Time-Consistent Prediction in Higher Education:A Framework for Preventing Data Leakage in Longitudinal Models

Tibor Fauszt *

Posted Date: 5 May 2026

doi: 10.20944/preprints202605.0186.v1

Keywords: learning analytics; dropout prediction; predictive modeling; data leakage; identity leakage; temporal modeling; longitudinal data; model evaluation; predictive validity

 doi:10.20944/preprints202605.0186.v1

Article

Time-Consistent Prediction in Higher Education: A Framework for Preventing Data Leakage in Longitudinal Models

Tibor Fauszt

Institute of Information Technology, University of Dunaújváros, Dunaújváros, Hungary; fausztt@uniduna.hu

Abstract

Data leakage represents a critical methodological challenge in machine learning–based predictive modeling, as it can inflate performance estimates and lead to misleading interpretations. In higher education contexts, where predictive models increasingly support institutional decision-making, the temporal and structural conditions under which predictions are generated and evaluated are often insufficiently specified. This study conceptualizes predictive modeling as a temporally formalized decision task and identifies four core design conditions: explicit specification of the prediction cutoff, temporal restriction of the information set, consistent definition of the at-risk population, and temporally coherent validation. The empirical analysis combines a structured review of recent dropout prediction studies with a controlled experimental demonstration based on longitudinal student data. The review shows that the joint formalization of these conditions remains uncommon, with many models relying on retrospective and temporally unspecified configurations. The experimental results demonstrate that improper validation in longitudinal data structures can produce systematic performance inflation, particularly through identity leakage, and that models with higher representational capacity exploit such leakage more effectively. These findings indicate that predictive performance cannot be interpreted independently of the temporal and structural definition of the prediction task. The proposed framework provides a methodological basis for evaluating predictive models in higher education and other domains where decisions depend on temporally grounded predictions.

Keywords: learning analytics; dropout prediction; predictive modeling; data leakage; identity leakage; temporal modeling; longitudinal data; model evaluation; predictive validity

1. Introduction

Machine learning–based predictive models have become widely used in a range of application domains, including Learning Analytics and Educational Data Mining, particularly in applications targeting early dropout prediction [1–4]. Numerous studies report high predictive performance across diverse data structures, feature sets, and classification algorithms [5–9]. The interpretability of these performance indicators depends on whether the modeling environment accurately reflects the temporal and decision context in which predictions are intended to be deployed, particularly with respect to the definition of the prediction task and the information available at the time of prediction.

Over the past decade, methodological research in predictive modeling has evolved from the early identification of data leakage [10] to broader concerns regarding reproducibility in machine learning research [11]. Increasingly, scholars have emphasized that high performance metrics alone do not establish predictive validity [12]. Reported results commonly rely on train–test splits or cross-validation procedures conducted within a single dataset. Although internal validation is methodologically accepted, its validity depends on strict enforcement of informational and temporal separation. When such separation is compromised, the training or evaluation process may incorporate information that would be unavailable at the time of real-world decision-making,

 doi:10.20944/preprints202605.0186.v1

including signals that reflect the post-outcome state of the target variable [10]. This phenomenon is referred to as data leakage and is widely recognized as a systemic issue in machine learning practice [10–16]. Such distortions often manifest as statistical optimism: models demonstrate strong performance within evaluation settings while exhibiting substantially reduced predictive performance under real-world deployment conditions [12,17–19]. However, these discussions typically treat leakage as a technical validation issue, while overlooking its structural origin in the specification of prediction tasks.

While methodological analyses of data leakage have gained increasing attention in the broader machine learning discourse, systematic empirical investigations within longitudinal predictive modeling remain limited, particularly in higher education contexts. Existing studies typically address leakage as an isolated validation issue, often focusing on distortions arising from improper train–test splits [20]. Fewer contributions examine leakage in relation to the temporal interpretation of prediction tasks and the structural characteristics of longitudinal data representation [4,21]. As a result, the relationship between data leakage and the temporal structure of prediction tasks remains insufficiently understood.

Dropout in higher education provides a clear example of a temporally extended and dynamic process in which relevant predictive information becomes available at different points in time. These characteristics require a longitudinal data representation. However, a widely adopted practice in the literature involves cross-sectional "flattening" of longitudinal data, often without explicit specification of the prediction time point [22]. Although this approach is not inherently incorrect, the absence of temporal anchoring shifts the modeling objective from forward-looking risk estimation toward retrospective separability of already realized outcomes [23,24]. As a result, model performance reflects retrospective separability rather than forward-looking predictive validity.

Different data representations are associated with distinct forms of leakage. In cross-sectional or aggregated structures, the primary risk arises from the inclusion of predictors that implicitly encode future information, leading to temporal leakage. In longitudinal data structures, improper handling and validation of repeated observations from the same individuals can induce structural leakage, commonly referred to as identity leakage [25–27]. For this reason, rigorous treatment of panel data structures is essential for valid model evaluation [28–31]. These forms of leakage are directly linked to violations of temporal and structural consistency in prediction design and therefore directly affect the interpretability of predictive performance.

The consequences of data leakage extend beyond technical considerations and directly affect the interpretability and practical use of predictive models. An overly optimistic predictive model may guide interventions toward students who do not belong to the actual risk group, while those genuinely at risk remain insufficiently supported. Such misalignment can result in the misallocation of institutional resources and may erode trust in predictive systems [32]. The methodological soundness of predictive systems therefore directly influences the quality of educational practice and the reliability of data-informed interventions [33]. Ensuring proper leakage control is thus integral to the validity of predictive modeling in decision-support contexts.

Predictive models differ substantially in their temporal orientation, particularly with respect to whether they support forward-looking decisions. Some approaches focus on estimating performance or task success within an ongoing learning situation [34]. While such models are predictive in a statistical sense, they do not necessarily correspond to a temporally grounded decision context. Dropout in higher education provides a clear example of a time-indexed phenomenon, where prediction inherently concerns a future risk state. As a result, such phenomena require explicit temporal formalization in predictive modeling.

