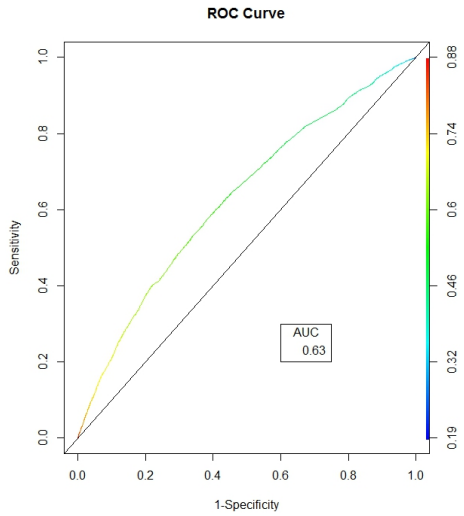

Figure 3: Model Performance (Source: Authors)

[4] Gerard J.A.Baarsa, et.al. (2017), "A Model to Predict Student Failure In The First Year of The Undergraduate Medical Curriculum", Health Professions Education,5–14.

[5] Woodman R. (2001), "Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region", M.Sc. Dissertation, Sheffield Hallam University, UK.

[6] Boero G., et.al, (2005), "An econometric analysis of student withdrawal and progression in post-reform Italian universities", Centro Ricerche Economiche Nord Sud, CRENoS Working Paper 2005/04.