Linear Regression Model in Student Prediction System

[I]Dr. B. Azhagusundari, [II] Dr John Grasias S, [III] G Sudha, [IV] G.Kanimozhi, [V] Dr. Radhika.V

1. Dr.B.Azhagusundari ,Associate Professor, Department of Computer Science NGM College Pollachi,azhagusundari@ngmc.org

2. Dr John Grasias S,Assistant Professor,Department of Computer Science ,AJK College of Arts and Science ,Coimbatore, johngrasias@gmail.com

3. G Sudha, Computer Instructor, Govt Girls Hrs Sec School Ekanampet Kanchipuram 631601 sudhasganapathi@gmail.com

4. G.Kanimozhi, Assistant Professor, Department of Computer Science, National College (Autonomous), Trichy, gajekani@gmail.com

5. Dr. Radhika.V, Department of Physics, Erode Sengunthar Engineering College College (AUTONOMOUS), Thudupathy, Perundurai – 638057,Tamil Nadu,radhikaesec@gmail.com

Abstract

Logistic Regression is a widely used statistical method for predicting a categorical dependent variable based on a set of independent variables. Recognized for its adaptability and frequent application, logistic regression is particularly effective in modeling binary and multinomial outcomes. This paper provides a clear and detailed exploration of the fundamental concepts of logistic regression and demonstrates its application in predictive analysis using student data. Through this practical example, the paper highlights the method's utility in identifying relationships and making informed predictions in student educational contexts.

Keywords: categorical variable, logistic regression, Prediction system, student data

1 Introduction

Let us apply a logistic regression to the example describe before to see how it works and how to interpret the results. Let us build a logistic regression model to include all descriptive variables.

The Logistic Regression(LR) statistic modeling technique is used when we have a binary outcome variable. For example: given the parameters, will the student pass or fail? Will it rain or not? etc.

In LR, continuous or categorical independent variables, we can use the logistic regression modeling technique to forecast the outcome when the outcome variable is binary. Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). In 1972, Nelder and Wedderburn described this model with an effort to present a means of using linear regression to the problems which were not straight suited for application of linear regression.

Wang et al. (2018) derived optimal subsampling probabilities that minimize the asymptotic mean squared error (MSE) of the subsampling-based estimator in the context of logistic regression. Drineas et al. (2011) developed an algorithm by giving out the data with randomized Hadamard transform and then using uniform subsampling to approximate LS estimates. Drineas et al. (2012) developed an algorithm to approximate statistical leverage scores that are used for algorithmic leveraging.

Logistic regression is actually an extension of linear regression. Linear regression analysis demands that the dependent variable is continuous. Where as Logistic regression is used to estimate the relationship between one or more independent variables and a binary (dichotomous) outcome variable.

2. Logistic Regression

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- The equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots .. + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}; 0\ for\ y = 0, and\ infinity\ for\ y = 1$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become: Log $\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots .. + b_nx_n$

The above equation is the equation for Logistic Regression.

2.1Types of Logistic Regression

There are three types of logistic regression algorithms:

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

○ **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"

○ **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

2.3 Classification of Student Dataset using Linear Regression Model

Classification of Student Dataset using Linear Regression Model consist of the following steps

2.3.1 Data Preparation

2.3.2 Fitting Logistic Regression to the Training set

2.3.3 Predicting the test result

2.3.4 Test accuracy of the result(Creation of Confusion matrix)

2.3.5 Visualizing the test set result.

Regno	Name	Mathematics-I	C Programming	Data Structure	Result
22CS01	Gowtham.S	60	56	45	Pass
22CS02	Kavikumar.K	60	67	78	Pass
22CS03	Kumaravel.K	89	78	89	Pass
22CS04	Madhan.S	89	67	89	Pass
22CS05	Manikandan P	90	78	90	Pass
----	-------	------	---	----	---
22CS29	Vinitha .P	67	65	67	Pass
22CS30	Vinothini.K	67	56	67	Pass

Table #1: student dataset

Above is the student dataset for first year computer science 30 students core paper marks. Using Logistic regression to Predict the **Result variable (Dependent Variable)** by using Mathematics-I, C Programming, Data structure and C Lab **(Independent variables)**.

2.3.1 Data Preparation: In this step, we will pre-process/prepare the data so that we can use it in our code efficiently ie extracting Independent and dependent Variable .