The aim of this study is to address this methodological gap by formulating a diagnostic framework for temporally consistent prediction in longitudinal settings, demonstrated in the context of higher education dropout. The framework systematizes key requirements that have appeared in the literature but are frequently addressed only implicitly, organizing them into an explicitly time-formalized structure. It explicitly articulates the core methodological conditions of temporally

 doi:10.20944/preprints202605.0186.v1

grounded predictive validity along four interrelated dimensions: (i) explicit specification of the prediction time point, (ii) temporal restriction of the information set, (iii) consistent delineation of the at-risk population, and (iv) validation procedures that ensure temporal and individual-level consistency. Within this perspective, data leakage becomes interpretable as a structural violation of temporal formalization and validation integrity.

Predictive modeling involves multiple interrelated layers, including ethical considerations, fairness, interpretability, and technical implementation. The present study focuses specifically on methodological foundations, with particular emphasis on the control of data leakage. The four articulated dimensions make explicit and structurally integrate requirements that have appeared in diverse forms within the literature. Together, they constitute a time-consistent prediction framework (referred to as TCDP in the dropout prediction context). Clarifying these foundational conditions provides a more stable basis for the evaluation and interpretation of predictive models [35].

The empirical focus of this study is primarily on institutional dropout models constructed from student-level administrative data. However, the proposed framework is not restricted to this specific data source or application context. The principles of temporal formalization, information availability, and validation consistency apply to any predictive task that aims to support a future decision context, including those in higher education. This includes, for example, the prediction of successful course completion when used to inform forward-looking interventions.

This study argues that data leakage in longitudinal prediction should not be understood merely as a technical flaw, but as a structural consequence of improperly specified prediction tasks. Building on this perspective, the study presents a focused literature review and a deliberately simplified empirical demonstration based on student data from a large public university. The empirical component illustrates how a commonly applied yet methodologically flawed validation strategy induces data leakage and distorts the predictive validity of reported performance metrics.

2. Conceptual and Methodological Framework

The proposed framework (referred to as TCDP in the dropout prediction context) specifies the minimal temporal and structural conditions required for valid predictive modeling in longitudinal settings.

The interpretability of predictive modeling is determined by the temporal context in which estimation occurs, the information available at the time of prediction, and the population to which the predictive claim applies [4,10,26]. When these conditions are not explicitly specified, the evaluation of model performance becomes inherently ambiguous, and the distinction between temporally valid forecasting and retrospective pattern discrimination becomes unclear [14,17,33].

This section conceptualizes prediction as a claim tied to a temporally situated decision context, where validity depends on the consistency of temporal and informational assumptions. Accordingly, we identify and systematize methodological components that determine under what conditions a predictive model achieves temporal validity and how violations of these conditions generate structural distortions, including manifestations of data leakage [11,13,19]. This perspective applies to any time-dependent predictive modeling task in which decisions are tied to a specific temporal context.

The prediction cutoff denotes the time point that separates the observable past from the future state to be predicted [10]. Every predictive claim is anchored in a specific decision context; without explicit temporal anchoring, the temporal reference of model outputs is undefined. By fixing this cutoff, the model establishes the point from which the outcome is defined as future-oriented, thereby specifying the temporal origin of the prediction.

2.1. Prediction as a Time-Formalised Task

(i) Explicit Specification of the Prediction Cutoff ($\boldsymbol{t_c}$)

 doi:10.20944/preprints202605.0186.v1

The information set comprises all variables and predictors that have been observed and recorded up to the specified cutoff time [15,36]. A predictive model is restricted to this set. Temporal restriction of the information set ensures that model inputs correspond to the actual informational state of the decision context at the time of prediction.

(ii) Temporal Restriction of the Information Set ($I(t_c)$)

The information set comprises all variables and predictors that have been observed and recorded up to the specified cutoff time [15,36]. A predictive model is restricted to this set. Temporal restriction of the information set ensures that model inputs correspond to the actual informational state of the decision context at the time of prediction.

(iii) Specification of the Risk Set (R(t_c))

The risk set consists of individuals who are active at the cutoff time and for whom the outcome of interest has not yet occurred [37]. Predictive claims are meaningful only within this population, as it is here that genuine uncertainty about the future outcome persists. By defining the risk set, the model specifies the domain within which the prediction is valid and makes explicit which individuals are subject to forward-looking risk estimation.

The three components outlined above jointly define a formal specification of a given decision configuration. Within this structure, a predictive claim is expressed as follows:

$$P(Y_{future} \mid I(t_c), R(t_c)), \tag{1}$$

where the outcome of interest is defined relative to the specified cutoff time and evaluated as a future state with respect to this temporal reference. This formulation makes explicit how prediction depends on temporally bounded information and a well-defined population.

(iv) Temporal and Cohort-Level Consistency (Temporal Hierarchy)

Temporal hierarchy defines the organizing principle of training and evaluation with respect to time. Its purpose is to preserve the temporal order of the data throughout the entire modeling process. Accordingly, training data must precede test data chronologically, observations linked to the same individual must not overlap across training and test sets, and information from later calendar periods must not be incorporated into earlier decision configurations [25,26]. This condition ensures that reported performance metrics reflect a temporally consistent estimation procedure. Under such validation, performance estimates correspond to the intended decision configuration rather than to temporal or individual-level dependencies embedded in the data structure, thereby reflecting a true forward-looking prediction scenario.

The first three components (i–iii) jointly define the formal structure of a decision configuration. The temporal hierarchy (iv) ensures the temporal consistency of its empirical estimation during model training and evaluation. The validity of a predictive model is determined by the alignment between the defined decision configuration and its empirical implementation. When this alignment is compromised, performance metrics remain computable, yet their interpretive scope no longer corresponds to the intended decision context. The formal definitions of the components are summarized in Table 1, and their structural relationships are illustrated in Figure 1.

