Comparative Analysis of Baseline Models, Ensemble Models and Deep Models for Prediction of Graduate Admission

Meet Vasani[1], Smit Patel[2] and Jasleen Kaur[3,*]

[1 – 2 – 3]School of Engineering, P P Savani University, Kosamba, Gujarat, India

*Corresponding Author: sidhurukku@yahoo.com

Abstract. As there is a rise in number of academic institutes all over the world, shortlisting the universities to apply has become a tedious task for aspiring graduate student. There are many crucial factors involved while submitting the application in any university. The cost involved while applying for any university is one of them. If student profile gets rejected, this may lead to wastage of time and money both. In order to handle this problem, we developed a machine learning based admission prediction system. This proposed prediction system considers various parameters such as English language score, university rank, Statement of purpose, letter of recommendation, Cumulative grade point average and research experience for predicting the chances of admission in a university using ensemble model. Total 13 machine learning based prediction models were trained and tested. These 13 models were divided into three categories: baseline models, ensemble models and deep models. Different performance parameters such as Root Mean Squared Error, Mean Squared Logarithmic Error and R2 were used. Execution time for each model was also recorded. In category wise comparison, ensemble models outperformed baseline models and deep models. While in total 13 models, multiple linear regression and ridge regression outperformed all other models (in terms of high R2 score, less root mean square error and less execution time) followed by gradient boosting regression and extra tree regression.

Keywords: Baseline models, deep models, ensemble models, graduate admission, prediction, regression and ridge regression.

1 Introduction

The world's job and business sectors are growing rapidly and persistently, in terms of skillset, and knowledge. Young individuals who want to succeed are always looking for higher education/degree to improve their skills and ability. In the last decade, there is an increase in number of individuals who are opting for higher education from esteemed foreign universities [1-3]. When it comes to taking decision about selection of universities, specific preparation is necessary. Various factors such as English language score, English language specific test, and consulting the experts for guidance are predominant factors in admission process. The Admission Committee of the respectful Universities accept/ reject a fellow's application on the basis of profile which comes under their criterion. If profile gets rejected, the entire resources becomes a complete waste, additionally the students will be demotivated and will have lost their crucial time to prepare for the worst case.

This fact motivated us to carry out this research work to predict the possibility of acceptance and rejection of application of student based on various factors which include GRE score, TOEFL score, undergraduate university ranking, SOP, LOR, CGPA and research experience using machine learning and deep learning. We proposed a deep learning-based methodology which helps students to predict their chances of admission in any esteemed university in the world.

Rest of paper is organized as follows. Section 2 presents related work carried out in proposed direction. Section 3 is presented by complete methodology followed for implementation of this work. Section 4 forecasts detailed results and analysis followed by Section 5 with concluding remarks.

2 Previous Work

This section provides a brief glimpse about machine learning application in various domains.

Mohan S. Acharya et al. presented a technique using various regression models such as Linear Regression, Decision Tree, Random Forest and Support Vector Regression [1]. Chithra Apporva D A et al. presents a machine learning based model using Linear Regression, K Nearest Neighbours, Random Forest and Ridge Regression and Linear Regression performs well [2]. Janani P. et al. uses Decision Tree algorithm-based technique to predict the possibility of admission in particular university on basis of input scores of students [3]. Sara Aljasmi et al. presents a technique to find chances of getting accepted in university using K Nearest Neighbours, Random Forest, Multi-layer Perceptron and Multiple Linear Regression [4]. Kanadpriya Basu et al. developed a system using data of liberal arts college in California which classifies whether a particular student will take admission in university or not [5].

