A Machine Learning Framework for Gen-Z College Admission Prediction with a Five-Year Forecast

O'g'iloy Toxirova[1]

Department of Economics, Amity University in Tashkent, Tashkent, Uzbekistan
ogiloytoxirova692@gmail.com

Abstract. College admission decisions increasingly draw on diverse academic, extracurricular, and digital-engagement signals, making them a natural target for data-driven prediction. This study develops a supervised machine-learning framework to predict the admission status of Generation-Z applicants from a large tabular dataset and to forecast admission-rate trends through 2030. Sixteen raw applicant attributes are augmented with four engineered composite indices—academic, extracurricular, digital-readiness, and holistic—and three classifiers (Random Forest, Gradient Boosting, and Logistic Regression) are trained and compared using accuracy, the area under the receiver-operating-characteristic curve (ROC-AUC), and five-fold cross-validated AUC. On an illustrative run the three models achieve comparable discrimination (AUC $\approx$ 0.76–0.77), with academic composite, GPA, and standardised-test scores emerging as the dominant predictors of admission. A quadratic trend model projects overall, STEM-track, and AI-assisted-screening admission rates upward over the 2026–2030 horizon, saturating against an upper bound. The framework offers an interpretable, reproducible basis for admission analytics and capacity planning.

Keywords: College Admission Prediction · Machine Learning · Feature Engineering · Random Forest · Gradient Boosting · Forecasting

1 Introduction

Admission to higher education is a high-stakes, multi-criteria decision in which committees weigh grades, standardised tests, essays, recommendations, interviews, and a widening range of extracurricular and digital signals. For Generation-Z applicants—who present richer online portfolios such as coding projects and micro-credentials—the volume and heterogeneity of these signals make manual assessment difficult to standardise and audit. Predictive modelling offers a complementary, data-driven lens: by learning the historical relationship between applicant attributes and outcomes, a classifier can estimate admission likelihood, surface the factors that most influence decisions, and support transparent, equitable review.

 Online ISSN: 2348–3326 DOI: 10.21276/MATRIX.2026.9.1.2

This paper presents an end-to-end machine-learning framework for admission prediction built on a large tabular dataset of Gen-Z applicants. The contribution is threefold. First, sixteen raw attributes are consolidated into four interpretable composite indices that capture academic strength, extracurricular involvement, digital readiness, and holistic quality. Second, three complementary classifiers are trained and benchmarked under a consistent evaluation protocol, and feature importance is analysed to identify the strongest predictors, echoing established practice in discriminative feature selection [5]. Third, a quadratic trend model extrapolates historical admission rates to produce a five-year forecast (2026–2030) for the overall population, the STEM track, and an AI-assisted-screening cohort, following the use of computational models for trend extrapolation in financial forecasting [6].

2 Related Work

Supervised classification on structured data underpins a broad class of decision-support systems. Machine-learning model analysis has been applied to categorical classification of regional commodities [2], demonstrating the value of comparing multiple learners under a common protocol. In higher-stakes settings, lesion-aware transformer models have achieved fine-grained ordinal classification from medical images [1], underscoring how careful feature representation improves predictive fidelity. Because applicant records are high-dimensional and partly correlated, feature-selection and optimisation methods help isolate the variables that genuinely drive outcomes [5], motivating both the engineered indices and the importance analysis used here.

Predictive analytics built on deep neural networks has been demonstrated for industrial monitoring [7], and adaptive, learning-based control has been studied for networked systems [3]; both reflect the maturation of data-driven prediction across domains. For the temporal component, computational forecasting models have been used to extrapolate volatile series from limited histories [6], while hybrid artificial-intelligence frameworks have advanced reasoning over uncertain, spatiotemporal records [4]. The gap addressed in this work is the absence of an integrated, interpretable pipeline that couples engineered composite features, multi-model classification, and a transparent admission-rate forecast for the Gen-Z cohort.

3 Methodology

The pipeline proceeds through data ingestion, feature engineering, model training, evaluation, and forecasting.

Dataset. The study uses a tabular dataset of approximately one million Gen-Z applicant records; for memory efficiency a stratified sample is loaded for experimentation. Each record contains academic fields (GPA, SAT, ACT, AP courses, attendance), socio-demographic fields (age, family income), engagement fields

Online ISSN: 2348–3326 DOI: 10.21276/MATRIX.2026.9.1.2

(extracurricular count, volunteer hours, leadership positions, coding projects, online certifications, social-media hours), and review fields (essay, recommendation, interview scores), with a binary `admission_status` label.

