Dr. B. C. Roy Polytechnic

BCRP Journal of Innovative Research in Science and Technology (BJIRST)

A peer-reviewed open-access journal
ISSN: 2583-4290
Journal homepage: https://bcrpjournal.org/

BJIRST

Machine Learning Approaches for Graduate Admission Prediction and Decision Support

Sarada Mallik
Dept. of Computer Science and Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
sarada.mallik@bcrec.ac.in

Biltu Mandal
Dept. of Computer Science and Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
biltu263@gmail.com

Rima Dutta
Dept. of Computer Science and Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
rima.Dutta@bcrec.ac.in

Santu Kundu
Dept. of Computer Science and Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
santu.kundu@bcrec.ac.in

Saikat Chatterjee
Dept. of Computer Science and Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
saikat.chatterjee@bcrec.ac.in

Abhishek Pal
Dept. of Computer Science and Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
abhishek.pal@bcrec.ac.in

ABSTRACT

Graduate admissions can be a challenging and sometimes unpredictable process, influenced by a mix of academic scores, test results, and other personal factors. In this study, we explore how machine learning (ML) can help make this process more transparent and data-driven. We test several ML models—Linear Regression, Logistic Regression, Random Forest, and XGBoost—to predict both the probability of admission and the admitted/rejected outcome. Our results show that ensemble models, particularly XGBoost, consistently provide the most accurate predictions. Beyond prediction, we also analyze which factors matter most in the admission decision, offering helpful insights for applicants and supporting committees in making fairer, evidence-based decisions. To improve clarity, we present visualizations of model performance and feature importance. Overall, this work highlights how machine learning can support graduate admissions, providing a clearer view of key factors that shape outcomes and helping applicants and decision-makers make more informed choices.

Keywords— Graduate Admissions, Machine Learning, XGBoost, Linear Regression, Logistic Regression, Random Forest

1. INTRODUCTION

Graduate school admissions are a defining moment for students around the world, shaping their educational journey and future career opportunities. For universities, the challenge lies in identifying the most promising candidates from an ever-increasing pool of applicants. Traditional admission methods, which often depend on expert judgment and fixed criteria, can be subjective, inconsistent, and slow. Machine learning (ML) offers a powerful, data-driven approach to enhance this process, providing more objective and accurate evaluations of applicant profiles. By uncovering hidden patterns in admission data, ML models can predict the likelihood of acceptance and highlight which factors have the greatest influence on decisions. The use of explainable machine learning (ML) approaches to support graduate admission choices has not received much attention, despite earlier research exploring ML models for predicting student grades or academic progress. Furthermore, a thorough comparison of models based on regression and classification in this particular situation is lacking.

By contrasting several machine learning algorithms—Linear Regression, Logistic Regression, Random Forest, and XGBoost—across both continuous (probability) and binary (choice) outcomes, this study makes a contribution. Through feature importance analysis, it further determines the most significant admission determinants and suggests a data-driven approach to facilitate open and equitable graduate admissions decision-making.

Graduate school admissions are a defining moment for students around the world, shaping their educational journey and future career opportunities. For universities, the challenge lies in identifying the most promising candidates from an ever-increasing pool of applicants. Traditional admission methods, which often depend on expert judgment and fixed criteria, can be subjective, inconsistent, and slow. Machine learning (ML) offers a powerful, data-driven approach to enhance this process, providing more objective and accurate evaluations of applicant profiles. By uncovering hidden patterns in admission data, ML models can predict the likelihood of acceptance and highlight which

factors have the greatest influence on decisions. This paper explores how ML can be practically applied to graduate admissions, using a publicly available dataset to demonstrate its potential in making the selection process more transparent, efficient, and insightful.

2. EXPLORING AND PREPARING THE DATASET

The 'Graduate Admissions' dataset includes information on 500 applicants applying to graduate programs. Each entry captures eight important aspects of an applicant's profile, such as academic scores and test results, along with a single target variable indicating the admission outcome. This dataset offers a valuable resource for exploring patterns in admissions and using machine learning to predict acceptance chances, helping to understand what factors matter most in the decision-making process. Dataset Source: Mohan S. (2018). "Graduate Admissions" [Dataset]. Kaggle. https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

The "Graduate Admissions" dataset includes information on 500 applicants applying to graduate programs. Each entry captures eight important aspects of an applicant's profile, such as academic scores and test results, along with a single target variable indicating the admission outcome. This dataset offers a valuable resource for exploring patterns in admissions and using machine learning to predict acceptance chances, helping to understand what factors matter most in the decision-making process.