A J Alvero et al. developed a college admission system which will be able to predict gender and household income of student from the narrative essay submitted for admission [6]. Sushruta Mishra et al. proposed a machine learning technique by which institute can enhance the quality of student's admission using K-Means clustering [7]. Simon Fong et al. applied a hybrid model of Neural Network and Decision Tree Classifier for predicting student's likelihood of getting admission in university [8]. Abdul Hamid M. Ragab et al. presented college admission system using two cascaded hybrid recommenders first recommender allocates study tracks for preliminary year students and second recommender allocates the specific school for students who breezed through the preliminary year tests effectively [9]. Shital Girase et al. developed a recommender system that comprehends data seeker's need and likewise produces recommendation through basic interface [10]. Sashank Sridhar et al. developed a system which could utilize information identified with past candidates of different colleges and their admit or reject status [11]. V. Raghavendran et al. presented a system based on investigating the dataset of various colleges for predicting the opportunity of being admitted in university using Multiple Linear Regression, Decision Tree Regression, Polynomial Regression and Random Forest Regression [12].

3 Methodology

Architecture of proposed methodology is depicted in Figure 1. Proposed system consists of two main phases: phase 1 and phase 2. The detail description of phase 1 and phase 2 is as follows:

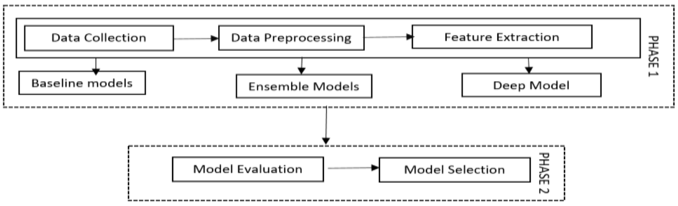

Fig. 1. Architecture of Admission prediction system

3.1 Phase 1:

It consists of following sub phases.

3.1.1 Data collection and understanding the Dataset

For this research work, we have utilized admission dataset from Kaggle [21].
It contains various attributes:
GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research and Chance of Admit.

Table 1. Dataset description

Total records	attributes	types of attributes
501	7	Simple Attributes
501	1	Derived Attribute

3.1.2 Data Pre – Processing

All the specified attributes were passed through various pre-processing phases. All the attributes were scaled. Scaling is a procedure which will level up all the values of the whole dataset, so that the model does not give dominance to a particular feature [22]. In this research work, scaling is carried out with the help of min-max scalar [22]:

3.1.3 Feature Extraction

For this research work, all the attributes were considered as features except the first column (i.e. Serial No.) because it does not add up as a feature. All features were divided into input and output features.

3.1.4 Model training

In total 13 models were trained and tested on this dataset. These models were divided into three categories: baseline models, ensemble models and deep models. The detail of each category is as follows:

1. Baseline Models: It includes multiple linear regression, Support vector Machine, Decision Tree regression, K-nearest Neighbhour, Lasso Regression, Ridge Regression, Elastic Net Regression [13-16].
2. Ensemble Learning Models: It includes Random Forest Regression, Gradient Boosting Regression, Ada Boost Regression and Extra Tree Regression [17-18].
3. Deep Learning Models: It includes Neural Network and Multi-layer perceptron [19-20].

3.2 Phase 2:

Performance evaluation of all the prediction model is carried out using RMSE, MSLE and R2 score. Final model selection is carried out using akaike information criteria (AIC), time and performance parameters.

4 Results

This section provides the result and analysis of the entire experimentation. Table 2 presents the results of baseline models, ensemble models and deep learning models respectively.

Table 2. Performance evaluation of baseline models, ensemble models and deep learning models

	Model	**RMSE Value**	**MSLE Value**	**R2 Score**	**Time (in milisec.)**
Baseline Learning Models	Multiple Linear Regression	0.003634	0.001382	0.825631	0.004246
	Ridge Regression	0.003642	0.001389	0.82524	0.004984
	K Neighbours Regression	0.004315	0.001655	0.792956	0.005667
	SVM	0.005061	0.001853	0.75718	0.008969
	Decision Tree Regression	0.007739	0.002977	0.628668	0.006028
	Lasso Regression	0.020903	0.00738	-0.00293	0.004321
	Elastic Net Regression	0.020903	0.00738	-0.00293	0.005145
	Average	0.003634	0.001382	0.825631	0.004246
Ensemble Learning Models	Gradient Boosting	0.004107	0.001574	0.802956	0.055536
	Extra Trees Regression	0.004161	0.001587	0.800365	0.14483
	Random Forest	0.004193	0.001613	0.798815	0.030146
	Ada Boosting	0.005082	0.001892	0.756177	0.08085
	Average	0.004386	0.001666	0.789578	0.077841
Deep Learning Models	Neural Network	0.004328	0.001649	0.792346	1.9771
	Multi-Layer Perceptron Regressor	0.0084	0.003216	0.59697	0.090733
	Average	0.006364	0.002433	0.694658	1.033917

