Nepal Journal of Mathematical Sciences (NJMS)
ISSN: 2738-9928 (*online*), 2738-9812 (*print*)
Vol. 4, No. 2, 2023 (August): 7-16
DOI: https://doi.org/10.3126/njmathsci.v4i2.59524
©School of Mathematical Sciences,
Tribhuvan University, Kathmandu, Nepal

Research Article
Received Date: July 3, 2023
Accepted Date: August 11, 2023
Published Date: August 30, 2023

High School Performance Based Engineering Intake Analysis and Prediction Using Logistic Regression and Recurrent Neural Network

Govinda Pandey[1],Nanda Bikram Adhikari[2], & Subarna Shakya[3]

[1,2,3]*Dept. of Electronics and Computer Engineering, Pulchowk Engineering Campus, Tribhuvan University, Nepal*

Email:[1]073*mscs*655.*govinda@pcampus.edu.np*,[2]*adhikari@ioe.edu.np*, [3]*drss@ioe.edu.np*

Abstract: *A student's high school performance is crucial for engineering admission in Nepal. Machine learning-based predictive models can provide valuable insights. This study aims to predict engineering entrance exam scores and admission probability based on high school academic records. In this study, we have used exam data from National Examination Board (NEB) and Institute of Engineering (IOE) containing grades, scores and results for over* 11,000 *students. Logistic Regression (LR) and Long-Short Term Memory (LSTM) models are implemented to predict pass/fail status and year-wise entrance score forecasting, respectively. In addition, the Prophet model analyzed trends in entrance score threshold averaging. The result shows that the logistic model achieved* 97% *accuracy in predicting pass/fail status and the LSTM network attained reasonable accuracy between* 65-85% *for score forecasting. The Prophet model accurately projected decreasing trends in threshold scores and admitted students' averages. Our model analyses provides actionable insights into student outcomes, complex patterns, and changing trends. Proactive interventions through upgraded curriculum, teacher training etc. could reverse declining enrolment.*

Keywords: *Education data mining, Intake prediction, Logistic regression, Long Short-Term Memory (LSTM), Student performance*

1. Introduction

In the realm of education, understanding and enhancing student performance is pivotal for fostering national development and stimulating economic growth. Educational institutions amass extensive datasets encompassing student activities, routines, backgrounds, and academic histories. However, a considerable portion of this data remains underutilized, primarily due to its sheer volume and complexity, as well as the institutions' capacity constraints in processing it. To harness the potential of this data for predictive and prescriptive purposes, the integration of advanced information technologies, notably data mining and machine learning, is imperative. The data on engineering entrance exam applicants and results over the past 5 years reveals concerning declines in both student interest and performance. Specifically, the number of applicants has steadily dropped from over 12,000 in 2017 to just 9,404 in 2022, indicative of reducing the popularity of engineering programs. However, even among those appearing for the exams, competitiveness and preparedness have worsened. The entrance score threshold has fallen from 52 down to 38 and average scores have declined from nearly 70 to around 62. This consistent downward trend in cut off marks and admitted student performance highlights deficiencies in pre-engineering preparation at the high school level. Students seem less academically equipped to handle the rigor of the entrance exams compared to previous years. The reasons could include deteriorating quality of schooling, increased opportunities abroad, or other competing fields attracting talent. Nonetheless, the data signals the need for interventions to boost interest in engineering and bolster high school teaching and resources. Addressing these gaps proactively through counselling, upgraded curriculum, teacher training, etc. can help reverse the concerning enrolment patterns. More effective high school preparation will translate into increased applicants and better performance.

This paper aims to bridge the existing knowledge gap between high school and engineering education by developing predictive models based on high school academic records. The scope includes students applying for engineering programs at the Institute of Engineering under Tribhuvan University. The objectives are to predict entrance exam scores and admission probability using machine learning techniques. The research questions are: 1) How accurately can high school performance predict engineering entrance outcomes? 2) What are the capabilities of different machine learning models for this predictive task? Machine learning techniques like logistic regression and LSTM networks are applied to high school and entrance datasets to uncover patterns and trends that can enable data-driven decision-making around admissions.

Data mining, also known as knowledge discovery in databases (KDD), employs a multitude of techniques and algorithms to extract valuable insights from vast datasets. When applied to the educational domain, termed as "educational data mining," these techniques can unveil patterns and correlations previously unseen. Algorithms such as decision trees, neural networks, linear regression, and random forests are particularly adept at predicting outcomes based on historical data, enabling educational stakeholders to anticipate student performance trends and act proactively.

For instance, a student's performance in high school, analysed holistically across various parameters-grade-wise and subject-wise results, demographics, school type, and more-can serve as a predictive indicator of their potential success in higher education. Especially in contexts where high school graduates aspire for competitive admissions in tertiary institutions, such predictive models can be invaluable. As a case in point, for admissions to engineering programs under Tribhuvan University, students are assessed through a rigorous computer-based entrance examination by the Institute of Engineering, which evaluates proficiency in subjects like Mathematics, Physics, Chemistry, and English, all grounded in the high school curriculum. Thus, a student's high school academic record becomes a significant predictor of their entrance score and subsequent success in the program.

