PPS - Placement Prediction System using Logistic Regression

Ajay Shiv Sharma[1], Swaraj Prince[2], Shubham Kapoor[3]
Information Technology Department,
Guru Nanak Dev Engineering College, Ludhiana.
Punjab, India.
Corresponding author email: 2ajayshiv@gmail.com*

Keshav Kumar[4]
Mechanical Engineering Department,
Guru Nanak Dev Engineering College, Ludhiana.
Punjab, India.

***Abstract* - This paper presents the development of placement predictor system (PPS) using logistic regression model. Based on the student scores in matriculation, senior secondary, subjects in various semesters of technical education and demographics, PPS predicts the placement of a student in upcoming recruitment session. The steps involved in designing and building logistic regression model is stated using the past academic and in-house placement data of Guru Nanak Dev Engineering College (GNDEC), Ludhiana. Machine learning parameterized approach is used to support research and analyze the students performance in previous sessions. The results are generated from an open source GNU Octave programming tool. The developed model has been applied to predict the placement of students at training and placement office (TPO). The testing of PPS brings about promising 83.33% accuracy. The learned parameters of the model gave insights into the placement process. Hence, the TPO decided to adopt this system to help them in informed decision making. This application endows the targeted group of students to boost their placement probability.**

***Index Terms* - Placement prediction, Logistic regression, Classification tool, Student performance analysis.**

I. INTRODUCTION

Every year across the country, the technical education enrollment figures are in millions through hundreds of universities and thousands of colleges. But only few of them land in good jobs as compared to total enrollments, despite the equal inputs made by institutions on all of them. There are a plethora of factors that determine the placement of a student. The Institution's accreditation status, campus activities, location, relationship with industry, field specializations are some of the influential factors although these factors do not qualify as students attributes because they cannot be directly controlled by the student whereas student sex, rural/urban status, performance in high school and college are the attributes which turn out to be the good predictors for placements.

Predictive modeling can be applied to number of different fields such as customer retention, stock trading, anomaly detection, finance and product marketing. These are some of the motivations to be applied in the field of student performance analysis. The predictive models can act as a tool to provide the information of students about their performance in the classroom and their chances of placement which in turn help the authorities to take informed decisions and maximize the results of the efforts made by the institutions.

II. LITERATURE REVIEW

In the year 2002, Luan used the predictive model to determine which student can easily pile up their courses and which student will take courses for longer time [1]. Sampath et al. developed a logistic regression model that predicts freshmen enrollments at George Mason University [2]. Minaei-Bidgoli et al., 2004 used to predict the final grades of students based on their web use feature and can distinguish those students that are at risk among the whole batch of students and hence, tutors can provide appropriate guidance in a timely manner [3]. Educational Data Mining (EDM) is used to improve the functioning of institutions by analyzing student performance. EDM research literature was provided by Romero and Ventura [4], in this field between 1995 and 2005, and by Baker and Yacef, for the period after 2005 [5]. In the year 2013 at Bulgarian Academy Of Science presented a paper using data mining methods for classification. The initial results from their implemented research project aimed to show the high potential of data mining applications for university management [6]. Kotsiantis and Pintelas, 2005 predicted the student marks (pass and fail classes) using the regression methods [7]. In 2012, Edin Osmanbegovic along with his co-author published his work on using data mining techniques for predicting student performance [8]. Similarly, in 2011, 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 [9].

III. PROBLEM FORMULATION

Every year the TPO of GNDEC faces the challenging task of placing final year students in various companies of their respective field while simultaneously maintaining the quality of companies recruiting students; in terms of job profile, working environment, salary packages, and

978-1-4799-6876-3/14/$31.00 ©2014 IEEE

opportunities for growth of the students. With the task to place maximum number of students becomes more daunting, especially when there are no concrete tools available to the TPO for getting insights of student's performance in the placement session. Due to absence of these tools, the steps taken by administration to boost the performance of students get wasted because they are implemented ambiguously on the whole bunch without any strategy. Inability to figure out the group of students who need the actual support, leads to poor results even when the efforts are done with the maximum force. Hence, a plan was laid out to appeal power of data mining and machine learning to build the predictive models using past academic and placement data at GNDEC.

IV. Methodology

This model categorizes the students on the basis of likelihood to get placed or not. Placement and academic data provided by GNDEC is used in this model. Data is cleaned and converted into numbers to use it for number crunching. Machine learning technique is used to design and implement a logistic classifier that predicts the probability of the student to get placed. Gradient descent algorithm is used to optimize the classifier to get the optimum values of parameters that minimizes the value of cost function. GNU Octave [10] is used to program the prediction model and is supplemented by visualization graphics generated from Tableau [11].

