Student Placement Analyzer: A Recommendation System Using Machine Learning

[1]Senthil Kumar Thangavel ,
Dept. of CSE
Amrita School of Engineering,
Coimbatore,
Amrita Vishwa Vidyapeetham
Amrita University, India
t_senthilkumar@cb.amrita.edu[1],

[2]Divya Bharathi P
Dept. of CSE
Amrita School of Engineering,
Coimbatore,
Amrita Vishwa Vidyapeetham
Amrita University, India
cb.en.p2cse15005@cb.students.amrita.edu [2],

[3,]Abijith Sankar,
Dept. of CSE
Amrita School of Engineering,
Coimbatore,
Amrita Vishwa Vidyapeetham
Amrita University, India
abijith.skn@gmail.com [3]

Abstract. One of the biggest challenges that higher learning institutions face today is to improve the placement performance of students. The placement prediction is more complex when the complexity of educational entities increase. Educational institutes look for more efficient technology that assist better management and support decision making procedures or assist them to set new strategies. One of the effective ways to address the challenges for improving the quality is to provide new knowledge related to the educational processes and entities to the managerial system. With the machine learning techniques the knowledge can be extracted from operational and historical data that resides within the educational organization's databases using. The dataset for system implementation contains information about past data of students. These data are used for training the model for rule identification and for testing the model for classification. This paper presents a recommendation system that predicts the students to have one of the five placement statuses, viz.,Dream Company, Core Company, Mass Recruiters, Not Eligible and Not Interested in Placements. This model helps the placement cell within an organization to identify the prospective students and pay attention to and improve their technical as well as interpersonal skills. Furthermore, the students in pre-final and final years of their B. Tech course can also use this system to know their individual placement status that they are most likely to achieve. With this they can put in more hardwork for getting placed in to the companies that belong to higher hierarchies.

Keywords: Decision Tree Classifier, Sci-kit Learn, Machine Learning, Prediction, Python

I. INTRODUCTION

The primary aim of students who join professional courses in higher learning institutions is to secure a well-paid job in a reputed organization. Professional education can be either completely technical or it can be managerial as well. Bachelors of Technology (B. Tech) provides technical education to students in various fields such as Computer Science and Engineering, Electronics and Communication Engineering, Civil Engineering Mechanical Engineering, etc. This degree is aimed at making students experts in state of the art conjectural as well as practical knowledge in various engineering branches. The prediction of placement status that B. Tech students are most likely to achieve will help students to put in more hard work to make appropriate progress in stepping into a career in various technical fields. It will also help the teachers as well as placement cell in an institution to provide proper care towards the improvement of students in the duration of course. A high placement rate is a key entity in building the reputation of an educational institution. Hence such a system has a significant place in the educational system of any higher learning institution. We used decision tree classifier within Scikit-learn-a machine learning module in python having simple and efficient data mining and data analytics capability- for the implementation of the system.

II.MATERIALS AND METHODS

Machine Learning

Machine Learning deals with the development, analysis and study of algorithms that can automatically detect patterns from data and use it to predict future data or perform decision making [1]. Machine learning does its functionality by creating models out of it [2]. Machine Learning has become widespread and has its applications in the field of bioinformatics, computer vision, robot locomotion, computational finance, search engine etc.

Decision Tree Learning

In real world problems, observations are made on entities associated with a problem so as to make inferences on the target value of those entities. This mapping is encompassed in a predictive model with the help of decision trees. This method of learning is referred to as Decision Tree Learning. This is one of the predictive modelling methods that can be found in the fields of data mining [3], machine learning and statistics. In this model we have made use of classification trees, a typical decision tree in which the predictor variable (target variable) takes on finite set of categorical values only. In this type of trees, the leaves represent the class labels and the branches represent the splitting path through which a decision travels from root to the leaf of the tree.

978-1-5090-4559-4/17/$31.00©2017IEEE

SCI-Kit Learn

Scikit-learn is an open source machine learning module in python [4] that is comprised of wide range of classification, clustering and regression algorithms in machine learning. Major algorithms featured are Naive Bayes, Decision Tree, Random Forests, Support Vector Machines, Logistic Regression, Gradient Boosting, k-Means and DBSCAN.This module is primarily aimed at solving supervised and unsupervised problems. It aims at making machine learning accessible to novices by providing an abstraction using a general-purpose high-level language. Ease of use, documentation, performance and API consistency are the key features of this module.

