Reconstructing A Framework for Predicting University Tuition Payment Trends Using Temporal Fusion Transformer

Zhiqun Zhang[1], Yaya Wang[2]
[1]Planning and Finance Division, Shandong University of Aeronautics, Binzhou, 256600, Shandong, China
[2]Department of Information Engineering, Binzhou Polytechnic, Binzhou 256600, Shandong, China
Corresponding author: Zhiqun Zhang (e-mail: twentymay@126.com)

ABSTRACT To improve the prediction accuracy and interpretability of complex university tuition payment time series, this paper proposes a reconstructed forecasting framework based on the Temporal Fusion Transformer. The framework encodes multi-source heterogeneous data, including static and dynamic variables such as student majors, economic status, scholarship disbursement plans, and real-time payment reminders, into a low-dimensional representation space. A sigmoid gated network dynamically weights informative variables and suppresses redundant features. LSTM modules and a specialized three-head attention mechanism for time patterns, group synchronization, and policy response are jointly used to model long-term dependencies and capture nonlinear payment fluctuations. Interpretable multi-step forecasts are generated by combining sequence outputs with attention weights. Experimental results show that the model achieves RMSE ≤ 0.08 and MAE ≤ 0.07 during peak payment periods, with R2 above 0.9 in five disturbance scenarios, demonstrating strong dynamic adaptability. Interpretability analysis yields a Pearson correlation coefficient of at least 0.68 and an information gain ratio of at least 0.7, confirming robust feature attribution and discriminative capability.

INDEX TERMS temporal fusion transformer, tuition payment prediction, heterogeneous data, gating mechanism, interpretable forecasting

I. INTRODUCTION

UNIVERSITY tuition payment behavior is a complex temporal process driven by the combined effects of institutional design (e.g., payment schedules and penalties), economic incentives (e.g., financial aid and installment options), and information access (e.g., clarity of billing communication). This dynamic interaction of multiple factors presents significant challenges to prediction, similar to other complex fields. such as how the timing of orders for high-tech machinery is influenced by government subsidies (institutional design/economic incentives) and clear market data (information access) [1,2]. Tuition collection not only involves financial processes but is also embedded in a collaborative system of scholarship and grant distribution, student status management, and digital services [3]. Modern universities are generally faced with the management reality of multiple payment batches, differentiated funding, and dynamic notification strategies. The payment completion rate is subject to the intertwined influences of individual economic capabilities, policy perception intensity, and the psychological effects of time nodes [4-6]. In this context, developing a prediction system capable of integrating multi-source heterogeneous data has become essential technical support for accurately forecasting capital flow and enabling proactive student service intervention. This requirement is analogous to the critical need within the manufacturing sector for predictive analytics that integrate diverse data—such as raw material futures, supply chain logistics, and historical production outputs—to accurately forecast cash flow and manage inventory for effective operational control.

Current methods for predicting university tuition fees have significant shortcomings: their modeling approach is severely out of touch with real-world application needs. Most systems treat payment rates as independent time series for extrapolation, ignoring the underlying generative mechanisms woven together by institutional cycles, individual attributes, and external interventions [7,8]. When the timing of scholarship receipt shifts or the payment deadline is adjusted, historical patterns quickly become ineffective, and the model struggles to capture such structural transitions [9]. Existing methods only integrate multi-source information superficially, with heterogeneous variables such as student financial status, notification frequency, and payment channel status being input with equal weight, lacking the ability to dynamically identify the time-varying importance of variables [10-12]. This makes it impossible for the model to distinguish between core driving factors and background noise at key decision nodes, resulting in a lack of directionality in the prediction results [13,14]. More significantly, the output side generally lacks transparency, leaving managers unable to determine whether an increase in forecasted values during a particular period is due to the cumulative effect of reminder text messages or the psychological impact of approaching deadlines [15,16]. This "black box" inference undermines the practical value of forecasts in resource allocation and risk early warning. Furthermore, different student groups exhibit significant heterogeneity in their response patterns: those with financial difficulties are more sensitive to notification delays, while those with greater financial security are more constrained by institutional constraints. A unified modeling framework struggles to account for these differences, leading to systematic biases in overall forecasts at the subgroup level [17,18]. Existing technical approaches have yet to establish a modeling paradigm that integrates rule perception, variable identification, and attribution visualization, hindering the evolution of forecasting models from statistical tools to decision support systems. This research aims to build an interpretable deep learning framework for predicting university tuition payment trends. The core goal is to overcome the bottleneck of dynamic modeling under the coupling of multiple heterogeneous factors and achieve a balance between high accuracy and strong interpretability. By reconstructing the input encoding structure of the Temporal Fusion Transformer, this paper hierarchically embeds student static attributes, payment rule variables, and dynamic event streams, ensuring semantic separation and alignment of different levels of information in a low-dimensional space. An innovative gated variable selection network design empowers the model to autonomously select key drivers at different time points, suppressing the interfering effects of irrelevant variables. At the sequence modeling level, the temporal memory properties of LSTMs are combined with a multi-head attention mechanism to synergistically capture the evolution of long-term trends and the instantaneous impact of sudden events, particularly enhancing the accuracy of modeling institutional effects. Finally, the attention weight matrix output by the decoder is used to generate a temporal trajectory map of variable contributions, allowing predictions to be traced back to specific policy actions or notification behaviors. The experimental framework is validated across four dimensions: short-term forecast error, disturbance responsiveness, group stability, and feature consistency, covering typical challenges in real-world management scenarios. Compared to existing methods such as ARIMA, LSTM, GRU, XGBoost, Informer, and Autoformer, this framework not only improves forecast accuracy, but more importantly, it provides a visual and interpretable basis for financial decision-making, enabling university finance departments to identify high-risk periods, optimize collection strategies, and assess the impact of policy adjustments. This capability offers a technical path for intelligent university finance that combines both predictive performance and transparency, much like how a similar analytics system could be used in the industry to visually map the influence of fluctuating prices or new regulatory standards on production capacity and inventory management.

II. RELATED WORK

In recent years, time series modeling techniques have been gradually explored in the field of education. Currently, the popular ARIMA (Autoregressive Integrated Moving Average Model) linear model is used to predict educational enrollment rates, but its adaptability to structural changes is limited, and its error increases significantly under disturbances [19,20]. Tree models such as XGBoost have been applied to student performance prediction and student growth calculation. Although they have made breakthroughs in static feature fusion, they have difficulty handling sequential dependencies and time alignment [21,22]. Among deep learning methods, GRU (gated recurrent unit) [23,24] and LSTM (long short-term memory) [25,26] structures have been applied to academic performance prediction and network monitoring. They have certain nonlinear fitting capabilities, but are unstable in terms of long-term dependencies and multivariate selection, and are particularly susceptible to noise interference in high-dimensional input environments. The application of the Transformer architecture provides new insights for modeling long-range dependencies. Informer [27,28] and Autoformer [29] reduce computational complexity through sparse attention mechanisms and have achieved success in applications such as power load and traffic flow; Zhao L proposed an Autoformer model optimization method to improve the accuracy of power load forecasting. Experimental results show that the average absolute error and mean square error of this method are reduced to 1.84 and 5.21, respectively, which is significantly better than the performance of a single model [30]. However, in the context of education payment, their ability to integrate static covariates and event markers is insufficient, failing to effectively distinguish the hierarchical relationship between institutional factors and individual responses. Previous studies have attempted to incorporate external variables as additional sequence inputs, but lack modeling of the dynamic evolution of variable importance, resulting in key signals being submerged in redundant information flows. Current approaches still suffer from gaps in key areas such as heterogeneous data fusion, dynamic feature selection, and decision logic visualization, and have yet to establish a systematic modeling paradigm for the unique mechanisms of university payment.

