Prediction of College Admission Scores Based on an XGBoost-LSTM Hybrid Model

Liangyu Xu

{xly971227@gmail.com}

Wuhan Institute of Technology, No. 206, Guanggu 1st Road, East Lake High-tech Development Zone, Wuhan, Hubei Province, China

Abstract. This study introduces a predictive model that combines XGBoost and Long Short-Term Memory (LSTM) networks for forecasting the minimum college admission scores in the Chinese college entrance examination system. By leveraging the strengths of LSTM in handling multivariate time series data and the efficiency of XGBoost in processing categorical data, the fusion model enhances prediction accuracy. The experimental results demonstrate that this hybrid model outperforms individual models in predicting college entrance scores, offering valuable support for educational planning and student decision-making.

Keywords: XGBoost, LSTM, Chinese college admission scores

1 Introduction

The establishment of college admission score thresholds significantly impacts the annual college admissions process. Accurately predicting these thresholds is crucial for the strategic planning of educational institutions and for informing students' decision-making. Current forecasting methods include the Score Difference Method, Average Ranking Method, and neural network algorithms like BP (Back Propagation) and LSTM (Long Short-Term Memory) networks.

The Score Difference Method predicts college admission scores by calculating the difference between a college's minimum admission score and the provincial control score for that year. This method forecasts current year scores by adding the calculated difference from previous years to the current year's provincial control score. While straightforward and easy to compute, this approach often results in lower accuracy due to its simplistic assumption that the year-over-year changes in admission scores and provincial control scores remain constant.

The Average Ranking Method employs a detailed approach to forecast college admission scores by utilizing the provincial rankings corresponding to each college's minimum admission score. It involves mapping the scores of the previous n-1 years to their respective provincial rankings. By calculating the average of these rankings, the method predicts the admission score for the nth year. This technique leverages historical data to provide a more nuanced prediction based on trends in ranking changes over time, offering a refined analysis compared to simple score-based predictions.

EIMT 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347687

The BP neural network is recognized for its robust nonlinear mapping capabilities, effectively realizing a function that maps input variables to output predictions. This network can be directly employed to forecast the minimum admission scores[1][2] and rankings[3]. On the other hand, the LSTM network, a significant variant of recurrent neural networks, is distinguished by its ability to learn long-term dependencies. This characteristic makes it particularly suitable for predicting college admission scores, where historical data trends play a crucial role[4].

Existing models either utilize BP neural networks to process score features and school categorical features for individual years or apply LSTM for time series score features without integrating these data types. Consequently, this paper enhances LSTM and pairs it with XGBoost to forecast both the lowest admission score and its percentile. Subsequently, XGBoost classifies these forecasts to produce the final predictions, thereby innovatively combining both data types for improved accuracy in admission score forecasting.

2 College Admission Score Prediction Model

2.1 XGBoost Model

XGBoost represents an enhanced ensemble learning algorithm derived from gradient boosting decision tree methodologies[5]. The principle underlying its predictions can be articulated as follows: the predicted value for each instance is the sum of the products of each sample and its corresponding weight, expressed by equation (1):

$$\hat{y}_i = \sum_j w_j x_{ij} \tag{1}$$

where j denotes the number of samples, w_j represents the weight, and x_{ij} refers to the sample data. In regression tasks, XGBoost sequentially integrates trees into the model to incrementally improve performance. This ensemble process is described by equation (2):

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

Here, $\hat{y}_i^{(t)}$ signifies the prediction of the model at iteration t, with $\hat{y}_i^{(0)} = 0$. $\hat{y}_i^{(t-1)}$ represents the cumulative prediction up to iteration t-1, and $f_t(x_i)$ is the tree added at iteration t. To mitigate the risk of overfitting as more nodes are added, a regularization term is incorporated within the objective function, defined by equation (3):

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

where γT is a penalty coefficient and $\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ constitutes the regularization term, with γ acting as the coefficient for the number of leaf nodes T. The objective function consists of the loss function and a regularization penalty, and is formulated by equation (4):

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)} + f_t(x_i)\right) + \Omega(f_t) + constant \tag{4}$$

Here, obj indicates the structural score, representing the maximum reduction in the objective when a tree structure is selected.