Background and Related Work

Machine Learning techniques has a significant role in deriving innovative knowledge in the educational field so as to help students for their better performance in placement. Many scientists across the world has done considerable amount of work in determining the methodologies for performance analysis and placement. Few of the relevant works in this field are listed out so as to obtain an idea on what has been done so far and what further growth is expected in this area of work.

Hijazi and Naqvi [11] conducted a study to find the factors affecting the academic performance of students. They made use of questionnaires to elicit information from students highlighting factors such as income factor, parents' educational background, size of the family, regularity of teachers, subject interest created by the teachers and student's interest in co-curricular activities. They used Pearson Correlation Coefficient to highlight the important factors and they found that mother's education and family income played an important role in students' academic performance.

Pal and Pal [6] conducted a study on student data that have information on their academic records and proposed a classification model to find an efficient method to predict student placements. They concluded that Naïve Bayes classifier is the best classification method for use in placements in comparison with Multilayer Perceptron and J48 algorithms.

Ramanathan, Swarnalatha and Gopal [7] conducted a study using sum of difference method for students' placement prediction. They used the attributes such as age, academic records, achievements etc. for the prediction. They concluded that based on their results higher learning institutions can offerits students a superior education.

Arora and Badal [8] conducted a study to predict student placements using data mining. They made predictions on MCA students in Ghaziabad in UP, considering parameters such as MCA result, Communication skills, programming skills, co-curricular activity participation, gender, 12^{th} result and graduation result. They concluded that their model based on decision tree algorithm can assist the placement cell and faculties in identifying set of students that are likely to face problem during final placements.

Elayidom, Idikkula and Alexander [9] designed a generalized data mining framework for placement chance prediction problems. They considered the students' Entrance Rank, Gender, Sector and Reservation Category to predict the branch of study that is Excellent, Good, Average or Poor for him/her using decision trees and neural networks.

Naik and Purohit [10] made a study to use prediction technique using data mining for producing knowledge about students of MCA course before admitting them.

Logistic Regression

Logistic regression [5] is a probabilistic view of classification. The approach is used when dependent variable is binary(dichotomous) and help in predicting a discrete outcome. Prediction is done using the probability scores. Logistic regression gives knowledge of the relationships among the variables.

In our project we have used logistic regression to find the probability of the student getting placed belonging to different departments by using the biglm package in R tool. Biglm processes the data in chunks at a time to perform the regression optimization and does not need larger memory allocations to the computer because it performs calculations on smaller data sets. The number of lines to be processed at any time is specified by chunk size and we have chosen a chunk size of 100. Biglm and RODBC package are required to perform calculations using database connectivity and biglm will identify the sql Query as a data frame.

Logistic regression can also be worked out in MS excel and the steps are as follows:

Step 1: Deciding a dependent variable which is placement variable in our case and independent variables which is gender, 12th mark, number of arrears, arrear history and CGPA.

Step 2: Pivot the independent variables.

Step 3: The value of constant, coefficient of A,B,C,D,E,F which are independent variables are found using regression analysis for different data using training data in excel

where

Y- Placement variable(dependent), A-Male, B-Female, C-12^{th} mark, D-number of arrears, E- Arrears history, F-CGPA.

Step 4: To find out the probability we are using the logit (L) value which is given by

Y = constant + A*(coefficient of A) + B*(coefficient of B) + C*(coefficient of C) + D*(coefficient of D) + E*(coefficient of E) + F*(coefficient of F).

Step 5: Then using these constant and coefficient values the probabilities of getting placed for the test data is

identified.
Probability of getting placed P(Y) = e^L / (1+e^L).
Step 6: The probability of getting placed for test data is identified.

Decision Learning Process

Several factors has to be considered for predicting the placement status a student would most likely to achieve at the end of placements in the final year. As such a prediction has a huge significance in identifying the prospective students who needs special attention to clear the placement procedures, the factors chosen should have the highest level of impact factor on the prediction.

Data Preparation

The dataset used for training as well as testing was obtained from the database of the placement cell in the institution. The training data sample had a size of 2205 tuples and the testing dataset had a size of 289.

Data Selection and Transformation

In this, only those factors which had a solid impact on the placement status prediction was chosen after doing a combination of forward selection and backward elimination. The predictor and response variables are shown in the table 1.

The domain value for only one variable needs to be explained as rest are self-explanatory

- COMP – The placement status the student is most likely to achieve. It is split into 5 class values: Dream Company- Companies offering CTC≥10 lpa, Core Company- Companies offering CTC≥4.5 lpa and CTC<10lpa, Mass Recruiter/ Common Company- Companies offering CTC<4.5 lpa, Not Interested- Those who want to go for higher studies, Not Eligible- Those who does not meet the placement eligibility criteria, Not Placed- Those who doesn't get placed despite of being eligible and interested in placements.