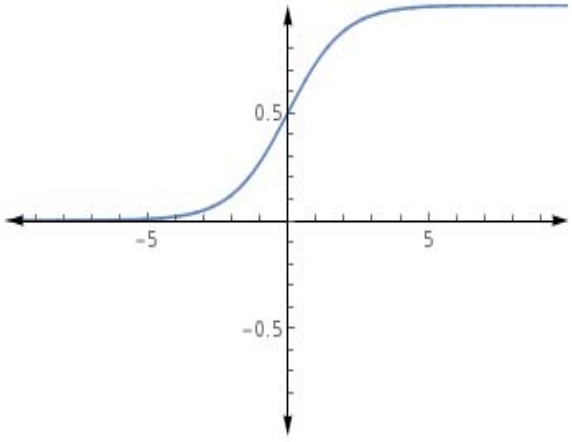

Fig. 1. Sigmoid Function

A. Logistic Regression

This section provides a brief background on the technique to predict the student's probability to get placed. The underlying outcome, namely placement predictor, is categorical (binary). It has yes (student placed) or no (student not placed) option to select from. Therefore, ordinary regression model cannot be used. Logistic regression is used in this case. Gradient descent algorithm is used to learn the parameters which calculate the probability a student to get placed. Logistic classifier is fed with the data made up of marks scored in secondary school examination, senior secondary school examination, first six semester subjects examination and demographic data. A logistic regression is the modified version of the linear regression. The prediction values are between -∞ to +∞ but the probabilities lie between 0 and 1.

The formula to get the value of prediction (z) is described by Eqn. 1 as :

$$z = \theta^T x \tag{1}$$

where,
z is dependent variable that represents prediction.
θ is vector of weights to the independent variables.
x is the vector of independent variables that represents the data.

A non-linear function namely Sigmoid function mentioned in Eqn. 2 illustrated with the help of Fig.1 drawn in GNU Octave is applied on a real number range to obtain the equivalent value between 0 and 1.

$$sigmoid(x) = \frac{1}{1+e^{(-x)}} \tag{2}$$

The hypothesis of the logistic classifier written in Eqn. 3 is derived by applying sigmoid function on predictions to get probabilities of classifier between 0 and 1.

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \tag{3}$$

To train the classifier , normally the difference between hypothesis of training example and actual outcome is squared mean over the whole training set. In case of logistic regression, the squared mean error will be a non-convex function because hypothesis is derived from the sigmoid function. Cost function J(θ) of the logistic regression model is not described as squared mean error, rather it uses logarithmic function to alter the generalized square mean cost function to make it convex as shown in Eqn. 4 . Convexity of J(θ) is necessary requirement for gradient descent algorithm to optimize the parameters θ in J(θ) to find the global optimum value to minimize the cost function J(θ) and avoid local minima.

$$J(\theta) = \frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))] \tag{4}$$

where,
$\mathbf{x^{(i)}}$ is any training example from training set.
$\mathbf{h_\theta(x^{(i)})}$ is hypothesis of training example $x^{(i)}$.
$\mathbf{y^{(i)}}$ is the actual outcome of the training example.
m is the number of training examples in the training set.

B. Gradient Descent Algorithm

Gradient descent algorithm minimizes the cost function J(θ) by finding the optimum values of weights θ. The minimization step is repeated over a number of iterations till the further decrease in values of weights is significantly less and it works in two steps. The iterative steps that gradient descent follows are explained with the help of Eqn. 5. First, the derivative of the cost function J(θ) is computed with respect to the weight of a particular feature θ_j and is multiplied by the learning rate (α) of the gradient descent algorithm. Second, the above computed value is subtracted from the previous value of weight θ_j and a final new value is updated in the same variable.

$$\textit{Repeat until convergence}$$
$$[\ \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}\ ] \tag{5}$$

All the weights are updated simultaneously by this algorithm to avoid the dependence of weights on one another. j represent the feature number where **J Є [0,n]**.

C. Describing Placement Data

The data used in model is only in-house placement data from 2009 to 2013 batch Information Technology (IT) branch of GNDEC. It consists of the marks obtained by students in secondary and higher secondary school examinations, subject examinations in first to sixth semester, sex and residential status. Aforementioned data are analyzed and used to train the logistic regression model. The Table I gives the list of all variables and descriptions that explains their data types and dependency status on other variables.

TABLE I. Variables used in the Regression Model.