Regarding XGBoost's application, it processes input features of schools x_s, all score features, the college admission score, and the percentage of the college admission score from the previous year $x_{n,t-1}$, y_{t-1}, y'_{t-1}, and other score features for year t excluding the prediction target $x_{n,t}$. Two models are trained to predict the college admission score and the percentage of the college admission score for the year t, yielding $\hat{y}_t^{(XGB)}$ and $\widehat{y'}_t^{(XGB)}$, see equation (5) and equation (6):

$$\hat{y}_t^{(XGB)} = XGB\left(x_s, x_{n,t-1}, y_{t-1}, y'_{t-1}, x_{n,t}\right) \tag{5}$$

$$\widehat{y'}_t^{(XGB)} = XGB'\left(x_s, x_{n,t-1}, y_{t-1}, y'_{t-1}, x_{n,t}\right) \tag{6}$$

2.2 LSTM model

The Long Short-Term Memory (LSTM) network, a particular variant of Recurrent Neural Networks (RNNs), was introduced by Hochreiter and Schmidhuber in 1997[6]. LSTM networks are particularly adept at handling time series data, demonstrating exceptional performance in scenarios involving long-term dependencies.

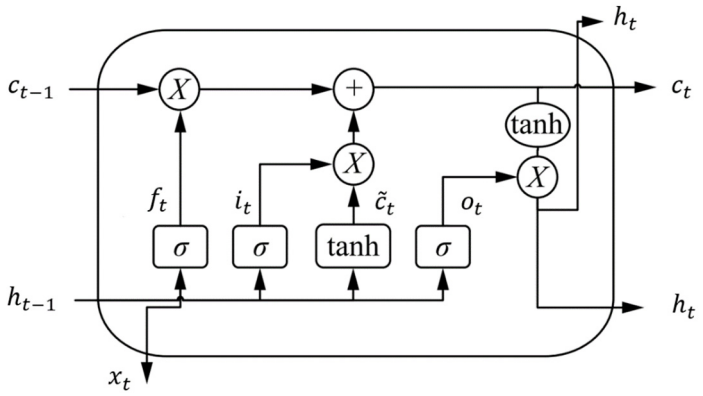

Figure 1. The architecture of an LSTM unit.

As shown in Figure 1, in LSTM, each unit controls the flow of information through a set of gating mechanisms. These gates include the forget gate f_t, the input gate i_t, and the output gate o_t, whose defining equations are seen in equations (7), (8), and (9) respectively. Here, h_{t-1} represents the output value of the LSTM at the previous moment. σ represents the sigmoid function.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{7}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

The forget gate decides how much of the cell state c_{t-1} from the previous moment is retained to the current moment c_t, and the input gate determines how much of the current network input x_t is saved to the cell state c_t. The calculation process of c_t is seen in equations (10) and (11).

$$\tilde{c}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (10)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (11)$$

The output gate controls how much of the cell state c_t is output to the current output value h_t of the LSTM, with the calculation formula of h_t seen in equation (12).

$$h_t = o_t * tanh(c_t) \quad (12)$$

Here, tanh represents the hyperbolic tangent function. W_f, W_i, W_C, W_o are weight matrices. b_f, b_i, b_C, b_o are biases.

2.3 Modified LSTM model

In the context of predicting college admission score, the input data can be categorized into school features and score features. School features encompass attributes related to the school, such as geographical location, type of institution, and ratings from educational websites. These characteristics undergo a one-hot encoding transformation, denoted as x_s. Score features include metrics such as the college admission score for different tiers, average scores, and highest scores of the current year. These features possess temporal information, with the score features of year t represented as $x_{n,t}$.

