Students Placement Prediction Model Using Logistic Regression

Manoj K Shukla
Student of Computer Engineering,
Atharva College of Engineering,
Mumbai, India

Pranay Rambade
Student of Computer Engineering,
Atharva College of Engineering,
Mumbai, India

Jay Torasakar
Student of Computer Engineering,
Atharva College of Engineering,
Mumbai, India

Rakesh Prabhu
Student of Computer Engineering,
Atharva College of Engineering,
Mumbai, India

Prof. Deepali Maste
Assistant Professor, Computer Engineering,
Atharva College of Engineering,
Mumbai, India

***Abstract*— This paper is implementation model of our previous paper as an extension of Logistic Regression Analysis as a Future Predictor System, International Journal of Technology Research and Application. Logistic model designing plays a key role in order to get correct predictions. This process includes selection of tuples for training data and their pre-known outcome often known as real data. This paper details the steps involved in actual designing and development of such model.**

Keywords*— Regression Analysis, Prediction, Logistic Regression, Data Mining, Logit Function, Machine Learning, Accuracy.

I. INTRODUCTION

While presenting this paper, we aim to extend our previous paper on analysis of logistic regression model [1]. Going through our previous paper, readers were able to the concept of logistic regression. A comparison can be made easily among the regression models. Now idea of current paper is focused on how to develop a working model of such complex and critical logistic system. We will consider the same example in calculation part i.e. Students Placement Prediction.

II. LITERATURE REVIEW

One of the most prominent work on prediction of placement for students has been cited by Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor and Keshav Kumar where they presented the development of placement predictor system (PPS) using logistic regression model. They used Machine learning technique to design and implement a logistic classifier that predicts the probability of the student to get placed along with Gradient Descent algorithm. [2]

S.Taruna and Mrinal Pandey implemented an empirical analysis on predicting academic performance by using classification techniques or mapping of data items into predefined groups and classes using supervised learning. They compared five classification algorithms namely Decision Tree, Naïve Bayes, Naïve Bayes Tree, K-Nearest Neighbour and

Bayesian Network algorithms for predicting students' grade particularly for engineering students using a four class prediction problem. [3]

Kotsiantis and Pintelas, 2005 predicted the student marks (pass and fail classes) using the regression methods and available previous data. A number of experiments have been conducted with six algorithms, which were trained using datasets provided by the Hellenic Open University. [4]

Saha and Goutam applied logistic regression method on the examination result data and analyzed the data under the University Grant Commission sponsored project entitled - Prospects and Problems of Educational development (Higher Secondary Stage) in Tripura - An in depth Study. [5]

Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque and Rashedur M Rahman predicted student performance considering more on academic records. They accuired higher by applying some algorithms like e.g., Naïve Bayes, Decision Tree and Neural Network. [6]

Zhiwu Liu and Xiuzhi Zhang used decision tree algorithm C4.5 to establish a classification rule and an analysis-forecasting model for students' marks. [7]

Hitarthi Bhatt, Shraddha Mehta and Lynette R. D'mello identified relevant attributes based on quantitative and qualitative aspects of a student's profile such as CGPA, academic performance, technical and communication skills and designed a model which can predict the placement of a student. [8]

III. PROBLEM STATEMENT

The general Placement Prediction System considers only academic performances in order to predict whether a student can be placed or not. Judging the student based only on his academic performances would be unfair for the student, since a student could be having good aptitude, technical and communication skills but unfortunately might not be good in academic performances.

It would wrong to judge a student based only on his academic performances, since Predicting the placement of a student needs a lot of parameters to be considered. But in order to get selected in campus interview, the student must be good in technical and aptitude skills. Of course academic performances are important but don't hold the highest importance in the outcome of student placement.

IV. EXISTING SYSTEMS AND THEIR GAPS

The current system generally uses only a single parameter to judge whether a student can be placed or not during the campus placements. Generally the parameter used to judge the strengths of the student, is the academic performances during the first three years of engineering.

But cracking an interview not only depends on the academic scores but also the awareness of student during the aptitude tests and interviews. Also some Data Mining algorithms, while calculating the probability of a student getting selected, sometimes interpret the result having a probability of more than 100% which is not feasible and denotes a wrong interpretation to the student.

Some algorithms give a negative probability which gives an wrong interpretation to the student. Judging the student only on the basis of academic grades is not enough. The other parameters like aptitude and technical tests should also be taken into consideration in order to determine the outcome for the student's future.

V. COMPARISION WITH PROPOSED SYSTEM

Proposed system follows the logistic model of regression, which removes the problems caused by linear regression like negative prediction and over-value prediction (more than 100% prediction value). Linear model works best for continuous set of data. But in reality, most of the data is available in discrete manner, which is the best case for logistic model. Proposed system supports additional R^2 (R-square) test. So error can be reduced and we can achieve high accuracy.

VI. SYSTEM OVERVIEW

A. *Process*

Logistic regression is a statistical method use analysing a training dataset in which there are one or more independent variables denoted by X_b (b=0 to N-1, ie N predictors) that determine an outcome ie The final prediction . The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success etc.) or 0 (FALSE, failure etc.).

$$\text{Logit}(p) = \pi = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots + \beta_0 X_{N-1}$$

.

Where "p" is probability of success.
The logit transformation is defined as the logged odds:

$$\text{Logit}(p) = \pi = \log(p/1\text{-}p) \qquad (1)$$

In Our case we will consider previous year placement data as training data which we will apply on our logistic mathematical equation. After mapping this data the final logistic model is used to predict the placement chances of next year student with help of thirteen same parameters which were used during model processing.