From table 2, it can be observed that, with RMSE value of 0.003634 and R^2 value of 0.825631, multiple linear regression performed better as compared to other baseline algorithms. Time taken to predict the result was reported to be 0.004246 ms for multiple linear regression algorithm. For multiple linear regression, independent variables are

GRE score, TOEFL score, University ranking, Statement of purpose, letter of recommendation, Cumulative grade point average and research experience whereas dependent variable is chance of admit. Average RMSE value for baseline models was reported to be 0.009457 and average R^2 value was reported to be 0.546258.

In second experimentation, different ensemble models were trained and tested for prediction task. From table 2, it can be observed that with lowest RMSE value and highest R^2 value, gradient boosting turns out to be the best among ensemble learning methods. RMSE value and R^2 value for gradient boosting was reported to be 0.004107 and 0.802956, respectively. Time taken to perform prediction using gradient boosting was 0.055536 ms. Parameters values taken into consideration for gradient boosting is number of random states with value 42. Average RMSE, MSLE, and R^2 value for ensemble learning models was 0.004386, 0.001666 and 0.789578, respectively.

In third experimentation, two different deep models were trained and tested. From table 2, it can be observed that with lowest RMSE value and highest R^2 value, neural network turns out to be the best among deep methods. RMSE value and R^2 value for neural network was reported to be 0.004328 and 0.792346, respectively. Time taken to perform prediction using gradient boosting was 1.9771 ms. Parameters values of neural network are number of units (neurons) is 11, activation is relu, input_dim is 7 and epochs is 50. Average RMSE, MSLE, and R^2 value for deep learning models was 0.006364, 0.002433 and 0.694658, respectively.

For comparative analysis of models based on time parameter, different performance parameters were plotted against time as depicted in figure 2, and 3. With high R^2 score and less execution time, Multiple Linear Regression and Ridge Regression performed best followed by Gradient Boosting Regression and Extra Tree Regression. With less RMSE value and R^2 value, multiple linear regression performed better followed by Gradient Boosting Regression and Extra Tree Regression

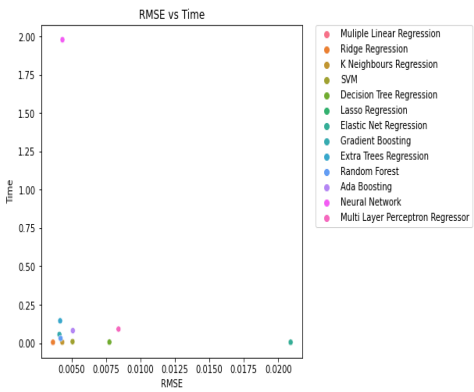

Fig. 2. Algorithms comparison with RMSE values and Execution Time.

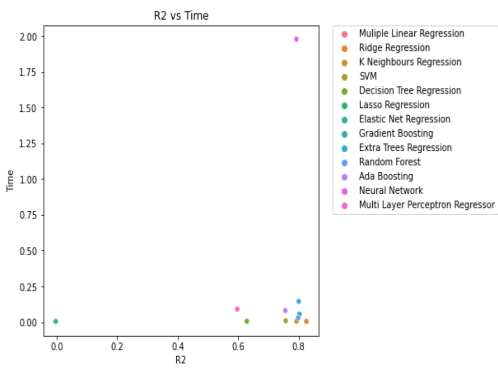

Fig. 3. Algorithms comparison with R2 scores and Execution Time.

4.1 Comparative Analysis of Baseline models, Ensemble Models and Deep Models

For comparative analysis of these three models, average values of different performance parameters was considered. Table 3 depicted the results obtained in three different model categories.

