IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

Placement Prediction Using Multiple Logistic Regression Method

Koushik Paul[1], Saheb karan[2], Siddhartha Kuri[3], Sulekha Das[4], Avijit Kumar Chaudhuri[5]

[1, 2, 3] B. Tech, Department of CSE, TEC Banipur, West Bengal, India

[4, 5]Assistant Professor, Department of CSE, TEC Banipur, West Bengal, India

Abstract: Standing in the early 21st century, the world has experienced various regression analysis such as Simple Linear regression, Multiple Linear regression, Logistic regression, Multiple Logistic regression etc. Multiple Logistic regression (MLR) or we can say multiple regression is one of them. A widely used statistical technique that allow predictions of systems with multiple explanatory(independent) variables.

In this paper, we collected the final year placement data of a university. Our main objective is to select the explanatory variables for predicting the placement results. Data that has been used in this research were taken from Kaggle website based on the college placements data compiled over 2 years.

Then the data will be analysed by using step by step multiple regression techniques. Here, we used train_test_split and 10_fold_cross_validation in our model.

Reference: - https://www.kaggle.com/tejashvi14/engineering-placements-prediction

Keywords: Multiple Logistic Regression, Placement predictor, Classification, Dataset, Machine Learning.

1. INTRODUCTION

Every year an enormous number of students enrol themselves in the technical field and a determined student always dreams of working in his or her desired sector or company. But it is seen that most of the student fail to get their dream jobs. Every year, the recruiting companies have their specific recruiting criteria. Based on these criteria students get placed every year. Now predicting the basic criteria of most companies can benefit the students. It can encompass the basic needs and requirements of the companies so that students can increase their overall performance throughout their academic. Institutions also analyse these criteria in order to groom and assist their students to bring out the best of them. Nowadays, a statistical analysis is widely used in various fields such as in science, medicine and in social sciences too. Regression is one of the most common statistical methods used. Multiple Logistic regression (MLR) is such a statistical technique that helps to determine a mathematical relation between the multiple independent variables and a single dependent variable. We aim to develop a placement predictor [1] that would predict the probability of the students getting placed (dependent variable) based on their skills (independent variables). The placement prediction is done by machine learning using stepwise multiple logistic regression analysis. Based on the analysis, it predicts the placement results of every student.

2. LITERACY REVIEW

Now the questions that come to our mind such as “What the basic relation between the dependent and independent variables must be?”, “Is there any possibility of making future oriented predictions for the dependent variable?”. With the help of regression analysis, we can really develop such relations with the dependent and independent variables and sought out the answers.

Making a placement management system will benefit the upcoming year students in a college by analysing the current year data. Also, by analysing the current year data students can develop a wide knowledge and uplift their skills before recruitment process starts.

Dr Satish Kumar et al. (2019) [2] have conducted a research work on predicting campus placement probability using binary logistic regression. They aimed to study the nature of campus placements and build a model to predict the probability of a random student that he or she will be placed or not. Also, they wanted to identify the factors that are influencing the placement chances of a student.

Data mining is one of those useful techniques commonly used by data scientists in order to analyse enormous amounts of data swiftly. S. Taruna et al. (2014) [3] performed an empirical analysis of classification techniques for the projection of overall academic completion using data mining.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

However,[4] a similar research study was conducted in Viet Nam at CTU and AIT for predicting the student performance. They used their predictions to identify the very good students for scholarships at AIT and the poor performing students for assistance at CTU.

There are various machine learning placement models that can be used to predict the placement results of final year students. Irene Treesa Jose et al. (2020) [5] aimed at developing a placement predictor using four different algorithms, namely KNN, Logistic Regression, SVM and Random Forest and to compare their accuracy to make a conclusion on which algorithm is the best fit for placement management system.

Ajay shiv Sharma et al. (2014) [6] developed a placement prediction system that predicts whether a student will be placed or not in the upcoming recruitment session based on the overall academic stats. They used the logistic regression model to analyse the past academic recruitments that had occurred.