This paper aims to bridge the existing knowledge gap between high school and engineering education by developing a predictive model based on high school academic records. Such a model can assist in identifying students at risk, guiding admission decisions, and formulating strategies to ensure every student's optimal academic progression.

2. Background Study

Predicting student entrance scores based on prior academic performance utilizes information extraction. Analysing student data, including exam scores, enables institutions to develop predictive models for identifying students needing extra support. Educational data mining explores predictive models for academic performance using machine learning techniques (Chen et al.,[2]). Educational institutions are amassing extensive datasets encompassing student activities, attendance patterns, geographical locations, family backgrounds, and more. Nevertheless, this wealth of data typically gets harnessed for generating basic queries and conventional reports that seldom reach the appropriate individuals in a timely manner to enable informed decision-making (Kabakchieva, [8]). Dien et al. study deep learning methods for student performance prediction, considering data preprocessing strategies (Dien et al., [4]). In Nepal, research explores hyper-parameter tuning for student grade prediction using neural networks (Rimal et al., [14]). GPA prediction employs Boruta algorithm and random forest with single and multiple-layer models. Artificial neural networks forecast student performance. Educational data mining evaluates classification algorithms for student success prediction(Gochhait & Rimal, [7]; Meghji et al., [10]; Naser et al., [11]).

Machine learning is crucial for education, enhancing retention, performance prediction, and curriculum design. Waheed et al. demonstrate deep neural networks outperform logistic regression and support vector machines. A hybrid 2D CNN model predicts academic achievement. A deep neural network predicts student performance effectively. High school GPA predicts college outcomes and future income. Prophet forecasting aids resource allocation using enrolment data (Bendangnuksung & Prabu, [1]; Enughwure & Ogbise, [5]; Marte, [9]; Patayon & Crisostomo, [12]; Poudyal et al., [13]; Waheed et al., [15]). Prophet methodology is a time series forecasting approach that uses a decomposable model with three main components: trend, seasonality, and holidays (Daraghmeh et al., [3]).To our knowledge, such modelling is not found in literature in the mentioned scope, thus this research attempts the same using the following methodology.

3. Methodology

The study methodology involved collecting student academic records from the National Examination Board (NEB) and Institute of Engineering (IOE) entrance exams. The NEB data contained high school (Grade 10 and 12) grades and GPAs, while the IOE data had entrance registration details and exam scores. After joining the datasets, feature extraction and selection was done to identify the most relevant input variables like PCL subject scores.

Data pre-processing steps included handling missing values, removing outliers, encoding categorical variables, and pivoting to summarize subject marks. The final dataset contained features such as academic year, gender, NEB grades, IOE entrance scores and results for over 11000 students. Exploratory analysis using summary statistics and visualizations provided insights into score distributions.The processed data was split 70:30 into train and test sets. Three models were developed - logistic regression to predict pass/fail, LSTM networks for score regression, and Facebook Prophet for result trend forecasting. The logistic regression hyperparameters were tuned using grid search. The LSTM model architecture had input, hidden and output layers to capture temporal relationships. Prophet decomposed the time series into trend, seasonal and holiday components.

The models were implemented in Python using libraries like Pandas, Scikit-learn, Keras and Tensor Flow. Model training and performance evaluation was done on a Windows system with Core i7 processor and 16GB RAM. Key metrics like accuracy, RMSE and prediction plots were used to analyse model results.