Table 3. Comparative analysis of three model categories

Model Category	RMSE Value	MSLE Value	R2 Score	Time (in millisec.)
Baseline models	0.009457	0.003431	0.546258	0.005623
Ensemble models	0.004386	0.001666	0.789578	0.077841
Deep Models	0.006364	0.002433	0.694658	1.033917

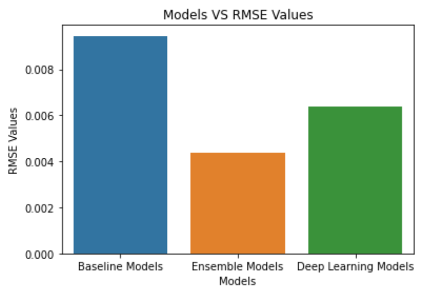

Fig. 4. Models comparison with RMSE values.

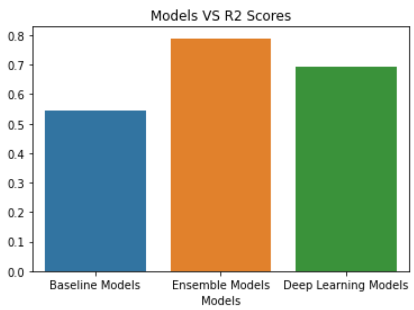

Fig. 6. Models comparison with R2 Scores.

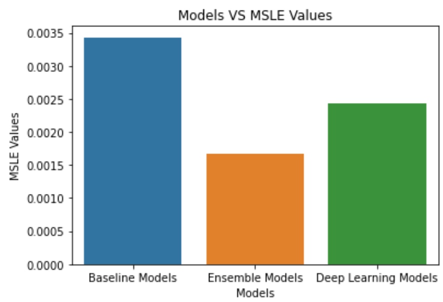

Fig. 5. Models comparison with MSLE values.

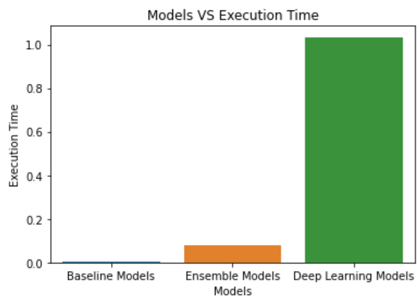

Fig. 7. Models comparison with Execution time.

From figure 4 and figure 5, it can be observed that while comparing models on basis of categories (Baseline Models, Ensemble Models and Deep Models), ensemble models performed best with less average Root Mean Squared Error and Mean Squared Logarithmic Error values compared to other two. From figure 6, it can be observed that, with high average R^2 score, ensemble models outperformed baseline and deep models. As depicted in figure 7, on basis of execution time, baseline models reported the lowest value.

5 Conclusion

In order to speed up the admission process and save the cost involved in admission process, we proposed a machine learning based approach for predict the chances of admission. This proposed system takes into consideration various parameters for prediction task. Total 13 models were trained and tested on admission dataset. For comparative analysis purpose, these models were divided into three categories: baseline, ensemble and deep models. Root Mean Squared Error, Mean Squared Logarithmic Error, R^2 and time parameters were used for evaluation of model. Multiple linear regression, gradient boosting and neural network outperformed all other models in each category. From all 13 models, multiple linear regression and ridge regression outperformed all other models, followed by gradient boosting and extra tree regression. RMSE value and R^2 value was reported to 0.003634 and 0.825631, respectively for multiple linear regression. From three model categories, ensemble models performed better as compared to baseline and deep models with lowest RMSE score and highest R^2 value.