Architectural Design:

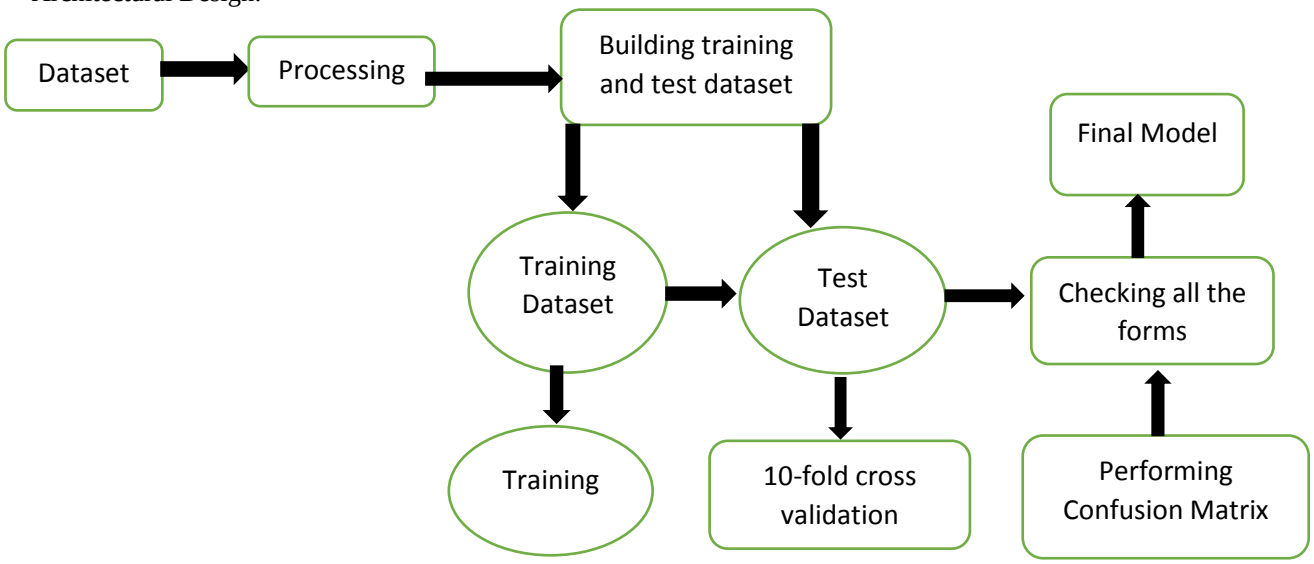

Figure 1: Architecture for Data Processing, Model Training, Prediction and Accuracy check.

3. METHODOLOGY

We all know that machine learning is encyclopaedic. There are many available algorithms that are used in a daily basis. In the world of handling massive amount data, statical methods of analysing them are most welcome. Data scientists use them to analyse, process and frame them to interpret the results to create an applicable plan for the organisations. Multiple Logistic Regression is one such technique where the data is analysed, and some independent variables are identified against a single dependent variable.

Data

In this paper, the data we have been working on has been taken from the Kaggle website. The dataset is based on the placement results of an unknown university of the years 2013 and 2014 whose reference can be found though the mentioned link.

"https://www.kaggle.com/tejashvi14/engineering-placements-prediction"

The dataset mentions the overall academic progress of the students from those years and their placement records upon their progress.

Table 1: Dataset Overview

ATTRIBUES	DEFINITION	MEAN	STANDARD DEVIATION
Internships	percentage of people under the poverty level	0.70	0.74
CGPA	percentage of households with farm or self-employment income in 1989	7.07	0.97
HistoryOfBacklogs	population for community	0.19	0.39

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

Hostel	percentage of people 25 and over with a bachelor's degree or higher education	0.27	0.44
PlacedOrNot	total number of violent crimes per 100K population	0.55	.5

Research Method

As mentioned earlier, we have used Multiple Logistic Regression (MLR), a statistical technique for regression analysis. Our first work was to find the independent variables which were making impact on the single dependent variable. Now as we have found the independent variables, namely- Internships (x_1), CGPA (x_2), HistoryOfBacklogs (x_3) & Hostel (x_4) and the dependent variable, namely- PlacedOrNot (y). We now construct a stepwise logistic relation between them.
As we are moving forward towards our final model, few steps need to be followed in MLR.