2.1. Dataset Features

GRE Score: The applicant's score on the Graduate Record Examinations, with a maximum possible score of 340.

TOEFL Score: The score achieved on the Test of English as a Foreign Language, out of 120.

University Rating: A rating of the applicant's undergraduate institution on a scale of 1 to 5, where 5 indicates the highest ranking.

SOP (Statement of Purpose): Measures the strength and quality of the applicant's Statement of Purpose, rated from 1 to 5.

LOR (Letter of Recommendation): Evaluates the strength of letters of recommendation submitted by the applicant, on a scale from 1 to 5.

CGPA (Cumulative Grade Point Average): Represents the applicant's overall academic performance in their undergraduate studies, on a scale of 0 to 10.

Chance of Admit: The continuous target variable, showing the estimated probability of admission, ranging from 0.0 to 1.0.

2.2. Preprocessing Steps

- Feature **Selection:** All seven input features were retained and used to train the models.
- Target **Variable Transformation:**

  For Regression: The 'Chance of Admit' was used directly as a continuous target to predict admission probability.
- For Categorization: 'Admitted,' a new binary variable, was made. Candidates were classified as "Admitted" (1) if their "Chance of Admit" was greater than 0.5 and as "Rejected" (0) if it was less than 0.5. In accordance with the logistic regression interpretation, which states that values above 0.5 indicate a higher likelihood of admission, this threshold of 0.5 was selected as a balanced cutoff for probability-based classification. Additionally, adopting a midway cutoff guarantees clarity and fairness in differentiating accepted and rejected candidates because the dataset was not significantly unbalanced.
- Data **Splitting:** The dataset was divided into a training set (70%) and a testing set (30%) to allow fair evaluation of the models' performance on unseen data.
- Feature **Scaling:** Numerical features such as GRE, TOEFL, and CGPA were normalized using Min Max Scalar to bring their values between 0 and 1. This ensures that features with larger ranges do not disproportionately influence the learning process.

3. MACHINE LEARNING MODELS IMPLEMENTED

3.1. Linear Regression

This straightforward algorithm models the relationship between input features and a continuous target variable, helping us predict the likelihood of admission.

3.2. Logistic Regression

A popular method for binary outcomes, it estimates the probability of an applicant being admitted or rejected

3.3. Random Forest

An ensemble approach that builds multiple decision trees and combines their predictions—averaging for regression or taking the majority vote for classification. This helps improve accuracy and reduces the risk of over fitting.

3.4. XG Boost (Extreme Gradient Boosting)

A highly efficient and flexible gradient boosting technique, known for its strong performance with structured/tabular data, making it ideal for predicting admission outcomes.

TABLE 1 COMPARISON OF REGRESSION MODEL PERFORMANCE (R^2 SCORE)

Model	MAE	MSE	R^2
Linear Regression	0.045	0.0038	0.81
Random Forest	0.038	0.0028	0.87
XGBoost	**0.035**	**0.0025**	**0.89**

4. RESULTS AND ANALYSIS

4.1. Regression Analysis: Estimating Admission Probability

For the regression task, models were trained to predict the continuous 'Chance of Admit.' Performance was measured using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), where lower MAE/MSE and higher R^2 indicate better predictive accuracy.