The Temporal Fusion Transformer (TFT), a neural network architecture designed specifically for multivariate time series forecasting, demonstrates potential for addressing these discontinuities. This framework utilizes components such as static/dynamic variable separation embedding, gated feature selection, and sequence-to-sequence attention decoding to achieve structured processing of heterogeneous inputs and identify critical paths [31]. Lim B's research shows that cold start time series prediction aims to deal with the problem of scarce historical data, and achieves prediction by mining the similarity between product features. Compared with traditional models, this method improves the accuracy of new product sales prediction by about 15% [32]. TFT has proven effective in scenarios such as runoff forecasting and tourism demand prediction, with its attention weight distribution tracing the temporal trajectory of variable contributions [33,34]. Similarly, in the field of dynamic design, Cid Montoya M used TFT to predict the time series of self-excited aerodynamic elastic forces on the bridge deck, and successfully achieved high-precision and interpretable prediction of nonlinear aerodynamic behavior [35]. These practices demonstrate that TFT has the ability to handle institutionally driven behavior sequences. However, existing applications mostly focus on design or flow prediction scenarios, and their input design is not suitable for the coupling of strong rule constraints and hierarchical response mechanisms in university payment. These studies still employ general-purpose attention mechanisms, which are unsuitable for structurally optimizing intervention logic in education management. This leads to a discrepancy between explanatory output and managerial intuition. Therefore, this paper proposes a domain-reconstruction of the TFT framework, embedding semantic encoding and attention constraints specific to education payment, to simultaneously enhance predictive performance and decision support capabilities. The framework developed in this study is primarily applicable to higher education institutions with the following characteristics: digital student information systems, multi-batch payment policies, and demand-or achievement-based financial aid systems. These features are typical management practices in many large public and private universities across North America, Western Europe, and East Asia.

III. CONSTRUCTION OF A FRAMEWORK FOR PREDICTING UNIVERSITY TUITION PAYMENT TRENDS BASED ON TFT

A. STRUCTURED CODING PROCESSING OF MULTI-SOURCE HETEROGENEOUS DATA

Figure 1 details the structured encoding process for multi-source heterogeneous data in university tuition payment prediction. Static attributes are mapped into dense vectors via an embedding layer and concatenated with normalized funding eligibility features to form a static feature tensor. Known time-varying covariates are projected into a time series tensor via hybrid encoding to ensure precise alignment of policy signals. Observed time-varying data is instance-normalized and embedded to preserve individual behavioral characteristics. These three types of tensors achieve semantic decoupling and gradient stabilization through a dimensionally aligned but path-isolated encoding strategy, providing a well-defined input foundation for subsequent modeling.

1) Semantic Separation Encoding of Static Attributes and Dynamic Variables

The data comes from a university's student payment management database from 2018 to 2023. It covers records from multiple systems, including student status, finance, and notification logs. After desensitization, it forms the original dataset for this study and is divided into training, validation, and test sets at a ratio of 7:2:1. Student basic information includes categorical fields such as grade, major, and student status. These variables remain stable throughout an individual's lifecycle and constitute fundamental factors influencing payment behavior. For this type of static attribute, a separate embedding layer performs nonlinear mapping, converting each category value into a dense vector representation of fixed dimension. The embedding operation is implemented using a learnable parameter matrix. During training, backpropagation is used to adjust the vector space distribution so that semantically similar categories appear clustered in the latent space. Let $c \in C$ be the student's major category, where C is the set of all majors. The embedding mapping function is:

$$\mathbf{e}_c = \mathbf{W}_{\text{emb}}^{(s)} \cdot \mathbf{v}_c \tag{1}$$

$\mathbf{v}_c$ is the one-hot representation of category c; $\mathbf{W}_{\text{emb}}^{(s)}$ is the learnable embedding weight matrix; and $\mathbf{e}_c$ is the output dense vector.

The embedded vectors are concatenated into a joint static feature tensor for subsequent network branches to use. This process avoids one-hot encoding of categorical variables, effectively compressing the input dimension and suppressing gradient perturbations caused by sparsity.

Financial data, including information on scholarship and grant distribution plans, tuition standards, and eligibility for exemptions, are time-varying covariates known a priori, and their temporal trajectories can be obtained in advance within the prediction domain. These variables contain a mixture of numerical and categorical types, and a channel-wise encoding strategy is implemented during processing. Categorical subsets, such as grant type and payment batch identifier, are also mapped through a dedicated embedding layer. Numerical subsets, such as amounts due and amounts already exempted, undergo standardization, using the mean and standard deviation derived from sliding window statistics to achieve numerical compression and ensure dimensional consistency. All known covariates are aligned on the time axis as equally spaced sequences and input into the model simultaneously with the target sequence. This encoding method ensures the temporal alignment accuracy of external intervention signals, enabling the model to accurately perceive the presence and intensity changes of policy actions at specific time steps.

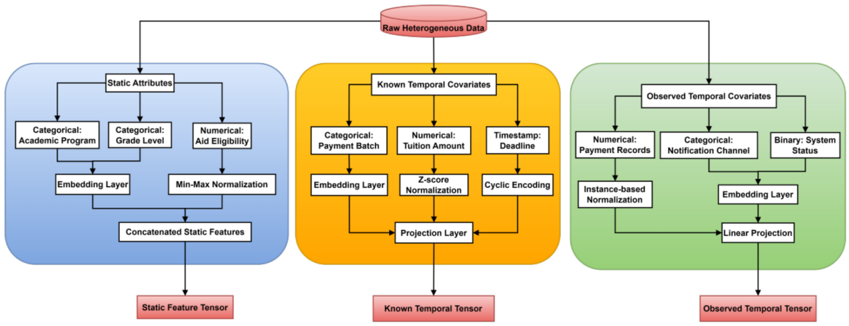

FIGURE 1. Structured encoding of multi-source heterogeneous data

2) Normalization and Linear Projection of Observed Time Series Variables

Observed time-varying covariates reflect unpredictable real-time behavioral flows, including actual payment records, number of payment reminders, and frequency of payment failure feedback. This data is complete only within the historical observation window and cannot be obtained in advance during the prediction phase. Raw observations exhibit significant magnitude differences and distribution shifts, and direct input leads to unbalanced gradient updates. A dynamic normalization mechanism is employed to scale numerical variables using the local statistics (mean and standard deviation) of each sample sequence, preserving the relative fluctuations in individual payment rhythms while eliminating interference from differences in absolute amounts between groups. For missing values, a mask flag is set, the embedding layer outputs a zero vector, and learnable bias compensation is applied to prevent spurious signals applied by interpolation. Assuming that the observed variable x_t is the raw value at time step t. Its normalized form is:

$$\hat{x}_t = \frac{x_t - \mu_x}{\sigma_x + \epsilon} \tag{2}$$

μ_x and σ_x are the mean and standard deviation of the variable in the training set, and ϵ is a small constant to prevent division by zero. The normalized variables are linearly transformed:

$$\mathbf{z}_t = \mathbf{W}_l \hat{x}_t + \mathbf{b}_l \tag{3}$$

$\mathbf{W}_l$ and $\mathbf{b}_l$ are learnable parameters, and $\mathbf{z}_t$ is the latent representation after projection.

Categorical observation variables, such as notification channel type (SMS, app push, email) and payment terminal device type, are each configured with a separate embedding table and mapped to a latent space of uniform dimensionality. All processed observation variables are stacked into a three-dimensional tensor by time step, with a temporal resolution that strictly matches the target payment completion rate sequence. Normalized numerical variables are mapped to the same latent space dimension using a trainable linear transformation layer, achieving homogeneous representation with the embedding vector. The linear transformation weights are adaptively adjusted during training to impart a discriminative scale to different variables in the initial representation. Ultimately, the static feature tensor, known covariate tensor, and observed covariate tensor are output to the corresponding input ports of the TFT backbone network, forming a hierarchical, semantically decoupled multi-source input structure. This architecture ensures semantic isolation of different types of variables at the model front end, preventing representation confusion caused by information aliasing and providing a clear input foundation for subsequent dynamic selection mechanisms.

Table 1 presents the structured encoding statistics of multi-source heterogeneous data in the university payment prediction framework, covering the dimensional mapping between static attributes and two types of time-varying covariates. This information presents the dimensional transformation of each variable category before and after embedding, the linear projection configuration of numeric variables, and the actual percentage of missing data. This information reflects the input layer's strategy for separating categorical and numeric features, embodies the encoding process's design to address high-dimensional sparsity and data incompleteness, and supports subsequent models for decoupling heterogeneous information.

