Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019)
IEEE Xplore Part Number:CFP19OSV-ART; ISBN:978-1-7281-4365-1

A Comparative Study on Machine Learning Algorithms for Predicting the Placement Information of Under Graduate Students

Tadi Aravind[1], Bhimavarapu Sasidhar Reddy[2], Sai Avinash[3], Jeyakumar G[4]
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India.
[1]cb.en.u4cse17062@cb.amrita.students.edu, [2]cb.en.u4cse17013@cb.amrita.students.edu,
[3]cb.en.u4cse17015@cb.amrita.students, [4]g_jeyakumar@cb.amrita.edu

Abstract—**As Machine Learning (ML) algorithms are becoming popular to solve challenging and interesting real world prediction problems around us, the interest level of student community has been increased in learning the principles of ML and its different algorithms. This includes by implementing the commonly known machine learning algorithms and tests them by solving simple prediction problems around the student community present in educational system. In this line, this paper proposes to solve the student placement prediction problem using linear regression model, K-neighbor regression model, decision tree regression model, XGBoost regression model, gradient boost regression model, light GBM regression model and random tree classifier model. This work is carried out in two phases. The Phase 1 is done on a simple data set and the Phase 2 is done with an extended data set with added additional features about the students. This research work presents the comparative performance analysis of these seven models by implementing them with these two data sets. The performance measurements considered in this study are prediction accuracy and the root mean square error (RMSE).**

Keywords—Machine learning; Learning model; Prediction; regression model; Student placement prediction.

I. INTRODUCTION

The principles of Machine Learning (*ML*) algorithms had becoming popular nowadays to solve variety of problem around us and to support us in many decision making processes. The web search engines like google, bing etc are popular because their systems learnt how to rank pages according to their usage by the user through complex learning algorithms in *ML*. There are spam detector algorithms in our daily usage applications like gmail, hot mail, proton mail etc., they do the work of classifying our mails and move the spam mails to spam folder location using the *ML* principles. In the process of data mining for automation, the task of constructing the medical records from the health care biological data also involves the learning algorithms. There are photo tagging applications in social network systems viz facebook and googlephotos where they automatically tag the person in a photo using certain high level *ML* based facial recognition algorithms which runs behind the application.

Considering the wide use of *ML* algorithms/models in all the scenarios, understanding the principles of each of the *ML* algorithms/models are inevitable. This paper is an attempt to demonstrate variety machine learning models in solving the student prediction problem. The experimental set up, the results obtained and the comparison of performance of the models used in this study are presented in this paper.

Rest of the paper is organized follows: Section 2 narrates few important related works, Section 3 discusses the details of the proposed work, Section 4 explains the results and discussion of the machine learning models and the Section 5 concludes the paper.

II. RELATED WORKS

This section summarizes the works in the state-of-the-art related to student placement prediction. In computer science education system programming plays an important role. The students are ranked according to their skill level in programming. The software companies also recruit and evaluate employees by their programming skills through some programming tests and contests, etc. A system to predict the placement and ranking of programming contests, which relieves teachers and recruiters from their burden, is proposed in [1].

Educational data mining is mining patterns from educational data [2]. This study includes the interesting research areas viz modeling the students' learning curve [3] and modeling the learning style [4]. This field also has the allied area of research such as modeling human behavior for predicting the memory process [5]. For taking important decisions or to assist educators, data mining techniques are used to discover useful information in educational environment. For example, a work showing how to predict the bad coding practices of the students is presented in [6]. The authors in [7] present a data mining methodology to analyze relevant information and produce different perspectives about student to monitor their activities. This study applied different classification algorithms on students' previous and current academic record. Based on that, a model is proposed to find an enhanced evaluation method for predicting the placement for students.

In [8], authors propose a placement prediction system to predict the likelihood of a student understudy in getting a job in IT organizations. This system uses Artificial Intelligence model of K-closest neighbours' arrangement. This work also compares different models like logistic regression and Support Vector Machine (*SVM*). The scholarly history of the understudy just as their range of abilities like programming aptitudes, relational abilities, investigative abilities, and collaboration are considered in this work.

In [9], authors directed an examination to foresee understudy placement prediction status utilizing two characteristics: area of interest and Cumulative Grade Point Average (CGPA). They utilized decision tree learning ad *SCI*-kit inclining in *AI* with the two parameters. This system 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 prediction helps the institute to focus attention on students based on their interest.