The improvement approach involves adding a parallel feedforward neural network to process school features, while an LSTM handles the time-series score features. The results from both parts are then concatenated and passed through fully connected layers to produce the final predicted score. The model structure is shown in Figure 2.

Specifically, the hidden layer *h* is obtained by processing the time-series score features $x_{n,t-2}$, $x_{n,t-1}$ from years t−2 and t−1, along with the admission score and admission score percentage y_{t-2}, y'_{t-2}, y_{t-1}, y'_{t-1} through an LSTM, as seen in equation (13).

$$h = LSTM\left(x_{n,t-2}, y_{t-2}, y'_{t-2}, x_{n,t-1}, y_{t-1}, y'_{t-1}\right) \quad (13)$$

School features and the current year's score features x_s, $x_{n,t}$ are processed through two fully connected layers to produce a vector *b*, as seen in equation (14):

$$b = FNN\left(x_s, x_{n,t}\right) \quad (14)$$

The vectors *h* and *b* are then concatenated and passed through a final fully connected layer to obtain the final predicted score for year t, represented by equation (15):

$$\hat{y}_t^{(LSTM)} = FC([h, b]) \quad (15)$$

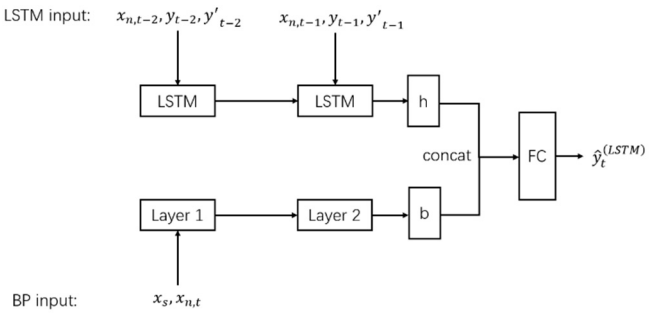

Figure 2. The architecture of modified LSTM model.

This methodology leverages the LSTM's capability to process sequential data and fully connected layers to handle non-temporal features, facilitating a comprehensive approach to predicting college admission score.

2.4 The stacked ensemble method

In the context of high school students selecting their university preferences based on entrance examination scores, both the scores themselves and their corresponding rankings can serve as reference points. Hence, both the score and the rank percentage can be utilized as targets for prediction. This study employs both XGBoost and an enhanced LSTM model to individually forecast the scores and the rank percentages, yielding four sets of predictive outcomes:

The college admission score predicted by the LSTM model, denoted as $\hat{y}_t^{(LSTM)}$;The percentage of the college admission score predicted by the LSTM model, denoted as $\widehat{y'}_t^{(LSTM)}$;The college admission score predicted by the XGBoost model, denoted as $\hat{y}_t^{(XGB)}$;The percentage of the college admission score predicted by the XGBoost model, denoted as $\widehat{y'}_t^{(XGB)}$.

To convert the rank percentages into college admission scores, a transformation is applied: $\widehat{y'}_t^{(LSTM)}$ to $\widehat{y''}_t^{(LSTM)}$ and $\widehat{y'}_t^{(XGB)}$ to $\widehat{y''}_t^{(XGB)}$. Subsequently, these four predictive outcomes, along with recent years' score features and school characteristics x_o, are used to train a classification XGBoost model. This model is tasked with selecting one of the four predictions as the final forecasted result, formalized by equation (16):

$$\hat{y}_t = f\left(\hat{y}_t^{(LSTM)}, \widehat{y''}_t^{(LSTM)}, \hat{y}_t^{(XGB)}, \widehat{y''}_t^{(XGB)}, x_o\right) \tag{16}$$

This ensemble method leverages the strengths of both LSTM and XGBoost models to enhance the accuracy of the final predictions by incorporating a diverse range of predictive insights and historical data.