Variable Name	Valid Range	Variable Type	IV/DV
10th Percentage	0-100	Numeric , Continuous	IV
12th Percentage	0-100	Numeric , Continuous	IV
Subject Marks	0 - 100	Numeric , Continuous	IV
Sex	0 , 1	Numeric , Discrete	IV
Residency (Rural /Urban)	0 , 1	Numeric , Discrete	IV
Placement Indicator	0 , 1	Numeric , Discrete	DV

These dependent variables (**DV**) are potential predictors of placement in future industrial sessions as they influence the probability of student to get placed. The outcome variable is the placement indicator which is categorical with the value **ZERO** (not placed) and **ONE** (placed). Missing data in independent variables (**IV**) is completed by mean values of the particular variable and it is not excluded from the model. Sex and residency status are transformed to appropriate numeric values.

Table II shows the demographic breakdown of the total students placed by TPO. Residential status and gender status is taken across and down the table respectively. Four categories are described below which show the placement status of a gender from a particular residential background. The percentages are with respect to total number of placements. As it can be observed from the data, around 91% of total placements are achieved by students from the urban background. Though in rural category, gender status doesn't showcase any trend but in urban status it is observed that gender status has a big role to play. It has been observed that female students are more likely to be placed than male students. Hence, residential and gender status play an important role in predicting the probability of students placement.

TABLE II. Variables used in the Regression Model

GENDER	RURAL	URBAN
FEMALE	4.55%	59.09%
MALE	4.55%	31.82%

V. RESULTS AND DISCUSSION

Research on the academic and placement data brings out the insights of students performance. The predictive model predicts the future outcomes of each student in future sessions of job placement. The parameters are learned by running gradient descent algorithm on training data that provided an idea about the most necessary features that are responsible for a positive outcome. As these features represent the subjects, TPO can use these learned parameters to know what are the hot skills of each industry which they are looking among students. Further, the training and testing accuracy of the algorithm was 98.93% and 83.333% respectively.

Fig. 2 shows the top ten skills of students of IT branch from GNDEC for which the companies, conducting drives look in them. A general tendency of a company is to hire the professionals of a particular skill set from a particular place. These top ten skills mined out of the data can be a game changer. Teachers and students can focus more on the particular set of skills and can achieve excellence. Also, this can help an institution to build a brand image in industry for that particular skill set. It is observed from the outcomes mined from data that students of GNDEC possess the skill of expertise level in system programming.

To verify this insight, it has been correlated with the job description data from TPO which shows that companies recruited a large number of students to employ them as system and system support engineers. TPO can also figure out the students who will not perform better in upcoming placement sessions and timely inputs can be made on this batch of students. Various workshops and expert sessions can be delivered to pre final year students to update them with the skills that are demanded in industry. Eventually, all the efforts put in answer to these insights, will increase the placement record by a huge factor. Resources can be used on essential matters and on the students who are in need of it, which will improve the overall productivity. Also, this model can predict the lagging skills of each student.