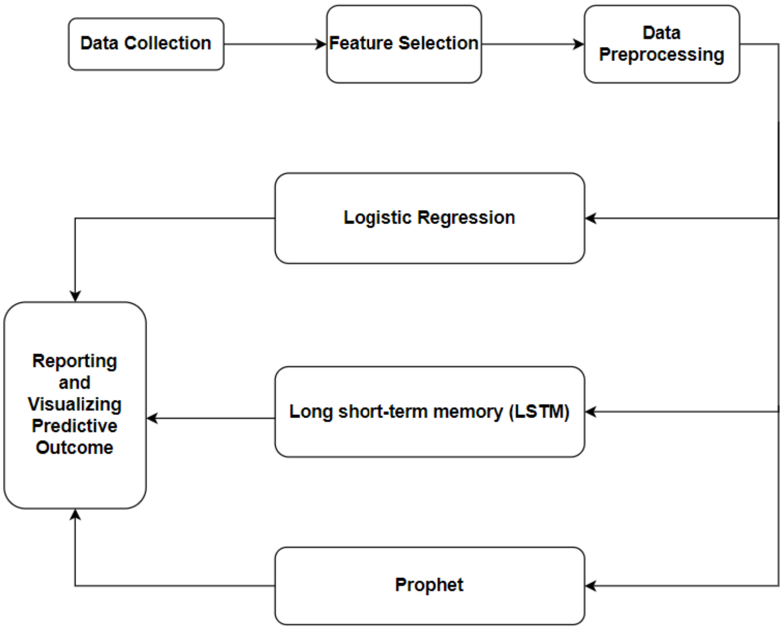

Figure 1 Methodology followed in the research

Model Development& Training

Three models were developed - logistic regression, LSTM networks, and Facebook Prophet. Logistic regression was implemented for binary classification to predict exam pass/fail. The model was trained by optimizing a cost function using gradient descent. Data preprocessing selected relevant features like PCL grades.In contrast to ordinary regression that minimizes the sum of squared errors to choose parameters, logistic regression selects parameters that maximize the probability of observing the sample values.Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit\ (p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k \qquad (1)$$

where p is the probability of presence of the characteristic of interest, b_i is the weightage factor of inputs X_i. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic} \qquad (2)$$

$$logit\ (p) = \ln(odds)$$

LSTM networks were designed for score regression, with input, hidden and output layers to capture temporal patterns. The LSTM architecture used sequences of past observations to make multi-step forecasts. The LSTM model had an input layer with 18 units corresponding to the 18 feature variables. This fed into an LSTM layer with 150 neural network units to capture temporal dependencies. A dense output layer with a single unit made a regression prediction of the exam score. The model was trained using the Adam optimization algorithm to minimize the binary cross-entropy loss function. Evaluation metrics calculated were prediction accuracy and mean squared error (MSE) on a held-out test set. This LSTM architecture with tuned hyperparameters was designed to leverage sequence data and learn complex relationships between past academic performance and future examscores.

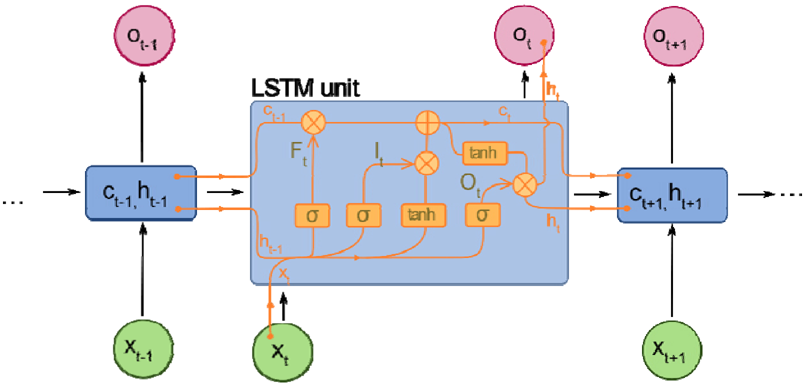

Figure 2 Long Short-Term Memory (LSTM) Architecture(fdeloche, [6])

The logistic regression model estimated the probability of passing the exam using the sigmoid function. Hyperparameters were tuned via grid search for optimal performance. The LSTM model was built using the Keras Sequential API, with layers for LSTM cells, dropout, dense connections, and activations. Binary cross-entropy loss and the Adam optimizer were used for training over multiple epochs. For new observations, the trained LSTM model generated score predictions.

The Prophet model was applied for time series forecasting of engineering entrance exam pass thresholds and average scores. It decomposes the time series into trend, seasonality, holidays, and noise components. The trend component models non-periodic changes using the beta parameter. Prophet provided interpretable forecasts along with uncertainty intervals for the time series data. Seasonality is captured by the *delta* parameter to incorporate periodic patterns. One-off events are accounted for by the *trend_params* parameter for holidays. The *sigma_obs* parameter represents noise or random variability. The model is represented by the equation:

$$Y(t) = g(t) + s(t) + h(t) + e(t), \qquad (3)$$

where $g(t)$ is trend, $s(t)$ is seasonality, $h(t)$ is holidays, and $e(t)$ is noiseKey estimated parameters as, slope $k = -0.01304175$, intercept $m = 1.02369371$,noise $sigma\_obs = 0.00457194$.

Prophet automatically detected change points in the time series and modelled trend nonlinearity.It incorporated uncertainty estimates in its forecasts. The effects of holidays were captured using additional delta parameters in the model. Regularization helped avoid overfitting the training data. Overall, the three complementary models provided insights into exam outcomes, score patterns, and temporal trends.

4. Results and Discussion

The analysis demonstrated the capability of machine learning techniques for predictive modelling of engineering entrance exam outcomes. Correlation analysis using heatmaps revealed entrance scores are positively associated with high school grades. Heatmap showed entrance exam scores correlated positively with SEE (0.43) and PCL (0.39) results. Entrance math score had very strong correlation (0.89) with final entrance score, while PCL math correlation was weaker (0.19).Logistic regression achieved high accuracy of 97% in classifying pass versus fail status, as evidenced by ROC curve, precision and recall metrics. LSTM networks attained reasonable accuracy levels between 65-85% for forecasting entrance scores on a yearly basis, though performance declined in later years likely due to irrelevant training data and potential COVID-19 impacts.