Table 1: Student Related Variables

Variables	Description	Possible Values
Dept	Department	{AE, CHEM, CIVIL, CSE, ECE, EEE, EIE, MECHANICAL}
Gender	Gender of Student	{M,F}
Board	12^{th} Board of studies	{SB-HP, SB-GJ, SB-MH. ST-AP, ST-TN, SB-BI, ST-KARNATAKA, CBSE, ICSE, SB-RJ, ST-KL}
Location	Location where 12^{th} is completed	Pin code
12^{th}	Marks percentage in 12^{th}	Marks in %
NA	Number of Standing Arrears	Integer
AH	Arrear History	Integer
CGPA	Cumulative Grade Point Average	Float value out of 10
COMP	Placement Status	{Dream Company, Core Company, Mass Recruiter/Common Company, Not Eligible, Not Interested, Not Placed}

Implementation of Machine Learning Model

Python is one of the best data analytics language used widely in the industry. It has sophisticated data mining and machine learning capabilities and is the best practical language for building products. This makes python a favourite in data processing. In data processing, there's often a trade-off between scale and sophistication, and Python has emerged as a compromise. IPython notebook and NumPy can be used as a scratchpad for lighter work, while Python is a powerful tool for medium-scale data processing. Python also has lot of advantages like rich data community, offering vast amounts of toolkits and features.Sci-kit learn is a greatly a sophisticated module available in python, which comprises of almost all major machine learning algorithms. The training dataset which contains the placement data of 2014 pass out batch is loaded to the python code, macros are attached to the variables for their easy processing and is fit to a decision tree classifier model using Scikit libraries. Once the modelling is completed, the test data is uploaded to the python, the variables are read using macros and is provided to a predict function available in Scikit learn. This produces the macro result corresponding to the placement status class with respect to the training data. Finally the macro result is mapped back to the placement status variable used in the model.

III.RESULTS

The result obtained on the test data is given in Table 2.

Table 2: Test Data Results

	Mass Recruiter/ Common Company	Core Company	Dream Company	Not Eligible	Not Interested	Not Placed	Total
Girls	35	30	24	6	5	14	114
Boys	63	26	27	23	9	27	175
Total	98	56	51	29	14	41	289

The accuracy of the system is measured against real life placement statuses. Since the real life placement status is a confidential data, the placement cell provided us with a dataset of size 60 to measure the accuracy. The real life placement status of 60 students are classified into the five class labels used by the system and then compared with the prediction made by the system. The comparison of results are given in table 3.

Table 3: Comparison of Results with Real World Data

Total Number of Instances	60
Number of instances correctly classified	43
Number of instancesincorrectly classified	17
Prediction Accuracy in Percentage	71.66
Running Time of the whole system(seconds)	10

The results predicted by the system is compared against predictions made by data analysis tools such as weka and datameer. The results of this analysis are given in table 4.

Table 4: Comparison of System Results with Different Data Analysis Tools

Tool	Total Number of Instances	Number of matching instances	Number of mismatching instances	Match Percentage
Datameer	289	197	92	68.16
Weka	289	244	45	84.42

To analyse the efficiency of decision tree learning, we compared the prediction made by our system across different algorithms in Weka tool. The results are given in table 5.

Table 5: Analysis of Different Classifiers

Algorithm	Time to build(seconds)	Total Number of Instances	Number of matching instances	Number of mismatching instances	Match Percentage
Classification Via Regression	0.49	289	148	141	51.21
Logistic Regression Classification	0.19	289	145	144	50.17
Decision Tree	.01	289	244	45	84.42
Metabagging Classifier	0.14		213	76	73.70
Naïve Bayes	0	289	128	161	44.29

Figure 1 and 2 are the graphical representation of the simulation results.

