AN EFFECTIVE ANALYSIS FOR STUDENT JOB OPPORTUNITIES USING MULTIPLE LINEAR REGRESSION MODEL

J.Umamageswaran[1], S.MohanRaj [2], Allamadhev [3], M. Rahul [4], Logesh [5]
[1] Assistant Professor,Department of Information Technology, R.M.K.Engineering college/
[2] Senior Network Engineer, NOKIA. [3,4,5] Department of Information Technology, R.M.K.Engineering college.

Abstract

Nowadays, the job opportunity for university and college graduates is becoming a disputed point for the people in India. With the increased growth of current technology, lot and lot of people have realized that only to uphold the balances and transparency about the information can provide more opportunities for many graduates. Normally admission and reputation is mainly depends on employability of the institute graduate. Hence, all college strives to strengthen the placement department. In this paper, we proposed an machine learning algorithm to analyse previous year's student's historical data and predict placement chance for the current students and the percentage placement chance of the institution. Furthermore, we have implemented an multi linear regression algorithms to predict the placement chance of students to each company. Data relating to this work collected from the same institution for which the prediction the placement chance and placement percentage need to be establishing from 2014 to 2018. The year 2017 is the test data and 2018 data is considered as current data. We have also applied appropriate data pre-processing methods are applied. This proposed model is compared with Simple Linear Regression. From the results obtained, it is found that the proposed algorithm predicts better in comparison with other algorithms.

Keywords: ***Data mining, prediction, Simple Linear Regression, Multiple Linear regression***

1. INTRODUCTION

It is well-known fact all over the world that the admission of the students in a university/college depends upon the placements. To determine the quality of any educational institution we consider the placement is one of the major factors. Hence, every educational institute aspires hard to provide effective placements to their students. An educational institution contains huge volume of student data. This data is affluence of information, but it is too difficult and large for any one person to understand in its entirety. It is essential task for finding of students who can get the offer based their performance.

It is responsible for every educational institution must help the students to be successful in life by providing a good job opportunity. However, we observed that every institution finds it very difficult to make every student to get job hence the purpose of this study is to predict job chances towards each company and thus find the placement percentage of the institution for the current academic year. This would help the college to analyse the status of the college in comparison with other college and respond appropriately to improve. The main objective of this paper is to predict students' chances of placing in each company using machine-learning algorithms.

2. RELATED WORKS

characteristics of data is an in education research. We can classify the students based on their knowledge and skills such as logical score, verbal score, and quantitative analysis score, programming score, and academic performance. We use these score as a data set and help us to predict the chance

Many researchers working to discover the best mining techniques for solving prediction problems. In this regard various research have been done. Few of the related works are shown below:

Jae H. Min et aI., 2001 [1] Applies support vector machines (SVMs) and used a grid-search technique using 5-fold cross validation to find out the optimal parameter values of kernel function of SVM, they applied SVM to bankruptcy prediction problem, and showed its attractive prediction power compared to the existing methods; l.A.K. Suykens et aI., 1998. [2]discussed a least squares version of support vector machine classifiers and illustrated that a least squares SVM with RBF kernel is readily found with excellent generalization performance and low computational cost, Tung-Kuang Wu et aI., 2008

[3] Apply two well-known artificial intelligence techniques, artificial neural network (ANN) and support vector machine (SVM), to the LD diagnosis problem; Guha, S et aI., 1999 [4]

Proposed a new concept of links to measure the similarity/proximity between a pair of data points with categorical attributes and developed a robust hierarchical clustering algorithm, KakotiMahanta et aI., 2005
[5] Prove that under certain conditions, the final clusters obtained by the algorithm are nothing but the connected components of a certain graph with the input data-points as vertices, AgnieszkaPrusiewicz et al., [6] 2010 proposal for services recommendation in online educational systems based on service oriented architecture are introduced, Christian Borgelt2005.

A new data structure for frequent item set mining algorithms. BalazsRacz, D 2004[8] described an implementation of a pattern growth based frequent item set mining algorithm. Data structure can accommodate top-down recursive approach, thereby further reducing memory need and computation Time, Ke Wang, Liu Tang et aI., 2002[9]propose an efficient algorithm, called TD-FP-Growth (the shorthand for Top-Down FP Growth), to mine frequent patterns, Sudheep Elayidom et aI., 2011 [10] attempt to help the prospective students to make wise career decisions using technologies like data mining using decision trees, Naive Bayes and artificial neural networks, Ajay Kumar Palet al,.2013[1l]suggested that Naive Bayes classifier has the potential to significantly improve the conventional classification methods for use in placement among all the machine learning algorithm tested, K. Pal et aI., 2013[12] describe the use of data mining techniques to improve the efficiency of academic performance in the educational institutions, B.K. Bharadwaj et aI., 2011 [13] the classification task is used on the student database to predict the students' division on the basis of the previous database;S. K. Yadav et aI., 2012[14]focusing upon methodologies for extracting useful knowledge from data and there are several useful KDD tools to extract the knowledge.