Mathematics-I	C Programming	Data Structure	Result
60	56	45	Pass
60	67	78	Pass
89	78	89	Pass
89	67	89	Pass
90	78	90	Pass
------	---	----	---
67	65	67	Pass
67	56	67	Pass

Table #2 : Data preparation

Now we will split the dataset into a training set and test set. Below is the code for it:

sno	Mathematics-I	C Programming	Data Structure
15	67	78	78
1	60	67	78
23	67	98	98
26	54	55	55
25	43	67	67
12	89	98	90
18	89	60	60
28	67	65	67
4	90	78	90
22	89	90	90
8	55	34	55
0	60	56	45
24	45	77	77
29	67	56	67
21	78	89	78
7	67	45	67
11	77	78	89

14	56	65	87

Table #3 :Training set

Sno	Mathematics-I	C Programming	Data Structure
2	89	78	89
19	90	60	60
17	56	36	67
13	36	67	78
27	56	65	67
5	98	76	98
10	38	67	78
3	89	67	89
16	78	76	98
6	77	54	77
9	66	67	32
20	78	89	78

Table #4: Testing Set

2.3.2 Fitting Logistic Regression to the Training set:

By Using student dataset, to train the dataset using the training set. For providing training or fitting the model to the training set in Table #3, and applying the **LogisticRegression** .

```
['Pass' 'Pass' 'Pass' 'Pass' 'Pass' 'Pass' 'Pass' 'Pass' 'Fail' 'Pass'

 'Pass' 'Fail']
```

2.3.4 Predicting the Test Result

Our model is well trained on the training set, so we will now predict the result by using test set data. Below is the code for it:

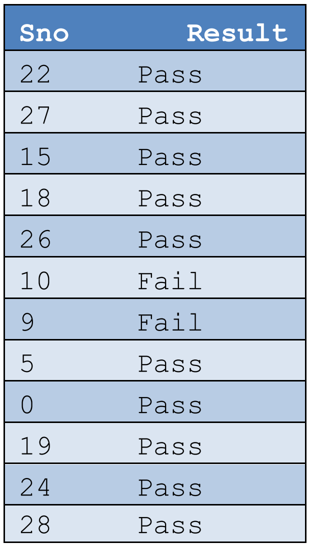
Logistic Regression predict([[50, 67, 14]])that is mathematics-I mark is 50,C Progrmming mark is 67 and Data structure mark is 14. After predict using logistic regression the result is **Fail**.
Logistic Regression predict([[60, 67, 64]])that is mathematics-I mark is 60,C Progrmming mark is 67 and Data structure mark is 64. After predict using logistic regression the result is **Pass**.

2.3.5 Test Accuracy of the result

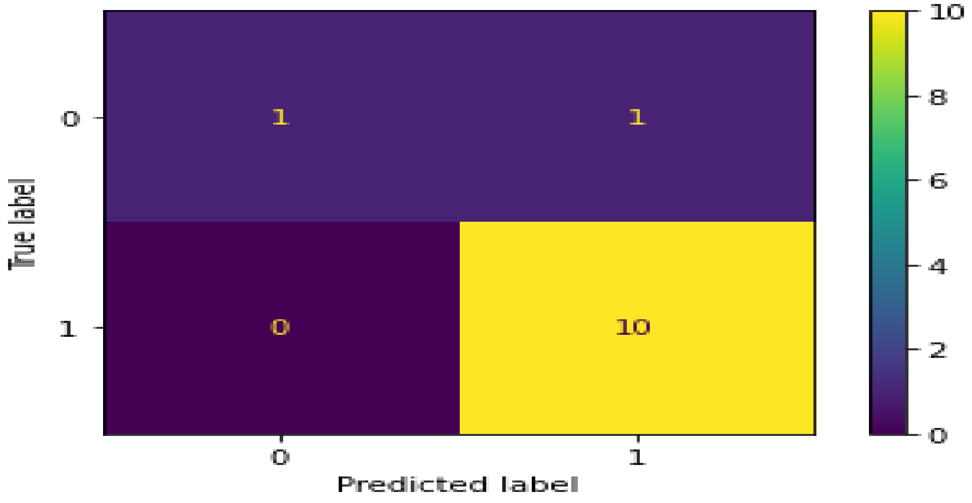
To create the confusion matrix here to check the accuracy of the classification.