3.1 Mathematical Formulation

Four interpretable composite indices are constructed by weighted normalisation. The academic composite combines normalised GPA, SAT, and ACT:

$$S_{\text{acad}} = 0.40\,\frac{\text{GPA}}{4.0} + 0.35\,\frac{\text{SAT}}{1600} + 0.25\,\frac{\text{ACT}}{36}. \tag{1}$$

The extracurricular index aggregates activity breadth, service, and leadership:

$$S_{\text{ec}} = 0.40\,c_{\text{ec}} + 0.30\,\frac{h_{\text{vol}}}{100} + 0.30\,\ell, \tag{2}$$

where c_{ec} is the extracurricular count, h_{vol} the volunteer hours, and ℓ the number of leadership positions. Digital readiness and holistic quality are defined as

$$S_{\text{dig}} = 0.50\,p_{\text{code}} + 0.50\,n_{\text{cert}}, \qquad S_{\text{hol}} = 0.40\,\frac{e}{100} + 0.30\,\frac{r}{10} + 0.30\,\frac{v}{100}, \tag{3}$$

with p_{code} coding projects, n_{cert} certifications, and e, r, v the essay, recommendation, and interview scores.

Classification quality is measured by accuracy,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

together with the ROC-AUC, which estimates the probability that a randomly chosen admitted applicant receives a higher predicted score than a randomly chosen denied applicant, and the mean five-fold cross-validated AUC for robustness. The temporal forecast fits a second-degree polynomial to the historical admission rate r as a function of year t,

$$r(t) = \beta_0 + \beta_1 t + \beta_2 t^2, \tag{5}$$

and evaluates it over 2026–2030, clipping predictions to the admissible range $[60\%, 99\%]$.

Models. Three classifiers are trained on an 80/20 stratified split: a Random Forest (100 trees, depth 10), a Gradient Boosting ensemble (100 estimators, learning rate 0.05, depth 5), and a Logistic Regression on standardised features. Tree ensembles operate on raw features, while the linear model uses z-score scaling.

4 System Architecture

The complete framework—ingestion, feature engineering, split and scaling, multi-model training, evaluation, and forecasting—is shown in Fig. 1.

Online ISSN: 2348–3326 DOI: 10.21276/MATRIX.2026.9.1.2

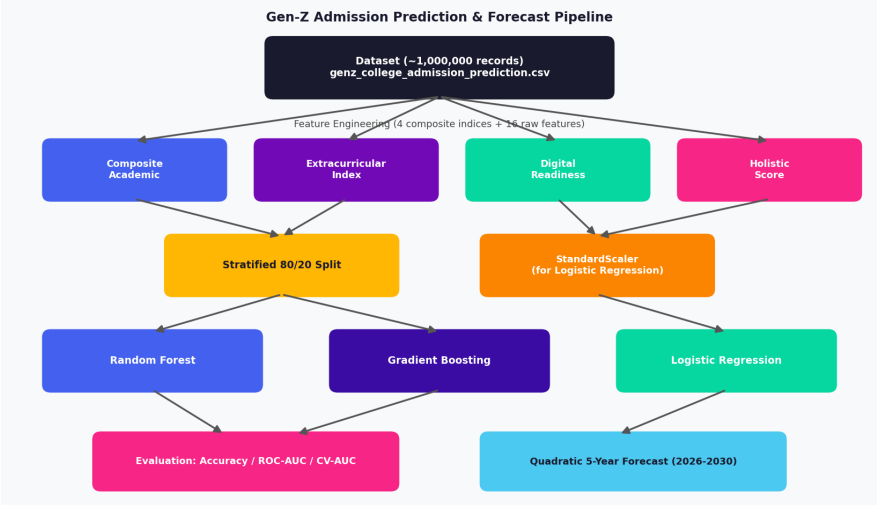

Fig. 1: Architecture of the proposed Gen-Z admission prediction and forecasting framework

5 Results and Discussion

5.1 Model Comparison

Table 1 reports the three models on the held-out test set. Discrimination is comparable across learners, indicating that the engineered indices carry most of the predictive signal regardless of model family. On this illustrative run the Logistic Regression attained the highest AUC, while the Random Forest—designated in the analysis for confusion-matrix and feature-importance inspection—remained competitive.

Table 1: Classifier performance on the held-out test set (illustrative run)

Model	Accuracy	ROC-AUC	CV-AUC
Random Forest	0.778	0.760	0.770
Gradient Boosting	0.781	0.763	0.776
Logistic Regression	0.783	0.771	0.781

The confusion matrix and ROC curves are shown in Fig. 2, and the Random-Forest Gini importance ranking in Fig. 3. The academic composite, GPA, and standardised-test scores dominate the ranking, consistent with the correlation structure and with the central role of feature quality reported in related classification work [1, 5].