3. PROPOSED MODEL

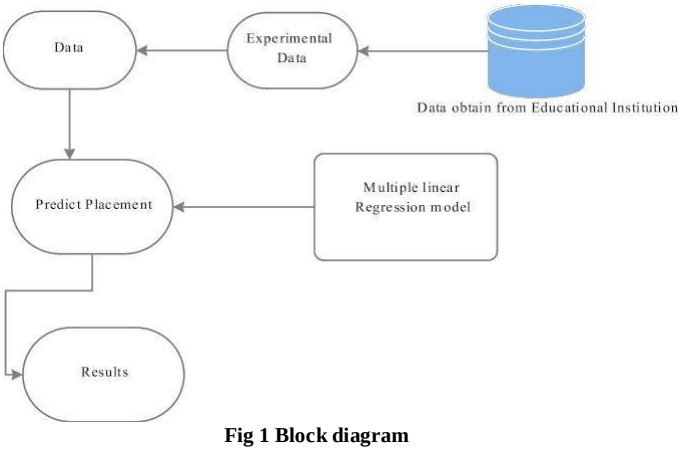

Fig 1 Block diagram

The algorithm of the proposed model, along with its computational processes for predicting placement chance given below:

Step 1: Data to be collection. The aim is to find the skillful students in the college under consideration viz., XX for the year 2016. In this college there were 1,434 students. These students hailed from various courses that were operative in the college. The courses are IT, CSE, MECH, ECE, and CIVIL Engineering.
Step 2: Predict employability chance, this step predicts placement chances of the student and percentage placement towards a company through the Proposed algorithm.

4. Data Description

Step 3: Evaluate the result The obtained result is compared to Simple Linear Regression.

The objective is to predict the Employability chance of students identified as proficient students in the college identified as XX. Basic requirement of any prediction problem is the existence of previous or past data based on which future is predicted. Data are collected from a college XX identified above, that offers various courses.

Data collection divided into three types.

- Historic Data: Collected for the duration of 10 years starting from 2009 to 2018
- Test data: Collected for the year 2017.
- Current data: Students identified as proficient students for the year 2018.

Table 1: Data Description

Variables	Description	Possible Values
Year	Year for which the data is entered	{int}
Reg-no	Register number of the student's	{int}
Branch	Branch (IT,CSE,CIVIL .... etc) of the student	{1,2,3,4,5….}
Percent	Over all Percentage of the students	{65,71,82, ... ,100}
Skills	Analytical, Verbal,Coding Ability	{50,60,70,.....100}
Company	Company student placed(Zoho,CTS,Infosys)	{Text}
Arrears	Backlogs in subjects	{1,2...5}
Effective score	Effective-score=percent+skills* 10, it shows the overall performance of the student.	{1,10,30….100}
Placed	Student Placed based on her performance	{Text}

- Year: Student-completed education. Data collected were from 2009-2018.
- Skills : it shows the overall Skills of the
- Reg-no: Register number of the student. It takes any integer values.
- Branch: represents the name of the Branch. It can take only text values ranging from A-Z
- Percent: various marks scored by students in subjects. It can take only the numeric values from 0 to 100.
-

student.. It can take only the numeric values from 0 to 100.

- Effective-score: it shows the overall performance of the student Formula to calculate Effective-score is as follows Effective score = percent + skills * 10 It can take only the numeric values from 0 to 200.
- Placed : Placed based on student performance. Value is taken in the form of Yes\No,

  IF Yes, Student placed,
  ELSE, Students Not Placed.

4. PROPOSED ALGORITHM

Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. We can define it as:

> Multiple Linear Regression is one of the important machine learning algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

The multiple linear regressions describes the relationship between one continuous dependent variable (y) and two or more independent variables (x_1, x_2, x_3… etc).
Note that it says CONTINUOUS dependent variable. Since y is the sum of beta, betax, betax etc etc, the resulting y will be a number, a continuous variable, instead of a “yes”, “no” answer (categorical).

Multiple linear regression is what you can use when you have a bunch of different independent variables!

Multiple regression analysis has three main uses.