TABLE 1. Structured coding statistics

Variable Category	Total Input Dimensions	Embedded Vector Dimensions	Numerical Transformation Dimensions	Missing Rate (%)
Static Variables	24	40	-	0.3
Known Time-Varying Covariates	18	14	4	1.7
Observed Time-Varying Covariates	22	22	4	12.4

B. *DYNAMIC SELECTION OF KEY VARIABLES UNDER THE GATING MECHANISM*

Figure 2 illustrates the architecture of a TFT-based gated variable selection network, specifically designed for dynamic multi-source feature selection in predicting university payment trends. The input consists of static attributes and time-varying covariates. After extracting time series features using LSTM, dynamic weights are generated through sigmoid gating, achieving adaptive masking in the feature space. Static context broadcasting is then fused with the filtered time-varying features to retain key drivers and suppress noise. The resulting output is an interpretable spatiotemporal representation, supporting heterogeneous analysis of payment behavior and attribution of policy effects.

1) Hidden State Extraction and Channel Compression of Temporal Representation

After encoding, multiple source variables form a high-dimensional input sequence. The driving force of each variable on payment behavior at different time points exhibits significant time-varying characteristics. To capture these dynamic correlations, a long short-term memory network is used to independently model the historical evolution path of each type of time-varying covariate. Each variable sequence is input into a shared-weight LSTM unit, which recursively updates its hidden state along the time dimension, integrating the complete context from the observation starting point to the current step. The hidden state $\mathbf{h}_t^{(i)}$ of the LSTM unit at time step t is determined by the input $\mathbf{x}_t^{(i)}$ and the previous state $\mathbf{h}_{t-1}^{(i)}$. Its update mechanism is:

$$\mathbf{h}_t^{(i)}, C_t^{(i)} = \mathrm{LSTMCell}\left(\mathbf{x}_t^{(i)}, \mathbf{h}_{t-1}^{(i)}, C_{t-1}^{(i)}\right) \tag{4}$$

$\mathbf{h}_t^{(i)}$ represents the i-th time-varying variable, $C_t^{(i)}$ represents the cell state, and the hidden states of all variables are concatenated into a matrix along the channel dimension.

The hidden state output reflects the cumulative dynamic characteristics of the variable over the course of the sequence, preserving key patterns such as trends, fluctuations, and mutations. The forget gate and input gate within the LSTM structure coordinately regulate information flow, automatically filtering persistent signals and suppressing transient noise perturbations, ensuring a balance between representation stability and responsiveness.

The hidden states of all variables are concatenated horizontally at each time step to form a high-dimensional joint representation matrix. This matrix contains the deep dynamic responses of all input variables at the current moment, but it suffers from information redundancy and channel overload. A learnable linear projection layer is applied to reduce the dimensionality of the joint matrix, mapping the original high-dimensional space into a compact latent space. The projection process is implemented by matrix multiplication, and the weight matrix is continuously optimized during backpropagation to retain the most discriminative combination directions. The reduced dimensionality vector serves as the input source of the gating network, providing a refined information carrier for subsequent weight generation. This operation reduces the computational load while enhancing the compactness of feature representation and preventing interference transmission from irrelevant variables in high-level models.

2) Generation of Dynamic Importance Weights and Variable Screening

The gating unit receives the dimension-reduced joint representation vector and assigns weights element-wise. This unit consists of a multilayer perceptron, using the Sigmoid function as the activation function instead of other nonlinear functions such as ReLU or tanh to ensure that the output value falls within the [0, 1] interval, forming interpretable soft weights that correspond to the relative importance of each variable at the current time step. The network output dimension is consistent with the total number of input variables, and each node corresponds to the weight value of a variable. The Sigmoid nonlinear transformation gives the weight a soft selection characteristic, allowing the model to give different attention to the same variable at different time points, achieving fine-grained control. The following formula calculates the weight vector α_t output by the gating network:

$$\alpha_t = \sigma\left(\mathbf{W}_g \cdot \mathrm{MLP}(\mathbf{H}_t)\right) \tag{5}$$

MLP (Multilayer Perceptron) is a multilayer perceptron, $\mathbf{W}_g$ is the projection weight, and σ is the Sigmoid function. The weighted variables are expressed as:

$$\widetilde{\mathbf{H}}_t = \sum_{i=1}^{N} \alpha_t^{(i)} \cdot \mathbf{h}_t^{(i)} \tag{6}$$

The weight generation process is driven by the global time context and responds to the changing rhythm of the external event sequence. For example, the weight values of "countdown days" and "accumulated number of notifications" are automatically increased as the deadline approaches.

The generated weight vector is multiplied channel-by-channel with the original encoded variable representation to complete weighted fusion. High-weight variables retain their original strength, while low-weight variables are significantly attenuated, equivalent to dynamic masking in the feature space. This operation is performed independently at each time step, forming a variable selection path that evolves over time. The weighted variable sequence is fed into the subsequent modeling module, carrying only the activated semantic information flow. The parameters of the gating mechanism are trained jointly with the entire model, and the gradient of the loss function can be back-propagated to the weight generation network to drive it to learn the optimal selection strategy. During training, the model gradually grasps the variable combinations that should be paid attention to in a specific time context.

Table 2 shows the key parameter configurations of the gated variable selection network. The hidden state dimension of the LSTM unit is set to 64, and the input dimension is 32, which is used to extract univariate time series features. The gated network uses a two-layer MLP (64→32 dimensions) to generate dynamic weights, and the output dimension is consistent with the number of variables (15).

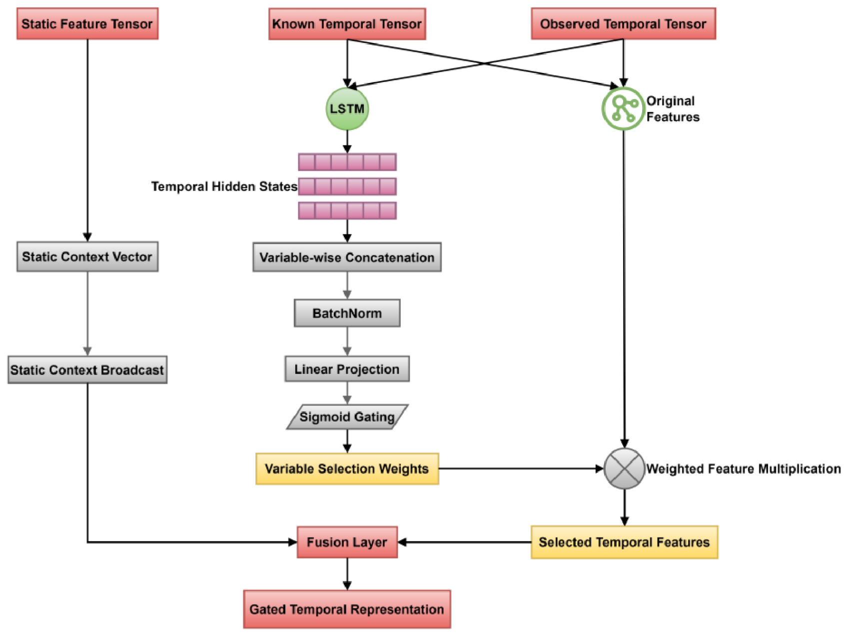

FIGURE 2. Gated variable selection network

The Adam optimizer (learning rate 1e-3) and batch size 128 are used in the training phase. The design implements soft selection through the Sigmoid activation function, ensuring that the model adaptively selects key variables, suppresses noise interference, and improves prediction robustness. The parameter values are verified by grid search to balance computational efficiency and feature selection accuracy.

TABLE 2. Gating variable selection network parameter configuration

Parameter Category	Parameter Name	Value/Dimension	Note
LSTM Unit	Hidden State	64	Per-variable encoding
LSTM Unit	Input Dimension	32	Encoded feature space
Gating Network	MLP Hidden Layers	[64, 32]	Sigmoid activation
Gating Network	Output Dimension	15 (number of variables)	Dynamic weights
Training	Learning Rate	1.00E-03	Adam optimizer
Training	Batch Size	128	Mini-batch training

C. MULTI-HEAD ATTENTION MECHANISM MODELING LONG-TERM DEPENDENCIES

Figure 3 illustrates the modeling process of university tuition payment time series using a multi-head attention mechanism in TFT. Input features are linearly projected to generate a query/key/value matrix. Three specialized attention heads capture the temporal pattern head, group synchronization head, and policy response head, respectively. Static student attributes are injected into each head via positional encoding, ensuring that long-term factors such as major category are included in the calculation. The softmax weight matrix represents attention peaks at key time points. The output layer fuses the multi-head results to generate contextualized features. The attention distribution can be directly mapped to specific policy actions, providing an interpretable basis for payment rate prediction.