Online ISSN: 2348–3326
DOI: 10.21276/MATRIX.2026.9.1.2

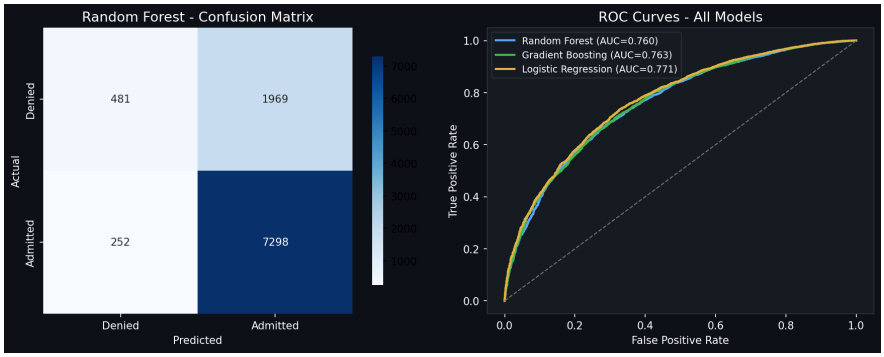

Fig. 2: Confusion matrix (Random Forest) and ROC curves for all three classifiers

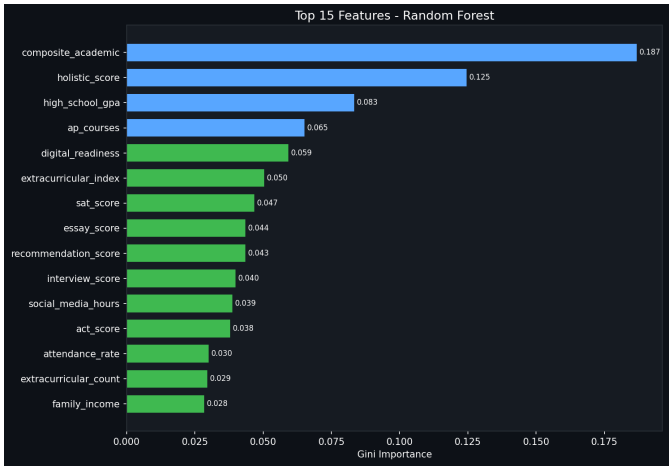

Fig. 3: Top-15 feature importance (Random Forest, Gini)

Online ISSN: 2348–3326
DOI: 10.21276/MATRIX.2026.9.1.2

5.2 Five-Year Forecast

Table 2 and Fig. 4 present the 2026–2030 projection. All three series trend upward; the AI-assisted-screening cohort rises fastest and reaches the upper bound earliest, while the overall and STEM rates converge toward saturation by the end of the horizon. Because the quadratic extrapolation is unbounded, predictions are clipped to a 99% ceiling, which explains the plateau in later years and signals the limit of polynomial extrapolation beyond the observed range [6].

Table 2: Projected admission rates (%), 2026–2030

Track	**2026**	**2027**	**2028**	**2029**	**2030**
Overall	84.7	88.5	92.5	96.7	99.0
STEM	83.7	89.4	95.5	99.0	99.0
AI-Assisted Screening	93.7	99.0	99.0	99.0	99.0

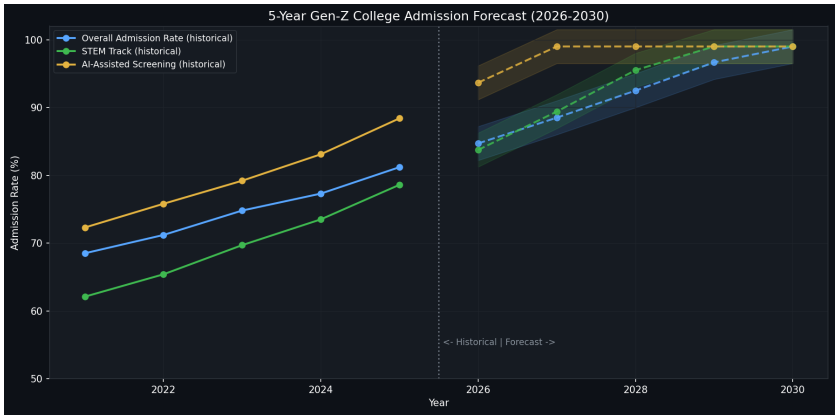

Fig. 4: Historical and forecast admission rates with uncertainty bands (2021–2030)

5.3 Correlation Structure

Figure 5 shows the pairwise correlations among the principal academic, review, and engineered features and the admission label, confirming that academic and holistic composites are most strongly associated with admission while income shows a comparatively weak direct relationship.