The authors of [10] conducted study to predict student placement status using parameters like *UCN*, tenth and *PUC*/Diploma, *CGPA*, technical and aptitude skills The study presented in [11] reveals that the chances of campus placement in influenced by the four predictors - *CGPA*, specialization in *PG*, specialization in *UG* and gender.

[12] Uses machine learning algorithms in *WEKA* and *R-studio*. It proposes a method for mining the student's performance based on various parameters to predict and analyze whether a student (he/she) will be recruited or not in the campus placement. The authors in [13] propose a placement prediction model to predict the chance of an undergrad student getting a job in the placement drive. The work presented in [14] demonstrates using data mining techniques to predict useful information from students' raw data. The algorithms such as KNN, naïve bayes, and decision tree are used in this study.

On considering different existing systems for student placement prediction task, this paper proposes to test the applicability of 7 different machine learning models for this task adding more number of parameters for accurate prediction. This work also compares their performance by the prediction accuracy and Root Mean Square Error (RMSE) on two different data sets.

III. The Proposed Research work

This work aims at predicting possible salary package of a student based on his historical data. The prediction model which are used in this study are linear regression model, k-neighbor regression model, decision tree regression model, XGBoost regression model, gradient boost regression model, lightGBM regression model and random tree classifier model. A summary of each of these models is presented below.

- *Linear regression model* - to find the linear relationship between target and one or more predictors.
- *K-Neighbor regression model* - to solve both classification and regression problems.
- *Decision tree regression model* - to predict a target by learning decision rules from features
- *XGBoost regression model* – it is a decision-tree-based ensemble *ML* algorithm, it uses a gradient boosting framework.
- *Gradient boost regression model* – to produce a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
- *LightGBM regression model* – is a regression model, to determine the impact of each tree on the final outcome.
- *Random tree classifier model* – it creates a set of decision trees from randomly selected subset of training set.

A. Validation of the models

For validation of the chosen models, the R_2*_Score* is calculated. This score usually lies between 0 and 1. The model securing R_2*_Score* closest to 1 is regarded as the best model. The R_2*_Score* is calculated using the equation (1).

$$R2\_Score = 1 - \ (SSres/SStol) \qquad (1)$$

where *SSres* is the sum of residual or errors calculated using the equation (2) and *SStol* is the sum of average calculated using the equation (3).

$$SSres = \sum(Y_i - Y_i')^2 \qquad (2)$$

Where Y_i is the observed value and Y_i' is the predicted value (values predicted as best fit line)

$$SStot = \sum\left(Average(Y_i, Y_i')\right)^2 \qquad (3)$$

A diagram to visualize the regression line and the errors between the observed and predicted values are shown in figure 1. For understanding the context of R_2*_Score* calculation.

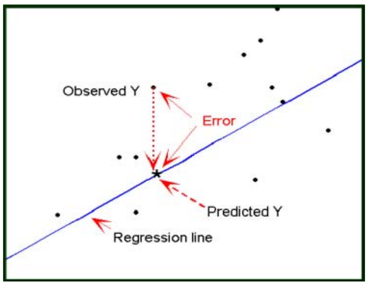

Fig. 1. Regression line passing through predicted values and shows residual or error.

Figure 2 shows the regression line passing through the average predicted outputs. Here the line passing through scattered points is the best fit line for average outputs, and the values present on the line in predicted values.

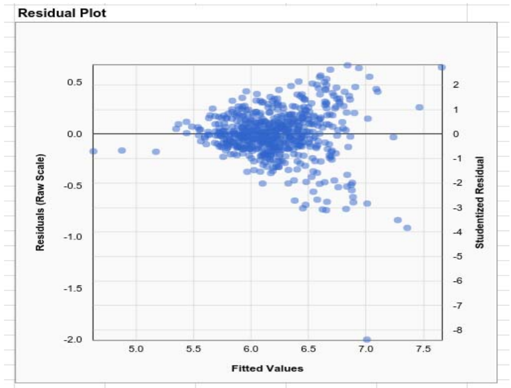

Fig. 2. Shows the regression line passing through average predicted outputs.