1) Modeling the Response of Time-Point Attention to Institutional Nodes

In the sequence-to-sequence architecture, the decoder receives the contextual representation output by the encoder at each prediction time step and selectively focuses on the historical input sequence through the time point attention mechanism. The attention weight is generated by the dot product calculation of the query vector (Query) and the key vector (Key), where the query comes from the current state of the decoder and the key and value are composed of the hidden layer output of the encoder at each time step. The softmax function normalizes the dot product result to generate a probability distribution in the time dimension, indicating the intensity of the model's attention to the historical moment when generating a prediction at a certain step. The time point attention weight is calculated as follows:

$$\mathbf{A}_t = \mathrm{softmax}\left(\frac{\mathbf{Q}_t\mathbf{K}^T}{\sqrt{d_k}}\right) \tag{7}$$

$\mathbf{Q}_t$ is the decoder query; $\mathbf{K}^T$ is the encoder key matrix; T is the historical sequence length; and d_k is the key vector dimension. The resulting context vector is then generated. This distribution does not evenly cover the entire historical window, but rather concentrates on key time points, such as the beginning of the semester, the date scholarships and grants are awarded, and seven days before the tuition payment deadline.

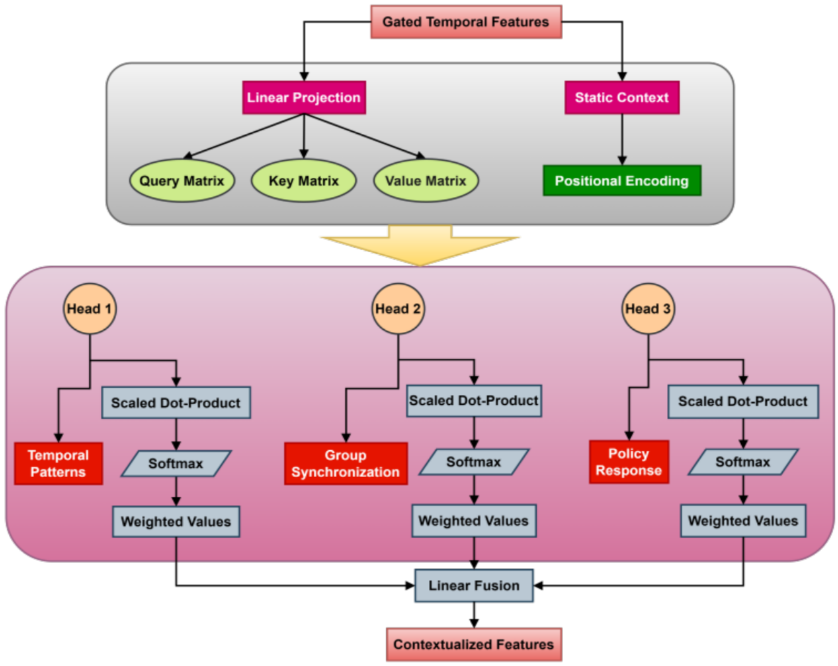

FIGURE 3. Modeling of university payment time series using the multi-head attention mechanism

The attention head scans patterns within the time span, identifying the historical segments most relevant to the current predicted position. The multilayered attention stack allows different layers to capture temporal dependencies at varying granularities, focusing shallowly on local fluctuations while integrating cross-period trends at deeper levels. Each layer's attention output is a weighted summed value vector, preserving the semantic information of moments of high attention while suppressing interference from irrelevant periods. This mechanism enables the model to accurately locate key intervention points in long sequences that influence the current payment rhythm, avoiding signal attenuation or misalignment caused by extended time spans.

2) Capturing Group Interaction Patterns with Multi-Head Self-Attention

Based on variable representation, a multi-head self-attention structure is applied to reveal the co-evolution of tuition payment behavior within student groups. This mechanism takes the encoded variable sequence as input and generates three matrices: query, key, and value through linear transformation. Attention operations are then performed in parallel in multiple independent subspaces. Each attention head uses different projection parameters to learn specific types of dependencies. For example, some heads focus on the delayed aggregation of economically disadvantaged groups, while others identify the synchronous upward trend among majors with high tuition rates. The outputs of each head are concatenated and fused through a linear layer to form a comprehensive representation, enhancing the model's ability to express complex interaction patterns. The self-attention mechanism calculates the similarity between individual representations and constructs an implicit association graph. When a certain group of students (such as those receiving scholarships) collectively pays their fees at a specific time, the corresponding time step triggers a high-attention response in multiple heads, forming a cross-sample pattern resonance. The model thus identifies the initiation nodes and diffusion paths of collective behavior.

The multi-head structure improves the model's ability to discern heterogeneous response patterns. Different heads spontaneously divide the work during training, with some focusing on capturing behavioral synchronization driven by temporal proximity and others identifying cross-group resonance driven by policy variables. The attention weight matrix is expanded across both time and variables to form a parseable behavioral response heat map.

D. JOINT OUTPUT OF PREDICTION RESULTS AND ATTENTION WEIGHTS

1) Synchronous Generation of Multi-Grained Prediction Values and Attention Streams

During the decoding phase, the model generates a time-step-by-time sequence of predictions for future payment completion rates, with output granularity covering multiple management periods, including the short, medium, and long term. Each prediction point is converted from the decoder's final hidden state via a linear mapping to a scalar value, representing the proportion of students expected to complete payment within that period. The prediction process employs a rolling decoding strategy, with the output of the previous time step serving as the input feature for the subsequent time step to maintain sequence continuity. Simultaneously, the time-point attention module within the decoder continuously records the contribution strength of each historical moment to the current prediction, forming a two-dimensional weight matrix for each time-history step. Let the query vector of the decoder at the prediction time step τ be $\mathbf{q}_\tau^T$, and the key vector of the encoder at the historical time step t be $\mathbf{k}_t$, then the attention weight at the time point is calculated as:

$$\alpha_{\tau,t} = \frac{\exp\left(\mathbf{q}_\tau^T \mathbf{k}_t / \sqrt{d_h}\right)}{\sum_{t'=1}^{T} \exp\left(\mathbf{q}_\tau^T \mathbf{k}_{t'} / \sqrt{d_h}\right)} \tag{8}$$

T is the length of the historical sequence, and $\alpha_{\tau,t}$ and τ represent the attention intensity of the t-th historical moment when the model generates the prediction at the τ-th step, which forms the attention distribution of the time dimension after normalization. The attention paths of static variables and dynamic covariates are extracted through a cross-layer tracking mechanism. The attention distribution of static features is derived from the interaction weights between the static encoding branch and the decoder state, reflecting the strength of the role of inherent attributes such as grade and major category in determining long-term trends. The attention of dynamic variables is determined by the degree of match between the time-varying input and the decoder query vector, distinguishing the driving differences between known covariates (such as scholarship programs) and observed variables (such as notification frequency) at different stages. After training, all attention streams are fixed as exportable numerical matrices and stored in conjunction with the prediction results to form a complete output package.

2) Inverse Mapping from Weights to Feature Space and Semantic Restoration

The attention weight matrix is parsed by the post-processing module and mapped back to the semantic level of the original input features. This process uses feature alignment restoration technology to reverse-link the weight distribution results in the high-dimensional latent space to specific input fields. For example, in a prediction step, the attention value corresponding to "Countdown Days" is 0.73, and "Accumulated SMS Reminders" is 0.61. The system automatically identifies the dominant role of these two fields in the current decision and annotates their path identifiers in the original data table. The weights of categorical variables are aggregated to their original classification level. Let the average attention intensity of variable v over all historical time steps be defined as:

$$\bar{\beta}_v = \frac{1}{T} \sum_{t=1}^{T} \gamma_v^{(t)} \tag{9}$$

$\gamma_v^{(t)}$ represents the cross-head attention weight of variable v at time step t (after multi-head fusion). $\bar{\beta}_v$ serves as a comprehensive importance indicator for the variable across the entire input sequence, which is used to rank and generate readable explanatory statements.