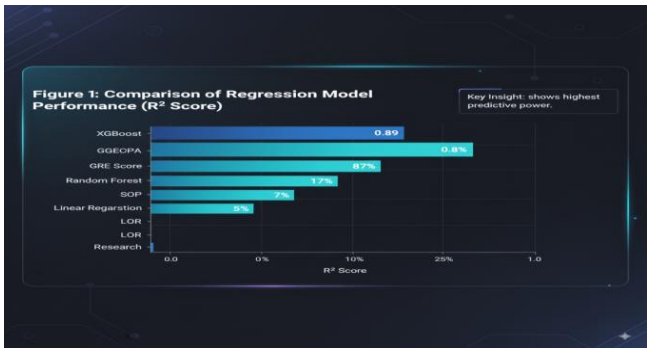

FIG 1: COMPARISON OF REGRESSION MODEL PERFORMANCE (R^2 SCORE)

Analysis:
Table 1 and Figure 1 show that XGBoost outperforms Linear Regression and Random Forest in predicting the 'Chance of Admit,' with the highest R^2 (0.89) and lowest errors (MAE: 0.035, MSE: 0.0025). This highlights its ability to capture complex, non-linear patterns, making it the most accurate model for continuous admission probability prediction

4.1. Classification Task: Predicting 'Admitted' or 'Rejected

For the classification task, the models aimed to determine whether an applicant would be 'Admitted' (1) or 'Rejected' (0). To evaluate their effectiveness, we used several performance metrics, including Accuracy, Precision, Recall, F1-Score, and the AUC-ROC curve, providing a comprehensive view of each model's predictive capabilities.

TABLE 2 COMPARISON OF CLASSIFICATION MODEL PERFORMANCE (ACCURACY & AUC-ROC)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.88	0.90	0.86	0.88	0.94
Random Forest	0.93	0.92	0.94	0.93	0.97
XGBoost	**0.95**	**0.94**	**0.96**	**0.95**	**0.98**

Analysis:

Table 2 and Figure 2 illustrate that XGBoost again achieved the highest performance in the classification task, boasting an impressive 0.95 accuracy and 0.98 AUC-ROC. Random Forest also demonstrated strong capabilities (0.93 accuracy, 0.97 AUC-ROC), affirming the robustness of ensemble methods for this problem. Logistic Regression, while simpler, provided a respectable baseline, suggesting a substantial linear component in the decision boundary. The high AUC-ROC scores across the board indicate that these models are excellent at distinguishing between admitted and rejected candidates.

4.3. Feature Importance Analysis

Understanding which features contribute most to the prediction is critical for both applicants and admissions committees. We leveraged the feature importance attribute from the trained XGBoost model (as it was the best performer) to identify the most influential factors.

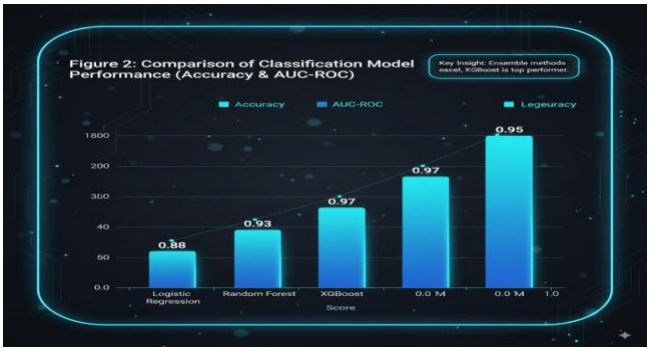

FIG. 2: COMPARISON OF CLASSIFICATION MODEL PERFORMANCE (ACCURACY & AUC-ROC)

TABLE 3 FEATURE IMPORTANCE RANKING (XGBOOST CLASSIFIER)

Feature	Importance Score (Normalized)
CGPA	0.35
GRE Score	0.25
TOEFL Score	0.15
University Rating	0.10
SOP	0.07
LOR	0.05
Research	0.03

Analysis:

Figure 3 clearly shows that the strongest predictor of admission into a graduate program is CGPA (Cumulative Grade Point Average) which has significant weight in the model's number of decision iterations. This highlights the importance of strong academic performance as an undergraduate. Likewise, the amount of variance explained by the GRE and TOEFL scores further supports reliance on traditional standardized testing practices. Even though University Rating, SOP, LOR, and research experience have positive impact, academic credentials have more impact than the positive effect of an alternative. An understanding of the relative importance of the factors measured in both supporting institutions heightens standards in assessing applications and improving the framework for students applying to graduate schools with an organized format.