STEP 1: Checking Assumptions

The first step of forecasting the model is to find the independent and dependent variable. After that we try to develop a logistic relation between the dependent & independent variables. We then split the data into three parts as 4/5, 2/3, 1/2 defined as training data and the rest as testing data.

Cross validation or one can say out of sample testing, is a method where we test and train various parts of the data individually and calculate the accuracy of the model in practice. Here we divided the dataset into 10 paths, each time we select a part out of the 10 as the testing data and the remaining a part as training parts.

Confusion matrix also unknown as an error matrix in a table that shows the overall performance of an algorithm or a clarification model. In the field of statistical analysis, a confusion matrix shows a set of test data for which the values are true or not.

		Predicted class: *P*	Predicted class: *N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 2: Overview of confusion matrix

Accuracy must be calculated for our model if means that how precisely or how close the measured value reflects the originals.

After accuracy, **Specificity** must be calculated. It refers to the test accuracy at identifying the probability of a negative test, provided the condition is absent.

Then comes **Sensitivity** which refers to the test accuracy at identifying the probability of a positive test, provided the condition is present.

Precision study refers to on how precisely or accurately, the model is measured. We develop precision investigations to check whether we are getting the correct results or not.

STEP 2: -Selecting the suitable method

There are various methods within MLR. We first develop the stepwise logistic relations between the dependent & independent variables then we split the data set into three fractions as 4/5 ,2/3 and 1/2 as the train test splitting followed by the 10-fold cross validations method.

STEP 3 – Developing equation of MLR and Confusion Matrix

Multinomial logistic regression model

a. The logit (logistic) regression model

The multinomial logistic regression [7] is fairly a generalization of a binary model. In general, logistic regression model is used to find the probability of an existing class such as yes or no based on the observation of a dataset.

It can be defined as a classification problem, where the output or target variable (y) is dependent on the given values or inputs (X) in a dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

For a response variable Y with two measurement levels (dichotomous) and explanatory variable X,
let: $\pi(x) = p\ (Y = 1 \mid X = x) = 1 - p\ (Y = 0 \mid X = x)$, the logistic regression model has logistic form for logit of this probability

$\text{Logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$, where the odds $= \frac{\pi(x)}{1-\pi(x)}$

The odds $= \exp(\alpha + \beta x)$, and the logarithm of the odds is called logit, so

$\text{Logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log[\exp(\alpha + \beta x)] = \alpha + \beta x$

The logit can be defined as the logarithm of the odds. The S – curve formed for $\pi(x)$ determines the parameter β with its rate of increase or decrease. If $(\beta > 0)$ the curve ascends and descends for $(\beta < 0)$.