Table 1. Formal Definitions of Temporally Consistent Prediction.

Component	Designation	Definition / Requirement
(i)	Cutoff (t_c)	The fixed time point at which prediction is issued.
(ii)	Information Set ($\mathcal{I}_{t_c}$)	Only variables observed at or before t_c are used ($t \leq t_c$).
(iii)	Risk Set ($\mathcal{R}_{t_c}$)	Individuals active at $t_c$for whom the outcome has not yet occurred

 doi:10.20944/preprints202605.0186.v1

(iv)	Temporal Hierarchy ($T_{train} < T_{test}$)	Preservation of temporal order during model training and evaluation: training data must precede test data chronologically; observations from the same individual must not overlap across sets; and information from later calendar periods must not be incorporated into earlier decision configurations.

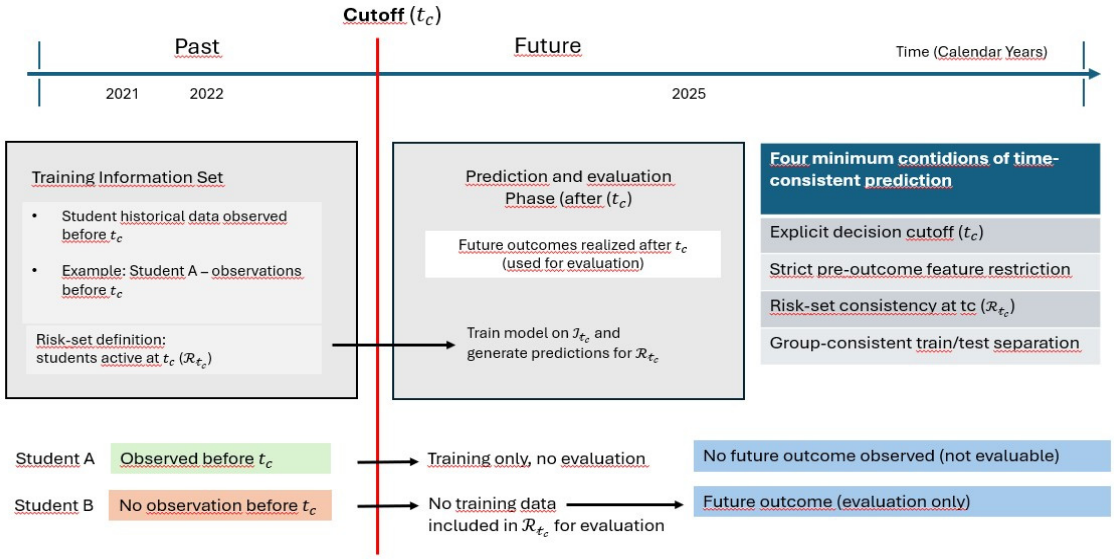

Figure 1. Time-consistent data splitting framework. The vertical red line marks the explicit decision cutoff ($\boldsymbol{t_c}$), separating past observations from future outcomes. Model training is restricted to historical data observed before $\boldsymbol{t_c}$ (training information set, $\boldsymbol{\mathcal{I}_{t_c}}$), while predictions are generated for a well-defined risk set consisting of students active at $\boldsymbol{t_c}$ ($\boldsymbol{\mathcal{R}_{t_c}}$). Model evaluation relies exclusively on outcomes realized after $\boldsymbol{t_c}$. The student-level illustration highlights that training and evaluation populations need not overlap, ensuring group-consistent separation and preventing temporal and identity-based data leakage.

The four methodological conditions summarized above jointly define the minimal requirements for temporally consistent prediction in longitudinal settings. Figure 1 illustrates their integrated operation within a unified prediction process, showing how each component relates to the decision cutoff. The figure presents a consolidated view of the fixed prediction time point, the training information set, the target risk population, and the temporal horizon of evaluation. This integrated representation clarifies how adherence to these conditions constrains data usage and structurally limits the emergence of different forms of data leakage.

2.2. Consequences of Violating the Core Conditions

The four core conditions defined in the previous subsection establish the temporal and conceptual boundaries of predictive validity. Violations of these conditions generate more than technical distortion; they fundamentally alter the epistemic status of the predictive claim. Under such circumstances, model outputs no longer correspond to forward-looking risk estimation within a defined decision context, but instead reflect structural separability embedded in the data. The following section summarizes the interpretive consequences associated with violations of each condition.

2.2.1. Absence of an Explicit Cutoff

When the prediction time point is not clearly specified, model outputs lack anchoring in a well-defined decision context. Under such circumstances, the predictive claim ceases to function as a

 doi:10.20944/preprints202605.0186.v1

forward-looking estimate and instead reduces to retrospective outcome discrimination. In this configuration, the model identifies patterns associated with an already realized end state, while the temporal reference required for decision-oriented prediction remains undefined.

2.2.2. Inconsistency of the Information Set

Information set inconsistency arises when predictors include variables that are not available at the time of prediction. Performance metrics then reflect the informational contribution of data that become observable only after the defined cutoff. Consequently, the predictive interpretation of the results is altered, as the the model extends beyond the temporally constrained information boundary of the decision context.

2.2.3. Violation of Risk-Set Consistency

Risk-set inconsistency occurs when the prediction population is not restricted to individuals who are active at the cutoff time and for whom the outcome has not yet occurred. Under such conditions, the model includes cases whose status is already determined. Reported performance reflects structural separability within the dataset rather than uncertainty about a future outcome. As a result, the validity domain of the prediction no longer aligns with the actual decision context.

2.2.4. Violation of Temporal Hierarchy

Disregarding temporal order during model training and evaluation disrupts the intended temporal structure of the estimation process. In longitudinal datasets, this configuration enables the model to re-identify stable individual- or system-level patterns across training and test sets. Reported performance therefore reflects dependencies embedded in the data structure rather than generalizable forward-looking prediction. Consequently, the apparent predictive accuracy overstates the model's capacity to operate under temporally consistent deployment conditions.