The restored weights are visualized as a graph, with the horizontal axis representing the prediction time point and the vertical axis representing the input variable. Color intensity indicates attention intensity. The graph reveals the evolution of attention patterns for specific variables over time. For example, "Award Arrival Status" remains consistently highlighted for three time steps after the award date, while "Online Payment Failure Marker" peaks during periods of system anomalies. These patterns directly correspond to the time anchors of management actions, forming an explanatory chain aligned with actual operation logs. The system further generates natural language summaries, converting key weights into readable sentences.

3) Structural Encapsulation and Interface Exposure of Interpretive Output

The model output module integrates a structured package of predicted values, confidence intervals, and attention maps, organized in a hierarchical JSON format to support front-end system calls and visual rendering. Each prediction record includes a timestamp, predicted value, error estimate, and a list of the main contributing variables and their weights. The main contributing variables are extracted using a threshold filtering mechanism, retaining only the top three weighted input fields to avoid information overload. The system provides an Application Programming Interface (API), allowing managers to query the decision-making basis of a specific prediction at a specific point in time or track the impact of each variable across the entire prediction sequence.

This output architecture ensures external auditability of the forecast logic. Financial managers can click on a specific day's forecast results through the interactive interface to instantly view the key inputs the model relies on and their relative importance rankings. When a sudden change in the forecast occurs, the system automatically labels the core variables driving the change, assisting in determining whether it represents a data anomaly, a policy adjustment, or a model misjudgment. The entire interpretation process is fully automated, requiring no manual intervention or post-processing modeling, ensuring that interpretation results are generated synchronously with the forecasted actions and are consistent and reliable. This mechanism transforms the internal reasoning process of the deep learning model into actionable management signals, supporting closed-loop decision-making from trend prediction to strategy adjustment.

IV. COMPARISON OF PREDICTION PERFORMANCE OF TFT FRAMEWORK IN DIFFERENT SITUATIONS

A. DYNAMIC FEATURE IMPORTANCE OVER TIME

The trained Temporal Fusion Transformer model performs forward inference on the entire time series. At each forecast time step, the feature weight vector output by the gated variable selection network is extracted. This vector contains the normalized importance scores of each input variable at the current moment. A complete weight time series matrix is collected by traversing all historical and forecast time steps. The weights are then aligned and smoothed to ensure that the time nodes are synchronized with actual management events. Finally, the weights of the six categories of variables are visualized over time, forming a dynamic feature importance evolution graph.

Figure 4 shows the dynamic weighting of six variables by the gating mechanism over the 15-week payment cycle, revealing the evolution of the dominant factors in university payment behavior over time. Initially (weeks 1–4), economic status is weighted the highest, reflecting the tendency of these students to pay early to avoid risk and demonstrating their high sensitivity to financial pressure. After awards and grants arrive in weeks 5–6, the weighting of award and grant disbursement briefly increases, but is quickly overwhelmed by reminders to pay. This suggests that while policy support has a priming effect, it lacks sustained driving force. SMS reminders saw three weight peaks (0.50, 0.52, and 0.55) in weeks 6, 9, and 12, demonstrating a concentrated notification rhythm, indicating that information delivery is a key external intervention in maintaining payment. The payment channel status surged to 0.65 in week 8, indicating a system failure causing widespread congestion. The model immediately shifted its focus, demonstrating its ability to rapidly respond to unusual events. As the deadline approaches, the weight of the number of days until the deadline jumps to 0.78, highlighting the strong constraints imposed by institutional timelines on behavior. Students generally prioritize their actions due to time pressure. The weight of major category remains extremely low (≤ 0.04), indicating that its contribution as a static attribute in dynamic prediction is limited. The overall trend suggests that payment behavior is driven by multiple factors. The gating mechanism successfully implements time-varying variable selection, avoids noise interference, and improves predictive accuracy and explanatory power.

B. PREDICTIONS ACROSS MODEL VARIABLES AND STUDENT SUBGROUPS

The trained TFT model is loaded, and forward inference is performed on the test set to generate a series of predicted payment completion rates for the next 14 days. The output of the comparison model under the same input is extracted, and the five predicted curves are aligned with the actual payment rates to ensure consistency in time steps and observation granularity. During the group stratification assessment phase, the sample is divided into three categories (Hard, Medium, and Rich) based on household economic status. The actual and predicted average completion rates for each group are calculated at the daily time granularity, preserving the original sequence alignment. Dotted and solid lines are drawn point by point to reflect the dynamic trajectory. All curves are rendered using the same coordinate axis range and style parameters to highlight the differences in trend patterns. Figure 5 shows the performance differences between the proposed TFT model and other comparison models in predicting trends in university tuition payment completion rates. The horizontal axis represents the prediction time step (days), and the vertical axis represents the tuition payment completion rate. The TFT model's prediction curve closely matches the true path and accurately tracks all change points, demonstrating its keen responsiveness to institutional milestones. In contrast, removing the gating mechanism or using only LSTM models exhibits significant lags, reflecting deficiencies in modeling dynamic variable selection and long-term dependencies. While the static weight model maintains a stable upward trend, it lacks a flexible adjustment mechanism and is unable to capture nonlinear transitions at key time points, resulting in a prediction error greater than 0.1. The differences in payment behavior among the three student groups and the model's predictions reveal that the economically disadvantaged group exhibits slow progress overall. While the model can accurately capture their lag characteristics, there is some error. The affluent group actively pays, and the predicted curve closely follows the true value. The clear separation between the different groups demonstrates that the model effectively captures attribute heterogeneity. Overall, the high consistency between the solid and dashed lines demonstrates the robust generalization of TFT across multiple subgroups. It not only reproduces overall trends but also accurately characterizes the tempo of group responses, providing a reliable basis for stratified interventions.

C. COMPARISON OF PREDICTION PERFORMANCE DURING THE PEAK PAYMENT PERIOD AT THE BEGINNING OF THE SEMESTER

In the performance evaluation of predictions during the peak payment period, it uses Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as quantitative metrics to measure the model's short- to medium-term prediction error for future payment completion rates. The evaluation covers four time granularities: 7 days, 14 days, 30 days, and 60 days. Each model performs rolling forecasts on the same historical sliding window, outputting a multi-step forward forecast sequence. Error calculations are based on a point-by-point comparison of true and predicted values on the test set. TFT is compared with ARIMA, Informer, Autoformer, XGBoost, and GRU. All model inputs are covariate aligned to ensure consistency of external information. Error statistics are calculated as the mean and standard deviation of five independent training runs.

Figure 6 compares the multi-step prediction errors of TFT and five other models during the peak payment period at the beginning of the semester, using RMSE and MAE as evaluation metrics. TFT performs best across all prediction timeframes, with RMSE below 0.08 and MAE below 0.07. This significant advantage is attributed to its ability to accurately capture institutional nodes. The attention mechanism enables the model to focus on key time anchors, effectively modeling nonlinear responses. While the errors of each model increase with increasing prediction