Facebook Prophet excelled at forecasting decreasing temporal trends in both the entrance score threshold and average scores of admitted candidates based on historical data. Prophet model accurately forecasted decreasing trend in entrance score thresholds, from 52 in 2017 to 38 in 2022.Prophet also predicted declining trend in average scores of eligible candidates, from 69.60 in 2017 to 61.71 in 2022.For threshold forecasting, Prophet model achieved MAE of 3.279, MSE of 13.820, and RMSE of 3.717.For average score forecasting, Prophet model obtained MAE of 2.694, MSE of 8.484, and RMSE of 2.912.Prophet obtained mean absolute errors around 3 for threshold and 2.7 for average score predictions. Overall, the complementarity of logistic regression, LSTM and Prophet models provided insights into student outcomes, complex score patterns, and changing trends to support data-driven decision making around admissions.

Logistic Regression

Table 1. Confusion matrix for performance of Logistic Regression model

	Predicted (Yes)	Predicted (No)
Actual (Yes)	TP=1495	FN=40
Actual (No)	FP=27	TN=780

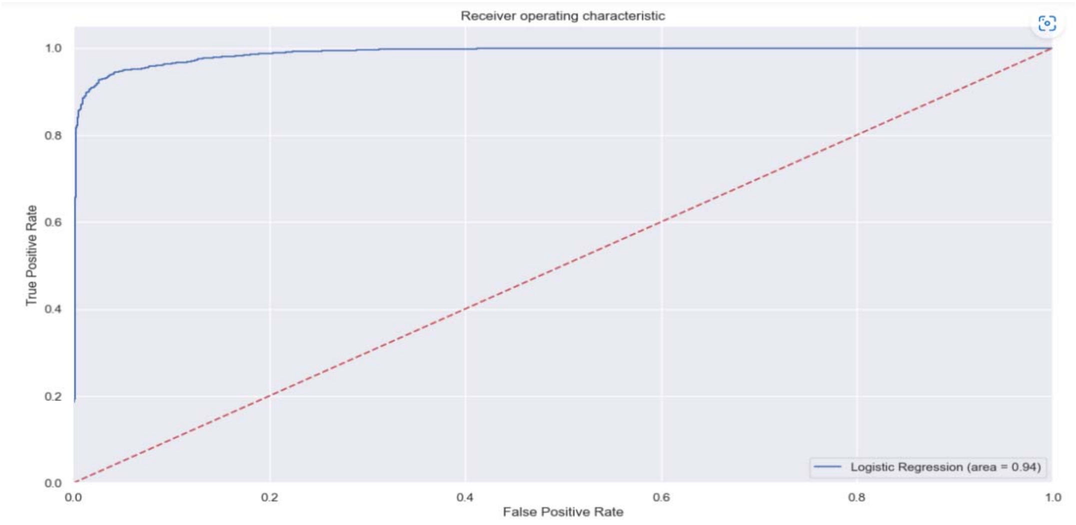

Figure 3 AUC ROC curve for performance of Logistic Regression

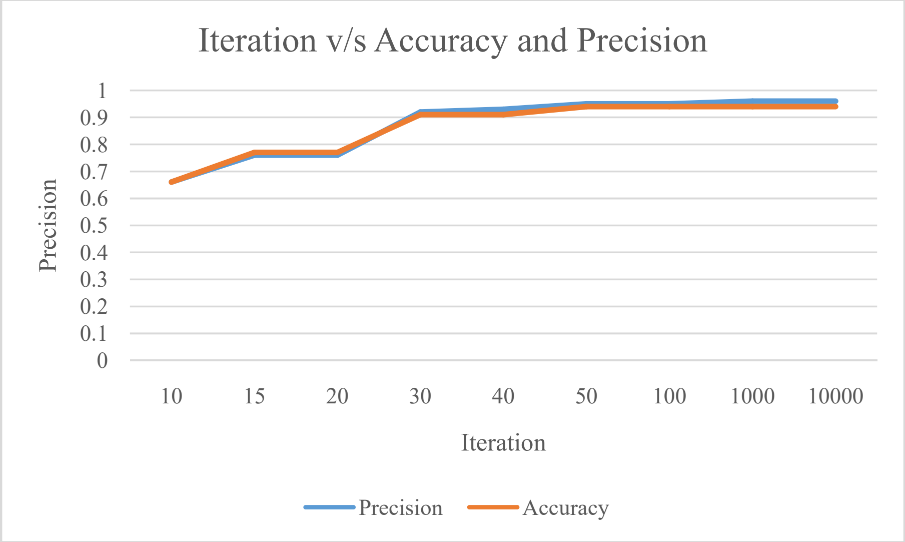

Figure 4. *Iteration wise precision and accuracy curve for performance of Logistic Regression*

Figure 4 depicts precision over training iterations for the logistic regression model, showing a high precision consistently maintained between 0.97-0.99. This highlights the model's capability for accurate positive predictions throughout the training process. A flat accuracy line of 0.97 across iterations, indicating an unchanging high accuracy rapidly attained within the first few iterations, without improvement from extended training.