- You can look at the strength of the effect of the independent variables on the dependent variable.
- You can use it to ask how much the dependent variable will change if the independent variables are changed.

- You can also use it to predict trends and future values.

Data Set - old students coding score ,verbal score , analytical score ,overall percentage , company they have placed
Input - current students coding score ,verbal score , analytical score ,overall percentage

Output - Chance of Company that can be placed

Note
Attribute Company Depend upon the independent variables(coding, verbal, analytical sore)

5. SIMULATION AND EMPIRICAL RESULTS

In the concept of statistical solving simple linear regression is an empirical approach and it can solve the tasks by considering the historical data set of the climate values or parameters. It can only consist of a single dependent variable and independent variable. In the simple linear regression model there can exist only two variables.

The representation of multiple linear regression will be like

$$Y=a_0+a_1x_1+ a_2x_2+ a_3x_3+ a_4x_4+---$$

..(1)

In which Y=dependent variable
x_1,x_2,x_3,x_4=independent variables
a_0,a_1,a_2,a_3,a_4=regression coefficients

In multiple linear regressions we can represent it in the form of mathematical equations by calculating slope and regression coefficients. The strength and direction of the association between the two variables can be estimated by using the regression coefficient formula. Similarly, there are various correlation coefficient formulas that can also be available in the mathematical and statistical evolution processing.

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

..(2)

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

..(3)

At this point, you should notice that all the terms from the one variable case appear in the two variable cases. In the two variable cases, the other X variable also appears in the equation. For example, X_2 appears in the equation for b_1. Note that terms corresponding to the variance of both X variables occur in the slopes. Also note that a term corresponding to the covariance of

verbal	coding	analytical	academics	backlogs	placed	Effective score
84	89	81	74	0	yes	82
83	85	88	85	1	yes	85.25
87	87	74	82	0	yes	82.5
87	94	89	77	0	yes	86.75
79	99	85	79	2	yes	85.5
77	98	73	60	5	no	77
88	96	74	67	3	no	81.25
89	83	87	75	4	no	83.5
76	95	89	70	5	no	82.5
81	88	80	94	3	no	85.75
78	83	86	88	4	no	83.75

X1 and X2 (sum of deviation cross-products) also appears in the formula for the slope.

The equation for *a* with two independent variables is:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

..(4)

Initially Consideration of Requirements for MLR

- Data set which is in the form of climate variables with numerical values.
- Database maintenance by Excel software.
- JDK software.
- The IDE software is called NetBeans.
- The package of java Jxl.jar.

Weka software.

Implementation steps

1) Defining the data set variables in excel.
2) Installing the NetBeans IDE along with JDK if JDK is not installed in your system.
3) Creation of project in NetBeans framework. Importing the jxl.jar package into the project library folder.
4) Declaring the variables and source file path in the by creating a file.
5) Implementing correlation coefficient calculation using programming.
6) Define regression equation.
7) Calculating slope value from using a simple regression equation and coefficient value.
8) Final forming the predicted equation with slope and coefficient values.
9) Implementing the equation with the climate parameters and comparing the actual values with the predicted values.
10) Calculate the RMSE using past data sets and obtained values.
11) Install the WEKA software.
12) Obtaining confusion matrix using WEKA.
13) Calculating specificity and sensitivity.

Finally implementing the accuracy formula using specificity and sensitivity.

Considered data set is:

Attributes: Reg-no, Branch, Verbal score, Analytical score, Coding score, Academic score, Backlog. **After preprocessing:** Reg-no, Branch is eliminated or pruned.

Table 2 : Specific record for company

6. EXPERIMENT RESULTS

In this section we present the experimental results for the regression methodologies simple regression and multiple regression, we have introduced briefly in the previous sections regarding regression concepts. The concept of regression can be implemented by calculating coefficient, slope and the considered data set . As well as the performance of predicting the future values can be calculated by using multiple linear regression algorithms. By using Net beans and weka framework, we implemented our project.