Taken together, the four structural control conditions define the interrelated dimensions of predictive validity: explicit specification of the decision time point, temporal restriction of the information set, delineation of the at-risk population, and chronological consistency of empirical estimation. When any of these conditions is compromised, performance metrics remain computationally obtainable, yet they no longer correspond to a clearly defined decision configuration. In such cases, reported predictive performance reflects structural or temporal inconsistencies embedded in the data, thereby limiting its alignment with the intended decision context. In this sense, predictive validity is not a property of the model alone, but of the consistency between problem specification and evaluation design.

2.3. Data Leakage as Structural Exposure

Violations of the structural control conditions defined above manifest as distinct empirical distortions in predictive modeling. Among these, data leakage represents a specific form of structural exposure that directly shapes reported performance metrics. Leakage occurs when model training or validation incorporates information that would not be available within the defined decision configuration, or when separation between individuals and across temporal periods is compromised.

In the literature, data leakage is typically described in technical forms, such as target leakage, preprocessing leakage, and feature selection leakage. The present framework reinterprets these manifestations as consequences of structural condition violations rather than treating them solely as isolated technical categories. In longitudinal data contexts, particularly in higher education, two forms are especially salient: temporal leakage and identity-based leakage.

2.3.1. Temporal Leakage

Temporal leakage arises when the relationship between predictors and the outcome violates the temporal constraints of the decision configuration. This occurs when the model incorporates

 doi:10.20944/preprints202605.0186.v1

information that was not available at the specified cutoff time, or when temporal order is not preserved during model training and evaluation.

Individual-level temporal leakage occurs when the model incorporates variables derived from post-cutoff observations, such as indicators aggregated over the full observation period or performance data from later semesters. In such configurations, the informational scope extends beyond the defined decision time point. This directly violates the temporal restriction of the information set and alters the interpretive basis of the predictive claim. Consider a model that predicts dropouts at the end of the second semester (t_c). If the predictor set includes cumulative GPA calculated over the entire study period, the model incorporates information that becomes available only after t_c. Although cumulative GPA is statistically informative, it encodes post-cutoff academic performance and therefore violates the temporal restriction of the information set. Reported performance metrics in this configuration reflect future academic outcomes rather than exclusively the informational state available at t_c.

Cohort-level temporal leakage occurs when training and test sets are constructed without chronological separation across successive calendar cohorts. In such configurations, the model exploits patterns arising from institutional, curricular, or regulatory changes that did not exist at the time of the defined decision point. This phenomenon is directly related to violations of temporal hierarchy, as the chronological consistency of empirical estimation is compromised. Consider a model predicting first-year dropout at the end of the second semester (t_c) using data from multiple academic cohorts. Consider a substantial curriculum reform introduced in 2021 that altered course structure and assessment rules. If students from both pre-reform and post-reform cohorts are randomly split across training and test sets, the model implicitly learn patterns associated with the reform itself. Reported performance reflects the model's ability to recognize cohort-specific structural differences rather than its capacity to generalize within a temporally consistent decision configuration. A similar issue arises when predicting dropout in the 2018 cohort while the training set contains observations from later cohorts (e.g., 2019–2021) that experienced different institutional environments. Even without explicit future variables, the absence of chronological separation enables the model to internalize structural patterns that postdate the defined decision horizon.

Temporal leakage is therefore typically associated with violations of the prediction cutoff, the temporal restriction of the information set, and the chronological integrity of model implementation.

2.3.2. Identity Leakage

Identity leakage occurs when multiple temporal observations of the same individual are simultaneously present in both the training and the test sets. In panel-structured datasets, this situation typically results from an inadequately specified validation strategy and reflects the absence of individual-level separation during data splitting.

Under such a configuration, model performance derives from the implicit recognition of individual-specific characteristics that remain stable over time. The reported metrics therefore reflect the detection of recurring patterns within the data structure, constraining generalizability to future cohorts or previously unseen individuals.

Identity leakage is associated with a violation of temporal hierarchy, as the separation of individuals and their observations is not properly enforced during training and evaluation. It also affects the scope of predictive validity when the boundaries between the decision population and the validation population become blurred.

For example, in a student dropout prediction context, identity leakage may occur when semester-level records of the same student are randomly split across training and test sets. In such a configuration, the model encounters earlier academic performance indicators of a student during training and later observations of the same individual during evaluation. The reported performance therefore reflects the model's ability to recognize student-specific patterns rather than its capacity to generalize to new students or future cohorts.

 doi:10.20944/preprints202605.0186.v1

2.3.3. The Impact of Data Leakage on the Interpretation of Predictive Performance

In models affected by data leakage or risk-set inconsistency, reported performance metrics—such as AUC or F1-score—are biased upward relative to true predictive performance. This statistical optimism [14] often manifests as a moderate and seemingly plausible improvement in performance, complicating the detection of underlying methodological flaws [18]. Consequently, performance metrics cannot be interpreted as direct indicators of predictive validity in isolation; their meaning is inseparable from the explicit temporal formalization of the predictive task [19].

The following section illustrates the empirical relevance of the proposed methodological framework by by examining the empirical consequences of violating the proposed methodological conditions. This shift highlights that performance evaluation is inseparable from problem specification in longitudinal predictive modeling.

3. Methodological Patterns in Dropout Prediction: A Literature-Based Analysis

This section examines recurring methodological constructions in recent empirical studies on higher education dropout prediction. The analysis does not aim to provide an exhaustive survey of the literature; rather, it focuses on identifying recurring model-building patterns in current publication practices. The analysis is structured along four focal dimensions: the temporal formalization of prediction, the configuration of the information set, the definition of the risk set, and the validation hierarchy. These dimensions correspond directly to the structural conditions defined in the proposed framework.