References

1. M. S. Acharya, A. Armaan and A. S. Antony; "A Comparison of Regression Models for Prediction of Graduate Admissions," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS) (2019).
2. Chithra Apoorva D A, Malepati ChanduNath, Peta Rohith, Bindu Shree.S, Swaroop.S; “Prediction for University Admission using Machine Learning”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue 6 (2020).
3. Janani P, Hema Priya V, Monisha Priya S; “Prediction of MS Graduate Admissions using Decision Tree Algorithm”, International Journal of Science and Research (IJSR) ISSN: 2319-7064, Volume 9, Issue 3 (2020).
4. Sara Aljasmi, Ali Bou Nassif, Ismail Shahin, Ashraf Elnagar; “Graduate Admission Prediction Using Machine Learning”, International Journal of Computers and Communications ISSN: 2074-1294, Volume 14 (2020).
5. Basu K, Basu T, Buckmire R, Lal N.; “Predictive Models of Student College Commitment Decisions Using Machine Learning” (2019).
6. A.J. Alvero, Noah Arthurs, anthony lising antonio, Benjamin W. Domingue, Ben Gebre-Medhin, Sonia Giebel, Mitchell L. Stevens; “AI and Holistic Review: Informing Human Reading in College Admissions”, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 200–206 (2020).

7. Sushruta Mishra, Soumya Sahoo, Brojo Kishore Mishra, Santosh Satapathy; “A Quality Based Automated Admission System for Educational Domain”, 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 221-223 (2016).
8. Simon Fong, Robert P. Biuk-Aghai; “An Automated University Admission Recommender System for Secondary School Students”, The 6th International Conference on Information Technology and Applications (2009).
9. Abdul Hamid M. Ragab, Abdul Fatah S. Mashat, Ahmed M. Khedra; “Design and Implementation of a Hybrid Recommender System for Predicting College Admission”, International Journal of Computer Information Systems and Industrial Management Applications, ISSN 2150-7988 Volume 6, pp. 35 – 44 (2014).
10. S. Girase, V. Powar and D. Mukhopadhyay; “A user-friendly college recommending system using user-profiling and matrix factorization technique”, 2017 International Conference on Computing Communication and Automation (ICCCA), pp. 1-5 (2017).
11. S. Sridhar, S. Mootha and S. Kolagati; “A University Admission Prediction System using Stacked Ensemble Learning” 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 162-167 (2020).
12. Raghavendran C.V., Pavan Venkata Vamsi C., Veerraju T., Veluri R.K.; “Predicting Student Admissions Rate into University Using Machine Learning Models”, Machine Intelligence and Soft Computing. Advances in Intelligent Systems and Computing, vol 1280. Springer, Singapore (2021).
13. G. Naga Satish, C.V. Raghavendran, M.D. Sugnana Rao, C. Srinivasulu; “House price prediction using machine learning” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9 (2019).
14. M. Frizzarin, I.C. Gormley, D.P. Berry, T.B. Murphy, A. Casa, A. Lynch, S. McParland; “Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods” Journal of Dairy Science, ISSN 0022-0302 (2021).
15. Husain H. Dawoodi, Manoj P. Patil; “Rainfall Prediction in North Maharashtra Region Using Support Vector Machine”, Turkish Journal of Computer and Mathematics Education Vol.12 No. 7, 1501-1505 (2021).
16. Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, Norafefah Mohd Sobri; “Supervised data mining approach for predicting student performance”, Indonesian Journal of Electrical Engineering and Computer Science, ISSN: 2502-4752, Vol. 16, No. 3, pp. 1584-1592 (2019).
17. Deokar, A.; “Using Visceral Adipose Tissue Measurements to build Classification Models for Gestational Diabetes Mellitus” Paper presented at 2020 Northeast Section Meeting, Online. 10.18260/1-2-020-36266 (2021).
18. Lee, S., Chung, J.Y.; “The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction” Applied Science, 9, 3093 (2019).
19. Shwetank Mishra, Satya Vijay Neurkar, Riddhesh Patil, Sharmila Petkar; “Heart Disease Prediction System”, International Journal of Engineering and Applied Physics (IJEAP) ISSN: 2737-8071, Vol. 1, No. 2 (2021).
20. Abu Zohair L.M.; “Prediction of Student’s performance by modelling small dataset size” International Journal of Educational Technology in Higher Education, 16, 27, 2019
21. https://www.kaggle.com/mohansacharya/graduate-admissions?select=Admission_Predict_Ver1.1.csv
22. Peshawa Jamal Muhammad Ali, Rezhna Hassan Faraj; "Data Normalization and Standardization : A Technical Report" Machine Learning Reports, 1(1) (2014).