3.2 Data Preparation and Exploration

The datasets for this experiment are stored in *.csv* files. The dataset 1 contains the following features of each student - roll_number, Cumulative Grade Point Average (CGPA), total_arrears, cleared_arrears, internship_attended, paid_internship, projects_done, articles_published, and annual_package. Among these, the target variable is annual_package and the others are predictor variables (input). The roll_number is used as categorical variable and it is neglected as it does not involve in predicting the output.

B. Treating Missing Value

Since the input values are differing from student to student and there is no dependency among them the missing values are replaced by 0.

C. Data Visualization

To understand the significance of data, summarizing huge data and presenting them in simple and easy to understand format is the usual practise. This, in turn, helps for communicating the information clearly and effectively. In this paper, all variables are presented using the bus charts for easy understanding. It is shown in Figure 3.

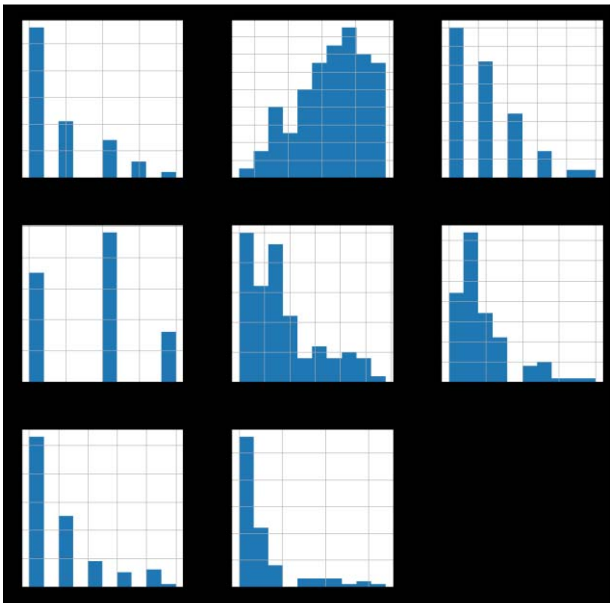

Fig. 3. Data visualization of the dataset.

IV. Results and Discussions

The Student placement prediction problem presented in this paper is to predict the annual_package for each student based on their values for the input features, using seven different machine learning models. The machine learning models considered in this study are

1. Linear regression model
2. K-neighbor regression model
3. Decision tree regression model
4. XGBoost regression model
5. Gradient boost regression model
6. LightGBM regression model
7. Random tree classifier model

A. Results obtained for Dataset-I

The models are trained using the variables in the dataset. The explanatory variables (independent/input variables) Cumulative Grade Point Average (CGPA), total_arrears, cleared_arrears, internship_attended, paid_internship, projects_done and articles_published are used for training this regression model to predict students' annual_pacakage. The observed values and predicted values along with the R_2_score and RMSE values of the linear regression model, K-neighbor regression model, decision tree regression model and XGBoost regression model are presented in Table 1.

The observed values and predicted values along with the R_2_Score and RMSE values for the gradient boost regression model, lightGBM regression model and random tree classifier model are presented in Table 2.

The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. The Table 1 shows all the supervised learning models and the Table 2 shows all the ensemble learning models. For dataset-I, both the supervised and ensemble learning models are implemented. The comparative results say that the K-neighbor regression model is the best model for prediction with the data set 1. The K-neighbor regression model has acquired the highest R_2_score of 0.94 and the lower RMSE of 103263.25.

TABLE 1. PERFORMANCE OF SUPERVISED LEARNING MODELS – DATASET-I

Sno	**Tested Values**	**Predicted Values**		
		Linear Regression Model	***K-Neighbor Regression Model***	***Decision Tree Regression Model***
1	280000	290038.62	189000	120000
2	200000	456306.47	353000	340000
3	400000	252499.26	442000	450000
4	160000	205284.22	260000	120000
5	1300000	1323431.24	11110000	1290000
	R_2_Score	**0.89**	**0.94**	**0.73**
	RMSE	**138160.38**	**103263.25**	**220703.42**