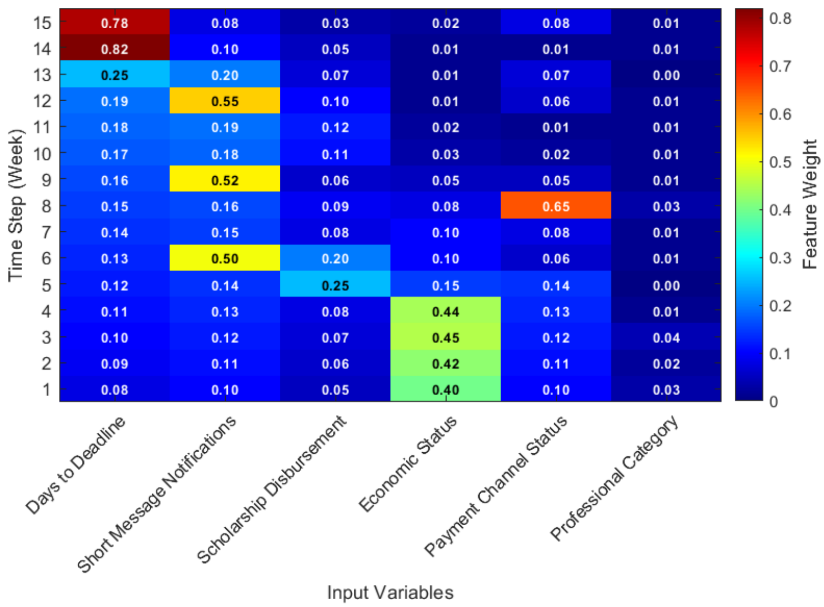

FIGURE 4. Dynamic feature importance (gate output weights) over time

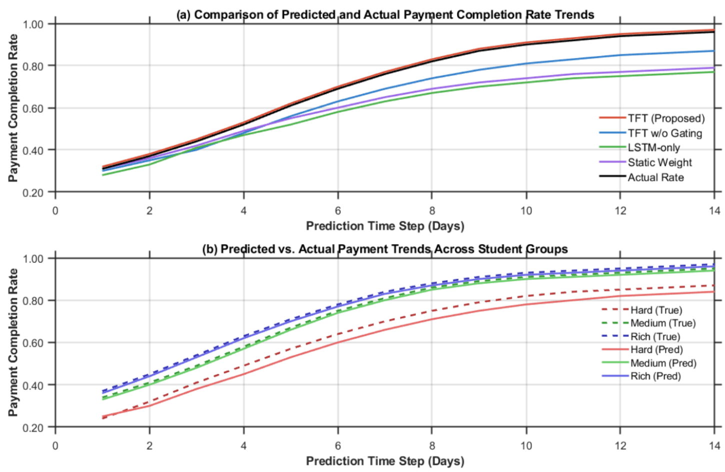

FIGURE 5. Predicted and Actual Payment Completion Trends Across Model Variables and Student Subgroups

step size, TFT exhibits a lower error, demonstrating its stability in modeling long-term dependencies. ARIMA and XGBoost lack the ability to incorporate dynamic covariates, resulting in consistently high errors. While GRU possesses temporal memory, it lacks responsiveness to multivariate interactions. Informer and Autoformer rely on sparse attention, making it difficult to distinguish core driving factors in high-dimensional, heterogeneous inputs. TFT suppresses redundant information through gated variable selection, retaining the highly weighted responses of key variables and forming a more discriminative representation path. The error bars reflect the stability of the model's predictions, and TFT's low standard deviation indicates good convergence in its training process and highly reproducible results. This performance validates the effectiveness of structured coding and dynamic weight allocation in complex educational time series scenarios.

D. DYNAMIC ADAPTABILITY ASSESSMENT UNDER MULTI-SOURCE DISTURBANCE

To assess dynamic adaptability in multi-source disturbance scenarios, five typical sudden events are designed: scholarship and grant disbursement delays, deadline changes, notices of partial invoicing, tuition adjustments, and payment failures. Each disturbance is injected into the original data stream at a deterministic time point, simulating structural shifts in real-world management scenarios. The disturbance amplitude is set based on historical fluctuations to ensure compliance with university financial operational standards. The model performs rolling forecasts within a 15-day window before and after the disturbance. MAPE (Mean Absolute Percentage Error) calculates the mean relative

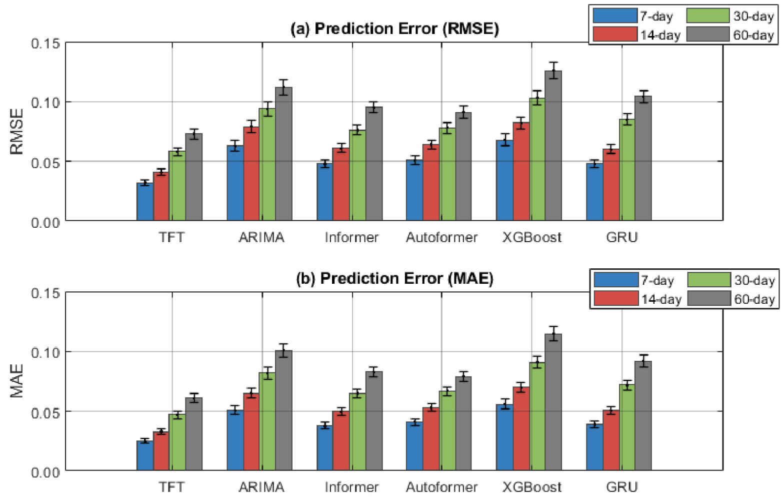

FIGURE 6. Comparison of prediction errors

deviation between the predicted and actual values seven days after the sudden change point, and R^2 measures the overall goodness of fit of the predicted series to the true trend. All models use the same historical training set, and no parameter updates are performed after the disturbance injection, rigorously testing their extrapolation robustness. The evaluation process is repeated five times to eliminate randomness, and the final indicators are averaged to ensure the reliability of the results.

Figure 7 shows the dynamic adaptability of six models under five typical types of regulatory disturbances, using MAPE and R^2 as dual-dimensional evaluation metrics. TFT exhibits the lowest forecast error and highest trend fit across all disturbance scenarios, with an MAPE no higher than 10.3% and an R^2 above 0.9, demonstrating its robustness to structural changes. ARIMA errors increase significantly across all events, reflecting the difficulty of linear models in capturing the nonlinear responses triggered by institutional interventions. While the XGBoost model is somewhat effective at integrating static features, it exhibits a delayed response at sequence mutation points, attributed to its inherent limitations in modeling temporal dependencies. Informer and Autoformer models have limited adaptability in different event-driven scenarios, and their attention mechanisms fail to fully activate short-term response pathways. GRU's memory mechanism is easily overwritten by noise under multivariate coupling interference. TFT, leveraging its gated variable selection and temporal attention synergy, can quickly identify key drivers and adjust weight distribution after a disturbance occurs, enabling accurate capture of mutation signals and trend reconstruction, demonstrating its deep modeling capabilities for the complex dynamics of university payment systems.

E. VERIFICATION OF FEATURE ATTRIBUTION CONSISTENCY AND FEATURE DISCRIMINATION ABILITY IN MULTIVARIABLE COUPLING SCENARIO

In a multivariate coupling scenario, five key driving variables are selected: Days to Deadline, Short Message Notifications, Scholarship Disbursement, Economic Status, and Payment Channel Status. The model's consistent identification of variable contributions is evaluated. Actual payment behavior sequences are used as the response variable. The Pearson correlation coefficient (Pearson r) is calculated between the attention weights or feature importance scores output by each model and the actual behavior to measure its dynamic attribution capability. Information Gain Radio (IG Radio) is calculated using the decision tree splitting criterion to assess the discriminative contribution of variables in the prediction path. All metrics are calculated on the test set using a sliding time window and averaged to improve stability. The focus model is evaluated to determine whether it appropriately weighted key variables during high-impact periods to avoid attribution drift.

Figure 8 compares the consistency of feature attribution and discriminative power of various models in a multivariate coupling scenario. The Pearson correlation coefficient (Pearson r) reflects the model's ability to dynamically identify key drivers. TFT exhibits significantly high correlations in "Days to Deadline" and "Short Message Notifications," demonstrating that its time-point attention mechanism accurately captures the response intensity of institutional nodes and intervention behaviors. The overall Pearson correlation coefficient is no less than 0.68. The information gain ratio shows that TFT maintains high discriminative power across different attributes, indicating that its gated variable selection network effectively incorporates individual heterogeneous information. The overall information gain ratio is no less than 0.7. In contrast, traditional models such as ARIMA and XGBoost have weak attribution capabilities for dynamic variables and struggle to distinguish core variables