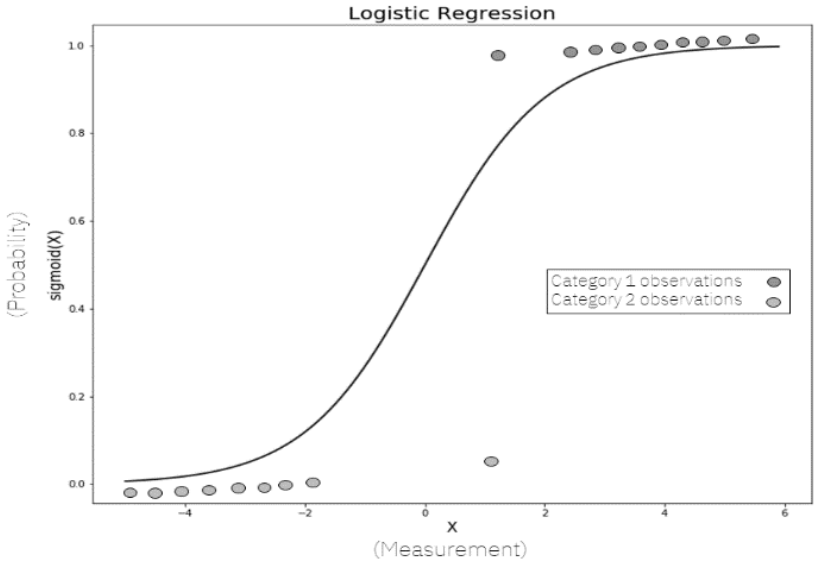

Figure 3: S – Curve of Logistic Regression

b. Multiple Logistic regression

Multiple logistic regression works on similar model building statistical techniques like Multiple Linear Regression. It aims to describe the relationship of the output to the inputs given and predict the consequences. The only distinguishable part between Multiple Logistic Regression and Multiple Linear Regression in that the dependent variable is dichotomous. Considering an example [8] where a person can know whether he/she will have a heart attack or not depending upon his/her body blood pressure, age and weight. The outcome is a binomial nominal variable i.e., heart attack vs no heart attack. The basic goal of Multiple logistic regression is to understand the functional relationship between the dependent and independent variables on what effects the probability of the outcome to change.

The logistic regression can be extending to models with multiple explanatory variables. Let k denotes number of predictors for a binary response Y by $x_1, x_2, \ldots\ldots x_k$, the model for log odds is

$\text{Logit}\ [P\ (Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$

And the alternative formula, directly specifying $\pi(x)$, is

$$\pi(x) = \frac{\exp\ (\alpha+\beta_1 x_1+ \ldots.+\beta_k x_k)}{1+\exp\ (\ \alpha+\beta_1 x_1+ \ldots.+\beta_k x_k)}$$

Here β refers to the impact of X_i in the odds for Y=1, controlled by other X_j.

If one has n independent observations with p – explanatory variables then to construct the logic, one of the categories must be considered as the base and the rest relative to it. Due to lack of ordering, any category may be used as k. Let π_j denote the multiple probability of an observation falling in the j^{th} category, to find the relationship between this probability and the p explanatory variables, $X_1, X_2, \ldots\ldots\ldots, X_p$, the Multiple logistic regression model then is

$$\log\left[\frac{\pi_j\ (x_i)}{\pi_k\ (x_i)}\right] = \alpha_{0i} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots\ldots + \beta_{pj}x_{pi}\ ,$$

Where j= 1, 2, …, (k-1), i = 1, 2, …, n. Since all the π's adds to unity, this reduces to

$$\log\left(\pi_j(x_i)\right) = \frac{\exp\ (\alpha_{0i} + \beta_{1j}\ x_{1i}+ \beta_{2j}x_{2i} + \ldots.+\beta_{pj}x_{pi})}{1+ \sum_{j=1}^{k-1} \exp(\alpha_{0i} + \beta_{1j}x_{1i} +\beta_{2j}x_{2i}+ \ldots.+\beta_{pj}x_{pi})}$$

For j = 1, 2, …, (k-1), the model parameters are estimated by the method of ML.
Practically, we use statistical software to do this fitting.

In this model, the hypothesis that is used:
H_0: None of the controlled variable X1, X2 and X3 is significantly related to Y

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

H_a: At least one of the controlled variable X1, X2 and X3 is significantly related to Y
The model of Multiple logistic regression can be represented as:

$$y = a + b_1x_1 + b_2x_2 + \cdots\cdots + b_nx_n$$

$$a = \frac{\sum y^* \sum x^2 - \sum x^* \sum (x^*y)}{n\sum x^2 - (\sum x)^2}$$

$$b_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{where } i = 1,2,3\ldots\ldots n$$

where,
y = PlacedOrNot = shows that a student gets placed or not
a = constant variable
b_1 = coefficient of first controlled variable
b_2 = coefficient of second controlled variable
b_3 = coefficient of third controlled variable
b_4 = coefficient of fourth controlled variable and so on
x_1 = Internships = No. of internships done by a student
x_2 = CGPA = Cumulative Grade Point Average
x_3 = HistoryOfBacklogs = No. of subjects in which a student failed to pass
x_4 = Hostel = Students who had completed their college while staying in the hostel

In the case of b_1, $\bar{x}$ is the mean of Internships. In the case of b_2, $\bar{x}$ is the mean of CGPA. In the case of b_3, $\bar{x}$ is the mean of HistoryOfBacklogs. In the case of b_4, $\bar{x}$ is the mean of Hostel.
In every case of b, $\bar{y}$ is the mean of PlacedOrNot.