3.1. Absence of an Explicit Prediction Cutoff

In a substantial proportion of the examined studies, the temporal reference point of prediction is not explicitly formalized. The target variable typically reflects the student's final status (graduation or dropout), while predictor variables incorporate information accumulated across the entire study trajectory [2,3,5,6,8]. Methodological descriptions frequently omit a clearly defined decision point to which the prediction is anchored. As a result, the reported performance metrics correspond to a retrospective classification framework rather than a forward-looking prediction task. The sample also includes studies that define multiple prediction time points [23] or apply explicitly time-based modeling logic [7,20]. However, these approaches remain a minority within the overall corpus. This pattern suggests that prediction is often implicitly treated as a post hoc classification task rather than a temporally grounded decision problem.

3.2. Predictive Framing and Retrospective Implementation

A substantial proportion of the reviewed publications emphasize early identification and timely intervention, while their technical implementation operates within a retrospective classification framework. The models rely on data from completed study trajectories to predict final status and perform ex post grouping of the full cohort [2,3,6]. This configuration constitutes a statistically valid classification task; however, the temporal formalization of the decision context remains unspecified. The prediction is not anchored in a clearly defined time-bound state but instead relies on retrospective processing of the full dataset. As a result, the apparent predictive capability reflects retrospective separability rather than actionable forward-looking prediction.

3.3. Information Set Configuration and Structural Inconsistency

The absence of an explicitly defined cutoff is frequently accompanied by ambiguity regarding the temporal boundaries of the information set. Numerous studies employ aggregated performance indicators (e.g., CGPA, accumulated credits, course results) without explicitly specifying the time point up to which these variables were available at the moment of prediction [3,5]. By contrast, time-anchored modeling approaches define predictors relative to a specified time point T, ensuring that only information available up to that point is included in the model [7,20]. This configuration clearly

 doi:10.20944/preprints202605.0186.v1

separates the prediction horizon from future information, thereby preserving the temporal validity of the predictive task.

3.4. Risk-Set Definition and Population Consistency

Many of the reviewed studies perform retrospective classification of the entire cohort without explicitly defining the population at risk at a given time point. As a result, the model does not generate predictions for students who are active at a specific time point; instead, it processes the full dataset ex post [2,6]. Explicitly risk-set–consistent implementations appear only in a limited number of studies [7,20], where the selection of active students and the construction of time-specific data matrices are clearly documented, enabling prediction to be aligned with a well-defined population at risk. In the absence of such specification, predictive performance reflects separability across the full cohort rather than risk estimation within an active population.

3.5. Validation Strategies and Temporal Hierarchy

In current validation practice, random train–test splits and k-fold cross-validation dominate [2,3,6]. These configurations do not temporally separate cohorts and therefore do not provide out-of-time evaluation. Cohort-based separation appears less frequently [5], as does explicitly time-hierarchical splitting [7,20]. Such approaches are more closely aligned with temporally interpretable predictive assessment.

Table 2 organizes the reviewed studies along the four dimensions of the proposed framework. The resulting matrix indicates that the simultaneous and explicit formalization of all four dimensions remains uncommon in current publication practices, while retrospective, aggregated, and temporally unspecified modeling configurations continue to dominate. This pattern reinforces the need for explicit temporal formalization in predictive modeling to ensure meaningful evaluation and interpretation.

Table 2. Methodological Configurations of Dropout Prediction Models Across the Four Structural Dimensions.

Study (Year)	Cutoff	Information Set	Risk-set	Temporal Hierarchy
Cho (2023)	✕	✕	✕	✕
Hassan (2024)	✕	✕	✕	✕
Villar (2024)	✕	✕	✕	✕
Okoye (2024)	✕	✕	✕	✕
Song (2023)	✕	✕	✕	✕
Arthana (2024)	✕	✕	✕	✕
Bouihi (2024)	✕	✕	✕	◐
Kabáthová (2021)	◐	✓	◐	✕
Goren (2024)	✓	✓	◐	◐
Vaarma (2024)	✓	✓	✓	✓
Barros (2023)	✓	✓	✓	✓

Notation: ✓ = explicitly formalized. ◐ = partially formalized refers to cases where temporal aspects are acknowledged but not consistently enforced, ✕ = not formalized.

Overall, the literature review reveals two distinct modeling orientations. The majority of publications approach dropout prediction as a classical classification task, focusing on the retrospective categorization of students' final status. Within this framework, predictor variables and validation procedures are are not explicitly anchored to a temporally defined decision context. Alongside this dominant orientation, a narrower but methodologically more reflective line of research has emerged. These studies explicitly link prediction to a specific time point, a temporally bounded information set, and cohort-based validation strategies.

The integrated application of the four structural dimensions cannot yet be regarded as common practice within the reviewed sample. The explicit specification of a cutoff, a temporally bounded

 doi:10.20944/preprints202605.0186.v1

information set, risk-set–consistent population handling, and time-hierarchical validation appear jointly in only a limited number of studies. In most models, predictive performance metrics are interpreted within configurations that do not distinguish forward-looking prediction from the retrospective reconstruction of complete study trajectories.

This pattern indicates that, beyond optimizing technical performance, explicit temporal and structural formalization of prediction design is required. An integrated framework that jointly specifies the cutoff, the information set, the risk set, and the validation hierarchy can ensure that reported results are interpretable not only in statistical terms but also within a defined decision context.

4. Empirical Demonstration: Identity Leakage in Longitudinal Data with Static Predictors

This section presents a controlled empirical demonstration of identity leakage as a structural consequence of improper validation in longitudinal prediction tasks. The experiment illustrates how improper train–test splitting in longitudinal data structures distorts predictive performance metrics. The model construction is intentionally simple to ensure that any observed performance differences can be directly attributed to the examined leakage mechanism. This simplification is a deliberate methodological choice: incorporating more complex predictors or model architectures would obscure the structural dynamics of identity leakage.