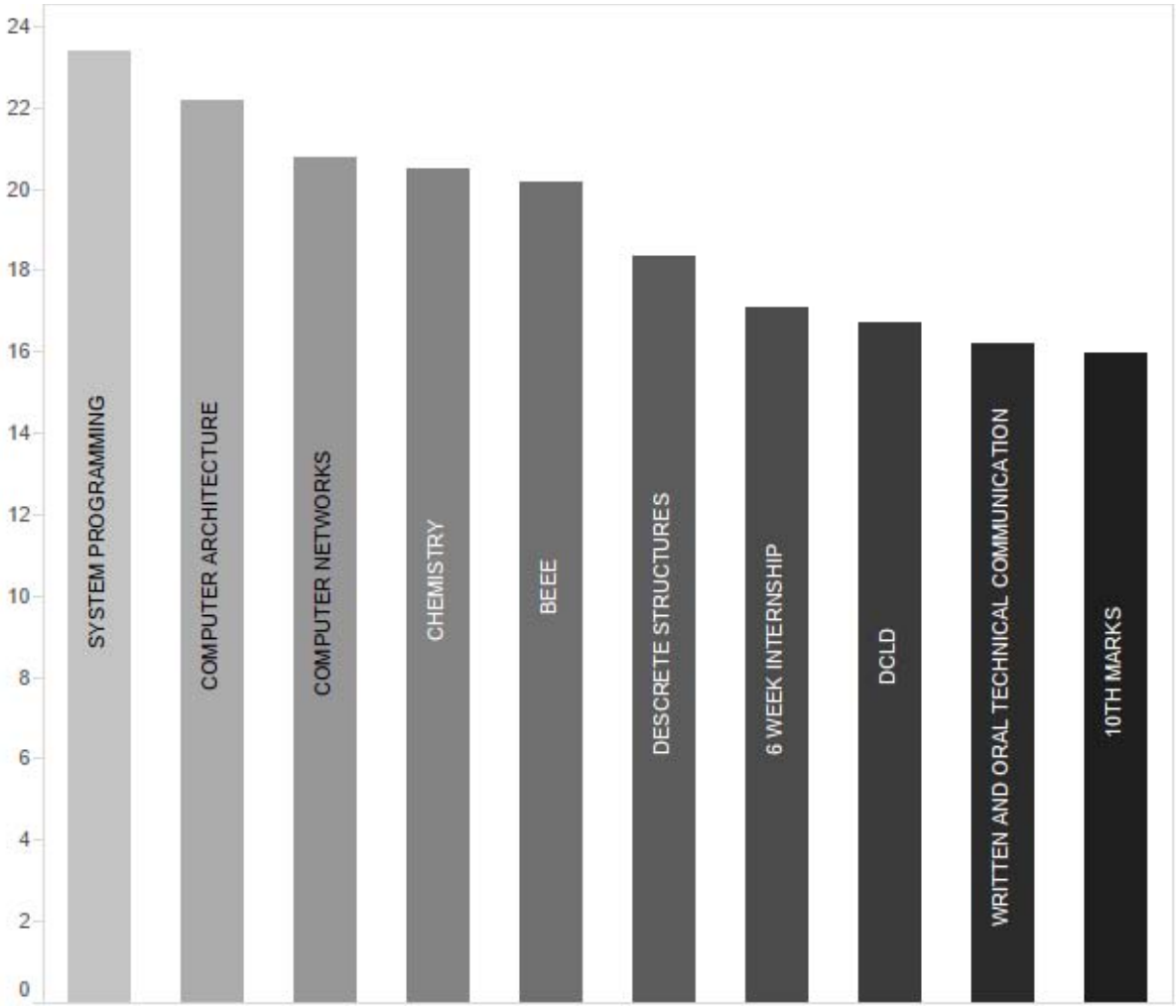

Fig. 2. Top ten hot skills of industry

VI. FUTURE SCOPE OF WORK

Till date, data of 2009-2013 IT batch is used to train the logistic classifier. The accuracy of the model is of average level with this data. Currently, this model can be used as a supplement method to get the insights for placements of students in upcoming session. As the data of every coming batch can be stored and cleaned to be fed to the learning classifier, more random training examples are obtained. These new training examples bring out new characteristics which will help the algorithm to learn different aspects for generating a high probability of student's placement, which is expressed in the form of newly learned parameters. These new parameters when applied to the classifier on test dataset, ultimately increases the accuracy of the model. Later on, some new features can be added to the training data that proves to be a good predictor. After repeated modifications in model and its data; more amount of data can be fed to algorithm, it can become most powerful tool to get insights of students placement in future with much more accuracy so that improvement in students performance can be achieved timely, to enhance the chances of getting placed.

ACKNOWLEDGMENT

The authors acknowledge Guru Nanak Dev Engineering College, Ludhiana for providing data support and other facilities to carry out this work. The authors are thankful for the valuable suggestions made by Dr. Kulvinder Singh Mann, Dean TPO. The authors are thankful to Director, GNDEC and Head of IT department for nominating this project work for 'Best B.Tech. Project' from the college which has been awarded by 'ISTE-PTU for Young Innovators Award' at 17th Annual Convocation of ISTE held at Gujarat on 10-11 October, 2014.

REFERENCES

[1] Luan, Jing. "Data Mining and Knowledge Management in Higher Education-Potential Applications" (2002).

[2] Vijayalaksmi Sampath, Andrew Flagel, Carolina Figueroa "A logistic regression model to predict freshmen enrollments" (2009).

[3] Minaei-Bidgoli, Behrouz, Gerd Kortemeyer, and William F. Punch. "Enhancing Online Learning Performance: An Application of Data Mining Methods1." Immunohematology 62, no. 150 (2004): 20-0.

[4] Romero, Cristóbal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33, no. 1 (2007): 135-146.

[5] Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM-Journal of Educational Data Mining 1, no. 1 (2009): 3-17.

[6] Kabakchieva, Dorina. "Predicting student performance by using data mining methods for classification." Cybernetics and Information Technologies 13, no. 1 (2013): 61-72.

[7] 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.

[8] Osmanbegović, Edin, and Mirza Suljić. "Data mining approach for predicting student performance." *Economic Review* 10, no. 1 (2012).

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

[10] www.gnu.org/software/ocatave. Accessed on 14 October, 2014.

[11] www.tableausoftware.com. Accessed on 14 October, 2014.