Now let's take,
TP= TRUE POSITIVE
TN= TRUE NEGATIVE
FP= FALSE POSITIVE
FN= FALSE NEGATIVE
Now,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1\_Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

4.RESULTS AND DISCUSSION

After analysing this model, we get the results that are given below.

Table 2: For 80-20% train-test split

ATTRIBUTES	VALUE RANGE (%)
Confusion Matrix	1087 505 252 531
Accuracy	68.13
Sensitivity	68.28
Specificity	67.82
Precision	81.18
Recall	68.28
F1_Score	74.17

Table 3: For 66-34% train-test split

ATTRIBUTES	VALUE RANGE (%)
Confusion Matrix	890 416 228 424
Accuracy	67.11
Sensitivity	68.15
Specificity	65.03
Precision	79.61
Recall	68.15
F1_Score	73.43

Table 4: For 50-50% train-test split

ATTRIBUTES	VALUE RANGE (%)
Confusion Matrix	703 326 179 275
Accuracy	65.97
Sensitivity	68.35
Specificity	60.57
Precision	79.73
Recall	68.34
F1_Score	73.6

Table 5: For 10-fold cross-validation

TEST CASES	ACCURACY	SENSITIVITY	SPECIFICITY	STANDARD DEVIATION	Recall	PRECISION	F1_SCORE
01	67.89	70.95	63.33	0.49	70.95	74.27	72.57
02	65.55	62.43	70.91	0.48	62.43	78.66	69.61
03	66.22	73.79	49.46	0.46	73.78	76.38	75.06
04	68.89	72.61	53.45	0.39	72.61	86.63	79.00
05	60.86	60.71	61.33	0.43	60.71	82.42	69.92
06	72.24	69.72	74.52	0.49	69.71	71.22	70.46
07	69.23	65.79	80.28	0.42	65.78	91.46	76.53
08	73.57	69.11	81.48	0.48	69.10	86.84	76.96
09	72.57	67.7	87.67	0.42	67.69	94.44	78.86
10	61.20	62.05	91.18	0.43	62.05	95.27	75.15

5.CONCLUSION

In this paper, Multiple Logistic regression (MLR) statistical technical has been used to develop a placement predictor. The overall data has been divided into three paths referred as train-test-split following up with 10-fold cross validation and developing the confusion matrix.

The recorded accuracy for the 4/5, 2/3, and 1/2 train-test-split are 68.13%, 67.11%, and 65.97 % respectively.

This model in proposed to predict the placement results of the students based on their academic performance and company preferences. Thus, the predictor shows the required criteria for getting recruited. This could really benefit the education institutions as they can artist their students in order to quality and meet the requirements of the companies. Also, the students can prepare themselves accordingly fulfilling the demands as presented.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11337

6.REFERENCES

[1] Irene Treesa Jose, Daibin Raju, Jeebu Abraham Aniyankunju, Joel James, Mereen, and Thomas Vadakkel, "Placement Prediction using Various Machine Learning Models and their Efficiency Comparison". International Journal of Innovative Science and Research Technology ISSN No: -2456-2165. Volume 5, Issue 5, May – 2020.

[2] D. Satish Kumar, Zailan Bin Siri, D.S. Rao, and S. Anusha, "Predicting Student's Campus Placement Probability using Binary Logistic Regression". International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019

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

[4] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy', "A Comparative Analysis of Techniques for Predicting Academic Performance", 37th ASEE/IEEE Frontiers in Education Conference, IEEE, 2007.

[5] Irene Treesa Jose, Daibin Raju, Jeebu Abraham Aniyankunju, Joel James, Mereen, and Thomas Vadakkel, "Placement Prediction using Various Machine Learning Models and their Efficiency Comparison". International Journal of Innovative Science and Research Technology ISSN No: -2456-2165. Volume 5, Issue 5, May – 2020

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

[7] Logistic Regression, Wikipedia, https://en.wikipedia.org/wiki/Logistic_regression#cite_note-1

[8] "Handbook of Biological Statistics~ John H. McDonald", Multiple Logistic Regression, https://www.biostathandbook.com/multiplelogistic.html