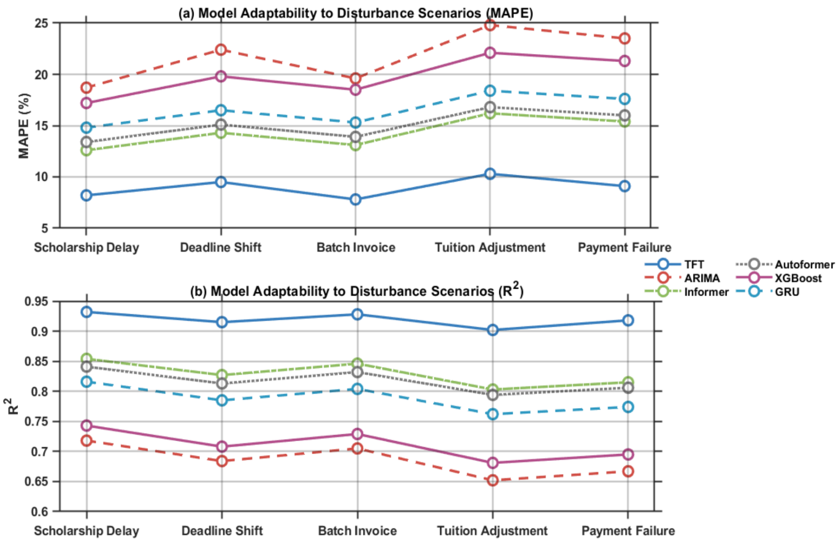

FIGURE 7. Dynamic adaptability under multi-source disturbances

from background noise. While Informer and Autoformer possess attention structures, they fail to model the rule-dependent constraints in educational management, resulting in low variable weights. GRU lacks an explicit feature selection mechanism, resulting in significant volatility in attribution results. Through structured coding and dynamic weight allocation, TFT achieves the unification of mechanism perception and attribution stability in complex coupled environments. Its output reflects not only prediction accuracy but also the ability to deeply model the logic behind university payment behavior.

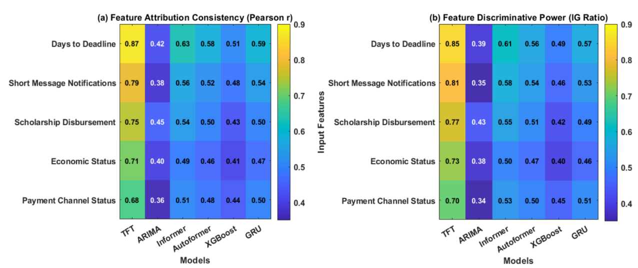

FIGURE 8. Feature attribution consistency and feature discrimination ability

F. TESTING THE PREDICTION STABILITY ACROSS DIFFERENT STUDENT GROUPS

To assess the model's predictive stability across different student groups, the sample is divided into Hard (financially difficult), Medium (average), and Rich (financially comfortable) groups based on their families' financial status. The average prediction error rate for each model is calculated independently for each subgroup. This error rate is defined as the ratio of the absolute value of the difference between the predicted and actual payment completion rates to the actual value, reflecting the degree of systematic deviation. The cross-period variance of the prediction error is also calculated to measure the model's long-term volatility stability.

Table 3 shows the predictive stability of six prediction models across student groups with varying economic status, spanning three subgroups: hard, medium, and rich. TFT maintains the lowest mean bias and smallest bias variance across all subgroups, ranging from 3.87% to 4.12% and 0.0015 to 0.0018, respectively, demonstrating its high adaptability to heterogeneous response patterns. Students with financial difficulties are often more sensitive to notification delays and changes in financial aid, and their behavior fluctuates significantly. Traditional models exhibit significant bias amplification and increased variance in this subgroup, reflecting their inability to model nonlinear responses to interventions. TFT dynamically focuses on key drivers through a gated variable selection mechanism, increasing the weighting of notification reach and

status among economically constrained groups, thereby suppressing noise interference. TFT maintains a balanced distribution of biases across the three groups, demonstrating its robust generalization and robustness in complex education management scenarios, effectively addressing the behavioral heterogeneity challenges posed by socioeconomic disparities.

TABLE 3. Prediction stability performance

Model	Group	Mean Deviation Rate (%)	Deviation Variance
TFT	Hard	4.12	0.0018
TFT	Medium	3.87	0.0015
TFT	Rich	3.95	0.0016
ARIMA	Hard	9.63	0.0082
ARIMA	Medium	7.41	0.0054
ARIMA	Rich	6.88	0.0047
Informer	Hard	6.75	0.0043
Informer	Medium	5.82	0.0036
Informer	Rich	5.39	0.0031
Autoformer	Hard	6.51	0.004
Autoformer	Medium	5.63	0.0034
Autoformer	Rich	5.24	0.0029
XGBoost	Hard	8.94	0.0075
XGBoost	Medium	7.26	0.0051
XGBoost	Rich	6.67	0.0044
GRU	Hard	7.18	0.0048
GRU	Medium	6.03	0.0038
GRU	Rich	5.72	0.0035

V. CONCLUSION

This paper addresses the common challenge in university payment trend forecasting—the difficulty in capturing trend fluctuations and the lack of interpretability in results—both of which are caused by the intricate coupling of multiple heterogeneous factors. This situation is highly similar to predicting market trends in the industry, where forecasting success depends on accurately disentangling the combined effects of varied factors like global raw material costs, shifting fashion cycles, and economic policies. It proposes a reconstruction framework based on the Temporal Fusion Transformer. This architecture achieves semantic decoupling of static attributes, known covariates, and observed variables through a hierarchical embedding mechanism, establishing a structured representation foundation at the input. A gated variable selection network autonomously identifies key driving factors in the temporal dimension, suppressing redundant information interference and enhancing the model's responsiveness to policy adjustments and emergencies. LSTM and multi-head attention collaborate to model long-term dependencies and time node effects, accurately capturing institutionally driven behavioral patterns. The attention weight distribution output by the decoder is mapped back to the original feature space, generating a traceable attribution path that reveals the management motivations behind the prediction results. Experimental validation encompasses dimensions such as prediction accuracy, perturbation adaptability, population stability, and feature interpretability. Results demonstrate that TFT significantly outperforms benchmark models such as ARIMA, XGBoost, GRU, and Informer in multi-step predictions during peak payment periods, achieving RMSE below 0.08 and MAE below 0.07, demonstrating outstanding error control capabilities. Under five types of disturbance scenarios, the MAPE remains below 10.3%, and the R^2 remains above 0.9, demonstrating stable model fitting performance and strong robustness. This research achieves a synergistic improvement in forecasting accuracy and stability, providing a high-performance and interpretable technical path for university financial systems, which supports crucial functions like capital flow prediction, risk warning, and intervention strategy optimization. This advancement offers insights that are equally valuable to the industry, where similar interpretable and stable predictive tools could be deployed to support supply chain finance, predict demand fluctuations, and optimize inventory investment strategies.

AUTHOR CONTRIBUTIONS

Conceptualization – Zhiqun Zhang and Yaya Wang; methodology – Zhiqun Zhang and Yaya Wang; investigation – Zhiqun Zhang and Yaya Wang; writing-original draft preparation – Zhiqun Zhang and Yaya Wang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. G. Marbun, J. A. Saputro, and H. Subiyakto, “Intention to use Virtual Account-based tuition payment system with UTAUT Model and Computer Self Efficacy: Initial Trust as Moderator,” IOSR Journal of Business and Management (IOSR-JBM), vol. 23, no. 10, pp. 9-18, 2021.

[2] A. Keykha, “Tuition Fees and Academic Decisions: Unpacking the Impact on the Students of Tehran University's Management Faculties,” Interdisciplinary Journal of Management Studies (Formerly known as Iranian Journal of Management Studies), vol. 17, no. 3, pp. 781-797, 2024, doi: 10.22059/ijms.2023.362881.676074.

[3] Z. Zafari, L. Goldman, K. Kovrizhkin, and P. Muennig, “Willingness-to-pay tuition and risk-taking proclivities among public health students,” Journal of American College Health, vol. 71, no. 9, pp. 2705-2710, 2023, doi: 10.1080/07448481.2021.1987249.

[4] L. Wulansari, A. Iskandar, S. Syahrul, M. Mansyur, M. Ramazanova, and D. Saputri, “Web-Based Tuition Payment Information System Design Using Codeigniter at SMA Negeri 2 Bungku,” Ceddi Journal of Education, vol. 4, no. 1, pp. 21-30, 2025, doi: 10.56134/cje.v4i1.138.

[5] R. R. Muis, S. N. Kemal, R. Rosantia, D. Cahyani, and M. Ali, “The Effect of Implementing the Smart Tihamah Application on the Level of Savings Participation and Ease of Payment of Education Fees (SPP) for MA Tihamah Putra Students,” Strata International Journal of Social Issues, vol. 2, no. 2, pp. 219-226, 2025.