4.1. Experimental Design and Data Structure

The demonstration is based on a longitudinal administrative dataset collected over multiple academic years at a Hungarian higher education institution (N = 1,851 students). The data were processed in pseudonymized form. The use of real institutional records ensured a realistic modeling context and preserved naturally occurring relationships among predictors.

The dataset covered six consecutive semesters. A separate predictive model was fitted for each semester ($t \in \{1,\ldots,6\}$), allowing us to examine how model performance evolved over time.

The data representation followed a panel structure: each row corresponds to a (StudentID, Semester) pair, meaning that multiple observations were available for each student. For each semester-specific model, only records available up to that semester were included (Semester ≤ t). Consequently, models fitted for later semesters are trained on larger datasets, as they incorporated records from earlier as well as newly observed semesters.

To isolate the structural mechanism of identity leakage, only static predictors were used—student characteristics that do not change over the course of study (e.g., gender, place of residence, age, year of birth). This design ensures that the demonstrated effect is not dataset-specific and can be reproduced in other longitudinal datasets with similar structural properties. The StudentID variable itself was excluded from the predictor set. As a result, the informational content of the input variables remains constant across semester-specific models; differences in model performance are therefore attributable solely to the increasing number of observations.

To examine the identity leakage mechanism, train–test splitting was intentionally performed at the record level using random allocation (70/30). The split ignored student-level grouping, allowing different semester observations of the same student to appear simultaneously in both training and test sets. As the number of records increased across semesters, the probability that a student's observations were distributed across both sets also increased. This configuration intentionally violates the temporal hierarchy condition in order to expose the effect of identity leakage.

To assess algorithm-specific sensitivity to the leakage mechanism, two classifiers with different representational capacities were applied: a linear logistic regression model and a nonlinear Gradient Boosting model. The objective is not to optimize predictive performance, but to examine how identity leakage affects the reported performance of algorithms with differing expressive power. Model performance was evaluated using row-level ROC-AUC, capturing predictive ranking across

 doi:10.20944/preprints202605.0186.v1

observations, including those belonging to the same individual, thereby allowing identity leakage to influence the evaluation.

All models were implemented using the scikit-learn library with default parameter settings. No extensive hyperparameter tuning was performed, as the purpose of the experiment was to isolate the structural effect of identity leakage rather than to optimize predictive performance. A fixed random seed was used to ensure reproducibility.

4.2. Results and Analysis of Performance Inflation

The results of the identity-leaky configuration across different cutoff points are summarized in Table 3. Student-level overlap between the training and test increases monotonically as the cutoff advances, approaching full overlap at later stages: while no overlap is observed at the first cutoff, the proportion of shared students progressively approaches full overlap at later cutoffs (Figure 2). This pattern follows directly from the longitudinal panel structure combined with record-level validation.

Table 3. Identity Leakage Magnitude and Predictive Performance Across Cutoff Points Under the Identity-Leaky Validation Configuration.

Cutoff	Student-level Overlap ratio	AUC Logistic Regression	AUC Gradient Boosting
1	0.00	0.62	0.66
2	0.68	0.63	0.72
3	0.92	0.64	0.76
4	0.97	0.64	0.77
5	0.99	0.65	0.77
6	1.00	0.65	0.78

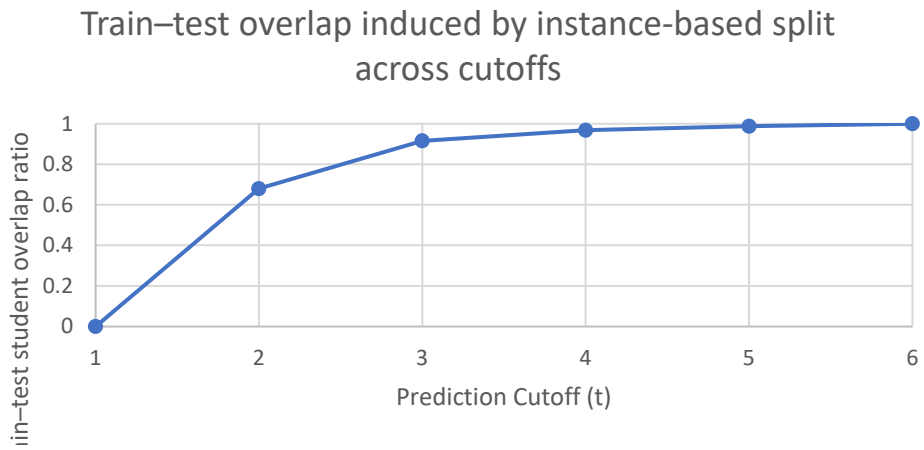

Figure 2. Growth of Student-Level Train–Test Overlap Across Successive Prediction Cutoffs.

The ROC-AUC performance metric exhibits distinct trajectories across the two applied algorithms. While the logistic regression model maintains relatively stable ROC-AUC values across cutoff points, the performance of the Gradient Boosting model increases progressively over time (Figure 3). This divergence indicates that the leakage-induced signal is more effectively exploited by nonlinear ensemble methods than by linear models with limited representational flexibility.