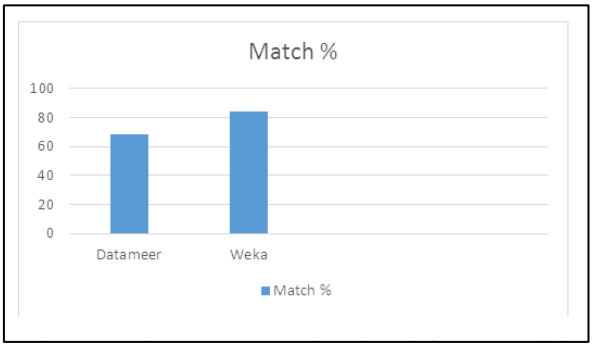

Fig.1. Match Percentage between System Results and Results from Data Analytics tools

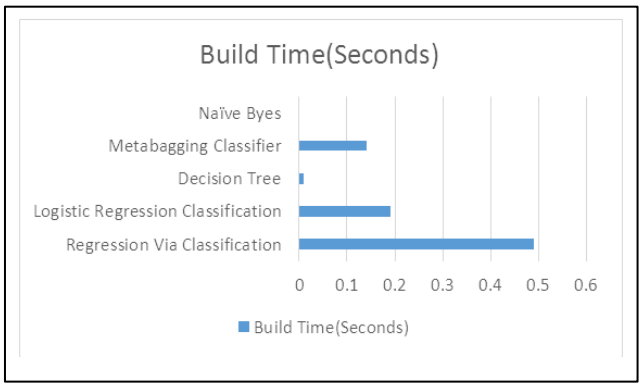

Fig 2: Build Time of different classifiers

IV. DISCUSSION

Out of the 289 students, the system predicted 33.91% students to get placed in mass recruiter or common company class. The number of students who would get in to Core Company and Dream Company are found to be in and around 18% each. System predicts 150% more boys than girls are most likely not to get placed in any tier of companies. These predictions when compared across the real life data of 60 students, gave 71.66% accuracy, which is a significant accuracy measure to consider a prediction system as reliable. The accuracy of the system will be improved once, the outliers in the test data are removed. The system has a considerable matching when compared to results obtained from weka tool. Since weka is a highly used data analysis tool, 84% matching with weka results further underscores the reliability of the system. The analysis of various algorithms in weka proves that decision tree classifier stands out with 0.01 seconds of running time and 84.42% accuracy. This proves the efficiency of the methodology employed in the system.

V.CONCLUSION

As a conclusion, we have met the objectives, which is to predict the placement status the students in Btech are most likely to have at the end of their final year placements. The accuracy of 71.66% with tested real life data indicates that the system is reliable for carrying out its major objectives, which is to help teachers and placement cell in an institution to find the prospective students and provide them with better coaching so as to excel in placement processes by various companies. The system helps in improving the placement rate of an institution thereby can act as a key element in improving the reputation of the institution. From the analysis, it is clear that the methodology used in system implementation is efficient enough to considerably improve the state of the art of classification technique that are employed so far in placement prediction field.

REFERENCES

[1]. Kohavi, R. and F. Provost(1998) Glossary of terms. Machine Learning 30:271-274.

[2]. Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer. ISBN0-387- 31073-8.

[3]. Rokach, L and O. Maimon (2008)Data mining with decision trees: theory and applications.World Scientific Pub Co Inc. ISBN 978-98127717711

[4]. Pedregosa, F , G. Varoquax, A. Gramfort, V. Michel, B. Thrion, O. Grisel, M.Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau(2011)Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12: 2825-2830

[5]. Abijith Sankar, P. Divya Bharathi, M. Midhun, K. Vijay, and T. Senthil Kumar, "A Conjectural Study on Machine Learning Algorithms," In Advances in Intelligent Systems and Computing published – Springer - from the proceedings of International Conference ICSCS 2015 2016, Vol.397, pp. 105-116.

[6]. Pal, A.K. and S. Pal (2013)Analysis and Mining of Educational Data for Predicting the Performance of Students. (IJECCE) International Journal of ElectronicsCommunication and Computer Engineering, Vol. 4, Issue 5, pp. 1560-1565, ISSN: 2278-4209, 2013.

[7]. Ramanathan, L., P. Swarnalathat and G.D. Gopal (2014)Mining Educational Data for Students' Placement Prediction using Sum of Difference Method. International Journal of Computer Applications 99(18): 36-39

[8]. Arora, R.K. and Dr. D. Badal (2014)Placement Prediction Through Data Mining.International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 7

[9]. Elayidom, S., S. M. IdikkulaandJ.Alexander(2011)A Generalized Data mining Framework for Placement Chance Prediction Problems. International Journal of Computer Applications(0975– 8887) Volume 31– No.3

[10]. Naik, N. and S. Purohit(2012) Prediction of Final Result and Placement of Students using Classification Algorithm. International Journal of Computer Applications (0975 – 8887) Volume 56-No.12

[11]. Hijazi, S.T. and R. S. M. M. Naqvi (2006)Factors affecting student's performance: A Case of Private Colleges, Bangladesh e-Journal of Sociology, Vol. 3, No. 1