[6] R. Stoyanova and S. Goranova, “The influence of a tuition fee increase on the drop-out rate of the nursing program,” Journal of Economy Culture and Society, vol. (63), pp. 55-66, 2021, doi: 10.26650/JECS2020-0013.

[7] D. Susilo, M. D. Nugroho, and D. Ruswanti, “Student Tuition Fee Management Using Web-Based Application System,” Tepian, vol. 4, no. 2, pp. 80-88, 2023, doi: 10.51967/tepian.v4i2.2643.

[8] M. Zainul, M. Ma'rifah, and S. Shaddiq, “The effect of promotion and tuition costs on the determination of higher options,” JPPI (Jurnal Penelitian Pendidikan Indonesia), vol. 9, no. 4, pp. 394-408, 2023, doi: 10.29210/020233123.

[9] S. Tjay, W. Dewi, and T. Widodo, “The Effect of Education Quality and Tuition Fees on School Selection Decisions Through Mediation of School Image at Citra Bangsa School Tangerang,” Indonesian Interdisciplinary Journal of Sharia Economics (IIJSE), vol. 8, no. 2, pp. 5133-5148, 2025, doi: 10.54371/jiip.v8i4.7556.

[10] A. Z. Wahyudi, A. F. S. Alfikri, and A. N. Sumidartini, "The Use of Paylater as a Single Tuition Payment Method by College Students in the Perspective of Maslahah," Perbanas Journal of Islamic Economics and Business, vol. 5, no. 2, pp. 145-153, 2025, doi: 10.56174/pjieb.v5i2.314.
[11] A. Miftahuddin, "University Students' Adoption of Cashless Payments in Uzbekistan: Behavior, Trust, and Challenges," Innovation Science and Technology, vol. 1, no. 5, pp. 33-42, 2025.
[12] H. Holly Wang, Y. Hua, and C. Wilson, "Do college students demand lower tuition for online learning? Empirical evidence from choice experiments during COVID-19," Education Economics, vol. 33, no. 2, pp. 293-310, 2025, doi: 10.1080/09645292.2024.2318215.
[13] K. Shanmugasamy, "The Future of Transactions: Exploring the Impact and Applications of Intelligent Payment Systems," International Journal of Science and Research (IJSR), vol. 13, no. 9, pp. 158-161, 2024, doi: 10.21275/SR24901223954.
[14] C. E. A. Ikegwuonu and S. Santa-Ramirez, "How Female Undergraduate Students' Holistic Experiences Predict Their Payment Methods," Youth, vol. 4, no. 2, pp. 759-786, 2024, doi: 10.3390/youth4020051.
[15] M. Tabrani, "Implementation of Prototype Method in School Payment Information System of SMP AL-Mushlis Karawang," Jurnal Teknologi dan Open Source, vol. 5, no. 1, pp. 64-72, 2022, doi: 10.36378/jtos.v5i1.2234.
[16] M. A. N. Agolmen, "Analysis of the Influence of Student Perceptions on the Use of Bri Mobile (Brimo) in Registering Tuition Fees," Winter Journal: Imwi Student Research Journal, vol. 5, no. 2, pp. 174-186, 2024, doi: 10.52851/wt.v4i2.59.
[17] H. Humphrey Muteru, F. H. Gatobu, and H. N. Munene, "Digital Payment Convenience and Student Personal Financial Management: The Tap-To-Pay Trap," Journal of African Interdisciplinary Studies, vol. 9, no. 6, pp. 52-60, 2025.
[18] K. T. Kim and J. Lee, "Financial well-being, anxiety and payment delinquency among student loan holders in the United States: Insights from the COVID-19 pandemic," International Journal of Bank Marketing, vol. 43, no. 2, pp. 424-445, 2025, doi: 10.1108/IJBM-05-2024-0277.
[19] S. Bouabdallah, "What beyond massification? Forecasting the Future of Enrollment in Algerian Higher Education Using ARIMA Approach," Journal of Development Research and Studies, vol. 10, no. 01, pp. 557-571, 2023.
[20] C. C. Chen, Y. Chen, and C. Y. Kang, "Estimating student-teacher ratio with ARIMA for primary education in fluctuating enrollment," ICIC Exp. Lett., Part B, Appl, vol. 12, no. 6, pp. 1-19, 2021.
[21] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interactive Learning Environments, vol. 31, no. 6, pp. 3360-3379, 2023, doi: 10.1080/10494820.2021.1928235.
[22] S. Tang and Z. Li, "Considerations in using XGBoost models with SHAP credit assignment to calculate student growth percentiles," Psychological Test and Assessment Modeling, vol. 64, no. 4, pp. 445-470, 2022.
[23] S. Lakshmi and C. P. Maheswaran, "Effective deep learning based grade prediction system using gated recurrent unit (GRU) with feature optimization using analysis of variance (ANOVA)," Automatika: Časopis za Automatiku, Mjerenje, Elektroniku, Računarstvo i Komunikacije, vol. 65, no. 2, pp. 425-440, 2024, doi: 10.1080/00051144.2023.2296790.
[24] A. Kumar and N. Sachdeva, "A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media," World Wide Web, vol. 25, no. 4, pp. 1537-1550, 2022, doi: 10.1007/s11280-021-00920-4.
[25] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, "A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance," Education and Information Technologies, vol. 29, no. 11, pp. 14365-14401, 2024, doi: 10.1007/s10639-023-12394-0.
[26] Z. Quan, J. Li, and W. Song, "An Improved LSTM Based Early Warning Model for Physical Education Network Teaching Achievements," Journal of Information Processing Systems, vol. 20, no. 6, pp. 793-800, 2024, doi: 10.3745/JIPS.04.0328.
[27] X. Fan and Y. Li, "An Improved Informer Network for Short-Term Electric Load Forecasting," Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), vol. 16, no. 5, pp. 532-540, 2023, doi: 10.2174/2352096516666230217113610.
[28] M. Liu, W. Wang, X. Hu, Y. Fu, F. Xu, and X. Miao, "Multivariate long-time series traffic passenger flow prediction using causal convolutional sparse self-attention MTS-Informer," Neural Computing and Applications, vol. 35, no. 34, pp. 24207-24223, 2023, doi: 10.1007/s00521-023-09003-z.
[29] S. Yu, J. Peng, Y. Ge, X. Yu, F. Ding, S. Li, et al., "A traffic state prediction method based on spatial–temporal data mining of floating car data by using autoformer architecture," Computer-Aided Civil and Infrastructure Engineering, vol. 39, no. 18, pp. 2774-2787, 2024, doi: 10.1111/mice.13179.
[30] L. Zhao, Z. Wu, and M. Du, "Improved Autuoformer Electricity load forecasting based on model fusion," Advances in Engineering Technology Research, vol. 11, no. 1, pp. 8-8, 2024, doi: 10.56028/aetr.11.1.8.2024.
[31] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748-1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.
[32] K. YEMETS, "Time series forecasting model for solving cold start problem via temporal fusion transformer," Computer Systems and Information Technologies, vol. (1), pp. 57-64, 2024, doi: 10.31891/csit-2024-1-7.
[33] G. Fayer, L. Lima, F. Miranda, J. Santos, R. Campos, V. Bignoto, et al., "A temporal fusion transformer deep learning model for long-term streamflow forecasting: A case study in the Funil Reservoir, Southeast Brazil," Knowledge-Based Engineering and Sciences, vol. 4, no. 2, pp. 73-88, 2023.
[34] B. Wu, L. Wang, and Y. R. Zeng, "Interpretable tourism demand forecasting with temporal fusion transformers amid COVID-19," Applied Intelligence, vol. 53, no. 11, pp. 14493-14514, 2023, doi: 10.1007/s10489-022-04254-0.
[35] M. Cid Montoya, A. Mishra, S. Verma, O. A. Mures, and C. E. Rubio-Medrano, "Aeroelastic force prediction via temporal fusion transformers," Computer-Aided Civil and Infrastructure Engineering, vol. 40, no. 15, pp. 2098-2129, 2025, doi: 10.1111/mice.13381.