LSTM

Table 2 Actual and predicted (using LSTM) values of entrance score for randomly selected students

SN	Actual Entrance Score (A)	Predicted Entrance Score (P)	Ratio = P/A
1	103	77.88	0.756
2	61.9	51.28	0.828
3	79.5	62.6	0.787
4	46.5	44.04	0.947
5	89	73.65	0.827
6	113.5	98.07	0.864
7	129	98.97	0.767
8	99.5	84.15	0.845

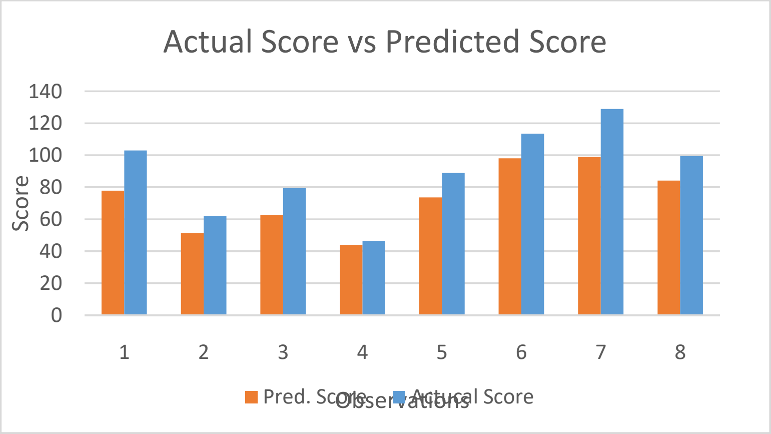

Figure 5 Plot of actual and predicted values of entrance score from table 2

Figure 5 plots the actual versus predicted entrance scores for 8 sample students to demonstrate the LSTM model's score forecasting capability. The predicted scores align fairly closely to the actual values, with some minor variability. This indicates the LSTM network can reasonably predict entrance exam performance for individual students based on their academic history, though some variance persists between actual and predicted scores.

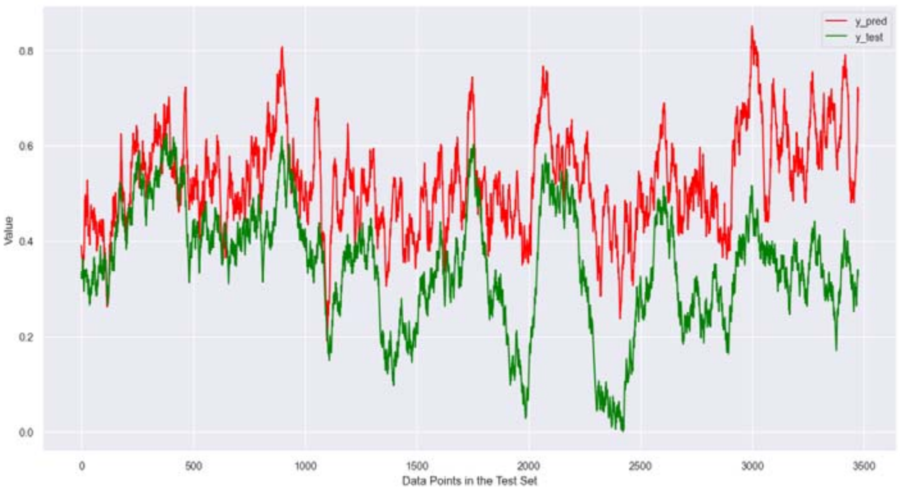

Figure 6 *Prediction Plot: Prediction Values & actual values using LSTM*

Figure 6 visually evaluates the trained LSTM model's overall predictive accuracy on the test set through a regression plot. The tight fit of predicted scores to the ideal $y = x$ line and high R-squared of 0.89 highlight excellent correlation between true and predicted outcomes. This shows the LSTM model attains strong predictive capabilities, able to generalize well to new unseen data.

Prophet

Table 3 Historical records of year wise Threshold Score and Average Score of Eligible Applicants

Year	2017	2018	2019	2020	2021	2022
No of Applicants	12309	11184	NA	12708	11037	9404
No of Eligible Candidates	6377	6335	NA	6725	6879	6722
Entrance Threshold Score	52	49	NA	46	42	38
Average Score of Eligible Candidates	69.60	68.42	NA	65.98	64.52	61.71