Sample Test Data

	const	verbal	coding	analytical	academics	backlogs
subiksha	1	90	90	80	65	2
seema	1	65	60	80	85	0
raghul	1	75	75	95	78	0
harini	1	75	85	80	55	4
divya	1	65	90	96	80	5

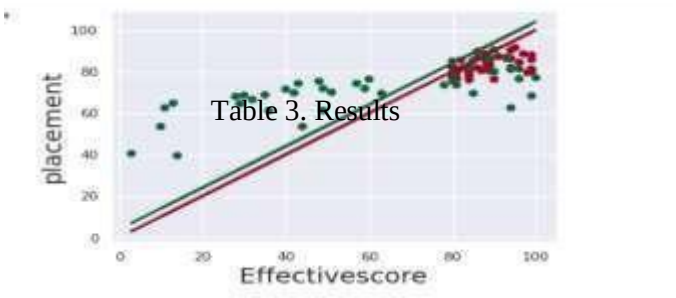

Table 3. Results

	const	verbal	coding	analytical	academics	backlogs	Predictions
subiksha	1	90	90	80	65	2	0.701997
seema	1	65	60	80	85	0	0.483827
raghul	1	75	75	95	78	0	0.724305
harini	1	75	85	80	55	4	0.219662
divya	1	65	90	96	80	5	0.012496

7. COMPARISON WITH LINEAR REGRESSION MODEL

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable

$y = a_0 + a_1 * x$ ## Lin Equ ..(5)

The motive of the linear regression algorithm is to find the best values for a0 and a1. Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted
value and the actual value

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \quad ..(6)$$

Minimization and Cost Function.

We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function. Now, using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

Here only one attribute can be selected as independent variable

Independent variable: Effective score

Effective score= ((verbal+coding+Analytical+Academics)/400)* 100

Dataset

Sno	Effective score	placed
1	77.5	yes
2	80	yes
3	75.75	yes
4	76.25	yes
5	86.5	yes
6	83.25	yes
7	81	yes
8	63.25	no
9	67.75	no
10	66.5	no
11	73.75	no
12	70.5	no
13	67.25	no
14	69.25	no
15	66.75	no

Output

8. CONCLUSION AND FUTURE ENHANCEMENT

In this project we have implemented the simple regression methodology, multiple regression and we predicted the values, the multiple regression error rate also less when comparing with simple linear regression. Finally concluding that multiple linear regressions can be more better than simple linear regression. Because simple linear regression can use only a single parameter [Effective score] to predict the outcome. But Multiple linear regression uses multiple parameters to predict the Placement chance over a company. Multiple linear regression algorithms is more accurate than simple linear regression.

Student Placement Predictor is a system, which predicts student placement status using machine-learning techniques. Many research papers are there related to the educational sector, all these papers mainly concentrate on student performance predictions. All these predictions help the institute to improvise the student performance and can come up with 100% results. Many of the previous research papers concentrate on a less number of parameters such as CGPA and Arrears for placement status prediction which leads to less accurate results, but proposed work contains many educational parameters to predict placement status which will be more accurate. The main motive of the paper is to predict students' chances of placing in each company.

In Future, we would like to focus to add some more parameters to predict more efficient placement status. We can also enhance the project by predicting some solutions or suggestions for the output generated by the system by using Reinforcement learning.

REFERENCE

[1] Student Placement Analyzer, A Recommendation System Using Machine Learning, 2017 International Conference on advanced computing and communication systems (ICACCS-2017), , Coimbatore, Jan 06-07,2017

[2] Prediction Model for Students Future Development by Deep Learning and Tensor Flow Artificial Intelligence Engine, 4th IEEE International Conference on Information Management. 2018

[3] Kohavi, R and F, Provost Machine Learning 30:271-27 (1998).

[4] Guha, S.; Rastogi, R.; Kyuseok Shim "ROCK: a robust clustering algorithm for categorical attributes"

[5] J.Umamageswaran, Dr. Balasubdra, Resource Allocation using Multi-Agent Based Approaches in Polynimbus Cloud Strategy: A Survey, International Journal of Control and Automation, vol. 12, issue. 6, pp. 371-389, 2017.

[6] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006.

[7] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Harlow, U.K.: Addison-Wesley, 1999.

[8] D. Gouscos, G. Mentzas, and P. Georgiadis, "PASSPORT, a novel architectural model for the provision of seamless cross-border e-government services," in Proc. Int. Conf. Database Expert Syst. Appl., Munich Germany, 2001, pp. 318–322.

[9] M. Janssen, R. Wagenaar, and J. Beerens, "Towards a flexible ICTarchitecture for multi-channel e-government service provisioning," in Proc. Hawaii Int. Conf. Syst. Sci., Big Island, HI, 2003, p. 148.

[10] [Sacco, "User-centric access to e-government information: e-Citizen discovery of e-services," in Proc. Int. Symp. Semantic Web Meets eGovernment, CA, 2006, pp.114–116.

[11] X. Fang and O. L. Sheng, "Designing a better Web portal for digital government: A Web-mining based approach," in Proc. Nat. Conf. Digit.GA, 2005, pp. 277–278.