3 Experimental Design and Results

3.1 Experimental Data

The experiment utilizes enrollment data from various universities in Hubei Province for the years 2016 to 2023, focusing exclusively on science streams and general categories, totaling 1998 entries. The input features for this study are categorized into two types:

School Features: This includes the type of school (e.g., engineering, comprehensive), level of enrollment (e.g., undergraduate, diploma), and geographical location (e.g., Beijing, Shanghai).

Score Features: This encompasses previous years' cut-off scores, first-tier and second-tier scores, highest scores, average scores, and the cumulative percentage of students for each score segment from a distribution table.

3.2 K-fold Cross-validation by Year

Three models are employed in the study: XGBoost and LSTM models for predicting the cut-off score and the percentage of the college admission score, along with another XGBoost

model that integrates these results. For predictions for the year t, only features from years up to t-1 and all features from year t except for the college admission score are available. A simple division of data into training sets from years before t and using year t data as the validation set would not suffice for training the ensemble XGBoost model due to lack of data. Hence, a K-fold cross-validation by year is utilized.

Data from 2016 to 2022 are subjected to K-fold cross-validation, with data from 2023 serving as the validation set. Given that the LSTM model requires data from the previous two years for feature input, years 2016 and 2017 are excluded from validation. Consequently, data from 2018 to 2022 are divided into 5 folds by year for cross-validation. For instance, data from 2019 to 2022 are used as the training set, with 2018 serving as the validation set.

Training data for the XGBoost classifier encompass the training sets from the 5-fold cross-validation along with model predictions. The final prediction for 2023 is obtained by averaging the predictions from four models trained through 5-fold cross-validation, which then serve as inputs to the final XGBoost classifier to yield the ultimate prediction outcome.

3.3 Experimental Results

Table 1 presents the Mean Absolute Error (MAE) of the predictive outcomes generated by four models. The MAE is calculated by converting the predicted result as a percentage into a score.

Table 1. Experimental results.

Model	K-fold MAE	Prediction MAE
XGBoost (Score)	10.79	13.61
XGBoost (Percentage to Score)	13.38	14.78
LSTM (Score)	15.46	14.48
LSTM (Percentage to Score)	13.19	14.11

A new XGBoost classifier was developed by integrating the outcomes of four previously predicted results along with additional features. This innovative classifier is designed to predict the model exhibiting the lowest error rate. The classification accuracy of the ensemble prediction model was recorded at 44.09%, and the final Mean Absolute Error (MAE) was determined to be 12.09. This achievement signifies a reduction of 1.52 in MAE when compared to the best-performing individual XGBoost model, underscoring the efficacy of the ensemble approach in enhancing predictive accuracy.

4 Conclusion

This study introduces an innovative approach to predicting university cut-off scores by integrating XGBoost and LSTM models through a stacked ensemble method. The proposed approach demonstrates potential in improving predictive accuracy, which is vital for educational planning and student guidance. Future efforts will focus on refining the model and expanding its applicability in other predictive contexts within the educational domain.

References

[1] Y. Zhang et al., "Research on the Prediction Method of the College Professional Admission Scores," 2022 International Seminar on Computer Science and Engineering Technology (SCSET), Indianapolis, IN, USA, 2022, pp. 406-409, doi: 10.1109/SCSET55041.2022.00098

[2] Hu Shuang. Research on the Prediction of College Entrance Examination Score Lines for the New College Entrance Examination Based on Artificial Neural Networks [D]. Dongbei University of Finance and Economics, 2022. DOI: 10.27006/d.cnki.gdbcu.2022.001456.

[3] Xu Zebin. Research on Predicting Admission Rankings of Colleges and Recommending Candidates' Preferences under the New College Entrance Examination [D]. Nanjing University of Posts and Telecommunications, 2023. DOI: 10.27251/d.cnki.gnjdc.2023.000140.

[4] Song Shiyu. Research and Application of Personalized Recommendation Method for College Entrance Examination Preferences Based on Machine Learning [D]. North China University, 2023. DOI: 10.27470/d.cnki.ghbgc.2023.000758.

[5] Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 785–794.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.