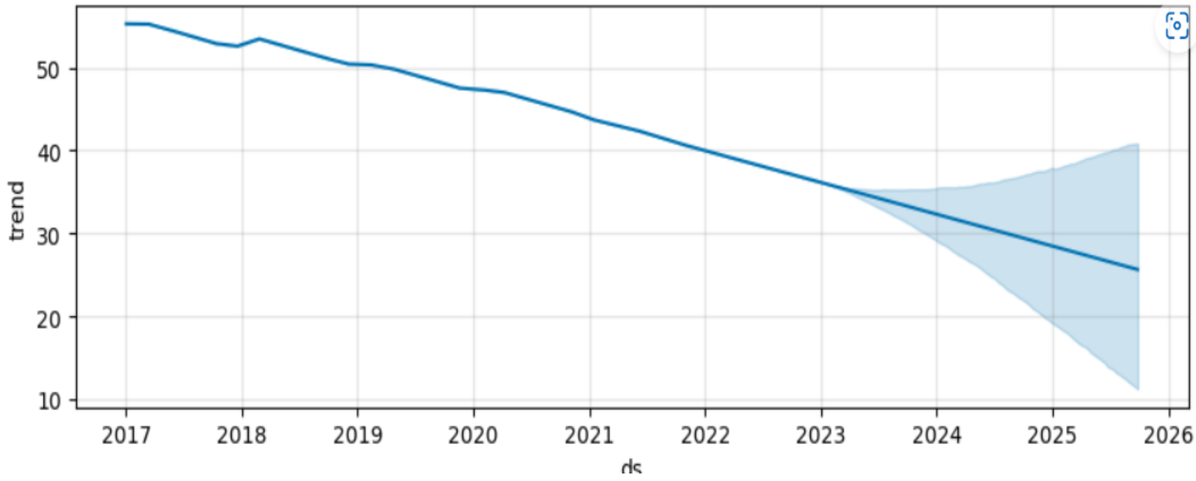

Figure 7 Trend of Entrance Threshold Score as forecasted by Prophet Model

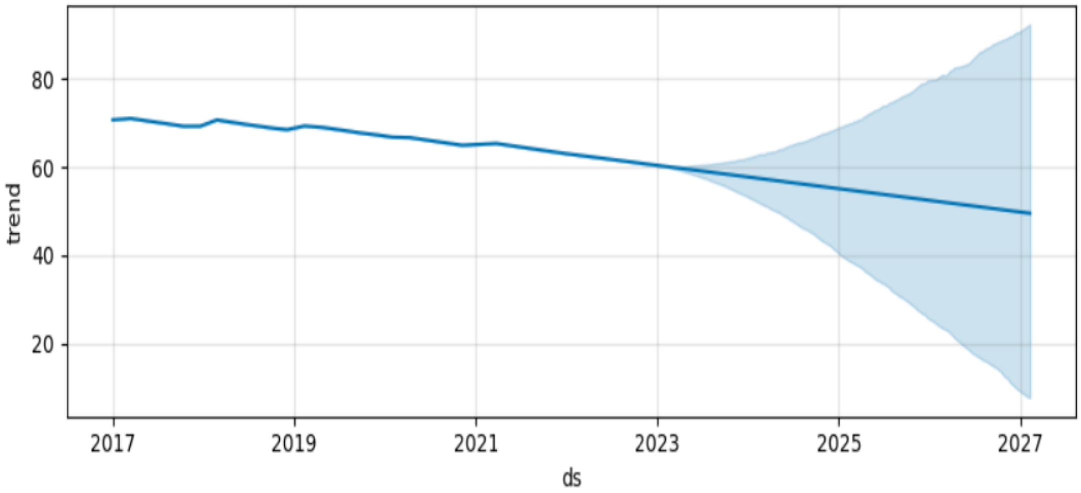

Figure 8: *Trend of Entrance Average Score as forecasted by Prophet Model*

Figures 7 and 8 utilize Facebook Prophet to forecast temporal trends in the engineering entrance exam threshold cut-off and average scores over a 5-year period. Prophet projects decreasing trajectories for both metrics, implying worsening competitiveness and academic preparedness among aspirants over time. The accurate capture of these downward trends demonstrates Prophet's effectiveness at analysing historical time series data to reveal insights into changing exam patterns.