TABLE 2. PERFORMANCE OF ENSEMBLE LEARNING MODELS – DATASET-I

Sno	**Tested Values**	**Predicted Values**			
		Gradient Boost Regression Model	***LightGBM Regression Model***	***Random Tree Classifier Model***	***XGBoost Regression Model***
1	280000	1.63E+05	168501.00	170000	168332.70
2	200000	3.59E+05	322692.12	450000	384510.53
3	400000	3.89E+05	295764.83	400000	339698.94
4	160000	1.73E+05	150258.56	350000	172094.16
5	1300000	1.29E+06	1045256.65	1290000	1289736.20
	R_2_Score	**0.86**	**0.87**	**0.86**	**0.85**
	RMSE	**158984.24**	**153664.13**	**155595.63**	**162093.60**

B. Results obtained for Dataset-II

The dataset-I is extended adding another two new parameters: programming_language and area_of_interest. The parameter Programming_Language means the programming language in which the student is comfortable for programming. The parameter area_of_interest means the domain of interest of the student. The count of students opted for different programming languages is shown in the Figure 4. A scatter plot depicting the annual salary package and the area of interest is presented in Figure 5. The obtained results of all the models are presented in Table 3 and Table 4. They represent the performance of the supervised learning models and the ensemble learning model, respectively, for the dataset 2.

The experimental results obtained for the dataset-II reveals that the linear regression model and K-neighbor regression Model are the equally best fit models, as they achieve the highest the R_2_score of 0.89. However, the RMSE value of linear regression model is smaller than the K-neighbor regression model.

It is interesting to note that the, though the R_2_Score of XGBoost regression model is smaller than that of linear regression model and K-Neighbor Regression model, it has secured very less RMSE value. However, it has produced negative predicted value for a case in the R_2_score calculation.

It is also observed from the results that the, among the ensemble learning models, the gradient boost model and XgBoost model are not suitable for prediction with dataset-II as their R_2_score calculation involved –ve values.

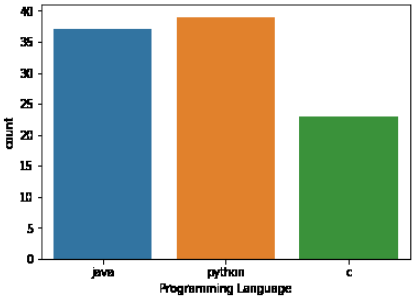

Fig. 4. Histogram of Programming Languages and Number of Students

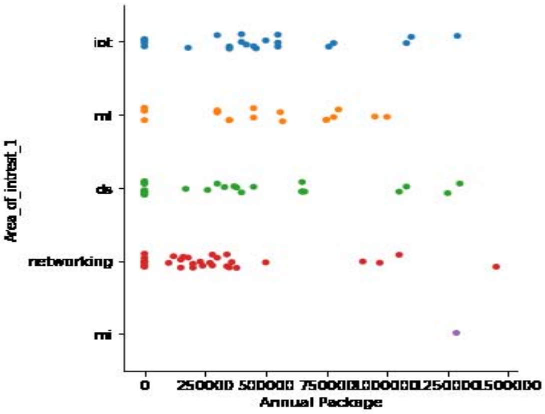

Fig. 5. Scatter plot of package and Area_of_Interest

TABLE 3. PERFORMANCE OF SUPERVISED LEARNING MODELS – DATASET-II

Sno	Tested Values	Predicted Values		
		Linear Regression Model	*K-Neighbor Regression Model*	*Decision Tree Regression Model*
1	280000	307389.70	433000	300000
2	230000	4644105	292000	450000
3	400000	413347.21	439000.56	420000
4	150000	9679.45	155000	0.00
5	1300000	1125441.97	1111000	1130000
	R_2_score	**0.89**	**0.89**	**0.75**
	RMSE	**111620.15**	**112905.04**	**169484.51**

TABLE 4. PERFORMANCE OF ENSEMBLE LEARNING MODELS – DATASET-II

Sno	Tested Values	Predicted Values			
		Gradient Boost Regression Model	*LightGBM Regression Model*	*Random Tree Classifier Model*	*XGBoost Regression Model*
1	280000	3.0061e+05	373621.44	300000	356526.20
2	200000	3.6169e+05	288690.20	170000	343827.44
3	400000	4.0406e+05	385851.34	350000	383409.7
4	150000	-1.9559e+02	29774.24	0	-5510.00
5	1300000	1.1062e+06	1046977.61	1080000	1216938.50
	R_2_score	**0.81**	**0.81**	**0.75**	**0.85**
	RMSE	**169484.51**	**148003.18**	**169631.95**	**12970.41**

V. CONCLUSIONS

A comparative study on implementing seven different machine learning models for student placement prediction problem is presented in this paper. The study is performed with two different data sets. For the dataset-I, the K-neighbor regression model is found to outperform other models with higher R_2_score (0.94) and lower RMSE value (103263.25). However for the dataset-II with increased number of features, the linear regression model and K-neighbor regression model stand first with good R_2_score (0.89). By the RMSE value of the XGBoost regression model is found to be very less for dataset-II (12970.41).