 doi:10.20944/preprints202605.0186.v1

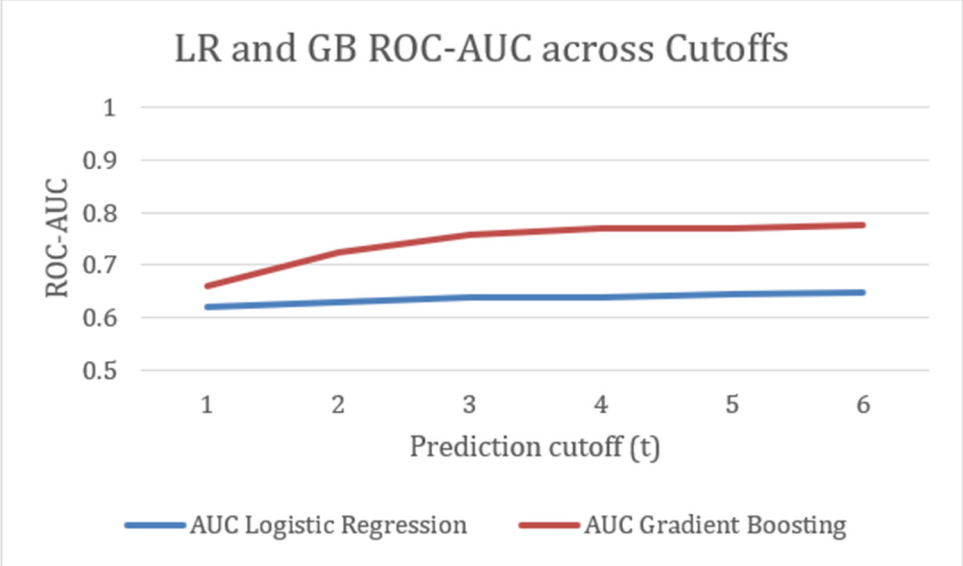

Figure 3. ROC-AUC of Logistic Regression and Gradient Boosting under identity leakage. The divergence illustrates algorithm-specific sensitivity to identity leakage.

The logistic regression model captures only a limited portion of the leakage-related signal, whereas the greater flexibility of Gradient Boosting enables the model to detect student-specific regularities ("identity signatures") that arise from the impermissible overlap between training and test sets. In this configuration, the ensemble model is able to align closely with the leaked signal, resulting in artificially inflated performance estimates.

Although the demonstration is based on data from a single institution, its purpose is not to derive institution-specific conclusions but to isolate a general methodological mechanism. The findings indicate that the interaction between longitudinal data structures and improper validation strategies generates structural bias independent of predictor content or institutional context. Identity leakage arises systematically when validation procedures fail to enforce strict student-level separation, enabling models to exploit leakage-induced signals. This mechanism highlights how model capacity interacts with such signals. Overall, the results demonstrate that evaluation outcomes in longitudinal models are highly sensitive to validation design and may reflect structural artifacts rather than predictive capability. This reinforces the need for structurally consistent validation to ensure meaningful interpretation of predictive performance.

5. Conclusion

The review of longitudinal predictive modeling practices, particularly in higher education, indicates that in a substantial proportion of dropout prediction studies, temporal handling, data representation, validation design, and cohort specification are not articulated within a unified and explicitly defined framework. In many cases, it remains unclear which temporal information underpins the prediction, which student population the model targets, and under what methodological conditions the reported performance metrics can be interpreted. This ambiguity constrains both cross-study comparability and the decision-support value of predictive claims.

The present study conceptualizes predictive modeling as a temporally formalized decision task and identifies four interdependent design conditions that jointly support conceptual and methodological coherence in longitudinal settings: explicit specification of a prediction cutoff, consistent definition of the risk set, student-level separation in training and testing, and coherent treatment of temporal cohorts. Together, these conditions establish a minimal interpretative standard for assessing the applicability of predictive performance metrics in decision-support contexts.

 doi:10.20944/preprints202605.0186.v1

Within this framework, the empirical analysis illustrates distinct data leakage mechanisms, with particular attention to temporal and identity-based leakage. The controlled demonstration shows that even in simple modeling settings, substantial performance gains may arise from identity leakage rather than genuine predictive signal. In such configurations, elevated AUC values reflect the magnitude of leakage embedded in the data structure rather than predictive capacity, indicating that increasing performance may be driven by leakage-induced signal rather than improved predictive validity.

The study establishes a diagnostic and interpretative framework for evaluating predictive models under temporal constraints. The empirical demonstration operationalizes this framework and shows how decisions related to temporal anchoring, cohort definition, risk-set specification, and validation strategy directly shape the interpretability of performance metrics. The primary contribution lies not in the generalizability of specific empirical results, but in clarifying the structural bias mechanisms that arise from data representation and validation design. Overall, the findings demonstrate that predictive performance cannot be interpreted independently of the temporal and structural definition of the prediction task, underscoring the need for temporally consistent validation in any domain where predictive models support forward-looking decisions.

Future work may extend the empirical validation of the proposed framework to publicly available datasets such as OULAD to further illustrate its applicability in standard benchmarking environments.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

References

1. Bognár, L.; Fauszt, T. Factors and conditions that affect the goodness of machine learning models for predicting the success of learning. *Computers & Education: Artificial Intelligence* 2022, 3, 100100. https://doi.org/10.1016/j.caeai.2022.100100.
2. Cho, C.-H.; Yu, Y.-W.; Kim, H.-G. A study on dropout prediction for university students using machine learning. *Applied Sciences* 2023, 13, 12004. https://doi.org/10.3390/app132112004.
3. Okoye, K.; Nganji, J.T.; Escamilla, J.; Hosseini, S. Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education. *Computers & Education: Artificial Intelligence* 2024, 6, 100205. https://doi.org/10.1016/j.caeai.2024.100205.
4. Rabelo A, Rodrigues MW, Nobre C, Isotani S, Zárate L (2024), "Educational data mining and learning analytics: a review of educational management in e-learning". *Information Discovery and Delivery*, Vol. 52 No. 2 pp. 149–163, doi: https://doi.org/10.1108/IDD-10-2022-0099
5. Bouihi, B.; Bousselham, A.; Aoula, E.; Ennibras, F.; Deraoui, A. Prediction of higher education student dropout based on regularized regression models. *Engineering, Technology & Applied Science Research* 2024, 14, 17811–17815. https://doi.org/10.48084/etasr.8644.
6. Hassan, M.A.; Muse, A.H.; Nadarajah, S. Predicting student dropout rates using supervised machine learning: Insights from the 2022 National Education Accessibility Survey in Somaliland. *Applied Sciences* 2024, 14, 7593. https://doi.org/10.3390/app14177593.
7. Vaarma, M.; Li, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society* 2024, 76, 102474. https://doi.org/10.1016/j.techsoc.2024.102474.
8. Villar, A.; Robledo Velini de Andrade, C. Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence* 2024, 4, 2. https://doi.org/10.1007/s44163-023-00079-z.
9. Ouyang, F., Zheng, L. & Jiao, P. Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Educ Inf Technol* **27**, 7893–7925 (2022). https://doi.org/10.1007/s10639-022-10925-9.