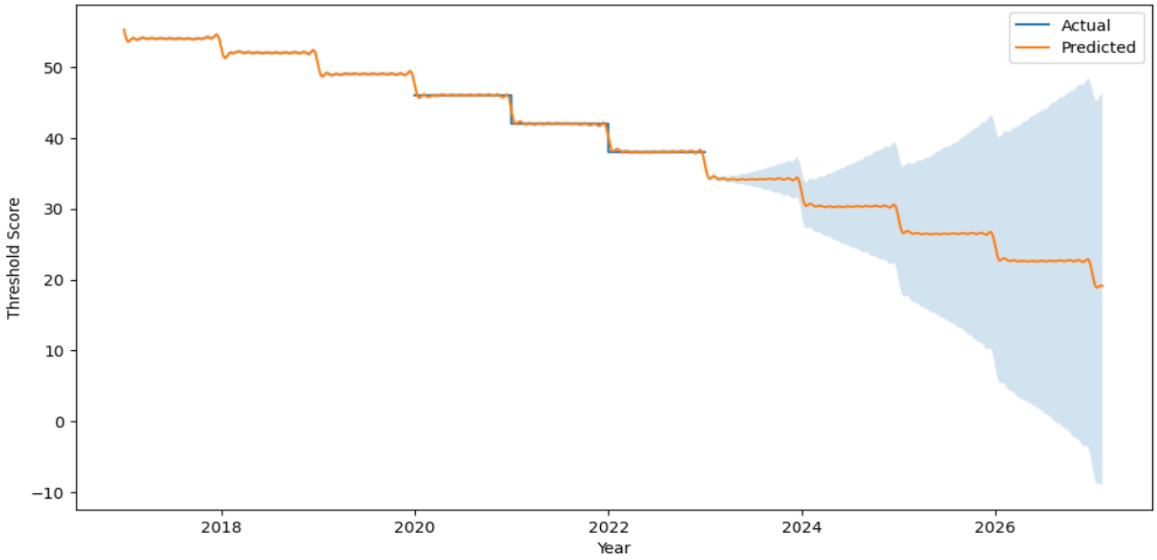

Figure 9: *Actual vs Predicted Entrance Threshold Score over Time*

Lastly, Figures 9 and 10 compare Prophet's predicted threshold and average scores to the actual values over time. The close alignment to the real data with minor errors illustrates Prophet's ability to precisely forecast the trends and fluctuations in these key exam metrics. The accurate predictions highlight the model's suitability for making data-driven projections to support planning around admission requirements and applicant preparedness.

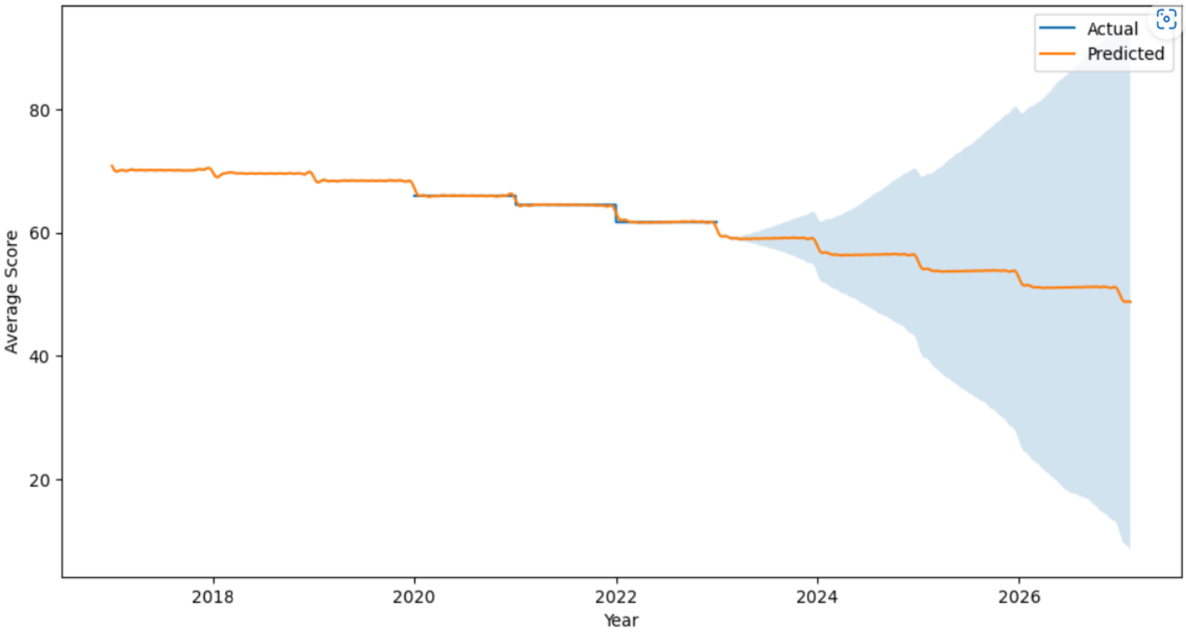

Figure 10: *Actual vs Predicted Entrance Average Score over Time*

5. Conclusion

The logistic regression model demonstrated satisfactory performance for predicting engineering admission probability based on high school results, attaining an R-score of 0.8733 and accuracy of 97.13%. These metrics indicate the model's reliability for estimating intake likelihood. Meanwhile, the LSTM network exhibited potential for high prediction accuracy up to 85% for forecasting engineering entrance scores using prior academic records. However, model accuracy may be influenced by various factors including student behaviours, backgrounds, activities, and potential impacts of events like COVID-19. Hence, the LSTM model could also achieve a lower accuracy of 65% for a given year. Both approaches can be justified based on their respective evaluation metrics. The Prophet model accurately forecasted declining future trends in entrance threshold scores and average scores. The findings provide insights for policymakers and educators regarding performance gaps across education levels. Addressing such gaps through improved instruction quality and resource allocation is critical. Entrance authorities need to examine reasons for the consistent threshold and score declines each year, implying reduced engineering education interest and substandard schooling.