Online ISSN: 2348–3326 DOI: 10.21276/MATRIX.2026.9.1.2

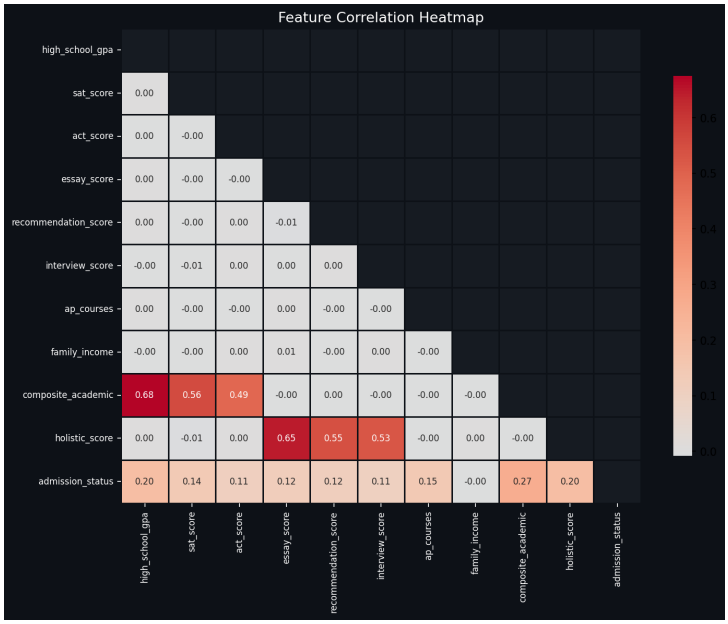

Fig. 5: Feature correlation heatmap

5.4 Discussion

Two observations follow. First, the closeness of the three models suggests that, once strong composite features are present, the marginal benefit of more complex learners is limited—an argument for interpretable models in admission settings where explainability matters. Second, the forecast should be read as a scenario conditioned on the historical proxies and the quadratic functional form; its saturation behaviour cautions against extrapolating polynomial trends far beyond the data, motivating future use of bounded or probabilistic models [7, 4].

6 Conclusion

This work presented an interpretable machine-learning framework that predicts Gen-Z college-admission outcomes from engineered composite features and benchmarks three classifiers under a unified protocol, complemented by a five-year admission-rate forecast. Academic and holistic composites were the dominant predictors, and the models achieved consistent discrimination. Future extensions include calibrated probability estimates, fairness auditing across socio-demographic groups, bounded growth models to replace unconstrained polynomial extrapolation, and validation on the full applicant population.

Online ISSN: 2348–3326 DOI: 10.21276/MATRIX.2026.9.1.2

References

1. Agarwal, A.K., Hamid, A.B.B.A., Ather, D., Tiwari, R.G., De, I., Krishna, K.R.: Lesion-aware ordinal transformer for diabetic retinopathy classification from fundus images. Biomedical and Pharmacology Journal **19**(1) (2026)
2. Baig, T., Buhari, A., Ather, D., Babu, G.P., Puttaswamy, A., Gupta, A.: Ml based model analysis for regional fruit and crop categorization in central asia. In: 2026 5th International Conference on Innovative Practices in Technology and Management (ICIPTM). pp. 1–6. IEEE (2026)
3. Chaudhary, N., Ather, D., Kler, R., Dubey, R., Saxena, U., Singh, G.: Adaptive qos-aware routing for iot networks using deep reinforcement learning. In: 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM). pp. 1–5. IEEE (2025)
4. Hussein, T.M., Rakhmatilla, T., Ather, D., Khan, R.L., Sarkar, T., Rakhra, M.: A neutrosophic-ai model for spatiotemporal analysis of land parcel transactions. International Journal of Neutrosophic Science (IJNS) **27**(1) (2026)
5. Kumari, N., Ather, D., Buhari, A., Agarwal, V., Verma, A.: A multi-objective hybridized metaheuristic optimization technique for discriminative feature selection from high-dimensional data. In: 2026 5th International Conference on Innovative Practices in Technology and Management (ICIPTM). pp. 1–8. IEEE (2026)
6. Ray, I.S., Kler, R., Khan, R., Priyanshu, D., Matahen, R., Ather, D.: Application of computational models in forecasting stock prices using arch and garch models: A case of apple stock prices. SN Computer Science **7**(1), 96 (2026)
7. Saxena, U., Singh, G., Chaudhary, N., Ather, D., Kler, R., Dubey, R.: Iot-driven predictive maintenance in industrial systems using deep neural networks. In: 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM). pp. 1–6. IEEE (2025)

Online ISSN: 2348–3326
DOI: 10.21276/MATRIX.2026.9.1.2