 doi:10.20944/preprints202605.0186.v1

10. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 2012, 6, 1–21. https://doi.org/10.1145/2382577.2382579.
11. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 2023, 4, 100804. https://doi.org/10.1016/j.patter.2023.100804.
12. Wen, J.; Thibeau-Sutre, E.; Diaz-Melo, M.; Samper-González, J.; Routier, A.; Bottani, S.; Colliot, O. Convolutional neural networks for Alzheimer's disease classification on MRI: The leakage conundrum and a framework for design and evaluation. *Nature Communications* 2020, 11, 2441. https://doi.org/10.1117/1.JMI.8.2.024503.
13. Apicella, A.; Isgrò, F.; Prevete, R. Don't push the button! Exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review* 2025, 58, 339. https://doi.org/10.1007/s10462-025-11326-3.
14. Cawley, G.C.; Talbot, N.L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 2010, 11, 2079–2107.
15. IBM. What is data leakage? 2023. Available online: https://www.ibm.com/topics/data-leakage (accessed on XX Month 2026).
16. Sayre, R.; Costello, R.P. Data leakage in health outcomes prediction: A systematic review. *Journal of Biomedical Informatics* 2022, 129, 104044. https://doi.org/10.2196/10969.
17. Hand, D.J. Classifier technology and the illusion of progress. *Statistical Science* 2006, 21, 2–14. https://doi.org/10.1214/088342306000000060.
18. Lipton, Z. C., & Steinhardt, J. (2019). Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, *17*(1), 45-77.
19. A. Abdullah, R. H. Ali, R. Koutaly, T. A. Khan and I. Ahmad, "Enhancing Student Retention: Predictive Machine Learning Models for Identifying and Preventing University Dropout," *2025 International Conference on Innovation in Artificial Intelligence and Internet of Things (AIIT)*, Jeddah, Saudi Arabia, 2025, pp. 1-6, doi: https://doi.org/10.1109/AIIT63112.2025.11082926.
20. Barros, B.M.; do Nascimento, H.A.D.; Guedes, R.; Monsueto, S.E. Evaluating splitting approaches in the context of student dropout prediction. *arXiv* 2023, arXiv:2305.08600. https://doi.org/10.48550/arXiv.2305.08600.
21. Bond, M.; Khosravi, H.; De Laat, M.; Bergdahl, N.; Negrea, V.; Oxley, E.; Pham, P.; Chong, S.W.; Siemens, G. A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education* 2024, 21, 4. https://doi.org/10.1186/s41239-023-00436-z.
22. Dake, D.K.; Buabeng-Andoh, C. Using machine learning techniques to predict learner drop-out rate in higher educational institutions. *Mobile Information Systems* 2022, 2022, 2670562. https://doi.org/10.1155/2022/2670562.
23. Gašević, D.; Dawson, S.; Siemens, G. Let's not forget: Learning analytics are about learning. *Computers & Education* 2015, 82, 64–71. https://doi.org/10.1007/s11528-014-0822-x.
24. Brooks, C.; Thompson, C. Predictive modelling in teaching and learning. In *Handbook of Learning Analytics*; Lang, C., Siemens, G., Wise, A., Gašević, D., Eds.; Society for Learning Analytics Research, 2017; pp. 61–73. https://doi.org/10.18608/hla17.005.
25. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Thuiller, W. Cross-validation strategies for data with temporal, spatial, or hierarchical structure. *Ecography* 2017, 40, 913–929. https://doi.org/10.1111/ecog.02881.
26. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Statistics Surveys* 2010, 4, 40–79. https://doi.org/10.1214/09-SS054.
27. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series forecasting. *Computational Statistics & Data Analysis* 2018, 120, 70–83. https://doi.org/10.1016/j.csda.2017.11.003.
28. Bollen, K.A.; Brand, J.E. A general panel model with random and fixed effects: A structural equations approach. *Social Forces* 2010, 89, 1–34. https://doi.org/10.1353/sof.2010.0072.

 doi:10.20944/preprints202605.0186.v1

29. Cerqua, A.; Letta, M.; Pinto, G. On the (mis)use of machine learning with panel data. *Oxford Bulletin of Economics and Statistics* 2025, 1–13. https://doi.org/10.1111/obes.70019.
30. Shmueli, G. To explain or to predict? *Statistical Science* 2010, 25(3), 289–310. https://doi.org/10.1214/10-STS330.
31. Wooldridge, J.M. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed.; MIT Press, 2010.
32. Knight, S.; Buckingham Shum, S.; Littleton, K. Epistemology, assessment, and learning analytics. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge (LAK '14), 2014; pp. 129–138. https://doi.org/10.1145/2460296.2460312.
33. Hofman, J.M.; Sharma, A.; Watts, D.J. Prediction and explanation in social systems. *Science* 2017, 355, 486–488. https://doi.org/10.1126/science.aal3856.
34. Fu, C.; Fang, Q. Curriculum-aware cognitive diagnosis via graph neural networks. *Information* 2025, 16, 996. https://doi.org/10.3390/info16110996.
35. Shmueli, G. Predictive Analytics in Information Systems Research. *MIS Quarterly* 2010, 35, 553–572.
36. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7.
37. Kleinbaum, D.G.; Kleinbaum, M. *Survival Analysis: A Self-Learning Text*, 3rd ed.; Springer, 2012. https://doi.org/10.1007/978-1-4419-6646-9.