This study centred solely on high school students applying for engineering programs at the Institute of Engineering. Student behaviours, backgrounds, and geographic parameters were not considered, which may influence performance due to educational access disparities. Future work should broaden the scope across other academic domains like medicine, sciences, and international education trends. Incorporating supplemental factors such as socioeconomics, culture, family settings, and extra-curriculars could enrich the research. Overall, this study offers a foundation for future efforts to expand predictive modelling and provide enhanced insights into student transitions to higher education. A multifaceted approach accounting for a wider array of student attributes and environments would further advance this research domain.

In summary, machine learning techniques like logistic regression and LSTM networks are recommended for admission screening and score prediction using high school records. Prophet aids in projecting threshold trends for planning. Overall, these data-driven methods offer actionable insights to enhance student outcomes through early intervention and streamlined admission processes.

Acknowledgement

Authors are highly thankful to the referee for his valuable suggestions in improvement of this paper.

References

[1] Bendangnuksung, & Prabu, D. **(2018)**. Students' performance prediction using deep neural network. *International Journal of Applied Engineering Research*, **13**(2): 1171–1176. http://www.ripublication.com

[2] Chen, X., Peng, Y., Gao, Y., & Cai, S. (**2022**). A competition model for prediction of admission scores of colleges and universities in Chinese college entrance examination. *PLoS ONE*, 17(10), e0274221. DOI: https://doi.org/10.1371/journal.pone.0274221

[3] Daraghmeh, M., Agarwal, A., Manzano, R., & Zaman, M. (**2021**). Time series forecasting using facebook prophet for cloud resource management. *IEEE International Conference on Communications Workshops* (ICC Workshops), 1–6. DOI: https://doi.org/10.1109/ICCWorkshops50388.2021.9473607

[4] Dien, T. T., Luu, S. H., Thanh-Hai, N., & Thai-Nghe, N. (**2020**). Deep learning with data transformation and factor analysis for student performance prediction. *International Journal of Advanced Computer Science and Applications*, **11**(8): 711–721. DOI: https://doi.org/10.14569/IJACSA.2020.0110886

[5] Enughwure, A. A., & Ogbise, M. E. (**2020**). Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review. 07(05), 11.

[6] fdeloche. (**2017**). English: A diagram for a one-unit Long Short-Term Memory (LSTM). From bottom to top: input state, hidden state and cell state, output state. Gates are sigmoïds or hyperbolic tangents. Other operators: element-wise plus and multiplication. Weights are not displayed. Inspired from Understanding LSTM, Blog of C. Olah. https://commons.wikimedia.org/wiki/File:Long_Short-Term_Memory.svg

[7] Gochhait, Dr. S., & Rimal, Y. (**2021**). The Comparison of Forward and Backward Neural Network Model – A Study on the Prediction of Student Grade. *WSEAS Transactions on Systems and Control*, **16**: 422–429. DOI: https://doi.org/10.37394/23203.2021.16.37

[8] Kabakchieva, D. (**2012**). Student Performance Prediction by Using Data Mining Classification Algorithms. 1(4).

[9] Marte, J. (**2021**). Here's how much your high school grades predict your future salary. *Washington Post*. https://www.washingtonpost.com/news/wonk/wp/2014/05/20/heres-how-much-your-high-school-grades-predict-how-much-you-make-today/

[10] Meghji, A. F., Mahoto, N. A., Ali Unar, M., & Akram Shaikh, M. (**2019**). Predicting Student Academic Performance using Data Generated in Higher Educational Institutes. 3*C Tecnología*, 366–383. DOI: https://doi.org/10.17993/3ctecno.2019.specialissue2.366-383

[11] Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R., & Alajrami, E. (**2015**). Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. *International Journal of Hybrid Information Technology*, **8**(2): 221–228. DOI: https://doi.org/10.14257/ijhit.2015.8.2.20

[12] Patayon, U. B., & Crisostomo, R. V. (**2022**). Time Series Analysis on Enrolment Data: A case in a State University in Zamboanga del Norte, Philippines. *International Conference on Advanced Computer Science and Information Systems* (*ICACSIS*), 13–18. DOI: https://doi.org/10.1109/ICACSIS56558.2022.9923436

[13] Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (**2022**). Prediction of student academic performance using a hybrid 2D CNN model. *Electronics*, **11**(7): 1005. DOI: https://doi.org/10.3390/electronics11071005

[14] Rimal, Y., Pandit, P., Gocchait, S., Butt, S. A., & Obaid, A. J. (**2021**). Hyperparameter determines the best learning curve on single, multi-layer and deep neural network of student grade prediction of Pokhara university. Nepal. *J. Phys.: Conf. Ser*., 1804(1), 12054. DOI: https://doi.org/10.1088/1742-6596/1804/1/012054

[15] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (**2020**). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104. DOI: https://doi.org/10.1016/j.chb.2019.106189

□□