Confusion matrix	**Predicted label**	
Actual Label	True Negative(TN)	False positive(FP)
	False negative(FN)	True Positive(TP)

- Precision quantifies the number of positive class predictions that actually belong to the positive class. ***Precision*** = $\frac{\boldsymbol{TP}}{(\boldsymbol{TP+FP})}$
- Recall quantifies the number of positive class predictions made out of all positive examples in the dataset. ***Recall*** = $\frac{\boldsymbol{TP}}{(\boldsymbol{TP+FN})}$
- F-Measure provides a single score that balances both the concerns of precision and recall in one number. ***F Score*** = $\frac{\boldsymbol{2*(Recall*Precision)}}{(\boldsymbol{Recall+Precision})}$
- *Accuracy* = ***(TP+TN)/(TP+FP+FN+TN)***



Sno	Result
22	Pass
27	Pass
15	Pass
18	Pass
26	Pass
10	Fail
9	Fail
5	Pass
0	Pass
19	Pass
24	Pass
28	Pass



2.3.6 Visualizing the training set result

Fitting Logistic Regression to the training set involves training a logistic regression model using the provided training data as shown in figure 1.1.

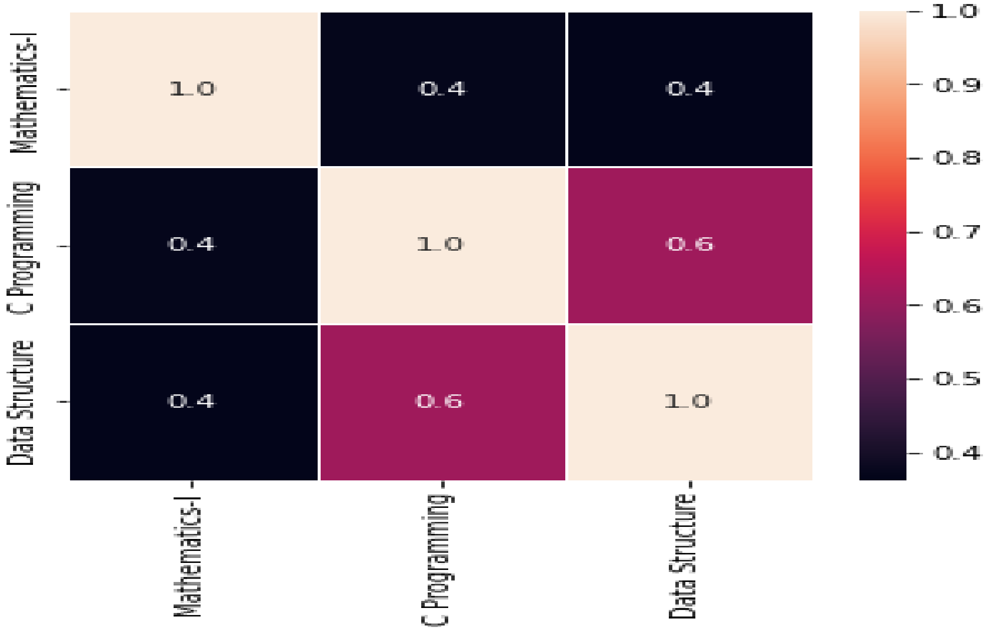


Figure 1.1

3. Conclusion

The study of the Linear Regression Model in Student Prediction System demonstrates the potential of linear regression as a robust statistical tool for analyzing and predicting student data based on various academic factors. By modeling the relationship between independent variables and dependent variables, the linear regression model provides actionable insights that can assist educators and administrators in making decisions.

4.References

1. Ranganathan, Priya, C. S. Pramesh, and Rakesh Aggarwal. "Common pitfalls in statistical analysis: logistic regression." *Perspectives in clinical research* 8, no. 3 (2017):

2. HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association, 113(522):829–844, 2018

3. P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation. Numerische Mathematik, 117:219–249, 2011.

4. P. Drineas, M. Magdon-Ismail, M.W. Mahoney, and D.P. Woodruff. Faster approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research, 13: 3475–3506, 2012.

5. Zabor, Emily C., et al. "Logistic regression in clinical studies." *International Journal of Radiation Oncology* Biology* Physics* 112.2 (2022): 271-277.

6. Harris, Jenine K. "Primer on binary logistic regression." *Family medicine and community health* 9.Suppl 1 (2021).