As this work is implemented as an initial attempt to understand the process of using machine learning models for student prediction problems, it is lacking in many aspects. The future work for enhancing this work includes adding more features to the data set and predicting few additional variables more relevant for student placement.

REFERENCES

[1] Ryosuke Ishizue, Kazunori Sakamoto, Hironori Washizaki and Yoshiaki Fukazawa, "Student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics", *Research and Practice in Technology Enhanced Learning*, Vol. 3, No. 7, 2018.

[2] Sreenath. K, Jeyakumar G, "Evolutionary Algorithm Based Rule(s) Generation for Personalized Courseware Construction in Educational Data Mining", *In proceeding of 2016 IEEE International Conference on Computational Intelligence and Computing Research*, 2016, pp. 609-615, 2016.

[3] Sucharitha.V, ReshmaReddy. R, Jeyakumar.G "Learning Interest Curve generation using PECS Agent based Model", *In Proceedings of International Conference on Communication and Computing*, pp. 152-160, 2014.

[4] Jeyakumar G, Raaghul K, Pavithra N, Kavya D and Aswathi B,"A Prototype for Student Learning Style Modelling Using Felder-Silverman Learning Style Model," *In Proceedings of International Conference on Smart Structures & Systems*, 2016, pp. 178 – 182.

[5] K.Roshini, B.Bavya, G.Jeyakumar, "RoBaJe – A Simulated Computational Model for Human Memory to Illustrate Encoding and Decoding of Information", *International Journal of Applied Engineering Research*, ISSN 0973-4562 Vol. 9, No. 24, pp. 26957-26970, 2015.

[6] Gowtham Deivanayagam .K, Gayathiri .D, Manikandan .A, Raghul Karthik K R, and Dr. G.Jeyakumar, "Learning to Identify Bad Coding Practice", *International Journal of Applied Engineering Research*, ISSN 0973-4562 Vol. 9, No. 20, pp. 6747-6755, 2014.

[7] Ajay Kumar Pal and Saurabh Pal, "Classification Model of Prediction for Placement of Student", *International Journal of Modern Education and Computer Science*, Vol 11. No. 49-57, 2013.

[8] Animesh Giri, M Vignesh V Bagavath, Bysani Pruthvi and Naina Dubey, "A Placement Prediction System using K-nearest neighbors' classifier", *In Proceedings of Second International Conference on Cognitive Computing and Information Processing,* 2016.

[9] Senthil Kumar Thangavel, Divya Bharathi P and Abhijith Shankar, “Student placement analyzer: A recommendation system using machine learning”, *In Proceedings of International Conference on advanced computing and communication systems*, 2017.

[10] Akash Prasad , Shreyas Harinath , Suma H S , Suraksha A , Tojo Mathew , “Student Placement Prediction Using Machine Learning”, *International Research Journal of Engineering and Technology(IRJET),* Vol. 06, No. 04, 2019.

[11] 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*, Vol. 8, No. 9, pp. 2278-3075, 2019.

[12] K.Sreenivasa Rao, N. Swapna an P.Praveen Kumar, “Educational data mining for student placement prediction using machine learning algorithm”, *International Journal of Engineering and Technology,* Vol. 7, No. 2, 2018.

[13] Agarwal, Krishnanshu, Ekansh Maheshwari, Chandrima Roy, Manjusha Pandey, and Siddharth Swarup Rautray, "Analyzing Student Performance in Engineering Placement Using Data Mining" *In Proceedings of International Conference on Computational Intelligence and Data Engineering*, pp. 171-181. Springer, Singapore, 2019.

[14] Mohammadi, Mehdi, Mursal Dawodi, Wada Tomohisa, and Nadira Ahmadi, "Comparative study of supervised learning algorithms for student performance prediction" *In 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 124-127. IEEE, 2019.