B. *Overall Model Fit*

The *null model* -2 Log Likelihood is given by $-2 * \ln(L_0)$ where L_0 is the likelihood of obtaining the observations if the independent variables had no effect on the outcome.

C. *Regression Coefficients*

The Regression Coefficient plays important role in prediction. The coefficient β is associated
with every predictor or parameter. This Coefficient computes the contribution of every predictor or independent variables in the Logistic Model. The more parameters apply on logistic model the more smooth sigmoidal curve will generate. The range of Sigmoidal curve is in between 0 to 1.

D. *Final Processing*

The final data obtained from students through their Online test is apply on Logistic Model which will give final placement prediction in terms of probability. The final equation obtained from initial processing is modifiable by reprocessing it with current predicted and observed data.

VII. ALGORITHM

For predicting a value based on history of data, it is necessary to train the prediction model. In our case, we chose logistic model as a predictor system so we shall be training logistic model using following algorithm.

Before starting with the steps of training, it is always good to understand the terms used in algorithm. Table given below gives the information about notations used in algorithm and their description.

Notation	Description
Training data set	Real set of data with known independent variables and their know outcome.
Weight factor	It is coefficient of regression, denoted as β_0, β_1, β_2, ….
logit	'z', exponential term used in the equation of logistic
Sigmoid function	Prediction calculated while training in terms of probability of getting success or failure. We have to code this success or failure as '1' or '0' respectively later.
rate_of_learning (good value always lies below 1)	The lower the learning rate higher the accuracy with large set of data.
Final Prediction coding	Final prediction obtained will always lie in range of [0,1] both inclusive. For the value lower than 0.5, it is considered as failure, and for higher than 0.5, it is considered as success.

Table. 1. Algorithm for training the coefficients of logistic model

Step I : START

Step II : For each training dataset i.e. rows

logit = 0;

1) foreach 'i' of weight factor in a row {

*logit = logit + weight[i]*x[i];*

}

2) Calculate sigmoid function of ***logit*** *// logit means 'z' here,*

$$Predicted = \frac{1}{1+e^{\wedge}(-logit)}$$

3) Update weight foreach 'j' weight {

// training coefficients of logistic

weight[j] +=rate_of_learning(Y-Predicted)*x[j];*

}

4) Update likelihood function *// not necessary for training*

$$like \mathrel{+}= Y * log\left(\frac{1}{1+e^{\wedge}(-logit)}\right) + (1-Y) * log\left(1-\frac{1}{1+e^{\wedge}(-logit)}\right)$$

Step III : Provide a sample test data in terms of x[0], x[1], …

1) foreach weight factor 'i' {

*z = z + weight[i]*x[i];*

}

2) $P = \frac{1}{1+e^{\wedge}(-z)}$ *// this is the prediction after training*

3) if (P<0.5) : failed (coded as 0)

else : successful (coded as 1)

Step IV : STOP

Fig. 1. Algorithm for training coefficient of logistic regression model

VIII. CONCLUSION

With above steps one can derive the coefficients of regression and can substitute the values in well formatted formulae to get real time predictions. The steps of algorithm are nothing but a simple pseudo-code implementation of calculation. It is easy to implement it in any programming language.

We can obtain the more accurate results with iterating the whole algorithm for 'N' number of times, where 'N' is the any integer number you prefer. Here optimistic value of 'N' is the number of times that your machine can afford to iterate the whole algorithm with considerable amount of time. i.e. time complexity is an important factor to be considered here. This complexity also depends upon the number of rows of data available to train. If large number of data is trained for large number of times, it costs higher CPU cycles.

IX. FUTURE SCOPE

There are many optimization algorithms available which can reduce the number of loops required for higher stability. One of them is known as Stochastic Gradient Decent Algorithm which is best for logistic function optimization. Newton's gradient method and hessian matrix methods can also give you the good value.

ACKNOWLEDGMENT

The authors are thankful to Prof. Deepali Maste, Department of Computer Engineering, Atharva College of Engineering, Mumbai and Mrs. Sinu Mathew, Assistant Professor, Atharva College of Engineering, Mumbai for their valuable suggestions and guidance while preparing this paper.

REFERENCES

[1] Jay Torasakar, Rakesh Prabhu, Pranay Rambade, Manoj Kumar Shukla – "Logistic Regression Analysis as a Future Predictor" International Journal of Technical Research and Applications.

[2] Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, Keshav Kumar,"PPS - Placement Prediction System using Logistic Regression" in 2014 IEEE International Conference on MOOC, Innovation and Technology in Education.

[3] S.Taruna , Mrinal Pandey ,"An Empirical Analysis of Classification Techniques for Predicting Academic Performance" in 2014 IEEE International Advance Computing Conference (IACC).

[4] Kotsiantis, Sotiris B., and Panayiotis E. Pintelas, "Predicting students marks in hellenic open university", in Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on, pp. 664-668. IEEE, 2005.

[5] Saha, Goutam, "Applying logistic regression model to the examination results data.,in Journal of Reliability and Statistical Studies 4, no.2(2011):1-13.

[6] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque and Rashedur M Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," in Decision Analytics (2015) 2:1 DOI 10.1186/s40165-014-0010-2(Springer Journel).

[7] Zhiwu Liu and Xiuzhi Zhang, "Prediction and Analysis for Students' Marks Based on Decision Tree Algorithm" in 2010 Third International Conference on Intelligent Networks and Intelligent Systems.

[8] Hitarthi Bhatt, Shraddha Mehta, Lynette R. D'mello, "Use of ID3 Decision Tree Algorithm for Placement Prediction," in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4785-4789.