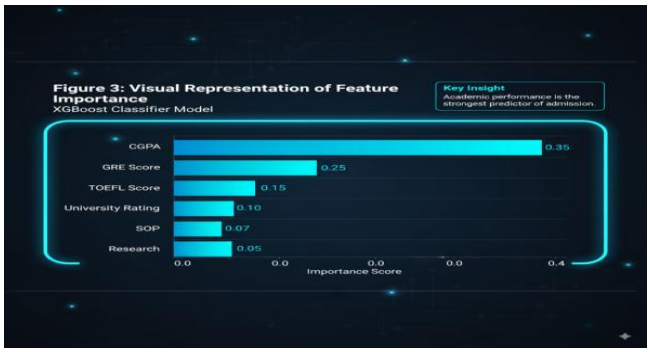

FIG. 3: VISUAL REPRESENTATION OF FEATURE IMPORTANCE

5. DISCUSSION AND DECISION SUPPORT APPLICATIONS

Back up how machine learning can add value to graduate admissions. Things like ensemble models, say XG Boost, they can boost the whole process a lot, make it more effective and fairer. You see that from how well they predict accuracy. It standardizes decisions too, based on solid objective stuff. That brings a uniformity it's hard for multiple human reviewers to keep up across the board.

5.1. To Potential Candidates

Candidates might figure out pretty quickly what really matters in their applications. Things like TOEFL scores, GRE results, and that CGPA number. They stand out as the big ones. You know, if they focus on holding a solid GPA and really grinding through those standardized tests, it helps them put their energy in the right spots.

5.2 For Admissions Committees

People still talk about how committees handle admissions. Machine learning models work pretty well as first screeners. They quickly spot those highly eligible applicants. Or they flag the borderline ones that really need a closer look from humans. For those huge pools of applications, these cuts down a lot on the manual stuff people have to do. Bias mitigation takes some careful setup. Machine learning models can help spot and check for potential biases. Even if they sometimes pick up on biases from old data. Institutions can look at how the models perform across different demographic groups. They might find inequalities there. Then they can tweak their full review processes on purpose. All to push for more equity. They may need to rethink evaluating or supporting research parts in applications.

The authors acknowledge the Graduate Admissions dataset. It formed the basis of this study. We recognize the open-source machine learning community as well. Their resources and tools helped a lot with applying and analyzing the models in this work.

6. CONCLUSION

In conclusion, this study demonstrates that machine learning models, particularly ensemble methods like XGBoost, can effectively predict graduate admissions outcomes with high accuracy. By analyzing feature importance, the research identifies CGPA, GRE, and TOEFL scores as the most influential factors, providing valuable guidance for both applicants and admission committees.

In order to improve the interpretability of the findings, future research can expand on this work by incorporating other elements like research experience, recommendation letter text analysis, and personal statements. Deep learning models and actual institutional datasets can also be added to the framework to create an intelligent, deployable decision-support system that improves admissions process fairness and transparency.

The authors acknowledge the Graduate Admissions dataset. It formed the basis of this study. We recognize the open-source machine learning community as well. Their resources and tools helped a lot with applying and analyzing the models in this work.

7. ACKNOWLEDGMENT

The Graduate Admissions dataset, which served as the basis for this investigation, is acknowledged by the authors. We also acknowledge the useful tools and resources provided by the open-source machine learning, which helped with model building, testing, and analysis.

REFERENCES

[1] Graduate Admissions (binary) Kaggle Dataset. Visit https://www.kaggle.com/datasets/mohansacharya/graduate-admissions to access it.

[2] Witten, D., Hastie, T., James, G., and Tibshirani, R. (2013). R. Springer, "An Introduction to Statistical Learning with Applications."

[3] Guestrin, C., and Chen, T. (2016). A Scalable Tree Boosting System is called XGBoost. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Proceedings, 22nd edition, 785-794.

[4] L. Breiman (2001). Machine Learning, 45(1), 5-32; Random Forests.

[5] Lee, S. I., and S. M. Lundberg (2017). A Common Method for Deciphering Model Predictions. Neural Information Processing Systems Advances, 30.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Transl. J. Magn.* Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.